DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER FAKULTAET FUER CHEMIE UND PHARMAZIE
DER LUDWIG-MAXIMILIANS-UNIVERSITAET MUENCHEN

True single-cell proteomics using advanced ion mobility mass spectrometry

Andreas-David Brunner

aus

Krefeld, Deutschland

2021

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsverordnung vom 28. November 2011 von Herrn Prof. Dr. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Martinsried, 01.06.2021, Brunner

Dissertation eingereicht am	04.05.2021
1. Gutachter:	Prof. Dr. Matthias Mann
2. Gutachter:	PD. Dr. Henrik Daub
Mündliche Prüfung am	26.05.2021

Summary

In this thesis, I present the development of a novel mass spectrometry (MS) platform and scan modes in conjunction with a versatile and robust liquid chromatography (LC) platform, which addresses current sensitivity and robustness limitations in MS-based proteomics. I demonstrate how this technology benefits the high-speed and ultra-high sensitivity proteomics studies on a large scale. This culminated in the first of its kind label-free MS-based single-cell proteomics platform and its application to spatial tissue proteomics. I also investigate the vastly underexplored 'dark matter' of the proteome, validating novel microproteins that contribute to human cellular function.

First, we developed a novel trapped ion mobility spectrometry (TIMS) platform for proteomics applications, which multiplies sequencing speed and sensitivity by 'parallel accumulation – serial fragmentation' (PASEF) and applied it to first high-sensitivity and large-scale projects in the biomedical arena. Next, to explore the collisional cross section (CCS) dimension in TIMS, we measured over 1 million peptide CCS values, which enabled us to train a deep learning model for CCS prediction solely based on the linear amino acid sequence. We also translated the principles of TIMS and PASEF to the field of lipidomics, highlighting parallel benefits in terms of throughput and sensitivity.

The core of my PhD is the development of a robust ultra-high sensitivity LC-MS platform for the high-throughput analysis of single-cell proteomes. Improvements in ion transfer efficiency, robust, very low flow LC and a PASEF data independent acquisition scan mode together increased measurement sensitivity by up to 100-fold. We quantified single-cell proteomes to a depth of up to 1,400 proteins per cell. A fundamental result from the comparisons to single-cell RNA sequencing data revealed that single cells have a stable core proteome, whereas the transcriptome is dominated by Poisson noise, emphasizing the need for both complementary technologies.

Building on our achievements with the single-cell proteomics technology, we elucidated the image-guided spatial and cell-type resolved proteome in whole organs and tissues from minute sample amounts. We combined clearing of rodent and human organs, unbiased 3D-imaging, target tissue identification, isolation and MS-based unbiased proteomics to describe early-stage β-amyloid plaque proteome profiles in a disease model of familial Alzheimer's. Automated artificial intelligence driven isolation and pooling of single cells of the same phenotype allowed us to analyze the cell-type resolved proteome of cancer tissues, revealing a remarkable spatial difference in the proteome.

Last, we systematically elucidated pervasive translation of noncanonical human open reading frames combining state-of-the art ribosome profiling, CRISPR screens, imaging and MS-based proteomics. We performed unbiased analysis of small novel proteins and prove their physical existence by LC-MS as HLA peptides, essential interaction partners of protein complexes and cellular function.

Table of contents

Abbreviations

AGC	automatic gain control
CCS	collisional cross section
CID	collision induced dissociation
CITE-seq	cellular indexing of transcriptomes and epitopes by sequencing
DDA	data dependent acquisition
DIA	data independent acquisition
DISCO-MS	3-dimensional imaging of solvent cleared organs coupled to mass spectrometry
DTIMS	drift-tube ion mobility spectrometry
EASI-tag	easily abstractable sulfoxide based isobaric tag
ES	electrospray
FAIMS	field-asymmetric ion mobility spectrometry
HCD	higher energy collisional dissociation
IMS	ion mobility spectrometry
LC	liquid chromatography
LFQ	label-free quantification
MRM	multiple reaction monitoring
MS	mass spectrometer
MS2	tandem mass spectrometry
ORF	open reading frame
PASEF	parallel accumulation – serial fragmentation
ppm	parts per million
PTM	post-translational modification
rf	radiofrequency
SCoPE-MS	single-cell proteomics by mass spectrometry
scRNA-seq	single-cell RNA sequencing
SCP	single-cell proteomics
SRIG	stacked ring ion guide
TIMS	trapped ion mobility spectrometry
TMT	tandem mass tag
TOF	time of flight
T-SCP	true single-cell proteomics

1. Introduction

1. 1. Life is just a more complex binary code

1.1.1. The central dogma of molecular biology

"The course of life processes in an organism show an admirable regularity and order that is unparalleled in inanimate matter. It is regulated by a highly ordered group of atoms, which make up only a tiny fraction of their totality in the cell."

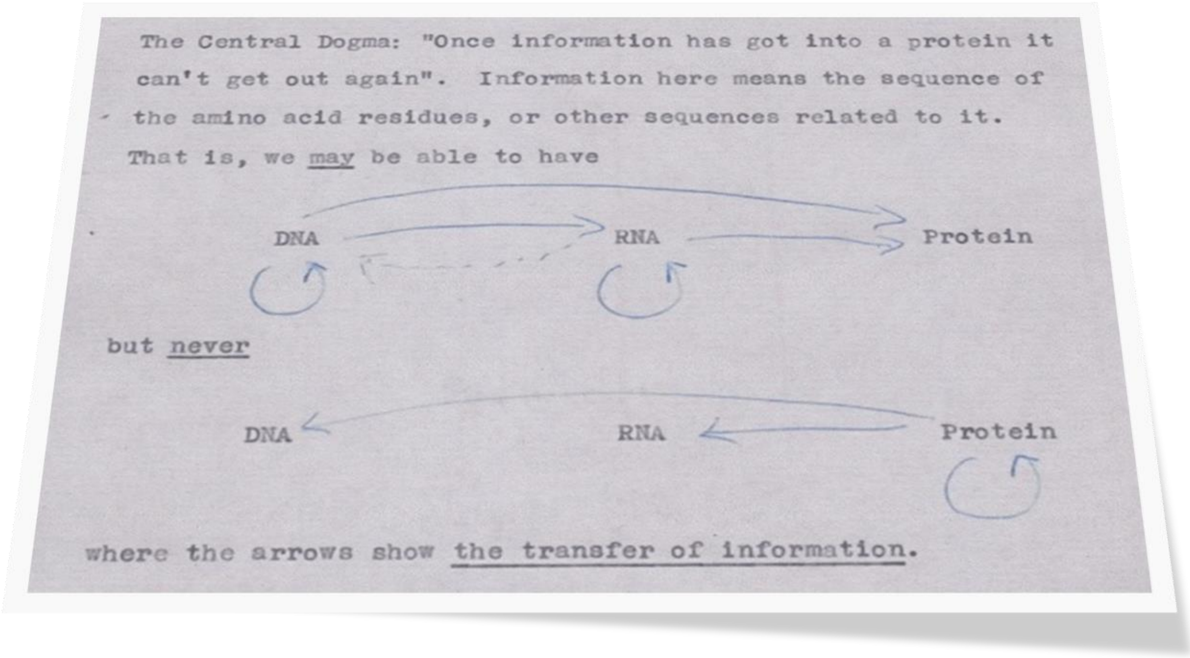
Erwin Schrödinger, *What is life?*, 1944

Understanding the fundamental principles that have enabled life to evolve throughout billions of years has been fascinating to humans for a long time. Only in 1859, Charles Darwin published his work 'On the Origin of Species', introducing the scientific theory that populations and species evolve over the course of generations by means of natural selection, challenging the dominant view of that time that God had created the world in seven days[1]. This theory also suggested that life has most likely evolved from a 'last universal common ancestor', the first life form on earth, and is still subject of studies today[2]. Only seven years later, Gregor Mendel discovered in pea experiments that phenotypic traits like color, height and shape can be inherited by offspring in a dominant or recessive manner and that this is passed on to the next generation. He stated that 'invisible factors' carry the phenotypic information, strengthening and expanding Darwin's theory[3].

Three years later in 1869, Friedrich Miescher isolated by accident a substance from white blood cell nuclei with a very high phosphorous content that furthermore resisted protein digestion. He also proved the existence of this molecular entity in a variety of nuclei isolated from different cells and consequently termed it 'nuclein'. He was also the first person to suggest that this substance could carry hereditary information, but rejected this hypothesis due to lack of experimental evidence[4]. Throughout the next decades it became well understood that the hereditary information, later coined 'genes' (Greek - 'birth') by the Danish botanist Wilhelm Johannsen, carry the information for phenotypic traits in any living organism. Still, it was not clear which substance is the carrier. In 1944, Oswald Avery and colleagues experimentally identified a distinct molecular entity, neither a protein nor a carbohydrate, that turned rather harmless pneumococcus species into pneumonia-inducing bacteria when co-cultivated with the pathogenic species. This groundbreaking work introduced the 'transforming principle' after isolating the yet unknown substance from literally 'twenty gallons of bacteria'. This was also the time, where the new hereditary information carrying entity was identified as a deoxyribonucleic acid[5].

Inspired by the described principle, Erwin Chargaff began to analyze the chemical constitution of DNA. He first discovered a regularity in DNA base distribution, where the number of guanines turned out to be equal to the number of cytosines, while the number of thymidines was equal to the number of adenines, and second, that the relative proportion of these two pairs varies between species[6]. At that time, it was well known that proteins exist in a three-dimensional structure, but it was still elusive what the structure of DNA looks like and how it is packed. Already in 1944, without knowing about the compositional regularity discovery by Erwin Chargaff, Erwin Schrödinger suggested the idea of an 'aperiodic crystal' that contained genetic information[7]. In 1952, Rosalind Franklin set out to elucidate the structure of the DNA experimentally via X-ray crystallography and produced several high-resolution DNA-fiber records hinting toward a helical structure. In 1953, James Watson and Francis Crick were able to solve the outstanding question of the phosphate-, sugar-, and base arrangement to a double-helical DNA-structure[8]. Since it was still not clear which molecular moiety stores the genomic information, Crick and Watson furthermore suggested a possible DNA-copying mechanism, which is necessary for the transfer of genetic information to descendants[9].

Their paper was the starting point for a race to decode the code of life laid down in the DNA sequence. Only three years later, Francis Crick proposed the 'The Central Dogma' of biology, in which he described how he imagines genes to be encoded by DNA, that RNA molecules are the most likely mediators of genetic information based on RNA-virus experiments and that RNA molecules could be turned into proteins[10]. This paper was historic in that it permanently altered how biology was seen, even though it only described the most likely theoretical flow of information at that time (Fig. 1).



The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have

DNA
RNA
Protein

but never

DNA
RNA
Protein

where the arrows show the transfer of information.

Figure 1 | Theory of the 'Central Dogma of Molecular Biology' as initially outlined by Francis Crick in 1956[11].

At that point, it was known that proteins consist of 20 amino acids and that DNA consists of four base pairs. To encode these, one would at least need a sequence of base pair triplets, which would give rise to a total of $4^3 = 64$ codons. A still remaining issue was the missing link of how to get from the DNA code to the protein level and which role RNA plays in this context. In 1960, experiments with bacteriophages, which inject their DNA into bacteria, giving rise to the fast appearance of short-lived small RNA molecules, which closely resemble the sequence of the bacteriophage genome, resolved this long-lasting question, using a simple beta-galactosidase reporter assay. The RNA mediating the information from DNA to protein level was called messenger RNA[12]. In 1965, Marshall Nirenberg succeeded in deciphering the *code of life* in a sequence of experiments, in which he introduced synthetic RNA of different base composition to an *E. coli* lysate capable of protein translation and the step-wise addition of a single radioactive amino acids including non-radioactive counterparts, while removing DNA enzymatically. Analyzing the amino acid composition and combinatorically linking it to the introduced synthetic RNA allowed him to break the 'code of life'[13]. Interestingly, until the introduction of Sanger sequencing for proteins, people did not know about the fact that proteins are defined by a determined amino acid sequence, which he first proved by the sequencing of insulin in the 1950s.

After *The Central Dogma* of molecular biology was experimentally proven and researchers were able to decipher the sequence of proteins and DNA, novel technologies had to be developed to be able to scale up these sequencing endeavors. The first of these techniques for DNA-sequencing was developed by Frederick Sanger in 1977 and termed *DNA sequencing with chain-terminating inhibitors*. This was rapidly followed by whole-genome sequencing of the first laboratory bacteriophage model *φX174* and the subsequent viral proteome prediction based on the deciphered genetic code[14,15].

This was the starting point for many exciting developments and researchers realized that there must be organism-specific blueprints encoded by at least four base pairs - just like a more complex binary code in computer science, giving rise to the functional level of proteins. It also suggested that life could in principle be programmable in the future, just like executable functions in informatics and that the field was poised to create new products, drugs and revolutionize medical diagnostics.

The fast-paced evolving field of *Genomics* turned to sequencing several organisms including *Haemophilus Influenzae*[16], *Mycoplasma genitalium*[17] and *Caenorhabditis elegans*[18] in 1995, *Saccharaomyces cerevisiae*[19] in 1996, *Drosophila melanogaster*[20] in 2000 and *Mus musculus*[21] in 2002. Soon, scientists realized the potential impact of *Genomics* on human health and disease and aimed to generate full sequence maps. This publicly funded effort commenced in 1990 under the leadership of Francis Collins and was called the *Human Genome Project*. In 1998, J. Craig Venter joined this race with a privately held company called *Celera* aiming to commercialize human genome sequencing and even patent readouts of genomic loci, which could be associated with health or disease. In February 2001, both groups published first drafts of the human genome, revealing a size of close to three billion nucleotides, thousands of single-nucleotide polymorphisms, novel aspects of gene and chromosome architecture and the estimate that there are more than 20,000 protein-coding genes to be found in the human genome[22,23].

Furthermore, the development of the polymerase chain reaction technology to amplify target DNA-sequences by Kary Mullis revolutionized the field of DNA research, enabling targeted gene amplification, sequencing, and cloning[24]. It also created the basis for many future sequencing technologies like next generation of sequencing approaches (NGS), which allowed parallel and scalable sequencing at much decreased overall costs - below 1,000 $ by now - leading the way into the clinic[25,26].

The entirety of the originally sought-after missing link between genes and proteins is called the *transcriptome* and represents all actively transcribed genomic loci including regulatory RNA and open reading frames, which will ultimately be translated into long strings of amino acids. Initially, the presence of the transcript was investigated by hybridization and so-called micro-array technologies. This also allowed the inference of single nucleotide polymorphisms (SNPs) in the corresponding gene, which is still state of the art in clinical applications. One major drawback of this technology is that it only probes for mRNA entities with a known sequence and therefore does not capture the entirety of the transcriptome[27]. Later, adapted NGS protocols also allowed the parallel sequencing of transcripts and the representation of the transcriptome as a whole, which enabled deep transcriptome analysis of tissues present within the human body, as one of a plethora of different applications as discussed later[28,29].

Understanding the entire dogma of molecular biology and having the capability to sequence whole genomes and transcriptomes gave rise to many game-changing discoveries and technologies. For example, the availability of genome sequencing and its exploration across the phylogenetic tree not only impacted diagnostics, but also lead to the finding of evolutionary conserved mechanisms resulting in disruptive technologies like CRISPR/Cas9 for genome engineering, which was awarded with the Noble prize in 2020[30–35]. Furthermore, elucidation of the interplay between genome, transcriptome and proteins also enabled researchers to extend the code of life with unnatural amino acids and introduce biorthogonal chemistry, engineer minimal life carrying only the most essential genes and even transplant artificial full genomes between organisms[36–38]. All these technological innovations happening within the last ~60 years have turned organisms into models for systems biologist and bioengineers, who can now program biological function in what is essentially just a more complex binary code.

1.1.2. The proteome and its higher order complexity

The development of technologies to investigate the first two layers of molecular biology, the *genome* and *transcriptome*, were immense breakthroughs that have enabled novel approaches in modern biology and clinical applications. Even though the importance of these omics technologies are undisputed, nearly all cellular processes are executed by proteins. The entirety of all proteins and their complex interplay is summarized as the *proteome*. Depending on the system it spans more than 10^{10} orders of magnitude in abundance distributed across approximately a total of 150 pg protein mass per cell[39,40]. The intrinsic importance of the proteome is also reflected by the fact that currently more than 40 % of laboratory tests in the clinic are protein readouts – in stark contrast, less than 0.5 % target nucleic acids (not taking into account the recently increased demand due to COVID-19)[41]. Antibody-based assays aiming for the quantification of target proteins are the gold-standard for fast and sensitive clinical readout and are widely used in basic research. The *Human protein atlas* project even aimed to raise highly specific antibodies against all known proteins with the goal to qualitatively and quantitatively describe the protein distribution across all human tissues, cells and body fluids. It demonstrated that protein expression levels are vastly different across analyzed matrices, or even specific to them in the sense that the amount of protein present was below the assays detection limit[42–44]. In recent years, many binder-based commercial assays are being developed, which promise scalability across large sample cohorts and the quantification of proteins across the full dynamic range. These include proximity-dependent DNA ligation assays from *Olink* and the aptamer-based technologies from *Somalogic*[45,46]. Briefly, both assays follow the same principle by binding to highly specific epitopes of the target protein, followed by PCR amplification of a specific oligonucleotide linked to the binder coupled to a quantitative readout (single binder for *Somamers* and two independent antibodies for *Olink*s). These assays are widely applied due to their multiplexing possibility, which resulted already in many examples of body fluid proteome expression analysis in clinical settings[47–49].

However, there are several drawbacks of antibody-based assays: first, they infer the protein expression levels indirectly; second, shared epitopes lead to disturbed quantification due to off target effects; third, proteins need to be in their native state to enable proper epitope binding; fourth, the proteome itself is highly dynamic in structure and function as described below and inferring accurate protein quantities from epitope targeting approaches is inherently challenging; fifth, they quantify only their target proteins, revealing only a part of the bigger picture; sixth, the detection of protein isoforms and posttranslational modifications adds another level of proteome complexity, which asks for the generation of even more highly specific antibodies[50].

The proteome by itself is immensely complex and cannot simply be summarized by the bioinformatically inferred ~20,000 human open reading frames (ORFs) as is done with 'black and white' rules like the presence of a translational start and stop codon, the presence of protein-coding exons and a strict cutoff of at least 100 amino acids (the latter under the assumption that smaller amino acid stretches cannot be functional since they are most likely not able to form a stable three-dimensional structure). Instead, studies have shown that there are ~26,000 additional small open reading frames present even in the yeast genome with a length of 2-99 amino acids, which can give rise to potentially functional microproteins[51]. Several technological advances including ribosome profiling, which allows to identify actively translated genomic regions, and ultra-high sensitivity mass spectrometry can experimentally prove their physical existence. Combined with phenotypic readouts and CRISPR-based genome engineering technologies, this enables the systematic investigation of these microproteins, highlighting their essential function in a cellular context[52–54] **(Article 10)**. Especially ribosome profiling turned out to be a very important technique for hunting down small novel ORFs, since it provides high-density ribosome maps enabling back-tracking of the actively translated genome locus, decreasing the total number of putative small ORFs dramatically and enables targeted genome engineering of these positions[55].

Furthermore, splicing events and so-called exon-shuffling increase the complexity of the proteome immensely together with posttranslational modifications including acetylation, ubiquitination, phosphorylation, glycosylation, amidation, methylation, citrullination and targeted proteolytic processing to regulate function [56]. For example, the tumor suppressor p53, a protein with a length of 393 amino acids is estimated to have more than 100 possible sites of modification[57]. Furthermore, PTMs can give rise to structural rearrangements of proteins, change their function in a signaling context and regulate the activity of bioactive peptides, which are highly modified short amino acid stretches, often involved in regulation of our behavior, and processed by a large hierarchical proteolytic network. The single ORF pro-glucagon, for example, is by now known to give rise to more than ten tissue-specific proteolytically processed bioactive peptides[58]. Furthermore, proteins are consistently proteolytically processed by the human leukocyte antigen presenting machinery, giving rise to intracellular proteins being presented as small peptides on the cell surface as class I or II peptides to CD8+ and CD4+ T-cells of the immune system, respectively, as checks to distinguish self from non-self[59,60]. This higher order complexity of the proteome is gaining more and more attention from the scientific community and is discussed under the umbrella of *proteoforms* or even *peptidoforms*. Taking these variations into account leads to proteome complexity estimates comprising several 100,000 proteins, which is still most likely a vast underestimation[61].

To fully capture the diversity of the proteome, one would need a technology that is unbiased in the sense of aiming to measure all proteins without prior knowledge about *proteome* composition, providing direct physical evidence of the proteins' presence. It should also be scalable, robust, quantitatively reproducible, and applicable to all biological matrices.

1.2. Mass spectrometry-based proteomics

The technology of choice with the capability to measure complete and quantitative proteomes as well as its structure and dynamics in an unbiased way is liquid chromatography coupled to mass spectrometry (LC-MS). Since the inception of electrospray ionization in the 1980s for transferring biomolecules from liquid into gas phase in their intact form and accurate charge state deconvolution of biomolecules, the field of LC-MS has been developing at a tremendous speed[62–64]. Mass spectrometers measure mass-to-charge ratios of ions dragged by the gas flow from ambient conditions into the vacuum of the system. This results in an inherently high specificity and the potential of extreme sensitivity – even down to detecting single ions[65,66]. By now, advanced sample preparation and computational workflows in conjunction with cutting-edge mass spectrometers have achieved the acquisition of comprehensive proteome profiles, the complement to the *transcriptomes* and *genomics*[67,68]. MS-based *proteomics* can also be used to reveal dynamic structural changes of proteins in response to metabolite or drug engagement, enabling the analysis of dynamic and functional changes of the proteome and paving the way for novel analytical approaches[69–72]. Since mass spectrometers can prove the physical existence of any ionizable biomolecule species, modern MS-based techniques can be applied to many other biomolecules besides proteins, including those that would elude the analysis capabilities of other techniques. Such samples include *in vivo* derived peptides from body fluids, peptides of the HLA system, but also metabolites, as well as lipids[59,73–75].

1.2.1. Bottom-up proteomics

Proteomics workflows are generally divided into three main pillars, which are sample preparation, LC-MS and data analysis. Depending on the size and processing of the proteomic sample, three approaches are pursued, which are called *top-down*, *middle-down*, and *bottom-up* proteomics workflows[76].
Top-down, as the name already states, aims to analyze proteins in their native constitution (usually in a size range of 10 to 30 kDa), conserving protein isoform information, amino acid sequence variants and all PTMs in their combinatorial arrangements. The main challenges to overcome are chromatographic separation of the proteins due to the complexity of proteoforms, their poor ionization decreasing overall sensitivity, low proteome coverage per sample and throughput. Besides this, the highly complex charge patterns require sophisticated algorithms and ultra-high resolution mass spectrometers to be able to obtain a high-quality isotope pattern for deconvolution[61,77,78]. *Middle-down* represents a fairly new field aiming to analyze rather large proteolytic fragments of proteins with a size between 3 to 10 kDa to obtain very high sequence coverage and also conserving PTM information, but again at the expense of throughput, proteome coverage and the need for novel bioinformatic solutions[76]. In contrast, in *bottom-up* proteomics (Fig. 2), the extracted and solubilized complex protein matrix of any given sample is proteolytically digested with a sequence-specific protease, mostly trypsin, to generate peptides at a median length of about 12 amino acids[79].

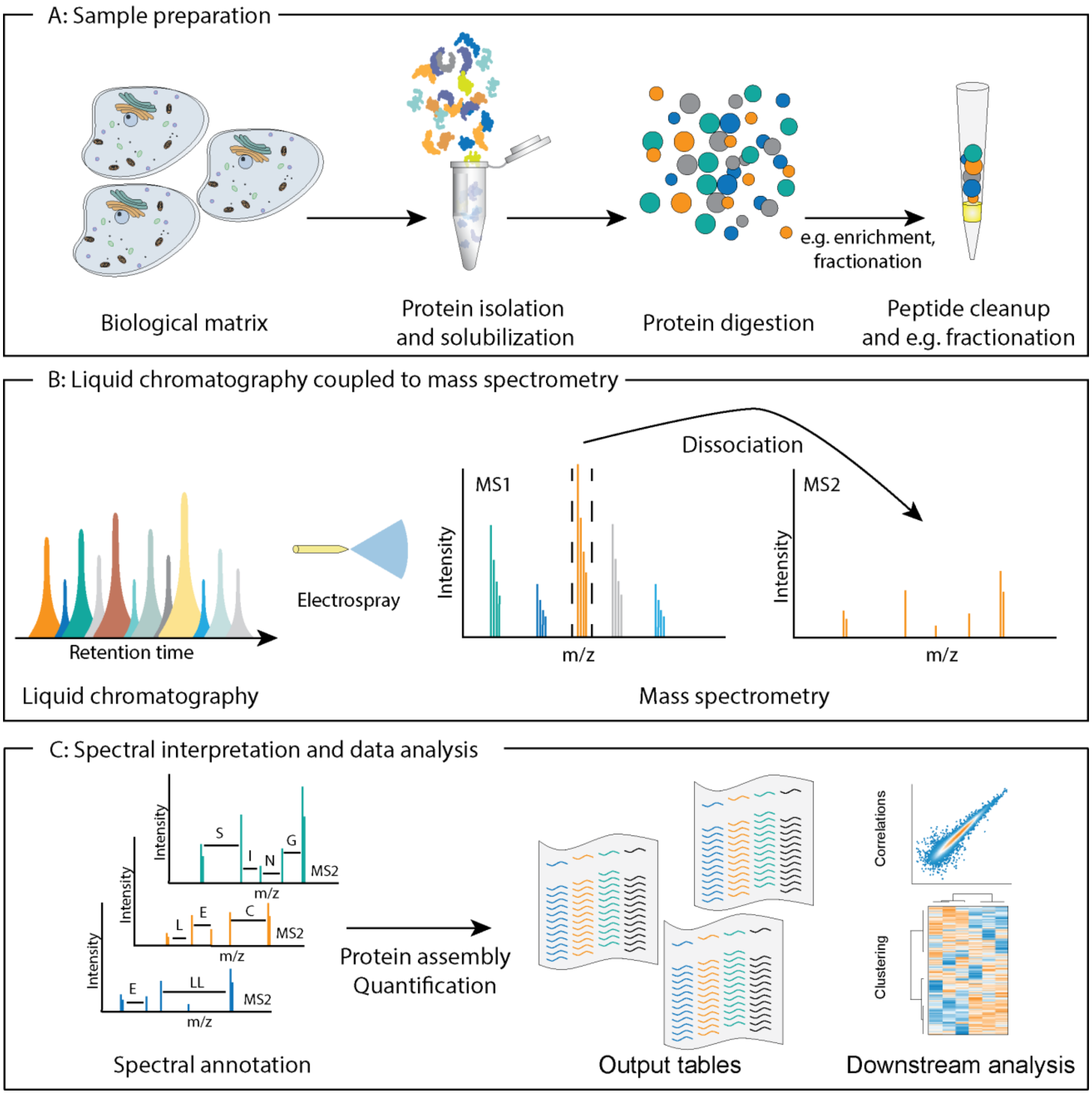


Figure 2 | Bottom-up proteomics. A, Proteins are obtained from a biological matrix, which could be of e.g. cellular origin. Proteins are then isolated and solubilized by chaotropic or detergent based solutions, proteolytically digested into peptides with a defined cleavage site, followed by optional fractionation to decrease proteome complexity per analysis or enriched for PTM and finally purified prior to LC-MS analysis. **B,** The peptide mixture is then separated by HPLC based on hydrophobicity, ionized and transferred into the vacuum of the mass spectrometer via electrospray. The mass spectrometer acquires full scans (MS1) to determine the peptide mass and fragment ion spectra (MS2) of either the TopN most abundant peptides per MS1 scan or from a pre-defined m/z range. **C,** Precursor masses and peptide fragment patterns within MS2 spectra are then compared to *in silico* generated peak lists, followed by spectral annotation, protein assembly, quantification and downstream bioinformatics analysis. (Adapted from Ref.[80])

This results in a very complex peptide mixture, which allows a very high proteome coverage, sensitivity and throughput, but at comprised protein sequence coverage often comprising the analysis of potentially important PTMs and their co-occurrence. To alleviate this issue, fractionation or subset enrichment techniques to decrease peptide mixture complexity can be applied, followed by online liquid chromatography to separate peptides by their hydrophobicity before electrospray ionization and analysis in the mass spectrometer[81–83]. *In silico* proteolytically digested proteins from a reference database and their respective theoretical peptide masses as well as fragment spectra, representing ORFs calculated from species-specific genome sequences, are then compared to the experimental spectra. Finally, these comparisons are scored by sophisticated computational frameworks, which includes reassembling proteins from peptides in a bottom-up-fashion, and quantified to enable downstream bioinformatics analysis[84–86]. Due to its high throughput and unbiased analysis capabilities for discovery proteomics, *bottom-up* proteomics is by far the most widely applied workflow in the proteomics community. It has reached many milestones including a proposed first draft of the human proteome (a compilation of MS2 spectra covering more than 70 % of expected human ORFs), showed its potential for diagnostic readouts, phyloproteomic analysis and large-scale interactome studies[87–91].

1.2.2. Sample preparation

The preparation of samples for MS-based proteomics analysis is crucial and has to be optimized for the sample type to be analyzed, the initial sample starting amount and also for the experimentally addressed question. Many sample preparation protocols have been developed over time, but there are three principles, which virtually all of them follow since more than a decade: (I) protein extraction and solubilization - including reduction of disulfide bonds and alkylation of free cysteins; (II) protein digestion by endoproteases; (III) peptide purification before LC-MS analysis.

1.2.2.1. Protein extraction and solubilization

The first step of sample preparation for LC-MS analysis is protein extraction and solubilization. Different classes of protein solubilizing agents are used in combination with mechanical and thermal treatment to extract proteins from their native environment and to stop residual enzymatic activity within the sample, preventing artificial modification, which could compromise the experimental outcome. Three major classes of protein solubilizing agents are used in proteomics, which are chaotropic agents like urea, thiourea, or guanidinium-chloride (GdmCl), detergent-based agents like sodium-dodecyl-sulfate (SDS) and sodium-deoxycholate (SDC), and organic solvent-based agents like acetonitrile (ACN)[92–95]. All of these agents have their advantages and disadvantages depending on the sample to be analyzed.

For example, SDS is presumably the most efficient protein solubilizing agent disrupting any protein structure, but interferes with tryptic digestion and ionization[96]. It is still used in many harsh tissue protocols and new approaches to remove SDS like Filter-Aided Sample Preparation (FASP), S-trap, Protein Aggregation Capture (PAC) on micro-particles, or the single-pot solid-phase-enhanced sample preparation (SP3) technology prior to proteome digestion have enabled a comeback even in automated and multiplexed applications[92,97–100]. GdmCl based protein extraction appears to be the most efficient chaotropic agent-based approach and can also be combined with heating steps in contrast to urea, which would result in carbamylation adducts of free amines at elevated temperatures[95,101,102]. Even though GdmCl is the chaotropic agent of choice, it is known to influence protein digestion efficiency at high molarities and consequently the sample has to be drastically diluted to keep 'missed cleavage rates' (percentage of peptides with internal cleavage sites) low and reproducible across sample preparations. SDC-based protocols recently found widespread application, because of its mild nature[90,103,104]. SDC interacts with exposed hydrophobic amino acid stretches under elevated temperature, stabilizing the unfolded state even when the sample returns to room temperature. This ensures highest digestion efficiency of the proteome since endoproteases are not structurally influenced. Organic solvent based approaches using ACN start to find their use especially in sample limited applications, since it can easily be removed by evaporation[105]. Furthermore, enzymatic digestion kinetics are elevated in many cases when using ACN concentrations of up to 20 %.

The extraction and protein solubilization step with detergents or chaotropic agents is in many cases performed in conjunction with the reduction of cysteine bridges and alkylation of free cysteins. This fully breaks down higher order protein structures, ensures a highly efficient digestion and downstream peptide identification. There are two commonly used reducing agents in proteomics sample preparation, namely dithiothreitol (DTT) and tris(2-carboxyethyl)phosphine (TCEP)[92,95]. DTT is a very mild reducing agent, works in a basic environment and results in a kinetically favored ring-structure when interacting with disulfide bridges[106]. TCEP, the stronger reducing agent of the two, is gaining more attention in modern proteomics sample preparation procedures since it is more stable, more reactive across a broader pH range and results in an irreversible breakdown of disulfide bridges[107]. Still, in our laboratory, we observed that prolonged elevated temperature in the presence of TCEP can lead to protein fragmentation, which can be an issue for sensitive *in vivo* peptidome analysis, and this favors DTT as reducing agent.

After reduction of disulfide bridges, the reactivity of cysteine thiol groups is quenched by the addition of alkylating agents. Two halogen-based alkylating agents - iodo- and chloro-acetamide - are predominantly used and they are added in excess during the reduction reaction[108]. Iodoacetamide (IAA) has been shown to be highly reactive and even alkylates off-targets like lysines[109]. This results in a comprised downstream peptide identification, which is the reason why many protocols switched to the less reactive chloroacetamide (CAA). Furthermore, due to the absence of free thiols on TCEP in contrast to DTT, TCEP can be used together with the less reactive CAA to enable a 'one-pot' reduction-alkylation reaction of disulfide bridges, while DTT would quench the alkylation reagent when added simultaneously and in equimolar stoichiometry[95].

A combination of mechanic and thermal disruption of rigid tissue for protein solubilization is in many cases mandatory to ensure the unbiased extraction of proteins. Many approaches have emerged of which sonication, bead-milling, grinding and blending are the most prominent ones and can also be used under semi-frozen conditions. Thermal disruption by boiling is also used in many applications where either sample-specific enzymes need to be inactivated or rigid tissue needs to be loosened up in combination with mechanic approaches. Highly optimized boiling protocols at above 90 °C in conjunction with excess tris concentrations have been developed for Formalin-Fixed Paraffin Embedded (FFPE) tissues to first break up formalin crosslinks and second quench free formalin[93,94].

In the course of my PhD project, I developed a protocol combining many of the described sample preparation advances to process solvent-cleared fully transparent tissues, which render completely solid, for a method called *DISCO-MS* **(Article 8)**. Interesting freeze-thaw-heat cycles together with sonication are emerging that are attractive for efficient protein extraction and solubilization in sample limited applications and prevent excess pipetting steps, which I also implemented for the sample preparation of single-cells **(Article 7)**[105,110].

1.2.2.2. Protein digestion by endoproteases

Enzymatic digestion is one of the most rate limiting and error-prone steps in the sample preparation process. The goal of protein digestion is to generate peptides from the sample, which are short, straightforward to be analyzed by the mass spectrometer and most likely unique per protein, and ionize well.

There is a plethora of different sequence specific enzymes, which could potentially be used in proteomics experiments. Due to the following characteristics, the by far two most used enzymes are trypsin and LysC[79]. First, trypsin always hydrolyses the amide bond C-terminally to a lysine or arginine and LysC only C-terminally to a lysine, yielding predominantly so-called *fully tryptic peptides*. Since bottom-up proteomics experiments are routinely performed in positive ionization mode, fully tryptic peptides yield multiply charged peptides due to the N-terminal amine group and the amine or guanidine group of the lysine or arginine side chain, respectively. Multiply charged tryptic peptides are highly advantageous for identification in tandem MS experiments since fragmentation experiments will mostly yield two singly charged fragments, which add up to the full precursor mass. Also, the combination of LysC and trypsin is favorable compared to Trypsin alone, because LysC is more efficient in cleaving lysine residues, yielding peptides with internal arginines to be cut by trypsin. This approach results in a very low missed cleavage rate, which is essential for label-free quantification approaches, ensuring accurate and sensitive protein quantification[111].

Alternative proteases such as LysN, AspN, GluC, chymotrypsin, or even chemical lysis by high formic acid concentrations do not have the above advantages, but can yield complementary peptides to increase protein sequence coverage[112,113].

1.2.2.3. Sample clean-up for LC-MS analysis

After digestion and prior to MS analysis, peptides need to be purified. This includes the removal of insoluble aggregates, detergents and chaotropic agents, and salts to prevent damaging or clogging of the liquid chromatography system and the analytical column. Sample cleanup is also essential to prevent analyte ionization suppression and the built-up of debris on hardware components of the MS, which can decrease ion transmission efficiency and consequently performance of the instrument. Essentially, the cleaner the injected sample, the higher the instrument up-time at highest performance and the more reproducible large-scale studies become.

State-of-the-art sample clean-up in bottom-up proteomics is done with so-called Stop-and-Go Extraction Tips (StageTip)[114,115]. StageTips are pipette tips filled with a small amount of chromatography material embedded in Teflon, also called solid phase extraction (SPE) material, which is pushed tightly to the very end of the tip. Two SPE materials are now commonly used in StageTip based cleanup procedures, which are C18 and styrenedivinylbenzene reversed phase sulfonate (SDB-RPS)[95]. After quenching of the digestion reaction under acidic conditions and pelleting of soluble fragments, the peptide solution is loaded onto StageTips and washed. C18 material allows aqueous washes only to remove salts, while SDB-RPS also allows isopropanol washes to remove lipids. SDB-RPS is also more efficient in retaining very hydrophilic peptides during the wash steps. In contrast, C18 allows for a more efficient recovery of the cleaned up and SPE-bound peptides with high acetonitrile concentrations under acidic conditions from the SPE material, while peptides can only be eluted from SDB-RPS under very basic conditions and high acetonitrile concentrations[95]. Both materials have their advantages and disadvantages with regards to sample cleanup and the experimenter has to decide which one to favor depending on the biological matrix to be analyzed. In general, our laboratory prefers the SDB-RPS cleanup, simply because the additional isopropanol wash allows for a cleaner peptide extract, which improves analytical column life-time and MS performance.

Recent developments also allow miniaturization of sample preparation in a single vial reactor, avoiding the transfer and aforementioned buffer-exchange steps[95]. This drastically simplifies sample preparation and decreases hands on time. Furthermore, miniaturized in-solution sample preparation also allowed the automation of large sample cohorts on liquid-handling platforms, streamlining the whole process and thereby increasing reproducibility within and across experiments[89,103].

Several sample preparation technologies for low cell-count experiments have been developed in the last years including micro-fabricated nanowell chips with a robotic nanoliter liquid handling system and in 384-well formats. All these approaches aim to reduce adsorptive peptides loss and improve digestion kinetics due to reaction volume miniaturization[110,116,117]. Still, one of the main remaining issues is the transfer of the peptides from the low cell-count experiment into the liquid chromatography system without losing the analytes. I overcame this challenge by coupling sample preparation to liquid chromatography as described in the next chapter and in **Article 7**.

1.2.2.4. Peptide fractionation and subset enrichment

Fractionation steps can be performed on the subcellular level, protein level and peptide level. Subcellular fractionation techniques usually aim to determine the proteomic makeup of cellular compartments, while protein or peptide level fraction aims to increase proteome coverage[118]. Since fractionation at the protein level usually suffers from lower resolution and is challenged by the solubility of proteins, peptide level fraction is the method of choice. Splitting the sample into several injections reduces complexity. This approach is especially useful for biological matrices with a very high dynamic range, or in other words, where only a small proportion of proteins occupy a large amount of the proteinogenic space as in body fluids or tissue samples[41]. The depth of proteome coverage by fractionation scales with the available sample amount, number of fractions to be analyzed and availability of measurement time[83,101]. Since on-line liquid chromatography separates peptides by their relative hydrophobicity under acidic conditions before being analyzed by the mass spectrometer, one would ideally want to couple it to orthogonal fractionation methods beforehand.

StageTips packed with either anionic or cationic SPE material allow the manual separation of bound peptides by their charge state resulting in more than 10,000 protein identifications in only six fractions per sample by means of strong cation/anion exchange (SCX/SAX)[95,119]. Sophisticated post-acquisition strategies allow merging of manually fractionated samples for qualitative and quantitative comparisons across fractionated samples, thus correcting partially for higher sample to sample variability or an automated fractionation procedure.

Another possibility is off-line basic reversed phase fractionation, which was first implemented on C18 StageTips. Here, peptides are eluted with an increasing ACN concentration under basic conditions into many fractions and pooled, or concatenated, keeping the largest distance with regards to ACN concentration levels[82]. This fractionation and concatenation principle was also automated by our group as a *loss-less spider fractionator* resulting in more than 12,000 protein identifications[83,101]. Extensive pre-fractionation of digested peptides by various enzymes already allowed us to reach a depth of proteome coverage which is in some ways on par with the comprehensiveness to which the transcriptome can be probed by next-generation sequencing[120]. I used this strategy in many projects of this PhD thesis to create representative peptide libraries, which were used to either transfer peptide identifications into single-shot measurements, or as spectral libraries for data independent acquisition analysis.

Besides the goal of obtaining the deepest representative proteome possible by reducing sample complexity, more specialized applications like the analysis of sub-stoichiometric global post-translational modifications of the proteome call for enrichment steps to reduce the background signal of unmodified peptides and increase the signal of the modified ones[56,121]. For example, several highly efficient protocols exist for phosphorylated or ubiquitylated peptides. The high affinity of the bivalent phospho-groups to immobilized metal cations (IMAC) or titanium dioxide (TiO2) is the basis for the standard phosphopeptide enrichment strategy[122]. Selective immunoprecipitation of tyrosine phosphorylated peptides is another powerful strategy to enrich only for very low abundant phosphopeptides directly involved in cell signaling[123]. This strategy can also directly be transferred to the enrichment of ubiquitylated peptides. Here, the remaining side-chain GGK-motif after tryptic digestion and removal of the covalent lysine-bound ubiquitin can be enriched by specific antibodies to investigate the ubiquitin system at scale[124,125]. Immunoprecipitation or -enrichment is also used to create global protein-protein interaction networks as presented for the human interactome in **Article 3** (Ref.[126,127]). Here, short parts of a fluorescent protein were directed to the protein of interest and fused N- or C-terminally in-frame by CRISPR-CAS9, while the complementary fluorescent protein part is expressed in trans[128]. Upon endogenous expression of the protein of interest, the fluorescent protein is reconstituted. Antibodies, which are either bead- or plate-conjugated, can then be used to immuno-precipitate the fusion protein and enrich for binding partners.

Furthermore, in the analysis of *in vivo* peptides or the entirety of the peptidome the digestion step is avoided completely. Here precipitation steps or molecular weight cutoff filters enrich for the endogenous peptides to be analyzed[73].

1.2.3. Liquid chromatography

Modern proteomics applications aim for the deepest proteome measurements at highest sensitivity, reproducibility and robustness at scale. Due to the inherently high dynamic range of the proteome and the presence of tens of thousands peptides per sample, mass spectrometers are routinely coupled to a reversed-phase high-performance liquid chromatography (RP-HPLC) system. In classical RP-HPLC, samples in solution are aspirated ('picked') with a sample syringe, transferred into a sample loop and pushed out onto an analytical column packed with a bed of (mostly hydrophobic C18) functionalized beads that the sample analytes interact with. Peptides are then eluted over time according to their physicochemical properties with an increasing proportion of organic solvent. LC-MS systems allow the on-line separation of the analyte and couple it directly to ES, which results in a decreased sample complexity entering the mass spectrometer at each time point[129].

One of the main performance determining LC components is the analytical column. Nanoflow columns are often produced in-house from cut fused silica slugs and usually filled with either monolithic material or functionalized beads. Due to many technical challenges of creating monolith columns, including polymerization reproducibility, most proteomics laboratories use particle-based column beds. Packed columns can then either be coupled to a separate emitter at orifice IDs of 20 µm and below, or the fused silica column can be laser-pulled to result in an emitter-like fine-structure before being packed[129–131].

Chromatography performance can be described by the Van Deemter equation summarizing the dependency of theoretical plate height on flow rate, eddy diffusion and mass transfer[132,133]:

$$H = A + \frac{B}{v} + \mathrm{C} * \mathrm{v}$$

H = Theoretical plate height

$\frac{B}{v}$ = Longitudinal diffusion

$C * v$ = Mass transfer

v = Linear velocity of the mobile phase

A = Eddy diffusion parameter

B = Longitudinal diffusion coefficient of analyte

The smaller the theoretical plate height, the higher the resolving power of the chromatographic system. The dimensions of the column are the main factor influencing theoretical plate height[133]. Different lengths, column IDs, bead IDs and functional groups on their surface, as well as porous surface IDs of the beads itself are parameters all need to be taken into account[134]. It is well-established that the sensitivity of the measurement increases with the lower ID of the column due to the fact that electrospray is more sensitive at the lower flow rates of smaller ID columns (see below)[135]. Furthermore, the more homogenous the column bead is packed and the smaller the bead ID, the more homogenous the sample analytes are transferred through the column preventing analyte distribution, known as Eddy dispersion. The mass transfer, a measure of the analyte moving between the stationary bead material and liquid phase, is largely dependent on the surface area of the bead material. Smaller ID beads and smaller ID bead surface porosity increase the theoretical stationary surface that the analyte interacts with and decrease mobile phase turbulence. This results overall in a more laminar flow and therefore sharper peaks. Furthermore, flow rates play a very important role. Since the same flow rate (volume per time unit) has a very different effect on columns of different dimensions, it is helpful to normalize to linear flow velocity (distance per hour, or speed of the mobile phase traveling through the column), which is defined as the volumetric flow rate per unit cross-sectional area. In general, the higher the flow rate, the less pronounced column bed imperfections are and therefore the lower the contribution of Eddy dispersion becomes. This also counteracts the longitudinal diffusion of the analyte within the column resulting in sharper peaks. The column temperature is also a very important factor to consider, since increased temperature positively influences the bidirectional mass transfer kinetics between the stationary and mobile phase, reduces column back pressure and therefore results in overall improved chromatographic resolution[136].

In our laboratory, the current gold standard for bottom-up proteomics are laser-pulled and in-house packed 50 cm columns with an ID of 75 µm packed with 1.9 µm ID beads with a 120 nm surface porosity. We run these columns at up to 60 °C to decrease backpressure at 300 nL/min flow rate resulting in a peak capacity of up to 1,000 in a 120 min gradient. However, the smaller the column and bead IDs get, the more challenging reproducible column bed packing becomes. Furthermore, the longer the columns get, the harder it is to get the column packed at all. Additionally, even small imperfections resulting from the tip pulling procedure can disturb electrospray efficiency. Finally, 'long pulls' (emitter tips, which are rather long and not conical) can make the column packing process challenging by itself, resulting in small but detrimental dead volumes at the column tip. They cause solvent mixing, peak broadening and decreased column resolution. Taking all these factors into account, packing of high performance columns can be considered as an art as much as a science, which researchers are still improving by faster and multiplexed column packing strategies[137,138]. High-performance small-dimensional LC columns can be easily overloaded with too high sample amounts, which in turn can result in very high backpressures that are detrimental for the LC pumps. This is especially an issue for binary pumping systems, where the elution gradient is mixed *in situ*, since very high backpressures can lead to solvent compression and admixture artefacts. We find that the performance decay of these columns is sharp, resulting in peak broadening and drastic analyte retention time variation across experiment, and often leading to a turnover in less than two weeks to keep measurements comparable.

Since the number of samples per proteomics study is increasing over time and the bottlenecks of current LC setups are well appreciated, the community is working on alternatives to keep the highest chromatography performance for sufficient time spans. An alternative to in-house pulled and bead-packed columns are commercial columns, which can be connected to a separate emitter with a close to zero dead volume. The most promising alternative to bead packed columns are solid silicon wafers, which are lithographically etched into a perfectly assembled array of micropillars[139,140]. This perfect column bed symmetry decreases Eddy dispersion and promises to create a perfect laminar flow of the analyte through the column. Interestingly, these *µPAC* columns appear to be very modular and can be engineered with different pillar density, shape, surface functionalization and column thickness. This allows the tailored and fully standardized creation of columns for many applications. They are also extremely reproducible with regards to analyte elution time point due to the incompressibility and perfect assembly of the pillars at potentially much longer life time[90]. These columns have already been used in large-scale pilot studies and also high-sensitivity applications, and appear to be ideal for many more applications in the near future[90,141].

Another direction, especially for high-throughput applications, where sample amounts are not very limited, are short higher ID columns (e.g. 15 cm length, 1 mm ID) operated at flow rates in the range of one or more microliters per minute. The chromatographic performance of these systems can be superb, resulting in very sharp analyte peaks, very high peak capacity and robustness, but at the expense of up to 100 to 1000-fold decreased sensitivity[142,143]. Since mass spectrometer scan speeds are increasing gradually calling for very sharp elution profiles, this approach will especially find its application in large-scale industry-like projects, where throughput and robustness are key, while there is sufficient sample amount. As one example, a recent proteomics study showed that sub-one minute gradients at a flow rate of 800 µL/min resulted in a rather high peak capacity and allowed in combination with ultra-fast mass spectrometry methods the robust analysis of large blood plasma sample cohorts at a proteome depth sufficient to draw novel biological conclusions in a biomedical setting[144].

The other extreme is liquid chromatography for ultra-high sensitivity applications down to the level of single cells. These ultra-high sensitivity applications call for very low flow rates to enable highest electrospray efficiency, a very low column ID to prevent analyte dispersion and longitudinal diffusion, while still keeping chromatographic resolution high. Many proof of principle studies report the use of in-house packed columns with IDs down to 20 µm with beads packed down to 1 µm ID at flow rates of less than 20 nL/min[145,146]. Since the production of these columns is very challenging, smallest imperfections are detrimental to overall performance, and the performance drop over time is sharp, no large scale or even routine studies have been reported with this type of columns so far. Furthermore, many binary pump systems (Figure 3A) have inherent issues mixing the gradient *in situ*, due to the very high backpressure and flow rate, resulting in reproducibility issues. Also, low sample amounts tend to adhere to sample-vials and disappear before being analyzed[147,148]. Additionally, autosampler-based LC systems dilute the sample into a large liquid reservoir, which leads to analyte dispersion and signal loss. Novel approaches transferring the analyte directly onto the chromatographic column by capillary forces have been suggested, but suffer from throughput and automation[149].

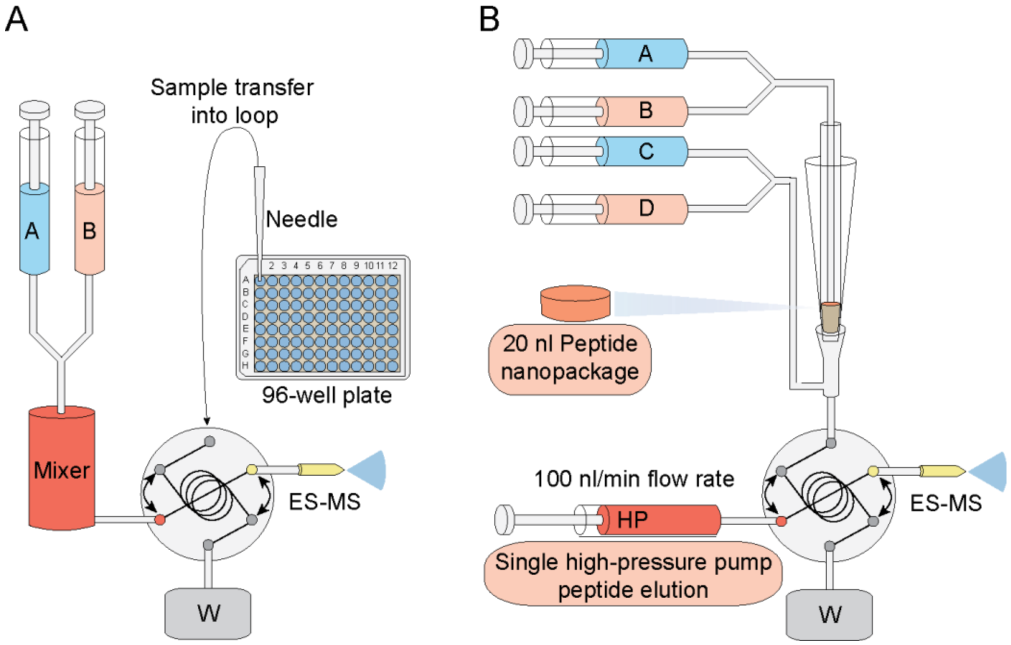


Figure 3 | Comparison of two LC pumping schemes. A, Binary high-pressure pumping scheme as used in state-of-the-art LC. **B**, Single-high-pressure pump setup as used in the *EvoSep One*.

A novel LC platform, called *EvoSep One* (Figure 3B), promises to solve many of the described bottlenecks for micro- and nano-flow applications, and even allows to directly couple sample preparation to LC analysis[150]. First, it aims to standardize LC by a high-performance industrial plug and play column and emitter setup. Second, samples without prior clean-up are loaded onto a disposable trap-column, which works just like a StageTip followed by elution of analytes with only up to 40 % ACN directly from the EvoTip into the system[114,151]. This prevents the introduction of very hydrophobic analytes and insoluble material, drastically reducing analytical column clogging and increasing lifetime. It also concentrates sample analytes at the interface of the SPE material, which is eluted as a very small volume peptide nanopackage and directly navigated to the start of the analytical column[105]. Third, analytes are eluted and stored at the very end of a preformed gradient, which is pushed out in a highly reproducible way by a single high-pressure pump. This alleviates the issue of *in situ* gradient admixture, sample dilution, and allows full control of flow rates across the gradient and gradient lengths itself. Furthermore, the choice of fixed column dimensions allows for standardized high-performance microflow applications with a flow rate of up to 2 µl/min at very high throughput or even true nanoflow applications down to only 25 nl/min for ultra-high sensitivity applications. I successfully applied the microflow gradient for ultra-robust high-speed interactome studies **(Article 3)** and realized the idea of a standardized *true nanoflow* setup with a ten-fold reduction in flow, allowing the analysis of true single-cells **(Article 7)**. Even though this LC system provides the opportunity to standardize high-performance liquid chromatography in the high-flow and true nano-flow segment, one of the weakest links is still the analytical column.

I am currently working on combining the advantages of the EvoSep LC system with the advantages of laser etched µPAC columns to create a robust full solution for LC, which is applicable to ultra-high sensitivity as well as non-sample limited high-throughput applications[139].

1.2.4. Mass spectrometry

Major advances in the fields of DNA and RNA sequencing happened due to methodological and technological developments[25,28]. This is also the case in the field of proteomics. Since our core technology besides liquid chromatography is the mass spectrometer, innovative hardware improvements are tightly coupled with an increase in proteome depth, throughput and performance robustness. In simple words, mass spectrometers detect the abundance of incoming ions and measure their accurate mass-to-charge ratio in vacuum[152]. Many parameters determine the performance characteristics of a mass spectrometer, such as mass resolving power, mass accuracy, sequencing-speed, duty cycle, sensitivity and dynamic range coverage (Box 1). Four main technical features are part of every modern mass spectrometer for proteomics applications: analyte ionization, ion filtering, ion fragmentation, mass analysis.

Box 1: Nomenclature in mass spectrometry

Mass resolving power

Measure of the ability to distinguish two peaks of different mass-to-charge ratios in a mass spectrum. It is calculated as m/Δm, while m is the actual mass of the peak and Δm is the mass deviation of the actual peak at Full Width at Half Maximum (FWHM) and is therefore a dimensionless quantity.

Mass accuracy

Measure describing the deviation of experimentally measured mass-to-charge and exact mass-to-charge ratios of an ion. It is often calculated as the root mean square value of technical repeat measurements and usually expressed in parts per million (ppm).

Sequencing-speed

Number of fragmentation scans the mass spectrometer can acquire at a given resolution per time frame, e.g. per second.

Duty cycle

Time proportion within an acquisition cycle spent on analyzing ions, or the proportion of incoming ion submitted for analysis.

Sensitivity

The signal intensity a mass spectrometer records for a fixed analyte concentration. The more sensitive a mass spectrometer becomes, the less analyte is needed to reach the same signal. Also related to the lower limit of detection, a measure for the minimum ion abundance required to detect a signal

Signal-to-Noise

The ratio of the analyte signal to the chemical and electric noise level measured on a blank.

Dynamic range

Fold change of the lowest and highest abundant ion quantified within a spectrum or across a measurement.

1.2.4.1. Analyte ionization - Moving analytes into the gas phase

One of the main challenges in MS-based proteomics is the transfer of sample analytes from ambient pressure into the first vacuum stage of the mass spectrometer. The transfer of non-volatile protein and peptide ions for biological mass spectrometry into gas phase was mentioned first in 1968 (Ref.[153]), but practically realized for the first time by the means of two soft-ionization techniques in the 1980s: matrix-assisted laser desorption/ionization (MALDI)[154–156] and electrospray (ES)[62,157].

In MALDI, analytes are co-crystalized with a matrix of molecules, which have a strong optical absorption in the UV or IR range enabling a rapid and efficient absorption of irradiation with a particular wave-length. A pulsed laser beam is directed towards the matrix with the embedded sample to locally excite the molecules. This results in the desorption of the analytes and ionization in the gas phase[158]. Due to the high-complexity of the proteome and the inability to easily couple MALDI to on-line separation techniques its application spectrum is much narrower than of ES. Still, MALDI imaging, a technique to reconstruct the analytes accessible from the surface of tissue sections in 2D, is gaining popularity in the biopharmaceutical industry and for example in large-scale drug screenings[159,160].

Electrospray is a very different approach, for transferring sample analytes from liquid into gas phase but was developed at the same time[62]. In ES, an electric field between the liquid flowing through the analytical column and the entrance of the mass spectrometer is applied, which charges the emitter tip to kilovolt potential. This leads to an electrostatic dispersion and desolvation of the sample analytes at the emitter tip[161]. Since sample analytes continuously leave the column at a fixed flow rate, this principle results in a continuous ion beam of LC-separated analytes entering the mass spectrometer. In more detail, a jet of highly charged droplets is ejected from the so-called Taylor cone at the emitter tip[161]. Subsequently, liquid dispersion and droplet formation results in a spray plume. Further evaporation of solvent molecules from the droplets increases their surface liquid tension and charge density until the Rayleigh limit[162]. This leads to a coulomb explosion forming even smaller charged droplets (note, however, that the details of this process are still not clear after many decades.) High temperatures of the carrier gas and the ion guides at the MS entrance, bridging the analyte transfer from subambient pressure into the first MS vacuum stage, assist further desolvation until analytes are close to completely desolvated. ES efficiency and therefore the transfer of sample analytes from liquid into gas phase strongly depends on the LC flow rate, since this determines the solvent amount leaving the emitter tip[163,164]. Increasing the flow rate increases initial droplet and Taylor cone size. High flow rates therefore decrease ES efficiency due to larger droplets that do not enter the MS, less efficient desolvation, ion suppression effects and the increased formation of singly charged cation adducts[152]. Nanoflow ES alleviates this issue with very narrow-bore analytical columns, allowing to reduce the flow rate into nanoflow regions, while still keeping an optimal linear velocity and increasing LC-MS sensitivity into the attomolar range[163]. In **Article 7**, we decreased the flow rate down to 100 nl/min to increase sensitivity by up to 10-fold compared to the microflow gradients that were the standard of this instrument. This was one of the main factors to enable true single-cell proteomics applications. As a matter of principle, the transfer of sample analytes from ambient pressure into the vacuum stages of the mass spectrometer always comes with a loss of sample analyte. A different ES Source, operating at pressures of about 30 Torr in the first vacuum stage of the mass spectrometer, called *sub-ambient pressure ionization* (SPIN-source), has been reported and would potentially side-step this issue[165]. Also, the addition of dimethyl-sulfoxide (DMSO) has been reported to positively impact the ES process, increasing the number of ions entering mass spectrometer[166,167]. A suggested model is that the addition of up to 10 % DMSO into the mobile phase of LC reduces the surface tension of droplets during ES, in particular for aqueous solvents. This would increase the likelihood of sample analytes being ionized since sequestration of sample analytes into charged nanoscale droplets is improved[167]. Once sample analyte ions enter the mass spectrometer, multiple ion optical elements refocus and guide ions through the instrument to the mass analyzer.

1.2.4.2. Mass isolation - The quadrupole

Quadrupoles are some of the simplest type of mass analyzers (not as simple as TOF) and consist of four cylindrical or hyberbolic rods, which are pairwise symmetrically aligned to the center of a square, with ions transmitted down the middle[152,168]. In operation, a quadrupolar field is generated by setting opposing rods to the same potential and adjacent rods to the opposite potential. Ions entering the quadrupole are exposed to the periodically alternating quadrupolar electric field, which results from a direct current (dc) superimposed on an alternating current (ac) run at radiofrequency (rf; Mhz)[152,169]. This means that ions are moving on oscillating trajectories through the quadrupole. Only a fraction of the ions with a distinct m/z will have a stable trajectory and hence be able to pass the quadrupole, while others with an oscillation larger than the inner diameter of the quadrupole will eventually hit the rods and get discharged. Setting defined potentials allows to isolate narrow m/z regions of interest[152,170]. Mathematically, the motion of ions through a quadrupolar field can be described as a function of the ac and dc potential and quadrupole geometry by the Matthieu equation, which identifies 'stability regions' for ions of a given m/z (Fig. 4).

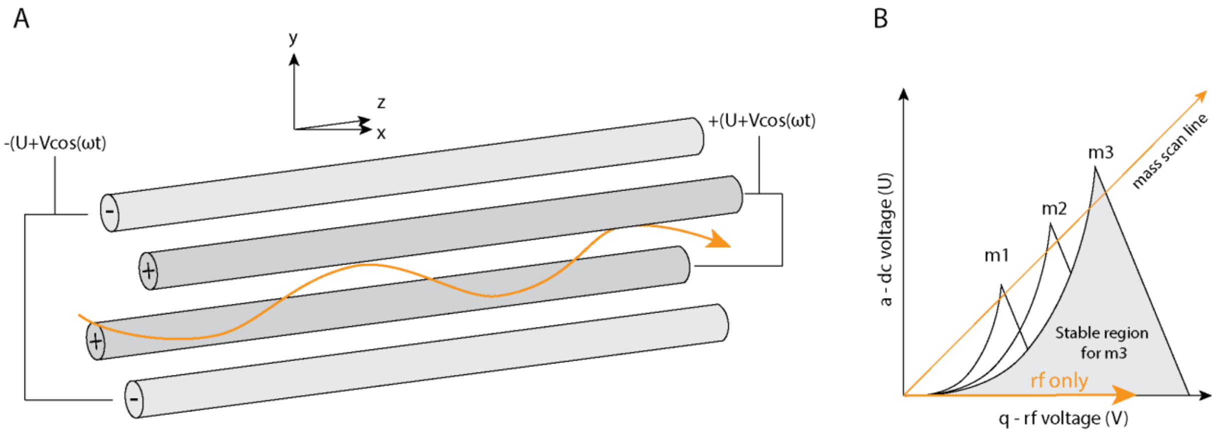


Figure 4 | A linear quadrupole mass filter. **A,** Schematic of a typical quadrupole, showing the parallel square orientation of all four rods and opposing potentials. Periodic attraction and repulsion occurs into the x- and y-plane, while analyte ions traverse in z-direction. **B,** Linear quadrupole scanning across m/z range highlighting transmission stability regions depending on the applied ac and dc potentials.

Quadrupoles operated in rf-only mode can be used to transmit very wide m/z ranges, which makes them attractive as ion guides or as 'switchable' ion filters in hybrid mass spectrometers[152]. The resolution of a quadrupole correlates with the number of oscillations ions perform in the quadrupole and is determined by the isolation width a quadrupole can offer without losing transmission efficiency of the target analyte ion. Modern quadrupoles allow the isolation of ions without signal compression at a width of below 2 m/z and with sub-ms switching times, which can be used in proteomics for the isolation of many peptide ions per second. To further increase resolution, which would enable even sharper ion isolation without transmission efficiency impairment, the kinetic energy of the ion passing the quadrupole can be lowered, the rf can be increased, rod-to-rod distance can be decreased, or the quadrupole length can be increased. However, some of these options come hand in hand with major manufacturing challenges.

Quadrupoles can also be used as ion storage devices (linear ion traps), with optical lenses positioned at the entrance and back of the quadrupole to create a potential well that prevents the ions from exiting the device. Another feature of quadrupoles (N = 4) is that they can be produced as higher order –poles such as hexa- (N = 6) or octapoles (N = 8), which have an increased ion transmission efficiency in rf-only mode and are therefore preferred at several stages of some MS as ion guides[152,170]. In some Orbitrap instruments, for instance, multipoles are used in a bent version to remove residual neutrals, where charged ions have a stable trajectory across a wide m/z range, while neutrals are tangentially ejected.

1.2.4.3. Ion fragmentation - The collision cell

Mass spectrometers in bottom-up proteomics perform two fundamental experiments per acquisition cycle. Peptides eluting from the column and entering the MS are first subjected to full scan (MS1) acquisition covering a broad m/z range of usually 100-1,700 by operating the quadrupole in rf-only mode. Precursors of interest are then isolated with either very narrow m/z ranges of about 2 m/z in data dependent, or consecutive, larger m/z ranges of for example 25 m/z in data independent acquisition, followed by mass analysis of their fragment ions (MS2). There are several fragmentation techniques for MS2 analysis, including collision-induced dissociation (CID), higher-energy C-trap dissociation (HCD), electron-capture dissociation (ECD), electron-transfer dissociation (ETD), and ultraviolet photo-dissociation (UVPD) all of which create distinct peptide fragment ion series' patterns from (usually) positively charged peptide precursors (Fig. 5)[171–173].

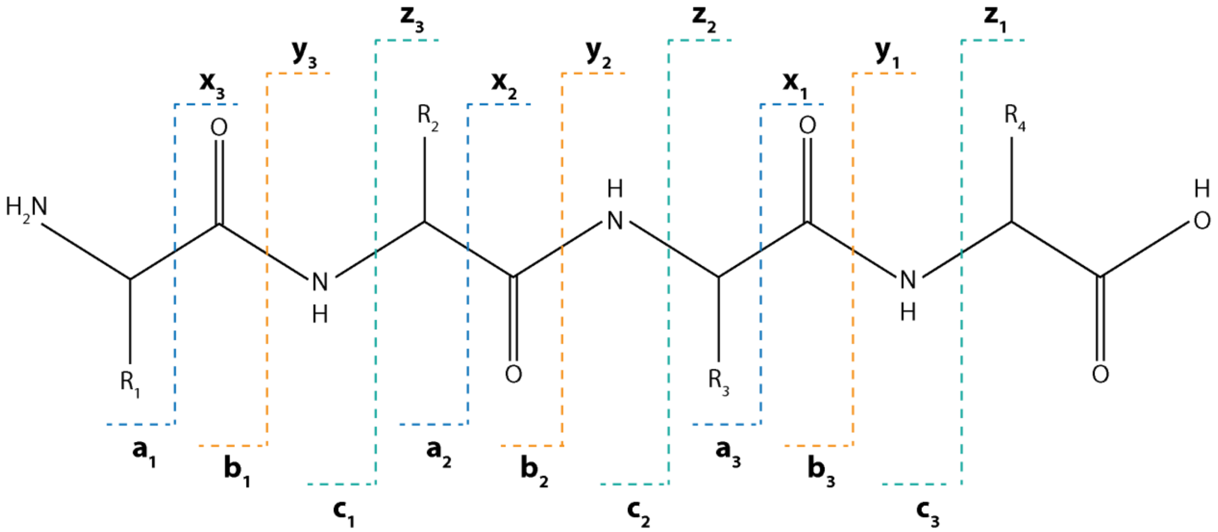


Figure 5 | Peptide backbone fragmentation pattern according to the Roepstorff-Fohlmann-Biemann nomenclature. Upon fragmentation of peptides, a/b/c fragments from the peptide N-terminus and x/y/z fragments from the peptide C-terminus are obtained.

Upon fragmentation of a positively charged peptide, fragments are created by breaking the peptide amide bond. Following the Roepstorff-Fohlmann-Biemann nomenclature, indicating the position of the dissociated bond, peptides yield a/b/c fragments when they include the N-terminus and x/y/z fragments when they include the C-terminus[174,175]. Peptide fragments are visible in the MS2 spectrum, as long as they preserve at least one positive charge from the precursor after fragmentation. All above mentioned fragmentation techniques have certain preferences for yielding a/x, b/y, c/z fragment ion pairs. The most widely used fragmentation technique in bottom-up proteomics are beam-type CID or HCD, which are actually minor variants of each other[176]. In CID, the ion beam is accelerated into a multipole run at rf-only and low pressure of a neutral gas ions like N_2, He is leaked into the multipole, which the ions collide with. High kinetic energy peptides (tens of eV) collide repeatedly with the neutral gas ions and the kinetic energy is converted into internal energy, destabilizing the chemical structure of the peptide. This ultimately leads to the breakage of the amide bond and the predominant formation of a b- and y-ion series' (Ref.[152,177]). Fragmentation efficiency is dependent on mass, charge stage and the constitution of the peptides. Furthermore, fragmentation energy can be controlled by altering the collision gas density as well as the kinetic energy of the peptides entering the collision cell. In triple quadrupole mass spectrometers (QQQ), the first quadrupole (Q1) is used as a mass filter, the second quadrupole (Q2) serves as a collision cell and the third quadrupole (Q3) serves again as a mass filter of a distinct fragmentation m/z range before ions hit the detector[152]. On Orbitrap instruments, beam-type CID was first performed by accelerating ions into the C-trap (normally used to trap ions before mass analysis in the Orbitrap) through a potential gradient - hence the term Higher energy C-trap Dissociation (HCD) and later on in a dedicated octapole ion trap[176]. Historically, this configuration overcame the problem of low trapping efficiency for low m/z fragment ions and poor fragmentation spectra for peptides with labile modifications as typically observed with resonant-excitation CID in ion traps, which used to be the only fragmentation mode for Orbitrap mass spectrometers. This development was essential for applications like the quantification of low molecular mass reporters in the Tandem Mass Tag (TMT) technology[178]. When the endoprotease trypsin is used for peptide generation, there will likely be at least one positive charge located on the N-terminal amine-group and one charge on the C-terminal side-chain of the lysine or arginine, while the 'mobile proton theory' suggests that the charge moves along the peptide backbone upon excitation[179]. This results in series' of b- and y-ion fragment pairs, which add up to the precursor mass (even though b-ions are often less stable than y-ions)[79]. In contrast, electron transfer dissociation (ETD) and electron capture dissociation (ECD) yield predominantly c- and z- fragments and are often used for top-down analysis as these techniques readily fragment the peptide backbone of highly charged ions while keeping posttranslational modifications on the side-chains intact[180–182]. UVPD is also mostly used in top-down analyses, yielding higher energy a- and x-ions that require higher energy, which was also demonstrated to be advantageous for glycosylation mapping[183–185]. These fragmentation techniques can provide complementary information to HCD or can even be combined with HCD, for example in EThcD.

1.2.4.4. Mass determination - Mass analyzers

Mass analyzers are 'the' essential hardware part of the MS. Many different types have been invented and further developed over time with very different physical working principles. All of them have their individual strengths and limitations. Since I contributed to the development of novel hardware and scan modes for a time-of-flight (TOF) based MS instrument and used an Orbitrap-based MS instrument for proteomics measurements during my PhD, I will focus on these two.

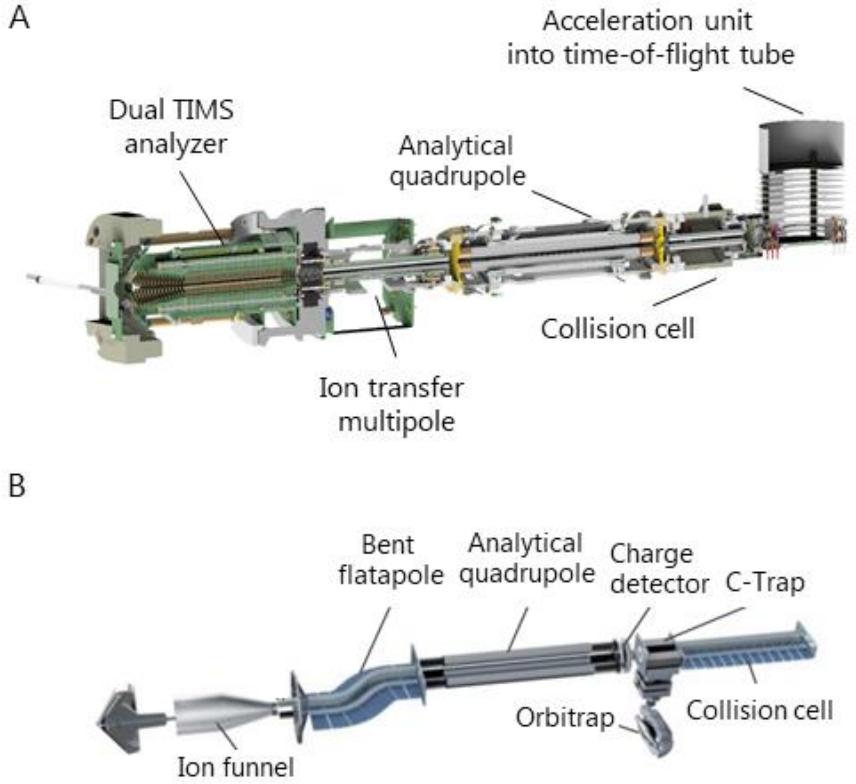


Figure 6 | Mass analyzers and how they are embedded in the ion path of hybrid MS instruments. A, The time-of-flight mass analyzer within the *timsTOF Pro* (Bruker Daltonik GmbH, adapted from Ref.[186]) used for method development and high sensitivity applications in this thesis. **B**, Orbitrap mass analyzer within the *Orbitrap Exploris* (Thermo Fisher Scientific GmbH, adapted from Ref.[187]).

Time-of-flight mass analyzers appeared the first time in 1946[188]. The first generation of instruments was designed for gas chromatography MS (GC-MS)[189]. Later, MALDI introduced a strong demand for TOF instrument, since its pulsed nature is suited to TOF and since its m/z range coverage is in theory unlimited, enabling the analysis of large molecules, which was dramatically demonstrated by the ES-TOF-based analysis of intact viruses[156,190]. In simple words, TOF instruments measure the time ions need to travel a fixed path length in a field-free high vacuum chamber after acceleration by a defined energy. This is summarized by the potential energy to kinetic energy conversion of a charged particle in an electric field. In this case, the acceleration voltage is translated into kinetic energy and finally translational motion according to the following equation[152]:

$$E_{el} = ezU = \frac{1}{2}mv^2 = E_{kin}$$

E_{el} = Potential energy of the electric field

e = Elementary charge

z = Charge stage of m

U = Acceleration voltage

m = Ion mass

v = Velocity of the ion

When ions start from rest as in MALDI, or if we assume that ions to be accelerated and extracted from a continuous directed beam were initially at rest with regards to the direction they are pushed (typically orthogonal to the direction of a continuous ion beam), its velocity can be calculated by[191,192]:

$$v = \sqrt{\frac{2ezU}{m}}$$

This highlights that the ions velocity after acceleration is inversely proportional to the square root of its mass. This allows to express the ion-specific flight time, which is the time an ion of unknown m/z needs to travel through the flight tube of a TOF instrument with a known length and acceleration voltage[152]:

$$t = \frac{s}{v} = \frac{s}{\sqrt{\frac{2ezU}{m}}}$$

s = Length of the TOF flight tube

t = Time

This finally enables the calculation of ion-specific m/z values for a beam of ions, which were accelerated into a field-free high vacuum TOF flight tube of known length at constant velocity[152]:

$$\frac{m}{z} = \frac{2eUt^2}{s^2}$$

Since ions with different m/z values will hit the detector at pico- to nanosecond scale differences, it is essential to be able to measure short time intervals with high accuracy, which was one of the resolution-limiting factors when TOF instruments initially appeared. This is now enabled by highly efficient electronics, whereas detector dynamic range is still a challenge[193]. At the end of the TOF flight tube, ions hit a detector, which is in most cases a microchannel plate (MCP)[194]. As an ion impinges on the detector surface, it releases one or more electrons with a given quantum efficiency and this signal is amplified into a cascade of electron multiplications, produced in a stacked microchannel plate. Today, MCP detectors compensate for the angular spread of the ion beam and can provide spatial resolution. The abundance of an ion hitting the detector is finally inferred from the number of electrons generated and transformed by fast analog-to-digital conversion in high dynamic range digitizers[195].

Modern TOF instruments are very attractive mass spectrometers, since their acquisition rate is very high with up to 1000 spectra/s, their unlimited m/z range in principle and intrinsically high and constant resolution. Since traditional TOFs very rapidly sample from a continuous ion beam, the signal-to-noise (S/N) of individually acquired mass spectra is not very high. Therefore, all single spectra acquired within a given time-frame can be summed up by an acquisition processing unit to drastically increase the S/N, which is common practice in the field[196]. An increase in ion density per push would in theory increase spectral S/N, which makes it attractive to introduce a trapping device fitting in between the chromatographic and TOF time-scale like a trapped ion mobility device[186,197,198].

As the mass resolution of a TOF analyzer is proportional to its total flight path length, a minimum path length is desired (roughly one to two meters). This also means insuring high vacuum to prevent collision scattering and maintain spatial focusing of ions with the same m/z values before hitting the detector, although not nearly as high as in the Orbitrap analyzers[199,200]. Reflectrons were developed in Russia in the 1970s, and cleverly compensate for slight differences in ion energy as described just below[201]. They also increase the flight path length, focusing spatially and now provide excellent overall TOF resolution of about 50,000. Reflectrons are ion mirrors, consisting of ring-shaped electrodes at increasing potentials located behind the field-free drift region and positioned opposed to the accelerator unit resulting in a V-shaped flight path. They correct for the initial spatial distribution of ions with the same m/z and slightly different kinetic energies in the accelerator unit. This principle results in an adjusted flight path length and spatial focusing[202]. After acceleration and initial drift, ions penetrate the reflectron until they reach a kinetic energy of zero, followed by acceleration into the opposite direction. This means that ions with slightly increased kinetic energy enter the reflectron deeper than ions with lower kinetic energy. To ensure that all ions are reflected within the homogenous portion of the electric field of the device, the reflection voltage is set to about 1.1x the pulse acceleration voltage.

Extending the principle of reflectrons also gave rise to so-called 'jig-saw' TOF designs, which position reflectrons at opposite positions to another to further increase drift length and ultimately resolution at the expense of compromised acquisition speed and sensitivity loss due to ion beam scattering[203]. To counteract this, electrostatic lenses are positioned halfway between the opposed reflectors for spatial focusing. In addition to reflectrons, modern TOFs use the *time-lag-focusing* principle to delay the extraction and orthogonal acceleration of the continuous ion beam right before the accelerator unit[196,204]. This results in a time-dependent ion focusing and pulsed acceleration of an ion package, which reduces the initial position acceleration effect further and consequently increases overall resolution. Combined, state-of-the art TOFs in proteomics reach a resolution between 20,000 and 100,000 across the entire m/z range[186,196].

The Orbitrap mass analyzer, introduced in 2000, is an ion trapping device in contrast to TOF mass analyzers [205]. Historical developments date back to 1923, where Kingdon used a wire along the axis of a cylindrical electrode enclosing the trapping volume with flanges to create an ion trapping device. This setup demonstrated that axial motion of ions along a wire, defined by a field curvature, can indefinitely capture ions and was called the Kingdon trap[206]. In 1981, Knight built on this principle and redesigned the Kingdon trap by introducing an increased radius at the center of the cylindrical electrode enabling storage and ejection of ions, but still with no m/z analysis[207]. Only in 2000 did Makarov realize that one could 're-purpose' the periodic back and forth movement of ions along the attractive central electrode in a perfectly symmetric Knight-like Kingdom trap to read out the m/z values of the analytes[205]. A stable trajectory along the Orbitrap is only possible, when rotation around the central axis with axial oscillations are combined. Makarov recognized that the axial frequency of the axial motion is a function of the ion m/z and is independent of their tangential velocity and spatial distribution[205]:

$$\omega_z = \sqrt{k\frac{ze}{m}}$$

ω_z = Axial oscillations (rad s^{-1})

z = Charge of the analyzed ion

m = Mass of the analyzed ion

e = Elementary charge

k = Constant

Thus, the axial oscillation frequency is inversely proportional to the square root of the m/z ratio of the ions to be analyzed. A differential amplifier connected to both halves of the outer electrode allows to detect and amplify the image current[152]. Ions with the same m/z value will move in phase, while diverging ions will move at lower or higher frequencies, which allows to deconvolute the multiplexed signals. The ion-specific frequency of its corresponding harmonic axial oscillations produces a characteristic sine wave for each m/z value. Finally, frequency domain signal translation via Fourier transformation allows sub-ppm mass accuracy measurements of analyzed ions[208]. The resolving power of the Orbitrap is directly linked to the time the ions oscillate within the Orbitrap and decreases inversely proportional to the square root of m/z. Depending on the requirements of the experimental application, the time spent within the Orbitrap (also known as *transient time*) is adjusted by the user, which results in a known resolving power and can be up to 1,000,000 at m/z 200 with a 3 sec transient on an advanced, hand-picked but otherwise standard Orbitrap instrument[209].

An essential step to couple the Orbitrap mass analyzer to continuous ion beam experiments and electrospray ionization was the development of the C-trap, which is a quadrupole bent into a C-shape and operated in rf-only mode[210,211]. Since the Orbitrap needs time to analyze a batch of ions at high resolution and sensitivity, the C-trap allows to partition the continuous ion beam by trapping a desired number of ions for subsequent introduction into the Orbitrap. As soon as the Orbitrap finished the analysis of the current ion package, it is purged and the C-trap moves the next ion package into the Orbitrap for analysis. Keeping in mind that the C-trap, like all ion trapping devices, has a maximum charge capacity, the continuous ion beam entering the mass spectrometer has to be cut off before that limit is reached – even when the mass analysis in the Orbitrap is not yet finished. This can result in the analysis of only a small percentage of the whole ion beam for sampling in an Orbitrap, depending on the desired transient length. To alleviate this issue to some degree, successive methods have increased the resolution of the orbitrap analyzer at a given transient time. The latest of these is a method called 'phase-constrained spectrum deconvolution method', which promises to half the timed needed for mass analysis at constant resolution[212].

The MS market in industry and academia is now dominated by Obitrap-based mass spectrometers due to their superior mass accuracy, dynamic range coverage and sensitivity compared to the TOF-based instruments available just a few years ago. The many Obitrap-based instrument iterations have brought an increase in speed, sensitivity, ion transmission efficiency, resolution and ease of handling[176,187,213–215]. However, TOF-based instruments appear to be ready to make a comeback due to their high speed of spectral acquisition and vast improvements in mass accuracy, dynamic range coverage and sensitivity[196]. Intriguingly, the combination of modern TOF instrument with a novel trapped ion mobility spectrometry device (TIMS) promises to open up entirely new vistas. We have shown that it increases sensitivity and sequencing speed compared to previous TOF instruments for proteomics experiments by at least 10-fold. It also allows close to 100 % ion beam utilization at full speed and much more[186,195,198]. Performance-wise, this instrument is currently a serious competitor for Orbitrap-based instruments, the first time this has happened in almost two decades.

1.2.5. Computational proteomics

The high-confidence identification, quantification and interpretation of acquired large-scale mass spectrometry data is a key challenge in proteomics. The human proteome alone, when subjected to a bottom-up proteomics workflow, could give rise to more than 600,000 tryptic peptides at a length cutoff of seven amino acids, even without missed cleavages or taking into account posttranslational modifications[216]. Analyzing tryptic proteome samples of human origin currently typically results in more than 120,000 MS2 fragment spectra on an Orbitrap platform and more than 700,000 MS2 fragment spectra on a TOF-based platform per 2 h run. Large-scale experiments have reported the analysis of more than 18 billion fragment ion spectra gathered across more than 16,000 LC-MS runs, which have to be identified and quantified in an automated way at high confidence, precision and accuracy[87].

By now, a plethora of software tools have been developed to enable proteomics analysis of such data sets. For data dependent acquisition, the gold standard in many laboratories including ours is still MaxQuant, developed in our group from 2006 onwards. Since MaxQuant was used for most of the projects in the course of this PhD and the same key concepts it relies on for protein identification and quantification are also used in other software solutions, I will mostly describe the computational workflow from raw data to output tables the way it is implemented in MaxQuant[217,218].

1.2.5.1. Protein identification

The first step of MaxQuant analysis is 'feature detection' where data objects are reconstructed from a 3-dimensional space (or by now 4-dimensional when using ion mobility) at the MS1 level. The key at this stage are high-resolution spectra where isotope patterns are fully resolved. Features describe ions analyzed by LC-MS, which are assembled as a function of retention time, m/z and intensity, and also ion mobility in case of a TIMS device coupled to a TOF mass analyzer (Fig. 7)[217,219].

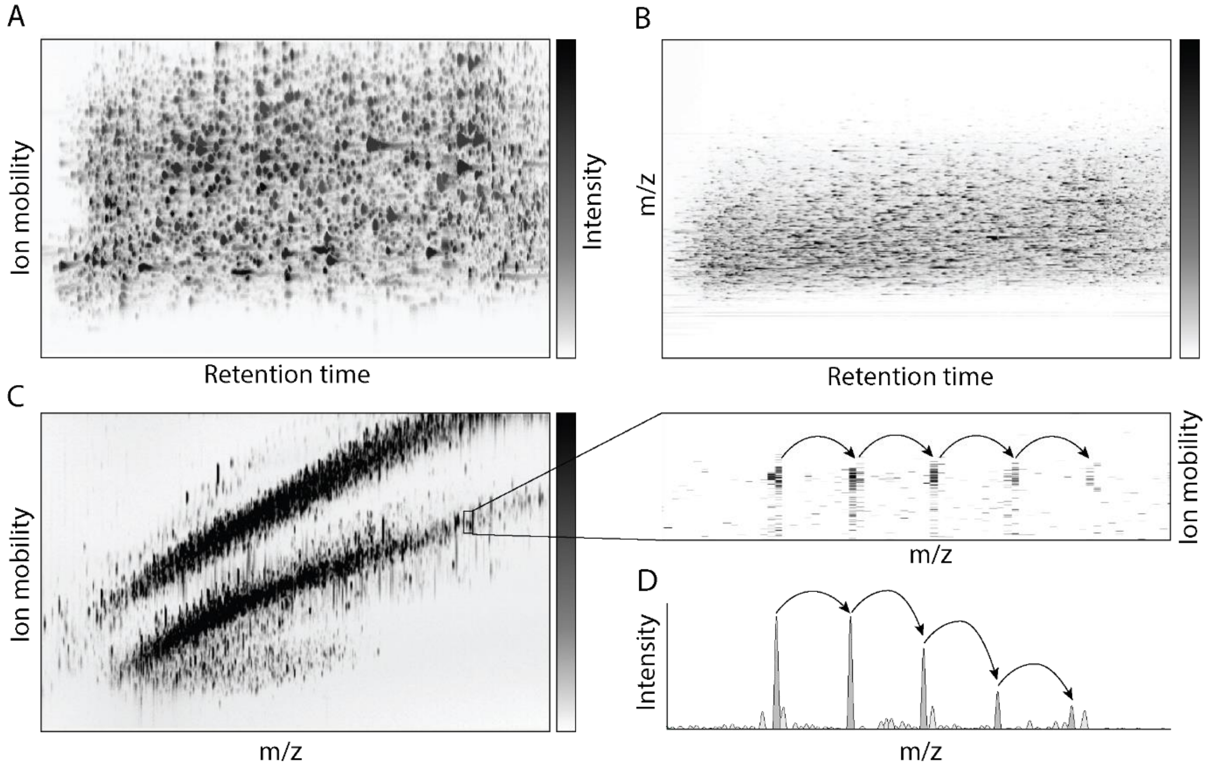


Figure 7 | Feature space in LC-MS analysis of complex proteomics samples. A, 2D feature space of RT versus ion mobility. **B**, Typical 2D feature space of RT versus m/z. **C**, A single ion mobility scan in the center of a gradient and zoomed view of an isotope pattern in m/z to ion mobility space. **D**, Mass spectrum of the corresponding isotope pattern.

Experimentally derived isotope patterns are fitted to an 'averagine model', which is an approximation of the expected isotope distribution for a distinct ion mass, followed by charge state assignment and de-isotoping[220]. Here, isotope patterns are collapsed to a single potential peptide precursor mass to speed up and enhance the next analysis steps. Co-eluting isotope patterns can be separated by higher-dimensional feature spaces (e.g. in ion mobility enhanced measurements)[219]. Even though high-resolution mass spectrometers have a very high mass precision, it does not mean that they are inherently accurate. This is due to static errors like hardware imperfections or dynamic errors like temperature shifts, which result in systematic measurement errors and calibration drifts in the 3D or 4D feature space. To correct for this in real-time, atmospheric polymethylsiloxanes were initially used as 'lock-masses', achieving low ppm mass accuracy[208]. Later, the 'software lock mass' approach was developed, which allows for a more robust post-acquisition strategy without using actual calibrants. Here, a first survey search at high mass tolerance for high confidence identifications is performed and thousands of high-confidence peptides are used for nonlinear global recalibration of the m/z dimension, resulting in sub-ppm mass accuracy[221]. Since non-linear shifts in the retention time are common, dedicated global correction algorithms were developed and in particular in MaxQuant a similarity-guided tree approach is used to circumvent the need for an alignment to a reference run[222].

After feature detection and recalibration at the MS1 level 3D- or 4D feature space, the next step in data processing is the peptide identification. Here, the fragmentation information on MS2 level of isolated precursors is used to determine the sequence of a peptide. In database search engines, peptides following cleavage rules mirroring the experimental protease are generated *in silico* from a protein sequence database, which is inferred from the genomic information of the experimental organism[223]. Then, masses corresponding to the expected fragment ions to be found in the MS2 spectra are calculated following the experimentally used fragmentation techniques (e.g. b- and y-ions for HCD/CID). Next, the search engine calculates a score for each experimental MS2 spectrum against all theoretical fragmentation spectra within a specified mass tolerance window and precursors mass, following many criteria like number of matching fragment ion masses. The highest scoring peptide spectrum match (PSM) remains as a possible peptide identity-verifying candidate spectrum[86,224]. The calculated high accuracy mass from the MS1 level is used to decrease the number of possible spectral comparisons by narrowing down the possible full length peptide mass of the match[217]. Due to the high proteome complexity, the best scoring PSM might still be a false-positive, which raises the issue of a proper false-discovery rate (FDR) control mechanism. In the target-decoy model, experimental spectra are not only compared to the target database, but also against a decoy database, which can only produce false-positive PSMs. Comparing the score distribution of all PSMs reveals that target and decoy database hits resemble two distinct Gaussian distributions, while the decoy database hits produce a low scoring one. Posterior error probabilities, taking into account the PSM identification score, number of variable modifications, charge stage, number of missed cleavages and peptide length, can now be calculated based on the target and decoy database distribution, which allows simple FDR-control by thresholding known false-positives within the dataset at 1 %. Decoy databases can be generated in several ways, while the most prominent one is a reversed version including terminal arginine or lysine swaps, providing nonsense peptides that do not occur in nature. Other models also exist, which are mostly machine learning based[225,226]. Interestingly, recent developments in deep learning and the prediction of fragment spectral intensities, retention times and collisional cross sections allowed to add more layers of information to increase the confidence in PSM calling by more than 10-fold depending on the application[227–230]. Especially very large search spaces with many more possible matches and non-standard peptide fragmentation patterns like unspecific digests or unexpected proteoforms drastically benefit from this in the future[231,232]. Currently, these models are mainly limited to unmodified and rather short peptides of less than 25 amino acids, but recent developments point into the direction that these approaches will be a driver of future innovations in the field of computational proteomics.

A different approach for peptide identification, based on MS2 level fragment ion information, is *de novo* sequencing. Here, exact fragment ion mass differences between adjacent fragments can correspond to unique amino acid masses and if e.g. a full b- and y-ion series exists, the full MS1 level precursor mass and amino acid sequence can in principle be reconstructed without the need for a sequence database search[233]. In hybrid approaches, incomplete *de novo* sequenced fragment ion spectra, resembling only a part of the peptide amino acid sequence (sequence tag) are combined with database search to increase sensitivity of the database search[84,234,235]. The advantage of this approach is that one knows the missing mass to the C- or N-terminus based on the precursor mass, which can only come about by a fixed combination of amino acids or a PTM. Such PTMs can then be localized by *in silico* placing them on all possible amino acids in the matching sequence.

Known PTMs of amino acids can be routinely identified by database searches, where a fixed mass corresponding to the PTM can be assigned as present or absent. A major challenge is that the number of different PTMs to be considered is limited due to combinatorial explosion of the *in silico* generated fragment ion spectra and the resulting search time needed for spectral comparisons. This issue is already alleviated to a large degree by novel search engines, which use fragment ion indexing to drastically speed up the whole search, but it still has its limits with regards to the size of the resulting database[236]. There is a trend towards using the massively parallel computing power available on graphics processing units (GPU) in the proteomics community and this promises a breakthrough in speeding up search calculations, just like it did in deep learning[237,238]. Two major recent developments promise a solution for the unbiased search of PTMs. First, MaxQuant offers the so-called dependent peptide search, which assumes that peptides tend to occur in a modified and unmodified state. Here, all unassigned MS2 spectra are compared to all already assigned unmodified spectra and investigated for systematic fragment mass shifts of the whole fragment ion series, which could be due to a PTM[239]. Another option are open-search algorithms, which do not increase the database size directly, but open up the MS1 precursor mass tolerance window to several hundred Dalton, while MS2 fragment mass tolerance is kept low. This allows to match mass shifted fragment ions series', due to a PTM or mutation, to still be matched to the unmodified counterpart in a database search[236,240]. However, the large search space itself decreases statistical power of peptide identification.

After peptide PSMs have been identified and filtered to 1 % FDR, they need to be assigned to proteins of an organism-specific reference proteome to reflect the qualitative and quantitative proteome of the analyzed sample[218]. These reference proteomes, including isoforms of open reading frames and much additional information about the proteins itself, are deposited in a curated form in the UniProt database[241]. Since bottom-up proteomics gives rise to many peptides, which can be either unique, or shared between proteins they could have arisen from. Dedicated methods have to be applied to solve this many-to-many relationship. In MaxQuant, *Occam's razor principle* is applied as a solution to the protein inference problem, which aims to find the smallest set of proteins that explain the observed peptides. Proteins that cannot be identified by unique peptides are combined into 'protein groups'. Since it is known that protein misidentifications tend to accumulate across large data sets, the FDR rate has to be controlled at the protein level, too. One possibility is to create a protein level score as the product of all PEPs from PSMs matching the protein and taking into account a factor for the total number of used PSMs. Since the protein FDR dominates the PSM FDR, retained PSM have FDRs far below the original of 1 %. Furthermore, high confidence PSM identifications can be transferred between experimental runs after recalibration and FDR filtering at the PSM and protein levels, as described above, to alleviate the stochastic nature of data dependent acquisition methods[242].

1.2.5.2. Protein quantification

The identification of peptide and their inferred proteins allows the description of the qualitative proteome within and between samples. The ability to quantify identified proteins precisely and accurately across a wide dynamic range is even more important, because it enables the investigation of the quantitative proteome distribution within a given sample and allows the comparison of single or protein group abundances across several conditions. Since the scale of MS-based proteomics studies has increased dramatically over the last few years and the technology is finally being applied to large-scale sample cohort studies comprising thousands of samples, robust identification and especially quantification is more important than ever[105,126,243].

Quantification of proteins is not a trivial task, since samples have to be submitted to the described multi-stage bottom-up proteomics workflows including sample preparation, liquid chromatography, electrospray ionization and mass spectrometric analysis. All of these steps can potentially introduce systematic or stochastic distortions on the true results. In sample preparation, pipetting errors of small volumes, variations in starting material, the amount of endoproteases used for digestion, and chemical modification are only some of the sources for initial variation. In liquid chromatography, minor differences in column length, solvent constitution and temperature can affect chromatographic resolution and peptide retention behavior, which makes quantitative comparisons across samples challenging. Furthermore, variation in electrospray ionization influences how many ions enter the mass spectrometer, which is affected by background ions, emitter fouling and spray voltage, if present. Mass spectrometers can also introduce variation by performance decay over time with regards to ion transmission efficiency and time-dependent calibration drifts. Quantitative proteomics is therefore a rather complex endeavor and a plethora of techniques have developed over time to enable precise and accurate protein abundance estimations[244–246]. They can be divided into two main categories, which are label-free and label-based techniques for the quantification of protein abundance on the MS1, MS2, or both levels (Fig. 8).

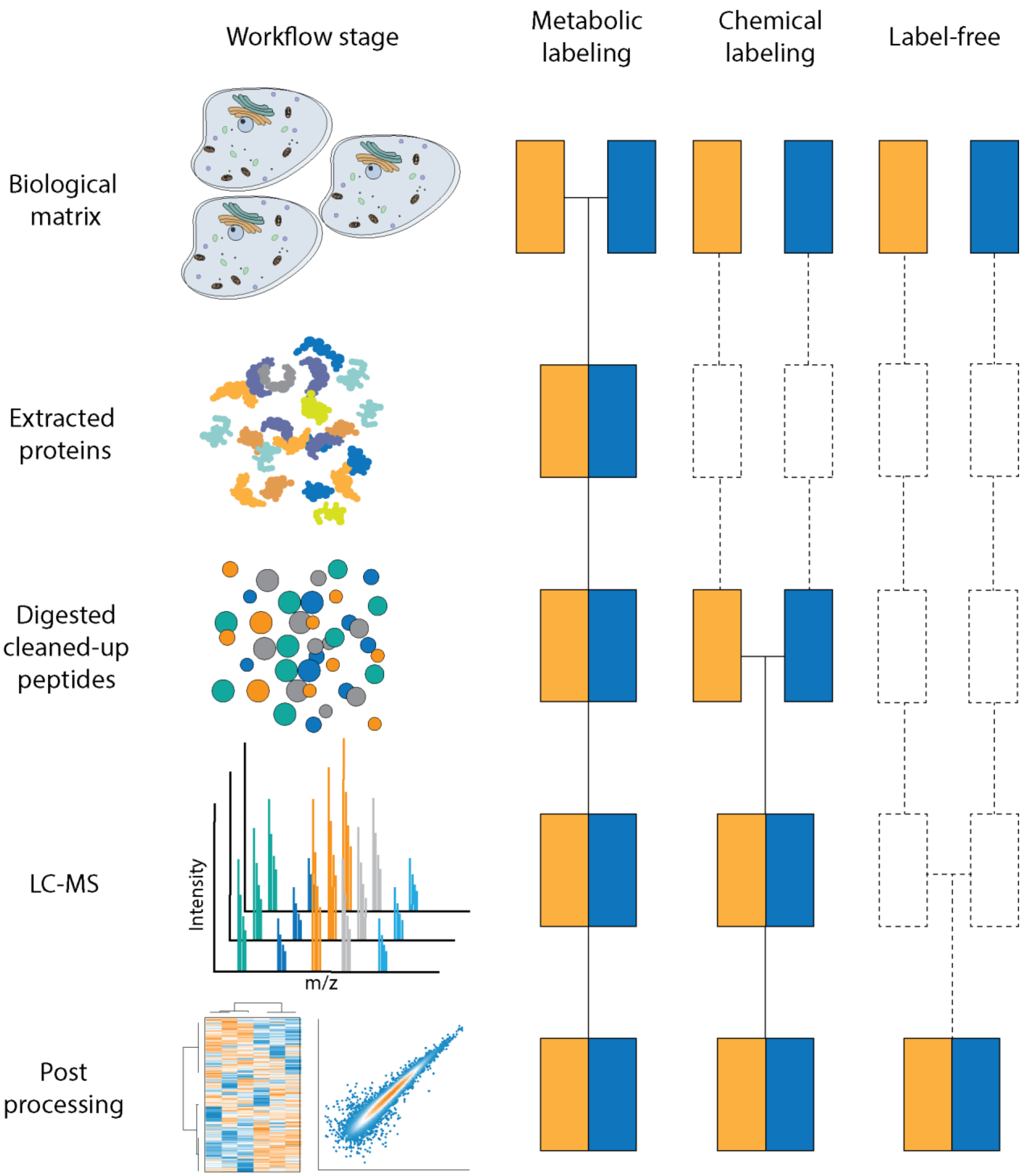


Figure 8 | Quantification strategies in bottom-up proteomics. Left, Stages of bottom-up proteomics. **Middle and right**, Orange and blue boxes indicate different samples; Horizontal lines indicate at which stage the samples are pooled; All steps before pooling can introduce systematic or stochastic quantitative bias. Dashed lines indicate potential bias introduction. (Adapted from Ref.[246])

In an ideal quantification experiment resulting in highest precision and accuracy, samples to be quantified experience exactly the same bottom-up workflow. This principle was largely implemented in 2002 with the introduction of *in vivo stable-isotope labeling by amino acids in cell culture*, short SILAC[247]. SILAC experiments use the natural occurrence of stable isotopes and in case of a cell culture experiments supplement the media with an either only heavy or light labeled version of an amino acid - in most cases lysine and arginine. After only five cell doublings, proteins of each respective cell are essentially labeled completely as either heavy or light[247]. Since trypsin digestion yields peptides with a C-terminal arginine or lysine, tryptic peptides are mass shifted exactly by the introduced natural isotope mass difference. The foundation for this approach is that isotopically labelled peptides have the same physicochemical properties when it comes to LC-MS analysis and behave the same throughout the workflow, but can be separated by their m/z difference in the mass spectrum. The most accurate quantification between conditions can be reached by mixing heavy and light labelled cells right after the experiment followed by a bottom-up proteomics workflow and calculation of relative MS1-level ratios, which circumvents any workflow introduced distortion[248]. SILAC has also been extended from cell culture to full organisms like *Mus musculus*, which resulted in fully heavy (13)C(6) or light (12)C(6) lysine labelled mice after four weeks of supplementation as exemplified by the analysis of several organs and blood cells[249]. Furthermore, SILAC was used to investigate protein turnover in a pulsed chase SILAC experiment[250,251] and found widespread application in PTM analysis[252]. Even though SILAC labeling presumably has the highest quantitative accuracy, it increases MS1 level spectral complexity by the factor corresponding to the number of introduced SILAC labels. This complicates spectral analysis, which usually results in a decreased proteome dynamic range coverage and together with the finite number of possible labels limits multiplexing capabilities[248]. Furthermore, the ideal heavy labelled amino acid remains lysine since it is essential and metabolic interconversion to proline is observed in case of heavy labelled arginines[253]. Although SILAC was initially limited to cell culture experiments, it was later extended to other biological systems that do not allow metabolic labelling. In one approach, called super-SILAC, heavy labeled pooled cell lysate spike-ins serve as a reference channel to the endogenous peptide version on MS1-level[254].

SILAC has been the gold standard for accurate quantification in proteomics for several years. Still, since it is labor intensive and expensive, metabolic labeling is not applicable to every sample type and in many cases more than two or three conditions need to be compared. Sample multiplexing by chemical labeling is an alternative approach to SILAC, which can be performed independently from sample origin and is usually introduced after protein digestion[246]. By now, several techniques have been developed, but *isobaric labeling* approaches like iTRAQ and foremost TMT have become very popular in the proteomics community[255–257]. These labels are constituted by a 'balancer group' and 'reporter group'. Using N-hydroxysuccinimide chemistry, free amine groups are labeled with a preference for side-chains rather than only the peptide N-terminus. Just like SILAC, the isobaric labels have the same physicochemical properties in LC-MS analysis, but they do not increase spectral complexity on the MS1 level. Stable ^{13}C and ^{15}N isotopes are distributed across the balancer and reporter group within an N-plex in a way that all N labels have the same mass. Upon fragmentation in MS2 experiments, label-derived low molecular mass reporters are produced, which split by the introduced isotope mass shift and are used for relative quantification across the N-plex. Furthermore, peptide ions carrying the balancer group are produced in minor proportion, known as complementary ions (TMTc in case of TMT labeling) and were shown to be promising for alternative quantification strategies[258]. The TMT principle by now allows the multiplexing of up to 16 samples in a single measurement by commercially available kits[178]. Within an N-plex, isobaric labeling approaches can result in deep proteome coverage with high quantitative accuracy when combined with fractionation, but the integration of several N-plexes suffers from a continuous increase in missing values and batch effects[259]. Also, a very high mass resolution of e.g. 128k for the 16-plex on MS2-level is needed to deconvolute the low molecular mass reporter ions, which limits it to very high-resolution mass spectrometers only. In addition, multiplexing reagents are very expensive, but miniaturization and highly optimized protocols decreased the cost per experiment by close to 10-fold[260]. Furthermore, since N-plexes are identical for all peptides within a sample, a major drawback is *ratio compression* due to co-eluting and co-isolated peptides, which disturb the low molecular mass reporter ion intensity[261]. This issue can be alleviated by MS3 methods, which repeat the fragmentation of the most intense MS2-derived peptide fragment ion, while co-isolated ions are fragmented and distributed across the full MS2 scan[262]. Decreased quadrupole isolation width has also been suggested[263]. Still, these approaches come with disadvantages like longer cycle times and decrease in proteome depth. A novel chemical labeling reagent called EASI-tag (easily abstractable sulfoxide-based isobaric-tag) promises very accurate quantification without described drawbacks of other multiplexing agents. First, EASI-tag very efficiently generates peptide-coupled reporter ions at high yield and rather low collision energies compared to TMT. Furthermore, the isolation of MS1 precursors with narrow and asymmetric quadrupole windows enables ^{12}C-only precursor isolation. This suppresses the signal from adjacent isotope peaks and enables ratio-compression free quantification of up to six multiplexed samples[264]. We implemented this particular scan approach in our MaxQuant.Live software, which is freely available to the community[265]. Still, there are some disadvantages like decreased identification rate, high costs and the availability of a 6-plex only.

In contrast to label-based quantification approaches, where the number of samples to be analyzed at high accuracy is very limited, label-free quantification (LFQ) approaches hold promise to keep up with the quantification of rapidly growing sample numbers in proteomics projects such as more than 1,000 plasma proteomes or more than 4,000 protein pulldown measurements[90,243,266]. They can also be applied to virtually any sample type and number. Due to its scalability, robustness and ease of use LFQ has turned into a method of choice for proteome quantification over the last few years and has become the gold standard in our laboratory[65]. Spectral counting concepts were among the first and most primitive LFQ approaches for protein quantification, as they simply count how often a peptide precursor was sent for MS2 sequencing[267,268]. Although a great improvement on purely qualitative measurements, these approaches are sensitive to experimental parameter changes like chromatographic peak width, dynamic exclusion and acquisition speed. More modern approaches calculate either MS1 intensities in data dependent, or precursor matched MS2 fragment ion intensities in data independent acquisition across the elution peak for more accurate quantitative estimates[222,269]. Since samples within a cohort are first identified within each run, followed by relative quantitative comparisons between the runs, any bias introduced by the bottom-up proteomics workflow is reflected at the raw intensity level, as discussed above. To enable most accurate and precise LFQ, all experimental sources of error have to be kept to a minimum. This can be done for example by automated sample preparation, robust liquid chromatography and mass spectrometers with a very flat or absent performance decay. Furthermore, sophisticated algorithms normalizing for many of these effects have been developed, which gather peptide intensity estimates into protein level estimates to allow for quantitatively very accurate comparisons[223]. In MaxLFQ for example, this is realized by calculating the median-fold change across many peptide pairs, resulting in a very robust quantification across many runs and very accurate quantitative estimates as shown by the means of a two-proteome experiment across several orders of magnitude dynamic range[270]. Inherently, MaxLFQ assumes that the majority of the proteome is stable between conditions. The more peptides are identified, the better LFQ becomes, since the number of data points across runs is automatically increased, which allows many more peptide pairings. Due to the semi-stochastic effect in data dependent acquisition scan modes, inconsistent quantification of peptides across many runs can occur, which can result in a comprised quantitative estimate. This is to some degree alleviated by the *matching between runs* algorithm[218]. Furthermore, novel data dependent and very fast data independent acquisition scan modes have proven to generate very high data completeness at high quantitative accuracy across hundreds of runs and even across laboratories[105,186,271–273]. Furthermore, the higher the resolution of a mass spectrometer, the more accurate extracted ion chromatograms become, which automatically improves label-free quantification accuracay[270]. Interestingly, novel approaches start to combine the quantitative information from MS1 and MS2 extracted ion chromatograms with dedicated LFQ algorithms similar to MaxLFQ to improve quantitative robustness even further.

After identification and quantification of experimental proteomes, a plethora of bioinformatics methods and software packages can be used to explore the data. Perseus and MSstats are only two state-of-the-art solutions[242,274]. A recent development for the interpretation of proteome data in a clinical setting is the clinical knowledge graph[275]. It aims to automate downstream analysis and integrate proteomics with clinical data in a graph database currently comprising more than 16 million nodes and 220 million relationships representing public databases, literature and experimental data.

1.3. Ion mobility spectrometry

Ion mobility spectrometry (IMS) describes the separation of ions by size-to-charge ratios based on their interactions, or collisions, with an inert buffer gas in an electric field. This promises the separation of isomers, isobars, and conformers in the analysis of biomolecules[75,276]. The first ion mobility spectrometry (IMS) related experiments can be traced back to the late 1800s with Thomson and Rutherford studying the mobility of ions formed by X-rays[277]. The first instruments in which ions were sent through electric fields, were developed in the early 1900s and they significantly improved the analytical capabilities of ion mobility spectrometry. This already allowed the creation of sharp ion species peaks[278]. Many parameters like pressure, temperature, electric field strength and the time ions spent within the device were subject of early investigation for their influence on ion mobility separation[279]. IMS has been coupled early to mass spectrometers to study gas-phase ion chemistry in the 1960s and has already been used in the early days of ES[280–282]. Since the success of MALDI and ES, IMS co-developed into an increasingly mature technique[62,156]. One of the important developments was the development of electrodynamic ion funnels, which re-focus ions undergoing jet expansion in ES and efficiently transfer them from ambient pressure into the first vacuum stage of the mass spectrometer where most of the IMS devices are located[283].

In a classical ion mobility or 'direct diffusion' experiment, the arrival time of ions migrating through a buffer gas under the influence of a homogenous electric field is measured. Under ideal conditions, e.g. no carrier gas contamination, the velocity or mobility of ions passing the field is proportional to the electric field strength and dependent on the intrinsic physical properties of the ions. This allows for the calculation of the ion mobility constant K [cm^2 V^{-1} s^{-1}] (Ref. [284]):

$$K = \frac{v}{E} = \frac{L^2}{Vt_d}$$

K = Ion mobility constant

v = Velocity of the ion

E = Electric field in the drift region

L = Length of the drift region

V = Total voltage drop from start to end region

Since the ion mobility of ions varies based on temperature and pressure, it is common practice to correct the ion mobility constant by standard temperature and pressure to a reduced mobility constant (K_0). This can be a quantitative indicator for an ions identity and is constant for a given compound in a distinct buffer gas[284]:

$$K_0 = K\frac{P}{760}\frac{273}{T}$$

K_0 = Reduced mobility constant

P = Pressure in the drift region

T = Buffer gas temperature

The reduced ion mobility constant is also fundamentally related to the collisional cross section value (CCS) of an ion through the already mentioned *Mason-Schamp equation*, which is a direct measure for the rotational average of the analyzed ion in the gas phase and directly proportional to the inverted reduced ion mobility constant (K_0)[285]:

$$\Omega = \frac{3Qe}{16}\frac{1}{K_0}\sqrt{\frac{2\pi}{\mu k_B T}}$$

Ω = CCS of the ion and drift gas molecules

K_0 = Reduced mobility constant

Q = Ion charge

e = Elementary charge

μ = Reduced mass of the ion and drift gas molecules

k_B = Boltzmann constant

T = Drift gas temperature

To determine accurate CCS values, it is essential that the electric field strength is kept the low field limit, which is defined by the maximum E/N, where E is the drift field strength and N the number density of the buffer gas, to keep the ions mobility independent of the drift field[286]. At an extreme, if the electric field strength is too high and the buffer gas density is too low, all ions would be pushed through the drift region without separation.

Due to the time-scale requirements of IMS (in the ms range), it ideally fits in between liquid chromatography (sec) and mass analyzer (especially time-of-flight mass spectrometers (µs)). It has the potential to improve a wide range of performance factors such as speed, selectivity and sensitivity when coupled to mass spectrometry. Furthermore, in this thesis, we could show that TIMS increases the peak capacity of an LC-MS setup on average by 10-fold and that the average number of peaks separated by the IMS device per time point is 10 (Ref.[216]). Importantly, the determination of ion mobility values increases the dimensionality for describing sample analytes from 3D (m/z, intensity, retention time) to 4D[186,219]. The fact that mobility correlates well with the mass or mass-to-charge ratio of an ion also makes it useful for the identification of unknown compounds by direct correlation curves, which we have made use of in **Article 4** (Ref.[75,287]).

Many different IMS devices have been developed over time with unique advantages and disadvantages depending on the application. There are three different types of ion mobility experiments, namely *temporally-dispersive*, *spatially-dispersive* and *confinement and selective release*[279]. The *spatially dispersive* IMS method, such as differential or field asymmetric IMS, sends ions on different drift paths and separates them in space based on their mobility, and thus act as ion filters. In contrast, *temporally-dispersive* methods, such as drift tube or traveling wave IMS, separate ions according to their mobility and ions will arrive sequentially in a time-resolved manner at the detector. In the *confinement and selective release* method, such as TIMS, ions are trapped in a low pressure region and selectively ejected based on their ion mobility. Four major types of ion mobility mass spectrometers have been commercialized: drift-time-, traveling wave-, differential- and trapped ion mobility spectrometry. All of these devices, except for differential IMS, allow to determine (directly or following calibration) collisional cross sections (CCS) as a measure for the rotational average of an ion conformation in the gas phase.

1.3.1. Drift tube ion mobility spectrometry

Drift tube IMS (DTIMS) devices belong to the *temporally-dispersive* category and are operated either under ambient or reduced pressure conditions. In DTIMS, ions are axially propagated through a static buffer gas with a uniform electric field. This leads to the separation of ions according to their ion mobility in the drift tube. Ions with a high ion mobility pass the drift tube earlier, while ions with a low ion mobility pass the device later (Fig. 9)[279].

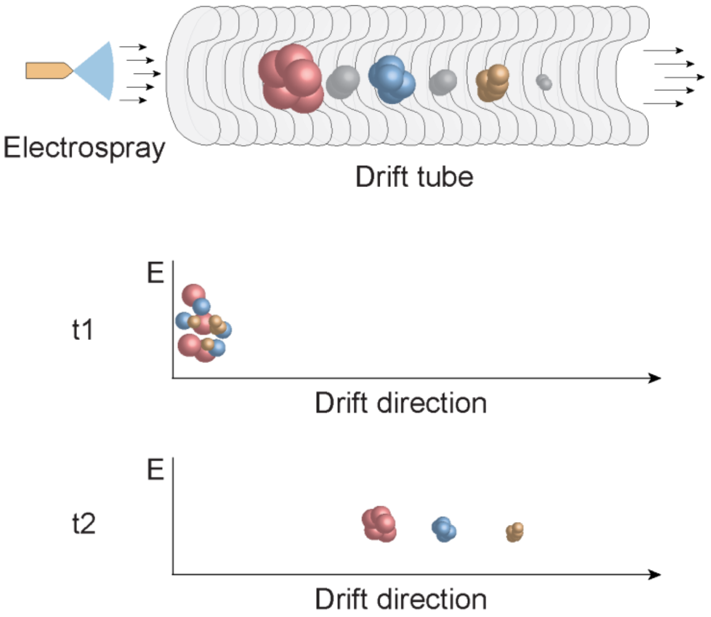


Figure 9 | Drift tube ion mobility spectrometry. Separation is achieved for ions dragged through a static drift gas by an electric field. Small ions pass the drift tube faster due to their higher mobility, while large ions pass the drift tube later due to their lower mobility. (Adapted from Ref.[279])

This strategy allows the direct calculation of ion mobility values of sample analytes and also collisional cross sections via *Mason-Schamp equation*[288]. The precise control of gas pressures and electronics of DTIMS instruments enables absolute CCS deviations across measurements of less than 0.5 %, highlighting the potential for precise CCS determination[289]. In contrast to other IMS devices, DTIMS results in a comprehensive ion beam representation, since all analyte mobilities are collected per pulse[290]. Stacked ring ion guides (SRIG) were included very early in the device manufacturing process of DTIMS instruments. SRIGs consist of metal rings stacked coaxially along a central line with small gaps in between the rings. They allow to maintain a very uniform electrostatic field to radially confine ions by two RF waveforms applied to two consecutive sets of plates at same frequency, but 180° out of phase. Furthermore, a direct current is axially applied to the SRIG to make ions propel through the SRIG. Their modularity allowed to assemble them at a length of more than one meter for ultra-high resolution DTIMS[291]. Due to long drift times, overall duty cycles tend to be rather low in these devices and in DTIMS in general. Endeavors to multiplex TOF pushes in the acceleration unit increased the duty cycle from a few % to up to 50 % (Ref.[292]). Furthermore, it is challenging to improve the resolving power of DTIMS instruments, since ions have to be kept in the low field limit so that the *Mason-Schamp equation* still holds true. Approaches to do so include an increase of the drift cell length and pressure in conjunction with an increased voltage drop across the device. This was shown to increase the resolving power to up to 250 ($\Omega/\Delta\Omega$)[293–295]. Even though prototype instruments have proven the applicability of DTIMS for proteomics applications, the need for very high voltages and rather long drift tubes hampered commercialization and acceptance in the community[296].

1.3.2. Travelling wave ion mobility spectrometry

Travelling wave ion mobility spectrometers (TWIMS) are operated in the first vacuum stage of the mass spectrometer at reduced pressure and belong to the *temporally-dispersive* category of IMS devices. This device, coupled to a time-of-flight mass analyzer, was first commercialized in 2006 by Waters as the *Synapt HDMS* and later the *G2* and *G2-Si* (Ref.[297,298]). However, in TWIMS, a pulsed direct current is applied to the electrodes and used to move ions through the drift gas region in 'voltage waves'. To radially confine the ions, RF-voltages of opposite phases, which result in an oscillating electric field, are applied to adjacent electrodes (Fig. 10)[299].

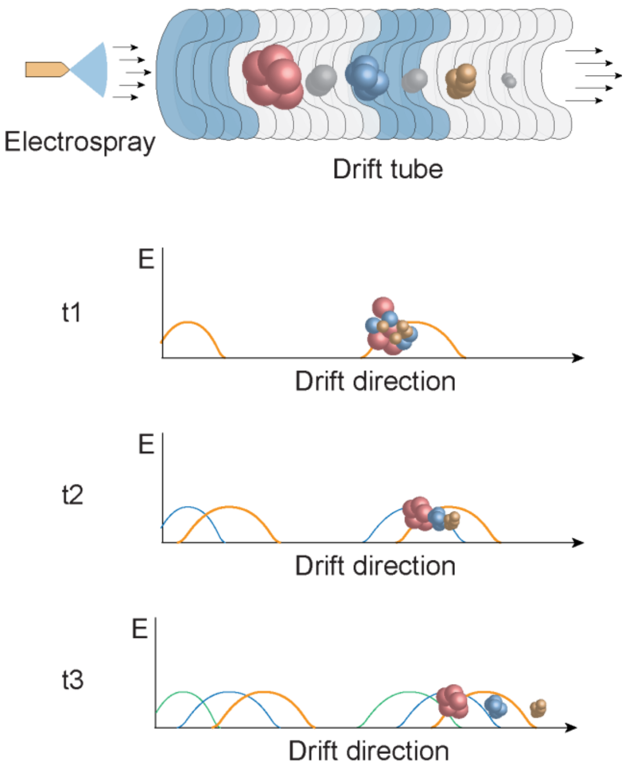


Figure 10 | Traveling wave ion mobility spectrometry. Separation is achieved by ions pushed axially into the device passing the drift tube region along an oscillating electric field. Small ions 'surf' the wave fast and pass the drift tube earlier due to their higher mobility, while large ions roll over the wave slower and pass the drift tube later due to their higher mobility (Adapted from Ref.[279]).

Ions are then separated at a fixed wave speed and magnitude. Alternation of the speed and travelling voltage wave magnitude allows to adjust mobility resolution ($\Omega/\Delta\Omega$) at a fixed drift tube length. Higher mobility ions 'get carried' by the wave and leave the drift region faster, while lower mobility ions will 'roll over' the wave, leaving the device last[276]. Even though TWIMS operates below the 'low-field limit', the separation principle differs from classical drift time IMS. TWIMS devices have to be calibrated with ions of known mobility to allow for ion CCS determination, since Mason Schamp only holds true for linear electric fields[285,300]. The effective 'wave length' was taken to the extreme by a cyclic ion mobility device introduced by Waters. Here, ions are orthogonally deflected into a 98 cm long cyclic device separating ions by traveling waves. This allows a user defined ion mobility resolution based on the time the ions spend within the device, resulting in up to 750 ($\Omega/\Delta\Omega$) resolution, albeit at a loss of sensitivity[301]. TWIMS, positioned in between LC and the mass analyzer, has been shown to be very beneficial for proteomics applications as exemplified by a reported increase of ~60% in peptide and protein identifications in *Escherichia Coli* whole cell lysates[302]. Furthermore, TWIMS positioned right behind the collision cell allows the separation of fragment ions by their ion mobility, also known as post-fragmentation IMS. Here, synchronization of orthogonal acceleration and ion mobility separation was used to increase the MS2 duty cycle, which doubled the number of protein identifications and increased sensitivity by up to 10-fold from a human cell line compared to non-synchronized TOF-pushes[303]. Although TWIMS appears to be a promising IMS technology, a number of issues, including mass accuracy distortion issues due to detector saturation and closed data formats of the vendor have prevented its widespread use in the proteomics community[302].

1.3.3. Differential or field asymmetric ion mobility spectrometry

Differential or field asymmetric ion mobility spectrometers (DMS; FAIMS) belong to the *spatially dispersive IMS* category and make use of different analyte ion mobilities in low-field and high-field operation[290,304]. They are operated under ambient pressure conditions, serve as mass filters and are positioned directly after ES when combined with LC-MS instruments. DMS devices can be constructed of two parallel electrodes with an electric field across them, while FAIMS devices have a cylindrical shape and ions are introduced perpendicular to the electric field (Fig. 11).

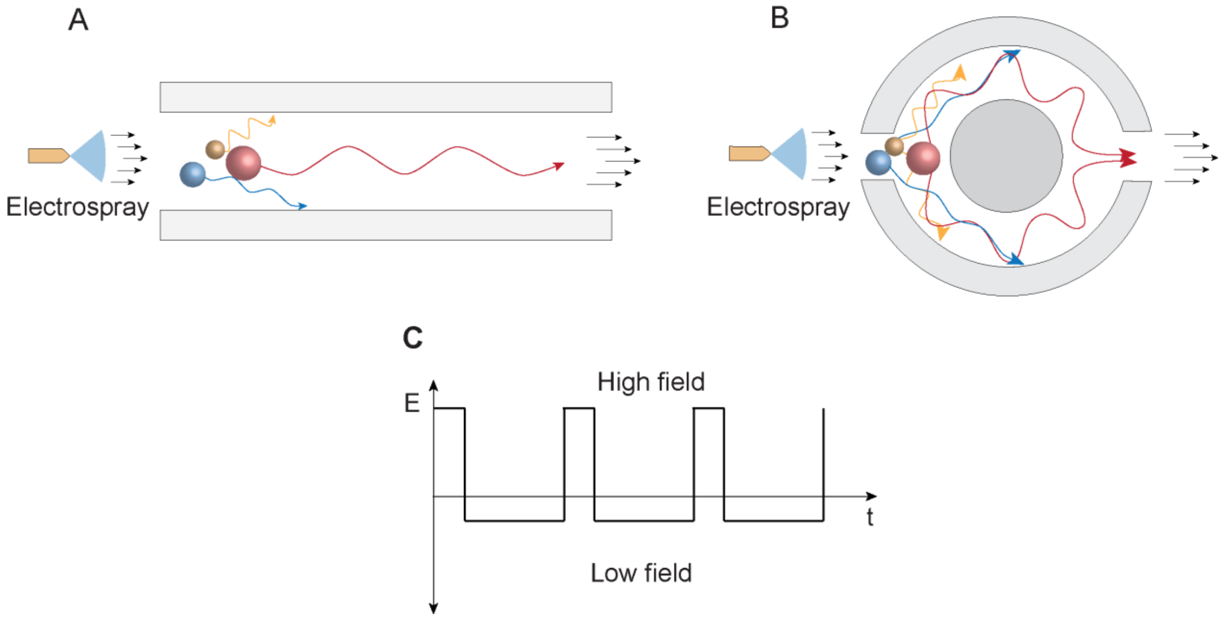


Figure 11 | Differential or field asymmetric ion mobility spectrometry. A, Differential ion mobility spectrometry setup. **B**, Field asymmetric ion mobility spectrometry setup. **C**, Ions are exposed to alternating low and high electric field strength to spatially disperse them as a function of their different ion mobilities in both fields. A DC compensation voltage is superimposed to transmit ion mobility ranges of interest[305,306].

To separate ions two voltages are applied: First, the dispersion voltage, which follows an alternating asymmetric waveform. The introduction of ions into this electric field causes them to drift towards the two electrodes at different rates, depending on whether the analyte is more mobile in the high- or low-field. The overall operation time at positive voltage is shorter than at negative voltage, while an equal 'voltage x time' product for each waveform is maintained. To counteract the drift towards either one of the electrodes, a superimposed DC voltage, also known as compensation voltage, refocuses the ions flight path through the device on a stable trajectory[276]. Ions with a high mobility travers the device quickly and need high CVs to prevent collision with the electrode, while ions with a low mobility need lower CVs to correct their trajectory, because they are less affected[276]. Adjusting the CVs results in a mass filter effect, enriching for analyte ions of interest, while chemical noise is reduced significantly. Due to its compact design and potentially advantageous effects for MS-based applications, FAIMS was recently re-introduced for the Orbitrap analyzers on high-end instruments[187,306]. Several use cases of FAIMS in proteomics have been shown since its introduction and rigorous investigations of its mass or charge state peptide filter capabilities have been published[187,306]. Gas phase fractionation in proteomics by alternating CVs can drastically increase overall proteome coverage and resulted in peptide identifications of more than 100,000 in 5 h DDA single-run analyses[306]. Furthermore, the use of a single CV optimized for proteome depth (although at compromised peptide coverage), allowed the identification of more than 5,000 proteins in as little as 21 min with DIA[187]. Even though FAIMS has been reported to have compromised sensitivity due to decreased target ion transmission in the early days, it has recently been shown to be useful for ultra-high sensitivity measurements down to the single cell level with protein identifications of up to 1,000 per cell[307,308]. This can most likely be explained by the efficient filtering of singly charged chemical noise by FAIMS and the consequently increased fill times and thus signal-to-noise level for peptides. Although many applications reportedly benefit from FAIMS, this technology has so far not been sufficiently robust for wide-spread applications in our laboratory. Unless several CVs are combined, it selectively loses many of the peptides present in the sample. Furthermore, since DMS and FAIMS operate above the low-field limit, CCS values cannot be determined, which leaves only the ion filter function. Still, it will be interesting to see in the future if fast CV-switch times can be achieved on a chromatographic time scale in a way that eluting peaks convoluted by several peptides benefit from the mass selection capabilities of FAIMS, which could increase proteome depth.

1.3.4. Trapped ion mobility spectrometry

Trapped ion mobility spectrometers belong to the class of *confinement and mobility-selective release* devices and TIMS is one of the latest developments in the field of IMS[279]. It was developed by Melvin Park and colleagues at Bruker and is operated in the first vacuum stage of the mass spectrometer when combined in a hybrid mass spectrometer[309,310]. TIMS devices reverse the principle of DTIMS by pushing the ions with a carrier gas into a weak electric field of increasing strength[290]. The latest TIMS devices consist of SRIGs and can be divided into three main regions, which are the entrance funnel, mobility analyzer section and exit funnel. The SRIG is furthermore divided into quadrants and a quadrupolar rf-field is applied, which radially confines ions entering the analyzer[279]. As ions are injected into the device by a continuous flow from the ion source, they are first focused by an electrodynamic funnel before entering the mobility analyzer[311]. A longitudinal increasing DC field is applied across the ring electrodes, which separates the ion beam into dense packages as a function of their size-to-charge ratio, or mobility, as they are axially trapped in regions of the analyzer, where the drag of the incoming gas flow is compensated by the electric force[305]. Upon decreasing the longitudinal DC potential across the analyzer, ion packages of the same mobility elute sequentially from high to low size-to-charge ratios, (or low to high mobility), refocused by an electrodynamic exit tunnel (Fig. 12).

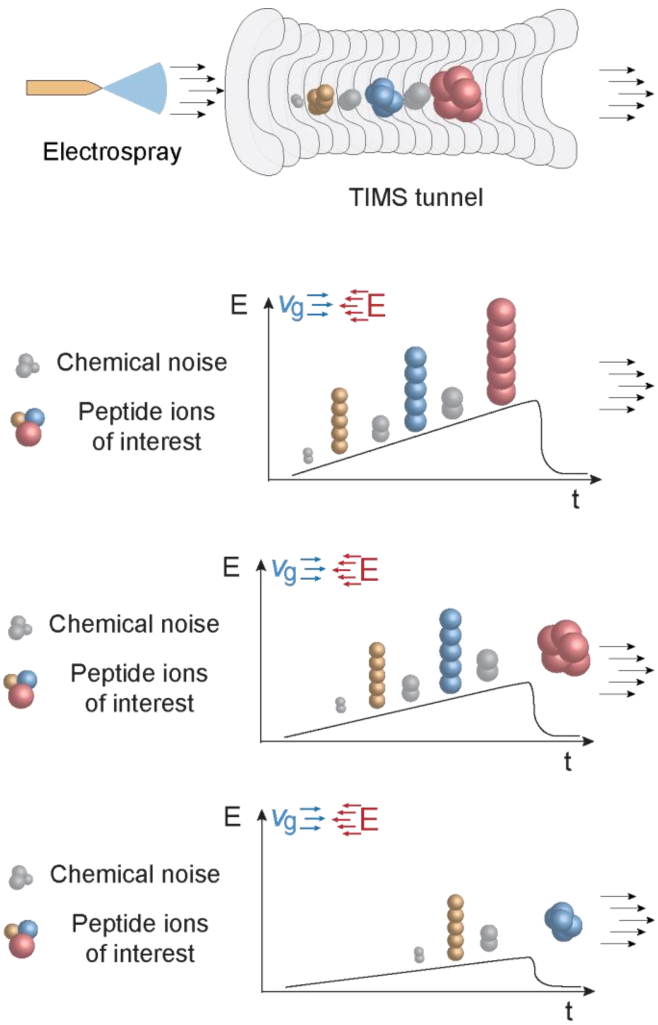


Figure 12 | Trapped ion mobility spectrometry. Sample analyte ions that are pushed into the TIMS tunnel by a steady-flow carrier gas (v_g) experience an increasing electric field strength (*E*) composed of a static component, which is superimposed on a quadrupolar rf-field to confine the ions trajectories. Ions are axially trapped at distinct ion mobility positions along the field gradient where the drift force is compensated by the electric field force. Dense ion packages are released upon lowering the electric field strength (*E*) (Adapted from Ref. [197]).

The resolution of TIMS depends on the ramp time - the time which is allocated to scan out the ions from the tunnel. The largest separation effect happens just at the exit of the TIMS device and can exceed a resolution of more than 250 Ω/ΔΩ[311,312]. Typical ramp times in proteomics and lipidomics analyses, which I have used in all articles of this thesis were between 50-100 ms and resulted in ion mobility peak widths less than 2 ms[195,216]. TIMS has been used to separate structurally very similar ions, where ions of the same charge state separate into 'charge state families', which appear along the mass-mobility correlation line. Interestingly, the MS2 analysis of target ions can yield virtually noise-free spectra **(Articles 1, 7)**. Recently, Melvin Park and colleagues also implemented a dual version of the TIMS device. This is possible due to the small footprint of the device itself of only about 0.8 cm inner diameter and 10 cm length[197]. The dual TIMS device circumvents one of the bottlenecks in IMS, pushing the duty cycle up to 100 % if operated in sync[198], where the first TIMS allows the accumulation of all incoming ions of the continuous ion beam, while the second TIMS device scans out the previous batch of ions. Furthermore, since TIMS can be operated within the low-field limit and the electric field potential within TIMS is linear, the *Mason-Schamp equation* allows, in principle, the direct calculation of CCS-values from the analysis[285,313]. With proteomics applications in mind, TIMS had been implemented into the ion path of a prototype based on the *Bruker Impact II*, just in between the ES-source and the analytical quadrupole. This hybrid MS was later marketed as *timsTOF Pro* by Bruker Daltonik. It allowed the first proof of principle implementation of a novel scan mode developed in our group and was called *Parallel accumulation serial fragmentation* (PASEF)[195].

A major focus of my PhD was to enable this instrumental setup for proteomics applications and make best use of its strengths such as ultra-high sensitivity, robustness and speed to advance bottom-up proteomics[126,186,314] **(Articles 2, 3, 5, 6, 7, 8, 9)**. Furthermore, we devised the novel scan mode diaPASEF, which combines the advantages of data independent acquisition and PASEF (Ref.[315]). I also implemented a next generation hybrid TIMS MS, which in conjunction with novel very low-flow liquid chromatography boosted the sensitivity of bottom-up proteomics by up to 100-fold. Together, these developments enabled the core of my PhD, the MS-based proteome analysis of single-cells at a depth of up to 1,500 proteins per cell – and ongoing improvements even pushed this to more than 2,000[105].

1.4. Data acquisition strategies

Mass spectrometry based proteomics has emerged to the method of choice for the in-depth analysis of proteomes in an unbiased way[65]. The number of open reading frames encoding proteins in humans exceeds 20,000 while the number of theoretically possible tryptic peptides easily exceeds 600,000 – even when not taking into account the plethora of possible posttranslational modifications[216,316]. The number of tryptic peptides to be sampled within a proteomics experiment is not only astronomically high, but it is also distributed across more than 10 orders of magnitude in many biological matrices. Furthermore, a single cell only contains approximately 150 pg of protein[39,41]. This means that sampling the proteome at depth is a tremendous challenge not only for the mass spectrometric technology but especially for the way how peptides are sampled, since it has been shown that a vast majority of the theoretically assessable peptides are simply not recorded in proteomics exeperiments[317]. Three methods of choice have developed and stood the test of time, namely *data dependent acquisition*, *targeted acquisition*, and *data independent acquisition*. Each of them has its unique way of sampling the proteome as discussed in detail below (Fig. 13).

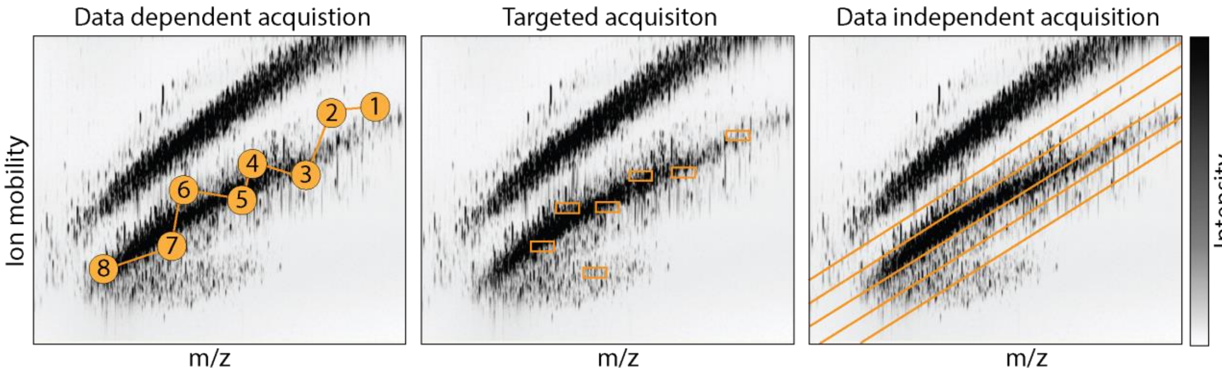


Figure 13 | Scan modes for data acquisition in proteomics on a timsTOF instrument. Left, Data dependent acquisition – The mass spectrometer selects suitable precursors as a function of intensity from MS1 survey scans for MS2 analysis (isolated precursors highlighted in orange). **Middle**, Targeted proteomics – Peptides to be targeted are constantly analyzed along the retention time and sent for either MS1 or MS2 analysis (target peptides highlighted in orange). **Right**, Data independent acquisition – The analytical quadrupole scans with a pre-defined width and window scheme across the mass range of interest after MS1 survey scan acquisition (orange boxes indicate a possible window scheme).

1.4.1. Data dependent acquisition

Data dependent acquisition (DDA) is currently still the by far most used method for unbiased proteome profiling. DDA belongs to the tandem MS category, where ions are subjected at least twice to MS analysis. A typical DDA experiment begins with the acquisition of a MS1 or survey scan in a defined m/z range (e.g. 100-1,700 m/z), which gives an overview of the peptides currently eluting from the liquid chromatography column into the MS. Then, based on the precursor ions identified in the survey scan, the quadrupole is moved to the m/z of interest for isolation with a narrow window (< 2 m/z) and fragmentation with a precursor-specific collision energy – This DDA scan cycle was initially executed manually[318]. DDA methods are also known as TopN methods as the instrument automatically determines a list of precursors from the MS1 ranked by their respective intensity values on the fly, followed by isolation and fragmentation of the N most intense precursors. This isolation strategy drastically increases the likelihood of identifying the MS2 fragmentation spectra, since the signal will be optimized. As TopN methods have to fit the chromatographic time scale by sampling the peak on MS1 level sufficiently often for reconstruction and quantification, the number N of precursors per TopN cycle has to be restricted. To prevent constant resequencing of abundant precursors, the concept of 'dynamic exclusion' was introduced[319]. Here, precursors, which have been subjected to MS2 sequencing, will be added to an 'exclusion list' for a distinct time (usually 30-40 sec) so that low-intensity precursors eluting within the same time-frame have the chance to also be subjected to MS2 analysis. Furthermore, singly charged species are usually excluded from MS2 analysis in bottom-up proteomics since trypsin cleavage results under normal conditions in at least doubly charged peptide species[79]. All these concepts optimize the MS analysis time and increase the likelihood of peptide identification with the goal to increase overall proteome depth. The concept behind DDA makes it readily compatible with label-free and label-based quantification/multiplexing at the MS1 level and also with MS2 label-based quantification strategies[247,256,264,270]. Due to the straightforward data interpretation, it is the method of choice to benchmark the intrinsic capabilities of MS instruments. It also scales well with the speed, sensitivity, and dynamic range, yielding a ground truth for instrumental performance.

However, there are also many limitations to this scan mode. First, successful identification of MS2 spectra benefits from the isolation and fragmentation of only one dominant target precursor species[317]. Too many co-isolated peptides will inevitably complicate the MS2 spectrum, which results in so-called chimera and decreases the likelihood of unique b- and y-ion series assignment. In consequence, peptide identification may be comprised. Still, algorithms like the 'second peptide search' can make use of chimeric spectra to re-search them after a first successful round of target precursor identification and *in silico* removal of those fragments[218]. MS3 analysis is another option to alleviate this issue and works by isolating fragmented ions from MS2 for a second round of fragmentation termed MS3. Here, the idea is that convoluted fragmentation patterns in MS2 are distributed across the full scan range, which allows for a 'pure' second round of isolation followed by fragmentation. This concept is especially successful for multiplexed analysis by tandem mass tags, however it also increases overall analysis time, and therefore comes at the expense of proteome depth and is currently incompatible with very short LC gradients[262]. Furthermore, DDA is a semi-stochastic scan mode, which means that the number of missing values will increase across many runs, especially in the low-intensity range[272]. This phenomenon is mainly due to speed limitations of MS platforms. Orbitrap-based analyzers achieve a DDA sequencing speed of up to 40 Hz at declining sensitivity. In contrast, we showed in **Article 1** that the timsTOF platform in conjunction with the *parallel accumulation serial fragmentation* (PASEF) scan mode can achieve a sequencing speed of more than 100 Hz without comprising measurement sensitivity[186,187].

In any case, the semi-stochasticity of DDA limits the quantification of proteins and hence the downstream interpretation of the data, because protein expression outputs have to be filtered stringently for high data completeness while remaining missing values have to be imputed by instrumental noise. This can hamper insights into biology or even be misleading. The *matching-between-runs* (MBR) algorithm aims to alleviate this problem by high-confidence transfer of re-calibrated and aligned MS1-level features, which have been identified by MS2-analysis in other runs within the same experiment[218]. This principle helps a lot with the above issues and was taken to the extreme with the advent of the *BoxCar* scan mode[320]. Still, it has been shown that MBR can inflate FDRs, which has only recently been addressed by a new algorithm for FDR-controlled feature transfer[321,322].

1.4.2. Targeted proteomics

In contrast to data dependent acquisition, which is used for the unbiased proteome analysis, targeted proteomics strategies aim to detect and quantify a single or a small set of predetermined fragment ions from precursor ions that are anticipated, but not necessarily detected in survey scans[323]. The relationship between a precursor ion and a specific fragment ion is termed a 'transition', while a targeting experiment itself has been called an 'assay'. Targeted proteomics approaches were originally used for the quantification of small molecules such as metabolites or drugs. They emerged as a promising technique for the precise and accurate quantification of target peptides with known fragmentation properties in complex backgrounds[324,325]. Targeted proteomics experiments were mainly performed on triple quadrupoles (QQQ), where Q1 acts as a mass filter, Q2 as a collision cell and Q3 to isolate fragment ions of interest[324,326]. Targeted proteomics assays can be divided into selected reaction monitoring (SRM), multiple reaction monitoring (MRM) and parallel reaction monitoring (PRM). SRM assays isolate and quantify a single transition, while MRM assays use several. PRM is an MRM assay performed on an Orbitrap mass spectrometer, which allows the readout of multiple transitions in the Orbitrap in parallel (as always the entire MS2 spectrum is recorded). The analysis of multiple fragments at high resolution drastically increases assay specificity compared to triple quadrupoles[327]. Furthermore, the presence of a C-trap in Orbitrap instruments in conjunction with sharp quadrupole isolation windows allows for long accumulation times to increase sensitivity of the assay several-fold[328].

Targeted proteomics has long been developed with the hope to make it a key technology to test biological hypotheses, reproducibility and validate biomarkers by scoring changes in their abundances in large sets of clinical samples[323]. In a clinical context, one can think of targeted MS as a single or multiplexed ELISA assay, but with much higher specificity, since it does not suffer from possible cross-reactivity with similar target epitopes. Repeatability and reproducibility can be excellent, even across laboratories for very challenging biological matrices like blood plasma[329]. Targeted proteomics can also be used for absolute quantification of peptide and proteins in conjunction with isotopically labeled reference peptides that are physicochemically identical to the 'light' endogenous peptides, but it is very difficult to guarantee a desired amount of the standard itself[330,331].

Many peptide/fragment ion features have to be considered to successfully set up a targeted assay. Peptides have to be selected by their proteotypicity, which means that they should be unique signature peptides of the target protein to be quantified[332,333]. Furthermore, their physicochemical properties are crucial, since signal response of different peptides can vary by as much as 100-fold[334]. Also, fragment ions should be selected based on least signal interference and robustness[335]. Ideally, properties of target peptides would have to be established only once for an instrument type and can be used steadily, which would make the transfer of assays between laboratories possible[329,336]. However, this can be very difficult in practice where even the same instruments in the same laboratory may need optimized assays. Another important factor to take into account is the instrument-specific dwell time, which is the time spent on actually acquiring the transition. Ideally, one wants to acquire as many transitions as possible for a particular peptide to achieve highest sensitivity. This requires a substantial dwell time. Consequently, very long cycle times and hence an insufficient number of data points to reconstruct the chromatographic peak at high resolution have to be avoided as they would compromise quantitative accuracy. This generates a tradeoff between limit of detection and number of transitions within an SRM experiment, and ultimately limits multiplexing.[334]

One of the newest methods in targeted proteomics is prm-PASEF, which is the implementation of multiplexed targeting on the timsTOF Pro[337]. It takes full advantage of the PASEF principle and the dual TIMS tunnel design, where several peptides can be targeted per ion mobility scan. This increases multiplexing capabilities at a fixed mass resolution of up to 60,000[195,198]. Furthermore, selectivity and sensitivity are drastically increased, since chemical noise and other analytes are distributed well across the ion mobility ramp, while incoming ions are focused into sharp ion mobility peaks.

As the minimum knowledge about the target peptide are retention time and m/z, the first of which is subject of variation between experiments, setting up target assays is very challenging. Real-time recalibration of these parameters, as exemplified by MaxQuant.Live, alleviates these issues and has allowed the global targeting of more than 25,000 peptides in only 120 min[265]. Furthermore, the tedious creation of libraries for targeting assays can likely be replaced by deep learning predicted libraries in the near future, which will make the experimental design of targeted proteomics assays less time consuming and more robust[216,227,228,230].

1.4.3. Data independent acquisition

Even though data dependent acquisition has evolved as the gold standard for unbiased proteome analysis, it has two major limitations. First, it is limited by analytical reproducibility due to its semi-stochastic sampling approach and the high complexity of the sample. Second, MS2 fragmentation can be triggered not exactly at the chromatographic elution apex, which can result in low-intense and potentially uninterpretable fragment ion spectra – especially for low abundant peptides. In contrast, targeted proteomics aims for highest data completeness and sensitivity, but is limited in terms of proteome coverage. Data independent acquisition (DIA) sets out to combine the advantages of both worlds. It aims for virtually complete data matrices at highest sensitivity and depth across hundreds of samples[269]. The term DIA was first mentioned in a publication from 2004 (Ref.[338]). Here, a method was proposed for the sequential isolation and fragmentation of precursor windows at 10 m/z width of a specified precursor range in an ion trap instrument. Peptide precursor quantities were reconstructed from the fragment ion spectrum level instead of from the full scan. Furthermore, the data structure consisted of a full 3D-record (retention time, fragment ion intensity, m/z) of fragment ion spectra across the run with intertwined full scans, only limited in resolution by the cycle time. It also promised a very high selectivity, quantitative robustness and data completeness. However, the mass spectrometric resolution and mass accuracy was orders of magnitude lower than possible today. A principal challenge of this scan mode is the loss of the direct relationship between precursor and fragment ion series due to highly convoluted and multiplexed MS2 spectra[339]. In the early days, DIA data were analyzed in analogy to DDA data by direct spectral comparisons with limited success[338]. A breakthrough for the analysis of this data type was the targeted extraction of precursors in a similar way as in SRM assays using prior knowledge of the peptides to be expected and their respective peptide query parameters (PQP) summarized in 'spectral libraries'. Spectral PQP information used for scoring during analysis originates either from project-specific or community-based single-shot and fractionated DDA measurements[269,340,341]. This approach was called 'peptide-centric' as opposed to 'spectrum-centric', where experimental spectra are compared to all theoretical *in silico* generated spectra within the search space. It is also marketed as 'SWATH-MS' by Sciex[269]. Since the initial inception of DIA, several other approaches or modifications to the method have appeared, including different windows sizes ranging from fragmenting the full scan at once down to windows of only 2.5 m/z[215,269,342,343]. Also, multiplexed (MSX) or intra-run offset window approaches as well as dynamic window width adjustment according to eluting precursor density were suggested and promise to increase selectivity[344,345].

Due to the highly complex data structure of DIA, rather slow instrument scan speeds and low resolution, its performance was for a long time inferior to DDA approaches. This changed with the development of more sophisticated software and faster instruments resulting in high proteome depth exceeding 4,000 protein identifications by SWATH-MS across many laboratories[346,347]. In particular, the implementation of DIA on the Q-Exactive Orbitrap instrument generation allowed fast acquisitions at increased resolution and much higher ion signals at the MS2 level due to relatively long injection times into the Orbitrap. Accompanying software for the analysis of these data led to very high performance[345]. Rigorous optimization of scan parameters was subsequently shown to exceed 8,000 and 11,000 protein identifications in single LC-MS runs at a data completeness of close to 100 % and robust quantification from cell lines[271,348]. In DIA, this level of performance has been so far only reached by highly sophisticated methods like 'BoxCar', which also make use of optimal Orbitrap filling[264]. Since quantitative robustness and high data completeness are essential in large clinical sample cohort studies, DIA was consequently successfully applied to blood plasma and cerebrospinal fluid studies exceeding a sample count of 1,000 (Ref.[89,243]). Due to the fact that DIA data contain a full run-record of fragment ion spectra, it is also very attractive for the analysis of posttranslational modification and site-localization, which has already been achieved but is still an active field of research[124,349].

Novel deep neural network strategies have recently been used to distinguish real signals from noise and new quantification and interference correction strategies have been developed that outperform classical software solutions in terms of protein recovery, data completeness and quantitative robustness[350]. Another very interesting development is the use of a scanning quadrupole[351]. Here, in contrast to stepping through the m/z region of interest, the quadrupole scans through a defined precursor mass range with a fixed isolation width. This means that fragment ion signals appear and disappear over time according to precursor isotopes entering and exiting the scan window. This additional dimension on top of retention time, m/z and intensity, promises to greatly add selectivity to traditional stepped DIA methods[340]. SWATH-DIA just recently was merged with this approach on a triple-quadrupole time-of-flight instrument, which now allows ultrafast DIA analyses with cycle times of as low as 280 ms in conjunction with short high-flow liquid chromatography (800 µL/min) but using a high sample load. This resulted in close to 2,000 protein identifications in 30 sec LC-MS runs from a human cell lysate and was also applied to clinical samples[144]. As DIA applications provide ever increasing numbers of reported proteins in studies comprising thousands of samples in drastically decreasing analysis time, sophisticated FDR models and benchmarks are required, which is still an active field of research[352]. Furthermore, since the computational prediction of PQPs by deep learning is becoming more and more accurate, it is only a matter of time that experimentally generated spectral libraries will be replaced by *in silico* generated ones[227,228,230,353].

In the course of this thesis, we implemented DIA on the *timsTOF Pro* and demonstrated that we can acquire very deep proteomes from long and short gradients at up to 100 % ion utilization due to the dual TIMS setup **(Article 6)**[198]. This is in stark contrast to traditional DIA, which may utilize only ~1.5 % (assuming the scan range is 800 m/z and quadrupole isolation width is 12.5 m/z). The implementation of diaPASEF and summing up of several consecutive diaPASEF scans within the same cycle was crucial to increase the signal-to-noise for ultra-high sensitivity applications - down to the level of true single-cell proteomes **(Article 7)**[105]. Furthermore, we showed that the prediction of CCS values is already accurate enough to replace experimentally determined CCS values with *in silico* generated ones without any compromise in diaPASEF data quality **(Article 5)**[216]. Currently, we are exploring possibilities of scan modes akin to scanning quadrupole implementation on the *timsTOF Pro*, however, using the correlation between peptide m/z and ion mobility to the full extent. As described above, this adds a fifth dimension to the inherent four dimensions (retention time, m/z, intensity, ion mobility), increasing selectivity even further while keeping up to 100 % ion utilization (assuming that the ion release from the TIMS ramp is in sync with the quadrupole scanning and that the quadrupole only performs a single scan per cycle).

Finally, since instrument electronics, communication interfaces to the mass spectrometer and software are becoming faster and faster, real-time data-acquisition software will become a reality in the future. Ultimately, the boundaries for distinct scan modes will disappear and 'Omni-methods' will emerge, which will allow very dynamic scan decisions during the run, as well as live quality control and data processing, followed by direct intervention to correct for these. 'MaxQuant.Live' **(Appendix)** is one of the latest software solutions, which are poised to make the dream of a very intelligent mass spectrometer soon come true[265,354].

1.5. Spatial and single-cell omics analysis

The discovery and naming of 'cells' dates back to 1665 when Robert Hooke inspected bottle cork slices under a light sheet microscope and realized that it was organized in well-ordered enclosed patterns, which reminded him of honeycomb or monastic cells[355]. Almost 200 years later in 1852, Robert Remak reported that every cell arises from a pre-existing one and in this way represent the minimum unit of life[356]. This cell division manifests as mitosis, following a hard-coded sequence of intermediate states. Further technological advances in microscopy allowed the description of cells in every greater detail, highlighting differences in morphology, location and spatial architecture with regards to how they are embedded in a tissue context. A typical human cell contains an estimated 6 billion base pairs of DNA, 600 million bases of mRNA and 150 pg of protein at a density of 300 g/l distributed across only 3,000 μm^3 volume (Ref.[39,357]). According to the 'central dogma of molecular biology', each cell constantly transcribes mRNA from the 'hardcoded' genome level and mRNA is in turn translated into proteins, the main drivers of cellular function[65]. Even though the analysis of samples in bulk, comprising a pool of cells, reveals impactful insights into health and disease, information about the individual contribution of cells is lost[104]. In a biomedical context, recent approaches have shown that the spatial context and the arrangement of cells, especially within the tissue proteome, is crucial to reveal direct treatment options[93,358]. This clearly indicates that cellular morphology and behavior correlates well with dynamic changes on the molecular level and that every cell is unique in its makeup to some degree. It occupies an exclusive position in space, carries unique differences in its genome and distinct changes in gene expression are induced by its environment.

Many imaging-based technologies like *fluorescent-activated cell sorting* (FACS) have emerged to reveal cellular phenotype heterogeneity, but are limited to a fixed number of fluorophores due to spectral overlap[359]. In addition, many advanced methods can track subcellular target proteome changes in individual cells by a combination of antibody labeling and live-cell imaging[360]. Spatially resolved *fluorescence in situ hybridization* (FISH) techniques enable tracking and counting single mRNA molecules in the cell[361]. Even though these technologies revealed novel biology, they do not create an unbiased picture of the cellular genome, transcriptome and proteome. It is therefore of highest importance to enable the analysis of all three layers of the *central dogma of biology* in an unbiased *omics* approach at single-cell resolution to fully understand the role of each single cell within a tissue and reconstruct their contribution to health and disease based on its heterogeneity.

1.5.1. Single-cell sequencing

Single-cell genome sequencing addresses the first layer of biological information and is by now applied to many research areas. For example, it is used to dissect the composition of microbial ecosystems and reveals its *dark matter* in metagenomics analyses[362]. It is also applied to detect gene mosaicism and allele frequencies in multicellular organisms to understand health and disease[363]. This is a tremendous challenge, since genes are mostly encoded as a single nucleotide stretch across the whole genome of the cell (not taking into account the presence of multiple chromosome sets and gene doublings). One of many modified techniques build on single cell genome sequencing and one, called ATAC-seq, can be used to study chromatin structure and epigenetics[364,365]. Here, a hyperactive transposase inserts primers into open chromatin, which allows the downstream amplification of accessible genome stretches.

Even though the genome represents the blueprint for every organism, it is rather static and does not necessarily allow for fast and dynamic changes in response to environmental changes. In contrast, the RNA-level does rapidly change and therefore single-cell RNA-sequencing is a very active field of research[366]. It allows to quantitatively describe gene expression levels, which is also very important in the context of proteomics, since it represents the most direct connection to the abundance levels of proteins. Current estimates suggest that 5,000-15,000 different genes are transcribed in a typical mammalian cell[357].

The feasibility of single-cell transcriptome analysis (scRNA-seq) was first demonstrated on only seven single cells in 2009 (Ref.[367]) where the aim was to elucidate changes in early embryonic development. Since then, many protocols have been published focusing on single-cell isolation automation, volume decrease in sample processing, chemistries to capture and amplify mRNA, increased multiplexed sequencing capabilities and reduced costs[368–372]. This scaled the number of single-cells routinely covered in scRNA-seq projects to several hundred thousand cells and even to whole bodies by now[373,374]. Single-cell sequencing was therefore justifiably selected to the method of the year by Nature Methods in 2013 (Ref.[375]). The main enabling technologies for the analysis of single-cells were the polymerase-chain reaction, cDNA library generation, large-scale combinatory nucleotide synthesis, and next generation sequencing, allowing for massively parallel sequencing of several billion short reads[28,376,377].

All scRNA-seq protocols follow the same scheme, starting with single-cell isolation and cell lysis[378]. Reverse transcription into cDNA and combinatorial nucleotide barcoding for quantification and cell-specific transcriptome labeling are the next crucial steps. Here, poly-T sequence primers designed to bind the poly-A tail of mRNAs are used for highly specific priming, followed by second strand synthesis and several rounds of full cDNA amplification. Poly-T primers can also carry short unique combinatory barcodes specific for each cell if cells will ultimately be pooled for sequencing. Last, the amplified fragments, which are now representative for the single-cell transcriptome and also called 'libraries', are fragmented, sequenced by NGS and aligned to a reference genome.

scRNA-seq techniques can be differentiated by the quantification approach and initial sample processing strategy. Quantification strategies fall into full-length sequencing methods, which allow for sequence variant calling and 3'-tag counting methods that predominantly cover the 3'-end of the mRNA. 3'-tag count approaches use so called unique molecular identifiers (UMI), which alleviate the amplification bias inherent to other techniques[379]. UMIs are unique short nucleotide sequences attached to the poly-T primer, which results in a unique barcode of every single transcript molecule – this means that the initial copies of the same transcript will each obtain a unique UMI and can be distinguished in the downstream analysis. Amplification bias is then reflected by a non-uniform presence of the UMI of each respective transcript and only unique UMIs are counted to determine the initial transcript copy number. Due to the improved robustness of this approach, the latest generation sequencing approaches like *10X Genomics* are exclusively UMI-based. Furthermore, template switching oligos have been introduced, which bind to the 5'-poly-guanidine cap of the mRNA and allow higher 5'-end sequence coverage[368,369]. Sample processing strategies are similarly divided into those based on 384-microwell plates and those using microfluidics chips[368,370,380].

Another very important factor that is specific for each protocol is its sensitivity. Sensitivity is a measure for the detection limit of single-cell mRNA and consequently also reflects the amplification chemistry and capture efficiency of each assay[381]. The more sensitive the methods are, the less *dropouts*, or zero entries for each gene are observed. Dropouts can either reflect a biological truth, since a gene is not expressed with even a single mRNA copy at the point when the experiment was performed, or the gene is not expressed at all - or the number of mRNA copies was simply too low to be captured efficiently. It is conceptually interesting – also in relation to my single-cell proteomics work - that some of the latest single-cell RNA sequencing technologies with very high overall sensitivity still report dropouts, which suggests that most of these reflect true biology[382]. This observation is also in agreement with the notion of transcriptional bursts[383].

Due to the accessibility and commercialization of single-cell RNA sequencing in the last years it has developed into an increasingly routine technology. It has allowed the study of circulating single tumor cell transcriptomes, direct gene expression differences and cell lineage characteristics[368]. Remarkably, the reconstruction of dynamic cell lineage branching across developmental stages, *RNA velocity* calculations, which is the rate of gene expression change for an individual gene at a given time point based on the ratio of its spliced and unspliced mRNA, and much more has been realised[384,385]. scRNA-seq has also proven itself in the analysis of many projects relevant to human health and disease[386,387]. The applications of this technology seem to be unlimited and has therefore been recognized with the breakthrough of the year award in 2018 by *Science* for tracking early development at the single-cell transcriptome level. It has also been combined with large-scale combinatory CRISPR screens and phenotypic readouts on the single-cell level in a technology termed perturb-seq[388]. Furthermore, several spatial scRNA-seq techniques are emerging, which reveal transcriptome dynamics, incorporating an additional dimension[389,390]. This approach was again awarded with the breakthrough method in 2020 by *Nature methods*[391].

The analysis of large-scale scRNA-seq data and the plethora of available techniques call for standardization in the downstream analysis to keep results consistent and reproducibe[392]. This is currently an important research topic and first tools have emerged that are already widely used by the community[393]. Further bioinformatics challenges in scRNA-seq are systematic biases introduced by technical differences between experiments, also known as batch effects[394]. This is especially of high interest for the global inter-laboratory effort called the *Human Cell Atlas initiative*, which aims to create comprehensive reference maps of all human cells and demands advanced data integration strategies for very large data sets, and for the *European LifeTime initiative* aiming for multi-omics integration of single-cell data to model disease progression[395,396]. Many more computational challenges have been recognized in large-scale scRNA-seq projects, which now drive the development of next generation bioinformatics tools[397].

Given the fact that a typical human cell contains on average only tens of each mRNA the question arises to what degree these small numbers of molecules can actually play a major role in cellular regulation and if they can give rise to functional cell states defining cell types. It is conceivable that proteins are the better proxy for this type of functional analysis[357]. Furthermore, even though scRNA-seq has developed into a mature technique, which creates comprehensive cell type maps based on differential mRNA signatures, it is not a direct record of the functional molecular level that is charged with executing functions in cells and gives rise to the phenotype – this is instead level of the proteome. Currently researchers implicitly use mRNA level quantifications as a proxy for protein expression values, but often neglect that there is a plethora of translationally regulatory mechanisms in between mRNA and final protein product. At the bulk level this has become more and more evident with mRNA expression levels often correlating only weakly (R ~ 0.2-0.4) with the protein level[251,398]. Importantly, we have shown that this is also true for the single-cell level **(Article 7)**. Many one-to-one correlations of mRNAs and their cognate proteins show no correlation or even anti-correlation in particular biological processes. This demonstrates that both molecular levels yield their own information and do not necessarily recapitulate each other's characteristics. This creates a fundamental need for the development of techniques that can measure the proteome of single cells, described next.

1.5.2. Single-cell proteomics

Several technologies such as cyTOF, CITE-seq and SCoPE-MS have been developed over time to determine aspects of the proteome of single cells, since proteins directly reflect the functional layer of the cell[65].

The technique termed *Cellular indexing of transcriptomes and epitopes by sequencing*, short CITE-seq combines protein-specific antibodies conjugated with a combinatorial nucleotide tag, which allows amplification and target-specific deconvolution for single cells – just like in scRNA-seq[399]. For now, this technology is limited to cell surface proteins, but has shown tremendous potential for the characterization of immune cell populations and tumor cells. Still, it suffers from amplification biases as described above and low proteome coverage[400,401]. Fundamentally, it is currently a targeted approach and reveals only a small part of the whole picture.

Single-cell mass cytometry, short cyTOF, combines a mass spectrometric readout with antibodies labelled with heavy isotopically pure elements[402]. After target binding, residual antibodies are washed off, followed by nebulization of each entire cell and injection into an argon plasma, which decomposes it to its atomic constituents. This ionizes the heavy metals whose characteristic isotope pattern can specifically be deconvoluted for antibody targets and their abundance. Even though the number of possible targets is currently limited to about 100, its throughput is immense, since it essentially only needs a single mass spectrum per cell as a readout. A recent study comprised more than 70 target protein identifications across more than 25 million cells in human breast cancer tumors, revealing the tumor ecosystem and heterogeneity at single-cell resolution[403,404].

Even though CITE-seq and cyTOF are limited by the number of available antibodies, their throughput is comparable to modern scRNA-seq methods. Recent developments move into the direction of 'cut' cell layers, followed by antibody labeling, which has the potential to reveal a quantitative picture of proteins in each particular plane. This approach could allow the reconstruction of full spatial proteomes across many stacked layers in 3D. However, their fundamental limitation is that they do not allow the full proteome description of single cells, which can only be addressed by MS-based proteomics strategies as described above.

MS-based proteomics is unbiased in the sense that it measures all proteins within its range of detection[65]. Thus, it would be highly desirable to apply this technology to single cells if the required sensitivity and robustness could be achieved. Bottom-up MS-based single-cell proteomics is an emerging technology and can be divided into two branches that drastically differ in the way of how single-cells are introduced into the MS and also how they are quantified – either label-based or label-free[110,149]. In contrast to sequencing-based single-cell approaches, proteomics does not have the capability of amplification, and therefore has to make do with processing and measurement of the originally present protein mass. This introduces severe challenges as one can only increase proteome coverage by increasing the direct sensitivity of the hardware and by sample handling – especially in label-free quantification approaches – but has the main advantage that PCR-induced biases are omitted. Label-based single-cell proteomics was introduced in 2018 as *Single Cell ProteOmics by Mass Spectrometry*, short SCoPE-MS[110]. This approach uses TMT-based multiplexing and single cells are labeled by different TMT isotope channels, to which a so-called booster-channel comprising the protein amount of several hundred cells is added. All differently labeled TMT samples are combined before LC-MS analysis, which results in a total of several hundred cells from the booster channel plus the single cells from the other channels. This means that the total signal on MS1 level stems largely from the booster channel. Upon fragmentation TMT labeled peptides produce low molecular weight mass reporter ions at the MS2-level. The peptide will ultimately be identified from the many cells in the booster channel, while the quantification of the single-cell derived peptide is meant to happen in the low molecular weight mass reporter region. This approach was initially reported to identify more than 1,500 proteins across the total data set and around 400 proteins from each cell[110]. Despite its conceptual attractiveness in that it stacks the peptide signal, SCoPE-MS unfortunately has many intrinsic issues that have prevented biological applications at least in its current form. First, it has been shown that the more cells used within the booster channel, the more overall proteins are identified and quantified - even though the total signal from the single cells remains the same. This can be explained by 'channel bleeding effects' intrinsic to TMT and here due to the dominant booster channel, where low abundant isotopes appear in the single-cell channels and result in an erroneous contribution to the quantified signal. Second, TMT ratio compression is a serious issue not only from the 'channel bleeding effect' but also from co-isolated chemical noise or background ions, which are in the same intensity range as the low molecular weight mass reporters from the single cells. This topic has been extensively discussed and simulated, leading to the recommendation to either completely eliminate the booster channel, or to only limit it to less than 20 cells in total[405]. Software and guidelines are being developed that aim to optimize MS acquisition parameters to prevent this phenomenon as much as possible[405,406]. Third, since

TMT can only multiplex up to 16 channels in total, many plexes have to be merged to enable the analysis of large single-cell proteome data sets. However, merging different TMT data sets has in practice been difficult since it is known to inflate false-positive identifications, missing values and results in batch effects[259]. Fourth, and related to the previous point, it is crucial to properly control the FDR on PSM and protein levels due to the reported false-positive inflation when merging several TMT-experiment. Keeping in mind that single-cell proteomics studies should eventually aim for study sizes of several thousand cells, FDR models have to be stringently applied. This has already been shown to be very important in bulk proteomics studies comprising hundreds of samples in data dependent and data independent acquisition scan modes[352,407]. In SCoPE-MS, researchers tend to increase PSM FDR to up to 3 % and do not apply any protein FDR calculation[110]. This is expected to vastly increase the overall FDR in the dataset, especially in conjunction with the known false-positive inflation from merging several TMT-experiments. Due to these limitations in quantitative accuracy and false-positive inflation, researchers have increasingly turned to LFQ-based approaches for single-cell proteomics.

In label-free single-cell proteomics single cells are injected one by one into the mass spectrometer, identifying and quantifying proteins based on tryptic peptides by conservative downstream FDR control and standardized software as in other bottom up experiments. However, this 'true single-cell' approach comes with the need for a drastic increase in overall sensitivity of the bottom-up proteomics workflow. Several groups, including ours, have miniaturized the sample preparation process in 384-microwell titer plates at below 10 μL volume or even in customized chips, which allows the handling of volumes down to 200 nL[116,149,408]. Furthermore, improvements in chromatography by scaling down the inner diameter of columns to reduce the radial analyte diffusion in conjunction with decreased flow rates have been reported to increase ES sensitivity and desolvation efficiency[145,146]. Extensive optimization of MS experiment parameters has also been investigated, including the use of matching between runs. The latter did increase the number of protein identifications up to 1,000 per cell, but can inflate false-positives[308,321].

Even though proof of principle for label-free single-cell proteomics has been demonstrated, a technology that provides quantitatively accurate MS proteomics data from true single cells to answer biological questions was still outstanding – and this challenge is the main topic of my PhD thesis.

2. Aims of the thesis

The overarching goal of my PhD thesis is the development of a novel mass spectrometry (MS) platform in conjunction with novel scan modes and a versatile and robust liquid chromatography (LC) platform, which overcomes current sensitivity and robustness limitations in MS-based proteomics. In the course of my PhD, I demonstrated how this technology benefits the high-speed and ultra-high sensitivity analysis of large-scale proteome studies in basic biology and biomedicine. This culminated in the first of its kind robust label-free MS-based single-cell proteomics platform and its application to spatial tissue proteomics. I also investigate the vastly underexplored 'dark matter' of the proteome, highlighting novel microproteins that contribute to human cellular function.

In the first project, a close collaboration with the research and development department at Bruker Daltonik, we aimed to develop a novel quadrupole time-of-flight (qTOF) platform in conjunction with a dual trapped ion mobility spectrometry (TIMS) device for proteomics applications. In the dual TIMS analyzer all incoming ions can be stored in the first TIMS, while ions within the second TIMS can be manipulated in their trajectory to enable an up to 100 % duty cycle. By synchronizing quadrupole switching events with the precursor elution profile from the TIMS device, which we call 'parallel accumulation – serial fragmentation', or short PASEF, we showed that this principle multiplies sequencing speed to up to 100 Hz at full sensitivity in routine proteomics applications. Furthermore, I demonstrated the capabilities for high-throughput and high-sensitivity proteomics **(Article 1)**.

Next, we aimed to apply this novel MS platform to projects which are in unique need of high-sensitivity and high-throughput. In a collaboration with the laboratory of Michele Solimena from the Paul Langerhans Institute at the TU Dresden, we aimed to dissect pancreatic islet cell proteomes across the diabetic continuum from living metabolically profiled human subjects. Pancreatic islet pools, isolated by laser capture microdissection, were subjected to high-sensitivity proteome analysis on the novel TIMS-qTOF platform revealing a substantial heterogeneity in islet isolates from diabetic subjects **(Article 2)**.

Furthermore, we aimed to use the high-throughput and robustness capabilities of the TIMS-qTOF platform to analyze the human protein interactome under endogenous expression levels and integrate the data with substantial imaging and localization studies. Finally, we combined the TIMS-qTOF MS with the *EvoSep One*, a very robust LC platform to measure more than 1,300 protein pulldowns in triplicates (~4,000 runs) in less than three months with no major interruption. Due to the high sensitivity and throughput capabilities of this setup, we were able to decrease the traditionally needed measurement time and initial sample amount by more than 10-fold **(Article 3)**.

Next, we aimed to translate the principles of TIMS and PASEF to the field of lipidomics, highlighting the same beneficial effects as described for the proteomics implementation. Furthermore, we show that the CCS dimension especially benefits lipidomics with regards to lipid class separation in the gas phase and isomer identification **(Article 4)**.
We also explored the collisional cross section (CCS) dimension, a nearly intrinsic property of biomolecules that the TIMS offers 'for free'. We set out to create a very large tryptic peptide CCS compendium from several laboratory organisms comprising more than 1 million CCS values, which enabled us to train a deep learning model for the prediction of CCSs for any tryptic peptide. We also described the fundamental behavior of peptides in the gas phase and the contribution of amino acid sequence composition as well as positional constitution **(Article 5)**.

The main aim of my PhD was the development of a robust ultra-high sensitivity LC-MS platform for the high-throughput analysis of single-cell proteomes to complement the achievements of single-cell RNA-sequencing in basic research and biomedical applications. We set ourselves the goal to increase the overall sensitivity of our LC-MS technology by up to 100-fold, which would enable us to inject single-cells – one by one – and quantify their proteomes in an unbiased manner. Together with the Bruker research and development team, we developed a brighter ion source and improved ion transmission through the TIMS-qTOF instrument, which increased the overall signal by more than 4-fold. Furthermore, we developed robust and very low flow gradients on the *EvoSep One* LC, which increased the signal by more than 10-fold compared to the standard microflow gradients at more than 40 single-cell measurements per day. We also coupled sample preparation to LC, which ensures highest single-cell peptide transfer efficiency and developed a novel scan mode combining the advantages of data independent acquisition and PASEF, which in principle allows up to 100 % ion utilization, while standard data independent acquisition methods usually make use of less than 3 % **(Article 6).** Altogether, this increased the sensitivity of our LC-MS technology by up to 100-fold and enabled the quantification of SCP to a depth of up to 1,400 proteins per cell. Comparisons to single-cell RNA sequencing data revealed fundamental insights such as that single cells have a stable core proteome, but Poisson noise-dominated transcriptome, emphasizing the need for both complementary technologies. **(Article7)**

Building on the capabilities of the SCP technology, we aimed to elucidate the image-guided spatial and cell-type resolved proteome in whole organs and tissues from minute sample amounts. We combined tissue clearing of rodent and human organs - rendering them fully transparent, unbiased 3D-imaging, pathological target tissue identification, isolation and MS-based unbiased proteomics. This revealed early-stage β-amyloid plaque proteome profiles in a familial Alzheimer's disease model **(Article 8)**.
We also aimed for the automated artificial intelligence driven isolation of pooled single-cells of the same phenotype from tissue sections. This allowed us to characterize cell-type specific spatial proteomes of cancer tissues, which would have otherwise been obscured in bulk measurements **(Article 9)**.

Last, we aimed for a systematic elucidation of pervasive functional translation of noncanonical human open reading frames, also known as the 'dark matter' of the proteome. Here, we combined state-of-the art ribosome profiling, CRISPR screens, imaging and MS-based proteomics. We highlight the unbiased analysis of small novel proteins and prove their physical existence by mass spectrometry as HLA-presented peptides, essential interaction partners of protein complexes and cellular function. **(Article 10)**

3. Articles

3.1. TIMS for high-sensitivity and high-speed proteome analysis

Great strides have been made in the development of liquid chromatography and mass spectrometry platforms to enhance the speed and sensitivity of proteome measurements. Hybrid quadrupole Time-Of-Flight (qTOF) mass spectrometers, like the *Bruker Impact II* have the capability to inherently scan precursor ions very fast at a constant and high resolution of up to 35,000 at excellent sub 1.5 ppm mass accuracy in routine bottom-up proteomics experiments[196]. A major bottleneck of current qTOF instruments is that this happens at compromised sensitivity per scan. Furthermore, the peptide complexity eluting into the mass spectrometer is very high and to increase peptide coverage during analysis, the sequencing rate has to be increased, again at the cost of reduced sensitivity[317]. Equipping the *Impact II* with a so-called Trapped Ion Mobility Spectrometer (TIMS) in the first vacuum stage of the instrument developed by Melvin Park and colleagues at Bruker promises to solve this dilemma[198,310]. Here, all ions introduced into the mass spectrometer face an opposed electric field gradient and are trapped at distinct positions corresponding to their IM within the TIMS device. Ions can then be accumulated for a chosen time period dictated by the speed of the chromatographic elution time of each precursor and their overall number. Since m/z and IM values are correlated, one can synchronize the elution of the precursors trapped per TIMS scan with sub-millisecond quadrupole switching events. In this way, when decreasing the electric field strength, the quadrupole isolates concentrated ion packages sequentially eluting from the TIMS device in sharp IM peaks. This principle, termed Parallel Accumulation SErial Fragmentation (PASEF), was recently introduced by our laboratory and promises to increase the sequencing speed by more than 10-fold per TIMS scan without a decrease in sensitivity[195].

When I started my PhD, Heiner Koch, Scarlet Beck, and Florian Meier co-developed and evaluated a first full implementation of the PASEF scan mode on an upgraded *Impact II* instrument equipped with a dual TIMS-tunnel and substantial firmware upgrades, the *Bruker timsTOF Pro*[195,196,198]. After Heiner Koch and Scarlet Beck left to join Bruker, I took over their responsibilities with regards to method development and implementation of the PASEF/TIMS-qTOF combination for proteomics analyses together with Florian Meier. In our paper, we evaluated the timsTOF Pro and the PASEF scan mode for proteomics applications and showed that the predictions described in the initial PASEF paper with regards to sequencing rates were actually realized in the current implementation with an astounding fragmentation rate of more than 100 per second. Careful optimization of this setup enabled us to identify more than 6,000 protein groups in 120 min measurement time, which was on par with the best performing instruments on the market at this time. We also showed that the instrument is well suited for high-sensitivity measurements, exemplified by the identification of more than 2,500 protein groups from only 10 ng of tryptic HeLa digest and for high-speed measurements, highlighted by the identification of more than 1,000 protein groups in less than 5 min in combination with the *EvoSep One* LC system. Quantitative reproducibility and accuracy were also very high across many replicate measurements and within mixed two-proteome experiments.

Due do the advantages in speed, robustness and sensitivity proteomics of the *timsTOF Pro* and the PASEF scan mode, I applied this technology in two main biological projects, which also showcase the benefits of this setup. First, I investigated the proteomic landscape of small numbers of pancreatic islets in the context of diabetes and second, I used it to study the HEK293T human protein-protein interactome.

In the first project, a collaborative effort within the RHADOPSY consortium with the laboratories of Michele Solimena from the TU Dresden and Mark Ibberson from the Swiss Institute of Bioinformatics, we asked if we can apply multi-omics analyses including RNA-sequencing and proteomics on small numbers of pancreatic islets. The goal was the in-depth description of molecular changes of pancreatic islets in diabetes and to integrate those with clinical data and lipidomics data from matching plasma samples. A major drawback of current pancreatic islet cell investigations in this field is that the pancreatic isolates are from terminal or post-mortem donors, which have also undergone severe pharmacological treatment perturbing molecular profiles, which can result in inconsistent conclusions[409]. Furthermore, proteomics studies in this field are hampered by technological challenges in sample preparation and the sensitivity of the LC-MS setup[410]. Additionally, donors often lack in-depth clinical records, which could for example be used for differential marker correlation analysis[409]. To minimize extraneous variability and reveal a more comprehensive picture of changes due to diabetes, we isolated pancreatic islets from pancreatectomized and metabolically profiled human living donors across the diabetic continuum by optimized laser capture microdissection (LCM) and subjected them to multiomics analysis including transcriptomics, metabolomics, and high-sensitivity proteomics analysis followed by the integration of clinical data.

We show that islets from diabetics have remarkably heterogeneous transcriptome and proteome profiles, while non-diabetic controls did not. Furthermore, we identified gene sets that are already dysregulated in pre-diabetic individuals. Furthermore, we applied an ultra-high sensitivity MS-based proteomics workflow on the newly developed timsTOF Pro instrument combined with miniaturized sample preparation to the analysis of pancreatic islets isolates, identifying up to 2,000 proteins per pancreatic islet pool. We demonstrate a progressive and disharmonic remodeling of pancreatic islets cells, challenging current hypotheses of linear trajectories toward precursor or trans-differentiation stages of T2D. Differential expression analysis revealed many potential biomarker candidates for T2D diabetic islets including the downregulation of the glucose-transporter Slc2a2. We describe for the first time that the sulfonylurea receptor ABCC8, whose role is to stimulate insulin secretion when glucose levels are high, was strongly reduced in islets of T2D. This could be a confounding or disease-associated effect of pharmacological treatment with anti-diabetic antagonists targeting this receptor. We also showed that the proteomes and transcriptomes of pancreatic islet cells are in general very different ($R < 0.3$), while the glycolytic enzyme AldoB is consistently up- and Slc2a2 is consistently down-regulated in diabetic islets on both mRNA and protein levels, which could make them suitable as prognostic markers for T2D. Furthermore, we show that T2D pancreatic islets tend to lose protein mass associated with the secretory pathway, which could contribute to the inability of beta-cells to efficiently secrete insulin. Since ACADS and ACADSB, proteins involved in promoting DNA methylation and inhibition of histone deacetylases by beta oxidation products of short-chain fatty acids including butyrate, were upregulated in T2D pancreatic islets, we reasoned that histone H3 and H4 lysine acetylation could be increased in T2D pancreatic islet cells. We confirmed this hypothesis by immunostaining, suggesting a change in gene expression programs in T2D.

In summary, we applied a multi-omics study on LCM isolated pancreatic islet pools from living metabolically profiled pancreatectomized patients spanning the glycemic continuum in diabetes. This identified many potentially novel biomarkers on different molecular layers, while transcriptome and proteome analyses revealed AldoB and Slc2a2 as shared markers. Importantly, AldoB correlates well with elevated HbA1c levels in blood – the clinical gold standard in tracking blood glucose levels and for the diagnosis of diabetes. Furthermore, lipidomics revealed a systematic upregulation of ceramides in plasma levels of diabetic. Our proteomics data set turned out to be especially rich in novel information. In addition to the decrease in protein mass of the secretory pathway, our data reveal a pharmacologically induced receptor downregulation, highlighted that pancreatic islet proteomes of T2D were highly heterogeneous while non-diabetic isolates were not, and enabled us to uncover an increased histone methylation in T2D pancreatic islets, hinting an active change in their gene expression program.

Since we were not able to assign detected proteome changes to the pancreatic islet cell types in this study, we are currently following this up with our recently developed *Deep Visual Proteomics* workflow, combining ultra-high sensitivity mass spectrometry with an artificial intelligence-driven imaging workflow for cell typing combined with Laser Capture Microdissection (LCM)-mediated cell type isolation. This will clarify if our findings in this study apply to all cell types found in pancreatic islets, or if they are specific to one of the major cell types and much more.

The measurement of very large sample-cohorts like the thousands of samples from protein-protein interactome studies needs ultra-robust and high-speed LC-MS systems to keep results qualitatively and quantitatively comparable, and to measure the project in a reasonable time. While developing the *timsTOF Pro* for routine proteomics applications, we observed that its performance was extremely stable irrespective of the injected sample type across thousands of runs. Furthermore, its high sensitivity makes it attractive for very large sample cohort measurements since sample and processing reagents can be kept to a minimum, drastically reducing overall project expenses. Combining the described MS advantages with the ultra-robust LC system *EvoSep One*[150] with fast turnover times, allowed us to measure 60 samples per day without any drop in performance and at highest sensitivity. In a second major project, I demonstrated these novel capabilities in the context of a close collaboration with the laboratory of Manuel Leonetti at the Chan Zuckerberg BioHub in San Francisco. We set out to decode the interaction and localization architecture of the human cellular proteome. This was a very ambitious goal and one of the main challenges was that it required the combination of several cutting-edge technologies to succeed. First, we used CRISPR to endogenously tag more than 1,300 proteins with a split mNeonGreen construct at their endogenous locus (either on the C- or N-terminus depending on structural accessibility), preserving endogenous protein expression in the HEK293T cell line employed[128]. Separately, we expressed the matching mNeonGreen construct to allow reconstitution of the fluorescent protein, which enables us to track cellular localization by 3D live-cell imaging. Furthermore, it provides a handle for immunoprecipitation followed by LC-MS (IP-MS) analysis for protein-protein interaction studies. We automated the imaging process for localization studies in Python, which enables scalable on-the-fly computer vision on a spinning disk confocal microscope to select desirable fields of view from 96-well cell culture plates and the reconstruction of 3D protein distribution in consecutive z-slices. This resulted in the most comprehensive image collection of live-cell protein localization to date comprising more than 5,000 3D stacks - on average four to five field of views per cell line. Interestingly, more than 50 % of our tagged proteins localized to multiple parts of the cell. Most of these proteins located to both the nucleus and cytoplasm, highlighting the importance of nuclear protein import- and export-machinery for protein localization. Next, we combined the advantages of our TIMS-qTOF/PASEF and the EvoSep liquid chromatography system to perform IP-MS experiment from the 1,300 cell lines in biological triplicates, resulting in more than 4,000 runs at very high robustness[150,195]. Compared to previous interactome studies at this scale, we decreased the overall data acquisition time at least by a factor of 15 to less than 3 months total[127]. Furthermore, the sensitivity of our LC-MS setup allowed miniaturization down to only $0.8x10^6$ cells of starting material per pulldown, which enabled us to cultivate cells in 12-well plates. This reduced the required starting amount for sample preparation by more than 10-fold, allowing multiplexed sample preparation and tremendous cost reduction[127]. Also, the use of digitonin for protein extraction enabled us to effectively preserve the native structure and properties of membrane proteins during sample preparation. However, we observed that digitonin tends to gradually accumulate in the analytical column and prevented the consecutive measurement of hundreds of samples. Using the *EvoSep One* LC-system alleviated the problem and allowed us to run the whole study on a single column-emitter setup - most likely because digitonin does not elute from the EvoTip, a Stage-tip like construct, which serves as a run-specific trap column[150,411]. Our protein-protein interaction network comprised 30,293 interactions distributed across 5,271 bait-proteins. Interestingly, amino acid sequence-based analysis revealed that proteins with highly disordered and hydrophilic domains tend to systematically have more interaction partners than proteins with helical and hydrophobic domains. Next, we integrated the bait localization information from the imaging data with the interaction partner data from the IP-MS experiments and trained a machine learning model to learn associations between patterns of localization and physically interactions. Indeed, we were able to show that localization patterns contain in many cases enough information to predict molecular interaction partners.

In summary, we delineated interacting protein families and elucidated cellular organization by superimposing physical and co-localization imaging data combining CRISPR-mediated genome engineering, automated confocal microscopy, high-speed/-sensitivity MS and data science. We uncovered that most proteins interact at low stoichiometry with low spatial overlap within the cell, emphasizing their role as 'molecular glue' of cellular proteome interaction networks. In contrast, high stoichiometry interactors tend to share similar localization patterns. We also reveal that membrane-related and RNA-binding interacting protein groups segregate from the global proteome, which emphasizes the power of combining localization and IP-MS interactome data. Membrane protein interactors segregate due to the spatial organization of the cell as expected, while the finding of segregating RNA-binding proteins was surprising. However, there is a growing body of evidence that proteins and RNA can form distinct condensates in cells sometimes termed phase-transitions, supporting our finding and suggesting that RNA itself could play a role in organizing function and localization of the cellular proteome[412]. Furthermore, we made all analytical tools and data developed in this project available to the community through our interactive website *OpenCell.czbiohub.org* presenting the quantitative cartography of human cellular organization at the proteome level.

The above applications - using the *timsTOF Pro* alone or in combination with the *EvoSep One* platform - proved its suitability to high-sensitivity, high-speed, and ultra-robust proteomics. Besides the analysis of the human interactome, we also applied this setup to *Saccharomyces cerevisiae* interactome mapping (manuscript in preparation) using every single known gene protein product as a bait – and approaching the ideal of a complete protein-protein interaction map. These papers only mark the start for further novel applications to biological and biomedical questionings. One could for example start to compare interaction networks upon perturbation or as a function of time. Likewise, one could analyze very large patient cohorts in the biomedical arena following the same rules for ultra-robust, fast and sensitive LC-MS. Advanced technological developments of this platform also allowed us to enter unexpected areas of MS-based proteome analyses with tremendous impact as described in the next chapters.

3.1.1. Article 1: Online PASEF with a novel TIMS

Online parallel accumulation – serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer

Molecular & Cellular Proteomics, December 01, 2018

Florian Meier[1], **Andreas-David Brunner[1]**, Scarlet Koch[2], Heiner Koch[2], Markus Lubeck, Michael Krause[2], Niels Goedecke[2], Jens Decker[2], Thomas Kosinski[2], Melvin A. Park[3], Nicolai Bache[4], Ole Hoerning[4], Jürgen Cox[5], Oliver Räther[2], Matthias Mann[1,6, #]

# Correspondence

[1]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[2]Bruker Daltonik GmbH, Fahrenheitstraße 4, 28359 Bremen, Germany

[3]Bruker Daltonics Inc., Manning Road, Billerica, Massachusetts 01821, USA

[4]Evosep Biosystems, Thriges Pl. 6, 5000 Odense, Denmark

[5]Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[6]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blogdamsvej 3B, 2200 Copenhagen, Denmark

Contribution

I contributed to the overall experimental design of the paper and accompanying data analysis. I performed all experiments to evaluate the mass spectrometry setup for high-speed proteomics in combination with the ultra-robust EvoSep One liquid chromatography platform and performed qualitative/quantitative benchmarking assessments of one-/two-proteome experiments. Furthermore, I evaluated the ion mobility reproducibility across many runs, scan times, and highlighted the instruments capability for ultra-high sensitivity proteomics down to only 10 ng of tryptic HeLa digest.

Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer

Authors

Florian Meier, Andreas-David Brunner, Scarlet Koch, Heiner Koch, Markus Lubeck, Michael Krause, Niels Goedecke, Jens Decker, Thomas Kosinski, Melvin A. Park, Nicolai Bache, Ole Hoerning, Jürgen Cox, Oliver Räther, and Matthias Mann

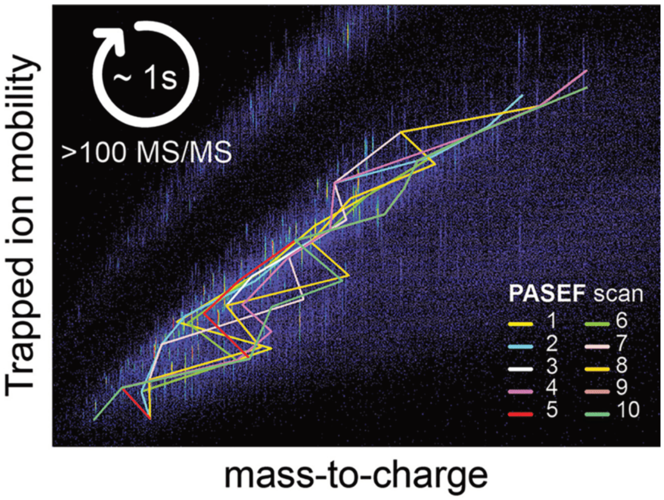
Correspondence

mmann@biochem.mpg.de

In Brief

PASEF multiplies the sequencing speed without any loss in sensitivity and is implemented in the timsTOF Pro instrument introduced here. Sequencing speeds above 100 Hz enable single run proteome analysis at a depth of 6000 proteins, making the instrument particularly attractive for rapid and highly sensitive proteomics. Collisional cross sections can be determined with up to 0.1% precision and acquired on a scale of 100,000s, which opens exciting areas for proteomics exploration.

Graphical Abstract



Highlights

- Online PASEF achieves greater than 100 MS/MS per second at full sensitivity.
- Accurate label-free quantification of over 6000 proteins in 2 h.
- High throughput demonstrated on 50 ng digests measured in 5 min.
- High-precision determination of 100,000 peptide collisional cross sections.

Meier et al., 2018, Molecular & Cellular Proteomics *17*, 2534–2545

https://doi.org/10.1074/mcp.TIR118.000900

⌘ Author's Choice

Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer*Ⓢ

Florian Meier‡, Andreas-David Brunner‡, Scarlet Koch§, Heiner Koch§, Markus Lubeck§, Michael Krause§, Niels Goedecke§, Jens Decker§, Thomas Kosinski§, Melvin A. Park¶, Nicolai Bache‖, Ole Hoerning‖, Jürgen Cox**, Oliver Räther§, and Matthias Mann‡ ‡‡§§

In bottom-up proteomics, peptides are separated by liquid chromatography with elution peak widths in the range of seconds, whereas mass spectra are acquired in about 100 microseconds with time-of-flight (TOF) instruments. This allows adding ion mobility as a third dimension of separation. Among several formats, trapped ion mobility spectrometry (TIMS) is attractive because of its small size, low voltage requirements and high efficiency of ion utilization. We have recently demonstrated a scan mode termed parallel accumulation - serial fragmentation (PASEF), which multiplies the sequencing speed without any loss in sensitivity (Meier *et al.*, PMID: 26538118). Here we introduce the timsTOF Pro instrument, which optimally implements online PASEF. It features an orthogonal ion path into the ion mobility device, limiting the amount of debris entering the instrument and making it very robust in daily operation. We investigate different precursor selection schemes for shotgun proteomics to optimally allocate in excess of 100 fragmentation events per second. More than 600,000 fragmentation spectra in standard 120 min LC runs are achievable, which can be used for near exhaustive precursor selection in complex mixtures or accumulating the signal of weak precursors. In 120 min single runs of HeLa digest, MaxQuant identified more than 6,000 proteins without matching to a library and with high quantitative reproducibility (R > 0.97). Online PASEF achieves a remarkable sensitivity with more than 2,500 proteins identified in 30 min runs of only 10 ng HeLa digest. We also show that highly reproducible collisional cross sections can be acquired on a large scale (R > 0.99). PASEF on the timsTOF Pro is a valuable addition to the technological toolbox in proteomics, with a number of unique operating modes that are only beginning to be explored. ***Molecular & Cellular Proteomics 17: 2534–2545, 2018. DOI: 10.1074/mcp.TIR118.000900.***

Jointly, proteins form a cellular machinery—the proteome—that orchestrates essentially all biological processes in health and disease. Studying it on a system-wide scale holds great promise to advance our understanding of cellular biology and disease mechanisms (1–3). However, as compared with genomics and transcriptomics technologies, proteomics still lags in terms of coverage, throughput, and sensitivity. Virtually complete measurements of mammalian proteomes have become possible (4), but have mostly involved laborious sample preparation workflows, days of measurement time and substantial amounts of starting material. Furthermore, current high-performance instrumentation often requires expert knowledge and extensive maintenance, which impedes widespread adaptation of proteomics in nonspecialized laboratories.

In bottom-up workflows, proteins are extracted from a biological sample of interest and enzymatically cleaved to make them more amenable to mass spectrometric (MS) analysis. The resulting complex peptide mixtures are typically separated via nano-flow liquid chromatography (LC)[1], ionized by electrospray and mass analyzed. In "data-dependent" or "topN" acquisition schemes, the mass spectrometer detects suitable peptide precursor ions in full scans (MS) and selects them for fragmentation in *N* consecutive MS/MS scans. High resolution and high mass accuracy analyzers detect hundreds of thousands of distinct molecular features in single LC-MS experiments, of which only a minority is identified and quantified (5). These co-eluting peptides with abundances ranging over many orders of magnitude present a formidable analytical challenge, which has constantly pushed the development of faster and more sensitive instrumentation over the last decades (1, 3, 6, 7).

From the ‡Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany; §Bruker Daltonik GmbH, Fahrenheitstr. 4, 28359 Bremen, Germany; ¶Bruker Daltonics Inc., Manning Road, Billerica, Massachusetts 01821; ‖Evosep Biosystems, Thriges Pl. 6, 5000 Odense, Denmark; **Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany; ‡‡NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark

Received June 4, 2018, and in revised form, October 30, 2018
Published, MCP Papers in Press, November 1, 2018, DOI 10.1074/mcp.TIR118.000900

Time-of-flight (TOF) instruments have several very desirable properties for the analysis of complex peptide mixtures and have consequently been employed in shotgun proteomics for a long time (8, 9). Instrumental performance has steadily improved over the years, and our groups have described shotgun proteome measurements at a mass resolution of more than 35,000 within about 100 μs on the *impact II* (10), the predecessor of the instrument that is the subject of this paper. The high acquisition rate of TOF instruments allows coupling them with very fast separation techniques, such as ion mobility spectrometry (IMS) (11–13). IMS separates ions in the gas phase based on their size and shape, or more precisely their collisional cross section (CCS, Ω), typically within 10s to 100s of milliseconds (14). As the ions emerge from the IMS device, they can be efficiently sampled in the ms or sub-ms time frame with TOF analyzers. Nested between LC and MS, the technology provides an additional dimension of separation (15–17) and can increase analysis speed and selectivity (18), also with highly complex proteomics samples (19–23). However, many implementations of IMS, such as drift tubes, are challenging because of the device sizes and high voltages involved and may also limit the proportion of the continuous incoming beam that can be utilized (12, 13, 24). Trapped ion mobility spectrometry (TIMS) (25, 26) reverses the concept of traditional drift tube ion mobility by bringing ions to a rest at different positions in an ion tunnel device, balanced in an electrical field against a constant gas stream (27). Once enough ions have been trapped and separated, lowering the electrical potential releases time-resolved ions from the TIMS device into the downstream mass analyzer. This design reduces the IMS analyzer dimensions to about 10 centimeters in length—allowing two of them to be implemented in series for 100% duty cycle operation (28). TIMS furthermore offers high flexibility in that users can tune the ion mobility resolving power ($\Omega/\Delta_{FWHM}\Omega$) up to 200 or higher by simply lowering the TIMS scan speed (29, 30).

We have recently introduced "Parallel Accumulation - SErial Fragmentation" (PASEF) (31), which synchronizes MS/MS precursor selection with TIMS separation. This acquisition scheme allows fragmentation of more than one precursor per TIMS scan and we demonstrated that PASEF increases the sequencing speed severalfold without loss of sensitivity. As precursor ions are accumulated in parallel, PASEF overcomes the diminishing returns of increasingly fast MS/MS acquisition, which otherwise necessarily implied less and less ions per spectrum. Our first iteration was implemented on a laboratory prototype, which required manual precursor programming and was limited by the speed of the electronics involved. Here, we describe the construction and investigate the proteomics performance of the first mass spectrometer that fully integrates the PASEF concept, the Bruker *timsTOF Pro*.

EXPERIMENTAL PROCEDURES

Cell Culture and Sample Preparation—Human cervical cancer cells (HeLa S3, ATCC, Manassas, VA) were grown in Dulbecco's modified Eagle's medium with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin-streptomycin (all Life Technologies Ltd., Paisley, UK). Escherichia coli (strain: XL1 blue) was cultured at 37 °C in LB medium until logarithmic phase (optical density = 0.5, λ = 600 nm). Cells were collected by centrifugation. Following a washing step with cold phosphate buffered saline, they were pelleted and flash frozen in liquid nitrogen and stored at −80 °C.

One-device cell lysis, reduction, and alkylation was performed in sodium deoxycholate (SDC) buffer with chloroacetamide (PreOmics GmbH, Martinsried, Germany) according to our previously published protocol (32). Briefly, the cell suspension was twice boiled for 10 min at 95 °C and subsequently sonicated for 15 min at maximum energy (Bioruptor, Diagenode, Seraing, Belgium). Proteins were enzymatically hydrolyzed overnight at 37 °C by LysC and trypsin (1:100 enzyme:protein (wt/wt) for both). To stop the digestion, the reaction mixture was acidified with five volumes of isopropanol with 1% trifluoroacetic acid (TFA). Peptides were de-salted and purified in two steps, first on styrenedivinylbenzene-reversed phase sulfonate (SDB-RPS), and second on C_{18} sorbent. The dried eluates were re-constituted in water with 2% acetonitrile (ACN) and 0.1% TFA for direct LC-MS analysis or high pH reversed-phase fractionation.

For the experiments with the Evosep One (see below), HeLa cell pellets were re-suspended and lysed in water/trifluoroethanol. Disulfide bonds were reduced with dithiothreitol and alkylated with iodoacetamide in ammonium bicarbonate buffer. Following tryptic digestion, the peptide mixture was de-salted and purified on C_{18} sorbent.

Peptide Fractionation—High pH reversed-phase fractionation was performed on an EASY-nLC 1000 (Thermo Fisher Scientific, Bremen, Germany) coupled to a "spider fractionator" (PreOmics) as detailed in ref (33). Purified peptides were separated on a 30 cm × 250 μm reversed-phase column (PreOmics) at a flow rate of 2 μl/min at pH 10. The binary gradient started from 3% buffer B (PreOmics), followed by linear increases to first 30% B within 45 min, to 60% B within 17 min, and finally to 95% B within 5 min. Each sample was automatically concatenated into 48 fractions in 90 s time intervals. The fractions were dried in a vacuum-centrifuge and reconstituted in water with 2% ACN and 0.1% TFA for LC-MS analysis.

Liquid Chromatography—An EASY-nLC 1200 (Thermo Fisher Scientific) ultra-high pressure nano-flow chromatography system was coupled online to a hybrid trapped ion mobility spectrometry - quadrupole time of flight mass spectrometer (*timsTOF Pro*, Bruker Daltonics, Bremen, Germany) with a modified nano-electrospray ion source (10) (CaptiveSpray, Bruker Daltonics). Liquid chromatography was performed at 60 °C and with a constant flow of 400 nL/min on a reversed-phase column (50 cm × 75 μm i.d.) with a pulled emitter tip, packed with 1.9 μm C_{18}-coated porous silica beads (Dr. Maisch, Ammerbuch-Entringen, Germany). Mobile phases A and B were water with 0.1% formic acid (v/v) and 80/20/0.1% ACN/water/formic acid (v/v/vol), respectively. In 120-min experiments, peptides were separated with a linear gradient from 7.5 to 27.5% B within 60 min, followed by an increase to 37.5% B within 30 min and further to 55% within 10 min, followed by a washing step at 95% B and re-equilibration. In 60 min separations, the gradient increased from 10 to 30% B within 30 min, followed by an increase to 40% B within 15 min and further to 57.5% B within 5 min before washing and re-equilibration. In 30 min separations, the initial 10–30% B step was 15 min, followed by a linear increase to 40% B (7.5 min) and 57.5% B (2.5 min) before washing and re-equilibration.

[1] The abbreviations used are: LC, liquid chromatography; CCS, collisional cross section; IMS, ion mobility spectrometry; PASEF, parallel accumulation–serial fragmentation; TIMS, trapped ion mobility spectrometry.

For some experiments we used the Evosep One (Evosep, Odense, Denmark), a new HPLC instrument employing an embedded gradient and capable of fast turnaround between analyses (34). Samples were eluted from Evotips at low pressure into the storage loop with a gradient offset to lower the percentage of organic buffer. Separation was performed on a customized 5.6 min gradient (200 samples/day method) at a flow rate of 2.0 μl/min on a 4 cm x 150 μm i.d. reversed-phase column packed with 3 μm C_{18}-coated porous silica beads (PepSep, Odense, Denmark).

The timsTOF Pro Mass Spectrometer—The *timsTOF Pro* is the successor to the *impact II* instrument, compared with which it features an additional ion mobility analyzer. However, the *timsTOF Pro* is a complete redesign in hardware and firmware. Apart from incorporating TIMS, the design goals included the achievement of similar or better mass resolution (>35,000) and improved robustness through a modified ion path.

In the experiments described here, the mass spectrometer was operated in PASEF mode. Desolvated ions entered the vacuum region through the glass capillary and were deflected by 90°, focused in an electrodynamic funnel, and trapped in the front region of the TIMS tunnel consisting of stacked printed circuit boards (PCBs) with an inner diameter of 8 mm and a total length of 100 mm. The PCB electrodes form a stacked multipole in the direction of ion transfer, in which an applied RF potential of 350 V_{pp} confined the trapped ions radially. The TIMS tunnel is electrically separated into two parts (dual TIMS), where the first region is operated as an ion accumulation trap that primarily stores all ions entering the mass spectrometer, while the second part performs trapped ion mobility analysis (28). As soon as the TIMS analysis is finished, all stored ions are transferred to the analyzer part and the storage region is filled again. If equal accumulation and analysis times are used in both TIMS regions, this enables operation at duty cycles close to 100%. Ion transfer between the two regions takes 2 ms and therefore does not affect the overall ion utilization for typical ramp and accumulation times around 25 to 200 ms.

In both TIMS regions, the RF field is superimposed (from entrance to exit) by an increasing longitudinal electrical field gradient, such that ions in the tunnel simultaneously experience a drag from the incoming gas flow through the capillary and a repulsion from the electrical field. Here, we used a flow of ambient laboratory air for ion mobility separation. Depending on their collisional cross sections and charge states, ions come to rest closer to the entrance of the tunnel (high ion mobility) or closer to its exit (low ion mobility). Trapped ion mobility separation was achieved by ramping the entrance potential of the second TIMS region from −207 V to −77 V. A single TIMS-MS scan is composed of many individual TOF scans of about 110 μs each. In the experiments reported here, we first systematically varied the ramp times from 25, 50, 100, 150, to 200 ms while keeping the duty cycle fixed at 100%. All further experiments were acquired with a 100 ms ramp and 10 PASEF MS/MS scans per topN acquisition cycle, except for the Evosep One experiments, which were performed with four PASEF MS/MS scans per cycle. In TOF mass spectrometry, signal-to-noise ratios can conveniently be increased by summation of individual TOF scans. Here, low-abundance precursors with an intensity below a ‘target value’ were repeatedly scheduled for PASEF-MS/MS scans until the summed ion count reached the target value (*e.g.* four times for a precursor with the intensity 5000 arbitrary units (a.u.) and a target value of 20,000 a.u.). We set the target value to 20,000 a.u. for all methods, except for the Evosep One experiments where it was set to 24,000 a.u.

MS and MS/MS spectra were recorded from *m/z* 100 to 1700. Suitable precursor ions for PASEF-MS/MS were selected in real time from TIMS-MS survey scans by a sophisticated PASEF scheduling algorithm (see also Results). A polygon filter was applied to the *m/z* and ion mobility plane to select features most likely representing peptide precursors rather than singly charged background ions. The quadrupole isolation width was set to 2 Th for $m/z < 700$ and 3 Th for $m/z > 700$, and the collision energy was ramped stepwise as a function of increasing ion mobility: 52 eV for 0–19% of the ramp time; 47 eV from 19–38%; 42 eV from 38–57%; 37 eV from 57–76%; and 32 eV for the remainder.

The TIMS elution voltage was calibrated linearly to obtain reduced ion mobility coefficients ($1/K_0$) using three selected ions of the Agilent ESI-L Tuning Mix (*m/z* 622, 922, 1222) (35).

Collisional cross sections were calculated from the Mason Schamp equation (36).

$$CCS = \frac{3ze}{16}\frac{1}{K_0}\sqrt{\frac{2\pi}{\mu k_b T}}$$

where z is the charge of the ion, e is the elemental charge, k_b is Boltzman's constant, μ is the reduced mass, and T the temperature (305 K). For all calculations, we assumed pure N_2 as the drift gas.

Data Analysis—Mass spectrometry raw files were processed with MaxQuant (37) version 1.6.1.13, which has been extended to incorporate the additional ion mobility dimension and adapted to handle the TIMS data format. This new version of MaxQuant is publicly available and will be described in detail separately (Cox and coworkers, *in preparation*). Briefly, it assembles four-dimensional isotope clusters - defined by *m/z*, retention time, ion mobility and intensity - from the TIMS-MS spectra and extracts ion mobility separated MS/MS spectra from the PASEF scans. Each MS/MS spectrum is assigned to its respective precursor ion by quadrupole isolation *m/z* and ion mobility values, and in case a precursor has been fragmented multiple times in one acquisition cycle, the respective spectra are collapsed to a single spectrum with increased signal-to-noise. For the four-dimensional feature detection in MaxQuant, only every third data point (TOF scan) in the TIMS dimension was considered (“TIMS step width” = 3), and the “TIMS half width” parameter was set to 4 TOF scans (equivalent to about 440 μs). The “TIMS mass resolution” parameter was set to 32,000 and MS/MS peaks with an intensity below 1.5 units were discarded.

The MS/MS spectra were matched to *in silico* derived fragment mass values of tryptic peptides from a reference proteome (Uniprot, 2016/05, HeLa: 91,618 entries including isoforms, *E. coli*: 4313 entries including isoforms) and 245 potential contaminants by the built-in Andromeda search engine (38). A maximum of two missing cleavages were allowed, the required minimum peptide sequence length was 7 amino acids, and the peptide mass was limited to a maximum of 4600 Da. Carbamidomethylation of cysteine residues was set as a fixed modification, and methionine oxidation and acetylation of protein N termini as variable modifications. The initial maximum mass tolerances were 70 ppm for precursor ions and 35 ppm for fragment ions. After re-calibration, precursor mass tolerances were individually adjusted by MaxQuant to less or equal 20 ppm for each isotope pattern based on the local precision of the mass measurement. We employed a reversed sequence library to control the false discovery rate (FDR) at less than 1% for peptide spectrum matches and protein group identifications.

Decoy database hits, proteins identified as potential contaminants, and proteins identified exclusively by one site modification were excluded from further analysis. Label-free protein quantification was performed with the MaxLFQ algorithm (39) requiring a minimum ratio count of 1. All other MaxQuant parameters were kept at their default values.

Mass spectrometric metadata, such as the information about PASEF-scheduled precursor ions, were directly accessed and extracted from the Bruker *.tdf* raw files with a SQLite database viewer (SQLite Manager, v0.8.3.1). Bioinformatic analysis and visualization was performed in either Python (Jupyter Notebook), Perseus (40) (v1.6.0.8) or the R statistical computing environment (41) (v3.2.1).

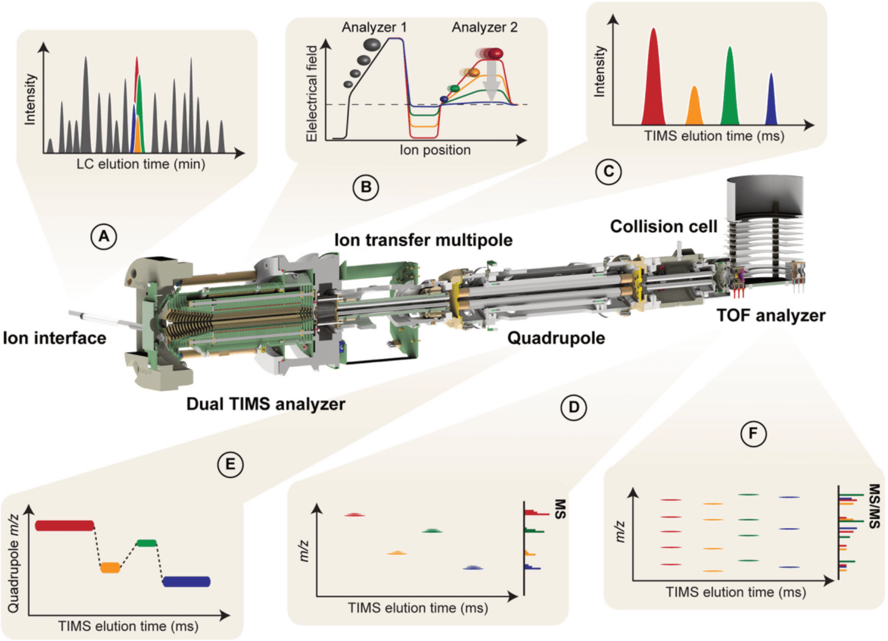


FIG. 1. **Online Parallel Accumulation - Serial Fragmentation (PASEF) with the timsTOF Pro.** *A*, Peptides eluting from the chromatographic column are ionized and enter the mass spectrometer through a glass capillary. *B*, In the dual TIMS analyzer, the first TIMS section traps and stores ion packets, and the second resolves them by mobility. *C*, *D*, Ion mobility separated ions are released sequentially from the second TIMS analyzer as a function of decreasing electrical field strength and yield mobility-resolved mass spectra. *E*, In PASEF MS/MS scans, the TIMS analyzer and the quadrupole are synchronized and the quadrupole isolation window switches within sub-milliseconds between mobility resolved precursor ions of different *m/z*. *F*, This yields multiple ion mobility resolved MS/MS spectra from a single TIMS scan, and ensures that multiple trapped precursor ion species are used for fragmentation. Non mobility-resolved MS and MS/MS spectra are projected onto the right axes in (*E*) and (*F*) for comparison.

Experimental Design and Statistical Rationale—The complete data set reported in this study comprises 108 raw files. Samples were grouped by mass spectrometric acquisition methods or, in case of the data for Fig. 5, by pipetting ratios. Replicate injections were performed to assess the technical reproducibility of the respective methods and their quantitative accuracy. The exact *N* numbers are $n = 4$ in Fig. 4; $n = 4$ in Fig. 5*A*–5*C*; $n = 5$ in Fig. 5*D*; $n = 3$ in Fig. 6*A*–6*B*; $n = 16$ in Fig. 6*C*; $n = 4$ in Fig. 7*A*–7*C* and $n = 1$ in Fig. 7*D*. To allow accurate external calibration of ion mobility values, we acquired experiments with different TIMS ramp times in batches. Dilution series were measured from low to high concentrations starting with blank runs to avoid carry over. This study does not draw biological conclusions, which is why process and biological replicates or controls were not performed. In the description of Fig. 5*D* the data were filtered for at least two valid values in each group (1:1 and 1:5 mixing ratios, respectively) and a one-sided two-sample *t* test was performed. Multiple-hypothesis testing was corrected by truncating *t* test significant hits at a permutation-based FDR threshold of 0.05 (250 randomizations) in the Perseus software.

RESULTS

Construction of a TIMS-QTOF Instrument with Online PASEF—The *timsTOF Pro* is a quadrupole time-of-flight (QTOF) mass spectrometer equipped with a second generation dual TIMS analyzer in the first vacuum stage (Fig. 1). This set-up spatially separates ion accumulation and ion mobility analysis into two sequential sections of the TIMS tunnel, so that these steps happen in parallel (28) (analyzer 1 and 2 in Fig. 1*B*). Within the limits of ion storage capacity, up to 100% of the ions that enter the mass spectrometer can therefore be utilized for mass analysis. Here, we typically accumulated ions for 25 to 200 ms, and transferred them into the second TIMS region within 2 ms. From this TIMS region they were released by decreasing the voltage gradient linearly within 25 to 200 ms (TIMS ramp time). Simulations show that most of the ion mobility separation happens near the top plateau close to the exit of the device (42–44) and we observed that leaving peptide ion packets had narrow ion mobility peaks with median half widths of about 2 ms or less (Fig. 1*C*). In TIMS, low mobility ions are released or 'eluted' first, followed by more mobile ions with smaller collisional cross sections relative to their charge. In addition to separating ions by shape and size, the time-focusing effect of TIMS increases signal-to-noise ratios about 50-fold (depending on the relative accumulation and ramp times) compared with the standard continuous acquisition mode because ion species are concentrated into narrow packets whereas the noise distributes across the ion mobility scan (28).

At the exit of the TIMS device, ions pass through the ion transfer multipole, the quadrupole mass filter and are then accelerated into the collision cell. From there, intact (MS scans) or fragment (MS/MS scans) ions are extracted into an orthogonal accelerator unit and pushed into the flight tube for mass analysis (Fig. 1*D*). The ions enter a V-shaped flight path through a two-stage reflectron and finally impinge on a multi-channel plate ion detector coupled to a 10-bit digitizer with a sampling rate of 5 Gigasamples/s, enabling high-resolution mass analysis ($R > 35,000$). With time-of-flight mass analyzers, as opposed to Fourier-transform mass analyzers, this resolving power is nearly constant over the entire *m/z* range and independent of the scan time. Of note, we observed that the re-designed ion transfer path - presumably mainly the 90-degree bent at the entrance of the TIMS device and the new quadrupole with increased inner diameter - had a positive effect on the instrument's robustness. This was evidenced by continuous operation of the instrument during its development for more than 1.5 years, in which time we only cleaned the ion transfer capillary but not the internals of the instrument.

In PASEF mode, MS/MS precursor selection by the quadrupole mass filter is synchronized with the release of ions from the TIMS device, which requires very fast switching times of the quadrupole to keep pace with the fast ion mobility separation and to maximize the number of precursors per TIMS scan (Fig. 1*E*). The *timsTOF Pro* electronics have been designed to meet these requirements. RF and DC voltages for mass selection are now calculated and set by a real-time field-programmable array, as opposed to a conventional and slower serial interface. This allows fully synchronized operation of TIMS and quadrupole with switching times of 1 ms or less. By setting the quadrupole to *N* different isolation windows, PASEF yields *N* ion-mobility-resolved MS/MS spectra for a single TIMS scan (Fig. 1*F*). Because all precursor ions are stored in parallel, the absolute ion count per MS/MS spectrum is equal to a conventional TOF MS/MS spectrum summed up over the accumulation time, giving rise to an *N*-fold increase in sequencing speed without sacrificing sensitivity. The maximum number of precursors per TIMS scan is not limited by the instrument electronics, but rather by the separation of precursors in the ion mobility dimension and by the efficient design of "switching routes" for precursor selection, which will be described next.

PASEF Precursor Selection in Real-time—In complex proteomics samples, such as whole cell lysates, hundreds to thousands of peptides elute at any time, presenting a challenge for optimal precursor selection even with the 10-fold higher sequencing speed offered by PASEF. Fortunately, precursors are now distributed in a two-dimensional (*m/z* and ion mobility) space in which an optimal route can be selected, like the "traveling salesman problem" in computer science. Even though exact solutions exist, for example by a brute-force method that simply iterates over all possible combinations, they cannot be computed on the LC time scale nor is it clear which peaks are most desirable to "visit". Instead, we here developed a heuristic algorithm that limits the computational time to about 100 ms in complex samples, and aims to maximize the number of precursors per acquisition cycle that can be successfully identified. This involves three dimensions: precursor *m/z*, signal intensity and ion mobility (Fig. 2). Our precursor search is offset by one acquisition cycle from ongoing data acquisition to avoid introducing any scan overhead time. In distributing precursors to PASEF scans, our algorithm accounts for the quadrupole switching time as well as the elution order of ion mobility peaks and prioritizes high-abundance precursors. In principle, the maximum coverage of eluting peptides should be achieved by using the PASEF speed advantage exclusively on unique precursor ions. However, this leads to many low abundant precursors being selected, and thus many low-quality MS/MS spectra. Common strategies to increase spectral quality are (1) increasing the ion accumulation time for selected, isolated ions (in trapping-based instruments) or (2) summing consecutive scans (in TOF-MS). In contrast, TIMS-PASEF accumulates all precursors upstream of selection whereas many precursors can be fragmented consecutively in each single PASEF scan. This enables deliberate and efficient re-scheduling of selected low-abundance precursor ions in subsequent PASEF scans. In post-processing, these individual spectra are summed to increase signal-to-noise. This "re-sequencing" is implemented in our precursor algorithm by a "target intensity" parameter, with which users can balance the desired spectral quality with the number of unique precursors. Other than that,

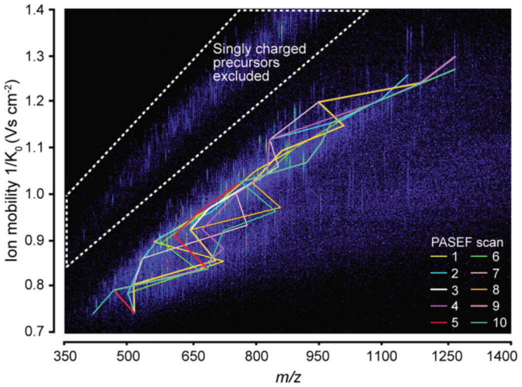


FIG. 2. **Real-time PASEF precursor selection in three dimensions.** Heat-map visualization of ion mobility resolved peptide ions at a single time point in an LC-TIMS-MS analysis of a HeLa digest. Connected lines indicate the *m/z* and mobility positions of all precursor ions selected for fragmentation in the following TIMS-PASEF scans (color-coded).

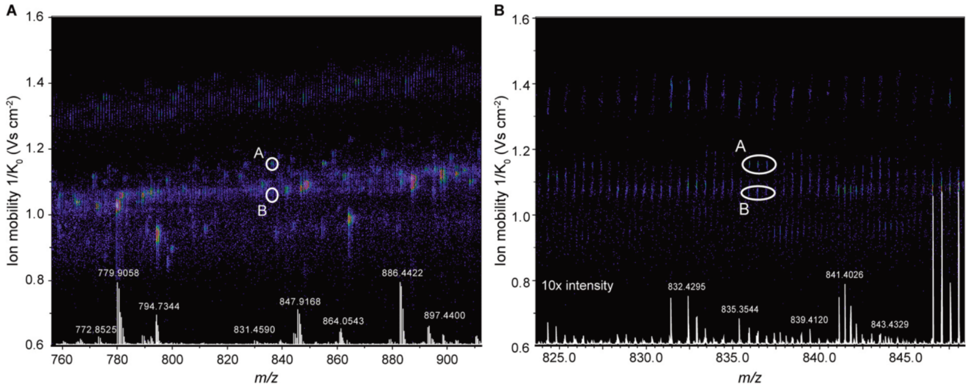


FIG. 3. **Trapped ion mobility separation of peptide precursor ions.** *A*, The two nearly isobaric peptide ions A and B were distinguished by their ion mobility and selected separately for fragmentation by the PASEF scheduling algorithm in an LC-TIMS-MS experiment of a HeLa digest. *B*, Zoomed view into the precursor *m/z* range. Non mobility-resolved MS spectra are projected onto the lower axis for comparison. The corresponding MS/MS spectra are shown in supplemental Fig. S1.

we excluded precursors dynamically after one sequencing event to not compromise proteomic depth. Singly charged species were readily excluded by their characteristic positions in the *m/z versus* ion mobility plane. The flow chart in supplemental Fig. S1 depicts the precursor selection algorithm in detail.

We tested the performance of our precursor selection algorithm in 120 min LC-TIMS-MS runs of HeLa digests. Fig. 2 shows a representative TIMS-MS survey scan in the middle of the LC gradient. From this 100 ms TIMS scan, our algorithm selected 50 unique precursor ions for fragmentation in the subsequent PASEF scans (color-coded) out of which 32 low-abundance precursors were repeatedly sequenced. All precursor ions were widely distributed in *m/z* and ion mobility space, indicating an efficient coverage of the entire precursor space. In total, 118 MS/MS spectra were acquired in this cycle, which equals a sequencing rate of more than 100 Hz. Because all precursors were accumulated for 100 ms, the total number of ions for each precursor corresponds to that of a 10 Hz MS/MS selection if no PASEF had been employed.

With the selection algorithm in place, we inspected hundreds of precursor identifications in our data sets. Often, the separation of precursors along the additional ion mobility dimension was crucial as illustrated in Fig. 3. In a projection of the data onto the *m/z* axis, no obvious precursor signals were present, even when enlarging the signal 10-fold relatively to the more abundant peaks. However, the precursor selection algorithm had found and fragmented two distinct isotope clusters in the ion mobility - *m/z* space, which were separately fragmented by PASEF and clearly identified (supplemental Fig. S2).

Single Run Proteomes—Next, we investigated the effect of different TIMS ramp times on precursor selection. Given a minimum selection and transition time for the quadrupole adjustment of a few ms, the overall number of achievable fragmentation events should be roughly similar for different TIMS ramp times as increasing ramp times allow fragmenting more precursors per PASEF scan - while acquiring less scans overall. To find a good balance for proteomics applications, we varied the TIMS ramp from 25 to 200 ms and kept the number of PASEF scans at 10 per acquisition cycle. We chose to operate the instrument at a near 100% duty cycle by setting the TIMS acquisition time equal to the ramp time.

With the slowest (and therefore highest mobility resolving) TIMS ramp, a median of 24 precursors were sequenced per scan (Fig. 4*A*). Faster ramp times resulted in nearly proportionately less precursors per PASEF scan, which was partially balanced by the overall higher number of scans per analysis. Interestingly, the 25 ms ramp was clearly inferior as it did not take full advantage of the PASEF principle, yielding only about 380,000 MS/MS spectra per run. With the 100 ms ramp over 600,000 MS/MS spectra were acquired in two hours (Fig. 4*B*). For comparison, acquiring the same number of MS/MS spectra without PASEF at the same sensitivity would have taken 11 times longer—almost 1 day. Remarkably, the instrument was sequencing at rates above 100 Hz during the entire time that peptides were eluting. As discussed above, we decided to use this extreme speed in part on resequencing low-abundance peptides to generate higher-quality summed spectra (Fig. 4*C*). On average, a given precursor ion was fragmented 2.2 times in 25 ms ramps and 1.9 times with 200 ms ramps. Overall, this resulted in up to 320,000 MS/MS spectra of unique precursor ions in a single run as detected by the real-time PASEF scheduling algorithm (Fig. 4*D*), although post-processing in MaxQuant combined many of these.

From 200 ng whole-cell HeLa digest per run, we identified on average 25,885 sequence-unique peptides in quadruplicate single runs with the 200 ms method, and about 32,000 with the faster 100 ms and 150 ms methods (Fig. 4*E*). The median peptide length was 14 amino acids. In all runs, the

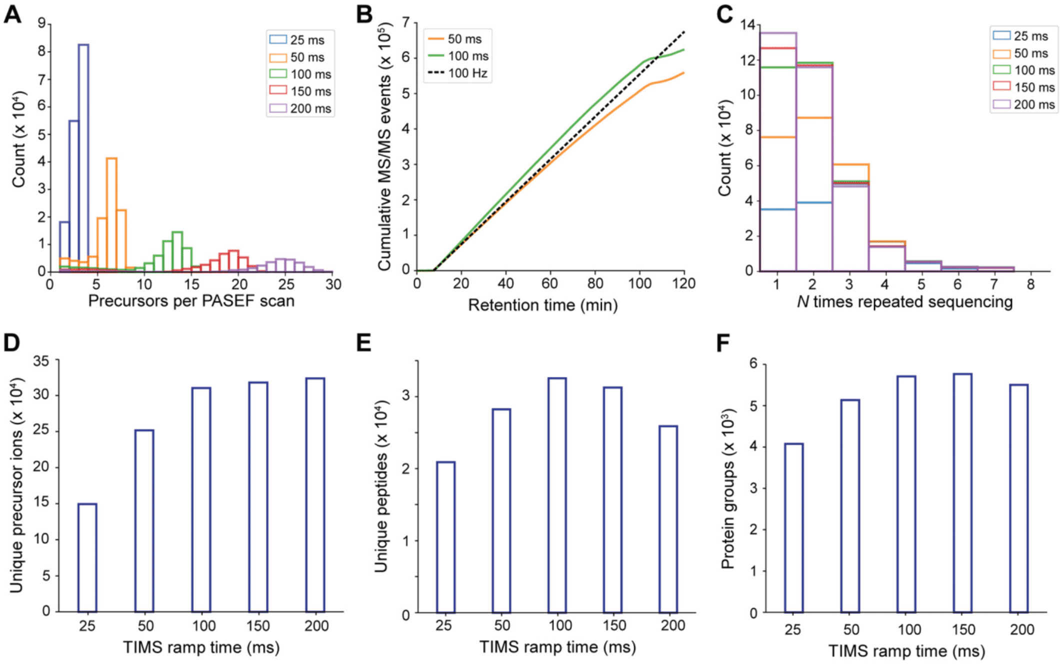


FIG. 4. **Single run analyses of a HeLa digest.** *A*, Number of selected precursor ions per PASEF scan with different TIMS ramp times in 120 min runs of 200 ng HeLa digests. *B*, Cumulative number of PASEF MS/MS spectra as a function of retention time for 50 ms and 100 ms TIMS ramps. The dashed line indicates the theoretical number of MS/MS spectra for a constant acquisition rate of 100 Hz starting at a retention time of 7.5 min. *C*, Number of repeated sequencing events for precursors with different ramp times. *D*, Number of unique precursor ions detected with different TIMS settings. *E*, Average number of sequence-unique peptides identified in single runs ($n = 4$) with different TIMS settings. *F*, Average number of protein group identifications in single runs ($n = 4$) with different TIMS settings.

median absolute peptide mass accuracy was below 1.5 ppm. The number of inferred protein groups at an FDR below 1% increased to an average of over 5700 protein groups per run at TIMS ramp times of 100 and 150 ms (Fig. 4*F*). With the 100 ms ramp, we identified in total 6090 protein groups (5230 with two or more peptides) with a median sequence coverage of 19.7%.

Label-free Proteome Quantification—A central task in proteomics is the accurate quantification of protein abundances across multiple biological samples. Label-free quantification (LFQ) is a popular method because of its simplicity and scalability to larger sample cohorts. Using the optimized 100 ms TIMS method we quantified on average 5575 protein groups in 2 h LC-MS time across quadruplicate injections. Run-to-run reproducibility was high with a median pairwise Pearson correlation coefficient of 0.979 between the four runs, with excellent linearity over four orders of magnitude in protein abundance (Fig. 5*A*). The median coefficients of variation (CVs) were 15.3% for the non-normalized peptide intensities and 7.2% at the protein level after MaxLFQ normalization (39) (Fig. 5*B*).

Quantitative accuracy in proteomics may be limited if proteins are inconsistently measured across the samples. In data-dependent acquisition schemes, this is partially because of semi-stochastic precursor selection - a consequence of the large number of co-eluting precursor candidates and the finite sequencing speed. We asked if the severalfold faster PASEF method as compared with standard shotgun acquisition methods would improve this situation even without transferring identifications by precursor mass ("matching between runs"). Indeed, PASEF alleviated the "missing value" problem and provided quantification values for 4972 proteins in four out of four runs (Fig. 5*C*). Only 279 low-abundance proteins were exclusively quantified in a single replicate. This translated into a data completeness of 91.5%, which compares favorably to standard data-dependent acquisition and is similar to data-independent acquisition schemes. We expect that transferring identifications between runs, as with the MaxQuant matching between runs feature, will lead to even more consistent protein quantification across samples.

To further benchmark the quantitative accuracy of our setup, we mixed tryptic digests from HeLa and *Escherichia coli* in 1:1 and 1:5 ratios and measured each sample in quin-

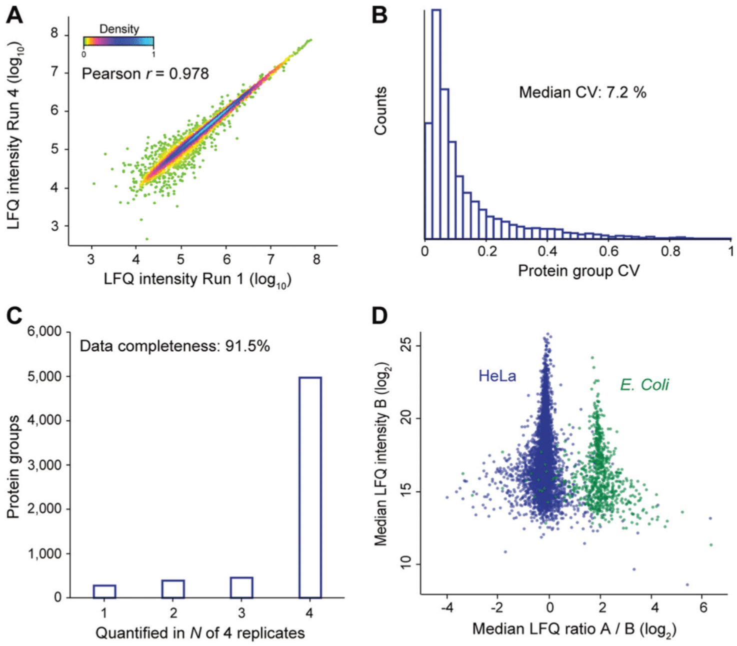


FIG. 5. **Label-free proteome quantification.** *A*, Pearson correlation of protein intensities in two replicate injections of a HeLa digest. *B*, Coefficients of variation (CVs) for protein quantities in four replicates (n = 5,811). *C*, Number of proteins quantified in N out of four replicates. *D*, Label-free quantification benchmark with whole-cell HeLa and *E. coli* digests mixed in 1:1 and 1:5 ratios (wt:wt). The scatterplot shows the median fold-change of 5407 human and 728 *E. coli* proteins in quintuplicate single runs.

tuplicate 120 min single runs. Overall, we quantified 6135 protein groups (5407 HeLa; 728 *E. coli*) with at least one valid value for both mixing ratios. Plotting the median fold-changes yielded two distinct clouds for HeLa and *E. coli* proteins, which were 4.3-fold separated in abundance, slightly less than the intended 5-fold mixing ratio (Fig. 5*D*). Both populations were narrow (σ(HeLa) = 0.44; σ(*E. coli*) = 0.77) relative to the expected fold-change and they had minimal overlap. Considering only the 5686 proteins with at least two valid values for each mixing ratio (5052 HeLa, 634 *E. coli*), a one-sided Student's t test returned 602 significantly changing *E. coli* proteins at a permutation-based FDR below 0.05. This represents an excellent sensitivity of ~95% and only 64 human proteins (1.3%) were falsely classified as changing. From these results, we conclude that the combination of TIMS and PASEF provides precise and accurate label-free protein quantification at a high level of data completeness.

High Throughput and Limited Sample Amounts—The performance characteristics discussed so far suggest that the instrument is particularly well suited for rapid and high sensitivity proteome analysis. To test this, we first reduced the peptide amount on column from 100 ng down to 10 ng HeLa digest per injection (Fig. 6*A*). With 100 ng on column and a 1 h gradient, we reproducibly identified 4513 protein groups, 79% of the proteome coverage with 200 ng in half the measurement time. Out of these, 3294 protein groups were quantified with a CV below 20%. At 50 ng, we quantified 4215 protein groups with high quantitative accuracy (median CV 9.8%), motivating us to inject even lower sample amounts. Remarkably, from only 10 ng HeLa digest, we still identified 2723 protein groups on average and 3113 in total (2159 with two or more peptides in at least one replicate). Assuming 150 pg protein per cell (45), this corresponds to the total protein amount of only about 60 HeLa cells, suggesting that TIMS-PASEF is well suited to ultrasensitive applications in proteomics. Even at this miniscule sample amount, quantitative accuracy remained high with a median peptide intensity CV of 9.7% and 1841 proteins quantified at a CV < 20%.

To investigate achievable throughput, we repeated our sensitivity experiments with a 30 min gradient (Fig. 6*B*). This is an attractive strategy as steeper gradients compress the ion current in narrower LC elution peaks (about 7 s FWHM for the 120 min gradient *versus* 4 s FWHM for the 30 min gradient), thereby providing higher ion counts in individual MS/MS scans. This is also evident from the observation that the target value was reached within an average of 1.5 repeats as compared with 2.0 and 1.7 for the 120 and 60 min gradients. Because of the very high sequencing speed of PASEF, reducing the measurement

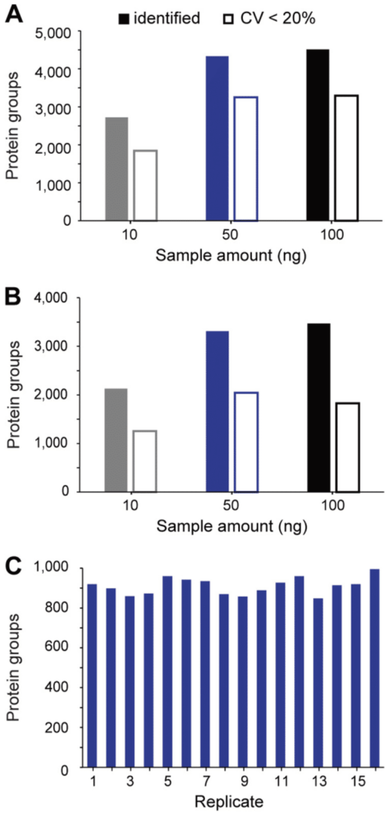


FIG. 6. **Rapid and sensitive HeLa proteome measurements.** *A*, Average number of protein groups identified and quantified with a CV <20% in 60 min single runs ($n = 3$). *B*, Average number of protein groups identified and quantified with a CV <20% in 30 min single runs ($n = 3$). *C*, Number of protein groups identified in sixteen replicate injections with the 5.6 min gradient.

time had only limited effect on proteome coverage. From 100 ng HeLa digest we identified on average 3470 protein groups in quadruplicate single runs, whereas 10 ng yielded 2128 protein groups, all with median CVs below 18%.

At the very short gradients made possible by the PASEF principle, throughput starts to be severely affected by the washing, loading and equilibration steps of the HPLC between injections. We therefore turned to the recently introduced Evosep One instrument, which features a preformed gradient, increasing robustness and largely eliminating idle time between injections (34). To explore the throughput limits of complex proteome analysis with PASEF, we made use of the "200 samples per day" method on the Evosep One, which consists of a 5.6 min gradient with 7.2 min total time between injections. Remarkably, in 16 replicates, 1231 proteins (910 with two or more peptides) were identified without any identification transfer from libraries and with only 50 ng of injected cell lysate (Fig. 6*C*). This combination of fast LC turnaround times with PASEF also holds great promise for rapid yet comprehensive analyses of less complex samples, for example protein interactomes, or the quantification of trace-level host cell proteins (HCPs) in recombinant biotherapeutics.

Large-scale Measurement of Peptide Collisional Cross Sections—In TIMS, the counteracting forces of a gas flow and an electrical field are used to separate the ions and to measure their mobility. Conceptually, this closely resembles the (inverted) situation in drift tube ion mobility, where ions are dragged by an electrical field through resting gas molecules. Because the underlying physics is identical, TIMS measurements are expected to correlate directly with classical drift tube ion mobility measurements and this has been established experimentally by Park and colleagues (42). Therefore, in contrast to other ion mobility setups (24), such as traveling-wave ion mobility (46) and differential ion mobility (47), TIMS can directly determine collisional cross sections by internal or external calibration.

We reasoned that the rapid measurement of tens of thousands of peptides demonstrated above, in combination with accurate CCS measurements, should allow generating a large-scale overview of the CCS dimension of peptides. We first explored the reproducibility with repeated injections of HeLa digest. Before the first injection, we calibrated the ion mobility dimension using reduced ion mobility values ($1/K_0$; V·s cm^{-2}) of phosphazine derivatives from the literature (35), which can be converted to ^{TIMS}CCS values using the Mason-Schamp equation (Experimental Procedures). Peptide ions can occur in multiple conformations (*e.g.* proline-containing peptides (48)), which results in multiple ion mobility peaks and complicates the analysis. For simplicity, we here only considered the most abundant feature reported by MaxQuant.

In four replicates, we generated 23,738 $1/K_0$ values of commonly identified peptide ions in all runs with a median CV of 0.1% and a median pairwise correlation coefficient > 0.99 (Fig. 7*A*). The average absolute deviation of ^{TIMS}CCS values across all four replicates was 0.2% (Fig. 7*B*). In our hands, this is at least 10-fold more reproducible than LC retention time, even on the same column and with the same gradient. Interestingly, the ^{TIMS}CCS measurements were also highly transferable across different TIMS ramp times (100 and 200 ms) as evident from a Pearson correlation coefficient of > 0.99 between them (Fig. 7*C*).

Having established precise ^{TIMS}CCS measurements in single runs, we next used loss-less high pH fractionation (33) to extend the scale of our data set. Measuring 48 fractions with 2 h gradients each resulted in 129,110 ^{TIMS}CCS values (filtered for $z > 1$) from 101,420 unique peptide sequences and about 9400 protein groups. In the *m/z versus* ^{TIMS}CCS plot, doubly, triply and higher-charged populations are clearly separated (Fig. 7*D*, Supplemental Table S1). Within each charge

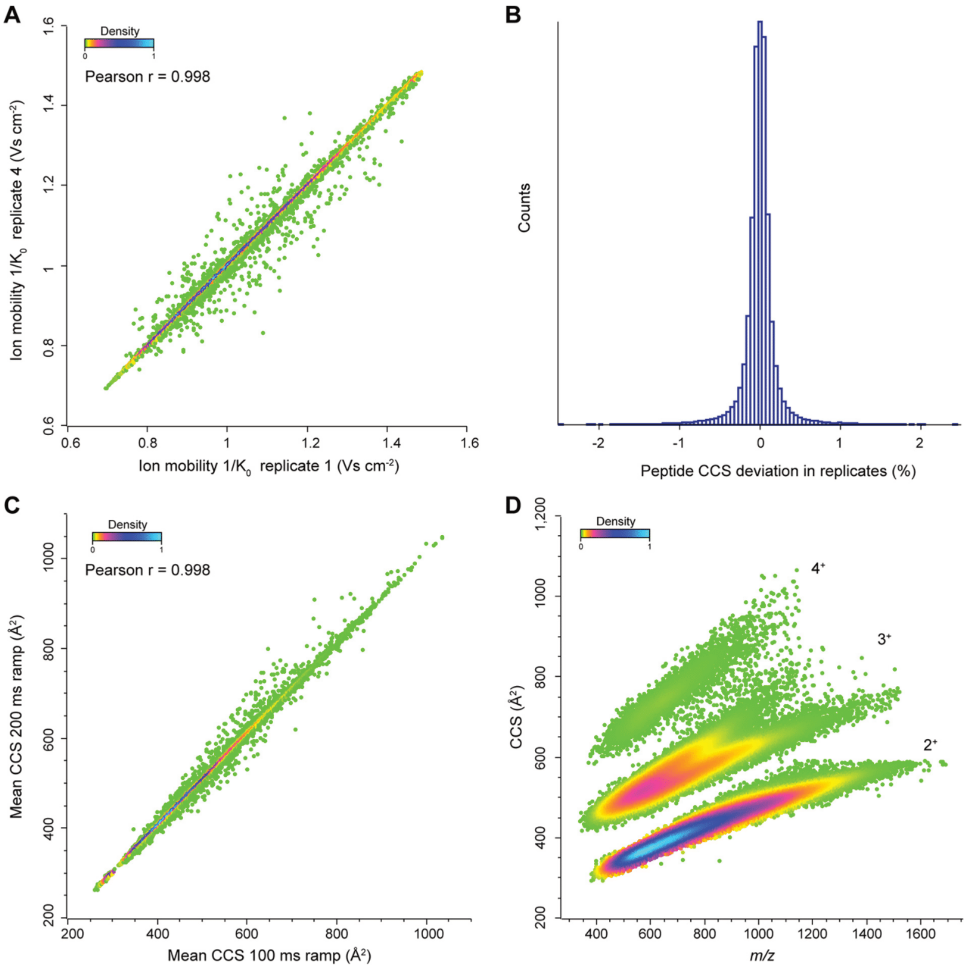


FIG. 7. **Large-scale and high-precision CCS measurements.** *A*, Pearson correlation of peptide ion mobilities in two replicate injections of a HeLa digest (100 ms TIMS ramps). *B*, Relative deviations of ^{TIMS}CCS values of all individual peptides from their mean of quadruplicate LC-MS runs (*n* = 144,363; 1,186 out of range). *C*, Pearson correlation of measured ^{TIMS}CCS values in two injections of a HeLa digest with different TIMS ramp times (100 and 200 ms TIMS ramps). *D*, Density distribution of 129,110 ^{TIMS}CCS values from human tryptic peptide ions as a function of *m/z*. The main populations are annotated with their respective charge states.

state, there is clear correlation between *m/z* and cross section and triply charged species split into two prominent subpopulations, as expected from the literature (49–51). However, the precision of the ^{TIMS}CCS determination is still more than 10-fold higher than the width of the ion mobility distribution for a given *m/z*. This results in additional peptide information that can be used for matching and identification.

DISCUSSION

Here, we have described the construction and evaluated the performance of a state-of-the-art quadrupole time-of-flight instrument with a trapped ion mobility device and deep integration of the PASEF principle. The novel Bruker *timsTOF Pro* successfully incorporates these building blocks in a robust and flexible manner, not only enabling shotgun-based PASEF operation but many other operation modes, which are still left to be explored.

The full implementation of PASEF in the hard- and firmware in an online format achieved results almost completely in line with those modeled and extrapolated from a laboratory prototype in our 2015 paper (31). This suggests that the physical operating principles are indeed directly translatable to proteomics workflows. In particular, the instrument routinely delivers sequencing rates above 100 Hz in complex proteome samples. In standard MS/MS acquisition schemes, such high fragmentation rates inevitably imply very short ion collection times and consequently poor spectrum quality. In contrast, PASEF leverages the full scan speed of TOF instruments with undiminished sensitivity as precursor ions are trapped and released as condensed ion packages by the time they are selected for fragmentation. This enabled the identification of about 6000 protein groups in single runs from a human cancer cell line with minimal input material, and with high quantitative accuracy.

Although we focused on label-free quantification in the current study, we expect that the high number of spectra per run will particularly benefit MS/MS-based quantification methods, for example isobaric labeling with TMT (52), iTRAQ (53) or EASI-tag (54). These approaches should additionally benefit from the ion mobility separation itself as it increases the purity of the isolation window and thereby reduces potential artifacts from co-eluting and co-isolated precursor ions.

The high speed and sensitivity of the *timsTOF Pro* allowed us to drastically decrease both measurement time and sample amount, which culminated in the identification of about 2100 proteins from only 10 ng HeLa digest in 30 min. This makes the instrument very attractive for proteomics studies with extremely low starting amounts, for example micro-dissected tumor biopsies, and for high throughput clinical applications of proteomics, in combination with robust and fast LC systems.

Finally, we demonstrated that TIMS-PASEF provides an efficient way to generate comprehensive libraries of peptide collisional cross sections, much beyond past reports (51). Such large-scale measurements could contribute to elucidating fundamental properties of modified and unmodified peptide ions in the gas phase and may eventually enable the *in silico* prediction of CCS values by deep learning algorithms. Furthermore, the very high precision of the CCS measurements with TIMS demonstrated here opens new avenues for spectral library-based identifications, in which the CCS parameter adds important evidence either on the MS level or, in data-independent acquisition strategies, also on the MS/MS level.

We conclude that the *timsTOF Pro* is a high-performance addition to the technology toolbox in proteomics, with many added opportunities enabled by TIMS-PASEF.

Acknowledgments—We thank our colleagues in the department of Proteomics and Signal Transduction and at Bruker Bremen and Bruker Billerica for discussion and help, in particular Drs. P. Geyer and I. Paron.

DATA AVAILABILITY

The raw datasets and the MaxQuant output files generated and analyzed throughout this study have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository (55) with the dataset identifier PXD010012 (https://www.ebi.ac.uk/pride/archive/). MaxQuant *msms.txt* and *.apl* files are provided for viewing of annotated spectra, including spectra of single peptide protein identifications.

* This work was partially supported by the German Research Foundation (DFG-Gottfried Wilhelm Leibniz Prize) granted to Matthias Mann and by the Max-Planck Society for the Advancement of Science.

S This article contains supplemental Figures and Table. The authors state that they have potential conflicts of interest regarding this work: S.K, H.K. M.L., M.K., N.G., J.D. M.P. and O. R. are employees of Bruker, the manufacturer of the timsTOF Pro. O.H. and N.B. are employees of Evosep. M.M. is an indirect investor in Evosep.

§§ To whom correspondence should be addressed: Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. Tel.: +49 89 8578-2557; Fax: +49 89 8578-2412; E-mail: mmann@biochem.mpg.de.

Author contributions: F.M., M.A.P., J.C., O.R., and M.M. designed research; F.M., A.-D.B., S.K., H.K., M.L., M.K., N.G., J.D., T.K., N.B., O.H., J.C., and O.R. performed research; F.M., A.-D.B., S.K., H.K., M.L., O.R., and M.M. analyzed data; F.M. and M.M. wrote the paper; N.B., O.H., and J.C. contributed new reagents/analytic tools.

REFERENCES

1. Altelaar, A. F. M., Munoz, J., and Heck, A. J. R. (2012) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48
2. Larance, M., and Lamond, A. I. (2015) Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* **16,** 269–280
3. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355
4. Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen, C. L., Nielsen, M. L., and Olsen, J. V. (2017) An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **4,** 587–599.e4
5. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793
6. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
7. Eliuk, S., and Makarov, A. (2015) Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8,** 61–80
8. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* **312,** 212–217
9. Han, X., Aslanian, A., and Yates, J. R. (2008) Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12,** 483–490
10. Beck, S., Michalski, A., Raether, O., Lubeck, M., Kaspar, S., Goedecke, N., Baessmann, C., Hornburg, D., Meier, F., Paron, I., Kulak, N. A., Cox, J., and Mann, M. (2015) The Impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol. Cell. Proteomics* **14,** 2014–2029
11. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., and Hill, H. H. (2008) Ion mobility-mass spectrometry. *J. Mass Spectrom.* **43,** 1–22
12. Cumeras, R., Figueras, E., Davis, C. E., Baumbach, J. I., and Gràcia, I. (2015) Review on Ion Mobility Spectrometry. Part 2: hyphenated methods and effects of experimental parameters. *Analyst* **140,** 1391–1410
13. May, J. C., and McLean, J. A. (2015) Ion mobility-mass spectrometry: time-dispersive instrumentation. *Anal. Chem.* **87,** 1422–1436
14. Eiceman, G. A., Karpas, Z., and Hill, H. H. J. (2013) *Ion Mobility Spectrometry* (CRC Press). 3rd Ed.
15. Valentine, S. J., Counterman, A. E., Hoaglund, C. S., Reilly, J. P., and Clemmer, D. E. (1998) Gas-phase separations of protease digests. *J. Am. Soc. Mass Spectrom.* **9,** 1213–1216
16. Srebalus Barnes, C. A., Hilderbrand, A. E., Valentine, S. J., and Clemmer, D. E. (2002) Resolving isomeric peptide mixtures: A combined HPLC/ion mobility-TOFMS analysis of a 4000-component combinatorial library. *Anal. Chem.* **74,** 26–36
17. Ewing, M. A., Glover, M. S., and Clemmer, D. E. (2015) Hybrid ion mobility and mass spectrometry as a separation tool. *J. Chromatogr. A*, 27–29

18. Lanucara, F., Holman, S. W., Gray, C. J., and Eyers, C. E. (2014) The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nat. Chem.* **6,** 281–294
19. Valentine, S. J., Plasencia, M. D., Liu, X., Krishnan, M., Naylor, S., Udseth, H. R., Smith, R. D., and Clemmer, D. E. (2006) Toward plasma proteome profiling with ion mobility-mass spectrometry. *J. Proteome Res.* **5,** 2977–2984
20. Baker, E. S., Livesay, E. A., Orton, D. J., Moore, R. J., Danielson, W. F., Prior, D. C., Ibrahim, Y. M., LaMarche, B. L., Mayampurath, A. M., Schepmoes, A. A., Hopkins, D. F., Tang, K., Smith, R. D., and Belov, M. E. (2010) An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies. *J. Proteome Res.* **9,** 997–1006
21. Geromanos, S. J., Hughes, C., Ciavarini, S., Vissers, J. P. C., and Langridge, J. I. (2012) Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Anal. Bioanal. Chem.* **404,** 1127–1139
22. Helm, D., Vissers, J. P. C., Hughes, C. J., Hahne, H., Ruprecht, B., Pachl, F., Grzyb, A., Richardson, K., Wildgoose, J., Maier, S. K., Marx, H., Wilhelm, M., Becher, I., Lemeer, S., Bantscheff, M., Langridge, J. I., and Kuster, B. (2014) Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell. Proteomics* **13,** 3709–3715
23. Distler, U., Kuharev, J., Navarro, P., Levin, Y., Schild, H., and Tenzer, S. (2014) Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **11,** 167–170
24. Cumeras, R., Figueras, E., Davis, C. E., Baumbach, J. I., and Gràcia, I. (2015) Review on Ion Mobility Spectrometry. Part 1: current instrumentation. *Analyst* **140,** 1376–1390
25. Fernandez-Lima, F. A., Kaplan, D. A., and Park, M. A. (2011) Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **82,** 126106
26. Fernandez-Lima, F., Kaplan, D. A., Suetering, J., and Park, M. A. (2011) Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* **14,** 93–98
27. Ridgeway, M. E., Lubeck, M., Jordens, J., Mann, M., and Park, M. A. (2018) Trapped ion mobility spectrometry: A short review. *Int. J. Mass Spectrom.* **425,** 22–35
28. Silveira, J. A., Ridgeway, M. E., Laukien, F. H., Mann, M., and Park, M. A. (2017) Parallel accumulation for 100% duty cycle trapped ion mobility-mass spectrometry. *Int. J. Mass Spectrom.* **413,** 168–175
29. Silveira, J. A., Ridgeway, M. E., and Park, M. A. (2014) High resolution trapped ion mobility spectrometry of peptides. *Anal. Chem.* **86,** 5624–5627
30. Ridgeway, M. E., Silveira, J. A., Meier, J. E., and Park, M. A. (2015) Microheterogeneity within conformational states of ubiquitin revealed by high resolution trapped ion mobility spectrometry. *Analyst* **140,** 6964–6972
31. Meier, F., Beck, S., Grassl, N., Lubeck, M., Park, M. A., Raether, O., and Mann, M. (2015) Parallel accumulation–serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14,** 5378–5387
32. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–324
33. Kulak, N. A., Geyer, P. E., and Mann, M. (2017) Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics*, mcp.O116.065136
34. Bache, N., Geyer, P. E., Bekker-Jensen, D. B., Hoerning, O., Falkenby, L., Treit, P. V., Doll, S., Paron, I., Müller, J. B., Meier, F., Olsen, J. V., Vorm, O., and Mann, M. (2018) A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell. Proteomics*, mcp.TIR118.000853
35. Stow, S. M., Causon, T. J., Zheng, X., Kurulugama, R. T., Mairinger, T., May, J. C., Rennie, E. E., Baker, E. S., Smith, R. D., McLean, J. A., Hann, S., and Fjeldsted, J. C. (2017) An interlaboratory evaluation of drift tube ion mobility-mass spectrometry collision cross section measurements. *Anal. Chem.* **89,** 9048–9055
36. Mason, E. A., and McDaniel, E. W. (1988) *Transport Properties of Ions in Gases* (John Wiley & Sons, Inc., New York, NY)
37. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
38. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
39. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction. *Mol. Cell. Proteomics*, M113.031591
40. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–740
41. R Development Core Team. (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria)
42. Michelmann, K., Silveira, J. A., Ridgeway, M. E., and Park, M. A. (2014) Fundamentals of trapped ion mobility spectrometry. *J. Am. Soc. Mass Spectrom.* **26,** 14–24
43. Silveira, J. A., Michelmann, K., Ridgeway, M. E., and Park, M. A. (2016) Fundamentals of trapped ion mobility spectrometry Part II: fluid dynamics. *J. Am. Soc. Mass Spectrom.* **27,** 585–595
44. Hernandez, D. R., Debord, J. D., Ridgeway, M. E., Kaplan, D. A., Park, M. A., and Fernandez-Lima, F. (2014) Ion dynamics in a trapped ion mobility spectrometer. *Analyst* **139,** 1913–1921
45. Volpe, P., and Eremenko-Volpe, T. (1970) Quantitative studies on cell proteins in suspension cultures. *Eur J Biochem* **12,** 195–200
46. Shvartsburg, A. A., and Smith, R. D. (2008) Fundamentals of traveling wave ion mobility spectrometry. *Anal. Chem.* **80,** 9689–9699
47. Buryakov, I. A., Krylov, E. V., Nazarov, E. G., and Rasulev, U. K. (1993) A new method of separation of multi-atomic ions by mobility at atmospheric pressure using a high-frequency amplitude-asymmetric strong electric field. *Int. J. Mass Spectrom. Ion Process.* **128,** 143–148
48. Counterman, A. E., and Clemmer, D. E. (2002) Cis-Trans Signatures of Proline-Containing Tryptic Peptides in the Gas Phase. *Anal. Chem.* **74,** 1946–1951
49. Valentine, S. J., Counterman, A. E., and Clemmer, D. E. (1999) A database of 660 peptide ion cross sections: use of intrinsic size parameters for bona fide predictions of cross sections. *J. Am. Soc. Mass Spectrom.* **10,** 1188–1211
50. Lietz, C. B., Yu, Q., and Li, L. (2014) Large-scale collision cross-section profiling on a traveling wave ion mobility mass spectrometer. *J. Am. Soc. Mass Spectrom.* **25,** 2009–2019
51. May, J. C., Morris, C. B., and McLean, J. A. (2017) Ion mobility collision cross section compendium. *Anal. Chem.* **89,** 1032–1044
52. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003) Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75,** 1895–1904
53. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
54. Virreira Winter, S., Meier, F., Wichmann, C., Cox, J., Mann, M., and Meissner, F. (2018) EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15,** 527–530
55. Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D56

3.1.2. Article 2: Multi-omics profiling of living human pancreatic islet donors

Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories toward type 2 diabetes

Nature metabolism, (Accepted)

Leonore Wigger[1, *], Marko Barovic[2, 3, 4, *], **Andreas-David Brunner[5, *],** Flavia Marzetta[1], Eyke Schöniger[2, 3, 4], Florence Mehl[1], Nicole Kipke[2, 3, 4], Daniela Friedland[2, 3, 4], Frederic Burdet[1], Camille Kessler[1], Mathias Lesche[6], Bernard Thorens[7], Ezio Bonifacio[3, 4, 8], Cristina Legido Quigley[9], Philippe Delerive[10], Andreas Dahl[6], Kai Simons[11], Daniela Aust[12, 13], Jürgen Weitz[14], Marius Distler[14], Anke M Schulte[15], Matthias Mann[5, #], Mark Ibberson[1, #], Michele Solimena[2, 3, 4, #]

** These authors contributed equally to this work*

# Correspondence

[1] Vital-IT, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[2]Department of Molecular Diabetology, University Hospital and Faculty of Medicine, TU Dresden, Dresden, Germany

[3]Paul Langerhans Institute Dresden (PLID), Helmholtz Center Munich, University Hospital and Faculty of Medicine, TU Dresden, Dresden, Germany

[4]German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany

[5]Max Planck Institute of Biochemistry, Martinsried, Germany

[6]Deep Sequencing Facility, CMCB Technology Platform, Dresden, Germany

[7]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

[8]DFG Center for Regenerative Therapies Dresden, Medical Faculty, TU Dresden, Dresden, Germany

[9]Steno Diabetes Center Copenhagen, Gentofte, Denmark

[10]Institut de Recherches Servier, Pôle d'Innovation Thérapeutique Métabolisme, Suresnes, France

[11]Lipotype GmbH, Dresden, Germany

[12]Department of Pathology, Medical Faculty, University Hospital Carl Gustav Carus, TU Dresden, Dresden

[13]NCT Biobank Dresden, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany

[14]Department of Visceral, Thoracic and Vascular Surgery, University Hospital Carl Gustav Carus, Medical Faculty, TU Dresden, Dresden, Germany

[15]Sanofi-Aventis Deutschland GmbH, Diabetes Research, Industriepark Höchst, Frankfurt am Main, Germany

Contribution

I contributed to the overall experimental and analytical design of the study and to the writing of the manuscript. I established and optimized the miniaturized sample preparation workflow in close consultation with the researchers isolating the pancreatic islets via laser capture microdissection in Dresden. Furthermore, I performed all proteomics experiments, analyzed accompanying data, performed RNA-sequencing to proteomics comparisons, prepared figures and wrote the proteomics part of the manuscript.

Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories toward type 2 diabetes

Leonore Wigger[1*], Marko Barovic[2,3,4*], Andreas-David Brunner[5*], Flavia Marzetta[1], Eyke Schöniger[2,3,4], Florence Mehl[1], Nicole Kipke[2,3,4], Daniela Friedland[,3,4], Frederic Burdet[1], Camille Kessler[1], Mathias Lesche[6], Bernard Thorens[7], Ezio Bonifacio[3,4,8], Cristina Legido Quigley[9], Philippe Delerive[10], Andreas Dahl[6], Kai Simons[11], Daniela Aust[12,13], Jürgen Weitz[14], Marius Distler[14], Anke M Schulte[15], Matthias Mann[5#], Mark Ibberson[1#], Michele Solimena[2,3,4#]

[1]Vital-IT, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; [2]Department of Molecular Diabetology, University Hospital and Faculty of Medicine, TU Dresden, Dresden, Germany; [3]Paul Langerhans Institute Dresden (PLID), Helmholtz Center Munich, University Hospital and Faculty of Medicine, TU Dresden, Dresden, Germany; [4]German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany; [5]Max Planck Institute of Biochemistry, Martinsried, Germany; [6]Deep Sequencing Facility, CMCB Technology Platform, Dresden, Germany; [7]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [8]DFG Center for Regenerative Therapies Dresden, Medical Faculty, TU Dresden, Dresden, Germany; [9]Steno Diabetes Center Copenhagen, Gentofte, Denmark; [10]Institut de Recherches Servier, Pôle d'Innovation Thérapeutique Métabolisme, Suresnes, France; [11]Lipotype GmbH, Dresden, Germany; [12]Department of Pathology, Medical Faculty, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany; [13]NCT Biobank Dresden, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany; [14]Department of Visceral, Thoracic and Vascular Surgery, University Hospital Carl Gustav Carus, Medical Faculty, TU Dresden, Dresden, Germany; [15]Sanofi-Aventis Deutschland GmbH, Diabetes Research, Industriepark Höchst, Frankfurt am Main, Germany.

* equal contribution

corresponding authors

Abstract

Existing studies do not sufficiently describe the molecular changes of pancreatic islet beta cells leading to their deficient insulin secretion in type 2 diabetes (T2D). Here we address this deficiency with a comprehensive multi-omics analysis of metabolically profiled pancreatectomized living human donors stratified along the glycemic continuum from normoglycemia to T2D. Islet pools isolated from surgical samples by laser-capture microdissection had remarkably heterogeneous transcriptomic and proteomic profiles in diabetics, but not in non-diabetic controls. Transcriptomics analysis of this unique cohort revealed islet genes already dysregulated in prediabetic individuals with impaired glucose tolerance. Our findings demonstrate a progressive but disharmonic remodeling of mature beta cells, challenging current hypotheses of linear trajectories toward precursor or trans-differentiation stages in T2D. Further, integration of islet transcriptomics and pre-operative blood plasma lipidomics data enabled us to define the relative importance of gene co-expression modules and lipids positively or negatively associated with HbA1c levels, pointing to potential prognostic markers.

Introduction

Type 2 diabetes (T2D) mellitus collectively defines a cluster of genetically complex pathological states characterized by persistent hyperglycemia, often leading to cardiovascular complications, kidney failure, retinopathy and neuropathies. Affecting more than 450 million people, with rising incidence rates over the past decades, this syndrome is a major threat for public health and society globally[1]. Common determinant and ultimate cause of T2D is the inability of pancreatic islet beta cells to secrete insulin in adequate amounts relative to insulin sensitivity, in the absence of evidence for their autoimmune destruction or a monogenetic deficit. Beta cell failure typically results from a lengthy process spanning many years. Remarkably, however, it can be rapidly reverted upon bariatric surgery or severe caloric restriction[2,3]. These observations argue against the occurrence of major beta cell apoptosis in T2D, especially since adult beta cells hardly replicate, while robust evidence of beta cell neogenesis after puberty is also lacking. Hence, the prevailing opinion is that persistent metabolic stress drives mature beta cells to phenotypically de-differentiate into progenitor cells or trans-differentiate into other islet endocrine cell types over time[4–6]. As the pathogenesis of beta cell dysfunction in T2D remains largely unclear, the diagnosis of this disease relies on accepted, but still surrogate parameters and cutoffs that have been primarily developed for clinical practice to optimize therapeutic interventions[7].

Insight into molecular alterations associated with impaired insulin secretion in T2D has been largely obtained from pancreatic islets isolated enzymatically from brain-dead or cadaveric subjects classified according to a categorical division into non-diabetic and diabetic, rather than on a continuum from euglycemia to steady hyperglycemia. This approach has multiple shortcomings[8]. Briefly, islet researchers do not generally have access to extensive clinical and laboratory information about the donors prior to their admission to an intensive therapy unit[9]. Moreover, the islet state is perturbed by the metabolic stress associated with a terminal condition and the related pharmacological treatments[10,11]. Enzymatic isolation of islets and their in vitro culture can further change their molecular profile[12,13]. In the attempt to overcome, at least in part, these limitations, we established a complementary platform for the procurement of islets which relies on the collection and analysis of pancreatic specimens from metabolically profiled living donors undergoing pancreatectomy for a variety of disorders[8,14]. We showed that this approach is very reproducible and scalable and provides a novel view on transcriptomic and functional alterations in pancreatic islets of subjects with T2D[15–17].

The aim of the present study has been to profile in greater detail gene expression changes occurring along the progression from euglycemia to long-standing T2D in human islets *in situ* and to integrate this knowledge with clinical traits, circulating lipid levels and the islet proteome, hence enabling inferences about the mechanisms driving islet dysfunction and the identification of potential biomarkers for it.

Results

Recruitment of a large cohort of living donors for islet and plasma omics data

To gain insight into the history of islet cell deterioration along the progression from normal glycemic regulation to T2D, we collected surgical pancreatic tissue samples from 133 metabolically phenotyped pancreatectomized patients (PPP). Eighteen were non-diabetic (ND), 41 had impaired glucose tolerance (IGT), 35 Type 3c Diabetes (T3cD) and 39 T2D (Fig. 1A and Fig. 1B). These group assignments were based on glycemic values at fasting and at the 2 h time point of an oral glucose tolerance test (OGTT) using the thresholds defined in the guidelines of the American Diabetes Association[7], or, when applicable, on a previously established diagnosis of T2D. In this cohort, 51.9% were males and the mean age was 65.36±11.54 years, with ND PPP being on average younger than the other three groups (Fig. 1C and Supplementary Table S1). The body mass index (BMI) was significantly lower in ND compared to IGT, T3cD and T2D PPP. The HbA1c value, as a parameter of longer-term glycemia, was 5.25±0.3 in ND, 5.75±0.42 in IGT, 6.29±0.95 in T3cD and 7.41±1.29 in T2D PPP (Fig. 1C and Supplementary Table S1). Moreover, based on histopathology, malignant tumors occurred in 50%, 60.97%, 74.29% and 69.23% of ND, IGT, T3cD, and T2D PPP, respectively (Supplementary Table S1).

Pancreatic islet gene expression and pathways drift progressively with glycemia deterioration

Gene expression profiles of islets isolated by laser capture microdissection (LCM) from resected and snap-frozen pancreas samples of ND, IGT, T3cD and T2D PPP were assessed by RNA sequencing. After removal of genes with low expression levels, the overall islet transcriptome encompassed 19,119 genes, of which 14,699±693 were present (raw read counts >0) in ND PPP, 14,967±455 in IGT PPP, 14,939±493 in T3cD PPP and 14,997±428 in T2D PPP. Genes with a fold change (FC)>1.5 and a false discovery rate (**FDR**)≤0.05 were considered to be differentially expressed (**DE**) between the groups. Pairwise group comparisons of IGT vs. ND, T3cD vs. ND and T2D vs. ND revealed an exacerbation of gene dysregulation with deterioration of glycemic control (Fig. 2A). Notably, no DE islet genes were identified between IGT vs. ND, while 161 and 650 DE genes were found between T3cD vs. ND and T2D vs. ND, respectively (Fig. 2A and Supplementary Table S2).

Restricting the transcriptomic analysis to libraries in which insulin (*INS*) was the most expressed gene resulted in the retention of islet datasets from 15 ND, 35 IGT, 21 T3cD and 24 T2D subjects, without dramatically affecting the overall composition of the cohort in regards to diabetes status and major descriptive parameters (Supplementary Table S3). Deconvolution analysis indicated that in 97.8% of retained samples the proportion of beta cells was >50% (Supplementary Fig. S1), supporting the choice of this strategy to discriminate samples especially enriched in beta cell transcripts. Despite the expected reduction in statistical power due to ~ 30% smaller size of this "restricted" cohort (92 samples retained from 133), the number of DE genes between islets of T2D vs. ND PPP increased by 51% to 984 (782 up, 202 down), and by 59% to 256 (209 up, 47 down) between islets of T3cD vs. ND PPP (Fig. 2A, Supplementary Table S4). Seven of the 984 DE genes are known risk genes for T2D, two upregulated (*SGSM2* and *BCL2*) and five downregulated (*RASGRP1, G6PC2, SLC2A2, ZMAT4* and *PLUT*)[18], while most of the remaining genes have not been previously reported to be altered in islets of subjects with T2D[14,19].

Among the DE genes in islets of T2D PPP, *INF2* and *AKR7L* were negatively correlated in a moderate fashion with duration of the disease measured in years (Spearman correlation coefficient -0.32 and -0.41 respectively), albeit they were both upregulated relative to islets of ND PPP. Most notably, this filtering step enabled, for the first time, the identification of 185 DE genes between islets of IGT vs. ND PPP. Most of these DE genes were upregulated (181/185), and 98 also dysregulated with the same directionality (97 up, 1 down) between islets of T2D vs. ND PPP. Intriguingly, and apparently at variance with previous eQTL findings[20], the T2D risk gene *ARAP1* and its neighboring gene *STARD10* were both upregulated and among the 77 genes dysregulated in islets of IGT PPP only. No islet cell type specific genes[21] were enriched in any of the differential expression analyses. Furthermore, no shift of islet cell type proportions with the progression of the disease was observed in the deconvolution analysis (Supplementary Fig. S1A).

For both the "restricted" and the full data set, heatmaps of gene expression levels in the four patient groups were prepared as a visual complement to the statistical analysis (Fig. 2B and Supplementary Fig. S2A). Despite the marked differences between the findings in the "restricted" and complete cohort, upregulation prevailed as the direction of gene dysregulation in both of them (Fig. 2A and Supplementary Fig. S2A). Based on these observations, pancreatic tissue sections of 5 ND and 5 T2D PPP with the "restricted" cohort were immunostained with antibodies specific for histone H3 and H4 lysine acetylation – an epigenetic modification associated with greater access of transcription factors to promoter sites resulting in increased gene expression. Notably, the immunoreactivity for both acetylated histones was remarkably increased in the islets, and also in the surrounding exocrine cells of T2D PPP, and to a lesser extent IGT PPP (not shown), compared to ND PPP (Fig. 2D).

Extracellular matrix and mitochondrial pathways are perturbed in T2D and IGT subjects

We further analyzed differentially expressed gene functions by gene set enrichment analysis using Gene Ontology terms and KEGG pathways (Fig. 2C, Supplementary Fig. S2B and Supplementary Tables S5 and S6). Results obtained from the different gene set collections cross-validated each other, since similar biological themes emerged. Islets of pre-diabetic and diabetic subjects displayed upregulation of islet genes that were functionally related to cell-extracellular matrix interaction, immune response and signaling pathways, while expression of genes related to RNA processing, protein translation and mitochondrial oxidative phosphorylation were downregulated. Importantly, the analysis performed on the "restricted" cohort, differently from the full dataset, also revealed that the strength of the enrichment increased with progression of the disease (Fig. 2C and Supplementary Fig. S2B). These data suggest that early dysregulation of gene pathways exacerbates with the decline of beta cell function.

Weighted gene co-expression network analysis identifies islet gene modules correlated with the elevation of HbA1c

To globally interpret transcriptomic data and identify sets of genes likely to be functionally related and co-regulated, we grouped genes based on similarities in their expression profiles into modules using a network-based approach[22]. In the cohort of 133 PPP, we identified 36 co-expressed gene modules, which were arbitrarily labeled M1 through M36. The expression profiles of the genes in each module were summarized by a module eigengene, or first principal component of the expression matrix. Module eigengenes were used to computationally relate modules to one another and to genes or clinical variables. Correlation between module eigengenes and diabetes-related clinical traits revealed modules M9 and M14 as those with the highest positive and negative correlation with HbA1c, respectively (Fig. 3A and Supplementary Table S7). The former consisted of a set of genes that showed similar patterns of increased expression in most PPP with T2D (Fig. 3B), while the latter was mostly composed of genes with coordinated down-regulation in diseased subject samples (Fig. 3C).

We next evaluated how close a gene was to a given module, denoted as module membership, by correlating its expression profile with the module eigengene. This analysis allowed us to identify highly connected genes or "hub" genes for HbA1c-related modules (Fig. 3D-E). These included genes that we had previously identified as differentially expressed in subjects with T2D[14,15], and which were correlated with HbA1c either positively, such as module M9 genes *ALDOB* (FC=8.45 with adj. *p*<0.001 in T2D vs. ND in "restricted" cohort) and *FAIM2* (FC=7.11 with adj. *p*<0.001 in T2D vs. ND in "restricted" cohort) or negatively, such as module M14 genes *SLC2A2* (FC=-2.77 with adj. *p*<0.001 in T2D vs. ND in "restricted" cohort) and *TMEM37* (FC=-1.73 with adj. *p*<0.001 in T2D vs. ND in "restricted" cohort). Interestingly, we (Supplementary Fig. S3A) and others[23] found *ALDOB* to be upregulated in islets from 13-week-old diabetic *db/db* mice compared to the heterozygous *db/+* littermate (Supplementary Fig. S3A) as well as in a mouse beta, but not alpha, cell line upon exposure to high glucose (Supplementary Fig. S3B). However, the overexpression of *ALDOB* in beta cells of T2D PPP could neither be verified by in situ hybridization using the RNAScope platform (data not shown), nor by immunofluorescence on tissue sections due to the cross-reactivity of the available anti-ALDOB antibody with other aldolase isoforms (Supplementary Fig. S3C).

Proteomics of LCM-isolated pancreatic islets reveals heterogenous profiles of T2D subjects and extends target identification

To verify and extend the transcriptomic data at the functional level of proteins, we analyzed the mass spectrometry (MS)-based proteomic profiles of LCM pancreatic islets from five ND and five T2D PPP (Supplementary Table S8). Using a very high sensitivity workflow on a novel trapped-ion mobility Time of Flight mass spectrometer[24], we identified 2,237±499 islet proteins for ND PPP and 1,819±412 islet proteins for T2D PPP (Figure 4A). Quantitative reproducibility between biological replicates was high with Pearson correlations ranging from 0.83 to 0.95 (Supplementary Fig. S4A). Principal component analysis (PCA) clustered the data into two distinct groups matching the clinical stratification (Fig. 4B). Interestingly, islets of ND PPP clustered closely, indicating a very similar proteome signature, while those of T2D PPP revealed substantial proteome heterogeneity among each other. Differential expression analysis confirmed that islets of T2D and ND PPP have very different proteomic profiles. The main differential drivers are well-characterized markers of pancreatic islet cells, including SLC2A2[25], and many proteins implicated in mitochondrial structure, translation, energy supply and amino acid or fatty metabolism such as YMEL1, MRPL12, BA3(C14orf159), ACADS and its paralogue ACADSB, which were highly depleted in islets of T2D PPP (Fig. 4C). Besides AKR7L, ACADS was the only other upregulated and differentially expressed gene in islets of both IGT and T2D PPP, while being also downregulated at the protein level. All differentially expressed mitochondrial proteins are encoded by the nuclear genome (Fig. S4B). Intriguingly, the level of the sulfonylurea receptor ABCC8 subunit[26] was also strongly reduced in islets of T2D PPP. This downregulation might be an effect secondary to pharmacological treatment, as three among these patients had been treated with anti-diabetic SUR1 antagonists glibenclamide (DP197), glimepiride (DP118) or mitiglinide (DP087) (Supplementary Fig. S4C). We found the glycolytic enzyme ALDOB to be on average four-fold upregulated in islets of T2D vs. ND PPP. This is consistent with our transcriptomic data (ALDOB FPKM: 76.16±50.82 in T2D PPP vs. 4.63±0.95 in ND PPP; p=0.03) and with previous[14,15] and our current WGCNA analyses. Other proteins robustly overexpressed in islets of T2D PPP included the alpha-L-fucosidase FUCA1 and the surface marker for hematopoietic stem cells THY1.

Next, we employed the proteomic ruler algorithm and annotations of subcellular localization to compare the protein mass distribution of major cellular compartments[27] (Fig. 4D). Islets of T2D PPP lost an estimated protein mass of 6% in the Golgi apparatus, 24% in the endoplasmic reticulum, and 27% in the mitochondria compared to those of ND PPP, while cytoskeleton protein mass was unchanged. Unsupervised hierarchical clustering of all 2,622 detected proteins, clustered the data according to clinical categories (Fig. 4E). One-dimensional gene ontology enrichment[28] revealed two distinct clusters whose protein intensity levels associated with the terms 'membrane attack complex' (p<2.18E-04) and 'Immunoglobulin C-domain' (p<2.68E-06) were enriched by 2.27-fold and 2.36-fold in islets of T2D vs. ND PPP, respectively. Proteins with the gene ontology-term 'differentiation' (p<3.09E-04) and 'mitochondrion' (p<2.19E-08) were 1.65 and 1.78-fold in islets of ND PPP.

T2D patients show decreased levels of plasma phospholipids and elevated levels of plasma (dihydro-)ceramides.

Our study encompassed two independently generated lipidomics data sets. First, shotgun lipidomics was performed on peripheral blood plasma samples of the aforementioned cohort (4 ND, 21 IGT and IFG, 13 T3cD and 17 T2D) (Supplementary Tables S9 and S10). Second, sphingolipid profiling was performed on peripheral blood samples of subjects within the cohort subjected to transcriptomic analysis (11 ND, 32 IGT and IFG, 26 T3cD and 32 T2D) (Supplementary Tables S11 and S12). Prior to data analysis, lipidomics samples from PPP with very high bilirubin values (>100 µmol/l) were removed to avoid bias in lipidomics profiles. All available samples from non-diabetic PPP (ND, as previously defined) and the subset of IGT PPP with HbA1c<6.0 were combined into one group, which is referred to here as ND for readability.

In shotgun lipidomics, 113 lipid species from 11 classes were included in the data analysis. When comparing T2D and T3cD to ND PPP, the majority of lipid classes displayed a remarkably homogeneous downward-trend of the individual lipid species they comprised (Fig 5A-B). Most prominently, plasma concentrations of lipids within the phosphatidylcholine (PC O-) class, a large class with 30 measured species, were lower in T2D versus ND PPP. Sixteen lipids of this class were significantly decreased (adjusted p<0.05) after adjusting for age and sex differences, with all of them showing at least a 1.4-fold change. Two lipid species from two smaller phospholipid classes (lysophosphatidylcholines (LPC) and phosphatidylinositols (PI)), and one from the sphingomyelin class (SM), were also significantly less abundant in T2D than in ND PPP (LPC 18:0;0: FC=-1.54, adj. p=0.03; PI 18:0;0/18:2;0: FC=-1.36, adj. p=0.04; SM 34:1;2:, FC=-1.24, adj. p=0.04). (Fig. 5A-B and Supplementary Table S13).

Next, we performed targeted sphingolipidomics on 14 distinct lipid species for very accurate plasma level estimation (ceramides, dihydroceramides and sphingoid bases). Plasma levels of ceramides d18:1/18:0 and d18:1/20:0 were increased in T2D compared to ND PPP (Cer d18:1/18:0: FC=1.34, p=0.02; Cer d18:1/20:0: FC=1.22, p=0.01). Of note, a similar trend towards elevation in T2D vs ND was also observed in the two dihydroceramide species having the same chain lengths as these ceramides (DH Cer d18:0/18:0: FC=1.44, p=0.05; DH Cer d18:0/20:0: FC=1.35, p=0.01). Thus, in our data set, plasma concentrations of ceramides and their precursor dihydroceramides appear to increase simultaneously in T2D.

Integrative data modelling identifies cell-matrix interaction, cell signaling and immune response as key pathways linked to pancreatic islet dysfunction

To identify a multivariate molecular profile that explains diabetes progression in the PPP cohort, we performed a large-scale integrative multi-omics analysis combining clinical data with islet transcriptomics and plasma lipidomics. Integration of transcriptomics and lipidomics data in the same model enables to weigh the relative importance of lipid and gene expression features in relationship to a chosen clinical trait. Hence, we explored the relationship between gene co-expression modules and plasma lipids by computing a consensus orthogonal partial least square (consensus OPLS)[29,30] model with HbA1c as the outcome. All three types of biological data, namely gene co-expression modules, lipids from shotgun analysis and sphingolipids from targeted analysis, contributed to the model (35%, 46.5% and 18.5%, respectively), suggesting that they help to explain HbA1c levels in a complementary way. Among them, different lipids and gene modules appear as the most relevant variables in the statistical modelling of HbA1c levels (Fig. 6A, 6B and Supplementary Table S14). Importantly, the model explained a large portion of data variance, highlighting a good fit with the experimental data (see Methods for more details). Among all considered biological data, the co-expression modules M1, M4, M8, M9, M30, M35 and M36 were the top predictive variables for high HbA1c levels, along with the two ceramide species C20 and C18. TAGs were also contributing, although to a lesser extent (Fig 6A, right hand side). Conversely, low levels of HbA1c were strongly related to the co-expression modules M12 and M14 (Fig 6A, left hand side). However, the majority of the predominant predictors for low HbA1c were lipid species, most importantly the PC O-class. This class was also found to be lower in T2D compared to ND patient groups in differential abundance analysis, as shown in Fig 5A. A number of SM, PI and PC lipid species were next in the importance ranking related to low HbA1c, followed by the gene co-expression module M29. These results suggest that the profile of patients with increased HbA1c is characterized by multiple molecular components, some of which represent signals that were not captured by differential abundance analyses comparing diabetes status groups nor by correlating gene co-expression modules individually to HbA1c. Most importantly, consensus OPLS multi-omics analysis pointed towards additional gene co-expression modules that may play a role in glucose dysregulation.

Next, we used the results from the integrative data modelling to infer a network of key altered biological pathways in dysfunctional beta cells. To this end, we pooled gene modules positively associated with HbA1c levels (M1, M4, M8, M9, M30, M35 and M36) (Fig. 6A) and assessed their overlap to KEGG pathways by over-representation analysis. We found that the biological themes underlying these genes were very similar to the pathways upregulated in T2D and IGT PPP and include cell-matrix interaction, cell signaling and immune response (Fig. 6C and Supplementary Table S15). The same strategy was used to identify pathways associated with genes from modules with a negative prediction score for HbA1c (M12, M14 and M29) (Fig. 6A), revealing an enrichment for metabolic pathways (Fig. 6C and Supplementary Table S15).

Of note, several islet genes dysregulated in T2D PPP were driving the enrichment of these pathways. These include, for example, ***ALDOB***, which stood out for its strong correlation to HbA1c levels (Fig. 3D and Fig. 6C). These genes, or the proteins encoded by them, should be regarded as putative candidate biomarkers for monitoring disease progression and therapeutic intervention.

Discussion

This study provides the most extensive dataset on islets *in situ* and plasma samples from the largest cohort of in-depth metabolically profiled living donors. Multi-omics data were generated using state-of-the-art approaches and integrated in a fashion not previously used in studies on islet dysregulation in relation to hyperglycemia in humans. Our transcriptomic and proteomic data from islets *in situ* of ND subjects represent a valuable reference for future investigations. Furthermore, we could identify for the first time a set of islet genes altered in their expression already in subjects with impaired glucose tolerance. This, in turn, enabled us to acquire an unprecedented cross-sectional overview of the progression of islet gene dysregulation in parallel with the continuous elevation of HbA1c values, beyond conventional thresholds for clinical classification of patients.

Pathways involved in RNA biology and especially in mitochondrial function emerged to be most negatively perturbed - a conclusion which in the case of the latter was strongly corroborated by the proteomic analysis, which enabled the identification of known and unknown differentially expressed proteins in islets of T2D PPP. In this context, we emphasize the downregulation of mitochondrial ACADS and its paralogue ACADSB, which catalyze the beta oxidation of short-chain fatty acids, including sodium butyrate. This finding is intriguing in view of the ability of this metabolite to broadly upregulate gene expression through inhibition of histone deacetylases. Unlike in previous studies on isolated islets from brain-dead organ donors[14,18], the vast majority of differentially expressed genes in islets of T2D, but also IGT and T3cD PPP were upregulated. Among those genes, *ALDOB* stands out being the one with the strongest correlation with the islet gene module M9, which in turn has the strongest correlation with elevated HbA1c. Since *ALDOB* is a marker of beta cell precursors, its overexpression could be interpreted as a sign that in T2D, mature beta cells revert back to an immature stage of differentiation, or that a compartment equivalent to the lifelong niche of virgin beta cells identified in adult mice[31] expands as a potential compensatory source of new beta cells. However, no additional disallowed gene of immature beta cells, markers of beta cell precursors or other islet cell types were dysregulated, while key determinants of mature beta cells, such as *PDX1*, *MAFA*, *NKX6.1* or *UCN3* were unchanged, at least at the transcriptomic level. Retention of fractions of major islet cell types (alpha, beta and delta) within the islet in T2D, consistent with recent imaging studies in samples from pancreatectomized subjects[17], was confirmed by deconvolution analysis. Our global unbiased proteomic analysis, which corroborated the upregulation of

ALDOB, further showed that the expression profile of islet cells in T2D PPP is very divergent, opposite to its remarkable homogeneity in islet cells of ND subjects. Hence, the regression of beta cells toward a de-differentiated state following a linear trajectory recapitulating their developmental path to maturation or their transdifferentiation into other islet cell types seems less likely than a disharmonic relaxation of constraints on gene expression. Such processes, although possibly reversible, could perturb the coordinated operation of islet cells, including beta cells. In line with this, Lawlor ***et al.*** reported no evidence of beta cell dedifferentiation/transdifferentiation and alterations in fractions of islet cells in the context of T2D upon sequencing of single islet cells from a small cohort of ND and T2D organ donors[32]. For the future it would be important to assess whether overexpression of ALDOB occurs indeed in beta cells and if it affects their glycolysis and metabolism, taking into account that its paralogue ALDOA, whose RNA and protein levels were unchanged, remains by far the predominant islet aldolase species. Attention may also be directed toward understanding whether impaired oxidative phosphorylation, as a likely outcome of the massively decreased expression of mitochondrial proteins, and thus energy balance homeostasis, accounts, at least in part, for the observed less restrained gene expression.

Similar to findings in other population-based studies on T2D[33,34], PC O- and LPC lipids were altered in our cohort of T2D PPP, thus supporting the general implications of our observations. In particular, we found that more than half of the PC O- class lipids (16 out of 30) and two of six LPC lipids were lower in T2D compared to ND PPP. In the present study we also found that several ceramides and dihydroceramides are elevated in T2D vs. ND, and whilst these increases were modest, these findings are consistent with those observed in several other recent studies[35–37], highlighting the importance of these lipids as potential biomarkers of beta cell function in T2D.

Finally, we use a data fusion method[29,30] to generate a model of how different molecular features (islet gene co-expression, plasma shotgun lipidomics and targeted sphingolipidomics) contribute to HbA1c levels in a continuum from healthy individuals to those with overt T2D. This model allowed us to measure the ***relative*** importance of different molecular components in explaining HbA1c variability, providing unique insights into the molecular profiles of individuals as they lose glycemic control towards development of T2D. To our knowledge this is the first time such an approach has been used in this field and we suggest that, by modelling multiple levels of information at the same time in deeply phenotyped populations such as the one presented here, we can gain a holistic view of the system and draw conclusions regarding key pathways, targets and biomarkers in metabolic and other diseases.

Data availability

RNA Sequencing data was deposited in the GEO database with GEO accession number (to be provided once the deposition process is completed)

The proteomics raw datasets and the MaxQuant output files generated and analyzed throughout this study were deposited at the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD022561 (https://www.ebi.ac.uk/pride/archive/).

Lipidomics data will be made publicly available shortly.

Acknowledgement

We wish to thank Leif Groop, Emma Ahlqvist, Stephan Speier and Triantafyllos Chavakis for discussion; Katja Pfriem for administrative assistance. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115881 (RHAPSODY). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. This work is further supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0097-2. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies. Work in the Solimena lab is also supported with funds from the German Ministry of Education and Research to the German Center for Diabetes Research (DZD).

Author contributions

J.W. and M.D., patient recruitment and surgery, provision of clinical data; E.S., N.K. and D.F., sample collection and processing, data entry; D.A., pathology; M.B., N.K. and E.S., patient database management and selection; A-D.B. and M.M., proteomics; M.L., A.D., RNA sequencing, C.L.Q., P.D., K.S., lipidomics and sphingolipidomics; L.W., M.B., A-D.B., F.Ma., F.Me., F.B. and C.K., analysis and integration of multi-omics data; E.B., autoantibody test; A.S., data in mouse tissue and cell lines; M.B., immunofluorescence stainings and antibody validation; B.T., D.A., J.W., A.S., M.M., M.I. and M.S., conceptual insights and provision of funds; L.W., M.B., A-D.B., F.Ma., F.Me., A.S., M.I., M.M. and M.S., writing of the manuscript. All authors read, revised and approved the final version of the manuscript.

Competing interests

The authors declare no conflicts of interest.

References

1. Saeedi, P. ***et al.*** Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **157**, (2019).

2. Mizera, M. ***et al.*** Type 2 Diabetes Remission 5 Years After Laparoscopic Sleeve Gastrectomy: Multicenter Cohort Study. *Obes. Surg.* 1–7 (2020). doi:10.1007/s11695-020-05088-w

3. Lim, E. L. ***et al.*** Reversal of type 2 diabetes: Normalisation of beta cell function in association with decreased pancreas and liver triacylglycerol. *Diabetologia* **54**, 2506–2514 (2011).

4. Talchai, C., Xuan, S., Lin, H. V., Sussel, L. & Accili, D. Pancreatic β cell dedifferentiation as a mechanism of diabetic β cell failure. *Cell* **150**, 1223–1234 (2012).

5. Wang, Z., York, N. W., Nichols, C. G. & Remedi, M. S. Pancreatic β cell dedifferentiation in diabetes and redifferentiation following insulin therapy. *Cell Metab.* **19**, 872–882 (2014).

6. Cinti, F. ***et al.*** Evidence of β-Cell Dedifferentiation in Human Type 2 Diabetes. *J. Clin. Endocrinol. Metab.* **101**, 1044–1054 (2016).

7. American Diabetes Association. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes-2020. *Diabetes Care* **43**, S14–S31 (2020).

8. Barovic, M. ***et al.*** Metabolically phenotyped pancreatectomized patients as living donors for the study of islets in health and diabetes. *Molecular Metabolism* **27**, S1–S6 (2019).

9. Poitout, V. ***et al.*** A call for improved reporting of human islet characteristics in research articles. *Diabetes* **68**, 209–211 (2019).

10. Ebrahimi, A. ***et al.*** Evidence of stress in β cells obtained with laser capture microdissection from pancreases of brain dead donors. *Islets* **9**, 19–29 (2017).

11. Toyama, H., Takada, M., Suzuki, Y. & Kuroda, Y. Activation of macrophage-

associated molecules after brain death in islets. *Cell Transplant.* **12**, 27–32 (2003).

12. Negi, S. *et al.* Analysis of Beta-Cell gene expression reveals inflammatory signaling and evidence of dedifferentiation following human islet isolation and culture. *PLoS One* **7**, 1–11 (2012).

13. Weir, G. C. Glucolipotoxicity, β-cells, and diabetes: The emperor has no clothes. *Diabetes* **69**, 273–278 (2020).

14. Solimena, M. *et al.* Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. *Diabetologia* **61**, 641–657 (2018).

15. Gerst, F. *et al.* The Expression of Aldolase B in Islets is Negatively Associated with Insulin Secretion in Humans. *J. Clin. Endocrinol. Metab.* **103**, 4373–4383 (2018).

16. Khamis, A. *et al.* Laser capture microdissection of human pancreatic islets reveals novel eQTLs associated with type 2 diabetes. *Mol. Metab.* **24**, 98–107 (2019).

17. Cohrs, C. M. *et al.* Dysfunction of Persisting β Cells Is a Key Feature of Early Type 2 Diabetes Pathogenesis. *Cell Rep.* **31**, (2020).

18. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

19. Taneera, J. *et al.* Identification of novel genes for glucose metabolism based upon expression pattern in human islets and effect on insulin secretion and glycemia. *Hum. Mol. Genet.* **24**, 1945–1955 (2014).

20. Carrat, G. R. *et al.* Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice. *Am. J. Hum. Genet.* **100**, 238–256 (2017).

21. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab.* **24**, 608–615 (2016).

22. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

23. Haythorne, E. *et al.* Diabetes causes marked inhibition of mitochondrial metabolism in

pancreatic β-cells. *Nat. Commun.* **10**, (2019).

24. Meier, F. *et al.* Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).

25. Thorens, B. GLUT2, glucose sensing and glucose homeostasis. *Diabetologia* **58**, 221–232 (2015).

26. Pipatpolkai, T., Usher, S., Stansfeld, P. J. & Ashcroft, F. M. New insights into KATP channel gene mutations and neonatal diabetes mellitus. *Nature Reviews Endocrinology* **16**, 378–393 (2020).

27. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* **13**, 3497–3506 (2014).

28. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13 Suppl 16**, (2012).

29. Boccard, J. & Rutledge, D. N. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Anal. Chim. Acta* **769**, 30–39 (2013).

30. Boccard, J. & Rutledge, D. N. Iterative weighting of multiblock data in the orthogonal partial least squares framework. *Anal. Chim. Acta* **813**, 25–34 (2014).

31. van der Meulen, T. *et al.* Virgin Beta Cells Persist throughout Life at a Neogenic Niche within Pancreatic Islets. *Cell Metab.* **25**, 911-926.e6 (2017).

32. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).

33. Wang-Sattler, R. *et al.* Novel biomarkers for pre-diabetes identified by metabolomics. *Mol. Syst. Biol.* **8**, (2012).

34. Suvitaival, T. *et al.* Lipidome as a predictive tool in progression to type 2 diabetes in

Finnish men. *Metabolism.* **78**, 1–12 (2018).

35. Wigger, L. *et al.* Plasma Dihydroceramides Are Diabetes Susceptibility Biomarker Candidates in Mice and Humans. *Cell Rep.* **18**, 2269–2279 (2017).

36. Haus, J. M. *et al.* Plasma ceramides are elevated in obese subjects with type 2 diabetes and correlate with the severity of insulin resistance. *Diabetes* **58**, 337–343 (2009).

37. Kopprasch, S. *et al.* Detection of independent associations of plasma lipidomic parameters with insulin sensitivity indices using data mining methodology. *PLoS One* **11**, (2016).

Material and methods

Cohort

Our cohort comprised 133 adult surgical patients from the University Hospital Carl Gustav Carus Dresden who after informed consent participated in this study over a period of 5 years. Based on the thresholds set by the American Diabetes Association[7] (ADA) for fasting glucose, HbA1c and 2-hour glycemia of an oral glucose tolerance test (OGTT) in the days immediately before surgery 18 of these patients were classified as non-diabetic (ND), 41 with impaired glucose tolerance (IGT), including 3 with impaired fasting glucose (IFG) only, 35 with Type 3c Diabetes (T3cD) and 39 with Type 2 Diabetes (T2D). A diagnosis of T3cD was made whenever the occurrence of diabetes was not recognized for longer than 1 year prior to the onset of the symptoms leading to surgery and the subject was negative for the presence of circulating autoantibodies against pancreatic islets, which were assessed as previously described[14]. In all analyses IFG and IGT subjects were merged in one group hereinafter labeled as IGT PPP. Medical and family history and relevant clinical biochemistry data available from the routine medical processing of the patients were retrieved from the hospital database and referring physicians. Patients who underwent neoadjuvant chemotherapy as well as those with endocrine neoplasms of the pancreas were excluded from this study.

Human pancreatic tissue and peripheral blood processing

Surgical tissue specimens were examined by a certified pathologist immediately after resection as per regular clinical procedures. Fragments of healthy pancreatic tissue from the resection margins were excised, snap frozen in liquid nitrogen and stored at -80ºC either natively or embedded in TissueTek OCT compound. Estimated warm and cold ischaemia time was on average 2 hours. Peripheral blood samples were stored at -80ºC in aliquots of full blood, plasma and serum.

Transcriptomics

Islet procurement and RNA isolation

Pancreatic tissue was sectioned in a cryostat and mounted on UV pre-treated Zeiss MembraneSlide 1.0 PEN slides. Laser capture microdissection (LCM) was done with a Zeiss Palm MicroBeam system using autofluorescence to identify islets, as previously described[38]. RNA was isolated from approximately 20x6μm3 of islet tissue using the Arcturus PicoPure RNA Isolation Kit. Only preparations with RNA Integrity Number ≥5 were used for RNA sequencing. The entire handling of the tissue samples was done in a strictly RNAse free environment.

Library preparation, RNA Sequencing and alignment

Sequencing libraries were prepared from bulk RNA using the Illumina SmartSeq protocol. Single ended 76bp sequencing was done with an Illumina HiSeq 2500 or Illumina HiSeq 500 at the Next Generation Sequencing Core Facility of the CMCB Dresden, with the target depth of 35 million fragments per library. From FASTQ files, purity-filtered reads were trimmed with Cutadapt to remove adapters and low-quality sequences (v. 1.8)[39]. Reads matching to ribosomal RNA sequences were removed with fastq_screen (v. 0.11.1)[40]. Remaining reads were further filtered for low complexity with reaper (v. 15-065)[41]. Reads were aligned against Homo sapiens GRCh38.92 genome using STAR (v. 2.5.3a)[42]. The number of read counts per gene locus was summarized with htseq-count (v. 0.9.1)[43] using Homo sapiens GRCh38.92 gene annotation. Quality of the RNA-seq data alignment was assessed using RSeQC (v. 2.3.7)[44].

RNA Sequencing quality control, processing and differential expression analysis

RNA Sequencing datasets were screened for exocrine contamination in an initial quality control (QC) step. Analysis of the absolute number of detected expressed genes, gene body coverage and cumulative gene diversity assessment flagged a number of libraries to be of insufficient quality for downstream analysis. Libraries were filtered for minimal expression by removal of genes with less than 5 mean raw reads. Reads were normalized for library size and transformed for variance stabilizing using tools from the DESeq2 Bioconductor package[45]. Further analysis revealed 41 libraries in which transcripts other than insulin (INS) displayed the highest normalized number of reads. Differential expression analysis across the clinical categories (ND, IGT, T3cD, T2D) was performed using limma function with voom approach from limma Bioconductor package[46,47] on both the full dataset of 133 libraries which passed the QC analysis as well as on the “restricted” dataset of 92 libraries featuring INS as the highest expressed gene based on the linear model with age, gender and BMI as covariates.

Gene set enrichment analysis of differentially expressed genes

Functional enrichment analyses of differentially expressed genes in IGT, T2D or T3cD compared to ND patients were performed by weighted gene set enrichment analysis (GSEA) on unfiltered gene lists ranked by decreasing differential expression test statistics. Gene Ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway collections were restricted to gene sets with a minimum and maximum sizes of 100 and 500, respectively. The enrichment scores were normalized by gene set size and their statistical significance was assessed by permutation test (n=1,000). GO enrichment analyses were carried out using the gseGO function from the R package clusterProfiler (version 3.10.1)[48]. GO terms enriched in at least one comparison were identified using *p* value and normalized enrichment score thresholds < 0.01 and > 2.5, respectively. Redundancy of enriched GO terms was removed using the clusterProfiler simplify function (selecting the most representative term by *p* value) and enrichment maps generated using the emapplot function from the R package enrichplot (version 1.2.0). KEGG pathway enrichment analyses were performed using the clusterProfiler gseKEGG function. Results were filtered based on a *p* value threshold < 0.01 and a normalized enrichment score threshold > 2. To simplify results visualization and interpretation, redundant KEGG pathways were also collapsed into fewer biological themes using the enrichment map visualizations.

Weighted Gene Correlation Network Analysis

Gene Co-expression Network Construction

The gene co-expression network was created following the weighted gene correlation network analysis (WGCNA) protocol as implemented in the WGCNA package in R (version 1.68)[22], as previously described[14]. WGCNA was performed on batch-corrected, normalized and variance stabilizing transformed expression data from the full cohort of 133 subjects. The co-expression network was constructed by calculating an adjacency matrix using Pearson correlation, pairwise complete observations and unsigned method. The soft-threshold parameter was optimized with the function pickSoftThreshold and the best threshold ($\alpha = 7$) selected by visual inspection. The adjacency matrix was then computed into a topological overlap matrix (TOM), converted to distances, and clustered by hierarchical clustering using average linkage clustering. Modules were identified by dynamic tree cut using the hybrid method and parameters minClusterSize=20 and deepSplit=2. Similar modules were merged using a module eigengene distance of 0.15 as the threshold.

Identification of co-expressed gene modules related to diabetes trait

We correlated the module eigengenes to clinical traits using Spearman correlation (pairwise complete observations) and calculated the corresponding *p* values using the cor and corPvalueStudent functions from the WGCNA package, respectively. Module-trait correlations were represented as heatmap using the labeledHeatmap function from the WGCNA package. The modules displaying the most positive or negative correlation to HbA1c were further analysed. Normalized and variance stabilizing transformed gene counts for selected modules were plotted as heatmap using the heatmap.2 function from the R gplots package (version 3.0.1.2). Rows (representing genes) were scaled and hierarchically clustered by Euclidean distances. Columns, representing patients, were custom ordered as described in the legend of figure 3. Module hub genes, such as highly connected genes within a module that could have a strong influence on a phenotypic trait, were identified as those with the highest correlation with the particular trait and the highest correlation with the module eigengene.

Significance of gene co-expression modules

We tested the significance of the co-expression modules by comparing their intramodular connectivity (connectivity between nodes within the same module, as computed by the WGCNA intramodularConnectivity function) to the background as follows. For each selected module of size N, we calculated a Z-score as:

$$Z=(k-\mu)/\sigma$$

where k is the intramodular connectivity and μ and σ are the mean and standard deviation of the intramodular connectivity from 1,000 randomly sampled modules of size N respectively. Empirical p values were then calculated as the fraction of random intramodular connectivity values ≥ to the observed intramodular connectivity. For the modules with the highest variable importance in projection score in the HbA1c multiblock model, all of the random intramodular connectivity values were below the observed intramodular connectivity, suggesting that these modules were more compact than modules assembled by randomly sampling the same number of genes from the expression data (Supplementary Table S7).

Functional profiles of gene modules with best prediction score for HbA1c

The clusterProfiler enrichKEGG function was used to test for the over representation of selected co-expressed gene modules in KEGG pathways using hypergeometric distribution. A *p* value threshold < 0.01 was used to identify enriched terms. Enrichment map visualizations were used to overcome gene set redundancy. Results were displayed as networks of enriched pathways and overlapping genes using cytoscape (version 3.5.1).

Deconvolution analysis

In all samples a cell proportions matrix was produced using the R package DeconRNASeq (v.1.26.0) on RPKM-transformed data. The signature file provided to DeconRNASeq comes from Xin et al. (2016)[21], Supplementary Table S2 (A), obtained using single-cell data. It was adapted to the human genome version 38 by excluding 15 obsolete genes.

Lipidomics

Sample availability and sample overlap with transcriptomics data

Pre-operative plasma lipidomics samples were obtained from a subset of the PPP cohort. Shotgun lipidomics analysis was performed on plasma from 55 PPP. These included 53 subjects who also had their islet transcriptomics profile included in this study plus two PPP who were not part of the transcriptomics analysis because the RNA-Seq data failed to pass the quality control. Moreover, targeted sphingolipid analysis was performed on plasma from 101 PPP. These included 98 PPP whose transcriptomics data was also included in this study plus three PPP whose RNA-Seq data was excluded for quality reasons. The number of samples in the two types of lipidomics analysis was smaller than in islet transcriptomic analysis because of the limited availability of plasma samples.

Shotgun lipidomics measurements

A streamlined mass-spectrometry (MS) -based platform for shotgun lipidomics developed by Lipotype GmbH (Dresden, Germany) was used for lipidomic profiling of patient plasma samples. Lipid extraction, internal standard addition and infusion into the mass spectrometer were performed as previously described[49]. The internal standard mixture contained: cholesterol D6 (chol), cholesterol ester 20:0 (CE), ceramide 18:1;2/17:0 (Cer), diacylglycerol 17:0/17:0 (DAG), phosphatidylcholine 17:0/17:0 (PC), phosphatidylethanolamine 17:0/17:0 (PE), lysophosphatidylcholine 12:0, (LPC) lysophosphatidylethanolamine 17:1 (LPE), triacylglycerol 17:0/17:0/17:0 (TAG) and sphingomyelin 18:1;2/12:0 (SM).

Samples were analyzed by direct infusion in a QExactive mass spectrometer (Thermo Scientific) in a single acquisition. Tandem mass-spectrometry (MS/MS) was triggered by an inclusion list encompassing corresponding MS mass ranges scanned in 1 Da increments. MS and MS/MS data were combined to monitor CE, DAG and TAG ions as ammonium adducts; PC, PC O-, as acetate adducts; and PE, PE O- and PI as deprotonated anions. MS only was used to monitor LPE as deprotonated anion; Cer, SM and LPC as acetate adducts and cholesterol as ammonium adduct.

Data post-processing and normalization were performed using an in-house developed data management system. Only lipid identifications with a signal-to-noise ratio >5 and a signal intensity 5-fold higher than in corresponding blank samples were considered for further analysis.

Targeted sphingolipid measurements

Ceramides (C16:0 cer, C18:0 cer, C18:1 cer, C20:0 cer, C22:0 cer, C24:0 cer and C24:1 cer), Dihydroceramides (C16:0 DHcer, C18:0 DHcer, C18:1 DHcer, C20:0 DHcer, C22:0 DHcer, C24:0 DHcer,C24:1 DHcer) and precursors (Sphingosine, Sphinganine, 1-Deoxysphinganine,1-Methyldeoxysphinganine, SB) were quantified in plasma by liquid chromatography tandem mass spectrometry (LC-MC/MC). In addition to samples, seven-point calibration curves and 3 levels of quality controls were made from pure standards in BSA 5%. Finally, reference plasma spiked with analytes at two different levels were prepared as additional QC samples.

After lipid chromatographic separation on a UPLC I-Class system (Waters), mass analysis was performed on an API 6500 system (Sciex) operating with an electrospray source in positive mode. General parameters were set as follows: curtain gas: N2 (35 PSI), Ion source gas 1: Air (50 PSI), Ion source gas 2: Air (50 PSI), ion source voltage: 5500 V, temperature: 300°C, collision gas: N2 (7). Scheduled multiple reaction monitoring (MRM) mode was used with a target scan time of 0.5s and an MRM detection window of 60s.

Data was acquired using Analyst 1.6.2 (Sciex) and data processing was performed with MultiQuant 3.0 (Sciex). Peak area of analyte and internal standard were determined by the MultiQuant 3.0 (Sciex) integration system. Analyte concentrations were determined using the internal standard method. The standard curves were generated from the peak area ratios of analyte/internal standard using linear regression analysis with 1/x2 weighting (except for C24 cer: quadratic regression analysis). Quantifications of analytes were accepted based on quality control samples. A tolerance of 25% and 30% was applied for accuracy and precision of QC samples and spiked plasma samples, respectively. All concentrations were reported in ng/mL.

Statistical analysis of shotgun lipidomics and targeted sphingolipid data

The statistical analyses of the shotgun lipidomics and targeted sphingolipid data sets were kept separate. Identical analysis steps were applied to the two data sets. Both sets had missing data values. Lipid species with ≥25% missing values across all available plasma samples were removed from the data set. This filtering resulted in 113 lipid species that were kept in the shotgun data set (523 were removed) and 14 in the targeted data set (4 were removed). For the lipids that remained in the data sets, missing values were imputed using a random forest approach, applying the function missForest from the R package missForest, with default parameters. In a next step, samples were filtered based on subject characteristics: individuals with bilirubin levels ≥100 µmol/l were removed before all analysis; moreover, individuals categorized as IGT with an HbA1c≥6% were excluded from the group comparisons in differential analysis, but they were retained in other analyses involving lipidomics data. In differential analysis, due to the limited number of available ND samples, the ND and the included IGT samples were combined into a single group for comparison with other sample groups, as described in the result section.

For differential analysis, linear models were applied, using the function lm from the R stats package. For each comparison between two sample groups, a linear model that included diabetes status as the main explanatory variable and age and sex as covariates was fitted to the data from the two groups. P values for diabetes status were adjusted across all included lipid species with the Benjamini-Hochberg method, separately for each comparison.

Integrative analysis of transcriptomics and lipidomics

Multiblock modeling

Consensus Orthogonal Partial Least Squares (OPLS) model was computed with the MATLAB 9 environment with combinations of toolboxes and in-house functions that are available at https://gitlab.unige.ch/Julien.Boccard/consensusopls. Modified RV-coefficients were computed with the publicly available MATLAB m-file[50]. KOPLS-DA was assessed with routines implemented in the KOPLS open source package[51]. Consensus OPLS modeling was performed on shotgun lipidomics, targeted sphingolipids and transcriptomics data tables, which were all autoscaled prior to the analysis. The Consensus OPLS model distinguishes variation of data that is correlated to Y response and those which is orthogonal to Y response. This eases the biological interpretation of results and enables the link between variation of variables and variation of the outcome while removing information coming from other sources of variation.

The model resulted in 3 components, of which 1 predictive latent variable and 2 orthogonal latent variables. The quality of the model was assessed by R^2 and Q^2 values, which define the portion of data variance explained by the model and the predictive ability of the model, respectively. The predictive component carried 11% of the total explained variance of global data (R^2X) and explained 51.7% of variation of HbA1c (R^2Y). This indicates that the model was able to explain a large part of variation of the response variable based on the different data matrices. The Q^2 value was computed by a K-fold cross validation (K=7), which led to a goodness of prediction of $Q^2 = 0.26$.

To ensure the validity of the model, a series of 1,000 permutation tests were carried out by mixing randomly the original Y response (HbA1c patient values). The true model Q2 value was clearly distinguished and statistically different from the random models distribution ($p<0.001$, mean=−0.1778, standard deviation (SD)=0.150, n=1,000). The variable relevance to explain the HbA1c variation was evaluated using the variable importance in projection (VIP) parameter, which reflects the importance of variables both with respect to the response and to the projection quality. The most relevant features were selected using a VIP threshold > 1.2.

Proteomics

Sample Preparation

Pooled pancreatic islet cells with an approximate surface area of 80,000 µm^2 were collected via Laser Capture Microdissection (LCM) onto adhesive cap tubes. Isolates were reconstituted in a 20 µl lysis buffer (PreOmics, Germany) and transferred into PCR tubes[52]. Samples were boiled at 95°C for 1min to denature proteins and reduce and alkylate cysteines without shaking in a thermocycler (Eppendorf GmbH) followed by sonication at maximum power (Bioruptor, Diagenode, Belgium) for 10 cycles of 30sec sonication and 30sec cooldown each. Sample liquid was briefly spun down and boiled again for 10min without shaking. 20µl of 100mM TrisHCl pH 8.5 (1:1 v/v) and 20ng Trypsin/LysC were added to each sample, followed by overnight digestion at 30°C without shaking. Next day, 40µl 99% Isopropanol 5% Trifluoroacetic acid (TFA) (1:1 v/v) was added to the solution and mixed by sonication. Samples were then subjected to stage-tip cleanup via styrenedivinylbenzene reversed-phase sulfonate (SDB-RPS). Sample liquid was loaded on one 14-gauge stage-tip plug. Peptides were cleaned up with 2x200µl 99% Isopropanol 5% TFA and 2x200µl 99% ddH2O 5% TFA in an in-house made Stage-tip centrifuge at 2,000xg, followed by elution in 40µl 80% Acetonitrile, 5% Ammonia and dried at 45°C in a SpeedVac centrifuge (Eppendorf, Concentrator plus) according to the 'in-StageTip' protocol (PreOmics, Germany). Peptides were resuspended in 0.1% TFA, 2% ACN, 97.9% ddH2O.

Liquid chromatography and mass spectrometry (LC-MS)

LC-MS was performed with an EASY nanoLC 1200 (Thermo Fisher Scientific) coupled online to a trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer (timsTOF Pro, Bruker Daltonik GmbH, Germany) via nano-electrospray ion source (Captive spray, Bruker Daltonik GmbH). Peptides were loaded on a 50cm in-house packed HPLC-column (75µm inner diameter packed with 1.9µm ReproSil-Pur C18-AQ silica beads, Dr. Maisch GmbH, Germany). Sample analytes were separated using a linear 120min gradient from 5-30% buffer B in 95min followed by an increase to 60% for 5min, and by a 5min wash at 95% buffer B at 300nl/min (Buffer A: 0.1% Formic Acid, 99.9% ddH2O; Buffer B: 0.1% Formic Acid, 80% CAN, 19.9% ddH2O). The column temperature was kept at 60°C by an in-house manufactured oven.

Mass spectrometry analysis was performed in a data-dependent PASEF mode with 1 MS1 survey TIMS-MS and 10 PASEF MS/MS scans per acquisition cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to 100ms each and we analyzed the ion mobility range from $1/K_0$ = 1.6 Vs cm^{-2} to 0.6 Vs cm^{-2}. Precursor ions for MS/MS analysis were isolated with 2Th windows for m/z<700 and 3Th for m/z>700 in a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at $1/K_0$=1.6 VS cm^{-2} to 20 eV at $1/K_0$=0.6 Vs cm^{-2}. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonik GmbH). Precursors for MS/MS were picked at an intensity threshold of 2.500 a.u. and resequenced until reaching a ‘target value’ of 20,000 a.u taking into account a dynamic exclusion of 40sec elution[24].

Proteomics raw file processing

Raw files were searched against the human Uniprot databases (UP000005640_9606.fa, UP000005640_9606_additional.fa) MaxQuant (Version 1.6.7), which extracts features from four-dimensional isotope patterns and associated MS/MS spectra[53]. False-discovery rates were controlled at 1% both on peptide spectral match (PSM) and protein level. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamidomethylation as fixed modification, while limiting the maximum peptide mass to

4,600 Da. Enzyme specificity was set to trypsin cleaving c-terminal to arginine and lysine. A maximum of two missed cleavages were allowed. Maximum precursor and fragment ion mass tolerance were searched as default for TIMS-DDA data, while the main search peptide tolerance was set to 20ppm. The median absolute mass deviation for the data was 0.68ppm. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs with a 0.7-min retention-time match window and a 0.05 $1/K_0$ ion mobility window[54]. Label-free quantification was performed with the MaxLFQ algorithm and a minimum ratio count of 1[55].

Bioinformatic analysis

Bioinformatics analysis was performed in Perseus (version 1.6.7.0 and 1.5.5.0) and GraphPad Prism (version 8.2.1)[56]. Reverse database, contaminant, and only by site modification identifications were removed from the dataset. Data were grouped by analytical replicates and filtered to at least 70% data completeness in one group. Missing values were imputed from a normal distribution with a downshift of 1.8 and a width of 0.3 and data were $\log_2$-transformed. To represent the data reproducibility and variability, a principal component analysis was performed on the median data of analytical replicate measurements of each individual. Clinically classified T2D and ND individuals were tested for differences in their mean by a two-sided Student's t-test with S0=0.1 and a Benjamini-Hochberg correction for multiple hypothesis testing at an FDR of 0.05 preserving grouping of each individuals analytical replicate measurements, and presented as volcano plot. We then normalized the data by row-wise z-scoring followed by hierarchical clustering using Euclidean as the distance parameter for column- and row-wise clustering. 1D gene ontology enrichments of clustered and systematically changed proteins were performed with regards to their cellular compartment and keywords assignment[28]. Log_2 transformed LFQ data were used for the calculation of intensity shifts of the enriched keyword or cellular compartment term for each of the displayed clusters. Total protein copy number estimation of the median LFQ intensities for patients clinically classified as non-diabetic and diabetic were calculated using the Perseus plugin 'Proteomic ruler'[27]. Median LFQ intensity values for all T2D and ND were calculated. We annotated protein groups for the leading protein ID with the human Uniprot fasta file (UP000005640_9606.fa) and estimated the protein copy number with the following settings: Averaging mode. 'All columns separately', Molecular masses: 'Average molecular mass', Detectability correction: 'Number of theoretical peptides', Scaling mode: 'Histone proteomic ruler', Ploidy: '2', Total cellular protein concentration: '200g/l'. Proteins were annotated with regards to their cellular compartment by gene ontology. We calculated the median protein copy number for the samples from T2D and ND PPP separately and multiplied it by its protein mass. To calculate the subcellular protein mass contribution, we calculated the protein mass proportion for the GOCC terms 'Nucleus', 'Mitochondrion', 'Cytoskeleton', 'Golgi apparatus', and 'Endoplasmic reticulum'. For calculating the organellar change between T2D and ND PPP, protein mass contributions of each organelle were normalized by its respective 'Nuclear part' contribution. Chromosomal annotation of significantly changed proteins between T2D and ND PPP was identified via Ensembl ID.

Antibody validation

Rabbit polyclonal anti-ALDOB antibody (Proteintech, Cat.No. 18065-1-AP) was tested for specificity by western blotting of protein extracts of *ALDOB*$^{-/-}$ MIN6 cells generated with a CRISPR/Cas9 system, as described[52]. The knock-out of *ALDOB* was verified by Sanger sequencing of the target locus.

Isolated mouse islet and cell line experiments

Mouse (C57Bl6, db/db and db/+ mice, 3 animals/strain, age 13 weeks) islets were cultured for 1 day post isolation. Islet beta MIN6s4 and alpha αTC1- clone 6 cell lines were harvested for RNA extraction using Qiagen RNeasy Mini Kit according to the manufacturer's instructions. After quality control, RNA samples were sequenced using the Illumina HiSeq 2000 platform and processed as previously described[45,57,58].

Immunofluorescence microscopy

Immunofluorescence staining was done on formalin-fixed paraffin embedded 5μm thick sections of human pancreatic tissue. Acetylated histone H3 and H4 were detected in separate sections using rabbit polyclonal antibodies (Merck Millipore Cat.No. 06-598 and 06-599, respectively). A mouse monoclonal anti-insulin antibody (Thermo Fisher Scientific Cat.No. 53-9769-82) was used for co-staining, to identify the beta cell areas. Images were acquired using a Nikon C2+ confocal microscope with a 60x oil immersion objective, with acquisition parameters normalized to a negative control sample.

Materials and methods references

38. Sturm, D. *et al.* Improved Protocol For Laser Microdissection Of Human Pancreatic Islets From Surgical Specimens. *J. Vis. Exp.* 2–7 (2013). doi:10.3791/50231

39. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1 Next Gener. Seq. Data Anal. - 10.14806/ej.17.1.200* (2011).

40. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).

41. Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41–49 (2013).

42. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

43. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

44. Wang, L., Wang, S. & Li, W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).

46. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

47. Smyth, G. K. *et al.* RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research* **5**, (2018).

48. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

49. Surma, M. A. *et al.* An automated shotgun lipidomics platform for high throughput, comprehensive, and quantitative analysis of blood plasma intact lipids. *Eur. J. Lipid Sci. Technol.* **117**, 1540–1549 (2015).

50. Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M. & Van Erk, M. J. Matrix correlations for high-dimensional data: The modified RV-coefficient. *Bioinformatics* **25**, 401–405 (2009).

51. Bylesjö, M., Rantalainen, M., Nicholson, J. K., Holmes, E. & Trygg, J. K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *BMC Bioinformatics* **9**, (2008).

52. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

53. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

54. Prianichnikov, N. *et al.* Maxquant software for ion mobility enhanced shotgun proteomics. *Mol. Cell. Proteomics* **19**, 1058–1069 (2020).

55. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

56. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**, 731–740 (2016).

57. Hu, J., Ge, H., Newman, M. & Liu, K. OSA: A fast and accurate alignment tool for RNA-Seq. *Bioinformatics* **28**, 1933–1934 (2012).

58. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2009).

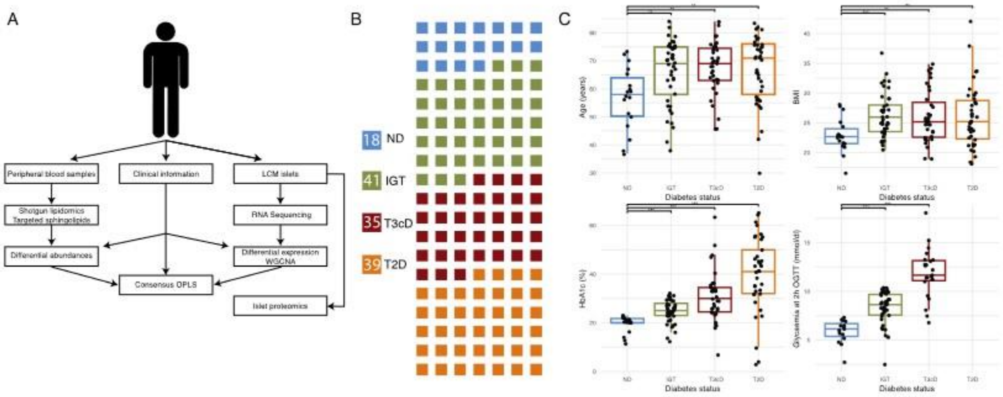


Figure 1: Overview of the experimental procedures and cohort characteristics. A) Experimental procedures overview. Clinical data and peripheral blood were collected preoperatively, and the snap-frozen surgical pancreatic tissue used for LCM of the islets of Langerhans. Blood samples were analyzed for lipidomics, while LCM islets for transcriptomics and proteomics. Omics datasets were individually evaluated in relationship to glycemic status   and further integrated with each other using Consensus Orthogonal Partial Least Squares (OPLS) analysis. B) Waffle plot showing the structure of the cohort in terms of glycemic/diabetes categories based on American Diabetes Association criteria. Absolute numbers for each category are given in the legend boxes. C) Boxplots of four major clinical parameters relevant for diabetes diagnosis and management. Statistically significant differences from ND PPP were determined using the Student's t-test (**p*<0.05; ***p*<0.01). LCM Laser Capture Microdissection, ND Non-diabetic, IGT Impaired Glucose Tolerance, T3cD Type 3c Diabetes, T2D Type 2 Diabetes.

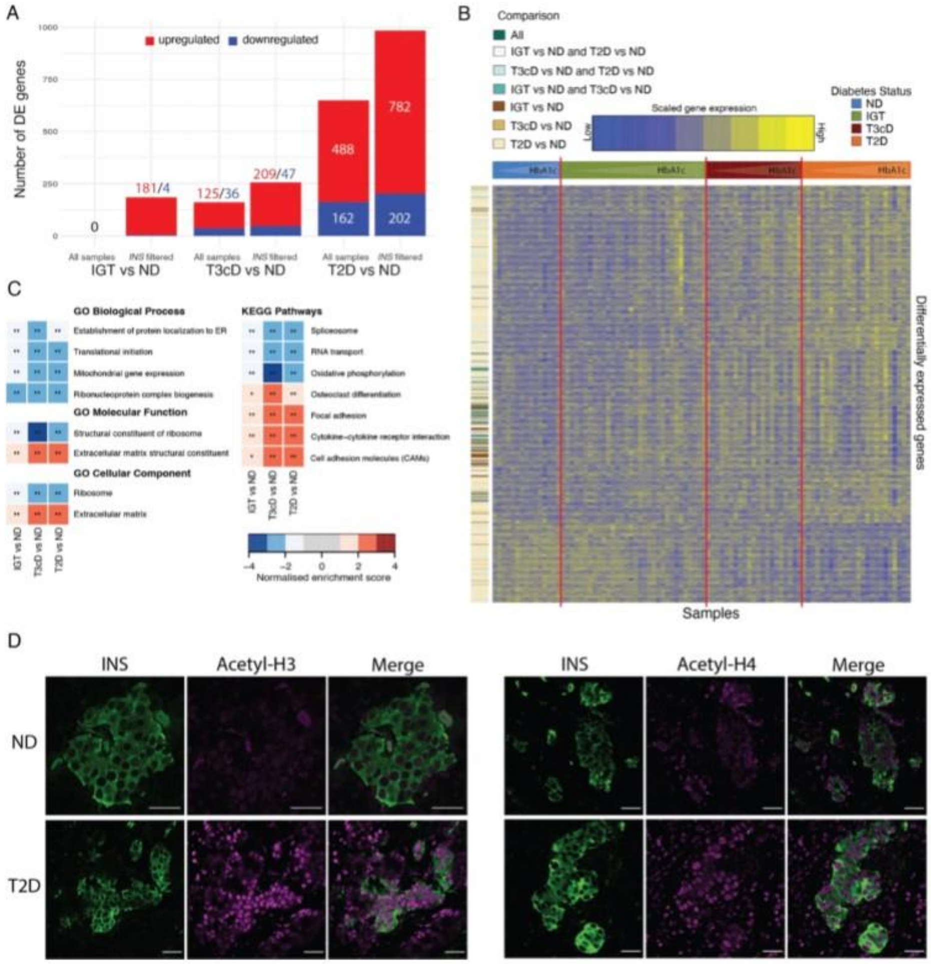


Figure 2: Transcriptional changes between non-diabetic, pre-diabetic and diabetic patients. A) Number of DE genes identified by comparing glycemic groups of PPP in the entire (all samples) or "restricted" cohort (*INS* filtered). B) Gene expression profile of DE genes in the "restricted" cohort. Columns represent patients grouped according to their glycemic status and ordered based on increasing HbA1c levels. Rows, representing DE genes (variance stabilizing transformation normalized counts), were clustered based on Euclidean distance. The colored side bar indicates in which comparisons a gene was identified as differentially expressed. C) Gene Set Enrichment Analysis of DE genes between IGT, T3cD or T2D and ND PPP in the "restricted" cohort. GO terms and KEGG pathways are colored according to the normalized enrichment score. Corresponding p-values are also indicated (*p<0.05, **p<0.01). D) Immunofluorescence for insulin (green), acetylated histones H3 (left) and H4 (right) (magenta) in representative samples of formalin fixed paraffin embedded pancreatic tissues from ND and T2D PPP. Scale bars correspond to 20µm. DE differentially expressed, ND Non-diabetic, IGT Impaired Glucose Tolerance, T3cD Type 3c Diabetes, T2D Type 2 Diabetes.

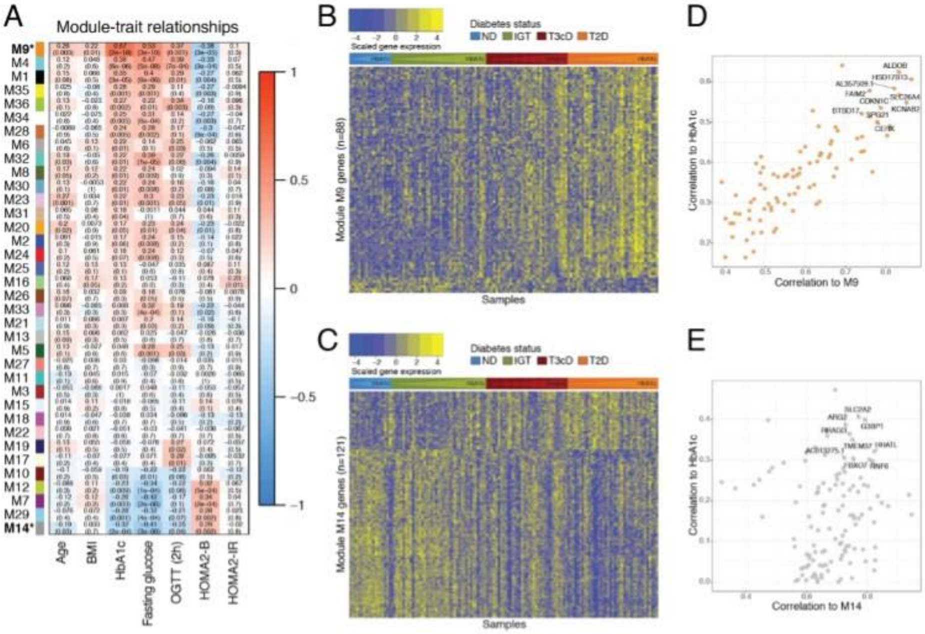


Figure 3: Identification of co-expressed gene modules related to diabetes traits. A) Correlation between module eigengenes and clinical traits including age, BMI, HbA1c, fasting glucose, OGTT at 2 hours, HOMA2-B and HOMA2-IR. Each cell contains the corresponding Spearman correlation coefficient and Student *p* value (in parenthesis). Cells are colored according to their correlation to clinical traits. Modules are ordered based on their correlation to HbA1c. B-C) Gene expression profiles of gene modules M9 (B) and M14 (C). Columns, representing PPP, were grouped according to their glycemic status and ordered based on increasing HbA1c levels. Rows, representing genes (variance stabilizing transformation normalized counts), were clustered based on Euclidean distance. D-E) Scatter plot of module membership vs. gene significance for HbA1c in modules M9 and M14. Genes with the highest module membership and gene significance ("hub genes") are labeled. ND Non-diabetic, IGT Impaired Glucose Tolerance, T3cD Type 3c Diabetes, T2D Type 2 Diabetes.

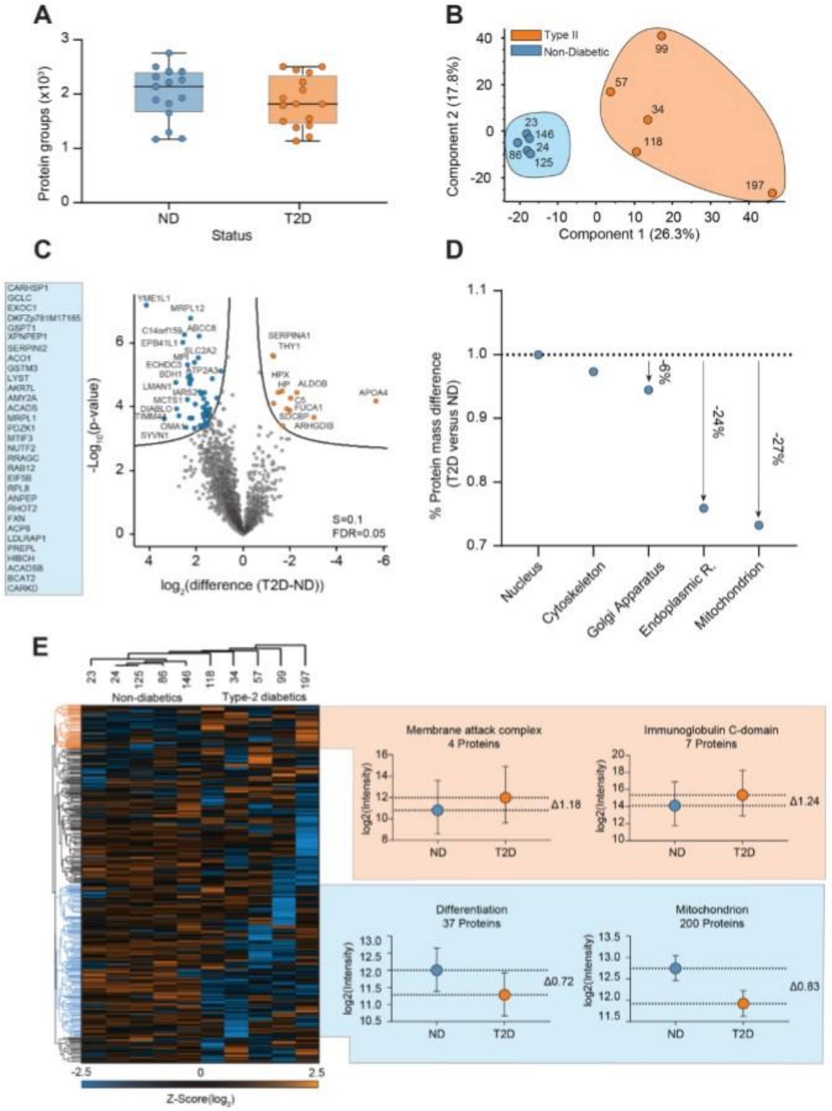


Figure 4: Proteomics Analysis. A) Number of identified proteins from pooled human pancreatic islet cells isolated by LCM from PPP classified as non-diabetic (ND, N=5) or with T2D (N=5). B) Principal Component Analysis (PCA) of all grouped pancreatic islet measurements (ND=blue, T2D=orange). C) Volcano plot comparing *p* values and $\log_2$-fold changes between islets of ND and T2D PPP. D) Percentage distribution of total protein islet mass and its contribution per organelle between ND and T2D PPP. The ND/T2D islet protein mass ratio in different organelles was normalized by the nucleus protein mass. E) Hierarchical clustering of all islet proteins identified in the T2D and ND PPP clusters. Log_2-transformed intensity values were normalized by z-scoring before the clustering followed by one-dimensional gene ontology enrichment for cellular compartment and keywords for each of the clusters. ND Non-diabetic, T2D Type 2 Diabetes.

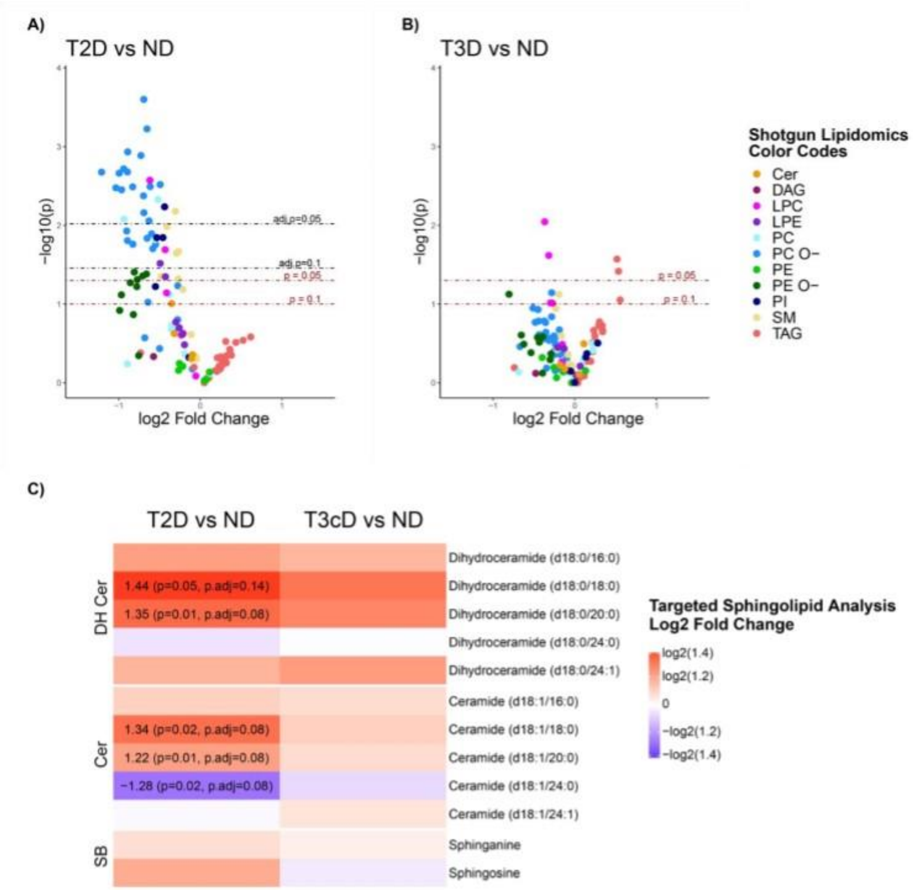


Figure 5: Results from lipidomics differential analysis. A-B) Shotgun lipidomics covering a variety of lipid classes: Ceramides (Cer), Diacylglycerols (DAG), Lysophosphatidylcholines (LPC), Lysophosphatidylethanolamines (LPE), Phosphatidylcholines (PC), Ether-linked Phosphatidylcholines (PC O-), Phosphatidylethanolamines (PE), Ether-linked Phosphatidylethanolamines (PE O-), Phosphatidylinositols (PI), Sphingomyelins (SM), Triacylglyerols (TAG). Volcano plots represent comparisons of plasma lipid levels between ND and T2D PPP. The X-axis shows direction and magnitude of the change; the Y-axis represents the statistical significance of the change. Each point is a lipid species, colored by lipid class to highlight class-specific trends. C) Targeted lipidomics on dihydroceramides (DH Cer), ceramides (Cer) and Sphingoid bases (SB). Each heatmap column represents the comparisons of plasma levels between ND and T2D PPP. Heatmap colors represent direction and magnitude of the change. Log$_2$ Fold Change: ratio of mean lipid concentration in the two groups, log$_2$ transformed. Statistical model used for all panels: linear regression with age and sex as covariates (p: *p* value); adjustment of *p* values across all lipid species by the Benjamini-Hochberg method (adj. *p*: adjusted *p* value). T2D Type 2 Diabetes, T3cD Type 3 Diabetes, ND & PD non-diabetic and pre-diabetic (with impaired fasting glucose and/or impaired glucose tolerance) with HbA1c<6.0.

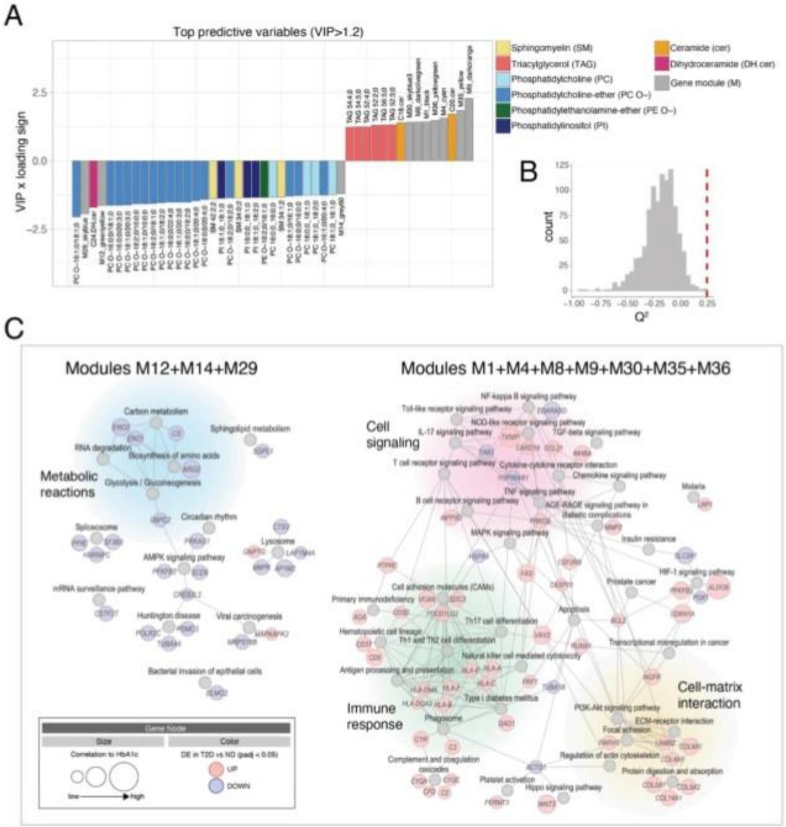


Figure 6: Multiblock data modeling of HbA1c. A) Barplot showing the variable importance in the multiblock consensus OPLS model. The Y-axis represents the importance scores for the predictors multiplied by the sign of the loadings on the predictive latent variable. Variables with importance in projection > 1.2 were selected. B) Statistical significance of the model through permutation test. C) Network representation of functional pathways enriched in modules with best prediction scores for HbA1c. Pathways are represented as gray nodes. Genes are represented as nodes sized based on their correlation to HbA1c and colored based on their differential expression in T2D vs. ND PPP. Only genes with significant differential expression (adj. *p*<0.05) in the "restricted" cohort are shown. VIP Variable Importance in Projection, DE Differentially expressed, ND Non-diabetic, T2D Type 2 Diabetes.

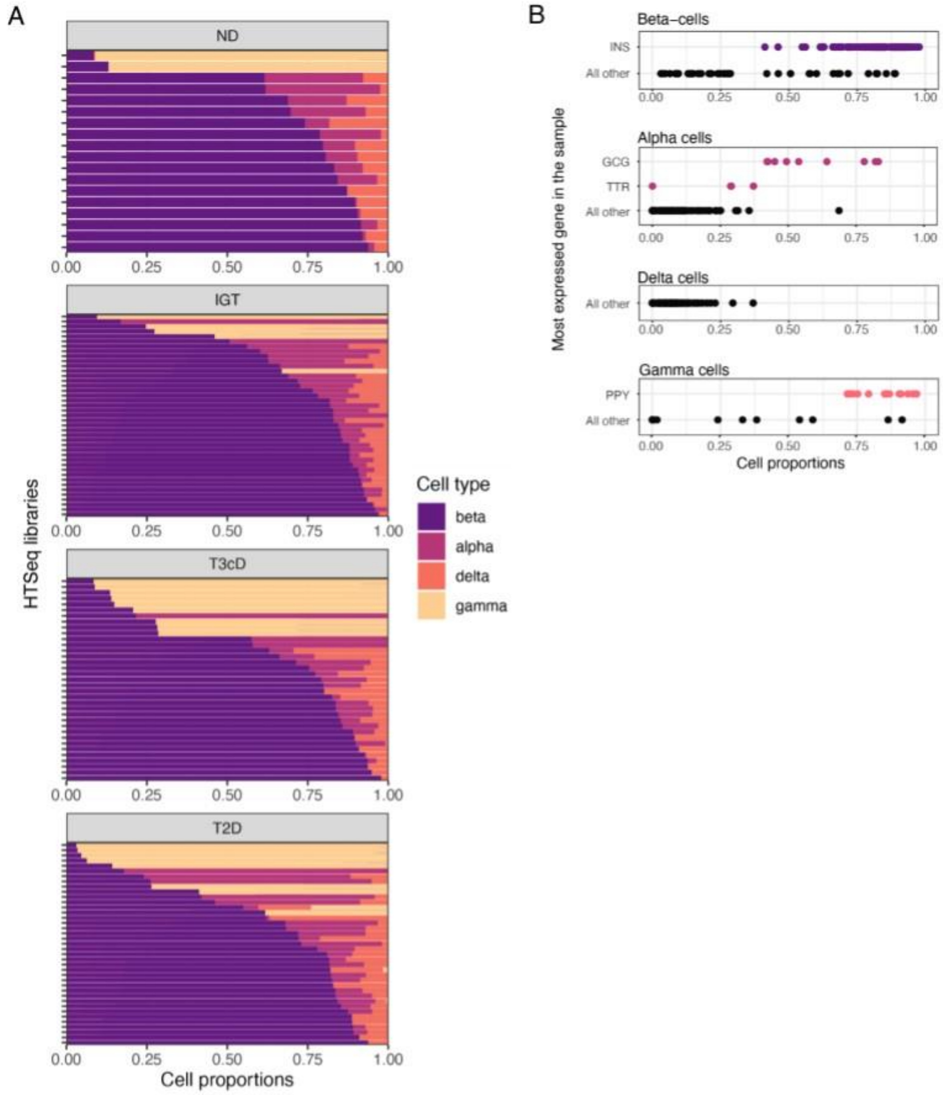


Figure S1: Deconvolution of cell types based on RNA-Seq data. A) Cell-type proportions by sample, as estimated with DeconRNASeq, panels faceted according to diabetes status. B) Sample distribution across each cell type proportion. Highlighted are samples presenting a cell type specific gene being the most expressed. Marker genes were *GCG* and *TTR* for alpha cells, *INS* for beta cells, *SST* for delta cells, and *PPY* for gamma cells.

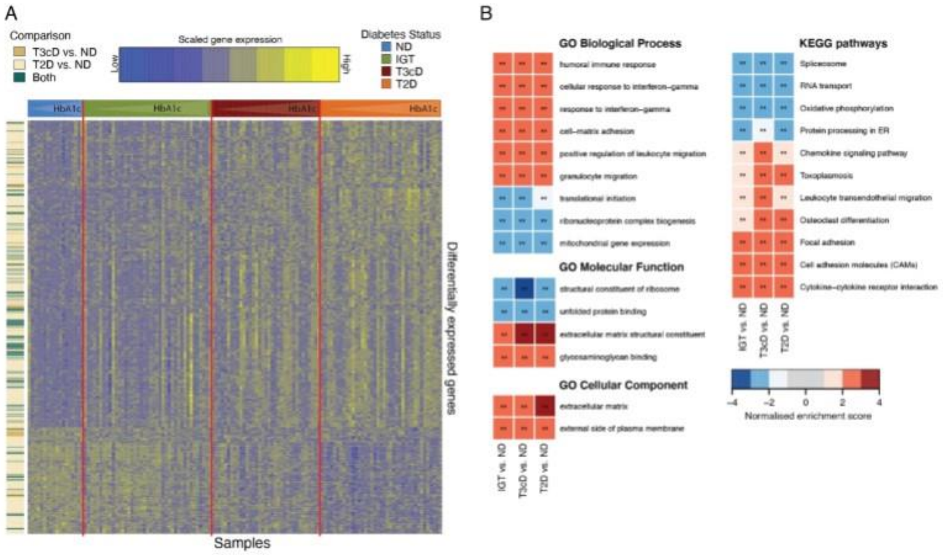


Figure S2: Differential gene expression analysis between glycemic groups in the entire cohort. A-B**)** Gene expression profile (A) and GSEA analysis (B) of DE genes between IGT, T3cD or T2D and ND PPP. Results are similar to those shown in Fig. 2BC, but obtained from the entire cohort of 133 PPP.

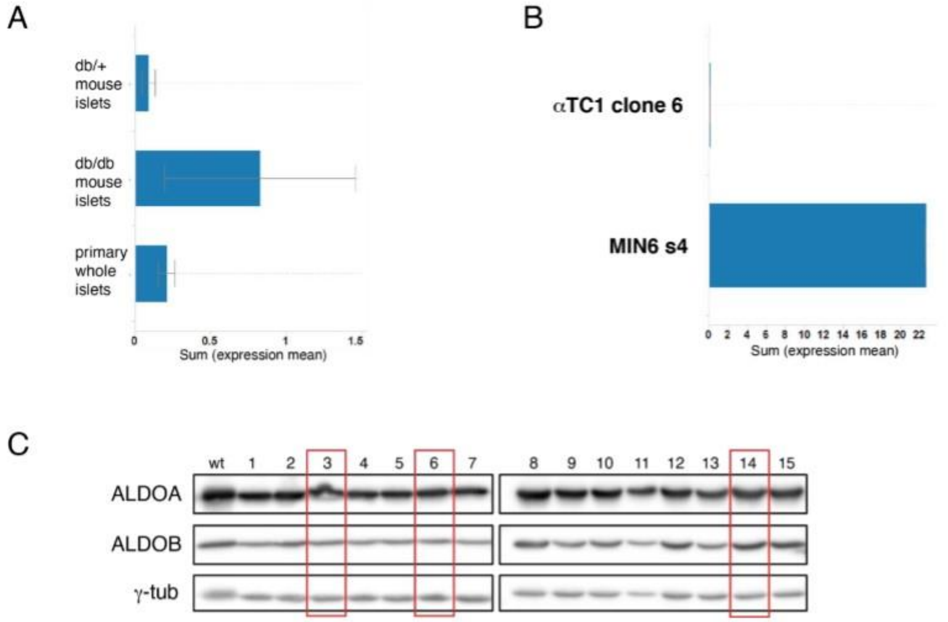


Figure S3: A-B) *ALDOB* expression (RNAseq, Illumina) in (A) islets from 13-week-old *db/db*, *db/+* mice and *C57Bl6* mice (3 animals/strain) or (B) mouse αTC1 clone 6 alpha and Min6s4 beta cell lines (n=4/cell line). C) Western blot of MIN6 single cell-derived clones with antibodies against ALDOB and ALDOA. Framed lanes mark *ALDOB* knockout clones as verified by site-specific sequencing.

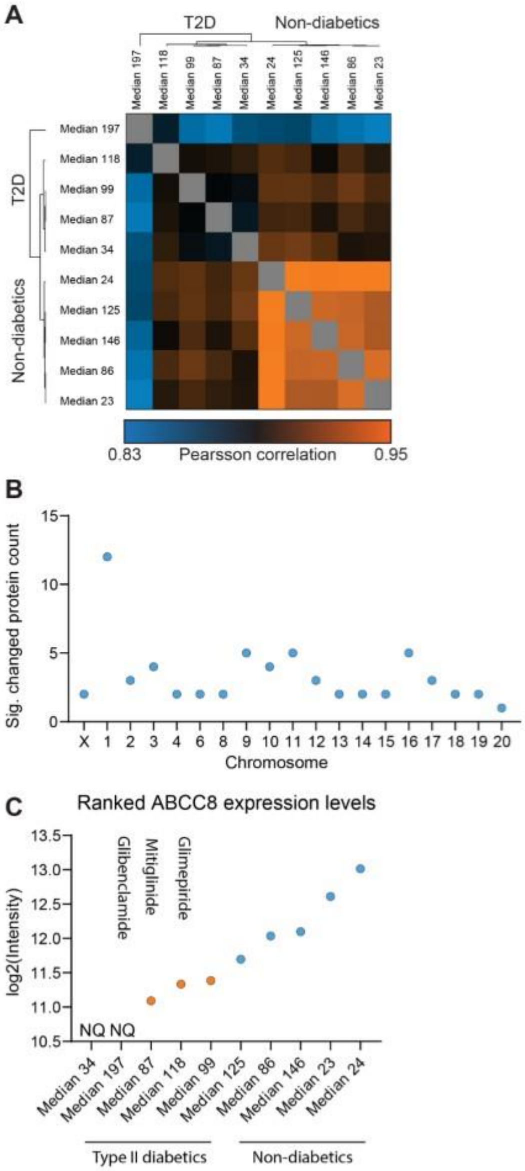


Figure S4: Hierarchical clustering of protein expression correlations in all biological replicates highlighting the technical and biological reproducibility of our proteome data set (A). Distribution of differentially expressed proteins between T2D and ND across chromosomes (B). Ranked ABCC8 protein expression levels across T2D and ND subjects. T2D are highlighted in orange, ND are highlighted in blue. Patient 118 was treated with Glimepiride; Patient 87 was treated with Mitiglinide; Patient 197 was treated with Glibenclamide (C).

3.1.3. Article 3: OpenCell

OpenCell: Proteome-scale endogenous tagging enables the cartography of human cellular organization

bioRxiv, March, 2021, (under review in Science)

Nathan H. Cho*, Keith C. Cheveralls*, **Andreas-David Brunner*,** Kibeom Kim*, André C. Michaelis*, Preethi Raghavan*, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y.S. Kim, Edna Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann, Manuel D. Leonetti#

** These authors contributed equally to this work*

# Correspondence

Contribution

In this project, I contributed to all aspects of the proteomics part, including experimental design, the establishment of the pulldown conditions and sample preparation for bottom-up proteomics analysis. I also combined the *timsTOF Pro* mass spectrometry platform with the *EvoSeop One* liquid chromatography system for a first of its kind study comprising several thousands of samples at highest robustness and sensitivity. Furthermore, I prepared and measured all >4,000 pulldowns, ensured highest quality control throughout the project, performed proteomics data analysis and contributed to manuscript writing.

‘OpenCell’: Proteome-scale endogenous tagging enables the cartography of human cellular organization

Nathan H. Cho*, Keith C. Cheveralls*, Andreas-David Brunner*, Kibeom Kim*, André C. Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y.S. Kim, Edna Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann#, Manuel D. Leonetti#

* equal contribution; # corresponding authors

Mapping the global proteome circuitry of the human cell is one of the central goals of the post-genomic era. Here, we combine high-throughput genome engineering of ~1,300 cell lines endogenously tagged with fluorescent protein fusions, 3D live-cell imaging, mass spectrometry (MS)-based high-speed interactomicsand advanced machine learning to decode the interaction and localization architecture of the human proteome. We delineate interacting protein families and facilitate unbiased biological discovery by unsupervised clustering, while hierarchical analyses of the interactome superimposed to localization uncover principles that template cellular organization. Furthermore, we discover that localization patterns alone are often enough to predict molecular interactions. ‘OpenCell’ is a global proteome-scale resource for human protein localization and interaction at endogenous expression levels. Our analytical methods are open-source and our data set is presented as an advanced interactive website (‘OpenCell’.czbiohub.org) to empower the community with the quantitative cartography of human cellular organization at proteome level.

The sequencing of the human genome has transformed our understanding of cell biology by defining the molecular parts list – the complete set of proteins encoded in the genome – that forms the canvas of cellular operation[1,2]. This paves the way for elucidating how the corresponding ~20,000 human proteins organize in space and time to define the cell's proteome architecture and function. Where does each protein localize within the cell? Can we comprehensively map how individual proteins assemble into larger units, whether as subunits of a molecular complex or as part of functionally inter-connected communities? Because of the complexity of the cell's molecular ecosystem, a full answer to these fundamental questions remains an ambitious challenge[3,4]. One aspect of this complexity is that cellular architecture is organized along different scales, so that multiple approaches must be combined for its complete description[5]. So far, two main features of the proteome's wiring diagram have been addressed separately: Molecular interactions and localization. In a series of seminal studies, human protein-protein interactions have been mapped using exogenous expression strategies with epitope tags coupled to immunoprecipitation-mass spectrometry (IP-MS)[6,7] or yeast two-hybrid (Y2H)[8], while protein localization has been charted using immuno-fluorescence in fixed samples[9]. However, the twin goals of systematically measuring human protein interactions under native expression regulation (which can be more precise than over-expression or Y2H[10]) and studying localization in live, unperturbed cells, remain unmet. By contrast, pioneering work in the budding yeast *S. cerevisiae* has demonstrated how libraries of endogenously tagged strains – made possible by the relative simplicity of homologous recombination in yeast[11] – can enable the systematic mapping of localization and interactions in a eukaryotic proteome[12–14]. Recent developments in gene editing, fueled by rapid advances in CRISPR technologies, now allow for similar strategies to be applied for the systematic interrogation of the human proteome[15,16].

Here, we combine genome engineering, advanced machine learning, live-cell microscopy and high-speed MS-based proteomics to generate 'OpenCell', a systematic map of localization and interactions across a large portion of the human proteome. We develop a high-throughput CRISPR-Cas9 pipeline to generate a library of ~1,300 human HEK293T cell lines harboring fluorescent protein fusions under native expression levels. Pairing confocal microscopy in live cells with IP-MS using the fluorescent tags for capture enables – for the first time – systematic measurements of localization and interactions from the same samples. For a quantitative description of cellular architecture, we encode each protein's interaction and localization signatures as vectors that allow data-driven comparisons, leveraging in particular a new machine learning model for image encoding. This analysis allows us to delineate communities of functionally related proteins by unsupervised clustering and to extract a hierarchical description of the organization of the human proteome. It also reveals that localization patterns measured by light microscopy often contain enough information to predict interactions at the molecular scale. We next illustrate how 'OpenCell' can be used for the exploration of the cellular protein interactome in space resulting in novel mechanistic insights. Finally, to facilitate access and exploration, we also present a fully interactive website ('OpenCell'.czbiohub.org) which includes an intuitive interface to explore our imaging dataset in 3D and the full connectivity of our protein interaction network. 'OpenCell' constitutes a proteome-scale, open-source resource coupled to a series of analytical examinations that augment the quantitative description of human cellular architecture.

The 'OpenCell' library

To systematically map localization and interactions across a large portion of the human proteome, we constructed a library of fluorescently tagged HEK293T cell lines by targeting ~1700 human proteins with the split-mNeonGreen2 system[17] (Fig. 1A). Split-fluorescent proteins (FP) constructs enable functional tagging with short sequences and greatly simplify CRISPR-based genome engineering[15], which allowed us to generate FP fusions directly into endogenous genomic loci and to preserve native expression regulation (Fig. 1B). Importantly, FPs enable both the study of protein localization by fluorescence microscopy and that of protein-protein interactions by acting as handles for IP-MS[15,18] (Fig. S1A, Methods). FP insertion sites (N- or C-terminus) were predominantly informed by literature curation or structural analysis. For each tagged target we isolated a polyclonal pool of CRISPR-edited cells, which was then characterized by live-cell 3D confocal microscopy, IP-MS, and genotyping of tagged alleles by next-generation sequencing (Fig. 1C). Notably, our high-throughput cell biology pipeline is supported by open-source software development and advances in instrumentation (Fig. 1C). In particular, we developed '*crispycrunch*', a software for CRISPR-based integration experiments enabling guide RNA selection and homology donor sequence design (github.com/czbiohub/crispycrunch). We also fully automated microscopy acquisition in Python to enable on-the-fly computer vision and selection of desirable fields of view (github.com/xxx). Furthermore, our mass-spectrometry protocols take advantage of the high sensitivity and speed of trapped ion mobility time-of-flight mass spectrometers (timsTOF)[19], which allowed miniaturization of IP-MS down to 0.8E6 cells of starting material (Fig. S1B, a >10-fold reduction compared to previous approaches[6,7]). Lastly, our quantitative approach for the analysis of interactions or localization patterns and the companion 'OpenCell'.czbiohub.org website (Fig. 1D) are further described in the next sections.

In total, we targeted 1728 genes, of which 1275 (74%) could be successfully detected by fluorescence and form the current version of our dataset. From these, we obtained paired IP/MS measurements for 1210 targets (95% of the publication set, Fig. 1E). RNASeq analysis revealed that "unsuccessful" targets shared significantly lower expression levels than successful ones (Fig. 1E, right panel), and a correlation between gene expression and fluorescence identified the expression threshold corresponding to our fluorescence detection limit (log[RNA tpm] = 1.5, Fig. 1F). This threshold corresponds to the median expression level in the HEK293T line (Fig. S1C), meaning that the top ~50% of expressed genes are detectable at endogenous level using current FPs. Because a single edited allele is sufficient to confer fluorescence (HEK293T are pseudo-triploid[20]), we used a stringent fluorescent cell sorting (FACS) strategy to significantly enrich for homozygous integrations (Fig. S1D). Our final cell library exhibited a median of 62% of mNG11-integrated alleles (68% among non-wt alleles, Fig. 1G). Because non-homologous end joining competes with homologous recombination for the repair of CRISPR-induced genomic breaks[21], a median 26% of all alleles in our library harbor non-functional mutations. These alleles do not support fluorescence and are therefore unlikely to impact downstream measurements, especially in the context of a polyclonal pool. The genotype information for all of our cell lines is available in Suppl. Table X and included on our website.

Because of technical constraints, protein function in human cells has so far mostly been profiled using overexpression methods. However, overexpression can have a multitude of adverse impacts on function and complicates mechanistic interpretation[16,22,23]. A few of these impacts are illustrated in Figure S1F: aberrant localization (SPTLC1), changes in organellar morphology (TOMM20) or masking effects (MAP1LC3B). We verified that our sorted cell populations maintained endogenous levels of protein expression by quantitative Western blotting of 12 tagged targets (Fig S1E). The median abundance level of the target proteins in engineered cells was 79% of that target's abundance in *wt* cells. Altogether, 'OpenCell' is a polyclonal library strongly enriched for homozygous insertions that supports the functional profiling of tagged proteins at near-endogenous expression levels, paving the way for re-examination of the human proteome under near-physiological conditions

The 'OpenCell' interactome

To map protein-protein interactions, we isolated tagged proteins ("baits") from cell lysates solubilized in digitonin, a very mild non-ionic detergent known to preserve the native structure and properties of membrane proteins[24]. Specific protein interactors ("preys") were identified from biological triplicate experiments using label-free bottom-up proteomics on a timsTOF instrument[19] (see Figure S2A for a detailed description of our statistical analysis, which builds upon established methods[6]). In total, the full interactome from our 1,210 'OpenCell' baits includes 25,212 interactions between 4,978 proteins (baits and preys, Fig. 2A). To assess the quality of our interactome, we estimated its precision and recall using reference data. To estimate recall, we used CORUM[25] – a set of protein interactions manually curated from the literature – and quantified the coverage of these interactions in our own data. To estimate precision, we quantified how many of our interactions involve protein pairs expected to localize to the same broad cellular compartment[26] (Fig. S2A, bottom panel). To benchmark "OpenCell" against other large-scale interactomes, we compared its precision and recall to Bioplex (overexpression of HA-tagged baits[7,27]), HuRI (Y2H[8]) and our own previous data (GFP fusions expressed from bacterial artificial chromosomes[6]) (Fig. S2B-D). We also calculated compression rates for each dataset as a measure of the overall inter-connectivity in the interaction networks[28] (Fig. S2E). Across all metrics, 'OpenCell' outperformed previous approaches. Together, these findings establish the high quality of our interactome, which likely reflects both our preservation of near-endogenous protein expression and the sensitivity of our mass spectrometry analyses.

To investigate the global properties of human protein interactions, we analyzed the distribution of the number of interactors for each tagged target (Fig. 2B). While this distribution approximates a power-law for moderate interaction counts (i.e., a linear relationship in log-log scale, Fig. 2B), our data highlights more targets with a large number of interactors ($N_{interaction}\geq 100$ in Fig. 2B) than expected from a "scale-free" model of protein interaction networks[29], as has been noted in other analyses[30]. We discovered that highly interacting proteins are not simply the most abundant (Fig. 2C), but rather exhibit specific biophysical signatures. A sequence-base analysis showed that highly interacting proteins are significantly less α-helical and less hydrophobic than other proteins, but more likely to contain disordered domains (Fig. 2D, 2E). Ontology analysis also revealed that the group is enriched for RNA-binding proteins (Fig. 2F). The propensity of highly interacting proteins for high intrinsic disorder has been recognized in Y2H datasets[31]. Intrinsic disorder is also a common property of proteins that form liquid condensates, which include a large number of RNA-binding proteins[32,33]. Interestingly, disorder has been proposed to be under positive selection in viral proteomes as a means for the relatively small number of proteins encoded in viral genomes to be able to interact pleiotropically with the host machinery[34]. Our data suggests that intrinsic disorder might also have been selected in human protein sequences to maximize numbers of interactions, and that the same biophysical properties that drive liquid phase transition may also play a more general role in shaping the human interactome.

A powerful way to interpret interactomes is to identify communities of interactors. These communities highlight complexes or functional pathways and also facilitate the assignment of protein function via "guilt-by-association"[7,13]. To this end, we applied unsupervised Markov clustering (MCL)[35] to the graph of interactions defined by our data (including all baits and preys). We first measured the stoichiometry of each interaction using a quantitative approach we previously established[6], and used it to weigh the edges in the interaction graph (Fig. 2G). A first round of stoichiometry-weighted MCL delineated inter-connected protein communities, outperforming clustering on the basis of connectivity alone (Fig. S2F). To further refine annotations, we subjected each MCL community (separately) to another round of clustering in which low-stoichiometry interactions were removed. The resulting sub-clusters outline core interactions within existing communities (Fig. 2G). An illustrative example of how this unsupervised approach allows us to delineate functionally related proteins is shown in Figure 2H: All subunits of the machinery responsible for the translocation of newly translated proteins at the ER membrane (SEC61/62/63) and of the EMC (ER Membrane Complex) are grouped within respective core interaction clusters, but both are part of the same larger MCL community. This mirrors the recently appreciated co-translational role of EMC for insertion of transmembrane domains at the ER[36]. Interestingly, additional proteins, which have only been recently shown to have a co-translational role, are found to be clustering with the translocon or EMC subunits. These include ERN1 (IRE1), a folding sensor[37], and CCDC47, a poorly characterized translocon interactor which, like EMC, regulates the biogenesis of membrane proteins at the ER[38,39]. This highlights how clustering can facilitate mechanistic exploration by grouping together proteins involved in related pathways. Overall, we identified 300 communities including a total of 2,097 baits and preys. A graph of interactions between communities reveals a richly inter-connected network (Fig. 2I), the structure of which outlines the global architecture of the human interactome (discussed further below).

The 'OpenCell' image dataset

To profile the localization of tagged proteins we used live-cell fluorescence microscopy on a spinning-disk microscope with a 63x 1.45NA objective under atmospheric control (37°C, 5% CO_2), and imaged the full 3D distribution of proteins in consecutive confocal z-slices. We fully automated microscopy acquisition in Python to enable scalability (Fig. S3A-B). In particular, we trained a computer vision model to identify fields of view (FOVs) with homogeneous cell density on-the-fly, significantly reducing uncontrolled experimental variation between images. Our resulting dataset contains a collection of 6375 3D stacks (5 different FOVs for each target) and includes paired imaging of nuclear morphology with Hoechst 33342, a cell-permeable DNA dye compatible with live-cell measurements.

We manually annotated localization patterns by assigning each protein to one or more of 15 separate cellular compartments such as nucleolus, centrosome or Golgi apparatus (see Figure 3A for the full list and example images). Because proteins often populate multiple compartments at steady-state[9], we annotated localizations using a three-tier grade system: grade 3 identifies the most prominent localization compartment(s), grade 2 represents clearly detectable but minor localizations, and grade 1 annotates weak localization patterns nearing our limit of detection (see Figure S4A for representative examples). Ignoring grade 1 annotations which are inherently less precise, we find that 55% of proteins in our library are multi-localizing and outline known functional relationships with, for example, clear connections between secretory compartments (ER, Golgi, vesicles, plasma membrane), or between cytoskeleton and plasma membrane (Fig. 3B). The most common pattern of multi-localization involves proteins found in both nucleus and cytoplasm (21% of our whole library), highlighting the importance of the nucleo-cytoplasmic import/export machinery in shaping global cellular function[40,41]. Importantly, because our split-FP system does not enable the detection of proteins in the lumen of organelles, multi-localization involving translocation across an organellar membrane (which is rare but does happen for mitochondrial or peroxisomal proteins) will not be detected in our data.

Extracting functional insights directly from cellular images is a major goal of modern cell biology and data science[42]. Because the function of a protein is tightly linked to its localization, we explored whether a quantitative comparison of localization signatures would allow us to delineate groups of co-functioning proteins. For this, we developed a machine learning model we describe in a companion study (ref. TBD). Briefly, our model is a variant of an autoencoder (Fig. 3C): a form of neural network that learns to vectorize an image through paired tasks of encoding (from an input image to a vector in a latent space) and decoding (from the latent space vector to a new output image).

After training, a consensus representation for a given protein can be obtained from the average of the encodings from all its associated images (Fig. 3C). One of the main advantages of this approach is that it is unsupervised: Training the model only involves minimizing differences between input and output, ensuring that the latent-space vectors capture the properties of the input images. Therefore, as opposed to other machine learning strategies that are trained to recognize pre-annotated patterns (for example, manual annotations of protein localization[43]), our method allows us to compare images without any *a priori* assumptions and to objectively measure the similarity between two localization signatures.

A UMAP representation of the localization encodings for the entire 'OpenCell' library is shown in Figure 3D. This map is organized in distinct territories that closely match manual annotations (Fig. 3D, highlighting mono-localizing proteins). This validates that our approach yields a quantitative representation of the biologically relevant information in our microscopy data. We then asked what degree of functional relationship could be inferred between proteins solely on the basis of their localization patterns. For this, we employed an unsupervised Leiden clustering strategy classically used to identify cell types in single-cell RNA sequencing datasets[44]. Strikingly, applying this data-driven approach identified groups of proteins that are closely related mechanistically (178 clusters in total, full list in Suppl. Table X). For example (Figure 3E), not only could our analysis separate P-body proteins (cluster #97) from other forms of punctuated cytoplasmic structures, but different vesicular trafficking pathways could also be unambiguously differentiated despite very similar localization patterns: the endosomal machinery (cluster #35), plasma membrane endocytic pits (cluster #167) or COP-II vesicles (cluster #40) were all delineated with high precision. Proteins involved in closely inter-related cellular functions were also found to cluster together: For example, the ER translocon clusters with the SRP receptor, EMC subunits and the OST glycosylation complex, all responsible for co-translational operations (cluster #22). Clustering performance was also high for non-organellar proteins, as shown in Figure S4B (cytoplasmic clusters) and Fig S4C (nuclear clusters). Altogether, our results show that the localization pattern of a given protein can characterize its function with high specificity, down to specific pathways, and that this information can be captured by unsupervised machine learning algorithms. While closely related proteins cluster tightly together, the existence of many separated groups also underlines the tremendous diversity of localization patterns across the full proteome. Example images from nuclear proteins offer a compelling illustrative example of this diversity (Fig. S4D).

Interactive community access at 'OpenCell'.czbiohub.org

To enable community to access and exploration of our multi-layered data set , we built the interactive web app 'OpenCell' that provides side-by-side visualizations of the confocal images and interaction network for every protein in our analysis (Figure 4). The app is organized around a 'target profile' page that displays all of the metadata, images, and interactions for a selected mNG11-tagged target (Fig. 4B). Confocal fluorescent images can be visualized either in 2D as a scrollable stack of z-slices, or in 3D via an interactive volume rendering module we developed (Fig. 4C). Our interface also allows the user to toggle between fluorescence channels (tagged protein and DNA stain) and to adjust image brightness and gamma levels. The interaction network, which consists of the target, its direct interactors, and the interactions between them, is organized by the communities and core clusters identified by MCL and is positioned directly adjacent to the image viewer (Fig. 4B). This side-by-side layout of images and interactions encourages the comparison between subcellular localizations and interaction signatures. The interactome visualization can be switched to reveal quantitative data for each pulldown in the form of volcano and stoichiometry plots (Fig. 4D). To enable the interactive navigation of the full interaction network, interactors in the network that are themselves 'OpenCell' targets are hyperlinked (from any visualization mode) to their corresponding target pages. Interacting proteins in the network that are not 'OpenCell' targets are likewise hyperlinked to a distinct 'interactor profile' page that displays information about all the pull-downs in which the interacting protein appears. Finally, to explore the dataset by localization pattern, a separate 'gallery' page displays a grid of thumbnail microscopy images for all targets in the library filtered according to a user-defined set of subcellular localizations (Figure 4A, right image). The app itself is built with modern web technologies and is supported by a relational database and a REST API that also allows for programmatic access to the underlying raw data.

Interactions vs. spatial relationships

IP-MS and microscopy examine the architecture of the cellular proteome at very different scales (molecular for IP-MS, pan-cellular for microscopy). Localization and interactions are linked: proteins must localize together to interact. But while the localization patterns of each interactor need to overlap, they do not need to match completely. Therefore, the degree to which interacting proteins also share the same overall spatial signature represents a global principle that shapes the cellular protein network. To quantify the similarity of localization of any two targets in 'OpenCell', we measured the Pearson correlation between their localization encodings; this gives a distance measure of similarity in "localization space" between two proteins (Fig. 5A). In parallel, similarities can be measured in "interaction space" by comparing the interaction profiles between any two proteins based on stoichiometry information (Fig. 5A). As expected, the distribution of localization vs. interaction similarities between all pairs of 'OpenCell' targets that were found to interact shows a positive correlation between the two parameters (Fig. 5B); however, it also highlights that the vast majority of interacting proteins are not particularly similar by either measure (large cloud around the origin of the graph). Examining the relationship between localization similarity and interaction stoichiometry further reveals two separate groups of interacting pairs (Fig. 5C): 1) the vast majority of interaction partners, which interact with low stoichiometry and whose spatial signatures do not specifically overlap (i.e., have low localization similarity, solid line in Fig. 5C), and 2) a smaller but well-delineated group of stoichiometric interactors, which share very similar localization patterns (dashed line). This result makes intuitive sense: interactions that are stoichiometric should be stable in space and time, and a high degree of co-localization is expected. Indeed, different subunits of known stable complexes share extremely similar localization signatures and form tight clusters in our image UMAP (Fig. 5D). But this intuitive result also has an important correlate: that highly similar localization patterns between two proteins can be used to infer close molecular interaction. In fact, looking at the entire set of 'OpenCell' target pairs (predicted to interact or not), proteins that share high localization similarities are also very likely to interact (Fig 5E). For example, target pairs with a localization similarity greater than 0.85 have a 58% chance of being direct interactors, and a 68% chance of being second-neighbors (i.e., sharing a direct interactor in common). Overall, this analysis demonstrates that a quantitative comparison of localization patterns can also make predictions about the molecular-level architecture of the proteome. It also further emphasizes the organization of the cell's proteome along two modes of interactions: small communities of high-stoichiometry protein groups, whose functions are intertwined to the point that their steady-state localization patterns are very similar, and a much larger set of low-stoichiometry, and presumably more dynamic interactions with much lower spatial overlap.

Biological discovery using interactomes or images

Unsupervised clustering of both localization and interaction signatures can be used to derive functional relationships between proteins, unambiguously linking together different forms of co-translational processes, for example. A direct application of this result is to help elucidate the cellular roles of poorly characterized human proteins[45]. We first mined our interactome for poorly studied proteins. We focused on the set of baits and preys found within MCL communities, reasoning that community association was a strong signal to map function via "guilt by association". As a simple metric for how well characterized a protein is, we quantified its occurrence in article titles and abstracts from PubMed. Empirically, we determined that proteins in the bottom 10th percentile of publication count (corresponding to less than 10 publications) were very poorly annotated (Fig. 6A). This set encompasses a total of 251 proteins for which our dataset offers specific mechanistic insights. For example, the poorly characterized NHSL1, NHSL2 and KIAA1522 proteins are all found as part of an interaction community centered around SCAR/WAVE, a large multi-subunit complex nucleating actin polymerization (Fig. 6B). Interestingly, all three proteins share sequence homology (Suppl. Doc. X) and are all homologous to NHS (Suppl. Doc. Y), a protein mutated in patients with Nance-Horan syndrome and that interacts with SCAR/WAVE components to coordinate actin remodeling[46]. This suggests that NHSL1, NHSL2 and KIAA1522 may also play a role in actin assembly. A recent mechanistic study made public after our initial prediction supports this hypothesis: NHSL1 was found to localize at the cell's leading edge and to directly bind SCAR/WAVE to negatively regulate its activity, reducing F-actin content in lamellipodia and inhibiting cell migration[47]. Importantly, the authors identified NHSL1's SCAR/WAVE binding sites, and we find these sequences to be conserved in NSHL2 and KIA1522 (Fig. 6B). Therefore, we propose that both NHSL2 and KIAA1522 are also direct SCAR/WAVE binders and new modulators of the actin cytoskeleton.

Our data also uncover a specific function for ROGDI, whose variants cause Kohlschuetter-Toenz syndrome,a recessive developmental disease characterized by epilepsy and psychomotor regression[48]. ROGDI appears in the literature because of its association with disease, but to our knowledge, no study specifically determined its molecular function until now. To delineate the function of ROGDI, we first observed that its interaction pattern closely matched that of three other proteins in our dataset: DMXL1, DMXL2 and WDR7 (Fig. 6C). This set exhibited a specific interaction signature related to v-ATPase, the lysosomal proton pump. All four proteins interact with soluble v-ATPase subunits (ATP6-V1), but not its intra-membrane machinery (ATP6-V0). Interestingly, DMXL1 and WDR7 have been shown to interact with V1 v-ATPase, and their knock-down compromises lysosomal re-acidification[49]. In fact, sequence analysis showed that DMXL1/2, WDR7 and ROGDI are homologous to proteins from yeast or Drosophila that have been involved in the regulation of assembly of the soluble V1 subunits onto the V0 transmembrane ATPase core[50,51] (Suppl. Fig. X). In yeast, Rav1 and Rav2 (homologous to DMXL1/2 and ROGDI, respectively) form the stoichiometric RAVE complex, a soluble chaperone that regulates v-ATPase assembly [51]. To further characterize the function of these proteins, we generated tagged cell lines for DMXL1/2, WDR7 and ROGDI and combined our IP-MS interactome analysis and imaging pipeline. Because of the low expression level of these proteins, imaging analysis proved difficult, and fluorescent fusions of DMXL2 and ROGDI did not lead to detectable fluorescence. However, pull-downs of DMXL1 and WDR7 confirmed a stoichiometric interaction between DMXL1/2, WDR7 and ROGDI (Fig. 6C, right panels). Interestingly, no direct interaction between DXML1 and DMXL2 was detected, suggesting that they might nucleate two separate sub-complexes. Therefore, our data uncovers the existence of a human RAVE complex comprising DMXL1/2, WDR7 and ROGDI, which likely acts as a chaperone for v-ATPase assembly. Altogether, NHSL1/2-KIAA1552 and DMXL1/2-WDR7-ROGDI illustrate how 'OpenCell' catalyzes new biological insights by combining quantitative analysis, literature curation and new functional data – including a direct mechanistic role for ROGDI, shedding light on the biology of Kohlschuetter-Toenz syndrome.

These examples highlight the power of mining interactome data for mechanistic predictions. However, having established that localization on its own could expose the molecular function of a given protein, we asked to which extent the function of an orphan protein could be characterized by imaging alone. FAM241A (or C4orf32) is a human protein without any functional annotation. Our initial interactome dataset placed FAM241A as part of a community centered around the OST complex, the transferase responsible for co-translational glycosylation (Fig. 6D). We generated an endogenous FAM241A fluorescent fusion and separately used imaging and mass-spectrometry to elucidate its function. Importantly, the autoencoder model we used to generate the localization encoding was not trained with images of this new target ("naïve" model). Strikingly, the quantitative distances between FAM241A and other 'OpenCell' targets measured using either localization or interaction signatures both identified FAM241A as a new OST subunit (Fig. 6D). Moreover, separate unsupervised clustering analyses using the two signatures both placed FAM241A in well-defined clusters with other OST subunits (Fig. 6D, right panels). Thus, the function of FAM241A could have been predicted with the same degree of precision by using either its interaction signature or its localization encoding. This proof-of-concept example establishes the potential of live-cell imaging as a specific readout of protein function, including the characterization of poorly studied human proteins.

Hierarchical structure(s) of proteome organization

Finally, we explored the global structure of our datasets by using hierarchical clustering to highlight the signatures patterning the proteome. Starting with the 300 interactome communities or the 178 localization clusters outlined above, we mapped the full hierarchical relationships underlying both our interactome and imaging sets. Specifically, we implemented an agglomerative clustering strategy based on node pair sampling[52], which we applied separately to the graph of interactions between interactome communities and to the graph connecting localization clusters derived from the corresponding UMAP adjacency matrix[53]. This resulted in fully connected hierarchical trees, as shown in Figure 7. Isolating groups of proteins at separate hierarchical layers reveals different levels of the proteome's organization. At an intermediate layer, the proteome can be delineated into sets of 16-18 separate "modules", each including a median of 95 proteins for the interactome (modules M1-M18, Fig. 7A) and of 86 proteins for the imaging dataset (modules N1-N16, Fig, 7B), respectively. At higher layer, each dataset can be divided into three "branches", which separately represent the core features that shape the proteome's global architecture from a molecular or spatial perspective. Gene ontology analysis underlined that modules and branches are enriched for specific cellular functions or compartments, which define unique functional signatures (labeled in Fig 7A,7B – full analysis in Figure S5 for branches and Suppl. Table X for modules).

Overall, the hierarchy of localization encodings reveals that, as expected, localization patterns are organized along the three foundational compartments of the eukaryotic cell: Nucleus, cytoplasm and membrane-bound organelles (Fig. 7B, Fig. S5G). Each localization branch is sub-divided into modules that correspond to discrete sub-cellular territories. For example, the organellar branch separates into the different components of the secretory pathway: ER, Golgi, endosome, lysosome or plasma membrane, mirroring the known spatial segregation between these compartments. The fact that this unsupervised data-driven clustering strategy can recapitulate known cellular compartments validates the performance of our approach. By contrast, the hierarchical analysis of the interactome graph reveals how the proteome is organized at molecular scale (Fig. 7A). The 18 modules separate the interactome into clear cellular functions such as transcription, splicing or vesicular transport. This reflects that functional pathways are templated by groups of proteins that physically interact, a principle also underscored by the overlap between genetic interactions and protein complexes in eukaryotes[54,55]. More interestingly, analysis of the high-level branches reveals a separation between three groups of proteins that differ in their functional and biophysical properties. Ontology enrichment highlights that proteins related to membrane processes (branch B) and RNA-binding (branch C) clearly segregate from the rest of the proteome in term of their interaction profiles (Fig. S5A-E). This functional segregation is correlated with different biophysical properties: Branch B proteins are significantly more structured, more hydrophobic and richer in aromatic residues than the rest of our dataset; conversely, branch C proteins have higher intrinsic disorder and higher isoelectric points (Fig. S5B-C). Overall, our data reveal that RNA-binding proteins form a specific molecular sub-group that shapes the global organization of the cell, similar to how association with membranes is a molecular feature that sets apart a large fraction of the proteome. Strikingly, the same group of proteins is known to form membrane-less organelles under stress conditions, especially through liquid-liquid phase transition processes supported by intrinsic disorder [32,33]. This suggests that the molecular and biophysical properties underlying liquid phase transition might also be a global driving force shaping the cellular proteome network under normal conditions.

Discussion

'OpenCell' combines innovations at four separate levels to augment the description of human cellular architecture. First, we describe an integrated experimental pipeline for high-throughput cell biology, fueled by scalable methods for genome engineering, live-cell microscopy and MS-based interaction proteomics. Second, we provide an open-source resource of well-curated localization and interactome measurements, easily accessible through an interactive web interface at 'OpenCell'.czbiohub.org. Third, we pioneer new analytical strategies for the representation and comparison of interaction or localization signatures (including a fully unsupervised machine learning approach for image encoding). And fourth, we demonstrate how our dataset can be used both for fine-grained mechanistic exploration (by elucidating the function of multiple proteins that were previously uncharacterized), as well as for investigating the core organizational principles that wire the proteome. In particular, we uncover two global features that shape cellular proteome architecture. First, we show that most proteins interact with low stoichiometry and distribute unequally within the cell, whereas high-stoichiometry interactors share very similar localization patterns. This reinforces the importance of low-stoichiometry interactions for defining the overall structure of the cellular network, not only providing the essential interactome "glue" [6], but also connecting different cellular compartments. Second, we reveal that two separate groups of interacting proteins segregate from the global proteome: Membrane-related and RNA-binding, both of which exhibit specific biophysical signatures (in particular hydrophobicity and high intrinsic disorder, respectively). That membrane-related proteins form a specific interaction group is perhaps not surprising as the two-dimensional membrane surface drives their sequestration within the three-dimensional cell. By contrast, the discovery of RNA-binding proteins as a separate sub-group is significant given the growing appreciation that RNA-binding proteins, together with RNAs themselves, can form separate liquid phases in the cytoplasm and nucleoplasm. Interestingly, intrinsic disorder is an important modulator of partition into liquid phases, as are specific protein-protein interaction domains[32,33,56,57]. Therefore, RNA-binding proteins might have evolved to multiply molecular interactions between themselves to create compartments that can concentrate a large functional variety of proteins (for example RNA processing factors, translation machinery or silencing complexes[58,59]) and enable the spatial specialization of cellular processes. In this context, it is interesting to consider a role for RNA itself as an organizer of the cellular proteome, as is for example the case for some non-coding RNAs whose function is to template molecular interactions to form nuclear bodies[60].

While 'OpenCell' joins a growing number of large-scale datasets[7–9,16,26,61] that contribute to the systematic and quantitative dissection of the human cell, it is unique in its analysis of both quantitative interactomes and live-cell localization of genome-edited proteins. However, the full description of human cellular architecture remains a formidable challenge, especially considering the vast diversity of cell types and cell states that shape human physiology. For example, the ~1,300 tagged proteins included in 'OpenCell' represent only ~7% of the full set of proteins encoded in the human genome. Moreover, our approach that combines split-FP systems and HEK293T – a cell line that is heavily transformed but easily manipulatable – is mostly constrained by scalability considerations. But given the current pace of technological advances, the scale of measurements required to match the full extent of cellular complexity might soon be within reach. In particular, advances in stem cell technologies enable the generation of cell libraries that can be differentiated into multiple cell types[16]. Also, innovations in genome engineering (for example, by modulating DNA repair[62]) pave the way for the scalable insertion of gene-sized payload (e.g., fluorescent proteins, HaloTag, degron systems), combination of multiple edits in the same cell (e.g., dual-tagged libraries for co-localization studies) or increased homozygosity in polyclonal pools.

Finally, 'OpenCell' provides a large set of open-source, quantitative and curated data that can be used as a proving ground for data science and algorithm development. Our own innovation in machine learning to encode images and localization signatures was made possible by the availability of a critical mass of high-quality live-cell images taken under uniform experimental conditions – itself facilitated by our collection of genome-edited lines that can be characterized repeatedly and in a native cellular context. Our results also demonstrate the power of unsupervised machine learning models to identify complex, but deterministic signatures from light microscopy images. In particular, we show that very detailed functional relationships can be inferred from similarities between localization patterns, including the prediction of molecular interactions. This opens exciting avenues for the use of imaging as a high-throughput, information-rich method for deep phenotyping and functional genomics. Since light microscopy is scalable and can be performed on live, unperturbed samples enabling measurements at the single-cell level, this offers many opportunities for the full quantitative description of cellular proteome diversity in normal physiology and disease.

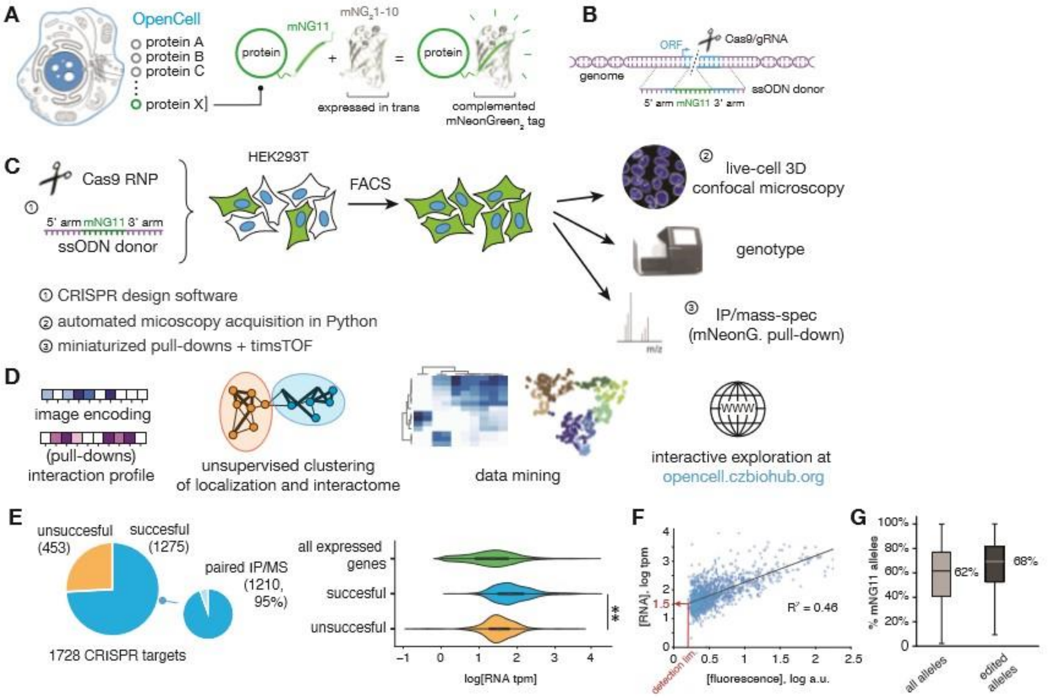


Figure 1: the OpenCell library. (A) Functional tagging with split-mNeonGreen2. **(B)** Endogenous tagging strategy: mNG11 fusion sequences are inserted directly within genomic open reading frames (ORFs) using CRISPR-Cas9 gene editing and homologous recombination with single-stranded oligonucleotides (ssODN) donors. **(C), (D)** The OpenCell experimental (C) and analytical (D) pipelines. See text for details. **(E)** Successful detection of fluorescence in the OpenCell library. Out of 1728 genes originally targeted, fluorescent signal was successfully detected for 1275 (left panel). Low expression level is the main obstacle to successful detection (right panel, showing the full distribution of RNA expression in transcripts per million reads – tpm – for all genes expressed in HEK293T vs. successfully or unsuccessfully detected OpenCell targets; **: p < 10-5, t-test). **(F)** Correlation between expression level (RNASeq tpm) and fluorescence intensity (as read by flow cytometry) for all successful OpenCell targets. A linear regression (solid line, Pearson R2 = 0.49) allows to estimate the expression threshold required for successful detection. **(G)** Genotype analysis of the polyclonal OpenCell library. A single allele is required for fluorescence, but our cell collection is enriched for homozygous insertions. In total, mNG11 insertions account for 62% (median) of all alleles, and 68% (median) of CRISPR-edited alleles (i.e., non-wild-type). Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent minimum and maximum values.

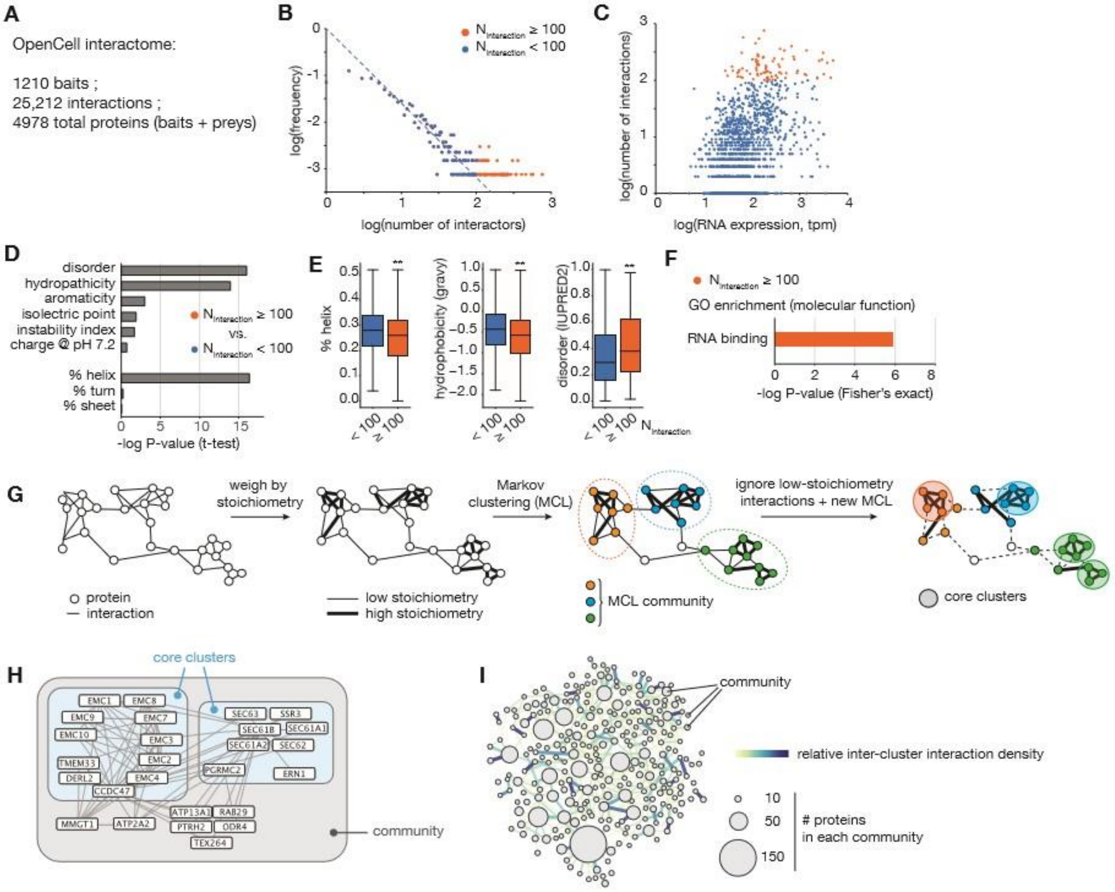


Figure 2: the OpenCell interactome. (A) Overall description of the interactome. **(B)** Frequency distribution of the number of interactors per targets on a log-log scale; the dotted line represents the linear fit for moderate interaction numbers ($4<N_{interaction}<100$). Highly interacting targets (>=100 interactors) are highlighted in orange. **(C)** Correlation between number of interactions and expression level (RNASeq tpm) for each target. High numbers of interactions are not restricted to highly expressed proteins. **(D)** Comparing biophysical properties of highly interacting (>=100 interactions) vs. other targets (<100 interactions). This analysis was performed by first breaking down the amino acid sequence of each target into 100 a.a. windows. Shown are t-tests comparing the set of 100-a.a. windows from both groups. **(E)** Full distribution of % helical 2^{ndary} structure, hydrophobicity or intrinsic disorder of the set of 100-a.a. windows from highly interacting vs. other targets. Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent minimum and maximum values. ** $p < 10^{-10}$ (t-test) **(F)** Gene ontology (GO, molecular function) enrichment analysis in highly interacting vs. other targets. **(G)** Unsupervised Markov clustering of the interactome graph. **(H)** Example of community and core cluster definition for the translocon/EMC community. **(I)** The complete graph of connections between interactome communities. The density of protein-protein interactions between communities is color-coded. The numbers of targets included in each community is represented by circles of increasing diameter.

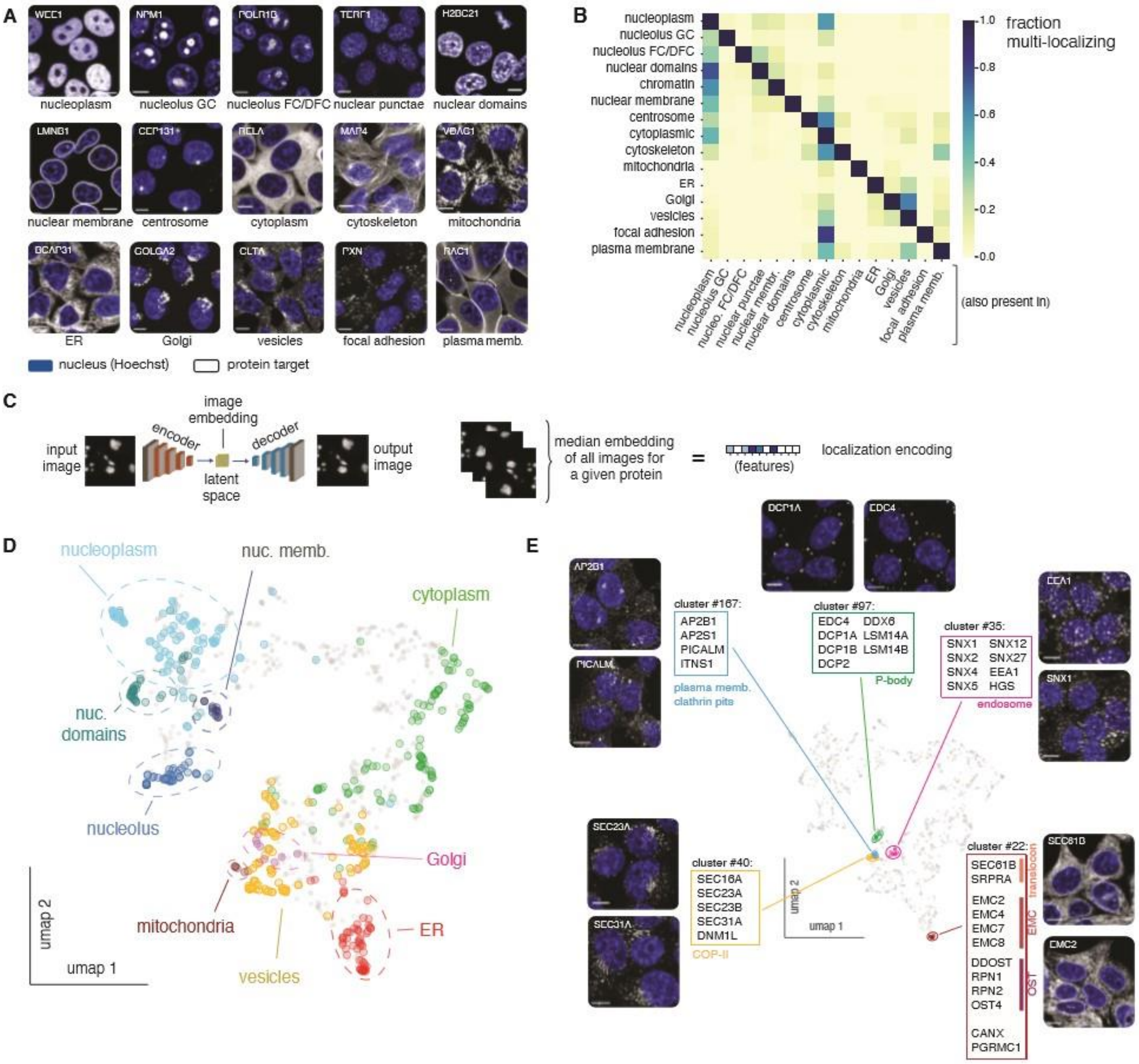
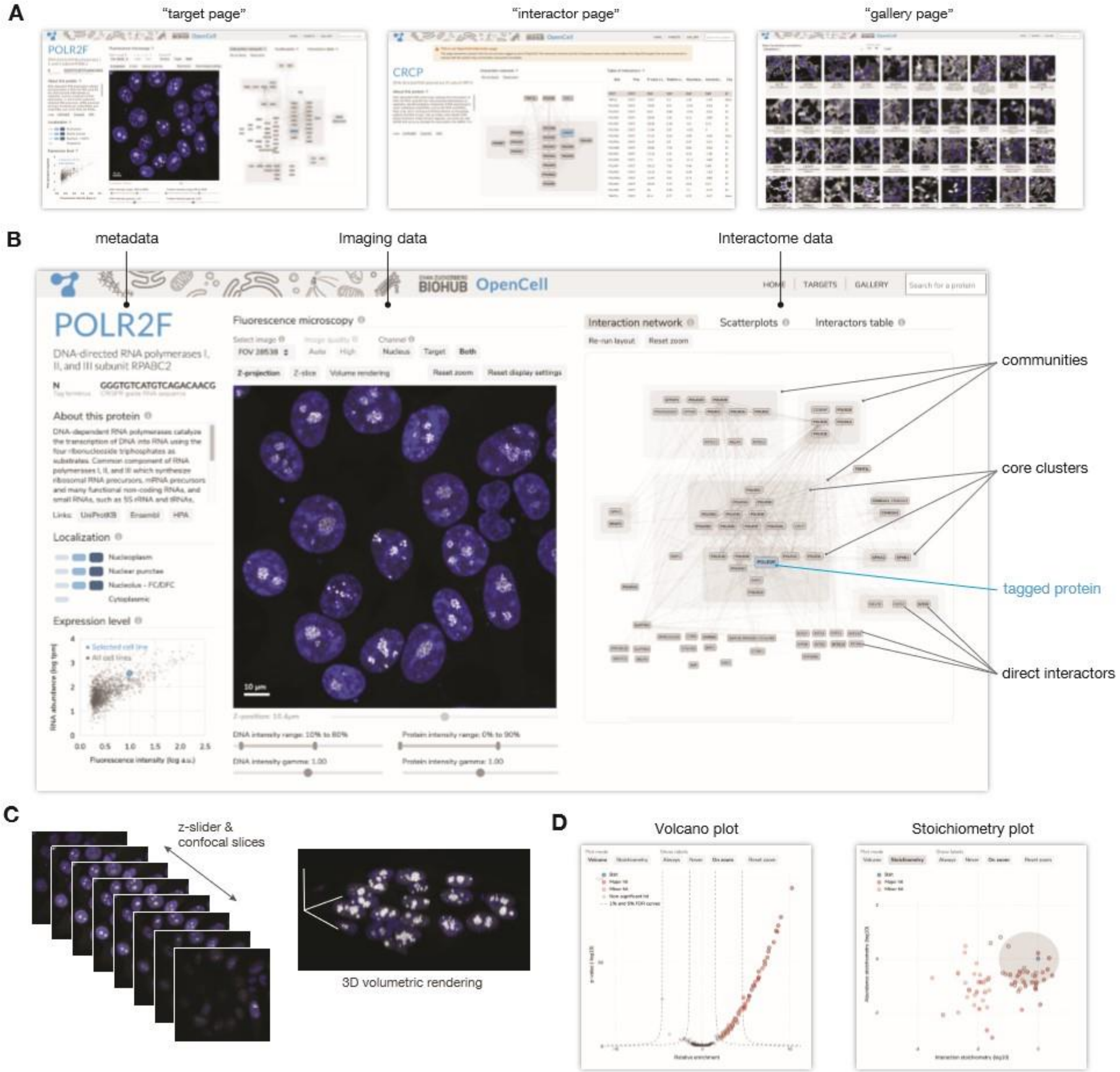


Figure 3: the OpenCell image collection. (A) The 15 cellular compartments segregated for annotating localization patterns. The localization of a representative protein belonging to each group is shown (greyscale, gene names in top left corners; scalebar: 10 µm). Nuclear stain (Hoechst) is shown in blue. **(B)** Fraction of multi-localization between cellular compartments. **(C)** Principle of localization encoding by unsupervised machine learning. See text for details. **(D)** UMAP representation of the OpenCell localization dataset, highlighting targets found to localize to a unique cellular compartment. **(E)** Examples of clusters delineated by unsupervised Leiden clustering of the localization dataset, highlighted on the localization UMAP.



POLR2F

DNA-directed RNA polymerases I, II, and III subunit RPABC2

N GGGTGTCATGTCAGACAACG

About this protein

DNA-dependent RNA polymerases catalyze the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. Common component of RNA polymerases I, II, and III which synthesize ribosomal RNA precursors, mRNA precursors and many functional non-coding RNAs, and small RNAs, such as 5S rRNA and tRNAs.

Links: UniProtKB Ensembl HPA

Figure 4: interactive data exploration at opencell.czbiohub.org. (A) The three principal pages of the OpenCell web app. From left to right: the target page, interactor page, and gallery page. **(B)** The target page consists of three columns. The leftmost column contains the functional annotation for the target from UniProt, links to other databases, our manually-assigned localization annotations, and measures of protein expression. The middle column contains the image viewer, and the rightmost column the interaction network. **(C)** The image viewer allows the user to scroll through the confocal z-slices using a slider or to visualize the z-stack in 3D as a volume rendering; in either mode, the user can pan and zoom by clicking, dragging, and scrolling. **(D)** The interaction network can be toggled with two alternative, complementary visualizations of the target's protein interactions: a volcano plot of relative enrichment vs. p-value and a scatterplot of interaction stoichiometry vs. abundance stoichiometry. In both the network view and the scatterplots, the user can click on an interactor to open the target or the interactor page for the corresponding protein.

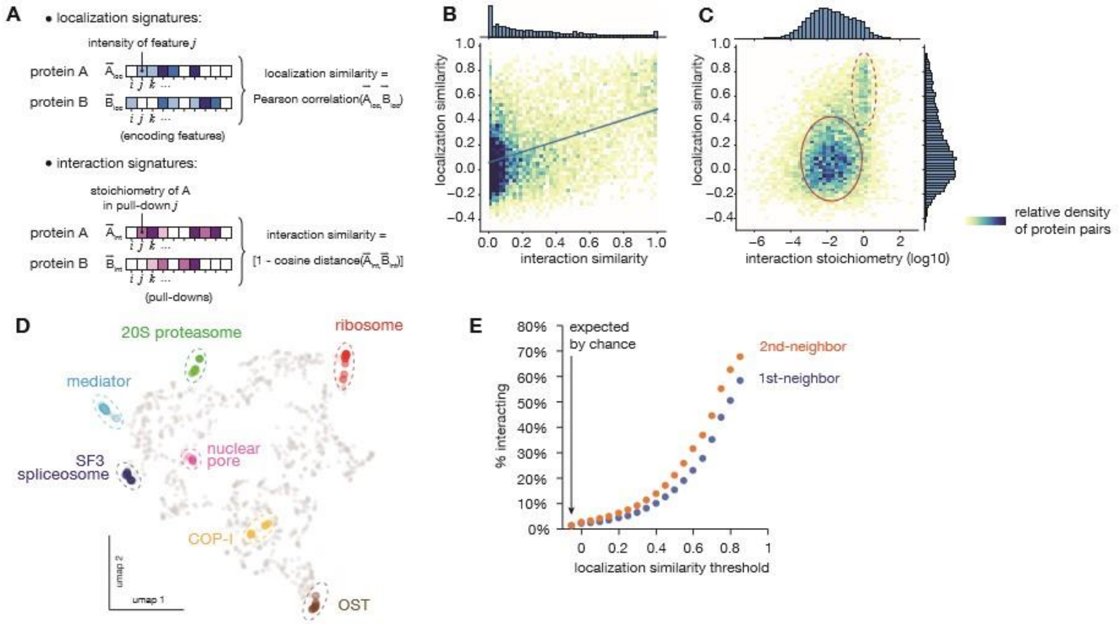


Figure 5: quantitative comparisons of localization and interaction signatures. (A) Measure of localization (top) and interaction (bottom) similarities between proteins. **(B)** Heatmap distribution of localization vs. interaction similarities between all interacting pairs of OpenCell targets. **(C)** Heatmap distribution of localization vs. interaction stoichiometry between all interacting pairs of OpenCell targets. The discrete sub-group of high-stoichiometry/high localization similarity pairs is outlined. **(D)** Localization patterns of different subunits from example stable protein complexes, represented on the localization UMAP (cf. Figure 3). **(E)** Frequency of direct (1st-neighbor) or once-removed (2nd-neighbor, having a direct interactor in common) protein-protein interactions between any two pairs of OpenCell targets sharing localization similarities above a given threshold (x-axis).

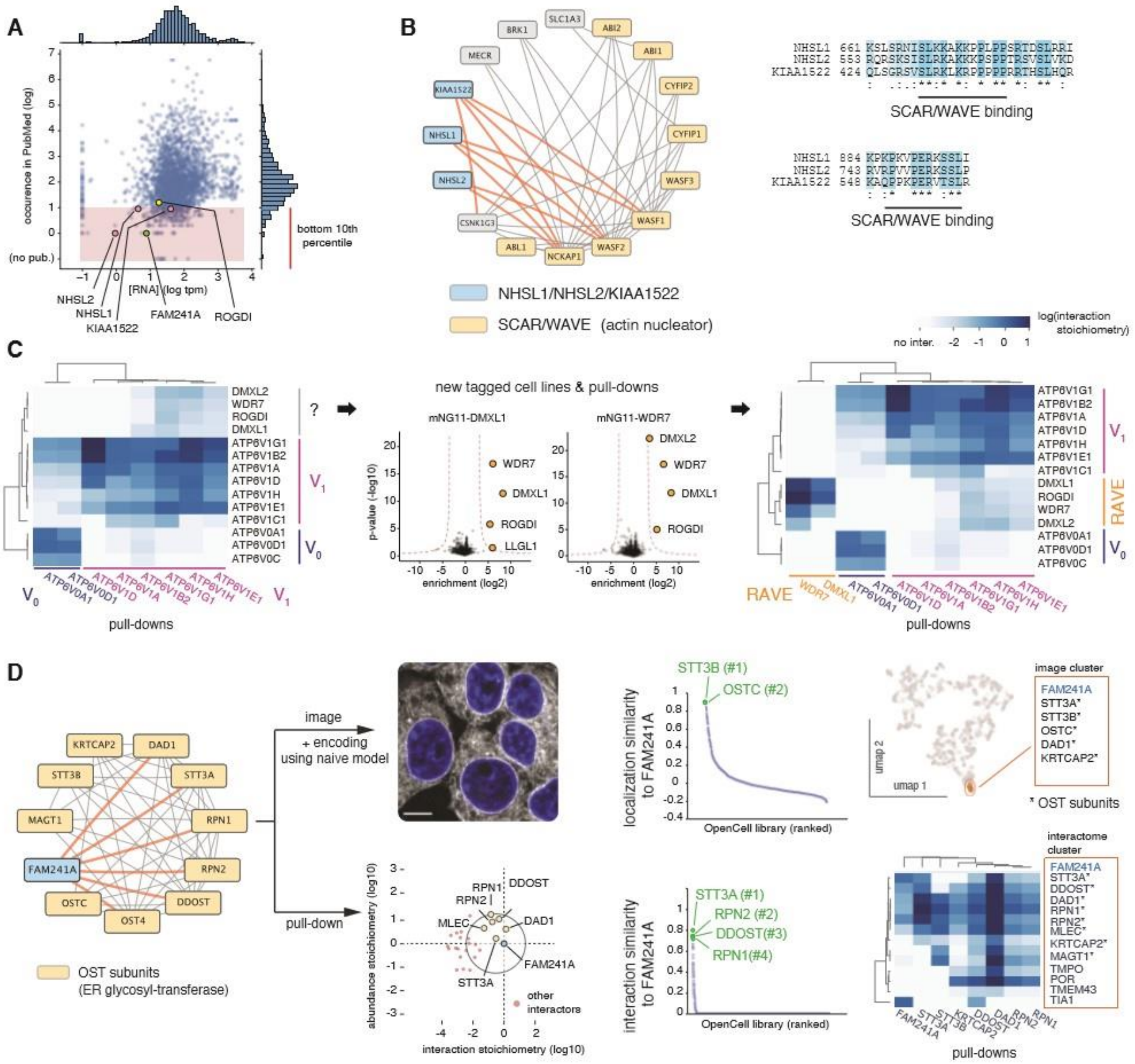


Figure 6: Biological discovery with OpenCell. (A) Distribution of occurrence in PubMed articles vs. RNA expression for all proteins found within interactome communities. The bottom 10th percentile of publication count (poorly characterized proteins) is highlighted. **(B)** NHSL1/NSHL2/KIAA1522 are part of the SCAR/WAVE community and share amino-acid sequence homology (right panel). **(C)** DMXL1/2, WDR7 and ROGDI form the human RAVE complex. Heatmaps represent the interaction stoichiometry of preys (lines) in the pull-downs of specific OpenCell targets (columns). See text for details. **(D)** Parallel identification of FAM241A as a new OST subunit by imaging or mass-spectrometry. See text for details.

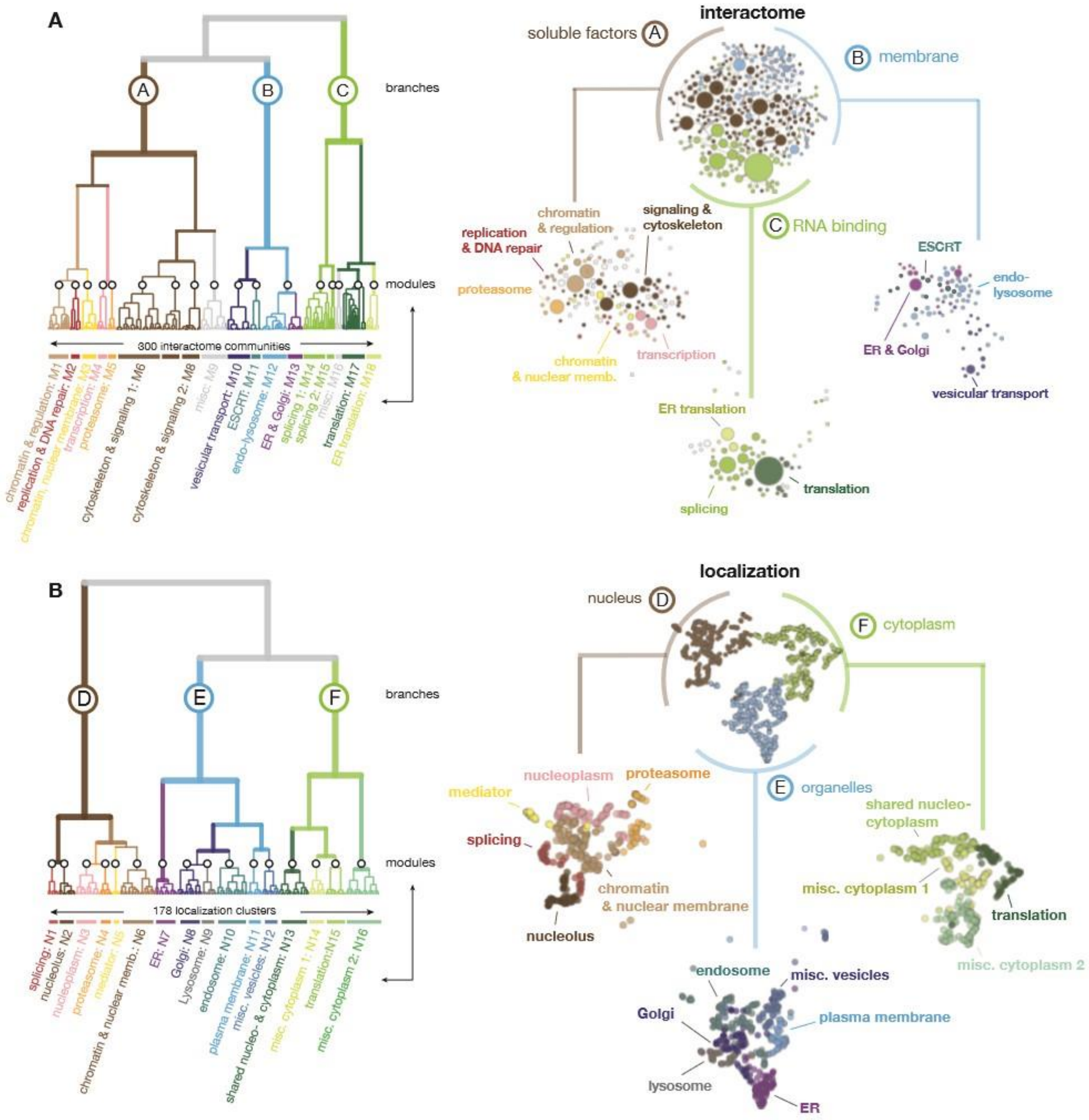


Figure 7: Full hierarchical structure of interactome and localization datasets. Dendrograms represent the hierarchical relationships connecting **(A)** the full set of protein communities identified in the OpenCell interactome (see Fig. 2) or **(B)** the full set of localization clusters identified in the OpenCell image collection (see Fig. 3). For each dataset, an intermediate layer of hierarchy separates 16-18 separate modules, while an upper hierarchical layer delineates three separate branches. Modules and branches are annotated on the basis of gene ontology enrichment analysis. Right-hand panels present the topological arrangement of branches (top) and modules (bottoms) in each dataset, highlighted from the the full graph of connections between interaction communities (“interactome”, see Fig. 2I) or from the OpenCell localization UMAP (“localization”, see Fig. 3D). The color codes between interactome and localization datasets are not directly comparable (i.e. same colors are not meant to represent the same exact set of proteins).

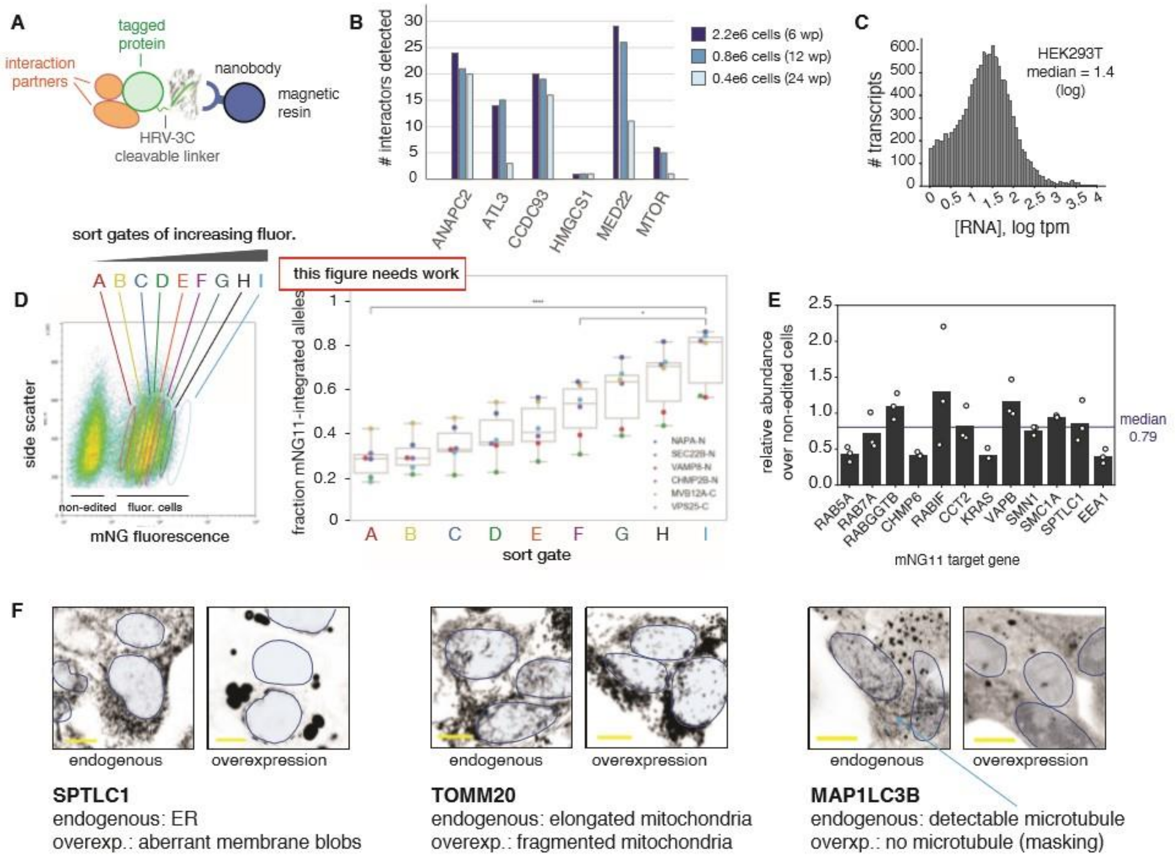


Figure S1: experimental pipeline (related to Figure 1). (A) IP/MS using FP capture. All mNG11 tagging constructs also include an HRV-3C cleavable linker for optional release from the capture resin. **(B)** Sensitivity of interaction proteomics detection on a timsTOF instrument. The number of interactors detected in pull-downs from 6 different targets is shown, varying the amount of input material. To balance sensitivity and scalability, 0.8e6 cells were used for high-throughput assays. **(C)** Distribution of transcript abundance in HEK293T. **(D)** Optimization of sorting strategy. Polyclonal cell pools were sorted using gates of increasing fluorescence (left panel) and genotyped to quantify the enrichment for mNG11-inserted alleles (right panel, showing data for 6 different target genes). This informed our final sorting strategy in which the top 1% of fluorescent cells (gate I) were selected. **(E)** Measurement of target protein abundance by quantitative Western blotting in final selected cell pools vs. parental cell line. **(F)** Examples of overexpression artifacts. Single z-slice confocal images are shown (scale bar: 10 μm). Endogenous and overexpression cells were not imaged using the same laser power, so that signal intensities are not directly comparable. Nuclei are shown as blue outlines (nuclei can be located in a different z-plane than the one shown).

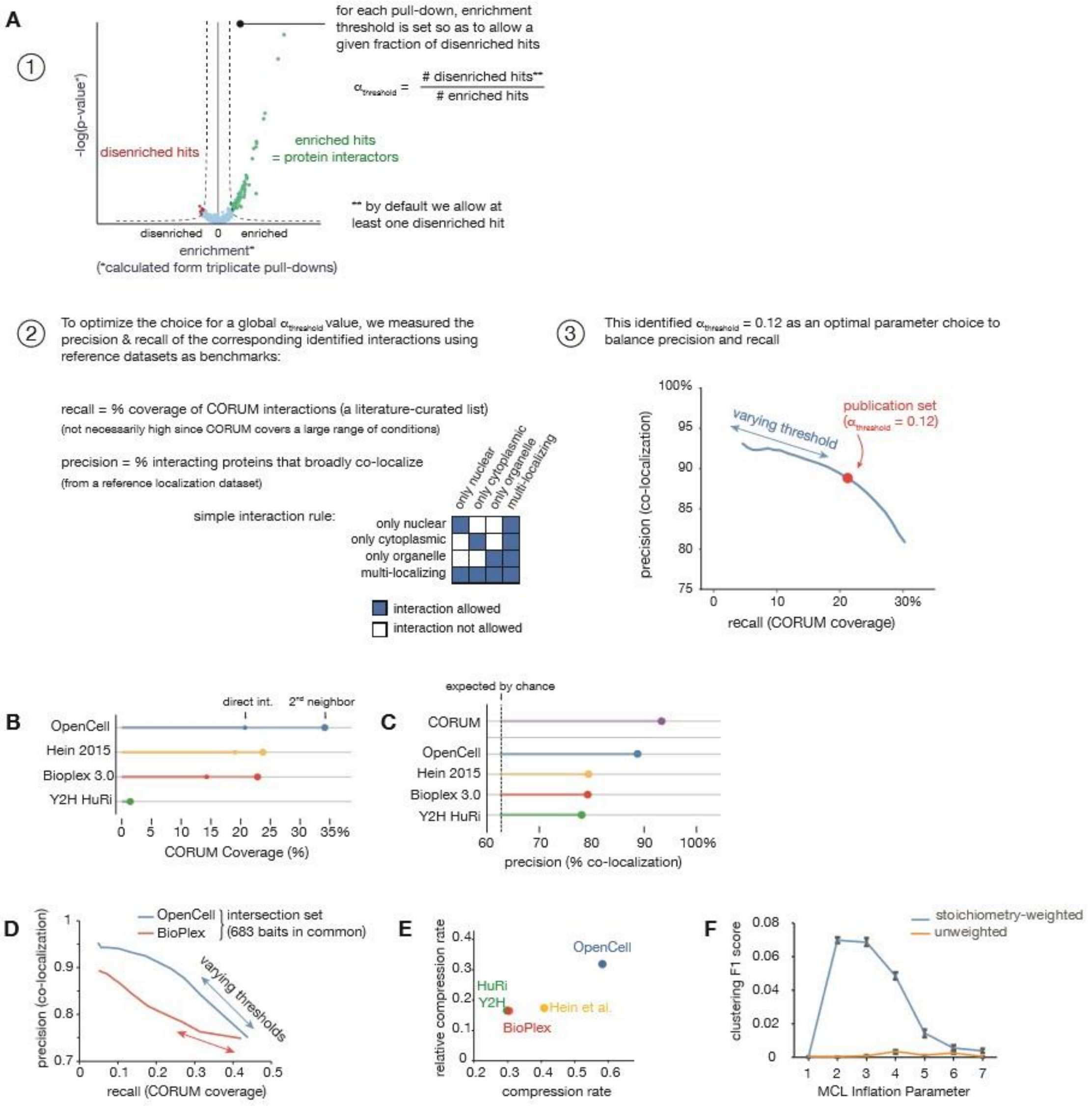


Figure S2: interactome analysis (related to Figure 2). (A) Strategy for defining enrichment threshold to define interactions. Our strategy builds upon methods described by Hein et al[6]. Here we use a quantitative approach to define enrichment thresholds dynamically for each replicate set, globally constrained by the parameter $\alpha_{threshold}$. To optimize parameter choice, we measured how precision (% co-localization) and recall (% CORUM coverage) of the corresponding interaction network varied with $\alpha_{threshold}$. This informed a final value of 0.12. **(B)** Comparing interaction recall (% CORUM coverage) of OpenCell vs. other large-scale interactomes, including direct or 2nd-neighbor interactions (i.e., sharing a direct interactor in common). **(C)** Comparing interaction precision (% co-localization) of OpenCell vs. other large-scale interactomes. CORUM interactions are shown as a reference. **(D)** Direct comparison of OpenCell vs. Bioplex 3.0. Both datasets use the same HEK293T cell line and share a large number (683) of baits in common. Precision and recall analysis by varying threshold for interaction detection is shown for the intersection set of 683 baits. **(E)** Compressibility analysis[28] of OpenCell vs. other large-scale interactomes. **(F)** MCL clustering performance (F1 score) using stoichiometry-weighted or unweighted interaction graphs, derived from CORUM interactions as described in Drew et al[64].

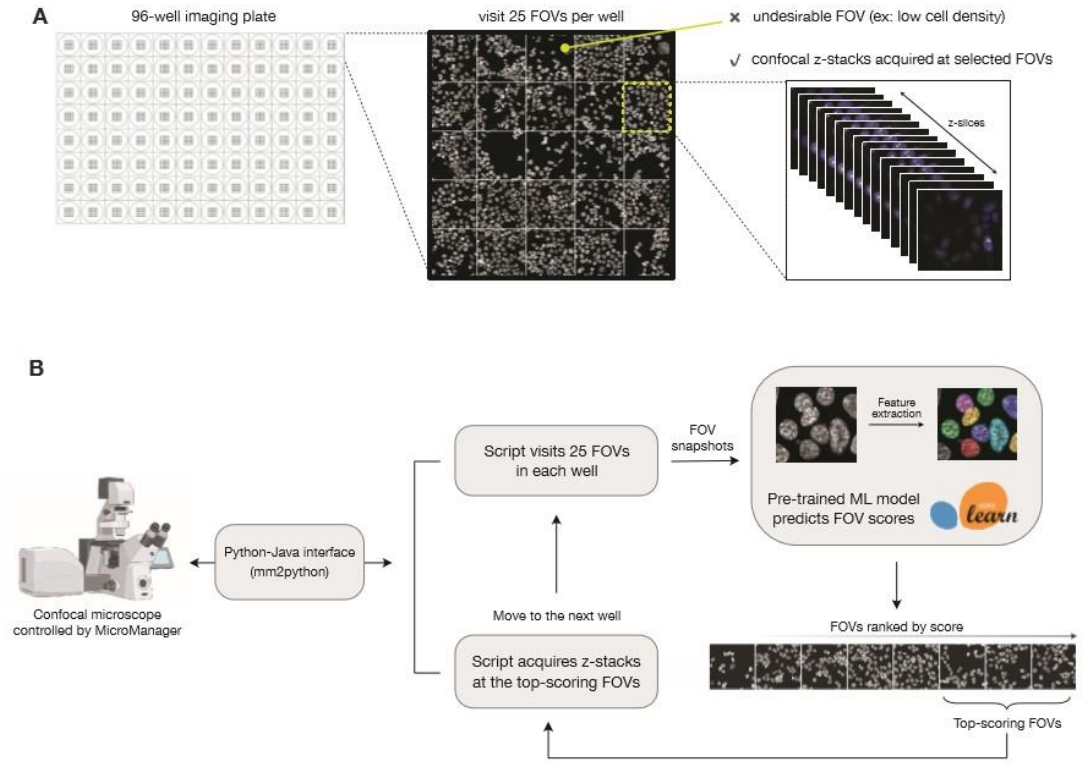


Figure S3: computer vision for automated microscopy acquisition (related to Figure 3). (A) To automate microscopy acquisition on 96-well plates and to limit experimental variability between imaging sessions (e.g., to limit variations in cell density) we paired an acquisition script, written in Python, with a pre-trained machine learning model to select fields of view (FOVs) on-the-fly during the acquisition. A total of 25 FOVs are sampled per well in a single z-plane, and desirable FOVs are selected for further 3D confocal acquisition on the basis of a score predicted by the pre-trained model. **(B)** Microscopy automation workflow. Microscope hardware is controlled by a Python-based acquisition script via an open-source MicroManager-Python bridge (mm2python; https://github.com/czbiohub/mm2python). This approach enables us to combine custom acquisition logic with the rich ecosystem of Python-based machine-learning packages. Here, we use the scikit-image package to extract features from each FOV snapshot, then use a pre-trained random-forest regression model (scikit-learn) to predict a quality score for the FOV. This process is not computationally expensive and requires less than a second; the FOV score can therefore be used immediately to determine whether the script should acquire a z-stack or else move on to the next position. To maximize the quality of our confocal z-stacks, however, we chose to visit and score all 25 FOVs in each well, then re-visit the top-scoring FOVs for confocal z-stack acquisition.

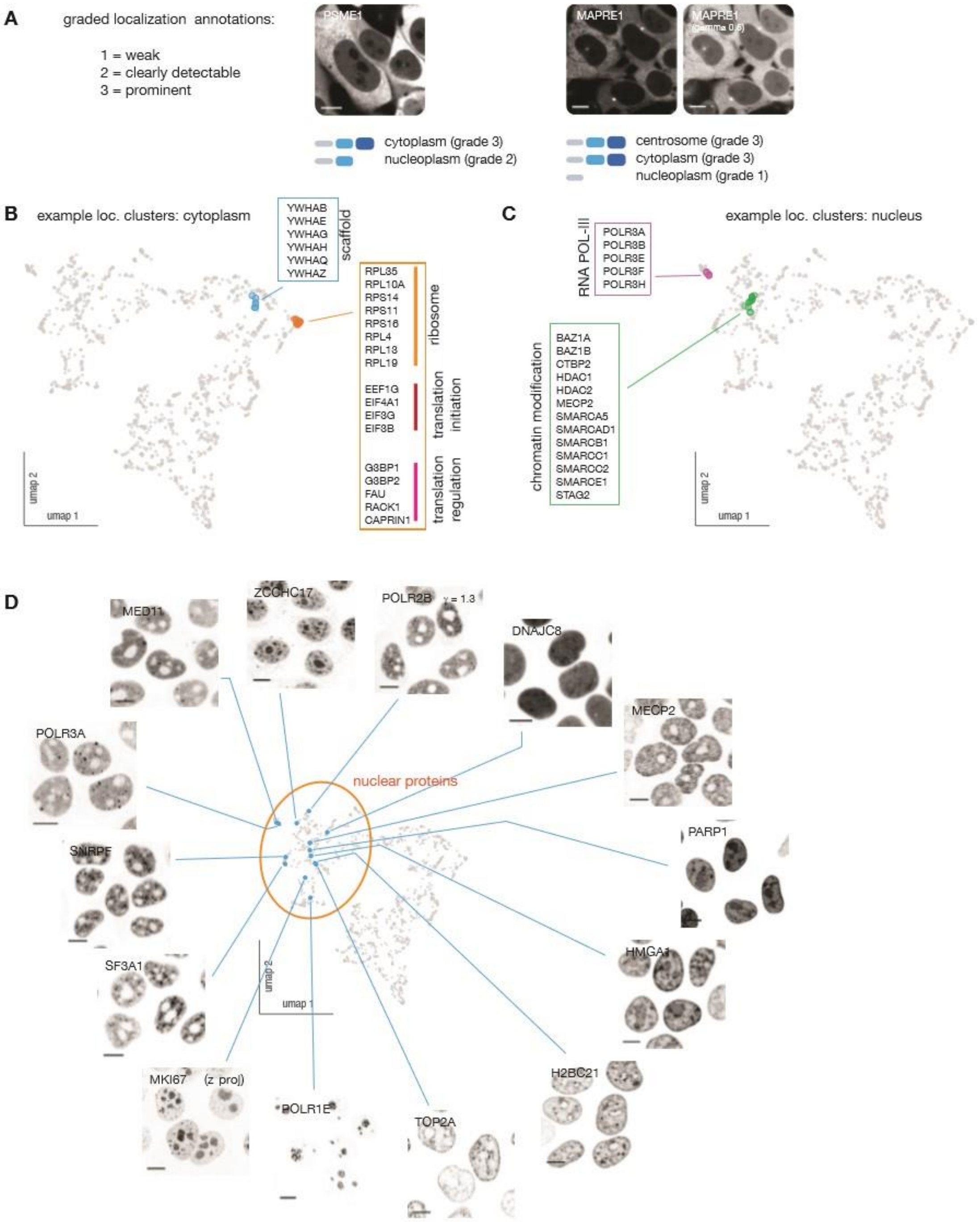


Figure S4: the OpenCell image dataset (related to Figure 3). (A) Principle of graded localization annotation (manual annotations). **(B)** Examples of cytoplasmic localization clusters. **(C)** Examples of nuclear localization clusters. **(D)** Representative images for 14 nuclear targets that exemplify the diversity of localization patterns across the proteome. Scale bar: 10 µm.

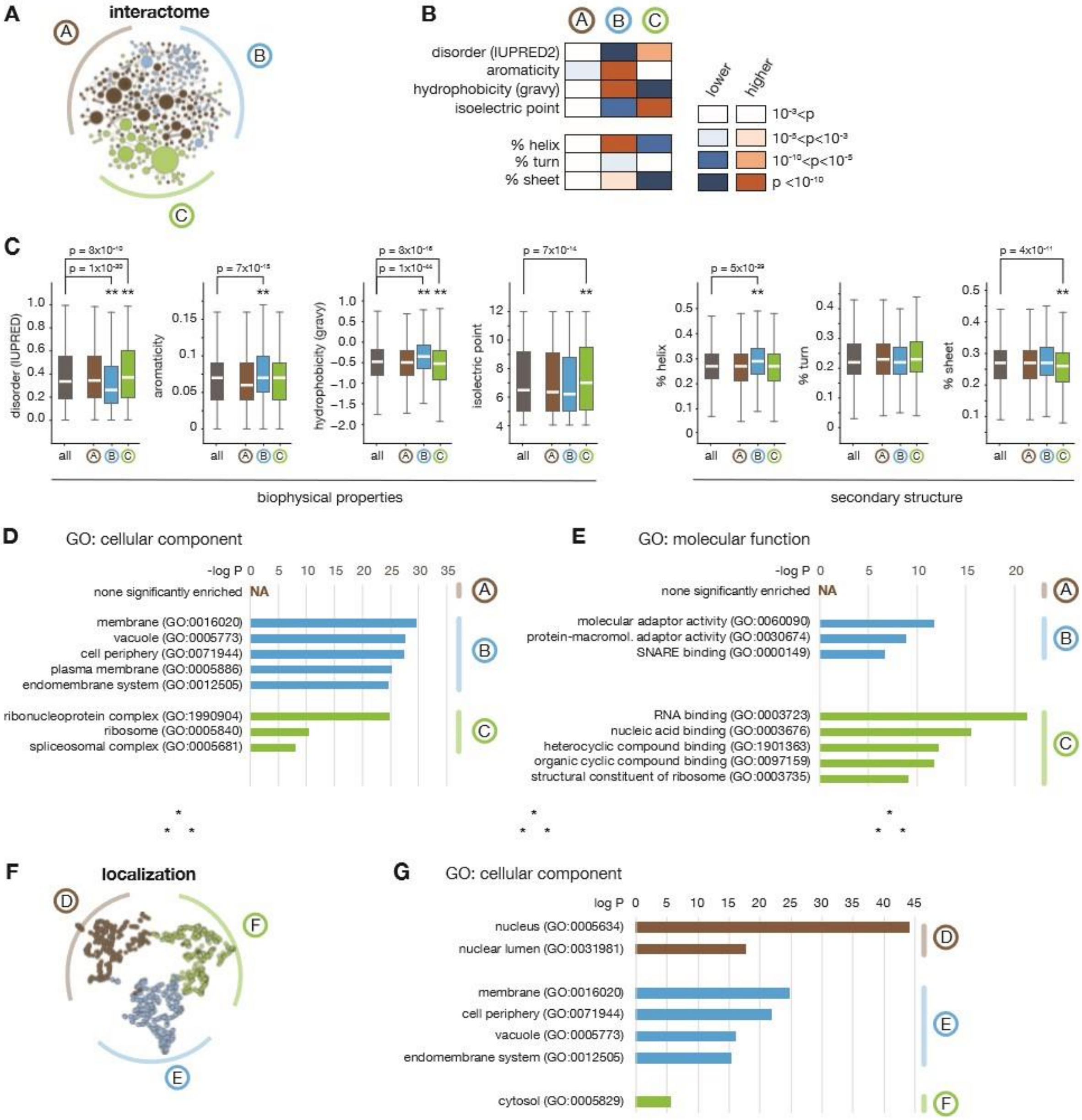


Figure S5: biophysical & ontology analysis of the main branches from interactome and localization hierarchies (related to Figure 7). (A) The three branches derived from the interactome hierarchy (see Figure 7A). **(B)** Heat-map representing significance testing of biophysical properties of protein sequences in the 3 branches. P-values were obtained using Student's t-test comparing proteins belonging to a specific hierarchical branch against all proteins in the three branches. **(C)** Box plot representing the significance testing of biophysical properties described in (B). Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent minimum and maximum values. Median is represented by a white line. ** p < 10-9 (Student's t-test), exact p-values are shown. **(D), (E)** Enrichment analysis of GO annotations in the hierarchical branches, testing GO term enrichment of proteins in each branch against all proteins in the interactome (Fisher's exact test, showing annotations enriched at p < 0.05 and excluding near-synonymous annotations). **(F), (G)**: same as (A) and (D) but for localization-based hierarchical branches (see Figure 7B).

REFERENCES

1. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

2. Hood, L. & Rowen, L. The Human Genome Project: big science transforms biology and medicine. *Genome Med* **5**, 79 (2013).

3. Nurse, P. & Hayles, J. The Cell in an Era of Systems Biology. *Cell* **144**, 850–854 (2011).

4. Mast, F. D., Ratushny, A. V. & Aitchison, J. D. Systems cell biologySystems cell biology. *The Journal of Cell Biology* **206**, 695–706 (2014).

5. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology* **20**, 285–302 (2019).

6. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).

7. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, (2017).

8. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 1–7 (2020).

9. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).

10. Mering, C. von *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).

11. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullin, C. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic acids research* **21**, 3329–3330 (1993).

12. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).

13. Collins, S. R. *et al.* Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP* **6**, 439–450 (2007).

14. Weill, U. *et al.* Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nature Methods* **15**, (2018).

15. Leonetti, M. D., Sekine, S., Kamiyama, D., Weissman, J. S. & Huang, B. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3501-8 (2016).

16. Roberts, B. *et al.* Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Molecular biology of the cell* mbc.E17-03-0209 (2017) doi:10.1091/mbc.e17-03-0209.

17. Feng, S. *et al.* Improved split fluorescent proteins for endogenous protein labeling. *Nature communications* **8**, 370 (2017).

18. Hubner, N. C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *The Journal of cell biology* **189**, 739–754 (2010).

19. Meier, F. *et al.* Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J Proteome Res* **14**, 5378–5387 (2015).

20. Lin, Y.-C. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature communications* **5**, 4767 (2014).

21. Lin, S., Staahl, B., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, (2014).

22. Doyon, J. B. *et al.* Rapid and efficient clathrin-mediated endocytosis revealed in genome-edited mammalian cells. *Nature cell biology* **13**, 331–337 (2011).

23. Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression. *Nat Methods* **10**, 715–721 (2013).

24. Thomas, J. A. & Tate, C. G. Quality Control in Eukaryotic Membrane Protein Overproduction. *J Mol Biol* **426**, 4139–4154 (2014).

25. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res* **47**, gky973- (2018).

26. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, 570 (2016).

27. Huttlin, E. L. *et al.* Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. *Biorxiv* 2020.01.19.905109 (2020) doi:10.1101/2020.01.19.905109.

28. Royer, L., Reimann, M., Stewart, A. F. & Schroeder, M. Network Compression as a Quality Measure for Protein Interaction Networks. *PLoS ONE* **7**, e35729 (2012).

29. Albert, R. Scale-free networks in cell biology. *Journal of Cell Science* **118**, 4947–4957 (2005).

30. Tanaka, R., Yi, T.-M. & Doyle, J. Some protein interaction data do not exhibit power law statistics. *Febs Lett* **579**, 5140–5144 (2005).

31. Haynes, C. *et al.* Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes. *Plos Comput Biol* **2**, e100 (2006).

32. Alberti, S. & Dormann, D. Liquid–Liquid Phase Separation in Disease. *Annu Rev Genet* **53**, 1–24 (2019).

33. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).

34. Xue, B. *et al.* Structural Disorder in Viral Proteins. *Chem Rev* **114**, 6880–6911 (2014).

35. Enright, A. J., Dongen, S. V. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575–1584 (2002).

36. Shurtleff, M. J. *et al.* The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *Elife* **7**, e37018 (2018).

37. Acosta-Alvear, D. *et al.* The unfolded protein response and endoplasmic reticulum protein targeting machineries converge on the stress sensor IRE1. *Elife* **7**, e43036 (2018).

38. McGilvray, P. T. *et al.* An ER translocon for multi-pass membrane protein biogenesis. *Elife* **9**, e56889 (2020).

39. Chitwood, P. J. & Hegde, R. S. An intramembrane chaperone complex facilitates membrane protein biogenesis. *Nature* **584**, 630–634 (2020).

40. Görlich, D. & Kutay, U. TRANSPORT BETWEEN THE CELL NUCLEUS AND THE CYTOPLASM. *Annu Rev Cell Dev Bi* **15**, 607–660 (1999).

41. Lusk, C. P. & King, M. C. The nucleus: keeping it together by keeping it apart. *Curr Opin Cell Biol* **44**, 44–50 (2017).

42. Meijering, E., Carpenter, A. E., Peng, H., Hamprecht, F. A. & Olivo-Marin, J.-C. Imagining the future of bioimage analysis. *Nat Biotechnol* **34**, 1250–1255 (2016).

43. Ouyang, W. *et al.* Analysis of the Human Protein Atlas Image Classification competition. *Nat Methods* **16**, 1254–1261 (2019).

44. Traag, V. A., Waltman, L. & Eck, N. J. van. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep-uk* **9**, 5233 (2019).

45. Stoeger, T., Gerlach, M., Morimoto, R. I. & Amaral, L. A. N. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS biology* **16**, e2006643 (2018).

46. Brooks, S. P. *et al.* The Nance–Horan syndrome protein encodes a functional WAVE homology domain (WHD) and is important for co-ordinating actin remodelling and maintaining cell morphology. *Hum Mol Genet* **19**, 2421–2432 (2010).

47. Law, A.-L. *et al.* Nance-Horan Syndrome-like 1 protein negatively regulates Scar/WAVE-Arp2/3 activity and inhibits lamellipodia stability and cell migration. *Biorxiv* 2020.05.11.083030 (2020) doi:10.1101/2020.05.11.083030.

48. Schossig, A. *et al.* Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. *Am J Hum Genetics* **90**, 701–707 (2012).

49. Merkulova, M. *et al.* Mapping the H+ (V)-ATPase interactome: identification of proteins involved in trafficking, folding, assembly and phosphorylation. *Scientific Reports* **5**, (2015).

50. Yan, Y., Denef, N. & Schüpbach, T. The Vacuolar Proton Pump, V-ATPase, Is Required for Notch Signaling and Endosomal Trafficking in Drosophila. *Dev Cell* **17**, 387–402 (2009).

51. Vasanthakumar, T. & Rubinstein, J. L. Structure and Roles of V-type ATPases. *Trends Biochem Sci* **45**, 295–307 (2020).

52. Bonald, T., Charpentier, B., Galland, A. & Hollocou, A. Hierarchical Graph Clustering using Node Pair Sampling. *Arxiv* (2018).

53. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Arxiv* (2018).

54. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).

55. Horlbeck, M. A. *et al.* Mapping the Genetic Landscape of Human Cells. *Cell* **174**, 953-967.e22 (2018).

56. Sanders, D. W. *et al.* Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. *Cell* **181**, 306-324.e28 (2020).

57. Yang, P. *et al.* G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell* **181**, 325-345.e28 (2020).

58. Markmiller, S. *et al.* Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590-604.e13 (2018).

59. Youn, J.-Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell* **69**, 517-532.e11 (2018).

60. Chujo, T. & Hirose, T. Nuclear Bodies Built on Architectural Long Noncoding RNAs: Unifying Principles of Their Construction and Function. *Mol Cells* (2017) doi:10.14348/molcells.2017.0263.

61. Go, C. D. *et al.* A proximity biotinylation map of a human cell. *Biorxiv* 796391 (2019) doi:10.1101/796391.

62. Riesenberg, S. *et al.* Simultaneous precise editing of multiple genes in human cells. *Nucleic acids research* **2**, 163 (2019).

63. Drew, K. *et al.* Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology* **13**, 932 (2017).

3.2. The collisional cross section universe in omics analysis

Following the successful introduction of the TIMS-qTOF platform (*timsTOF Pro, Bruker Daltonik GmbH*) to bottom-up proteomics and its application to high-throughput interactome measurements and high-sensitivity applications in pancreatic islet research, we aimed to transfer these achievements to the field of lipidomics.

The current state of the art in this omics field entails the identification of lipids based on the accurate mass measurement at the MS1 level and to integrate it with MS2 or even extend it to MS3 fragmentation information, when the first two levels are not sufficient. Furthermore, either direct infusion or high-flow chromatography systems are used to increase sample throughput with the negative effect of compromised analytical sensitivity resulting in a need for more initial sample.[413]

First, we reasoned that combining nanoflow chromatography with the PASEF principle on the timsTOF Pro should increase speed and sensitivity of the analysis. Indeed, we showed that this combination increased sensitivity more than 100-fold to the attomomole-range at a 16-fold increased sequencing speed reaching more than 100 Hz while keeping full resolution of co-eluting isomers. Next, we showcased the applicability of our setup to several biological matrices including blood plasma, liver tissue biopsies and cell lines, reaching system saturation with as little as 0.05 µl blood plasma, 10 µg liver tissue, or 2,000 HeLa cells. All major lipid classes like glycerophospholipids, mono-/di-/tri-acyl-glycerols, sterol lipids, ceramides, glycosphingolipids, and phosphosphingolipids were covered in our measurements.

Second, since the chemical structure of lipids is well known to determine their collisional cross sections, we reasoned that the TIMS device, positioned between chromatography and analytical quadrupole, could allow us to separate lipid isomers and yield an additional dimension for identification[414]. This was indeed the case. We investigated the correlation of lipid mass and ion mobility and showed that lipid isomers can be separated in routine measurements at a resolution of up to 200 CCS/ΔCCS, while keeping an up to 100 % ion beam utilization and up to 50-fold increased signal to noise ratio compared to not using TIMS. Furthermore, we showed that the intra- and inter-laboratory variability of CCS determination was well below 0.3 % with Pearson correlations of greater than 0.99. Comparisons to published CCS compendia acquired with drift tube systems, ion mobility devices that revert the principle of TIMS, also highlighted the high accuracy of our measurements ($r = 0.999$)[415]. We investigated the ion mobility-enhanced lipidomics 4D space and found that each lipid class occupies a specific area or volume in the conformational lipid landscape, reflecting structural differences in chemical composition. Since features of these lipid classes, only diverging by e.g. the number of carbon atoms in their chains, perfectly line up in the multidimensional space, we were able to infer the composition of unidentified lipids, which followed the positional rules in the 4D cuboid comprised of intensity, CCS, retention time, and m/z.

In summary, we developed a nanoflow lipidomics workflow that takes full advantage of TIMS and PASEF in terms of sensitivity, sequencing speed and ion mobility resolution. It is applicable to a wide range of biological samples and it appears to be attractive for sample limited applications such a biopsies and small cell populations. Furthermore, we assume that all the analytical advantages demonstrated for lipidomics should be transferable to metabolomics and that the high precision and accuracy of the 4D feature space including the CCS as additional dimension should also be suited well for machine learning approaches.

In the next study we explored the CCS peptide universe. We reasoned that the robustness of the timsTOF instrument itself and the high precision of CCS determination (CV <0.1 %), should allow us to infer all CCS values of peptides encountered in bottom-up proteomics – the distinct rotational average of a peptides gas-phase conformation.

To fully capture the conformational diversity of peptides in the gas phase, we acquired CCS values for more than 400,000 unique peptide sequences resulting in more than 2,000,000 peptide CCS values from five model organisms (*HeLa*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, and *Saccharomyces cerevisiae)*. We digested isolated proteins with three different proteases (LysC, LysN, trypsin) yielding complementary peptide termini. Furthermore, we separated those samples into 24 fractions to increase peptide analysis depth. This experimental design allowed us to ask and answer several fundamental questions.

First, we investigated the precision of the CCS determination at scale and compared it to publically available data sets from drift tubes devices. We showed that our initial observation of highly reproducible CCS value determination in repetitive measurements holds true at scale across several instruments (CVs < 1%). We found that systematic TIMS tunnel pressure-dependent CCS shifts, can be corrected by linear alignment. This highlights that ion mobility values are largely independent of experimental circumstances and can be of high value for peptide identification, similar to the molecular mass.

Second, it allowed us to address the long-standing question of amino acid sequence and positional contribution to the CCS value of peptides, and if amino acids with high occurrence frequencies for particular secondary structures also determine structures in the gas phase. Indeed, we found CCS values of peptides with a high proportion of amino acids known to be in alpha helices (Q, E, A, L, M, H) to be larger, while peptides with a high proportion of amino acids (D, N, G, S, P) known to be in beta sheets to be smaller. These trends are well reflected by the overall hydrophobicity of the peptide itself and summarized as the GRAVY score, a measure for the compositional hydrophobicity of the peptide[416]. Peptides with a high GRAVY score (high hydrophobicity) tend to be larger than peptides with a low GRAVY score (low hydrophobicity). Furthermore, we found that the more prolines a peptide sequence contains, the smaller the CCS values, most likely due do its inability of to donate hydrogen bonds for 2D structure stabilization and increased flexibility. We also showed that the position of histidines within a peptide drastically influences the CCS value, most likely due its peptide net charge contribution and its position along the peptide sequence. The closer the histidine is located to the terminal positively charged lysine, the smaller the peptides cross section is and vice versa. Interestingly, these experimental observations were also recapitulated by the Shapley Additive Explanation (SHAP) analysis of CCS values predicted by our deep recurrent neural network described below[417].

Third, we were able to increase the number of peptide-specific CCS values to a level that allowed us to train a deep recurrent neural network consisting of a three-layered bi-directional long short-term memory (LSTM) network followed by a two-layer multilayer perceptron (MLP) for CCS value regression. This resulted in a generalized prediction model of CCS values solely based on the peptide sequence. Comparing network derived CCS values with experimental values of the synthetic ProteomeTools[418] peptide resource demonstrated a prediction accuracy with a 1.4 % median relative error. Replacing experimentally derived CCS values within a spectral library used for diaPASEF analysis by our predicted ones did not compromise identification, suggesting that predicted CCS aware spectral libraries will be able to replace the tedious process of creating those experimentally in the future.

In summary, CCS values can now be predicted for any tryptic peptide and organism. We illustrated this by the prediction of CCS values for the human peptide universe, comprising 616,948 unique tryptic peptides to form a basis for advanced proteomics workflows that make full use of the additional information. Furthermore, this project is only a starting point for the prediction of a peptide ions shape in the gas phase and can be extended to other peptide classes like modified and cross-linked peptides, or even to the peptidome - the set of *in vivo* processed peptides. Most importantly, we set the stage for the prediction of fully predicted ion mobility aware spectral libraries for data independent acquisition analysis, which could improve the specificity in database searches, especially in challenging analyses with a very large search spaces and will benefit targeting approaches.

3.2.1. Article 4: TIMS and PASEF enable high-sensitivity lipidomics

Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts

Nature communications, January 16, 2020

Catherine G. Vasilopoulou[1], Karolina Sulek[2], **Andreas-David Brunner[1]**, Ningombam Sanjib Meitei[3], Ulrike Schweiger-Hufnagel[4], Sven W. Meyer[4], Aiko Barsch[4], Matthias Mann[1, 2, #], Florian Meier[1, #]

# Correspondence

[1]Max Planck Institute of Biochemistry, Martinsried, Germany
[2]NNF Center for Protein Research, Copenhagen, Denmark
[3]PREMIER Biosoft, Indore, India
[4]Bruker Daltonik GmbH, Bremen, Germany

Contribution

In this project, I contributed to the experimental design of the paper and performed experiments.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-14044-x OPEN

Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts

Catherine G. Vasilopoulou[1], Karolina Sulek[2], Andreas-David Brunner[1], Ningombam Sanjib Meitei[3], Ulrike Schweiger-Hufnagel[4], Sven W. Meyer[4], Aiko Barsch[4], Matthias Mann[1,2*] & Florian Meier[1*]

A comprehensive characterization of the lipidome from limited starting material remains very challenging. Here we report a high-sensitivity lipidomics workflow based on nanoflow liquid chromatography and trapped ion mobility spectrometry (TIMS). Taking advantage of parallel accumulation–serial fragmentation (PASEF), we fragment on average 15 precursors in each of 100 ms TIMS scans, while maintaining the full mobility resolution of co-eluting isomers. The acquisition speed of over 100 Hz allows us to obtain MS/MS spectra of the vast majority of isotope patterns. Analyzing 1 μL of human plasma, PASEF increases the number of identified lipids more than three times over standard TIMS-MS/MS, achieving attomole sensitivity. Building on high intra- and inter-laboratory precision and accuracy of TIMS collisional cross sections (CCS), we compile 1856 lipid CCS values from plasma, liver and cancer cells. Our study establishes PASEF in lipid analysis and paves the way for sensitive, ion mobility-enhanced lipidomics in four dimensions.

[1] Max Planck Institute of Biochemistry, Martinsried, Germany. [2] NNF Center for Protein Research, Copenhagen, Denmark. [3] PREMIER Biosoft, Indore, India. [4] Bruker Daltonik GmbH, Bremen, Germany. *email: mmann@biochem.mpg.de; fmeier@biochem.mpg.de

Disentangling the lipid composition of biological model systems and clinical samples in a robust and high throughput manner promises novel insight into basic biology, as well as the onset and progression of disease[1–4]. Lipid extracts from biological sources can be analyzed either directly via high-resolution mass spectrometry[5,6] or via online liquid chromatography coupled to mass spectrometry (LC-MS)[7]. Lipid identifications base on accurate mass and the MS^2 or MS^3 fragmentation pattern, which is increasingly facilitated by recent software developments and ever growing reference databases[8–11]. Established LC-MS lipidomics workflows separate lipids at flow rates in the higher micro- or milliliter per minute range, which ensures high sample throughput and robustness, but also compromises sensitivity. As the available sample amount becomes a limiting factor, for example with small tissue biopsies from biobanks or small cell sub-populations, it is increasingly attractive to employ nanoflow chromatography[12–14].

MS technology has greatly improved and state-of-the-art high-resolution Orbitrap or time-of-flight (TOF) instruments transmit ions very efficiently and achieve low- to sub-ppm mass accuracy[15,16]. The high acquisition speed of TOF analyzers makes them compatible with very fast separation techniques such as ion mobility spectrometry (IMS)[17,18]. Nested in-between LC and MS, IMS provides an additional dimension of separation based on the ions' shape and size (collisional cross section, CCS). This is particularly interesting for lipidomics, as it provides an opportunity to separate otherwise unresolved isomers[19–22]. Furthermore, the chemical structure of lipids is closely linked to the CCS, which allows predictions by machine learning and could facilitate lipid identification[23–26].

Trapped ion mobility spectrometry (TIMS) is a relatively new form of IMS that inverts the separation principle of classical drift tube ion mobility[27–31]. Ions entering the TIMS analyzer are positioned in an electrical field by the drag of a gas flow. Lowering the electrical force releases ions from the TIMS device separated by their ion mobility, while the IMS resolution is proportional to the user-defined ramp time. It can be tuned to over 200 CCS/ΔCCS, for example to separate isomeric lipids with distinct double bond positions or geometries[32]. We recently introduced a MS scan mode termed parallel accumulation serial fragmentation (PASEF) that synchronizes TIMS with MS/MS precursor selection[33]. In proteomics, PASEF increases MS/MS scan rates more than tenfold, importantly, without the loss of sensitivity that is otherwise inherent to faster acquisition rates[34].

Here, we explore whether the PASEF principle can be transferred to lipidomics. We build on nanoflow chromatography to establish a rapid PASEF lipidomics workflow capable of comprehensively analyzing low sample amounts. To investigate the potential of the additional TIMS dimension, we set out to compile a high-precision lipid CCS library from body fluids, tissue samples, and human cell lines.

Results

Development of the nanoflow PASEF lipidomics workflow. We aimed to develop a rapid workflow that enables global lipid analysis in a straightforward manner (Fig. 1). We adapted an MTBE lipid extraction protocol[35] that is applicable to common biological sample types, such as body fluids, tissue, as well as cell lines (Fig. 1a) and requires only a few manual liquid handling steps that could easily be automated in the future. We found that our extraction protocol scales well from small sample volumes (1 µL blood plasma) to relatively large cell counts (5e5 HeLa cells) and can be performed in <1 h.

We loaded the lipid extracts directly onto a C_{18} column and eluted them within 30 min, for a total of little more than 1 h analysis time per sample when using both positive and negative ionization modes (Fig. 1b). Retention times were reproducible with median CVs of 0.54% in replicate injections prior to alignment (Supplementary Data 1) and chromatographic peak widths were in the range of 3–6 s full width at half maximum (FWHM), at least two orders of magnitude slower than TIMS ion mobility analysis (100 ms) and the acquisition speed of high-resolution TOF mass spectra (~100 µs). The timsTOF Pro mass spectrometer (Bruker Daltonics) features a dual TIMS analyzer that allows to utilize up to 100% of the incoming ion beam[29] (Methods). In this mode, ions are accumulated in the first TIMS analyzer while another batch of ions is separated by ion mobility in the second TIMS analyzer. TIMS closely resembles classical drift tube ion mobility, however, ions arrive at the mass analyzer in the inverse order, which means low-mobility (and high-mass) ions are released first, followed by ions with higher mobility (and lower mass). In our experiments with 100 ms TIMS scan time, the ion current accumulated during 100 ms was concentrated into ion mobility peaks of 2–3 ms FWHM, which should lead to a 50-fold increase in signal-to-noise as compared with continuous acquisition. These peak widths equate an ion mobility resolution of 40–50 CCS/ΔCCS. The ion mobility-resolved mass spectra can be illustrated in two-dimensional heat maps, from which suitable precursor ions are selected for fragmentation in data-dependent MS/MS mode (Fig. 1b). With PASEF, multiple precursors are fragmented in each TIMS ramp by rapidly switching the quadrupole (see below). As the collision cell is positioned after the TIMS analyzer in the ion path, fragment ions occur at the same ion mobility position as their precursor ions in MS1 mode.

We make use of this information in the post-processing (Fig. 1c) to connect MS/MS spectra to their corresponding MS features extracted from the four-dimensional data space (retention time, *m/z*, ion mobility, intensity). Finally, we rely on established lipidomics software to automatically assign lipids to the spectra based on diagnostic fragment ions and database matching (Methods). Our workflow automatically converts ion mobility to CCS values and records them for all detected MS1 features and thus all identified lipids.

Evaluating PASEF in lipidomics. The central element of our workflow is the PASEF acquisition method. PASEF takes advantage of the temporal separation of ions eluting from the TIMS device to select multiple precursors for MS/MS acquisition in a single TIMS scan[33]. To illustrate, Fig. 2a, b shows representative MS1 heat maps of co-eluting lipids from an LC-TIMS-MS analysis of human plasma. The ions are widely spread in *m/z* vs. ion mobility space, while higher mass roughly correlates with lower ion mobility. In standard MS/MS mode, the quadrupole isolates a single precursor mass throughout the entire TIMS scan time (red dots in Fig. 2a). However, the targeted precursor completely elutes during about 6 out of 100 ms, and thus over 90% of the acquisition time is effectively not used. In PASEF mode, the quadrupole instead switches its mass position within ~1 ms to capture as many precursors as possible (red dots in Fig. 2b). In this example, 16 precursors were selected during a single PASEF scan, which translates into a 16-fold increased MS/MS acquisition rate of over 100 Hz. Importantly, this does not come at a loss in sensitivity because the full precursor ion signal of the 100 ms accumulation time is captured.

In a 30 min analysis of plasma, we found that on average 15 precursors were fragmented per PASEF scan (Fig. 2c), confirming that the PASEF principle is transferable to lipidomics. In total, we acquired 187,177 MS/MS spectra—15-fold more than without PASEF. This fragmentation capacity greatly exceeds the number of expected lipids and, in principle, allows to acquire MS/MS

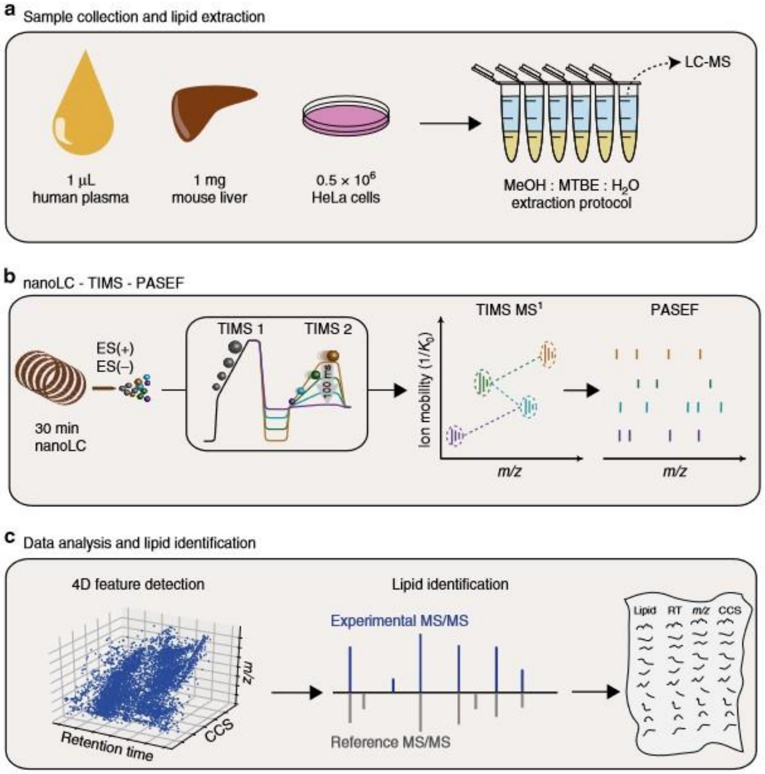


Fig. 1 Nanoflow lipidomics with trapped ion mobility spectrometry. **a** Lipids from various biological sources, such as body fluids, tissues and cells, are analyzed using a single MeOH:MTBE extraction. **b** The crude extract is injected into a nanoflow liquid chromatography (LC) system coupled online to a high-resolution TIMS quadrupole time-of-flight mass spectrometer (timsTOF Pro). In the dual TIMS analyzer, ions are accumulated in the front part (TIMS 1), while another batch is released as a function of ion mobility from the TIMS 2 analyzer. PASEF synchronizes precursor selection and ion mobility separation, which allows fragmenting multiple precursors in a single TIMS scan at full sensitivity. **c** Features are extracted from the four-dimensional (retention time, *m/z*, ion mobility, intensity) data space and assigned to PASEF MS/MS spectra for automated lipid identification and compilation of comprehensive lipid CCS libraries. MeOH = methanol, MTBE = methyl-tert-butyl ether, CCS = collisional cross section.

spectra for every suitable isotope pattern detected in a single lipidomics LC run. Here, we chose to fragment low-abundance precursors repeatedly to increase their signal-to-noise ratios in a summed spectrum. On average, precursors were fragmented two times as indicated by the acquisition engine.

We evaluated the performance of our PASEF method with lipid extracts from human plasma, mouse liver, and HeLa cells (Fig. 2d and Supplementary Fig. 1). In all three sample types, the 4-dimensional feature detection yielded 8900–13,400 MS1 features above the intensity threshold and after collapsing multiple adducts. In standard TIMS-MS/MS mode, on average 5.5% of these were fragmented. This fraction increased up to 11.5-fold with PASEF and, in both negative and positive ionization mode, about 65% of all features had corresponding MS/MS spectra. Overall, PASEF increased the number of identified lipids across all runs on average 3.6-fold (Supplementary Fig. 2). To further investigate whether PASEF is fast enough to acquire MS/MS spectra of close to all informative lipid features in a short time, we extended the LC gradients to 60 and 90 min (Supplementary Fig. 3). Indeed, 97% of all lipids identified with the three times longer gradient were already identified in the 30 min PASEF run, which confirms our hypothesis and suggests that even shorter runtimes could be explored.

Comprehensive and accurate lipid quantification. Having ascertained that PASEF achieves a very high MS/MS coverage of lipidomics samples, we investigated our automated data analysis pipeline in more detail (Fig. 3a). Starting from the thousands of 4D features detected in all replicate injections of human plasma, mouse liver and HeLa cells, we kept those with assigned MS/MS scans for further analysis. PASEF spectra are resolved by ion mobility and the software extracts the MS/MS spectra at the ion mobility position of the respective precursor ion. We then searched all MS/MS spectra considering four lipid categories with the respective lipid classes and subclasses. This yielded 653–1595 annotations for each sample and ionization mode. We manually inspected all automatically annotated MS/MS spectra to filter potential false positives based on the observed fragmentation pattern (Methods). Finally, we grouped adducts, isomers, and co-eluting peaks that were separated by their ion mobility but could not be distinguished based on their MS/MS spectra. However, we

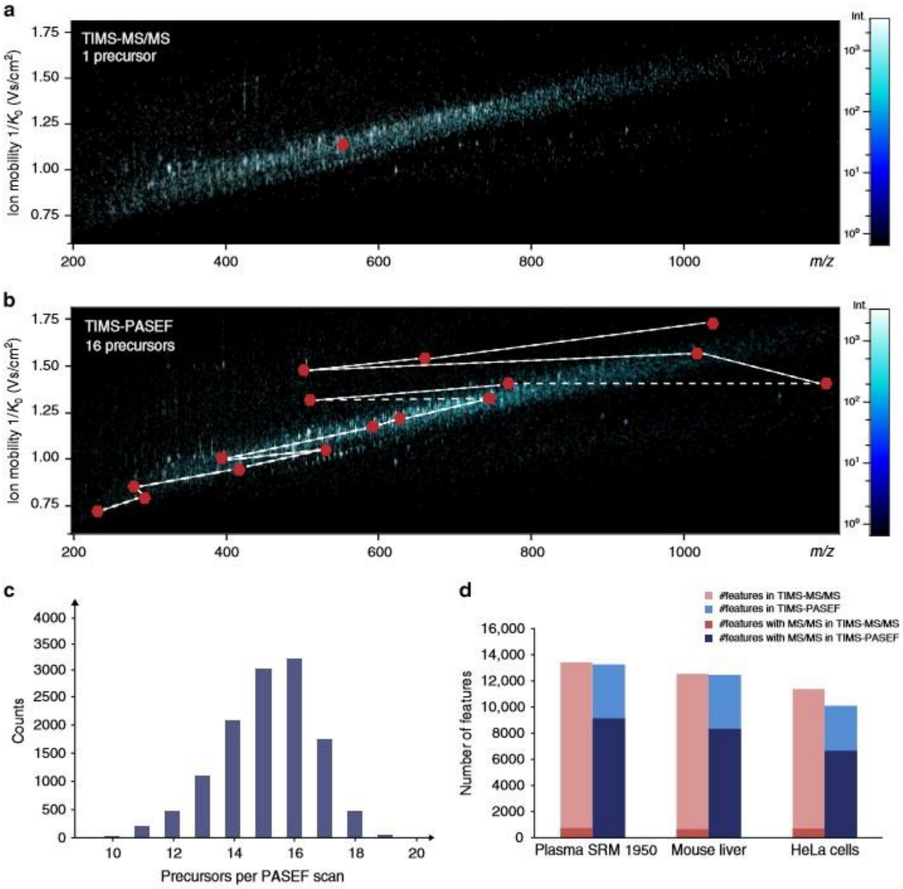


Fig. 2 Evaluating PASEF in lipidomics. Heat-map visualization of a representative trapped ion mobility resolved mass spectrum of human plasma at an elution time of 9.2 min. Red dots indicate precursors selected for MS/MS fragmentation in the subsequent 100 ms PASEF scan in **a**, standard TIMS-MS/MS mode and **b** TIMS-PASEF mode. The dashed line indicates the positioning of the quadrupole. **c** Distribution of the number of precursors per PASEF scan in an LC-MS analysis of human plasma lipid extract ($n = 1$). **d** Total number of 4D features extracted from 30 min runs of human plasma ($n = 5$), mouse liver ($n = 5$), and human cancer cells ($n = 5$) in positive ion mode without (TIMS-MS/MS, red) and with PASEF (TIMS-PASEF, blue). The fraction of features with assigned MS/MS spectra is indicated by a darker color.

kept separate lipids with same MS/MS-based annotation but eluting in close proximity as these are potential isomers. Removing duplicates resulted in 460–879 identified unique lipids per experiment. Combining both ionization modes, we identified 1108 unique lipids from the equivalent of 0.05 µL plasma per injection, 976 unique lipids from 10 µg liver tissue and 1351 unique lipids from ~2000 HeLa cells, with a median absolute precursor mass accuracy of 1.06 ppm (Supplementary Data 2–4). The identified lipid species covered all major lipid classes such as glycerophospholipids (PC, PE, PA, PS, PI, PG), oxidized glycerophospholipids, monoacyl-, diacyl- and triacyl-glycerols, sterol lipids, ceramides, glycosphingolipids, and phosphosphingolipids.

This comprehensive lipid coverage from relatively small sample amounts motivated us to investigate our sensitivity limit in more detail. Starting from the concentration above, we diluted the lipids extracted from human SRM 1950 plasma over three orders of magnitude in seven steps. With a 10-fold dilution, we were still able to identify 526 lipids in positive mode and this number dropped below 400 only at >100-fold dilution (Supplementary Fig. 4). We reasoned that this sensitivity is partially due to our nanoflow chromatography setup as opposed to conventional high-flow systems. In fact, a direct comparison indicated a 100-fold lower sensitivity limit with nanoflow in both ionization modes. Injecting the same amounts of plasma lipids on column, we identified three to six times more lipid species with the nanoflow setup (Supplementary Fig. 4).

In biological or clinical studies, quantitative accuracy is at least as important as cataloging the lipid composition and may be compromised if lipids are sparsely detected across samples. We hypothesized that the speed of PASEF and the improved signal-to-noise of TIMS should lead to very reproducible quantification. Indeed, we observed that 816 out of the total 976 identified lipids in liver tissue were quantified in five out of five replicates (Fig. 3b), resulting in a data completeness of 95.4%. The median

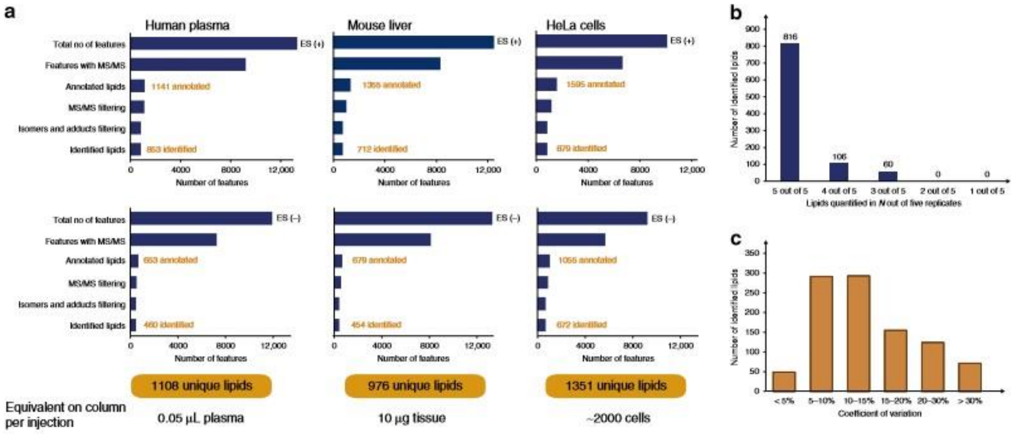


Fig. 3 Lipid identification and quantification. **a** Sequential data analysis steps from the total number of detected 4D features to unique lipids for human plasma, mouse liver, and human cancer cells in both ionization modes. **b** Number of lipids quantified in *N* out of five replicate injections of liver tissue extract. **c** Coefficients of variation for 976 lipids quantified in at least three out of five replicate injections of liver tissue extract.

coefficient of variation (CV) was 12.3%, and 80% of all quantified lipids had a CV below 20% (Fig. 3c and Supplementary Data 5).

From these results, we conclude that our nanoflow PASEF lipidomics workflow covers the lipidome comprehensively with high quantitative accuracy, while requiring only minimal sample amounts.

Coverage of the human plasma lipidome. To assess the lipid coverage of our PASEF workflow in more detail, we analyzed the human plasma Standard Reference Material (SRM 1950), a pool from 100 individuals from the United States in the age range of 40–50 years, provided by the National Institute of Diabetes and Digestive and Kidney Diseases and the National Institute of Standards (NIST)[36]. This sample has served as a reference for many lipidomics studies, establishing a range of detectable lipid species and their absolute concentrations[37]. The LIPID MAPS consortium recently compiled consensus results from 31 laboratories, each of which followed their in-house analysis workflow (Bowden et al.[38]). In an effort to disentangle the human plasma lipidome, Quehenberger et al.[39] employed class-specific analysis strategies to quantify about 500 lipids from >1 mL NIST SRM 1950 plasma.

Taking these two studies as a reference, we first compared the number of identified lipids in each lipid category based on the short name annotation (Fig. 4a and Supplementary Data 6, 7). Starting from 1 µL plasma and with a single extraction protocol, our PASEF workflow detected many more glycerolipids and glycerophospholipids, exceeding both studies three- to four-fold. At the same time, 87 and 83% of all glycerolipids and glycerophospholipids reported in the Bowden et al. study[38] were also present in our dataset. Similarly, we retrieved 65 and 49% of all lipids from these two abundant plasma lipid categories reported by Quehenberger et al.[39]. We observed a two-fold gain for sphingomyelins, again with a high overlap of 77 and 60% with both reference studies. Analysis of ceramides typically requires specific extraction methods and this category was therefore underrepresented in ours as well as in the Bowden et al. study[38] relative to the class-specific analysis by Quehenberger et al.[39]. From another analytically challenging class of lipids, sterol lipids, we still detected 33 species in the human plasma reference sample.

To further investigate the sensitivity of our method, we mapped all identified lipids onto absolute plasma concentrations reported in ref. [38] (Fig. 4b and Supplementary Data 7). We quantified about 80% of the lipids covering the full abundance range from about 0.01 up to 1000 nmol/mL. For example, we achieved full coverage of the triacylglycerols and also quantified less abundant lipids such as phosphatidylethanolamines comprehensively. Even though coverage was sparser in the lowest abundance range, we quantified the least abundant lysophosphatidylcholine (LPC 22:1) with a reference concentration of 0.013 nmol/mL. Since we injected only 1/20th of the lipids extracted from 1 µL plasma in each replicate, this translates into a sensitivity in the attomol range for the entire workflow.

Accuracy and precision of online lipid ^{TIMS}CCS measurements. In addition to generating MS/MS spectra for almost all detectable precursors with PASEF, TIMS measures the ion mobility of all identified and unidentified lipids. We calibrated all TIMS measurements to reduced ion mobility values using well-characterized phosphazine derivatives and converted them to collisional cross sections (^{TIMS}CCS) using the Mason–Schamp equation[29,40]. Because they have the same underlying physics, ^{TIMS}CCS can be directly related to drift tube experiments[30,31,41] and should result in high quality and highly reproducible collisional cross section data.

First, we investigated a mixture of commercially available lipid standards (Differential Ion Mobility System Suitability Lipidomix Kit, Avanti) on four timsTOF Pro instruments in two independent laboratories (in Bremen and Munich, Germany) (Fig. 5a). The measured ^{TIMS}CCS values for all 22 lipid ions (12 distinct lipids) clustered closely around their median values with a median CV of 0.35%. The median intra-instrument variability in five replicate injections ranged from 0.10 to 0.17% and the median intra-laboratory CV was between 0.18 and 0.21% in both laboratories. The ^{TIMS}CCS values were also highly reproducible between laboratories with an average inter-laboratory CV of only 0.35% and did not reveal any lipid-class specific biases (Supplementary Data 8).

To test whether the high quality of ^{TIMS}CCS values manifests in complex biological samples, we first investigated all detectable features in all three sample types regardless of their identification

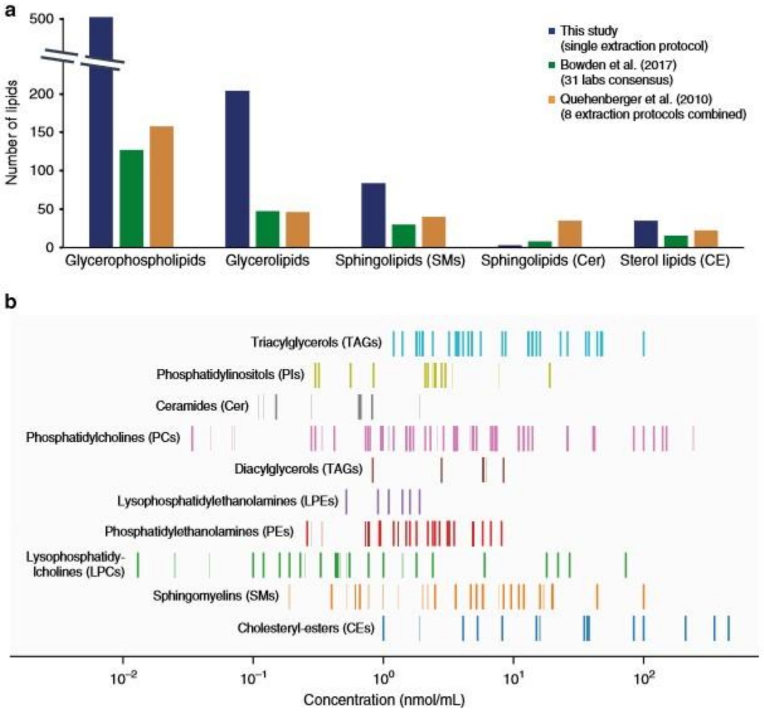


Fig. 4 Analysis of 1 µL NIST SRM 1950 human plasma. a Number of identified lipids from major lipid classes in this study and two reference studies from the same standard material[38,39]. **b** Mapping of lipids identified with our PASEF lipidomics workflow to absolute plasma concentrations reported in [38]. Vertical lines indicate the abundance range of reported lipids from different lipid classes and dark color indicates commonly identified lipids in both studies.

as a lipid. Plotting the $^{\mathrm{TIMS}}$CCS values across repeated injections on one instrument revealed excellent reproducibility (Pearson $r >$ 0.99) (Supplementary Fig. 5 and Supplementary Data 9), which motivated us to measure lipid extracts from the NIST SRM 1950 plasma on all four instruments. Considering only $^{\mathrm{TIMS}}$CCS of lipids identified in all experiments, we found median CVs < 0.1 to 0.45% in repeated injections on the same instrument (Fig. 5b and Supplementary Data 10) and similar intra-laboratory CVs of 0.15–0.45%. Overall, CCS measurements from both laboratories agreed within 0.38% on average and were highly correlated with a Pearson correlation coefficient $r > 0.99$ for all pair-wise comparisons (Fig. 5c and Supplementary Data 11).

Having ascertained highly reproducible $^{\mathrm{TIMS}}$CCS measurements in complex samples, we next investigated the accuracy of our results by comparing it with different methods and instrumentation. Our dataset shared 149 and 28 lipid identifications (based on the short name annotation) with recent reports from the Zhu[24] and McLean[23] laboratories, which both employed drift tube ion mobility analyzers to establish high-precision reference data. Our comparison revealed a very high correlation (Pearson $r = 0.999$) and 98% of all values were within ±1% deviation centered at zero (Fig. 5d and Supplementary Data 12). The median absolute deviation was 0.18%, which is very well in the range of recently reported inter-laboratory variability for standard compounds measured with a commercial drift tube analyzer[42].

The high reproducibility of lipid ion mobility measurements makes them also very attractive for machine learning approaches. The Zhu laboratory has developed a support vector regression model that predicts lipid CCS values from SMILES structures[24] and which is implemented in the Bruker MetaboScape software in a modified version (Methods). Even though the model was trained on independent data from a different instrument type, predicted and experimental $^{\mathrm{TIMS}}$CCS values correlated very well (Pearson $r = 0.996$) and the relative deviations were normal distributed with 95% of the values within ±2% deviation for lipid classes that were contained in the initial training data (Fig. 5e and Supplementary Data 13). The median absolute deviation was 0.44% and based on the experimental precision demonstrated above, we expect that machine learning models trained directly on TIMS data yield even more accurate predictions.

The TIMS lipidomics landscape. Data generated by our TIMS lipidomics workflow span a three-dimensional data space in which each feature is defined by retention time, *m/z* and CCS, with intensity as a 4th data dimension. To explore this data space, we compiled all measurements from human plasma, mouse liver and HeLa cells acquired in both ionization modes. The total dataset comprises CCS values of 1856 unique lipids (positive mode), representing the four major lipid categories and 15 lipid classes (Supplementary Data 14). To make our dataset fully accessible, we provide Supplementary Data 14 in a format that follows the standard lipid nomenclature guidelines by the LIPID MAPS consortium[43] and the Lipidomics Standards Initiative (https://lipidomics-standards-initiative.org/).

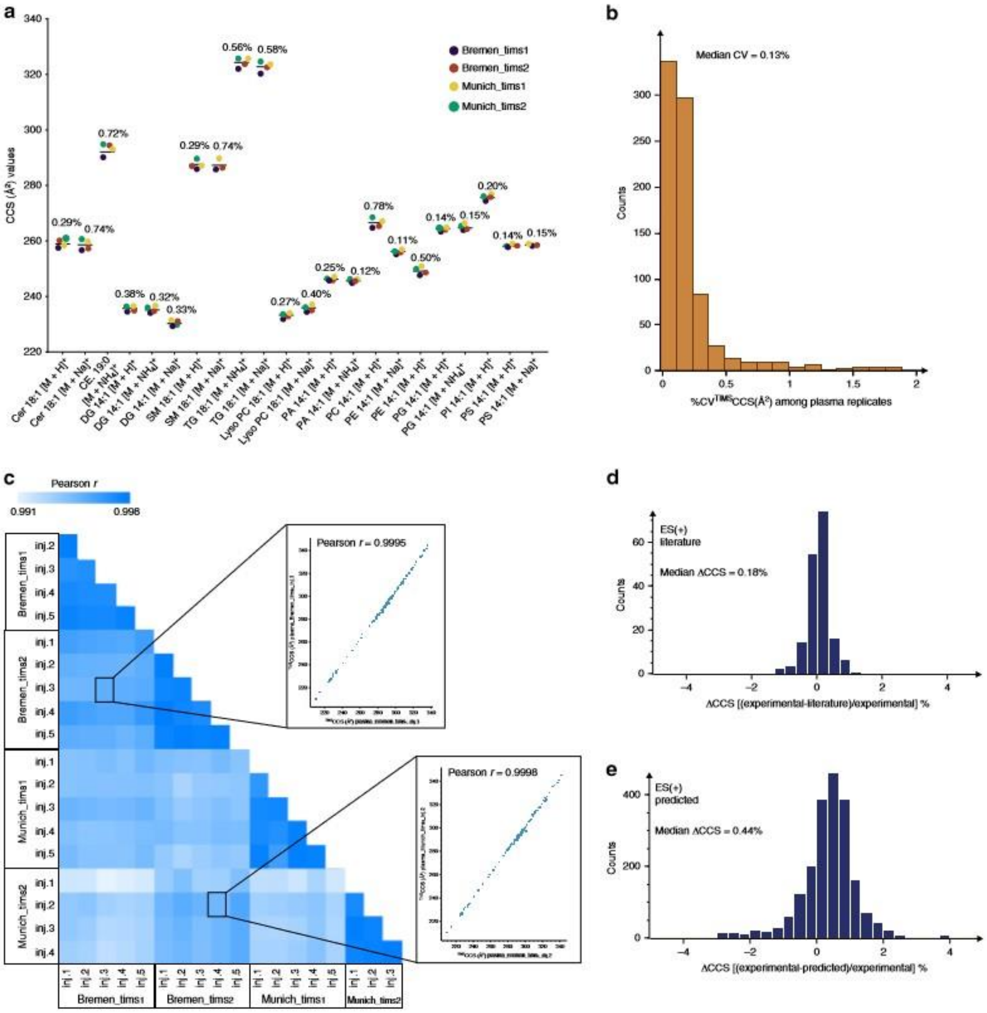


Fig. 5 Precise and accurate determination of lipid ^{TIMS}CCS values. a Cross-instrument and cross-laboratory ^{TIMS}CCS measurement of a mixture of standard compounds (source data provided in Supplementary Data 8). Data labels indicate the coefficient of variation (CV) ($n = 4$ instruments). **b** CVs of ^{TIMS}CCS values for lipids commonly identified in replicate injections of a human plasma sample ($n = 5$ replicates, $n = 1$ instrument). **c** Pair-wise correlation of lipid ^{TIMS}CCS values from human plasma SRM 1950 measured on four different timsTOF Pro instruments. Relative deviation of experimental ^{TIMS}CCS values in this study (**d**) from literature reports[23,24] and **e** machine learning predictions[24].

The investigation of the correlation of lipid mass and ion mobility has been a long term interest in ion mobility spectrometry-based lipidomics[23,24]. TIMS and PASEF provide a very efficient way to extend the scope of such studies to complex biological samples. Figure 6a shows a three-dimensional representation of all identified lipids in all three sample types in positive ionization mode color-coded by their respective classes. Each lipid class occupies a discrete space in the conformational landscape, which reflects the structural differences in their chemical composition. Hydrophilic lipids, such as monoacyl- and low molecular weight diacylglycerophospholipids (*m/z* 400–600) that elute first in reversed-phase chromatography, distribute in the CCS dimension from 204 to 253 Å^2. The second half of the LC gradient is dominated by the large population of glycerolipids and glycerophospholipids, which are often co-eluting, but distinct in mass and CCS. For example, diacylglycerols (DAG) and triacylglycerols (TAG) differ by one acyl chain and occupy a different CCS space shifted by 54.7 Å^2. Similarly, the head groups can strongly influence the ion mobility of lipids, as exemplified by PIs and PGs with the same acyl chain composition (Fig. 6a and Supplementary Fig. 6).

A key feature of our workflow is that each detected MS feature is precisely positioned in the multi-dimensional data cuboid, while the speed of PASEF ensures that most of these features are associated with MS/MS information. We hypothesized that the combined information can be leveraged to assign putative lipid

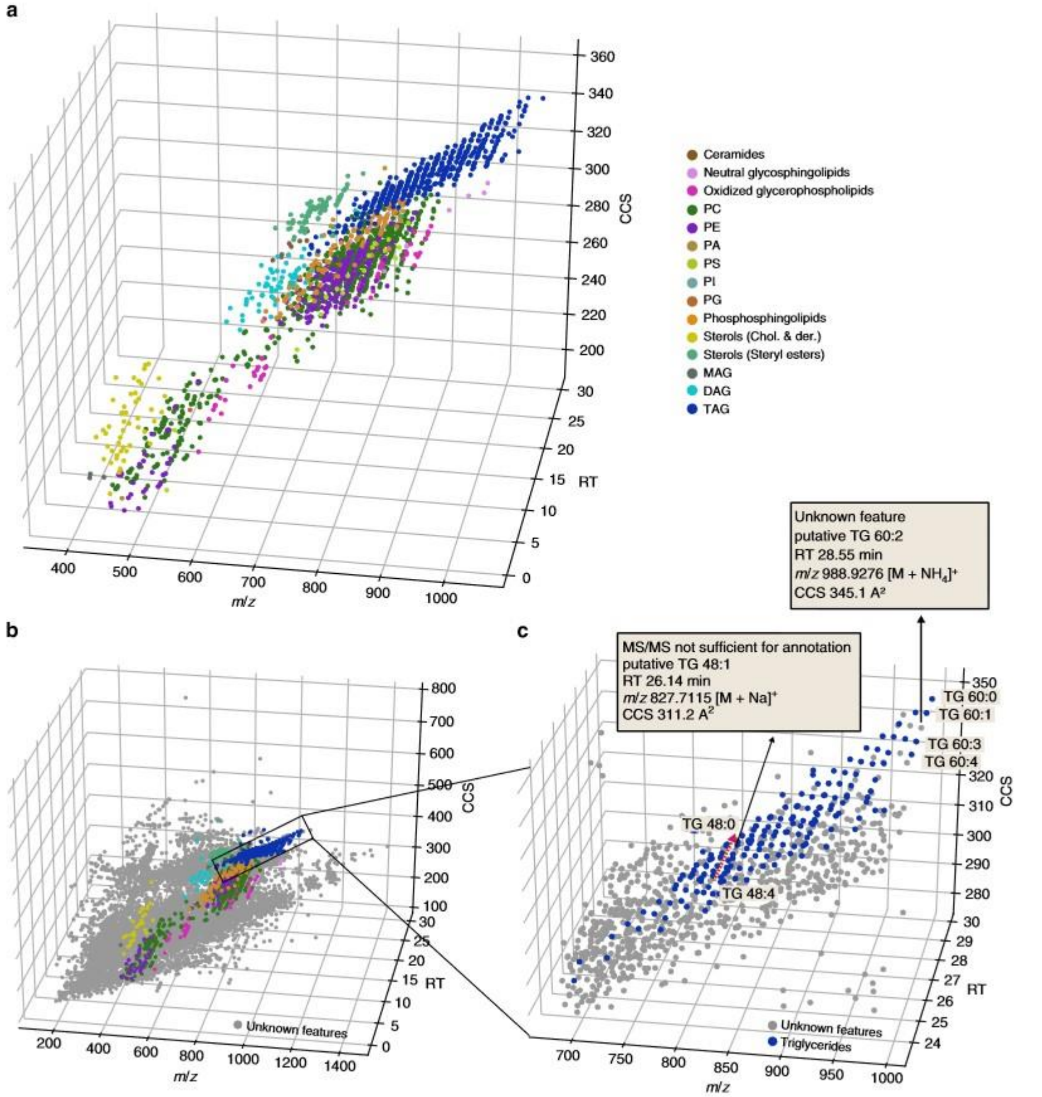


Fig. 6 The conformational landscape of lipid ions in TIMS. a Three-dimensional (RT, *m/z*, CCS) distribution of 1856 lipids from various classes from three biological samples (plasma, liver, HeLa) in positive ion mode. **b** Overlay of unidentified (gray) and identified MS features detected in a human plasma sample. **c** Zoom into the data cuboid and putative assignment of two previously unidentified lipids based on their relative position in the data space. PC = Phospatidylcholine, PE = Phospatidylethanolamine, PA = Phosphatidic acid, PI = Phospatidylinositol, PG = Phospatidylglycerol, PS = Phospatidylserine, MAG = Monoacylglycerol, DAG = Diacylglycerol, TAG = Triacylglycerol.

identifications for features that would otherwise have remained unidentified. To test this, we overlaid all detected features with MS/MS information on top of all identified lipids (Fig. 6b). Zooming into the distinct space occupied by triglycerides revealed the conformational fine-structure of this lipid class, which results in clusters of lipids with the same acyl chain composition (Fig. 6c). Within each cluster, the lipids are differentiated by their degree of unsaturation as the addition of a double bond decreases the CCS value almost linearly. This enabled the identification of features that were not fully characterized by the available MS/MS information. As an example, we putatively assigned the MS feature at retention time 26 min, *m/z* 827.7115 ($\Delta = 1.9$ ppm) and CCS 311.2 Å^2 as TG 48:1 based on the relative position in the 3D space. This is further supported by the predicted CCS value of 308.4 Å^2, which deviates <1% from our experimental value. Similarly, we derived a putative assignment for TG 60:2, which had escaped identification due to a low-quality MS/MS spectrum in this particular experiment.

Discussion

Trapped ion mobility spectrometry is a particularly compact and efficient ion mobility setup, in which ions are held against an incoming gas flow and released as a function of their size and shape. The Bruker timsTOF Pro incorporates two TIMS analyzers in the front part of a high-resolution quadrupole TOF mass spectrometer and fully supports our recently introduced PASEF scan mode. In this study, we developed a nanoflow lipidomics workflow that takes full advantage of TIMS and PASEF.

In 100 ms TIMS scans, readily compatible with fast chromatography, we made use of the ion mobility separation to fragment on average 15 precursors per PASEF scan by rapidly switching the mass position of the quadrupole. Even though this translates into MS/MS acquisition rates above 100 Hz, the ion count per spectrum, and thus the sensitivity, is determined by the TIMS accumulation time (here 100 ms or 10 Hz). In principle, this allows to acquire MS/MS spectra for all detectable isotope patterns in short LC-MS runs, which would otherwise require much longer gradients or multiple injections and advanced acquisition strategies. While many of the acquired spectra remained unidentified in our current data analysis pipeline, the PASEF acquisition strategy generates very comprehensive digital MS/MS archives of all samples that can be mined with alternative and novel search algorithms at any time in the future.

Our results indicate that the nanoflow PASEF lipidomics workflow is readily applicable to a broad range of biological samples, such as body fluids, tissue samples and cell cultures. With a single extraction step, we quantified thousands of lipids with very high accuracy and reproducibility from as little as micrograms of tissue or only a few thousand cells. This makes our workflow very attractive for sample-limited applications, such as lipid analysis from biopsies, micro dissected tissue or sorted cell populations.

An important application of lipidomics is the investigation of body fluids, for example blood plasma. Our analysis of a human reference sample was in good agreement with previous reports and greatly surpassed the coverage of glycerol- and glycerophospholipids, using only a fraction of the analysis time and sample amount. Based on these results, we estimate a limit of detection in the attomole range for these analyte classes.

Our online PASEF lipidomics workflow positions each lipid in a four-dimensional space with a precision of 1 ppm for masses, <0.2% for CCS and about 1% for retention times. We exemplified that this precision and accuracy can be leveraged to facilitate lipid assignment in addition to the comprehensive MS/MS information generated by PASEF. To further explore this data space, we compiled a library of over 1800 high-precision lipid CCS values directly from unfractionated biological samples. Our dataset largely extends the number of reported lipid CCS values and provides a basis for emerging machine learning techniques to predict CCS values more accurately and for a broader range of lipid classes.

We conclude that TIMS and PASEF enable highly sensitive and accurate lipidomics, and generate comprehensive digital archives of all detectable species along with very precise ion mobility measurements—a wealth of information which awaits full exploration and application. We also note that all the analytical advantages demonstrated for lipidomics should carry over to metabolomics in general, an area that we are currently exploring.

Methods

Chemicals and biological samples. 1-Butanol (BuOH), iso-propanol (IPA), ortho-phosphoric acid, formic acid, methanol (MeOH), and water were purchased from Fisher Scientific (Germany), and methyl tert-butyl ether (MTBE) from Sigma Aldrich (Germany) in analytical grade or higher purity. The standard lipid mixture Differential Ion Mobility System Suitability Lipidomix Kit was purchased from Avanti Polar Lipids, Inc (product no. 330708). The human plasma reference standard NIST SRM 1950 was obtained from Sigma Aldrich. Human cancer cells (HeLa, human, ACC57, DSMZ) were cultured in Dulbecco's modified Eagle's medium (DMEM), with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin–streptomycin (all from PAA Laboratories, Germany) and collected by centrifugation. The cell pellets were washed, frozen in liquid nitrogen and stored at −80 °C. Mouse liver was dissected from an individual male mouse (strain: C57BL/6) and snap frozen immediately. Animal experiments were performed in compliance with the ethical and institutional regulations of the Max Planck Institute of Biochemistry for animal testing and research, and have been approved by the government agencies of Upper Bavaria.

Lipid extraction. Lipids were extracted using an adapted MTBE protocol[35]. Plasma samples were thawed on ice and the sample preparation was performed at 4 °C. 200 µL cold MeOH were added to 1 µL of blood plasma and vortexed for 1 min. Subsequently, 800 µL of cold MTBE were added and the sample was mixed for another 6 min before adding 200 µL water. To separate the organic and aqueous phases, we centrifuged the mixture at 10,000 g for 10 min at 4 °C. The upper organic phase was collected and vacuum-centrifuged to dryness. To extract lipids from mouse liver, we first homogenized 1 mg of tissue in methanol and followed the extraction protocol described above. To extract lipids from ~5×10^5 HeLa cells, they were lysed after addition of MTBE by sonication (Bioruptor, Diagenode, Belgium). The dried lipid extracts from all samples were reconstituted in BuOH: IPA:water in a 8:23:69 ratio (v/v/v) with 5 mM phosphoric acid (nanoflow LC)[14] or in MeOH: Dichloromethane 9:1 (v/v) (high-flow LC) for LC-MS analysis.

Liquid chromatography. An Easy-nLC 1200 (Thermo Fisher Scientific) ultra-high pressure nanoflow chromatography system was used to separate lipids on an in-house reversed-phase column (20 cm × 75 µm i.d.) with a pulled emitter tip, packed with 1.9 µm C_{18} material (Dr. Maisch, Ammerbuch-Entringen, Germany). The column compartment was heated to 60 °C and lipids were separated with a binary gradient at a constant flow rate of 400 nL/min. Mobile phases A and B were ACN: H_2O 60:40% (v/v) and IPA:ACN 90:10% (v/v), both buffered with 0.1% formic acid and 10 mM ammonium formate. The 30 min LC-MS experiment started by ramping the mobile phase B from 1 to 30% within 3 min, then to 51% within 4 min and then every 5 min to 61, 71 and 99%, where it was kept for 5 min and finally decreased to 1% within 1 min and held constant for 2 min to re-equilibrate the column. The total LC runtime was ~40 min including time for re-filling of LC pumps and sample loading before the start of the analytical gradient.

The gradient was extended proportionately for 60 min and 90 min experiments. We injected 1 µL in positive and 2 µL in negative ion mode on column and each sample was injected five times in both ionization modes.

For the high-flow experiments (flow rate 0.4 mL/min), an Elute-HT UHPLC system (Bruker Daltonics Bremen, Germany) was used with an Intensity C_{18} column (100 mm × 2.1 mm, 1.9 µm beads) (Bruker, Daltonics, Germany) heated to 55 °C. Mobile phases and gradient were the same as with the nano-flow setup. The injection volume was 5 µL and each sample was injected five times in both ionization modes.

Trapped ion mobility—PASEF mass spectrometry. The nanoLC was coupled to a hybrid trapped ion mobility-quadrupole time-of-flight mass spectrometer (timsTOF Pro, Bruker Daltonics, Bremen, Germany) via a modified[16] nano-electrospray ion source (Captive Spray, Bruker Daltonics). For a detailed description of the mass spectrometer, see Refs. [33,34]. Briefly, electrosprayed ions enter the first vacuum stage where they are deflected by 90° and accumulated in the front part of a dual TIMS analyzer. The TIMS tunnel consists of stacked electrodes printed on circuit boards with an inner diameter of 8 mm and a total length of 100 mm, to which an RF potential of 300 V_{pp} is applied to radially confine the ion cloud. After the initial accumulation step, ions are transferred to the second part of the TIMS analyzer for ion mobility analysis. In both parts, the RF voltage is superimposed by an electrical field gradient (EFG), such that ions in the tunnel are dragged by the incoming gas flow from the source and retained by the EFG at the same time. Ramping down the electrical field releases ions from the TIMS analyzer in order of their ion mobility for QTOF mass analysis. The dual TIMS setup allows operating the system at 100% duty cycle, when accumulation and ramp times are kept equal[29]. Here, we set the accumulation and ramp time to 100 ms each and recorded mass spectra in the range from *m/z* 50–1550 in both positive and negative electrospray modes. The ion mobility was scanned from 0.6 to 1.95 Vs/cm^2. Precursors for data-dependent acquisition were isolated within ± 1 Th and fragmented with an ion mobility-dependent collision energy, which was linearly increased from 25 to 45 eV in positive mode, and from 35 to 55 eV in negative mode. The overall acquisition cycle of 0.4 s comprised one full TIMS-MS scan and three PASEF MS/MS scans. Low-abundance precursor ions with an intensity above a threshold of 100 counts but below a target value of 4000 counts were repeatedly scheduled and otherwise dynamically excluded for 0.2 min. TIMS ion charge control was set to 5e6. The TIMS dimension was calibrated linearly using four selected ions from the

Agilent ESI LC/MS tuning mix [m/z, $1/K_0$: (322.0481, 0.7318 Vs cm^{-2}), (622.0289, 0.9848 Vs cm^{-2}), (922.0097, 1.1895 Vs cm^{-2}), (1221,9906, 1.3820 Vs cm^{-2})] in positive mode and [m/z, $1/K_0$: (666.01879, 1.0371 Vs cm^{-2}), (965.9996, 1.2255 Vs cm^{-2}), (1265.9809, 1.3785 Vs cm^{-2})] in negative mode.

Dilution series experiment. Plasma SRM 1950 was extracted on the same day and injected in both nano- and high-flow LC systems as described above. For positive mode experiments with the high-flow system, 30 µL plasma was extracted and reconstituted in 300 µL reconstitution solvent to inject a final volume of 5 µL (translating into 0.5 µL plasma on column). For the nanoflow separation, 1 µL plasma was extracted and reconstituted in 20 µL reconstitution solvent to inject a final volume of 1 µL (translating into 0.05 µL plasma on the column). These stock samples were then sequentially diluted in 1:3.3, 1:10, 1:33, 1:100, 1:333, and 1:1000 ratios (vol:vol). For the high-flow experiment in negative mode, 30 µL plasma was extracted and reconstituted in 30 µL to inject a final volume of 5 µL (5 µL plasma on column). For the nanoflow experiment, 1 µL plasma was extracted and reconstituted in 20 µL to inject a final volume of 2 µL (0.1 µL plasma on column). The samples were diluted to the final concentrations as above. All samples were injected in five replicates.

Data analysis and bioinformatics. The mass spectrometry raw files were analyzed with MetaboScape alpha version 5.0 (Bruker Daltonics, Germany). This version contains a novel feature finding algorithm (T-ReX 4D) that automatically extracts data from the four-dimensional space (m/z, retention time, ion mobility and intensity) and assigns MS/MS spectra to them. Masses were recalibrated with the lock masses m/z 622.028960 (positive mode) and m/z 666.019887 (negative mode) and the ion mobility dimension was recalibrated using the ions of the tuning mix as above. Feature detection was performed using an intensity threshold of 500 counts in positive mode and 200 counts in the negative mode. The minimum number of data points in the 4D TIMS space was set to 100, or 50 when using recursive feature extraction.

Lipid annotation of detected molecular features with assigned MS/MS spectra was performed using the high-throughput lipid search (HTP) function of SimLipid v6.05 software (PREMIER Biosoft, Palo Alto, USA). The lipid search comprised four lipid categories, Glycerolipids (GL), Glycerophospholipids (GP), Sphingolipids (SP) and Sterol lipids (SL) and TAG, DAG, PA, PC, PE, PG, PI, PS, Ceramides, Sphingomyelins, Neutral Glycosphingolipids, Steryl esters, Cholesterols, and Derivatives, as well as oxidized glycerophospholipid classes. PE and PC lipids with ether- and plasmalogen- substituents were considered. Lipid species from TAG and sterol classes were not considered for the negative mode MS/MS database search. Glycerophospholipids were only considered if containing an even number of carbons on at least one of the fatty acid chains. We searched for $[M+H]^+$, $[M+Na]^+$, and $[M+NH_4]^+$ ions in positive mode, and $[M–H]^-$, $[M+Cl]^-$, $[M–CH_3]^-$, $[M+HCOO]^-$ and $[M+AcO]^-$ in negative mode. The precursor ion and MS/MS fragment mass tolerances were set to 5 and 10 ppm, respectively.

The initial search results were filtered to ensure that lipids were annotated based on high-quality MS/MS spectra with fragment ions corresponding to structure specific characteristic ions. To this end, we manually inspected the SimLipid results and removed potential false positives and refined lipid annotations based on head-groups and/or fatty acyl composition as follows. In positive mode, we rejected unlikely PC lipid ion species such as $[PC+NH_4]^+$ and $[PC+Na]^+$ if the corresponding $[M+H]^+$ ion was not observed, and if $[M\text{-}59+Na]^+$ (neutral loss of $(CH_3)_3N$) and $[M\text{-}183+Na]^+$ (neutral loss of phosphocholine) fragments were not detected. Lipids from GP, ST, and SP categories were required to have their corresponding head group diagnostic ions e.g., m/z 369.3516 for cholesterol esters, m/z 184.073 for PC lipid species, as well as the neutral loss of 141 Th for PEs. Neutral glycosphingolipids and ceramides were rejected if the structure-specific N ″-type fragments were not annotated. However, lipid species from sterol classes were accepted if the precursor ion was the base peak in the MS/MS spectrum. TG/DG lipids with three or two unique fatty acid chains were reported only if at least two/one fatty acid chain fragment ion were/was detected. In negative mode, we rejected all lipid annotations for which we did not detect at least one characteristic fragment ion corresponding to one of the fatty acid chains.

We report lipid identifications with increasing level of fragment ion evidence using the following nomenclature: (i) a short name (e.g., PC 32:1) to indicate mass-resolved lipid molecular species, (ii) a long name for composition-resolved identifications where the symbol @ indicates that this particular acyl-chain is not fully characterized by fragment ions (e.g., Cer d18:1_26:0@), and (iii) a long name where head group and fatty acyl-chain composition are fully characterized (e.g., PG 16_1:16:1). Note that sn1/sn2/sn3 chain assignments, positions of the double bonds, as well as cis/trans isomers are not evident from our data and therefore not annotated.

Lipid CCS values were predicted in MetaboScape using SMILES from LipidMaps based on a support vector machine learning approach by Zhou et al.[24]. Mass spectrometric metadata such as the PASEF frame MS/MS information were extracted from the.tdf files using an SQLite database viewer (SQLite Manager v3.10.1). Further data analysis and visualization was performed in Python 3 (Jupyter Notebook) and Perseus (v1.6.0.8)[44].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry raw files have been uploaded to the MASS Spectrometry Interactive Virtual Environment (MassIVE) and are accessible via the identifier MSV000083858. Source data for all figures are provided in the Supplementary Data 1–14. All other data are available from the corresponding authors on reasonable request.

Received: 29 May 2019; Accepted: 12 December 2019;
Published online: 16 January 2020

References

1. Shevchenko, A. & Simons, K. Lipidomics: coming to grips with lipid diversity. *Nat. Rev. Mol. Cell Biol.* **11**, 593–598 (2010).
2. Röhrig, F. & Schulze, A. The multifaceted roles of fatty acid synthesis in cancer. *Nat. Rev. Cancer* **16**, 732–749 (2016).
3. Parker, B. L. et al. An integrative systems genetic analysis of mammalian lipid metabolism. *Nature* **567**, 187–193 (2019).
4. Han, X. & Gross, R. W. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry. *J. Lipid Res.* **44**, 1071–1079 (2003).
5. Schwudke, D., Schuhmann, K., Herzog, R., Bornstein, S. R. & Shevchenko, A. Shotgun lipidomics on high resolution mass spectrometers. *Cold Spring Harb. Perspect. Biol.* **3**, 1–13 (2011).
6. Ejsingr, C. S. et al. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. *PNAS* **106**, 2136–2141 (2008).
7. Cajka, T. & Fiehn, O. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. *Trends Anal. Chem.* **61**, 192–206 (2014).
8. Hutchins, P. D., Russell, J. D. & Coon, J. J. LipiDex: an integrated software package for high-confidence lipid identification. *Cell Syst.* **6**, 621–625.e5 (2018).
9. Kind, T. et al. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods* **10**, 755–758 (2013).
10. Sud, M. et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35**, 527–532 (2007).
11. Hartler, J., Triebl, A., Ziegl, A., Trötzmüller, M. & Gerald, N. Deciphering lipid structures based on platform-independent decision rule sets. *Nat. Methods* **14**, 1171–1174 (2018).
12. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**, 1–8 (1996).
13. Park, S. M. et al. Lipidomic perturbations in lung, kidney, and liver tissues of p53 knockout mice analyzed by nanoflow UPLC-ESI-MS/MS. *J. Proteome Res.* **15**, 3763–3772 (2016).
14. Danne-Rasche, N., Coman, C. & Ahrends, R. Nano-LC/NSI MS refines lipidomics by enhancing lipid coverage, measurement sensitivity, and linear dynamic range. *Anal. Chem.* **90**, 8093–8101 (2018).
15. Eliuk, S. & Makarov, A. Evolution of orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem.* **8**, 61–80 (2015).
16. Beck, S. et al. The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol. Cell. Proteom.* **14**, 2014–2029 (2015).
17. Cumeras, R., Figuerasa, E., Davis, C. E., Baumbach, J. I. & G., I. Trying to detect gas-phase ions? Understanding ion mobility spectrometry: part 2: hyphenated methods and effects of experimental parameters. *Analyst* **140**, 1391–1410 (2015).
18. Cumeras, R., Figueras, E., Davis, C. E., Baumbach, J. I. & Gracia, I. Review on ion mobility spectrometry. Part 1: current instrumentation. *Analyst* **140**, 1376–1390 (2015).
19. Paglia, G. et al. Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Anal. Chem.* **87**, 1137–1144 (2015).
20. Blažević, I. et al. Increasing compound identification rates in untargeted lipidomics research with liquid chromatography drift time-ion mobility mass spectrometry. *Anal. Chem.* **90**, 10758–10764 (2018).
21. Rainville, P. D. et al. Ion mobility spectrometry combined with ultra performance liquid chromatography/mass spectrometry for metabolic phenotyping of urine: effects of column length, gradient duration and ion mobility spectrometry on metabolite detection. *Anal. Chim. Acta* **982**, 1–8 (2017).
22. Lintonen, T. P. I. et al. Differential mobility spectrometry-driven shotgun lipidomics. *Anal. Chem.* **86**, 9662–9669 (2014).
23. Leaptrot, K. L., May, J. C., Dodds, J. N. & McLean, J. A. Ion mobility conformational lipid atlas for high confidence lipidomics. *Nat. Commun.* **10**, 985 (2019).

24. Zhou, Z., Tu, J., Xiong, X., Shen, X. & Zhu, Z. J. LipidCCS: prediction of collision cross-section values for lipids with high precision to support ion mobility-mass spectrometry-based lipidomics. *Anal. Chem.* **89**, 9559–9566 (2017).
25. Soper-Hopper, M. T. et al. Collision cross section predictions using 2-dimensional molecular descriptors. *Chem. Commun.* **53**, 7624–7627 (2017).
26. Picache, J. A. et al. Collision cross section compendium to annotate and predict multi-omic compound identities. *Chem. Sci.* **10**, 983–993 (2019).
27. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion-. Mobil. Spectrom.* **14**, 93–98 (2011).
28. Fernandez-Lima, F. A., Kaplan, D. A. & Park, M. A. Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **82**, 126106 (2011).
29. Ridgeway, M. E., Lubeck, M., Jordens, J., Mann, M. & Park, M. A. Trapped ion mobility spectrometry: a short review. *Int. J. Mass Spectrom.* **425**, 22–35 (2018).
30. Michelmann, K., Silveira, J. A., Ridgeway, M. E. & Park, M. A. Fundamentals of trapped ion mobility spectrometry. *J. Am. Soc. Mass Spectrom.* **26**, 14–24 (2014).
31. Silveira, J. A., Michelmann, K., Ridgeway, M. E. & Park, M. A. Fundamentals of trapped ion mobility spectrometry part II: fluid dynamics. *J. Am. Soc. Mass Spectrom.* **27**, 585–595 (2016).
32. Jeanne Dit Fouque, K. et al. Effective liquid chromatography-trapped ion mobility spectrometry-mass spectrometry separation of isomeric lipid species. *Anal. Chem.* **91**, 5021–5027 (2019).
33. Meier, F. et al. Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14**, 5378–5387 (2015).
34. Meier, F. et al. Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteom.* **17**, 2534–2545 (2018).
35. Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A. & Schwudke, D. Lipid extraction by methyl- tert -butyl ether for high-throughput lipidomics. *J. Lipid Res.* **49**, 1137–1146 (2008).
36. Simón-Manso, Y. et al. Metabolite profiling of a NIST standard reference material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal. Chem.* **85**, 11725–11731 (2013).
37. Burla, B. et al. MS-based lipidomics of human blood plasma: a community-initiated position paper to develop accepted guidelines. *J. Lipid Res.* **59**, 2001–2017 (2018).
38. Bowden, J. A. et al. Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950–metabolites in frozen human plasma. *J. Lipid Res.* **58**, 2275–2288 (2017).
39. Quehenberger, O. et al. Lipidomics reveals a remarkable diversity of lipids in human plasma. *J. Lipid Res.* **51**, 3299–305 (2010).
40. Mason, E. & McDaniel E.W. Transport Properties of Ions in Gases (John Wiley and Sons, 1988).
41. Chai, M., Young, M. N., Liu, F. C. & Bleiholder, C. A transferable, sample-independent calibration procedure for trapped ion mobility spectrometry (TIMS). *Anal. Chem.* **90**, 9040–9047 (2018).
42. Stow, S. M. et al. An interlaboratory evaluation of drift tube ion mobility-mass spectrometry collision cross section measurements. *Anal. Chem.* **89**, 9048–9055 (2017).
43. Fahy, E., Cotter, D., Sud, M. & Subramaniam, S. Lipid classification, structures and tools. *Biochim. Biophys. Acta* **1811**, 637–647 (2011).
44. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).

Acknowledgements

We thank our colleagues in the Department of Proteomics and Signal Transduction for discussion and help, in particular D. Voytik, A. Strasser, J. Mueller, I. Paron, and C. Deiml. We acknowledge our colleagues H. Neuweger, N. Kessler, and S. Wehner at Bruker Daltonics in Bremen for their support in data processing and analysis. This work was partially supported by the German Research Foundation (DFG–Gottfried Wilhelm Leibniz Prize), the Novo Nordisk Foundation (Grant agreement NNF14CC0001), the European Commission (FP7 GA ERC-2012-SyG_318987–ToPAG), and the Max-Planck Society for the Advancement of Science.

Author contributions

C.G.V., K.S., M.M., and F.M. designed the research project; C.G.V., K.S., A.D.B., S.M., A.B., and F.M. performed the experiments; C.G.V., K.S., N.S.M., U.S.H., S.M., A.B., and F.M. analyzed the data; N.S.M. as well as U.S.H., S.M., and A.B. contributed analytical tools; C.G.V., M.M., and F.M. wrote the paper.

Competing interests

The following authors state that they have potential conflicts of interest regarding this work: U.S.H., S.M., and A.B. are employees of Bruker, the manufacturer of the timsTOF Pro, and N.S.M. is employee of PREMIER Biosoft, the vendor of the SimLipid software. The other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-14044-x.

Correspondence and requests for materials should be addressed to M.M. or F.M.

Peer review information *Nature Communications* thanks John McLean, and the other, anonymous, reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

3.2.2. Article 5: Deep learning the CCS peptide universe

Deep learning the collisional cross sections of the peptide universe from a million experimental values

Nature communications, February 19, 2021

Florian Meier[1, 2, *], Niklas D. Köhler[3, *], **Andreas-David Brunner[1, *],** Jean-Marc H. Wanka[3], Eugenia Voytik[1], Maximilian T. Strauss[1], Fabian J. Theis[3, #], Matthias Mann[1, 4, #]

** These authors contributed equally to this work*

# Correspondence

[1]Department for Proteomics and Signal transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[2]Functional Proteomics, Jena University Hospital, Jena, Germany

[3]Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

[4]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Denmark

Contribution

In this project, I contributed to all aspects of the paper including the overall experimental design and establishment of its goals. I performed the sample preparation of all presented model organisms with different proteases, including the optimization of peptide isolation and fractionation conditions. Furthermore, I acquired all data for the presented model organisms and the ProteomicsTOOLS peptide set. I also ensured highest quality of the data including downstream analysis and evaluation. This resulted in a total of more than 400 high-quality raw data files comprising more than 2,500,000 peptide collisional cross sections. I also contributed to data analysis, literature research and manuscript writing.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21352-8 OPEN

Deep learning the collisional cross sections of the peptide universe from a million experimental values

Florian Meier[1,5,6], Niklas D. Köhler[2,6], Andreas-David Brunner[1,6], Jean-Marc H. Wanka[2], Eugenia Voytik[1], Maximilian T. Strauss[1], Fabian J. Theis[2,3✉] & Matthias Mann[1,4✉]

The size and shape of peptide ions in the gas phase are an under-explored dimension for mass spectrometry-based proteomics. To investigate the nature and utility of the peptide collisional cross section (CCS) space, we measure more than a million data points from whole-proteome digests of five organisms with trapped ion mobility spectrometry (TIMS) and parallel accumulation-serial fragmentation (PASEF). The scale and precision (CV < 1%) of our data is sufficient to train a deep recurrent neural network that accurately predicts CCS values solely based on the peptide sequence. Cross section predictions for the synthetic ProteomeTools peptides validate the model within a 1.4% median relative error ($R > 0.99$). Hydrophobicity, proportion of prolines and position of histidines are main determinants of the cross sections in addition to sequence-specific interactions. CCS values can now be predicted for any peptide and organism, forming a basis for advanced proteomics workflows that make full use of the additional information.

[1] Department Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2] Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany. [3] Department of Mathematics, TU München, Munich, Germany. [4] NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [5] Present address: Functional Proteomics, Jena University Hospital, Jena, Germany. [6] These authors contributed equally: Florian Meier, Niklas D. Köhler, Andreas-David Brunner. ✉email: fabian.theis@helmholtz-muenchen.de; mman@biochem.mpg.de

The combination of ion mobility spectrometry (IMS) and mass spectrometry (MS) extends conventional liquid chromatography–mass spectrometry (LC–MS) by an extra dimension of separation, increasing peak capacity, selectivity, and depth of analysis[1–5]. Recent advances have greatly improved the sensitivity of commercially available IMS devices and the technology is now set for a broader application in MS-based proteomics[6–10].

IMS separates ions in the gas phase (typically in the mbar pressure range) based on their size and shape within milliseconds. This time scale allows recording full ion mobility spectra between typical chromatographic peaks (seconds) and the acquisition pulses of time-of-flight (TOF) instruments (~100 µs). We have recently integrated trapped ion mobility spectrometry (TIMS)[11,12], a relatively new and particularly compact ion mobility device, with a high-resolution quadrupole TOF mass analyzer[10,13,14]. In MS/MS mode, this opens up the possibility to step the precursor selection window as a function of ion mobility, allowing the fragmentation of multiple precursors during a single TIMS scan[13]. We termed this novel scan mode parallel accumulation-serial fragmentation (PASEF) and demonstrated that it increases MS/MS rates more than ten-fold without any loss in sensitivity as is otherwise inherent to faster scanning rates[10,15].

An intriguing feature of the combination of TIMS and PASEF is that it should allow the acquisition of ion mobility values on a very large scale. Such data have previously been measured on a case by case basis by classical drift tube IMS, in which a weak electric field drags ions through an inert buffer gas[16–18]. Larger ions collide more frequently with gas molecules and hence traverse the drift tube with a lower speed as compared with their smaller counterparts. In TIMS the physical process is the same, except that the setup is reversed with the electric field holding ions stationary against an incoming gas flow, prior to their controlled release from the device by lowering the electric field[19,20]. In both cases, the measured ion mobility (reported as the reduced ion mobility coefficient K_0) can be used to derive a collisional cross section (CCS), which is the rotational average of an ion's gas-phase conformation[21,22]. The CCS intrinsically depends on the ion structure, which is also illustrated by the fact that different classes of biomolecules (e.g., metabolites, carbohydrates, peptides) show different trends in their ion mobilities as a function of molecular mass[23]. Interestingly, conformations also vary within a compound class - to the extent that isobaric peptide sequences can be distinguishable by their different CCS[24,25].

The link between the amino acids of a peptide and its measured cross section has the potential to increase the confidence in its identifications through reference or predicted CCS values. This has motivated researchers to develop various (machine learning) models based on amino acid-specific parameterization and physicochemical properties[16,26–29]. However, as comprehensive experimental data are not available, predicting the full complexity of the peptide conformational space remains elusive. Furthermore, it is not clear which properties should be considered to best parameterize such models and make them generalizable. We reasoned that a combination of very large and consistent datasets acquired by PASEF with state of the art deep learning methods would address both challenges. Due to their inherent flexibility and their ability to scale to large datasets, deep learning methods have proven very successful in genomics[30,31] and more recently in proteomics for the prediction of retention times and fragmentation spectra[32–35].

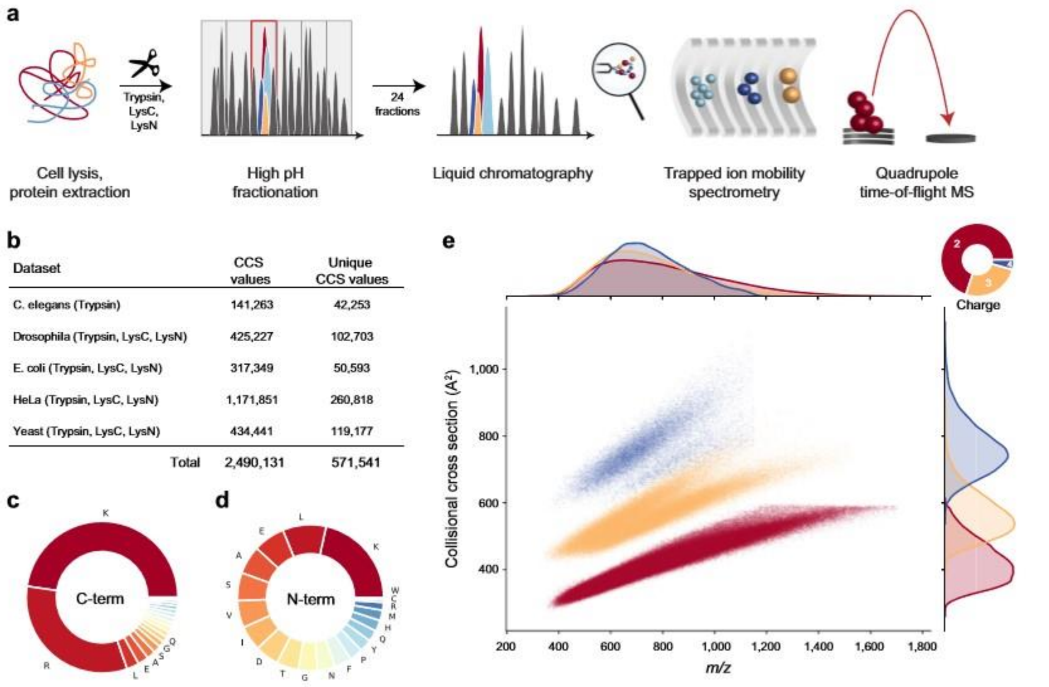
We here set out to explore the nature and utility of the peptide CCS space in proteomics by first measuring a very large dataset of CCSs by TIMS-TOF PASEF across five different biological species. Building on this dataset, we develop and train a bi-directional recurrent neural network with long short-term memory (LSTM) units to predict CCS values for any peptide sequence in the tryptic peptide universe. Interpreting our network based on recent approaches from explainable AI allows us to investigate the nature of the underlying relationship between linear peptide sequence and peptide cross section.

Results

Construction of a very large-scale peptide CCS dataset. To fully capture the conformational diversity of peptides in the gas phase, we generated peptides from whole-cell proteomes of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *HeLa*, and *budding yeast* using up to three different enzymes with complementary cleavage specificity (trypsin, LysC, and LysN). To increase the depth of our analysis, we split peptide mixtures into 24 fractions per organism and analyzed each of them separately with PASEF on a TIMS-quadrupole TOF MS (Methods; Fig. 1a). As this is the same setup we used before, we combined our new experimental data with our previously reported dataset from a tryptic HeLa digest[10].

In total, we compiled 360 LC-MS/MS runs and processed them in the MaxQuant software[36,37]. This resulted in about 2.5 million peptide spectrum matches and 426,845 unique peptide sequences at globally controlled false discovery (FDR) rates of less than 1% at the peptide and protein levels for each organism and enzyme. MaxQuant links each peptide spectrum match to a four-dimensional (4D) isotope cluster (or 'feature') in mass, retention time, ion mobility, and intensity dimension. For each of these, the ion mobility value is determined as the intensity-weighted average of the corresponding mobilogram trace and can be converted into an ion-neutral CCS value using the Mason-Schamp equation[21]. Some peptides occur in more than one conformation and have multiple peaks in an LC-TIMS-MS experiment, but for simplicity we here chose to keep only the most abundant feature per charge state (Supplementary Fig. 1).

Overall, our dataset comprises over two million CCS values, which we collapsed to about 570,000 unique combinations of peptide sequence, charge state and, if applicable, side chain modifications such as oxidation of methionine (Fig. 1b). Peptide sequence lengths ranged from 7 up to 55 amino acids with a median length of 14. The trypsin and LysC datasets contributed 79% of the peptide sequences (C-terminal R or K), whereas LysN peptide (N-terminal K) accounted for the remaining 21%. Within the two classes of peptides, the proportion of the terminal amino acids conformed to their expected frequencies from the database (Fig. 1c, d). Due to our selection of enzymes, peptides should have at least one basic amino acid. Consequently, singly charged ions were a small minority (2%), which we excluded from further analysis. We detected 69% of the peptides in the doubly charged, and 25% in the triply charged and 4% in the quadruply charged state. Plotting the mass-to-charge (*m/z*) vs. CCS distribution of all peptides separates them by their charge state over the *m/z* range 400–1700 Å^2 and 300–1000 Å^2 in cross section (Fig. 1e). Within each charge state, *m/z* and CCS were correlated in accordance with previous observations in smaller datasets[10,18,23,38–40]. Overall, 95% of all tryptic peptides were distributed within ±8% around power-law trend lines for each charge state (Supplementary Fig. 2). Interestingly, the deviation increases with charge state and mass—to the extent that there are two distinct sub-populations for charge state 3—perhaps due to the increased amino acid variability and structural flexibility in longer sequences. Our data show that peptides occupy about one-quarter of the 2D *m/z*-mobility space, whereas a fully orthogonal 2D separation would occupy the full space. Assuming an average ion mobility resolution of 60, this translates into an at least ten-fold increased analytical peak capacity as compared with only MS (Supplementary Fig 3).



Dataset	CCS values	Unique CCS values
C. elegans (Trypsin)	141,263	42,253
Drosophila (Trypsin, LysC, LysN)	425,227	102,703
E. coli (Trypsin, LysC, LysN)	317,349	50,593
HeLa (Trypsin, LysC, LysN)	1,171,851	260,818
Yeast (Trypsin, LysC, LysN)	434,441	119,177
Total	2,490,131	571,541

Fig. 1 Large-scale peptide collisional cross section (CCS) measurement with TIMS and PASEF. a Workflow from extraction of whole-cell proteomes through digestion, fractionation, and chromatographic separation of each fraction. The TIMS-quadrupole TOF mass spectrometer was operated in PASEF mode. **b** Overview of the CCS dataset in this study by organism. **c** Frequency of peptide C-terminal amino acids. **d** Frequency of peptide N-terminal amino acids. **e** Distribution of 559,979 unique data points, including modified sequence and charge state, in the CCS vs. *m/z* space color-coded by charge state. Density distributions for *m/z* and CCS are projected on the top and right axes, respectively. Source data are provided as a Source Data file.

Evaluating the accuracy, precision, and utility of TIMS CCS measurements. Peak capacity indicates how many peptides can be analytically resolved from each other. However, for their identification it is sufficient to determine their apex positions with adequate precision. In MS-based proteomics, accurate measurement of the peptide mass greatly reduces the number of candidates in database searches[36], and the retention time can likewise be employed as a filter, as is typically done in the analysis of data-independent acquisition (DIA) experiments[41]. We reasoned that ion mobility values should be precise and reproducible as they are based on gas-phase interactions and defined electric fields, in contrast to chromatographic retention times, which depend on surface interactions that vary according to sample matrices and over time. We therefore investigated the precision, accuracy and added benefit of ion mobility measurements in our dataset.

First, we calculated correlation coefficients for retention times and CCS values from pair-wise overlapping tryptic peptides in the 168 LC-MS/MS runs that had the highest number of shared peptides across organisms. Depending on evolutionary distance, this number ranged from none to hundreds and these formed the basis of our calculations. We obtained two triangular half-matrices of color-coded Pearson correlation coefficients—one for the retention time correlations and one for CCS (upper and lower part of Fig. 2a, respectively). Correlation values were generally above 0.9 for both retention time and cross section, although experiments were done over several months on three different instruments. However, correlations of CCS values were systematically higher than those for retention times, for example, the median correlation for the HeLa runs between June 2018 and May 2019 is $r = 0.990$ for retention times and $r = 0.995$ for cross sections (based on 1264 peptides per pairwise comparison on average). Further, the upper triangle of the heatmap shows patches of similar color, unlike the mirrored positions in the lower triangle (Fig. 2a). This indicates chromatographic batch-effects resulting in non-linear shifts or changes in the peptide elution order. In contrast, the absence of similar patterns in the CCS comparisons supports our starting hypothesis that the ion mobility is largely independent of experimental circumstances.

Closer inspection of the variation in CCS values revealed mostly linear shifts, which do not affect the correlation coefficient. These shifts were only in the range from absolute 0 to 40 Å^2 (median 9.4 Å^2) even for very distant measurements, and they are mainly due to variations of the gas flow in the TIMS tunnel. Importantly, a linear alignment based on a few peptide CCS values almost completely corrects for these shifts (Methods, Fig. 2b). With such an alignment, CCS values can be compared across disparate datasets, which we did for all analyses shown here. Across the 347,885 peptide CCS values measured at least in duplicate, the median coefficient of variation (CV) was 0.4%, which highlights the excellent reproducibility of TIMS CCS measurements also over longer periods of time and across instruments (Fig. 2c). This may even be improvable as suggested by our previously reported CVs of 0.1% for replicate injections of a whole-proteome digest on a single instrument[10]. Reassuringly, we found an excellent correlation of $^{\text{TIMS}}\text{CCS}_{\text{N2}}$ values and drift

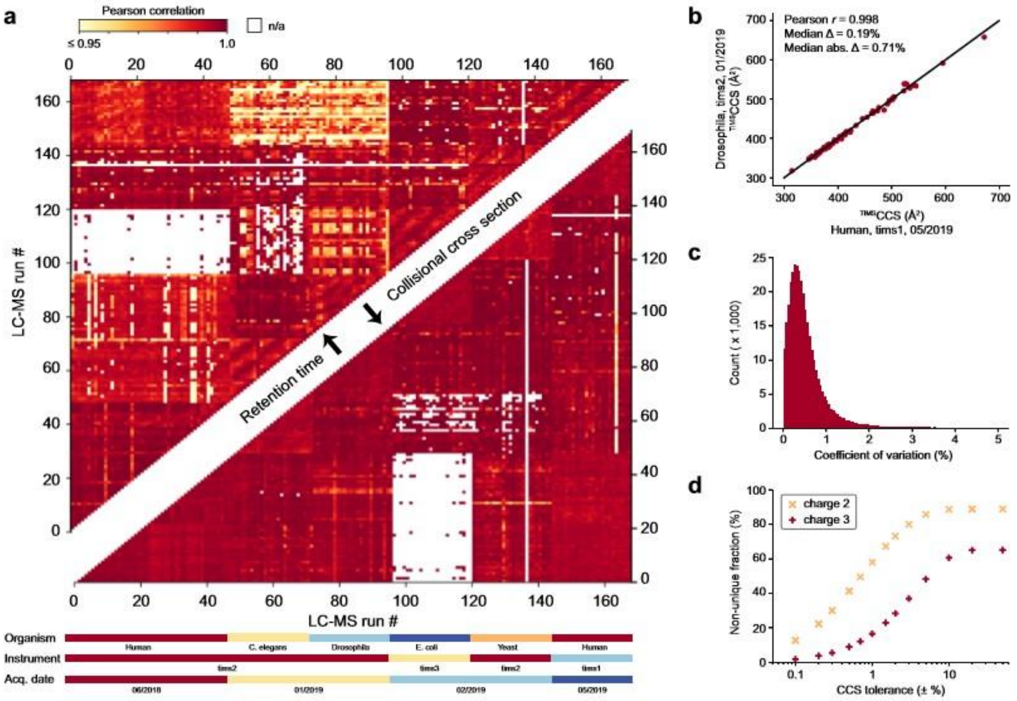


Fig. 2 Precision, accuracy, and utility of experimental peptide CCS values. a Color-coded pairwise Pearson correlation values of peptide retention time (upper triangular matrix) and CCS values (lower triangular matrix) between the 168 LC-MS/MS runs of fractionated tryptic digests. Experimental metadata are indicated below the *x*-axis. White (n/a) indicates less than 5 data points for pairwise comparison. **b** CCS values of shared tryptic peptides independently measured in two typical LC-MS runs of fractions from Drosophila and HeLa ($n = 68$). **c** CVs of repeatedly measured peptide CCS values in the full dataset ($n = 347,885$ peptides). **d** Specificity of combined peptide *m/z* and CCS information for doubly and triply charged peptides with C-terminal arginine or lysine ($n = 324,246$ and 112,015) with a fixed *m/z* tolerance of ±1.5 ppm and as a function of CCS tolerance. For details, see main text and Methods.

tube ion mobility experiments[42–44] (Pearson $r = 0.997$) with an average absolute deviation <2% (Supplementary Fig. 4).

To investigate the utility of the additional CCS information for peptide identification, we returned to Fig. 1e and defined tolerance windows in *m/z* and CCS dimensions for each peptide with C-terminal arginine or lysine as expected in tryptic digests (identified by MS/MS at an FDR < 1%). We then determined the fraction of windows in this map that were exclusively occupied by a single peptide, meaning a unique match between experimental measurement and our large peptide dataset (Fig. 2d). We set the mass tolerance at the median mass accuracy (±1.5 ppm) and varied the CCS tolerance separately for doubly and triply charged peptides, because they occupy different cross section areas (Methods). Without the CCS information, at ±50% tolerance, about 90% of the doubly charged and 67% of the triply charged peptides had at least one other peptide within 1.5 ppm distance ('non-unique'). The fraction of unique peptides increased once the CCS window was restricted to less than ±10%, in accordance with the roughly 20% spread of CCS values in Fig. 1e. Within three standard deviations (±1.5%) of the measured CCS values, about two-thirds of the doubly charged and 75% of the triply charged species were unique and these fractions increased progressively for narrower CCS windows. We thus conclude that ion mobility can substantially reduce the number of potential peptides that need to be considered, benefiting peptide identification or MS1 level feature matching. At current CCS value accuracy, this is about a factor of two to three. As Fig. 2d also shows, an increase in accuracy down to 0.1% could make the large majority of peptides unique (56% for 2^+ and 90% and 3^+ in a ±0.5% CCS window).

Dependence of CCS values on linear sequence determinants. Having investigated the accuracy and utility of peptide CCS values, we asked whether a dataset of this scale could also shed a light on potential substructures in the *m/z* vs. ion mobility space and the relationships between linear peptide sequences and their corresponding gas-phase structures. In the *m/z* vs. CCS space of Fig. 1e, more compact conformations appear below and more extended confirmations appear above the overall trend lines for CCS values as a function of *m/z*.

We first explored whether amino acids with preferences for secondary protein structures[45], would also effect peptide ion structures in the gas phase and form clusters in this global view (Supplementary Fig. 5). This is a long-standing interest in ion mobility research and detailed studies of model peptides revealed that in particular helical structures can be stable in the gas phase[46–48]. Mapping the amino acids in each peptide sequence that favor helices in proteins, we found a tendency toward higher CCS with an increasing fraction of A, L, M, H, Q, and E. This suggests that such peptides, indeed, have a propensity to adopt

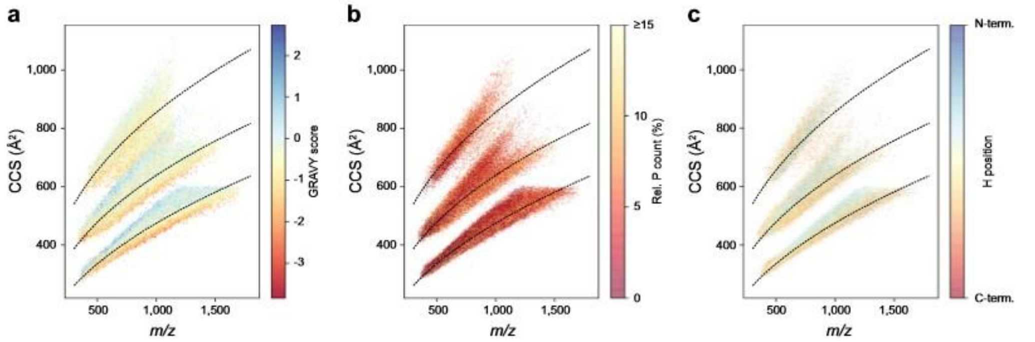


Fig. 3 A global view on peptide cross sections. **a** Mass-to-charge vs. collisional cross section distribution of all peptides in this study colored by the GRAVY hydrophobicity index ($n = 559{,}979$). **b** Subset of peptides with C-terminal arginine or lysine colored by the fraction of prolines in the linear sequence ($n = 452{,}592$). **c** Histidine-containing peptides of (**b**) colored by the relative position of histidine ($n = 171{,}429$). Trend lines (dashed) are fitted to the overall peptide distribution to visualize the correlation of ion mass and mobility in each charge state.

extended helical rather than more compact globular structures. In contrast, peptides with a high fraction of amino acids favoring turn structures (G, S, D, N, and P) tended to more compact conformations. Note, however, that these are subtle, population-wide effects. An interesting result was that peptides with <10% of the mostly non-polar amino acids V, I, F, T, and Y (favoring sheet structures in proteins) formed a narrow band of compact gas-phase conformations.

Such tendencies have been ascribed to intra-molecular interactions such as coulombic repulsion, charge solvation and hydrogen bonding[47–51]. We reasoned that the hydrophobicity of peptides could thus be a good indicator of these interactions in a global view. Indeed, the GRAVY score[52], a commonly used index of hydrophobicity, highlighted distinct areas of the *m/z* vs. ion mobility space and within the CCS value distributions of each charge state, the peptides below the trend line had lower GRAVY scores than those above (Fig. 3a). The two major subgroups of the triply charged peptides also followed this trend in that hydrophobic peptides had a higher propensity to be in the upper population and vice versa. Interestingly, and perhaps counter-intuitively, this correlation was less apparent when comparing the relative bulkiness of amino acid residues even though these properties are related (Supplementary Fig. 6). These results are, however, in line with early work in ion mobility, indicating that non-polar amino acids contribute over-proportionately to the peptide CCS value[26,53] and stabilize helices in the absence of solvent[47]. When rotationally averaged, this results in larger, effective cross sections.

To resolve structural trends at the level of individual amino acids, we visualized their relative distribution in the same 2D space. Proline is unique due to its cyclic structure, which results in an inability to donate hydrogen bonds and to disruption of secondary structures in proteins. We found that peptides with more prolines had somewhat smaller CCS values on a global scale (Fig. 3b). In line with the above reasoning, this could be explained by a disruption of extended conformations and preference for globular ones.

A peptide's CCS value is not only determined by its amino acid composition, but also by its amino acid sequence. As a large-scale example of this, we generated complementary peptide sequences with lysine either at the N-terminus (LysN digestion) or at the C-terminus (LysC digestion). As described before[39], the two peptide populations are most distinct in triply charged species (Supplementary Fig. 7). Comparing 43,463 complementary sequences of doubly charged peptides, we found changing CCS values in the range of −5% up to +10% with a slight median shift of about 1% toward higher CCS values for peptides with C-terminal lysine. The 14,388 triply charged species split in two sub-populations, with one maximum at about +1% similar to the doubly charged species and a second maximum at a shift of about +8%. This indicates that for the latter, switching the position of lysine from the C- to the N-terminus destabilizes the extended conformation. Assuming that the LysC peptides have a more extended conformation due to charge repulsion of the terminal charges, this again conforms to the above considerations.

We next investigated such effects in histidine-containing tryptic peptides, by color-coding them by their relative histidine position in the linear sequences (Fig. 3c). Peptides with histidines close to the N-terminus are more likely to adopt an extended conformation and peptides with histidines closer to the C-terminal lysine or arginine are more compact in the gas phase. This again emphasizes that the internal charge distribution and the ability to solvate charges intra-molecularly have a strong influence on peptide CCS values.

Although our analysis revealed interesting general trends and suggested common principles, it is challenging to combine them into robust models that rationalize the trends and determine the CCS value of a given peptide from its linear sequence. More importantly, peptide CCS values do not lend themselves to global ab initio calculations as this is beyond the capabilities of computational chemistry. To that end, we next turned to deep learning.

Deep learning accurately predicts peptide CCS values. To construct an accurate CCS predictor that can incorporate these large-scale peptide measurements, we decided to employ a flexible deep learning model. We set out to define a network architecture that is capable of learning a non-linear mapping function connecting the linear amino acid peptide sequence with associated charge states to the experimentally measured CCS value with the following properties: (i) Exploit the sequential structure of the data where each peptide is encoded as a string of amino acid sequences; (ii) Account for the influence of an amino acid in the context of the entire peptide sequence; and (iii) Process peptide sequences of arbitrary length. An architecture fulfilling those properties is a bi-directional LSTM network on top of the raw sequence followed by a two-layer multilayer perceptron (MLP)

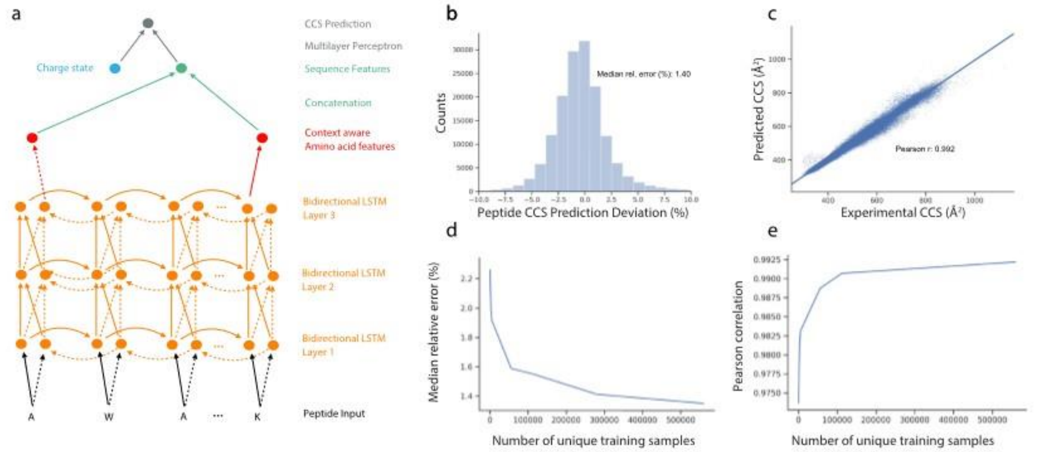


Fig. 4 Deep learning peptide CCS values. a Architecture of the neural network. Bi-directional long short-term model (LSTM): (i) amino acid sequence input, (ii) vectorization of amino acid information for processing, (iii) bi-directional LSTM layers, (iv) reduction to fixed length peptide feature vector by concatenating the last output neurons of both directional LSTMs, and (v) CCS prediction. **b** Relative deviation of predicted CCS values from an independent experimental validation dataset of synthetic peptides from the ProteomeTools project. **c** Correlation of predicted versus experimental CCS values ($n =$ 155,004). **d** Dependence of the median relative error on training dataset size. **e** Same for Pearson correlation coefficient. Source Data are provided as a Source Data file.

(Fig. 4a, Methods). Similar models have already proven successful in proteomics[32,34,35]. The bi-directional LSTM layers enable the model to interpret each amino acid in the context of neighboring amino acids, while the following concatenation layer reduces the resulting N (sequence length) vectors into a single set of 256 features, together encoding the properties of the entirety of the peptide sequence. Together with the charge state, this vector constitutes the input to the MLP module for the final CCS value regression. The entire architecture is implemented with differentiable modules and is end-to-end trainable. We trained our model with the set of 559,979 unique CCS values from our experimental data of the five organisms.

Machine learning models, in particular deep learning models, can easily be over-fitted, resulting in poor generalization performance on new datasets. While holding out samples within the dataset helps, for a more rigorous safeguard, we acquired an independent additional dataset from the synthetic ProteomeTools peptides[54]. This yielded 155,004 unique peptide sequences as an external test set, which was never seen by the model during training. In this test set, our model reached a high accuracy with a 1.4% absolute median deviation and a Pearson correlation coefficient of 0.992 (Fig. 4b, c). For the subset of doubly charged peptides the median absolute deviation was 1.2%, and for triply and quadruply charged species it was 1.8% and 2.0%, respectively (Supplementary Fig. 8). Presumably as a result of an increasing number of accessible conformations, we found that the median absolute deviation increased from 1.2% for CCS values <400 Å^2, to 1.5% for CCS values between 400 and 800 Å^2 ($n = 129{,}710$) and 2.2% for 2580 peptides with CCS values >800 Å^2 (Supplementary Fig. 9). Of all predicted CCS values, 90% were within ±4.0% deviation from the experimental data. In comparison, the experimental median absolute deviation between tryptic peptides from ProteomeTools and endogenous peptides was 0.6% ($r =$ 0.995, $n = 54{,}914$).

In our ProteomeTools data we also found a subset of 7% of the peptide sequences, for which MaxQuant identified at least one secondary feature with a CCS difference >2% relative to the most abundant feature. As we trained our model with CCS values of the latter, it is expected to predict the CCS value of the main conformation in such cases. However, for peptides with a more compact secondary conformation, we observed a bias toward lower CCS values and vice versa (Supplementary Fig. 10). Future prediction models may therefore benefit from considering multiple conformations, in particular for longer peptides and higher charge states.

To independently validate the accuracy of our predictions in a real-world example, we replaced experimental CCS values in a spectral library for DIA, built from the 24 HeLa fractions, with our predictions. We then used the experimental and the predicted libraries individually to re-analyze a triplicate diaPASEF experiment of a whole-proteome HeLa sample[55]. Targeted data analysis in the Spectronaut[56] software makes use of library values to score peptide signals and to restrict the data extraction window in the ion mobility dimension, thereby removing interfering signals from precursors with similar mass and retention time, but different ion mobility. The software automatically performs an alignment of the diaPASEF experiment to the library and optimized the median ion mobility extraction window to 0.07 and 0.09 Vs cm^{-2} for the experimental and predicted library, respectively. The median absolute deviation of peptide ion mobility values were 0.74% and 0.93%. Overall, the experimental and predicted libraries performed very similarly, resulting in 7766 (experimental) and 7685 (predicted) identified protein groups on average (Supplementary Fig. 11).

Given that datasets in hundreds of thousands may still not be seen as large in deep learning, we next investigated the dependency between model accuracy in the test set and training dataset size (Fig. 4d, e). We observed a monotonous improvement in relative prediction accuracy as well as in the Pearson correlation with growing training dataset size. The model error decreased from 1.91% median relative error at 5600 samples to 1.42% for a set of 279,990 training samples, reflecting a substantial decrease in relative error of more than 20%. In contrast, moving from 279,990 samples to the full set of

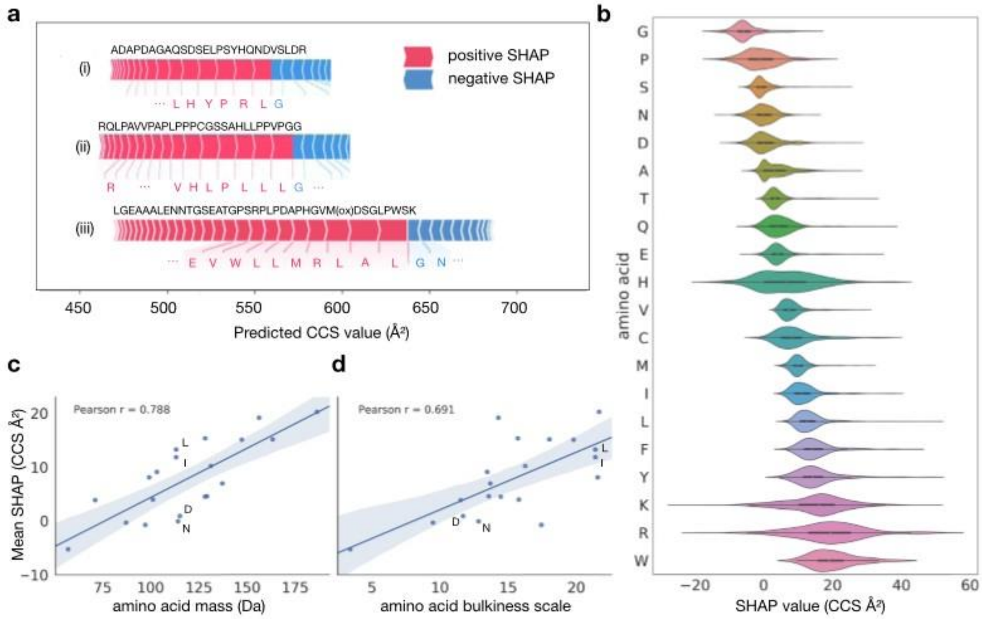


Fig. 5 Explainable artificial intelligence reveals context-dependent amino acid contributions. a Example peptide sequences with SHAP value attributions of the most influential amino acids in the linear sequence. **b** Amino acid-specific SHAP value distributions over the test dataset. Data are presented as violin plots showing kernel density estimates and boxplots with the following elements: median (center), 25th and 75th percentiles (lower and upper box limits), the 1.5× interquartile range (whiskers); $n = 100{,}000$ sequences. **c** Correlation between amino acid mass and mean SHAP value. **d** Correlation between amino acid bulkiness[59] and mean SHAP value.

559,979 samples resulted in a relative improvement of only 1.4% to a median relative error or 1.4%. These diminishing returns in accuracy of prediction indicated that the number of CCS values was sufficient—at least for currently achievable data quality.

Resolving amino acids contributions. Deep learning models are often deemed black boxes, as they are powerful predictors but learned relationships are typically hard to interpret. To make our model interpretable in relation to our experimental findings and to extract further molecular insights we calculated Shapley Additive Explanation (SHAP)[57,58] values for each amino acid in each sequence. In this case, SHAP values indicate the influence of a specific amino acid on the peptide CCS value by comparing it to reference values determined by randomly sampling sequences. This allowed us to interpret the CCS prediction for a peptide sequence by determining the individual, contextual attribution of each amino acid (Methods).

Figure 5a illustrates our analysis of sequence-specific amino acid SHAP values for three representative peptide sequences. In the regular tryptic peptide sequence (i), arginine and leucine with long side-chains shifted the prediction value to larger CCS as compared with a random sequence, while the smaller glycine contributed less than average. In the atypical peptide sequence (ii), the attribution of leucine was similar, however, the attribution of arginine was largely reduced in the N-terminal position. The context-dependent attribution of each amino acid was also evident from the long peptide sequence (iii), indicating a relatively large contribution of the small amino acid alanine to the prediction value. Interestingly, in this particular sequence, glutamic acid had a positive attribution, whereas asparagine somewhat reduced the prediction value, despite the fact that both are similar in size and mass.

Plotting the aggregated SHAP value distribution over the entire test dataset for each individual amino acid, showed the expected relative order in terms of their average contribution (Fig. 5b): light and small amino acids such as glycine and proline had smaller SHAP values, whereas large and bulky amino acids such as tryptophan, arginine and lysine had larger attributions on average. In line with this observation, the average SHAP values correlated well with the amino acid mass and bulkiness[59], as indicated by Pearson correlation coefficients of 0.79 and 0.69, respectively (Fig. 5c, d). Deviations from these correlations, for example, for asparagine, aspartic acid, leucine, and isoleucine, which all have similar mass, could be explained by differences in their bulkiness and hydrophobicity, in line with our experimental results above. Collectively, these results highlight that our deep learning model learned plausible features, extracting related physical quantities on the level of individual amino acids automatically from the training data, even though we solely used the linear peptide sequence as an input.

Beyond the average values, the contribution of individual amino acids to a CCS prediction had vastly different values depending on their position in a sequence (Fig. 5b). Whereas the contributions of glycine, serine, glutamic acid, and methionine were quite constant, those of lysine, arginine, and histidine nearly varied over the entire range of observed SHAP values. In particular for histidine, this agrees with our empirical observation that the position in the linear sequence had a distinct effect on the cross section (Fig. 3c). We thus conclude that our model resolves substantial structural effects for some of the amino acids within each sequence to provide a very accurate CCS estimate for the entire peptide.

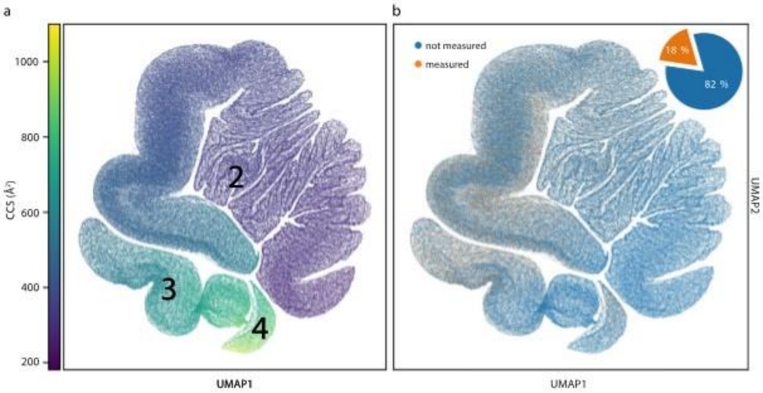


Fig. 6 The human peptide CCS universe. a Two-dimensional UMAP representation of 616,948 unique tryptic peptide sequences colored by their predicted CCS value. **b** Same UMAP plot. Peptide sequences with experimental values in this study are highlighted in orange (18%).

Human whole-proteome level CCS prediction. The human proteome gives rise to 616,948 unique tryptic peptide sequences (considering a minimum length of 7 amino acids and no missed cleavages), of which we measured about 18% in the course of this study. To investigate the entire peptide universe and to create a reference database of all tryptic peptides in the human organism, we next used our trained deep learning model to predict CCS values for the remaining 82%. Given the importance of charge in ion mobility and the fact that it does not follow from the linear sequence in a trivial manner, we first trained a second deep learning model on our experimental training data to also predict the charge state (Methods). We then fed each human peptide sequence together with its predicted charge state into the trained CCS model, resulting in a virtually complete compendium of human peptide CCS values (Supplementary Data 1).

To provide a bird's-eye view of the structure of these data, we visualized the data manifold learned in the last layer of the neural network, in which each sequence is described by a vector of 256 neural network features. These features represent all information relevant to the prediction and were used to regress the final CCS values. However, the data manifold is too high dimensional to be directly accessible to human interpretation, hence we used a non-linear dimension reduction algorithm (Uniform Manifold Approximation and Projection, UMAP[60]) to visualize the data in a 2D space. In this view, each point represents a single peptide sequence and each local structure represents classes of peptides with similar features. Distances in this space can be interpreted as similarities between sequences in terms of the features extracted by the network, meaning that sequences with similar gas-phase properties are close to each other. Figure 6a reveals that the neural network organized the data in three connected manifolds, in which the sequences are ordered in terms of their associated CCS value, starting with low CCS values (<300 Å^2) in the first cluster and increasing to high values (>900 Å^2) in the third cluster. Similar to the representation in *m/z* vs. CCS space, we found that the main clusters were directly associated with the charge state and, within each charge state, there were apparent local structures.

Importantly, our experimental CCS values are distributed across the entire predicted peptide universe (orange and blue points in Fig. 6b), with very high densities in the CCS regions 400−800 Å^2, and lower densities in the region below 300 Å^2. This reassures that the depth of our experimental dataset was sufficient to sample the full feature space, and therefore suggests that our model can be applied to predict CCS values of any tryptic peptide sequence with similar high accuracy.

Discussion

Technological advances have rekindled the interest in IMS, which is now about to become mainstream in proteomic laboratories. Differential ion mobility spectrometers act as filters, only allowing selected ions to enter the mass spectrometer. In contrast, TIMS allows to measure ion mobility values and to derive CCS values that reflect an ion's size and shape. To investigate the benefit of this additional information in proteomics and making use of the speed and sensitivity of PASEF, we measured over two million CCS values of about 500,000 unique peptide sequences from five biological species. This covers a substantial proportion of the peptide space and is by far the most comprehensive dataset of CCS values to date.

This scale allowed us to first assess the analytical benefits of CCS values, which turn out to correspond to a roughly ten-fold increase in separation power. We further established that at an accuracy of 1%, the number of possible precursors of a peptide in a proteomics experiment decreases about two- to three-fold. Such an accuracy can be achieved with a simple linear re-calibration across distant measurements and different instruments. With this re-calibration, CCS values essentially become intrinsic properties of a molecule—meaning they do not depend on external circumstances—similar to their molecular weights, and unlike their retention times. In this regard, we note ongoing research on minimizing ion heating effects in TIMS measurements, as this may also influence the observed cross section or result in fragmentation before MS/MS, depending on instrument settings and space-charge effects[61–64]. However, results presented here and in other studies[15,22,65,66] indicate that $^{\mathrm{TIMS}}$CCS values are generally in excellent agreement with the current gold-standard drift tube ion mobility.

The scale and uniformity of our dataset makes it a valuable resource to investigate fundamentals of peptide gas-phase structures in detail. Beyond the well-known correlation of CCS values with peptide mass, they also correlated with physicochemical amino acid properties such as hydrophobicity, while the contribution of certain amino acids varied based on their position in the sequence. While this scale allowed us to compare a multitude of different peptide sequences, a limitation of our analysis is that we considered only one CCS per peptide and charge state for simplicity. However, ions from a single peptide may occur in multiple gas-phase conformations that can be resolved by IMS[50]. Even more information could thus be derived by resolving the ion-mobility fine structure, for example, of higher charge states[51] or proline-containing peptides[67]. As peptide CCS values in the gas phase are fully determined by their linear amino acid sequences, we reasoned that they should also be predictable with high accuracy. Indeed, after training our state-of-the-art deep learning model on our extensive dataset, it achieved a median accuracy of about 1% for independently measured synthetic peptides, close to the experimental uncertainty. Our model generalized very well to the extent that it accurately predicted CCS values even for unseen peptides, such as those from the 'missing genes' subset in ProteomeTools[54]. Adding even more data values would have diminishing returns, however, prediction accuracy could be further improved with even more consistent measurements and higher ion mobility resolution or by considering multiple conformations. To obtain a sufficient number of CCS values for deep learning, we trained and validated our model with complex samples of proteolytic digests and pooled synthetic peptides. In the future, this work could be complemented with manual investigation of isolated peptides, for example, to study mobility peak shapes and multiple conformations in more detail and independent of MS feature detection algorithms or other factors.

We also interrogated our deep learning model with regard to the determinants of its predictions with Shapley Additive Explanation (SHAP). Amino acids greatly differ in the extent to which their CCS contribution depends on their sequence context —ranging from almost none to a rather wide positive or negative contribution compared to an average amino acid. This highlights how our model, indeed, learned underlying principles. These could readily be extended to other peptide classes, such as modified[68] or cross-linked[69] peptides, using transfer learning[70], with little additional experimental effort.

Our study complements recent efforts in predicting properties of peptides on the basis of their sequences alone, especially those using deep learning for retention times and MS/MS spectra intensities[32,34,35]. Taken together, almost any peptide property relevant to proteomics workflows can now be predicted accurately, even in an ion mobility setup. Conceptually, this allows the community to nearly fully reconstruct the expected experimental values of a MS-based proteomics experiment, given a list of identified and quantified peptides. In more narrow terms, there is great potential to render time- and cost-intensive experimental libraries largely dispensable as exemplified here for diaPASEF. The CCS model presented here further extends the capabilities of such strategies to make full use of the ion mobility dimension. Similarly, predicted CCS values open up the possibility to reuse comprehensive community libraries such as the Pan Human library[71] for ion mobility-enhanced DIA or targeted workflows. We further envision that the combination of predicted CCS, retention time, and MS/MS spectra may improve scoring in database searches and narrow down the list of candidates. This is especially important in challenging applications such as peptidomics or proteomics of microbiomes[34] that have a very large search space. To foster its application and further developments, we make the source code available for training and predictions, in addition to the ready-to-use predictions of the human peptide universe included here.

Methods

Sample preparation. The human HeLa cell line (S3, ATCC), *C. elegans* (N2 wild-type), *D. melanogaster* (CantonS), *E. coli* (XL1 Blue), and *Saccharomyces cerevisiae* (BY4741) were cultivated following standard protocols. All animal experiments were performed in compliance with the institutional regulations of the Max Planck Institute of Biochemistry and the government agencies of Upper Bavaria. Whole organisms were first grinded in liquid nitrogen and cell pellets were directly suspended in lysis buffer with chloroacetamide (PreOmics, Germany) to simultaneously lyse cells, reduce protein disulfide bonds, and alkylate cysteine side chains[72]. The samples were boiled at 95 °C for 10 min and subsequently sonicated at maximum power (Bioruptor, Diagenode, Belgium). Protelytic digestion was performed overnight at 37 °C by adding either (i) equal amounts of LysC and trypsin, (ii) LysC, or (iii) LysN in a 1:100 enzyme:protein (wt/wt) ratio. The resulting peptides were de-salted and purified via solid-phase extraction on styrenedivinylbenzene reversed-phase sulfonate (SDB-RPS) sorbent according to our 'in-StageTip' protocol (PreOmics). The dried eluates were reconstituted in water with 2% acetonitrile (ACN) and 0.1% trifluoroacetic acid (TFA) for further analysis. The synthetic ProteomeTools[54] peptides were reconstituted in the same buffer. To make the data comparable and reusable, we spiked iRT standards (Biognosys) into all samples.

High-pH reversed-phase fractionation. Peptide fractionation was performed at pH 10 on an EASY-nLC 1000 (Thermo Fisher Scientific, Germany) using a 30 cm × 250 µm C_{18} reversed-phase column (PreOmics). Approximately 50 µg of peptides were separated at a flow rate of 2 µL min^{-1} with a binary gradient starting from 3% B, which was linearly increased to 30% B within 45 min, to 60% B within 17 min, and to 95% B within 5 min before re-equilibration. Fractions were collected into 24 wells by switching the rotor valve of an automated concatenation system[73] (Spider fractionator, PreOmics) in 90 s intervals. Peptide fractions were vacuum-centrifuged to dryness and reconstituted in water with 2% ACN and 0.1% TFA.

Liquid chromatography and mass spectrometry. LC–MS was performed on an EASY-nLC 1200 (Thermo Fisher Scientific) system coupled online to a hybrid TIMS-quadrupole TOF mass spectrometer[10] (Bruker Daltonik timsTOF Pro, Germany) via a nano-electrospray ion source (Bruker Daltonik Captive Spray). Approximately 200 ng of peptides were separated on an in-house 45 cm × 75 µm reversed-phase column at a flow rate of 300 nL min^{-1} in an oven compartment heated to 60 °C. The column was packed in-house with 1.9 µm C_{18} beads (Dr. Maisch Reprosil-Pur AQ, Germany) up to the laser-pulled electrospray emitter tip. Mobile phases A and B were water and 80%/20% ACN/water (v/v), respectively, and both buffered with 0.1% formic acid (v/v). To analyze fractionated peptides from whole-proteome digests, we used a gradient starting with a linear increase from 5% B to 30% B over 95 min, followed by further linear increases to 60% B and finally to 95% B in 5 min each, which was held constant for 5 min before returning to 5% in 5 min and re-equilibration for 5 min. The pooled synthetic peptides were analyzed with a gradient starting from 5% B to 30% B in 35 min, followed by linear increases to 60% B and 95% in 2.5 min each before washing and re-equilibration for a total of 5 min.

The mass spectrometer was operated in data-dependent PASEF[13] mode with 1 survey TIMS-MS and 10 PASEF MS/MS scans per acquisition cycle. We analyzed an ion mobility range from $1/K_0 = 1.51$ to 0.6 Vs cm^{-2} using equal ion accumulation and ramp time in the dual TIMS analyzer of 100 ms each. Suitable precursor ions for MS/MS analysis were isolated in a window of 2 Th for $m/z < 700$ and 3 Th for $m/z > 700$ by rapidly switching the quadrupole position in sync with the elution of precursors from the TIMS device. The collision energy was lowered stepwise as a function of increasing ion mobility, starting from 52 eV for 0–19% of the TIMS ramp time, 47 eV for 19–38%, 42 eV for 38–57%, 37 eV for 57–76%, and 32 eV until the end. We made use of the m/z and ion mobility information to exclude singly charged precursor ions with a polygon filter mask and further used 'dynamic exclusion' to avoid re-sequencing of precursors that reached a 'target value' of 20,000 a.u. The ion mobility dimension was calibrated linearly using three ions from the Agilent ESI LC/MS tuning mix (m/z, $1/K_0$: 622.0289, 0.9848 Vs cm^{-2}; 922.0097, 1.1895 Vs cm^{-2}; and 1221.9906, 1.3820 Vs cm^{-2}). All experimental parameters with relevance to the measurement of CCS values are summarized in Supplementary Table 1.

Data processing. MS raw files were analyzed with MaxQuant[36,37] version 1.6.5.0, which extracts 4D isotope patterns ('features') and associated MS/MS spectra. The built-in search engine Andromeda[74] was used to match observed fragment ions to theoretical peptide fragment ion masses derived from in silico digests of a reference proteome and a list of 245 potential contaminants using the appropriate digestion rules for each proteolytic enzyme (trypsin, LysC or LysN). We allowed a maximum of two missing values and required a minimum sequence length of 7 amino acids while limiting the maximum peptide mass to 4600 Da. Carbamidomethylation of cysteine was defined as a fixed modification, and oxidation of methionine and acetylation of protein N-termini were included in the search as variable modifications. Reference proteomes for each organism including isoforms were accessed from UniProt (*Homo sapiens*: 91,618 entries, 2019/05; *E. coli*: 4403 entries, 2019/01; *C. elegans*: 28,403 entries, 2019/01; *S. cerevisiae*: 6049 entries, 2019/01; *D. melanogaster*: 23,304 entries, 2019/01). The synthetic peptide library (ProteomeTools[54]) was searched against the entire human reference proteome. The maximum mass tolerances were set to 20 and 40 ppm for precursor and fragment ions, respectively. False discovery rates were controlled at 1% on both the peptide spectrum match and protein level with a target-decoy approach. The analyses were performed separately for each organism and each set of synthetic peptides ('proteotypic set', 'SRM atlas', and 'missing gene set'). To demonstrate the utility of CCS prediction, we re-analyzed three diaPASEF experiments from Meier et al.[55] with Spectronaut 14.7.201007.47784 (Biognosys AG), replacing experimental ion mobility values in the spectral library with our predictions. Singly charged peptide precursors were excluded from this analysis as the neural network was exclusively trained with multiply charged peptides.

Bioinformatic analysis. Bioinformatic analysis of the MaxQuant output files and data visualization was performed with Python version 3.6 employing the following packages: NumPy, pandas, SciPy[75], Biopython[76], Matplotlib, and Seaborn. Decoy database hits were excluded from the analysis as well as peptide features assigned with zero intensity values. Peptides can adopt multiple conformations, both in the liquid and in the gas phase. For simplification, we here selected only the most abundant feature for each modified peptide sequence and charge state per LC-TIMS-MS run. To account for experimental drifts in the measurements of $^{\mathrm{TIMS}}$CCS values over time, we performed a hierarchical clustering (similar to[37]) and aligned all experiments by calculating pair-wise linear offsets ($y = x + b$) going from the closest to the most distant nodes. Multiple measurements of the same modified peptide and charge state in different LC-MS experiments were merged to one unique CCS value by calculating the mean. To perform nearest neighbor analysis in the m/z vs. CCS space, we represented the data in a Kd-tree structure using the Chebyshev distance metric to define a rectangular area with a given mass and CCS tolerance surrounding a node of interest.

Deep learning model for CCS prediction. The deep learning model takes a raw (modified) peptide sequence as input. First, each amino acid gets one-hot encoded into a 26-dimensional vector representation for processing. This one-hot encoding also is applied to the elements '(ox)' and '(ac)', resulting in a total feature vector with dimension $L \times 26$ with L being the length of a given peptide. This vector is connected to a two-layer bi-directional recurrent network with LSTM[77] units with 500 hidden nodes each, which extract context-based features for each individual amino acid. This feature embedding gets further reduced to a global 256-dimensional peptide feature vector by concatenating the last output neurons of both the LSTM networks aggregating from left or right over the sequence. This peptide feature vector is further concatenated with additional charge state of the sequence and then is fed to a logistic regression layer which regresses the expected CCS value for the sequence. The most significant hyperparameters, namely: number of hidden neurons, number of layers were chosen by running a small search in a first preliminary step on a validation set consisting of 10% of the training data. The combination of recurrent layers with the concatenation step allows the model architecture to process peptide sequences with arbitrary lengths. The final model is end-to-end optimized by an ADAM Optimizer on 559,979 unique CCS values (modified peptide sequence and charge state) and validated on 155,004 holdout peptides from the synthetic ProteomeTools library. The full framework is implemented using Python with TensorFlow[78] as the autograd library, enabling the neural network optimization. On an i7-4930K CPU machine equipped with an NVIDIA Geforce 1080 our model was trained within 8 h and the prediction of single peptide CCS values takes approximately 1 ms.

Deep learning model for peptide charge state prediction. To predict the most probable (most abundant) charge state from the linear peptide sequence, we built a charge prediction neural network which has the identical structure as our CCS prediction model. It takes the raw peptide sequence as input following the same one-hot encoding procedure and predicts a single associated charge value. We trained the charge prediction model on the same 559,979 unique training values and validated it on the holdout set of 155,004 peptides from ProteomeTools. The charge prediction model reaches a final accuracy of 93.5% for predicting the three observed charge states 2, 3, and 4.

Analysis of amino acid feature attribution of the learnt network. For a given sequence and its CCS prediction, every amino acid is associated with a SHAP value[57,58]. This SHAP value quantifies how the presence of the amino acid influences the final prediction. By the summation-to-delta property, the SHAP values are constrained in a way such that the sum of all SHAP values in a sequence results in the final CCS prediction. SHAP values are a unification of multiple existing approaches[79–83] for explaining predictions by feature attribution. For interpreting the predictions of our model we use the DeepExplainer from the official SHAP implementation (https://github.com/slundberg/shap). The DeepExplainer approximates SHAP values and is based on DeepLift[84]. Here the importance of individual features is approximated by comparing the model output for an input that contains the specific feature value to the model output where the feature is set to a reference value. A crucial step for this approach is to define the reference values. In our case, the inputs are sequences of one-hot-encoded amino acids and we use 128 randomly chosen background sequences from the dataset in order to define meaningful reference values for all neurons. In order to capture non-linearities, the DeepLift approach approximates feature attributions for every neuron in the model. It starts at the output layer and propagates the values to the input by backpropagation, which is called applying the chain rule for multipliers in the original publication[81]. Applying this approach to the input sequences in our CCS model we are able to capture the SHAP value for an individual amino acid in a peptide sequence.

Visualization of learnt network representation of the human proteome. To visualize the 256-dimensional neural network feature space, we apply the UMAP[60] algorithm, which is a dimension reduction technique for general non-linear dimension reduction and it assumes uniform distribution of the data on a Riemannian manifold. Under certain conditions this manifold can be modeled with a fuzzy topological structure. The 2D embedding, which is used for visualization is found by searching for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure. Therefore, pairwise similarities between peptide sequences in the high-dimensional NN space approximately resemble positions in the low-dimensional embedding visualization.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The MS raw files and associated MaxQuant output files generated and analyzed throughout this study have been deposited at the ProteomeXchange Consortium via the PRIDE partner repository[85] with the dataset identifier PXD019086. The previously acquired HeLa data is available through the dataset identifier PXD010012. The diaPASEF raw files are available through the dataset identifier PXD017703. *H. sapiens* (taxon identifier: 9606), *S. cerevisiae* (taxon identifier: 559292), *D. melanogaster* (taxon identifier: 7227), *E. coli* (taxon identifier: 83333) and *C. elegans* (taxon identifier: 6239) proteome databases were downloaded from UniProt [https://www.uniprot.org]. Source data are provided with this paper.

Code availability

The source code of our deep learning model and data analysis scripts are available on GitHub (https://github.com/theislab/DeepCollisionalCrossSection and https://github.com/mannlabs/DeepCollisionalCrossSection).

Received: 18 May 2020; Accepted: 22 January 2021;
Published online: 19 February 2021

References

1. McLean, J. A., Ruotolo, B. T., Gillig, K. J. & Russell, D. H. Ion mobility–mass spectrometry: a new paradigm for proteomics. *Int. J. Mass Spectrom.* **240**, 301–315 (2005).
2. Baker, E. S. et al. An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies. *J. Proteome Res.* **9**, 997–1006 (2010).
3. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H. Ion mobility-mass spectrometry. *J. Mass Spectrom.* **43**, 1–22 (2008).
4. Distler, U. et al. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **11**, 167–170 (2014).
5. Helm, D. et al. Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell. Proteom.* **13**, 3709–3715 (2014).
6. Pfammatter, S. et al. A novel differential ion mobility device expands the depth of proteome coverage and the sensitivity of multiplex proteomic measurements. *Mol. Cell. Proteom.* **17**, 2051–2067 (2018).
7. Hebert, A. S. et al. Comprehensive single-shot proteomics with FAIMS on a hybrid orbitrap mass spectrometer. *Anal. Chem.* **90**, 9529–9537 (2018).
8. Bekker-Jensen, D. B. et al. A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Mol. Cell. Proteom.* **19**, 716–729 (2020).
9. Yu, Q. et al. Benchmarking the orbitrap tribrid eclipse for next generation multiplexed proteomics.*Anal. Chem. Anal. Chem.* **92**, 6478–6485 (2020).
10. Meier, F. et al. Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteom.* **17**, 2534–2545 (2018).
11. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion. Mobil. Spectrom.* **14**, 93–98 (2011).
12. Fernandez-Lima, F. A., Kaplan, D. A. & Park, M. A. Note: integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **82**, 126106 (2011).
13. Meier, F. et al. Parallel accumulation–serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14**, 5378–5387 (2015).

14. Ridgeway, M. E., Lubeck, M., Jordens, J., Mann, M. & Park, M. A. Trapped ion mobility spectrometry: a short review. *Int. J. Mass Spectrom.* **425**, 22–35 (2018).
15. Vasilopoulou, C. G. et al. Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. *Nat. Commun.* **11**, 331 (2020).
16. Valentine, S. J., Counterman, A. E. & Clemmer, D. E. A database of 660 peptide ion cross sections: use of intrinsic size parameters for bona fide predictions of cross sections. *J. Am. Soc. Mass Spectrom.* **10**, 1188–1211 (1999).
17. Tao, L., McLean, J. R., McLean, J. A. & Russell, D. H. A collision cross-section database of singly-charged peptide ions. *J. Am. Soc. Mass Spectrom.* **18**, 1232–1238 (2007).
18. May, J. C., Morris, C. B. & McLean, J. A. Ion mobility collision cross section compendium. *Anal. Chem.* **89**, 1032–1044 (2017).
19. Michelmann, K., Silveira, J. A., Ridgeway, M. E. & Park, M. A. Fundamentals of trapped ion mobility spectrometry. *J. Am. Soc. Mass Spectrom.* **26**, 14–24 (2014).
20. Silveira, J. A., Michelmann, K., Ridgeway, M. E. & Park, M. A. Fundamentals of trapped ion mobility spectrometry part II: fluid dynamics. *J. Am. Soc. Mass Spectrom.* **27**, 585–595 (2016).
21. Mason, E. A. & McDaniel, E. W. *Transport Properties of Ions in Gases* (John Wiley & Sons, Inc., 1988).
22. Gabelica, V. et al. Recommendations for reporting ion mobility mass spectrometry measurements. *Mass Spectrom. Rev.* **38**, 291–320 (2019).
23. May, J. C. et al. Conformational ordering of biomolecules in the gas phase: nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer. *Anal. Chem.* **86**, 2107–2116 (2014).
24. Wu, C., Siems, W. F., Klasmeier, J. & Hill, H. H. Separation of isomeric peptides using electrospray ionization/high-resolution ion mobility spectrometry. *Anal. Chem.* **72**, 391–395 (2000).
25. Srebalus Barnes, C. A., Hilderbrand, A. E., Valentine, S. J. & Clemmer, D. E. Resolving isomeric peptide mixtures: a combined HPLC/ion mobility-TOFMS analysis of a 4000-component combinatorial library. *Anal. Chem.* **74**, 26–36 (2002).
26. Shvartsburg, A. A., Siu, K. W. M. & Clemmer, D. E. Prediction of peptide ion mobilities via a priori calculations from intrinsic size parameters of amino acid residues. *J. Am. Soc. Mass Spectrom.* **12**, 885–888 (2001).
27. Wang, B., Valentine, S., Plasencia, M., Raghuraman, S. & Zhang, X. Artificial neural networks for the prediction of peptide drift time in ion mobility mass spectrometry. *BMC Bioinformatics* **11**, 182 (2010).
28. Shah, A. R. et al. Machine learning based prediction for peptide drift times in ion mobility spectrometry. *Bionformatics* **26**, 1601–1607 (2010).
29. Wang, B. et al. Prediction of peptide drift time in ion mobility mass spectrometry from sequence-based features. *BMC Bioinformatics* **14**, S9 (2013).
30. Zou, J. et al. A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
31. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
32. Zhou, X. X. et al. PDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
33. Ma, C. et al. Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal. Chem.* **90**, 10881–10888 (2018).
34. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
35. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
36. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
37. Prianichnikov, N. et al. MaxQuant software for ion mobility enhanced shotgun proteomics.*Mol. Cell. Proteomics* **19**, 1058–1069 (2020).
38. Valentine, S. J., Counterman, A. E., Hoaglund, C. S., Reilly, J. P. & Clemmer, D. E. Gas-phase separations of protease digests. *J. Am. Soc. Mass Spectrom.* **9**, 1213–1216 (1998).
39. Lietz, C. B., Yu, Q. & Li, L. Large-scale collision cross-section profiling on a traveling wave ion mobility mass spectrometer. *J. Am. Soc. Mass Spectrom.* **25**, 2009–2019 (2014).
40. Taraszka, J. A., Counterman, A. E. & Clemmer, D. E. Gas-phase separations of complex tryptic peptide mixtures. *Fresenius. J. Anal. Chem.* **369**, 234–245 (2001).
41. Ludwig, C. et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).
42. Bush, M. F., Campuzano, I. D. G. & Robinson, C. V. Ion mobility mass spectrometry of peptide ions: effects of drift gas and calibration strategies. *Anal. Chem.* **84**, 7124–7130 (2012).
43. Stow, S. M. et al. An interlaboratory evaluation of drift tube ion mobility-mass spectrometry collision cross section measurements. *Anal. Chem.* **89**, 9048–9055 (2017).
44. Picache, J. A. et al. Collision cross section compendium to annotate and predict multi-omic compound identities. *Chem. Sci.* **10**, 983–993 (2019).
45. Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* **17**, 4277–4285 (1978).
46. Jarrold, M. F. Peptides and proteins in the vapor phase. *Annu. Rev. Phys. Chem.* **51**, 179–207 (2000).
47. Jarrold, M. F. Helices and sheets in vacuo. *Phys. Chem. Chem. Phys.* **9**, 1659 (2007).
48. Wyttenbach, T., Pierson, N. A., Clemmer, D. E. & Bowers, M. T. Ion mobility analysis of molecular dynamics. *Annu. Rev. Phys. Chem.* **65**, 175–196 (2014).
49. McLean, J. R. et al. Factors that influence helical preferences for singly charged gas-phase peptide ions: the effects of multiple potential charge-carrying sites. *J. Phys. Chem. B* **114**, 809–816 (2010).
50. Pierson, N. A., Chen, L., Valentine, S. J., Russell, D. H. & Clemmer, D. E. Number of solution states of bradykinin from ion mobility and mass spectrometry measurements. *J. Am. Chem. Soc.* **133**, 13810–13813 (2011).
51. Xiao, C., Pérez, L.M. & Russell, D.H. Effects of charge states, charge sites and side chain interactions on conformational preferences of a series of model peptide ions.*Analyst* **140**, 6933–6944 (2015).
52. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
53. Valentine, S. J., Counterman, A. E., Hoaglund-Hyzer, C. S. & Clemmer, D. E. Intrinsic amino acid size parameters from a series of 113 lysine-terminated tryptic digest peptide ions. *J. Phys. Chem. B* **103**, 1203–1207 (1999).
54. Zolg, D. P. et al. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).
55. Meier, F. et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
56. Bruderer, R. et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteom.* **14**, 1400–1410 (2015).
57. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
58. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
59. Zimmerman, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201 (1968).
60. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
61. Morsa, D. et al. Effective temperature and structural rearrangement in trapped ion mobility spectrometry. *Anal. Chem.* **92**, 4573–4582 (2020).
62. Bleiholder, C., Liu, F.C. & Chai, M. Comment on effective temperature and structural rearrangement in trapped ion mobility spectrometry: TIMS enables native mass spectrometry applications.*Anal. Chem.* **92**, 16329–16333 (2020).
63. Naylor, C. N., Ridgeway, M. E., Park, M. A. & Clowers, B. H. Evaluation of trapped ion mobility spectrometry source conditions using benzylammonium thermometer ions. *J. Am. Soc. Mass Spectrom.* **31**, 1593–1602 (2020).
64. Yu, F. et al. Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant. *Mol. Cell. Proteom.* **19**, 1575–1585 (2020).
65. Silveira, J. A., Ridgeway, M. E. & Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* **86**, 5624–5627 (2014).
66. Hernandez, D. R. et al. Ion dynamics in a trapped ion mobility spectrometer. *Analyst* **139**, 1913–1921 (2014).
67. Counterman, A. E. & Clemmer, D. E. Cis–trans signatures of proline-containing tryptic peptides in the gas phase. *Anal. Chem.* **74**, 1946–1951 (2002).
68. Glover, M. S. et al. Examining the influence of phosphorylation on peptide ion structure by ion mobility spectrometry-mass spectrometry. *J. Am. Soc. Mass Spectrom.* **27**, 786–794 (2016).
69. Steigenberger, B. et al. Benefits of collisional cross section assisted precursor selection (caps-PASEF) for cross-linking mass spectrometry. *Mol. Cell. Proteom.* **19**, 1677–1687 (2020).
70. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9 (2016).
71. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031 (2014).
72. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

73. Kulak, N. A., Geyer, P.E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
74. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
75. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
76. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
77. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
78. Abadi, M. *et al.* TensorFlow: large-scale machine learning on heterogeneous distributed systems. *OSDI'16: Proc. 12th USENIX Conf. Operating Systems Design and Implementation* 265–283 (USENIX, 2016).
79. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).
80. Datta, A., Sen, S. & Zick, Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. *IEEE Symp. Security and Privacy (SP)* 598–617 (IEEE, 2016). https://doi.org/10.1109/SP.2016.42.
81. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).
82. Lipovetsky, S. & Conklin, M. Analysis of regression in game theory approach. *Appl. Stoch. Model. Bus. Ind.* **17**, 319–330 (2001).
83. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the predictions of any classifier. *Proc. 2016 Conf. North American Chapter of the Association for Computational Linguistics: Demonstrations* 97–101 (ACL, 2016).
84. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *ICML'17: Proc. 34th Int. Conf. Machine Learning* (eds. Precup, D. & Whye Teh, Y.) Vol. 70, 3145–3153 (ACM, 2017).
85. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

Acknowledgements

We thank our colleagues in the department of Proteomics and Signal Transduction for help and discussions, and in particular I. Paron, B. Splettstößer, J. Müller, and A. Strasser for technical assistance. We acknowledge the ProteomeTools project led by B. Küster for providing a clone of the synthetic peptide library. This work was partially supported by the German Research Foundation (DFG-Gottfried Wilhelm Leibniz Prize granted to M.M., grant# MA 1764/2-1) and by the Max-Planck Society for the Advancement of Science. F.J.T. acknowledges support by the BMBF (grant #L031L0214A, grant# 01IS18036B and grant# 01IS18053A) and by the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI (grant number: ZT-I-PF-5-01) and sparse2big (grant number ZT-I-007).

Author contributions

F.M., A.B., and M.M. designed the proteomics experiments. F.M. and A.B. performed the experiments. F.M., A.B., and M.M. analyzed the data and interpreted the results. E.V. and M.T.S. contributed to the data analysis. N.D.K., with contributions from F.J.T., designed and developed the deep learning model as well as the prediction interpretation and visualization pipeline. J.M.W. performed neural network training runs and supported N.D.K. in integrating the feature attribution functionality. F.M, N.D.K., F.J.T., and M.M. wrote the manuscript. F.J.T. and M.M. supervised the project.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc. and Dermagnostix. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21352-8.

Correspondence and requests for materials should be addressed to F.J.T. or M.M.

Peer review information *Nature Communications* thanks Aivett Bilbao, Zheng-Jiang Zhu and the other, anonymous, reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.3. True single-cell proteomics on a TIMS-qTOF platform

Single-cell technologies are revolutionizing biology but were until now mainly limited to imaging and deep sequencing. Since proteins are the main drivers of cellular function rather than nucleic acids or metabolites, we reasoned that it would be highly valuable to analyze the proteomes of single cells, complementing the achievements in the single-cell sequencing arena[65,378,419]. Many approaches, including multiplexing and the usage of a booster channel consisting of several hundred cells as described above, promised to enable this[110,408,420]. However, that came at the expense of drastic quantitative distortion and the inflation of false-positive identifications[405]. To pave the way for true unbiased single-cell proteomics at highest qualitative and quantitative reproducibility, we had to address fundamental technological limitations across the bottom-up proteomics workflow.

First, we optimized conditions to capture live single-cells by FACS sorting and coupled it to sample preparation to quantitatively 'freeze' the proteome. We realized this by sorting the cells into a weak-organic reservoir comprising only 1 µL of volume, followed by thermal and mechanical steps to fully dissociate the single-cells in the 384-well format. Furthermore, since single cells comprise on average only 150 pg of protein, we had to keep the digestion kinetics high by titrating digestion enzyme concentration accordingly. This allowed us to process single-cells at a volume of less than 2 µL. Next, we enabled the loss-less transfer of the minute amounts of peptides onto a StageTip, which we realized by repeated solvation of the pellet in pure formic acid. Loading of the StageTip at a centrifugation speed at 600 xg for 1 min was crucial to concentrate the single-cell peptides at the solid phase extraction (SPE) material surface to create a *nanopackage*. Furthermore, the StageTip itself allowed us not only to concentrate the peptides, but also to remove residual salt, which could otherwise result in suppression of the ES signal.

Second, we had to make liquid chromatography extremely sensitive, robust and reproducible, since even smallest imperfections such as of the analytical column or accompanying connectors, could disrupt the transfer of the 150 pg protein digest of a single cell into the mass spectrometer[39]. To do so, we decided to use the *EvoSep One* LC system designed for robustness at microflow conditions over thousands of runs[150]. We reasoned that the system would have two key advantages for single-cell proteomics, one of which is the peptide *nanopackage* principle described above. Before analytical separation of the single-cell material, the peptide *nanopackage* is eluted from the EvoTip into the LC system. Here it stacks up at the head of the analytical column, preventing any peptide dilution and LC-related analyte loss. The second advantage should be that the preformed gradient is pushed out by a single high-pressure pump, thereby preventing the admixture of the elution gradient as can happen in binary pump systems at very low flow rates. We found that this principle enables robust flow rates down to 25 nl/min and results in proportionally increased overall ES efficiency due to its concentration dependency. Establishing this concept enabled us to run hundreds of single-cell proteome analyses at a throughput of more than 40 cells per day and at a 100 nl/min flow rate without any drop in performance. The performance increase was about 10-fold compared to a microflow gradient, as anticipated.

Third, we realized that the sensitivity of our *timsTOF Pro* mass spectrometer - although superb due to virtually noise-free spectra in TIMS and concentrated ion packages - was not sufficient for the analysis of true single cells. Since ionization efficiency, ion transfer efficiency into the vacuum system and ion utilization of the instrument govern MS sensitivity, we teamed up with Bruker Daltonik to construct an *alpha prototype instrument*, including an ion source with at least 4-fold increased brightness and improved ion transfer efficiency throughout the instrument.

Fourth, we developed a novel data independent scan mode called diaPASEF for parallel accumulation-serial fragmentation combined with data-independent acquisition by hijacking the instruments firmware. Common to DIA, this scan mode is especially attractive for the analysis of large sample collections, e.g. hundreds of single-cell proteomes, due to its non-stochastic nature, resulting in high measurement reproducibility and data completeness. However, in DIA, precursor ions are recursively isolated by the quadrupole and concurrently fragmented to generate convoluted fragment ion spectra from many precursors, resulting in a great challenge for subsequent analysis[269,271]. To reduce this spectral complexity, isolation window sizes are often decreased at the expense of reduced ion current sampling down to less than 1%[215]. In contrast, diaPASEF allows ion sampling efficiency of up to 100% when precursor elution from the TIMS ramp is synchronized with the quadrupole isolation window and when scan time in the TIMS tunnel two is equal to the accumulation time in TIMS tunnel one[198]. Note that the addition of ion mobility separation to the chromatographic and mass separation results in a four dimensional data cuboid (taking intensity values into account) containing all fragment ions of all precursors across each run. This should also result in reduced spectral complexity, increased overall sensitivity and enable improved algorithmic scoring due to the introduction of ion mobility as an additional dimension, and the fact that fragmentation ions share the same ion mobility position as their parental ion.

We demonstrated the applicability of diaPASEF to deep proteome measurements in the context of sample saturation and long gradients for high-speed applications and benchmarked its quantitative accuracy. This included optimizing the window placement of diaPASEF to balance selectivity, sensitivity and precursor coverage as a function of chromatographic performance and gradient length.

Taking all these parameters into account, we designed a fast diaPASEF scan mode for single-cell proteomics consisting of one MS1 scan and three subsequent diaPASEF scans covering the m/z range from 400-1,000 m/z at a duty cycle of 12.5 % resulting in a cycle time of 2.5 sec. In downstream processing, we observed that the summation of several subsequently acquired diaPASEF scans increases S/N levels in the 4D data cuboid, resulting in a higher spectral library recovery, especially for very low sample amounts. In my hands, three consecutive diaPASEF scans were optimal, which still allowed efficient sampling of the chromatographic peak.

Together these developments allowed us to sample ions at an estimated 40- to 100-fold increased sensitivity, as highlighted by the robust identification of more than 3,200 protein groups from only 1 ng of HeLa digest compared to only about 800 protein groups from a similar input amount on the previous standard setup. Benchmarking single-cell dilution experiments in DDA mode identified up to 1,000 protein groups per cell at excellent quantitative reproducibility (R = 0.92) and a steady increase to more than 2,000 protein group identifications for six-cell measurements while maintaining the expected dynamic range increase with increasing cell counts. I then applied our true single-cell proteome workflow to the analysis of a cell cycle arrested and released culture of cells, identifying up to 1,400 protein groups per single cell. Furthermore, raw total single-cell peptide signals reflected cell size gain across cell cycle progression. Normalized quantitative data allowed the prediction of cell-cycle states and highlighted known and potentially novel cell-cycle state markers. In contrast to other methods, signals in raw spectra were still readily visualized. Next, we compared our single-cell proteomics data to publicly available single-cell RNA-sequencing data from two different technologies (dropSEQ, smartSEQ) in an attempt to obtain insights into fundamental differences between the proteome and transcriptome levels[370,380]. We found that single-cells have a very high protein expression completeness - in contrast to RNA levels - and that quantitative protein expression correlations between single-cell proteomes are very high. Furthermore, we asked to what degree single-cell RNA sequencing data could serve as a proxy for protein measurements and showed that single-cell proteomes are very different to RNA levels. This finding implies distinct RNA and protein abundance regulation mechanisms, dissection of which is only possible when integrating both layers of information. We also discovered that single cells have a quantitatively and qualitatively stable, and functionally essential core proteome comprising members of the folding machinery, nucleic acid helicases, cellular structure determining proteins and the translation initiation/elongation machineries. This is in stark contrast to the mRNA level, which is qualitatively and quantitatively volatile across single cells because of the very low median number of messages per gene[383,419].

Although these technological developments and findings only mark the beginning of an exciting new area of research, we believe that they are a milestone for the elucidation of single-cell proteomes and their integration with single-cell RNA-sequencing studies. There are many opportunities to increase overall workflow sensitivity and for improving data analysis and modelling tools, similar to the rapid advances in single-cell RNA sequencing technologies over the last ten years. Since our workflow is also compatible with chemical multiplexing, but at much higher sensitivity, it should also be possible to multiplex single-cell measurements without the need for a booster channel, minimizing reporter ion compression. Finally, these developments are not only limited to single-cell proteomics. We imagine that this platform will find its use in many ultra-high sensitivity settings including PTMs from minute sample amounts and very small tissue isolates from FFPE material as described in the next chapter. We also imagine it to be applied to the lipidomics and metabolomics analysis of single cells, which could finally enable their MS-based multiomics investigation in conjunction with single-cell sequencing methods.

3.3.1. Article 6: diaPASEF – PASEF combined with DIA

diaPASEF – parallel accumulation-serial fragmentation combined with data-independent acquisition

Nature Methods, November 30, 2020

Florian Meier[1, 2], **Andreas-David Brunner[1]**, Max Frank[3], Annie Ha[3], Isabell Bludau[1], Eugenia Voytik[1], Stephanie Kaspar-Schoenefeld[4], Markus Lubeck[4], Oliver Raether[4], Nicolai Bache[5], Ruedi Aebersold[6, 7], Ben C. Collins[6, 8, #], Hannes L. Röst[3, #], Matthias Mann[1, 9, #]

# Corresponding author

[1] *Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany*
[2] *Functional Proteomics, Jena University Hospital, Jena, Germany*
[3] *Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada*
[4] *Bruker Daltonik GmbH, Bremen, Germany*
[5] *EvoSep Biosystems, Odense, Denmark*
[6] *Department of Biology, Institute for Systems Biology, ETH Zurich, Zurich, Switzerland*
[7] *Faculty of Science, University of Zurich, Zurich, Switzerland*
[8] *School of Biological Sciences, Queen's University of Belfast, Belfast, UK*
[9] *NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark*

Contribution

I contributed to the conceptualization of the scan mode and its adjustments for high-sensitivity, high-throughput and also to the deep proteome measurements. Furthermore, I performed experiments, benchmarks, method optimization and analysis of the raw data with a focus on tuning the method for ultra-high sensitivity applications.

nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-00998-0

diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition

Florian Meier[1,2], Andreas-David Brunner[1], Max Frank[3], Annie Ha[3], Isabell Bludau[1], Eugenia Voytik[1], Stephanie Kaspar-Schoenefeld[4], Markus Lubeck[4], Oliver Raether[4], Nicolai Bache[5], Ruedi Aebersold[6,7], Ben C. Collins[6,8] ✉, Hannes L. Röst[3] ✉ and Matthias Mann[1,9] ✉

Data-independent acquisition modes isolate and concurrently fragment populations of different precursors by cycling through segments of a predefined precursor *m/z* range. Although these selection windows collectively cover the entire *m/z* range, overall, only a few per cent of all incoming ions are isolated for mass analysis. Here, we make use of the correlation of molecular weight and ion mobility in a trapped ion mobility device (timsTOF Pro) to devise a scan mode that samples up to 100% of the peptide precursor ion current in *m/z* and mobility windows. We extend an established targeted data extraction workflow by inclusion of the ion mobility dimension for both signal extraction and scoring and thereby increase the specificity for precursor identification. Data acquired from whole proteome digests and mixed organism samples demonstrate deep proteome coverage and a high degree of reproducibility as well as quantitative accuracy, even from 10 ng sample amounts.

Mass spectrometry–based proteomics, like other omics technologies, aims for an unbiased, comprehensive and quantitative description of the system under investigation[1–3]. Proteomics workflows have become increasingly successful in the characterization of complex proteomes in great depth[4,5]. Application to large sample cohorts requires a high degree of reproducibility and data completeness, which makes data-independent acquisition (DIA) schemes particularly attractive[6,7]. In contrast to data-dependent acquisition (DDA), in which particular precursors are sequentially selected, in DIA, groups of ions are recursively isolated by the quadrupole and concurrently fragmented to generate convoluted fragment ion spectra composed of fragments from many different precursors[8–10]. Although DIA guarantees that each precursor in a predefined mass range is fragmented once per cycle, spectral complexity poses a great challenge to subsequent analysis[11]. Narrower isolation windows result in less complex spectra, but this increases the total number of windows and hence the DIA cycle times needed to cover the entire mass range[9,12,13]. Moreover, as every precursor is isolated only once per cycle, the ion sampling efficiency at the mass-selective quadrupole for DIA methods is limited to 1–3% with typical schemes of 32 or 64 windows.

The addition of ion mobility separation to the chromatographic and mass separation should increase sensitivity and reduce spectral complexity[14–17]. The trapped ion mobility spectrometer (TIMS) is a particularly compact mobility analyzer in which ions are captured in an ion tunnel, between the opposing forces of the gas flow from the source and the counteracting electric field[18–20]. Trapped ions are then sequentially released as a function of their mobility as the electric potential is lowered. In proteomics, ramp times typically range from 50 to 100 ms, in between chromatographic peak widths (seconds) and the time-of-flight (TOF) spectral acquisition (approximately 100 µs per pulse). In a TIMS-quadrupole-TOF configuration, the mobility separation can be synchronized with the quadrupole mass selection in a method termed parallel accumulation–serial fragmentation (PASEF)[21]. Given that multiple precursors are mass selected and fragmented during a single TIMS scan, PASEF achieves a more than tenfold increase in sequencing speed in DDA, without the loss of sensitivity that is otherwise inherent in very fast fragmentation cycles[22,23]. This is because the precursor ion current is compressed into narrow ion mobility peaks and, with two TIMS in series, ions can be accumulated and mobility analyzed in parallel[24].

Here, we investigate whether the PASEF principle can be extended to DIA, which would combine the advantages of this acquisition method with the inherent efficiency of PASEF. To realize this vision, we modified the mass spectrometer to support 'diaPASEF' acquisition cycles. Building on open-source software[25], we perform targeted extraction of fragment ion traces from the four-dimensional data space for peptide quantification. We explore the performance of the diaPASEF principle in typical proteomics applications such as single-run proteome analysis and label-free quantification, as well as in the characterization of very limited sample amounts.

Results

The diaPASEF principle. In the timsTOF Pro instrument (Bruker Daltonik), peptides separated by liquid chromatography are ionized, introduced into the mass spectrometer and immediately trapped in a dual TIMS device (Fig. 1a). Mobility-separated ions reach the orthogonal accelerator, from which rapid TOF pulses result in high-resolution mass spectra (resolution >35,000 over

[1]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]Functional Proteomics, Jena University Hospital, Jena, Germany. [3]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada. [4]Bruker Daltonik GmbH, Bremen, Germany. [5]Evosep Biosystems, Odense, Denmark. [6]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [7]Faculty of Science, University of Zurich, Zurich, Switzerland. [8]School of Biological Sciences, Queen's University of Belfast, Belfast, UK. [9]NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark. ✉e-mail: ben.collins@qub.ac.uk; hannes.rost@utoronto.ca; mmann@biochem.mpg.de

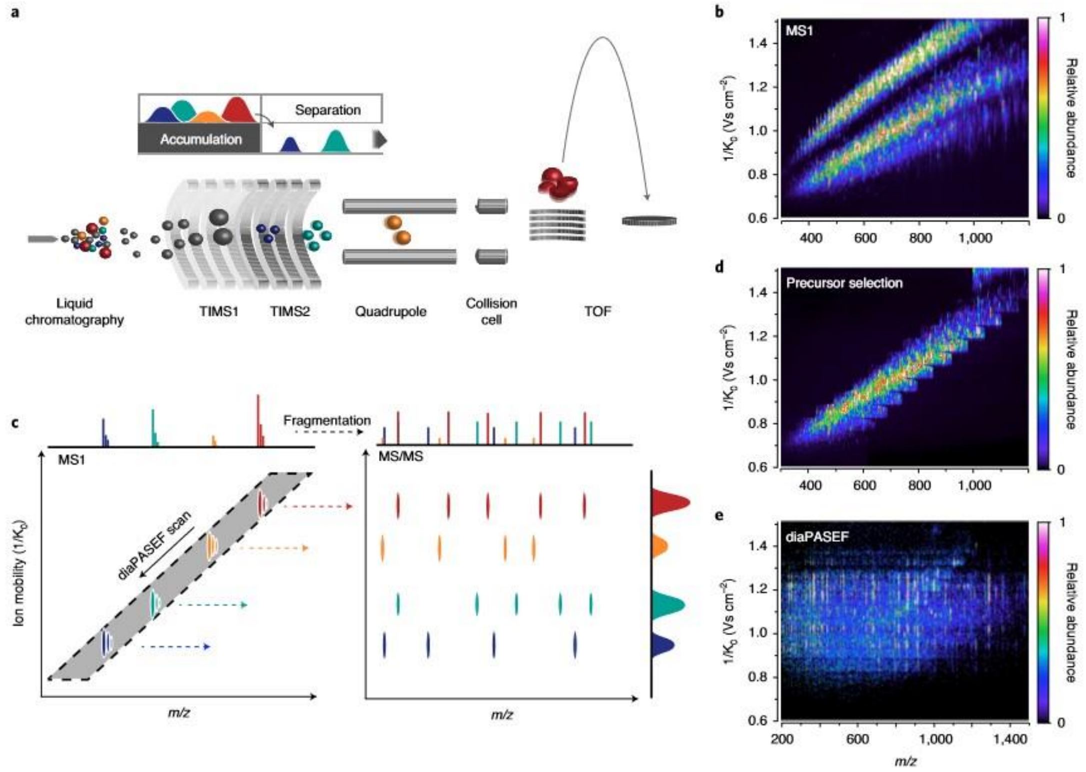


Fig. 1 | The diaPASEF acquisition method. a, Schematic ion path of the timsTOF Pro mass spectrometer. **b**, Correlation of ion mobility and *m/z* in a tryptic digest of HeLa cell lysate. **c**, In diaPASEF, the quadrupole isolation window (gray) is dynamically positioned as a function of ion mobility (arrow). In a single TIMS scan, ions from the selected mass ranges are fragmented to record ion mobility-resolved MS/MS spectra of all precursors. **d**, Implementation of diaPASEF precursor selection with a stepped quadrupole isolation scheme. **e**, Representative example of a single diaPASEF scan with the precursor selection scheme from **d** (Supplementary Fig. 1).

the entire mass range). For peptide ions of a given charge state, ion mobilities and masses are correlated (Fig. 1b). We reasoned that this feature could be used to isolate precursor mass windows for DIA without losing the ions outside the respective windows, in contrast to other DIA acquisition schemes. Given that low-mobility (typically high *m/z*) ions are trapped near the TIMS exit, they are released first, and the mass-selective quadrupole therefore needs to be first positioned at high *m/z*. As higher mobility (typically decreasing *m/z*) ions are sequentially released from the TIMS, the quadrupole mass isolation window should slide down to lower *m/z* values to fully transmit the ion cloud (Fig. 1c). To approximate this ideal diaPASEF scan, we stepped the isolation window as a function of TIMS ramp time (Methods), covering the vast majority of precursors of the 2^+ and 3^+ charge states (Fig. 1d and Supplementary Fig. 1). Implementation of this principle required firmware able to synchronize collision energies with the mass selection (Methods). Note that the fragment ions in each DIA window are detected at the exact ion mobility position of the precursor (Fig. 1e). Over the chromatographic elution of a precursor, the intensities of its fragments follow the precursor intensity in time (*z* direction). The signal traced out by the set of fragments of an individual precursor is a set of very flat ellipsoids (*x* or *m/z* dimension), spreading in the ion mobility direction (*y* direction) and elongated in the retention time dimension (*z* dimension). For the entire experiment, this leads to a data cuboid in four-dimensional space, containing all fragment ions of all precursors over the entire elution time, with signal intensity as the fourth dimension.

Quantification of the increase in ion sampling efficiency. To explore the diaPASEF principle in practice, we measured a tryptic digest of BSA and compared the signals obtained across the DDA, DIA and diaPASEF acquisition methods. As a typical example, the peptide DLGEEHFK eluted over 9 s (Fig. 2a). In DDA, the doubly charged precursor was accumulated at the beginning of the elution peak once for 100 ms before fragmentation. This is approximately 1% of the total elution time and much less than 1% of the entire precursor ion population, as estimated by the relative peak area. In DIA, with a comparably fast cycle time of 1.6 s, the peptide was fragmented seven times over its elution profile. This is sufficient to reconstruct the chromatographic peak shape, but still captured only a small proportion of the total ion signal (less than 5%). By contrast, the diaPASEF scheme (Supplementary Fig. 2) sampled the fragments in each scan for a total of more than 100 times, which resulted in a nearly complete record of the fragments at every time

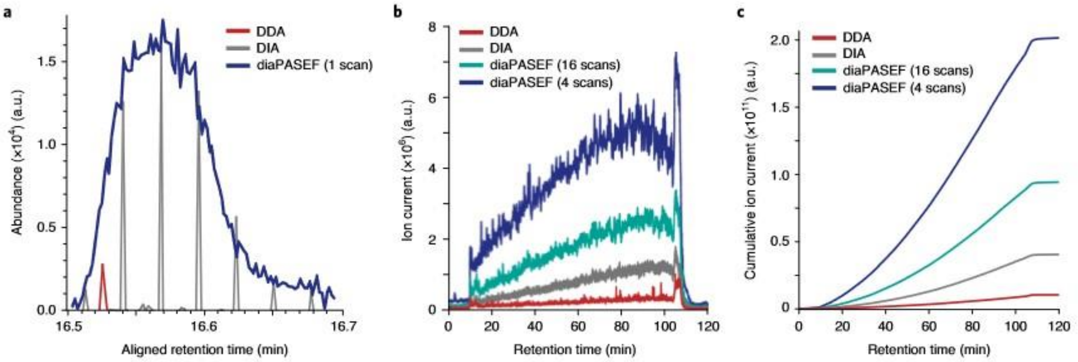


Fig. 2 | Efficiency of different data acquisition methods. **a**, Extracted fragment ion chromatograms of the y_6 ion of the doubly charged DLGEEHFK peptide precursor in a 45 min liquid chromatography–mass spectrometry analysis of BSA digest acquired with typical DDA and DIA methods as well as with a close to 100% duty cycle diaPASEF method shown in Supplementary Fig. 2. a.u., arbitrary units. **b**, Detected ion current from multiply charged precursors in single-run analyses of 200 ng HeLa digest acquired with DDA, DIA and two diaPASEF schemes (Supplementary Figs. 3,4). To extract the ion current after quadrupole isolation, no collision energy was applied and the ion current in the expected peptide space was summed for each TIMS scan. The plot shows the rolling average of 60 TIMS scans. **c**, Same as in **b**, but for cumulative ion current.

point (96% efficiency in terms of acquisition time because of the full scans acquired in between diaPASEF cycles).

We next studied the ion sampling efficiency for a HeLa cell tryptic digest. To address the very high density of fragment ions in the data cuboid, we chose a scheme with four diaPASEF scans, each isolating approximately one-fourth of all precursors with 50 *m/z* isolation windows, and another scheme with 16 diaPASEF scans and 25 *m/z* isolation windows (Supplementary Figs. 3,4). To compare acquisition schemes, no collision energy was applied and we extracted the total ion current of isolated precursors in the expected peptide space in the *m/z*–ion mobility plane (Methods). In DIA, the sampled fraction of the ion current was approximately threefold higher than in DDA, whereas the four-scan diaPASEF scheme further increased the accumulated peptide ion current by a factor of five compared with DIA (Fig. 2b,c). We conclude that the diaPASEF principle yields the expected increase in data acquisition efficiency in both simple and complex proteomes.

Targeted data extraction in four dimensions. To identify and quantify peptides from this novel data structure, we developed Mobi-DIK (ion mobility DIA analysis kit; Fig. 3a). The workflow is based on the targeted extraction of sets of fragment ions of a specific precursor from the acquired dataset over chromatographic elution time, followed by statistical scoring. Mobi-DIK extends this targeted data analysis principle for DIA[26] (as implemented in the OpenSWATH software suite[25]) to diaPASEF. First, ion mobility–enabled spectral libraries are generated from data-dependent PASEF runs using, for instance, the MaxQuant[27,28] output. The spectral library is processed using OpenMS tools[29,30], which we here extended to support ion mobility. Calibration between the assay library and experimental data is automatically performed in *m/z*, retention time and ion mobility dimensions using a set of high-confidence peptides (Methods). The Mobi-DIK package uses the vendor interface to query diaPASEF raw data, convert them to mzML files, and link the isolation windows to individual TOF scans. The algorithm then uses the targeted extraction paradigm for DIA data to construct four-dimensional data cuboids with a user-defined width in *m/z* (ppm), retention time (s) and ion mobility ($\mathrm{V\,s\,cm^{-2}}$). These are projected onto the retention time and ion mobility axes to obtain fragment ion chromatograms and mobilograms for each precursor-to-fragment transition in the spectral library. Restricting the ion mobility extraction width removes signals from co-eluting peptides in the same precursor mass window that have different ion mobility (Fig. 3b and Supplementary Figs. 5–7). Through investigation of transitions in a single-run analysis of HeLa digest, we found that when the ion mobility extraction window was narrowed to $0.06\,\mathrm{V\,s\,cm^{-2}}$, this resulted in an average fourfold increase in the signal-to-noise ratios (Supplementary Fig. 8). Note that the acquisition scheme already removes interfering ions with very different ion mobility such as singly charged species, therefore the true gain in signal-to-noise ratio (compared with the respective value of a DIA experiment without ion mobility) is even higher.

From all projected traces, we next pick peak groups along the chromatographic dimension using established OpenSWATH modules. This step selects putative peak candidates and scores them based on their chromatographic co-elution, goodness of library match and correlation with the precursor profile[25]. For Mobi-DIK, we extended these modules by ion mobility scores. Through use of the high precision of TIMS ion mobility measurements (<1% in replicates of complex samples[22]), a discriminatory score is computed based on the difference between the library and the experimental ion mobility. Additionally, we extract full ion mobilograms for each fragment ion to score the mobility peak shape as well as the peak consistency between all fragment ions. In line with the increased signal-to-noise ratios, the corresponding 'MS/MS signal-to-noise' score increased proportionately with narrower ion mobility extraction windows (Fig. 3c). As a result, in a single-run analysis of a full proteome digest (see below), targeted extraction in the ion mobility dimension (combined with ion mobility–aware scoring) increased peptide identifications by 22% compared with a naive analysis (Fig. 3d).

Single-run proteome analysis. To investigate diaPASEF in a typical DIA experiment, we first built a project-specific library from 24 high-pH reversed-phase peptide fractions of a HeLa digest with data-dependent PASEF, which consisted of 135,671 target precursors and 9,140 target proteins. For sample amounts on column of at least 200 ng and liquid chromatography–mass spectrometry runs of 120 min, we reasoned that a diaPASEF method with a somewhat lower duty cycle, but higher precursor selectivity, should

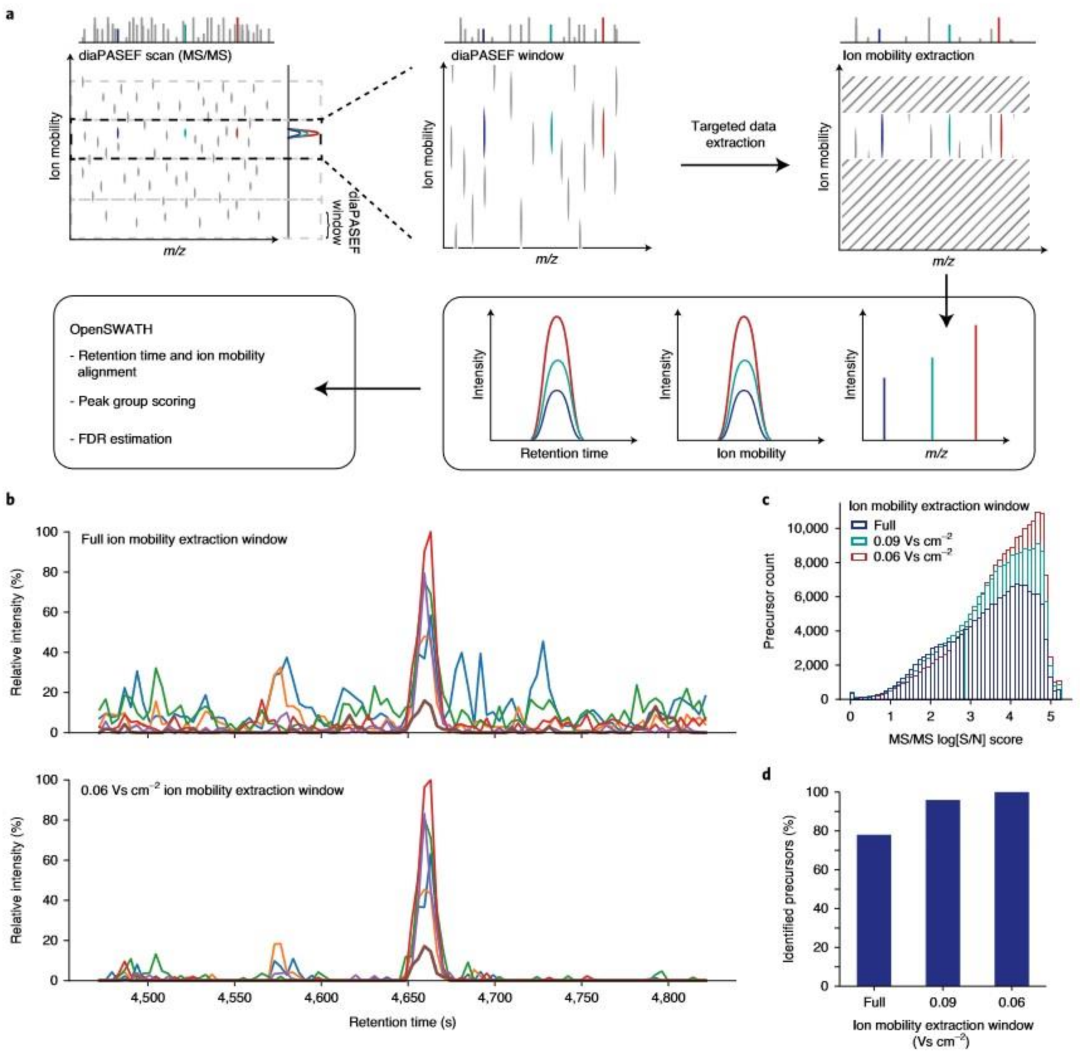


Fig. 3 | Ion mobility-aware targeted data extraction. **a**, Steps in the Mobi-DIK workflow to extract fragment ion chromatograms from diaPASEF scans with restricted ion mobility windows and ion mobility-enhanced peak group scoring in OpenSWATH. Colors (red, blue and cyan) indicate fragment ions from a precursor of interest; gray indicates background signals. **b**, Example fragment ion chromatograms of DGLLIIGVHSAK (color-coded) extracted with (bottom panel) and without (top panel) restriction in the ion mobility dimension from a single-run diaPASEF experiment of HeLa digest. Fragment ions: y_4, orange; y_6, green; y_7, red; y_8, purple; y_9, brown; b_4, blue. **c**, Removal of interfering signals from co-eluting precursors in the same diaPASEF window in triplicate diaPASEF analysis of a HeLa digest. Histograms of the MS/MS signal-to-noise (S/N) scores of identified precursors for different ion mobility extraction windows. $n = 158{,}603$ (full), 194,197 (0.09 Vs cm^{-2}) and 202,218 (0.06 Vs cm^{-2}). **d**, Percentage of detected peptide precursors in triplicate diaPASEF runs from **c** at an FDR of 1% as a function of the ion mobility extraction window.

be beneficial. We devised a method with four windows in each 100 ms diaPASEF scan and 25 *m/z* precursor isolation windows (Supplementary Fig. 4). Eight of these scans covered the diagonal scan line for doubly charged peptides in the *m/z*–ion mobility plane and a second parallel scan line ensured coverage of triply charged species. To reduce potential artifacts from reduced ion transmission at the edges of the diaPASEF windows, we overlapped these scan lines in the ion mobility dimension (Supplementary Fig. 9). The theoretical coverage of library precursor ions was 99.5% and 92.1% for doubly and triply charged peptides in the analyzed *m/z* range 400–1,200, respectively.

In triplicate runs, we detected a total of 80,580 peptide precursors (with 1% precursor and protein false discovery rates (FDRs)), and on average 67,312 peptide precursors per run (Fig. 4a and

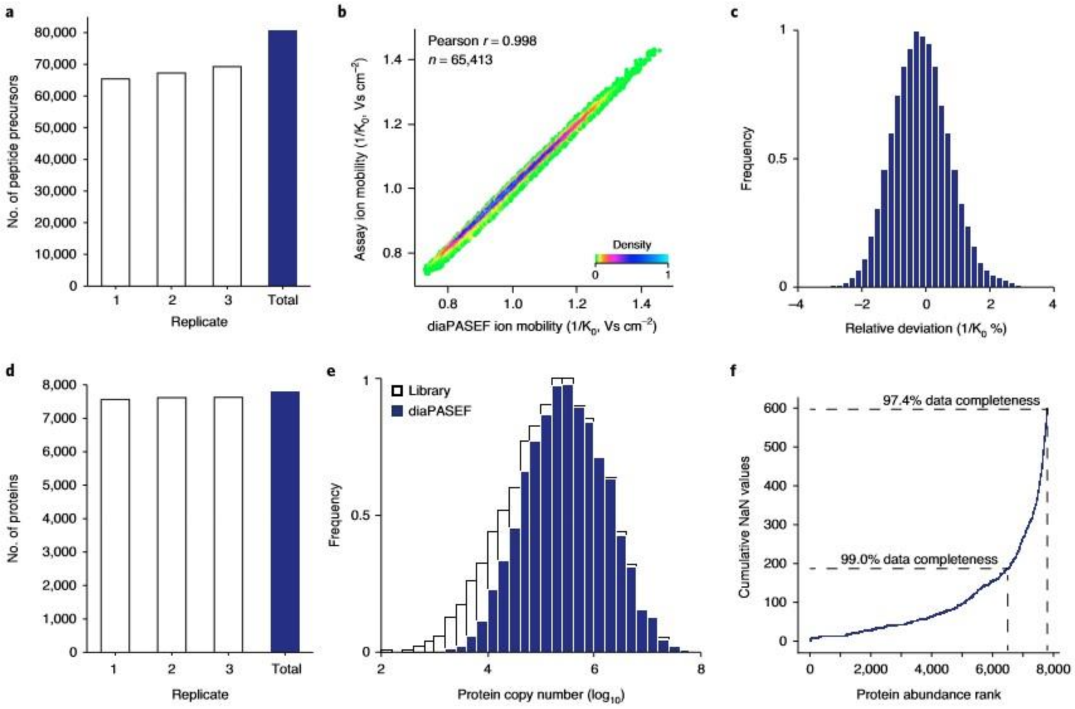


Fig. 4 | Single-run HeLa proteome analysis with diaPASEF. a, Number of peptide precursor ions in triplicate injections of 200 ng HeLa digest with 120 min gradients using the 16-scan diaPASEF scheme shown in Supplementary Fig. 4. **b**, Correlation of precursor ion mobility in a single diaPASEF run with that in the assay library. **c**, Relative deviation of ion mobility values in a single diaPASEF run from the precursor ion mobility in the library. **d**, Number of quantified proteins. **e**, Estimated copy numbers of proteins contained in the assay library and detected with diaPASEF in triplicate single runs. **f**, Cumulative number of missing protein quantification data points (NaN) in the three replicate injections as a function of decreasing protein abundance. Data completeness was calculated as the fraction of valid values in the (number of replicates × abundance rank) matrix.

Supplementary Fig. 10). The ion mobility values in the diaPASEF runs were highly correlated with the library values ($r > 0.99$, Fig. 4b), and the median absolute deviation of the fragment ion mobility values in diaPASEF from those in the library runs was 0.6% (Fig. 4c). The median summed absolute fragment mass deviation was 6.6 ppm and the median absolute retention time deviation was 17 s. Together, these values define the precision of the position of each precursor and its fragments in the diaPASEF data cuboid.

Overall, 66,998 unique peptide sequences were identified at an FDR of 1%, from which 7,601 proteins per run on average and 7,800 proteins in total were inferred using only proteotypic peptides as mapped in the low-redundancy Swiss-Prot database and at a global protein FDR of 1% (Fig. 4d and Supplementary Fig. 11). The quantified proteins spanned a dynamic range of approximately four orders of magnitude, as estimated by protein copy numbers derived from the library (Fig. 4e). Of these, 7,348 proteins (94%) were quantified in all three replicates, 307 in two and only 145 proteins in a single replicate, resulting in a virtually complete data matrix (Fig. 4f) with a median coefficient of variation of 7.7%.

Label-free quantification benchmark. Next, we set up a two-proteome experiment. We spiked 200 ng HeLa samples with approximately 45 ng and 15 ng of a tryptic yeast digest, respectively, and measured both samples in triplicate single runs as above. Mobi-DIK analysis using a combined human and yeast library quantified a total of 82,808 human and 7,483 yeast unique peptide sequences from 101,395 human and 7,992 yeast peak groups, for which 7,943 human and 2,250 yeast proteins were inferred. Although the low-abundance yeast spike-in constituted only 7% of the sample, we quantified 7,697 human and 1,394 yeast proteins in at least two replicates in both samples. Their protein abundance ratios split into two distinct populations according to the mixing ratios (median 2.7-fold, Fig. 5). In line with the quantitative precision demonstrated above, the human population clustered precisely around the 1:1 ratio throughout the full abundance range ($\sigma(\log_2) = 0.22$). The low-abundance yeast spike-ins were quantified with a somewhat lower overall precision ($\sigma(\log_2) = 0.70$), although quantitatively similar to human proteins in the same abundance range. We therefore conclude that the label-free diaPASEF workflow precisely and accurately quantifies changes in protein abundance.

Adaptation of diaPASEF to high-throughput and high-sensitivity proteomics. The diaPASEF schemes can be optimized to balance selectivity (narrower isolation windows), sensitivity (higher mass spectrometry efficiency, fewer diaPASEF scans) and precursor coverage (Fig. 6a). Fast chromatographic methods typically require shorter mass spectrometry cycle times to achieve a sufficient number of data points for accurate quantification. Hence, we devised

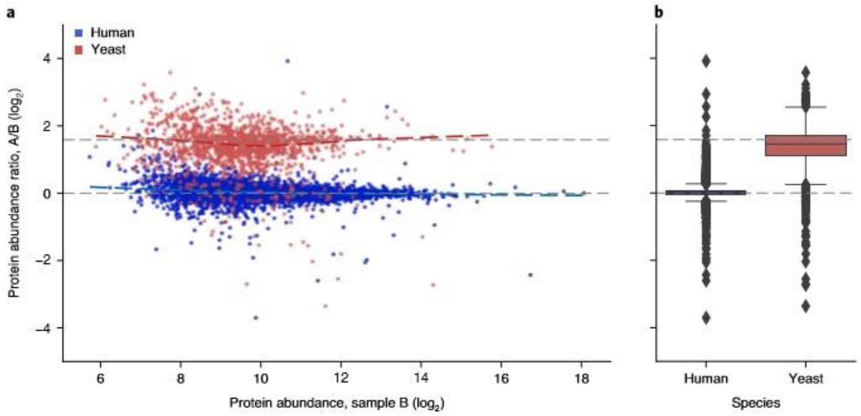


Fig. 5 | Label-free protein quantification benchmark. a, HeLa digest was spiked with approximately 45 ng (sample A) and 15 ng (sample B) yeast digest, and analyzed in triplicate 120 min diaPASEF single runs each, using the 16-scan diaPASEF scheme (Supplementary Fig. 4). log-transformed ratios are plotted as a function of protein abundance for $n=7,697$ human and $n=1,394$ yeast proteins. Dashed gray lines indicate the expected ratio. LOESS regression lines are dashed and colored by species. **b**, Boxplots of the data in **a**, showing the median ratio (center line), the 25th and 75th percentiles (lower and upper box limits, respectively), the 1.5× interquartile range (whiskers) and the outliers (diamonds).

an acquisition scheme that focused on a narrower precursor range with a 0.9 s cycle time (Supplementary Fig. 12). To test this scheme, we turned to a liquid chromatography system with fast turnaround times and predefined, standardized gradients for the analysis of 60, 100 and 200 samples per day (Evosep One)[31]. In triplicate analysis of 200 ng HeLa with the 60 samples per day method (21 min gradient), we quantified on average 4,813 proteins per run and 5,183 in total with a median coefficient of variation of 5.8% (Fig. 6b,c). Remarkably, 4,255 proteins were quantified with a coefficient of variation of <20%. When the throughput was increased to 100 and 200 samples per day, more than 4,000 and 3,000 proteins in triplicate were still quantified, respectively. At 200 samples per day, the median coefficient of variation increased to only 10.3%, which indicates that an even faster diaPASEF method could be viable.

When the number of diaPASEF scans is lowered and the quadrupole isolation width is increased, diaPASEF can be tuned to utilize a higher fraction of the incoming ion beam and still achieve a high precursor selectivity because of the ion mobility separation. To demonstrate this concept, we analyzed only 10 ng of HeLa digest in triplicate 120 min single runs and used a diaPASEF scheme that samples approximately 25% of the ion current of a given precursor (Supplementary Fig. 3). Compared with the standard method, the high duty cycle increased the detected fragment ion signal on average by approximately fourfold and resulted in a more precise quantification of the common peptides, in particular for low-abundance peptides (Fig. 6d). Although the method covers a narrower precursor space, we quantified on average approximately 13,000 peptides with each method and, in effect, the high-sensitivity method extended the detection range of peptides approximately fourfold at the lower end (Fig. 6e). The standard diaPASEF method already quantified on average 3,538 proteins per injection of 10 ng HeLa digest, which highlights the intrinsic high sensitivity of diaPASEF and of the TIMS-QTOF setup. The high-sensitivity method further increased this to 3,835 proteins on average. Cumulatively, we quantified 4,310 proteins in triplicates of 10 ng injections, of which 3,909 were quantified in at least two replicates (Fig. 6f). The increased quantitative precision at the peptide level also translated into higher precision at the protein level, resulting in median coefficients of variation of 9.0% and 11.2% for the high-sensitivity and the standard methods, respectively (3,132 and 2,690 proteins quantified with a coefficient of variation of <20%). However, at higher sample amounts, narrower quadrupole windows were more beneficial. With the high-sensitivity and the standard diaPASEF methods we quantified 4,755 and 4,833 proteins, respectively, from 50 ng samples with a coefficient of variation of <20% (median coefficients of variation of 5.1% and 7.3%), and 5,396 and 5,626 proteins, respectively, from 100 ng samples with a coefficient of variation of <20% (median coefficients of variation of 4.4% and 5.7%, respectively).

Discussion

Here, we have developed and demonstrated a PASEF workflow in a TIMS-TOF mass spectrometer that implements the DIA principle. To make use of the correlation between the ion mobility and the m/z of peptides, precursors are trapped and then released in synchronization with the quadrupole position in our diaPASEF scheme, which results in almost complete sampling of the precursor ion beam. This is in contrast to DDA methods, which convert only a very small fraction (generally much less than 1%) of the incoming ion beam into fragments, and even to typical DIA workflows, which convert a few per cent of the ion beam at best. For less complex mixtures, we achieve close to 100% of the theoretical maximum, whereas for more complex mixtures, it was beneficial to use the quadrupole to decrease spectral complexity and increase selectivity, and thereby to some extent reduce the fraction of total available ions sampled. Note that results could be further improved by the use of brighter electrospray sources and the minimization of ion losses that may occur along the ion path through the instrument and during mass selection. On the predecessor QTOF instrument (Bruker impact II) we found a >80% ion transmission up to the collision cell and an overall detection probability of approximately 10% for ions transferred into the vacuum[32], and an ion trapping efficiency of approximately 70% has been reported for the TIMS device[24].

To extract information using spectral library-based targeted data analysis, we extended the OpenSWATH tool developed for DIA applications to efficiently make use of the ion mobility dimension for library matching, to provide full FDR control and excellent quantification.

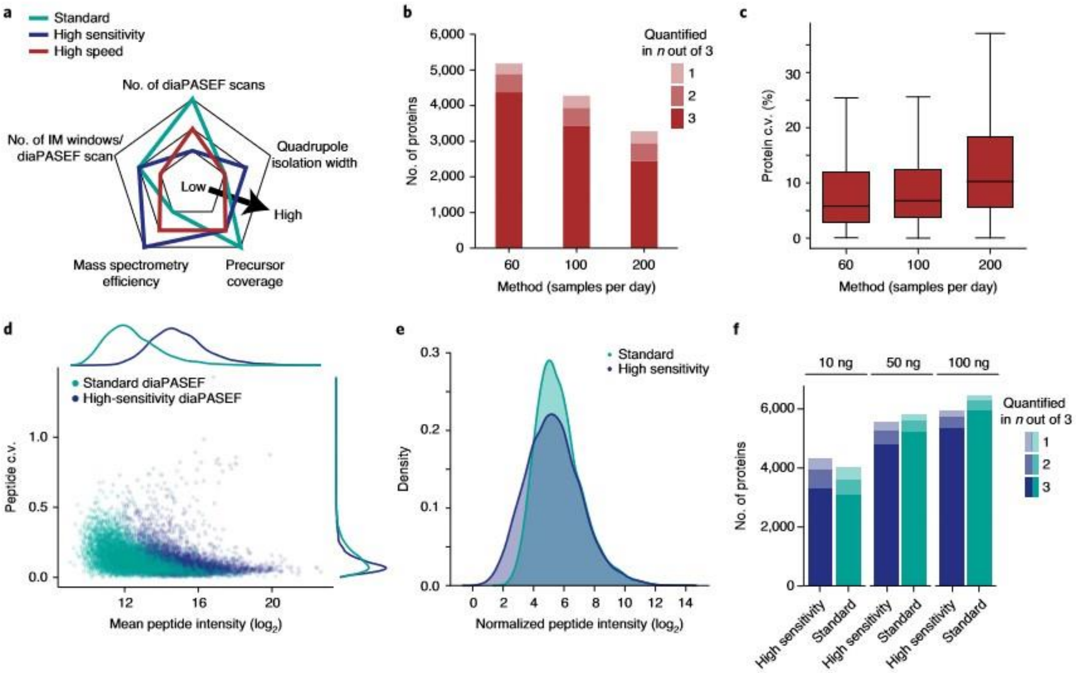


Fig. 6 | High-throughput and high duty cycle diaPASEF analysis. a, Schematic of the parameter space in designing diaPASEF acquisition schemes. IM, ion mobility. **b**, Quantified proteins in *n* out of three replicate injections of 200 ng HeLa digest using different Evosep One liquid chromatography methods (Methods) and a rapid diaPASEF acquisition scheme (Supplementary Fig. 12). **c**, Coefficients of variation of protein abundances measured in at least two replicates. Boxplots show the median (center line), 25th and 75th percentiles (lower and upper box limits, respectively) and the 1.5× interquartile range (whiskers). $n = 4{,}884$ (60 samples per day), 3,940 (100 samples per day) and 2,939 proteins (200 samples per day). **d**, Mean peptide intensity and coefficient of variation of shared peptide precursors in triplicate injections of 10 ng HeLa digest with two different diaPASEF acquisition schemes (Supplementary Figs. 3,4) and a 120 min gradient. Kernel density estimates of peptide intensities and coefficients of variation are presented in the top traces and right traces, respectively. **e**, Peptide intensity distribution for both experiments in **d**, normalized to the most abundant peptide in each. **f**, Quantified proteins in *n* out of three replicate injections of 10 ng, 50 ng and 100 ng HeLa digest as in **d**.

Even in this first implementation, we achieved deep proteome coverage of more than 7,000 proteins in single, 2 h experiments from 200 ng HeLa peptide sample on column with a high degree of reproducibility. Our two-proteome experiment verifies that the quantitative accuracy of the method is in line with previous strategies even when substantially constrained by the lower loading amount of yeast (15 ng). Even more remarkably, we detected more than 4,000 proteins in triplicate injections of only 10 ng HeLa peptide mass on column. This result points to a perhaps unexpected advantage of diaPASEF, namely that the high ion sampling also fully translates into higher sensitivity. Likewise, the very short cycle time of our new scan mode was found to be advantageous for short gradients, which is an increasingly important attribute because large-scale biological and clinical studies require very large throughput. Given that DIA methods record chromatographic profiles for each fragment ion, they are also increasingly attractive for site-specific analysis of modified peptides[33,34]. With diaPASEF, such strategies could additionally benefit from the separation of positional isomers in the ion mobility dimension[35]. For the future, we imagine that both hardware and software can still be greatly optimized to further increase the amount and quality of the information contained in and extracted from the extremely rich four-dimensional diaPASEF data cuboids. For example, advanced data acquisition schemes could sample the correlation of precursor mobility and *m/z* more precisely if the isolation window width is varied or the quadrupole is scanned rather than moved in discrete steps. Furthermore, we note that applications of diaPASEF are not restricted to peptides but could equally well be extended to metabolites, lipids or other compound classes[23].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-00998-0.

Received: 24 February 2020; Accepted: 15 October 2020;
Published online: 30 November 2020

References

1. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2012).
2. Larance, M. & Lamond, A. I. Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* **16**, 269–280 (2015).

3. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
4. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599.e4 (2017).
5. Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).
6. Röst, H. L., Malmström, L. & Aebersold, R. Reproducible quantitative proteotype data matrices for systems biology. *Mol. Biol. Cell* **26**, 3926–3931 (2015).
7. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35 (2015).
8. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* **33**, 452–470 (2014).
9. Ludwig, C. et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).
10. Gillet, L. C., Leitner, A. & Aebersold, R. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annu. Rev. Anal. Chem.* **9**, 449–472 (2016).
11. Bilbao, A. et al. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15**, 964–980 (2015).
12. Bruderer, R. et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteom.* **16**, 2296–2309 (2017).
13. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Mol. Cell. Proteom.* **19**, 1088–1103 (2020).
14. McLean, J. A., Ruotolo, B. T., Gillig, K. J. & Russell, D. H. Ion mobility–mass spectrometry: a new paradigm for proteomics. *Int. J. Mass Spectrom.* **240**, 301–315 (2005).
15. Distler, U. et al. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **11**, 167–170 (2014).
16. Helm, D. et al. Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell. Proteom.* **13**, 3709–3715 (2014).
17. Ewing, M. A., Glover, M. S. & Clemmer, D. E. Hybrid ion mobility and mass spectrometry as a separation tool. *J. Chromatogr. A* **1439**, 3–25 (2016).
18. Fernandez-Lima, F. A., Kaplan, D. A. & Park, M. A. Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **82**, 126106 (2011).
19. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* **14**, 93–98 (2011).
20. Ridgeway, M. E., Lubeck, M., Jordens, J., Mann, M. & Park, M. A. Trapped ion mobility spectrometry: a short review. *Int. J. Mass Spectrom.* **425**, 22–35 (2018).
21. Meier, F. et al. Parallel accumulation–serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14**, 5378–5387 (2015).
22. Meier, F. et al. Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteom.* **17**, 2534–2545 (2018).
23. Vasilopoulou, C. G. et al. Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. *Nat. Commun.* **11**, 331 (2020).
24. Silveira, J. A., Ridgeway, M. E., Laukien, F. H., Mann, M. & Park, M. A. Parallel accumulation for 100% duty cycle trapped ion mobility-mass spectrometry. *Int. J. Mass Spectrom.* **413**, 168–175 (2017).
25. Röst, H. L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).
26. Gillet, L. C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteom.* **11**, O111.016717 (2012).
27. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
28. Prianichnikov, N. et al. MaxQuant software for ion mobility enhanced shotgun proteomics. *Mol. Cell. Proteom.* **19**, 1058–1069 (2020).
29. Rosenberger, G. et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).
30. Röst, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
31. Bache, N. et al. A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics.*Mol. Cell. Proteom.* **17**, 2284–2296 (2018).
32. Beck, S. et al. The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol. Cell. Proteom.* **14**, 2014–2029 (2015).
33. Searle, B. C., Lawrence, R. T., MacCoss, M. J. & Villén, J. Thesaurus: quantifying phosphopeptide positional isomers. *Nat. Methods* **16**, 703–706 (2019).
34. Bekker-Jensen, D. B. et al. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* **11**, 787 (2020).
35. Glover, M. S. et al. Examining the influence of phosphorylation on peptide ion structure by ion mobility spectrometry-mass spectrometry. *J. Am. Soc. Mass Spectrom.* **27**, 786–794 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Sample preparation. The human cancer cell line (HeLa S3, ATCC) was cultured in Dulbecco's modified Eagle's medium with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin–streptomycin. Cells were collected by centrifugation, washed with phosphate-buffered saline, flash-frozen in liquid nitrogen and stored at −80 °C. Cell lysis, reduction and alkylation were performed in lysis buffer with chloroacetamide (PreOmics) as reported previously[36]. In brief, the cell suspension was heated to 95 °C for 10 min and subsequently sonificated to further disrupt cells and shear nucleic acids. Proteins were enzymatically cleaved overnight by adding equal amounts of Lys-C and trypsin in a 1:100 (wt/wt) enzyme:protein ratio. De-salting and purification were performed according to the PreOmics iST protocol on a styrene divinylbenzene reversed-phase sulfonate (SDB-RPS) sorbent. Purified peptides were vacuum-centrifuged to dryness and reconstituted in double-distilled water with 2 vol% acetonitrile (ACN) and 0.1 vol% trifluoroacetic acid (TFA) for single-run LC-MS analysis or fractionation.

To evaluate the quantitative accuracy of diaPASEF, we performed a two-proteome experiment with HeLa and yeast. Furthermore, to evaluate the achievable sample throughput we analyzed HeLa samples using the Evosep One liquid chromatography system. For these experiments, purified and predigested yeast standard was purchased from Promega and resuspended in 0.1 vol% formic acid; whole HeLa cell pellets were purchased from CIL Biotech and lysed using trifluoroethanol[37]. In brief, the cell suspension was kept on ice for 10 min and subsequently incubated for 20 min at 56 °C. We used 200 mM dithiothreitol to reduce proteins at 90 °C (20 min), and 200 mM iodoacetamide to alkylate cysteine residues during 90 min at room temperature (21 °C). Proteins were enzymatically cleaved overnight by adding trypsin in a 1:100 (wt/wt) enzyme:protein ratio. The proteome digests were de-salted and purified on a solid phase extraction cartridge (Empore C_{18} SPE cartridge, Sigma Aldrich). Samples were washed with 0.1 vol% formic acid and subsequently eluted with 50 vol% ACN in 0.1 vol% formic acid. Purified and dried peptides were reconstituted in 0.1 vol% formic acid for injection. For the two-proteome experiment, the purified peptides from HeLa and yeast were combined as follows: sample A consisted of 200 ng human and 45 ng yeast proteins per LC-MS injection, and sample B of 200 ng human and 15 ng yeast proteins per LC-MS injection. For the Evosep experiments, approximately 200 ng peptides was loaded onto Evotips (EV2001, Evosep) in accordance with the manufacturer's instructions.

High-pH reversed-phase fractionation. To generate a comprehensive library of HeLa precursor and fragment ions, peptides were fractionated at pH 10 with a 'spider fractionator' coupled to an EASY-nLC 1000 chromatography system (Thermo Fisher Scientific) as described previously[38]. Approximately 50 µg purified peptides were separated on a 30 cm C_{18} column in 96 min and automatically concatenated into 24 fractions by shifting the exit valve every 120 s. The fractions were vacuum-centrifuged to dryness and reconstituted in double-distilled water with 2 vol% ACN and 0.1 vol% TFA for LC-MS analysis. To generate spectral libraries for the Evosep and two-proteome experiments, 100 µg purified peptides from yeast and from HeLa digests were each fractionated at pH 10 on a reversed-phase column (Waters Acquity CSH C18 column, 1.7 µm, 2.1 × 150 mm) using a Dionex Ultimate 3000 system (Thermo Fisher Scientific). For mass spectrometric analysis, the freeze-dried fractions were reconstituted in 0.1% formic acid and placed in the autosampler or loaded onto Evotips.

Liquid chromatography. Nanoflow reversed-phase chromatography was performed on an EASY-nLC 1200 system (Thermo Fisher Scientific). Peptides were separated in 120 min at a flow rate of 300 nl min^{-1} on a 50 cm × 75 µm column with a laser-pulled electrospray emitter packed with 1.9 µm ReproSil-Pur C_{18}-AQ particles (Dr. Maisch). Mobile phases A and B were water with 0.1 vol% formic acid and 80:20:0.1 vol% ACN:water:formic acid, respectively. The fraction of B was linearly increased from 5% to 30% in 95 min, followed by an increase to 60% in 5 min and a further increase to 95% in 5 min before re-equilibration.

For the two-proteome experiment, we used a nanoElute liquid chromatography system (Bruker Daltonics). Peptides were separated in 120 min at a flow rate of 400 nl min^{-1} on a commercially available reversed-phase C_{18} column with an integrated CaptiveSpray Emitter (25 cm × 75 µm, 1.6 µm, IonOpticks). Mobile phases A and B were 0.1 vol% formic acid in water and 0.1 vol% formic acid in ACN, respectively. The fraction of B was linearly increased from 2% to 25% in 90 min, followed by an increase to 35% in 10 min and a further increase to 80% in 10 min before re-equilibration.

For proteome analyses with fast gradients, we used an Evosep One liquid chromatography system[31] and analyzed the samples with the predefined 60, 100 or 200 samples per day methods (Evosep RC.Net 1.3 plugin). For the 60 and 100 samples per day methods, we used an 8 cm × 150 µm column with 1.5 µm C_{18} beads (EV1109, Evosep) and for the 200 samples per day method, we used a 4 cm × 150 µm column with 1.9 µm C_{18} beads (EV1107, Evosep). Mobile phases A and B were 0.1 vol% formic acid in water and 0.1 vol% formic acid in ACN, respectively.

Mass spectrometry. Liquid chromatography was coupled online to a hybrid TIMS quadrupole TOF mass spectrometer (Bruker timsTOF Pro) via a CaptiveSpray nano-electrospray ion source. A detailed description of the instrument is available in ref. [22]. The dual TIMS analyzer was operated at a fixed duty cycle close to 100% using equal accumulation and ramp times of 100 ms each. We performed DDA in PASEF mode with 10 PASEF scans per topN acquisition cycle. Singly charged precursors were excluded by their position in the m/z–ion mobility plane, and precursors that reached a target value of 20,000 arbitrary units were dynamically excluded for 0.4 min. The quadrupole isolation width was set to 2 m/z for $m/z < 700$ and to 3 m/z for $m/z > 700$. TIMS elution voltages were calibrated linearly to obtain the reduced ion mobility coefficients ($1/K_0$) using three Agilent ESI-L Tuning Mix ions (m/z 622, 922 and 1,222).

To perform DIA, we extended the instrument control software (Bruker otofControl v6) to define quadrupole isolation windows as a function of the TIMS scan time (diaPASEF). The instrument control electronics were modified to allow seamless and synchronous ramping of all applied voltages. We tested multiple schemes for data-independent precursor windows and placement in the m/z–ion mobility plane and defined up to eight windows for single 100 ms TIMS scans, as detailed earlier. Acquisition schemes for the diaPASEF methods used herein are shown in Supplementary Figs. 1–4,12. To limit the number of MS1 scans, we repeated diaPASEF in acquisition schemes; for example, each of the four diaPASEF scans was done twice in the high-sensitivity scheme, and this resulted in one MS1 and eight diaPASEF scans per acquisition cycle. In both scan modes, the collision energy was ramped linearly as a function of the mobility from 59 eV at $1/K_0 = 1.6\ \text{V s cm}^{-2}$ to 20 eV at $1/K_0 = 0.6\ \text{V s cm}^{-2}$. To visualize the isolation of precursor ions in Fig. 1d and analyze the ion current from multiply charged precursors (likely peptide precursors) in Fig. 2, we set the collision energy to 5 eV to prevent fragmentation. In the BSA experiment, we distributed the 14 diaPASEF windows to one TIMS scan each and defined 14 × 50 Th precursor isolation windows from m/z 325 to 1,025. In the HeLa DIA experiment, we defined 32 × 25 Th isolation windows from m/z 400 to 1,200. To adapt the MS1 cycle time in diaPASEF, we set the repetitions to 2 in the 16-scan diaPASEF scheme and to 4 in the 4-scan diaPASEF scheme in these experiments.

Spectral library generation. To generate spectral libraries for targeted data extraction, we first analyzed high-pH reversed-phase fractions acquired in DDA mode with MaxQuant v1.6.5.0 or 1.6.7.0, which extracts four-dimensional features on the MS1 level (retention time, m/z, ion mobility and intensity) and links them to peptide spectrum matches. We had acquired the 120 min HeLa library previously for the purpose of predicting ion mobility cross-sections by deep learning[39]. The maximum precursor mass tolerance of the main search was set to 20 ppm and de-isotoping of fragment ions was deactivated. Other than that, we used the default 'TIMS-DDA' parameters. Tandem mass spectrometry spectra were matched against an *in silico* digest of the appropriate Swiss-Prot proteome database (human, 20,402 entries; *Saccharomyces cerevisiae*, 6,721 entries) and a list of common contaminants. The minimum peptide length was set to 7 amino acids, and the peptide mass was limited to 4,600 Da. Carbamidomethylation of cysteine residues was defined as a fixed modification, and methionine oxidation and acetylation of protein N-termini were defined as variable modifications. The FDR was controlled at <1% at both the peptide spectrum match level and the protein level. The Mobi-DIK software package builds on OpenMS tools to compile spectral libraries in the standardized TraML or pqp formats from the MaxQuant output tables and retains the full ion mobility information for each precursor-to-fragment ion transition. Only proteotypic peptides with precursor $m/z > 400$ were included in the library; they were required to have a minimum of six fragment ions with $m/z > 350$ and to be outside the precursor mass isolation range. We generated separate, project-specific libraries for the 120 min HeLa experiments, the two-proteome experiment and the Evosep experiment (60 samples per day method).

Targeted data extraction. To analyze diaPASEF data, we developed an ion mobility DIA analysis kit (Mobi-DIK) that extracts fragment ion traces from the four-dimensional data space, as detailed earlier. Repeated diaPASEF scans were merged. Raw data were automatically re-calibrated using curated reference values in m/z, retention time and ion mobility dimensions (387 peptides for linear and 3,184 peptides for non-linear alignment). We applied an outlier detection in each dimension before calculating the final fit function to increase robustness. Peak picking and subsequent scoring functionalities in the Mobi-DIK software build on OpenSWATH[25] modules. For diaPASEF, we extended these modules to also consider the additional ion mobility dimension. OpenSWATH (revision: e0b987a) was run with the following parameters: min_coverage = 0.1 (0.3 in Fig. 3), RTNormalization:alignmentMethod = LOWESS, RTNormalization:lowess:span = 0.01, Scoring:TransitionGroupPicker:PeakPickerMRM:sgolay_frame_length = 11, Scoring:stop_report_after_feature = 5, rt_extraction_window = 250, Scoring:Scores:use_ion_mobility_scores, mz_correction_function = quadratic_regression_delta_ppm, use_ms1_traces, mz_extraction_window = 25, mz_extraction_window_unit = ppm, mz_extraction_window_ms1 = 25, mz_extraction_window_ms1_unit = ppm, irt_mz_extraction_window_unit = ppm, irt_mz_extraction_window = 40, Calibration:ms1_im_calibration, ion_mobility_window = 0.06, irt_im_extraction_window = 99, RTNormalization:NrRTBins = 8, RTNormalization:MinBinsFilled = 4. All other parameters were set to default values. PyProphet was used to train an XGBoost classifier for target-decoy separation by first creating one concatenated and subsampled OpenSwath output for each set of three replicate injections of the same acquisition strategy and sample amount. The classifier was subsequently applied to score all samples, with FDR controlled to <1% at the peak group level per sample, and at both the global peptide and global protein levels. For the two-proteome experiment, the protein FDR was set to <1%, and TRIC alignment[40] was performed using a peak-group level seed q value threshold of 0.01 and extension q value threshold of 0.05. In the case of two overlapping diaPASEF windows, the analysis was performed separately for the individual windows, and for FDR estimation the highest scoring peak group was selected. Protein abundances were estimated using an R implementation of the MaxLFQ[41] algorithm for DIA termed *iq* (v1.9) with default parameters[42]. Potential contaminants were excluded from further analysis.

Bioinformatics. Output tables from the Mobi-DIK data analysis pipeline were further analyzed and visualized in the R statistical computing environment v4 or in Python v3.6. Ion chromatograms shown in Fig. 2a were extracted from raw data files with the Bruker DataAnalysis software. To estimate the peptide precursor ion current sampled with different acquisition methods in Fig. 2b, we extracted tandem mass spectrometry spectra directly from the raw data files using an SQL interface. Given that the isolated precursors were not fragmented in this experiment, we were able to restrict the analysis to likely multiply charged peptide ions by their position in the ion mobility–m/z space. For this, we empirically estimated a line separating singly from multiply charged species and discarded all signals with $1/K_0 \geq 0.0009 * m/z + 0.48$. Protein copy numbers were estimated with the Proteomic Ruler[43] Perseus[44] (v1.6.0.8) plugin from the MaxQuant output table.

Statistics. Summary statistics such as coefficients of variation were calculated based on replicate injections of the same sample ($n = 3$ technical replicates) to indicate the technical variation of the mass spectrometry method.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry raw data and spectral libraries generated and analyzed during the current study have been deposited with the ProteomeXchange Consortium via the PRIDE[45] partner repository with the dataset identifier PXD017703. *Homo sapiens* (taxon identifier: 9606) and *S. cerevisiae* (taxon identifier: 559292) proteome databases were downloaded from https://www.uniprot.org. Source data are provided with this paper.

Code availability

Code is available under the three-clause BSD license on https://github.com/OpenMS/OpenMS and https://github.com/Roestlab/dia-pasef.

References

36. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
37. Wang, H. et al. Development and evaluation of a micro- and nanoscale proteomic sample preparation method. *J. Proteome Res.* **4**, 2397–2403 (2005).
38. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteom.* **16**, 694–705 (2017).
39. Meier, F. et al. Deep learning the collisional cross sections of the peptide universe from a million training samples. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.19.102285 (2020).
40. Röst, H. L. et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13**, 777–783 (2016).
41. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteom.* **13**, 2513–2526 (2014).
42. Pham, T. V., Henneman, A. A. & Jimenez, C. R. iq: an R package to estimate relative protein abundances from ion quantification in DIA-MS-based proteomics. *Bioinformatics* **36**, 2611–2613 (2020).
43. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteom.* **13**, 3497–3506 (2014).
44. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
45. Vizcaino, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).

Acknowledgements

This work was partially supported by the German Research Foundation (DFG-Gottfried Wilhelm Leibniz Prize granted to M.M., grant no. MA 1764/2-1) and by the Max Planck Society for the Advancement of Science (M.M.). This work was partially supported by the Government of Canada through Genome Canada (grant no. 15411) and by the Canadian Institutes for Health Research (H.L.R.). B.C.C. was supported by a Swiss National Science Foundation Ambizione grant (no. PZ00P3_161435). R.A. was supported by the Swiss National Science Foundation (grant no. 3100A0-688 107679) and the European Research Council (ERC-20140AdG 670821). We thank our colleagues in the Department of Proteomics and Signal Transduction (Max Planck Institute of Biochemistry) and at Bruker Daltonik for discussions and help; in particular J. Müller, A. Strasser, C. Deiml and I. Paron for technical support.

Author contributions

F.M., R.A., B.C.C., H.L.R. and M.M. conceptualized and designed the study; F.M. and M.M. conceived the acquisition mode; H.L.R. conceived the data analysis software; F.M., A.-D.B., S.K.-S., M.L., O.R., N.B. and B.C.C. performed experiments; A.H. and M.F. contributed to the software development; F.M., A.-D.B., M.F., A.H., I.B., E.V., S.K.-S., B.C.C., H.L.R. and M.M. analyzed the data; F.M., R.A., B.C.C., H.L.R. and M.M. wrote the manuscript.

Competing interests

S.K.-S., M.L. and O.R. are employees of Bruker Daltonik. N.B. is an employee of and M.M. a shareholder in Evosep Biosystems. All other authors have no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-00998-0.

Correspondence and requests for materials should be addressed to B.C.C., H.L.R. or M.M.

Peer review information Arunima Singh was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

3.3.2. Article 7: Ultra-high sensitivity MS quantifies single-cell proteomes

Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation

bioRxiv, December 23, 2020 (Under review in Nature)

Andreas-David Brunner[1], Marvin Thielert[1], Catherine G. Vasilopoulou[1], Constantin Ammar[1], Fabian Coscia[2], Andreas Mund[2], Ole B. Hoerning[3], Nicolai Bache[3], Amalia Apalategui[4], Markus Lubeck[4], Sabrina Richter[5,6], David S. Fischer[5,6], Oliver Raether[4], Melvin A. Park[7], Florian Meier[1, 8], Fabian J. Theis[5,6], Matthias Mann[1,2, #]

# *Corresponding author*

[1] *Proteins and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany*

[2] *NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 220 Copenhagen, Denmark*

[3] *EvoSep Biosystems, Thriges Pl. 6, 5000 Odense, Denmark*

[4] *Bruker Daltonik GmbH, Fahrenheitstraße 4, 28359 Bremen, Germany*

[5] *Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstraße 1, Neuherberg, Germany*

[6] *TUM School of Life Sciences Weihenstephan, Technical University of Munich, Alte Akademie 8, 85354 Freising, Germany*

[7] *Bruker Daltonics Inc., 40 Mannin Road, Billerica, MA 01821, United States of America*

[8] *Functional Proteomics, Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany*

Contribution

I developed a full single-cell proteomics workflow in which we measured true single cell proteomes – one by one. Sample preparation included optimization of miniaturization and its coupling to the *EvoSep One* LC platform. Furthermore, since ES is concentration dependent, I re-purposed or 'hijacked' the *EvoSep One* microflow system together with our collaboration partners to run robust true nanoflow gradients. Additionally, I benchmarked the novel and modified *timsTOF Pro*, tuned it to highest sensitivity and coupled it to the novel *EvoSep One* true nanoflow gradients. The diaPASEF scan mode optimization, all single-cell benchmarking and cell cycle experiments, as well as downstream analysis was performed by me. Comparisons to similar single-cell RNA sequencing data was led by me with help from our collaboration partners. Moreover, I wrote the manuscript draft and co-edited it to its final form.

Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation

Andreas-David Brunner[1], Marvin Thielert[1], Catherine G. Vasilopoulou[1], Constantin Ammar[1], Fabian Coscia[2], Andreas Mund[2], Ole B. Hoerning[3], Nicolai Bache[3], Amalia Apalategui[4], Markus Lubeck[4], Sabrina Richter[6, 7], David S. Fischer[6, 7], Oliver Raether[4], Melvin A. Park[5], Florian Meier[1, 8], Fabian J. Theis[6, 7], Matthias Mann[1, 2, *]

[1]Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[2]NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark

[3]EvoSep Biosystems, Thriges Pl. 6, 5000 Odense, Denmark

[4]Bruker Daltonik GmbH, Fahrenheitstraße 4, 28359 Bremen, Germany

[5]Bruker Daltonics Inc., 40 Manning Road, Billerica, MA 01821, United States of America

[6]Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstraße 1, Neuherberg, Germany

[7]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Alte Akademie 8, 85354 Freising, Germany

[8]Functional Proteomics, Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany

*Corresponding author: mmann@biochem.mpg.de

Abstract

Single-cell technologies are revolutionizing biology but are today mainly limited to imaging and deep sequencing[1–3]. However, proteins are the main drivers of cellular function and in-depth characterization of individual cells by mass spectrometry (MS)-based proteomics would thus be highly valuable and complementary[4,5]. Chemical labeling-based single-cell approaches introduce hundreds of cells into the MS, but direct analysis of single cells has not yet reached the necessary sensitivity, robustness and quantitative accuracy to answer biological questions[6,7]. Here, we develop a robust workflow combining miniaturized sample preparation, very low flow-rate chromatography and a novel trapped ion mobility mass spectrometer, resulting in a more than ten-fold improved sensitivity. We accurately and robustly quantify proteomes and their changes in single, FACS-isolated cells. Arresting cells at defined stages of the cell cycle by drug treatment retrieves expected key regulators such as CDK2NA, the E2 ubiquitin ligase UBE2S, DNA topoisomerases TOP2A/B and the chromatin regulator HMGA1. Furthermore, it highlights potential novel ones and allows cell phase prediction. Comparing the variability in more than 430 single-cell proteomes to transcriptome data revealed a stable core proteome despite perturbation, while the transcriptome appears volatile. This emphasizes substantial regulation of translation and sets the stage for its elucidation at the single cell level. Our technology can readily be applied to ultra-high sensitivity analyses of tissue material[8], posttranslational modifications and small molecule studies to gain unprecedented insights into cellular heterogeneity in health and disease.

Main

In single-cell analysis, biological variability can directly be attributed to individual cells instead of being averaged over an ensemble or complex tissue[9]. While microscopy has always been single-cell based, specialized deep sequencing technologies have achieved this for systems biological approaches[10–12]. At the level of proteins, the functional actors of cells, single cells are currently studied by antibody-based technologies, which are by necessity directed against previously chosen targets[13–15]. In contrast, mass spectrometry (MS)-based proteomics is unbiased in the sense that it measures all proteins within its range of detection[4,16]. Thus, it would be highly desirable to apply this technology to single cells if the required sensitivity and robustness could be achieved. Previous approaches that employed chemical multiplexing of peptides have labeled a small number of single cells but combined them with a dominant booster channel for MS-analysis[17,18], which can hamper signal deconvolution[6,19]. Alternatively, proof of principle has been demonstrated for unlabeled approaches using sophisticated sample preparation methods in pico-liter devices[7,20,21]. However, a technology that provides quantitatively accurate MS proteomics data from true single-cells (T-SCP) and answers biological questions is still outstanding.

Noise-reduced quantitative mass spectra

We recently introduced parallel accumulation – serial fragmentation (PASEF), a mass spectrometric acquisition scheme in which peptide ions are released from a trapped ion mobility (TIMS) device into the vacuum system in concentrated packages[22,23]. Chemical noise is widely distributed as a result of its heterogeneous nature and the ten-fold increased peak capacity due to TIMS (**Fig. 1a, b**)[24]. These precursors can be fragmented in a highly sensitive manner, either in data dependent (ddaPASEF) or data independent (diaPASEF) mode, resulting in very high ion utilization and data completeness[25]. To explore sensitivity limits, we measured a dilution series of HeLa cell lysate from 25 ng down to the equivalent of a few single cells on a quadrupole time-of-flight instrument (TIMS-qTOF). This identified more than 550 proteins from 0.8 ng HeLa lysate with the dda acquisition mode and a conservative MaxQuant analysis (**Fig. 1c**)[26]. Proteins were quantified with the linear signal response expected from the dilution factors (**Fig. 1d**). Furthermore, quantitative reproducibility in replicates at the lowest level was still excellent (R = 0.96, **Fig. 1e**). Given that the protein amount of a single HeLa cell is as low as 150 pg[27], and accounting for inevitable losses in sample preparation including protein digestion, we estimated that we would need to increase sensitivity by at least an order of magnitude to enable true single cell proteomics.

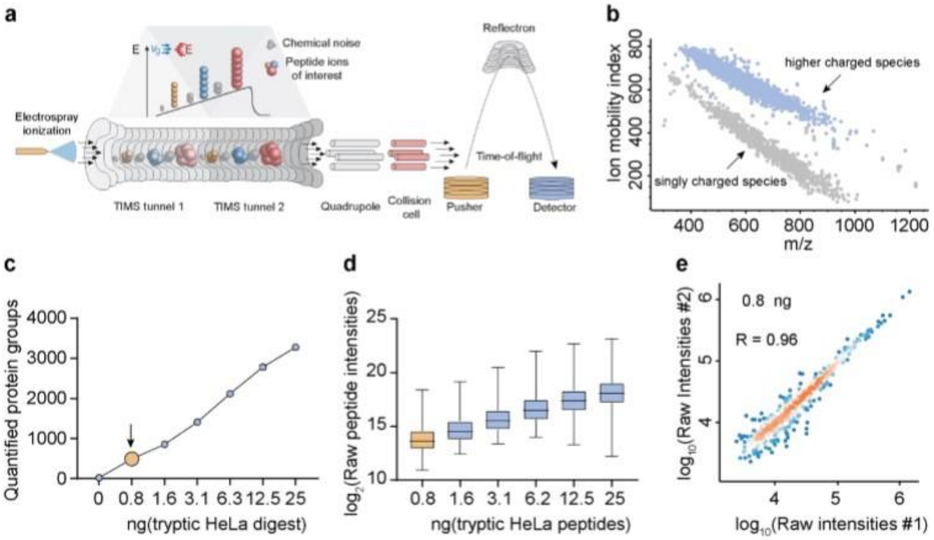


Figure 1: TIMS enables virtually noise-free spectra and ultra-high sensitivity proteomics. a, b, TIMS-qTOF principle separating singly charged background peaks from multiply charged peptide precursor ions, making precursor ions visible at extremely low signal levels (0.8 ng HeLa digest). **c,** Quantified proteins from a HeLa digest dilution series from 25 ng peptide material down to 0.8 ng (arrow), roughly corresponding to the protein amount contained in three HeLa cells. **d,** Linear quantitative response curve of the HeLa digest experiment in c. **e,** Quantitative reproducibility of two successive HeLa digest experiments at the lowest dilution (technical LC-MS/MS replicates).

True single-cell proteome analysis

Three main factors govern MS sensitivity: ionization efficiency, transfer efficiency into the vacuum system and ion utilization by the instrument[28]. We first constructed an instrument with a brighter ion source, introduced different ion optic elements and optimized parameters such as detector voltage. Together, this led to a more than 4-fold higher ion current (**Fig. 2a**). Next, we FACS sorted zero, one and up to six single HeLa cells in quadruplicate into individual 384-wells, processed them separately and analyzed them on this modified mass spectrometer. This resulted on average in 843, 1,279 and 1,890 identified proteins for one, two and six cells, respectively. Note that this analysis benefited from transferring peptide identifications on the MS1 level, as expected from extremely low sample amounts (**Fig. 2b**). Quantification accuracy was high when comparing single cells, not much reduced from comparing six cells (**Fig. 2 c, d**). A rank order abundance plot revealed that the measured single-cell proteome preferentially mapped to the higher abundant part of the six-cell proteome, indicating that proteome coverage depended deterministically on overall LC-MS sensitivity (**Fig. 2e**). Inspecting shared peptides between the single-cell and six-cell experiment showed that clearly interpretable precursor isotope patterns were still present at high signal-to-noise levels even at single-cell level following the cell count intensity ratio trend (**Fig. 2f**).

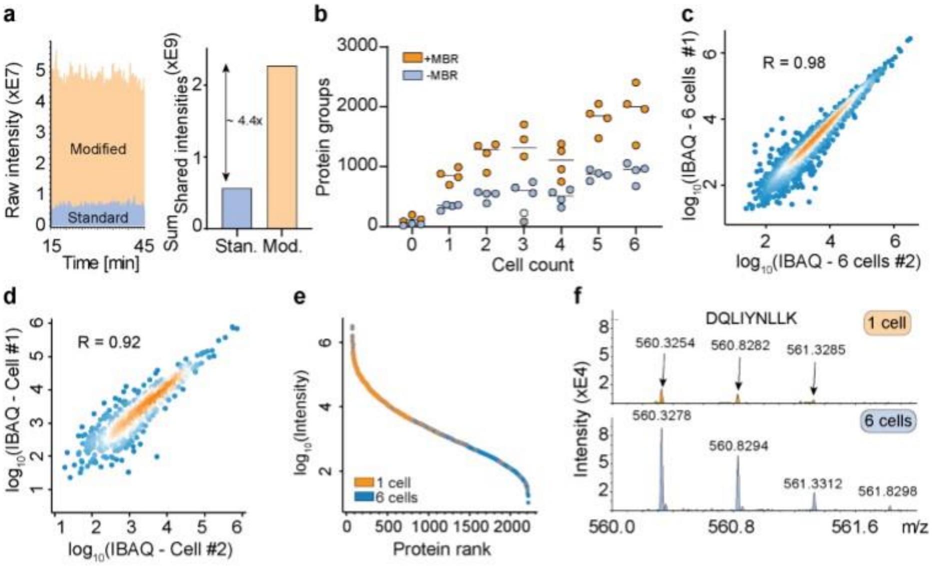


Figure 2: A novel mass spectrometer allows the analysis of true single-cell proteomes. a, Raw signal increase from standard versus modified TIMS-qTOF instrument (left) and at the evidence level (quantified peptide features in MaxQuant) (right). **b,** Proteins quantified from one to six single HeLa cells, either with 'matching between runs' (MBR) in MaxQuant (orange) or without matching between runs (blue). The outlier in the three-cell measurement in grey (no MBR) or white (with MBR) is likely due to failure of FACS sorting as it identified a similar number of proteins as blank runs. **c,** Quantitative reproducibility in a rank order plot of a six-cell replicate experiment. **d,** Same as C for two independent single cells. **e,** Rank order of protein signals in the six-cell experiment (blue) with proteins quantified in a single cell colored in orange. **f,** Raw MS1-level spectrum of one precursor isotope pattern of the indicated sequence and shared between the single-cell (top) and six-cell experiments (bottom).

More than 10-fold sensitivity increase

As electrospray (ES) is concentration dependent, sensitivity increases with decreasing flow-rate, however, very low flow systems are challenging to operate robustly and are consequently not widely available[28–30]. We recently described a chromatography system that decouples sample loading and gradient formation from the LC-MS run and operates at a standardized flow rate of 1 μL/min for high reproducibility[31]. This flow is fully controlled by a single pump instead of the binary gradients produced by other systems. We found that it worked robustly at flow rates down to 25 nL/min but standardized on 100 nL/min, which enabled stable operation for the entire project with the same column-emitter setup (**Extended data Fig. 1a, b**). ES sprayer diameter and gradient length were optimized for turnover, minimizing carry-over and stability.

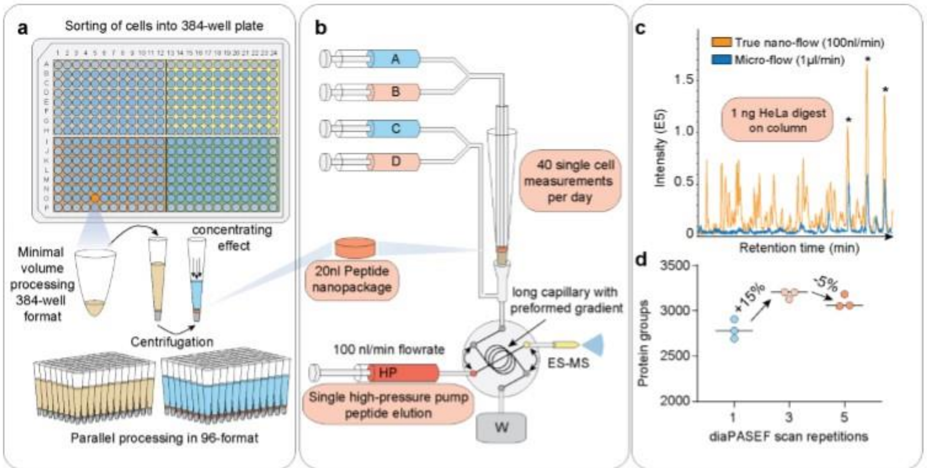

a
Sorting of cells into 384-well plate
Minimal volume processing 384-well format
concentrating effect
Centrifugation
Parallel processing in 96-format
b
A
B
C
D
40 single cell measurements per day
20nl Peptide nanopackage
long capillary with preformed gradient
100 nl/min flowrate
HP
ES-MS
Single high-pressure pump peptide elution
W
c
True nano-flow (100nl/min)
Micro-flow (1µl/min)
1 ng HeLa digest on column
Intensity (E5)
1.5
1.0
0.5
0
Retention time (min)
d
Protein groups
3500
3000
2500
2000
+15%
-5%
1
3
5
diaPASEF scan repetitions

Figure 3: Miniaturized sample preparation coupled to very low flow chromatography and diaPASEF. a, Single cells are sorted in a 384-well format into 1 µL lysis buffer by FACS with outer wells serving as qualitative and quantitaive controls. Single cells are lysed and proteins are solubilized at 72 °C in 20 % acetonitile, and digested at 37 °C. Peptides are concentrated into 20 nL nanopackages in StageTips in a 96-well format. **b,** These tips are automatically picked and peptide nanopackages are eluted in a sub-100 nL volume. After valve switching, the peptide nanopackage is pushed on the analytical column and separated fully controlled by the single high-pressure pump at 100 nL/min. **c,** Basepeak chromatogram of the standardized nano-flow (100 nL/min, orange) and micro-flow (1 µL/min, blue) gradients with 1 ng of HeLa digest on the StageTip. Asterices indicate polyethylene glycole contaminants in both runs. **d,** Nano-flow (100 nL/min) and short gradient diaPASEF method combined. Summation of 1 to 5 diaPASEF scan repetitions was used to find the optimum for high-sensitivity measurements at 1 ng of HeLa digest.

MS-based T-SCP requires loss-less sample preparation by protein isolation and solubilization, followed by tryptic protein digestion and peptide purification ready for MS-analysis[20,21,32]. We found that small volumes of weak-organic solvents in conical 384-well plates provided a versatile and automatable environment for efficient cell lysis and protein digestion in minimal volumes (**Fig. 3a**). Briefly, single cells were sorted into wells containing 1 µL lysis buffer, followed by a heating step and further addition of buffer containing digestion enzymes to a total of 2 µL, all in an enclosed space. Peptides were concentrated in StageTip[33] devices into

20 nL nanopackages, from which they were eluted in minimal volumes (**Fig. 3b**). To benchmark the effect of reduced flow rate and the concentrated peptide nanopackage elution, we directly compared signal traces of the normal 1 µL/min to the 100 nL/min set up. For 1 ng peptide material this resulted in a ten-fold increase in signal (**Fig. 3c**). To achieve high data completeness between hundreds of single-cell measurements, we next replaced ddaPASEF by diaPASEF, in which fragment level matching is further supported by ion mobility data[25]. We found that combining subsequent diaPASEF scan repetitions further improved protein identification numbers. Together, the very low flow chromatography and this diaPASEF acquisition mode resulted in the highly reproducible identification and quantification of more than 3,300 HeLa proteins from only 1 ng **(Fig. 3d)**, a drastic increase from the 550 identified in our initial set-up from a similar amount. Data completeness was at 94% and coefficient of variation (CV) less than 10 % for the selected scan repetition mode (**Extended data Fig. 1b**). This demonstrates that diaPASEF provides its advantages also at extremely low sample amounts, prompting us to adopt this acquisition mode for the single-cell workflow in the remainder of this work.

T-SCP dissects arrested cell-cycle states

The cell cycle is an important and well-studied biological process that has frequently been used as a test case in single-cell studies[34,35]. To investigate if our proteomics workflow could detect biological responses to drug perturbation at the single-cell level, we treated HeLa cells with thymidine and nocodazole to produce four cell populations enriched in specific cell cycle stages (231 cells; **Fig. 4a**). We quantified up to 1,441 proteins per single-cell and 1,596 overall using a HeLa dia spectral library with about 34,000 precursors. This number ranged from a median of 611 in G1 to 1,263 in G1/S, 962 in G2, and 1,106 in G2/M (**Fig. 4b**). To estimate the total protein amount per cell, we summed all protein signals based on their identifying peptides. Judged by protein amount, G2 cells were approximately 1.5-fold larger than G1 cells; thus T-SCP correctly reflected the proliferation state, while highlighting a substantial heterogeneity within each that would have been hidden in bulk sample analysis (**Fig. 4c**). To be able to directly compare single-cell proteomes and cancel out protein abundance differences attributed to varying total protein amounts and identifications of each cell, we normalized our data set by local regression for all proteins with at least 15 % completeness across cells[36]. Furthermore, we stringently filtered our data set for at least 500 protein identifications per cell and more than 20% observations for each protein across remaining single cells (**Extended data Fig. 2a**). The proteomes of the different cell cycle states grouped together in a Principal Component Analysis (PCA) plot (**Fig. 4d**). In addition to these drug-perturbed cells, we measured more than 200 untreated ones from two independent cell culture batches to increase the overall numbers and representative proteome variability of single-cells measured. The T-SCP data set covered proteins assigned to many cellular compartments, membranes and biological processes involved in biological regulation, metabolism, transport, and signal transduction at high quantitative accuracy despite severe systematic perturbation introducing stark biological variation and proteome remodeling (**Extended data Fig. 2b, c; Extended data Table 1**).

Next, we asked whether single-cell proteome measurements can be used to assign cellular states, similar to how single-cell RNA-sequencing (scRNA-seq) measurements have frequently been applied to cell type and state discovery, highlighted by cellular atlas projects[9]. In previous proteomics studies, cell populations had been enriched for cell cycle states and sets of regulated proteins had been extracted[34,35]. We here selected cell cycle stage marker proteins as the top 60 most differentially expressed in either the G2/M-, G1- and S-phase protein set from Geiger *et al.*[35], as it used similar drug treatment on bulk populations and investigated how likely cells from different cell cycle stages could be distinguished (**Extended data Table 2**). We used these marker proteins to set up cell cycle stage specific scores indicating the likelihood to belong to the respective phase previously used for scRNA-seq cell cycle stage predictions. This model clearly distinguished cells from G2/M and G1/S and also other comparisons (**Fig. 4e, Extended data Fig. 2d**)[37].

Next, we investigated the differentially expressed proteins between the drug arrested cell cycle stage transition G2/M and G1/S. Among the significantly regulated proteins was a large number of known cell cycle regulators, some of which are highlighted (**Figure 4f; Extended data Table 3)**. Quantitative MS data at the fragment ion level was highly significant for these as illustrated by the cell cycle regulator CDKN2A and further examples (FDR $< 10^{-14}$, **Fig. 4g**; **Extended data Fig. 3**). We also exemplary inspected precursor isotope patterns of CDKN2A on the raw MS1 and fragment ion series on the raw MS2 spectrum level highlighting the quality of ion signal and noise distribution by TIMS (**Extended data Fig. 4**). Our single-cell data set also highlighted proteins not previously associated with the cell cycle and the G2/M transition. For instance, the putative pseudogene NACAP1 was clearly identified and regulated (FDR $< 10^{-14}$, **Fig. 4h, Extended data Fig. 5**). It might have escaped previous detection because of its small size (213 amino acids) as we have noticed previously[38].

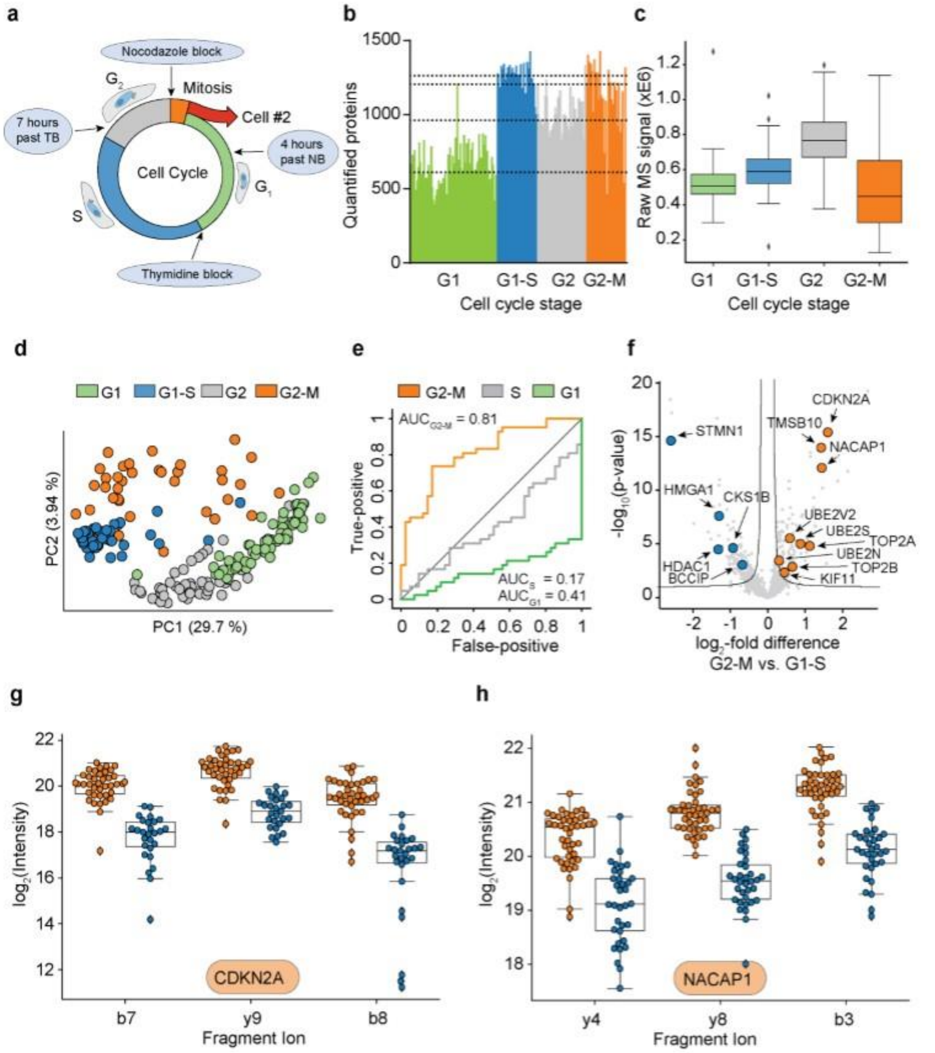


Fig. 4: T-SCP correctly quantifies cell cycle states. a, Arresting single cells by drug perturbation. **b,** Numbers of protein identifications across 231 cells in the indicated cell cycle stages as enriched by the drug treatments in A. **c,** Boxplot of total protein signals of the single cells in B after filtering for at least 500 protein identifications per cell and 20% data completeness per protein across cells (G1: n = 69; G1-S: n = 41; G2: n = 51; G2-M: n = 42). **d,** PCA of single-cell proteomes of B. **e,** Receiver Operator Curve (ROC) for the prediction of G2/M cells against G1/S based on G2/M marker proteins with the indicated area under the curve (AUC) score. The other two curves, based on S and G1 marker proteins, respectively, indicate the inverse predictive power of these scores. **f,** Volcano plot of quantitative protein differences in the two drug arrested states. Arrows point towards colored significantly regulated key proteins of interest. **g,** Quantitative fragment ion level data for CDKN2A and its peptide ALLEAGALPNAPNSYGR (FDR < 10^{-14}). **h,** Quantitative fragment ion level data for the pseudogene NACAP1 and its peptide IEDLSQEAQLAAAEK (FDR < 10^{-14}).

SC proteomes compared to transcriptomes

Given our set of more than 430 single-cell proteomes, we compared the T-SCP measurements after filtering with similar single-cell RNA sequencing data (scRNA-seq)[39,40]. To achieve technology-independent insights, we selected assays from two widespread scRNA-seq technologies, Drop-seq[41] and the lower throughput SMART-sSq2[42], on the same cellular system. The Drop-seq assay is based on unique molecular identifiers (UMIs) to control for amplification biases in library preparation, whereas the SMART-Seq2 assay is not UMI-controlled. Note that MS-based proteomics inherently does not involve any amplification and is not subject to associated artifacts.

Despite subtle differences, HeLa cell culture should reflect a characteristic global distribution of gene and protein expression states[43]. This assumption would allow us to assess self-consistency of the measurement technologies. First, we computed the distribution over all pairwise correlation coefficients of cells within a technology[44]. We found that in the proteome measurement, cells have higher correlation on average than in the droplet-based method and similar correlation to the SMART-Seq2 method (**Extended data Fig. 6a**). On average in 54 % of the 1,596 proteins observed by MS-based proteomics, cells had non-zero expression values and the protein expression completeness per cell followed a normal distribution (**Fig. 5a**). For SMART-Seq2 this dropped to 28 % and to 7 % in the droplet-based protocol. Both single-cell RNA-sequencing technology data sets followed a bimodal gene completeness frequency distribution, while single-cell proteomes do not. (**Extended data Fig. 6b**). Next, we investigated whether there were fundamental limitations of the detection in the protein measurements. Such effects are discussed for scRNA-seq measurements as "drop-out events" or "zero-inflation", although they are now much reduced in UMI-based protocols[44]. We identified signs of such detection limits as bimodality in the lower abundance range of the protein measurements (**Extended data Fig. 6c, d**), suggesting that our single-cell protein analysis could benefit from imputation or tailored likelihood-based parameter estimation methods[45].

For bulk measurements, transcript levels generally correlate moderately with the corresponding protein levels, however, this correlation strongly depends on the biological situation[46]. At the single-cell level this effect is further convoluted by technical differences in the measurement technologies. We asked to what degree scRNA-seq measurements could be used as a proxy for protein measurements in our data and found that protein measurements separate strongly from RNA in a principal component analysis (**Fig. 5b).** Furthermore, single-cell transcript expression levels correlate well across scRNA-seq technologies, but not with single-cell protein measurements (**Fig. 5c, Extended data Fig. 6e**). This suggests that single-cell protein and RNA levels are very different, re-emphasizing that protein measurements yield complementary information to RNA measurements and do not simply re-iterate similar gene expression states. This implies distinct RNA and protein abundance regulation mechanisms on both modalities, dissection of which would not be possible with RNA measurements alone.

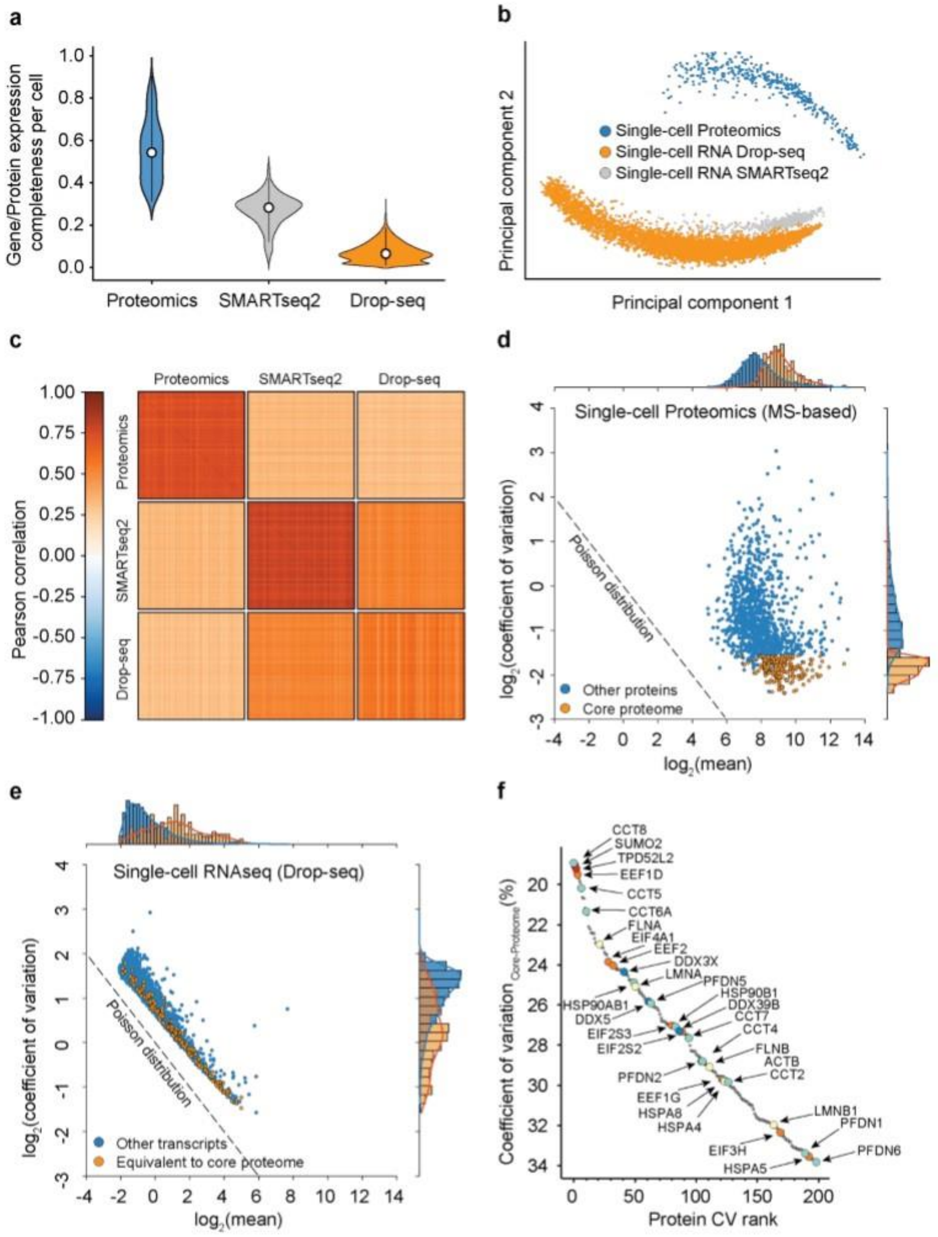


Fig. 5: Single cells have a stable core proteome but not transcriptome. a, Gene or protein expression completeness per cell for either T-SCP (Cells x Proteins: 398 x 1,596), SMARTseq2 (Cells x Genes: 720 x 24,990), and Drop-seq (Cells x Genes: 5,383 x 41,161). **b,** Principal component analysis of single-cell gene and protein expression measurements. **c,** Heat map of cell-cell correlations across individual cells measured by proteomics and by both transcriptome technologies. **d,** Coefficient of variation of protein levels as a function of mean expression levels with the 'core proteome' colored in orange. **e,**

Coefficient of variation of transcript expression values as a function of mean expression levels with transcripts corresponding to the core proteome in orange. **f,** Rank order abundance plot for the core proteome with color coded protein classes (Red: SUMO2 and TDP52L2 proteins; Turquoise: Chaperonin and folding machinery associated proteins. Orange: Translation initiation and elongation; Yellow: Structural proteins; Blue: DEAD box helicase family members).

T-SCP reveals a stable core proteome

Prompted by the divergent correlation values between the proteome and transcript levels, we next investigated the variability of gene expression as a function of abundance. For protein expression measurements, coefficients of variation were very small across covered abundances, independent of the data completeness (**Extended data Fig. 7a**). This is consistent with a model in which the covered proteome is stable and probed deterministically across its full dynamic range. In contrast, the same analysis for UMI-controlled and not UMI-controlled scRNAseq data revealed a much higher overall transcriptome variability, as measured by the coefficient of variation of single-cell RNA-seq compared to protein measurements (**Extended data Fig. 7a, c**). Remarkably, this difference is already very apparent with the current sensitivity of MS-based proteomics, which will surely increase in the future. Comparing single-cell proteome measurements with six-cell proteomes (**Fig. 2c**) suggests that a moderate increase in MS sensitivity would reveal a large part of the proteome to be quantitatively stably expressed. Based on these observations, we defined a 'core-proteome' subset in the MS-based proteomics data by selecting the top 200 proteins with the lowest CVs of the proteins shared between at least 70% of the more than 430 single-cells, including the drug perturbations (**Extended data Table 4**). Interestingly, these proteins where distributed well across the covered dynamic range of the proteome (**Fig. 6d**). Strikingly, we found the corresponding transcripts of the core proteome to be distributed across the full range of CVs in single-cell transcriptome data (**Fig. 6e, Extended data Fig. 7b, c**). The core proteome highlighted proteins frequently used for normalization such as HSP90 and ACTB, providing a positive control (**Fig. 5f**). The CV rank plot of the core proteome also reveals a diverse set of proteins, including representatives of translation initiation and elongation, folding machineries, nucleic acid helicases, as well as cellular structure determining proteins. Interestingly, we also identify TPD52L2 as one of the most stable proteins, which in turn is described as one of the most abundant proteins in HeLa cells[47] and SUMO2, which is known for its involvement in a plethora of essential regulatory cellular processes, suggesting a stable cellular SUMO2-pool even during stark proteome remodelling[48].

Outlook

The T-SCP pipeline combines miniaturized sample preparation coupled to very low flow liquid chromatography and a novel mass spectrometer resulting in at least one order of magnitude sensitivity gain at highest robustness for the analysis of single cells. We quantify cellular heterogeneity following targeted perturbation, which enables the direct analyses of drug-responses in single-cell hierarchies on the proteome level. Furthermore, the comparison of single-cell RNA- and proteome level revealed that the proteome is stable while the transcriptome is more stochastic, highlighting substantial regulation of translation and setting the stage for its elucidation at the single cell level.

Although mainly demonstrated here for single-cell total proteome measurements, the sensitivity gain achieved in our workflow will be advantageous in any situation that is sample limited. This includes investigation of other compound classes such as metabolites or drugs, post-translational modifications from small numbers of cells or from *in vivo* material, measurements directly from paraffin embedded formalin fixed (FFPE) pathology specimens, which we are already pursuing[8].

Our workflow is also compatible with chemical multiplexing with the advantage that the booster channel causing reporter ion distortions could be omitted or reduced. Furthermore, there are many opportunities for increasing overall sensitivity, including even brighter ion sources, improved chromatography and better data analysis and modeling tools, similar to the rapid recent advances in the scRNAseq field.

Acknowledgements. A.-D.B. acknowledges support from the International Max Planck Research School for Life Sciences – IMPRS-LS. S.R. is supported by the Helmholtz Association under the joint research school "Munich School for Data Science" - MUDS. D.S.F. acknowledges support from a German Research Foundation (DFG) fellowship through the Graduate School of Quantitative Biosciences Munich (QBM). F.C. acknowledges the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie individual fellowship under grant agreement 846795). We thank our colleagues in the Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, at the Center for Protein Research at Copenhagen University, and at Bruker Daltonik for discussions and help. In particular, we thank I. Paron, A. Piras and C. Deiml for technical support and J.B. Müller for column production. We also thank Mark Ridgeway, Craig Whitehouse, Andreas Brekenfeld, Niels Goedecke and Christoph Gebhardt of Bruker for their contribution to the development, installation and trouble-shooting of the modified mass spectrometer. We are grateful for the FACS-sorting support by the Imaging Core Facility at the Max Planck Institute of Biochemistry, in particular Martin Spitaler and Markus Oster.

Author contributions. A.-D. B. and M.M. conceptualized and designed the study. A.-D.B., M.C.T., F.C., and A.M. performed experiments. M.A.P. and O.R. designed the new mass spectrometer. A.-D.B., O.B.H. and N.B. conceived the new EvoSep gradient. O.B.H., N.B., A.-D.B. and M.C.T. designed the new EvoSep gradient and optimized it for proteomics performance. D.F. and F.J.T. conceptualized the single cell modeling. A.-D.B., S. R., D. F., F.J.T., C. A., M.C.T., O.B.H., N.B., C.V., F.M. and M.M. analyzed the data; A.-D.B. and M.M. wrote the manuscript.

Competing interests. M.L., O.R., A.A. and M.A.P. are employees of Bruker Daltonik. O.B. H. and N.B. are employees of EvoSep Biosystems. M.M. is an indirect shareholder in EvoSep Biosystems. F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc. and Dermagnostix. All other authors have no competing interests.

References

1. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).
2. Smith, Z. D., Nachman, I., Regev, A. & Meissner, A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nat. Biotechnol. Vol.* **28**, (2010).
3. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
4. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
5. Slavov, N. Unpicking the proteome in single cells. *Science.* **367**, 512–513 (2020).
6. Cheung, T. K. *et al.* Defining the carrier proteome limit for single-cell proteomics. *Nat. Methods* (2020). doi:10.1038/s41592-020-01002-5
7. Liang, Y. *et al.* Fully Automated Sample Processing and Analysis Workflow for Low-Input Proteome Profiling. *Anal. Chem.* (2020). doi:10.1021/acs.analchem.0c04240
8. Mund, A. *et al.* AI-driven Deep Visual Proteomics defines cell identity and heterogeneity Proteomics Program. *bioRxiv* (2021). doi:10.1101/2021.01.25.427969
9. Regev, A. *et al.* The human cell atlas. *Elife* **6**, (2017).
10. Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. *Nat. Protoc.* **12**, 1198–1228 (2017).
11. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science.* **343**, 776–779 (2014).
12. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
13. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science.* **347**, (2015).
14. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
15. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
16. Larance, M. & Lamond, A. I. Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* **16**, 269–280 (2015).
17. Tsai, C. F. *et al.* An Improved Boosting to Amplify Signal with Isobaric Labeling (iBASIL) Strategy for Precise Quantitative Single-cell Proteomics. *Mol. Cell. Proteomics* **19**, 828–838 (2020).
18. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: Mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation 06 Biological Sciences 0601 Biochemistry and Cell Biology 06 Biological Sciences 0604 Genetics. *Genome Biol.* **19**, 1–12 (2018).
19. Brenes, A., Hukelmann, J., Bensaddek, D. & Lamond, A. I. Multibatch TMT reveals false positives, batch effects and missing values. *Mol. Cell. Proteomics* **18**, 1967–1980 (2019).
20. Williams, S. M. *et al.* Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography–Mass Spectrometry for High-Throughput Single-Cell

Proteomics. *Cite This Anal. Chem* **92**, 10588–10596 (2020).

21. Li, Z. Y. *et al.* Nanoliter-Scale Oil-Air-Droplet Chip-Based Single Cell Proteomic Analysis. *Anal. Chem.* **90**, 5430–5438 (2018).
22. Meier, F. *et al.* Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14**, 5378–5387 (2015).
23. Meier, F. *et al.* Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).
24. Meier, F. *et al.* Deep learning the collisional cross sections of the peptide universe from a million training samples. *bioRxiv* (2020). doi:10.1101/2020.05.19.102285
25. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
26. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, (2008).
27. Volpe, P. & Eremenko-Volpe, T. Quantitative Studies on Cell Proteins in Suspension Cultures. *Eur. J. Biochem.* **12**, 195–200 (1970).
28. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**, 1–8 (1996).
29. Emmett, M. R. & Caprioli, R. M. Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. *J. Am. Soc. Mass Spectrom.* **5**, 605–613 (1994).
30. Greguš, M., Kostas, J. C., Ray, S., Abbatiello, S. E. & Ivanov, A. R. Improved Sensitivity of Ultralow Flow LC-MS-Based Proteomic Profiling of Limited Samples Using Monolithic Capillary Columns and FAIMS Technology. *Anal. Chem.* **92**, 14702–14712 (2020).
31. Bache, N. *et al.* A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
32. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).
33. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
34. Ly, T. *et al.* Proteomic analysis of cell cycle progression in asynchronous cultures, including mitotic subphases, using PRIMMUS. *Elife* **6**, (2017).
35. Aviner, R., Shenoy, A., Elroy-Stein, O. & Geiger, T. Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. *PLoS Genet.* **11**, 1005554 (2015).
36. Callister, S. J. *et al.* Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **5**, 277–286 (2006).
37. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene

expression data analysis. *Genome Biol.* **19**, 15 (2018).

38. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science.* **367**, 140–146 (2020).

39. Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M. & Rajewsky, N. The transcriptome dynamics of single cells during the cell cycle. *Mol. Syst. Biol.* **16**, (2020).

40. Hu, W. er *et al.* HeLa-CCL2 cell heterogeneity studied by single-cell DNA and RNA sequencing. *PLoS One* **14**, e0225466 (2019).

41. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

42. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

43. Liu, Y. *et al.* Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **37**, 314–322 (2019).

44. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 147–150 (2020).

45. Risso, D., Perraudeau, F., Gribkova, S. & Dudoit, S. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* (2018). doi:10.1038/s41467-017-02554-5

46. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* doi:10.1038/s41576-020-0258-4

47. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).

48. Gareau, J. R. & Lima, C. D. The SUMO pathway: Emerging mechanisms that shape specificity, conjugation and recognition. *Nature Reviews Molecular Cell Biology* **11**, 861–871 (2010).

Methods

Sample preparation for bulk dilution experiments. For all benchmark experiments purified peptides from bulk HeLa cells were used. HeLa was cultured in Dulbecco's modified Eagle's medium at 10 % fetal bovine serum, 20 mM glutamine and 1% penicillin–streptomycin. Cells were collected by centrifugation, washed with phosphate-buffered saline (PBS), flash-frozen in liquid nitrogen and stored at −80 °C. Cells were resuspended in PreOmics lysis buffer (PreOmics GmbH) and boiled for 20 min at 95 °C, 1500 rpm to denature, reduce and alkylate cysteins, followed by sonication in a Branson, cooled down to room temperature and diluted 1:1 with 100 mM TrisHCl pH 8.5. Protein concentration was estimated by Nanodrop measurement and 500 µg were further processed for overnight digestion by adding LysC and trypsin in a 1:50 ratio (µg of enzyme to µg of protein) at 37 °C and 1500 rpm. Peptides were acidified by adding 1 % trifluoroacetic acid (TFA) 99 % isopropanol (IprOH)in a 1:1 ratio, vortexed, and subjected to StageTip[1] clean-up via styrenedivinylbenzene reverse phase sulfonate (SDB-RPS). 20 µg of peptides were loaded on two 14-gauge StageTip plugs. Peptides were washed two times with 200 µL 1 % TFA 99 % IprOH followed by 200 µL 1 % TFA 99 % IprOH in an in-house-made StageTip centrifuge at 2000 xg and elution with 100 µL of 1 % Ammonia, 80 % acetonitrile (ACN), 19% ddH2O into PCR tubes and finally dried at 60 °C in a SpeedVac centrifuge (Eppendorf, Concentrator plus). Peptides were resuspended in 0.1 % TFA, 2 % ACN, 97.9 % ddH_2O.

Sample preparation for single-cell experiments. HeLa cells were cultured as described above. Supernatant was removed, cells were detached with trypsin-treatment, followed by strong pipetting for cell aggregate dissociation. Cells were washed three times with ice-cold Phosphate-buffered Saline (PBS), pelleted by centrifugation, and the supernatant was removed. For fluorescent-activated cell-sorting (FACS), DAPI was added and sorting performed on the DAPI-negative live cell population. Single cells were sorted into 384-well plates containing 1µl of 20 % acetonitrile (ACN), 100 mM TrisHCl pH 8.5, centrifuged briefly, sealed with aluminum foil and frozen at -80 °C until further use. When needed, single-cell containing 384-well plates were incubated for 30 min at 72 °C in a PCR cycler, followed by 5 min sonication (Elmasonic P) at 37 kHz and room temperature. Protein digestion was performed overnight at 37 °C in a PCR cycler after adding 1 µL of 20 % ACN, 100 mM TrisHCl pH 8.5, 1 ng trypsin/lysC mix. For the peptide bulk and cell count dilution experiments, peptides were resuspended in 4 µL of 2 % ACN, 0.1 % TFA, 97.9 % ddH_2O and injected directly via NanoLC. For all other single cell experiments, samples were dried in a SpeedVac, resuspended in 20 µL 0.1 % formic acid (FA), 99.9 % ddH2O (Buffer A) before transfer into activated EvoTips. These were activated following the standard EvoSep protocol[2]. Then, 50 µL buffer A was added to each EvoTip followed by centrifugation at 200 xg for 1 min. The sample was transferred into the EvoTip, followed by centrifugation at 600 xg for 1min, and two centrifugation steps after adding 50 µL buffer A. Last, 150 µL buffer A was added to each EvoTip and spun for 30 sec at 300 xg.

Cell cycle experiments. The drug-perturbed cell cycle arrest experiment was designed to enrich cells in four cell-cycle stages – G1, the G1/S-transition, G2 and the G2/M-transition. HeLa cells were grown to approximately 30 % confluence as described above, washed and treated for 24 h with 5 mM thymidine, released for 4.5 h and treated again with 5 mM thymidine, or 0.1 µg/mL nocodazole for 13 h. Cells of the G1/S phase (thymidine block) or G2/M phase (nocodazole block) were washed in PBS, trypsinated, subjected to strong pipetting to dissociate cell aggregates and ice-cold PBS washes before DAPI-negative single live cell FACS sorting. A second set of G1/S phase and G2/M phase blocked cells was washed and cultured for 7 h or 2.5 h to enrich early G2 and G1-phase HeLa cells. These were washed with PBS, trypsinated and subjected to DAPI-negative single live cell FACS sorting into 384-well plates pre-loaded with 1 µL 20 % acetonitrile, 100 mM TrisHCl pH 8.5 lysis buffer. Furthermore, we prepared presumable unsynchronized cells sets from two independent cell cultures and subjected them to sample preparation as described below.

High-pH reversed-phase fractionation. To generate a deep library of HeLa precursors for all data-dependent benchmark experiments, peptides were fractionated at pH 10 with the spider-fractionator[3]. 50 µg of purified peptides were separated on a 30 cm C_{18} column in 96 min and concatenated into 24 fractions with 2 min exit valve switches. Peptide fractions were dried in a SpeedVac and reconstituted in 2 % ACN, 0.1 % TFA, 97.9 % ddH_2O for LC-MS analysis.

Liquid-chromatography. For the initial benchmark experiments with HeLa bulk dilution and the cell count dilution, liquid chromatography analysis was performed with an EASY nanoLC 1200 (Thermo Fisher Scientific). Peptides were loaded on a 45 cm in-house packed HPLC-column (75 µm inner diameter packed with 1.9 µm ReproSil-Pur C18-AQ silica beads, Dr. Maisch GmbH, Germany). Sample analytes were separated using a linear 60 min gradient from 5-30 % B in 47.5 min followed by an increase to 60 % for 2.5 min, by a 5 min wash at 95 % buffer B at 300 nL/min and re-equilibration for 5 min at 5 % buffer B (Buffer A: 0.1 % Formic Acid (FA), 99.9 % ddH2O; Buffer B: 0.1 % FA, 80 % ACN, 19.9 % ddH2O). The column temperature was kept at 60 °C by an in-house manufactured oven.

For all other proteome analyses, we used an EvoSep One liquid chromatography system[4] and analyzed the single-cell proteomes with a novel 35 min stepped pre-formed gradient eluting the peptides at 100 nL/min flow-rate. We used a 15 cm × 75 µm ID column with 1.9 µm C18 beads (EvoSep) and a 10 µm ID electrospray emitter (Bruker Daltonik). Mobile phases A and B were 0.1 % FA in water and 0.1 % FA in ACN, respectively.

Both LC systems were coupled online to a modified trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer (timsTOF Pro, Bruker Daltonik GmbH, Germany) via a nano-electrospray ion source (Captive spray, Bruker Daltonik GmbH).

Construction of a novel mass spectrometer with higher sensitivity. We modified our ion source to draw more ions into the vacuum system of the instrument by modifying the glass capillary that conducts gas and ions between the ionization region at atmospheric pressure and the first pumping region of the mass spectrometer. Additional gas is eliminated via an extra pumping stage. Novel prototype ion optics, a high-pressure ion funnel and a radio frequency (RF) multipole confine the ions and transport them to the next vacuum region where the analysis by trapped ion mobility mass spectrometry (TIMS) occurs. The glass capillary is oriented orthogonal to the high-pressure funnel so that neutral contaminants and droplets are first directed away from the funnel by the gas flow. Furthermore, the high-pressure funnel and RF multipole are oriented orthogonal to the TIMS, maintaining the gas dynamics of our original design. Remaining neutral contaminants are guided away from the TIMS entrance. To accommodate the increased ion current, the TIMS analyzer was updated to a new stacked ring (SRIG) design. We use a higher order RF field in the ion accumulation region to create a larger effective ion storage volume than the low order fields of previous designs. A low order quadrupolar field is maintained in the analyzer region to compress the ions towards the analyzer axis during elution to maintain high mobility resolution. The transition between the high order and low order parts of the device was optimized compared to prior designs to further improve peak shape and ion mobility resolution. This results in about a factor of three gain in ion capacity and therefore about a factor of three in the instrument's dynamic range.

Mass spectrometry. Mass spectrometric analysis was performed either in a data-dependent (dda) or data-independent (dia) PASEF mode. For ddaPASEF, 1 MS1 survey TIMS-MS and 10 PASEF MS/MS scans were acquired per acquisition cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to either 50/100/200 ms each and we analyzed the ion mobility range from $1/K_0$ = 1.6 Vs cm^{-2} to 0.6 Vs cm^{-2}. Precursor ions for MS/MS analysis were isolated with a 2 Th window for $m/z < 700$ and 3 Th for $m/z > 700$ in a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at $1/K_0$ = 1.6 VS cm^{-2} to 20 eV at $1/K_0$ = 0.6 Vs cm^{-2}. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonik GmbH). Precursors for MS/MS were picked at an intensity threshold of 1.500 arbitrary units (a.u.) and re-sequenced until reaching a 'target value' of 20.000 a.u. considering a dynamic exclusion of 40 s elution. For DIA analysis, we made use of the correlation of Ion Mobility (IM) with m/z and synchronized the elution of precursors from each IM scan with the quadrupole isolation window. We used the described 100ms ddaPASEF method for the acquisition of a HeLa bulk single-shot library for the single-cell experiments and the short gradient diaPASEF method as described in Meier *et al* [5], but performed up to 5 consecutive diaPASEF cycles before the next MS1-scan (see main text). The collision energy was ramped linearly as a function of the IM from 59 eV at 1/K0 = 1.6 Vs cm^{-2} to 20 eV at 1/K0 = 0.6 Vs cm^{-2}.

Raw data analysis. ddaPASEF data for tryptic HeLa digest dilution series and the cell count experiment were analyzed in the MaxQuant environment (version 1.6.7) and searched against the human Uniprot databases (UP000005640_9606.fa, UP000005640_9606_additional.fa), which extracts features from four-dimensional isotope patterns and associated MS/MS spectra[6,7]. False-discovery rates were controlled at 1% both on peptide spectral match (PSM) and protein levels. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamidomethylation as fixed modification, while limiting the maximum peptide mass to 4,600 Da. Enzyme specificity was set to trypsin cleaving c-terminal to arginine and

lysine. A maximum of two missed cleavages were allowed. Maximum precursor and fragment ion mass tolerance were searched as default for TIMS-DDA data. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs (MBR) with a 0.7-min retention-time match window and a 0.05 $1/K_0$ ion mobility window in case of the single cell-count dilution experiment into a deep ddaPASEF library consisting of 24 fractionations of tryptic HeLa digest. These data were also searched without matching between runs to access the MBR-mediated identification increase. Either intensity-based absolute quantification (IBAQ) or label-free quantification was performed with the MaxLFQ algorithm and a minimum ratio count of one[8].
For all other single-cell experiments, we used a small precursor library consisting of 34,195 precursors mapped to 28,235 peptides and 4,536 protein groups, which was acquired with the 100ms ddaPASEF method described above and generated with the Spectronaut software (version 14.10.201222.47784; Biognosys AG, Schlieren, Switzerland)[9]. A minimum of three fragments per peptide, and a maximum of six fragments were included. All single-cell measurements were searched against the human UniProt reference proteome (UP000005640_9606.fa, UP000005640_9606_additional.fa) of canonical and isoform sequences. Searches used protein N-terminal acetylation and methionine oxidation as variable modifications. We generated one decoy precursor per precursor in the spectral library and used a conservative normal distribution estimator approach for p-value estimation. Protein intensities were normalized using the "Local Normalization" (Q-value = 0.15) algorithm based on a local regression model[10]. A protein and precursor FDR of 1% was used. Default settings were used for other parameters.

Visualization and FDR estimates of fragment ion intensities. Quantitative fragment ion profiles were generated from the Spectronaut output table via the "F.PeakArea" column. Only fragment ions used for quantification in Spectronaut were included (EG.UsedForProteinGroupQuantity = True, EG.UsedForPeptideQuantity = True, F.ExcludedFromQuantification = False). To cancel out cell-size dependent abundance changes, one normalisation factor was estimated per cell, using fold-change based normalization of the whole dataset, as described in the MS-EmpiRe method, which we also used for FDR control[11]. The intensities were log2 transformed and subsequently visualized.

Proteomics downstream data analysis. Proteomics data analysis was performed in the Perseus environment (version 1.6.7, 1.5.5)[12], GraphpadPrism (version 8.2.1) and Python (version 3.8.2). MaxQuant output tables were filtered for 'Reverse', 'Only identified by site modification', and 'Potential contaminants' before further processing. Ontologies for the biological process and cellular compartment assignment for proteins was performed with the mainAnnot.homo_sapiens.txt.gz followed by categorical counting across all proteins for each of the ontologies and counts were exemplary visualized as frequency plot. For single-cell analysis, if not otherwise specified, the Spectronaut data output was filtered first for at least 500 protein observations per cell and at least 20% quantification events across rows and log(x+1)-transformed. The single-cell run with the analysis number 306 was removed since it was identified as full quantitative outlier via PCA analysis. For correlation analysis of two protein expression vectors, transformed gene or protein quantification events of two cells were plotted against each other replacing missing values by zeros. For principal component analysis (PCA), missing values were imputed from a normal distribution with a width of 0.3 standard deviations that was downshifted by 1.8 standard deviations. Differential expression analysis by two-sided unpaired t-test was performed on two groups filtered for at least 50% row-wise quantification events within one group. False-discovery rate control due to multiple hypothesis testing was performed by a permutation-based model and SAM-statistic with an S_0-parameter of 0.3. For cell-size estimation based on raw MS-signal, intensity outputs within cell cycle resolved single-cell proteomics results were summed up and visualized as boxplots. The core proteome was calculated by filtering the whole single-cell proteomics data set for at least 70% quantification events for each protein followed by selection of the top 200 proteins with the smallest coefficient of variation across the dataset.

Single-cell protein and RNA comparison and dropout statistics. The SMART-Seq2[13] data set measured 720 HeLa cells in 3 different batches with a total of 24,990 expressed genes. The Drop-seq[14] data set contained 3 batches with a total of 5,665 cells and 41,161 expressed genes. We performed the single cell analysis with scanpy v1.6.0[15]. If not stated otherwise, we used standardized filtering across all datasets, removed cells with less than 500 genes expressed and removed genes detected in less than 20% of the remaining cells, resulting in 9,900 transcripts in 720 cells in the SMART-Seq2 dataset and 5,000 transcripts and 5,383 cells measured with Drop-seq technology. Ratios of non-zero entries in the scRNAseq datasets and the number of identified proteins in our data are summarized as violin plots. To investigate data completeness across covered dynamic range, we computed the data completeness as a function of the mean log(x+1)-transformed protein abundance of all non-zero/-NaN entries. We included the expected data completeness based on the assumption that missing values are purely due to shot-(Poisson)-noise as 1-exp(-x). For correlation analysis, the RNA abundance entries were linearly scaled to sum to the mean cell size of the respective dataset per cell (230,881.33 for SMART-Seq2 and 6,948.1 for Drop-Seq) followed by log(x+1)-transformation of all abundance entries. Correlation values between the expressions of two cells were computed as the Pearson correlation on the 1073 genes that were shared in all 3 datasets. Entries of missing protein abundance values were excluded from the specific computation. For the PCA plot of technological comparisons, the gene coverage intersection of all technologies (1,073) was isolated, NaNs were replaced by zeros, and expression values were linearly scaled to 1E6 followed by log(1+x) transformation. In coefficient of variation (CV) versus CV plots comparing different technologies as well as the mean versus CV analysis (including the core proteome analysis) and the CV distribution boxplots, RNA expression vectors were scaled to the mean cell size of that measurement technology and mean and CV values were computed per gene under the assumption that single-cell RNA-sequencing data are not zero-inflated[16]. CV (Proteomics) versus CV (RNA-seq) plots show the comparison of cv values of proteins/genes that were shared between all datasets.

Cell Cycle State Prediction. Cell cycle predictions were performed using the scanpy method score_genes[15] based on three sets of proteins that are specifically expressed in the G1- (MARCKS, LMO7, KRT1, GDA, KRT2, HIST1H1E, KRT18, HNRNPA1, DBNL, OGT, CHCHD3, CD44, DBN1, NASP, TARDBP, SH3BGRL3, PODXL, SUMO2, ZYX, STMN1, BAG3, TRIM28, PGRMC1, COASY, EFHD2, SPTAN1), S- (NOLC1, ATP2A2, CANX, CPT1A, TMX1, CKB, SLC25A3, ATP1B3, SLC16A1, MT-CO2, EPHX1, SRPRB, CYB5R3, TECR, LETM1, ANP32B, NUP205), or G2/M-phase (TOP2A, TOP2B, HMGB1, EIF5B, TMSB10, NCAPD2, EIF3D, ANP32A, SELENBP1, BAZ1B, RCC2, S100A4, FASN, HINT1, DKC1, LUC7L2, AARS, KPNA2, CKAP5), respectively. The cell phase specific protein sets were selected based on the z-scored fold-change ratios provided in Geiger *et al.*[17]. The top60 highest differentially expressed genes were selected, but only the aforementioned ones were also identified in our single-cell proteomics data. This scoring method yields the average expression on the provided set of genes minus the average expression on a reference set of genes for each cell. The reference set is chosen to mirror the average expression of the target gene set. For this analysis, cells and genes were filtered, log(x+1)-transformed and missing values replaced by zeros. Plotted are the ROC curves for the three scores corresponding to the three sets of characteristic proteins (G1, S and G2M) used individually to discriminate between the cells of two cell cycle stages.

Data availability. All mass spectrometry raw data, libraries and outputs from each particular search engine analyzed in this study have been deposited to the ProteomeXchange Consortium via the PRIDEpartner repository and are made available to the reviewers.

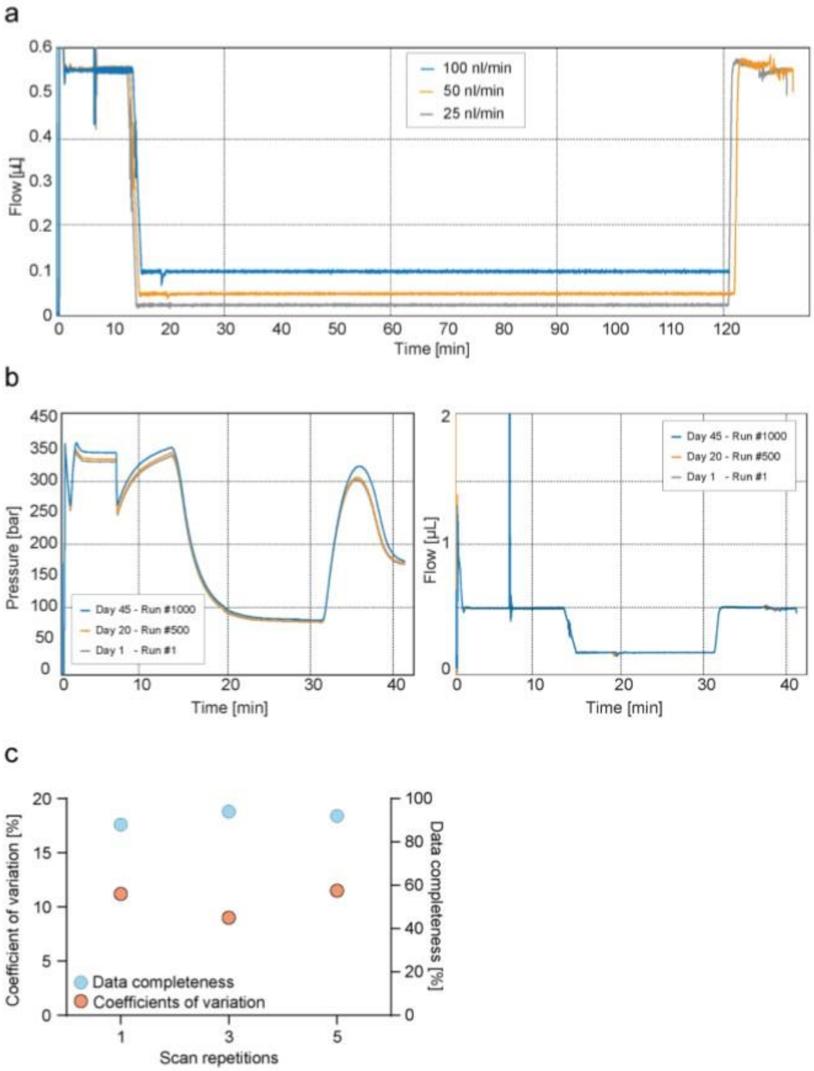
Methods references

1. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
2. Sample loading protocol for Evotips. https://www.evosep.com/wp-content/uploads/2020/08/Sample-loading-protocol_A6_v4_28.02_WEB.pdf.
3. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
4. Bache, N. *et al.* A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
5. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
6. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, (2008).
7. Prianichnikov, N. *et al.* MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics. *Mol. Cell. Proteomics* **19**, 1058–1069 (2020).
8. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
9. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (2015).
10. Callister, S. J. *et al.* Normalization approaches for removing systematic biases

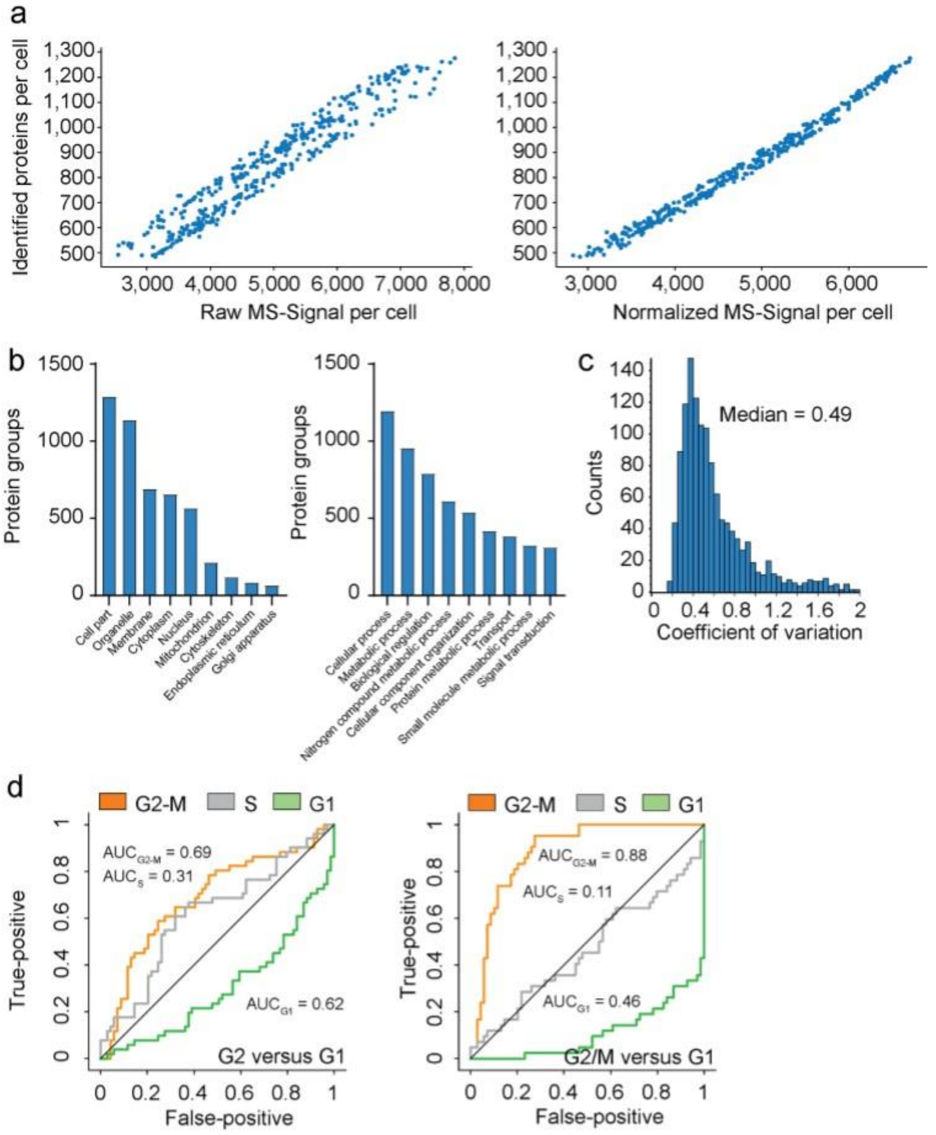
associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **5**, 277–286 (2006).

11. Ammar, C., Gruber, M., Csaba, G. & Zimmer, R. MS-EmpiRe utilizes peptide-level noise distributions for ultra-sensitive detection of differentially expressed proteins. *Mol. Cell. Proteomics* **18**, 1880–1892 (2019).
12. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* vol. 13 731–740 (2016).
13. Hu, W. er *et al.* HeLa-CCL2 cell heterogeneity studied by single-cell DNA and RNA sequencing. *PLoS One* **14**, e0225466 (2019).
14. Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M. & Rajewsky, N. The transcriptome dynamics of single cells during the cell cycle. *Mol. Syst. Biol.* **16**, (2020).
15. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
16. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* vol. 38 147–150 (2020).
17. Aviner, R., Shenoy, A., Elroy-Stein, O. & Geiger, T. Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. *PLoS Genet.* **11**, 1005554 (2015).

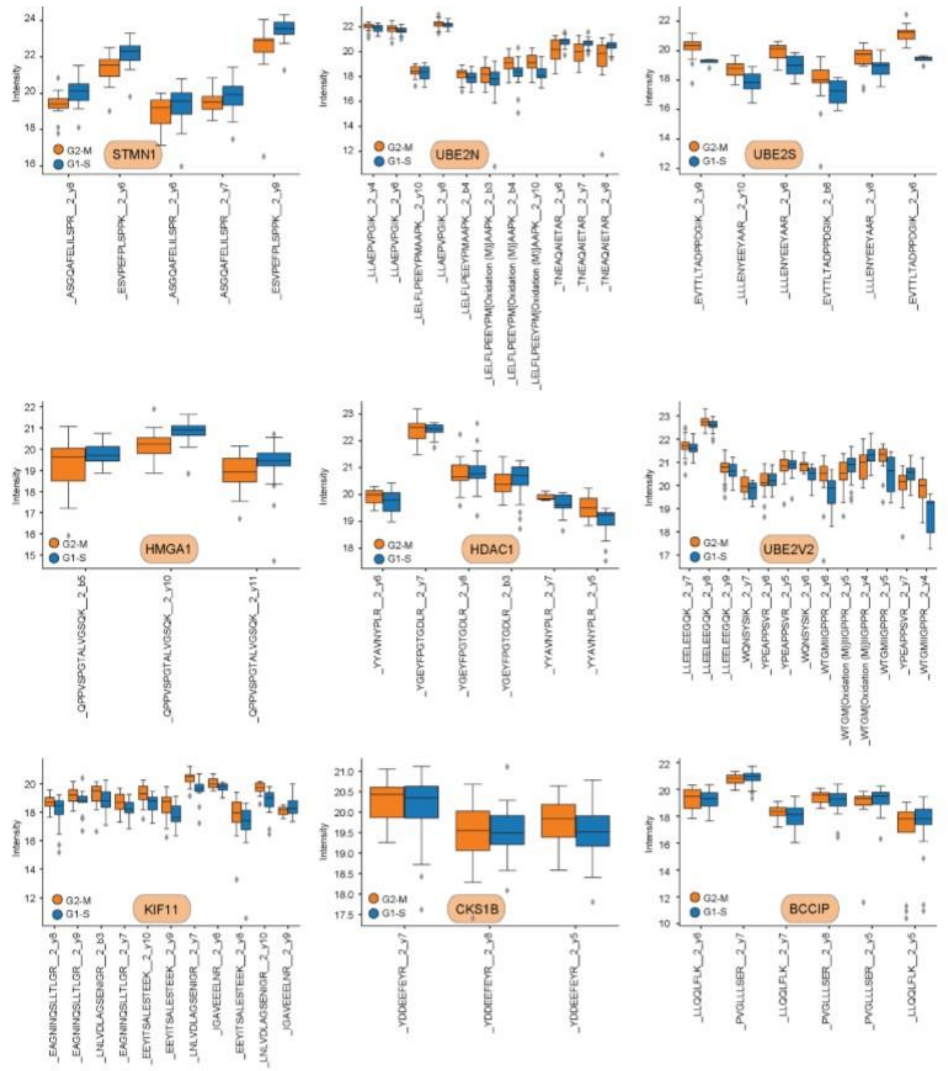
Extended data Figures



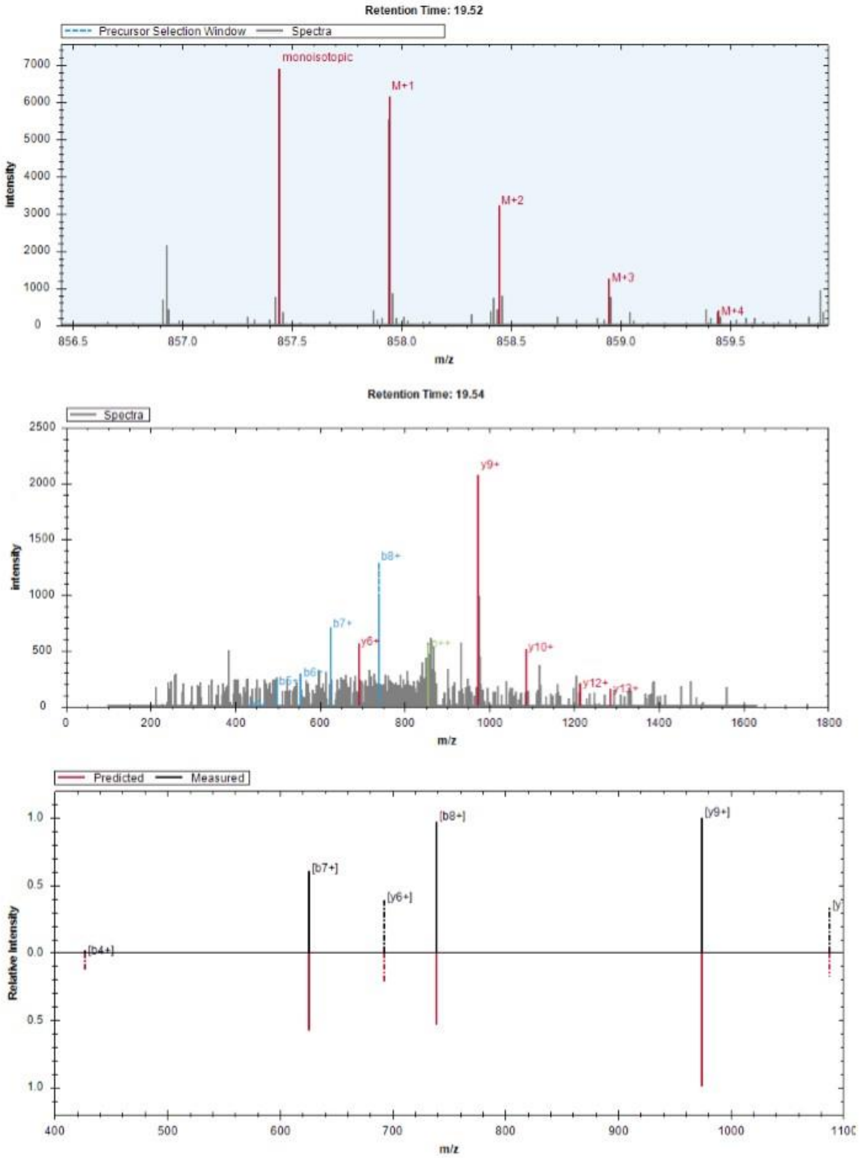
Extended data Figure 1: Robust very low flow rate chromatography and performance evaluation of diaPASEF for ultra-high sensitivity proteomics. a, True nanoflow at 25, 50, and 100 nl/min flow rate on the EvoSep One liquid chromatography system. **b,** Standardized 100 nl/min true nano flow gradient on the EvoSep One liquid chromatography system. Pressure (Left) and flow profile (right) of the gradient of more than 1000 consecutive runs (Day 1 – Run #1 = gray; Day 20 – Run #500 = orange; Day 45 – Run #1,000 = blue). **c,** Data completeness (Blue) and coefficient of variation (Orange) evaluation of different diaPASEF consecutive scan repetitions merged for the analysis of 1 ng tryptic HeLa digest. Scans were varied from one, three, and five repetitions.



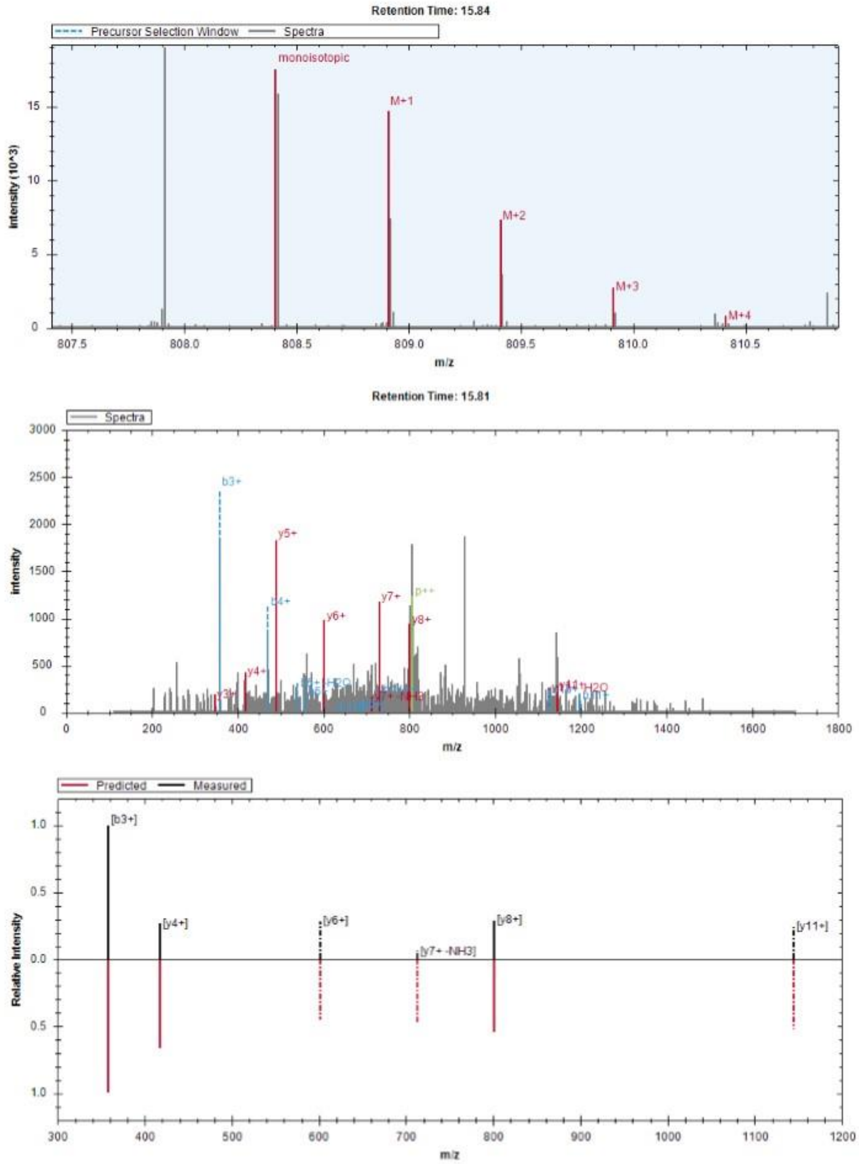
Extended data Figure 2: Qualitative and quantitative investigation of T-SCP data. a, Raw log(x+1)-transformed intensity values of proteins per cell plotted against the number of identified proteins per cell (Left) and after normalization by local regression to cancel out those differences to enable downstream analysis (Right). **b,** Category count of gene ontology annotations for cellular compartment and biological process terms. Exemplary, category count terms are shown for the cellular compartment (left) and biological process (right) for the more than 430 single-cell proteomics data set. **c,** Frequency plot for coefficient of variation occurrence within the 420 single cell proteomics data set. **d,** Cell cycle stage prediction for G2 versus G1 phase cells (left) and G2/M versus G1 phase cells (right) using the top 60 most differentially expressed proteins reported by Geiger *et al.* as input.



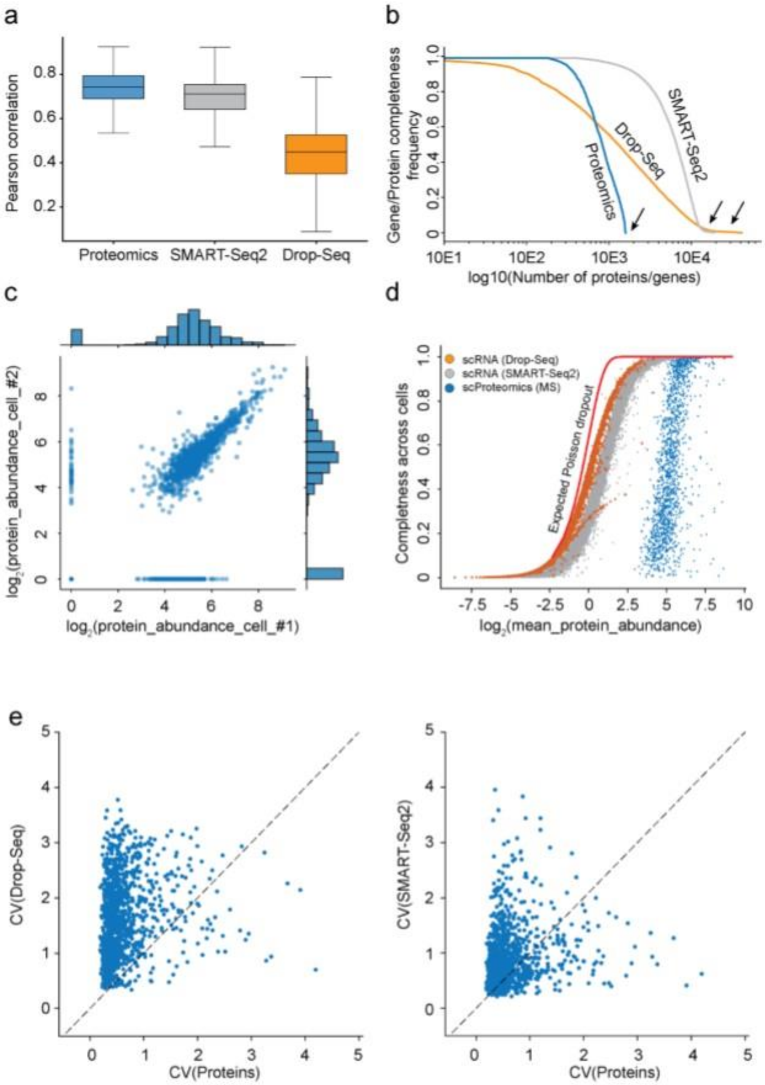
Extended data Figure 3: Fragment ion intensities of peptides for several differentially expressed proteins (UBE2S, FDR = 2.96E-10; STMN1, FDR = 0.00159; HMGA1, FDR = 0.00664; BCCIP, FDR = 0.92523; HDAC1, FDR = 0.83649; UBE2V2, FDR = 0.5152; KIF11, FDR = 2.38E-6; CKS1B, FDR = 0.60734; UBE2N, FDR = 0.7521) in the comparison of nocodazole (G2-M transition) and thymidine (G1-S transition) treated cells. Boxplots represent the intensity distribution of indicated peptide fragment ion intensities (G2-M: red; G1-S: blue).



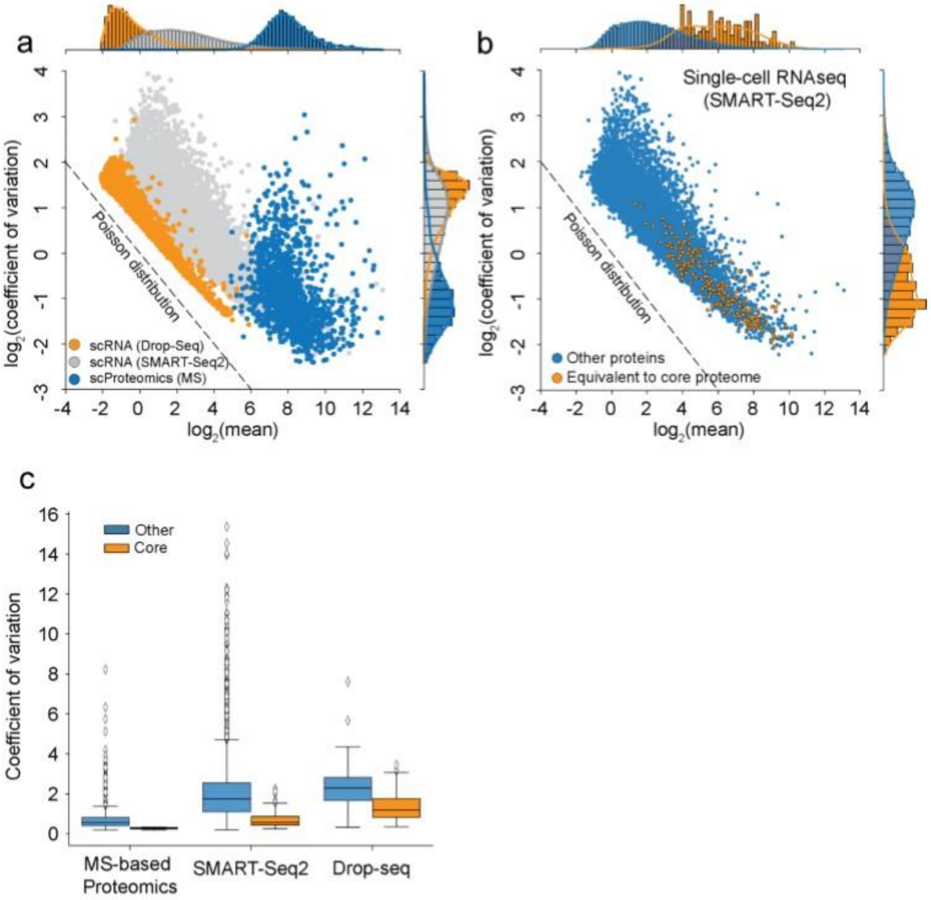
Extended data Figure 4: Exemplary raw MS1 isotope pattern level for the CDKN2A peptide ALLEAGALPNAPNSYGR at MS1 spectrum elution apex (Retention time: 19.52 min; top) and raw MS2 b- (blue) and y- (red) fragment ion series at elution apex (p++ = Unfragmented precursor-ion signal; Retention time: 19.54 min; middle). MS2 intensity correlation of predicted and experimental fragment ion series and intensity values of the same peptide (bottom).



Extended data Figure 5: Exemplary raw MS1 isotope pattern level for the NACAP1 peptide IEDLSQEAQLAAAEK at MS1 spectrum elution apex (Retention time: 15.84 min; top) and raw MS2 b- (blue) and y- (red) fragment ion series at elution apex (p++ = Unfragmented precursor-ion signal; Retention time: 15.81 min; middle). MS2 intensity correlation of predicted and experimental fragment ion series and intensity values of the same peptide (bottom).



Extended data Figure 6: Comparison of single-cell RNA-sequencing and proteomics data. a, Pearson correlation of observations for each cell within each of the technologies (MS-based proteomics, SMART-Seq2 RNA-sequencing, droplet-based RNA-sequencing). **b,** Gene or protein expression frequency occurrence as a function of completeness higher than X for all three technologies (Proteomics, dropSEQ, SMART-Seq2). Arrows indicate a bimodal distribution for single-cell RNAseq data in both technologies, while absent in Proteomics. **c,** Scatter plot of two independently measured single-cell proteome expression values. **d,** Data completeness across single cells as a function of mean protein abundance for MS-based single-cell proteomics and both single-cell RNA-sequencing (Drop-Seq, SMART-Seq2). Expected poison dropout distribution shown in red. **e,** The coefficient of variation of a gene measured by either Drop-Seq technology (left) or SMART-Seq2 (right) compared to the coefficient of variation of the corresponding protein measured by MS-based single-cell proteomics.



Extended data Figure 7: Single cells have a stable core proteome, but not transcriptome. a, Coefficient of variation distribution as a function of mean gene or protein intensities for either Drop-Seq (orange), SMART-Seq2 (gray), or MS-based single-cell proteomics (blue). Expected Poisson distribution shown as dashed line. **b,** Coefficient of variation of single-cell RNA-sequencing (SMART-Seq2) levels as a function of mean expression levels with the 'core proteome' colored in orange and non-'core proteome' genes in blue. Expected Poisson distribution shown as dashed line. **c,** Coefficients of variation for each protein or gene expressed across the data set for either MS-based single-cell proteomics, SMART-Seq2 and Drop-Seq dataset. Proteins or genes of the 'core-proteome' are shown in orange, others in blue.

3.4. Image-guided spatial and cell-type resolved proteomics

The development of the ultra-high sensitivity mass spectrometry workflow as described before enabled the analysis of single-cell proteomes upon drug perturbation, which opens the door for the elucidation of drug-response assays at the level of single cells in a biomedical setting. Today, single-cell sequencing allows the mapping of cell types and states, uncovering a tremendous complexity[395]. However, proteins are the drivers of cellular function and single-cell proteomics will therefore extensively complement the elucidation of cellular heterogeneity, leading to a better understanding of developmental and pathological processes[65]. Keeping in mind that many pathologies start in unknown areas of organs, cell types, or even subcellular regions, it would be highly desirable to combine our ultra-high sensitivity MS single-cell technology with unbiased imaging approaches to automatically locate the pathological region, excise it without contaminating surrounding tissue and subject it to unbiased proteome analysis. This would allow us to connect the visual dimension in 3D with the molecular phenotype in its native environment, determine drivers in health and disease, and suggest treatment options to eliminate diseased tissue or even restore function. We realized this vision in two independently developed approaches called *DISCO-MS: Proteomics of spatially located target tissues in whole organs* and *DVP: Deep Visual Proteomics*, which I imagine to be combined in the future:

DISCO-MS, a very close collaborative effort with the laboratory of Ali Ertürk at the Helmholtz center Munich, combines solvent-based tissue clearing, whole-organ imaging by light-sheet microscopy, automated image analysis powered by deep learning and ultra-high sensitivity mass spectrometry. In solvent-based tissue clearing, whole rodent bodies or isolated organs are rendered translucent by several shrinkage-mediating organic solvent extraction steps resulting in a rigid sample, which can then be subjected to unbiased imaging, followed by *in silico* 3D image reconstruction[421]. Since many pathologies arise in unknown regions, whole-organ imaging holds promise to locate them in an unbiased way, followed by laser capture microdissection isolation and unbiased proteomics analysis.
First, we asked if proteomes of solvent-cleared tissues artificially changed qualitatively or quantitatively upon the diverse and harsh clearing steps and how they compare to their fresh or PFA-fixed counterparts. To do so, I developed a bottom-up proteomics sample preparation workflow aiming for highest protein recovery. Indeed, fresh and paraffin -fixed (PFA) tissue subjected to several clearing techniques yielded quantitatively and qualitatively very comparable proteomes at a depth of up to 6,000 proteins. Deep proteome comparisons at 8,000 proteins in mouse brains showed high quantitative reproducibility with Pearson correlations of up to 0.94. We also verified the applicability of the

protocol to other rodent organs besides brain (lung, heart) with a very different proteome dynamic range and archival human brains, obtaining equal proteome quality. Furthermore, we showed that summed abundance changes of protein groups annotated by gene ontology for 'cellular compartment' were well below 15 % between fresh and cleared organs, highlighting high proteome preservation. The exception was 'Blood microparticle' proteins which can easily be explained as rodents undergo cardiac blood-flushing and PFA-fixation before the actual clearing process starts. Interestingly, even aggregated membrane protein abundances changed less than 3 % change in cleared organs compared to fresh counterparts, suggesting that DISCO-MS could identify novel surface markers for drug targeting. Astonishingly, rodent whole-organ clearing is also very reproducible on the proteome level across biological replicates with median CVs of 20 % across the full dynamic range, just like their fresh organ counterparts.

Next, we investigated if we could isolate pathological regions of interest by LCM, which have been located in the cleared organ by AI-guided image analysis followed by MS-based proteomics. If done manually, it would have taken years of hands on work to finish the imaging analysis and even then, very early-stage pathologies would most likely have been missed. Two challenges had to be solved to enable this workflow, first the reliable dissection of small tissue regions comprising a volume of less than 100 cells and second to obtain a proteome from only a few nanograms of dissected rigid solvent-cleared tissue. Both hurdles were taken by literally reversing the last tissue clearing protocol steps to rehydrate the sample and combining this with miniaturization of the proteomics sample preparation. We explored DISCO-MS first on a mild traumatic brain injury mouse model, isolating locally inflamed tissue regions and their counterparts in controls. This revealed very distinct proteomes and recapitulated biomarkers known to be involved in injury and tissue recovery in the mTBI isolates.

We then turned to a familial Alzheimer's disease (FAD) mouse model to measure all early-stage Aβ-plaques in an unbiased way, followed by their equally unbiased proteome analysis using DISCO-MS. First, since the locations of early-stage plaque are unknown, we automatically localized them in whole FAD mouse brains in week 5 after birth by imaging combined with deep learning and showed that they initially preferentially localized towards the hippocampal region. Interestingly, we were able to detect them already at a volume of 2,000-3,5000 μm^3, highlighting the sensitivity of our methods. We then isolated some of these early-stage plaques as well as non-FAD littermate control tissue at corresponding coordinates and subjected them to MS-based proteome analysis. This quantified more than 1,900 proteins across replicates, determined the core proteome of early-stage beta-amyloid plaques including known and potentially novel biomarkers comprising oligopeptidases/-isomerases, calcium binding protein family members, while confirming that Amyloid-beta deposition plays a key role in early Aβ plaque development. Furthermore, we highlighted a substantial compositional proteome heterogeneity across deposits[422]. Together, this provided valuable insights into the initial stages of AD in this mouse model.

DISCO-MS successfully combines unbiased imaging of whole rodent bodies or organs with the unbiased proteome analysis of AI-determined and LCM-isolated target tissue regions. It also yields qualitative and quantitative proteomics data nearly indistinguishable from uncleared samples in both rodent and human tissue, confirmed many known and revealed potentially novel biomarkers in two disease models. This sets the stage for the elucidation of very early-stage pathologies and accompanying therapeutic interventions. One remaining drawback of the current implementation of this technology is that pathological target regions are isolated in a small volume comprising cells, extracellular matrix and other depositions. This inevitably results in a merged proteome of all constituents within this volume, but is addressed by the technology described next.

Deep Visual Proteomics combines high-resolution microscopy, deep learning-based image recognition, cell segmentation and identification of cell phenotypes, coupled to automated LCM-based isolation of single cells or cell states followed by the ultra-high sensitivity proteomics workflow developed and described above. This concept promises to link protein abundance to complex cellular and subcellular phenotypes while preserving the spatial meta-information of cell cultures and FFPE-embedded tissue thin sections.

The key challenge to realize DVP was the AI-driven accurate definition of single cell boundaries and cell classes from high-resolution whole-slide images, which is the specialty of our close collaborators, the Peter Horvath group at the Hungarian Academy of Sciences. Furthermore, the transfer of the resulting coordinates from the scanning to a laser microdissection microscope turned out to be a major challenge. We solved these issues in the software called BIAS (Biology Image Analysis Software), which combines image preprocessing, deep learning-based image segmentation, machine-learning driven morphological and neighborhood feature extraction, plus phenotype classification. It also allows to transfer defined features, or cell contours from a scanning to a laser microdissection microscope to physically isolate cells at full accuracy by the integration and conversion of microscope data formats. After establishing this workflow, we benchmarked it in several applications. Since deep learning methods require large training datasets and we are limited by the size of high-quality training data, we used project-specific image style transfer to synthesize artificial microscopy images. This allowed us to train the cell segmentation algorithm called NucleAIzer to define cells and cellular compartments of interest at highest accuracy[423,424].

In a first proof of principle application of DVP, we aimed to characterize the proteome of minute numbers of phenotypically different cells (80-100) and nuclei (250-300) of an unperturbed U2OS cell line. Replicate analyses of cells and nuclei showed high quantitative reproducibility (R = 0.96), while cellular and nuclei proteomes were different (R = 0.84) proving the idea of analyzing distinct proteome subsets by direct LCM-mediated biological or cellular fractionation.

Next, we asked if DVP can define distinct nuclei classes based on morphological differences in nuclear area, perimeter, form factor, solidity and DNA staining intensity. The goal was to investigate if those phenotypes are reflected by their proteomic makeup. Indeed, we were able to define six distinct classes, which we isolated and analyzed by MS-based proteomics revealing quantifiable proteomic differences in these subcellular phenotypes. Interestingly, this experiment revealed that these morphological differences of the nucleus reflects in many cases the cell cycle progression state and even uncovered a characteristic nuclear deformation in the context of cell migration. Furthermore, we identified many differentially expressed uncharacterized proteins, which could be associated with a potential cellular function driving the actual phenotype differences.

In a second experiment, we asked if DVP allows the precise and unbiased proteomic profiling of distinct cells or subcellular compartments preserving their spatial context in tissue, since this would enable the analysis of archived pathological samples. To do so, we developed an immunohistochemical protocol, stained archived tissue thin sections of salivary gland acini cell carcinoma tissue sections with EpCAM to define cell boundaries and subjected it to our BIAS software for cell segmentation. We aimed to directly compare normal appearing to malignant cell isolates only with DVP, which should reveal distinct proteomic signatures without admixing unrelated cell populations as in current studies. Indeed, within group quantitative reproducibility was very high (R > 0.96), while between groups correlations were much lower (R = 0.8). We also found that protein markers of the healthy tissue were downregulated in neoplastic tissue, confirming our approach.

In a third experiment, we asked if DVP can also resolve different states of the same cell type simply based on spatial differences in a pathological setting, thereby highlighting differences in proteome signatures. We turned to 18-years old archived thin sections of melanoma tissue from which we isolated cells double-positive for melanoma markers from the inner tumor and out tumor mass with close proximity to the stroma. Indeed, central and peripheral proteomes were very different, while e.g. peripheral melanoma cells revealed an upregulation of favorable prognostic proteins for immune related processes.

In summary, we established a workflow that can automatically excise up to 700 cell contours per hour and it allows to correlate cellular phenotypes with the proteome level in an unbiased way. This enables the investigation of rare cell types and even subcellular structures in their natural tissue environment.

I imagine that in the future, DISCO-MS and DVP could merge for many applications. This would allow us to define unknown pathological target regions by unbiased macroscopic whole-organ imaging, microscopic isolation of the target tissue and ultimately cell type- or structure-resolved isolation by LCM followed by ultra-high sensitivity proteomics without the admixture of surrounding constituents. Both technologies have the potential to already discover disease-specific and therapeutically relevant proteins on the basis of target tissue (DISCO-MS) or cell-type (DVP) resolved proteomics and could serve as starting points in drug development.

3.4.1. Article 8: DISCO-MS

DISCO-MS – Proteomics of spatially identified tissues in whole organs

bioRxiv, April, 2021 (To be submitted)

Harsharan Singh Bhatia[1,2]*, **Andreas-David Brunner[3]*,** Zhouyi Rong[1,2], Hongcheng Mai[1,2], Marvin Thielert[3], Rami AL-Maskari[2,6], Johannes Christian Paetzold[6], Florian Kofler[6,7], Mihail Ivilinov Todorov[1,2,4,] Mayar Ali[1], Bjoern H Menze[6,], Muge Molbay[1,2,8], Zeynep Ilgin Kolabas[1,2,4], Doris Kaltenecker[2,9], Ingo Bechmann[10], Fabian J Theis[11], Stephan Müller[12], Stefan Lichtenthaler[4,5,12,13], Matthias Mann[3,14#], Ali Ertürk[1,2,4,5#]

** These authors contributed equally to this work*

# Correspondence

[1]Insititute for Tissue Engineering and Regenerative Medicine (iTERM), Helmholtz Zentrum München, 85764 Neuherberg, Germany

[2]Institute for Stroke and Dementia Research, Klinikum der Universitaẗ Mu̇nchen, Ludwig-Maximilians University Munich, Munich, Germany

[3]Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, 82152 Martinsried, Germany

[4]Graduate School of Neuroscience (GSN), 82152 Munich, Germany

[5]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

[6]Center for Translational Cancer Research (TranslaTUM) of the TUM, 80798 Munich, Germany; Image-Based Biomedical Modeling, Department of Informatics, Technical University of Munich, Munich, Germany

[7]Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany

[8]Munich Medical Research School (MMRS), Munich, Germany

[9]Institute for Diabetes and Cancer, Helmholtz Center Munich, Neuherberg, Germany

[10]Institute of Anatomy, University of Leipzig, 04109 Leipzig, Germany

[11]Institute of Computational Biology, Helmholz Zentrum München, Neuherberg, Germany; TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany; Department of Mathematics, Technical University of Munich, Garching, Munich, Germany

[12]German Center for Neurodegenerative Diseases (DZNE), Munich, Germany

[13]Neuroproteomics, School of Medicine, Klinikum Rechts der Isar, Technical University, Munich, Germany

[14]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Contribution

I developed the sample preparation workflow to enable bottom-up proteomics from solid solvent-cleared organs and tissue resulting in qualitative and quantitative data nearly indistinguishable from fresh of PFA-fixed tissue. Furthermore, I optimized this workflow for the processing of laser capture microdissection (LCM) isolates comprising a volume of less than 100 cells. I also transferred the knowledge in ultra-high sensitivity proteomics analysis of true single-cells to this project and acquired high quality data in an Alzheimer's and mild traumatic brain injury disease model. I performed all proteomics data analysis and related figures. Furthermore, I helped to conceptualize and design the study, contributed to the overall experimental design and helped to write the manuscript.

DISCO-MS: Proteomics of spatially identified tissues in whole organs

Harsharan Singh Bhatia[1,2]*, Andreas-David Brunner[3]*, Zhouyi Rong[1,2,9], Hongcheng Mai[1,2,9], Marvin Thielert[3], Rami Al-Maskari[2,6], Johannes Christian Paetzold[6], Florian Kofler[6,8], Mihail Ivilinov Todorov[1,2], Mayar Ali[1,4], Muge Molbay[1,2,9], Zeynep Ilgin Kolabas[1,2,4], Doris Kaltenecker[2,10], Stephan Müller[11,12], Stefan Lichtenthaler[4,5,11,12], Bjoern H Menze[6,7], Fabian J Theis[13,14], Matthias Mann[3,15#], Ali Ertürk[1,2,4,5#]

[1]Insititute for Tissue Engineering and Regenerative Medicine (iTERM), Helmholtz Zentrum München, 85764 Neuherberg, Germany.

[2]Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians University Munich, Munich, Germany.

[3]Department for Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, 82152 Martinsried, Germany.

[4]Graduate School of Neuroscience (GSN), 82152 Munich, Germany.

[5]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany.

[6]Center for Translational Cancer Research (TranslaTUM) of the TUM, 80798 Munich, Germany; Image-Based Biomedical Modeling, Department of Informatics, Technical University of Munich, Munich, Germany.

[7]Department for Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

[8]Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany.

[9]Munich Medical Research School (MMRS), Munich, Germany.

[10]Institute for Diabetes and Cancer, Helmholtz Center Munich, Neuherberg, Germany.

[11]German Center for Neurodegenerative Diseases (DZNE), Munich, Germany.

[12]Neuroproteomics, School of Medicine, Klinikum Rechts der Isar, Technical University, Munich, Germany.

[13]Institute of Computational Biology, Helmholz Zentrum München, Neuherberg, Germany

[14]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany; Department of Mathematics, Technical University of Munich, Garching, Munich, Germany.

[15]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark.

*These authors contributed equally to this work

One Sentence Summary:
A spatial proteomics technology enabling unbiased molecular analysis of target tissues after unbiased imaging of whole organs

#Correspondence:
erturk@helmholtz-muenchen.de & mmann@biochem.mpg.de

Abstract

Recent advances in imaging, transcriptomics, and proteomics now allow molecular profiling of complex tissues. However, methods for the unbiased molecular analysis of tissue regions identified by global imaging are lacking. Here, we present DISCO-MS, a technology utilizing solvent-based tissue clearing, whole-organ imaging by light-sheet microscopy, automated image analysis powered by deep learning, and ultra-high sensitivity mass spectrometry. DISCO-MS yields qualitative and quantitative proteomics data indistinguishable from uncleared samples in both rodent and human tissue. We identified pathological tissue regions in mild traumatic brain injury and Alzheimer's disease mouse models, revealing their compositional heterogeneity and novel biomarkers. DISCO-MS thus enables quantitative, unbiased proteome analysis of target tissues guided by unbiased imaging of entire organs, providing new diagnostic and therapeutic opportunities for complex diseases, including neurodegeneration.

At their early stages, many diseases start with a modest pathology in mostly unknown tissue regions making them hard to identify and characterize. For example, early changes in dementia may include the activation of few local inflammatory cells, changes in the microvasculature, and the appearance of just a few initial amyloid-beta plaques in unknown brain regions (*1*). Such small regional changes are extremely hard to be identified using standard histology, limiting our ability to investigate initial stages of the disease for early diagnostic and therapeutic opportunities. Recent advances in tissue clearing technologies allow fluorescence imaging of complete biological tissues including mouse organs and whole bodies, as well as intact human organs (*2–4*). After entire tissues are rendered transparent, end-to-end laser scanning microscopy reveals their cellular and sub-cellular level details. Leveraging artificial intelligence (AI)-guided image analysis, years of imaging work can now be completed within days, allowing to capture even tiny changes in cellular structures, which would be missed in otherwise pre-selected tissue sections (*5*)(*6*). However, visually pinpointing these regions alone does not answer molecular level, mechanistic questions. Here, we report a new technology in which we apply ultra-high sensitivity MS-based proteome analysis to target tissues comprising less than 100 cells after their identification by organic solvent-based organ clearing and imaging (3-Dimensional Imaging of Solvent Cleared Organs profiled by Mass Spectrometry: DISCO-MS). Using DISCO-MS, we analyzed proteomes of rare pathological regions with known spatial location and morphology in whole mouse brain scans, which provides unique opportunities for spatial-molecular profiling of intact organs.

MS-based proteomics of solvent-cleared tissue

Tissue clearing enables rapid imaging of intact organs at the cellular level without sectioning. It is a chemical process that relies on tissue permeabilization and subsequent extraction of different biomolecules including water (in organic-solvent-based DISCO clearing methods) and lipids (in most clearing methods) (*2*). We asked how the proteome of solvent-cleared tissues changes after these diverse and harsh extraction steps. Toward this goal, we turned to MS-based proteomics, which can provide unbiased in-depth insights into the composition, structure, and function of the entirety of expressed proteins (*7*).

We first explored the feasibility of subjecting solvent-cleared rigid tissue to proteomics sample preparation workflow regarding protein recovery, qualitative and quantitative reproducibility. Using fresh-frozen tissues as controls, we started with 3DISCO and uDISCO tissue clearing methods, two commonly used organic solvent-based transparency methods that are quick and known to provide the highest tissue and organ transparency (*2*). We tested several protein solubilization approaches based on: Sodiumdeoxycholate (SDC) alone, 2,2,2-Trifluoroethanol (TFE) + Sodiumdodecylsulfate (SDS), and SDS + SDC (See methods for in-depth description). Remarkably, the combination of SDS-based protein solubilization and tissue pulverization, several boiling steps, and acetone-based precipitation/washing, followed by SDC-resolubilization and protein digestion, yielded qualitatively and quantitatively very similar proteomes among fresh and these two organic solvent-based clearing methods. We identified up to 5,500 proteins across conditions with Pearson correlation coefficients between 0.89 and 0.99. Note that this minimal quantitative variability also reflects biological differences as the cleared brains came from different mice (**Fig. S1**).

While 3DISCO and uDISCO work well for fluorescent dye imaging, the signal of endogenous fluorescence proteins such as EGFP can decay before panoptic imaging (*8*). To avoid this, we recently developed vDISCO, which uses fluorescent dye-conjugated nanobodies to stabilize and enhance endogenous fluorescence protein signals (*9*). As vDISCO includes several additional tissue labeling and clearing steps that might change the proteome, we tested our proteomics workflow also on tissues subjected to this technique. Identification of more than 6,000 proteins across all clearing conditions including vDISCO and replicates with Pearson correlations ranging from 0.85 to 0.94 confirmed the feasibility of our workflow (**Fig. 1A**). Next, we asked if we can also analyze archival human brain tissues with the same protocol. To this end, using 3DISCO, uDISCO and SHANEL methods (*3*), we cleared human brain tissues stored in formalin for more than five years and subjected them to our workflow. Encouragingly, we identified more than 5,000 proteins in all clearing conditions compared to PFA-fixed controls at high quantitative reproducibility (R = 0.91-0.96) (**Fig. 1B**). We concluded that our new sample preparation workflow allowed the MS-based proteome analysis of cleared mouse and human organs at qualitative depth and quantitative accuracy comparable to fresh and PFA-fixed control samples.

High proteome yield in vDISCO cleared tissues

Next, we examined the proteome of cleared tissue in-depth to investigate potential protein depletions introduced by the clearing process. Three biological replicates of vDISCO-cleared mouse brains and three fresh-frozen control brains (C57BL/6J) were subjected to our sample preparation workflow for bottom-up proteomics, separated into 16 fractions each and subjected to MS-based proteome analysis and data mining (**Fig. 2A**). This workflow identified close to 8,000 proteins in total across conditions and biological replicates with a high quantitative reproducibility between conditions across the full dynamic range (R = 0.94; **Fig. 2B, Fig. S2**). Furthermore, coefficients of variation within fresh and clearing conditions were below 0.2, highlighting that vDISCO clearing yields proteomes which are qualitatively and quantitatively in the same range as fresh tissue and very reproducible across biological replicates (**Fig. S2**). The only altered gene ontology (GO) term was 'Blood microparticle' proteins, which was expected as tissues were perfused, while other GO keywords, such as 'Aging', 'Neurogenesis', 'Immunity', 'Wound healing', 'Virus-Host', 'Neurodegeneration', and 'Receptor' are quantitatively and qualitatively preserved (**Fig. 2C, Fig. S3**). Cellular proteomes can be quantified by the proteome ruler algorithm, which uses the fixed ratio of total histones to the genome (*10*) allowing us to relate gene ontology annotations of subcellular localization to protein mass distributions between fresh and vDISCO mouse brains for membrane, organelle, and cytoskeleton terms. We found that all percentage protein mass differences were well below 15% across sub-terms and the protein mass change associated with membrane-related terms was below 3% (**Fig. 2D**). In summary, even the strongest organic solvent-based tissue clearing approach, vDISCO, yields very similar proteomes compared to the fresh tissue.

Proteomes of micro-dissected tissues imaged in 3D

After establishing a high-quality and reproducible MS-compatible sample preparation workflow for solvent-cleared tissues, we turned to the proteome analysis of tiny target tissue regions previously imaged and located in 3D. Here, we encountered two major challenges: 1) reliable dissection of small tissue regions identified by 3D-imaging of cleared tissues, and 2) measuring a deep proteome from only a few nano-grams of dissected and rigid solvent-cleared tissue. To solve the first problem, we developed a series of steps to render cleared rigid tissue soft for precise cryosectioning and laser capture microdissection (LCM) without deformation. In short, we reversed the clearing protocol, rehydrated the cleared tissue stepwise, and quickly cryo-preserved it using isopentane in a sucrose bed. This workflow avoided rupturing of the tissue during cryosectioning and allowed us to laser micro-dissect tissue regions as small as 0.0005 mm^3 corresponding to approximately 60 cells in volume. Next, we miniaturized our sample preparation workflow and then performed MS-based proteomics analysis on a modified trapped ion mobility mass spectrometry platform developed to the highest sensitivity as we recently described for the analysis of single FACS sorted cells in *Brunner* et. al (*11*).

To explore the potential of our technology in clinically relevant applications, we first used a mild traumatic brain injury (mTBI) mouse model to identify and analyze proteomes of brain regions depicting discrete local inflammation. mTBI, including concussions, are common and can lead to long-term comorbidities such as sleep disorders, neuropsychiatric disorders, and even early onset of dementia (*12*). They are characterized by chronic inflammation, which can induce neurodegeneration in selected brain regions, particularly along the stretched axonal tract (*13*). We used a repetitive mTBI injury model on CX3CR1-EGFP mice (**Fig. 4SA−D**), in which all microglia are labelled with an EGFP-fusion construct. The ClearMap quantification approach (*14*) readily identified activated microglia with enlarged morphology in diverse brain regions, especially along the axonal tracts including the optic tract and the corpus callosum (**Fig. S4E−J**). Applying the same mTBI injury model on Thy1-GFP-M reporter mice (expressing GFP only in neurons), we confirmed the axonal abnormalities in the same brain regions (**Fig. 4SK−N**). We then used DISCO-MS workflow on isolated regions of interest (ROIs) including locally activated microglia with known spatial information (**Fig. 3A,B**). Analyzing three ROIs from the optic tract as small as 0.0005 mm^3 compared to corresponding regions in sham control animals, we quantified up to 1400 proteins per ROI. Principal Component Analysis (PCA) separated the proteomes of ROIs between mTBI and controls (**Fig. 3C**). Several proteins related to axonal damage and repair were upregulated in the mTBI ROIs including Stmn1 (32-fold increase) and Ncan (30-fold increase) (**Fig. 3D**). We found 602 common proteins in all ROIs of mTBI and sham. Comparing ROIs from mTBI among themselves, we found a common proteome signature comprising of a total of 717 proteins (**Fig. 3E**). We further validated the enrichment of Stmn1 and Ncan in mTBI brain tissues by immunofluorescence (**Fig. 3F-H**). Our data demonstrate that DISCO-MS is a powerful approach to obtain unbiased proteomics information on heterogeneous tissue regions with known spatial location.

Scalable and robust pathology identification using deep learning

One of the early hallmarks of Alzheimer's disease (AD) pathology is the accumulation of amyloid-beta (Aβ) plaques in the brain parenchyma (*15*). We imagined that unbiased detection of all Aβ plaques, followed by their equally unbiased proteome analysis using DISCO-MS would provide valuable insights to the initial stages of AD. To explore this hypothesis, we used the 5xFAD mouse model of AD and aimed for the identification of Congo red-labelled Aβ plaques in young mouse brains.

As the locations of these initial plaques are unknown, we developed a deep learning approach to rapidly and reliably identify all Aβ plaques in whole mouse brain scans. In short, our network architecture is based on U-Net, a well-established approach for biomedical image analysis (*16*). As the loss function, we used an equally weighted combination of Dice and binary cross entropy, and the Ranger optimizer, which combines Rectified Adam, gradient centralization and LookAhead (**Fig. 4A**). We applied suitable data augmentation protocols for training and implemented test time augmentation. To assess our segmentation quality, we calculated a wide range of voxel-wise and Aβ plaque segmentation metrics. Our deep learning architecture for automated Aβ plaque detection showed high performance in volumetric accuracy (0.99±0.00), volumetric (0.71±0.06) and surface (0.94±0.03), as well as overall Dice scores (0.89±0.09) per Aβ plaque. After segmenting all plaques in the entire brain using deep learning (**Fig. 4B,C**), we registered our data to the Allen brain atlas to obtain region-wise quantifications for more than a thousand brain subregions (*17*)(*18*). We then grouped them into the major brain regions defined by the Allen mouse brain ontology for the simplicity.

Our deep learning model identified Aβ plaques already in six weeks old mouse brains, much earlier than any previous study suggested (*19*). Some of the main brain regions with initial plaques were retro hippocampal region (310 plaques), medulla (motor area, 90 plaques), molecular layer of cerebellar cortex (67 plaques), fiber tracts (48 plaques), subiculum areas (31 plaques), visual area (27 plaques) and hippocampal formation (16 plaques) (**Fig. 4D**). We also analyzed the volume of detected Aβ plaques in these brain areas and observed the largest plaques in the temporal association, ectorhinal and auditory areas with an average volume between 2000-3500 μm^3 (**Fig. 4E**). Moreover, brain regions with the highest density of Aβ plaques were retro hippocampal and the subiculum areas (**Fig. 4F**). We confirmed our finding of Aβ plaques in six-week-old mice in the same brain regions by immunohistochemistry with anti-amyloid beta monoclonal antibodies (**Fig. S5A**). Additional experiments highlighted the absence of Aβ plaques at five weeks of age, whereas plaques were evident at seven weeks (**Fig. S5B**).

After deep learning-based identification of Aβ plaques in the 5xFAD mouse model, we isolated four ROIs (volume: ~0.0005 mm^3) from the hippocampal region (vs. corresponding brain regions from the control mice) and subjected them to MS-based proteomics (**Fig. 5A–D)**. We compared more than 1,900 proteins across replicates and PCA separated the ROIs with Aβ plaques from the control brain regions (**Fig. 5E**). Differential expression analysis revealed that many well-characterized AD-associated proteins were enriched in 5xFAD ROIs including the amyloid-beta precursor protein (*15*)(*20*) (32-fold increase) and the thimet oligopeptidase 1 (8-fold increase) (**Fig. 5F**). Apart from known and well-established AD-related proteins (*21*)(*22*), we also detected less characterized proteins in early-stage Aβ plaques such as a member of the calcium-binding protein family S100a11. We confirmed the presence of S100a11 and Thop1 in early-stage plaques of 5xFAD brain slices by immunofluorescence (**Fig. S6**). Next, we asked how similar the proteome in our ROIs with early Aβ plaques are. Plaques with more than 1200 protein identifications each shared a total of 768 proteins, defining the core proteome of early-stage Aβ plaque formation (**Fig. 5G**). An abundance rank plot of the shared early-stage Aβ plaques core proteome revealed several members of the Ywhaz (14-3-3) and the S100a protein family. We also found many other proteins known to be involved in AD such as Mapt, Snca and Park7. Illustrating the specificity of MS-based proteomics, we identified two isoforms of Mapt, namely Mapt-4 (Tau C) and Mapt-5 (Tau D) in the early-stage Aβ plaque ROIs (**Fig. 5H**). Our proteomics data also suggest early-stage Aβ plaque variability (**Fig. 5I**) with regards to well-characterized AD proteins (Mapt, Tmed10, Park7, Snca, App), proteins of the S100a family (a6, a4, a13, a1, a10, a11), peptidases (Thop1, Ppia), proteins of the Ywha family (14-3-3, q, g, e, b, h, z) and other structure-determining proteins including Nefm and Map2.

Thus, DISCO-MS recovered many known markers of Aβ plaques in AD and revealed potentially novel ones, while confirming that Aβ deposition plays a key role in early plaque development. Our data are consistent with the idea that S100a family members could be another driving factor in early-stage Aβ plaque development (*23*). Taken together, DISCO-MS allowed us to pinpoint early Aβ plaques in whole brains and analyze the proteomic makeup of isolated regions of interest including identified Aβ aggregates (**Movie S1**).

Discussion

Deciphering tissue heterogeneity is essential to better understand developmental and pathological processes. Enormous progress has been made in single cell transcriptomic technologies throughout the last years (*24*). Furthermore, novel ultra-high sensitivity MS-based approaches have emerged, enabling the analysis of proteomes down to the level of several labeled and pooled or even true single cells (*11*) (*25–28*). A major leap for both technologies would be the integration of proteomics and transcriptomics data with their spatial information (*29*)(*30*). Towards this goal, powerful technologies emerged for the molecular and spatial characterization on tissue slices (*29*, *31–38*). Furthermore, selected proteins and RNA molecules could be identified in cleared tissues (*9*, *39*, *40*). However, a bottom-up proteomics workflow for the unbiased detection of complete proteome profiles from cleared tissues has thus far been challenging. As many pathologies start in yet unknown regions of organs, this would pave new avenues for disease characterization at the earliest stage. We addressed this challenge by developing and applying DISCO-MS for unbiased proteome analysis of tiny tissues regions identified by panoptic imaging of whole organs and *in silico* 3D reconstructions using AI. Notably, we identified initial brain regions that form Aβ plaques in the brains of very young AD mouse model. As we also provided tissue proteome content with these tiny plaques, our work can guide future studies to dissect out initial mechanisms of AD pathology (all the proteome data are available at PRIDE repository, see methods for details). Our technology performs equally well on both rodent and human tissues and yields qualitative and quantitative proteomics data nearly indistinguishable from uncleared samples, even for the harshest organic solvent-based tissue clearing approaches. Although tissue clearing presumably removes the lipidic cast of membrane proteins, we observed that the plasma membrane protein gene ontology class was hardly affected, suggesting that DISCO-MS could be used to identify novel surface markers for drug targeting. Using DISCO-MS, we first identified pathological tissue regions throughout the entire brain in mTBI and AD mouse models. Then, we revealed novel proteins differentially expressed in mTBI brains and a core proteome signature of early-stage amyloid-beta plaques.

Notably, the DISCO-MS technology presented here is versatile across labeling, and solvent cleared methods for whole organs. It is not only applicable to reporter mouse lines but can also be utilized where reporter lines are currently absent. In those cases, deep tissue labeling with antibodies and/or dyes can be performed against the antigen of interest, imaged as a whole organ, and subjected to MS-based proteome profiling. DISCO-MS should be of great interest to researchers in possession of archived solvent cleared organs and imaging data, where molecular data is missing. Now, this method will further enable understanding the molecular basis of a pathological milieu in these organs. In conclusion, we present a spatial unbiased proteome profiling technology comprising complete 3D imaging data of whole organs, enabling unbiased identification of interesting tissue regions for subsequent molecular characterization.

References and Notes

1. H. Braak, D. R. Thal, E. Ghebremedhin, K. Del Tredici, Stages of the Pathologic Process in Alzheimer Disease: Age Categories From 1 to 100 Years. *J Neuropathol Exp Neurol*. **70**, 960–969 (2011).

2. H. R. Ueda, A. Ertürk, K. Chung, V. Gradinaru, A. Chédotal, P. Tomancak, P. J. Keller, Tissue clearing and its applications in neuroscience. *Nat Rev Neurosci*. **21**, 61–79 (2020).

3. S. Zhao, M. I. Todorov, R. Cai, R. A. -Maskari, H. Steinke, E. Kemter, H. Mai, Z. Rong, M. Warmer, K. Stanic, O. Schoppe, J. C. Paetzold, B. Gesierich, M. N. Wong, T. B. Huber, M. Duering, O. T. Bruns, B. Menze, J. Lipfert, V. G. Puelles, E. Wolf, I. Bechmann, A. Ertürk, Cellular and Molecular Probing of Intact Human Organs. *Cell*. **180**, 796-812.e19 (2020).

4. M. Belle, D. Godefroy, G. Couly, S. A. Malone, F. Collier, P. Giacobini, A. Chédotal, Tridimensional Visualization and Analysis of Early Human Development. *Cell*. **169**, 161-173.e12 (2017).

5. D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, K. Smith, B. Revaz, B. Finnbogason, A. Szantner, E. Lundberg, Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat Biotechnol*. **36**, 820–828 (2018).

6. E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, D. Van Valen, Deep learning for cellular image analysis. *Nat Methods*. **16**, 1233–1246 (2019).

7. R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function. *Nature*. **537**, 347–355 (2016).

8. D. S. Richardson, J. W. Lichtman, Clarifying Tissue Clearing. *Cell*. **162**, 246–257 (2015).

9. R. Cai, C. Pan, A. Ghasemigharagoz, M. I. Todorov, B. Förstera, S. Zhao, H. S. Bhatia, A. Parra-Damas, L. Mrowka, D. Theodorou, M. Rempfler, A. L. R. Xavier, B. T. Kress, C. Benakis, H. Steinke, S. Liebscher, I. Bechmann, A. Liesz, B. Menze, M. Kerschensteiner, M. Nedergaard, A. Ertürk, Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull-meninges connections. *Nat. Neurosci.* **22**, 317–327 (2019).

10. J. R. Wiśniewski, M. Y. Hein, J. Cox, M. Mann, A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards. *Molecular & Cellular Proteomics*. **13**, 3497–3506 (2014).

11. A.-D. Brunner, M. Thielert, C. G. Vasilopoulou, C. Ammar, F. Coscia, A. Mund, O. B. Hoerning, N. Bache, A. Apalategui, M. Lubeck, S. Richter, D. S. Fischer, O. Raether, M. A. Park, F. Meier, F. J. Theis, M. Mann, "Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation" (preprint, Systems Biology, 2020), , doi:10.1101/2020.12.22.423933.

12. J. A. Langlois, W. Rutland-Brown, K. E. Thomas, The incidence of traumatic brain injury among children in the United States: differences by race. *J Head Trauma Rehabil*. **20**, 229–238 (2005).

13. A. Ertürk, S. Mentz, E. E. Stout, M. Hedehus, S. L. Dominguez, L. Neumaier, F. Krammer, G. Llovera, K. Srinivasan, D. V. Hansen, A. Liesz, K. A. Scearce-Levie, M. Sheng, Interfering with the Chronic Immune Response Rescues Chronic Degeneration After Traumatic Brain Injury. *J Neurosci*. **36**, 9962–9975 (2016).

14. N. Renier, E. L. Adams, C. Kirst, Z. Wu, R. Azevedo, J. Kohl, A. E. Autry, L. Kadiri, K. Umadevi Venkataraju, Y. Zhou, V. X. Wang, C. Y. Tang, O. Olsen, C. Dulac, P. Osten, M. Tessier-Lavigne, Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. *Cell*. **165**, 1789–1802 (2016).

15. M. Meyer-Luehmann, T. L. Spires-Jones, C. Prada, M. Garcia-Alloza, A. de Calignon, A. Rozkalne, J. Koenigsknecht-Talboo, D. M. Holtzman, B. J. Bacskai, B. T. Hyman, Rapid appearance and local toxicity of amyloid-beta plaques in a mouse model of Alzheimer's disease. *Nature*. **451**, 720–724 (2008).

16. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (2015) (available at http://arxiv.org/abs/1505.04597).

17. Q. Wang, S.-L. Ding, Y. Li, J. Royall, D. Feng, P. Lesnar, N. Graddis, M. Naeemi, B. Facer, A. Ho, T. Dolbeare, B. Blanchard, N. Dee, W. Wakeman, K. E. Hirokawa, A. Szafer, S. M. Sunkin, S. W. Oh, A. Bernard, J. W. Phillips, M. Hawrylycz, C. Koch, H. Zeng, J. A. Harris, L. Ng, The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*. **181**, 936-953.e20 (2020).

18. M. I. Todorov, J. C. Paetzold, O. Schoppe, G. Tetteh, S. Shit, V. Efremov, K. Todorov-Völgyi, M. Düring, M. Dichgans, M. Piraud, B. Menze, A. Ertürk, Machine learning analysis of whole mouse brain vasculature. *Nat Methods*. **17**, 442–449 (2020).

19. A. Boza-Serrano, Y. Yang, A. Paulus, T. Deierborg, Innate immune alterations are elicited in microglial cells before plaque deposition in the Alzheimer's disease mouse model 5xFAD. *Sci Rep*. **8**, 1550 (2018).

20. S. F. Lichtenthaler, Ectodomain shedding of the amyloid precursor protein: cellular control mechanisms and novel modifiers. *Neurodegener Dis*. **3**, 262–269 (2006).

21. B. Meckelein, H. A. de Silva, A. D. Roses, P. N. Rao, M. J. Pettenati, P. T. Xu, R. Hodge, M. J. Glucksman, C. R. Abraham, Human endopeptidase (THOP1) is localized on chromosome 19 within the linkage region for the late-onset alzheimer disease AD2 locus. *Genomics*. **31**, 246–249 (1996).

22. G. Pollio, J. J. M. Hoozemans, C. A. Andersen, R. Roncarati, M. C. Rosi, E. S. van Haastert, T. Seredenina, D. Diamanti, S. Gotta, A. Fiorentini, L. Magnoni, R. Raggiaschi, A. J. M. Rozemuller, F. Casamenti, A. Caricasole, G. C. Terstappen, Increased expression of the oligopeptidase THOP1 is a neuroprotective response to Abeta toxicity. *Neurobiol Dis*. **31**, 145–158 (2008).

23. J. S. Cristóvão, C. M. Gomes, S100 Proteins in Alzheimer's Disease. *Front. Neurosci.* **13**, 463 (2019).

24. T. Stuart, R. Satija, Integrative single-cell analysis. *Nat Rev Genet*. **20**, 257–272 (2019).

25. A. Mund, F. Coscia, R. Hollandi, F. Kovács, A. Kriston, A.-D. Brunner, M. Bzorek, S. Naimy, L. M. Rahbek Gjerdrum, B. Dyring-Andersen, J. Bulkescher, C. Lukas, C. Gnann, E. Lundberg, P. Horvath, M. Mann, "AI-driven Deep Visual Proteomics defines cell identity and heterogeneity" (preprint, Systems Biology, 2021), , doi:10.1101/2021.01.25.427969.

26. T. K. Cheung, C.-Y. Lee, F. P. Bayer, A. McCoy, B. Kuster, C. M. Rose, Defining the carrier proteome limit for single-cell proteomics. *Nat Methods*. **18**, 76–83 (2021).

27. B. Budnik, E. Levy, G. Harmange, N. Slavov, SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol*. **19**, 161 (2018).

28. Y. Cong, K. Motamedchaboki, S. A. Misal, Y. Liang, A. J. Guise, T. Truong, R. Huguet, E. D. Plowey, Y. Zhu, D. Lopez-Ferrer, R. T. Kelly, "Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per mammalian cell" (preprint, Systems Biology, 2020), , doi:10.1101/2020.06.03.132449.

29. S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, E. Z. Macosko, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. **363**, 1463–1467 (2019).

30. X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, G. P. Nolan, F.-A. Bava, K. Deisseroth, Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*. **361** (2018), doi:10.1126/science.aat5691.

31. P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016).

32. S. C. van den Brink, A. Alemany, V. van Batenburg, N. Moris, M. Blotenburg, J. Vivié, P. Baillie-Johnson, J. Nichols, K. F. Sonnen, A. Martinez Arias, A. van Oudenaarden, Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*. **582**, 405–409 (2020).

33. Y. Takei, J. Yun, S. Zheng, N. Ollikainen, N. Pierson, J. White, S. Shah, J. Thomassie, S. Suo, C.-H. L. Eng, M. Guttman, G.-C. Yuan, L. Cai, Integrated spatial genomics reveals global architecture of single nuclei. *Nature*. **590**, 344–350 (2021).

34. S. Alon, D. R. Goodwin, A. Sinha, A. T. Wassie, F. Chen, E. R. Daugharthy, Y. Bando, A. Kajita, A. G. Xue, K. Marrett, R. Prior, Y. Cui, A. C. Payne, C.-C. Yao, H.-J. Suk, R. Wang, C.-C. (Jay) Yu, P. Tillberg, P. Reginato, N. Pak, S. Liu, S. Punthambaker, E. P. R. Iyer, R. E. Kohman, J. A. Miller, E. S. Lein, A. Lako, N. Cullen, S. Rodig, K. Helvie, D. L. Abravanel, N. Wagle, B. E. Johnson, J. Klughammer, M. Slyper, J. Waldman, J. Jané-Valbuena, O. Rozenblatt-Rosen, A. Regev, I. Consortium19¶, G. M. Church, A. H. Marblestone, E. S. Boyden, Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*. **371** (2021), doi:10.1126/science.aax2656.

35. Spatially resolved, highly multiplexed RNA profiling in single cells | Science, (available at https://science.sciencemag.org/content/348/6233/aaa6090).

36. D. Mahdessian, A. J. Cesnik, C. Gnann, F. Danielsson, L. Stenström, M. Arif, C. Zhang, T. Le, F. Johansson, R. Shutten, A. Bäckström, U. Axelsson, P. Thul, N. H. Cho, O. Carja, M. Uhlén, A. Mardinoglu, C. Stadler, C. Lindskog, B. Ayoglu, M. D. Leonetti, F. Pontén, D. P. Sullivan, E. Lundberg, Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature*. **590**, 649–654 (2021).

37. Tissue-based map of the human proteome | Science, (available at https://science.sciencemag.org/content/347/6220/1260419).

38. W.-T. Chen, A. Lu, K. Craessaerts, B. Pavie, C. Sala Frigerio, N. Corthout, X. Qian, J. Laláková, M. Kühnemund, I. Voytyuk, L. Wolfs, R. Mancuso, E. Salta, S. Balusu, A. Snellinx, S. Munck,

A. Jurek, J. Fernandez Navarro, T. C. Saido, I. Huitinga, J. Lundeberg, M. Fiers, B. De Strooper, Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. *Cell*. **182**, 976-991.e19 (2020).

39. B. Yang, J. B. Treweek, R. P. Kulkarni, B. E. Deverman, C.-K. Chen, E. Lubeck, S. Shah, L. Cai, V. Gradinaru, Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell*. **158**, 945–958 (2014).

40. A. Blutke, N. Sun, Z. Xu, A. Buck, L. Harrison, S. C. Schriever, P. T. Pfluger, D. Wiles, T. Kunzke, K. Huber, J. Schlegel, M. Aichler, A. Feuchtinger, K. Matiasek, S. M. Hauck, A. Walch, Light sheet fluorescence microscopy guided MALDI-imaging mass spectrometry of cleared tissue samples. *Sci Rep*. **10**, 14461 (2020).

Acknowledgements: A.-D.B. acknowledges support from the International Max Planck Research School for Life Sciences – IMPRS-LS. We thank our colleagues in the Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, for discussions and help. In particular, we thank I. Paron, A. Piras and C. Deiml for technical support and J.B. Müller for column production, and I. Bechmann and H. Steinke for human tissues.

Funding: This work was supported by the Vascular Dementia Research Foundation, Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the framework of the Munich Cluster for Systems Neurology (EXC 2145 SyNergy – ID 390857198). Furthermore, NVIDIA supported this work via the GPU Grant Program.

Author contributions: H.S.B., A.-D. B., M.M. and A.E. conceptualized and designed the study. H.S.B., A.-D.B., Z.R, H.M, M.C.T., M.M, Z.I.K performed experiments. H.S.B, Z.R, H.M performed mice experiments, solvent-based organ clearing, light sheet imaging procedures and stitching of data. H.S.B. developed the LCM-based isolation procedure for target tissues from solvent-cleared organs. A.-D.B. developed the sample preparation workflow for proteomics analysis. A.-D.B. and M.C.T. performed mass spectrometry-based proteomics analysis. A.-D.B. performed proteomics data analysis. R.A., J.C.P., F.K., M.A., B.M. and F.T. developed deep learning models. M.A., F.T. performed data analysis. M.I.T. performed atlas registration of 5xFAD brains. H.S.B. and D.K. performed ClearMAP analyses. S.M. and S.L. helped with the prototyping experiment. H.S.B., A.-D.B., M.C.T., analyzed the data. H.S.B., A.-D.B., M.M. and A.E. wrote the manuscript.

Competing interests: A.E., H.S.B., A.-D.B., M.M., filed a patent on some of the technologies presented in this work. F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc. and Dermagnostix. All other authors have no competing interests.

Data availability:

All mass spectrometry raw data, libraries and outputs from each particular search engine analyzed in this study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD025316.

Reviewer account details:

Link: https://www.ebi.ac.uk/pride/login

Username: reviewer_pxd025316@ebi.ac.uk

Password: qw3cimng

Materials and Methods:

Animals. We used the following animals in the study: mixed gender CX3CR1-eGFP, Thy-1-GFPM, 5xFAD and C57Bl6/J from Jackson Laboratory. The animals were housed under a 12/12 hours light/dark cycle. The animal experiments were conducted according to institutional guidelines: Klinikum der Universität München / Ludwig Maximilian University of Munich and after approval of the Ethical Review Board of the Government of Upper Bavaria (Regierung von Oberbayern, Munich, Germany) and the Animal Experiments Council under the Danish Ministry of Environment and Food (2015-15-0201-00535) and following the European directive 2010/63/EU for animal research. All data are reported according to the ARRIVE criteria. Sample sizes were chosen based on prior experience with similar models.

Human samples. Intact human brains were taken from human body donors with no known neuropathological diseases. The donors gave their informed and written consent to explore their cadavers for research and educational purposes, when still alive and well. The signed consents are kept at the Anatomy Institute, University of Leipzig, Germany. Institutional approval was obtained in accordance to the Saxonian Death and Funeral Act of 1994. The signed body donor consents are available on request.

Repeated closed head mild traumatic brain Injury (mTBI). Before mTBI, tin foil was taped and tightened to the U-shaped stage made of clear plastic container (38 x 27 x 27 cm^3)

containing a sponge collection (38 x 25 x 15 cm3). Then, mice were pre-treated with buprenorphine (1:15 saline, 50ul/20mg, ip) and anesthetized with 4% isoflurane using 1.0 L min^{-1} air until non-responsive to a paw or tail pinch. To ensure the head acceleration and rotation following the head impact, the mice were placed under impact tip on the tin foil, which contains holes according to the shape of a mouse and can support the body weight of the mice. The mice were kept under light anesthesia with continued 2% isoflurane. mTBI was produced using a stereotaxic impactor device with a 5-mm round tip coated with 1mm thick rubber, which can preserve an intact skull after impact. Impact tip was placed and covered the scalp's area from just behind the eyes to the midline of the ears, the center of the tip at approximately midway along the sagittal suture. The injury was produced without skin incision (velocity of 5 m/sec, depth of 0.5mm, and dwell time of 0.1 sec). The mouse would be removed quickly from the collection sponge and transferred to the recovery box maintained at 32 °c. In total, mice received four hits with a 48 hours interval in seven days. Sham mice receive identical handling and exposure to the same time length of anesthesia as the mTBI mice but received no impact.

Perfusion and tissue preparation. Mice were anesthetized using a combination of midazolam, medetomidine and fentanyl (MMF) (1mL/100g of body mass for mice; i.p.). As soon as the animals did not show any pedal reflex, they were intracardially perfused with cold heparinized 0.1 M PBS (10 U/mL of Heparin, Ratiopharm; 100-125 mmHg pressure using a Leica Perfusion One system) for 5-10 minutes at room temperature until the blood was washed out, followed by ice-cold 4% paraformaldehyde (PFA) in 0.1 M PBS (pH 7.4) (Sigma) for 10 minutes. Then, the brains were extracted and post-fixed in 4% PFA for 1 day at 4 °C and later washed with 0.1 M PBS for 10 minutes 3 times at room temperature. The whole brain clearing or nanoboosting procedure was started immediately. For the collection of fresh frozen samples, animals were sacrificed by cervical dislocation and brains were quickly snapped frozen in liquid nitrogen and stored in -80 °C until further processing.

Congo red labeling of whole brains of 5xFAD animals. Whole brains were dehydrated with gradual addition of methanol in PBS (50% x1, 80% x1, 100% x2, each for 1 hr). Overnight bleaching with 5% hydrogen peroxide in methanol was done at 4 °C. Brains were then gradually rehydrated in 100%, 80%, 50% methanol in PBS (1 hr for each step, followed by 2 additional washes in PBS). Detergent washing was then performed in PBS with 0.2% Triton X-100 for 2 hr, brains were incubated overnight at 37°C in PBS with 0.2% Triton X-100 and 0.3 M glycine, followed by blocking in PBS with 0.2% Triton X-100 and 6% goat serum for 7 days. Following blocking, the tissue was washed for 1 hr twice in PBS with 0.2% Tween-20 and 10 µg/ mL heparin (PTwH). Next, brains were incubated with 10 µM Congo Red (Sigma, C6277) at 37°C in PTwH for 5 days. After that brains were washed in PTwH for 2 days with periodic solution changes and gradually dehydrated using 3DISCO clearing as described next.

Clearing of brains using DISCO methods. We followed the 3DISCO and uDISCO passive clearing protocol as described previously. In brief, dissected brains were placed in 5 ml tubes (Eppendorf, 0030 119.401) and covered with 4.5 mL of clearing solution. All incubation steps were performed in a fume hood with gentle shaking or rotation, with the samples covered with aluminum foil to keep them in dark. To clear the samples using 3DISCO, gradient of tetrahydrofuran (THF) in distilled water (vol/vol %), 2 hours for each step, was used as 50, 70, 90, 100 and overnight 100 % THF; after dehydration, the samples were incubated for 45 min in dichloromethane (DCM, Sigma, 270997), and finally in BABB (benzyl alcohol + benzyl benzoate 1:2, Sigma, 24122 and W213802) until transparency. Next for uDISCO a gradient of tert-butanol (Sigma, 360538) in distilled water (vol/vol %) was used as 50, 70, 90, 100 twice at 32°C for 12 hours each step, followed by immersion in DCM for 45 minutes at room temperature and finally incubated with the refractive index matching solution BABB-D15 containing 15 parts BABB, 1 part diphenyl ether (DPE) (Alfa Aesar, A15791) and 0.4% Vol vitamin E (DL-alpha-tocopherol, Alfa Aesar, A17039), for at least 6 hours at room temperature until achieving transparency.

vDISCO whole-brain passive immunostaining, clearing and imaging. Passive vDISCO was performed on dissected organs as performed by Cai R et al. First, the post-fixed brains were pretreated with permeabilization solution containing 1.5% goat serum, 0.5% Triton X-100, 0.5 mM of Methyl-beta-cyclodextrin, 0.2% trans-1-Acetyl-4-hydroxy-L-proline and 0.05% Sodium Azide 0.1 M for 2 days at 37°C with gentle shaking. Subsequently, the brains were incubated in 4.5 mL of this same permeabilization solution plus the nanobooster Atto647N conjugated anti-GFP (1:600, which is ~5-8 µg of nanobooster in 4.5 ml) for CX3CR1-eGFP and Thy-1-GFPM brains for 12-14 days at 37°C with gentle shaking, then brains were washed for 2 hours 3 times and once overnight with the washing solution (1.5% goat serum, 0.5% Triton X-100, 0.05% of sodium azide in 0.1 M PBS) at room temperature and in the end washed for 2 hours 4 times with 0.1 M PBS at room temperature. The immunostained brains were cleared with 3DISCO clearing first they were put in the Eppendorf 5 ml tubes and then incubated at room temperature with gentle shaking in 4.5 mL of the following gradient of THF in distilled water (vol/vol %), 2 hours for each step: 50, 70, 90, 100 THF and overnight 100 % THF; after dehydration, the samples were incubated for 45 min in DCM, and finally in BABB until transparency. During all the clearing steps, the tubes were wrapped with aluminum foil to keep them in dark**.**

SHANEL sample preparation and clearing. Archived human brain samples were obtained in PFA which were stored for a long period of time (>5 years) at 4 °C and subjected to our previously published SHANEL clearing protocol with some modifications(*3*). Briefly, samples were dehydrated with EtOH /dH_2O series at RT: 50%, 70%, 100% for 1 h for each step. Then incubated with 10 mL DCM/MetOH (2:1 v/v) (freshly prepared) for 6 h at RT followed by rehydration with EtOH /dH_2O series at RT: 100%, 70%, 50%, dH_2O for 1 h each step then incubated with 0.5M acetic acid (30 mL/L) at RT for 2 h, then wash with dH_2O twice for 15 min and then incubated with 4M guanidine hydrochloride (382.12 g/L), 0.05M sodium acetate (4.1g/L), 2% v/v Triton X100 in dH_2O, (measure pH: 6.0) at RT for 2 h, then wash with dH_2O twice for 15 min each and wash with PBS twice for 15 min each. Afterwards samples were incubated with 10% CHAPS, 25% N-Methyldiethanolamine in dH_2O at 37 °C for 4 hours and then washed with dH_2O twice for 15 min each. Since we did not perform any deep antibody labeling in these samples, we started clearing these samples without prior blocking or antibody labeling steps. Clearing was done with THF in water with dilutions (vol/vol%) of 50%, 70 %, 90%, 100%, 1hr each, 100% overnight, DCM 45 min and incubated in BABB until the samples were transparent.

Behavioral assessment

Barnes Maze. Briefly, a maze consisting of a surface bright circular platform with an escape black box can be recessed and located at the bottom of one of the 20 holes. Visual shapes were placed on 3 walls of the room as cues. For all trials, mice were placed in a cylinder black start chamber in the center of the maze for 10 second. After the chamber lifted and the test started, mice were given 3 minutes to locate and enter the target box during the spatial acquisition time. For a period of 4 days, 4 trials were given per day with an inter-trial of 15 minutes. The trial ended when the mouse entered the escape box or after 3 minutes had elapsed. Mice were allowed to remain in the escape box for 1 minute. A system (Ethovision XT) was used to continually track and record the movement of the mice. Escape latency was measured as the time taken for the mouse to enter the box.

Assessment of motor function. The mice were given 3 trials training per day for 3 days to walk along a 1 cm diameter and 100 cm long wood beam with a goal box on the end of the beam before mild TBI. The height of the beam was placed 1 m above ground. The latency that it takes walking to cross the beam after 8 weeks mild TBI were recorded. Mice that were unable to cross the beam would be removed in the training.

Immunofluorescence and Confocal microscopy. Briefly, mice were sacrificed after 8 weeks of injury or at six weeks of age following transcardial perfusion with PBS and with 4 % cold PFA. Brains were post-fixed in 4% PFA at 4 °C overnight. Either frozen sections or cleared-rehydrated frozen sections were treated with 0.2% Triton X-100 in PBS for 15 min, blocked for 1 hour at room temperature with 10% serum in PBST. Then incubation with primary antibodies Iba1 (1:1000, Wako, 019-19741), Stathmin 1 (1:300, Novus, NBP1-76798), Neurocan (1:300, abcam, ab31979), S100a11 (1:300, R&D, MAB5167), Thop1 (1:300, Novus, NB400-146), MOAB2 (1:1000, Novus, NBP2-13075), at 4°C for overnight and Alexa conjugated secondary antibodies (1:1000, Goat anti-rabbit IgG Alexa fluor 647, Invitrogen, A21245; Goat anti-Mouse IgG Alexa Fluor 488, A11029; Goat anti-Mouse IgG Alexa Fluor 594, A11032; Goat anti-Rat IgG Alexa Fluor 594, A11007; Goat anti-Rat IgG Alexa Fluor 488, A11006;) were incubated for 1 hour at room temperature. Slices were mounted after being stained with Hoechst 33342 (Invitrogen). Images were acquired with 10x, 40x, and 63x objective of confocal microscope (ZEISS LSM880).

Light-sheet microscopy and image processing. Single plane illuminated (light-sheet) image stacks were acquired using an Ultramicroscope II (LaVision BioTec), featuring an axial resolution of 4 μm with following filter sets: ex 470/40 nm, em 535/50 nm; ex 545/25 nm, em 605/70 nm; ex 640/40 nm, em 690/50 nm. Whole brains were imaged individually using high magnification objectives: 4x objective (Olympus XLFLUOR 4x corrected/0.28 NA [WD = 10 mm]), LaVision BioTec MI PLAN 12x objective (0.53 NA [WD 10 = mm]) coupled to an Olympus revolving zoom body unit (U-TVCAC) kept at 1x. High magnification tile scans were acquired using 20% overlap and the light-sheet width was reduced to obtain maximum illumination in the field. Processing, data analysis, 3D rendering and video generation for the rest of the data were done on an HP workstation Z840, with 8 core Xeon processor, 196 GB

RAM, and Nvidia Quadro k5000 graphics card and HP workstation Z840 dual Xeon 256 GB DDR4 RAM, nVidia Quadro M5000 8GB graphic card. We used Imaris (Bitplane), Fiji (ImageJ2) and Vision 4D (Arivis) for 3D and 2D image visualization. Tile scans were stitched by Fiji's stitching plugin49.

ClearMap quantification. To quantify microglia distribution in whole brains of mTBI and sham animals, we used ClearMap. As the script was originally developed for quantification of the cFos^{+} cells, to comply with the offered method, we did the following pre-processing steps on our microglia data using Fiji before ClearMap:

1. Background equalization to homogenize intensity distribution and appearance of the microglia cells over different regions of the brain, using pseudo-flat-field correction function from Bio-Voxxel toolbox.

2. Convoluted background removal, to remove all particles bigger than relevant cells. This was done with the median option in the Bio-Voxxel toolbox.

3. Two-dimensional median filter to remove remaining noise after background removal. The filter radius was chosen to ensure the removal of all particles smaller than microglia cells.

4. Unshapen mask to amplify the high-frequency components of a signal and increase overall accuracy of the cell detection algorithm of ClearMap.

After pre-processing, ClearMap was applied by following the original publication and considering the threshold levels that we obtained from the pre-processing steps. As soon as the quantification was completed, the data was exported as an Excel file for further analysis.

Deep learning analyses. The segmentation of the stained Aβ plaques represents a key step towards a reliable quantification thereof. We develop a customized, three-dimensional deep learning approach to optimize segmentation of Aβ plaques in the whole brains of 5xFAD animals. Our network architecture is inspired by the well-established U-Net architecture. The used loss function is an equally weighted combination of Dice and binary cross entropy loss. We use the Ranger optimizer, which combines Rectified Adam, gradient centralization and LookAhead. Our dataset consists of 98 image volumes (300x300x300 voxel) from one Alzheimer brain; where 34 volumes include Aβ plaques and 64 volumes do not contain any plaques. An ensemble of experts including the scientist who imaged the brains labeled all images. We randomly sample our training set of 85 volumes (21 with AD plaques, 64 without AD plaques), our validation set of seven volumes and our separate test set of six volumes. During training and testing we applied suitable data augmentation protocols. In order to assess the quality of our segmentation we calculate a wide range of voxel-wise and Aβ plaque wise segmentation metrics. Based on the reliable segmentation of individual Aβ plaques we continue towards a statistical evaluation of the number and size of Aβ plaques per brain region. First, we register all of our brains to the Allen brain Atlas, enabling a single voxel assignment to brain structures. For whole brain segmentation we extract the single connected components, which represent our individual segmented Aβ plaques and calculate their total size in voxels as a biomarker. Using this registration and biomarker, we calculate per brain region statistics for the presence and size of Aβ plaques across the whole brain.

Optimization of cleared tissue for cryopreservation and sectioning. After acquiring the whole brain images from CX3CR1-eGFP, 5xFAD, C57BL/6J mice the brains were further optimized for cryopreservation and sectioning. The course of tissue clearing and imaging in BABB makes the tissue brittle and hard for further processing, Thus, to solve this hurdle we re-hydrated the samples with respective clearing solutions to be able to process samples for cryosectioning. Thereafter, samples were washed with PBS twice for 15 min each and cryopreserved overnight with 30 % sucrose solution in 4 °C. In order to avoid any ice crystals formation, samples were further embedded in Optimal cutting temperature compound (OCT compound) under the chilled isopentane container placed on dry ice. Samples were stored in -80 °C until cryosectioning.

Laser- capture microdissection. For the microdissection of cells and plaques, we used the PALM MicroBeam system (Zeiss). PALM MicroBeam uses a focused laser beam to cut out and isolate the selected specimen without contact. The laser catapult isolates the region of interest fast and uncontaminated in the adhesive cap mounted in the RoboMover. Briefly, after cryosectioning, sections were mounted on the polyethylene naphthalate (PEN, Zeiss) slides and were either stored at -80 °C in 50 ml falcon tubes filled with molecular sieves (Sigma-Aldrich) or processed further for serial dehydration with ethanol and air dried for 15 min under the hood. Cells in the optic tract from mTBI/Sham brains and amyloid-beta plaques from the 5xFAD and from respective WT brain regions were microdissected by laser pressure catapulting (LPC) UV laser capture microdissection system (Palm Zeiss Microlaser Technologies, Munich, Germany) consisting of an inverted microscope with a motorized stage, an ultraviolet (UV) laser and an X-Cite 120 fluorescence illuminator (EXFO). The microdissection process was visualized with an AxioCam ICc camera coupled to a computer and was controlled by Palm RoboSoftware (Zeiss, Germany). Approximately an area of 200x200 µm (corresponding to 40-60 cells) were cut by laser using 20x objective (LD Plan-Neofluar 20x/0.4 corr M27) and catapulted against gravity into the adhesive cap. Tissues were quickly lysed in 20 µl of lysis buffer, spinned down and kept in dry ice or stored in -80 °C. To avoid any uncertainties in capturing ROI, each time after catapulting as well as after lysing and spin down, adhesive cap was properly examined under the camera.

Optimization of DISCO cleared sample preparation for mass spectrometry analysis. Several conditions and combinations of solubilizing agents for the isolation of proteins from tissue cleared mouse brain, heart, and lung samples were initially evaluated for protein extraction efficiency, peptide recovery, and qualitative and quantitative reproducibility keeping fresh or PFA-fixed as reference. Our goal was to establish a workflow that recovers proteomes that are as similar as possible to non-cleared tissue and is universal for all tissue clearing techniques.

Cleared organs or cryosections were removed from the refractive index matching solution BABB and washed five times with 1x PBS solution. The organ was then flash-frozen and pulverized in a Covaris CP02. Afterwards, the samples were resuspended in different protein solubilizing solutions (6 % Sodiumdodecylsulfate, 500 mM TrisHCl, pH 8.5 (SDS buffer); 2 % Sodiumdeoxycholate, 100 mM TrisHCl pH 8.5, 10 mM Tris-(2-carboxyethyl)-phosphin (TCEP), 40 mM Chloroacetamide (SDC buffer); 50 % Trifluoroethanole, 100 mM TrisHCl, pH 8.5 (TFE buffer), followed by protein extraction at 95 °C, 1.000rpm for 45 min. Then the samples were subjected to sonication (Branson) at maximum frequency for 30 cycles at 50% output, followed by another heating step at 95 °C, 1.000 rpm for 45 min. From here on, processing steps diverged for each protocol.

Proteins solubilized in the SDS buffer were precipitated with ice-cold Acetone at 80 % v/v ratio overnight at -80 °C, followed by centrifugation at max. g for 15 min at 4 °C. The supernatant was removed, the pellet was washed with 5 ml ice-cold 80 % v/v Acetone/ddH2O, followed by 30 min precipitation on dry ice. The acetone wash steps were repeated two times for a total of three washes. Proteins solubilized in the TFE buffer, were subjected to solvent evaporation in a speedvac at 45 °C until dryness before further processing.

In case of SDS-SDC or TFE-SDC protocol, in which SDS or TFE protein extraction was coupled to an SDC-based protein digestion, SDS- or TFE-solubilized proteins were resuspended in 1ml of SDC buffer and heated to 95 °C at 1.000 rpm for 10 min to denature proteins, reduce cysteine bridges and alkylate free cysteine residues. Afterwards, samples were sonicated for 15 cycles each 30sec at max power in a Bioruptor, followed by another heating step for 10 min at 95 °C, 1.000 rpm in a Thermoshaker.

SDC-only, SDS-SDC, TFE-SDC solubilized protein solutions were cooled down to room temperature, diluted 1:1 with 100 mM TrisHCl, pH 8.5, followed by protein concentration estimation by Nanodrop. Extracted and solubilized proteins were digested overnight at 37 °C and 1.000 rpm, with Trypsin and LysC at a protein to enzyme w/w ratio of 1:50. Next day, Trypsin and LysC were added again at a protein to enzyme w/w ratio of 1:50 and proteins were digested further for 4 h at 37 °C, 1.000 rpm. Resulting peptides were acidified with 1 % TFA 99 % Isopropanol in a 1:1 ratio and vortexed, followed by centrifugation at 22.000 xg RT to pellet residual particles. The supernatant was transferred into a fresh tube and subjected to stage-tip clean-up via SDB-RPS. 20 µg of peptides were loaded on two 14-gauge stage-tip plugs. Peptides were washed twice with 200 µL 1 % TFA 99 % ddH2O followed by 200 µL 1 % TFA 99 % Isopropanol in an in-house-made Stage-tip centrifuge at 2000 xg. Peptides were eluted with 100 µL of 5 % Ammonia, 80 % CAN into PCR tubes and dried at 45 °C in a SpeedVac centrifuge (Eppendorf, Concentrator plus). Peptides were resuspended in 0.1% TFA, 2% ACN, 97.9% ddH2O.

After evaluation of protein extraction efficiency, all sample preparation for Fresh, PFA-fixed, uDISCO-, 3DISCO-, Shanel-cleared tissue was performed following the SDS-SDC protocol. For LCM sample preparation, LCM samples were caught on PCR tubes with adhesive caps and successful isolation was verified by visual inspection. 20µl of SDS-buffer was added to each tube. The tube was closed and vortex for 30 sec, followed by centrifugation for 5 min in a table-top centrifuge to 'catch' the LCM sample in the protein solubilization buffer, which was confirmed afterwards by visual inspection. Sample preparation was performed as described for the SDS-SDC protocol, except for the following modifications: No shaking during cooking steps; Instead of a Branson sonicator, a Bioruptor was used for each sonication step; No Covaris CP02 was used for crushing the sample; Acetone precipitation was performed at 100 µl total volume; SDC resuspension and protein digestion was performed at a 20 µl volume.

High-pH reversed-phase fractionation. To generate a deep library of experiment-specific precursors, peptides were fractionated at pH 10 with the spider-fractionator. 50 µg of purified peptides were separated on a 30 cm C$_{18}$ column in 96 min and concatenated into 24 fractions with 2 min exit valve switches. Peptide fractions were dried in a SpeedVac and reconstituted in 2 % ACN, 0.1 % TFA, 97.9 % ddH$_2$O for LC-MS analysis.

Liquid chromatography and mass spectrometry (LC-MS). LC-MS was performed on an EASY nanoLC 1200 (Thermo Fisher Scientific) coupled online either to a quadrupole Orbitrap mass spectrometer (Q Exactive HFX, Thermo Fisher Scientific), or a trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer (timsTOF Pro, Bruker Daltonik GmbH, Germany) via nano-electrospray ion source (Captive spray, Bruker Daltonik GmbH). Peptides were loaded on a 50 cm in-house packed HPLC-column (75 µm inner diameter packed with 1.9 µm ReproSil-Pur C18-AQ silica beads, Dr. Maisch GmbH, Germany). Sample analytes were either separated using a linear 100min gradient from 5-30 % B in 80 min followed by an increase to 60 % for 4 min, and by a 4 min wash at 95 %, a decrease to 5 % B for 4 min, and a re-equilibration step at 5 % B for 4 min, or separated on a linear 120 min gradient from 5-30 % B in 90 min followed by an increase to 60 % for 10 min, and by a 5 min wash at 95 %, a decrease to 5 % B for 5 min, and a re-equilibration step at 5 % B for 5 min (Buffer A: 0.1 % Formic Acid, 99.9 % ddH2O; Buffer B: 0.1 % Formic Acid, 80 % CAN, 19.9 % ddH2O). Peptides derived from laser capture microdissection and matching libraries were separated using a linear 70 min gradient from 3-30 % B in 45 min followed by an increase to 60 % for 5 min, an increase to 95 % in 5min, followed by 5 min at 95 % B, a decrease to 5 % B for 5 min, and an equilibration step at 5 % B for 5 min. Flow-rates were constant at 300 nl/min. The column temperature was kept at 60 °C by an in-house manufactured oven.

Mass spectrometry analysis for the evaluation of sample preparation on a Q Exactive HFX was performed in data dependent scan mode. For full proteome measurements, MS1 spectra were acquired at 60.000 resolution and a m/z range of 300-1.650 with an automatic gain control (AGC) target of 3E6 ions and a maximum injection time of 20 ms. The top 15 most intense ions with a charge of two to eight from each MS1 scan were isolated with a width of 1.4 m/z, followed by higher-energy collisional dissociation (HCD) with a normalized collision energy of 27 % and a scan range of 200-2000 m/z. MS/MS spectra were acquired at 15,000 resolution with an AGC target of 1E5, a minimum AGC target of 2.9E3, and a maximum injection time of 28 ms. Dynamic exclusion of precursors was set to 30 s.

Deep proteomes and comparisons of clearing conditions with the SDS-SDC protocol were acquired on a standard timsTOF Pro in a data-dependent PASEF mode with 1 MS1 survey TIMS-MS and 10 PASEF MS/MS scans per acquisition cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to 100 ms each and we analyzed the ion mobility range from $1/K_0$ = 1.6 Vs cm^{-2} to 0.6 Vs cm^{-2}. Precursor ions for MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 in a total m/z range of 100-1.700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at $1/K_0$ = 1.6 VS cm^{-2} to 20 eV at $1/K_0$ = 0.6 Vs cm^{-2}. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonik GmbH). Precursors for MS/MS were picked at an intensity threshold of 2.500 a.u. and re-sequenced until reaching a 'target value' of 20.000 a.u taking into account a dynamic exclusion of 40 sec elution.

Peptides derived from LCM samples were acquired on a timsTOF Pro modified for highest ion transmission and sensitivity, as described in Brunner *et al.*, in a data-dependent PASEF mode with 1 MS1 survey TIMS-MS and 5 PASEF MS/MS scans per acquisition cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to 50 ms each and we analyzed the ion mobility range from $1/K_0$ = 1.6 Vs cm^{-2} to 0.6 Vs cm^{-2}. Precursor ions for MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 in a total m/z range of 100-1.700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at $1/K_0$ = 1.6 VS cm^{-2} to 20 eV at $1/K_0$ = 0.6 Vs cm^{-2}. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonik GmbH). Precursors for MS/MS were picked at an intensity threshold of 1.500 a.u. and re-sequenced until reaching a 'target value' of 20.000 a.u taking into account a dynamic exclusion of 40 sec elution.

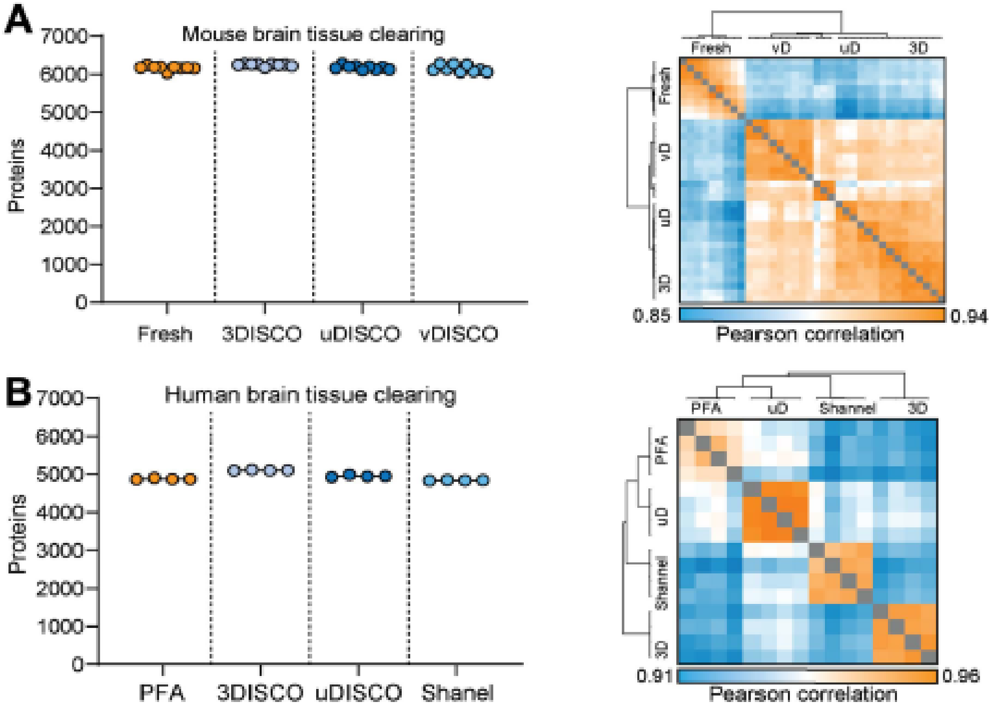
Data processing. Raw files were either searched against the mouse Uniprot databases (UP00000589_10090.fa, UP00000589_10090_additional.fa) or human Uniprot databases (UP000005640_9606.fa, UP000005640_9606_additional.fa using the MaxQuant version 1.6.7.0 which extracts features from four-dimensional isotope patterns and associated MS/MS spectra. False-discovery rates were controlled at 1 % both on peptide spectral match (PSM) and protein level. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamidomethylation as fixed modification, while limiting the maximum peptide mass to 4.600 Da. Enzyme specificity was set to trypsin cleaving c-terminal to arginine and lysine. A maximum of two missed cleavages were allowed. Maximum precursor tolerance in the first search and fragment ion mass tolerance were searched as default for TIMS-DDA data. Main search tolerance was set to 20 ppm. The median absolute mass deviation for the data set was 1.57 ppm. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs with a 0.7 min retention-time match window and a 0.05 $1/K_0$ ion mobility window. Label-free quantification was performed with the MaxLFQ algorithm and a minimum ratio count of 1.

Proteomics downstream data analysis. Proteomics data analysis was performed in the Perseus environment (version 1.6.7.0), Prism (GraphPad Software, version 8.2.1). MaxQuant output tables were filtered for ‘Reverse’, ‘Only identified by site modification’, and ‘Potential contaminants’ before further processing. Protein and peptide identifications were reported after filtering as described above. Proteome correlations across technical/analytical/biological replicates were performed after $\log_{10}$-transformation. Coefficients of variation (CVs) were calculated across the full data set or within experimental groups on raw intensity levels for shared observations of more than one. Hierarchical clustering was performed in Perseus with default parameters and Pearson correlation as distance parameters. Before differential expression analysis, data were filtered for at least two observations in one group to be compared, followed by $\log_2$-transformation and imputation from a normal distribution modelled as the data set with a downshift of 1.8 standard deviations and a width of 0.3 standard deviations. Deep proteomes of biological replicates from fresh or vDISCO cleared tissue were tested for differences by a two-sided t-test. False-discovery rate control due to multiple hypothesis testing was performed by a permutation-based model and SAM-statistic with an S_0-parameter of 0.2 and an FDR of 0.01. Ontologies for cellular compartment assignment and keywords was performed with the mainAnnot.Mus_musculus.txt.gz followed by $\log_2$-fold difference frequency counts for the terms ‘Extracellular space’, ‘Blood microparticle’, ‘Neurodegeneration’, ‘Aging’, ‘Neurogenesis’, ‘Receptor’, ‘Virus-Host’, ‘Immunity’, Wound healing’ and ‘Cell migration’. 1D enrichment analysis was performed on the two-sided t-test difference and only enriched terms with a size of larger than ten were displayed in the comparison of fresh versus vDISCO deep proteomes. CVs rank plots were calculated within each of the deep proteome groups and plotted against the median abundance of each protein within each group after $\log_{10}$-transformation.

For the calculation of systematic ontology-related protein mass shifts, total protein copy number estimations of the deep fresh and vDISCO cleared proteomes of biological replicates were calculated using the Perseus plugin ‘Proteomic ruler’. Protein copy numbers were calculated with the following settings: Averaging mode. ‘All columns separately’, Molecular masses: ‘Molecular weight [kDa]’, Scaling mode: ‘Histone proteomic ruler’, Ploidy: ‘2’, Total cellular protein concentration: ‘200g/l’. Proteins were annotated with regards to their cellular compartment by gene ontology from the mainAnnot.mus_musculus.txt.gz. For protein mass estimates, we multiplied the resulting protein copy number by its protein mass for each conditional replicate and summed up all protein masses to obtain the total protein mass for each representative proteome reflecting 100% of the protein mass. To calculate the subcellular protein mass contribution, we calculated the protein mass proportion for the GOCC terms related to the cytoskeleton: ‘Actin filament’, ‘Intermediate filament’, ‘Centrosome’, ‘Microtubule’; Membranes: ‘Cytoplasm’, ‘Plasma membrane’, ‘Membrane’; Organelles: ‘Mitochondrion’, ‘Nucleus’, ‘Endoplasmic reticulum’, ‘Golgi apparatus’. For calculating the organellar change between the respective Fresh and vDISCO sub-proteomes, individual protein mass contributions were normalized by its total proteome mass first, followed by ratio calculation to obtain the percentage shift of protein mass between Fresh and vDISCO brains.

For principal component analysis (PCA) of both LCM applications (mTBI and FAD), data were grouped according to their condition, filtered for at least 760 or 900 proteins for the FAD or mTBI experiment respectively and at least 2 observations within one of the two conditions, column-wise median normalized, and missing values were imputed from a normal distribution with a width of 0.3 standard deviations that was downshifted by 1.8 standard deviations. Differential expression analysis for the FAD and mTBI experiment was performed by two-sided Welch’s t-test on LFQ or IBAQ data respectively. False-discovery rate control due to multiple hypothesis testing was performed by a permutation-based model and SAM-statistic with an S_0-parameter of 0 or 0.2 and an FDR of 0.3 or 0.5 for the mTBI and FAD comparison, respectively.

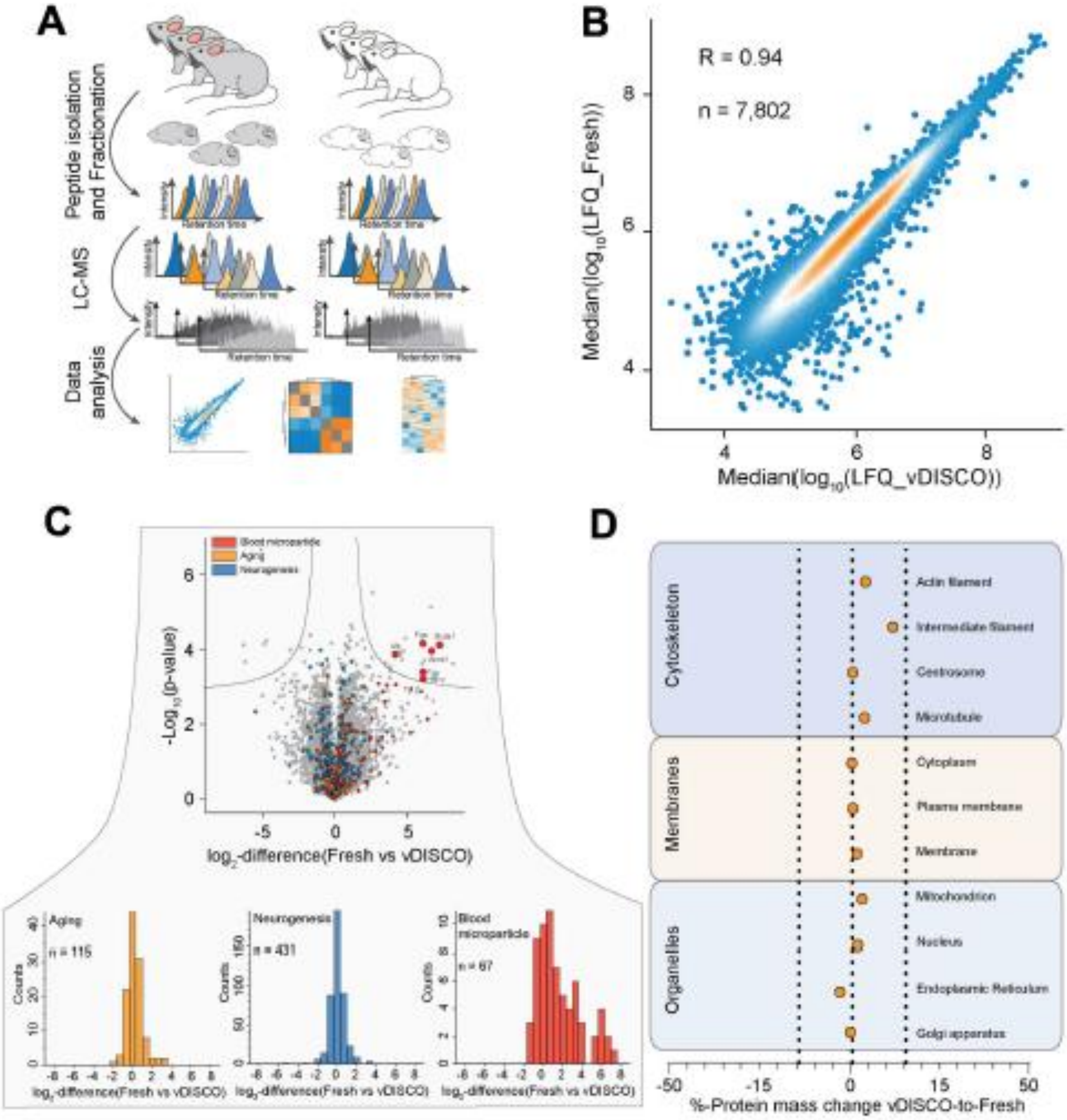
Figure 1: Proteome of cleared rodent and human tissues



(A) Proteome analysis from mouse brain tissues after three different organic solvent-based tissue clearing methods (3DISCO, uDISCO, vDISCO) vs. fresh tissue controls. Protein identifications and proteome correlations across all clearing techniques and fresh tissue are shown (N=3 biological replicates, n=9 total experimental replicates per condition).

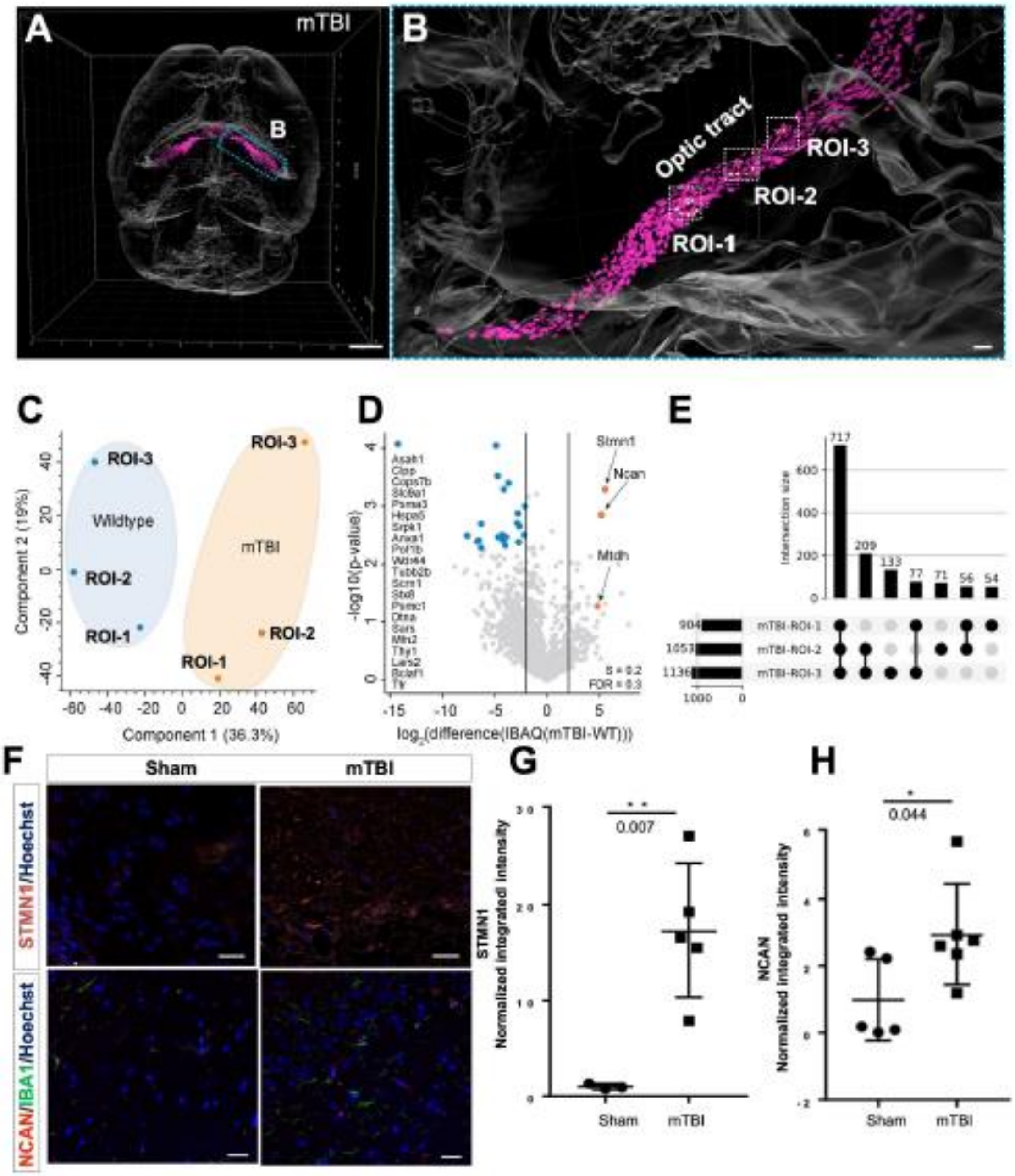
(B) Archived human brain cortex blocks cleared with 3DISCO, uDISCO and SHANEL methods and number of detected protein groups with proteome correlations across all clearing techniques are compared with the detected numbers in PFA fixed blocks (*n*=4 experimental replicates).

Figure 2: Deep proteome analysis after vDISCO tissue clearing



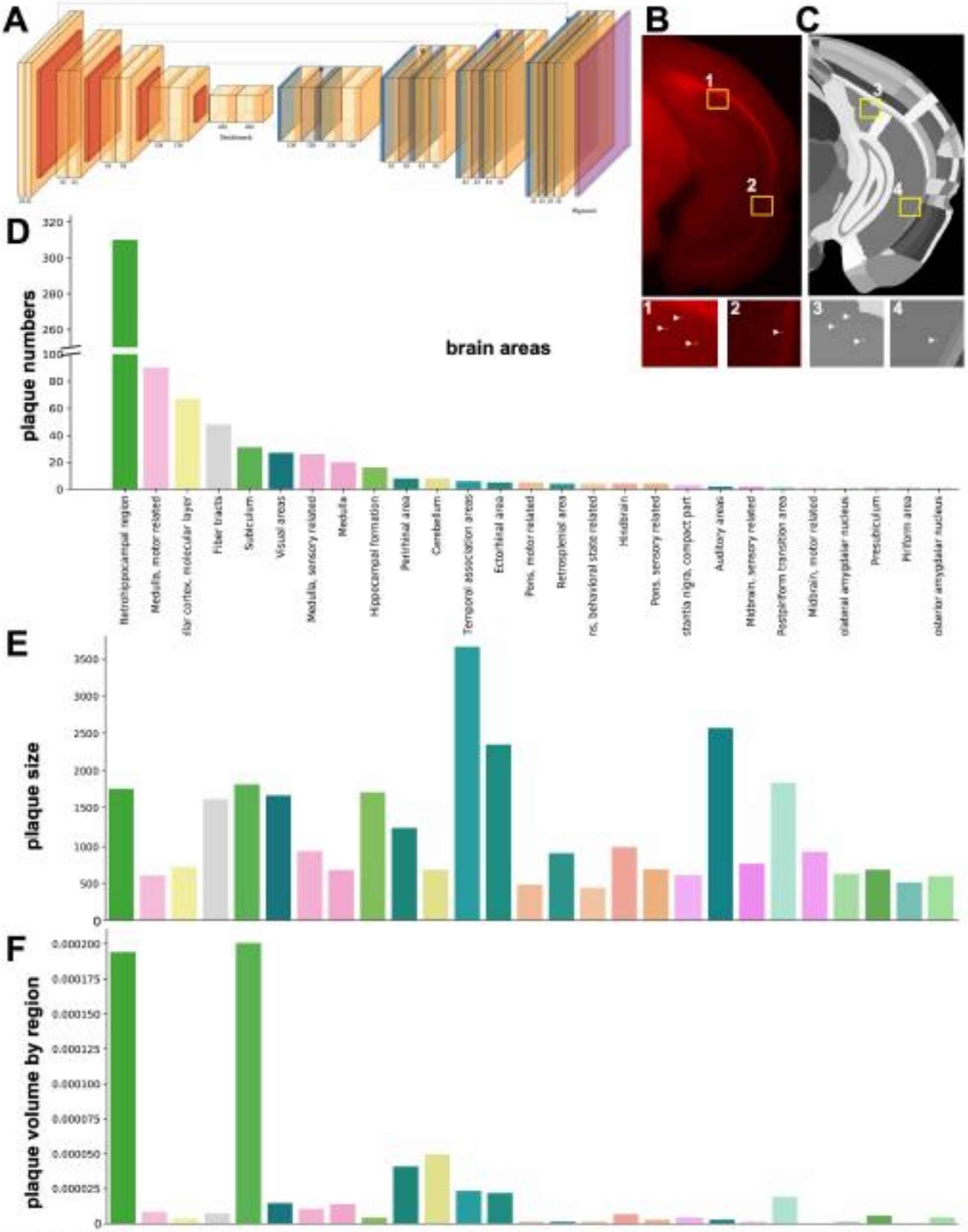
(A) Experimental design for deep preteome analysis after vDISCO clearing vs. snap-frozen fresh tissues. **(B)** Quantitative reproducibility of vDISCO-cleared vs. fresh samples. **(C)** Differential expression analysis of vDISCO-cleared vs. fresh sample proteomes highlighting the expected change in 'blood microparticle' due to blood-flushing step for tissue clearing in contrast to fresh samples with retained-blood. Otherwise, proteins in other gene ontology groups were unchanged (See also **FigureS3**). **(D)** Percentage change of protein mass distribution between vDISCO-cleared vs. fresh biological triplicates. Percentage changes are shown as a median change within one group for organelles, membranes and cytoskeleton gene ontology terms (N=3 biological replicates, n=9 total experimental replicates per condition).

Figure 3: DISCO-MS reveals effects of mTBI in discrete regions of whole brain



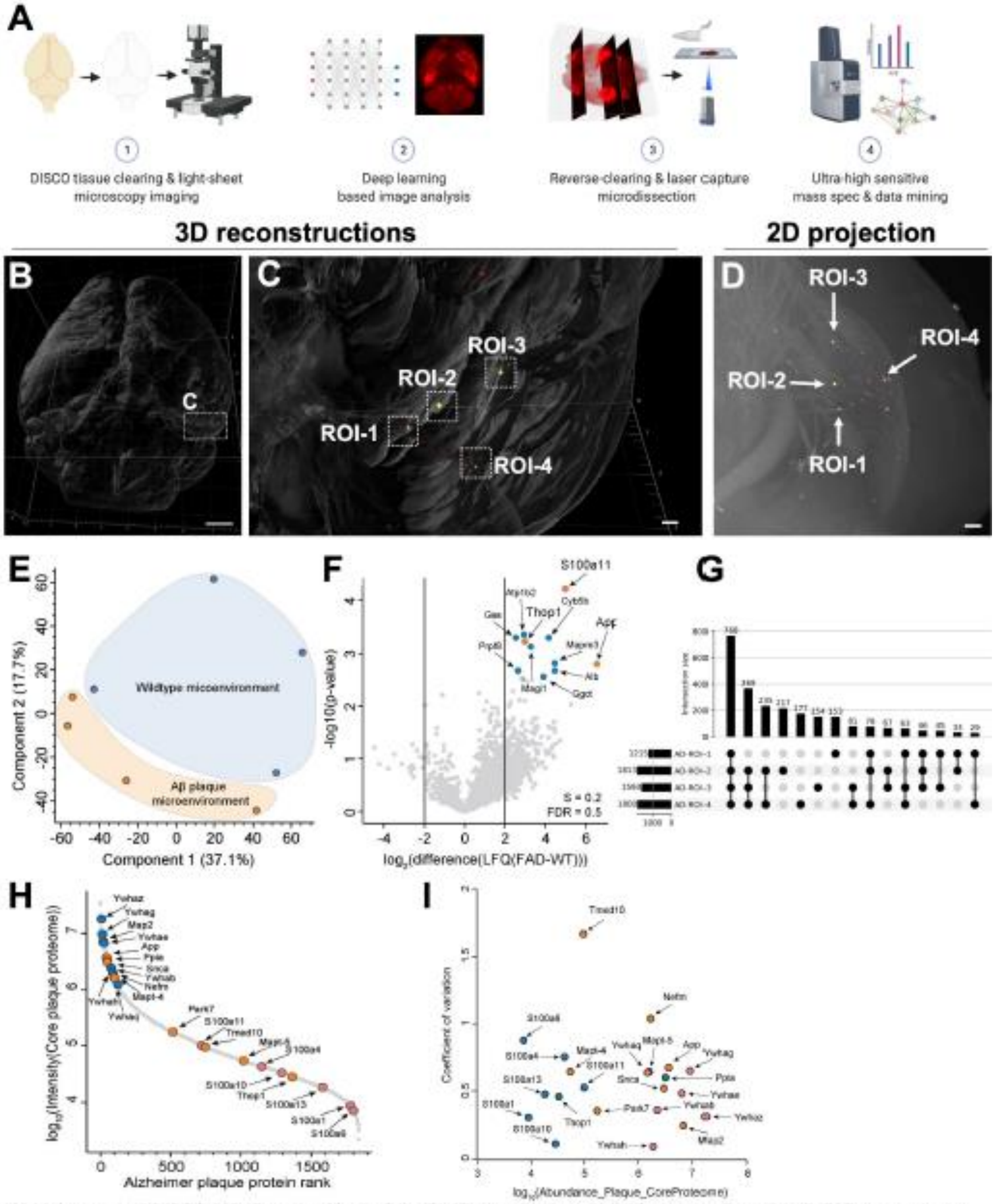
(A) 3D-reconstruction of an exemplary CX3CR1$^{GFP/+}$ mouse brain after mTBI. Segmented microglia shown in magenta. Scale bar, 500 µm. **(B)** High-magnification view of region shown in **(A)**, highlighting substantial increase in activated microglia along optic tract region. 3 neighboring ROIs along optic nerve were identified, laser captured and subjected to proteomic analyses. Scale bar, 100 µm. **(C)** PCA plot showing the distribution of individual ROIs from control mTBI vs. ROIs from control (sham surgery) with the same spatial location. **(D)** Volcano plot showing the significant enrichment of proteins. **(E)** The number of shared and unique set of proteins in ROIs in mTBI. **(F)** Histological validation of top 2 proteins in optic tract: 1) stathmin (STMN1) shown in red and nuclear marker Hoecst dye in blue; 2) neurocan (NCAN) shown in magenta along with microglia marker (IDA1) in green and Hoechst dye in blue. Scale bars, 20 µm. **(G, H)** Intensity quantification of STMN1 immunostaining signal (*P* = 0.007, *n*=3 animals per group, total 9 sections) and NCAN immunostaining signal (*P* = 0.044, *n*=3 animals, total 11 sections) in mTBI vs. sham controls, respectively (unpaired two-sided Student's t-test, data presented as ± SD).

Figure 4: Deep learning analysis of plaques in whole 5xFAD mouse brains



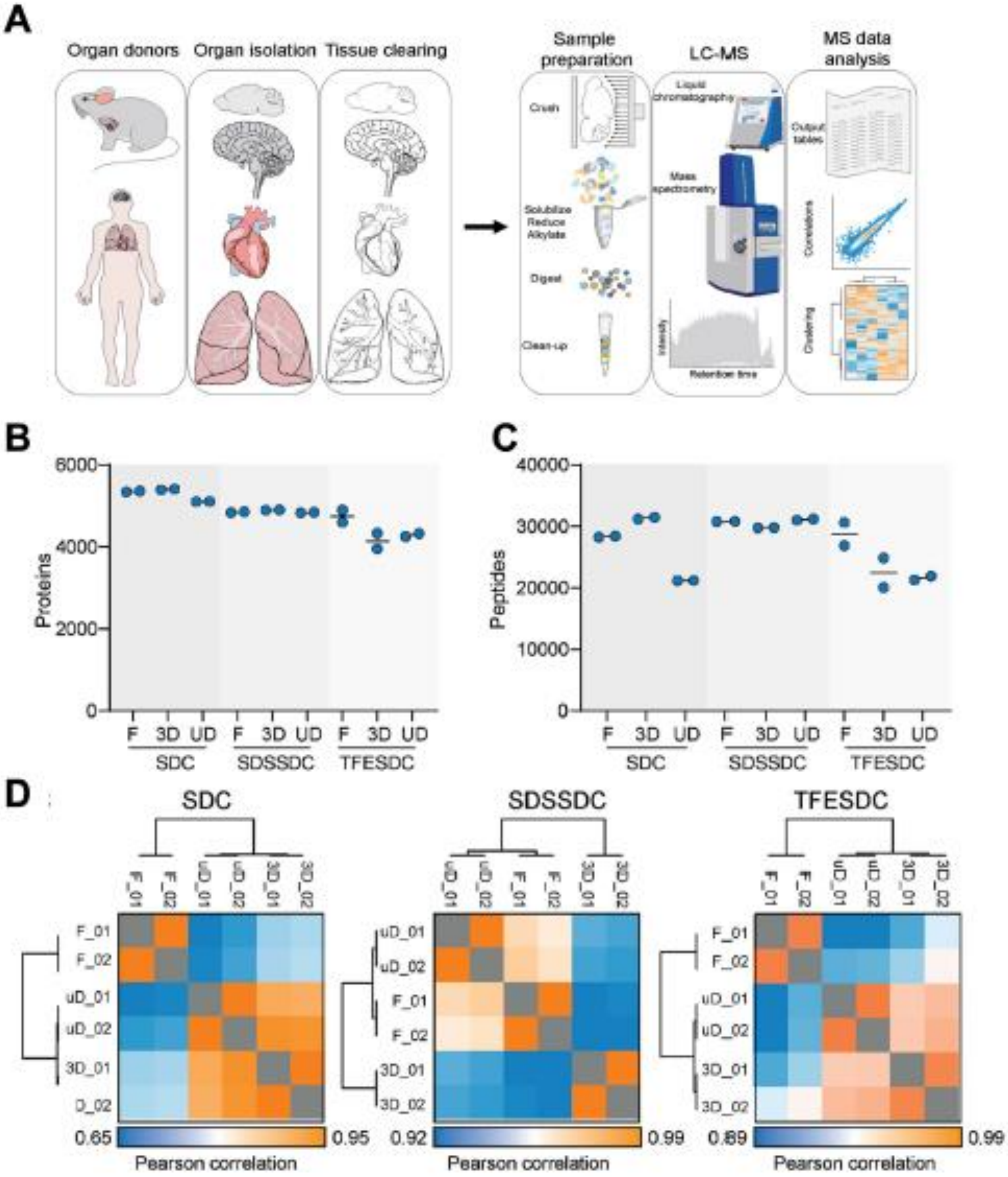
(A) 3D U-Net architecture including layer information and feature sizes. **(B)** Maximum intensity projection of the hippocampal region of the brain with segmented plaques (enlarged in numbered-boxes for better visualization, plaques are pointed by white arrowheads). **(C)** After deep learning-based quantification, the data was registered to Allen brain atlas to obtain region-wise quantification. The atlas regions are randomly grey color-coded to reveal all annotated regions available. **(D)** The number of the plaques in the major brain areas. **(E)** The average plaque size per brain region. **(F)** The density of plaques: the ratio of plaque volume per brain region volume.

Figure 5: DISCO-MS unravels the single plaque proteome in AD mouse model



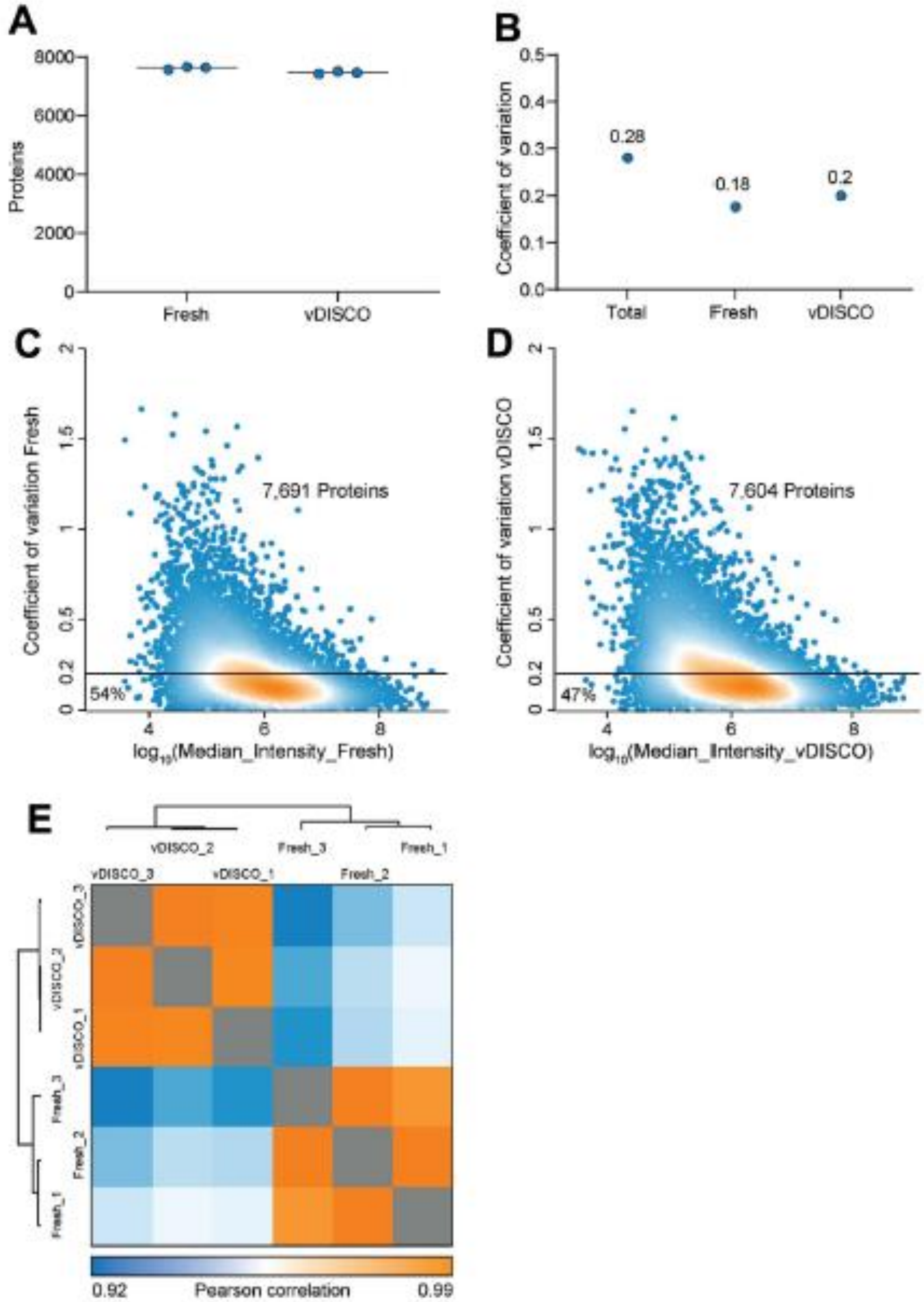
(A) Major steps of DISCO-MS (See also **Movie S1**). **(B)** 3D visualization of all Aβ plaques (in red) in 5xFAD mouse brains at 6 weeks age (*n*=4 experimental replicates). Scale bar, 500 µm. **(C)** Enlarged view of regions marked in **(B).** 4 different ROIs from hippocampus -each containing single plaque- selected and isolated for mass spectrometric measurements. Scale bar, 100 µm. **(D)** 2D projection of selected ROIs. Scale bar, 100 µm. **(E)** PCA plot of ROIs' proteome from 5xFAD vs. the same regions from wildtype controls. **(F)** Volcano plot showing the significantly enriched proteins. **(G)** The number of shared and unique set of proteins in 5xFAD ROIs. **(H)** Rank order of core protein signals in a single plaque microenvironment. **(I)** Log$_{10}$ abundance distribution of selected proteins and protein families as a function of their coefficient of variation (CV) across the core Aβ plaque proteomes. Dynamic range coverage is up to four orders of magnitude. CVs indicate variability in the shared plaque core proteome, among proteins known to play a role in Alzheimer's disease.

Figure S1: Optimization of sample preparation from cleared tissues



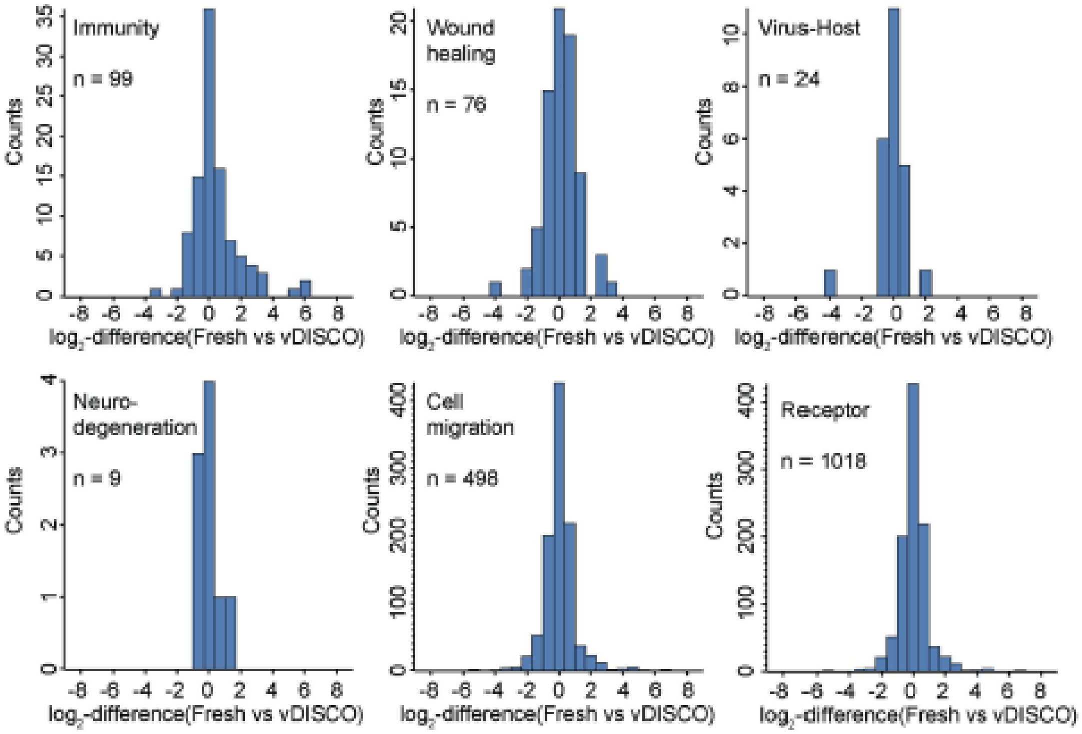
(A) The workflow for optimization of DISCO-MS from clearing bulk tissue to mass spectrometry. Organs can be isolated from any organism followed and cleared by any organic solvent-based tissue clearing. The cleared tissues then subjected to sample preparation workflow we developed for the mass spectrometry analysis. In short: the tissues were solubilized, reduced and alkylated, digested into tryptic proteins and cleaned up ready for liquid chromatography coupled to mass spectrometry (LC-MS) analysis. **(B)** Protein identifications across analytical duplicates for SDC, SDSSDC, or TFESDC preparations coming from fresh vs. cleared mouse brains (3D, uD: 3DISCO and uDISCO clearing methods, respectively). **(C)** Peptide levels of the samples shown in **(B)**. **(D)** Proteome correlation matrices for measurements presented in **(B)** and **(C)**. High Pearson correlations indicate very similar proteomes across conditions in SDSSDC and TFESDC.

Figure S2: Quantitative assessment of proteome in vDISCO-cleared and fresh mouse brain tissues in biological triplicates



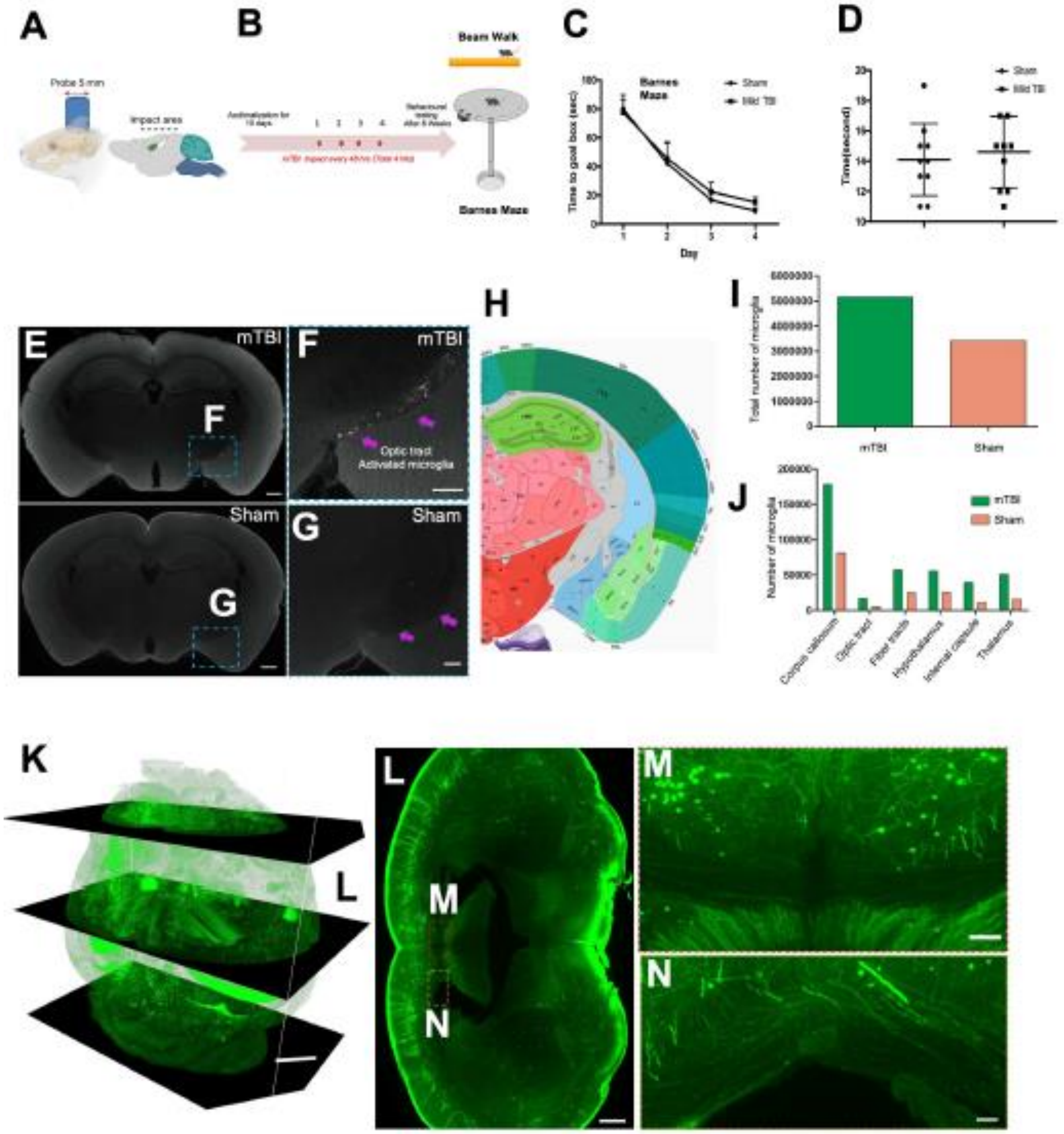
(A) Proteins identified across all three biological replicates for either fresh or vDISCO-cleared tissue. **(B)** Coefficients of variation (CV) for either the total data set including fresh and vDISCO cleared tissue, or fresh/vDISCO only. Note that CVs across biological replicates are low and that CVs across biological triplicates are very similar for fresh and for vDISCO highlighting that proteome of vDISCO-cleared organs is highly reproducible. **(C)** Abundance to CV rank plot for either fresh tissue (left; 7,691 proteins in total; 54% of all proteins are below CV = 0.2) or **(D)** vDISCO cleared tissue (right; 7,604 proteins in total; 47% of all proteins are below CV = 0.2). **(E)** Protein intensity correlation plot for all six biological replicates (3x fresh and 3x vDISCO-cleared).

Figure S3: Fold-changes of gene ontologies between fresh and vDISCO cleared biological triplicate mouse brains



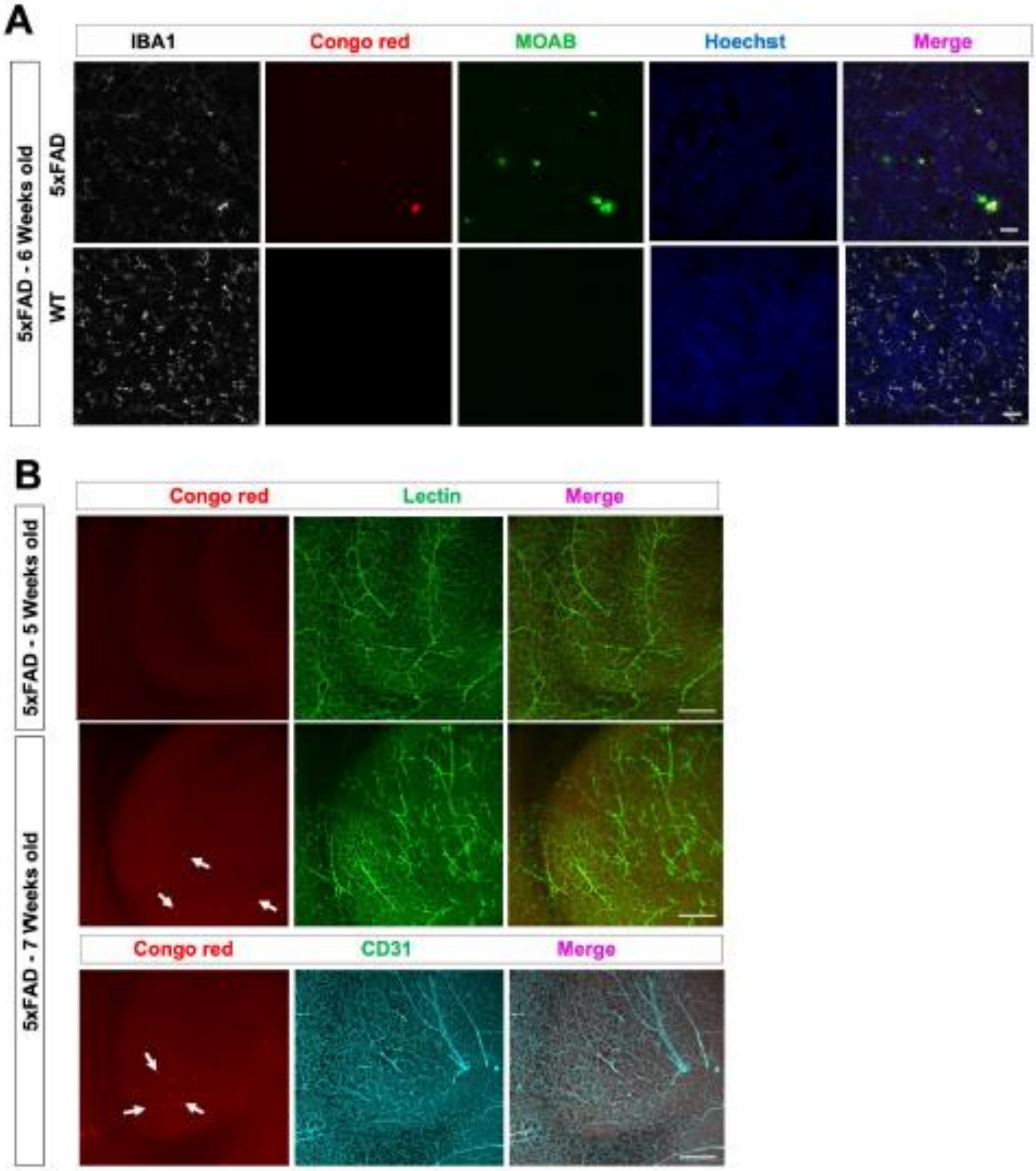
Log2-fold changes for the terms 'Immunity' (99 proteins), 'Wound healing' (76 proteins), 'Virus-Host' (24 proteins), 'Neurodegeneration' (9 proteins), 'Cell migration' (498 proteins) and 'Receptor' (1,018 proteins) between fresh and vDISCO-cleared biological triplicates.

Figure S4: mTBI model validation by behavior and axonal morphology



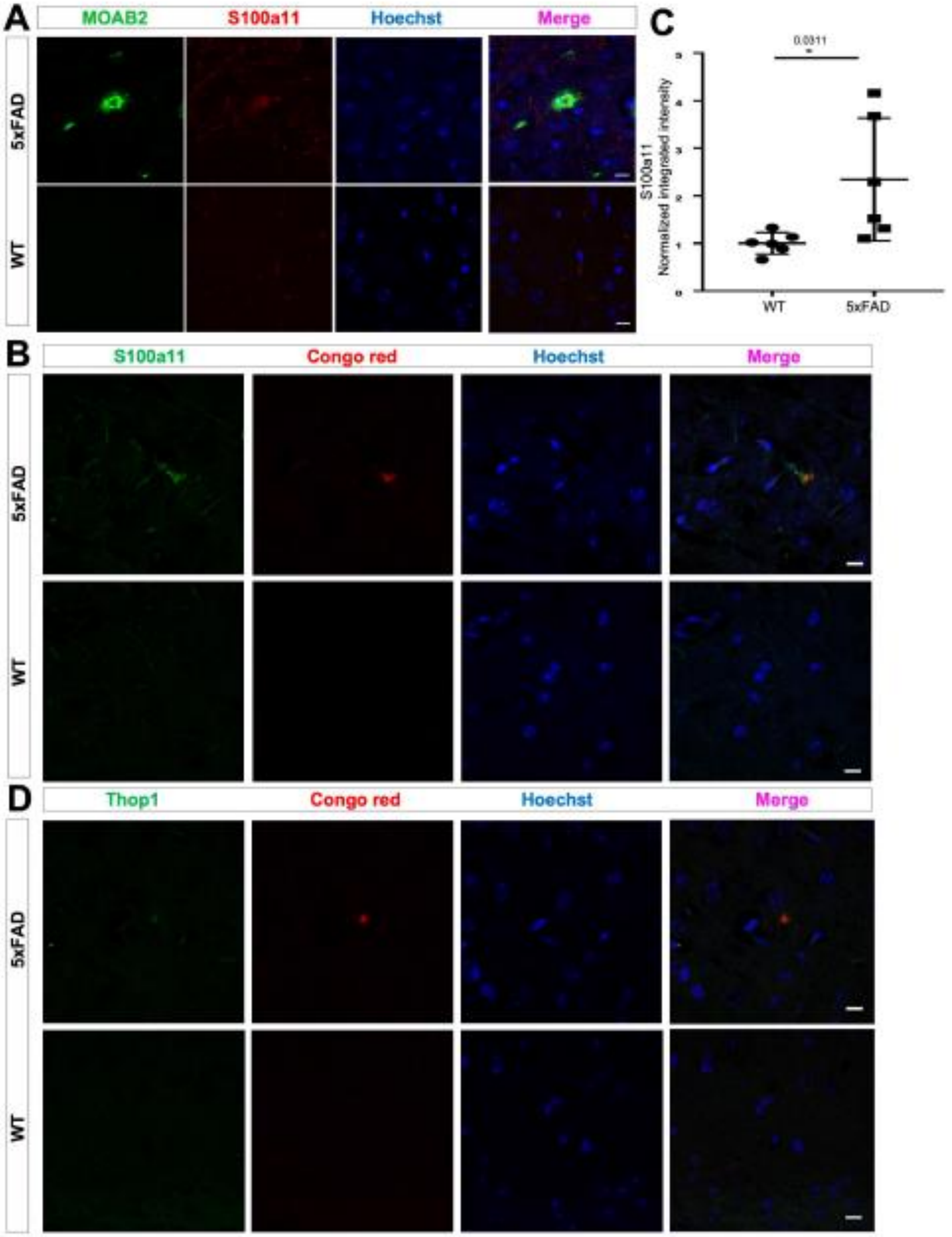
(A) The mTBI impact point on the intact skull is shown. **(B)** Schematic plan of repetitive mTBI experimental mouse model (red points indicate each impact ime). **(C)** Barnes maze test in sham vs. mTBI animals (*n*=10 animals per group). **(D)** Beam walk test in sham and mTBI animals (*n*=10 animals per group). No significant behavior change detected-confirming the "mild" nature of our TBI model. **(E)** Coronal optical slices showing optic nerve with activated microglia. Scale bar, 400 µm. **(F, G)** High magnification image of optic tract in mTBI brain **(F)** vs. the same region from sham control brain **(G)** showing the activated microglia morphology in mTBI brain compared to control brain. Scale bar, 200 µm. **(H)** Corresponding brain regions (coronal view) shown in Allen Brain atlas. **(I)** Quantification of total number of microglia in mTBI vs. Sham animals using ClearMap method. **(J)** Quantification of microglia numbers in mTBI vs. sham mice. Only the regions with major changes are shown. **(K)** 3D view of whole brain from a Thy1-GFP-M after mTBI. Scale bar, 1000 µm. **(L)** 2D orthoslice showing the axonal swellings in corpus callosum (white matter areas). Scale bar, 500 µm. **(M, N)** High magnification images marked in **(L)**. Scale bar, M, 100 µm and N, 50 µm.

Figure S5: Histological validation of Aβ plaques in 5xFAD brains



(A) Tissue histology validation of Congo red plaque staining with a plaque-specific monoclonal antibody plaques (MOAB, green). Furthermore, the microglia were stained using IBA1 antibody (in white) and nuclei using Hoechst dye (in blue). Microglia activation around the plaques of 5xFAD mouse brain is apparent. Scale bars, 20 µm. **(B)** No plaques detected in 5 weeks old 5xFAD mice by Congo red (in red) labeling along with lectin (in green) labeling of vasculature, which provides anatomical information. In contrast, many plaques detected in 7 weeks old 5xFAD mouse brains (pointed by white arrows), whose vasculature were co-stained with lectin (in green) or CD31 antibody (in cyan). Scale bars, 200 µm.

Figure S6: Histological validation of DISCO-MS hits



(A) Histological validation of DISCO-MS hit S100a11 (red) in hippocampal region using antibody immunostaining in 6 weeks old mice. The plaques were co-labeled using the MOAB2 antibody (in green). **(B)** We confirmed the S100a11 signal also around Congo red-labeled plaques in 6 weeks old mice. **(C)** Intensity quantification of S100a11 (*N*=3 animal per group from total 12 sections; unpaired two-sided Student's t-test; *p* = 0.0311; data are presented as average ± SD). **(D)** Histological validation of DISCO-MS hit Thop1 (in green) in hippocampal region of 5xFAD animals along with Congo red-labeled (in red) plaques in 6-weeks old mice. Scale bars, 10 μm.

3.4.2. Article 9: Deep visual proteomics

AI-driven deep visual proteomics defines cell identity and heterogeneity

bioRxiv, January, 2021 (Under review in Nature)

Andreas Mund[1, *, #], Fabian Coscia[1, *, #], Réka Hollandi[4], Ferenc Kovács[4, 5], András Kriston[4, 5], **Andreas-David Brunner[6],** Michael Bzorek[7], Soraya Naimy[7], Lise Mette Rahbek Gjerdrum[7, 13], Beatrice Dyring-Andersen[1, 8, 14], Jutta Bulkescher[3], Claudia Lukas[2, 3], Christian Gnann[9], Emma Lundberg[9, 10, 11], Peter Horvath[4, 5, 12, #], Mathias Mann[1, 6, #]

** These authors contributed equally to this work*

# Correspondence

[1]Proteomics Program, NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[2]Synthetic and Systems Biology Unit, Biological Research Centre, Eötvös Loránd Research Network, Szeged 6726, Hungary

[3]Single-Cell Technologies Ltd, Szeged 6726, Hungary

[4]Protein Signaling Program, NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[5]Protein Imaging Platform, NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[6]Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, 82152 Martinsried, Germany

[7]Department of Pathology, Zealand University Hospital, Roskilde, Denmark

[8]Department of Dermatology and Allergy, Herlev and Gentofte Hospital, University of Copenhagen, Hellerup, Denmark

[9]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, 17121, Sweden.

[10]Department of Genetics, Stanford University, Stanford, CA 94305, USA.

[11]Chan Zuckerberg Biohub, San Francisco, San Francisco, CA 94158, USA.

[12]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 00014, Finland

[13]Institute for Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

[14]Leo Foundation Skin Immunology Research Center, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Contribution

In this project, I contributed to the experimental design of the paper and transferred the achievements in ultra-high sensitivity mass spectrometry for single-cell proteomics to our sibling group in Copenhagen. Furthermore, I ensured the sound data acquisition of cell-type resolved pooled single cell proteomes on the novel hardware and supervised the mass spectrometry experiments.

AI-driven Deep Visual Proteomics defines cell identity and heterogeneity

Andreas Mund[1,*,#], Fabian Coscia[1,*], Réka Hollandi[4], Ferenc Kovács[4,5,], András Kriston[4,5], Andreas-David Brunner[6], Michael Bzorek[7], Soraya Naimy[7], Lise Mette Rahbek Gjerdrum[7,13], Beatrice Dyring-Andersen[1,8,14], Jutta Bulkescher[3], Claudia Lukas[2,3], Christian Gnann[9], Emma Lundberg[9,10,11], Peter Horvath[4,5,12,#], Matthias Mann[1,6,#,§]

[1]Proteomics Program, [2]Protein Signaling Program, and [3]Protein Imaging Platform, NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[4]Synthetic and Systems Biology Unit, Biological Research Centre, Eötvös Loránd Research Network, Szeged 6726, Hungary

[5]Single-Cell Technologies Ltd, Szeged 6726, Hungary

[6]Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, 82152 Martinsried, Germany

[7]Department of Pathology, Zealand University Hospital, Roskilde, Denmark

[8]Department of Dermatology and Allergy, Herlev and Gentofte Hospital, University of Copenhagen, Hellerup, Denmark

[9]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, 17121, Sweden.

[10]Department of Genetics, Stanford University, Stanford, CA 94305, USA.

[11]Chan Zuckerberg Biohub, San Francisco, San Francisco, CA 94158, USA.

[12]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 00014, Finland

[13]Institute for Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

[14]Leo Foundation Skin Immunology Research Center, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

*Authors contributed equally

Correspondence

§ Lead contact: mmann@biochem.mpg.de

ABSTRACT

The systems-wide analysis of biomolecules in time and space is key to our understanding of cellular function and heterogeneity in health and disease[1]. Remarkable technological progress in microscopy and multi-omics technologies enable increasingly data-rich descriptions of tissue heterogeneity[2,3,4,5]. Single cell sequencing, in particular, now routinely allows the mapping of cell types and states uncovering tremendous complexity[6]. Yet, an unaddressed challenge is the development of a method that would directly connect the visual dimension with the molecular phenotype and in particular with the unbiased characterization of proteomes, a close proxy for cellular function. Here we introduce Deep Visual Proteomics (DVP), which combines advances in artificial intelligence (AI)-driven image analysis of cellular phenotypes with automated single cell laser microdissection and ultra-high sensitivity mass spectrometry[7]. DVP links protein abundance to complex cellular or subcellular phenotypes while preserving spatial context. Individually excising nuclei from cell culture, we classified distinct cell states with proteomic profiles defined by known and novel proteins. AI also discovered rare cells with distinct morphology, whose potential function was revealed by proteomics. Applied to archival tissue of salivary gland carcinoma, our generic workflow characterized proteomic differences between normal-appearing and adjacent cancer cells, without admixture of background from unrelated cells or extracellular matrix. In melanoma, DVP revealed immune system and DNA replication related prognostic markers that appeared only in specific tumor regions. Thus, DVP provides unprecedented molecular insights into cell and disease biology while retaining spatial information.

The deep visual proteomics concept

The versatility, resolution and multi-modal nature of modern microscopy delivers increasingly detailed images of single cell heterogeneity and tissue organization[8]. However, a pre-defined subset of proteins is usually targeted, far short of the actual complexity of the proteome. Taking advantage of recent developments in mass spectrometry (MS)-based technology, especially dramatically increased sensitivity, we here set out to enable the analysis of proteomes within their native, subcellular context to explore their contribution to health and disease. We developed a concept called Deep Visual Proteomics (DVP) that combines high-resolution imaging, artificial intelligence (AI)-based image analysis for single-cell phenotyping and isolation with a novel ultra-sensitive proteomics workflow[7] (Fig. 1). A key challenge in realizing the DVP concept turned out to be the accurate definition of single cell boundaries and cell classes as well as the transfer of the AI defined features into proteomic samples, ready for analysis. To this end, we introduce the software 'BIAS' (Biology Image Analysis Software), which coordinates scanning and laser microdissection microscopes. This seamlessly combines data-rich imaging of cell cultures or archived biobank tissues (formalin-fixed and paraffin-embedded, FFPE) with deep learning-based cell segmentation and machine learning-based identification of cell types and states. Cellular or subcellular objects of interest are selected by the AI alone or after instruction and subjected to automated laser microdissection and proteomic profiling. Data generated by DVP can be mined to discover novel protein signatures providing molecular insights into proteome variation at the phenotypic level with full spatial meta-information. We show below that this concept provides a powerful, multi-layered resource for researchers with applications ranging from functional characterization of single cell heterogeneity to spatial proteomic characterization of disease tissues with the aim of assisting clinical decision-making.

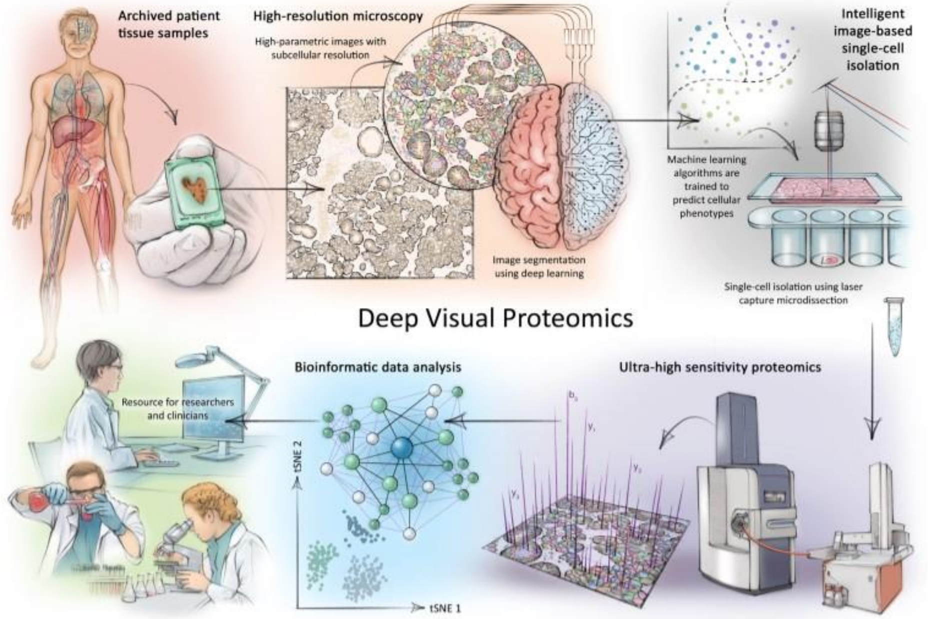

Archived patient tissue samples
High-resolution microscopy
High-parametric images with subcellular resolution
Image segmentation using deep learning
Intelligent image-based single-cell isolation
Machine learning algorithms are trained to predict cellular phenotypes
Single-cell isolation using laser capture microdissection
Deep Visual Proteomics
Bioinformatic data analysis
Ultra-high sensitivity proteomics
Resource for researchers and clinicians
tSNE 2
tSNE 1

Fig.1: Deep Visual Proteomics concept and workflow
Deep Visual Proteomics (DVP) combines high-resolution imaging, artificial intelligence (AI)-guided image analysis for single-cell classification and isolation with a novel ultra-sensitive proteomics workflow[7]. DVP links data-rich imaging of cell culture or archived patient biobank tissues with deep learning-based cell segmentation and machine learning based identification of cell types and states. (Un)supervised AI-classified cellular or subcellular objects of interests undergo automated laser microdissection and mass spectrometry (MS)-based proteomic profiling. Subsequent bioinformatic data analysis enables data mining to discover protein signatures providing molecular insights into proteome variation in health and disease states at the level of single cells. DVP serves as resource for researchers and clinicians.

The image processing and single cell isolation workflow

The microscopy-related aspects of the DVP workflow build on state-of-the-art high-resolution and whole-slide imaging as well as machine learning and deep learning (ML and DL) for image analysis. For the required pipeline, further advances in our image analysis software were needed, as well as downstream, automated, rapid single-cell laser microdissection.

First, we used scanning microscopy to obtain high-resolution whole-slide images and developed a software suite for integrative image analysis termed 'BIAS' (Methods). BIAS allows the processing of multiple 2D and 3D microscopy image file formats, supporting the major microscope vendors and data formats. It combines image preprocessing, deep learning-based image segmentation, feature extraction and machine learning-based phenotype classification. Building on a novel deep learning-based algorithm for cytoplasm and nucleus segmentation[9], we undertook several optimizations to implement pre-processing algorithms to maintain high quality images across large image datasets. Deep learning methods require large training datasets, a major challenge due to the limited size of high-quality training data[10]. To address this challenge, we used NucleAIzer[9] and applied project-specific image style transfer to synthetize artificial microscopy images resembling real images. This approach is inherently adaptable to different biological scenarios such as new cell and tissue types or staining techniques. We trained a deep learning model with these synthetic images for specific segmentation of the cellular compartment of interest (e.g. nucleus or cytoplasm, Fig. 2A), and benchmarked it against two leading deep learning approaches: unet4nuclei[11] and Cellpose[12] and a widely-used adaptive threshold- and object splitting-based method[13]. Notably, our deep learning algorithms for cell and nucleus segmentation of cell cultures and tissues showed the highest accuracy (Fig. 2A, Suppl. S1). For interactive cellular phenotype discovery, BIAS performs phenotypic feature extraction taking into account morphology and neighborhood features based on supervised and unsupervised machine-learning (Fig 2B, Methods). Importantly, we can combine feature-based phenotypic classification with biomarker expression levels from antibody staining for precise cell classification.

To physically extract the cellular features discovered with BIAS, we developed an interface between scanning and laser microdissection microscopes (currently ZEISS PALM MicroBeam and Leica LMD6 & 7) (Fig. 2C). BIAS transfers cell contours between the microscopes, preserving full accuracy. Laser microdissection has a theoretical accuracy of 70 nm using a 150x objective and in practice we reached 200 nm. After optimization the LMD7 allows the excision of 700 collected high-resolution contours per hour, with full remote and automated operation (Methods). To prevent potential laser-induced damage, contours can be excised with a definable offset (Fig. 2C, D, video 1,2). In summary, BIAS successfully unifies scanning and laser microdissection microscopy on the basis of AI-driven image analysis.

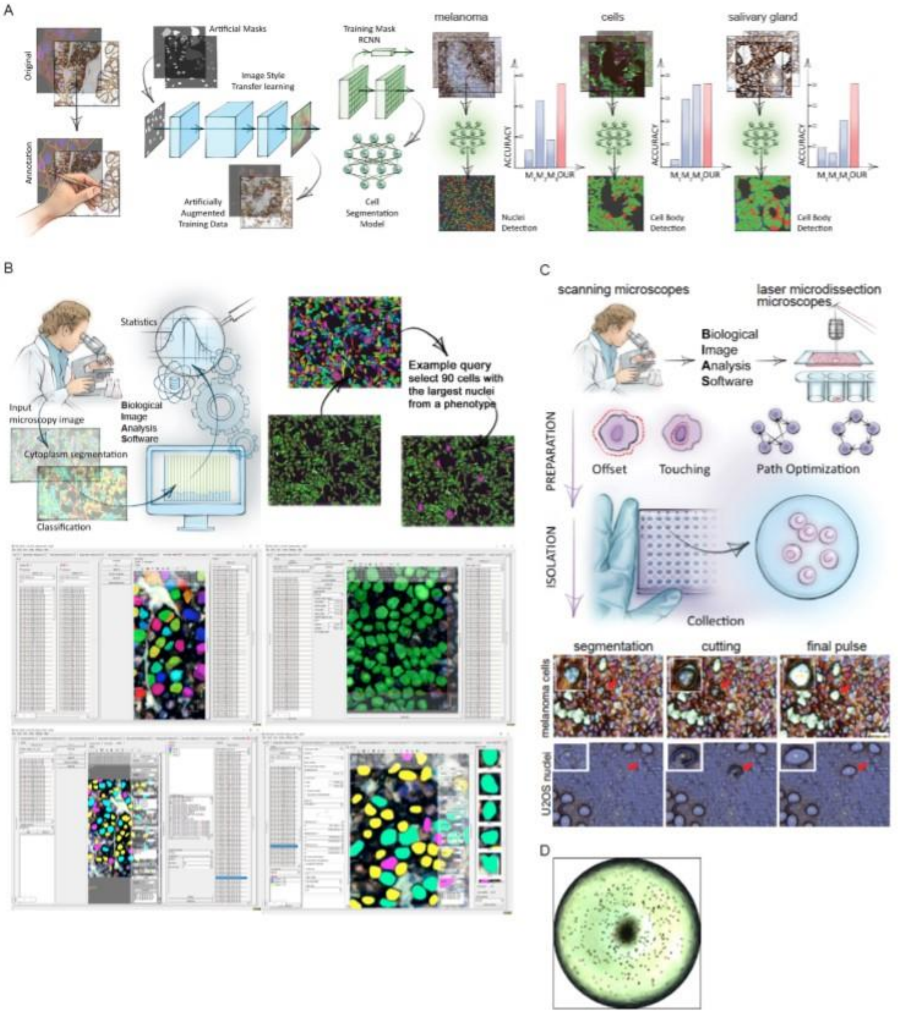


Figure 2. BIAS for integrative image analysis and automated LMD single-cell isolation

A. Left: AI-driven nuclei and cytoplasm segmentation of normal appearing and cancer cells and tissue using image style transfer learning in the Biological Image Analysis Software (BIAS), developed here. Right: We benchmarked the accuracy of its segmentation approach using the F1 metric and compared results to three additional methods M1-M3. Visual representation of the segmentation results: green areas correspond to true positive, blue to false positive and red to false negative. **B.** BIAS allows the processing of multiple 2D and 3D microscopy image file formats. Examples for image preprocessing, deep learning-based image segmentation, feature extraction and machine learning-based phenotype classification. **C.** BIAS also serves as the interface between the scanning and a laser microdissection microscope, allowing high accuracy transfers of cell contours between the microscopes. Upper panel: conceptual overview of cutting functions, cutting offset with respect to the object of interest and optimal path finding. Lower panel: Practical illustration of the functions in the upper panel. **D.** Captured single nuclei can be quality controlled in collection wells

DVP defines single cell heterogeneity at the subcellular level

To determine if DVP can characterize functional differences between phenotypically distinct cells, we applied our workflow to an unperturbed cancer cell line (FUCCI - U2OS cells[14]). After deep learning-based segmentation for nuclei and cell membrane detection, we isolated 80-100 single cells or 250-300 nuclei per experiment (Fig. 2A, 3A, B). Although the analysis of small numbers of tissue cells by MS has been a long-standing goal, transfer, processing and analysis of these minute samples pose formidable analytical challenges[15] which we addressed in turn. We processed samples on the basis of a recently developed workflow for ultra-low sample input[7,16], that omits any sample transfer steps and ensures decrosslinking in very low volumes (Methods). We found that samples could be analyzed directly from 384 wells without any additional sample transfer or clean-up. MS measurements were performed with a data independent acquisition method using the parallel accumulation – serial fragmentation acquisition method an additional ion mobility dimension and optimal fragment (diaPASEF) ion usage on a newly developed mass spectrometer[17,7]. Replicates of cell and nucleus proteomes demonstrated a robust workflow with high quantitative reproducibility (Pearson r = 0.96). Proteomes of whole cells were very different from those of nuclei alone, as in subcellular proteomics experiments based on biochemical separation[18] (Fig. 3C, Fig. S2A). This was likewise reflected in the bioinformatic enrichment analysis, with terms like plasma membrane, mitochondrion, nucleosomes and transcription factor complexes being highly significant (FDR < 10^{-5}) (Fig. 3D).

To address if morphological differences between nuclei are also reflected in their proteomes, we used an unsupervised phenotype finder model to identify groups of morphologically distinct nuclei based on nuclear area, perimeter, form factor, solidity and DNA staining intensity (Fig. 3E). ML found three main nuclei classes (27-37% each) and also discovered three rare ones (2-4% each) (Fig. 3F). The resulting six distinct nuclei classes had visible differences in size and shape. Class 1 represented mitotic states while the remaining five were in interphase with varying feature weighting (Fig. 3G, H). For subsequent analysis, we focused on those five nuclei classes of unknown state. In principal component analysis (PCA), replicates of the respective proteomes clustered closely and the more frequent classes (2, 3 and 5) grouped together (Fig. 3I). The rare classes 4 and 6, which were mainly driven by their unique morphologies (large nuclei (> 4n) and 'bean-shaped', respectively) separated in component 1 and 2 from this group. To verify and quantify this observation, we compared each cell class proteome to a proteome of all nuclei in a field of view. This revealed that the rarest cell classes had the largest numbers of differentially expressed proteins compared to unclassified proteomes (Fig. 3J, S2B). These results demonstrate that the differences visible by microscopy translate into quantifiable proteomic differences and highlights that subcellular phenotypes are linked to distinct proteome profiles.

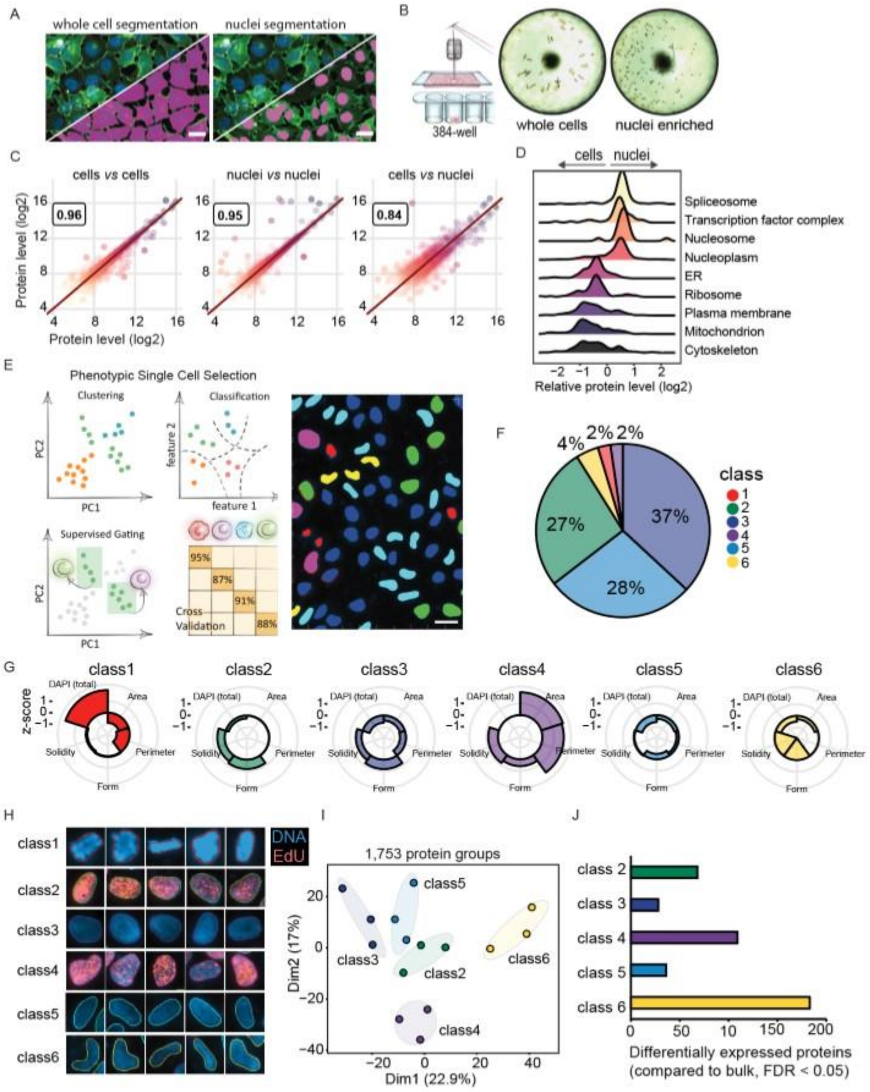


Fig.3: DVP defines single cell heterogeneity at the subcellular level

A. Deep learning-based segmentation of whole cells and nuclei in BIAS of DNA (DAPI) stained U2OS FUCCI cells. Scale bar =20μm **B.** Automated laser microdissection of whole cells and nuclei into 384-well plates. Images show wells after collection. **C.** Quantitative proteomic results of whole cell and nuclei replicates, and comparison between whole cells and nuclei. **D.** Relative protein levels (X-axis) of major cellular compartments between whole cell and nuclei specific proteomes. Y-axis displays point density. **E.** Left: Conceptual workflows of the phenotype finder model of BIAS for machine learning-based classification of cellular phenotypes. Right: Results of unsupervised ML-based classification of six distinct U2OS nuclei classes based on morphological features and DNA staining intensity. Colors represent classes. Scale bar = 20μm. **F.** Relative proportions of the six nuclei classes. **G.** Phenotypic features used by ML to identify six distinct nuclei classes. Radar plots show z-scored relative levels of morphological features (nuclear area, perimeter, solidity and form factor) and DNA staining intensity (total DAPI signal). **H.** Example images of nuclei from the six classes identified by ML. Blue color shows DNA staining intensity and red color 5-ethynyl-2'-deoxyuridine (EdU) staining intensity to identify cells undergoing replication. Represented nuclei are enlarged for visualization and do not reflect actual sizes **I.** Principal component analysis (PCA) of five interphase classes based on 1,753 protein groups after data filtering. Replicates of classes are highlighted by ellipses with a 95% confidence interval. **J.** Number of differentially expressed proteins compared to unclassified nuclei (bulk). Proteins with an FDR less than 0.05 were considered significant.

We next asked if the proteomic differences across the five nuclei classes could give clues to the functional differences between the interphase states (Fig. 3E, H). The 361 significantly differentially expressed proteins across classes were enriched for nuclear and cell cycle related proteins (e.g. 'DNA unwinding involved in replication' and 'condensation of prophase chromosomes'), implicating the cell cycle as a strong biological driver (Fig. 4A, B). To confirm this, we compared our data to a single-cell imaging dataset including 574 cell cycle regulated proteins [19] and found that they were 2.3-fold enriched in our significantly regulated proteins (FDR $< 10^{-6}$). In addition, nuclear area, one of the driving features between our classes, increases from G1 to S/G2 cells during interphase (Fig. 3G, Fig. S3A-C), supporting the importance of the cell cycle in defining the nuclei classes.

We were intrigued that our single cell type proteomes contained a number of uncharacterized proteins, offering an opportunity to associate them with a potential cellular function. Focusing on three open reading frame (ORF) proteins remaining after data filtering, two of them - C7orf50 and C1orf174 - showed class specific expression patterns ($p < 0.01$, Fig. S3D). C7orf50 was most highly expressed in the nucleoli of classes 2 and 4 nuclei, which showed S/G2 specific characteristics (Fig. 3H, S3D, E), suggesting that its expression is cell cycle regulated. Indeed, we confirmed higher levels of C7orf50 in G1/S and S/G2 compared to G1 phase cells (Fig. S3E). As cell cycle regulated proteins tend to be over-represented in cancer prognostic associations[19], we investigated C7orf50 in the human pathology atlas[20] and found high expression was associated with favorable outcome in pancreatic cancer (Fig. S3G, $p < 0.001$). Its interaction, co-expression and co-localization with the protein LYAR ('Cell Growth-Regulating Nucleolar Protein') suggests a functional link between them and a potential role in cell proliferation and cancer growth (Fig. S3F, H).

Although our data revealed the cell cycle to be a strong driver of nuclei classification, class 6 showed a pronounced proteomic signature independent of known cell cycle markers (Fig. 4C, D). In these rare, bean-shaped nuclei, cytoskeletal and cell adhesion proteins (e.g. VIM, TUBB, ACTB and ITGB1) were upregulated, suggesting that they derived from migrating cells undergoing nuclear deformation, reminiscent of what has recently been described in the context of cell invasion[21, 22]. Note that we classified nuclei from 2D images but laser cutting isolates them in 3D, thus samples also probe morphology-driven protein re-localization around the nucleus as exemplified by class 6 nuclei.

These cell culture experiments establish that DVP correlates cellular phenotypes, heterogeneity and dynamics with the proteome level in an unbiased way. Common and rare phenotypes can be studied in their natural environment, preserving the native biological variation within cancer cells to link phenotypes to proteomic make-up.

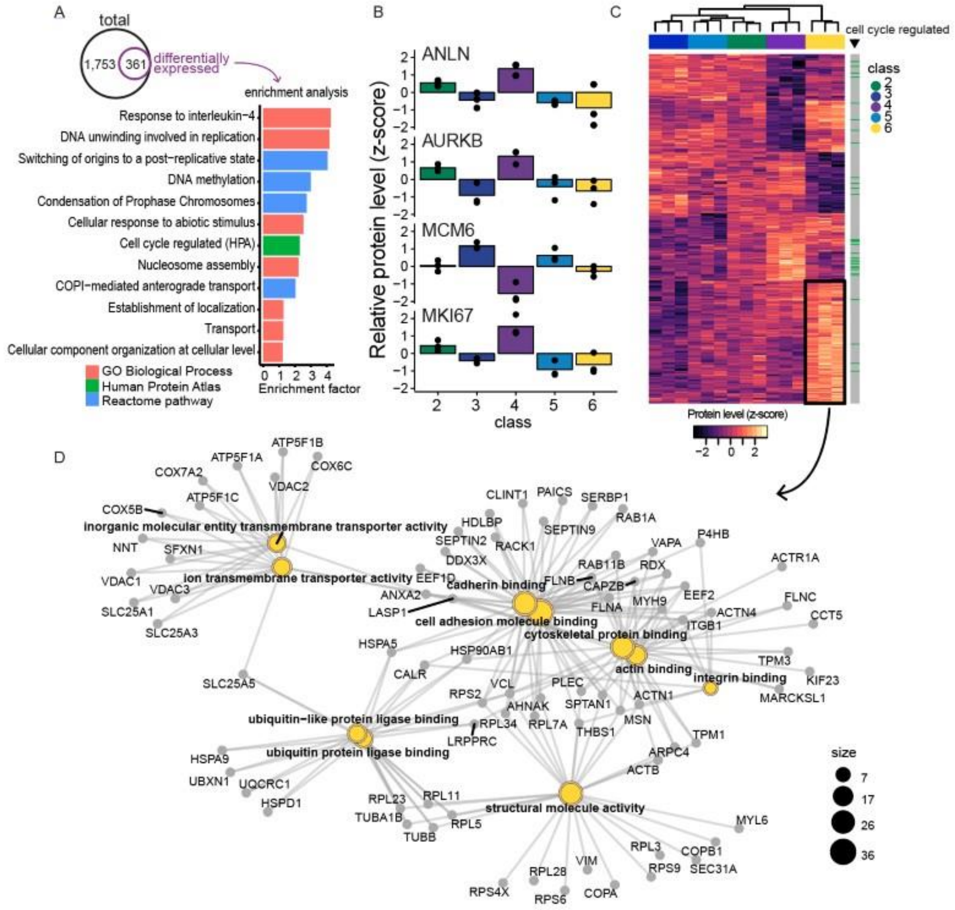


Fig.4: DVP links cellular phenotypes to functional protein networks in U2OS cancer cells

A. Bioinformatic enrichment analysis of proteins regulated between the five nuclei classes. Significant proteins (361 ANOVA significant, FDR < 0.05, $s_0 = 0.1$) were compared to the set of unchanged proteins based on Gene Ontology Biological Process (GOBP), Reactome pathways and cell cycle regulated proteins reported by the Human Protein Atlas[20]. A Fisher's exact test with a Benjamini-Hochberg FDR of 0.05 was used. **B.** Relative protein levels (z-score) of known cell cycle markers across the five nuclei classes. **C.** Unsupervised hierarchical clustering of all 361 ANOVA significant protein groups. Cell cycle regulated proteins reported by the HPA are shown in green in the right bar. Nuclei classes are shown in the column bar. Proteins upregulated in class 6-nuclei (yellow) are indicated by the black box. **D.** Enrichment analysis of proteins upregulated in class 6-nuclei (black box in panel C). Relationship between the top 10 most significantly enriched GO terms and proteins are shown. Node sizes represent the number of genes in each category.

DVP applied to cancer tissue heterogeneity

We next explored if DVP could provide unbiased proteomic profiling of distinct cell classes in their individual spatial environments at high resolution. Billions of patient samples are collected routinely during diagnostic workup and biobanked in the archives of pathology departments around the world[23]. Therefore, the precise proteomic characterization of single cells in their spatial and subcellular context from these samples could have great clinical impact as an extension of the emerging field of digital pathology[24] . We selected archived tissue of a salivary gland acinic cell carcinoma, a rare and understudied malignancy of the epithelial secretory cells of the salivary gland that arises from their epithelial secretory cells. First, we developed an immunohistochemical staining protocol on glass membrane slides for routine histopathology and stained the tissue for EpCAM to outline the cellular boundaries for segmentation and feature extraction by BIAS (Methods). This revealed normal appearing and neoplastic regions with different cellular composition. While the former was mainly comprised of acinar, duct and myoepithelial cells, the carcinoma was dominated by uniform tumor cells with round nuclei and abundant basophilic cytoplasm (Fig. 5A, B).

To identify disease specific protein signatures, we wished to directly compare the normal appearing acinar cells with the malignant cells, rather than admixing it with varying proportions of unrelated cells. To this end, we classified acinar and duct cells from normal parotid gland tissue based on their cell-type specific morphological features and isolated single cell classes for proteomic analysis (Fig. 5C and S4A). Bioinformatics of the resulting proteomes revealed strong biological differences between these neighboring cell types, reflecting their distinct physiological functions. Acinar cells, which produce and secrete saliva in secretory granules, showed high expression of proteins related to vesicle transport and glycosylation along with known acinar cell markers such as α-amylase (AMY1A), CA6 and PIP (Fig. S4B). In contrast, duct cells, which are rich in mitochondria to deal with the energy demand for transcellular saliva secretion[25], indeed expressed high levels of mitochondria and metabolism related proteins (Fig. S4B). For comparison, we exclusively excised malignant and benign acinar cells from the various regions within the same tissue section. Interestingly, the proteomes of acinar cells clustered together regardless of disease state, demonstrating a strong cell of origin effect (Fig. S4C). Building on this foundation, we analyzed six replicates of normal appearing and nine of neoplastic regions, which showed excellent within-group quantitative reproducibility (Pearson $r > 0.96$). Correlation between normal and cancer were lower, reflecting disease and cell-type specific proteome changes (r = 0.8, Fig. 5D, E). Acinar cell markers in the carcinoma were significantly downregulated, consistent with previous reports[25]. We discovered an up-regulation of interferon-response proteins (e.g. MX1, HLA-A, HLA-B, SOD2) and the proto-oncogene SRC, a well-known therapeutic target[26], together with a treasure-trove of other proteins. This highlights the ability of DVP to discover disease-specific and therapeutically relevant proteins on the basis of cell-type resolved tissue proteomics.

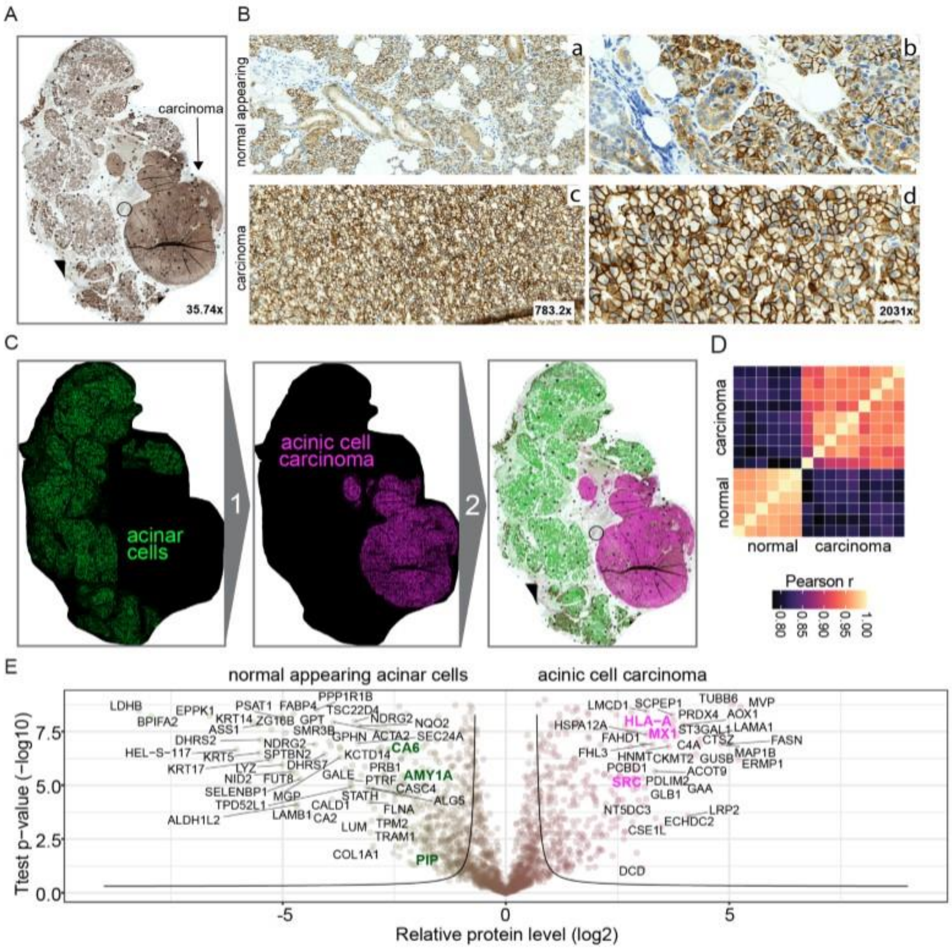


Fig.5: DVP applied to archived tissue of a rare salivary gland carcinoma

A. Immunohistochemical staining of an acinic cell carcinoma of the parotid gland by the cell adhesion protein EpCAM. **B.** Representative regions from normal appearing tissue (upper panels a and b) and acinic cell carcinoma (lower panels c and d) from A. **C.** DVP workflow applied to the acinic cell carcinoma tissue. Deep learning-based single cell detection of normal appearing (green) and neoplastic (magenta) cells positive for EpCAM. Cell classification based on phenotypic features (form factor, area, solidity, perimeter, EpCAM intensity). **D.** Proteome correlations of replicates from normal appearing (normal, n=6) or cancerous regions (cancer, n=9). **E.** Volcano plot of pairwise proteomic comparison between normal and cancer tissue. T-test significant proteins (FDR < 0.05, s_0 = 0.1) are highlighted by black lines. Proteins higher in normal tissue are highlighted in green on the left of the volcano including known acinic cell markers (AMY1A, CA6, PIP). Proteins higher in the acinic cell carcinoma are on the right in magenta, including the proto-oncogene SRC and interferon-response proteins (MX1, HLA-A).

To investigate if DVP could resolve different states of the same cell type, we next applied it to melanoma, a highly aggressive and heterogeneous cancer associated with poor outcomes in advanced stages[27,28]. We chose an FFPE-preserved tissue specimen archived 18 years ago, as we and others have found such tissues to be amenable to MS-based proteomics, a key advantage over transcriptomics[16,29].

Melanoma heterogeneity is driven by distinct tumor cell subpopulations and interactions with their tumor microenvironment (TME or tumor stroma), influencing disease progression, treatment response and patient survival[30]. We therefore asked if DVP could identify disease-relevant proteome signatures by comparing cancer cells from the inner tumor mass to those that were in close proximity to the stroma (Fig. 6A). To profile only melanoma cells, we isolated cells double-positive for the melanoma markers SOX10 and CD146 (Fig. 6A). Proteomes from the central tumor (central) and tumor-stroma border (peripheral) region were distinct in unsupervised hierarchical clustering and PCA (Fig. 6B). Peripheral cells showed strong BRAF, integrin and immune system related signatures, whereas in central cells up-regulated proteins with functions in DNA replication, regulation of p53 activity and mRNA splicing (FDR < 0.05, Fig. 6C, D). Prognostically relevant genes for subcutaneous melanoma have previously been reported by transcriptomics but without spatial context[31,32]. When we interrogated these markers in our data sets, we found that the outer region had significantly up-regulated favorable prognostic proteins for immune related processes (e.g. HLA-B, TAPBP, p-value < 0.01). In contrast, unfavorable proteins showed the inverse trend with higher levels in the central region (e.g. MCM3/6, CDK11A/B, DHX9) (Fig 6E, p-value < 10^{-5}). Furthermore, our differential analysis highlighted 'signaling by GCPR', 'extracellular matrix organization' and a number of other relevant proteins associated with these and other cancer-relevant functions. These results show the power of DVP to quantify the spatial variability of the disease-related proteome and suggests the potential to improve molecular disease subtyping to guide clinical decision-making.

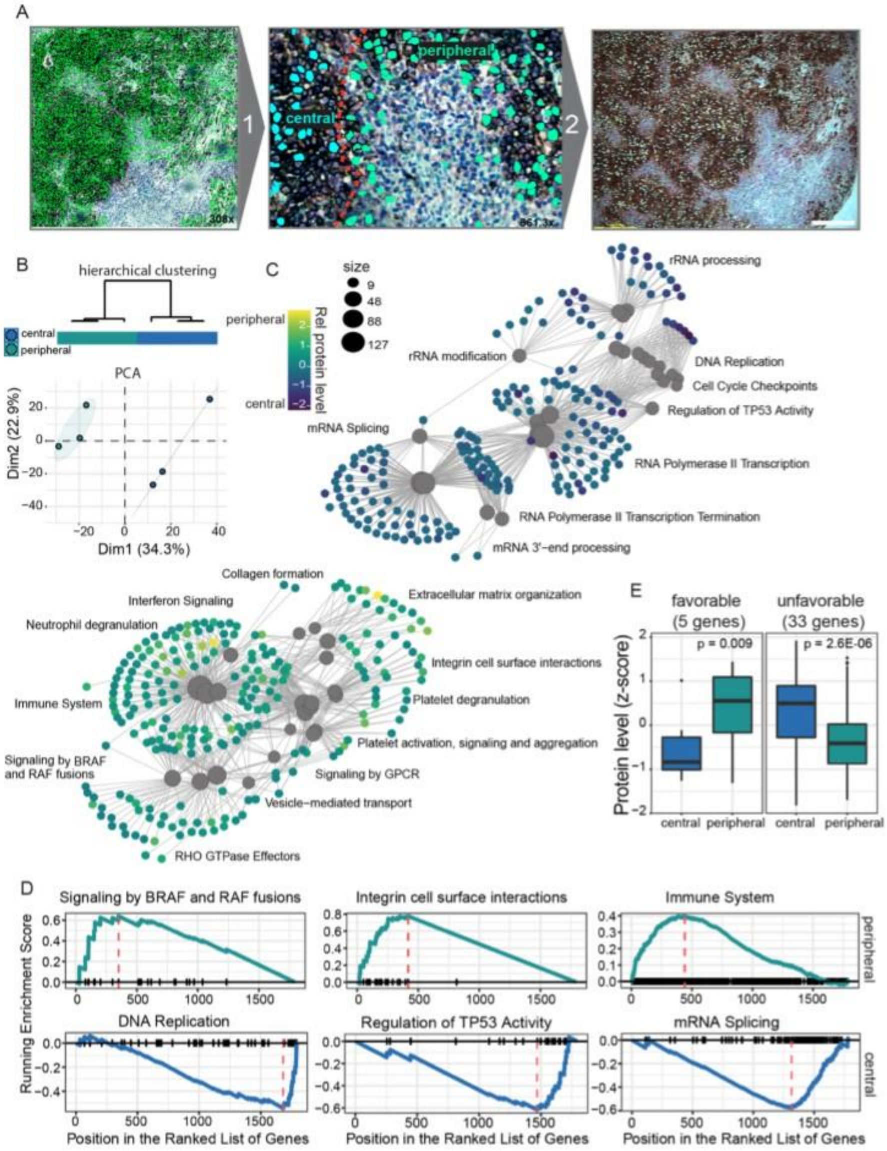


Fig.6: DVP applied to archived melanoma tissue

A. DVP applied to melanoma immunohistochemically stained for the melanoma markers SOX10 and CD146. After AI-guided single cell segmentation (left panel), SOX10 and CD146 double-positive melanoma cells were classified according to proximity to the tumor stroma (middle panel). Cells from the central tumor mass (central, marked in turquoise) and close to the tumor stroma (peripheral, marked in light green) were isolated by automated LMD (right panel) and subjected to MS-based proteomics. **B.** Upper panel: Dendrogram of unsupervised hierarchical clustering of central (n = 3) and peripheral (n = 3) melanoma cells. Lower panel: Principal component of central and peripheral melanoma cell proteomes. **C.** Enrichment analysis of proteins differentially regulated between central and peripheral melanoma cells. Relationship between all significantly enriched GO terms and proteins are shown. Node sizes represent the number of genes in each category. Node colors indicate relative log2 transformed protein levels. **D.** Gene Set Enrichment Analysis (GSEA) plot of significantly enriched pathways for central and peripheral cells. **E.** Relative protein levels comparing central and peripheral cells for favorable (left, p-value < 0.009) and unfavorable (right, p-value < 10^{-5}) genes reported by[31,32].

Outlook

The DVP pipeline combines high resolution microscopy with new developments for image recognition, automated laser microdissection and ultra-sensitive MS-based proteomics in a robust way. Our examples demonstrate a wide range of applications, from cell culture to pathology and in principle, any biological systems that can be microscopically imaged is amenable to DVP.

A single slide can encompass hundreds of thousands of cells or more and many such slides could rapidly be scanned to isolate very rare cell states or interactions between cells. Likewise, DVP should be uniquely suited to study the proteomic composition and post-translational modifications in the extracellular matrix surrounding particular cell constellations. The resolution of excision is limited to the width of a laser beam, which is sufficient to excise individual chromosomes[33], but cells could also be interrogated by super-resolution microscopy or highly-multiplexed imaging to better delineate precise and subtle cell states as part of their classification.

In conclusion, DVP marries increasingly powerful imagining technologies with unbiased proteomics, with a plethora of applications in basic biology and biomedicine. At the conceptual level, this technology integrates cell biology as studied by microscopy with unbiased 'omics' type analyses and in particular MS-based proteomics. Furthermore, the visual information gives specific context that is helpful in interpreting the proteomics data. For the field of oncology, DVP encompasses digital pathology but integrates and extends the information in stainings against a few pre-defined markers to thousands of proteins making up a cellular context.

ACKNOWLEDGMENT

The authors thank Martin Rykær, Jeppe Madsen (NNF CPR Mass Spectrometry Platform, University of Copenhagen) and Lylia Drici (NNF CPR Proteomics Program) as well as Johannes Mueller (MPIB Munich) for technical assistance. We acknowledge Florian Hoffmann, Christoph Greb, and Falk Schlaudraff from Leica for technical support and Ede Migh, Tivadar Danka, and Mária Kovács for fruitful scientific discussions. Thomas Hartig Braunstein, Pablo Hernandez-Varas, and Clara Prats from the Core Facility of Integrated Microscopy for microscopy support. We thank Jiri Lukas for scientific support and guidance.

This work was supported by grants from the Novo Nordisk Foundation (grant agreement NNF14CC0001 and NNF15CC0001) and The Max-Planck Society for the Advancement of Science, and the Chan-Zuckerberg Initiative for partial funding of the cell cycle work (grant CZF2019-002448) to E.L., M.M. and P.H. F.C acknowledges the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie individual fellowship under grant agreement (846795). BDA acknowledges support from the Lundbeck Foundation (R252-2017-1414) and the Novo Nordisk Foundation (NNF20OC0065720). PH, RH, FK and AK acknowledge support from the LENDULET-BIOMAG Grant (2018-342), from the European Regional Development Funds (GINOP-2.3.2-15-2016-00006, GINOP-2.3.2-15-2016-00026, GINOP-2.3.2-15-2016-00037), from the H2020 (ERAPERMED-COMPASS, DiscovAIR).

We acknowledge Dr. Sayuri Ito and Dr. Hisao Masai (Tokyo Metropolitan Institute of Medical Science) for providing the stable U-2 OS FUCCI cell line. LCK-GFP plasmid was a gift from Steven Green (Addgene plasmid #61099).

AUTHOR CONTRIBUTIONS

Conceptualization, A.M. F.C., P.H. and M.M.; Methodology, A.M., F.C., A.D.B, M.B., B.D.A, M.M.; Software, R.H., F.K., A. K. and P.H.; Investigation, A.M., F.C., R.H.,; Formal Analysis, A.M., F.C., and R.H.; Writing - Original Draft, A.M., F.C., P.H. and M.M.; Writing - Review & Editing, all authors; Resources, all authors.; Data Curation, L.M.R.G., M.B., S.N., A.M., F.C., R.H., F.K., A.K., P.H.; Visualization, A.M., F.C., and R.H.; Project Administration, A.M., and P.H.; Supervision, M.M.; Funding Acquisition, F.C., P.H., E.L., and M.M.

COMPETING INTERESTS

P.H. is the founder and a shareholder of Single-cell technologies Ltd., a biodata analysis company that owns and develops the BIAS software.

REFERENCES

1. Rajewsky, N. *et al.* LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386 (2020).
2. Method of the Year 2019: Single-cell multimodal omics. *Nature methods* vol. 17 1 (2020).
3. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
4. Zhu, C. *Single-cell multimodal omics: the power of many*. www.nature.com/naturemethods doi:10.1038/s41592-019-0691-5.
5. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
6. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 9–12 (2020).
7. Brunner, A.-D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *bioRxiv* 2020.12.22.423933 (2020) doi:10.1101/2020.12.22.423933.
8. Hériché, J.-K., Alexander, S. & Ellenberg, J. Integrating Imaging and Omics: Computational Methods and Challenges. *Annu. Rev. Biomed. Data Sci.* **2**, 175–197 (2019).
9. Hollandi, R. *et al.* nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Syst.* **10**, 453-458.e6 (2020).
10. Smith, K. & Horvath, P. Active learning strategies for phenotypic profiling of high-content screens. *J. Biomol. Screen.* **19**, 685–695 (2014).
11. Caicedo, J. C. *et al.* Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods* **16**, 1247–1253 (2019).
12. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2020).
13. Carpenter, A. E. *et al.* CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7, R100 (2006).
14. Sakaue-Sawano, A. *et al.* Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression. *Cell* **132**, 487–498 (2008).
15. Altelaar, A. F. M. & Heck, A. J. R. Trends in ultrasensitive proteomics. (2012)

doi:10.1016/j.cbpa.2011.12.011.

16. Coscia, F. *et al.* A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *J. Pathol.* **251**, 100–112 (2020).

17. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).

18. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology* vol. 20 285–302 (2019).

19. Mahdessian, D. *et al.* Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *bioRxiv* 543231 (2020) doi:10.1101/543231.

20. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, 1260419–1260419 (2015).

21. Venturini, V. *et al.* The nucleus measures shape changes for cellular proprioception to control dynamic cell behavior. *Science (80-.).* **370**, (2020).

22. Arias-Garcia, M., Rickman, R., Sero, J., Yuan, Y. & Bakal, C. The cell-cell adhesion protein JAM3 determines nuclear deformability by regulating microtubule organization. *bioRxiv* 689737 (2020) doi:10.1101/689737.

23. Kokkat, T. J., Patel, M. S., McGarvey, D., Livolsi, V. A. & Baloch, Z. W. Archived formalin-fixed paraffin-embedded (FFPE) blocks: A valuable underexploited resource for extraction of DNA, RNA, and protein. *Biopreserv. Biobank.* **11**, 101–106 (2013).

24. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, e253–e261 (2019).

25. Zhu, S., Schuerch, C. & Hunt, J. Review and updates of immunohistochemistry in selected salivary gland and head and neck tumors. *Arch. Pathol. Lab. Med.* **139**, 55–66 (2015).

26. Kim, L. C., Song, L. & Haura, E. B. Src kinases as therapeutic targets for cancer. *Nature Reviews Clinical Oncology* vol. 6 587–595 (2009).

27. Gershenwald, J. E. *et al.* Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA. Cancer J. Clin.* **67**, 472–492 (2017).

28. Balch, C. M. *et al.* Final version of 2009 AJCC melanoma staging and classification. *J. Clin. Oncol.* **27**, 6199–6206 (2009).

29. Piehowski, P. D. *et al.* Residual tissue repositories as a resource for population-based

cancer proteomic studies. *Clin. Proteomics* **15**, 1–12 (2018).

30. Hanahan, D. & Weinberg, R. a. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

31. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

32. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science (80-.).* **357**, (2017).

33. Yang, H. yan, Wang, J. & Kang, X. yang. Meiotic chromosome preparation techniques of pollen mother cells for laser micro-dissection in Populus spp. *For. Stud. China* **12**, 74–78 (2010).

34. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

SUPPLEMENTARY FIGURES

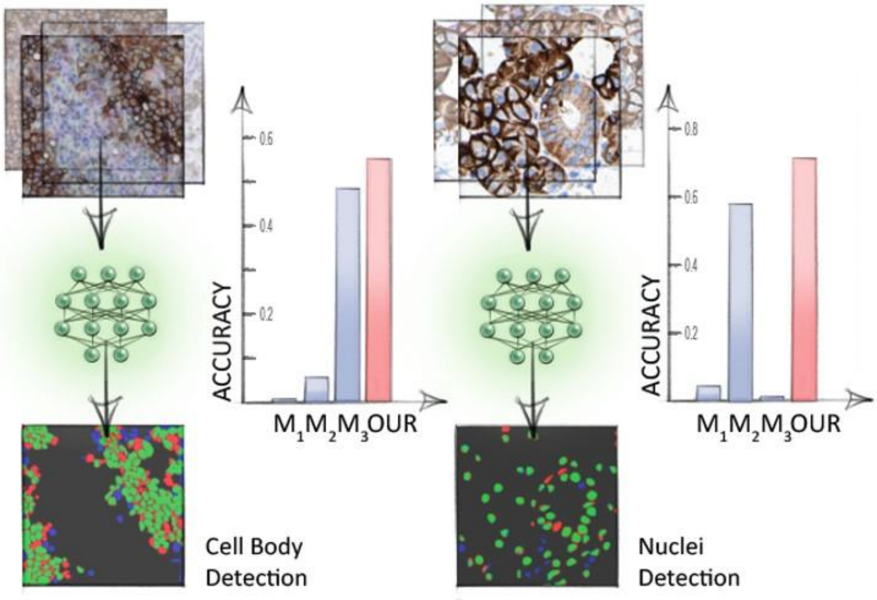


Fig. S1: Benchmarking of segmentation algorithm

Cell body and nuclei segmentation of melanoma (left) and salivary gland tissue (right) using the Biological Image Analysis Software (BIAS). We benchmarked the accuracy of our segmentation approach using the F1 metric and compared results to three additional methods M1-M3. unet4nuclei (M_1)[11], conventional adaptive threshold- and object splitting-based application (M2)[13], CellPose (M3)[12]. Visual representation of the segmentation results: green areas correspond to true positive, blue to false positive and red to false negative. Data provided in Table S1.

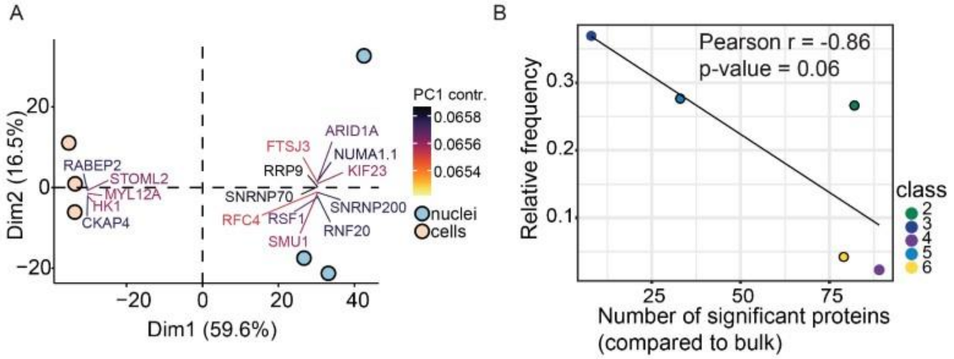


Fig. S2: PCA and loadings of cell culture classes at sub-cellular level and number of significantly changed proteins vs. class abundance

A. Principal component analysis (PCA) of whole cell (n = 3) and nuclei proteomes (n = 3) based on 1,993 quantified protein groups after data filtering for no missing values. Proteins with the strongest contribution to PC1 are highlighted. **B.** Correlation between number of significantly regulated proteins per nuclei class vs relative class proportion. A linear model was fitted to the data showing an inverse correlation with Pearson r = -0.86 (p-value = 0.06).

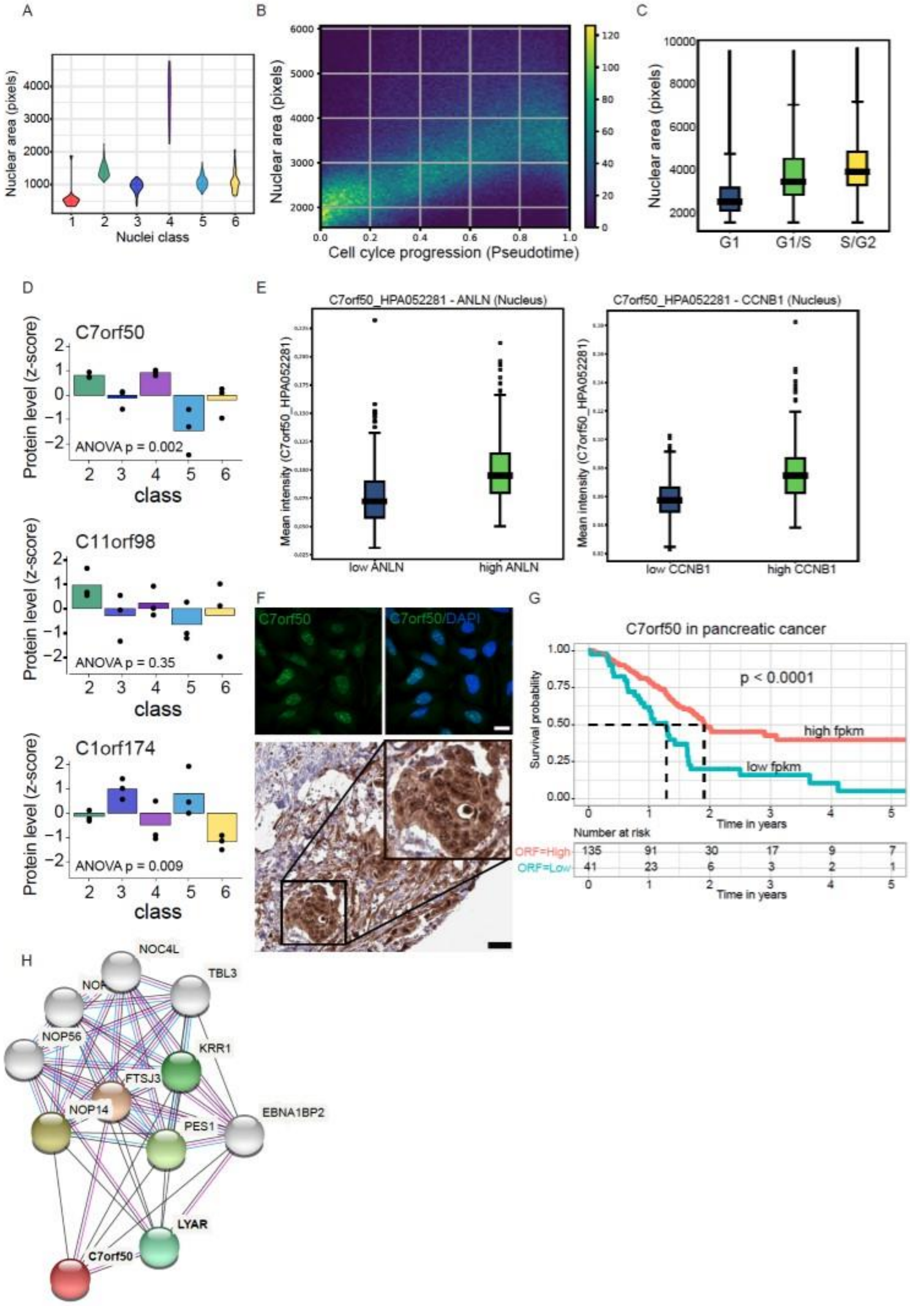

A
Nuclear area (pixels)
1000
2000
3000
4000
1 2 3 4 5 6
Nuclei class
B
Nuclear area (pixels)
2000
3000
4000
5000
6000
0.0 0.2 0.4 0.6 0.8 1.0
Cell cylce progression (Pseudotime)
0 20 40 60 80 100 120
C
Nuclear area (pixels)
2000 4000 6000 8000 10000
G1
G1/S
S/G2
D
C7orf50
Protein level (z-score)
ANOVA p = 0.002
2 3 4 5 6
class
C11orf98
ANOVA p = 0.35
C1orf174
ANOVA p = 0.009
E
C7orf50_HPA052281 - ANLN (Nucleus)
Mean intensity (C7orf50_HPA052281)
low ANLN
high ANLN
C7orf50_HPA052281 - CCNB1 (Nucleus)
low CCNB1
high CCNB1
F
C7orf50
C7orf50/DAPI
G
C7orf50 in pancreatic cancer
Survival probability
1.00 0.75 0.50 0.25 0.00
p < 0.0001
high fpkm
low fpkm
0 1 2 3 4 5
Time in years
Number at risk
ORF=High 135 91 30 17 9 7
ORF=Low 41 23 6 3 2 1
H
NOC4L
NOP
TBL3
NOP56
KRR1
FTSJ3
NOP14
PES1
EBNA1BP2
LYAR
C7orf50

Fig. S3: DVP discovers uncharacterized proteins with potential clinical relevance

A. Violin plots showing nuclear area in pixels of the 6 nuclei classes identified by ML. **B.** Nuclear area in pixels of U2OS FUCCI cells in relation to the cell cycle pseudotime[19]. Color code indicates point density. **C.** Nuclear area of three major cell cycle states G1, G1/S and S/G2 determined by fluorescently tagged CDT1 and GMNN intensities and Gaussian clustering. **D.** Relative protein levels of the three ORF proteins C7orf50, C11orf98 and C1orf174 across the 5 nuclei classes. C7orf50 and C1orf174 were significantly differentially regulated ($p < 0.01$). **E.** Mean intensities of immunofluorescently stained C7orf50 and the cell cycle markers ANLN and CCNB1 in U20S cells. C7orf50 levels were quantified in nuclei with low and high ANLN and CNNB1 intensities. **F.** Upper panel: Representative immunofluorescence images of C7orf50 and DNA (DAPI) stained U2OS cells[20,32]. Scale bar is 20 µm. Note, C7orf50 is enriched in nucleoli. Lower panel: Immunohistochemistry of a C7orf50 stained pancreatic adenocarcinoma (https://www.proteinatlas.org/ENSG00000146540-C7orf50/pathology/pancreatic+cancer#img). Image credit: Human Protein Atlas. Scale bar is 40µm. **G.** Kaplan-Meier survival analysis of pancreatic adenocarcinoma (https://www.proteinatlas.org/ENSG00000146540-C7orf50/pathology/pancreatic+cancer) based on relative C7orf50 RNA levels (FPKM, number of Fragments Per Kilobase of exon per Million reads)[32]. RNA-seq data is reported as median FPKM, generated by The Cancer Genome Atlas (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Patients were divided into two groups based on C7orf50 levels with n=41 low and n=135 high patients. A log-rank test was calculated with $p = 0.0001$. **H.** String interactome analysis for C7orf50. A high confidence score of 0.7 was used with the five closest interactors highlighted by color[34].

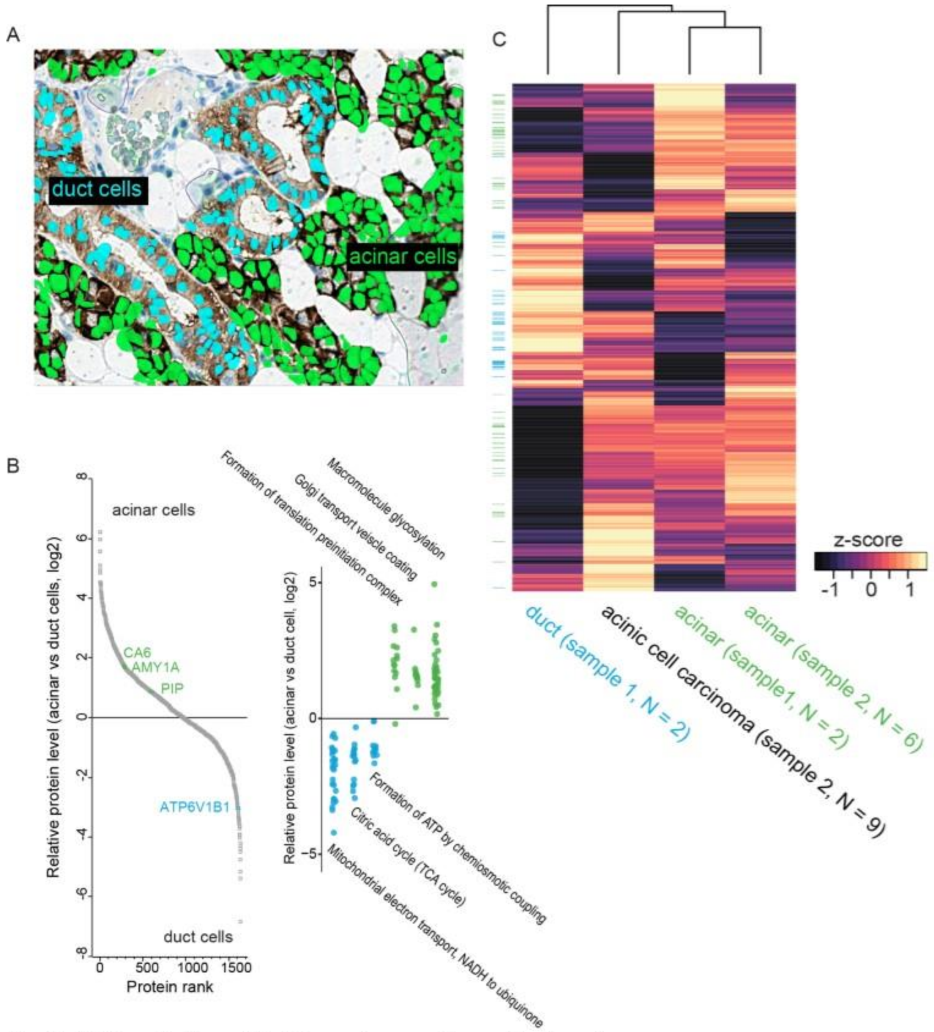


Fig. S4: DVP applied to archival tissue of a rare salivary gland carcinoma

A. Immunohistochemical staining of normal salivary gland stained for the cell adhesion protein EpCAM. Supervised (random forest) ML was trained to identify acinar (green) and duct cells (turquoise). **B.** Left panel: Quantitative proteomic comparison between acinar and duct cells from tissue in A with known cell type specific markers highlighted (https://www.proteinatlas.org/humanproteome/tissue/salivary+gland). Right panel: Relative protein levels of selected pathways that were significantly higher in acinar or duct cells. **C.** Unsupervised hierarchical clustering of acinar and duct cell proteomes from two different patients together with acinar cell carcinoma cells. Note that normal acinar cells of two different tissues clustered together. Duct cells clustered furthest away. Prior to clustering, protein levels from different sample groups (duct cell tissue #1, acinar cell tissue #1, acinar cell tissue #2, carcinoma tissue #2) were averaged and z-scored. Bar on the left shows differentially expressed pathways from panel B with acini and duct specific proteins in green and turquoise, respectively.

MATERIAL & METHODS

Patient samples and ethics

We collected archival FFPE tissue samples of salivary gland acinic cell carcinoma and melanoma from the Department of Pathology, Zealand University Hospital, Roskilde, Denmark. The study was carried out in accordance with the institutional guidelines under approval by the local Medical Ethics Review Committee (SJ-742), the Data Protection Agency (REG-066-2019) and in agreement with Danish law (Medical Research Involving Human Subjects Act).

In accordance with the Medical Ethics Review Committee approval, all FFPE human patient tissue samples were exempted from consent as these studies used existing archived pathological specimens and not human subjects directly. Human tissue specimens were assessed by a board-certified pathologist.

Cell lines

The human osteosarcoma cell line U2OS was grown in Dulbecco's modified Eagle's medium (high glucose, GlutaMAX) containing 10% FBS and penicillin-streptomycin (Thermo Fisher Scientific).

The U2OS FUCCI (Fluorescent Ubiquitination-based Cell Cycle Indicator) cells were kindly provided by Dr. Miyawaki[1]. These cells are endogenously tagged with two fluorescent proteins fused to the cell cycle regulators CDT1 (mKO2-hCdt1+) and Geminin (mAG-hGem+). CDT1 accumulates during the G1 phase while Geminin accumulates in S and G2 phases allowing cell cycle monitoring. The cells were cultivated at 37 °C in a 5.0 % CO2 humidified environment in McCoy's 5A (modified) medium GlutaMAX supplement, (Thermo Fisher, 36600021, MA, USA) supplemented with 10% fetal bovine serum (FBS, VWR, Radnor, PA, USA) without antibiotics.

U2OS cells stably expressing a membrane-targeted form of eGFP were generated by transfection with plasmid Lck-GFP (Addgene #61099[2]) and culturing in selection medium (DMEM medium containing 10% FBS, Penicillin-Streptomycin, 400 µg/ml Geneticin) under conditions of limited dilution to yield single colonies. A clonal cell line with homogenous and moderate expression levels of Lck-eGFP at the plasma membrane was established from a single colony.

All cell lines were tested for mycoplasma (MycoAlert, Lonza) and authenticated by STR profiling (IdentiCell Molecular Diagnostics).

Tissue preparation for immunohistochemistry

Immunohistochemical staining on membrane slides:

Membrane PEN slides 1.0 (Zeiss, Göttingen, Germany, cat. #415190-9041-000) were treated with UV light for 1 h and coated with APES (3-aminopropyltriethoxysilane) using Vectabond reagent (Vector Laboratories, Burlingame, CA, USA, cat. #SP-1800-7) according to the manufacturer's protocol. FFPE tissue sections were cut (2.5 µm), air-dried at 37 ºC overnight and heated at 60 ºC for 20 min to facilitate better tissue adhesion. Next, sections were deparaffinized, rehydratrated and loaded wet on the fully automated instrument Omnis (Dako, Glostrup, Denmark). Antigen retrieval was conducted using Target Retrieval Solution pH 9 (Dako, cat # S2367) diluted 1:10 and heated for 60 minutes at 90 °C. Single stain for EpCAM (Nordic Biosite, Copenhagen K, Denmark, clone BS14, cat. #BSH-7402-1, dilution 1:400) and sequential double stain for SOX10/CD146 (SOX10; Nordic Biosite, clone BS7 cat. #BSH-7959-1, dilution 1:200 and CD146; Cell Marque, Rocklin, CA, USA, clone EP54, cat. #AC-0052, dilution 1:400) was performed and slides were incubated for 30 min (32 °C). After washing and blocking of endogenous peroxidase activity, the reactions were detected and visualized using Envision FLEX+ High pH kit (Dako, cat #GV800+GV809/GV821) according to manufacturer's instructions. In the double stain, Envision DAB+ (Dako, cat # GV825) and Envision Magenta (Dako, cat. #GV900) Substrate Chromogen Systems was used for visualization of CD146 and SOX10, respectively. Finally, slides were rinsed in water, counterstained with Mayer's hematoxylin, and air-dried without mounting.

Immunofluorescence staining

Cells were first incubated with 5-ethynyl-2 deoxyuridine (EdU) for 20 min, then fixed for 5 min at room temperature with 4 % paraformaldehyde and washed three times with PBS. Cells were then permeabilized with PBS/0.2% Triton-X for 2 min on ice and washed three times with PBS. Cells were then stained with an EdU labeling kit (Life Technologies) and counterstained with Hoechst 33342 for 10 min. Slides were mounted with GB mount (GBI Labs # E01-18).

96-well glass bottom plates (Greiner Sensoplate Plus, Greiner Bio-One, Germany) were coated with 12.5 µg/ml human fibronectin (Sigma Aldrich, Darmstadt, Germany) for 1 h at RT. Immunocytochemistry was carried out following an established protocol[3]. 8,000 U2OS cells were seeded in each well and incubated in a 37 °C and 5% CO_2 environment for 24 hours. Cells were washed with PBS, fixed with 40 µl 4% ice-cold PFA and permeabilized with 40 µl 0.1 Triton X-100 in PBS for 3x5 min. Rabbit polyclonal HPA antibodies targeting the proteins of interest were diluted in blocking buffer (PBS + 4% FBS) at 2-4 µg/ml along with marker primary antibodies (see just below) and incubated overnight at 4°C. Cells were washed with PBS for 4x10 min and incubated with secondary antibodies (goat anti-rabbit Alexa488 (A11034, Thermo Fisher), goat anti-mouse Alexa555 (A21424, Thermo Fisher), goat anti-chicken Alexa647 (A21449, Thermo Fisher)) in blocking buffer at 1.25 µg/ml for 90 min at RT. Cells were counterstained in 0.05 µg/ml DAPI for 15 min, washed with for 4x10min and mounted in PBS.

Primary antibodies used:
For C7orf50 cell cycle validation:
mouse anti ANLN at 1.25 ug/ml (amab90662, Atlas Antibodies)
Mouse anti CCNB1 at 1 ug/ml: (610220, BD Biosciences)

High-resolution microscopy

Images of immunofluorescence-labeled cell cultures were acquired using an AxioImager Z.2 microscope (Zeiss, Germany), equipped with widefield optics, a 20×, 0.8-NA dry objective, a quadruple-band filter set for Hoechst, FITC, Cy3 and Cy5 fluorescent dyes. Widefield acquisition was performed using the Colibri7 LED light source and an AxioCam 702 mono camera with 5,9 mm/px. Z-stacks with 19 z-slices were acquired at 3 mm increments to capture the optimal focus plane. Images were obtained automatically with the Zeiss Zen 2.6 (blue edition) at non-saturating conditions (12-bit dynamic range).

IHC images from salivary gland and melanoma tissue were obtained using the automated slide scanner Zeiss Axio Scan.Z1 (Zeiss, Germany) for brightfield microscopy. Brightfield acquisition was obtained using the VIS LED light source and a CCD Hitachi HV-F202CLS camera. PEN slides were scanned with a 20×, 0.8-NA dry objective yielding a resolution of 0,22 mm/px. Z-stacks with 8 z-slices were acquired at 2 mm increments to capture the optimal focus plane. Color images were obtained automatically with the Zeiss Zen 2.6 (blue edition) at non-saturating conditions (12-bit dynamic range).

Widefield fluorescence microscopy for validation of cell cycle dependent C7orf50 expression

Cells were imaged on a Leica Dmi8 widefield microscope equipped with a 0.80 N/A 40x air objective and a Hamamatsu Flash 4.0 V3 camera using the LAS X software. The segmentation of each cell was performed using the Cell Profiler software[4] using DAPI for nuclei segmentation. The mean intensity of the target protein and the cell cycle marker protein was measured in the nucleus. The cells were grouped into the G1 and G2 phases of the cell cycle by using the 0.2 and 0.8 quantile of ANLN or CCNB1 intensity levels in the nucleus and cell cycle dependent expression of C7orf50 was validated by comparing differences in expression levels between G1 and G2 cells.

Laser microdissection

To excise cells or nuclei we used the Leica LMD7 system, which we had adapted for automated single cell automation. High cutting precision was achieved using a HC PL FLUOTAR L 63x/0.70 (tissue) or 40x/0.60 (cells cultures) CORR XT objective. We used the Leica Laser Microdissection V 8.2.3.7603 software (adapted for this project) for full automated excision and collection of more than 700 contours per hour.

Leica LMD 7 cutting accuracy (Leica R&D, patent EP1276586)

For 150x objective: $\frac{10}{150} = 0.07\mu\text{m}$

Biological Image Analysis Software (BIAS) introduction

A typical image analysis workflow in BIAS consists of multiple images processing steps, including image preprocessing, object segmentation, contour post-processing, feature extraction and statistical analysis, supervised or unsupervised machine learning methods for phenotype classification, cell selection and cell extraction (by a selected or supported micro-dissection microscope). For each workflow step BIAS provides conveniently customizable modules.

Images were captured with a Zeiss Axio Scan.Z1 or AxioImager Z.2 microscope, both are supported by the analysis software with preservation of correct spatial information (spatial topology and size). Other types of microscopes with similar support include 3DHistech, Hamamatsu, GE IN Cell, Molecular Devices ImageXpress Micro, Leica SP, Perkin Elmer Opera and Operetta. It is also possible to import standard image files with editable image orientation and resolution. Illumination-correction algorithm, primarily CIDRE[5] (within BIAS) were applied where it was necessary to solve the frequent 'vignetting' effect observable in raw microscopy images.

Image preprocessing was followed by deep learning-based nucleus and cell segmentation modules (see segmentation methods and accuracy evaluation) further refined by unary and binary morphological operators (e.g.: dilation, erosion, cavity filling, addition and subtraction). For example, subtraction can be used to calculate the cytoplasm-only region from cell and nuclei masks.

Results of different segmentation algorithms may be connected by a linking module to form complex structures, e.g. an abstract cell object might be constructed from a segmented nucleus, cytoplasm and proteins where each component can be analyzed individually or as a whole. Objects were forwarded to the feature extraction modules, configurable to extract properties from the selected image channels and cell components. A multitude of features can be retrieved from the image and contour data, such as shape (e.g.: area, perimeter, form factor, solidity etc.), intensity (e.g.: min, max, mean, total) or texture (e.g.: Haralick features) and represented in a feature matrix[6]. Features to be extracted may vary by experiments according to their specific requirements each, potentially containing up to hundreds or a few thousands per cell depending on the configuration. Features from the neighboring regions of each cell can be incorporated as well, to further improve accuracy where local neighborhoods might also contain valuable information for the cell phenotypes (such as in tissues)[7]. The resulting feature matrices can be analyzed internally or exported to a 3rd party tool[8]. Subsequently reimport of extended feature matrices into BIAS is also possible to extend the statistical capabilities or simply to visualize the data in the plate overview.

Internal analysis tools include simple, value-based statistics, manual gating, automatic feature space clustering and interactive supervised machine learning; additionally, these may also be combined. Final cell selection can depend on simple, value-based statistics or complex queries searching in multiple feature and classification matrices.

With manual gating, two features can be represented in a two-dimensional coordinate system and with cluster centers defined manually, samples are displayed at their actual position in the coordinate system. K-Means clustering can automatically find a fixed number of cluster centers in the feature space with an arbitrary number of dimensions.

During supervised machine learning, the biologist defines the phenotypes of interest and provides training samples (usually around a hundred samples for each class). Training is iterative and interactive, refinable and new phenotypes may be identified using an active learning technique[9]. A cross validation tool is provided to continuously monitor accuracy, so that when a satisfactory threshold is reached, all other cells in the whole experiment are classified. Different machine learning approaches are implemented in the BIAS software for various experimental needs. Such methods are 1) Multilayer perceptron (a feedforward artificial neural network), 2) Support-vector machine (separates the feature space by hyperplanes between the training samples), 3) Random forest (a number of decision trees trained to separate the training data into classes) and 4) Logistic regression (a statistical model that determines the probability of passing or failing the criteria of a certain class). These classification algorithms can analyze and classify tens of thousands of cells in a matter of seconds.

Results of the feature extraction and classification phases can be summarized in the statistics module, in addition it also provides an interface where custom queries can be written in SQL language and executed on the cell information database containing feature and classification data for all cells in the experiment. Cross queries between different classification and feature matrices are also supported. Query templates and wizards are provided for the most common questions.

The queries might be as simple as e.g. listing N items that have the highest value in a selected feature, or rather complex e.g. to calculate the sum and ratio of the areas of cells belonging to different classes). Results can be represented in graphs, heatmaps or used to filter cells for capturing by a suitable microscope.

The visualization tool supports a virtually unlimited number of channels with adjustable intensity window, gamma and look-up-table settings. An interactive, zoomable overview of the whole experiment (let it be a slide or a plate) can be displayed, reflecting the changes in visualization or data processing real-time. The results of all processing steps could be overlaid on it, such as segmentation masks, feature heatmaps or phenotype classification as color-coded segments, etc. A schematic display of the multi-well plate helps the navigation, enabling manual selection of cross-field areas for isolation.

The isolation and collection module uses a registration algorithm based on a marker or point-of-interest (POI) to connect the coordinate systems of the source and the isolation microscopes as well as to transfer the contour points to that of the isolation microscope. The tool supports sorting cells into different collectors and also cutting components in order (e.g. cutting nuclei into a microplate well or collection cap first then remaining cytoplasm into another). Cells of interest can be selected manually or using the statistics module. To preserve object integrity, it allows the user to define cutting offsets or exclude touching regions thereby preventing undesirable laser-induced damage.

Segmentation methods and accuracy evaluation

NucleAIzer[10] models were integrated into BIAS and customized for these experiments by retraining and refining the nucleus and cytoplasm segmentation models. Firstly, style transfer[11] learning was performed as follows. Given a new experimental scenario such as our melanoma or salivary gland tissue sections stained immunohistochemically, the acquisition of which produces such an image type no annotated training data exists for, preventing efficient segmentation with even powerful deep learning methods. With an initial segmentation or manual contouring by experts (referred to as annotation) a small mask dataset is acquired (masks represent e.g. nuclei) which is used to generate new mask images such that the spatial distribution, density and morphological properties of the generated objects (e.g. nuclei) are similar to those of the annotated images. The initial masks and their corresponding microscopy images are used to train a style transfer model that learns how to generate the texture of the microscopy images on the masks marking objects: foreground to mimic e.g. nuclei and background for surrounding e.g. tissue structures. Parallelly, artificial masks of either nucleus or cytoplasm objects were created and input to the style transfer network that generated realistic-looking synthetic microscopy images with the visual appearance of the original experiment. Hence, with this artificially created training data (synthetic microscopy images and their corresponding, also synthetic masks) their applied deep learning segmentation model, Mask R-CNN, is prepared for the new image type and can accurately segment the target compartments.

We benchmarked the accuracy of the segmentation approach on a fluorescent LCK-U2OS cell line as well as tissue samples of melanoma and salivary gland, and compared results to three additional methods, including two deep learning approaches: unet4nuclei (denoted as M_1 on Fig. 2A and S1)[12] and CellPose (M3)[13], alongside a widely-used, conventional adaptive threshold- and object splitting-based application (M2)[4]. We note that M_1 is not intended for cytoplasm segmentation (see details in[12] and below). Segmentation accuracy according to the F1 metric is displayed as bar plots (Fig. 2A, S1), while visual representation in a color-coded manner is also provided.

unet4nuclei[12] is optimized to segment nuclei on cell culture images, and 2) CellPose[13] is an approach intended for either nucleus or cytoplasm segmentation on various microscopy image types, while CellProfiler[4] is a conventional threshold- and object splitting-based software broadly used in the bioimage analysis community. unet4nuclei as its name suggests is primarily intended for nucleus segmentation and uses a U-Net-based network after pre-processing of input images, then post-processes detected objects. Cellpose uses a vector flow-representation of instances and its neural network (also based on U-Net) predicts and combines horizontal and vertical flows. unet4nuclei has successfully been applied in nucleus segmentation of cell cultures, while CellPose is able to generalize well on various image modalities even outside microscopy and can be used to segment nuclei and cytoplasms.

We evaluated our segmentation performance (and comparisons) according to the F1-score metric calculated at 0.7 IoU (intersection over union) threshold. IoU also known as Jaccard index was calculated from the overlapping region of the predicted (segmented) object with its corresponding ground truth (real) object at a given threshold (see formulation below). True positive (TP), false positive (FP) and false negative (FN) objects were counted accordingly, if they had IoU greater than the threshold t (in our case 0.7), to yield the F1-score at this threshold (see formulation below). Considering the mean F1-scores measured, we conclude that the applied deep learning-based segmentation method[10] available in BIAS produced segmentations on both nucleus and cytoplasm level in a higher quality than the compared methods; see results on Fig. 2A and S1.

$$Jaccard\ index = \frac{|x \cap y|}{|x \cup y|} = \frac{|x \cap y|}{|x| + |y| - |x \cap y|}$$

$$precision\ (t) = \frac{TP(t)}{TP(t) + FP(t)}$$

$$recall\ (t) = \frac{TP(t)}{TP(t) + FN(t)}$$

$$F1\ score\ (t) = 2 \cdot \frac{precision\ (t) \cdot recall\ (t)}{precision\ (t) + recall\ (t)}$$

Our evaluation results of nucleus and cell body segmentation on melanoma-, salivary gland tissues and U2OS cells is presented in (Supplementary Table 1).

sample	method			
	M1	M2	M3	OUR
U2OS cyto	0.0667*	0.5994	0.7205	**0.7336**
Melanoma nuc	0.1126	0.4386	0.1801	**0.5498**
Melanoma cyto	0.0058*	0.0549	0.4859	**0.5536**
Salivary gland nuc	0.0476	0.5830	0.0160	**0.7158**
Salivary gland cyto	0.0991*	0.0703	0.2312	**0.4111**

Supplementary Table 1. F1-scores of the compared segmentation methods on our samples. The methods are as follows: M1 is unet4nuclei[12], M2 is CellProfiler[4], M3 is Cellpose[13], while OUR refers to nucleAIzer[10] (implemented in BIAS). High scores are highlighted in bold. Asterisks mark that M1 is intended for nucleus segmentation but was applied to segment cytoplasm.

Sample preparation for mass spectrometry

Cell culture (nuclei or whole cells) and tissue samples were collected by automated laser microdissection into 384-well plates (Eppendorf 0030129547). For the collection of different U2OS nuclei classes (Fig. 3 and 4), we normalized nuclear size differences (resulting in different total protein amounts) by the number of collected objects per class. On average, we collected 267 nuclei per sample. For FFPE tissue samples of salivary gland and melanoma, (2.5 µm thick section cut in microtome) an area of 80,000 – 160,000 µm2 per sample was collected, an estimated number of 100-200 cells based on the average HeLa cell volume of 2,000 µm3 (BNID 100434).

20µl of ammonium bicarbonate (ABC) were added to each sample well and the plate closed with sealing tape (Corning, CLS6569-100EA). Following vortexing for 10 s, plates were centrifuged for 10 min at 2000g and heated at 95C for 30 min (cell culture) or 60 min (tissue) in a thermal cycler (Biorad S1000 with 384-well reaction module) at a constant lid temperature of 110 °C. 5 µl 5x digestion buffer (60% acetonitrile in 100 mM ABC) was added and samples heated at 75 °C for another 30 min. Samples were shortly cooled down and 1 µl LysC added (pre-diluted in ultra-pure water to 4 ng/µl) and digested for 4 h at 37 °C in the thermal cycler. Subsequently, 1.5 µl trypsin was added (pre-diluted in ultra-pure water to 4ng/µl) and incubated overnight at 37 °C in the thermal cycler. Next day, digestion was stopped by adding trifluoroacetic acid (TFA, final concentration 1% v/v) and samples vacuum-dried (approx. 1.5 h at 60 °C). 4 µl MS loading buffer (3% acetonitrile in 0.2% TFA) was added, the plate vortexed for 10s and centrifuged for 5 min at 2000g. Samples were stored at -20 °C until LC-MS analysis.

High-pH reversed-phase fractionation

We used high-pH reversed-phase fractionation to generate a deep U2OS cell precursor library for data-independent (DIA) MS analysis (below). Peptides were fractionated at pH 10 with the spider-fractionator[14]. 30 µg of purified peptides were separated on a 30 cm C18 column in 100 min and concatenated into 12 fractions with 90 s exit valve switches. Peptide fractions were vacuum-dried and reconstituted in MS loading buffer for LC-MS analysis.

LC-MS analysis

Liquid chromatography mass spectrometry (LC-MS) analysis was performed with an EASY-nLC-1200 system (Thermo Fisher Scientific) connected to a modified trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer with about five-fold higher ion current[15] (timsTOF Pro, Bruker Daltonik GmbH, Germany) with a nano-electrospray ion source (Captive spray, Bruker Daltonik GmbH). The autosampler was configured for sample pick-up from 384-well plates.

Peptides were loaded on a 50 cm in-house packed HPLC-column (75µm inner diameter packed with 1.9µm ReproSilPur C18-AQ silica beads, Dr. Maisch GmbH, Germany).

Peptides were separated using a linear gradient from 5-30% buffer B (0.1% formic acid, 80% ACN in LC-MS grade H2O) in 55 min followed by an increase to 60% for 5 min and 10 min wash at 95% buffer B at 300nl/min. Buffer A consisted of 0.1% formic acid in LC-MS grade H2O. The total gradient length was 70 min. We used an in-house made column oven to keep the column temperature constant at 60 °C.

Mass spectrometric analysis was performed essentially as described in Brunner et al.[15], either in data-dependent (ddaPASEF) (Fig. 5 and 6) or data-independent (diaPASEF) mode (Fig. 3 and 4). For ddaPASEF, 1 MS1 survey TIMS-MS and 10 PASEF MS/MS scans were acquired per acquisition cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to 100 ms each and we analyzed the ion mobility range from 1/K0 = 1.6 Vs cm-2 to 0.6 Vs cm-2. Precursor ions for MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 in a total m/z range of 100-1.700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at 1/K0 = 1.6 VS cm-2 to 20 eV at 1/K0 = 0.6 Vs cm-2. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonik GmbH). Precursors for MS/MS were picked at an intensity threshold of 1.000 arbitrary units (a.u.) and resequenced until reaching a 'target value' of 20.000 a.u taking into account a dynamic exclusion of 40 s elution. For DIA analysis, we made use of the correlation of Ion Mobility (IM) with m/z and synchronized the elution of precursors from each IM scan with the quadrupole isolation window. The collision energy was ramped linearly as a function of the IM from 59 eV at 1/K0 = 1.6 Vs cm−2 to 20 eV at 1/K0 = 0.6 Vs cm−2. We used the ddaPASEF method for library generation[16].

Data analysis of proteomic raw files

Mass spectrometric raw files acquired in ddaPASEF mode (Fig. 5 and 6) were analyzed with MaxQuant (version 1.6.7.0)[17,18]. The Uniprot database (2019 release, UP000005640_9606) was searched with a peptide spectral match (PSM) and protein level FDR of 1%. A minimum of seven amino acids was required including N-terminal acetylation and methionine oxidation as variable modifications. Due to omitted reduction and alkylation, cysteine carbamidomethylation was removed from fixed modifications. Enzyme specificity was set to trypsin with a maximum of two allowed missed cleavages. First and main search mass tolerance was set to 70 ppm and 20 ppm, respectively. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs (MBR) with a 0.7-min retention-time match window and a 0.05 1/K0 ion mobility window. Label-free quantification was performed with the MaxLFQ algorithm[19] and a minimum ratio count of one. For diaPASEF raw file analysis (Fig. 3 and 4), we used a hybrid library approach combining a 12-fraction high-pH reversed-phase fractionated precursor library from U2OS cells with the directDIA search of the diaPASEF raw files. The hybrid library consisted of 178,948 precursors, 127,049 peptides, 9,954 protein groups and was generated with the Spectronaut software (version 14.5.200813.47784, Biognosys AG, Schlieren, Switzerland) under default settings. Search parameters were according to default settings. Protein intensities were normalized using the 'Local Normalization' (Q-value complete) algorithm in Spectronaut based on a local regression model. A protein and precursor FDR of 1% was used. Decoy hits and proteins, which did not pass the Q-value threshold, were filtered out prior to data analysis.

Bioinformatic analysis

Proteomics data analysis was performed with Perseus[20] and within the R environment (https://www.r-project.org/). MaxQuant output tables were filtered for 'Reverse', 'Only identified by site modification', and 'Potential contaminants' before data analysis. Data was stringently filtered to only keep proteins with 30% or less missing values (those displayed as 0 in MaxQuant output). Missing values were imputed based on a normal distribution (width = 0.3; downshift = 1.8) prior to statistical testing. Principal component analysis was performed in R. For multi-sample (ANOVA) or pairwise proteomic comparisons (two-sided unpaired t-test), we applied a permutation-based FDR of 5% to correct for multiple hypothesis testing. An s_0 value[21] of 0.1 was used for the pairwise proteomic comparison in Fig. 5E. Pathway enrichment analysis was performed in Perseus (Fig. 4A, Fisher's exact test with Benjamini-Hochberg FDR of 0.05) or ClusterProfiler[22] (Fig. 4D and 5C, D). For all ClusterProfiler analyses, an FDR filter of 0.05 was used. Minimum category size was set to 20 and maximum size to 500.

SUPPLEMENTARY REFERENCES

1. Sakaue-Sawano, A. *et al.* Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression. *Cell* **132**, 487–498 (2008).
2. Benediktsson, A. M., Schachtele, S. J., Green, S. H. & Dailey, M. E. Ballistic labeling and dynamic imaging of astrocytes in organotypic hippocampal slice cultures. *J. Neurosci. Methods* **141**, 41–53 (2005).
3. Stadler, C., Skogs, M., Brismar, H., Uhlén, M. & Lundberg, E. A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **73**, 1067–1078 (2010).
4. Carpenter, A. E. *et al.* CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
5. Smith, K. *et al.* CIDRE: An illumination-correction method for optical microscopy. *Nat. Methods* **12**, 404–406 (2015).
6. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
7. Toth, T. *et al.* Environmental properties of cells improve machine learning-based phenotype recognition accuracy. *Sci. Rep.* **8**, 1–9 (2018).
8. Piccinini, F. *et al.* Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data. *Cell Syst.* **4**, 651-655.e5 (2017).
9. Smith, K. & Horvath, P. Active learning strategies for phenotypic profiling of high-content screens. *J. Biomol. Screen.* **19**, 685–695 (2014).
10. Hollandi, R. *et al.* nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Syst.* **10**, 453-458.e6 (2020).
11. Isola, P., Zhu, J. Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* vols 2017-Janua 5967–5976 (Institute of Electrical and Electronics Engineers Inc., 2017).
12. Caicedo, J. C. *et al.* Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods* **16**, 1247–1253 (2019).
13. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2020).
14. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
15. Brunner, A.-D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. doi:10.1101/2020.12.22.423933.
16. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
17. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
18. Prianichnikov, N. *et al.* MaxQuant software for ion mobility enhanced shotgun proteomics. doi:10.1101/651760.
19. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–26 (2014).
20. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* (2016) doi:10.1038/nmeth.3901.
21. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied

to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5116–21 (2001).

22. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

3.5. MS-based proteomics proves noncanonical human ORF translation

The *in silico* proteome is bioinformatically inferred from the genome itself, following a distinct set of rules. These rules include the presence of a traditional start and stop codon, as well as base and amino acid conservation across genomes of the same and different species. Furthermore, those defined 'open reading frames' (ORFs) are routinely filtered to a minimum length of e.g. 100 amino acids, because it is difficult to assign coding regions below this threshold. It is also assumed that proteins or peptides need a certain amino acid length to fold into stable structures, which are required for independent function.[51]

Ribosome profiling, a method that 'freezes' ribosomes actively translating mRNAs, allows the reconstruction of nascently translated regions by recovering so-called ribosome density maps[55]. Briefly, the more ribosomes are identified to be located across a continuous stretch of mRNA, the higher the chance that this sequence is currently translated into a continuous amino acid sequence of a functional peptide or protein. Ribosome profiling ideally complements mass spectrometry, since it provides an additional ORF search space to find peptides in and because active translation or physical existence of the product can be directly proven by MS analysis. Our collaborators, the Jonathan Weissman group at UCSF, used this technology to reveal that active translation can occur outside of the above mentioned classical ORF-defining boundaries. Specifically, there is non-AUG translation initiation, upstreamORF (uORF) localization (located before the 5' start codon of a known ORF) known to regulate translation of the actual downstream ORF in cis, alternative start codons within a known ORF, or even small proteins encoded on thought to be long noncoding RNAs[425]. Many small proteins that do not follow the classical ORF annotation guidelines have been found and described by now[53,54]. However, a systematic evaluation of those hits and especially the opportunity to link them to function, including proof of physical existence to narrow down the number of true positive novelORF hits, had been lacking.

In this study, we combined ribosome profiling, advanced ORF-rating, mass spectrometry, large-scale CRISPR screens and sequencing to systematically identify noncanonical protein coding sequences (novelORFs) of the genome that are essential for cellular growth and whose disruption elicits robust transcriptomic and phenotypic changes in human cells.

First, we investigated the genome wide translation by ribosome profiling across several cell lines and annotated all ORFs including features like ribosome density accumulation at start and stop codons, three-nucleotide periodicity, and harringtonine-induced ribosome stalling at the translation initiation site[426]. This led to the identification of several thousand novel ORFs that were either unknown or already described, verifying our approach. Strikingly, these novel ORFs correlated with the hallmarks of active translation - high ribosome density maps and three nucleotide periodicity features of well characterized protein coding regions.

Next, we asked if the novel ORFs affect cell growth as a potentially functional peptide and turned to CRISPR-mediated 'barcoded knockins' at novelORF positions in combination with deep or single-cell RNA-sequencing as functional readout in several cell lines. Indeed, we found hundreds of our novel ORF hits to affect cell growth using indexed sgRNA (barcoded knockins) accumulation as readout after ten cell doublings – If the novelORF does not affect the growth phenotype it accumulates in the total population and vice versa[388]. Interestingly, novel ORFs affecting the phenotype had a higher conservation score than novel ORFs that did not affect the phenotype in our assay. Furthermore, canonical ORFs had a higher conservation score than novelORFs affecting the phenotype in our cellular screen. Disruption of novelORFs systematically affected known proteins involved in essential cellular processes like glycosylation and novelORF expression in trans could rescue the knockout phenotype.

Importantly, we wished to directly detect novelORF proteins by mass spectrometry, which is by itself challenging since the shorter a protein is the lower the chance of harboring a tryptic peptide with the required ionization efficiency for identification and quantification. We identified several novelORF proteins after high stringency filtering and manual inspection of mass spectra. We also asked if we can identify novelORFs that are presented as HLA class I peptides, since every protein has the chance to contribute to the antigen repertoire and also has to be counter-selected in the thymus during development. Indeed, we found several hundred novelORF peptides to be presented by the HLA system, highlighting that they undergo processing and presentation in an essential cellular system just like any other protein. Furthermore, we showed that many of these novelORF HLA class I peptides bind with high affinity to characterized allotypes, specific for the investigated cell lines.

We also tagged many novelORF hits found to impact the cellular growth phenotype with a minimally disruptive 16 amino acid long splitNeonGreen-tag (mNG11), which reconstitutes to a functional GFP by expression of the matching split-GFP in trans[128]. This prevents functional disruption of the very short novelORF derived peptides and allowed us to evaluate stable expression of the novelORF proteins, their cellular localization and also to perform MS-based interactomics studies to investigate interaction partners. Indeed, several novelORFs localized to distinct cellular compartments and we identified protein interaction partners for many, which co-localize to the same compartments. Interestingly, we also found uORFs that integrate into the main ORF complex and co-localize. There were also cases where uORFs even localized to distinct compartments independent of the main ORF. This reveals dependency and independency mechanisms to the main ORF. Strikingly, the knockout of the MIEF1 uORF induced differential expression of mitochondrial fusion and fission genes, while overexpression induced a fragmented mitochondrial phenotype hinting increased fission. As expected, the MIEF1 uORF knockout resulted in elongated, tubular mitochondria and therefore increased fusion, while exogenous expression rescued this phenotype.

In summary, our results highlight a yet to be fully explored novel protein and peptide diversity encoded by the human genome. Novel microproteins affect the cellular growth phenotype, localize to distinct cellular compartments, are part of protein complexes, and engage with the human leukocyte antigen system. Our data also indicate a role for upstream novelORF encoded peptides that regulate the downstream-encoded ORF, thereby challenging the mono-cistronic gene assumption, suggesting that bi-cistronic expression could be a general phenomenon in mammalian genomes and highlights an unexpected complexity of the human proteome.

This exciting project showed that the combination of several high-throughput screening technologies on different molecular levels allowed us to reveal novel biology. It also demonstrated that there is much room for technological developments to fully grasp and understand the biological complexity in the cell. Since novelORFs tend to be rather small with ~60 amino acids in length on average and consequently yield at most only a few of tryptic peptides of which only some will ionize efficiently, alternative approaches like top down peptide analysis or peptidomics could be more successful in physically identifying novelORFs. This is a topic we are currently working on.

3.5.1. Article 10: Pervasive translation of noncanonical human ORFs

Pervasive functional translation of noncanonical human open reading frames

Science, March 06, 2020

Jin Chen[1,2], **Andreas-David Brunner[3]**, J. Zachery Cogan[1,2], James K. Nuñez[1,2], Alexander P. Fields[1,2], Britt Adamson[1,2], Daniel N. Itzhak[4], Jason Y. Li[4], Matthias Mann[3,5], Manuel D. Leonetti[4], Jonathan S. Weissman[1,2, #]

# Correspondence

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94158, USA
[2]Howard Hughes Medical Institute, University of California, San Francisco, CA 94158, USA
[3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany.
[4]Cell Atlas Initiative, Chan Zuckerberg Biohub, San Francisco, CA 94158, USA.
[5]Clinical Proteomics Group, Proteomics Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen 2200, Denmark.

Contribution

In this project, I contributed to the experimental design, the writing of the paper and prepared all proteomics related figures. I optimized and performed all proteomics experiments including deep proteome, pulldown and HLA peptidome analyses. Furthermore, I established a high-confidence framework for the identification of microproteins encoded by novel open reading frames, many of which were thought to be long non-coding RNA and proved their physical existence as e.g. protein complex members, HLA presented peptides in cancer cells and their localization to distinct cellular compartments.

MOLECULAR BIOLOGY

Pervasive functional translation of noncanonical human open reading frames

Jin Chen[1,2], Andreas-David Brunner[3], J. Zachery Cogan[1,2], James K. Nuñez[1,2], Alexander P. Fields[1,2]*, Britt Adamson[1,2]†, Daniel N. Itzhak[4], Jason Y. Li[4], Matthias Mann[3,5], Manuel D. Leonetti[4], Jonathan S. Weissman[1,2]‡

Ribosome profiling has revealed pervasive but largely uncharacterized translation outside of canonical coding sequences (CDSs). In this work, we exploit a systematic CRISPR-based screening strategy to identify hundreds of noncanonical CDSs that are essential for cellular growth and whose disruption elicits specific, robust transcriptomic and phenotypic changes in human cells. Functional characterization of the encoded microproteins reveals distinct cellular localizations, specific protein binding partners, and hundreds of microproteins that are presented by the human leukocyte antigen system. We find multiple microproteins encoded in upstream open reading frames, which form stable complexes with the main, canonical protein encoded on the same messenger RNA, thereby revealing the use of functional bicistronic operons in mammals. Together, our results point to a family of functional human microproteins that play critical and diverse cellular roles.

Efforts to bioinformatically discover and annotate protein-coding open reading frames (ORFs) in genomes, termed coding sequences (CDSs), have traditionally relied on rules such as amino acid conservation and homology, translation initiation from an AUG start codon, and minimum length (i.e., 100 amino acids) (*1*). These rules have been widely adopted on the basis of the assumption that short peptides are unlikely to fold into stable structures to perform functions. However, the generality of these rules has been challenged. For example, the ribosomal protein RPL41 is a 25–amino acid (aa) peptide and both sarcolipin (SLN, 31 aa) and phospholamban (PLN, 52 aa) bind to and regulate the sarcoplasmic Ca^{2+} transporter SERCA (*2*, *3*). Additionally, MYC can be translated from a noncanonical start codon CUG (*4*), which demonstrates that non-AUG initiation can produce functional proteins. Recent studies have added a handful of examples of short proteins, or microproteins (also called micropeptides or just peptides), performing diverse functions (*5*–*18*), some encoded on transcripts annotated as long noncoding RNAs (lncRNAs). Finally, upstream ORFs (uORFs), located in the 5′ untranslated regions of mRNAs, have long been implicated in cis-acting translational control of the main, canonical CDS (*19*–*21*), though it has remained unclear whether they can generate stable, functional peptides.

Systematic identification of functional short CDSs remains challenging. Recent ribosome profiling (deep sequencing of ribosome-protected fragments) and mass spectrometry (MS) studies have identified thousands of previously unannotated CDSs (*22*–*25*) across bacteria, yeasts, viruses, and mammalian cells. However, for most cases, the cellular functions of these identified CDSs or their peptide products remain unexplored. We reasoned that the advent of CRISPR and its ability to precisely disrupt protein-coding regions (*26*), when combined with ribosome profiling, provides an opportunity to define and empirically characterize the functional protein-coding capacity of a given genome. In this work, we applied various types of approaches—including ribosome profiling, MS, and multiple CRISPR-based techniques—to systematically discover noncanonical CDSs encoded in the human genome and validate their critical roles in diverse cellular pathways.

To annotate potential CDSs comprehensively and accurately, we first investigated genome-wide translation by ribosome profiling across multiple cell types and conditions, including human induced pluripotent stem cells (iPSCs), iPSC-derived cardiomyocytes, human foreskin fibroblasts (HFFs), and HFFs infected with cytomegalovirus (*27*, *28*) (fig. S1A). We leveraged the ORF-RATER algorithm to annotate ORFs (*27*), incorporating multiple lines of evidence to identify ORFs undergoing active translation. This included consideration of the accumulation of ribosome densities at the start and stop codons, three-nucleotide periodicity, and additional experimental results, such as data from harringtonine-treated cells in which ribosomes are stalled at initiation sites (*27*). In iPSCs and cardiomyocytes, in addition to 9490 annotated CDSs (62% of the identified CDSs), we identified 3455 distinct, noncanonical CDSs (22%, i.e., with no in-frame overlap with previously annotated CDSs) and 2466 variant CDSs of annotated proteins (16%) in our high-statistical confidence set (Fig. 1A and materials and methods) (*27*). Among the distinct CDSs, 818 were CDSs on transcripts lacking prior protein-coding annotations ("new", i.e., lncRNAs), 2342 were upstream CDSs (i.e., uORFs or start overlaps: CDSs that overlap annotated start codons in a different reading frame), and only 13 were downstream CDSs. Similar numbers of CDSs were present in HFFs (fig. S1B), with 75% of the CDSs shared between the two cell types. Of the distinct CDSs, 96% are less than 100 aa in length, and 36% of the CDSs use non-AUG start codons (Fig. 1, B and C; see also fig. S2 for further characterizations).

Multiple lines of evidence suggest that the noncanonical CDSs are actively translated. The average ribosome density (metagene) of the lncRNA CDSs and of the translated uORFs closely mirrors footprints from that of annotated coding regions with strong three-nucleotide periodicities, a hallmark of active translation, as exemplified by traces from the lncRNA *LINC00998* transcript and a uORF of *ARL5A* (Fig. 1, D and E, and fig. S3). Our analysis also successfully recapitulated well-characterized short ORFs, such as the uORF on *ATF4* (*29*) and the recently discovered lncRNA-encoded microproteins MOXI/mitoregulin (*11*, *12*) and NoBody (*10*). Bona fide lncRNAs, such as *XIST*, *HOTAIR*, and *NEAT1*, were not identified to be protein coding (fig. S3E). Moreover, many of the CDSs were differentially translated during iPSC differentiation or viral infection (fig. S3F), providing evidence for translational control in different cell states.

MS-based proteomics in iPSCs and major human leukocyte antigen class I (HLA-I) peptidomics confirmed the stable expression of hundreds of noncanonical CDS peptides (Fig. 1F and figs. S4 and S5). HLA-I peptidomics identified 240 noncanonical peptides, which suggests that these peptides enter the HLA-I presentation pathway and contribute to the antigen repertoire and possible immunogenicity (Fig. 1F) (*30*). HLA-I prediction analysis cross-validated strong binding ($K_d \leq 50$ mM, where K_d is the dissociation constant) of noncanonical CDS HLA-I peptides to their respective allotypes (fig. S6) (*30*). MS-based proteomics using tryptic digestion identified far fewer noncanonical peptides, which may be due to challenges in detecting the trypsin-digested products from short, noncanonical CDSs or possibly to more rapid turnover of these noncanonical peptides (fig. S7).

To test whether translation of the noncanonical CDSs is important for cell growth and potentially yields functional peptides, we measured the growth phenotypes resulting from CRISPR-mediated ORF knockout in

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94158, USA. [2]Howard Hughes Medical Institute, University of California, San Francisco, CA 94158, USA. [3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany. [4]Cell Atlas Initiative, Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [5]Clinical Proteomics Group, Proteomics Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen 2200, Denmark.
*Present address: GRAIL, Inc., Menlo Park, CA 94025, USA.
†Present address: Department of Molecular Biology and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.
‡Corresponding author. Email: jonathan.weissman@ucsf.edu

pooled screens (*26*). We designed a Cas9 ORF single guide RNA (sgRNA) library to specifically knock out thousands of the noncanonical CDSs identified by ribosome profiling (Fig. 2A and materials and methods) (*31, 32*), targeting 1098 uORFs, 613 lncRNA CDSs, 352 extensions of annotated coding regions, 283 start overlaps, and 7 downstream CDSs. We performed pooled Cas9 knockout screens in iPSC and K562 chronic myeloid leukemia cells expressing Cas9 and the sgRNA library, akin to conventional pooled screens for essential proteins (*26, 31*). We measured sgRNA abundance in the cell populations shortly after library transduction and after 10 additional population doublings by deep sequencing to quantify the fitness defect conferred by each sgRNA. We then calculated a phenotype score (γ) and

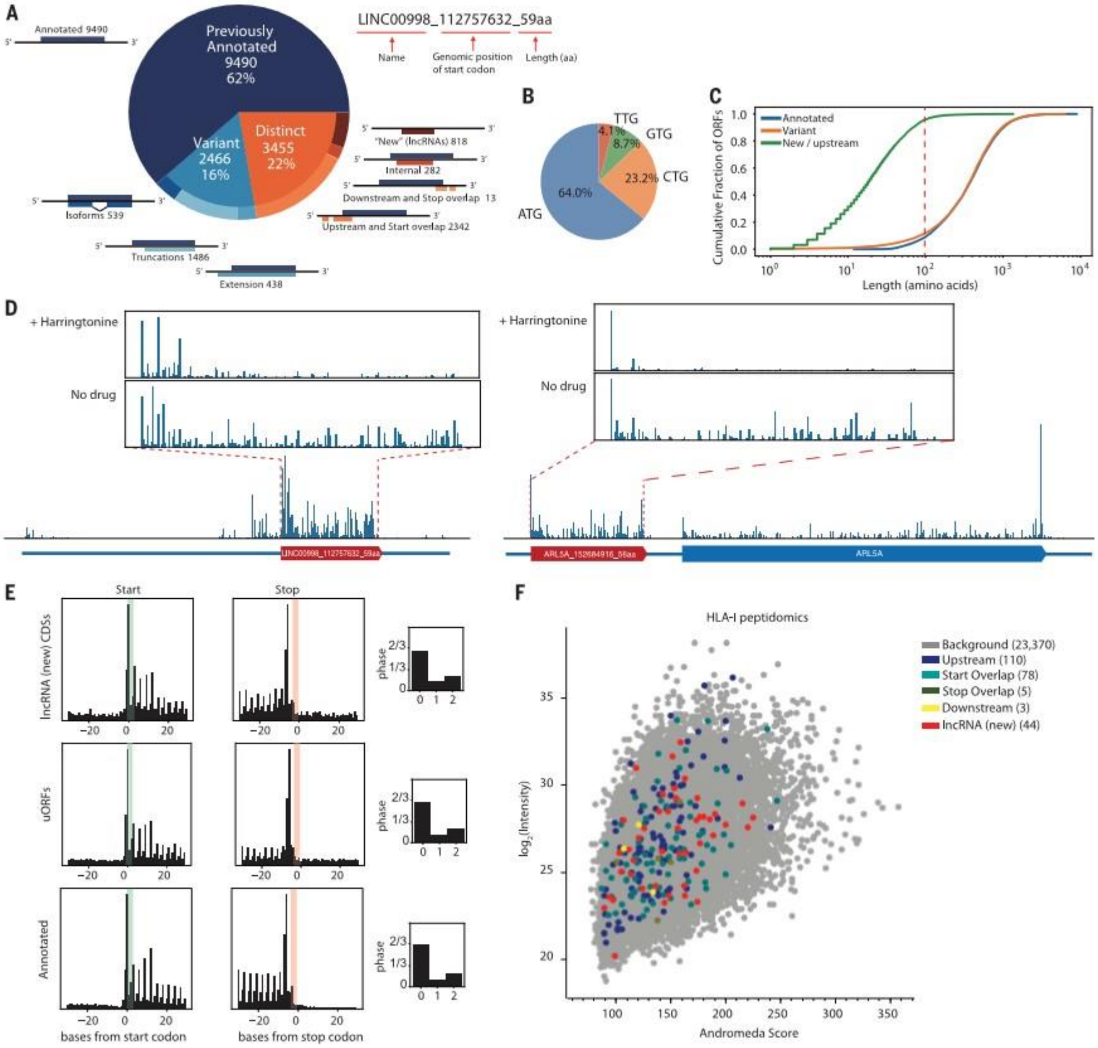
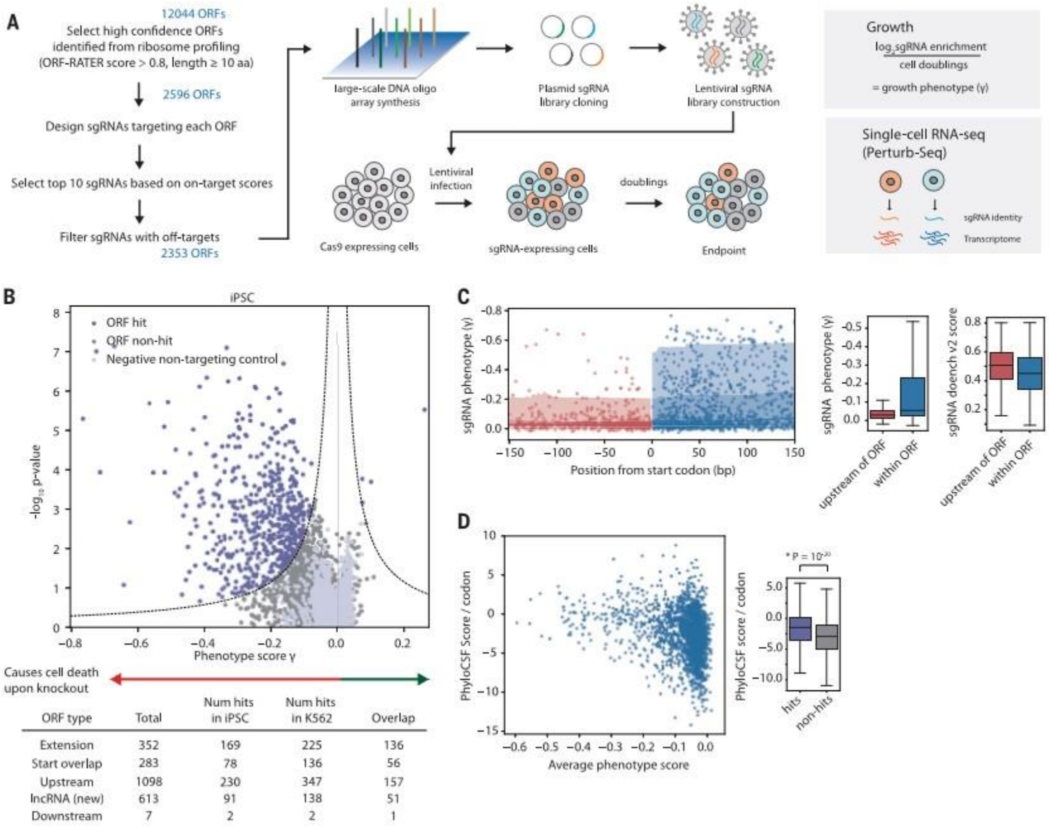


Fig. 1. Ribosome profiling and MS reveal translation of unannotated CDSs. (**A**) ORF-RATER analysis of ribosome profiling data: 62% are previously annotated coding sequences, whereas 16% are variants of canonical coding sequences that share portions of the coding sequence and 22% are distinct from annotated coding sequences. The naming convention of the identified ORFs is shown on the right. (**B**) Start-codon usage of the identified CDSs. (**C**) Cumulative distribution of CDS length. For distinct CDSs, 96% are smaller than 100 amino acids. (**D**) Example ribosome profiling traces of a lncRNA peptide from *LINC00998* and a uORF peptide from *ARL5A* displaying the hallmarks of translation, including peaks of density around the start codon following harringtonine treatment and three-nucleotide periodicities along the coding region. (**E**) Metagene analysis shows that the signatures of translation, including three-nucleotide periodicity in the expected reading frame, for uORFs and lncRNA CDSs are similar to those for annotated coding regions. (**F**) Identification of >200 noncanonical CDS peptides from HLA-I peptidomics, cross-validating their existence across the whole abundance range, with a mean Andromeda score of 141 compared with a total mean Andromeda score of 144. See materials and methods for further details.



ORF type	Total	Num hits in iPSC	Num hits in K562	Overlap
Extension	352	169	225	136
Start overlap	283	78	136	56
Upstream	1098	230	347	157
lncRNA (new)	613	91	138	51
Downstream	7	2	2	1

Fig. 2. Genome-scale CRISPR screens to identify functional, non-canonical CDSs. (**A**) Schematic of CRISPR library design and screening strategies, either by growth screens or Perturb-seq. For growth screens, frequencies of cells expressing a given sgRNA are determined by next-generation sequencing, and phenotype scores are quantified with the formula shown. For Perturb-seq, single-cell transcriptomes and sgRNA identities were obtained by single-cell RNA-seq. (**B**) Volcano plot summarizing knockout phenotypes and statistical significance (determined by Mann-Whitney *U* test) for ORFs targeted in the pooled screen in iPSCs. Each dot represents a targeted ORF, and ORF hits are labeled in purple, with a more negative phenotype score indicating a stronger growth defect. See materials and methods for further details. (**C**) Plot of the sgRNA phenotypes and distance from the start codon across all ORF hits. sgRNAs targeting the genome immediately upstream of the ORF (shown in red) have significantly lower phenotype scores than sgRNAs targeting within the ORF (shown in blue). Note the axis is increasingly negative (stronger) phenotype. The sgRNA phenotypes are quantified by the boxplot to the right. The difference is not because of differences in sgRNA on-target efficiencies, as quantified by the Doench v2 score. (**D**) The PhyloCSF score per codon (higher scores are more conserved across the Euarchontoglires) is generally higher for ORF hits (*$P = 10^{-20}$, Kolmogorov-Smirnov test) and ORFs with a stronger phenotype. Note that lack of a growth phenotype does not necessarily imply a low PhyloCSF score.

confidence (*P* value) for each ORF from the relative enrichment or depletion of sgRNAs targeting a particular ORF (Fig. 2B and materials and methods). In iPSCs, our screen identified >500 ORF knockout hits that resulted in statistically significant phenotypes. The hits include 169 genes that are variants of annotated proteins, 78 start overlap hits, 230 uORF hits, 91 lncRNA CDS hits, and 2 downstream CDS hits. iPSC and K562 cells had 401 shared hits, suggesting housekeeping or general cellular roles as well as CDSs that may play cell-specific functions (fig. S8). A fraction of the uORF hits do not have main, canonical CDSs with fitness defects upon knockout. This suggests an independent function of the uORFs or that disruption of the uORFs leads to increases in main CDS expression, which results in the growth phenotype (fig. S8E). Thus, unannotated CDSs with important functions across multiple cell types are an abundant feature of the genome.

Several lines of evidence further suggested that our screen reported specifically on the phenotypes of the selected ORFs. First, the phenotypes of control sgRNAs targeted directly upstream of each ORF in the genome (Fig. 2C) are significantly weaker than those of sgRNAs targeted within the ORF ($P = 10^{-26}$, Mann-Whitney test). Second, sgRNA phenotypes are independent of distance to other annotated proteins, splice sites, or transcriptional start sites (fig. S9A). Functionally, ORF hits are, on average, more phylogenetically conserved with a higher conservation score than non-hits (PhyloCSF score per codon, $P = 10^{-20}$, Kolmogorov-Smirnov test; Fig. 2D) (*33*), and they have other distinguishing sequence features (e.g., enrichment for Kozak consensus

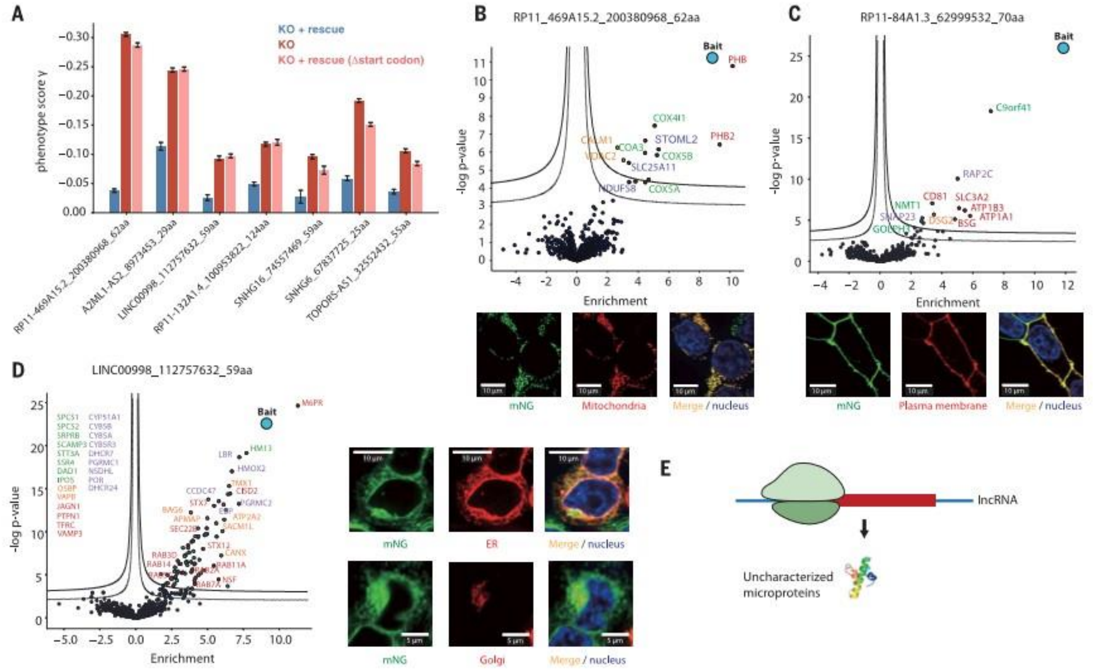
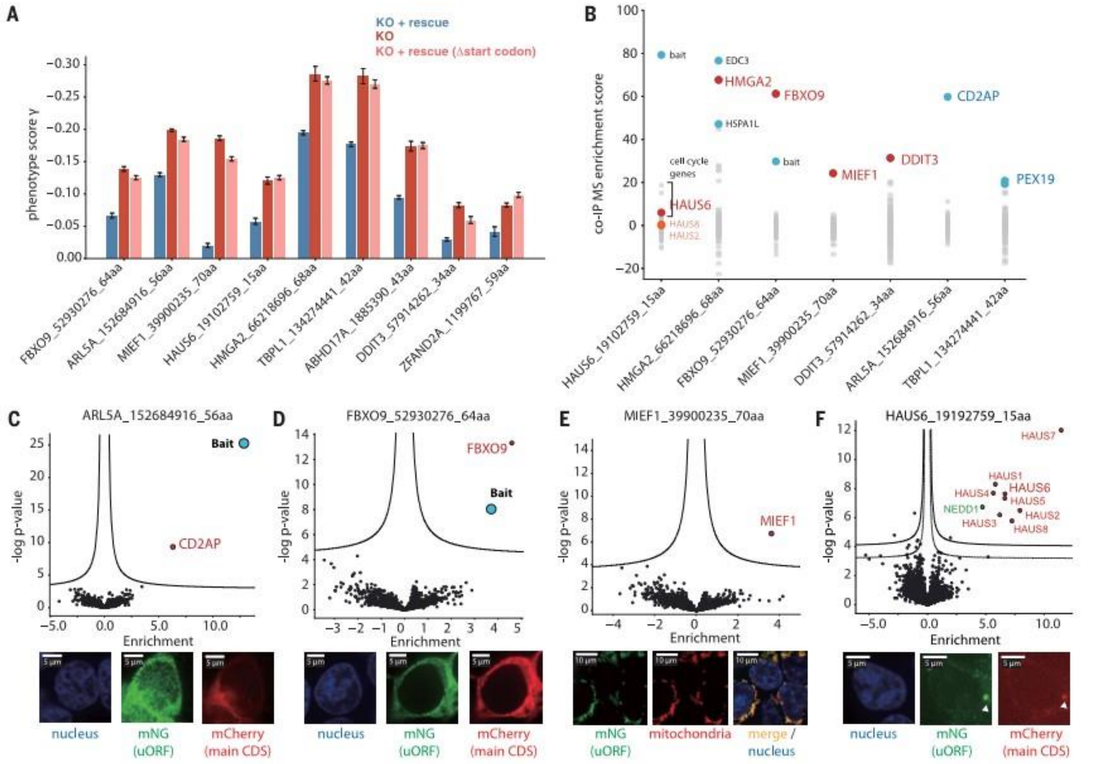


Fig. 3. Short lncRNA CDSs encode functional microproteins. (**A**) Rescue of lncRNA CDS knockout growth phenotypes by ectopic expression of the transcript encoding the peptide, as well as controls in which the initiating start codon is removed (Δstart codon). Error bars represent standard deviation of triplicates. $P < 0.05$ for all comparisons between knockout (KO) and KO + rescue. (**B** to **D**) Microscopy images and volcano plots of the co-IP MS of three example lncRNA-encoded microproteins tagged with mNG11, expressed ectopically (in the native transcript context) in a HEK293T cell line expressing mNG1-10. Green is mNG, red is the indicated organelle localization, and blue is Hoechst 33342, which stains for the nucleus. Scale bar dimensions are labeled. Statistically significant interactors are shown in the top, right corner of the volcano plots. Thick threshold line is 1% FDR (false discovery rate), and the thin threshold line is 5% FDR. The bait (the tagged peptide) is labeled in blue. The interactors are colored according to their functional groups. (**E**) lncRNA-encoded microproteins are uncharacterized proteins that may play important regulatory roles in cells.

sequence) (fig. S10). However, the noncanonical CDSs, on average, have lower PhyloCSF scores compared with canonical proteins (fig. S2B). Finally, sgRNAs targeting ORF hits versus nonhits have indistinguishable off-target and on-target scores (fig. S9B) (*32*). We then performed validation follow-ups with individual sgRNAs, which recapitulated the growth phenotypes from our genome-scale screen (fig. S8D). Sequencing of the targeted genomic regions revealed insertions and deletions (indels) of <50 base pairs (bp) (fig. S9, C and D). Together, these analyses independently support the conclusion that our screen phenotypes result specifically from the disruption of the target ORFs.

To survey function of the noncanonical CDSs at scale, we combined CRISPR screening with single-cell RNA sequencing (Perturb-seq) (*34*, *35*). Disruptions of the various noncanonical CDSs resulted in broad and diverse changes in RNA-sequencing profiles across a variety of critical pathways, suggesting that the candidate CDSs play diverse cellular roles (fig. S11). As an example, disruption of the CDS on *LINC00998* resulted in differentially expressed genes related to glycosylation ($P < 10^{-10}$), suggesting a function at the Golgi or endoplasmic reticulum (ER). The transcriptional phenotype also allowed us to functionally profile CDSs that are not essential for robust growth (fig. S11C). Furthermore, we found that CRISPR-targeted transcripts did not show detectable changes in abundance that might result from processes such as nonsense-mediated decay, which indicates that the phenotypes we observed were not caused by decreasing the abundance of the entire transcript (fig. S11D). Thus, similar to screens for essential protein-coding genes (*26*, *31*), our screen for noncanonical CDSs required for robust cell growth underestimated the true number of functional CDSs in the genome. This finding further underscores the pervasiveness of functional, unannotated CDSs in the genome that affect a wide range of cellular activities.

We next explored the functional role of the peptides encoded by the noncanonical CDSs identified from our screen, focusing first on lncRNA CDSs. For seven lncRNAs, we ectopically expressed the transcript encoding for the peptide and found, in all cases, that knockout-induced growth defect was partially or completely rescued. This rescue was abrogated by the removal of the initiating start codon (Δstart codon) (Fig. 3A), which suggests an essential role of the peptide itself in cell growth. To further investigate the specific functions of the noncanonical microproteins, we adopted a split-fluorescent protein approach using mNeonGreen (mNG), in which we fused each peptide with a minimally disruptive 16-aa tag (mNG11). Coexpression of the tagged peptide with the remainder of the mNG protein (mNG1-10) results in a fluorescence signal upon complementation (*36*, *37*). This creates



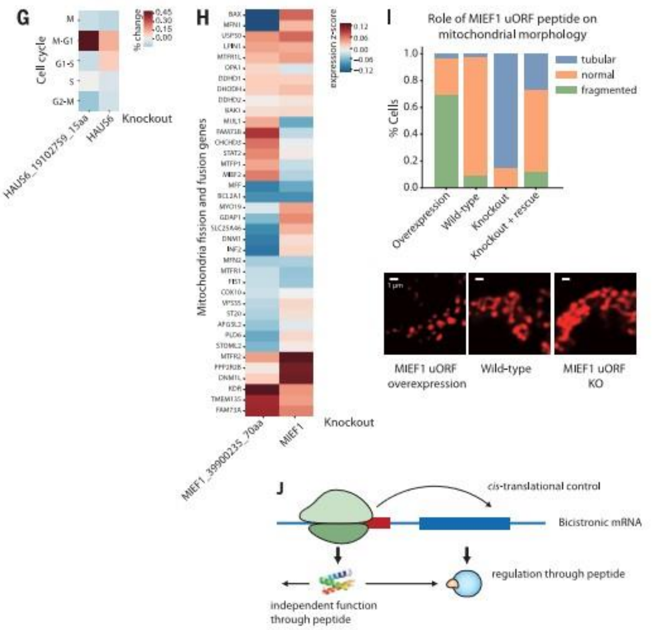


Fig. 4. Bicistronic mRNAs can encode uORF peptides that function in trans. **(A)** Rescue of uORF knockout growth phenotypes by ectopic expression of a transcript encoding the uORF peptide alone, as well as controls in which the initiating start codon is removed (Δstart codon). Error bars represent standard deviation of triplicates. $P < 0.05$ for all comparisons between KO and KO + rescue. **(B)** Summary of co-IP MS interactions of uORF peptides tagged with mNG11, showing five uORF peptides that interact with their downstream-encoded protein (shown in red). The tagged peptides are expressed alone ectopically (in the native transcript context) in an HEK293T cell line expressing mNG1-10. Other statistically significant interactors are shown in blue. **(C** to **E)** Example volcano plots of co-IP MS in (B) reveal statistically significant interactors with uORF peptides. Threshold line is 1% FDR. The bait (the tagged peptide) is labeled in blue. For microscopy in (C) and (D), the main, canonical protein tagged with mCherry (red) is coexpressed. For (E), the mNG11-tagged MIEF1 uORF peptide (green) localizes to the mitochondria (red). **(F)** Volcano plot of co-IP MS from endogenously mNG11-tagged HAUS6 uORF. For microscopy, the mNG11-tagged uORF is expressed alone ectopically (green), and the canonical HAUS6 tagged with mCherry (red) is coexpressed. **(G)** Percent change for each cell cycle state for HAUS6 knockout (KO) and HAUS6 uORF KO, compared with control cells. **(H)** Transcriptome response of the MIEF1 uORF KO compared with the main CDS KO from Perturb-seq. **(I)** Quantification of mitochondria morphology upon MIEF1 uORF peptide overexpression and knockout, as well as rescue of knockout phenotype. Representative microscopy images of the different mitochondria morphologies are shown to the right. **(J)** Possible model of uORF peptide functions and regulatory roles in cells.

both a fluorescent reporter to detect stable expression and cellular localization and a handle for coimmunoprecipitation (co-IP) and MS to define interaction partners (*36*) (fig. S12A). We probed the functions of six essential lncRNA CDSs and found that five of them formed specific complexes that were consistent with their subcellular localization. For example, the 62-aa peptide encoded by lncRNA *RP11_469A15.2* specifically localized to the mitochondria. The peptide has a predicted transmembrane domain and coimmunoprecipitates with the cytochrome c oxidase (COX) complex and the mitochondrial Prohibitin complex (Fig. 3B). Moreover, the 70-aa peptide encoded by *RP11-84A1.3* localizes to the plasma membrane and interacts with various cell surface proteins (Fig. 3C). Third, the 59-aa peptide encoded by lncRNA *LINC00998*, which contains two predicted transmembrane domains, localizes specifically to both the ER and Golgi and coimmunoprecipitates with lysosomal and vesicular transport proteins (Fig. 3D). Finally, the 55-aa peptide encoded on *TOPORS-AS1* and the 124-aa peptide on *RP11-132A1.4* also form functional complexes consistent with their cellular localization (fig. S12, C and D, and fig. S13). Consistent with prior studies (*5–18*), these examples demonstrate that lncRNAs can encode uncharacterized proteins, and they highlight the need to fully extend the annotation of lncRNAs and the proteome.

We next explored the functional effects of uORF translation, which is complicated by the fact that phenotypes can, in principle, be mediated by the peptide product (*24*, *38–41*), the effect of uORF translation on expression of the main, canonical CDS (*20*), or both. To distinguish between these possibilities, we first separately tagged the uORF and the main CDS and used Western blot to confirm the independent expression of uORF peptides from the canonical protein (fig. S14). Furthermore, we established that ectopic expression of a transcript encoding only the uORF peptide could at least partially rescue the growth phenotype caused by disruption of the endogenous uORF. In all cases this rescue is dependent on the initiating start codon in the ectopically expressed message, which demonstrates that the rescue is the result of production of the expressed peptide (Fig. 4A). Consistent with this, in all cases we tested, deleting the start codon for the uORFs only minimally increased (~20% to 60%) the expression of the main CDS. This suggests that the growth defect observed is mediated by the peptide and not by increased expression of the canonical protein (fig. S14, E and F). Taken together, these findings establish that uORFs could function through the peptide they produce independently of any cis-regulatory effects.

To explore the functions of uORF-encoded microproteins, we examined their localization and protein binding partners by tagging the uORF peptides with mNG11. Out of the 10 uORF peptides further tested by co-IP MS, we failed to detect statistically significant interaction partners for three of the tagged peptides. Two peptides, encoded by the uORFs of *TBPL1* and *ARL5A*, localize generally to the cytoplasm, whereas the main CDS proteins exhibit different cellular localization patterns. Consistent with our observed cellular localizations, these two uORF peptides specifically immunoprecipitate proteins with functions that are independent of the main CDS protein (Fig. 4, B and C, and fig. S12). Thus, these uORF peptides and their main CDS protein have independent functions.

We found that 5 of the 10 uORF peptides colocalized and formed a stable physical complex with the downstream-encoded, canonical protein on their shared mRNA. These include *MIEF1*, *DDIT3*, *FBXO9*, *HMGA2*, and *HAUS6* (Fig. 4, B, D, and E, and fig. S12). In all cases, we expressed the tagged peptides in their native transcript context but without the downstream CDS, thereby eliminating the possibility of stop codon read-through. We further confirmed this interaction by co-IP of the canonical protein and immunoblotting for the uORF peptide (fig. S12F), as well as with endogenously tagged clonal lines (fig. S15 and Fig. 4F). This physical interaction between the proteins encoded by the uORF and the canonical CDS on the same transcript is notable (*39*, *42*, *43*) because it implies an additional layer of regulation beyond the propensity of uORFs to modulate translation of downstream CDSs.

Next, we further investigated the function of uORF-expressed microproteins in *HAUS6* and *MIEF1*. In both cases, disrupting the uORF led to minimal increase in the expression of the main CDS protein, and the ectopic expression of a peptide-encoding transcript rescued the knockout-induced growth phenotype (Fig. 4A and fig. S14). mNG11-tagged HAUS6 uORF expressed from its endogenous locus efficiently pulled down key components of the HAUS6 complex, localized to the centrosome, and knockout of the uORF caused cells to arrest in the G1 stage, consistent with the role of HAUS6 microtubule attachment to the kinetochore and central spindle formation (Fig. 4, F and G, fig. S12, and fig. S15). Similarly, the MIEF1 uORF peptide localized to the mitochondria, consistent with the localization of the MIEF1 protein (Fig. 4E), which regulates mitochondrial fission and fusion (*44*). The MIEF1 uORF peptide knockout induced differential expression of mitochondrial fusion and fission genes, with a transcriptional signature that was distinct from that seen in the knockout of the MIEF1 protein (Fig. 4H). We observed that overexpression of the MIEF1 uORF peptide alone induced a fragmented mitochondrial phenotype (increased fission), whereas a clonal knockout of the MIEF1 uORF (with the sequence disrupted but nonetheless preserving an upstream ORF; see fig. S15) resulted in a tubular and more elongated mitochondrial phenotype (increased fusion). Notably, this knockout morphology could be rescued by the exogenous expression of the MIEF1 uORF peptide (Fig. 4I). Together, our results indicated a possible role of the uORF-encoded peptide in regulating the downstream-encoded protein, thereby challenging the monocistronic assumption about mammalian genomes. We speculate that this type of genomic architecture may be general, opening the doors to investigation of the cooperative and regulatory nature of bicistronic human mRNAs. Indeed, a number of stress-regulated alternate translation initiation factors can modulate translation initiation site choice and uORF usage, which suggests that regulation of bicistronic expression could play roles in both normal biology and diseases states (*21*, *45*).

We described a strategy that combines ribosome profiling, MS-based proteomics, microscopy, and CRISPR-based genetic screens to discover and characterize widespread translation of functional microproteins and define the protein-coding potential of complex genomes. We identified a subset of lncRNAs that can encode stable, functional proteins, which suggests that they may be misannotated RNAs or potentially have dual roles at the RNA and protein levels. Furthermore, we provided examples of uORFs encoding functional peptides, highlighting the diverse cellular roles that uORFs may play beyond translational control. We also identified uORF-encoded peptides binding to the downstream-encoded protein on the same mRNA. Thus, our data highlight a previously unappreciated complexity of the functional mammalian proteome, as well as the full spectrum of antigens presented by the HLA system.

REFERENCES AND NOTES

1. M. A. Basrai, P. Hieter, J. D. Boeke, *Genome Res.* **7**, 768–771 (1997).
2. A. Odermatt *et al.*, *Genomics* **45**, 541–553 (1997).
3. D. H. MacLennan, E. G. Kranias, *Nat. Rev. Mol. Cell Biol.* **4**, 566–577 (2003).
4. S. R. Hann, M. W. King, D. L. Bentley, C. W. Anderson, R. N. Eisenman, *Cell* **52**, 185–195 (1988).
5. R. Jackson *et al.*, *Nature* **564**, 434–438 (2018).
6. T. Kondo *et al.*, *Science* **329**, 336–339 (2010).
7. B. R. Nelson *et al.*, *Science* **351**, 271–275 (2016).
8. D. M. Anderson *et al.*, *Cell* **160**, 595–606 (2015).
9. E. G. Magny *et al.*, *Science* **341**, 1116–1120 (2013).
10. N. G. D'Lima *et al.*, *Nat. Chem. Biol.* **13**, 174–180 (2017).
11. C. S. Stein *et al.*, *Cell Rep.* **23**, 3710–3720e8 (2018).
12. C. A. Makarewich *et al.*, *Cell Rep.* **23**, 3701–3709 (2018).
13. A. Matsumoto *et al.*, *Nature* **541**, 228–232 (2017).
14. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, *J. Biol. Chem.* **289**, 10950–10957 (2014).
15. P. Bi *et al.*, *Science* **356**, 323–327 (2017).
16. J.-Z. Huang *et al.*, *Mol. Cell* **68**, 171–184.e6 (2017).
17. Q. Zhang *et al.*, *Nat. Commun.* **8**, 15664 (2017).
18. A. Pauli *et al.*, *Science* **343**, 1248636 (2014).
19. T. G. Johnstone, A. A. Bazzini, A. J. Giraldez, *EMBO J.* **35**, 706–723 (2016).
20. G. L. Chew, A. Pauli, A. F. Schier, *Nat. Commun.* **7**, 11663 (2016).

21. S. R. Starck *et al.*, *Science* **351**, aad3867 (2016).
22. N. T. Ingolia *et al.*, *Cell Rep.* **8**, 1365–1379 (2014).
23. N. T. Ingolia, L. F. Lareau, J. S. Weissman, *Cell* **147**, 789–802 (2011).
24. S. A. Slavoff *et al.*, *Nat. Chem. Biol.* **9**, 59–64 (2013).
25. A. A. Bazzini *et al.*, *EMBO J.* **33**, 981–993 (2014).
26. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, *Science* **343**, 80–84 (2014).
27. A. P. Fields *et al.*, *Mol. Cell* **60**, 816–827 (2015).
28. N. Stern-Ginossar *et al.*, *Science* **338**, 1088–1093 (2012).
29. K. M. Vattem, R. C. Wek, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11269–11274 (2004).
30. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, M. Mann, *Mol. Cell. Proteomics* **14**, 658–673 (2015).
31. L. A. Gilbert *et al.*, *Cell* **159**, 647–661 (2014).
32. A. R. Perez *et al.*, *Nat. Biotechnol.* **35**, 347–349 (2017).
33. M. F. Lin, I. Jungreis, M. Kellis, *Bioinformatics* **27**, i275–i282 (2011).
34. B. Adamson *et al.*, *Cell* **167**, 1867–1882.e21 (2016).
35. P. Datlinger *et al.*, *Nat. Methods* **14**, 297–301 (2017).
36. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3501–E3508 (2016).
37. S. Feng *et al.*, *Nat. Commun.* **8**, 370 (2017).
38. Z. Ji, R. Song, A. Regev, K. Struhl, *eLife* **4**, e08890 (2015).
39. S. Samandi *et al.*, *eLife* **6**, e27860 (2017).
40. A. Rathore *et al.*, *Biochemistry* **57**, 5564–5575 (2018).
41. V. Delcourt *et al.*, *Mol. Cell. Proteomics* **17**, 2402–2411 (2018).
42. D. Bergeron *et al.*, *J. Biol. Chem.* **288**, 21824–21835 (2013).
43. C. F. Lee, H. L. Lai, Y. C. Lee, C. L. Chien, Y. Chern, *J. Biol. Chem.* **289**, 1257–1270 (2014).
44. R. Yu *et al.*, *Sci. Rep.* **7**, 880 (2017).
45. A. Sendoel *et al.*, *Nature* **541**, 494–499 (2017).

ACKNOWLEDGMENTS

We thank the Weissman laboratory members, particularly T. M. Norman, M. Jost, M. J. Shurtleff, M. A. Horlbeck, L. A. Gilbert, M. Y. Hein, S. E. Torres, C. R. Liem, D. A. Santos, J. M. Replogle, and A. Xu. We also thank B. R. Conklin and M. P. Olvera (Gladstone Institute) for iPSC culturing; N. Cho (Chan-Zuckerberg Biohub) for endogenous tagging; and S. E. O'Leary and I W. Lin for discussions. **Funding:** This work was funded by NIH (RM1 HG009490), DARPA (HR0011-17-2-0043), and the Chan-Zuckerberg Initiative. J.S.W. is a HHMI Investigator. A.-D.B. and M.M. are supported by the Max Planck Society. J.C. is funded by the Jane Coffin Childs Memorial Fund for Medical Research and the NIH K99/R00 Pathway to Independence Award (K99 GM134154). J.K.N. is a fellow of the Hanna H. Gray Fellows Program. Oligonucleotide pools were courtesy of the Innovative Genomics Institute. **Author contributions:** J.C. designed and performed all experiments and analyzed and interpreted all the data with guidance from J.S.W. A.-D.B. and M.M. performed the proteomic MS and analysis. J.Z.C. and J.K.N. assisted with sample preparation, experiments, and key discussions. A.P.F. performed preliminary experiments and provided key analytical pipelines. B.A. contributed to Perturb-seq experiments. M.D.L. and J.Y.L. performed the pull-downs. M.D.L. designed the endogenous tagging methods. D.N.I. performed MS for the pull-downs. J.C. and J.S.W. conceived the study and wrote the manuscript with input from all authors. **Competing interests:** J.S.W. consults for and holds equity in KSQ Therapeutics and Maze Therapeutics and consults for 5AM Ventures. B.A. is an advisory board member and has restricted stock in Celsius Therapeutics, Inc. **Data and materials availability:** Raw sequencing data are deposited on the National Center for Biotechnology Information Gene Expression Omnibus database with accession number GSE131650. MS-based proteomics data are deposited to the ProteomeXchange Consortium via the Proteomics Identifications Database (PRIDE) partner repository with the dataset identifier PXD014031. Processed data are included as supplementary tables.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6482/1140/suppl/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S8
References (*46*–*70*)

View/request a protocol for this paper from *Bio-protocol*.

26 May 2019; resubmitted 22 November 2019
Accepted 13 January 2020
10.1126/science.aay0262

4. Discussion and outlook

Electrospray, the enabling technology we are using in nearly every MS-based proteomics experiment, only emerged 30 years ago[62,64]. Since then, the field has rapidly evolved and by now allows measuring proteomes at unprecedented throughput **(Article 3)**, depth **(Article 5, 10)** and sensitivity **(Articles 7, 8, 9)**. The technology is starting to be so robust and universally applicable that it is finally poised to impact the study of human health and disease as well as to tackle fundamental biological questions at spatial and single cell resolution[101,103]. To continue this developmental trajectory, sample preparation, liquid chromatography, mass spectrometry and bioinformatics technology have to continuously improve.

Every bottom-up experiment starts with the extraction of proteins, followed by proteolytic digestion and peptide clean up. Sophisticated isolation methods now enable the automated processing of bio-fluids, tissues and cell culture[100,103]. Even the processing of single cells comprising only about 150 pg of protein is rapidly advancing[39]. Miniaturization on glass slides or specialized chips and the use of MS-compatible protein solubilization agents side-stepping peptide clean-up allows the processing in less than 1 µl volume[116,149,408]. One of the largest remaining challenges was to transfer the single-cell derived tryptic peptides into the mass spectrometer. Our solution to this dilemma was to merge sample preparation with liquid chromatography using an EvoTip to concentrate single-cell derived tryptic peptides in a nanopackage that can readily be moved through the LC system **(Article 7)**. This is in contrast to solutions from other groups, which tend to be much more complex and therefore less robust for any ultra-high sensitivity application. The EvoTip itself has also proven to be beneficial for the large-scale human interactome project **(Article 3)**, since it serves as a concentration device and at the same time as disposable trap-column. This allowed us to use digitonin as a protein solubilizing agent, which is known for its superior membrane protein solubilizing characteristics[427].

Another challenge in single-cell and ultra-high sensitivity applications is the robust, high-resolution and high-throughput liquid chromatography itself[428]. The system has to ensure that the sample is transferred without any loss to the analytical column and limit radial dilution, since electrospray is concentration dependent[163]. Our approach to this challenge was to repurpose and standardize the *EvoSep One* platform, which is known for its robustness and reproducibility across hundreds of runs at microliter flow rate[150]. We optimized the sample transfer route to minimize sample loss. We made use of the fact that the preformed gradient is pushed out under the control of a single high-pressure pump, which in principle allows us to elute single-cell derived peptides at an arbitrarily low flow rate while all known benefits from the microflow gradients with regards to robustness and reproducibility remain. We standardized on a 100 nl/min flow rate to increase desolvation efficiency and ion transfer into the mass spectrometer, while keeping a high throughput of 40 samples per day. Standardization here was key to enable method optimization at very low sample loads and to perform comparative single-cell proteome studies.

Very low flow rates usually come at the expense of broader peaks, since the linear velocity of the mobile phase decreases when column inner diameters (ID) are kept constant. To compensate for this, researchers have turned to smaller ID columns, which comes at the expense of decreased overall robustness and reproducibility[145,146]. In practice, this renders them incompatible with large-scale studies. In our single-cell proteomics study, arguably the most performance limiting component is the rather large ID column (75 µm), which we retained since we experienced robustness issues with smaller IDs. A possible solution to this could be the next generation of chip-based columns, also called µPAC[139]. These columns can be manufactured in any ID with perfectly arranged micropillars, ensuring highest reproducibility and chromatographic performance[140]. We are currently investigating the combination of the chip-based µPACs with the *EvoSep One* nanoflow gradients to improve chromatography for ultra-high sensitivity applications and also for large-scale microflow approaches, where robustness and reproducibility are key.

Next up is the crucial electrospray process, which currently happens under ambient pressure conditions and therefore requires the transfer of gas-phase ions into the first vacuum stage of the mass spectrometer. This comes at the expense of analyte loss, especially at higher flow rates[429]. Sub-ambient pressure ionization appears to be a promising solution to elevate measurement sensitivity by making all ions available for analysis in the mass spectrometer[165]. This is a technology, which could enter the field sooner rather than later and could fit well into the ion path of the presented TIMS-qTOF instrument **(Article 1)**. This setup promises to further multiply performance in single-cell proteomics applications together with standardized very low flow *EvoSep One* chromatography.

Even though we already used a novel TIMS-qTOF instrument with a brighter ion source and the diaPASEF scan mode for ultra-high sensitivity applications down to the level of single cells **(Articles 6, 7, 8, 9)**, mass spectrometry technology and scan modes itself will always be subject of continuous improvement. The combination of diaPASEF **(Article 6)** with a continuously moving quadrupole window in a way that fully synchronizes the elution profile of trapped ions from the TIMS tunnel, can replace the stepped window acquisition to create a five-dimensional scan mode (m/z, intensity, ion mobility, retention time, quadrupole scan speed) at up to 100 % ion utilization. Note that this is in stark contrast to 'scanning quadrupole DIA' without TIMS, which typically samples only a few percent of the ion beam[144,351]. This would enable very fast cycle times and allow the measurement of large sample cohorts as in single-cell proteomics or clinical studies at unprecedented speed. Furthermore, increasing the capacity of the TIMS tunnel and the detector dynamic range of the TIMS-qTOF instrument **(Article 1)** would increase overall proteome dynamic range coverage in highly complex samples.

The advances in mass spectrometry and scan modes always depend on the implementation of highly sophisticated software. The arrival of deep learning in proteomics already allows the prediction of all dimensions in mass spectrometry, including CCS values **(Article 5)**[227,228,230]. It is set to obviate the need for experimental spectral libraries in DIA experiments and is beneficial for proteogenomic applications with large search spaces[232,353]. However, its full implementation in a dedicated search engine is still missing. Such software could be especially advantageous for ultra-high sensitivity applications where fragment ion spectra are often of low quality due to the compromised signal-to-noise. It will also be interesting to see how these models will transfer to post-translational modification searches and non-tryptic peptides derived from the human leukocyte antigen system.

Single-cell analysis and spatially resolved omics analyses are currently two of the 'hottest' topics in research[375,391]. The mRNA of a single cell was first analyzed in 2009 and since then many workflow developments have continued to revolutionize the field[366]. Due to the robustness and throughput of scRNA-seq of by now several hundred thousand cells per study and its promise to impact health and disease, consortia like the *Human cell atlas* and the *LifeTime initiative* aim to map all single-cell transcriptomes of the human body[395,396]. Interestingly, it appears that the latest scRNA-seq techniques have reached a technological plateau in terms of depth – only 12 years after the inception of scRNA-seq[366,382]. Keeping in mind that mRNA is only a mediator of the flow of molecular information and often regulated in transcriptional bursts, it may not be ideal to use this layer of information to define cell types[383]. This could explain the need for very large data sets in the field.

In contrast, from a biological perspective, a large proportion of the proteome has to be stable to enable cellular function and cell types are most likely defined by a subset of signature proteins, which give rise to their distinct phenotype **(Article 7)**. The field of single-cell proteomics is just rising and for now only very specialized laboratories have the knowledge and technology to perform these experiments. Two main approaches are currently emerging, which are the label-based multiplexed and the label-free **(Article 7)** analysis of single cells. Both have their distinct advantages and disadvantages, mainly in terms of throughput and quantitative accuracy. It will be interesting to see where the community develops in the future and which approach will take over.

Since scRNA-seq has the main goal of cell typing by aggregating the mRNA profiles of thousands of cells, it could also be possible to isolate cells by their phenotype itself and pool them before the MS-based analysis as we have shown in the concept of deep visual proteomics (DVP) **(Article 9)**. This drastically decreases project measurement time, but at the expensive of limiting proteome resolution compared to the analysis of single-cell proteomes and requires that the cellular phenotype correlates highly with its proteomic makeup. In our own work, we are interested to see if DVP can elucidate the contribution of pancreatic islet cells to the emergence of diabetes, which we have until now performed on complete islets **(Article 2)**. In a clinical setting, DVP would use tissue thin sections from a pathology laboratory. The unbiased localization of pathologies in whole organs, especially at an early stage, is very challenging. We showed that this can be realized by solvent-based whole organ clearing followed by whole organ imaging **(Article 8)**. We established a protocol to isolate thin sections of the pathology itself and subject them to unbiased proteomics analysis, a concept we call DISCO-MS. The combination of DISCO-MS and DVP could complete the cycle of unbiased analysis from whole organ analysis, pathology localization, target tissue isolation and cell-type or even subcellular resolved proteome analysis.

Classic bottom-up proteomics searches are performed against databases comprising more than 20,000 proteins and an optimized bottom-up proteomics strategy for the analysis of human cellular proteomes has already detected more than 14,000 protein isoforms and 12,200 protein-coding genes[101]. Still, those proteins are inferred from the genome following cutoff rules like the presence of classic start and stop codons at a minimum length of more than 100 amino acids to reduce possible noise[51]. These assumptions mean that proteomics data base searches are not able to capture proteins below this cutoff, which leaves the 'dark matter' of the proteome untouched. We have shown **(Article 10)** that a plethora of proteins exist below the 100 amino acid cutoff and execute essential functions in human cell lines. Even though we identified many actively translated open reading frames of presumably non-coding genomic regions by ribosome profiling[55], many of these proteins do not yield tryptic peptides and consequently evaded our traditional bottom-up proteomics approach. Therefore, it will be of highest importance to develop novel approaches to solve this dilemma by the implementation of dedicated workflows that aim to identify small proteins or peptides in their *in vivo* constitution. This would again come with challenges in sample preparation, LC-MS and especially bioinformatics approaches, since these peptides will have very different properties as compared to tryptic peptides in addition to the vastly increased computational search space.

In conclusion, it is an exciting time with a burst of technological breakthroughs in sample preparation, instrumentation and computational proteomics. Sample preparation is becoming automated and applicable to virtually every biological matrix and can be tailored to any experimental question. Instrumentation has become very robust, fast and sensitive, which finally allows the acquisition of thousands of samples within a single project. Also, computational proteomics is rapidly advancing and most importantly made available to the community as open-source packages. It appears that all these developments will enable MS-based proteomics to make even more important discoveries in basic biology and biomedicine. These will not only result from classical proteome profiling. Functional studies of proteome dynamics at unprecedented depth and sensitivity, as well as the integration with other layers of information like the genome, transcriptome, imaging and less explored modalities such as the metabolome and *in vivo* processed peptides will have a tremendous impact.

5. References

1. Darwin, C. *The origin of species*. (1859).
2. Theobald, D. L. A formal test of the theory of universal common ancestry. *Nature* **465**, 219–222 (2010).
3. Gregor Mendel. Experiments in Plant Hybridisation. (1866).
4. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics* **122**, 565–581 (2008).
5. Avery, O. T., Macleod, C. M. & Mc Carty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* **179**, 378–384 (1994).
6. CHARGAFF, E. *et al.* Nucleotide composition of pentose nucleic acids from yeast and mammalian tissues. *J. Biol. Chem.* **186**, 51–67 (1950).
7. Schrödinger, E. *What is Life?* (1944).
8. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
9. Watson, J. D. & Crick, F. H. C. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).
10. Crick, F. CSHL Archives Repository | On Protein Synthesis.
11. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol.* **15**, (2017).
12. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576–581 (1961).
13. Bernfield, M. R. & Nirenberg, M. W. RNA codewords and protein synthesis. *Science (80-.).* **147**, 479–484 (1965).
14. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
15. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φx174 DNA. *Nature* **265**, 687–695 (1977).
16. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).* **269**, 496–512 (1995).
17. Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science (80-.).* **270**, 397–403 (1995).
18. Waterston, R. & Sulston, J. The genome of Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 10836–10840 (1995).
19. Goffeau, A. *et al.* Life with 6000 genes. *Science (80-.).* **274**, 546–567 (1996).
20. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science (80-.).* **287**, 2185–2195 (2000).
21. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
22. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
23. Craig Venter, J. *et al.* The sequence of the human genome. *Science (80-.).* **291**, 1304–1351 (2001).
24. Kary Mullis, S. B. *et al.* PROCESS FOR AMPLIFYING, DETECTING, AND/OR-CLONING NUCLEIC ACID. (1986).
25. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16–18 (2008).
26. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad.*

Sci. U. S. A. **106**, 19096–19101 (2009).

27. Hoheisel, J. D. Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews Genetics* **7**, 200–210 (2006).

28. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

29. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, (2019).

30. Bolotin, A., Quinquis, B., Sorokin, A. & Dusko Ehrlich, S. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).

31. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science (80-.).* **315**, 1709–1712 (2007).

32. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* **337**, 816–821 (2012).

33. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2579–E2586 (2012).

34. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (80-.).* **339**, 823–826 (2013).

35. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* **339**, 819–823 (2013).

36. Lartigue, C. *et al.* Genome transplantation in bacteria: Changing one species to another. *Science (80-.).* **317**, 632–638 (2007).

37. Chin, J. W. *et al.* An expanded eukaryotic genetic code. *Science (80-.).* **301**, 964–967 (2003).

38. Fottner, M. *et al.* Site-specific ubiquitylation and SUMOylation using genetic-code expansion and sortase. *Nat. Chem. Biol.* **15**, 276–284 (2019).

39. Volpe, P. & Eremenko-Volpe, T. Quantitative Studies on Cell Proteins in Suspension Cultures. *Eur. J. Biochem.* **12**, 195–200 (1970).

40. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549 (2011).

41. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13**, 942 (2017).

42. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, (2015).

43. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* **356**, (2017).

44. Uhlén, M. *et al.* The human secretome. *Sci. Signal.* **12**, (2019).

45. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).

46. Fredriksson, S. *et al.* Protein detection using proximity-dependent DNA ligation assays. *Nat. Biotechnol.* **20**, 473–477 (2002).

47. Arunachalam, P. S. *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science (80-.).* **369**, 1210–1220 (2020).

48. Williams, S. A. *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).

49. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

50. Baker, M. Blame it on the antibodies. *Nature* **521**, 274–276 (2015).

51. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: Beautiful needles in the haystack. *Genome Research* **7**, 768–771 (1997).

52. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science (80-.).* **367**, 140–146 (2020).

53. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nature Chemical Biology* **11**, 909–916 (2015).

54. D'Lima, N. G. *et al.* A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).

55. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (80-.).* **324**, 218–223 (2009).

56. Doll, S. & Burlingame, A. L. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chemical Biology* **10**, 63–71 (2015).

57. Murray-Zmijewski, F., Slee, E. A. & Lu, X. A complex barcode underlies the heterogeneous response of p53 to stress. *Nature Reviews Molecular Cell Biology* **9**, 702–712 (2008).

58. Drucker, D. J., Habener, J. F. & Holst, J. J. Discovery, characterization, and clinical development of the glucagon-like peptides. *Journal of Clinical Investigation* **127**, 4217–4227 (2017).

59. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14**, 658–673 (2015).

60. Racle, J. *et al.* Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).

61. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).

62. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).

63. Mann, M., Meng, C. K. & Fenn, J. B. Interpreting Mass Spectra of Multiply Charged Ions. *Anal. Chem.* **61**, 1702–1708 (1989).

64. Mann, M. The ever expanding scope of electrospray mass spectrometry—a 30 year journey. *Nature Communications* **10**, 1–3 (2019).

65. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).

66. Kafader, J. O. *et al.* Measurement of Individual Ions Sharply Increases the Resolution of Orbitrap Mass Spectra of Proteins. *Anal. Chem.* **91**, 2776–2783 (2019).

67. De Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).

68. Hebert, A. S. *et al.* The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).

69. Piazza, I. *et al.* A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172**, 358-372.e23 (2018).

70. Piazza, I. *et al.* A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. *Nat. Commun.* **11**, 1–13 (2020).

71. Cappelletti, V. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**,

545-559.e22 (2021).

72. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science (80-.).* **346**, (2014).

73. Secher, A. *et al.* Analytic framework for peptidomics applied to large-scale neuropeptide identification. *Nat. Commun.* **7**, (2016).

74. Parker, B. L. *et al.* Multiplexed temporal quantification of the exercise-regulated plasma peptidome. *Mol. Cell. Proteomics* **16**, 2055–2068 (2017).

75. Vasilopoulou, C. G. *et al.* Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. *Nat. Commun.* **11**, (2020).

76. Pandeswari, P. B. & Sabareesh, V. Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Advances* **9**, 313–344 (2019).

77. Smith, L. M. & Kelleher, N. L. Proteoform: A single term describing protein complexity. *Nature Methods* **10**, 186–187 (2013).

78. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).

79. Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).

80. Hein, M. Y., Sharma, K., Cox, J. & Mann, M. Proteomic Analysis of Cellular Systems. *Handb. Syst. Biol.* 3–25 (2013). doi:10.1016/B978-0-12-385944-0.00001-0

81. Villén, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **3**, 1638 (2008).

82. Ruprecht, B., Zecha, J., Zolg, D. P. & Kuster, B. High pH reversed-phase micro-columns for simple, sensitive, and efficient fractionation of proteome and (TMT labeled) phosphoproteome digests. in *Methods in Molecular Biology* **1550**, 83–98 (Humana Press Inc., 2017).

83. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).

84. Mann, M. & Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **66**, 4390–4399 (1994).

85. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).

86. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).

87. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

88. Niu, L. *et al.* Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Mol. Syst. Biol.* **15**, (2019).

89. Bader, J. M. *et al.* Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Mol. Syst. Biol.* **16**, (2020).

90. Müller, J. B. *et al.* The proteome landscape of the kingdoms of life. *Nature* **582**, 592–596 (2020).

91. Gregersen, M. SPIN - Species by Proteome INvestigation. *bioRxiv* 2021.02.23.432520 (2021).

doi:10.1101/2021.02.23.432520

92. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).

93. Coscia, F. *et al.* Multi-level Proteomics Identifies CT45 as a Chemosensitivity Mediator and Immunotherapy Target in Ovarian Cancer. *Cell* **175**, 159-170.e16 (2018).

94. Coscia, F. *et al.* A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *J. Pathol.* **251**, 100–112 (2020).

95. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

96. Rundlett, K. L. & Armstrong, D. W. Mechanism of Signal Suppression by Anionic Surfactants in Capillary Electrophoresis-Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **68**, 3493–3497 (1996).

97. Hailemariam, M. *et al.* S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *J. Proteome Res.* **17**, 2917–2924 (2018).

98. Batth, T. S. *et al.* Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Mol. Cell. Proteomics* **18**, 1027–1035 (2019).

99. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).

100. Müller, T. *et al.* Automated sample preparation with SP 3 for low-input clinical proteomics. *Mol. Syst. Biol.* **16**, e9111 (2020).

101. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **4**, 587-599.e4 (2017).

102. McCarthy, J. *et al.* Carbamylation of proteins in 2-D electrophoresis - Myth or reality? *J. Proteome Res.* **2**, 239–242 (2003).

103. Geyer, P. E. *et al.* Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst.* **2**, 185–195 (2016).

104. Doll, S. *et al.* Rapid proteomic analysis for solid tumors reveals LSD1 as a drug target in an end-stage cancer patient. *Mol. Oncol.* **12**, 1296–1307 (2018).

105. Brunner, A. D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *bioRxiv* 2020.12.22.423933 (2020). doi:10.1101/2020.12.22.423933

106. Cleland, W. W. Dithiothreitol, a New Protective Reagent for SH Groups. *Biochemistry* 480–482 (1964).

107. Han, J. C. & Han, G. Y. A procedure for quantitative determination of tris(2- carboxyethyl)phosphine, an odorless reducing agent more stable and effective than dithiothreitol. *Anal. Biochem.* **220**, 5–10 (1994).

108. Crankshaw, M. W. & Grant, G. A. Modification of Cysteine. *Curr. Protoc. Protein Sci.* **3**, 15.1.1-15.1.18 (1996).

109. Nielsen, M. L. *et al.* Iodoacetamide-induced artifact mimics ubiquitination in mass spectrometry. *Nature Methods* **5**, 459–460 (2008).

110. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).

111. Glatter, T. *et al.* Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *J. Proteome Res.* **11**, 5145–5156 (2012).

112. Cristobal, A. *et al.* Toward an Optimized Workflow for Middle-Down Proteomics. *Anal. Chem.* **89**, 3318–3325 (2017).

113. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* **11**, 993–1006 (2016).

114. Rappsilber, J., Ishihama, Y. & Mann, M. Stop And Go Extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).

115. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

116. Williams, S. M. *et al.* Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Anal. Chem.* **92**, 10588–10596 (2020).

117. Li, Z. Y. *et al.* Nanoliter-Scale Oil-Air-Droplet Chip-Based Single Cell Proteomic Analysis. *Anal. Chem.* **90**, 5430–5438 (2018).

118. Borner, G. H. H. Organellar Maps Through Proteomic Profiling - A Conceptual Guide. *Molecular & cellular proteomics : MCP* **19**, 1076–1087 (2020).

119. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* **18**, 1819–1831 (2015).

120. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24**, 11–17 (2015).

121. Olsen, J. V & Mann, M. Status of large-scale analysis of posttranslational modifications by mass spectrometry. *Molecular and Cellular Proteomics* **12**, 3444–3452 (2013).

122. Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B. & Aebersold, R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat. Methods* **4**, 231–237 (2007).

123. Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23**, 94–101 (2005).

124. Hansen, F. M. *et al.* Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. *Nat. Commun.* **12**, (2021).

125. Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell* **44**, 325–340 (2011).

126. Cho, N. H. *et al.* OpenCell: proteome-scale endogenous tagging enables the cartography of human cellular organization. doi:10.1101/2021.03.29.437450

127. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).

128. Leonetti, M. D., Sekine, S., Kamiyama, D., Weissman, J. S. & Huang, B. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3501–E3508 (2016).

129. Wilson, S. R., Vehus, T., Berg, H. S. & Lundanes, E. Nano-LC in proteomics: Recent advances and approaches. *Bioanalysis* **7**, 1799–1815 (2015).

130. Nguyen, D. T.-T., Guillarme, D., Rudaz, S. & Veuthey, J. L. Fast analysis in liquid chromatography using small particle size and high pressure. *J. Sep. Sci.* **29**, 1836–1848 (2006).

131. Gerber, F. *et al.* Practical aspects of fast reversed-phase high-performance liquid chromatography using 3 µm

particle packed columns and monolithic columns in pharmaceutical development and production working under current good manufacturing practice. *J. Chromatogr. A* **1036**, 127–133 (2004).

132. Klinkenberg, A. & Zuiderweg, F. J. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. *Chem. Eng. Sci.* **5**, 271–289 (1956).

133. Gritti, F. & Guiochon, G. The van Deemter equation: Assumptions, limits, and adjustment to modern high performance liquid chromatography. *J. Chromatogr. A* **1302**, 1–13 (2013).

134. Gritti, F. & Guiochon, G. The current revolution in column technology: How it began, where is it going? *J. Chromatogr. A* **1228**, 2–19 (2012).

135. Petersson, P., Frank, A., Heaton, J. & Euerby, M. R. Maximizing peak capacity and separation speed in liquid chromatography. *J. Sep. Sci.* **31**, 2346–2357 (2008).

136. Xiang, Y., Liu, Y. & Lee, M. L. Ultrahigh pressure liquid chromatography using elevated temperature. *J. Chromatogr. A* **1104**, 198–202 (2006).

137. Kovalchuk, S. I., Jensen, O. N. & Rogowska-Wrzesinska, A. FlashPack: Fast and simple preparation of ultrahigh-performance capillary columns for LC-MS. *Mol. Cell. Proteomics* **18**, 383–390 (2019).

138. Müller-reif, J. B., Hansen, F. M., Schweizer, L., Treit, P. V & Geyer, P. E. A new parallel high-pressure packing system enables rapid multiplexed production of capillary columns. *bioRxiv* 1–17 (2021). doi:10.1101/2021.02.26.433033

139. Op De Beeck, J. *et al.* On the advantages of radially elongated structures in microchip-based liquid chromatography. *Anal. Chem.* **85**, 5207–5212 (2013).

140. De Malsche, W., Op De Beeck, J., De Bruyne, S., Gardeniers, H. & Desmet, G. Realization of 1 × 10 6 theoretical plates in liquid chromatography using very long pillar array columns. *Anal. Chem.* **84**, 1214–1219 (2012).

141. Stadlmann, J. *et al.* Improved Sensitivity in Low-Input Proteomics Using Micropillar Array-Based Chromatography. *Anal. Chem.* (2019). doi:10.1021/acs.analchem.9b02899

142. Bian, Y. *et al.* Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nat. Commun.* **11**, (2020).

143. Bian, Y. *et al.* Robust Microflow LC-MS/MS for Proteome Analysis: 38 000 Runs and Counting. *Anal. Chem.* **93**, 3690 (2021).

144. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nat. Biotechnol.* 1–9 (2021). doi:10.1038/s41587-021-00860-4

145. Greguš, M., Kostas, J. C., Ray, S., Abbatiello, S. E. & Ivanov, A. R. Improved Sensitivity of Ultralow Flow LC-MS-Based Proteomic Profiling of Limited Samples Using Monolithic Capillary Columns and FAIMS Technology. *Anal. Chem.* **92**, 14702–14712 (2020).

146. Cong, Y. *et al.* Improved Single-Cell Proteome Coverage Using Narrow-Bore Packed NanoLC Columns and Ultrasensitive Mass Spectrometry. *Anal. Chem.* **92**, 2665–2671 (2020).

147. Stejskal, K., Op De Beeck, J., Dürnberger, G., Jacobs, P. & Mechtler, K. Ultra-sensitive nanoLC-MS using second generation micro pillar array LC technology with Orbitrap Exploris 480 and FAIMS PRO to enable single cell proteomics. *bioXriv* (2021). doi:10.1101/2021.02.10.430648

148. Stejskal, K., Potěšil, D. & Zdráhal, Z. Suppression of peptide sample losses in autosampler vials. *J. Proteome Res.* **12**, 3057–3062 (2013).

149. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian

cells. *Nat. Commun.* **9**, 1–10 (2018).

150. Bache, N. *et al.* A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).

151. Binai, N. A. *et al.* Rapid analyses of proteomes and interactomes using an integrated solid-phase extraction-liquid chromatography-MS/ms system. *J. Proteome Res.* **14**, 977–985 (2015).

152. Gross, J. H. *Mass Spectrometry - A Textbook: Third Edition.* (Springer-Verlag, 2017).

153. Dole, M. *et al.* Molecular beams of macroions. *J. Chem. Phys.* **49**, 2240–2249 (1968).

154. Karas, M., Bachmann, D. & Hillenkamp, F. Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules. *Anal. Chem.* **57**, 2935–2939 (1985).

155. Karas, M., Bachmann, D., Bahr, U. & Hillenkamp, F. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *Int. J. Mass Spectrom. Ion Process.* **78**, 53–68 (1987).

156. Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Analytical Chemistry* **60**, 2299–2301 (1988).

157. Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* **88**, 4451–4459 (1984).

158. Zenobi, R. & Knochenmuss, R. Ion formation in maldi mass spectrometry. *Mass Spectrom. Rev.* **17**, 337–366 (1998).

159. Norris, J. L. & Caprioli, R. M. Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chemical Reviews* **113**, 2309–2342 (2013).

160. Aichler, M. & Walch, A. MALDI Imaging mass spectrometry: Current frontiers and perspectives in pathology research and practice. *Lab. Investig.* **95**, 422–431 (2015).

161. Taylor, G. Disintegration of water drops in an electric field. *Proc. R. Soc. London. Ser. A. Math. Phys. Sci.* **280**, 383–397 (1964).

162. Rayleigh, Lord. XX. On the equilibrium of liquid conducting masses charged with electricity. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **14**, 184–186 (1882).

163. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**, 1–8 (1996).

164. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).

165. Page, J. S., Tang, K., Kelly, R. T. & Smith, R. D. Subambient pressure ionization with nanoelectrospray source and interface for improved sensitivity in mass spectrometry. *Anal. Chem.* **80**, 1800–1805 (2008).

166. Meyer, J. G. & Komives, E. A. Charge state coalescence during electrospray ionization improves peptide identification by tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **23**, 1390–1399 (2012).

167. Hahne, H. *et al.* DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods* **10**, 989–991 (2013).

168. Paul, W., Reinhard, H. P. & von Zahn, U. Das elektrische Massenfilter als Massenspektrometer und Isotopentrenner. *Zeitschrift für Phys.* **152**, 143–182 (1958).

169. Douglas, D. J. Linear quadrupoles in mass spectrometry. *Mass Spectrom. Rev.* **28**, 937–960 (2009).

170. Dawson, P. H. Quadrupole mass analyzers: Performance, design and some recent applications. *Mass Spectrom. Rev.* **5**, 1–37 (1986).

171. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *J. Proteome Res.* **11**, 5479–5491 (2012).

172. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology* **12**, 483–490 (2008).

173. R. Julian, R. The Mechanism Behind Top-Down UVPD Experiments: Making Sense of Apparent Contradictions. *J. Am. Soc. Mass Spectrom.* **28**, 1823–1826 (2017).

174. Biemann, K. *MASS SPECTROMETRY OF PEPTIDES AND PROTEINS*. (1992).

175. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biol. Mass Spectrom.* **11**, 601–601 (1984).

176. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).

177. Mitchell Wells, J. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods in Enzymology* **402**, 148–185 (2005).

178. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* (2020). doi:10.1038/s41592-020-0781-4

179. Boyd, R. & Somogyi, Á. The mobile proton hypothesis in fragmentation of protonated peptides: A perspective. *J. Am. Soc. Mass Spectrom.* **21**, 1275–1278 (2010).

180. Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society* **120**, 3265–3266 (1998).

181. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9528–9533 (2004).

182. Mikesh, L. M. *et al.* The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1764**, 1811–1822 (2006).

183. Cotham, V. C. & Brodbelt, J. S. Characterization of Therapeutic Monoclonal Antibodies at the Subunit-Level using Middle-Down 193 nm Ultraviolet Photodissociation. *Anal. Chem.* **88**, 4004–4013 (2016).

184. Reilly, J. P. Ultraviolet photofragmentation of biomolecular ions. *Mass Spectrom. Rev.* **28**, 425–447 (2009).

185. Ko, B. J. & Brodbelt, J. S. Comparison of glycopeptide fragmentation by collision induced dissociation and ultraviolet photodissociation. *Int. J. Mass Spectrom.* **377**, 385–392 (2015).

186. Meier, F. *et al.* Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).

187. Bekker-Jensen, D. B. *et al.* A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Mol. Cell. Proteomics* **19**, 716–729 (2020).

188. Ligon, W. V. Molecular analysis by mass spectrometry. *Science (80-.).* **205**, 151–159 (1979).

189. Gohlke, R. S. & McLafferty, F. W. Early gas chromatography/mass spectrometry. *J. Am. Soc. Mass Spectrom.* **4**, 367–371 (1993).

190. Fuerstenau, S. D. *et al.* Mass spectrometry of an intact virus. *Angew. Chemie - Int. Ed.* **40**, 541–544 (2001).

191. Coles, J. & Guilhaus, M. Orthogonal acceleration - a new direction for time-of-flight mass spectrometry: Fast, sensitive mass analysis for continuous ion sources. *Trends Anal. Chem.* **12**, 203–213 (1993).

192. Dawson, J. H. J. & Guilhaus, M. Orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **3**, 155–159 (1989).

193. Guilhaus, M. Special feature: Tutorial. Principles and instrumentation in time-of-flight mass spectrometry. Physical

and instrumental concepts. *J. Mass Spectrom.* **30**, 1519–1532 (1995).

194. Ladislas Wiza, J. Microchannel plate detectors. *Nucl. Instruments Methods* **162**, 587–601 (1979).

195. Meier, F. *et al.* Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14**, 5378–5387 (2015).

196. Beck, S. *et al.* The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol. Cell. Proteomics* **14**, 2014–2029 (2015).

197. Ridgeway, M. E., Lubeck, M., Jordens, J., Mann, M. & Park, M. A. Trapped ion mobility spectrometry: A short review. *International Journal of Mass Spectrometry* **425**, 22–35 (2018).

198. Silveira, J. A., Ridgeway, M. E., Laukien, F. H., Mann, M. & Park, M. A. Parallel accumulation for 100% duty cycle trapped ion mobility-mass spectrometry. *Int. J. Mass Spectrom.* **413**, 168–175 (2017).

199. Toyoda, M., Okumura, D., Ishihara, M. & Katakuse, I. Multi-turn time-of-flight mass spectrometers with electrostatic sectors. *J. Mass Spectrom.* **38**, 1125–1142 (2003).

200. Schuerch, S., Schaer, M., Boernsen, K. O. & Schlunegger, U. P. Enhanced mass resolution in matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometry. *Biol. Mass Spectrom.* **23**, 695–700 (1994).

201. Mamyrin, B., Karataev, V., Shmikk, D. & Zagulin, V. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Sov. J. Exp. Theor. Phys.* **37**, 45 (1973).

202. Cotter, R. J. Time-of-Flight Mass Spectrometry for the Structural Analysis of Biological Molecules. *Anal. Chem.* **64**, (1992).

203. Yavor, M. *et al.* Planar multi-reflecting time-of-flight mass analyzer with a jig-saw ion path. in *Physics Procedia* **1**, 391–400 (2008).

204. Wiley, W. C. & McLaren, I. H. Time-of-flight mass spectrometer with improved resolution. *Rev. Sci. Instrum.* **26**, 1150–1157 (1955).

205. Makarov, A. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Anal. Chem.* **72**, 1156–1162 (2000).

206. Kingdon, K. H. A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Phys. Rev.* **21**, 408–418 (1923).

207. Knight, R. D. Storage of ions from laser-produced plasmas. *Appl. Phys. Lett.* **38**, 221–223 (1981).

208. Olsen, J. V. *et al.* Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010–2021 (2005).

209. Denisov, E., Damoc, E., Lange, O. & Makarov, A. Orbitrap mass spectrometry with resolving powers above 1,000,000. *Int. J. Mass Spectrom.* **325–327**, 80–85 (2012).

210. Hardman, M. & Makarov, A. A. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal. Chem.* **75**, 1699–1705 (2003).

211. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annual Review of Analytical Chemistry* **8**, 61–80 (2015).

212. Grinfeld, D., Aizikov, K., Kreutzmann, A., Damoc, E. & Makarov, A. Phase-constrained spectrum deconvolution for fourier transform mass spectrometry. *Anal. Chem.* **89**, 1202–1211 (2017).

213. Michalski, A. *et al.* Ultra high resolution linear ion trap orbitrap mass spectrometer (orbitrap elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11**, 1–11 (2012).

214. Scheltema, R. A. *et al.* The Q exactive HF, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field orbitrap analyzer. *Mol. Cell. Proteomics* **13**, 3698–3708 (2014).

215. Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **17**, 727–738 (2018).

216. Meier, F. *et al.* Deep learning the collisional cross sections of the peptide universe from a million training samples. *Nat. Commun.* **12**, 1185 (2020).

217. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

218. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

219. Prianichnikov, N. *et al.* Maxquant software for ion mobility enhanced shotgun proteomics. *Mol. Cell. Proteomics* **19**, 1058–1069 (2020).

220. Senko, M. W., Beu, S. C. & McLaffertycor, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **6**, 229–233 (1995).

221. Cox, J., Michalski, A. & Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* **22**, 1373–1380 (2011).

222. Mueller, L. N. *et al.* SuperHirn - A novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7**, 3470–3480 (2007).

223. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: The protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).

224. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

225. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).

226. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

227. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).

228. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).

229. Zhou, X. X. *et al.* PDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **89**, 12690–12697 (2017).

230. Tarn, C. & Zeng, W.-F. pDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning. *Anal. Chem.* **93**, 5815–5822 (2021).

231. Verbruggen, S. *et al.* PROTEOFORMER 2.0: Further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Mol. Cell. Proteomics* **18**, S126–S140 (2019).

232. Verbruggen, S. *et al.* Spectral prediction features as a solution for the search space size problem in proteogenomics. *Mol. Cell. Proteomics* 100076 (2021). doi:10.1016/j.mcpro.2021.100076

233. Ma, B. *et al.* PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).

234. Chi, H. *et al.* Open-pFind enables precise, comprehensive and rapid peptide identification in shotgun proteomics. *bioRxiv* 285395 (2018). doi:10.1101/285395

235. Devabhaktuni, A. *et al.* TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat. Biotechnol.* **37**, 469–479 (2019).

236. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).

237. Baumgardner, L. A., Shanmugam, A. K., Lam, H., Eng, J. K. & Martin, D. B. Fast parallel tandem mass spectral library searching using GPU hardware acceleration. *J. Proteome Res.* **10**, 2882–2888 (2011).

238. Bittremieux, W., Laukens, K. & Noble, W. S. Extremely fast and accurate open modification spectral library searching of high-resolution mass spectra using feature hashing and graphics processing units. *bioRxiv* (2019). doi:10.1101/627497

239. Savitski, M. M., Nielsen, M. L. & Zubarev, R. A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteomics* **5**, 935–948 (2006).

240. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **33**, 743–749 (2015).

241. Consortium, T. U. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35**, D193–D197 (2007).

242. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**, 731–740 (2016).

243. Bruderer, R. *et al.* Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Mol. Cell. Proteomics* **18**, 1242–1254 (2019).

244. Ong, S. E. & Mann, M. Mass Spectrometry–Based Proteomics Turns Quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).

245. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: A critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).

246. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry* **404**, 939–965 (2012).

247. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).

248. Ong, S. E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1**, 2650–2660 (2007).

249. Krüger, M. *et al.* SILAC Mouse for Quantitative Proteomics Uncovers Kindlin-3 as an Essential Factor for Red Blood Cell Function. *Cell* **134**, 353–364 (2008).

250. Schwanhäusser, B., Gossen, M., Dittmar, G. & Selbach, M. Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* **9**, 205–209 (2009).

251. Schwanhüusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

252. Blagoev, B., Ong, S. E., Kratchmarova, I. & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.* **22**, 1139–1145 (2004).

253. Ong, S. E., Kratchmarova, I. & Mann, M. Properties of 13C-substituted arginine in stable isotope labeling by

amino acids in cell culture (SILAC). *J. Proteome Res.* **2**, 173–181 (2003).

254. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7**, 383–385 (2010).

255. Bantscheff, M. *et al.* Robust and sensitive iTRAQ quantification on an LTQ Obitrap mass spectrometer. *Mol. Cell. Proteomics* **7**, 1702–1713 (2008).

256. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).

257. Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).

258. Wühr, M. *et al.* Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* **84**, 9214–9221 (2012).

259. Brenes, A., Hukelmann, J., Bensaddek, D. & Lamond, A. I. Multibatch TMT reveals false positives, batch effects and missing values. *Mol. Cell. Proteomics* **18**, 1967–1980 (2019).

260. Zecha, J. *et al.* TMT labeling for the masses: A robust and cost-efficient, in-solution labeling approach. *Mol. Cell. Proteomics* **18**, 1468–1478 (2019).

261. Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* **12**, 3586–3598 (2013).

262. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).

263. Savitski, M. M. *et al.* Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on orbitrap-type mass spectrometers. *Anal. Chem.* **83**, 8959–8967 (2011).

264. Winter, S. V. *et al.* EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15**, 527–530 (2018).

265. Wichmann, C. *et al.* MaxQuant.live enables global targeting of more than 25,000 peptides. *Mol. Cell. Proteomics* **18**, 982–994 (2019).

266. Geyer, P. E. *et al.* Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12**, 901 (2016).

267. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).

268. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4**, 1265–1272 (2005).

269. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).

270. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

271. Bruderer, R. *et al.* Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics* **16**, 2296–2309 (2017).

272. Röst, H. L., Malmström, L. & Aebersold, R. Reproducible quantitative proteotype data matrices for systems biology. *Molecular Biology of the Cell* **26**, 3926–3931 (2015).

273. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).

274. Choi, M. *et al.* MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).

275. Santos, A. *et al.* Clinical knowledge graph integrates proteomics data into clinical decision-making. *bioRxiv* 2020.05.09.084897 (2020). doi:10.1101/2020.05.09.084897

276. Lanucara, F., Holman, S. W., Gray, C. J. & Eyers, C. E. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nat. Chem.* **6**, 281–294 (2014).

277. Thomson, J. J. & Rutherford, E. XL. On the passage of electricity through gases exposed to Röntgen rays. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **42**, 392–407 (1896).

278. Tyndall, A. M. & Starr, L. H. The mobility of ions in air. Part IV.—Investigations by two new methods. *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character* **121**, 172–184 (1928).

279. May, J. C. & McLean, J. A. Ion mobility-mass spectrometry: Time-dispersive instrumentation. *Anal. Chem.* **87**, 1422–1436 (2015).

280. McAfee, K. B. & Edelson, D. Identification and mobility of ions in a townsend discharge by time-resolved mass spectrometry [8]. *Proceeding Phys. Soc.* **81**, 382–384 (1963).

281. Dole, M. *et al.* Gas phase macroions. *Macromolecules* **1**, 96–97 (1968).

282. McDaniel, E. W., Martin, D. W. & Barnes, W. S. Drift tube-mass spectrometer for studies of low-energy ion-molecule reactions. *Rev. Sci. Instrum.* **33**, 2–7 (1962).

283. Kim, T. *et al.* Design and implementation of a new electrodynamic ion funnel. *Anal. Chem.* **72**, 2247–2255 (2000).

284. Fernández-Maestre, R., Harden, C. S., Ewing, R. G., Crawford, C. L. & Hill, H. H. Chemical standards in ion mobility spectrometry. *Analyst* **135**, 1433–1442 (2010).

285. Mason, E. A. & Schamp, H. W. Mobility of gaseous lons in weak electric fields. *Ann. Phys. (N. Y).* **4**, 233–270 (1958).

286. Mason, E. A. & McDaniel, E. W. Transport Properties of Ions in Gases. *Transp. Prop. Ions Gases* (1988). doi:10.1002/3527602852

287. Griffin, G. W., Dzidic, I., Carroll, D. I., Stillwell, R. N. & Horning, E. C. Ion Mass Assignments Based on Mobility Measurements Validity of Plasma Chromatographic Mass Mobility Correlations. *Anal. Chem.* **45**, 1204–1209 (1973).

288. Ahmed, A. *et al.* Application of the Mason-Schamp equation and ion mobility mass spectrometry to identify structurally related compounds in crude oil. *Anal. Chem.* **83**, 77–83 (2011).

289. May, J. C. *et al.* Conformational ordering of biomolecules in the gas phase: Nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer. *Anal. Chem.* **86**, 2107–2116 (2014).

290. Dodds, J. N. & Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **30**, 2185–2195 (2019).

291. Dugourd, P., Hudgins, R. R., Clemmer, D. E. & Jarrold, M. F. High-resolution ion mobility measurements. *Rev. Sci. Instrum.* **68**, 1122–1129 (1997).

292. Belov, M. E., Buschbach, M. A., Prior, D. C., Tang, K. & Smith, R. D. Multiplexed ion mobility spectrometry-orthogonal time-of-flight mass spectrometry. *Anal. Chem.* **79**, 2451–2462 (2007).

293. Kanu, A. B., Gribb, M. M. & Hill, H. H. Predicting optimal resolving power for ambient pressure ion mobility

spectrometry. *Anal. Chem.* **80**, 6610–6619 (2008).

294. Baker, E. S. *et al.* Ion Mobility Spectrometry-Mass Spectrometry Performance Using Electrodynamic Ion Funnels and Elevated Drift Gas Pressures. *J. Am. Soc. Mass Spectrom.* **18**, 1176–1187 (2007).

295. Kirk, A. T., Raddatz, C. R. & Zimmermann, S. Separation of isotopologues in ultra-high-resolution ion mobility spectrometry. *Anal. Chem.* **89**, 1509–1515 (2017).

296. Baker, E. S. *et al.* Enhancing bottom-up and top-down proteomic measurements with ion mobility separations. *Proteomics* **15**, 2766–2776 (2015).

297. Pringle, S. D. *et al.* An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass Spectrom.* **261**, 1–12 (2007).

298. Giles, K., Williams, J. P. & Campuzano, I. Enhancements in travelling wave ion mobility resolution. in *Rapid Communications in Mass Spectrometry* **25**, 1559–1566 (John Wiley & Sons, Ltd, 2011).

299. Shvartsburg, A. A. & Smith, R. D. Fundamentals of traveling wave ion mobility spectrometry. *Anal. Chem.* **80**, 9689–9699 (2008).

300. Gelb, A. S., Jarratt, R. E., Huang, Y. & Dodds, E. D. A study of calibrant selection in measurement of carbohydrate and peptide ion-neutral collision cross sections by traveling wave ion mobility spectrometry. *Anal. Chem.* **86**, 11396–11402 (2014).

301. Giles, K. *et al.* A Cyclic Ion Mobility-Mass Spectrometry System. *Anal. Chem.* **91**, 8564–8573 (2019).

302. Shliaha, P. V., Bond, N. J., Gatto, L. & Lilley, K. S. Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. *J. Proteome Res.* **12**, 2323–2339 (2013).

303. Helm, D. *et al.* Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell. Proteomics* **13**, 3709–3715 (2014).

304. Shvartsburg, A. A., Tang, K. & Smith, R. D. Optimization of the design and operation of FAIMS analyzers. *J. Am. Soc. Mass Spectrom.* **16**, 2–12 (2005).

305. Cumeras, R., Figueras, E., Davis, C. E., Baumbach, J. I. & Gràcia, I. Review on Ion Mobility Spectrometry. Part 1: Current instrumentation. *Analyst* **140**, 1376–1390 (2015).

306. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **90**, 9529–9537 (2018).

307. Prasad, S., Belford, M. W., Dunyach, J. J. & Purves, R. W. On an aerodynamic mechanism to enhance ion transmission and sensitivity of faims for nano-electrospray ionization-mass spectrometry. *J. Am. Soc. Mass Spectrom.* **25**, 2143–2153 (2014).

308. Cong, Y. *et al.* Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per mammalian cell. *Chem. Sci.* **12**, 1001–1006 (2021).

309. Fernandez-Lima, F. A., Kaplan, D. A. & Park, M. A. Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **82**, 126106 (2011).

310. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* **14**, 93–98 (2011).

311. Michelmann, K., Silveira, J. A., Ridgeway, M. E. & Park, M. A. Fundamentals of trapped ion mobility spectrometry. *J. Am. Soc. Mass Spectrom.* **26**, 14–24 (2014).

312. Silveira, J. A., Ridgeway, M. E. & Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* **86**, 5624–5627 (2014).

313. Chai, M., Young, M. N., Liu, F. C. & Bleiholder, C. A Transferable, Sample-Independent Calibration Procedure for Trapped Ion Mobility Spectrometry (TIMS). *Anal. Chem.* **90**, 9040–9047 (2018).

314. Wigger, L. *et al.* Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories toward type 2 diabetes. *bioRxiv* (2020). doi:10.1101/2020.12.05.412338

315. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).

316. Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536 (2004).

317. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10**, 1785–1793 (2011).

318. McLafferty, F. W. Tandem mass spectrometry. *Science (80-.).* **214**, 280–287 (1981).

319. Zhang, Y., Wen, Z., Washburn, M. P. & Florens, L. Effect of dynamic exclusion duration on spectral count based quantitative proteomics. *Anal. Chem.* **81**, 6317–6326 (2009).

320. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448 (2018).

321. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. Label-free quantification with FDR-controlled match-between-runs. *bioRxiv* (2020). doi:10.1101/2020.11.02.365437

322. Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J. Proteome Res.* **18**, 4020–4026 (2019).

323. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **28**, 710–721 (2010).

324. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* (2008). doi:10.1038/msb.2008.61

325. Zweigenbaum, J. & Henion, J. Bioanalytical high-throughput selected reaction monitoring-LC/MS determination of selected estrogen receptor modulators in human plasma: 2000 samples/day. *Anal. Chem.* **72**, 2446–2454 (2000).

326. Yost, R. A. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Anal. Chem.* **51**, 16 (1979).

327. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11**, 1475–1488 (2012).

328. Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteomics* **11**, 1709–1723 (2012).

329. Addona, T. A. *et al.* Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **27**, 633–641 (2009).

330. Malmström, J. *et al.* Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **460**, 762–765 (2009).

331. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6940–6945 (2003).

332. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).

333. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev.*

Mol. Cell Biol. **6**, 577–583 (2005).

334. Picott, P., Aebersold, R. & Domont, B. The implications of proteolytic background for shotgun proteomics. *Mol. Cell. Proteomics* **6**, 1589–1598 (2007).

335. Sherwood, C. A. *et al.* Correlation between y-type ions observed in ion trap and triple quadrupole mass spectrometers. *J. Proteome Res.* **8**, 4243–4251 (2009).

336. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9**, 429–434 (2008).

337. Lesur, A. *et al.* Highly multiplexed targeted proteomics acquisition on a TIMS-QTOF. *Anal. Chem.* **93**, 1383–1392 (2021).

338. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).

339. Bern, M. *et al.* Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82**, 833–841 (2010).

340. Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).

341. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 1–15 (2014).

342. Geiger, T., Cox, J. & Mann, M. Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **9**, 2252–2261 (2010).

343. Panchaud, A. *et al.* Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Anal. Chem.* **81**, 6481–6488 (2009).

344. Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* **10**, 744–746 (2013).

345. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (2015).

346. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 1–12 (2017).

347. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* **32**, 219–223 (2014).

348. Muntel, J. *et al.* Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol. Omi.* **15**, 348–360 (2019).

349. Rosenberger, G. *et al.* Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat. Biotechnol.* **35**, 781–788 (2017).

350. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).

351. Moseley, M. A. *et al.* Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. *J. Proteome Res.* **17**, 770–779 (2018).

352. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent

acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).

353. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* **11**, 1–10 (2020).

354. Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J. & Mann, M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol. Cell. Proteomics* **11**, 1–11 (2012).

355. Hooke, R. *Micrographia, or, Some physiological descriptions of minute bodies made by magnifying glasses :with observations and inquiries thereupon /by R. Hooke ... Micrographia, or, Some physiological descriptions of minute bodies made by magnifying glasses :with observations and inquiries thereupon /by R. Hooke ...* (Printed by Jo. Martyn and Ja. Allestry, printers to the Royal Society ... , 2011). doi:10.5962/bhl.title.904

356. The diary of Robert Hooke, M.A., M.D., F.R.S., 1672-1680,. (1935). Available at: https://www.worldcat.org/title/diary-of-robert-hooke-ma-md-frs-1672-1680-transcribed-from-the-original-in-the-possession-of-the-corporation-of-the-city-of-london-guildhall-library/oclc/726891. (Accessed: 24th April 2021)

357. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Publ. Gr.* (2014). doi:10.1038/nmeth.2769

358. Eckert, M. A. *et al.* Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts. *Nature* **569**, 723–728 (2019).

359. De Rosa, S. C., Brenchley, J. M. & Roederer, M. Beyond six colors: A new era in flow cytometry. *Nat. Med.* **9**, 112–117 (2003).

360. Mahdessian, D. *et al.* Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* **590**, 649–654 (2021).

361. Levsky, J. M., Shenoy, S. M., Pezo, R. C. & Singer, R. H. Single-cell gene expression profiling. *Science (80-.).* **297**, 836–840 (2002).

362. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

363. Francis, J. M. *et al.* EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* **4**, 956–971 (2014).

364. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

365. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

366. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

367. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

368. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

369. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

370. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.

Cell **161**, 1202–1214 (2015).

371. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2**, 666–673 (2012).

372. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, (2016).

373. Almanzar, N. *et al.* A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).

374. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-.).* **357**, 661–667 (2017).

375. Method of the Year 2013. *Nat. Methods* **11**, 1 (2014).

376. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).

377. Okayama, H. & Berg, P. High-Efficiency Cloning of Full-Length cDNA. *Mol. Cell. Biol.* **2**, 161–170 (1982).

378. Saliba, A. E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).

379. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).

380. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1100 (2013).

381. Svensson, V. *et al.* Power analysis of single-cell rnA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).

382. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 147–150 (2020).

383. Larsson, A. J. M. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).

384. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

385. Fischer, D. S. *et al.* Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* **37**, 461–468 (2019).

386. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* **26**, 1271–1279 (2020).

387. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352**, 189–196 (2016).

388. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

389. Lee, Y. *et al.* XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci. Adv.* **7**, eabg4755 (2021).

390. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681.e18 (2020).

391. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).

392. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, 8746 (2019).

393. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

394. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq

batch correction. *Nat. Methods* **16**, 43–49 (2019).

395. Regev, A. *et al.* The human cell atlas. *Elife* **6**, (2017).

396. Rajewsky, N. *et al.* LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* 1–10 (2020). doi:10.1038/s41586-020-2715-9

397. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 53 (2020).

398. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* **21**, 630–644 (2020).

399. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

400. Gutierrez-Arcelus, M. *et al.* Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nat. Commun.* **10**, 1–15 (2019).

401. Tirosh, I. & Suvà, M. L. Deciphering human tumor biology by single-cell expression profiling. *Annual Review of Cancer Biology* **3**, 151–166 (2019).

402. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780–791 (2016).

403. Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **177**, 1330-1345.e18 (2019).

404. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).

405. Cheung, T. K. *et al.* Defining the carrier proteome limit for single-cell proteomics. *Nat. Methods* **18**, 76–83 (2021).

406. Huffman, R. G., Chen, A., Specht, H. & Slavov, N. DO-MS: Data-Driven Optimization of Mass Spectrometry Methods. *J. Proteome Res.* **18**, 2493–2500 (2019).

407. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* **14**, 2394–2404 (2015).

408. Hartlmayr, D. *et al.* An automated workflow for label-free and multiplexed single cell proteomics sample preparation at unprecedented sensitivity. *bioRxiv* doi:10.1101/2021.04.14.439828

409. Barovic, M. *et al.* Metabolically phenotyped pancreatectomized patients as living donors for the study of islets in health and diabetes. *Mol. Metab.* **27**, S1–S6 (2019).

410. Waanders, L. F. *et al.* Quantitative proteomic analysis of single pancreatic islets. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18902–18907 (2009).

411. Ishihama, Y., Rappsilber, J. & Mann, M. Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. *J. Proteome Res.* **5**, 988–994 (2006).

412. Youn, J. Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol. Cell* **69**, 517-532.e11 (2018).

413. Cajka, T. & Fiehn, O. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. *TrAC - Trends in Analytical Chemistry* **61**, 192–206 (2014).

414. Paglia, G. *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Anal. Chem.* **87**, 1137–1144 (2015).

415. Stow, S. M. *et al.* An Interlaboratory Evaluation of Drift Tube Ion Mobility-Mass Spectrometry Collision Cross Section Measurements. *Anal. Chem.* **89**, 9048–9055 (2017).

416. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

417. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *arXiv* **30**, (2017).

418. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).

419. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics* **10**, 317 (2019).

420. Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* **22**, (2021).

421. Molbay, M., Kolabas, Z. I., Todorov, M. I., Ohn, T. & Ertürk, A. A guidebook for DISCO tissue clearing. *Mol. Syst. Biol.* **17**, (2021).

422. Chen, G. F. *et al.* Amyloid beta: Structure, biology and structure-based therapeutic development. *Acta Pharmacol. Sin.* **38**, 1205–1235 (2017).

423. Smith, K. & Horvath, P. Active learning strategies for phenotypic profiling of high-content screens. *J. Biomol. Screen.* **19**, 685–695 (2014).

424. Hollandi, R. *et al.* nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Syst.* **10**, 453-458.e6 (2020).

425. Starck, S. R. *et al.* Translation from the 5' untranslated region shapes the integrated stress response. *Science (80-.).* **351**, aad3867–aad3867 (2016).

426. Fields, A. P. *et al.* A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **60**, 816–827 (2015).

427. Stetsenko, A. & Guskov, A. An overview of the top ten detergents used for membrane protein crystallization. *Crystals* **7**, 197 (2017).

428. Shishkova, E., Hebert, A. S. & Coon, J. J. Now, More Than Ever, Proteomics Needs Better Chromatography. *Cell Systems* **3**, 321–324 (2016).

429. Wilm, M. Principles of electrospray ionization. *Molecular and Cellular Proteomics* **10**, (2011).

6. Appendix

6.1. Article 11: MaxQuant.Live

MaxQuant.Live enables global targeting of more than 25,000 peptides

Molecular & Cellular Proteomics, February 12, 2019

Christoph Wichmann[1], Florian Meier[2], Sebastian Virreira Winter[2], **Andreas-David Brunner[2]**, Jürgen Cox[1, #], Matthias Mann[1, 3, #]

# Correspondence

[1] Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[2] Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[3] NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark

Author's Choice

MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides

Christoph Wichmann‡, Florian Meier§¶, Sebastian Virreira Winter§, Andreas-David Brunner§, Jürgen Cox‡**, and Matthias Mann§¶‖

Mass spectrometry (MS)-based proteomics is often performed in a shotgun format, in which as many peptide precursors as possible are selected from full or MS1 scans so that their fragment spectra can be recorded in MS2 scans. Although achieving great proteome depths, shotgun proteomics cannot guarantee that each precursor will be fragmented in each run. In contrast, targeted proteomics aims to reproducibly and sensitively record a restricted number of precursor/fragment combinations in each run, based on prescheduled mass-to-charge and retention time windows. Here we set out to unify these two concepts by a global targeting approach in which an arbitrary number of precursors of interest are detected in real-time, followed by standard fragmentation or advanced peptide-specific analyses. We made use of a fast application programming interface to a quadrupole Orbitrap instrument and real-time recalibration in mass, retention time and intensity dimensions to predict precursor identity. MaxQuant.Live is freely available (www.maxquant.live) and has a graphical user interface to specify many predefined data acquisition strategies. Acquisition speed is as fast as with the vendor software and the power of our approach is demonstrated with the acquisition of breakdown curves for hundreds of precursors of interest. We also uncover precursors that are not even visible in MS1 scans, using elution time prediction based on the auto-adjusted retention time alone. Finally, we successfully recognized and targeted more than 25,000 peptides in single LC-MS runs. Global targeting combines the advantages of two classical approaches in MS-based proteomics, whereas greatly expanding the analytical toolbox. ***Molecular & Cellular Proteomics 18: 982–994, 2019. DOI: 10.1074/mcp.TIR118.001131.***

Mass spectrometry (MS)-based proteomics has matured into a versatile and widely used analytical tool in the life sciences (1–3). State-of-the-art workflows cover the proteome of model organisms to near-completeness and sensitivity extends into the attomole range (4–7). Still, many, and in particular low-abundance, proteins escape accurate and reproducible quantification across large sets of biological samples, which hinders wider applications of proteomics in systems biology and translational medicine (8, 9). In part, this is because of the complexity of bottom-up proteomics samples. Enzymatic digestions of protein extracts are comprised of millions of peptide species and, even with liquid chromatography, many peptides co-elute, spanning several orders of magnitude in abundance (10). In data-dependent acquisition schemes (DDA)[1], the mass spectrometer acquires as many MS2 spectra as possible to maximize the number of peptide identifications. As high-abundance ions are more likely to yield high-quality MS2 spectra, precursors are typically prioritized for isolation and fragmentation by their abundance and dynamic exclusion is employed to prevent their resequencing. This topN approach has been the method of choice for unbiased and comprehensive proteomic studies for many years. However, given the enormous number of precursor candidates and unavoidable run-to-run variabilities, a specific peptide may not be fragmented in every run. Data-independent acquisition (DIA) aims to avoid stochasticity by repeatedly cycling through fixed isolation windows (11), however, it increases spectral complexity and may diminish dynamic range and by its nature does not address all application scenarios .

In contrast to DDA and DIA, the goals of targeted proteomics methods are to analyze a limited number of selected proteins of interest with maximum sensitivity and reproducibility (12). Rather than selecting precursors based on the MS1 scans, the instrument is instructed to continuously fragment certain predefined precursor-fragment mass combinations during scheduled time windows, in which the targeted peptides are expected.

Traditionally, these experiments have been performed on triple quadrupole instruments, even before the advent of proteomics (13, 14). However, today high-resolution time-of-flight or Orbitrap mass analyzers are gaining popularity. Instead of recording one or a few specified precursor-fragment

From the ‡Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany; §Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany; ¶NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark

Received October 12, 2018, and in revised form, February 4, 2019
Published, MCP Papers in Press, February 12, 2019, DOI 10.1074/mcp.TIR118.001131

transition, these instruments acquire complete MS2 spectra and therefore monitor all fragment ions simultaneously, which increases specificity and quantitative accuracy (15, 16). In practice, setting up a robust targeted proteomics experiment with a desired coverage remains challenging as the number of targets needs to be balanced with acquisition speed and sensitivity (17, 18). Selecting too many targets may unduly reduce the acquisition time for each of them, whereas specifying narrow LC elution windows increases the risk of missing a peptide entirely as the retention times cannot be estimated very accurately beforehand. Reports in the literature generally employ minute-wide monitoring windows to target tens and sometimes hundreds peptides and proteins. Further, despite the creation of community-wide MRM peptide libraries (19–21), these assays are typically reestablished and optimized in each laboratory.

To address some of the above limitations, Domon and co-workers spiked-in isotope-labeled variants of the peptides of interest to trigger "pretargeting" and targeting events more precisely (22). Coon and colleagues used the expected elution order of peptides to bias DDA toward peptides of interest (23). Building on the MaxQuant software suite (24, 25) our own group developed the MaxQuant-RealTime framework, which identified peptides within milliseconds, providing a basis to implement intelligent data acquisition methods in different research scenarios (26). Although these concepts and potential applications are very promising in principle, the uptake of the underlying software packages was limited.

Proteomics post-processing algorithms generally contain a mass recalibration step as well as retention time alignment, which can be used to transfer identifications between runs (24, 27, 28). For instance, MaxQuant can achieve sub-parts-per-million (ppm) mass accuracies and absolute retention time deviations below 30s by nonlinear recalibration and alignment. We reasoned that approaching such an accuracy in real-time could dramatically improve our ability to predict the appearance of a very large number of peptides. This would drastically reduce the monitoring time for each peptide and might allow extending the targeting concept to a global, proteome-wide scale.

To realize this vision, we developed the freely available software MaxQuant.Live, which interacts with any Thermo Fisher Q Exactive mass spectrometer via the redesigned instrument application programing interface (IAPI) (29). Scan modules ("apps") can be plugged into the MaxQuant.Live core application on the acquisition computer, allowing straightforward implementation and modification of standard acquisition schemes as well as advanced data acquisition strategies based on live data analysis.

EXPERIMENTAL PROCEDURES

Cell Culture and Sample Preparation—We cultured the human HeLa cancer cell line (HeLa S3, ATCC, Manassas, VA) in Dulbecco's modified Eagle's medium (DMEM) with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin-streptomycin added (all Life Technologies Ltd., Paisley, UK). Metabolic stable isotope labeling (30) was performed in arginine- and lysine-free DMEM, fortified with arginine and lysine with natural isotope abundances (light channel) or stable-isotope labeled arginine-10 and lysine-8 (Cambridge Isotope Laboratories, Tewksbury, MA) as previously described (31). Cells were collected by centrifugation, washed twice with cold phosphate-buffered saline, pelleted and stored at −80 °C.

We lysed the cells and reduced and alkylated the proteins in a single reaction vial with sodium deoxycholate (SDC) buffer containing chloroacetamide (PreOmics GmbH, Martinsried, Germany) following our previously published protocol (32). Cells were suspended in the SDC buffer and boiled for 10 min at 95 °C. To disrupt remaining cellular structures and shear nucleic acids, we sonicated the suspension for 15 min at full power (Bioruptor, Diagenode, Seraing, Belgium). The crude protein extracts were enzymatically digested with LysC and trypsin (1:100, enzyme wt/protein wt) overnight at 37 °C before stopping the reaction with 5 volumes of isopropanol/1% trifluoroacetic acid (TFA). Peptide micro-purification and de-salting was performed on styrenedivinylbenzene-reversed phase sulfonate StageTips. Following sequential washing steps with isopropanol/1% TFA and water with 0.1% TFA, peptides were eluted with 80% acetonitrile (ACN) containing 1% ammonia. The vacuum dried eluates were reconstituted in water with 2% ACN and 0.1% TFA for further analysis.

Liquid Chromatography and Mass Spectrometry (LC-MS)—In single LC-MS runs, ~500 ng of purified whole-cell digests were analyzed with an EASY-nLC 1200 nanoflow chromatography system (Thermo Fisher Scientific, Bremen, Germany) coupled online to a hybrid quadrupole Orbitrap mass spectrometer (Thermo Q Exactive HF-X (33)). The peptides were separated at 60 °C on a 50 cm long column (75 μm inner diameter) packed with 1.9 μm porous silica beads (Dr. Maisch, Ammerbuch-Entringen, Germany), and electrosprayed from a laser-pulled silica emitter tip at 2.4 kV. Mobile phases A and B were water with 0.1% formic acid (v/v) and 80/20/0.1% ACN/water/formic acid (v/v/v). To elute the peptides at a constant flow rate of 300 nL/min, a binary gradient was ramped from 5% to 30% B within 95 min, followed by an increase to 60% B within 5 min and further to 95% B for washing. After 5 min, the organic content was decreased to the starting value within 5 min and the column was reequilibrated for another 5 min.

Standard top15 DDA methods were generated with the graphical Thermo Xcalibur method editor. Full MS scans in the mass range from *m/z* 300 to 1650 were acquired with a 128 ms transient time corresponding to a resolution of 60,000 at *m/z* 200. The target value for the automatic gain control (AGC) algorithm was set to 3×10^6 charges, which was typically reached within about 1 ms during the elution of peptides. Precursor ions for MS2 scans were isolated with a ±0.7 Th window centered on the precursor mass and fragmented with higher energy collisional dissociation (HCD) (34) at a normalized collision energy (NCE) of 27. MS2 spectra were acquired with a resolution of 15,000 at *m/z* 200, and the maximum ion injection time and the AGC target were set to 25 ms and 1×10^5 charges, respectively. Only precursors with assigned charge states $>=2$ and $<=5$ were considered, and previously sequenced precursors were dynamically excluded for 30 s.

Acquisition Software—MaxQuant.Live (Version 0.1) was continuously running in its "listening mode" on the acquisition computer waiting for a signal to load and execute a scan protocol from the library. To schedule batches of LC-MS runs, we used the sequence list from Xcalibur whose entries contain settings for the LC device as

[1] The abbreviations used are: DDA, data-dependent acquisition; GUI, graphical user interface; IAPI, instrument application programing interface; NCE, normalized collision energy; SIM, selected ion monitoring; pmSIM, predictive multiplexed selective ion monitoring.

well as the method for the mass spectrometer. While using the normal LC-settings we constructed the instrument method in such a way that it encodes the start signal for MaxQuant.Live to load a given scan protocol and take over the control of the mass spectrometer for the whole run on starting.

Scan protocols for the different targeting strategies were all specified using the targeting app that is included in MaxQuant.Live. As initial settings for peptide recognition, we chose by default a mass tolerance of ±10 ppm, a retention time tolerance of ±3 min and an intensity threshold value of 10^{-5} from the expected intensity of the target. To calculate the corrections, the adaptive correction includes the peptides recognized within the last 3 min but retains a minimum of the last 100. The correction automatically started 6 min after the first peptide was recognized and the mass tolerances were set to 4.5 ppm.

Breakdown Curves—As an example of advanced acquisition schemes enabled by MaxQuant.Live, we studied the large-scale and automated acquisition of HCD fragmentation characteristics of peptides. In triplicate runs of a HeLa digest, we targeted 1000 precursors, using 10,000 endogenous peptides for real-time corrections. On recognition in real-time, precursors were isolated with a ±0.2 Th window and repeatedly fragmented with increasing collision energy in ten steps from NCE 18 to 36. Other than that, the MS parameters were set as above. Target peptides were selected randomly from the top 50% abundance quantile of peptides identified in a standard DDA run of the same digest after removal of contaminants and reverse hits and filtering for the most abundant evidence of each unique peptide sequence. The tolerances for the real-time correction were the default values listed before.

Predictive Multiplexed Selective Ion Monitoring (pmSIM)—Light and heavy labeled tryptic Hela lysates were mixed in a ratio of 4:1 and 500 ng were injected on column. DDA raw files were analyzed using MaxQuant to identify light to heavy SILAC peptide ratios. To generate a targeting list for MaxQuant.Live, the "evidence" output file was filtered for modified sequence duplicates, missed cleavages, keeping only unmodified peptides with a sequence length less than 25 amino acids, a retention length less than 2 min, no modifications and a charge state of 2. Peptides for retention correction were additionally filtered for >10 and < 80 min retention time, after which the top 5,000 most intense light channel peptides were selected. The fifty peptides for selected ion monitoring (SIM) were randomly chosen from a list fulfilling the following criteria: retention time 20–70 min, no reported L/H ratio, an intensity of zero in the heavy channel. The initial retention time tolerance was ±10 min and the final value was 1.5-fold of the elution time standard deviation.

MaxQuant.Live pmSIM experiments were performed with a 1 Th isolation window and a +0.2 Th offset and acquired with a resolution of 120,000 at *m/z* 200. The heavy and light channels were multiplexed in a single scan. A maximum of 1 × 10^5 ions were collected in each channel with a maximum ion injection time of 48 and 192 ms for the light and heavy channel, respectively.

Data analysis of the pmSIM experiment was performed with the Skyline (35) (Version 4.1.0.18169) and XCalibur (3.1.66.10) software suites. The SIM targeting raw file was split into SIM and MS1 scans and analyzed independently.

Large-scale Targeting—To build a reference DDA dataset, 500 ng of tryptic Hela digest were measured in triplicate and raw files were analyzed with MaxQuant. The matching between runs feature was activated using the default settings. Peptide identities as well as their mass, charge state, retention time and intensity were extracted from the evidence output file and used to generate targeting lists for MaxQuant.Live. Only peptides with a retention time between 10 and 100 min that were identified by MS/MS or matching in all three replicates were candidates for the targeting lists and any hits from the reverse decoy library and potential contaminants were excluded from the selection. To generate the targeting lists, 100, 1000, 5000, 10,000, 20,000 or 30,000 peptides were randomly selected from all peptides fulfilling the above criteria, ensuring a uniform distribution of targets over the whole abundance range. For real time correction, we selected the 10,000 most abundant peptides identified by MS/MS or matching in all three replicates with a retention length less than 30 s. The tolerances for real-time correction were the default values listed above. To demonstrate the functionality of the real-time correction we performed an additional run with 20,000 targeting peptides, in which the minimal mass tolerance was 4.5 ppm and the retention time windows size was 2.5-fold of the standard deviation of the peptide elution times. Here, the correction was calculated from all the peptides that were recognized within the last minute but at least the last 20.

Full MS scans in the mass range from *m/z* 300 to 1650 were acquired with a resolution of 60,000 at *m/z* 200 and an automatic gain control (AGC) target of 3 × 10^6. Target peptides had a "Life Time" (max. time between recognition and fragmentation) of 1000 ms and were isolated for MS2 scans with a +- 0.2 Th window centered on the precursor mass and fragmented with a NCE of 27. MS2 spectra were acquired with a resolution of 15,000 at *m/z* 200, and the maximum ion injection time and the AGC target were 110 ms and 1 × 10^5 charges, respectively. Raw files of the targeting runs were analyzed in MaxQuant together with the standard DDA runs using the matching between runs feature. Coefficients of variation (CVs) were calculated for the targeted peptides between replicate measurements of standard DDA and targeting runs. To normalize for variations in total sample between injections, intensities were median normalized before calculation of CVs.

Proteomics Data Processing—Shotgun proteomics raw data acquired with either the standard user interface or MaxQuant.Live were processed with MaxQuant (24) (version 1.6.1.13) using the default settings if not stated otherwise. The built-in Andromeda search engine (25) scored MS2 spectra against fragment masses of tryptic peptides derived from a human reference proteome containing 95,057 entries including isoforms (UniProt, release 2018/06) and a list of 245 potential contaminants. We required a minimum peptide length of 7 amino acids and limited the search space to a maximum peptide mass of 4600 Da and two missed cleavage sites. Carbamidomethylation of cysteine was specified as a fixed modification, and methionine oxidation and acetylation at protein N termini as variable modifications. MaxQuant uses individual mass tolerances for each peptide, whereas the initial maximum precursor mass tolerances were set to 20 ppm in the first search and 4.5 ppm in the main search, and the fragment mass tolerance was set to 20 ppm. The false discovery rate was controlled with a target-decoy approach at less than 1% for peptide spectrum matches and less than 1% for protein group identifications.

Bioinformatics—Post-processing was performed in either Perseus (36), the R computational environment (37) or the Python programming language. Potential contaminants, reverse database hits and proteins identified by only one modified peptide were excluded from the analysis.

Experimental Design and Statistical Rationale—In this study, we developed the MaxQuant.Live software for MS data acquisition, which enables classical acquisition schemes as well as methods that are more elaborate. We evaluated the technical feasibility of large-scale peptide recognition. Sample sizes were chosen to allow assessing technical variations with replicate LC-MS injections of aliquots from the same sample preparation batch. The n numbers were 1, 1, 3, 3, 3 for the experiments in Figs. 2, 3, 4, 5, and 6, respectively. Statistical testing, control samples and blinding were not applicable, and no data points were excluded. No targeting experiments em-

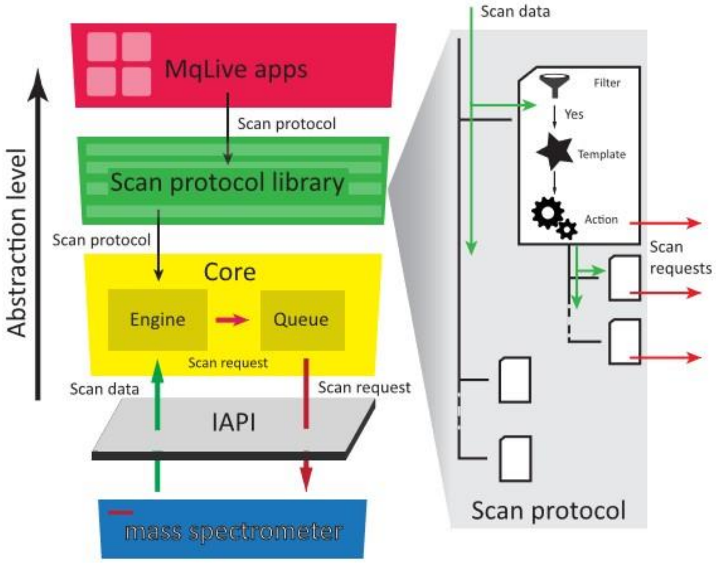


FIG. 1. **Architecture of MaxQuant.Live and the logic of its scan protocols.** The core of our software (yellow box) handles the real-time control of the mass spectrometer using the IAPI by Thermo Fisher. Its engine processes the scan data according to a scan protocol, which specifies a data-acquisition strategy trough a decision tree logic (right). Scan protocols for different applications are collected in a library and can be generated by small applications ('apps').

ployed internal standards and are therefore classified as "Tier 3" according to the MCP guidelines. Target peptides were selected randomly as detailed above without considering modification states or uniqueness to proteoforms.

RESULTS

Here we describe the development of a software framework termed MaxQuant.Live for real-time monitoring of mass spectrometric data and controlling of the data acquisition. We demonstrate its usability and performance in terms of scanning speed using a reimplemented topN method. We demonstrate that thousands of peptides of interest can be detected and immediately selected for deeper analysis, greatly extending the toolbox for targeted proteomics. To explore the current limits of our technology, we targeted over 25,000 peptides in a single experiment.

Design and Functionality of MaxQuant.Live—A few years ago, Thermo Fisher Scientific developed an IAPI that enables fast, bidirectional communication between a Q-Exactive mass spectrometer and an outside. We developed a software module, written in the C# programming language, containing functionality for advanced data acquisition and analysis in real-time, which communicates with the mass spectrometer through the IAPI (Experimental Procedures). We termed the program MaxQuant.Live because it forms a bridge between intelligent data acquisition and downstream analysis in the MaxQuant environment. In one direction, the IAPI transmits every measured mass spectrum to our software on the fly and in the other direction it enables MaxQuant.Live to send scan commands to the instrument every time it is ready to accept new instructions. Fig. 1 illustrates the interplay between the core module of MaxQuant.Live and the mass spectrometer enabled by the IAPI. The engine in MaxQuant.Live executes a run-specific scan protocol (see below) which contains the acquisition strategy for the current LC-MS run and which is loaded from the scan protocol library. The scan requests generated by the engine are stored in the local scan queue before they are pushed sequentially to the MS instrument.

In case the scan queue is empty it periodically sends fallback scan requests to prevent the instrument from running idle or changing its operation status. This design of the scan queue ensures that the core module of MaxQuant.Live keeps control of the instrument during the entire run once a scan protocol has been loaded from the library and while the instrument is connected to the IAPI. Because of this generic additional abstraction layer, our core module is independent of the attached IAPI and could also be combined with instrument control libraries of other mass spectrometers.

The scan protocol specifies the acquisition concept for an LC-MS run. It implements an abstract logic (right panel in Fig. 1) which makes use of a decision tree, a common construct in computer science that has previously been applied in proteomics to select optimal fragmentation modes by Coon and co-workers (39). The decision tree simplifies the development of new acquisition strategies and generates a cascade of scan requests on the basis of the incoming scan containing the mass spectrum and the associated metadata. Every node of the scan protocol tree consists of three components: A filter, a scan template and data-dependent actions. The filter checks if meta- and spectral data of the incoming scan match particular features. If the check is negative, processing of this node and its children is stopped and the scan protocol tree proceeds to the next item. If the check is positive, for instance because the incoming scan is of type MS1 or contains specific ions of interest, then a new tailored scan request is created based on properties defined in the scan template. This comprises settings for the quadrupole, the collision cell and the mass analyzer. The third component, the data-dependent action, then establishes the connection between the incoming scan data and the settings of the next scan request. Based on its stored data and the incoming data, it chooses particular actions, such as selecting a particular precursor for isolation in the quadrupole, followed by acquisition of a fragmentation spectrum at a particular energy. Only the values that are different from the default template are overwritten. In the simple example of a topN method, the data-dependent action would be restricted to setting the position of the isolation window. After the incoming scan has successfully run through a scan protocol node, it is passed to its children, which may implement additional logic by themselves and trigger further scan requests.

Although scan protocols allow an easy and flexible way to develop acquisition strategies on a high abstraction level, using them is complex and difficult for a nonspecialized mass spectrometry laboratory. For this reason, MaxQuant.Live includes a series of small programs (apps) that can automatically generate scan protocols based on predefined acquisition strategies. We have created an app store for MaxQuant.Live that allows easy access to a collection of apps for different data acquisition strategies, which we have developed and tested in our group. In addition to the strategies described in this publication, BoxCar (40) acquisitions and support for the EASI-tag method (41) are already included.

Usability and Performance of the Software Package—Our ambition was to make MaxQuant.Live very robust and fast, so that any mass spectrometry laboratory can use it for their workflows, without affecting ease of use or throughput. We further aimed to make it universally available and supportable in the long term, like the other parts of the MaxQuant ecosystem.

The graphical user interface (GUI) of MaxQuant.Live unifies the control over all our software components in a user-friendly way, starting from the instrument connection to the scan protocol library and the apps for creation of new acquisition schemes (Fig. 2*A*). The user can start a scan protocol from the library which then triggers MaxQuant.Live to take control of the mass spectrometer until the end of the run where it switches back to idle mode. It does not interfere with the vendor's software and it can be continuously active in "listening mode." In this way, acquisition can seamlessly switch between Xcalibur and MaxQuant.Live.

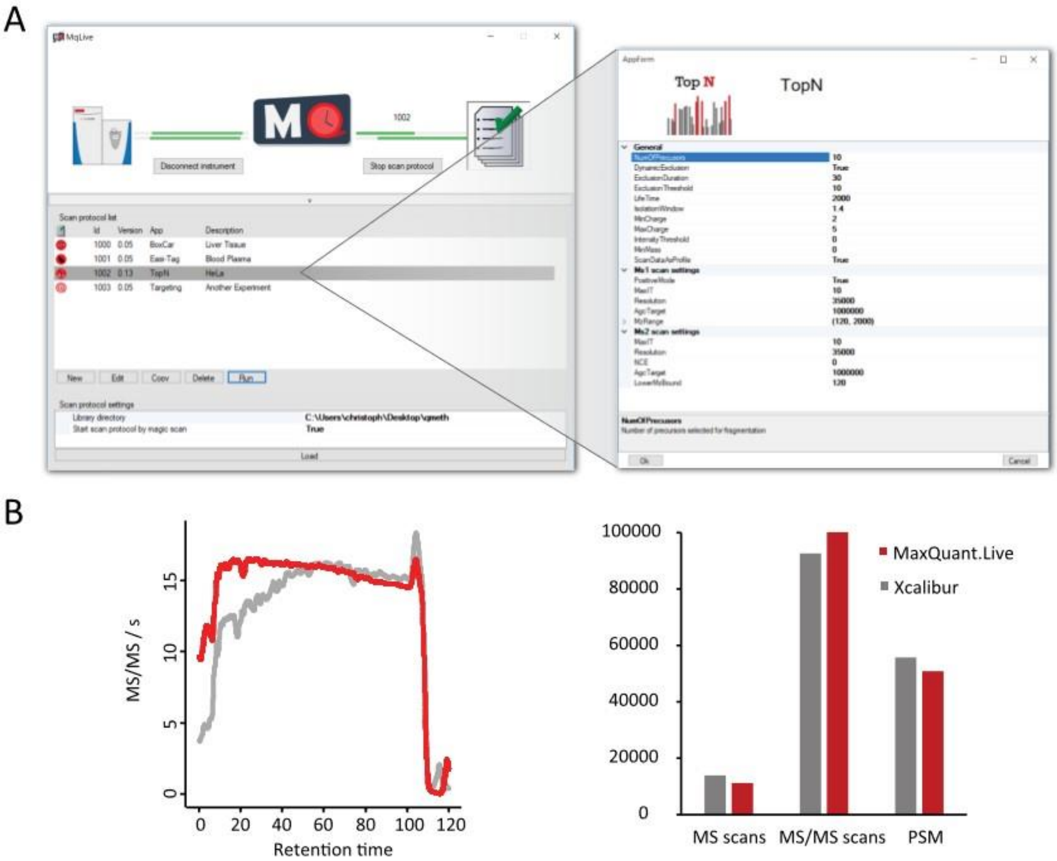
The user initially selects the app from the built-in app store for the desired workflow. MaxQuant.Live allows the creation of new scan protocols without knowledge of the underlying decision tree structure using simple GUIs that break down the complexity of each scan protocol into a small number of required settings. Fig. 2*A* illustrates this for the GUI of the topN app, which reimplements the standard data-dependent acquisition scheme as a benchmark example. After the user specifies the parameters, the app generates the corresponding scan protocol and adds it to the scan protocol library. The GUI also allows editing a scan protocol within the app at later time points to modify the acquisition strategy.

In our topN implementation, the peak selection can be restricted to specific charge states and intensity values/ranges to focus fragmentation on preferred classes. As in Xcalibur, resequencing of precursors can be prevented. Additionally, the relevant instrument parameters for the MS1 survey as well as the MS2 fragmentation scans can be specified in the GUI.

To benchmark the acquisition speed of a mass spectrometer under the control of MaxQuant.Live, we performed standard HeLa LC-MS/MS runs using our implementation of the top15 method. MaxQuant.Live achieved at least as many MS2 scans per second over the full 120 min gradient as the vendor's software (Fig. 2*B*). (The faster speed at the beginning of the gradient is likely because the Xcalibur peak selection algorithm uses a different intensity threshold.) This indicates that both MaxQuant.Live and the IAPI are extremely fast, and do not impose any relevant overhead in acquisition compared with direct control by Xcalibur. In particular, the total number of MS2 scans and peptide spectrum matches (PSMs) is not compromised, creating a solid basis for more intelligent acquisition schemes.

Three-dimensional Adaptive Control for Peptide Recognition in Real-time—Having established a fast and robust framework for data-dependent acquisition, we next set out to accurately recognize eluting peptides in real-time at a very large scale. This is challenging because hundreds of precursor ions elute at any given time in complex proteome analysis and the elution time for every peptide subtly shifts from run to run. As a result, existing "inclusion lists" and "exclusion lists" are in practice limited to a relatively small number of precursors. In contrast, the ability to detect large numbers of peptides should enable MaxQuant.Live to take data-dependent decisions about the next scan operations in real-time and thereby to realize more intelligent acquisition strategies.

MaxQuant.Live includes a powerful app that implements diverse strategies to target specific precursor ions in an LC-MS run. They build on a real-time feature detection algorithm combined with adaptive nonlinear corrections in the retention time, *m/z* and intensity dimensions.



MqLive

Scan protocol list

Id	Version	App	Description
1000	0.05	BoxCar	Liver Tissue
1001	0.05	Easi-Tag	Blood Plasma
1002	0.13	TopN	HeLa
1003	0.05	Targeting	Another Experiment

Scan protocol settings

Library directory	C:\Users\christoph\Desktop\gmeth
Start scan protocol by magic scan	True

AppForm

TopN

General	
NumOfPrecursors	10
DynamicExclusion	True
ExclusionDuration	30
ExclusionThreshold	10
LifeTime	2000
IsolationWindow	1.4
MinCharge	2
MaxCharge	5
IntensityThreshold	0
MinMass	0
ScanDataAsProfile	True
Ms1 scan settings	
PositiveMode	True
MaxIT	10
Resolution	35000
AgcTarget	1000000
MzRange	(120, 2000)
Ms2 scan settings	
MaxIT	10
Resolution	35000
NCE	0
AgcTarget	1000000
LowerMzBound	120

NumOfPrecursors
Number of precursors selected for fragmentation

FIG. 2. **Ease of use and acquisition speed of MaxQuant.Live.** *A*, Graphical user interface containing functionalities to create and manage scan protocols. *B*, Benchmarking the acquisition speeds of MaxQuant.Live *versus* the vendor's software (Xcalibur). In both top15 implementations, the instrument was acquiring MS2 spectra at nearly the maximum rate throughout the run (left panel). The number of MS and MS2 scans as well as the peptide spectrum matches (PSM) are comparable (right panel).

In the list of centroid *m/z* values received from the instrument, our software determines isotope patterns, which are then compared with a list of precursor ions of know mass, charge state, intensity and estimated retention time. For a potential match, the ion intensity must first exceed a threshold, which is a user-defined percentage of the expected intensity. This is calculated from the first two isotopic peaks, which are assumed to conform to the averagine-model (42), as is the case in MaxQuant post-processing, and which must have the expected mass-to-charge difference within a user-defined tolerance of several ppm. MaxQuant.Live sets the recorded ion intensity to zero if either of the two peaks are missing or if there is an interfering peak before the presumed monoisotopic peak.

The second condition of the recognition algorithm requires that the precursor elutes within a certain time window around its expected retention time. Depending on a variety of external factors, peptide elution times can shift by several minutes between any two LC-MS runs, with the consequence that "retention time windows" are generally set to several minutes. For very large numbers of targeted precursors, this would lead to too many potential matches to eluting features. To tackle this problem, we extended the recognition algorithm by an adaptive nonlinear correction of the observed retention time shifts that is inspired by the "match between runs" approach of MaxQuant (27). Briefly, we use a subset of easily recognizable peptides to continuously minimize the median differences between the observed and the expected retention times. Because of applying dynamic corrections we can dramatically shrink the tolerances for the elution time values that are used in the ion recognition. In a typical run, the interval containing 95% of the expected precursors shrinks from sev-

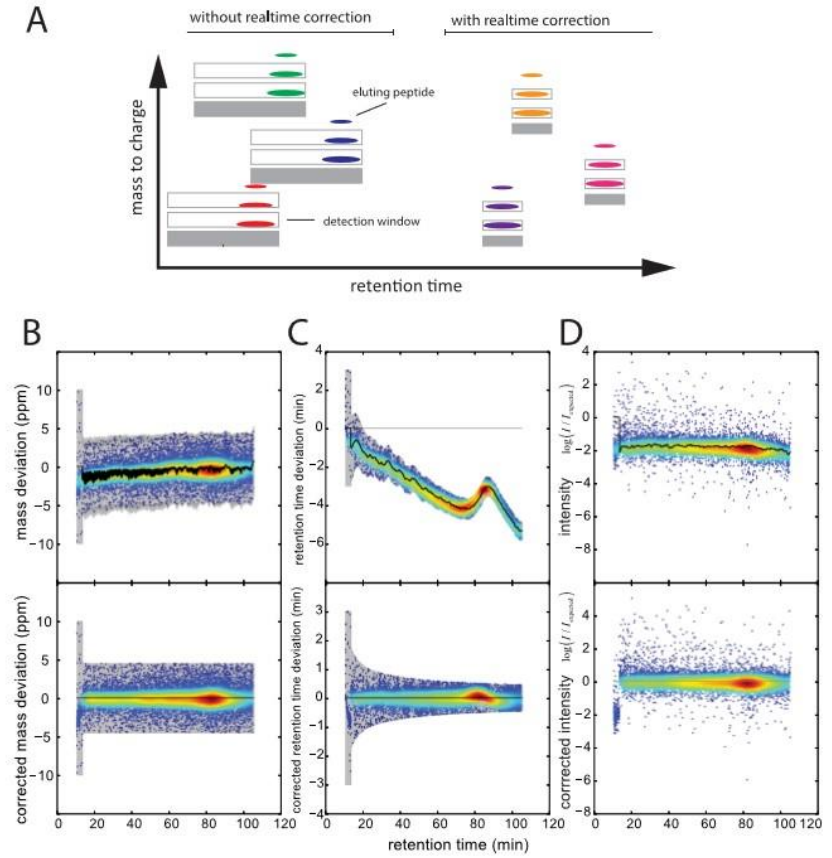


FIG. 3. **MaxQuant.Live targeting application.** *A*, Real-time peptide recognition expects the first two isotopic peaks within a dynamic retention time and mass-to-charge tolerance window as indicated by the gray boxes. Our adaptive correction approach continuously corrects observed global shifts of the elution time, mass calibration and peptide intensity and reduces the tolerances to minimum values. *B–D*, application of the dynamic global corrections (black lines) during an LC-MS run (upper row) drastically narrows the recognition algorithm tolerances (gray areas in *B* and *C*) and the scaling of the peptide intensities to the values observed in the reference run.

eral minutes to less than 1 min (Fig. 3*A*). Although similar "dynamic corrections" have been applied by us and others before, MaxQuant.Live achieves high robustness and precision by using a very large number of peptide precursors for real time correction (up to thousands).

Like the retention time alignment, mass accuracy can be greatly improved with the help of subsets of peptides that serve as internal calibrants (43). Based on the same principle as above, we therefore continuously recalibrate the mass scale, achieving severalfold improvements in real time mass accuracy. However, in contrast to MaxQuant, our mass correction applies to each entire spectrum, rather than being peptide specific. For the example in Fig. 3*B*, the a priori mass window could be reduced from a maximum mass deviation ±10 ppm to ±4.5 ppm which is the same maximum value as used in MaxQuant post-processing.

Signal intensity is the third dimension of precursor features and its adaptive control accounts for day to day differences in sensitivity of the LC-MS set up. Given the signals of the reference peptide population, an overall scaling factor is applied to make the recorded signal intensities comparable to the ones in the targeting list generated from a reference run. In our experiments we noticed that this scaling factor varied between different runs, for example as a result of varying sample amounts on column, but only little within a single run (Fig. 3*C*).

Targeted Acquisition of Breakdown Curves—Robust and precise peptide recognition in real-time should open various

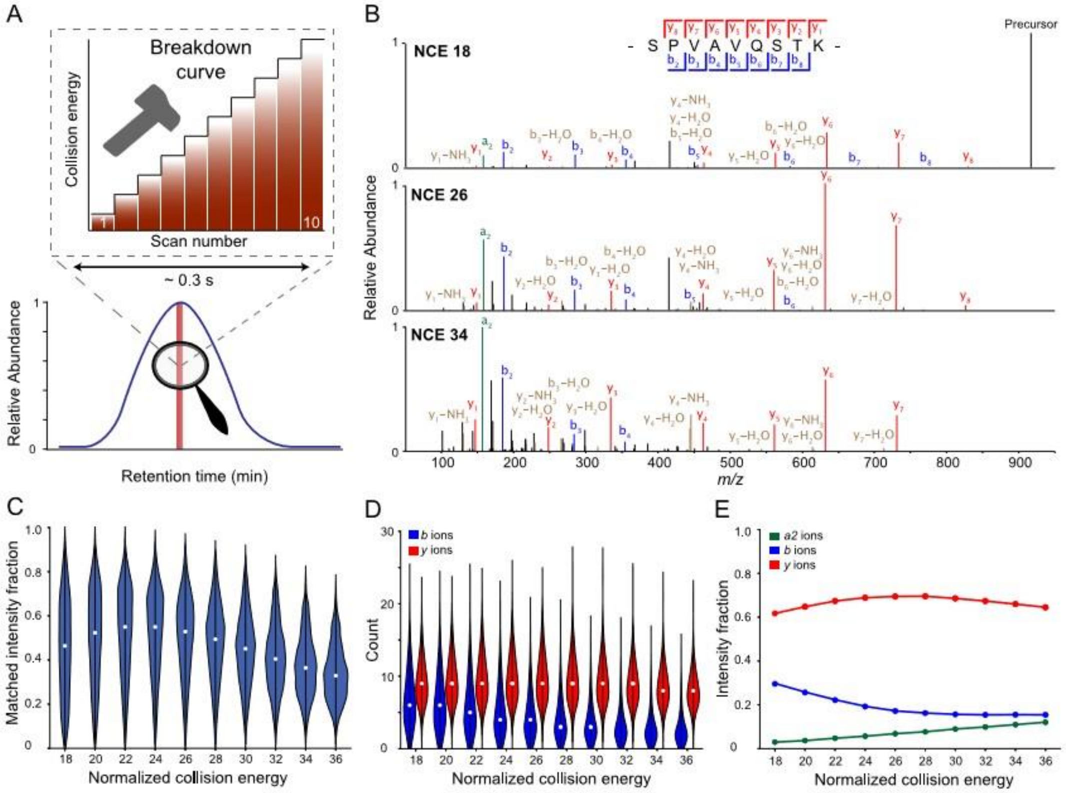


FIG. 4. **Automated acquisition of peptide breakdown curves.** *A*, Extracted ion chromatogram of a targeted peptide. On detection, MaxQuant.Live acquires repeated MS2 scans of this precursor with increasing collision energies. *B*, Exemplary spectra from a single breakdown curve. *C*, Fraction of the total MS2 ion current annotated by the Andromeda search engine as a function of the normalized collision energy. *D*, Number of identified b and y ions as a function of the normalized collision energy. *E*, Median summed intensity of a2, b, and y ions relative to the sum of all identified fragment ions. n = 962 peptides.

opportunities for advanced analysis of selected peptides. To demonstrate this, we chose to generate "breakdown curves," which are useful to determine optimal collision energies of peptides or to determine the structure of metabolites. We directed MaxQuant.Live to detect a subset of 1000 peptides in a complex HeLa background and fragment each of them with increasing collision energies. Using 10,000 abundant background peptides for our adaptive real-time correction, the monitoring time for each of the peptides of interest was reduced to less than 4 min in the 120 min runs. Notably, the median absolute retention time deviation was only 0.2 min after recalibration in all three replicates (supplemental Fig. S1). Together with the sub-ppm mass accuracy, this allowed us to successfully acquire breakdown curves for 962 of the 1000 targeted peptides. Fig. 4*A* illustrates the method for a specific target peptide (SPVAVQSTK). We used ten different collision energies from NCE 18 to 36 at a mass resolution of 15,000 at *m/z* 200, which translates into a net analysis time of only 0.3 s per breakdown curve. Three example spectra for low, middle and high collision energies are annotated in Fig. 4*B*. At NCE 18, the spectrum was dominated by the precursor ion, indicating incomplete fragmentation. Despite the relatively low abundance of fragment ions, we observed the complete y ion series (y_1–y_8) as well as the complementary b_2 to b_8 ion series. At NCE 26, the precursor ion was completely fragmented. Increasing the NCE further yielded low-mass immonium ions and many internal fragment ions, which escaped automatic scoring with the Andromeda search engine (44). The possibility to target thousands of peptides enables global analysis of peptide fragmentation. To illustrate uses of this capability, we plotted the fraction of the fragment ion current that has been identified as a function of the collision energy (Fig. 4*C*). This value peaked at NCE 22–24, presumably because of the less frequent generation of internal fragment ions. Generally, we noted a wide distribution for the peptide specific optimal NCEs, highlighting sequence-dependent differences in the fragmentation efficiency even with the normalized collision energies. Next, we investigated the energy-dependent generation of b and y ions (Fig. 4*D*). B ions were preferably generated at lower collision energies, whereas the number of annotated y ions increased with higher collision energy. Over 60% of the annotated ion current was accounted for by y ions throughout all NCEs, while the relative abundance of b ions was decreasing (Fig. 4*E*). Interestingly, the fraction of the a2 fragment ion in the characteristic a2 -b2 ion pair that is formed instead of the b1 ion, increased up to 15% of the annotated fragment ion current at higher collision energies.

Predictive Multiplexed Selective Ion Monitoring (pmSIM)—In the example above, the MS1 signal of a targeted peptide triggered the acquisition of MS2 scans. However, MS1 spectra can be incomplete in that low abundance precursors may be present in some but not other runs. Thus, instead of relying on the MS1 trigger signal and motivated by the high accuracy of the real-time retention time alignment described above, we next predicted the elution of target peptides based on the endogenous background population (Fig. 5*A*).

To demonstrate our approach, we set up a SILAC (30) experiment in which heavy and light whole-cell HeLa digests were mixed in a 1:4 ratio. In DDA, the limited dynamic range of the full scan resulted in many missing heavy-to-light ratios in the low intensity range (Fig. 5*B*) and overall, MaxQuant reported ratios for only ~60% of the identified peptides. The sensitivity can be boosted dramatically by isolating and selectively accumulating narrow *m/z* ranges, which results in improved MS1 (SIM) or, when fragmented, MS2 quantification (termed parallel reaction monitoring "PRM," when used in targeting studies (15)). Without adaptive retention time alignment, such scans must be repeated over a time range large enough to account for the typical shifts and fluctuations in the elution times. This typically results in a very large overhead of scans for each targeted peptide, limiting the total number that can be studied in a single LC-MS run. Here, we used a "predictive multiplexed SIM" (pmSIM) method to measure heavy and light SILAC peptides simultaneously.

We selected 5000 high abundant peptides as correction peptides for the adaptive real-time corrections (Fig. 5*A*). The correction algorithm of MaxQuant.Live dynamically centered the observation time ranges around the peptide elution times (Fig. 5*C*, colored lines and circles, respectively), yielded an accurate prediction of the time range in which each peptide was expected to elute. This resulted in two times smaller window sizes compared with the initial values. The comparison of the window sizes with the deviations of the peptide apex times from the predictions (Fig. 5*C*, histogram) shows that the time windows could have been chosen smaller. It should be noted that our settings were very conservative and the number of target peptides could be much higher.

To validate the accuracy of our prediction algorithm, we selected 50 peptides from the low abundance range with missing ratios from our SILAC HeLa study (Fig. 5*B*). We then used the MaxQuant.Live targeting app to specify an acquisition method that executed SIM scans of the corresponding ion pairs repeatedly over the expected elution time range. The pmSIM strategy correctly quantified the ratios for the targeted peptides close to the expected value of 4:1 with a median CV of 8.2% (Fig. 5*D*). This is notable, because none of the heavy labeled peptides was quantified at the MS1 level before. The example in Fig. 5*E* shows the increase in sensitivity by comparing the MS1 with the corresponding SIM scan. In the SIM scan, the injection time for the previously unrecorded heavy peptides is 400 times larger than the injection time of the full scan, drastically improving the quantitative accuracy.

Highly Efficient Proteome Quantification—The examples shown so far demonstrate the ability of MaxQuant.Live to perform a specific and sophisticated analysis of a limited number of peptides of interest. The fact that the underlying peptide recognition algorithm can in principle deal with an unlimited number of peptides, makes applications feasible that target a substantial proportion of the total set of precursor ions. We reasoned that this generically boosts the reproducibility of peptide fragmentation events between LC-MS runs compared with the topN method.

We implemented our strategy using the targeting app of MaxQuant.Live and generated sets with different numbers of targeted peptides, which were randomly selected from triplicate MS analysis of tryptic HeLa lysates using a standard top15 method in Xcalibur (Fig. 6*A*). For every set of peptides, we performed triplicate LC-MS runs in which MS2 scans were triggered if one of the peptides was recognized by our algorithm in the MS1 scans. The number of peptides that were fragmented and correctly identified afterward by MaxQuant is shown in Fig. 6*B* for all six sets of peptides. Although nearly every targeted peptide was hit in at least one of the runs (green line), this fraction decreases for those hit in at least two (blue line) and all three runs (red line), respectively. The quantification precision was comparable to standard DDA runs with coefficients of variation between the triplicate measurements ranging from 10.4%–12.5% (supplemental Fig. S2).

These results indicate that our strategy can target very large numbers of peptides even though some stochasticity remains between the acquisitions in the different runs. This is likely because of some peptides not being recognized by our algorithm at the MS1 level. An analysis of the of the initial topN and the targeting raw data files using MaxQuant with the matching between runs feature showed very similar results. Thus, the well-established feature detection of MaxQuant could not find significantly more spectral features at the MS1 level from the targeting list, even given full information after complete analysis. This suggests that these peptides, which were selected from the original topN runs, are not "visible" at all in the MS1 scans of the targeting runs. A comparison

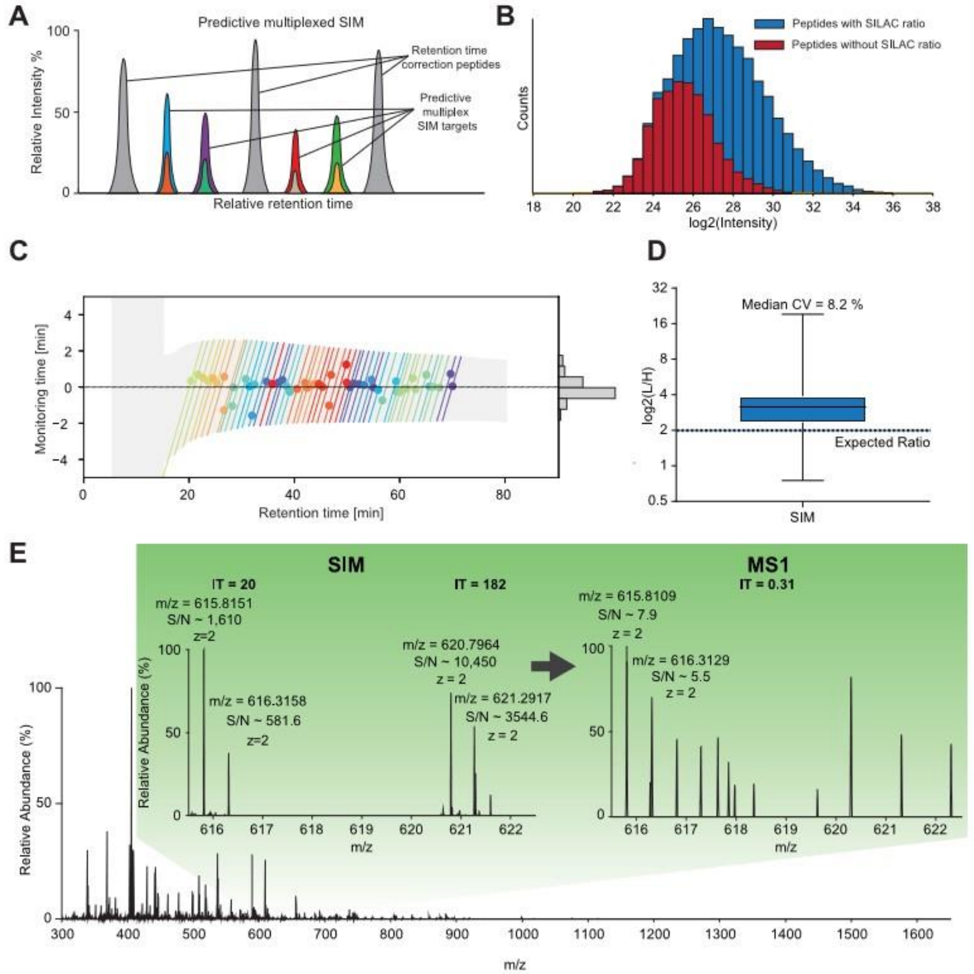


FIG. 5. **Predictive multiplexed SIM (pmSIM).** Peptides employed for live retention time correction are distributed across the gradient and are highly abundant, compared with the low abundant targeted peptides and their SILAC partner needed for quantification (*A*). *B*, Total log2 intensity abundance range of all peptide identifications from the standard run with a light to heavy SILAC ratio of 4:1. In blue, all peptide identifications with identified ratios are highlighted. In red, all peptide identifications in the light channel without any reported SILAC ratios are highlighted. *C*, Watch-time of the predictive multiplexed SIM scan-mode in the targeting experiment. *D*, Resulting SILAC ratios after SIM-targeting using Skyline (35). *E*, Example of a very low abundant target peptide compared in the original MS1 and after pmSIM demonstrating ~40-fold increase of S/N.

of the intensity histograms of the peptides that were successfully targeted and identified by MS/MS in all three runs (Fig. 6*C*, blue) to the ones that were hit in less than three runs (green) shows a slight shift of the intensity distribution to lower values. We therefore suspect that the spectral noise thresholding employed by Xcalibur set the corresponding signals to zero even though they only slightly dipped below acceptance criteria, something we have also noticed when boosting the dynamic range in the BoxCar acquisition scheme (40). A possible solution to tackle this problem of the acquisition side would be to boost the peptide intensities by using BoxCar scans for the peptide recognition instead of the MS1 scans.

To counter the effect of peptide features missing because of thresholding in our analysis and make the data between the two triplicate measurements comparable, we further normal-

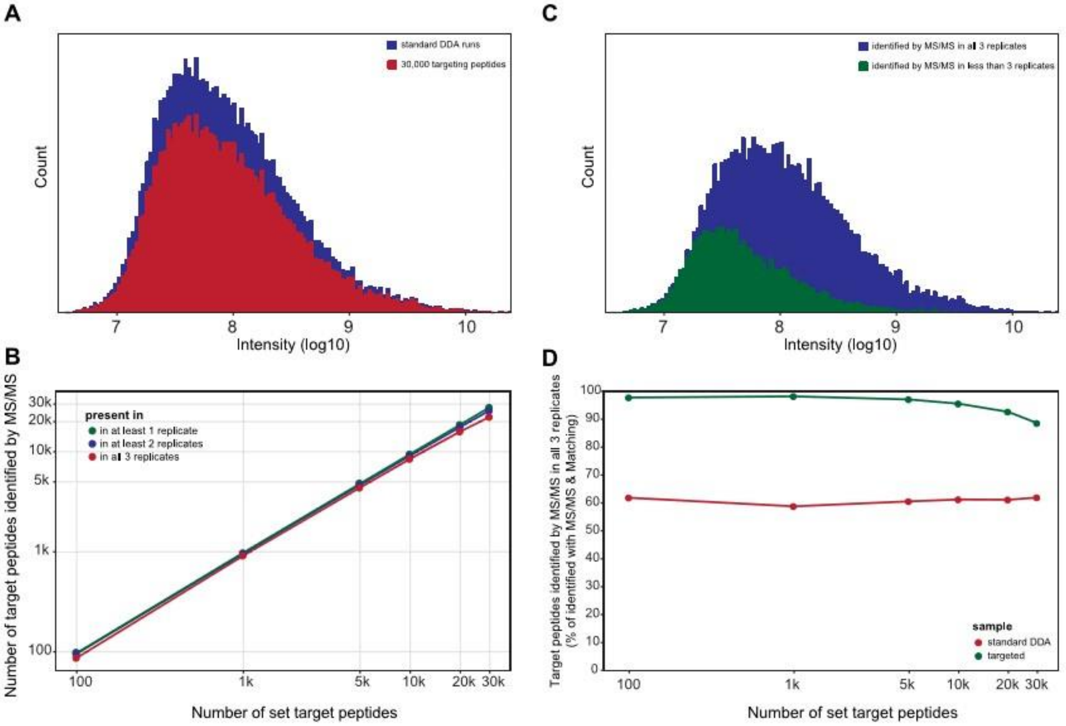


FIG. 6. **Reproducible identification of over 20,000 peptides of interest.** Tryptic HeLa lysates were analyzed by either a standard DDA method designed in the Xcalibur method editor or by the MaxQuant.Live targeting method with 100, 1000, 5000, 10,000, 20,000 or 30,000 targeted peptides. Triplicate injections were performed and sequenced peptides were identified by MaxQuant with or without the matching between runs function activated. *A*, Selected targets were uniformly distributed over the whole intensity range of peptides identified in previous triplicate standard DDA runs. *B*, The number of targets correctly identified by MS/MS. *C*, Intensity distribution of the correctly identified sequences in MaxQuant.Live runs with 30,000 targets. *D*, Percentage of correctly identified targets by MS/MS for the MaxQuant.Live method compared with the identifications by MS/MS in the standard topN method and matching between runs in all three triplicate injections.

ized the number of peptides that were successfully targeted in all triplicate measurements to the number of targeted peptides identified by the matching between runs feature (Fig. 6*D*). The standard top15 method reached about 61% success rate, regardless of the number of peptide precursors. In contrast, in our targeting runs, the normalized percentage of successfully targeted peptides (green line) was 95% and only slightly lower for >20,000 targeted peptides. This effect is presumably related to the fact that the number of co-eluting peptides increases beyond what can be fragmented sequentially by the mass spectrometer in the time available. Although this is not a conceptual limitation of our large-scale targeting approach, it is an opportunity to be addressed by instrument improvements.

DISCUSSION AND CONCLUSION

In bottom-up proteomics, data dependent acquisition and targeted approaches have co-existed for many years. DDA has been and remains the method of choice for initial characterization of proteomes under study. Conversely, there are many applications, where only a restricted number of peptides is of interest, but these need to be measured consistently over many samples. Although both approaches have become more powerful with the general advances in instrumentation and proteomics technologies, DDA is still not powerful enough to subsume targeted analyses. Conversely, targeted methods have been difficult to establish in a robust manner, enough for clinical use, for instance, especially when monitoring more than a few dozens of peptides.

Here, we made use of the recently developed fast and robust IAPI of the Thermo instruments to interface with the acquisition process in real time. MaxQuant.Live makes use of experimental information as they are acquired to direct the acquisition in a more intelligent way. We have implemented different mass spectrometric acquisition schemes in the form of small built-in apps. We demonstrated that this workflow is highly performant as it can easily replicate the standard topN methods, for instance, without loss of quality. In targeted schemes, MaxQuant.Live continuously recalibrates the signal coming from the mass spectrometer in retention time, mass and intensity dimensions, allowing a much better prediction of the identity of eluting peptide features than possible previously. This allows, for instance, selecting any subgroup of hundreds of peptides to be targeted for accurate quantification (exemplified by our predictive multiplexed SIM method). In-depth analysis of the fragmentation patterns of large numbers of individual peptides is another valuable addition to the proteomics toolbox, which can be used to optimize precursor-fragment transitions for PRM or to pinpoint and localize modifications of low-abundance proteins. MaxQuant.Live ensures that all peptides are reliably acquired at all collision energies (as opposed to stochastic precursor selection with DDA) and in single runs. In our own group, we have already applied such a strategy to characterize the fragmentation of a novel isobaric tag (41). Even if these methods take much longer than standard fragmentation for the selected peptides, they still do not substantially contribute to overall measuring time. This means that sophisticated measurements could be done on peptides of interest, while still recording the overall proteome. Although we randomly chose peptides, one could, for instance, select a specific class of post-translation modifications, peptides that distinguish between isoforms or any other highly informative class of interest. The resulting, "enhanced," data sets could also become valuable sources for imminent machine-learning approaches (20, 28).

Building on the precise recalibration in MaxQuant.Live, we demonstrate that the scale of such experiments can be readily extended to over 25,000 peptides of interest with very high reproducibility. Conceptually, MaxQuant.Live bridges the approaches of classical shotgun and targeted proteomics. On one hand, when high numbers of peptides are targeted, it resembles shotgun experiments and could be interpreted as a very thorough and much more efficient implementation of inclusion lists (45). On the other hand, when smaller target numbers are addressed, the acquisition strategy resembles classical targeting, however, with more flexibility and much greater robustness because of real-time recalibration. To express the conceptual nature of our approach we call it "global targeting," as it combines desirable aspects of classical shotgun and targeted proteomics.

To analyze such data we therefore resorted to existing tools from both approaches. In the pmSIM experiments, we manually inspected mass spectra and elution profiles of the few target peptides. When we extend this approach to hundreds or thousands of targets, we recommend using established tools for statistically sound identification of targeted peptides such as target-decoy strategies (46). Finally, in experiments with high target peptide numbers we followed a "spectrum centric approach" (47) for precursor identification, like regular shotgun proteomics experiments.

In our global targeting experiments, we observed coefficients of variation that lie in the expected range for classical shotgun experiments at very high numbers of targeted precursors and very close to classical targeting experiments, when these numbers are somewhat reduced. This means that our method unifies the two approaches not only in terms of identifications, but also allows to implicitly define the desired accuracy (within instrument capabilities) via the number of targets beforehand. The same applies to sensitivity, which is limited by the instrument's capabilities and the measurement time per peptide. Global targeting allows the operator to optimally balance target numbers and sensitivity.

We have made MaxQuant.Live freely available and hope that it will stimulate the community into exploring this exciting direction.

Acknowledgments—We thank our colleagues in the Max-Planck Institute of Biochemistry and at Thermo Fisher Scientific for discussion and help, in particular Drs. A. Kreutzmann, D. Mourad, K. Ayzikov, F. Grosse-Coosmann, S. Horning, and A. Makarov.

DATA AVAILABILITY

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (38) partner repository with the dataset identifier PXD011225.

* This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 686547, by the German Research Foundation (DFG-Gottfried Wilhelm Leibniz Prize) granted to Matthias Mann and by the Max-Planck Society for the Advancement of Science.

[S] This article contains supplemental Figures.

‖ To whom correspondence may be addressed. E-mail: mmann@biochem.mpg.de.

** To whom correspondence may be addressed. E-mail: cox@biochem.mpg.de.

Author contributions: C.W., F.M., and M.M. designed research; C.W., F.M., S.V.W., and A.-D.B. performed research; C.W., F.M., S.V.W., A.-D.B., and M.M. analyzed data; C.W., F.M., S.V.W., A.-D.B., and M.M. wrote the paper; J.C. contributed new reagents/analytic tools; J.C. supervised Dr. Christoph Wichmann.

REFERENCES

1. Lössl, P., van de Waterbeemd, M., and Heck, A. J. (2016) The diverse and expanding role of mass spectrometry in structural and molecular biology. *EMBO J.* **35,** 2634–2657
2. Larance, M., and Lamond, A. I. (2015) Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* **16,** 269–280
3. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355
4. de Godoy, L. M., Olsen, JV, Cox, J, Nielsen, ML, Hubner, NC, Fröhlich, F, Walther, TC, and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–1254
5. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., and Coon, J. J. (2014) The one hour yeast proteome. *Mol. Cell. Proteomics* **13,** 339–347
6. Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen, C. L., Nielsen, M. L., and Olsen, J. V. (2017) An optimized shotgun

strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4,** 587–599.e4

7. Altelaar, A. M., and Heck, A. J. (2012) Trends in ultrasensitive proteomics. *Curr. Opin. Chem. Biol.* **16,** 206–213
8. Röst, H. L., Malmström, L., and Aebersold, R. (2015) Reproducible quantitative proteotype data matrices for systems biology. *Mol. Biol. Cell* **26,** 3926–3931
9. Geyer, P. E., Holdt, L. M., Teupser, D., and Mann, M. (2017) Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13,** 942
10. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793
11. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., and Aebersold, R. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol.* **14,** e8126
12. Picotti, P., and Aebersold, R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9,** 555–566
13. Kondrat, R. W., McClusky, G. A., and Cooks, R. G. (1978) Multiple reaction monitoring in mass spectrometry/mass spectrometry for direct analysis of complex mixtures. *Anal. Chem.* **50,** 2017–2021
14. Yost, R. A., and Enke, C. G. (1978) Selected ion fragmentation with a tandem quadrupole mass spectrometer. *J. Am. Chem. Soc.* **100,** 2274–2275
15. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., and Coon, J. J. (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11,** 1475–1488
16. Bourmaud, A., Gallien, S., and Domon, B. (2016) Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications. *Proteomics* **16,** 2146–2159
17. Picotti, P., Rinner, O., Stallmach, R., Dautel, F., Farrah, T., Domon, B., Wenschuh, H., and Aebersold, R. (2010) High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* **7,** 43–46
18. Zauber, H., Kirchner, M., and Selbach, M. (2018) Picky: a simple online PRM and SRM method designer for targeted proteomics. *Nat. Methods* **15,** 156–157
19. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9,** 429–434
20. Zolg, D.P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H. C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., Deutsch, E. W., Aebersold, R., Moritz, R.L., Wenschuh, H., Moehring, T., Aiche, S., Huhmer, A., Reimer, U., vKuster, B. (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat Methods* **14,** 259–262
21. Kusebauch, U., Campbell, D. S., Deutsch, E. W., Chu, C. S., Spicer, D. A., Brusniak, M. Y., Slagel, J., Sun, Z., Stevens, J., Grimes, B., Shteynberg, D., Hoopmann, M. R., Blattmann, P., Ratushny, A. V., Rinner, O., Picotti, P., Carapito, C., Huang, C. Y., Kapousouz, M., Lam, H., Tran, T., Demir, E., Aitchison, J. D., Sander, C., Hood, L., Aebersold, R., and Moritz, R. L. (2016) Human SRMAtlas: A resource of targeted assays to quantify the complete human proteome. *Cell* **166,** 766–778
22. Gallien, S., Kim, S. Y., and Domon, B. (2015) Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM). *Mol. Cell. Proteomics* **14,** 1630–1644
23. Bailey, D. J., McDevitt, M. T., Westphall, M. S., Pagliarini, D. J., and Coon, J. J. (2014) Intelligent Data Acquisition Blends Targeted and Discovery Methods. *J. Proteome Res.* **13,** 2152–2161
24. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
25. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *Proteome Res J* **10,** 1794–1805
26. Graumann, J., Scheltema Ra Zhang, Y., Cox, J., and Mann, M. (2012) A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol. Cell. Proteomics* **11,** M111.013185
27. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11,** M111.014050
28. Sinitcyn, P., Rudolph, J. D., and Cox, J. (2018) Computational methods for understanding mass spectrometry-based shotgun proteomics data. *Annu. Rev. Biomed. Data Sci.* **1,** 207–234
29. Kuehn A *et al.* (2013) Customized real-time control of benchtop orbitrap MSin *Proceedings of the 61st ASMS Conference on Mass Spectrometry and Allied Topics* Poster MP377
30. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1,** 376–386
31. Ong, S. E., and Mann, M. (2006) A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1,** 2650–2660
32. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–324
33. Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hogrebe, A., Harder, A., and Olsen, J. V. (2018) Performance evaluation of the QExactive, H. F.-X for shotgun proteomics. *Proteome Res. J.* **17,** 727–738
34. Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712
35. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968
36. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–740
37. R Core Team (2008) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* ISBN 3-900051-07-0, URL http://www.R-project.org
38. Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D456
39. Swaney, D. L., Mcalister, G. C., and Coon, J. J. (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **5,** 959–964
40. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15,** 440–448
41. Virreira Winter, S., Meier, F., Wichmann, C., Cox, J., Mann, M., and Meissner, F. (2018) EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15,** 527–530
42. Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **6,** 229–233
43. Cox, J., Michalski, A., and Mann, M. (2011) Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* **22,** 1373–1380
44. Neuhauser, N., Michalski, A., Cox, J., and Mann, M. (2012) Expert system for computer-assisted annotation of MS/MS Spectra. *Mol. Cell. Proteomics* **11,** 1500–1509
45. Schmidt, A., Claassen, M., and Aebersold, R. (2009) Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr. Opin. Chem. Biol.* **13,** 510–517
46. Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Brusniak, M. Y., Hengartner, M. O., and Aebersold,. R. (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **8,** 430–435
47. Ting, Y. S., Egertson, J. D., Payne, S. H., Kim, S., MacLean, B., Käll, L., Aebersold, R., Smith, R. D., Noble W. S., and MacCoss M. J. (2015) Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol. Cell. Proteomics* **14,** 2301–2307

6.2. Article 12: Proteomics of human brown and white adipocytes

Proteomics-based comparative mapping of the secretomes of human brown and white adipocytes reveals EPDR1 as a novel batokine

Cell Metabolism, November 05, 2019

Atul S. Deshmukh[1, 2, 3, 12], Lone Peijs[3, 4, 12], Jacqueline L. Beaudry[5], Naja Z. Jespersen[4], Carsten H. Nielsen[6, 7], Tao Ma[3], **Andreas-David Brunner[1],** Therese J. Larsen[4], Rafael Bayarri-Olmos[8], Bjargav S. Prabhakar[2], Charlotte Helgstrand[9], Mai C.K. Severinsen[4], Birgitte Holst[10], Andreas Kjaer[6], Mads Tang-Christensen[9], Annika Sanfridson[9], Peter Garred[8], Gilbert G. Privé[11], Bente K. Pedersen[4], Zachary Gerhart-Hines[3], Soren Nielsen[4], Daniel J. Drucker[5], Matthias Mann[1, 2, 13, #], Camilla Scheele[3, 4, 13, 14, #]

# Correspondence

[1]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[2]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark

[3]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark

[4]The Centre of Inflammation and Metabolism and the Centre for Physical Activity Research, Rigshospitalet, University of Copenhagen, Copenhagen 2100, Denmark

[5]The Lunenfeld-Tanenbaum Research Institute, Mt. Sinai Hospital, Department of Medicine, University of Toronto, Toronto, ON M5G 1X5, Canada

[6]Department of Clinical Physiology, Nuclear Medicine and PET and Cluster for Molecular Imaging, Department of Biomedical Sciences, Rigshospitalet and University of Copenhagen, Copenhagen 2200, Denmark

[7]Minerva Imaging ApS, Copenhagen 2200, Denmark

[8]Laboratory of Molecular Medicine, Department of Clinical Immunology, Rigshospitalet, University Hospital of Copenhagen, Copenhagen 2100, Denmark

[9]Global Drug Discovery, Novo Nordisk A/S, Måløv 2760, Denmark

[10]Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark

[11]Princess Margaret Cancer Centre, Toronto, ON M5G 1L7, Canada

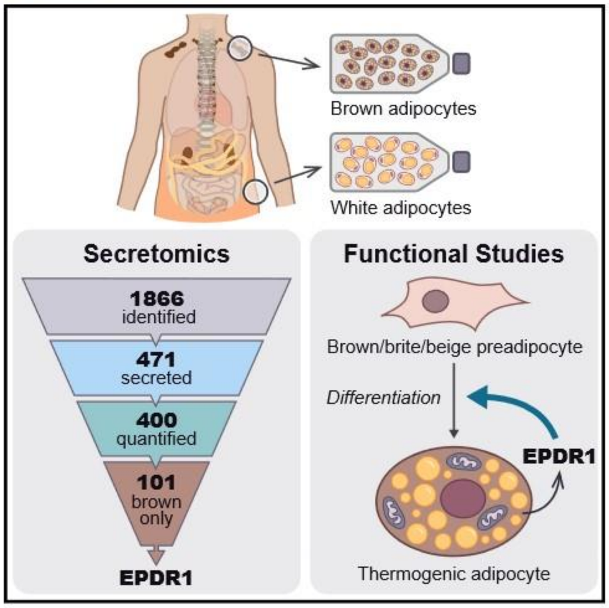
.

Short Article

Cell Metabolism

Proteomics-Based Comparative Mapping of the Secretomes of Human Brown and White Adipocytes Reveals EPDR1 as a Novel Batokine

Graphical Abstract



Authors

Atul S. Deshmukh, Lone Peijs,
Jacqueline L. Beaudry, ...,
Daniel J. Drucker, Matthias Mann,
Camilla Scheele

Correspondence

matthias.mann@cpr.ku.dk (M.M.),
cs@sund.ku.dk (C.S.)

In Brief

Deshmukh et al. describe the human brown fat secretome and identify novel candidate batokines with potential effects on human metabolism. One such batokine is EPDR1, shown here to play a role in brown fat commitment.

Highlights

- Identification of the first human brown fat secretome
- Comparative analysis of the secretomes of human brown and white adipocytes
- 101 proteins were exclusively identified in the secretome of brown adipocytes
- EPDR1 is a novel batokine important for brown fat commitment

Deshmukh et al., 2019, Cell Metabolism *30*, 963–975
November 5, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.cmet.2019.10.001

CellPress

Cell Metabolism
Short Article

Proteomics-Based Comparative Mapping of the Secretomes of Human Brown and White Adipocytes Reveals EPDR1 as a Novel Batokine

Atul S. Deshmukh,[1,2,3,12] Lone Peijs,[3,4,12] Jacqueline L. Beaudry,[5] Naja Z. Jespersen,[4] Carsten H. Nielsen,[6,7] Tao Ma,[3] Andreas D. Brunner,[1] Therese J. Larsen,[4] Rafael Bayarri-Olmos,[8] Bhargav S. Prabhakar,[2] Charlotte Helgstrand,[9] Mai C.K. Severinsen,[4] Birgitte Holst,[10] Andreas Kjaer,[6] Mads Tang-Christensen,[9] Annika Sanfridson,[9] Peter Garred,[8] Gilbert G. Privé,[11] Bente K. Pedersen,[4] Zachary Gerhart-Hines,[3] Søren Nielsen,[4] Daniel J. Drucker,[5] Matthias Mann,[1,2,13,*] and Camilla Scheele[3,4,13,14,*]

[1]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[2]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark
[3]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark
[4]The Centre of Inflammation and Metabolism and the Centre for Physical Activity Research, Rigshospitalet, University of Copenhagen, Copenhagen 2100, Denmark
[5]The Lunenfeld-Tanenbaum Research Institute, Mt. Sinai Hospital, Department of Medicine, University of Toronto, Toronto, ON M5G 1X5, Canada
[6]Department of Clinical Physiology, Nuclear Medicine and PET and Cluster for Molecular Imaging, Department of Biomedical Sciences, Rigshospitalet and University of Copenhagen, Copenhagen 2200, Denmark
[7]Minerva Imaging ApS, Copenhagen 2200, Denmark
[8]Laboratory of Molecular Medicine, Department of Clinical Immunology, Rigshospitalet, University Hospital of Copenhagen, Copenhagen 2100, Denmark
[9]Global Drug Discovery, Novo Nordisk A/S, Måløv 2760, Denmark
[10]Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark
[11]Princess Margaret Cancer Centre, Toronto, ON M5G 1L7, Canada
[12]These authors contributed equally
[13]Senior author
[14]Lead Contact
*Correspondence: matthias.mann@cpr.ku.dk (M.M.), cs@sund.ku.dk (C.S.)
https://doi.org/10.1016/j.cmet.2019.10.001

SUMMARY

Adipokines secreted from white adipose tissue play a role in metabolic crosstalk and homeostasis, whereas the brown adipose secretome is less explored. We performed high-sensitivity mass-spectrometry-based proteomics on the cell media of human adipocytes derived from the supraclavicular brown adipose and from the subcutaneous white adipose depots of adult humans. We identified 471 potentially secreted proteins covering interesting categories such as hormones, growth factors, extracellular matrix proteins, and proteins of the complement system, which were differentially regulated between brown and white adipocytes. A total of 101 proteins were exclusively quantified in brown adipocytes, and among these was ependymin-related protein 1 (EPDR1). EPDR1 was detected in human plasma, and functional studies suggested a role for EPDR1 in thermogenic determination during adipogenesis. In conclusion, we report substantial differences between the secretomes of brown and white human adipocytes and identify novel candidate batokines that can be important regulators of human metabolism.

Context and Significance

In this paper, researchers from the University of Copenhagen focused on identifying the entire spectrum (or "secretome") of proteins released by human white adipocytes and the heat-producing and energy-burning brown adipocytes. The authors found a substantial difference in the secretomes; for example, indicating a higher anti-inflammatory capacity in brown adipocytes and a higher ability of plasticity in white adipocytes. 101 proteins were exclusively quantified in the secretome of brown adipocytes. One of these proteins was EPDR1, which was found to be important for brown fat cell development. This study further provides a catalog of molecules that could be involved in regulating human metabolism, and possibly leading to the discovery of novel drug targets for obesity and its associated diseases.

INTRODUCTION

Adipose tissue is a major regulator of whole-body energy homeostasis by communication with the brain and other organs (Stern et al., 2016). Well-established mediators of adipocyte-derived crosstalk include leptin and adiponectin, which are produced and secreted by white adipose tissue (WAT) and contribute to the regulation of whole-body energy homeostasis (Ahima et al., 1996; Scherer et al., 1995). Adipokines derived from brown adipose tissue (BAT), known as batokines (Villarroya et al., 2017), are less investigated, especially in humans. BAT differs from WAT in its heat-producing capacity, providing an energy-consuming process, which is turned on or off in response to sympathetic activation (Cannon and Nedergaard, 2004). Considering the functional differences in WAT and BAT, batokines represent a poorly explored source of metabolic regulators. Some of the identified batokines have been described as having a hormonal function, enhancing BAT activity, improving glucose metabolism, or mediating browning of white fat (Lee et al., 2014; Stanford et al., 2013; Svensson et al., 2016). Batokines could also be represented by growth factors acting in an autocrine or paracrine manner by regulating BAT differentiation (Villarroya et al., 2017). Mapping of the human BAT secretome has previously been restricted by the lack of representative human BAT cell models as well as the challenges associated with measuring the secretome in cell culture media. However, the development of advanced secretomics technology as well as non-immortalized human BAT cell models has now made such experiments possible (Deshmukh et al., 2015; Jespersen et al., 2013; Meissner et al., 2013). In the current study, we investigate the secretomes of human brown and white adipocytes using high-resolution mass spectrometry (MS)-based proteomics. We mine the results for novel candidates with the potential for intercellular communication and perform follow up studies on Mammalian ependymin-related protein 1 (EPDR1), shown here to be a modulator of energy homeostasis and thermogenic commitment.

RESULTS AND DISCUSSION

The Secretome of Human Brown and White Adipocytes

Supraclavicular fat precursor cells (termed brown throughout this manuscript) were derived from five adult humans, as previously reported (Jespersen et al., 2013). The cell cultures were matched with subcutaneous fat precursor cells (termed white throughout this manuscript) with equal differentiation capacity as assessed by visual estimation of lipid droplet accumulation and expression of the adipocyte differentiation marker, fatty-acid-binding protein 4 (*FABP4*) (Figures 1A–1C). Mitochondrial uncoupling protein 1 (*UCP1*) was higher expressed in brown adipocytes, both at baseline (Figure 1D) and following stimulation with NE (Figure 1E). The robust *UCP1* expression underscores that these brown fat cell cultures represent an appropriate model for the subsequent secretome studies. We used high-resolution MS (Michalski et al., 2011) and automated computational analysis in MaxQuant (Cox and Mann, 2008) (Figure 1F), and detected 1,866 protein groups (protein entries that were distinguishable by MS of their identified peptides) in total in human brown and white adipocyte conditioned media (Figure 1G; Table S1). Of these, 471 were predicted to be secreted proteins, including 340 classically and 131 non-classically secreted proteins (Figure 1G; Table S1). Following our computational filtering for secreted proteins, the intracellular location protein category had decreased, whereas the categories related to protein secretion, such as glycoprotein, signal, and extracellular locations, were increased (Figure 1H). The intracellular protein portion was still high after the bioinformatics filtering for signal, secreted, and extracellular protein annotations. However, these proteins had "multiple annotations," and besides the intracellular annotation, they either had a signal peptide (72.5%) or had a second annotation as extracellular proteins (27.5%) (Figure 1H; Table S1). This is consistent with the idea that proteins may carry out different functions at different cellular locations (Huberts and van der Klei, 2010). Most of the secreted proteins were detected in cell culture media from both brown and white adipocytes (Figure 1G). However, vascular endothelial growth factor A (VEGF-A) was only identified in media from brown adipocytes, while leptin was exclusively identified in media from white adipocytes (Figure 1G; Table S1). This provides proof of concept as leptin is one of the most studied white fat-derived hormones (Stern et al., 2016), while VEGF-A is a growth factor previously reported as important for the development of functional BAT (Park et al., 2017; Shimizu et al., 2014). The presence of a signal peptide suggest secretion through the classical ER-Golgi pathway, yet hundreds of intracellular proteins are thought to be secreted through various non-classical pathways, including secretion from exosomes and microvesicles (Huberts and van der Klei, 2010; Nickel and Rabouille, 2009). Mapping of our dataset with Vesiclepedia (Kalra et al., 2012) and ExoCarta (Keerthikumar et al., 2016) revealed that almost all proteins overlapped with secreted proteins found in these two databases (Figure 1I).

Comparative Mapping of Human Brown versus White Adipocyte Secretomes

We applied label-free quantification based on the MaxLFQ algorithm, which has proven robust and allows a comparison of an arbitrary number of samples simultaneously (Cox et al., 2014). The secretomes were highly correlated within groups (median Pearson correlation for the brown adipocytes = 0.86; Figure 2A; Table S1).

The secretomes of brown and white adipocytes were sufficiently separated to classify them as distinct entities as visualized by a principal component analysis (PCA) (Figure 2B). In our search for novel batokines, we filtered for proteins that were annotated with gene ontologies (GOs) for growth factor activity or hormone activity. This identified six hormones (Figure 2C) and eight growth factors (Figure 2D; Table S2). We calculated the fold change for these proteins between unstimulated brown and white adipocytes (woNE brown/white), unstimulated and NE-stimulated brown adipocytes (brown NE/woNE), and unstimulated and NE-stimulated white adipocytes (white NE/woNE). Among the hormones was the well-described adipokine adiponectin (ADIPOQ) (Scherer et al., 1995; Stern et al., 2016) and Fibrillin-1 (FBN1), the precursor protein for the recently discovered adipose hormone asprosin (Duerrschmid et al., 2017; Romere et al., 2016). Among the growth factors, we observed that granulins (GRN) and hepatoma-derived growth factor

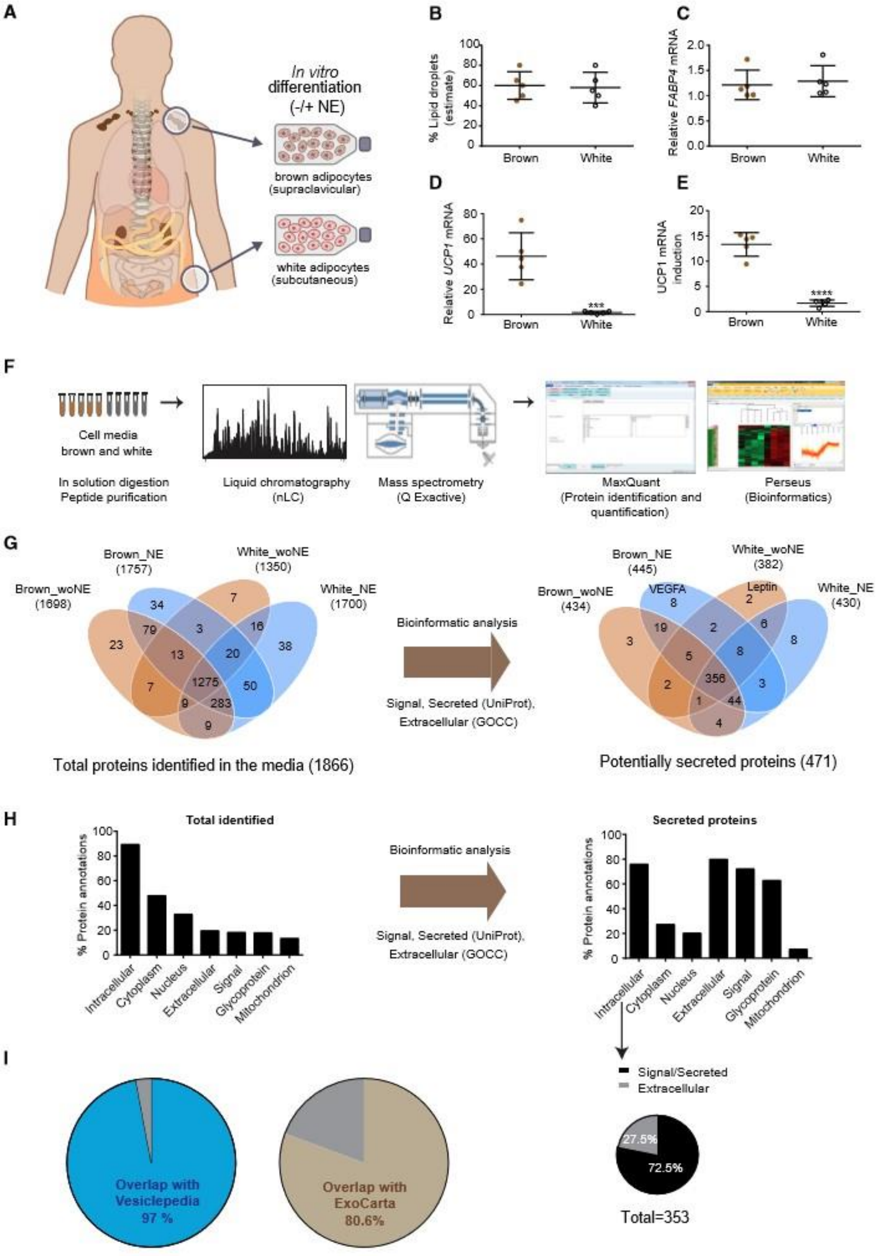


Figure 1. Model and Proteomics Workflow for Generating the Secretome of Human Brown and White Adipocytes

(A and B) (A) Mature adipocytes (brown adipocytes from n = 5 human donors; white adipocytes from n = 5 human donors) included in the study were characterized for (B) lipid droplet accumulation (estimated visually by phase contrast microscopy).

(legend continued on next page)

(HDGF) were more abundant in the brown adipocyte media than the white adipocyte media and could be considered as potential batokines. Granulins are cleaved into nine chains and have been described as autocrine growth factors important for wound healing (He et al., 2003). Granulins have not previously been described as batokines. However, one of the nine chains, acrogranin, also known as progranulin, was identified as a WAT-derived adipokine mediating high-fat diet (HFD)-induced insulin resistance in mice (Matsubara et al., 2012) and circulating levels of PRGN is associated with systemic insulin sensitivity in patients with metabolic syndrome (Li et al., 2014). HDGF promotes mitogenic activity through DNA-binding mediated transcriptional repression (Yang and Everett, 2007) but has also been detected in the extracellular region (Nüße et al., 2017). Taken together, our secretome analyses identified both established adipokines and novel candidate batokines and adipokines.

When comparing NE-induced secretomes in both brown and white adipocytes, we surprisingly observed many ribosomal proteins or proteins involved in translational processes (Table S2). We therefore decided to focus on the proteins secreted from brown and white adipocytes without norepinephrine stimulation in our search for novel batokines. A PCA plot including all proteins quantified under non-stimulated conditions revealed a clear separation between brown and white adipocytes (Figure 2E). A two-sample t test returned 143 differentially regulated proteins of which 106 were more abundant in cell media from brown adipocytes, while 37 were more abundant in cell media from white adipocytes (Figures 2F and 2G; Table S2). Fischer exact test (FDR = 0.02) for enrichment on significantly different protein in the background of total quantified proteins (1,113) returned only two categories: "Extracellular (GOCC)" and "secreted (UniProt Keyword)." This analysis suggests that the majority of differentially regulated proteins in brown and white adipocyte cell culture media are truly secreted proteins (Figure 2G). Several white-adipocyte-selective secreted proteins were extracellular matrix (ECM) associated, including transforming growth factor beta 1 (TGFB1), which has been associated with diabetes risk (Kim et al., 2013), and tenascin (TNC), an ECM glycoprotein with proinflammatory effects, which is highly expressed in WAT of obese patients and in murine models of obesity (Kim et al., 2013). We observed a more than 2-fold higher secretion of TNC and COL18A1 in cell media from white adipocytes compared to brown adipocytes. Cartilage intermediate layer protein 1 (CILP) and collagen alpha-1(XII) chain (COL12A1) were additional extracellular matrix proteins accumulating in the white adipocyte cell media (Figures 2G and 2H; Table S2). Interestingly, a recent comparative secretome study of murine interscapular brown and inguinal white adipocytes found that murine brown adipocytes secrete more ECM proteins (Ali Khan et al., 2018). These data suggest that the differences in ECM protein secretion might be either depot specific or species specific and provide a rationale for future investigations to understand the biological role of these secreted collagens and other ECM-associated proteins in the context of further defining the differences in WAT compared to BAT.

In brown adipocyte cell media, we detected highly abundant proteins with diverse functions. Parathymosin (PTMS) and prothymosin alpha (PTMA) were more than 4-fold more abundant in cell media from brown adipocytes than white adipocytes (Figure 2F). These relatively small proteins (11–12 kDa) were quantified with more than five unique peptides in cell media from brown adipocytes (Table S2) and have been shown to be involved in proliferation and regulation of immune function (Hannappel and Huff, 2003; Samara et al., 2017). The biological role of PTMS and PTMA in brown fat has not been explored. Although these proteins are not annotated as secreted, it is possible that they are secreted via exosomes or microvesicles from brown adipocytes. Interestingly, the clathrin-related proteins CLTA and CLTB, coating proteins of intracellular vesicles (Pearse, 1976), were also among the proteins with higher abundance in cell media from brown adipocytes compared to white adipocytes. This suggests that brown adipocytes make use of alternative secretion pathways such as vesicular trafficking more frequently than white adipocytes. Further, we observed that mitochondrial creatine kinase U-type (CKMT1) was more abundant in the media from brown adipocytes. This protein is involved in BAT mitochondrial energy metabolism (Kazak et al., 2015) and has gained interest as a human BAT-selective protein (Müller et al., 2016).

Among the proteins that were secreted in higher amounts from the brown adipocytes was a negative regulator of the innate immune system, complement factor H (CFH) (Figures 2F, 2I, and 2J). CFH inhibits the alternative pathway of the complement system by binding to C3b and enhancing the degradation of the alternative C3 convertase (C3bBb). In addition, CFH acts as a cofactor to complement factor I (CFI), a central inhibitor of all three complement system pathways, i.e., the classical, lectin, and alternative pathways (Zipfel and Skerka, 2009). Therefore, our MS data of the adipocyte media suggest that brown adipocytes might have higher anti-inflammatory capacity compared to white adipocytes. To verify our findings, we performed western blots on cell culture media from the brown and white adipocytes. In concordance, CFH immunoreactive protein was detected at relatively greater abundance in the media from brown adipocytes compared to white adipocytes (Figure 2K), and mRNA expression of *CFH* was higher in brown adipocytes versus white adipocytes (Figure 2L).

(C) *FABP4* mRNA expression.
(D) *UCP1* mRNA expression.
(E) *UCP1* mRNA induction, calculated as fold change between unstimulated and norepinephrine (NE) stimulated adipocytes (4 h stimulation).
Data in (B)–(E) are mean ± SEM; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
(F) Proteomics workflow.
(G) Overlap of the total amount of identified proteins between the different conditions, before and after filtering for proteins predicted to be secreted.
(H) Categorization of identified proteins before and after filtering for proteins predicted to be secreted.
(I) Overlap between identified secreted proteins and databases of secreted proteins; Vesiclepedia and ExoCarta.
See also Figure S1 and Table S1.

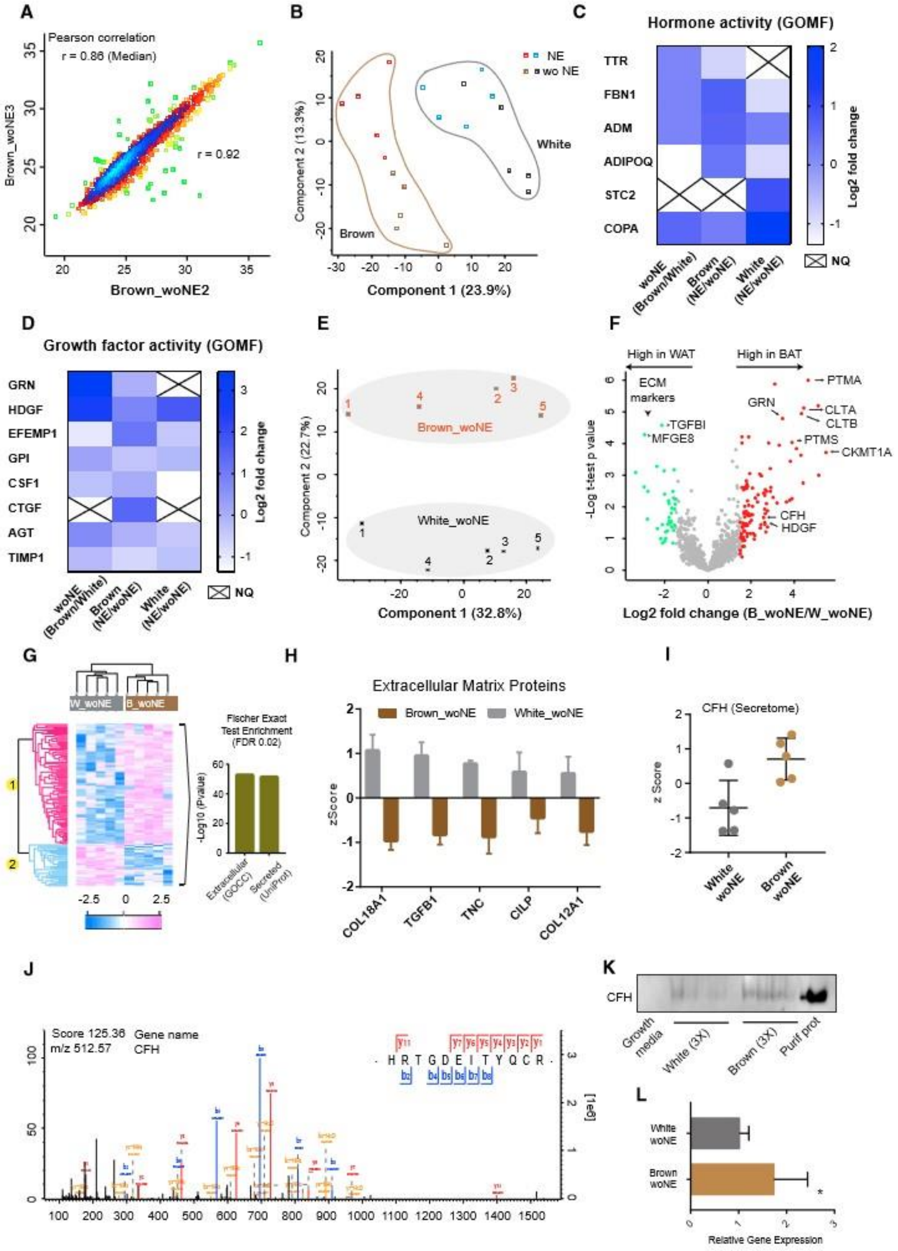


Figure 2. The Quantified Secretomes of Human Brown and White Adipocytes

(A) Representative Pearson correlation between two samples within the same group (brown adipocytes derived from two different individuals without NE stimulation).

(legend continued on next page)

A Distinct Secretome of Brown Adipocytes Identifies EPDR1 as a Candidate Batokine

Data-dependent quantitative proteomics data may contain a high percentage of missing values and systemic evaluation of the differentially expressed proteins is therefore important. For the comparative mapping described above, we imputed the missing values to best simulate the Gaussian distribution of low abundant proteins when comparing the secretomes of brown and white adipocytes, and we quantified more than 1,000 proteins expressed in cell media from both cell types, allowing for fold change comparison. However, we were concerned about the possibility that the imputation might mask differential regulation of low abundant proteins, even though the imputed values are downshifted. Thus, to avoid exclusion of any secreted protein that were selectively higher in the secretomes of either brown or white adipocytes, we investigated their selective secretomes separately. Using this approach, we exclusively quantified 101 proteins in the cell culture media from brown adipocytes, and 37 in the cell culture media from white adipocytes all these proteins were considered potentially secreted according to our criteria (Figure 3A; Table S2). Among the 101 proteins secreted from brown adipocytes was mammalian ependymin-related protein 1 (EPDR1) (Figure 3B). EPDR1 had previously been shortlisted in a TMT-labeling-based secretomics study of PR/SET domain 16 (PRDM16)-induced browning in mice (Svensson et al., 2016). We found EPDR1 to be exclusively quantified in BAT cell media (Figures S2A and S2B), while the imputation (Figure S2B) explains why EPDR1 was not annotated as differentially expressed in the volcano plot comparing the secretomes of white and brown adipocytes (Figure S2C). Consistent with these findings, *EPDR1* mRNA levels were higher in brown adipocytes compared to white adipocytes (Figure 3C). The *EPDR1* gene encodes three transcript variants, translated to three protein isoforms, of which two include an N-terminal signal peptide for secretion. We detected peptides unique to the secreted isoform 1, confirming its presence in our samples. Peptides mapping to isoform 2 (secreted) and isoform 3 (not secreted) were also detected, but these peptides also mapped to isoform 1; thus, their presence in our samples could be neither confirmed nor excluded (Figure 3D). These results do not exclude a low-level secretion of EPDR1 from white adipocytes. To further assess this, we performed a targeted proteomics of EPDR1. Notably, only one peptide could be quantified. Importantly, this peptide was unique to the secreted isoform 1, and the quantification confirmed that EPDR1 secretion is higher in cell media from brown adipocytes compared to media from white adipocytes (Figures S2D and S2E). We thus conclude that EPDR1 is selectively secreted from brown adipocytes yet also produced by white adipocytes, albeit likely in minimal amounts. Isoform 1 has been assigned as the canonical sequence, and as our data indicated that this isoform was detected in the brown adipocyte cell media, we decided to focus on this variant in our downstream experiments. As revealed by the crystal structure, EPDR1 has hydrophobic binding grooves and can interact with liposomes, suggesting a potential role in lipid transport (Wei et al., 2019).

To investigate the role of EPDR1 in brown adipocytes, we performed small interfering RNA (siRNA)-mediated knockdown. As the protein was expressed already at the onset of differentiation (Figure 3E), we transfected human brown adipocytes at day 0 of differentiation and measured *EPDR1* mRNA levels after 24 h and again after the full 12-day differentiation program. The siRNA knockdown was highly efficient and had a sustained effect throughout differentiation (Figure 3E). *EPDR1* mRNA levels were higher in the mature adipocytes than the early differentiation state (Figure 3E). The knockdown was also confirmed at the protein level (Figure S2F).

We observed no visual changes in lipid droplet accumulation or mitochondrial content following knockdown (Figure S2H). However, we found a decrease in the metabolic response to adrenergic signaling in brown adipocytes transfected with EPDR1 siRNA, as NE-induced proton leak was decreased (Figures 3F and 3G). In concordance, NE-induced upregulation of the thermogenic markers *DIO_2*, *UCP1*, *PPARGC1A*, *$PPAR\alpha$*, and *CKMT1* was blunted in these cells (Figure 3H). We validated these data using two additional siRNA oligos, also targeting *EPDR1*, and could reproduce the effects on *UCP1* regulation (Figure S2G). To further assess the effects of EPDR1 knockdown, we measured a range of thermogenic and adipogenic markers. Whereas many thermogenic and adipogenic markers remained unchanged following EPDR1 knockdown (Figures S2I and S2J), we found that the blunted thermogenic activation was accompanied by a reduced expression of GLUT4 (Figure 3I) and an increased expression of *COL1A2* and *$PDGFR\alpha$*, suggesting a potential increase in an undifferentiated, fibroblastic state (Sun et al., 2017) (Figure 3J). Interestingly, *CITED1*, a beige fat marker (Sharp et al., 2012), was also increased (Figure 3J).

To further understand the metabolic deficiency in the cells differentiating with reduced amounts of EPDR1, we performed

(B) PCA plot, including all proteins quantified in brown and white adipocytes with (NE) and without NE (wo NE) stimulation.

(C) Quantified proteins in the datasets with hormone activity, as annotated with the gene ontology molecular function (GOMF) term. NQ, not quantified.

(D) Quantified proteins in the datasets with growth factor activity, identified with GOMF term.

(E) PCA plot, including all quantified proteins in unstimulated brown and white adipocytes.

(F) Volcano plot depicting a two-sample t test between all quantified proteins in unstimulated brown and white adipocytes.

(G) Heatmap of all differentially regulated proteins between unstimulated brown and white adipocytes, as determined by two-sample t test.

(H) ECM proteins in white adipocytes compared to brown adipocytes.

(I) Secreted amounts of complement factor H (CFH). Values represent *Z* score from the mass spectrometry analysis of cell media.

Data in (B)–(I) are based on n = 5 brown adipocytes and n = 5 white adipocytes, biological replicates, all from separate human donors.

(J) Mass spectrometry trace of CFH.

(K) Western blot of CFH in brown and white adipocyte culture media from brown adipocyte cultures (n = 3) and white adipocyte cultures (n = 3); biological replicates, all from separate human donors.

(L) *CFH* relative mRNA levels in white (n = 5) and brown (n = 5) adipocytes, all derived from separate human donors. Data are mean ± SEM; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

See also Table S2.

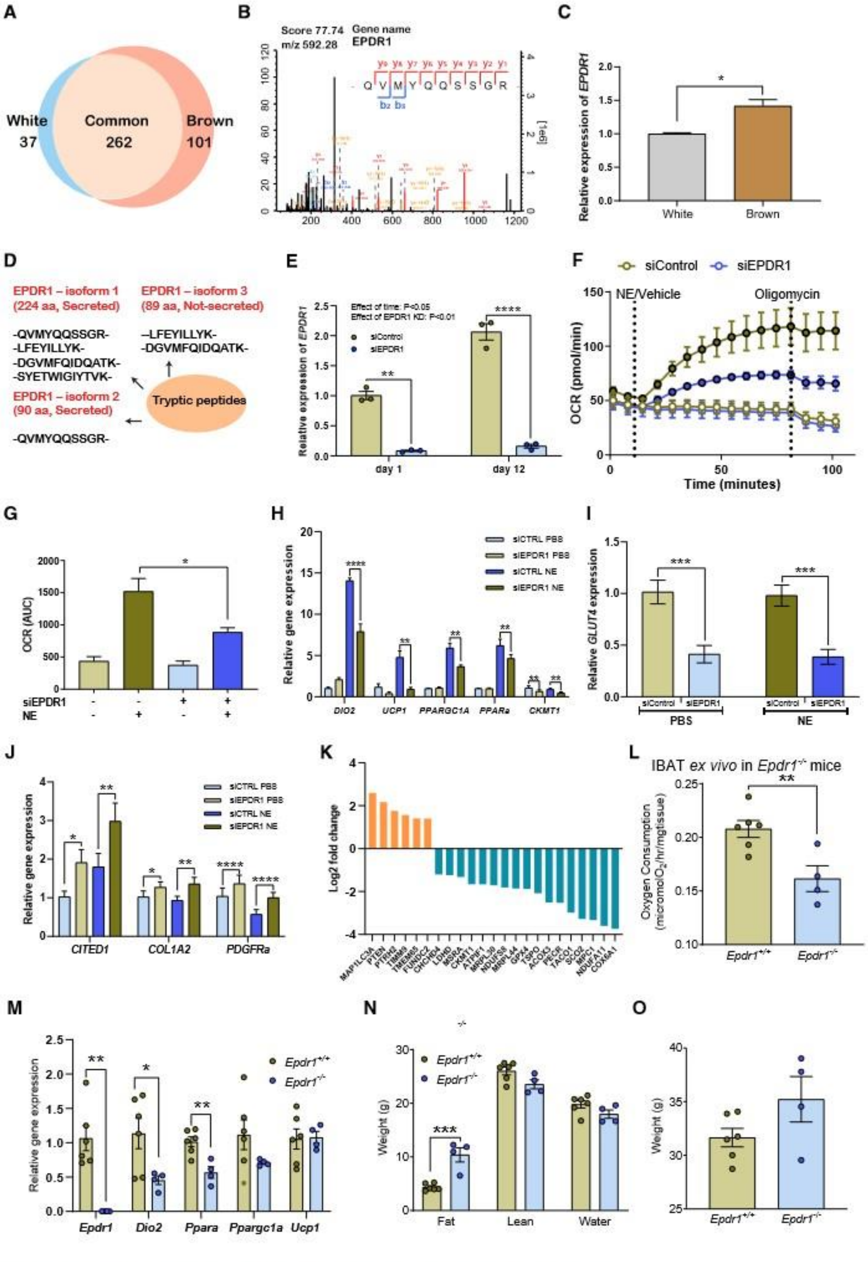


Figure 3. Identification and *In Vitro* Characterization of EPDR1, a Novel Batokine Candidate

(A) Venn diagram illustrating quantified proteins annotated as secreted.

(B) Example of EDPR1 unique peptides sequenced by mass spectrometry.

(legend continued on next page)

a cellular proteomics analysis on brown adipocytes derived from four human subjects and transfected at day 0 of differentiation with either siRNAs targeting EPDR1 or a non-targeting siRNA control. Cells were differentiated *in vitro* and harvested on day 12. We quantified 3,164 proteins in total of which 47 proteins were upregulated, and 53 proteins were downregulated following *EPDR1* knockdown (Table S3). Among the downregulated proteins were EPDR1 and GLUT4, consistent with our mRNA data (Table S3). To relate the proteomics data to the deficiency in NE- induced proton leak, we investigated whether mitochondria-associated proteins were among the regulated proteins. Indeed, the list of regulated proteins included 23 mitochondrial proteins (UniProt GO and Keywords annotation) of which 6 proteins were upregulated, whereas 17 were downregulated (Figure 3K). Among the downregulated molecular entities were key mitochondrial proteins, such as NDUFA11 and NDUFS8, core subunits of complex I in the mitochondrial respiratory chain; COX6A1, the terminal oxidase in mitochondrial electron transport; and CKMT1A, a key protein in creatine kinase-dependent thermogenesis (Kazak et al., 2015). Taken together, these data indicate that adipocytes lacking EPDR1 during differentiation exhibit a hampered mitochondrial phenotype, reducing the functional capacity of the mature brown adipocyte.

To further address this, we examined the status of interscapular brown fat (iBAT) in a previously uncharacterized Epdr1$^{-/-}$ whole-body knockout mouse. Interestingly, we found that *ex vivo* oxygen consumption was reduced in iBAT from Epdr1$^{-/-}$ compared to wild-type mice at 20 weeks of age (Figure 3L). This deficiency in thermogenic function was accompanied by a reduction in *Dio2* and *Pparα* mRNA levels in the *Epdr1*$^{-/-}$ mice compared to wild-type mice, whereas *Ucp1* and *Ppargc1a* mRNA levels remained unchanged (Figure 3M). Moreover, Epdr1$^{-/-}$ mice had a pronounced accumulation of body fat compared to control mice (Figure 3N), while total body weight was not different between groups (Figure 3O). These data, while encouraging, and directionally consistent with the totality of the EPDR1 data, should be interpreted with caution, as only a small number of mice were available for analysis and control mice were not littermates but were age- and gender-matched wild-type C57BL/6J.

Taken together, these data suggest that reduced levels of EPDR1 during differentiation results in incomplete brown fat commitment and thus propose a role for EPDR1 in thermogenic differentiation, perhaps as part of an auto- or paracrine circuit.

EPDR1 Affects Metabolism Independently of BAT Activity

Further characterization of the *Epdr1* whole-body knockout mouse revealed that at 9–11 weeks of age, the Epdr1$^{-/-}$ mice had a lower oxygen consumption than wild-type mice, which was mostly pronounced in the dark phase (Figure 4A). Interestingly, this was accompanied by a decrease in physical activity (Figure 4B) without differences in food intake (Figure 4C). These data suggested that the metabolic effects of mice lacking *Epdr1* in all tissues were not restricted to BAT.

To further investigate the effects of EPDR1 on whole-body metabolism, we produced recombinant EPDR1 (human isoform 1) (Figure S3A) and injected 14-week-old C57BL/6NRj mice with EPDR1 protein at 2 mg/kg. Mice were acclimatized to thermoneutrality prior to the injection, allowing us to assess whether EPDR1 had any effect on metabolism independently of cold stimulation. The injection was given just prior to the dark period, when BAT activity is naturally increased by murine circadian rhythm (Gerhart-Hines et al., 2013). Hence, we studied the effect of EPDR1 without cold stimulation but when BAT was metabolically primed. We recorded whole-body metabolism throughout the dark period using indirect calorimetry. Subcutaneous injection of EPDR1 protein resulted in an increase in oxygen consumption during a 12-h period as compared with control mice injected with PBS (Figure 4D). A corresponding increase in energy expenditure was observed, while respiratory exchange ratio (RER) remained unchanged (Figure S3B). Whereas no difference in locomotor activity was detected (Figure 4E), the increase in oxygen consumption was followed by a subsequent increase in food intake (Figure 4F), exemplifying the fine-tuned synergy between energy expenditure and energy intake (Contreras et al., 2014, 2017; Sutton et al., 2014). Based on the reduced

(C) EPDR1 isoform 1 mRNA expression in a representative white adipocyte culture (n = 3 independent experiments) and a representative brown adipocyte culture (n = 3 independent experiments).

(D) Tryptic peptides sequenced and mapped to EPDR1 isoforms.

For the downstream cell experiments in (E)–(J), a representative brown adipocyte culture was transfected with siRNAs (n=3 independent experiments).

(E) *EPRD1* mRNA expression at day 0 and at day 12 of differentiation, with and without knockdown of *EPDR1* (n = 3).

(F) Oxygen consumption rate in human brown adipocytes at day 12 of differentiation with or without *EPDR1* knockdown at day 0 and with or without NE stimulation (n = 3).

(G) Oxygen consumption rates following oligomycin treatment. The area under the curve for the three time points (88.09, 94.8, 101.47) following addition of oligomycin is presented for the four groups. An unpaired t test was performed to test differences between norepinephrine-stimulated leak in the siControl treated cells compared to the siEPDR1-treated cells.

(H) Relative mRNA expression of thermogenic markers *DIO*$_2$, *UCP1*, *PPARGC1A*, *PPARα*, and *CKMT1* (n = 3).

(I) Relative mRNA expression of *GLUT4* (n = 3).

(J) Relative mRNA expression of *CITED1*, *COL1A2*, and *PDGFRa*.

(K) Mitochondrial proteins differentially regulated in brown adipocytes at day 12 of differentiation following knockdown of EPDR1 at day 0 (n = 4 represents brown adipocytes derived from 4 different human donors).

(L) *Ex vivo* oxygen consumption in iBAT of Epdr1$^{-/-}$ mice (n = 4) and age-matched Epdr1$^{+/+}$ controls (n = 6).

(M) Relative mRNA expression of thermogenic markers *Dio2*, *Ucp1*, *Ppargc1a*, and *Pparα* in iBAT of *Epdr1*$^{-/-}$ mice (n = 4) and age-matched *Epdr1*$^{+/+}$ controls (n = 6).

(N) Weight of fat and lean mass in in iBAT of *Epdr1*$^{-/-}$ mice (n = 4) and age-matched *Epdr1*$^{+/+}$ controls (n = 6). Data are presented as mean ± SEM; *p < 0.5; **p < 0.01; ***p < 0.001; ****p < 0.0001.

See also Figure S2 and Table S2.

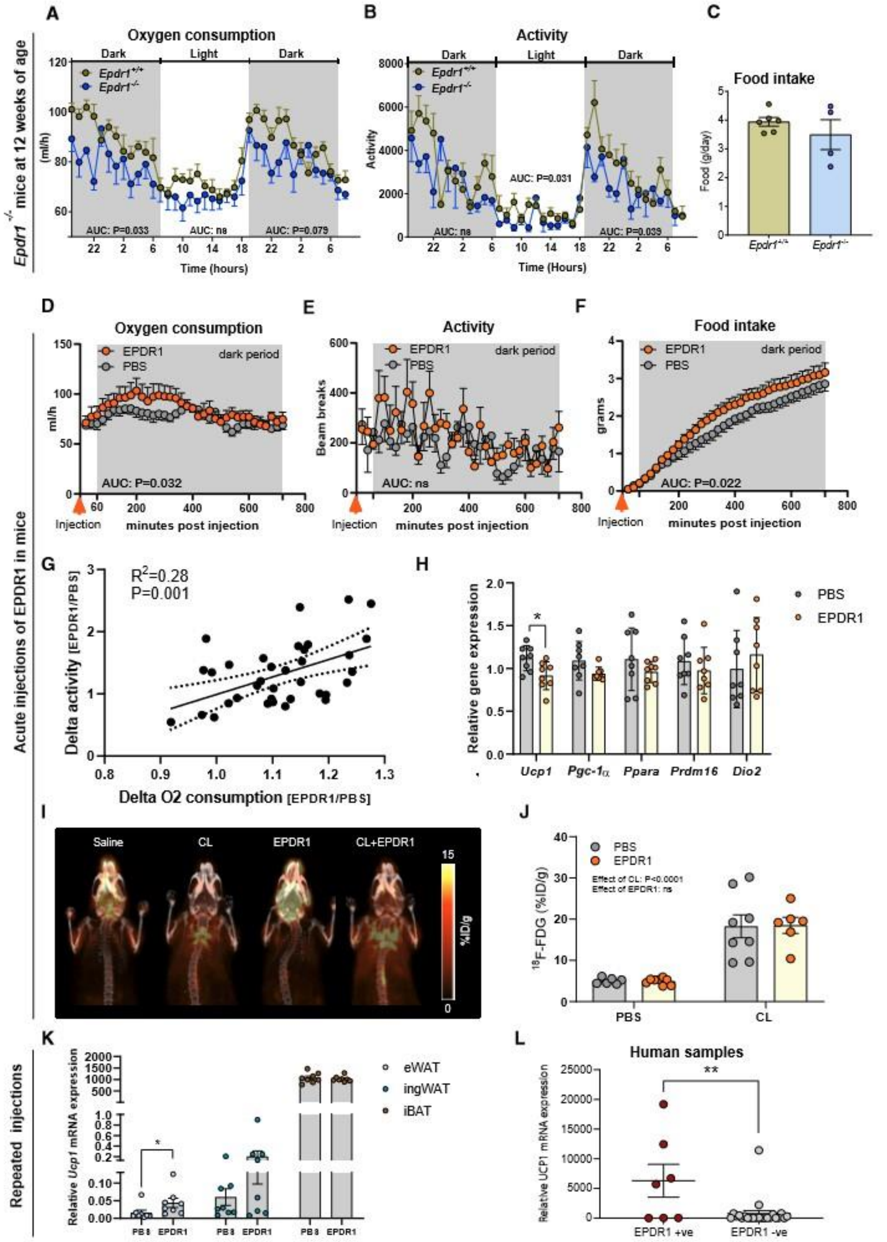


Figure 4. Effects of EPDR1 on Whole-Body Metabolism

(A) Oxygen consumption in *Epdr1*$^{-/-}$ (n = 4) or age-matched *Epdr1*$^{+/+}$ controls (n = 6).

(B) Locomotor activity in *Epdr1*$^{-/-}$ (n = 4) or age-matched *Epdr1*$^{+/+}$ controls (n = 6).

(legend continued on next page)

physical activity observed in the Epdr1$^{-/-}$ mice, we performed a linear regression analysis of the EPDR1-induced increases in oxygen consumption and physical activity, respectively. This revealed a correlation (R^2 = 0.28, p = 0.001), and the EPDR1-induced increase in oxygen consumption could thus at least partly be explained by elevated physical activity (Figure 4G). Consistent with this finding, we found no induction of genes linked to thermogenic activation (Figure 4H). In fact, we found a modest downregulation of *Ucp1*, while the rest of the markers were unchanged (Figure 4H).

Acute BAT activation in humans can be quantified by performing a PET/CT scan using a radioactive glucose tracer (Chen et al., 2016). This method can also be applied in anesthetized mice (Sustarsic et al., 2018). Although our data did not support that the increase in oxygen consumption was due to increased BAT activity, we decided to investigate this further using [18F] FDG-PET/CT scanning. Moreover, we aimed to assess whether a potential activation was interacting with a sympathetic tone. Therefore, we performed an experiment where EPDR1 was injected into C57BL/6NRj mice simultaneously with either PBS vehicle or the β3-agonist CL-316,243 (CL). Uptake of a glucose tracer ([18F] FDG) was then assessed by PET/CT in anesthetized animals. We observed a substantial increase in FDG uptake in mice that were injected with CL, but no further activity was detected with EPDR1 injections (Figures 4I and 4J). Consistent with the gene expression data obtained from the mice in the metabolic chambers, we observed a downregulation of *Ucp1*, and in this case, *Prdm16* was also reduced whereas other thermogenic markers including *Dio$_2$*, *Ppargc1a*, and *Pparα* remained unchanged (Figure S3C). In conclusion, EPDR1 affected metabolism but without enhancing BAT activity. Importantly, EPDR1 is expressed in other tissues, including the brain (Wei et al., 2019), possibly explaining the observed effects on metabolism.

To assess whether treatment with EPDR1 had any long-term effects on whole-body metabolism, we fed C57BL/6NRj mice a HFD and injected them with EPDR1 once daily for 21 days. We observed no difference in metabolic phenotype between the groups receiving EPDR1 compared to the group receiving PBS (Figure S3D). These mice were housed at room temperature, which is an appreciable cold stimulation and therefore could have masked EPDR1-dependent effects. Interestingly, we did observe an increase in *Ucp1* expression in WAT of mice that had been injected with EPDR1 protein (Figure 4K). Consistent with the lack of whole-body metabolic effects, the upregulation of *Ucp1* in the WAT following repeated injections of EPDR1 was very modest. This needs to be interpreted in the light of the mechanism of white fat browning. It has been reported that subpopulations of thermogenic adipocytes can be induced within epididymal (Petrovic et al., 2010) or inguinal (Seale et al., 2011) adipose tissues. It has further been demonstrated that thermogenic precursor cells co-exist with white fat precursor cells in the white fat depot (Wu et al., 2012). Therefore, measuring the whole tissue mRNA levels will likely dilute the upregulation of *Ucp1* limited to a thermogenic subpopulation. The reason that *Ucp1* expression in iBAT is not increased following the chronic stimulation is likely due to the room temperature housing conditions, which already result in efficient BAT recruitment (Sanchez-Gurmaches et al., 2018). Finally, we investigated whether EPDR1, as a secreted protein, could be detected in human plasma using a commercially available ELISA kit with a detection level of 30 ng/mL. We found that circulating EPDR1 was detected in seven out of thirty adult humans (Table S4). Interestingly, the subjects that were positive for EPDR1 had higher levels of *UCP1* mRNA in their deep neck brown adipose tissue, consistent with the possibility that EPDR1 could be secreted into the circulation perhaps from metabolically active BAT in humans (Figure 4L).

Collectively, we here provide the first comprehensive analysis of the human brown adipocyte secretome compared to the white adipocyte secretome in a basal state and following an acute NE stimulation. Our results reinforce that brown and white adipocytes have distinct secretory profiles and metabolic functions. We identify a large number of novel candidate batokines. Among several interesting candidates, we focused on the role of a novel human batokine, EPDR1, which is selectively secreted from brown adipocytes. We demonstrate that EPDR1 is vital for development into a functional thermogenic adipocyte, and our data further indicate that EPDR1 can act in an endocrine fashion. In conclusion, our data illuminate the human BAT secretome and provides a promising source of novel metabolic regulators that could serve as an important resource for future studies, including evaluation of potential drug targets for mediating improved metabolic control.

(C) Food intake in *Epdr1*$^{-/-}$ (n = 4) or age-matched *Epdr*$^{+/+}$ controls (n = 6).

(D) Oxygen consumption following EPDR1 protein (n = 8) or vehicle (n = 8) injection in C57Bl/6NRj mice at thermoneutrality (TN) during the dark period.

(E) Locomotor activity following EPDR1 (n = 8) or vehicle (n = 8) injection in C57Bl/6NRj mice at TN.

(F) Food intake following EPDR1 (n = 8) or vehicle (n = 8) injection in C57Bl/6NRj mice at TN.

(G) Linear regression analysis of the EPDR1-induced increases in oxygen consumption and physical activity.

(H) Relative mRNA expression of thermogenic markers in iBAT of C57/Bl/6NRj mice following EPDR1 (n = 8) or vehicle (n = 8) injection.

(I) Representative images of maximum intensity projection following injection of recombinant EPDR1 with and without injection of CL 316,243 at thermoneutrality and during FDG-PET/CT scanning (n = 6–8/group, as one mouse in the EPDR1-injected group was euthanized due to a paralyzed leg, and two mice in group in the EPDR1 + CL group were excluded due to bad injections).

(J) Quantification of maximum intensity in mice injected with or without recombinant EPDR1 and with and without CL 316,243 (n = 6–8 as described in I).

(K) Relative mRNA expression of *Ucp1* in eWAT, ingWAT, and iBAT of C57/Bl/6NRj mice on an HFD at room temperature following daily injections of EPDR1 or vehicle injections for 21 days (n = 8). The effects were tested using a two-way ANOVA, detecting an effect of tissue ($p < 0.0001$) and an effect of EPDR1 ($p < 0.05$). The tissue driving the effect of EPDR1 injections was further examined by using a Sidak's multiple comparisons test of which the p value is shown in the figure.

(L) EPDR1 protein levels were assessed in human plasma samples, and subjects were divided into two groups based on detectable levels (EPDR1+ve, n = 7) or no detection (EPDR1-ve, n = 23). An unpaired t test was used to assess differences. Relative gene expression of *UCP1* in deep neck surgical biopsies was measured using qPCR and plotted in the two groups. Data are presented as mean ± SEM; *$p < 0.5$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.

See also Figure S3.

Limitations of Study

Regarding the secretome analysis, methodological limitations could have resulted in that low abundant secreted proteins were not detected. Moreover, as we could exclude that some of the identified proteins originated from fragmented cells, we designed a computational pipeline to filter out proteins that were not likely to be secreted. This might, however, have resulted in removal of proteins that were secreted. On the other hand, we might have included proteins that were fragmentation products. Nevertheless, if any cell fragmentation had occurred, it is reasonable to expect that it occurred to the same extent between white and brown adipocytes as culturing and incubation conditions were identical. Our secretome analysis is done in mature adipocytes, but levels of (brown) adipokines differ during differentiation (Zhong et al., 2010). Analyses of more time points in the differentiation process could provide additional future insights into the secretory differences between white and brown adipocytes. We identify EPDR1 as a novel batokine, important for brown fat determination, whereas the exact pathway through which EPDR1 acts remains to be explored.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
  - Disclosure of Limited Availability of Biological Material
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human Supraclavicular (Brown) and Subcutaneous (White) Fat Precursor Cells
  - Human Subjects
  - Mouse Models
- METHOD DETAILS
  - Human Primary Adipocyte Culture Conditions
  - siRNA Mediated Knockdown of EPDR1 in Adipocytes
  - RNA Isolation and Quantitative Real-Time PCR of Adipocytes
  - Oxygen Consumption Measurements in Adipocytes
  - Lipid and Mitochondrial Staining
  - ELISA
  - EPDR1 Production
  - Assessment of the Purity of EPDR1 Recombinant Protein
  - Small Animal [18F] FDG PET/CT Imaging
  - RNA Isolation and Quantitative Real-Time PCR following EPDR1 Injection
  - Indirect Calorimetry following EPDR1 Injection
  - Body Composition, Food Intake and Indirect Calorimetry in EPDR1$^{-/-}$ Mice
  - *Ex Vivo* Oxygen Consumption
  - RNA Isolation and Quantitative Real-Time PCR in EPDR1$^{-/-}$ Mice
  - Secretome Analysis by Mass Spectrometry
  - Sample Preparation for Secretome and Cellular Proteome
  - LC MS/MS Analysis
  - Computational MS Data Analysis
  - Predictive Multiplexed Selective Ion Monitoring (pmSIM)
  - Western Blot Analysis on Cell Media
  - Western Blot Analysis on Cell Lysate
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cmet.2019.10.001.

ACKNOWLEDGMENTS

We thank Jacob Steen Petersen and Birgitte Andersen from Novo Nordisk A/S for valuable scientific discussion, and Kirsten L. Meeske and Charlotte Wittrup from Minerva Imaging ApS for help with the scanning and data analysis. The Centre for Physical Activity Research (CFAS) is supported by TrygFonden (grant IDs 101390 and ID 20045). During the study period, the Center of Inflammation and Metabolism (CIM), Rigshospitalet, was supported by a grant from the Danish National Research Foundation (DNRF55). CIM/CFAS is a member of DD2, the Danish Center for Strategic Research in Type 2 Diabetes (the Danish Council for Strategic Research, grant nos. 09-067009 and 09-075724). Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center, based at the University of Copenhagen, Denmark, and partially funded by an unconditional donation from the Novo Nordisk Foundation (https://cbmr.ku.dk/) (grant no. NNF18CC0034900). Novo Nordisk Foundation Center for Protein Research (NNF14CC001) (https://www.cpr.ku.dk/) is supported by an unconditional grant from the Novo Nordisk Foundation to University of Copenhagen. M.M. was also supported by Max-Planck-Society for the Advancement of Science. D.J.D. was supported by a Banting and Best Diabetes Centre-Novo Nordisk Chair in Incretin Biology and CIHR operating grant 154321. Z.G.H. was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement no. 639382). N.Z.J. is supported by a research grant from the Danish Diabetes Academy, which is funded by the Novo Nordisk Foundation (grant no. NNF17SA0031406). P.G. and R.B.-O. were supported by the Danish Research Foundation of Independent Research (grant no. DFF-6110-00489). The study was further funded by a shared research grant from Novo Nordisk A/S to C.S. and M.M. and by a grant to C.S. from the Novo Nordisk Foundation (grant no. NNF18OC0034378).

AUTHOR CONTRIBUTIONS

C.S. and M.M. supervised the study. C.S., A.S.D., M.M., L.P., S.N., D.J.D., Z.-G.H., B.K.P., G.P., P.G., A.S., M.T.C., A.K., and B.H. performed hypothesis generation, conceptual design, data analysis, and manuscript preparation. L.P., A.S.D., J.L.B., N.Z.J., C.H.N., T.M., A.D.B., T.J.L., R.B-.O., B.S.P., C.H., and M.C.K.S. conducted experiments and data analysis. All authors edited and approved the final manuscript.

DECLARATION OF INTERESTS

D.J.D. has served as an advisor or consultant or speaker within the past 12 months to Forkhead Biotherapeutics, Heliome Inc., Intarcia Therapeutics, Kallyope, Merck Research Laboratories, Novo Nordisk Inc., Pfizer Inc., and Sanofi Inc. C.H., M.T.C., and A.S. are employees and shareholders at Novo Nordisk A/S. Z.G.H. works, in some capacity, for Embark Biotech Aps developing therapeutics for the treatment of diabetes and obesity.

Received: September 18, 2018
Revised: April 26, 2019
Accepted: October 2, 2019
Published: October 24, 2019

REFERENCES

Ahima, R.S., Prabakaran, D., Mantzoros, C., Qu, D., Lowell, B., Maratos-Flier, E., and Flier, J.S. (1996). Role of leptin in the neuroendocrine response to fasting. Nature *382*, 250–252.

Ali Khan, A., Hansson, J., Weber, P., Foehr, S., Krijgsveld, J., Herzig, S., and Scheideler, M. (2018). Comparative secretome analyses of primary murine white and brown adipocytes reveal novel adipokines. Mol. Cell. Proteomics *17*, 2357–2370.

Bradley, A., Anastassiadis, K., Ayadi, A., Battey, J.F., Bell, C., Birling, M.C., Bottomley, J., Brown, S.D., Bürger, A., Bult, C.J., et al. (2012). The mammalian gene function resource: the International Knockout Mouse Consortium. Mamm. Genome *23*, 580–586.

Cannon, B., and Nedergaard, J. (2004). Brown adipose tissue: function and physiological significance. Physiol. Rev. *84*, 277–359.

Chen, K.Y., Cypess, A.M., Laughlin, M.R., Haft, C.R., Hu, H.H., Bredella, M.A., Enerbäck, S., Kinahan, P.E., Lichtenbelt, Wv, Lin, F.I., et al. (2016). Brown adipose reporting criteria in imaging studies (BARCIST 1.0): recommendations for standardized FDG-PET/CT experiments in humans. Cell Metab. *24*, 210–222.

Contreras, C., Gonzalez, F., Fernø, J., Diéguez, C., Rahmouni, K., Nogueiras, R., and López, M. (2014). The brain and brown fat. Ann. Med. *47*, 150–168.

Contreras, C., Nogueiras, R., Diéguez, C., Rahmouni, K., and López, M. (2017). Traveling from the hypothalamus to the adipose tissue: the thermogenic pathway. Redox Biol. *12*, 854–863.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics *13*, 2513–2526.

Deshmukh, A.S., Cox, J., Jensen, L.J., Meissner, F., and Mann, M. (2015). Secretome analysis of lipid-induced insulin resistance in skeletal muscle cells by a combined experimental and bioinformatics workflow. J. Proteome Res. *14*, 4885–4895.

Duerrschmid, C., He, Y., Wang, C., Li, C., Bournat, J.C., Romere, C., Saha, P.K., Lee, M.E., Phillips, K.J., Jain, M., et al. (2017). Asprosin is a centrally acting orexigenic hormone. Nat. Med. *23*, 1444–1453.

Gerhart-Hines, Z., Feng, D., Emmett, M.J., Everett, L.J., Loro, E., Briggs, E.R., Bugge, A., Hou, C., Ferrara, C., Seale, P., et al. (2013). The nuclear receptor Rev-erbα controls circadian thermogenic plasticity. Nature *503*, 410–413.

Hannappel, E., and Huff, T. (2003). The thymosins. Prothymosin alpha, parathymosin, and beta-thymosins: structure and function. Vitam. Horm. *66*, 257–296.

He, Z., Ong, C.H., Halper, J., and Bateman, A. (2003). Progranulin is a mediator of the wound response. Nat. Med. *9*, 225–229.

Huberts, D.H.E.W., and van der Klei, I.J. (2010). Moonlighting proteins: an intriguing mode of multitasking. Biochim. Biophys. Acta *1803*, 520–525.

Jespersen, N.Z., Larsen, T.J., Peijs, L., Daugaard, S., Homøe, P., Loft, A., De Jong, J., Mathur, N., Cannon, B., Nedergaard, J., et al. (2013). A classical brown adipose tissue mRNA signature partly overlaps with Brite in the supraclavicular region of adult humans. Cell Metab. *17*, 798–805.

Kalra, H., Simpson, R.J., Ji, H., Aikawa, E., Altevogt, P., Askenase, P., Bond, V.C., Borràs, F.E., Breakefield, X., Budnik, V., et al. (2012). Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. PLoS Biol. *10*, e1001450.

Kazak, L., Chouchani, E.T., Jedrychowski, M.P., Erickson, B.K., Shinoda, K., Cohen, P., Vetrivelan, R., Lu, G.Z., Laznik-Bogoslavski, D., Hasenfuss, S.C., et al. (2015). A creatine-driven substrate cycle enhances energy expenditure and thermogenesis in beige fat. Cell *163*, 643–655.

Keerthikumar, S., Chisanga, D., Ariyaratne, D., Al Saffar, H., Anand, S., Zhao, K., Samuel, M., Pathan, M., Jois, M., Chilamkurti, N., et al. (2016). ExoCarta: a web-based compendium of exosomal cargo. J. Mol. Biol. *428*, 688–692.

Kelstrup, C.D., Bekker-Jensen, D.B., Arrey, T.N., Hogrebe, A., Harder, A., and Olsen, J.V. (2018). Performance evaluation of the Q Exactive HF-X for shotgun proteomics. J. Proteome Res. *17*, 727–738.

Kim, Y., Kim, H.D., Youn, B., Park, Y.G., and Kim, J. (2013). Ribosomal protein S3 is secreted as a homodimer in cancer cells. Biochem. Biophys. Res. Commun. *441*, 805–808.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Larsen, T.J., Jespersen, N.Z., and Scheele, C. (2019). Adipogenesis in primary cell culture. In Handbook of Experimental Pharmacology (Springer), pp. 73–84.

Lee, P., Linderman, J.D., Smith, S., Brychta, R.J., Wang, J., Idelson, C., Perron, R.M., Werner, C.D., Phan, G.Q., Kammula, U.S., et al. (2014). Irisin and FGF21 are cold-induced endocrine activators of brown fat function in humans. Cell Metab. *19*, 302–309.

Li, H., Zhou, B., Xu, L., Liu, J., Zang, W., Wu, S., and Sun, H. (2014). Circulating PGRN is significantly associated with systemic insulin sensitivity and autophagic activity in metabolic syndrome. Endocrinology *155*, 3493–3507.

Matsubara, T., Mita, A., Minami, K., Hosooka, T., Kitazawa, S., Takahashi, K., Tamori, Y., Yokoi, N., Watanabe, M., Matsuo, E., et al. (2012). PGRN is a key adipokine mediating high fat diet-induced insulin resistance and obesity through IL-6 in adipose tissue. Cell Metab. *15*, 38–50.

Meissner, F., Scheltema, R.A., Mollenkopf, H.J., and Mann, M. (2013). Direct proteomic quantification of the secretome of activated immune cells. Science *340*, 475–478.

Michalski, A., Damoc, E., Hauschild, J.P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011). Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop Quadrupole Orbitrap Mass Spectrometer. Mol. Cell. Proteomics *10*, M111.011015.

Müller, S., Balaz, M., Stefanicka, P., Varga, L., Amri, E.Z., Ukropec, J., Wollscheid, B., and Wolfrum, C. (2016). Proteomic analysis of human brown adipose tissue reveals utilization of coupled and uncoupled energy expenditure pathways. Sci. Rep. *6*, 30030.

Nickel, W., and Rabouille, C. (2009). Mechanisms of regulated unconventional protein secretion. Nat. Rev. Mol. Cell Biol. *10*, 148–155.

Nüße, J., Blumrich, E.M., Mirastschijski, U., Kappelmann, L., Kelm, S., and Dietz, F. (2017). Intra- or extra-exosomal secretion of HDGF isoforms: the extraordinary function of the HDGF-A N-terminal peptide. Biol. Chem. *398*, 793–811.

Park, J., Kim, M., Sun, K., An, Y.A., Gu, X., and Scherer, P.E. (2017). VEGF-A - expressing adipose tissue shows rapid beiging and enhanced survival after transplantation and confers IL-4-independent metabolic improvements. Diabetes *66*, 1479–1490.

Pearse, B.M.F. (1976). Clathrin: a unique protein associated with intracellualar transfer of membrane by coated vesicles. Proc. Natl. Acad. Sci. USA *73*, 1255–1259.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. *47*, D442–D450.

Petrovic, N., Walden, T.B., Shabalina, I.G., Timmons, J.A., Cannon, B., and Nedergaard, J. (2010). Chronic peroxisome proliferator-activated receptor gamma (PPARG) activation of epididymally derived white adipocyte cultures reveals a population of thermogenically competent, UCP1-containing adipocytes molecularly distinct from classic brown adipocytes. J. Biol. Chem. *285*, 7153–7164.

Romere, C., Duerrschmid, C., Bournat, J., Constable, P., Jain, M., Xia, F., Saha, P.K., Del Solar, M., Zhu, B., York, B., et al. (2016). Asprosin, a fasting-induced glucogenic protein hormone. Cell *165*, 566–579.

Samara, P., Karachaliou, C.E., Ioannou, K., Papaioannou, N.E., Voutsas, I.F., Zikos, C., Pirmettis, I., Papadopoulos, M., Kalbacher, H., Livaniou, E., et al. (2017). Prothymosin alpha: an Alarmin and more. Curr. Med. Chem. *24*, 1747–1760.

Sanchez-Gurmaches, J., Tang, Y., Jespersen, N.Z., Wallace, M., Martinez Calejman, C., Gujja, S., Li, H., Edwards, Y.J.K., Wolfrum, C., Metallo, C.M., et al. (2018). Brown fat AKT2 is a cold-induced kinase that stimulates ChREBP-mediated de novo lipogenesis to optimize fuel storage and thermogenesis. Cell Metab. *27*, 195–209.e6.

Scheltema, R.A., and Mann, M. (2012). SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. J. Proteome Res. *11*, 3458–3466.

Scherer, P.E., Williams, S., Fogliano, M., Baldini, G., and Lodish, H.F. (1995). A novel serum protein similar to C1q, produced exclusively in adipocytes. J. Biol. Chem. *270*, 26746–26749.

Seale, P., Conroe, H.M., Estall, J., Kajimura, S., Frontini, A., Ishibashi, J., Cohen, P., Cinti, S., and Spiegelman, B.M. (2011). Prdm16 determines the thermogenic program of subcutaneous white adipose tissue in mice. J. Clin. Invest. *121*, 96–105.

Sharp, L.Z., Shinoda, K., Ohno, H., Scheel, D.W., Tomoda, E., Ruiz, L., Hu, H., Wang, L., Pavlova, Z., Gilsanz, V., et al. (2012). Human BAT possesses molecular signatures that resemble beige/Brite cells. PLoS One *7*, e49452.

Shimizu, I., Aprahamian, T., Kikuchi, R., Shimizu, A., Papanicolaou, K.N., MacLauchlan, S., Maruyama, S., and Walsh, K. (2014). Vascular rarefaction mediates whitening of brown fat in obesity. J. Clin. Invest. *124*, 2099–2112.

Stanford, K.I., Middelbeek, R.J., Townsend, K.L., An, D., Nygaard, E.B., Hitchcox, K.M., Markan, K.R., Nakano, K., Hirshman, M.F., Tseng, Y.H., et al. (2013). Brown adipose tissue regulates glucose homeostasis and insulin sensitivity. J. Clin. Invest. *123*, 215–223.

Stern, J.H., Rutkowski, J.M., and Scherer, P.E. (2016). Adiponectin, leptin, and fatty acids in the maintenance of metabolic homeostasis through adipose tissue crosstalk. Cell Metab. *23*, 770–784.

Sun, C., Berry, W.L., and Olson, L.E. (2017). PDGFRα controls the balance of stromal and adipogenic cells during adipose tissue organogenesis. Development *144*, 83–94.

Sustarsic, E.G., Ma, T., Lynes, M.D., Larsen, M., Karavaeva, I., Havelund, J.F., Nielsen, C.H., Jedrychowski, M.P., Moreno-Torres, M., Lundh, M., et al. (2018). Cardiolipin synthesis in brown and beige fat mitochondria is essential for systemic energy homeostasis. Cell Metab. *28*, 159–174.e11.

Sutton, A.K., Pei, H., Burnett, K.H., Myers, M.G., Rhodes, C.J., and Olson, D.P. (2014). Control of food intake and energy expenditure by Nos1 neurons of the paraventricular hypothalamus. J. Neurosci. *34*, 15306–15318.

Svensson, K.J., Long, J.Z., Jedrychowski, M.P., Cohen, P., Lo, J.C., Serag, S., Kir, S., Shinoda, K., Tartaglia, J.A., Rao, R.R., et al. (2016). A secreted slit2 fragment regulates adipose tissue thermogenesis and metabolic function. Cell Metab. *23*, 454–466.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Villarroya, F., Cereijo, R., Villarroya, J., and Giralt, M. (2017). Brown adipose tissue as a secretory organ. Nat. Rev. Endocrinol. *13*, 26–35.

Wei, Y., Xiong, Z.J., Li, J., Zou, C., Cairo, C.W., Klassen, J.S., and Privé, G.G. (2019). Crystal structures of human lysosomal EPDR1 reveal homology with the superfamily of bacterial lipoprotein transporters. Commun. Biol. *2*, 52.

Wu, J., Boström, P., Sparks, L.M.M., Ye, L., Choi, J.H.H., Giang, A.-H.H., Khandekar, M., Virtanen, K.A.A., Nuutila, P., Schaart, G., et al. (2012). Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. Cell *150*, 366–376.

Yang, J., and Everett, A.D. (2007). Hepatoma derived growth factor binds DNA through the N-terminal PWWP domain. BMC Mol. Biol. *8*, 101.

Zhong, J., Krawczyk, S.A., Chaerkady, R., Huang, H., Goel, R., Bader, J.S., Wong, G.W., Corkey, B.E., and Pandey, A. (2010). Temporal profiling of the secretome during adipogenesis in humans. J. Proteome Res. *9*, 5228–5238.

Zipfel, P.F., and Skerka, C. (2009). Complement regulators and inhibitory proteins. Nat. Rev. Immunol. *9*, 729–740.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit Polyclonal anti-UCC1 (EPDR1), 1:1000	Thermo Fisher Scientific	Cat#PA5-50404; RRID: AB_2635857
Mouse Monoclonal anti-α-Tubulin (clone DM1A), 1:1000	Sigma-Aldrich (Merck)	Cat#T9026; RRID: AB_477593
Mouse polyclonal anti-CFH, 1.8 μg/ml	This paper	N/A
Biological Samples		
Human supraclavicular brown adipocytes and subcutaneous white adipocytes	Jespersen et al., 2013	N/A
Human supraclavicular brown adipose tissue	This paper	N/A
Human plasma samples	This paper	N/A
Chemicals, Peptides, and Recombinant Proteins		
Seahorse XF base medium	Agilent	Cat#103335-100
FGF-1	Immunotools	Cat# 11343557
Transferrin	Sigma-Aldrich (Merck)	Cat#T8158
T3	Sigma-Aldrich (Merck)	Cat#T5516
Rosiglitazone	Sigma-Aldrich (Merck)	Cat#R2408
Dexamethazone	Sigma-Aldrich (Merck)	Cat#D4902
IBMX	Sigma-Aldrich (Merck)	Cat#I5879
DMEM/F-12, HEPES, no phenol red	Thermo Fisher Scientific	Cat#11039-047
FBS	Thermo Fisher Scientific	Cat#10270106
Opti-MEM I Reduced Serum Medium	Thermo Fisher Scientific	Cat#31985062
Lipofectamine RNAiMAX Transfection Reagent	Thermo Fisher Scientific	Cat#13778150
Sodium pyruvate	Sigma-Aldrich (Merck)	Cat#5280
L-Glutamine	Sigma-Aldrich (Merck)	Cat#G5126
D-(+) Glucose	Sigma-Aldrich (Merck)	Cat#G7021
MitoTracker Red CMXRos	Thermo Fisher Scientific	Cat#M7512
NucBlue Fixed Cell ReadyProbes Reagent	Thermo Fisher Scientific	Cat#R37606
PBS, pH 7.4	Thermo Fisher Scientific	Cat#10010049
PowerUp SYBR Green Master Mix	Thermo Fisher Scientific	Cat#A25777
TaqMan Universal PCR Master Mix	Thermo Fisher Scientific	Cat#4305719
BODIPY 493/503	Thermo Fisher Scientific	Cat#D-3922
NORadrenalin SAD	Amgros I/S	Cat#745661
Insulin (Actrapid)	Novo Nordisk	Cat#A10AB01
EPDR1 protein	Novo Nordisk	N/A
High fat diet for rodents with lard (60% kJ fat)	Ssniff	D12495
Critical Commercial Assays		
High-Capacity cDNA Reverse Transcription kit	Thermo Fisher Scientific	Cat#4374966
Direct-zol RNA miniprep kit	Zymo Research	Cat#R2063
Seahorse XFe96 FluxPak	Agilent	Cat#102416-100
EPDR1 ELISA kit	MyBioSource	Cat#MBS9316185
Seahorse XF Cell Mito Stress Test Kit	Agilent	Cat#103015-100
Bio-Rad Protein Assay	Bio-Rad	Cat#5000006
Deposited Data		
Raw and processed MS data	ProteomeXchange Consortium	ProteomeXchange: PXD008541

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Mouse: Male wild-type C57Bl/6NRj	Janvier	N/A
Mouse: Female C57Bl/6NRj	Janvier	N/A
Mouse: Male *Epdr1*$^{-/-}$	Toronto Centre for Phenogenomics (TCP)	N/A
Mouse: wild-type C57Bl/6N	Toronto Centre for Phenogenomics (TCP)	N/A
Oligonucleotides		
TaqMan primers	Life Technologies	Table S5
SybrGreen primers	TAG Copenhagen	Table S5
Short interfering RNAs	Dharmacon	Table S5
Software and Algorithms		
Graphpad Prism 8.0 for statistical analysis	https://www.graphpad.com/	N/A
MaxQuant	https://maxquant.org/	Free
Perseus	http://www.coxdocs.org/doku.php?id=perseus:start	Free

LEAD CONTACT AND MATERIALS AVAILABILITY

Requests for reagents and resources should be directed to the Lead Contact, Camilla Scheele (cs@sund.ku.dk).

Disclosure of Limited Availability of Biological Material

We hereby disclose that the availability of biological material including the human cell cultures and tissue, is dependent on specific permission from the Danish Data Protection Agency and on the researcher's adherence to the specifications of this permission.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Supraclavicular (Brown) and Subcutaneous (White) Fat Precursor Cells

Brown fat precursor cells were isolated from the supraclavicular adipose depot of a cohort of adult humans (n = 21), as previously reported along with a subset of the cell cultures (Jespersen et al., 2013). These non-immortalized cell cultures were differentiated *in vitro* and the five cultures (from different individuals) displaying the best differentiation capacity in terms of lipid accumulation and the highest induction of UCP1 expression in response to norepinephrine (NE) were included in the study (Figure S1). The five brown fat precursor cultures were derived from two women and three men. White fat precursor cells were obtained from the subcutaneous abdominal region of three women and two men, with equal distribution of age and body mass index (BMI) as the donors of the brown fat cell cultures. All subjects provided written informed consent. The Scientific-Ethics Committees of the Capital Region and Copenhagen and Frederiksberg Municipalities Denmark approved the study protocols, journal numbers H-A-2009-020, H-A-2008-081, and (KF) 01-141/04, respectively, and the studies were performed in accordance with the Helsinki declaration.

Human Subjects

The human tissue and plasma samples used for this study are a subset from a cohort of subjects undergoing surgery for benign goiter. All subjects provided written informed consent prior to participation. The Scientific-Ethics Committees of the Capital Region of Denmark approved the study protocol and amendments, and the study was performed in accordance with the Helsinki declaration journal number H-1-2014-015. A subset of 30 non-obese normal glucose tolerant subjects were included in the current study. Surgical biopsies were obtained from the deep neck region and were snap-frozen in liquid nitrogen until RNA isolation was performed. Subject characteristics of EPDR1 positive and EPDR1 negative subjects are presented in Table S4.

Mouse Models

EPDR1 Injections

Animal studies were approved by the Animal Experimentation Inspectorate of the Danish Ministry of Justice no 2014-15-0201-00181. Mice were raised under Specific Pathogen Free (SPF) conditions. Acute injections: The animals were single housed and maintained at a 12-h light-dark cycle in temperature (30°C–32°C) and humidity (50–60%) controlled cabinets, with free access to standard chow (Altromin 1314F, pellets) and tap water. Chronic injections: The animals were maintained at a 12-h light-dark cycle in temperature (20°C–22°C) and humidity (50–60%) controlled cabinets, with free access to tap water and high fat diet (60 kJ% fat) from 6 weeks of age. At 22 weeks of age, the animals were single housed and placed in calorimetric chambers. Estimation of sample sizes were based on previous studies with similar metabolic readouts (Sustarsic et al., 2018).

Epdr1 *Knockout Mouse Model*

Animal studies were approved by the Toronto Centre for Phenogenomics Mt. Sinai Hospital, in Toronto. Mice were raised under SPF conditions. The mouse line C57BL/6N-Epdr1<tm1a(NCOM)Mfgc>/Tcp was generated at the Toronto Centre for Phenogenomics (TCP) and obtained from the Canadian Mouse Mutant. Repository as part of the NorCOMM2 project from NorCOMM ES cells (Bradley et al., 2012). Male mice generated from this mouseline were fed a regular standard rodent chow (18% kcal from fat, 2018 Harlan Teklad, Mississauga, ON) for 20 weeks. Epdr1-/- mice were gender and age matched to wild-type mice generated at the TCP facility that were not littermate controls.

METHOD DETAILS

Human Primary Adipocyte Culture Conditions

The protocols for isolation and differentiation of human fat precursor cells has been thoroughly described and discussed (Larsen et al., 2019). Cells were plated and maintained in DMEM F12 (Thermo Fisher Scientific) containing 10% FBS, 1% Penicillin/Streptomycin and 1nM FGF1 (Immunotools). Two days post confluence (designated day 0) the cells were induced to differentiate by replacing the medium with DMEM/F12 with 10 µg/ml Transferrin, 2 nM T3, 100 nM Insulin, 100 nM Dexamethasone, 200 nM Rosiglitazone and 540 µM 3-Isobutyl-1-methylxanthine (IBMX). At day 3, the medium was replaced with the same medium composition, with the exemption of IBMX. The cells were refed with new medium at day 6 and day 9 with the exemption of Rosiglitazone. At day 12, cells were considered fully differentiated mature adipocytes. Secretome analysis was performed on supraclavicular (n=5) and subcutaneous (n=5) cells. For secretome analysis, fully differentiated cells were washed with DMEM/F12 and subsequently incubated for 2 h in DMEM/F12 containing 1% penicillin-streptomycin. Serum-starved cells were stimulated with 10 µM norepinephrine (NE) (Sigma-Aldrich) for 4 h. Cell culture media (2 ml) was collected for secretome analysis while cells were harvested for RNA analysis. To address whether cell viability was affected by NE treatment, cells were stained according to the manufacturer's protocol, applying a LIVE/DEAD fluorescent assay (Thermo Scientific) visualized using an EVOS FL fluorescent microscope (Thermo Scientific).

siRNA Mediated Knockdown of EPDR1 in Adipocytes

At day 0 of the differentiation program, adipocytes were transfected with 22.2 nM siRNA targeting EPDR1 or a non-targeting pool (Dharmacon) (Table S5) using Lipofectamine RNAiMAX Reagent (Thermo Fisher Scientific) according to the manufacturer's instructions. The cells were then cultured as described in the previous section. To investigate whether our siRNA mediated knockdown of EPDR1 (using a pool targeting four different sites on the EPDR1 mRNA transcript and a non-targeting siRNA control) targeted the secreted isoforms of EPDR1 (transcript variant 1 and transcript variant 2), we designed qPCR primers specifically targeting either of these transcripts. A unique qPCR assay for the intracellular transcript variant 3 could not be designed, but a qPCR assay for assesment mRNA of all three transcript variants of EPDR1 was designed (Figures S4A–S4C; Table S5). To ensure that the knockdown was specific for *EPDR1* we also measured mRNA expression of *SFRP4*, which is located in the same locus as EPDR1, but transcribed from the opposite strand (Figure S4D). We validated the effects of EPDR1 knockdown on NE-induced thermogenic gene expression by transfecting with two additional siRNAs (Dharmacon) (Table S5), each targeting only one site of the mRNA molecule (Figure S2G). Finally, to investigate whether the observed reduction in induction of thermogenic gene expression following NE- stimulation in brown adipocytes following knockdown of *EPDR1* was specific for brown adipocytes, we also measured thermogenic gene expression in human white adipocytes (Figures S4E–S4G).

RNA Isolation and Quantitative Real-Time PCR of Adipocytes

Total RNA from human adipocytes was isolated with TRIzol (Thermo Fisher Scientific), according to the manufacturer's recommendations. RNA concentrations were measured using Nanodrop 1000, and 250 ng of RNA was used as input for subsequent cDNA synthesis with High Capacity cDNA Synthesis Kit (Thermo Fisher Scientific). Primer sequences can be found in Table S5. Target mRNA was normalized to PPIA and calculated using the comparative delta-delta-Ct method.

Oxygen Consumption Measurements in Adipocytes

Mitochondrial respiration rates were assessed using the XFe96 Extracellular Flux Analyzer (Agilent Technologies). Cells were plated at confluence (7000cells/well) and transfected and differentiated as described above. The cell medium was replaced with Seahorse Base medium without phenol red, supplemented with 1 µM L-Glutamine, 2 µM Na_2PO_4 and 25 mM Glucose pH 7.4 1 h prior to respiration measurements. Oxygen consumptions rates were followed under basal conditions, NE (1 µM) injection and finally oligomycin injection (20 µM).

Lipid and Mitochondrial Staining

Mature adipocytes were incubated with DMEM/F12 containing 0.2µM MitoTracker Red CMXRos (Thermo Fisher Scientific) for 20 min. Cells were then washed 3 times with PBS and fixed with 4% formaldehyde (Sigma-Aldrich) for 15 min. Fixed cells were washed again 3 times with PBS, and lipid were stained with 0.5 mM Bodipy (Thermo Fisher Scientific) for 20 min. Subsequently nuclei were stained with NucBlue Fixed ReadyProbe Reagent (Thermo Fisher Scientific) for 7 min. Cells were washed 3 times with PBS and visualized with EVOS FL imaging system (Thermo Fisher Scientific).

ELISA

EPDR1 Plasma concentrations were determined using a commercially available ELISA kit from Mybiosource. Briefly, 50 μl plasma were loaded in duplicates and EPDR1 abundance was determined. The detection limit of the kit was 30 ng/ml. UCP1 mRNA levels were determined using qPCR. Subject characteristics of EPDR1 positive and EPDR1 negative subjects are presented in Table S4.

EPDR1 Production

Production of recombinant human EPDR1 (UNIPROT entry Q9UM22, isoform 1, amino acids 39-224) was performed by transfecting expi293F cells growing in suspension culture in Expi293 Expression medium (ThermoFisher Scientific, cat# A1435101) with a mix of ExpiFectamine 293 Reagent (ExpiFectamine 293 Transfection Kit, ThermoFisher Scientific, cat# A14525) and plasmid DNA encoding the relevant sequence. Transfection Enhancers 1 and 2 from the ExpiFectamine 293 Transfection Kit were added the day after transfection. The cell culture was harvested 4 days after transfection. The recombinant EPDR1 secreted into the culture medium contained a short N-terminal sequence compatible with affinity purification and the protein was purified by affinity chromatography followed by size exclusion chromatography according to standard chromatographic methods. The purified protein solution was sterilized by filtration through a 0.2 mm filter unit.

Assessment of the Purity of EPDR1 Recombinant Protein

The purity of the protein was analyzed by SDS-PAGE. Briefly, Histidine tagged EPDR1 was loaded (6 μg) on SDS PAGE either directly or after three freeze thawing cycles (3 μg). The gel was stained with Coomassie blue stain (Figure S3A). A silver staining was also performed. Briefly, 2.92 ng and 3.18 ng protein from batch 2 and batch 3 was loaded on 4-12% Bis-Tris Nu PAGE gels (Thermo Fischer Scientific). Silver staining was done using SilverXpress, Silver staining kit (Thermo Fisher Scientific) and performed as recommended by the manufacturer (Figure S3A). The protein purity was also checked by using size exclusion chromatography (SEC) and protein identity was confirmed by LC MS-MS (Figure S3A). The protein solution was analyzed for endotoxin levels: batch 1: 0.059 EU/mg (used for indirect calorimetry study) and batch 2: 0.055 EU/mg (used for daily injections of EPDR1 over 21 days) batch 3: 0.68 EU/mg (used for [18F] FDG PET/CT imaging). All batches were confirmed to be suitable for animal studies. To further ensure that the observed increases in oxygen consumption following acute EPDR1 injection in C57BL/6NRj mice housed at thermoneutrality, was specific for EPDR1 and were not associated to artifacts coming from protein preparation, we measured metabolic parameters following injection of two other proteins. The two proteins (GM2A and SBSN) were also specifically quantified in the culture media from brown adipocytes and proteins were prepared/purified in the same manner as EPDR1. All further experimental conditions were also identical between setups (Figures S4H–S4K).

Small Animal [18F] FDG PET/CT Imaging

Eight-week-old female mice were group housed at thermo neutrality on a 12:12 hour light/dark cycle (lights on at 9 PM and off at 9 AM). [18F] FDG was administered intravenously between 10 AM and 2 PM. The average radioactive dose was 4.8 MBq (range: 3.6–5.8 MBq). EPDR1 (2 mg/kg) or PBS was administered subcutaneously 180 minutes prior to FDG injection, and CL 316,243 (1 mg/kg) or PBS was administered subcutaneously 15 min prior to FDG administration. Animals were fasted and housed in the dark at 29°C following injection of EPDR1.

Small animal PET/CT (Inveon Multimodality PET/CT scanner; Siemens) was performed 1 hour after FDG administration. Mice were anaesthetized by sevoflurane 40 minutes after FDG injection until the end of the imaging session. Heating was applied in order to maintain normal body temperature. PET data were acquired in list mode for 240s, and images were reconstructed using a 3-dimensional maximum a posteriori algorithm with CT-based attenuation and scatter correction. CT images were acquired using 360 projections, 65 kV, 500 mA, and 430 ms exposure and reconstructed with an isotropic voxel size of 0.210 mm. Images were analyzed using the Inveon software (Siemens). Quantitative analysis of the [18F]FDG uptake was performed by manually drawing region of interests over the areas containing iBAT based on the CT images. The FDG uptake was expressed as % injected dose per gram tissue (%ID/g). Animals were euthanized after the imaging session and iBAT, iWAT, Soleus muscle and heart were excised, weighted, submerged in RNAlater for subsequent RNA isolation and gene expression analysis.

RNA Isolation and Quantitative Real-Time PCR following EPDR1 Injection

Total RNA was extracted from iBAT and WAT depots that was stored in RNAlater. The tissues were processed in Tri Reagent (Qiagen) using a TissueLyser II system (Qiagen) and RNeasy Lipid Tissue Mini Kit according to instructions from the manufacturer. RNA concentrations were measured using Nanodrop 2000, and 500 ng of RNA was used as input for subsequent cDNA synthesis with High Capacity cDNA Synthesis Kit (Thermo Fisher Scientific). Primer sequences can be found in Table S5. Target mRNA was normalized to peptidylprolyl isomerase A (PPIA) and results were calculated using the comparative delta-delta-Ct method.

Indirect Calorimetry following EPDR1 Injection

Twelve-week (acute injections) or twenty-week (chronic injections) old male C57BL/6NRJ mice (Janvier Labs, France) were used for metabolic measurements. All animals were acclimatized in habituation cages one week before they were placed in the TSE phenomaster (TSE systems GmbH, Bad Homburg, Germany) system. Here they were acclimatized for another 4 days prior to the initiation of the injection experiments and measurements. Mice were randomized according in two groups of eight mice taking their weight into account. For the acute injection study, EPDR1 (2 mg/kg in PBS) or PBS as vehicle was administered subcutaneously 20 min prior to the dark cycle. Metabolic measurements were followed for a total of 43 hours, and the initial 12-hour dark period was used to assess treatment effects. For the chronic injection study, EPDR1 (2 mg/kg in PBS) or PBS as vehicle was administered subcutaneously in the middle of the light cycle. Metabolic measurements were followed for a total of 21 days and used to assess treatment effects.

Body Composition, Food Intake and Indirect Calorimetry in EPDR1$^{-/-}$ Mice

Body composition was measured using Echo nuclear magnetic resonance system (Echo Medical Systems, Houston, TX) MRI, in the non-fasted state around 10 am at 20 weeks of age. Food intake was averaged over a 3-day period where mice were housed individually. To measure oxygen consumption (ml/hr), carbon dioxide production, respiratory exchange ratio, activity, and energy expenditure (kcal/hr) the Comprehensive Lab Animal Monitoring System (CLAMS; Columbus Instruments, Columbus OH) was used when mice were 9-11 weeks of age.

Ex Vivo Oxygen Consumption

iBAT was immediately harvested from non-fasted animals after CO2 euthanasia. iBAT was split into 3 depots of similar size (~10 mg of scaled weight) and kept at DMEM high-glucose media at 37°C. Tissues were then washed in filtered respiration buffer (PBS, 0.02% fatty acid free BSA, 25-mM glucose, 0.01% (vol/vol) of 100 mM Na pyruvate (Sigma)), minced with scissors, re-suspended in 1-ml respiration buffer and placed into a Mitocell chamber (MT200A, Strathkelvin Instruments, North Lanarkshire, Scotland) also kept at 37°C with a Clark electrode (Strathkelvin). Recordings were normalized to tissue scaled weight and each data point represents an average oxygen consumption from 3 depots from each animal.

RNA Isolation and Quantitative Real-Time PCR in EPDR1$^{-/-}$ Mice

Total RNA was extracted from iBAT that was harvested immediately from mice after CO2 euthanasia in the non-fasted state and stored at -80°C until processed in Tri Reagent (Molecular Research Center, Cincinnati, ON) using a TissueLyser II system (Qiagen, Germantown, MD). First strand cDNA was synthesized from DNase I-treated total RNA using the SuperScript III and random hexamers (Thermo Fisher Scientific, Markham, ON). Gene expression levels were quantified by real-time PCR using a QuantStudio System and TaqMan Gene Expression Master Mix and Assays (Thermo Fisher Scientific). Primer/probes were purchased from Thermo Fisher Scientific (Table S5). qPCR data were analyzed by 2-DeltaDeltaCt method, and expression levels for each gene were normalized to Tbp (TATA-box-binding protein).

Secretome Analysis by Mass Spectrometry

A high-resolution mass spectrometry (MS)-based approach was used to detect a highly complex secreted protein mixture from human primary brown and white adipose cell cultures. Serum free cell supernatants were trypsin digested, and the resulting peptide mixtures were directly analyzed in a single-run LC-MS format. We performed liquid chromatography with 2 h gradient and analyzed peptides on bench top quadrupole-Orbitrap instrument with very high sequencing speed and high mass accuracy in MS and MS/MS modes (Michalski et al., 2011). Label-free quantification of the MS data is performed in the MaxQuant environment while bioinformatics analysis was done with Perseus software (Tyanova et al., 2016).

Sample Preparation for Secretome and Cellular Proteome

Secretome analysis of the conditioned media from brown and white adipocytes was performed as described before (Deshmukh et al., 2015) with slight modifications. Proteins in conditioned media were denatured with 2 M urea in 10 mM HEPES pH 10 by ultrasonication on ice and acetone precipitated (overnight). Protein pellets were washed with 80% acetone and suspended in urea (6 M) and thiourea (2 M) buffer (pH 8). Proteins were reduced with 10 mM dithiothreitol for 40 min followed by alkylation with 55 mM iodoacetamide for 40 min in the dark. Proteins were digested with 0.5 μg LysC (Wako) for 3 h and digested with 0.5 μg trypsin for 16 h at room temperature. The digestion was stopped with 0.5% trifluoroacetic acid, 2% acetonitrile. Peptides were desalted on reversed phase C18 StageTips. The peptides were eluted using 20 μl of 60% acetonitrile in 0.5% acetic acid and concentrated in a SpeedVac. Concentrated peptides were acidified with 2% acetonitrile, 0.1% trifluoroacetic acid in 0.1% formic acid. Samples for cellular proteome of BAT and WAT cells were prepared according to iST protocol (Kulak et al., 2014).

LC MS/MS Analysis

The peptides from the cell culture media were analyzed using LC-MS instrumentation consisting of an Easy nanoflow UHPLC (Thermo Fischer Scientific) coupled via a nanoelectrospray ion source (Thermo Fischer Scientific) to a Q Exactive mass spectrometer (Thermo Fischer Scientific) (Michalski et al., 2011). Peptides were separated on a 50-cm column with 75-μm inner diameter packed in-house with ReproSil-Pur C18-aq 1.9 μm resin (Dr. Maisch). Peptides were loaded in buffer containing 0.5% formic acid and eluted with a 160 min linear gradient with buffer containing 80% acetonitrile and 0.5% formic acid (v/v) at 250 nL/min. Chromatography and column oven (Sonation GmbH) temperature were controlled and monitored in real time using SprayQC (Scheltema and Mann, 2012). Mass spectra were acquired using a data dependent Top8 method, with the automatic switch between MS and MS/MS. Mass spectra were acquired in the Orbitrap analyzer with a mass range of 300-1650 m/z and 70 000 resolution at m/z 200. HCD peptide fragments were acquired with a normalized collision energy of 25. The maximum ion injection times for the survey scan and the MS/MS scans were 20 and 220 ms, and the ion target values were set to 3e6 and 1e5, respectively. Data were acquired using Xcalibur software. Peptides from BAT and WAT cells were analyzed using LC-MS instrumentation consisting of an Easy nanoflow UHPLC coupled via a nanoelectrospray ion source to a latest generaton Q Exactive mass spectrometer (HFX) (Kelstrup et al., 2018). Peptides were separated on a 50 cm column with 75 μm inner diameter packed in-house with ReproSil-Pur C18-aq 1.9 μm resin (Dr. Maisch). Peptides were loaded in buffer containing 0.5% formic acid and eluted with a 100 min linear gradient with buffer containing 80% acetonitrile and 0.5% formic acid (v/v) at 350 nL/min. Mass spectra were acquired using a data-dependent Top15 method, with the automatic switch between MS and MS/MS. Mass spectra were acquired in the Orbitrap analyzer with a mass range of 300-1650 m/z and 60 000 resolution at m/z 200. HCD peptide fragments were acquired with a normalized collision energy of 27. The maximum ion injection times for the survey scan and the MS/MS scans were 20 and 25 ms, and the ion target values were set to 3e6 and 1e5, respectively.

Computational MS Data Analysis

The raw files for secretome and cellular proteome were analyzed in the MaxQuant environment (Tyanova et al., 2016). The initial maximum allowed mass deviation was set to 6 ppm for monoisotopic precursor ions and 20 ppm for MS/MS peaks. Enzyme specificity was set to trypsin, defined as C-terminal to arginine and lysine excluding proline, and a maximum of two missed cleavages was allowed. A minimal peptide length of six amino acids was required. Carbamidomethylcysteine was set as a fixed modification, while N-terminal acetylation and methionine oxidation were set as variable modification. The spectra were searched by the Andromeda search engine against the human UniProt sequence database with 248 common contaminants and concatenated with the reversed versions of all sequences. The false discovery rate (FDR) was set to 1% for peptide and protein identifications. The peptide identifications across different LC-MS runs were matched by enabling the 'match between runs' feature in MaxQuant with a retention time window of 30 s. If the identified peptides were shared between two or more proteins, these were combined and reported in protein group. Contaminants and reverse identifications were removed from further data analysis. Protein quantification was based on the Max LFQ algorithm integrated into the MaxQuant software (Cox et al., 2014).

Bioinformatics analysis was performed with the Perseus software (Tyanova et al., 2016) (http://www.perseus-framework.org). Categorical annotation was supplied in the form of KEGG pathways, Keywords (UniProt), and Gene Ontology (GO) (biological process (BP), molecular function (MF) and cellular component (CC)). All annotations were extracted from the UniProt database. To define the secretome of brown and white adipocytes, we applied previously described computational workflow on all proteins identified in the media (Deshmukh et al., 2015). Briefly, signal peptide-containing proteins were categorized as 'classical' secreted proteins while protein annotated to 'extracellular location' (GOCC) or 'secreted' (UniProt, Keywords) were classified as 'non-classical' secreted proteins. While comparing proteins identified in the cell media from brown or white adipocytes with Vesiclepedia (Kalra et al., 2012) and ExoCarta databases(Keerthikumar et al., 2016), we included only evidence at the protein level.

The global comparative analysis was performed on LFQ intensities. We used very stringent criteria for MaxLFQ-based quantification (min ratio count 2 in MaxQuant). Moreover, we included only those proteins which were quantified at least 2 times in at least one group (i.e. Brown_woNE vs White_woNE) Due to the randomness of peptide sampling in shotgun proteomics, the quantification of several proteins is missing for some samples. The data was imputed to fill missing abundance values by drawing random numbers from a Gaussian distribution. These parameters have been tuned in order to simulate the distribution of low abundant proteins best. To investigate differences between brown and white adipocyte secretomes, we compared brown and white adipocyte media proteome under non-stimulated (woNE) conditions while the effect of NE-stimulation was investigated by comparing stimulated and non-stimulated conditions (Brown_woNE vs Brown_NE/ White_woNE vs White_NE). These comparisons were made using a two-sample t-test in Perseus with FDR 0.05. Proteins which were regulated by 1.5-fold (Log$_2$) were considered as significantly different proteins (Table S2). Two-sample t-test (volcano plot), hierarchical clustering and annotation enrichment were based on label-free quantitation of the samples. Hierarchical clustering of significantly different proteins was performed after Z-score normalization. We then performed Fisher exact test on significantly different proteins (background total quantified proteins), testing for enrichment or depletion of any annotation term in the cluster compared to the whole matrix.

Downstream analysis of global cellular proteome was performed in Perseus software. The statistical comparisons were made using a two-sample t-test in Perseus with FDR 0.05. Protein abundances were compared siCtr vs siEPDR1 in human brown adipocytes. While comparing siCtrl vs siEPDR1 in human brown adipocytes, proteins with >1.2 log$_2$ FC were considered as significantly differentially regulated.

Predictive Multiplexed Selective Ion Monitoring (pmSIM)

Predictive multiplexed selected ion monitoring (pmSIM) targeting in MaxQuantLive (version 1.0) for peptide and protein quantification relies on the real-time recalibration of retention time and mass accuracy based on background peptides and the identification of the heavy labeled counterpart of the to be quantified endogenous peptide of interest. Therefore, the heavy labeled EPDR1 peptide SYETWIGIYTVK was equally spiked into ready to inject brown and white fat secretomes and analyzed by data dependent acquisition on an orbitrap QE-HFX platform followed by MaxQuant (1.6.7) to identify retention time, intensity, and m/z of background peptides and the EPDR1 peptide SYETWIGIYTVK for pmSIM targeting of the endogenous counterpart. A total of 2743 shared realtime correction peptides were selected from the 'evidence' output file, removing potential contaminants, reverse database hits, modified sequences, miss cleaved peptides and peptides eluting in a time frame of <10 and >90min. For realtime correction, the initial retention time tolerance was ± 20 min and the final retention scale factor was set to three times the standard deviation of the recorded elution time with a mass tolerance of ± 9 ppm, and an intensity threshold of 10^{-5}. MaxQuant.Live pmSIM experiments of three biological replicates for brown (BAT5, BAT13, BAT19) and two for white fat (WAT11a, WAT13a) were performed with a 1st isolation window and a +0.2 Th offset and acquired with a resolution of 120,000 at *m/z* 200. The heavy and light channels were multiplexed in a single scan. A maximum of 5 x 10^5 ions were collected in each channel with a maximum ion injection time of 120 ms for the light and heavy channel, respectively.

Data analysis of the pmSIM experiment was performed with the Skyline (Version 4.2.0.19009) and XCalibur (3.1.66.10) software suites. For data analysis light and heavy channel intensities for the EPDR1 peptide SYETWIGIYTVK were extracted from Skyline. Light channel intensities were normalized against the total background peptide MS1 intensity to take sample specific properties into account followed by normalization against the heavy channel. Fold change difference calculation was performed on the median white fat intensities.

Western Blot Analysis on Cell Media

For complement factor H, we validated proteomics data using western blot analysis on conditional media. The conditioned media was concentrated (3 different Brown adipocyte cell strains and 3 different White adipocyte cell strains) through two centrifugation rounds using an Amicon Ultra-4 3K and 10 K filter devices (Merk Millipore, USA). Purified proteins (FB Hycult HC2129, FH CompTech A137, FI CompTech A138) were used as positive control. Growth media was used as negative control. Precision Plus Protein Blue (Bio-Rad, USA) was used as molecular weight (MW) marker. The CFH antibody was produced in-house by Prof. Peter Garred's research group using mouse hybridomas and was used in a concentration of 1.8 μg/ml. The secondary antibody was Rabbit α-mouse HRP (Dako P0260 Lot# 00062101), utilized at a dilution of 1:10,000.

Western Blot Analysis on Cell Lysate

SiRNA mediated knockdown of EPDR1 was validated using western blot analysis on cell protein lysate. In short, the mature differentiated cells were washed twice with ice cold PBS and the lysed in lysis buffer containing 20 mM tris-HCl, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton-X, 2.5 mM $Na_4P_2O_7$, 1 mM β-glycerophosphate, 1mM Na_3VO_4, Complete Mini, Protease Inhibitor Cocktail. Lysates were centrifuged for 15 min at 13.000 g at 4°C. Protein concentration was determined using the Bio-Rad Protein (Bio-Rad, California, USA). 3.6 μg of protein lysate was loaded on Bis-Tris SDS-page gels and subject to electrophoresis, using iBright Prestained Protein ladder (Thermo Fisher Scientific) to determine MW of detected bands. Proteins were transferred to PVDF membranes by semi-dry transfer for 7 min with Pierce Power blot cassette. Membranes were blocked in FSG and incubated for 16 h with anti-EPDR1 at 1:1000 in 1% FSG (Thermo Fisher Scientific) or anti-α-Tubulin (Sigma-Aldrich) (1:1000 in 1% FSG). Bands were detected with IRDye secondary antibodies (LI-COR) at 1:5000 and visualized using the Odyssey Fc Imaging System (LI-COR).

QUANTIFICATION AND STATISTICAL ANALYSIS

Secretome data was collected from five separate brown adipocytes cultures derived from five different human donors and five separate white adipocytes cultures derived from five different human donors, thus representing biological replicates. Statistical analysis of proteomics data was performed using Perseus software (Tyanova et al., 2016). The details of the proteomics data analysis can be found in the STAR Methods in the section "Computational MS data analysis". Statistical analyses of the rest of the experiments were performed with GraphPad Prism software. No methods were used to further determine whether the data met assumptions of the statistical approach. For the cell experiments in Figures 3E and 3H–3J, a representative brown fat culture was utilized and technical replicates (n=3) from independent experiments are presented. To account for a putative batch to batch variation, effects of EPDR1 knockdown and norepinephrine was assessed using a Mixed-effects analysis in Graphpad Prism 8, using repeated measurements for both EPDR1 knockdown and norepinephrine analysis. In case of a significant overall effect of EPDR1 knockdown, specific effects were assessed with Sidaks multiple comparison's test for which p-values are presented in the graphs. In the interpretation of the qPCR data, it should be considered that no correction for multiple testing was performed. All 21 genes measured was reported. Genes assessed in the data set presented in Figure 3 that were not significantly regulated are presented in Figures S2I and S2J using the same statistical approach. Additional statistical details, including exact value, description and exclusion of n as well as definition of center, and dispersion and precision measures, are specified in the figure legends. A p-value below 0.05 was considered significant.

DATA AND CODE AVAILABILITY

The assession number for the mass spectrometry proteomics data reported in this paper is ProteomeXchange: PXD008541. The data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al., 2019) partner repository.

7. Acknowledgements

My deepest gratitude goes to the following people:

First of all, I want to thank Matthias Mann for being such a great mentor and supervisor. Thank you for being so approachable on a daily basis and being so enthusiastic about mass spectrometry, proteomics and so much more. Also, thank you for creating such a vibrant cutting-edge atmosphere in the department, which lets us all grow at the speed of light, still leaves room to pursue the craziest ideas and sending me around the globe! ☺

Alison Dalfovo and Theresa Schneider for your untiring and amazing support!! An important person once said: “Without you, we can shut up shop!” ☺

Florian Meier for all the support, amazing projects we pushed together, being so approachable and a great teacher. I will never forget our first conversation: FM: “Do you like to work with large machines?”

Heiner and Scarlet Koch for being so helpful in the initial phase of my PhD enabling such an easy transition to MS-based proteomics and all the support at Bruker.

The Italian gang comprising Igor, Antonio and Matteo – Thanks for just being who you are. You make every day in the department even better ☺

Patricia and Marvin for creating such a nice atmosphere and pushing timsTOF-based proteomics to the next level! – “Was isch des??” - “Cantuccini?” ☺

Katerini, our Greek princess, smiling trouble, and Lipidomics queen – “Lipidomics does not need more standardization, it needs more Vasilopoulou!” – A FrapOuzo to that! Nothing more to say.

Fabian Coscia and Andreas Mund for the fun and exciting times, the introduction to really deep proteomes one can actually spatially dissect and visualize in beautiful heatmaps! ☺

André for being such an interactive person and introducing me to alcohol-free “beer” and cold tea!

Özi for all the Turkish sweets, fun times and making my start in the department easy!

Lisa, Maria, Ashok, Fynn, Jane for all the amazing conversations and fun times!

Maks – Maksimalkraft; Kinder Maksiking; Dönerstag, Maksmalstift; ELEMU; Pflanziska!

Lili, Alberto, Michi and all the other people from Copenhagen!

All colleagues from Bruker Daltonik and EvoSep, in particular Markus Lubeck, Oliver Raether, Amalia Apalategui, Nicolai Bache, Ole Bjeld Horning and everyone else – A lot would have not been possible without you!

All my collaboration partners from all across the world, especially Jin Chen, Michele Solimena, Manuel Leonetti and Dan Itzhak.

Everyone in the Mann department for creating such an amazing atmosphere!

I would like to thank my family for all your support, no matter what I did, where I was and which decisions I made – It is you who made all this possible and it is impossible to express how grateful I am.

Johanna, my love, thank you for being the most valuable part of this whole journey since 2012 ☺

All my friends.