Data Acquisition Methods for Next-Generation Mass Spectrometry-Based Proteomics

Florian Meier

aus

Neunkirchen (Saar), Deutschland

2018

II

Erklärung

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

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Martinsried, 21.08.2018 _____

Florian Meier

Dissertation eingereicht am   03.07.2018

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: Prof. Dr. Bernhard Küster

Mündliche Prüfung am   23.07.2018

IV

Summary

Of all biomolecules, proteins are arguably the most important gears in our cellular machinery when it comes to biological function. Mass spectrometry (MS)-based proteomics has become the method of choice to study proteomic systems in a global and unbiased manner. Yet, it still trails 'next-generation' genomics and transcriptomics technology in terms of coverage, throughput and sensitivity. In this thesis, I present three MS acquisition strategies that break through longstanding technological limitations and facilitate comprehensive and high-throughput proteomics.

Isobaric labeling enables quantification of multiple samples in a single analysis and thereby increases throughput. In a first project, I established a tailored acquisition strategy for a new generation of isobaric labels termed EASI-tag. As opposed to previous technologies, the EASI-tag method is interference-free and therefore allows multiplexed and accurate quantification, also on widely used tandem mass spectrometers.

A core subject of my PhD was establishing ion mobility spectrometry as an additional dimension of separation in MS-based proteomics. This work builds on a high-resolution and high-speed quadrupole time-of-flight platform, which was equipped with a trapped ion mobility (TIMS) device. Making use of the sensitivity and flexibility of the TIMS device, we developed 'parallel accumulation – serial fragmentation' (PASEF), which effectively multiplies peptide sequencing speed and sensitivity by first storing all ions and then switching the quadrupole isolation window synchronously with ion mobility separation. The PASEF method has now become an integral part of a full-fledged commercial proteomics instrument.

The third method, termed BoxCar, addresses a fundamental limitation of Orbitrap mass analyzers, which are the main workhorses in proteomics laboratories worldwide. For reasons detailed in this thesis, in practice, less than 1% of all ions are used for mass analysis in full scans. BoxCar increases this fraction up to a factor of 10 by dividing the entire mass range into multiple narrow segments or 'boxes'. This thesis establishes an increase in dynamic range of the mass analysis of about one order of magnitude, which allowed detection of 10,000 proteins in 100 min from mouse brain tissue at the MS1-level. The method is particularly beneficial for analyses that are limited by a large protein abundance range, as is often the case in a clinical context. In our laboratory, BoxCar has already dramatically improved proteomics studies of human heart and plasma samples.

In summary, this PhD thesis provides the basis for next-generation MS acquisition methods that promise to elevate proteomics to the next level, closer to the ultimate goal of complete and ubiquitous proteomes. Importantly, the fundamental concepts and methods developed here are generic and can seamlessly be applied in other MS-based omics fields facing similar challenges, fox example metabolomics.

Table of Contents

Abbreviations

AGC	automatic gain control
CCS	collisional cross section
CID	collision induced dissociation
DDA	data-dependent acquisition
DIA	data-independent acquisition
DMS	differential mobility spectrometry
DSSO	disucccinimidyl sulfoxide
DTIMS	drift-tube ion mobility spectrometry
EASI-tag	easily abstractable sulfoxide based isobaric tag
ESI	electrospray ionization
FAIMS	field-asymmetric ion mobility spectrometry
HCD	higher energy collisional dissociation
IIT	ion injection time
IMS	ion mobility spectrometry
K0	ion mobility coefficient
LC	liquid chromatography
LFQ	label-free quantification
MRM	multiple reaction monitoring
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NHS	N-hydroxysuccinimide
PASEF	parallel accumulation - serial fragmentation
ppm	parts-per-million
PTM	post-translational modification
rf	radiofrequency
Th	Thomson (unit of m/z)
TIMS	trapped ion mobility spectrometry
TMT	tandem mass tag
TOF	time-of-flight
TWIMS	traveling wave ion mobility spectrometry

I. Introduction

A Systems Biology Perspective on Proteome Diversity and Complexity

The decoding of the complete human genome more than a decade ago[1–3] marked a milestone in science that has revolutionized our understanding of biology, medicine and evolution[4]. This breakthrough was preceded by sequencing of the yeast[5], C. elegans[6], and drosophila genomes[7]. While these pioneering efforts to reveal the base pair sequence of nucleic acids built on early work from Frederick Sanger and colleagues[8], the successful completion of the human genome project propelled the development of higher throughput and more cost-efficient 'next-generation' sequencing (NGS) technologies[9,10]. Massive parallelization and two-dimensional optical image detection now readily allow studying genome variation on a population-wide scale[11] as well as in single cells[12]. At the same time, the cost per analysis has decreased to a level that fosters applications in clinical practice[13] and makes it affordable even for the general public to identify personal health risk factors in their genomes[14]. Yet, knowing the sequence of base pairs alone does not reveal how our genetic 'blueprint' is translated into a biological phenotype as long as the functional assignment to the encoded elements remains elusive[15]. Studies of the 'transcriptome'[16] therefore aim to elucidate the complex relationship between genes and the first step of gene expression by mapping the variety of transcribed molecules, including protein-coding messenger RNA (mRNA), and identifying splicing variants as well as post-transcriptional modifications[17–19]. Disentangling factors that drive the expression of specific genes and thus the fate of a cell, and how gene expression adapts to internal or external factors has become a central focus of biological research[4]. However, transcripts are not the final products of gene expression, but rather represent intermediate steps to proteins, which - from a functional perspective - are the key actors in any living cell **(Fig. 1)**. Collectively, proteins build a fascinating and highly complex machinery - the proteome[20] - that orchestrates essentially all of cell and organismal biology. Proteins precisely regulate such disparate processes as gene expression and energy conversion, act jointly in multifaceted signaling cascades and steer inter-cellular trafficking. However, the relationship between the abundance of transcripts and proteins remains incompletely
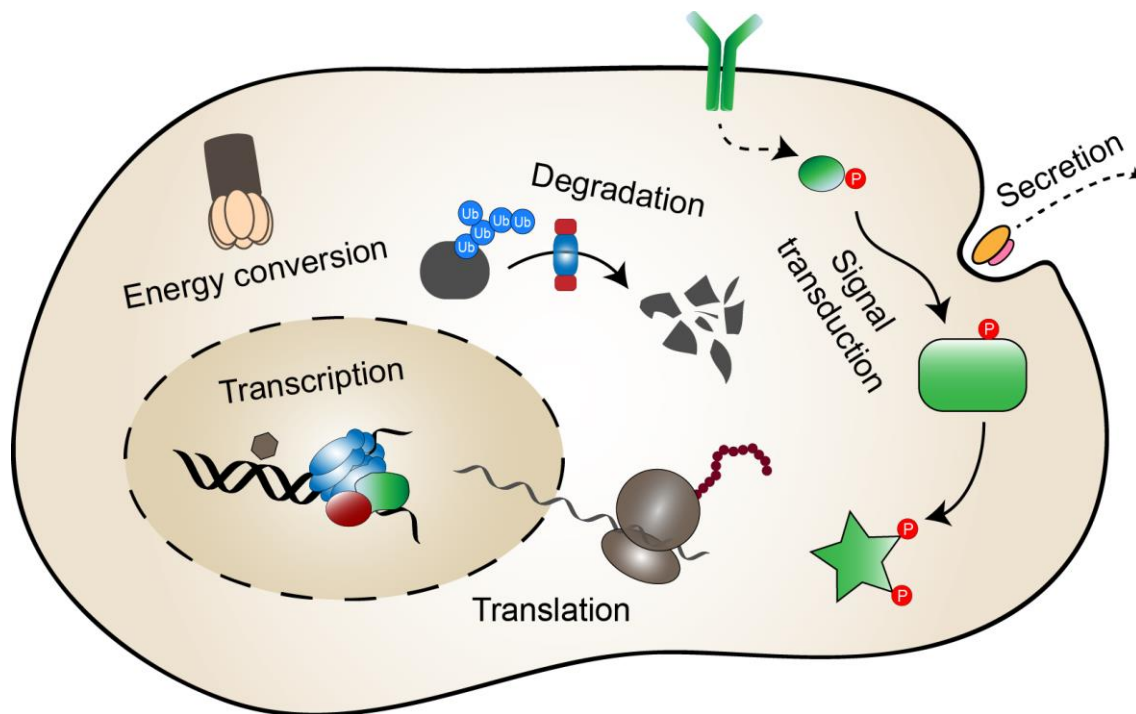
Figure 1 | Proteomics in systems biology. The operation of a mammalian cell is determined by the joint action of the hundreds of billion protein constituents and their interactions with other biomolecules. The ultimate goal of proteomics is to monitor and understand the dynamics of the proteome system, including protein synthesis, degradation, post-translational modification and localization.

understood, and the transcript levels alone are not sufficient to predict the phenotype[21–24]. Therefore, to gain a detailed understanding of health and disease at the molecular level, a direct analysis of the proteome is highly desirable.

At any given time, a typical mammalian cell expresses over 10,000 different genes and, across various tissue and cell types, transcriptional evidence for more than 85% of the roughly 20,000 predicted human protein-coding genes has been found[15,25]. A single mammalian cell comprises an estimated 10^9 to 10^{11} protein molecules (about 150 pg, ref [26]), while the abundance of individual proteins ranges from tens to over ten million copies per gene[27]. The different abundance levels imply major analytical challenges for the detection of low-abundance proteins as over 90% of the cellular protein mass is made up of just a few hundred proteins. This proportion can be even worse in tissue samples with heterogeneous cell types[28], or body fluids[29,30] with highly abundant transporter proteins. However, given the many different and highly specialized cells in our body, the overall number of different proteins is surprisingly low. In fact, an unexpected finding of the 'human protein atlas' project[31,32], with the ultimate goal to image the distribution of

all proteins in all human tissues, was that the majority of proteins are not specific to a certain tissue, but rather expressed across many tissues at different levels of abundance. This raises the question what else determines the identity and function of a cell and highlights the importance of protein abundances and the need to not only catalogue the protein content of a cell, but also quantify proteins precisely.

Importantly, the number of expressed proteins alone provides only a glimpse of the complexity of the (human) proteome. Genomic variations, alternative RNA processing and editing, and post-translational modifications (PTMs) vastly expand the proteomic landscape by giving rise to multiple 'proteoforms'[33] of each expressed gene[34]. To illustrate, already a single protein with five acetylation, six phosphorylation sites, a potential N-terminal acetylation and one single nucleotide polymorphism, could theoretically be present in over 8,000 different proteoforms. From an evolutionary perspective, this variety of gene expression products opens up the possibility to explore a much wider space of protein structures and functions than would have been possible by genetic variations alone[35]. Reversible and fast modifications of molecular switches such as phosphorylation sites are key to exchange information between or within cells[36,37], and the fact that a large class of drugs target kinases[38] underlines the importance of PTMs. While the aforementioned examples emphasize the importance of single proteins and modification states, most proteins do not act in an isolated manner. They are part of highly dynamic protein complexes or macromolecular assemblies, in which subtle changes in one protein may influence others directly or indirectly[39].

Capturing the proteome system in its entire complexity, including protein expression, abundance, localization, modification state and interaction with other biomolecules, is the central goal of proteomics – and an enormous technological challenge.

Mass Spectrometry-Based Proteomics

Historically, biochemical methods have focused on the isolation and reconstruction of a biological entity, typically a protein, to study its function, one at a time. Today, the detection and quantification of proteins is still routinely based on the recognition of target proteins by antibodies in conjunction with light-emitting reporters[40]. Due to the specificity of many antibody-antigen interactions[41], the established nature of the assays and the outstanding sensitivity of light detection, this technology is still the gold standard for known target proteins. Examples include the visualization of protein localization in pathology[42,43] or high-throughput clinical diagnostics in the form of enzyme-linked immunoassays (ELISAs)[44]. However, antibody-based approaches can be compromised by cross-reactions with proteins other than the one of interest, and monoclonal antibodies typically target only a single epitope, making them sensitive to structural changes caused by post-translational modifications or changing experimental conditions[45,46]. Moreover, the development of a specific antibody takes a lot of resources, while reproducible production is only possible for monoclonal antibodies and remains an expensive and laborious task.

The shortcomings of antibody-based methods and the need for system-wide measurements of ideally all expressed proteins in a single experiment akin to genomics and transcriptomics, has propelled the development of an alternative technology – mass spectrometry (MS)-based proteomics[47,48]. As opposed to antibodies, it is largely unbiased and offers the possibility to study proteomes in a comprehensive and hypothesis-free manner.

Bottom-up Proteomics

The most common MS-based proteomics workflow is the so-called 'bottom-up' approach (**Fig. 2**), which has evolved over the years through breakthroughs in all its constituent parts, beginning with sample preparation and ending in computational analysis. It has now become universally applicable to essentially all kinds of proteomics samples and research questions[47,48].
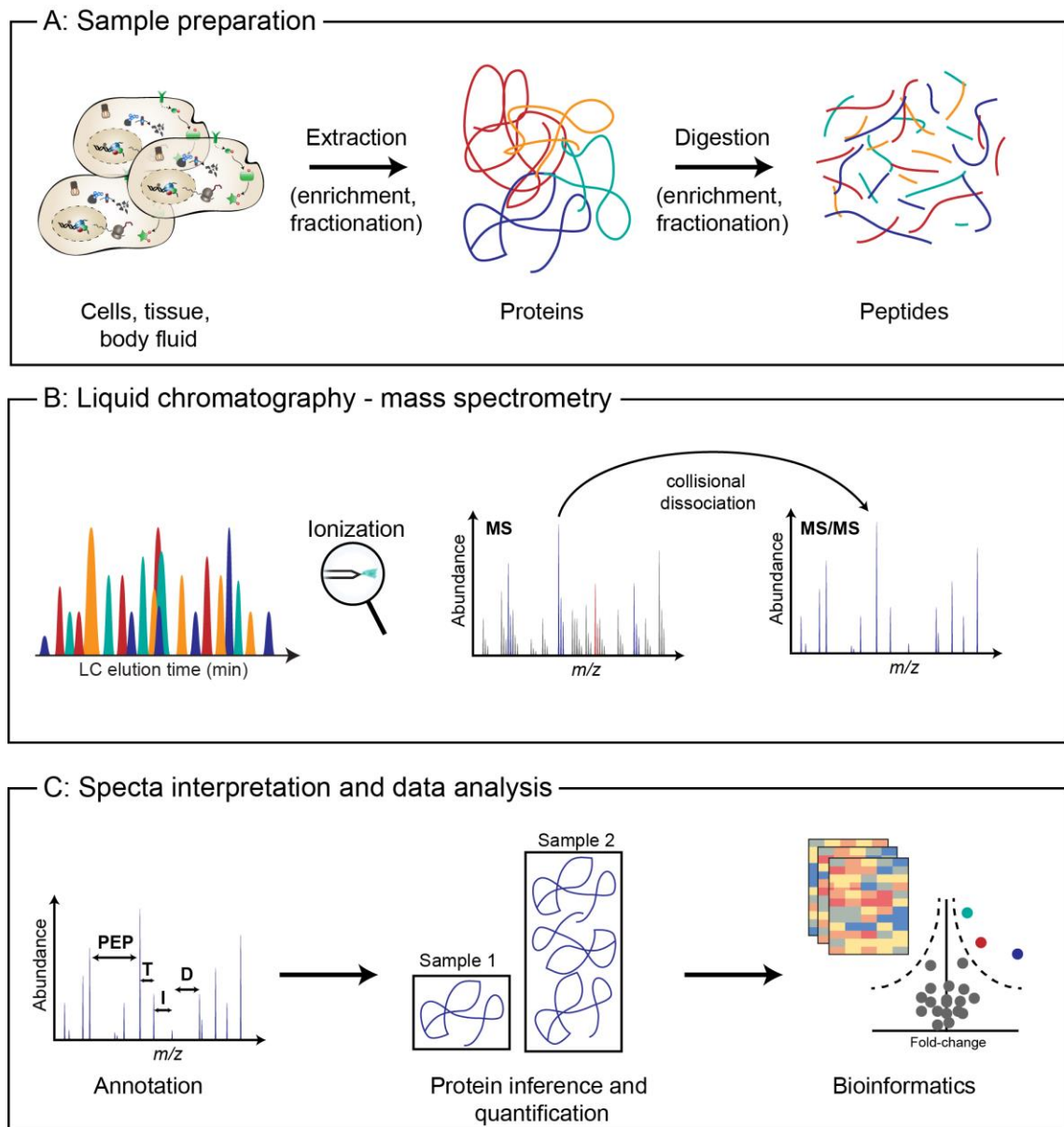
Figure 2 | The bottom-up proteomics workflow. A: Proteins are extracted from a biological sample of interest and solubilized prior to enzymatic digestion with trypsin or other proteases. Fractionation or enrichment of target proteins or PTMs can be performed either at the protein or peptide level. **B:** Purified peptides are then separated via liquid chromatography at nanoliter per minute flow rates and electrosprayed into the mass spectrometer. The MS acquires full spectra of all eluting peptides (MS spectra) or fragment mass spectra (MS/MS spectra) of selected precursor ions. **C:** The resulting spectra are interpreted by comparison of the observed precursor and fragment ion masses with *in silico* databases. Identified peptides are mapped onto a protein sequence and ultimately, bioinformatics tools are employed to analyze quantitative differences between sample cohorts. Adapted from ref. 58.

The term 'bottom-up' refers to the underlying working principle: proteins are extracted and enzymatically cleaved by sequence-specific proteases[49]. The cleavage products, peptides, are separated via liquid chromatography and analyzed in a mass spectrometer[50,51].

Protein identity is inferred *in silico* based on the specificity of the cleavage site and by comparing the observed mass spectral signature of a peptide to all possible peptides from a reference genome[52]. Simplified sample handling, reduction of complexity, and the much superior amenability of peptides to MS analysis as compared to entire proteins, generally more than compensate for the loss of information about protein structures and co-occurring modifications. The alternative 'top-down' approach omits the proteolytic cleavage and analyzes whole proteins, in some cases without disrupting their native tertiary and quaternary structures[53], however, at the expense of throughput, sensitivity and proteome coverage[54,55]. An in-between alternative, the so-called 'middle-down' approach, overcomes some of these downsides by employing proteases that produce longer peptides and therefore retain some co-occurring modifications[56,57]. While in principle very attractive, this approach introduces new limitations and, to date, it also cannot compete with bottom-up in terms of sensitivity and proteome coverage.

Any proteomic workflow starts with the extraction and solubilization of proteins from the biological sample of interest, for example cell lysates, tissue homogenates or body fluids[58]. Modern methods readily overcome the challenges introduced by the poor solubility of membrane proteins via soluble peptides that they give rise to after digestion, and employ buffers that are readily compatible with mass spectrometry[59]. The typical bottom-up workflow involves reduction and alkylation of disulfide bonds to prevent dimerization and thus to facilitate the proteolytic cleavage. Trypsin, the most commonly employed protease, cleaves at the C-terminal sides of lysine and arginine[60]. As we will see later, this yields peptides with favorable properties as they are relatively short yet specific enough to be mapped to a reference genome[52], and as they can bear a positive charge at both the N-terminus and the C-terminal side chain. Alternative proteases such as chymotrypsin, GluC, and AspN or chemical proteolysis, are less popular, but can provide complementary information and increase overall protein sequence coverage[61,62]. Prior to MS analysis of the resulting peptide mixture, salts and remaining detergents have to removed, which is typically performed via solid-phase extraction on C_{18} coated particles or similar materials[63–65]. Miniaturization and careful selection of cross-

compatible reagents and conditions now allow performing all of the aforementioned steps in a single vial reactor[66]. These recent developments avoid loss-prone transfer and buffer exchange steps, and have tremendously simplified sample preparation. Notably, as only liquid-handling steps are involved, the protocols can readily be automated with robotic systems, which enables high-throughput processing of clinical samples in just a few hours from receiving the sample to MS analysis[67].

These principles also extend to the analysis of PTMs, which typically involves additional enrichment steps to overcome sensitivity challenges in the detection of low abundant and sub-stoichiometrically modified peptides[68]. Numerous protocols exist to enrich peptides with a specific modification over their non-modified counterparts. For example, analysis of protein phosphorylation exploits the affinity of the bivalent phosphate group to titanium dioxide (TiO_2) beads or immobilized metal cations (IMAC) or diverse phospho-specific antibodies[69,70]. An interesting strategy is used to study ubiquitination sites on a proteome-wide scale: an antibody targets the di-glycine remnant of the covalent ubiquitin modification after tryptic digestion[71,72]. Enrichment is also possible at the protein level if a suitable antibody is available. This strategy has proved very powerful to explore tyrosine phosphorylation, which is otherwise hard to detect, although pY antibodies are usually used at the peptide level as well[36,73]. Immunopurification or –enrichment in combination with bottom-up proteomics is also commonly applied to reveal interaction partners of specific target proteins, an approach that can in principle define entire 'interactomes'[74]. The latter often employs GFP-tagging of gene products and enrichment via anti-GFP antibodies, which makes this technology applicable to entire proteomes[75,76], for example in combination with the CRISPR-Cas9 technology[77].

After enzymatic digestion, the resulting peptide mixture is separated via reversed-phase liquid chromatography at low pH. Here, peptides interact with a non-polar stationary C_{18} phase that is covalently bound to porous silica beads and a binary aqueous/organic mobile phase. During the analysis, the proportion of the organic component, typically acetonitrile, increases gradually so that peptides elute from the column in order of increasing hydrophobicity. The development of nano-flow chromatography, with flow
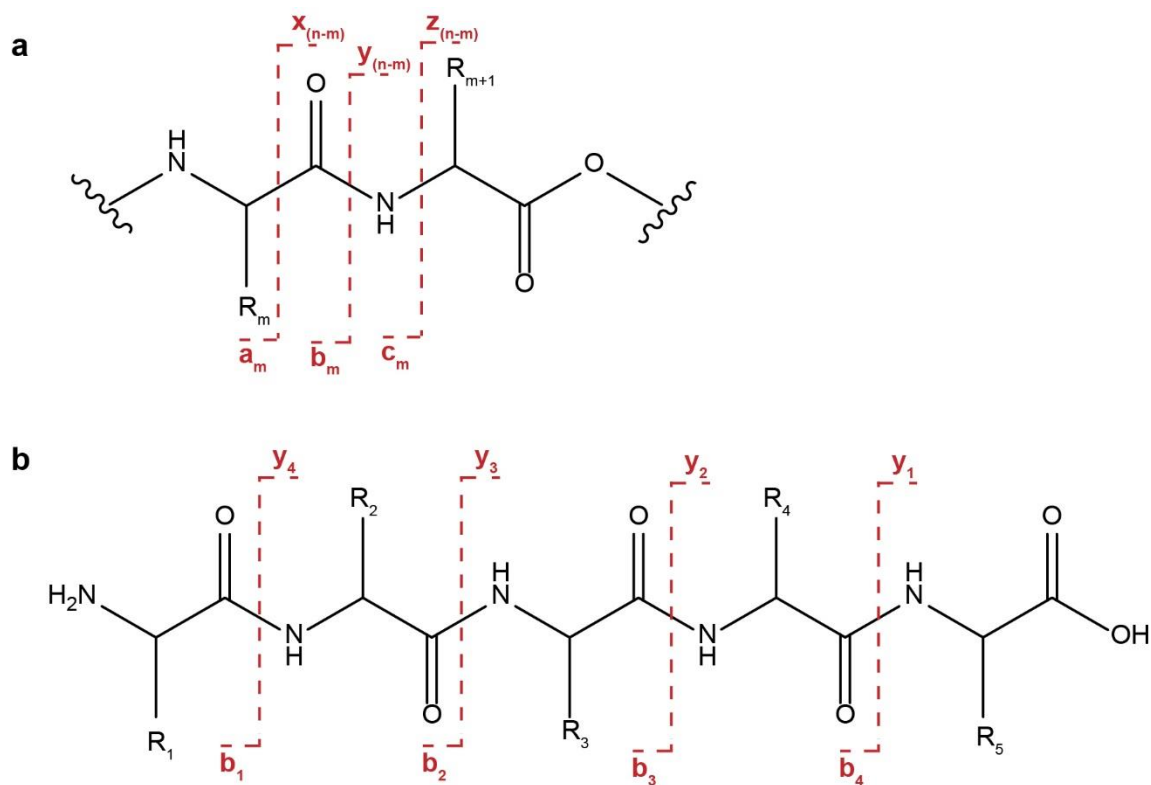
Figure 3 | Fragmentation of peptide ions in the mass spectrometer. a, Roepstorff-Fohlmann-Biemann nomenclature of peptide backbone fragmentation[78]. When the charge is retained on the C-terminal side of the peptide, x, y, and z fragments are obtained, and conversely, retaining the charge on the N-terminal side yields a, b, and c ions. n is the length of the peptide sequence in amino acids and m the position of the amino acid relative to the N-terminus. **b**, In proteomics practice, collision induced dissociation is widely used, which yields primarily b- and y-type ions by breakage of the lower-energy peptide bond. The fragment ion nomenclature is exemplified for a peptide with 5 amino acids. Adapted from ref. 79.

rates in the range of hundreds of nanoliters per minute and very narrow columns (typically 75 um inner diameter), provided a boost in sensitivity and selectivity[47]. The column is interfaced 'online' with the mass spectrometer to allow direct mass analysis of the eluting peptides (LC-MS). In proteomics, as further detailed below, the mass spectrometer performs two fundamental experiment types: the first one yields a survey spectrum of all detectable peptides (MS scan), and the second yields fragmentation spectra of selected peptide precursor ions (MS/MS scan). It turns out that positive peptide ions follow a very specific fragmentation scheme that allows straightforward annotation of spectra and prediction of the fragment ion masses from a given amino acid sequence (**Fig. 3**)[78,79]. The exact type of ions generated depends on the used fragmentation method. Collisional

induced dissociation (CID) in an ion trap and higher energy collisional dissociation (HCD)[80] are widely used and yield primarily *b* and *y* ion series. Alternative fragmentation methods produce different ion types, which can provide additional information under certain circumstances. Electron transfer dissociation (ETD), for example, yields *c* and *z* fragments[81], and ultra-violet photodissociation (UVPD) can also yield higher-energy *a* and *x* ions[82,83]. Multiple MS acquisition methods for proteomics have evolved (see also below); however, these two experiments or scan modes (MS and MS/MS) are a shared feature of all MS-based proteomics workflows.

In light of the inherent complexity of proteomics samples, pre-fractionation steps are often employed to decrease sample complexity prior to MS analysis[84–86]. In principle, fractionation is possible at different points in the workflow: at the level of cellular compartments, intact proteins or peptides. As only a few proteins typically constitute a large proportion of the total protein mass and tend to mask low-abundance proteins in the analysis, fractionation techniques are most efficient if they in effect reduce the dynamic range (which is the ratio from the highest to the least abundant protein) in a fraction prior to MS analysis. Fractionating a sample by cellular compartments or intact proteins usually does not separate high and low abundant proteins due to the relatively low resolution and is in any case limited to soluble proteins. This explains the popularity of peptide fractionation, which is readily applicable to hydrophobic proteins and is conveniently combined with manual peptide purification[65] or performed with standard chromatography systems[87], either offline or directly interfaced with LC-MS analysis (as in the 'MudPIT' technology[88]). A current trend is to use reversed-phase peptide fractionation at basic pH[89,90], which shows good orthogonality with low pH LC-MS[91] and can be combined with sophisticated concatenation schemes to maximize LC-MS utilization[92–94]. Recent advances in our laboratory have increased sensitivity by minimizing sample losses and reducing manual pipetting steps[95]. However, while fractionation increases the depth of proteome analysis, it also increases the number of sample preparation steps, requires higher starting amounts and multiplies the measurement time and therefore the cost.

The publication of the first complete MS-based proteomic map of a complex organism, yeast, in 2008 marked a milestone in proteome research[96]. This pioneering effort was a technological *tour de force* building on extensive fractionation steps prior to LC-MS analysis. Ten years later, the multi-faceted technological developments in the field now make the analysis of yeast proteomes at comparable depth a matter of minutes or hours

instead of weeks[97,98]. As the technology improved, researchers tackled increasingly complex proteomes[99], including nematodes[100] and diverse mouse or human cell lines[101]. Increased throughput fostered comparative studies of tissue proteomes across entire animals[102], and provided novel insight into cell-type and anatomy-resolved proteomes of murine liver[103] and brain tissue[28]. In 2014, two research groups independently presented first draft maps of a human proteome, indicating that the goal of a complete human proteome is within reach[104,105]. Along the lines of previous in-depth proteome studies and relying on multi-dimensional protein and peptide fractionation, the Pandey group profiled 30 human cell and tissue types in over 2,000 LC-MS experiments. In addition to data generated in the author's laboratory, Küster and co-workers employed the wealth of freely accessible community data, and assembled their initial draft proteome from 16,857 LC-MS experiments in total. This project is constantly updated and curated in ProteomicsDB[106], which counts 15,721 human proteins identified from over 43 million MS/MS spectra at the time of writing. This data suggests a 'core proteome'[101,107] of about 10,000 to 12,000 that is ubiquitously expressed with varying abundance across many tissue and cell types and readily accessible by standard proteomics workflows. It follows that many proteins are confined to specific cell types and organs, for example proteins of the immune system. Such insight into the human proteome is a first step toward answering the question, which of the about 20,000 human protein-coding genes are actually translated into proteins, and can further provide valuable information for the emerging field of proteogenomics[108]. Remarkably, Olsen and co-workers recently reported a comparable proteomic depth of about 14,000 protein isoforms and 12,200 protein-coding genes in HeLa cells alone, employing multiple proteases and rapid measurements of a large number of peptide fractions to increase the effective sample load for mass analysis[109].

It appears that we are now entering an era of proteomics covering the protein-coding genome to near-completeness in various biological samples, with constantly growing study sizes[110]. Numerous large-scale investigations have been initiated to link genomic variations with proteomics phenotypes, for example in cancer biopsies[22,23,111], or to profile body fluid proteomes of hundreds or thousands of individuals[67,112]. However, so far such studies are only being planned or are limited to highly specialized laboratories because very deep proteome coverage still requires sample fractionation, which entails lower throughput, higher starting amounts and can complicate quantification. A grand challenge

in MS-based proteomics is thus to provide the maximum proteome coverage, with highly accurate quantification in thousands of samples, and ideally in less than one to two hours per sample and instrument.

Computational Proteomics

The success of bottom-up proteomics relies on the ability to match the observed fragmentation pattern of a peptide to its amino acid sequence, and to map the peptide sequence on a protein sequence with high confidence. A typical proteomics experiment on state of the art instrumentation now yields over 100,000 MS/MS spectra from a single sample in less than two hours and on a single instrument. Analyzing such an amount of data would clearly not be possible without the continued development of bioinformatics software[113].

A multitude of software tools have emerged over time, however, the aim of this section is to briefly introduce the underlying principles focusing on those relevant to the
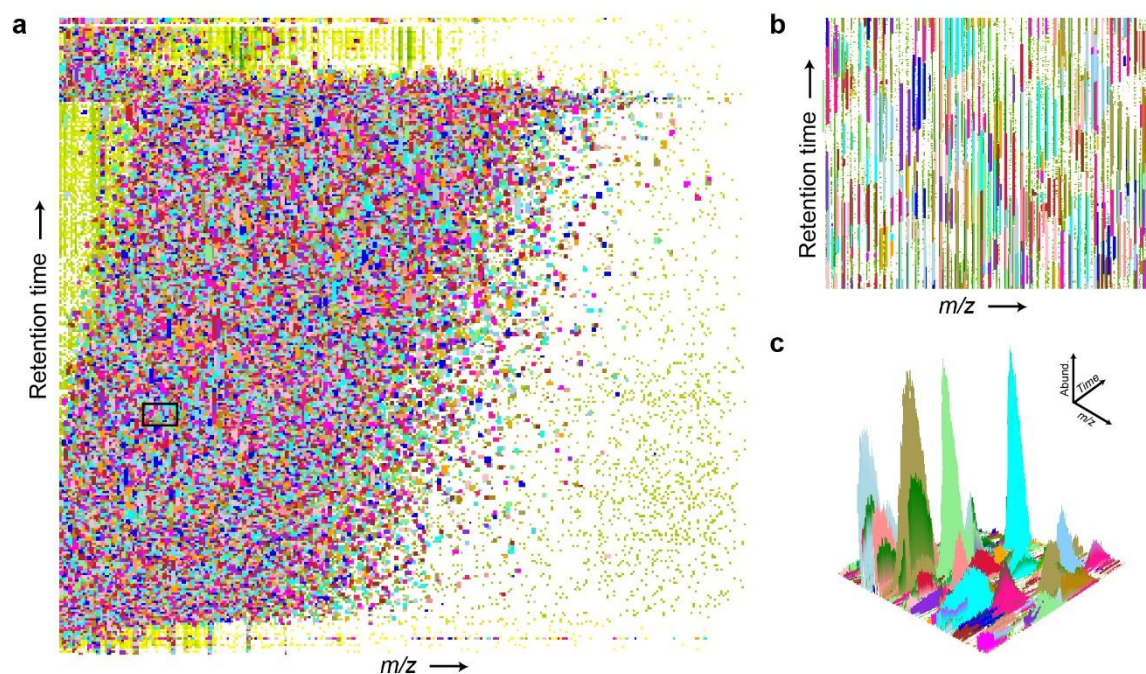
Figure 4 | MaxQuant feature detection in LC-MS runs of proteomics samples. a, In a single experiment hundreds of thousands of isotope patterns are detected and assembled into three-dimensional features (here indicated by color). **b**, A zoomed view into a small window of the feature map shown in a. **c**, Three-dimensional representation of b, showing the chromatographic elution profiles of selected peptide ions detected by high-resolution mass spectrometry.

MaxQuant software package[114], which has been primarily used throughout this thesis. For a more in-depth review, readers are referred to a recent publication from Cox and co-workers[115].

In the first analysis step, the detected MS signals are assembled into three-dimensional peaks or 'features' in mass, intensity and retention time dimensions (**Fig. 4**). As further described below, very fast time-of-flight mass spectrometers allow introducing ion mobility spectrometry as an additional dimension of separation, in which case feature detection is extended to four dimensions. Isotope patterns are typically fitted to an 'averagine' model[116], which represents the average peptide isotopic composition for a given mass. To simplify downstream analysis, the software collapses the detected isotope peaks ('de-isotoping'). Key to the success of MaxQuant is the highly accurate determination of masses from high-resolution mass spectra[117]. In the absence of un-resolved interferences, high-end instrumentation achieves mass accuracies in the lower parts-per-million ($\Delta m/m$) range, which is however, susceptible to systematic drifts over time. Lock-masses, which are ubiquitously present ions with known mass, can be used to accurately re-calibrate the mass spectra either in real-time or in post-processing and thus largely eliminate systematic mass errors[118,119]. Even more accurate masses can be achieved with 'software lock-masses' as implemented in MaxQuant[120]. Instead of one or two lock masses, the algorithm uses thousands of peptides confidently identified in a first pass search to correct time- and mass-dependent mass errors and thereby routinely achieves sub-ppm accuracies. Similarly, these identifications can also be used to align retention times of different runs, which is, as explained further below, crucial for reproducible (label-free) protein quantification across many samples.

In standard data-dependent acquisition experiments, the acquired MS/MS spectra are next assigned to the detected peptide features, the 'precursors', based on the raw data information from the MS scans. Because of the predictable fragmentation patterns of peptides, observed fragment ions can be compared to *in silico* derived fragment masses of tryptic peptides obtained by 'digesting' a reference database containing all protein-coding regions of a genome[52]. This is the database search strategy most commonly used in proteomics and restricting the precursor mass estimate by sub-ppm mass accuracies greatly improves search sensitivity[114]. Peptides spectrum matches (PSMs) are scored by the number of matching fragment ion masses and, if applicable, additional peptide properties (see for example ref.[121]). To control for false discoveries, the probability

12

scoring is also performed on a decoy database[122]. This has the advantage that the combined search result can be cut off when 'non-sense' hits account for a desired proportion of all hits, typically at a threshold corresponding to an estimated maximum false discovery rate of 1%. Alternatively, machine learning algorithms can stratify true and false positive identifications[123]. Either way, database strategies imply that only peptides present in the search database (including modifications) can be identified. Adding too many potential modifications leads a 'combinatorial explosion' of the search space – increasing the computation time and decreasing the statistical sensitivity. The search space explosion does not apply to *de novo* search algorithms, which identify peptide sequences starting from the measured amino acid mass differences[124,125], however, they generally have lower sensitivity. In-between these extremes, there are several hybrid strategies that can favorably combine the strengths of both approaches[126–128]. An interesting recent development is the application of deep learning tools to predict fragment ion intensities, which provides an additional layer of information that can be used for scoring and improve the identification success[129]. This technology is still under development, and currently limited to non-modified peptides, but it has great potential and will excel with higher data quality and computational power. Further improvements in this area are highly needed, given that typically only 25 to 60% of the acquired MS/MS lead to peptide identifications, even though many more spectra appear to be of high quality upon visual inspection[130,131].

As the biological goal of proteomics is primarily to identify and quantify proteins rather than peptides, we have to assemble peptide PSMs into proteins in the next step[52]. The commonly used human reference proteome (UniProt)[132] counts over 90,000 protein entries including isoforms, and its proteins have an average length of 480 amino acids. This opens up many combinatorial possibilities to match a given peptide sequence to proteins, and many (in particular short) peptides are non-unique. To avoid these non-unique peptides inflating the number of identified proteins, MaxQuant, along with other software solutions, applies Occam's razor principle to find the smallest subset of proteins that explains the entirety of observed peptides in an experiment. Similar to controlling the false discovery rate at the peptide level, it is important to avoid accumulation of false protein identifications, which could otherwise skew the analysis and ultimately lead to incorrect biological conclusions[133,134]. MaxQuant does this by calculating a combined score of the best PSMs for each protein, which can then be sorted in descending order and

truncated at a user-defined false discovery rate (typically 1%), similar to the cut-off for peptide identifications[135].

Having curated a high-quality list of protein identifications (and quantifications), the result files can be analyzed with various bioinformatics methods and software packages. One example is Perseus[136] that is closely tied to the MaxQuant software, or the R computational environment or the multitude of available Python packages. These software packages also connect the results to meta-information, such as gene ontology terms[137], known complex members or biological function.

Quantification Strategies

Identifying gene expression products from various biological specimens in specific conditions has now become more feasible than ever and recently culminated in the identification of more than 14,000 proteins from a single human cancer cell line in a single experiment[109]. In addition to cataloguing all expressed proteins, a central goal of proteomics is the accurate quantification of protein abundances across multiple biological samples[138]. In fact, biomarkers are likely present in both, health and disease states, however, they will be elevated in either one of them[30]. Reliable quantification of subtle changes is a tremendous challenge and requires accurate and highly precise methods. Protein quantities can be expressed in absolute levels, e.g. "*ng/mL in blood plasma*", or in relative levels between two samples, e.g. "*two-fold more abundant in condition x than in y*". In proteomic practice, relative quantification is more common, mainly for reasons of simplicity and straightforward application to any situation[48,139,140].

MS-based proteomics employs multi-step workflows and delicate high-end instrumentation, and each of these steps can introduce systematic or stochastic biases into the analysis. Alterations in the protein signal can arise, amongst other sources, from different starting amounts, different efficiencies in the protein extraction or the proteolytic digestion, inaccuracies of the LC autosampler, fluctuations in electrospray ionization and time-dependent drifts in instrument performance. Most of these errors can be corrected by the addition of stable isotope-labeled references, assuming that they behave just like their non-labeled endogenous counterparts throughout sample preparation and analysis steps. This is usually a valid assumption because labeling makes peptides distinct in mass (typically through introduction of ^{13}C and ^{15}N isotopes) and thus distinguishable in the

14
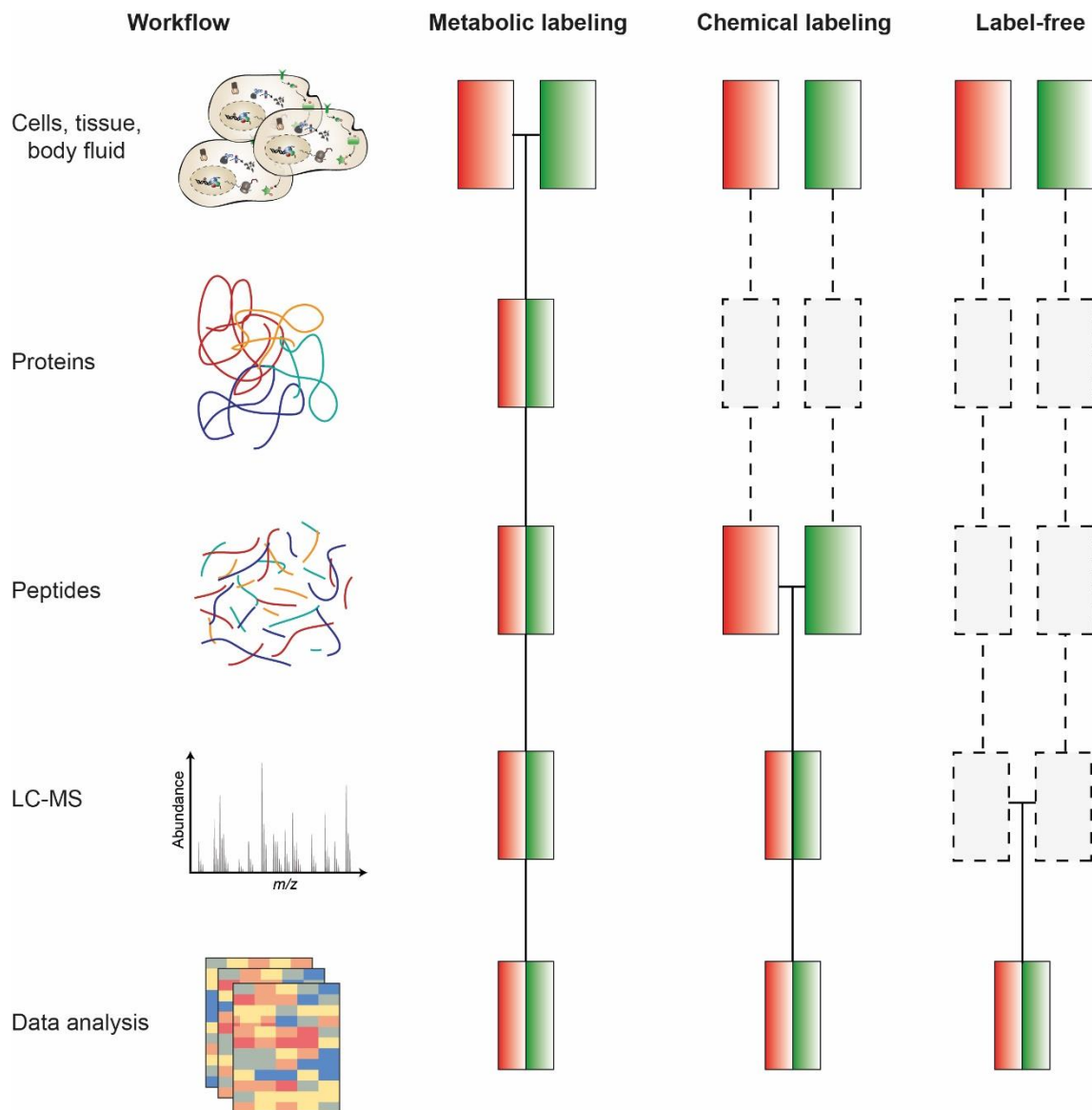
Figure 5 | Common workflows for quantitative bottom-up proteomics. Red and green boxes represent two samples of interest, which are processed from top to bottom. Vertical lines indicate the step at which both samples are combined with metabolic labeling, chemical labeling and the label-free approaches, respectively. Steps in which quantitative biases due to sample handling can be introduced are indicated by dashed lines. Adapted from ref. 140.

mass spectrometer, while they remain physio-chemically nearly identical. Protein quantification strategies can therefore be classified in two main groups: stable isotope label-based quantification (either MS or MS/MS-based) and label-free quantification[138–140]. **Figure 5** provides an overview of the quantitative workflows described in the following and **Figure 6** illustrates the quantitative read-out from the MS signal in each case.

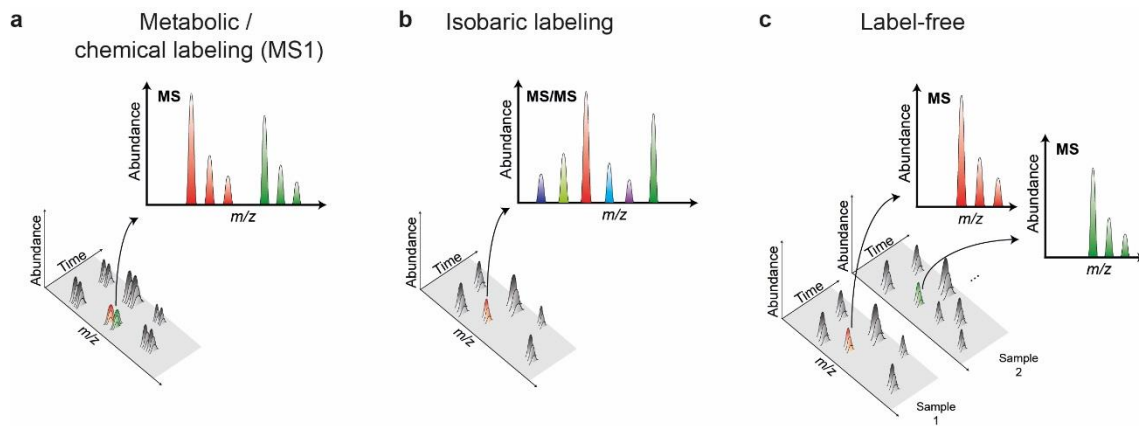**Figure 6 | Quantification in bottom-up proteomics. a**, In stable-isotope labeling experiments with for example SILAC or dimethyl-labeling, samples are mixed and the relative quantification is read-out on the MS-level. **b,** In TMT and iTRAQ experiments, the peptide ions from different samples are indistinguishable on the MS-level and the relative quantification is performed on the MS/MS-level. **c,** In label-free experiments, samples are measured separately and quantified by comparing the relative intensities between runs.

Stable isotope label-based strategies can further be classified by the sample preparation step at which the labeled reference channel is admixed (**Fig. 5**). At the earliest, stable isotopes can already be incorporated metabolically into the sample of interest. A historically important method, 'stable isotope labeling by amino acids in cell culture' (SILAC) for *in vivo* labeling of cell lines[141] was introduced in 2002. SILAC largely eliminates biases introduced during sample processing and analysis, and therefore offers outstanding accuracy[142]. After mixing the samples and tryptic digestion, the relative MS signals of stable isotope labeled peptides and their light counterparts represent the relative protein abundances in the samples. Common workflows add labeled arginine and lysine to the culture medium, and thereby ensure that every peptide is present in isotope-labeled and non-labeled states after tryptic cleavage. SILAC is still a popular method for pulse-chase experiment to study protein synthesis and degradation[143–145] and, remarkably, researchers have extended SILAC even to label entire living animals[102,146]. The invention of 'super-SILAC' broadened the application of SILAC to samples such as human tissues that are not directly accessible to metabolic labeling, by admixing cell lysates from cultured cells as a labeled reference channel[147]. However, the incorporation of N stable isotope channels multiplies the complexity of the MS signal by a factor of N. This complicates the quantitative analysis and generally lowers the proteome coverage as

16

compared to single-channel analysis. For this reason, SILAC is in practice limited to multiplexing only two or three samples (light, medium, heavy)[148,149]. More recent developments, such as neutron encoded amino acid labeling (NeuCode) avoid added complexity at the MS level by labeling with the delta masses of ^{13}C and ^{15}N isotopes (6 mDa), which are only resolved at very high mass resolution)[150,151]. However, to date, these methods have not overcome their limitations in terms of proteomic depth and number of quantitative channels.

Even though SILAC and similar approaches provide superior accuracy, the need for metabolic labeling is limiting. In contrast, chemical labeling at the peptide level is applicable to digested proteins from any source[138]. In principle, various functional groups of peptides are modifiable using established peptide chemistry, although most protocols target the primary amine (–NH2) groups at peptide N-termini and lysine side chains. Similar to SILAC, dimethylation with stable isotope labeled formaldehyde (^{13}C and D) introduces a specific peptide mass shift in each channel and therefore allows mixing the labeled digests and analyzing them in one LC-MS run[152]. The relative ratios in each channel (typically two or three) directly reflect the relative protein abundances. In practice, the toxic chemicals and the slightly different chromatographic properties of deuterium as compared with hydrogen[153] have hampered wide adoption of dimethyl labeling, whose proteome coverage suffers from the multiplication of the MS signal in a similar way to SILAC. An increasingly popular alternative are so-called 'isobaric labeling' methods with iTRAQ[154] and especially TMT[155]. They use established N-hydroxysuccinimide (NHS) chemistry to specifically label reactive amine groups. In contrast to the methods above, isobaric labels are not distinguishable by mass at the MS level and therefore do not convolute the MS signal. This is very attractive as it does not affect the complexity and therefore the coverage of the proteome. Isobaric labels consist of a 'reporter group' and a 'balancer group'[154]. Distributing stable isotopes between both groups in a way that the overall mass does not change allows encoding of different quantitative channels (**Fig. 7a,b**). Mixing labeled peptides from different samples and fragmenting the label in an MS/MS experiment yields distinguishable mass signals for each quantitative channel and retrieves the quantitative information (**Fig. 7c**). Similar to NeuCode, ^{13}C to ^{15}N differences can be used to encode additional channels[156], currently
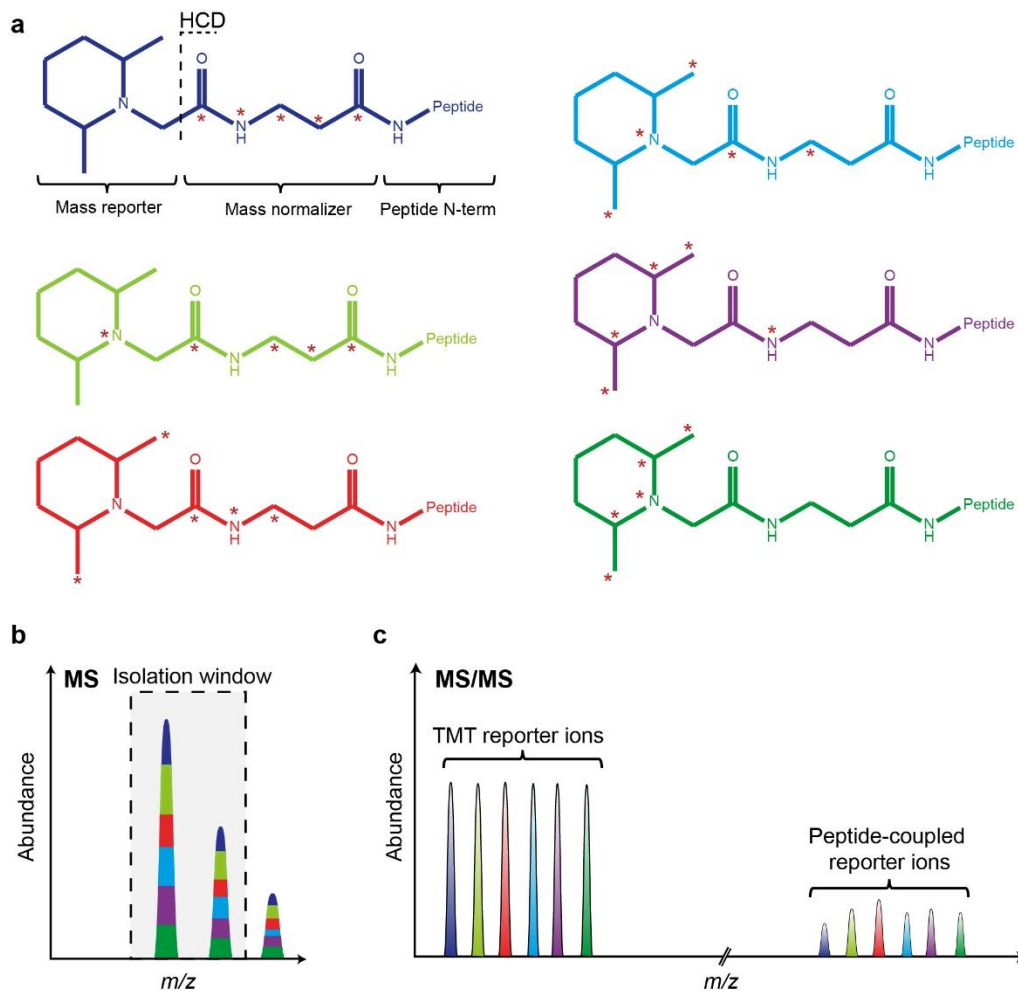
Figure 7 | Isobaric labeling with TMT. a, Chemical structure of the six-plex TMT molecule. The dashed line indicates the fragmentation site with higher energy collisional dissociation (HCD). Red asterisks indicate the distribution of stable isotope labels (^{13}C and ^{15}N) in the mass reporter and normalizer groups, respectively. The molecule is covalently bound to the peptide N-terminus or a side-chain amine group. **b,** The same peptides from different peptides (indicated by color) are isobaric and co-isolated for fragmentation. **c,** The relative abundances of the low mass reporter ions indicates directly the protein abundance in the multiplexed samples. Alternatively, the high *m/z* 'complementary' reporter ions can be quantified.

allowing to multiplex up to 11- with commercially available kits. Within a set of isobaric labels, these methods provide very accurate and complete measurements. However, in cohorts with more than 11 samples, entire sets of quantitative values can be missing and optimal strategies for inter-batch normalization are still a subject of ongoing research. In almost all isobaric labeling strategies, the low-mass reporter ions are quantified, however, the design of the labels imply that these reporter ions are the same for all peptides. As a result of imperfect isolation of precursor ions, co-isolated peptides distort the signal of

the target peptide – an effect often referred to as 'ratio compression'[157]. Attempts to address this issue range from minimizing the width of the isolation window[158,159], computational signal deconvolution[160], gas phase separation[161] and gas phase purification[162], to repeated fragmentation of peptide fragment ions (MS^3)[163]. MS^3 achieves the most accurate ratios, even though none of these methods resolves the problem to satisfaction. The major drawback of the MS^3 method is that the additional fragmentation step inevitably lowers sensitivity, and that it requires more expensive MS instrumentation with MS^3 capabilities. To increase the sensitivity, multiple fragment ions can be selected simultaneously for further fragmentation ('SPS-MS^3' method)[164]. However, simply by chance, this can again result in co-fragmentation of fragment ions originating from different peptides and thus distorted ratios. This effect is particularly prominent in case of low-abundance peptides. A method termed TOMAHAQ uses prior knowledge of theoretical fragment masses to select exclusively fragment ions of the target peptide[165]. Although interesting, it increases the complexity of the experiment and is limited to pre-defined targeted peptides. A more generally applicable strategy aims to quantify the peptide-coupled remnant rather than the low-mass reporter ions ('complementary reporter ions')[166,167]. This approach is very attractive as it does not require MS^3 and retains the precursor peptide information, allowing accurate quantification even of co-isolated peptides. First iterations employed commercially available TMT reagents, which were initially designed to generate the low-mass reporter ions with high yield. It turns out that these labels generate the peptide-coupled reporters only inefficiently and the quantitative signal is furthermore convoluted by the abundance of natural isotopes[166]. Both of these limitations have been addressed in the course of this thesis by a re-designed label and a tailored MS-acquisition strategy.

In label-free quantification strategies, each sample is analyzed separately and the relative protein abundances are inferred from the signal intensities in each individual run[140]. As it does not involve stable isotope labels, it is universally applicable to all sample types, including human body fluids and tissues. The absence of a labeling step simplifies the sample preparation workflow, lowers the costs and in principle enables straightforward scaling up to 100s or 1,000s of samples. For these reasons, label-free quantification (LFQ) has become one of the most popular methods for protein quantification[48]. However, processing and analyzing each sample separately may introduce biases due to variations in sample handling and instrument performance (**Fig. 5**)[168]. Correcting for workflow- and

instrument-related alterations in signal intensities is very challenging and has fostered the development of more reproducible sample preparation workflows and more sophisticated computational algorithms[115,140,169,170]. In data-dependent acquisition methods (see below) the mass spectrometer selects peptides for fragmentation based on abundance, which introduces a bias toward high-abundance peptides. Spectral counting approaches use this to infer protein quantities indirectly from the number of fragmentation events related to a protein of interest[171]. Clearly, this is only a very rough estimate of a protein's abundance, yet surprisingly reliable and very simple to analyze[172]. More sophisticated approaches infer protein abundances from the peptide signal integrated over their chromatographic elution peaks[173,174]. MS intensity-based method cover a much larger dynamic range and provide superior linearity as the signal is not quantized by the number of fragmentation events[175–177]. While the relative quantification of peptide intensities is rather straightforward, inference of protein abundances is not trivial[52]. Intuitively, one can use summed peptide intensities as a proxy for protein abundance[178]. However, quantification in particular of low-abundance proteins becomes more robust when the median fold-change of peptide pairs is calculated rather than their summed intensity. Cox implemented the latter strategy in the MaxQuant software (termed MaxLFQ')[179]. A common assumption used for normalization of inter-run variability is that the vast majority of protein abundances does not change in a biological experiment. This generally holds true and enables accurate quantification of small fold-changes in label-free experiments[140]. LFQ typically also achieves the most comprehensive proteome coverage of the methods described here and can be applied to very large data sets. However, in shotgun proteomics experiments of increasing size, quantitative accuracy can be limited if proteins are inconsistently quantified across samples, as 'non available (*N/A*)' values interfere with statistical testing[180]. This problem is often referred to as the 'missing value' problem and is addressed by novel data-acquisition methods as detailed below and diverse software solutions (see also below). In data-dependent acquisition schemes, missing values can occur due to semi-stochastic precursor selection, which results in not every peptide being identified in each run. However, identifications can be transferred between runs following rigorous retention time alignment and based on high-resolution and high-accuracy mass measurements[101,135,181]. This greatly diminishes the effect of missing values, as long as the peptide precursor signal is detected and mass resolved.

Mass Spectrometry Instrumentation for Proteomics

The core of proteomics technology is the mass spectrometer itself, and developments in the field have always been closely tied to technological advances in mass spectrometry[47,48]. Mass spectrometers measure the mass, or more precisely the mass-to-charge ratio (*m/z*), and the abundance of ions (**Box 1**). The MS measurement is performed in high vacuum (down to 10^{-10} mbar in the analyzer region of an Orbitrap instrument). The transfer of intact, non-volatile compounds, e.g. large biomolecules with masses exceeding 10 kDa, from the liquid or solid phase into the gas phase and then into the MS instrument was an insurmountable hurdle for researchers in the early days of biological mass spectrometry. In the 1980s, matrix-assisted laser desorption/ionization (MALDI)[182–184] and electrospray ionization (ESI)[185–188] provided *"wings for molecular elephants"* (John B. Fenn, Nobel lecture[189]) and revolutionized MS for large biomolecules, which was awarded with a share of the Nobel Prize in 2002. In MALDI, the analyte is embedded in a crystalline, light-absorbing matrix. Local excitation with a pulsed (UV-) laser beam creates a hot plume of desorbed ion clusters which can be mass analyzed, typically by time-of-flight instruments[190]. Mass spectrometry imaging[191,192] ingeniously uses the high spatial resolution of laser ionization and has recently rekindled interest in this method. Generally, applications of MALDI in proteomics are limited, mainly because the requirement for dried matrix spots prevents on-line coupling with liquid chromatography,

---

Box 1 | Glossary of mass spectrometry terms.

Duty cycle: The percentage of time in an acquisition cycle spent on actual analysis, and in particular the fraction of all ions used for mass analysis.

Dynamic range of an MS scan or run: Ratio from the highest to the least abundant peak in a mass spectrum or data file.

Isotope cluster: Set of ions with the same monoisotopic elemental composition, but different actual isotopic composition due to the distribution of ^{13}C vs. ^{12}C, for example.

Mass accuracy: Deviation of the measured mass-to-charge ratio to the known exact value typically given as a root mean square value in parts-per-million [ppm].

Mass-to-charge ratio (m/z): The mass of an ion divided by its number of charges.

Monoisotopic mass: The mass of a molecule that contains only the most abundant isotopes.

Box 2 | Survey of mass analyzers used in MS-based proteomics.
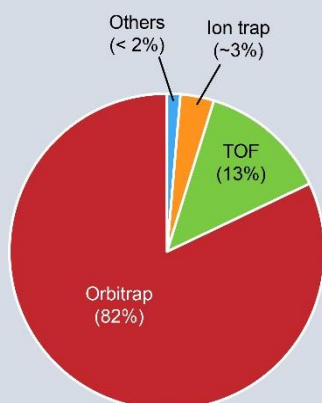
ProteomeXchange is a widely used platform for proteomics data submission and dissemination. At the time of writing, 5,437 publicly available datasets were deposited in ProteomeCentral. Categorizing by the mass analyzer used according to the submission details into Orbitrap, time-of-flight (TOF), ion trap and others (Fourier-transform ion cyclotron resonance MS and triple quadrupole) leads to the pie chart depicted.

http://proteomecentral.proteomexchange.org, Date accessed: 2018/06/14.

the prevalently used separation technique in the field. ESI by its nature yields a continuous beam of ions from liquid samples and has become the standard ionization method in LC-MS. Ion formation starts from charged droplets ejected from a Taylor cone[193] upon application of an electric field between the liquid flowing out of a capillary and the entrance of the mass spectrometer. Successive evaporation of solvent molecules yields smaller and smaller droplets until singly or multiply charged and desolvated analyte ions remain[194]. Decreasing the size of the initial droplets, by miniaturization and nL/min flow rates, increases the yield of ions and the transmission efficiency, leading to very high sensitivity in the range of attomoles and even below[195,196].

Once the ions enter the mass spectrometer, they are focused and guided through a succession of electrodynamic lenses and fields until they are ultimately mass analyzed. The quest for the 'ideal mass analyzer'[197] has produced numerous types of mass spectrometers with distinct performance metrics[198]. In proteomics, the Orbitrap mass analyzer has become the market leader soon after its introduction in 2005, taking over from time-of-flight instruments (**Box 2**). Both types of mass analyzers are almost always combined with an up-front quadrupole mass filter.

The Quadrupole Mass Filter

A linear quadrupole mass analyzer transmits ion through a square assembly of four cylindrical or hyperbolic rods[199–202]. Ions entering this assembly experience a quadrupolar electrodynamic field, which is generated by superimposed AC and DC potentials on all rods (**Fig. 8a**). The potential is applied pairwise to opposite rods in a way that the sign of the electric charge alternates periodically. Thus, a positively charged ion entering the quadrupole in *z* direction is first attracted by a negatively charged rod and, in the next moment, repulsed by the very same rod of now positive charge. This forces traversing ions onto oscillating trajectories in the *x*- and *y*-directions, which can be mathematically described by the Matthieu equations for an ion of given *m/z* as a function of the quadrupole geometry and the applied voltages[203]. Note that ions will only be transmitted successfully if the amplitude of their motion in both directions is less than the diameter of the quadrupole, or in other words, if they do not hit one of the rods. This offers the possibility to transmit individual ion species or *m/z* ranges of interest by appropriate AC and DC voltages (**Fig. 8 b,c**)[201,202]. Typically, they are in the range of 0.1-10 kV, and the AC component oscillates at radiofrequency (rf). In modern mass spectrometers, suitable sets of values can be calculated in real-time and applied in less than a millisecond, allowing the selection of hundreds of different ions per second. Quadrupole mass filters are typically operated at 'unit resolution', i.e. they transmit 1 Th wide *m/z* windows, but they can also be tuned to isolate 5- to 10-fold narrower windows at the expense of transmission
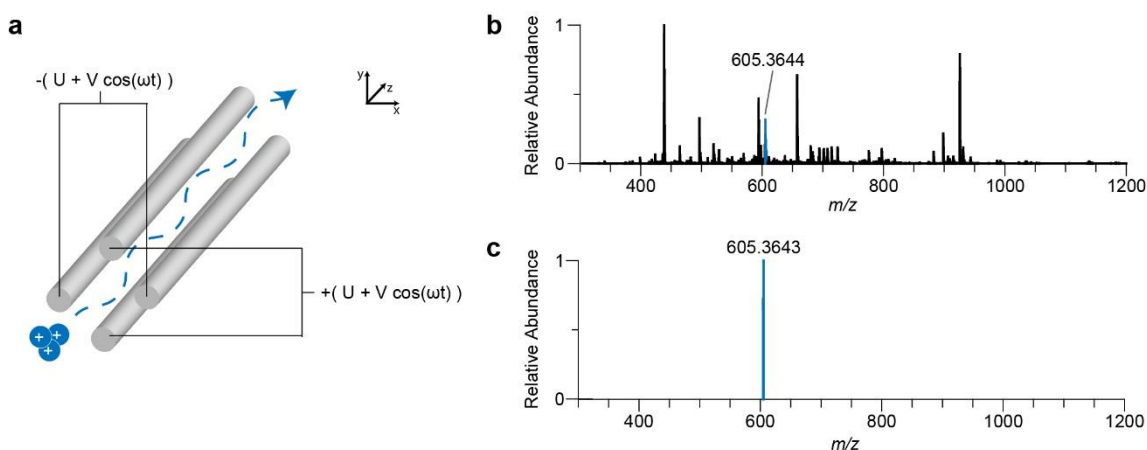
Figure 8 | The quadrupole mass filter. a, Construction and applied voltages. The blue line indicates the simplified trajectory of an ion species traversing the quadrupolar electric field. Mass ranges of interest can be filtered by adjusting the AC (*V*) and DC (*U*) voltages. Adapted from ref. 198. **b,** Full mass spectrum of a complex sample acquired with a hybrid quadrupole-Orbitrap mass spectrometer. **c,** As in b, but now isolating a narrow mass range centered around *m/z* 605.4 with the quadrupole mass filter.

efficiency. Another feature of the quadrupole (or higher *2N*-poles, such as hexa- and octopoles) is its ability to efficiently transmit very broad *m/z* ranges if no DC potential is applied ('rf-only' mode)[202]. This is often used to guide ions, for example through different vacuum stages or even on a bent flight path. In quadrupole 'collision cells', accelerated ions collide with inert gas molecules, typically N_2, and dissociate into smaller fragments which can then be mass analyzed to elucidate the precursor structure. A standard configuration combines three quadrupoles in a row, where the first one is used to filter a mass of interest – the precursor ions -, which are then fragmented in the succeeding collision cell, and the resulting fragment ions are mass analyzed by scanning the entire mass range in a third quadrupole[204,205].

The speed and high transmission efficiency of the quadrupole mass filter, together with its relatively low-cost, compact setup and ease of maintenance, have contributed to its popularity. It is one of the most widely used mass spectrometers, either as a stand-alone or as part of hybrid instruments. However, quadrupole mass filters are not suitable for high resolution and high accuracy mass analysis. In proteomics, triple quadrupole mass spectrometers are mainly employed for high-sensitivity quantification of targeted molecules and hybrid mass spectrometers with quadrupole mass filters are an integral part of essentially all proteomics laboratories.

The Orbitrap Mass Analyzer

"*Electrostatic axially harmonic orbital trapping*" for high resolution mass analysis was invented by Alexander Makarov at the turn of the millenium[206]. The 'Orbitrap' concept derives from the Kingdon trap[207], an ion trapping device assembled from a metal wire and a surrounding cylindrical electrode. The actual design has undergone several iterations and the commercial Orbitrap analyzer now features a spindle-like central electrode and an outer barrel-like electrode. However, the operational principle has remained the same[208–210]: To trap ions, an attractive electric potential is applied to the central electrode and, perhaps counter-intuitively, ions entering the trap on a trajectory offset with respect to the central electrode, do not hit the latter, but are rather trapped in an orbital motion around it (**Fig. 9**). This requires extremely fast switching of the kV trap potential during the time the ion package enters the Orbitrap (Voltage ramp in **Fig. 9**).
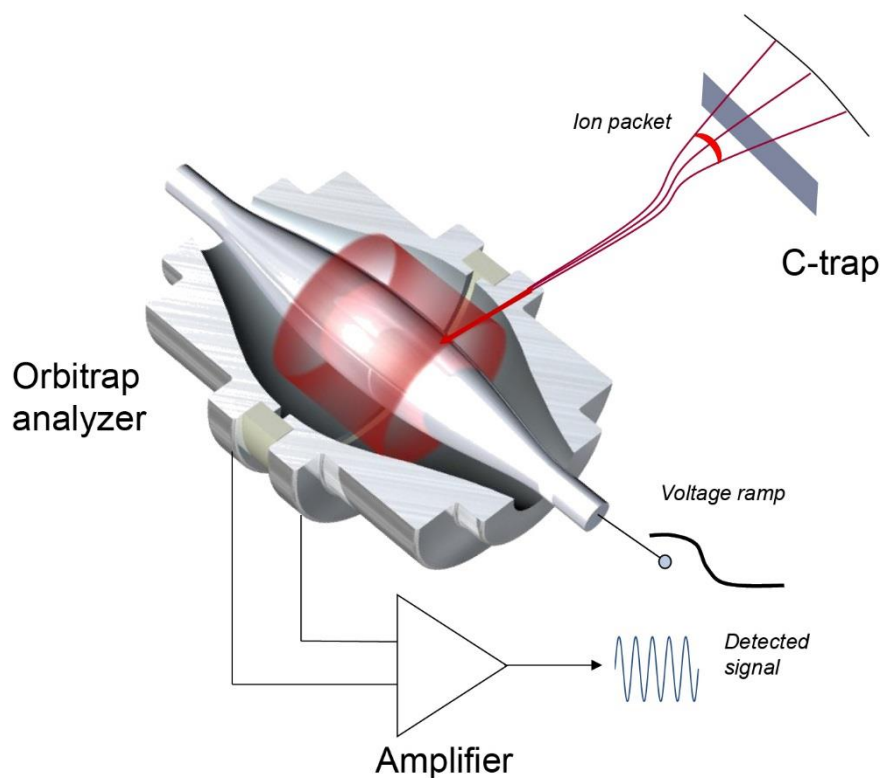
Figure 9 | The Orbitrap mass analyzer. Construction and operation of the C-trap and the Orbitrap mass analyzer. Red lines indicate the ion trajectories. During the voltage ramp, ions are accelerated into the Orbitrap mass analyzer and start to orbit around the central spindle electrode. The image current induced by the axial movement of the ions is amplified and detected as a function of time. Reprinted with permission from Thermo Fisher Scientific.

In the Orbitrap, a precisely engineered quadro-logarithmic electric field induces the complex spiral movement of trapped ions. It can be shown that the trajectories are dependent on *m/z*, which, in principle, allows deriving the mass of an ion by the current induced by its motion. However, the rotational motions further depend on the initial kinetic energy distribution and exact position of the ions, and are therefore not suitable for mass analysis. In contrast, the axial movement *along* the central electrode is independent of the starting conditions and can be expressed as a simple harmonic oscillation with the axial frequency ω [rad s^{-1}]:

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

where *e* is the elementary charge, z the number of charges, *m* the mass of the ion, and k a constant.

Makarov was the first to realize this opportunity for high resolution mass analysis, with very high accuracy and sensitivity[206]. Splitting the outer electrode in two halves allows to detect and amplify the differential image current induced by the axial movement of the ion cloud. Ions of the same *m/z* will move in phase, while ions of different *m/z* will oscillate at higher or lower frequencies. The sum of all individual components generates the total image current that is detected as a function of time, often termed a 'transient'. Fourier transformation decomposes the highly convoluted signal into its constituting frequencies and ultimately yields a high resolution mass spectrum. The resolution *R* of the mass analysis is thus directly linked to the ability to resolve two frequencies with a small difference $\Delta\omega$ and decreases inversely proportional to the square root of *m/z*:

$$R = \frac{\omega}{2\,\Delta\omega} = \frac{1}{2\,\Delta\omega}\,\sqrt{k\frac{ze}{m}} = \frac{m}{\Delta m}$$

Notably, recording the image current for a longer time increases the resolving power and gives the user an effective handle to balance resolving power with acquisition speed. For example, the current generation of high-field Orbitraps achieves a mass resolution of 15,000 at *m/z* 200 with only a 32 ms transient, and 120,000 at *m/z* 200 with a 256 ms transient.

A key step in the evolution of the Orbitrap mass analyzer has been the invention of the 'C-trap', essentially a bent rf-only quadrupole[210–212]. The C-trap decouples the Orbitrap mass analyzer from all preceding steps, such as ionization and mass filtering, and ensures precise alignment of the timings and control of the starting conditions for each ion packet, including the initial kinetic energy and the angular spread. Moreover, the C-trap accumulates and stores an ion packet, while another one is mass analyzed, which makes the Orbitrap compatible with continuous ion beams, e.g. generated by electrospray ion sources. However, the fact that ion packets rather than continuous ion beams are analyzed has important practical implications. It makes the overall percentage of the incoming ion beam that can be used for mass analysis dependent on the charge capacity of the C-trap with detrimental effects in cases where its maximum charge capacity is reached faster than the time needed for mass analysis. In this thesis, I study the consequences of this limitation in the context of MS-based proteomics and present a method termed 'BoxCar' that samples the ion current up to 10-fold more efficiently.

To enable fragment ion analysis with the Orbitrap mass analyzer, it has been combined with linear ion traps and quadrupole mass filters. The development of 'higher-energy collisional dissociation' (HCD)[80] either in the C-trap or in a dedicated multipole (termed HCD cell) enabled high resolution mass analysis of low *m/z* fragment ions, which typically escape detection with ion trap fragmentation. Recent advances in Orbitrap technology were primarily aimed at increasing the speed and sensitivity of the analysis, driven by the increasing demands of the MS-based proteomics community[210,213–215]. This encompassed reduced overhead times in ion handling and more efficient ion transmission in the front part of the instrument with an electrodynamic ion funnel and increasing the diameter of the ion entrance. Recently, the 'phase-constrained spectrum deconvolution method' (ΦSDM)[216] has been introduced, improving the signal processing and surpasses the uncertainty limits of Fourier transformation, offering an about three- to five-fold higher resolution at the same transient recording time.

Today, Orbitrap instruments are the workhorses in proteomics laboratories. The Q Exactive series combines a quadrupole mass filter with the Orbitrap mass analyzer and allows streamlined analysis of proteomics samples. The high-end Lumos series additionally incorporates a linear ion trap, which allows for very sophisticated MS experiments, such as the repeated fragmentation of fragment ions (MS^n experiments).

The Time-of-Flight Mass Analyzer

In contrast to the Orbitrap, time-of-flight (TOF) mass analyzers date back to the 1940s[217,218]. **Figure 10** shows a contemporary hybrid quadrupole TOF instrument, which was used in parts of the present thesis for proteomic measurements[219]. TOF analyzers build on a fundamental principle of physics, the conversion of electric potential energy (E_{el}) into kinetic energy (E_{kin}) and thus translational motion[198]. Ions are accelerated into a high vacuum field-free drift region by a pulse of high voltage *U:*

$$E_{el} = ez\,U = \frac{1}{2}mv^2 = E_{kin}$$

where *e* is the elementary charge, *z* the number of charges and *m* the ion mass.
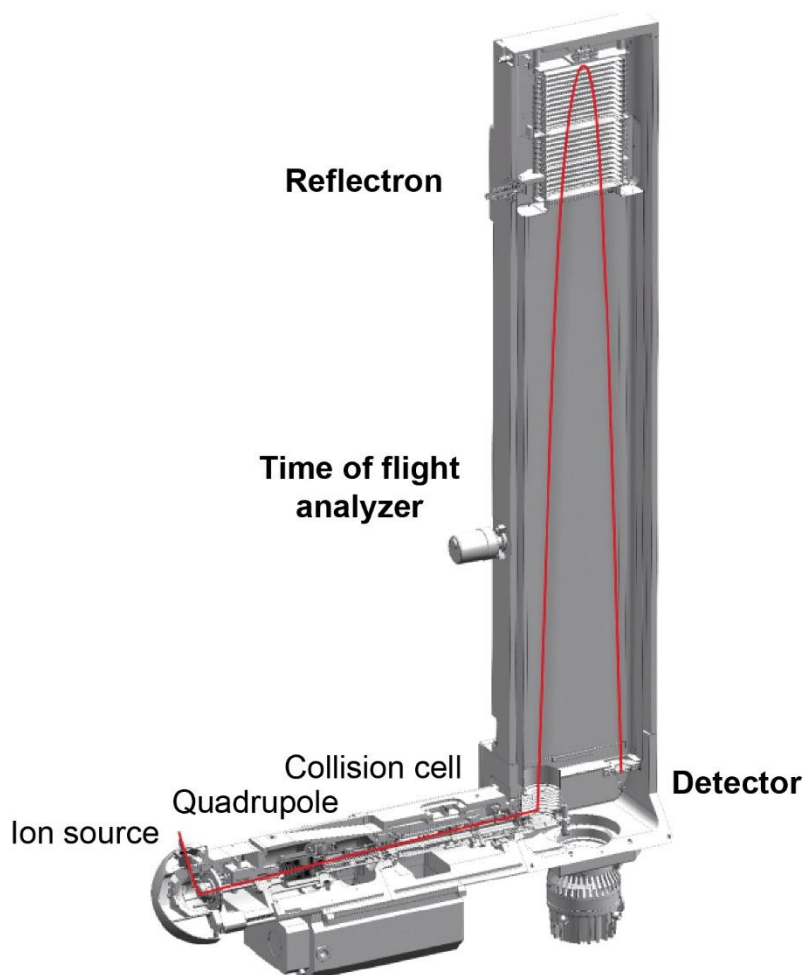
Figure 10 | The time-of-flight (TOF) mass analyzer. Construction of a hybrid quadrupole TOF mass spectrometer. Ions entering the vacuum are guided through the quadrupole and collision cell. An orthogonal accelerator pushes ions into the flight tube, where they enter a reflectron and finally impinge on a detector. Adapted from ref. 219.

Assuming that the initial kinetic energy of the ion was zero, and knowing the length l of the drift region, the flight time t of an ion is given by a direct function of its mass m:

$$t = \frac{l}{\sqrt{2eU}} \sqrt{\frac{m}{z}}$$

Typical flight paths are in the range of 1-3 m and ions are accelerated with 5-10 kV, resulting in flight times of only 10-100s µs. In contrast to a quadrupole mass analyzer and similar to the Orbitrap, all ions are analyzed together and full mass spectra are obtained from a single pulse. When combined with ESI sources, the continuous ion beam is diverted into an orthogonal accelerator[220,221] (**Fig. 10**), giving now rise to 10^3-10^4 spectra

28

per second. The high spectral rate is a distinct attraction of the TOF analyzer, allowing sampling a much larger proportion of the ion beam as compared with trapping-based instruments, which promises very high sensitivity, dynamic range and quantitative accuracy. However, in practice, the spectral rate is lowered as almost always multiple scans have to be accumulated to achieve sufficient signal-to-noise ratios[219,222]. Note that this is not due to the detection system, which is capable of counting and detecting single ions at high speed, but rather the result of inefficient ion transmission from the ion source to the detector.

From the fundamental equation above further follows that there is no upper mass limit and the resolving power is independent from *m/z*, but increases proportionally to the length of the flight path. In practice, the resolving power is mainly limited by geometric constrains and the energy distribution. Modern TOF analyzers employ microchannel plate ion detectors and digitizers (analog-to-digital converters) converting the time-resolved ion signal approximately every 200 picoseconds (see also below[223]).

A breakthrough in resolving power contributed by Russian scientists decades ago, was the invention of an 'ion mirror', a so-called reflectron, which forces ions onto a V-shaped flight path[224]. In addition to increasing the effective flight path by a factor of two, without expanding the instrument's footprint, the reflectron accounts for different starting conditions of the ions in the accelerator unit. This includes different starting times, locations, as well as the initial kinetic energy. A reflectron is a series of ring electrodes with an increasing repulsive potential, into which the ions enter, come to rest and are then accelerated out of the reflectron in the opposite direction. Ions with higher kinetic energy penetrate deeper into the reflectron as compared with their lower-energy counterparts and these slightly different flight paths largely correct for the initial energy spread, which reduces the mass peak widths significantly. State-of-the-art TOF mass analyzers reach resolutions in the range of 30,000-70,000 throughout the entire mass range and low ppm mass accuracies. The concept of reflectrons can also be extended to multiple reflections, for example W-shaped or higher-order flight paths[225–227]. These setups promise even higher resolution as a result of the multiplied flight paths. Conversely, longer flight times also imply lower spectrum acquisition rates and increased ion losses, resulting from scattering events and Coulombic repulsion, leading to an overall lower sensitivity.

TOF instruments have been widely used in proteomics[228], however, the invention of the Orbitrap largely banished them from proteomics laboratories. Recent increases in sensitivity and resolution and the unique advantage of fast scanning speeds, are creating renewed interest in TOF instruments[219]. The very fast scanning speed in combination with high resolution and mass accuracy also make it possible to nest ion mobility spectrometry between liquid chromatography and TOF mass analysis – a promising yet underexplored opportunity for MS-based proteomics[229].

Ion Mobility Spectrometry

Ion mobility spectrometry (IMS) separates ions in the gas phase by their mobility in an electrical field[230]. Classical ion mobility experiments are performed in 'drift tubes' (**Fig. 11**), in which ions are dragged through an inert gas at relatively low pressure by an electrostatic field gradient[231]. Ions with larger collisional cross sections (CCS, Ω) collide more often with residual gas molecules, and are therefore retained more than their smaller counterparts, which arrive earlier at the detector. The mobility of ions in an electrical field E is expressed as the field-normalized mobility coefficient K [$cm^2V^{-1}s^{-1}$], with the known drift length d and the measured drift time t_d:

$$K = \frac{d}{t_d E}$$

As the density of the residual gas, and thus the frequency of ion-neutral collisions, depends on the temperature T and the pressure p, K is further normalized to $T = 273\ K$ and $p = 760\ Torr$ (reduced ion mobility coefficient, K_0):

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

The collisional cross section Ω of ions colliding with neutral gas molecules can be calculated from K_0 using the Mason-Schamp equation[232],:

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

where z is the charge of the ion, e the elementary charge, μ the reduced mass, k_b Boltzmann's constant and T the temperature.
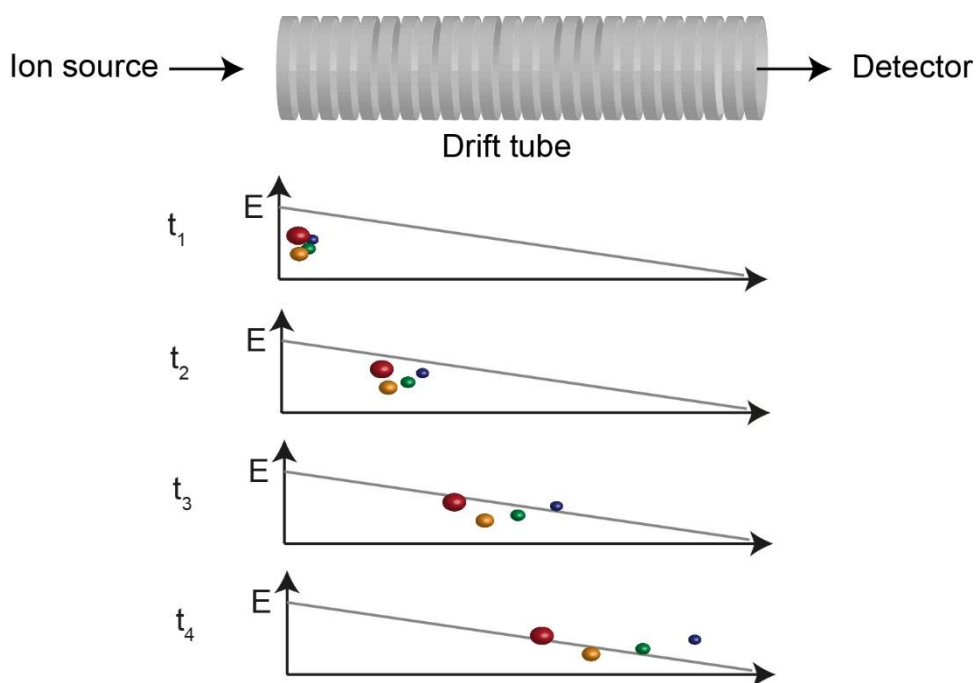
30

Figure 11 | Drift tube ion mobility spectrometry. The drift tube is constructed from stacked ring electrodes building an electrostatic field (E) as indicated in the diagrams. The diagrams further illustrate the separation of ions with different collisional cross sections at hypothetical time points in a drift tube experiment. Adapted from ref. 237.

In this thesis, the inverted reduced ion mobility coefficient $1/K_0$ is used for convenience due to its direct proportionality with the CCS (and thus roughly with m/z). Collisional cross sections reported in here refer to N_2 as the collision partner, even though other inert gases with different size and polarizability are also possible[233]. Notably, the models above are only valid with low electrical field strength, and increasing the field strength above a critical limit can alter the selectivity[231,232,234].

The history of IMS[235–239] (and MS) dates back to the late 19th century when Rutherford and Thomson studied ionization processes[240], and Zeleny[241] and Langevin[242] constructed apparatuses in which ions drifted in electrical fields. Later, Tyndall and others[243–245] pursued more mature versions of ion mobility spectrometers and in the early 1960s, McDaniel[246], McAfee and Edelson[247] pioneered the hybridization of IMS with mass spectrometry. Since then, a multitude of IMS devices has emerged, either stand-alone or coupled to MS, and IMS has become an integral part of the product range for many mass spectrometry vendors[238]. The millisecond time-scale of ion mobility analysis falls in

between typical up-front separation (seconds) and TOF mass analysis (microseconds) times, which makes IMS a promising third dimension to separate complex biological samples[248].

May and McLean[237] categorized the various ion mobility techniques based on their underlying separation principles: (1) time-dispersion, (2) space-dispersion, and (3) trapping and selective release. Representative and commercially available devices for each category are briefly discussed below.

Drift tube IMS (DTIMS) falls into the first category as ions arrive at the detector separated in time according to their relative ion mobility (**Fig. 11**). Most drift tubes are constructed from stacked-ring electrodes enabling long distances (up to two meters length[249]) between the ion source and the detector for high-resolution separation of ions with similar gas phase structures, such as isomeric peptides[250–252]. Ion clouds enter the drift tube through a pulsed ion gate and migrate in the direction of the electrical field. Even though it is tempting to increase the resolution by longer drift tubes, the increased footprint and the high voltage (> 1kV) requirements hamper routine use outside specialized laboratories. Moreover, axial diffusion and ion losses at the exit lower the sensitivity[253], together with the decreased duty cycle implied by long drift times. Radio-frequency confinement[254], ion funnels[253], and multiplexed injections[255] mitigate these obstacles to some extent, allowing the measurement of complex biological samples with adequate sensitivity[256]. A prototype instrument built by Smith and colleagues and described in ref. [256] has also shown promise to enhance proteomics measurements[257], further highlighting the potential of IMS.

Traveling wave ion mobility (TWIMS)[258] is another time-dispersive IMS type and is commercially available from Waters since 2006[259–261]. Here, radially-confined ions 'surf' through a stacked-ring ion guide on a traveling voltage wave (**Fig. 12**)[262]. Ions experience repeated pushes from these waves, which moves more mobile ions quickly through the device, while low-mobility ions collide more frequently with residual gas and therefore migrate slower. The ion mobility resolution depends on the speed and amplitude of the traveling wave, and resolutions ($\Omega/\Delta\Omega$) of 40 to 50 were reported[260,263]. Notably, the Mason-Schamp equation does not hold true under the conditions of constantly changing electrical fields, which somewhat complicates the interpretation of cross sections obtained by traveling wave experiments beyond the observed separation effect[264–266]. In one commercial configuration, the ion mobility device is placed downstream of the
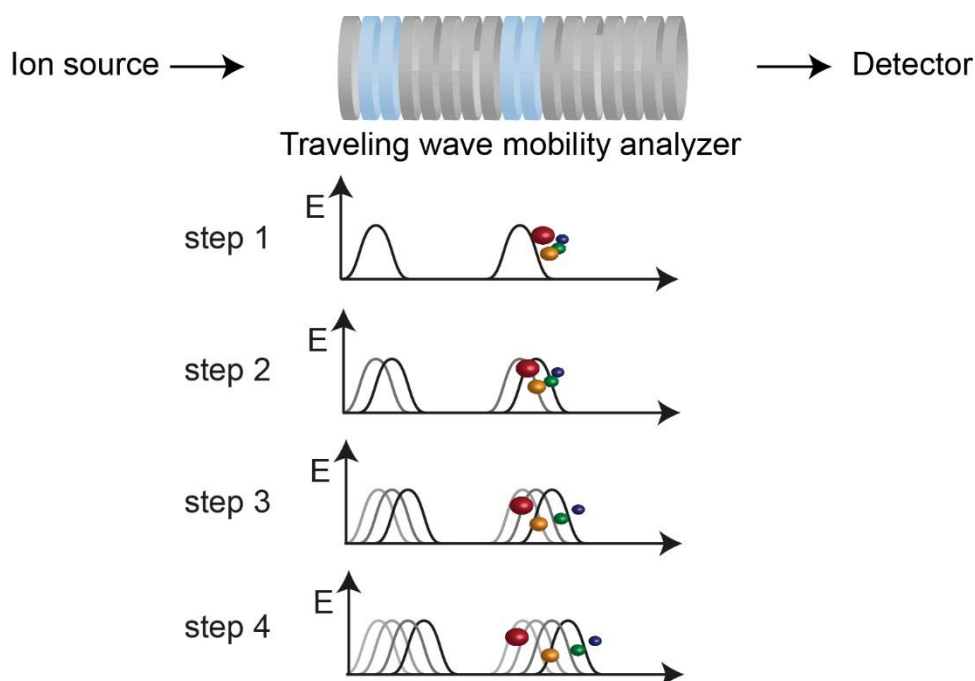
Figure 12 | Traveling wave ion mobility spectrometry. In the traveling wave analyzer there is an RF-potential between adjacent pairs of stacked ring electrodes. In addition, a DC potential (*E*) is applied to some of these electrodes (blue) and systematically offset in the following analysis step, generating traveling wave potentials. Ion mobility separation is induced by ions 'surfing' these waves as illustrated in the diagrams. Adapted from ref. 237.

quadrupole mass filter, which allows to separate fragment ions by their ion mobility. This has been advantageously used in proteomics by Küster and colleagues to increase the duty cycle of MS/MS scans by synchronizing the orthogonal accelerator with the ion mobility separation[267]. In data-dependent analysis of human cell lysates, this increased the number of protein identifications by a factor of two as compared with asynchronous operation. A disadvantage of data-independent analysis in general (see below), is that collision energies cannot be adjusted to individual precursor ions as done in data-dependent analyses where the mass and charge of precursor ions is known upon fragmentation. Tenzer and co-workers reported dramatically improved performance in data-independent proteomics after adjusting the collision energy as a function of ion mobility[268]. However, while promising, the TWIMS technology in general has not been widely adopted in the proteomics community, not least for reasons of underdeveloped post-processing software and compromised accuracy in the quantification of high-abundance peptides due to detector saturation with condensed ion packages eluting from the TWIMS device[269].
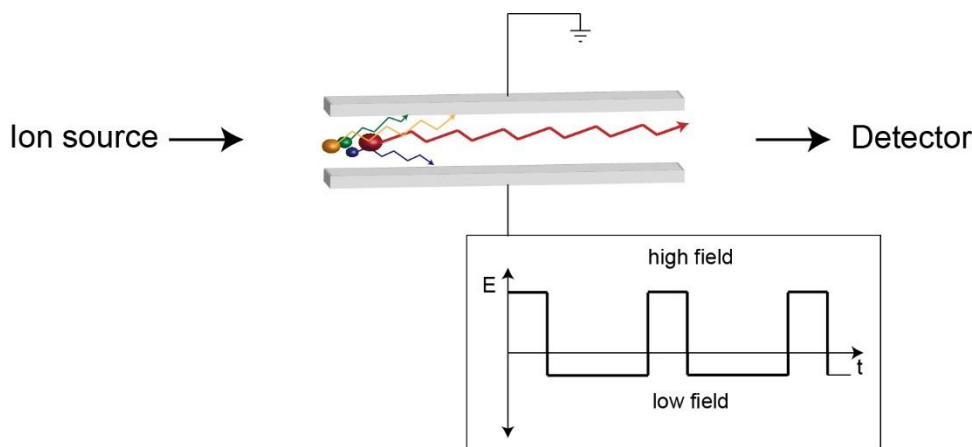
Figure 13 | Differential ion mobility spectrometry. An oscillating electric field (E) is applied to two planar electrodes. Ions experience alternating high and low electrical field strengths and are spatially dispersed as a function of their different ion mobility coefficients in both fields. A DC compensation voltage can be superimposed to select ions of interest. Adapted from refs. 238, 270.

Differential ion mobility spectrometry (DMS)[270] and field asymmetric ion mobility spectrometry (FAIMS)[271] are space-dispersive types of IMS. In contrast to DTIMS and TWIMS, this technology employs high field strengths and is based on the different mobility of ions in low-field and high-field conditions[272]. A DMS device comprises a pair of electrodes that are typically less than 2 mm apart, with ions migrating through this gap (**Fig. 13**). Alternating high and low electric fields are applied perpendicular to the ion path, in a way that the area under the curve is zero. This means that if there was no difference between low- and high-field mobility, all ions would stay on a stable trajectory. Instead, they are pushed stepwise to either one of the electrodes depending on whether they are more mobile in high or low field strengths. Interestingly, the difference in the high- and low-field ion mobility is analyte-dependent, which causes ions to spread perpendicular to the flight direction in a planar setup. Superimposing a direct current 'compensation voltage' forces selected ions of interest on a stable trajectory, while others impinge on the electrodes. The same principle applies to the FAIMS device, which is cylindrical instead of planar. Although a very high resolving power can be achieved under special circumstances[273], FAIMS and DMS devices are typically employed as low-resolution filters directly after electrospray ionization. In combination with trapping-based mass spectrometers, FAIMS can lower the fraction of background ions and therefore increase the dynamic range, as already demonstrated in proteomics applications[274–276]. However, decreased transmission of the targeted ion species has been

34

an issue so far[277]. Also, the compensation voltage does not allow conclusions about the collisional cross section and implies that only a fraction of the total ion population can be analyzed at a time[278].
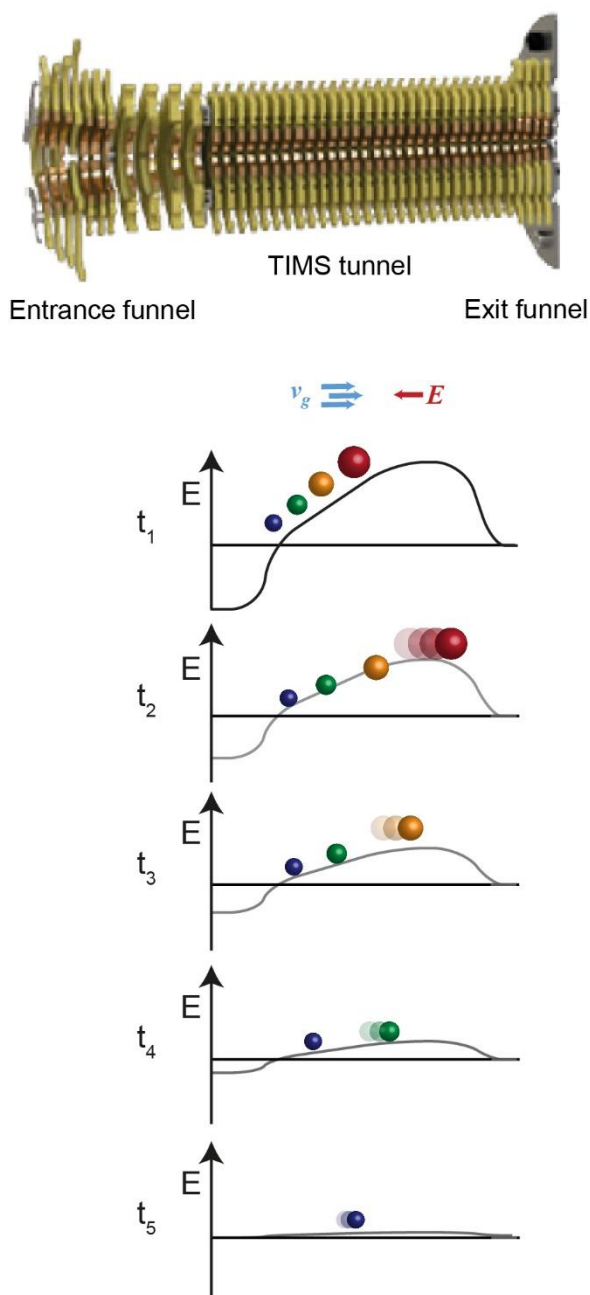
Figure 14 | Trapped ion mobility spectrometry. Ions are trapped in the TIMS tunnel build from stacked printed circuit boards with an inner diameter of 8 mm at a length of about 10 cm. A quadrupolar field is applied to each plate by two RF phases shifted by 180°. The RF field is superimposed by a DC potential (E) as indicated in the diagrams. In addition, trapped ions experience a drag from the incoming gas stream (v_g). Lowering the electrical field release ions to the downstream mass analyzer. Reprinted and adapted with permission from Bruker Daltonic.

Trapped ion mobility spectrometry (TIMS) is a relatively new IMS technique and was developed by Melvin Park and colleagues at Bruker[279,280]. The TIMS device traps ions and releases them as a function of their ion mobility[281]. TIMS is incorporated in the first vacuum stage of a mass spectrometer, where an electrodynamic ion funnel focuses incoming ions into a contiguous ion tunnel, which is the actual TIMS analyzer. Ions come to rest at a position defined by the counteracting forces of an electrical field and the drag of the incoming gas flow (**Fig. 14**). In this sense, TIMS reverses the concept of DTIMS. The mobility range of the trapped ions is determined by the upper and lower voltages of the analyzer (exit and entrance of the device, respectively), and this principle can also be used to selectively accumulate an ion species of interest to increase the sensitivity and ion mobility resolution[282–284]. After the initial accumulation and trapping step, lowering the electrical field strength releases ions from the TIMS device in the order of their ion mobility into the exit funnel, from where they are focused into the downstream mass analyzers. Note that, in contrast to DTIMS and TWIMS, ions with larger collisional cross sections (lower ion mobility) are trapped closer to the funnel exit and therefore arrive first at the detector, whereas ions with smaller collisional cross sections (higher ion mobility) arrive later. Interestingly, most of the ion mobility separation is effected at the electric plateau near the exit[285–287] and by slowing down the scan out time, high ion mobility resolution $> 200 \ \Omega/\Delta\Omega$ has been achieved[288,289]. In almost all applications, linear scan functions are used to release ions from the TIMS device. However the simplicity and flexibility of the device also allow non-linear scanning modes, which can increase the ion mobility resolution in mobility regions of interest[290]. Typical accumulation times in proteomics applications are in the range of 20 to 100 ms and yield ion mobility peak widths less than 2 ms (see below). Its unique design allows operating TIMS at relatively low potential differences (about 130 V) and reduces the dimensions of the device to about 10 cm in length and 8 mm inner diameter[281]. This implies a high sensitivity and furthermore allows to install two analyzers in a row, which increases the duty cycle up to 100%[291]. The physical similarity to classical DTIMS means that the same equations apply and offers a straightforward way for the determination of collisional cross sections[281]. The use of TIMS in proteomics is a core subject of the present thesis and the unique characteristics and operation possibilities of the device will be discussed in more detail in the results section.

Data Acquisition Strategies for Proteomics

Bottom-up proteomics generates hundreds of thousands of peptide species from biological samples such as whole-cell lysates or tissue sections[292]. Due to the large abundance range of proteins, post-translational modifications and differences in the ionization efficiency[27,47,293,294], peptide abundances span a magnitude huge range, easily exceeding 10 orders of magnitude even without considering peptide modifications for body fluids such as plasma[29,30,295]. With limited sample amounts and in the absence of signal amplification akin what is used for RNA or DNA sequencing, sampling the proteome reproducibly and comprehensively in a practical timeframe presents a paramount analytical challenge. At the level of MS data acquisition strategies the contemporary experimental toolbox of proteomics researchers addressing this problem comprises three main types shown in **Figure 15** – each with distinct advantages and disadvantages[48].
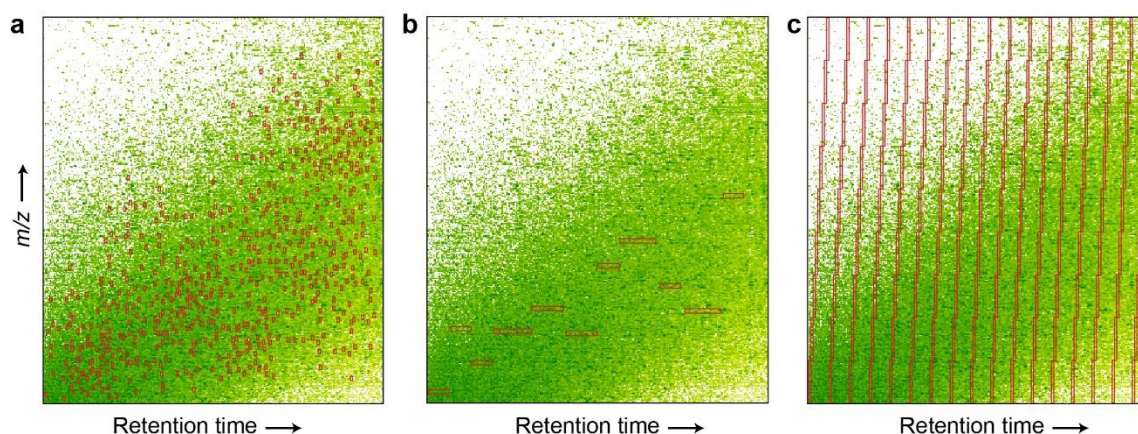
Figure 15 | Data acquisition methods in proteomics. a, In data-dependent acquisition (DDA) methods, the MS selects suitable precursors from survey scans for fragmentation with narrow isolation windows. **b**, In targeted proteomics, peptides of interest are constantly monitored for the time of the expected LC elution peak. **c,** In data-independent acquisition (DIA), the MS cycles repeatedly through pre-defined precursor windows covering the entire mass range of interest. Red boxes indicate the position of the quadrupole isolation window (not to scale). Adapted from ref 115.

Data-Dependent Acquisition

The most widely used acquisition method for unbiased large-scale and comprehensive proteome studies is data-dependent acquisition (DDA), also often referred to as 'topN' strategy. In DDA, the mass spectrometer constantly cycles through a full scan acquisition (survey scan) and N MS/MS scans[50,296]. To ensure a sufficient number of sampling points across a chromatographic peak, cycle times are optimally around one second. Precursors for the MS/MS scans are selected in real-time (hence 'data-dependent') based on the survey spectrum. As indicated by the name topN, the precursor selection algorithms prioritize higher abundant signals, as they are more likely to yield high-quality MS/MS spectra. Even though this concept is well-established, precursor detection algorithms are still subject of active research[297]. Tryptic peptides mostly bear two or more charges, which allows excluding singly charged species (which are predominantly chemical background ions) from the sequencing events and thus making more efficient use of the available analysis time. To further increase proteome coverage, already-sequenced precursors are excluded from re-sequencing for a certain time period in the range of chromatographic peak widths, a process termed 'dynamic exclusion'[298,299]. More sophisticated data-acquisition schemes that make use of real-time data-processing are largely underexplored, but are likely to evolve with the next generation of instrument application interfaces and increasing computational power[300,301]. DDA is compatible with label-based and label-free quantification strategies on the MS level as well as with MS/MS-based quantification with isobaric labeling.

The success of a DDA experiment largely depends on the rate and quality of the acquired MS/MS spectra, as well as the 'purity' with which precursor can be selected[292]. For this reason, proteome coverage has steadily improved with advances in the MS technology[48,228,302]. Almost 20 years ago, only relatively low sequencing rates of about 3 to 5 precursors per second were practicable and only very low resolution mass spectra could be obtained[51,303]. As the sensitivity of the mass spectrometers improved, more MS/MS spectra could be acquired per time unit[97,213,304]. Today, state-of-the-art mass spectrometers routinely operate at sequencing rates of 20 Hz or above and thereby quantify thousands of proteins in single runs[215,297]. When combined with up-front peptide fractionation, virtually complete proteomes can be measured[109]. Similarly, the total number of fragmented precursor ions for a given sample (and thus the proteome coverage) can be increased by injecting the same sample repeatedly, and in each injection selecting

precursors from only a subset of the entire mass range, a technique referred to as gas-phase fractionation[298,305,306]. However, even with very fast mass spectrometers and sequencing rates exceeding 40 Hz, only a fraction of all detectable precursors is fragmented, and ever faster rates inevitably imply a lower acquisition time per precursor and therefore a lower number of ions in each MS/MS spectrum. In addition, once all abundant precursors have been fragmented, less abundant peptides will be selected, further increasing the demand for sensitivity – a problem addressed in the present thesis through TIMS and the "PASEF" method. Incomplete sequencing, in particular of low-abundance species, also introduces stochasticity in the precursor selection process, which results in not every peptide being fragmented and identified in every run. This situation impedes statistical testing and biological interpretation of the results, because 'missing values' in the data matrix need to be imputed or even worse, proteins with too few data points need to be discarded[307]. To a certain extent, this problem is alleviated by transferring identifications between runs based on accurate precursor mass and after retention time alignment[101,135,181,308].

Targeted Proteomics

In some experiments, rather than exploring the proteome on a large scale, it may be sufficient to study a pre-defined 'targeted' subset of peptides or proteins by mass spectrometry, conceptually similar to ELISAs[168,309]. This is the case in some forms of hypothesis driven research, for example to validate biomarkers[310,311] or to screen activated biological pathways through 'sentinel' proteins[312]. While less comprehensive in terms of proteome coverage, targeted proteomics aims for highly sensitive, rapid and reproducible protein quantification. The MS acquisition method used is rooted in selected (or multiple) reaction monitoring (SRM or MRM), which has been applied in small molecule research for decades, and was initially developed on low-resolution triple quadrupole instruments[204,313,314]. The well-defined fragmentation pathways of peptides allow selecting a peptide precursor in the first quadrupole, collide it with inert gas in the second quadrupole (collision cell), and quantify the signal by constantly monitoring the signal of a selected fragment ion in the third quadrupole[315]. In targeted proteomics jargon, pairs of precursor and fragment ions are referred to as 'transitions'. As opposed to DDA, where precursor ions are selected on-the-fly and for single MS/MS scans, in targeted proteomics

the mass spectrometer is programmed to scan a pre-defined transition repeatedly for either the entire time of an experiment or in an elution time window where the peak is expected. With modern instrumentation, this approach is extended to 10-100s target proteins per run[316,317]. Due to the possibility to selectively monitor ions even in the complex background of tryptic digests, MRM can be quite sensitive and cover a reasonable dynamic range as exemplified in human plasma[318]. In the same manner, Aebersold and co-workers reported the quantification of yeast proteins spanning the whole abundance range with targeted proteomics[319,320]. Current workflows frequently make use of high-resolution mass analyzers such as the Orbitrap to read out all fragment ions (or transitions) in one MS/MS scan, referred to as parallel reaction monitoring (PRM), drastically increasing specificity[321]. Of note, in quadrupole-Orbitrap mass analyzers, the quadrupole can select one or multiple narrow mass windows for accumulation and storage in the C-trap, without immediately overfilling it[322]. In targeted analyses, this allows accumulating ions for a longer time than would have been possible with a standard full scan, often resulting in a several-fold increase in sensitivity[323]. In an attempt to increase the number of targets per run, researchers have employed trigger signals to reduce the 'monitoring time' in which targeted peptides are expected to elute from the column to seconds rather than minutes[324]. Targeted proteomics combined with isobaric labeling also shows promise to increase throughput, and can be extremely sensitive if the ion accumulation time is set to a large value[165]. Challenges inevitably arise due to the inherent complexity of proteomics samples and the resulting risk of interfering signals. This has fostered the development of software to automatically assess data quality and validity, in particular with large data sets[325,326]. The multitude of potential transitions and interferences complicates experimental design, and while balancing the number of targets, acquisition speed, chromatographic separation and sensitivity, the optimal choice of peptides can vary in-between experimental setups. This turns the selection of suitable peptides and transitions into a tedious process[327]. Therefore, in particular with human samples, the curation of well-characterized, quantifiable unique peptides for each protein ('proteotypic') and high-quality transitions remains a work in progress[328]. In addition to in-house libraries and the mining of community data, this motivated the development of large-scale *in vitro*[329] or synthetic[330] peptide libraries. In parallel, software tools are developed to assist users in streamlined assay development based on these libraries[331].

Data-Independent Acquisition

In recent years, data-independent acquisition (DIA) methods have gained attention as they promise to combine the advantages of targeted proteomics – sensitivity and reproducibility – with the comprehensive proteome coverage of DDA methods[332]. In DDA, the goal is to identify a single peptide sequence from each spectrum (with exceptions for co-isolated precursors[121]), which is ultimately limited by the number of MS/MS spectra that can be acquired per experiment. In principle, the number of identifications can be multiplied by (purposely) acquiring a combined spectrum for multiple peptide precursors ('multiplexing')[333]. However, depending on the number of simultaneously fragmented peptides, this generates increasingly convoluted MS/MS spectra, which are challenging to identify with conventional database search algorithms, resulting in lower sensitivity and identification success rates[334–337]. In contrast, DIA acquires multiplexed MS/MS theoretically from all eluting peptides by cycling through a pre-defined acquisition cycle rather than switching the precursor isolation window based on the observed ion signal. The term 'data-independent acquisition' was coined in 2004, when Yates and co-workers described a method in which an ion trap mass spectrometer constantly cycled through successive 10 Th windows covering a mass range from m/z 400 to 1,400 in total. Since then, a multitude of DIA methods have been described, ranging from fragmenting the entire precursor mass range at once[338–340] to selecting narrow windows with isolation widths similar to DDA[341–343]. Wider isolation windows enable faster acquisition cycles and cover a wider mass range with the downside of increasing spectral complexity, whereas narrower isolations contain fewer precursors and importantly can increase the dynamic range, although at the cost of increased cycle time[344]. In the beginning, the primary analysis strategy for this type of data was either to generate 'pseudo-MS/MS' spectra by grouping fragment ion signals together based on their correlation in the retention time dimension, or to search them directly with established DDA search engines[345]. A paradigm change occurred in 2012, when Aebersold and co-workers from Sciex proposed a different data analysis strategy with elements borrowed from targeted proteomics rather than the DDA search engine approaches. The novelty of this method termed "SWATH-MS"[346], was the 'targeted' or 'peptide-centric' data extraction strategy, meaning that the presence of a peptide of interest is inferred by matching to previously generated knowledge, as opposed to identifying a peptide from each spectrum ('spectrum-centric')[347]. Information about

peptide ions are stored in 'spectral libraries', which include the precursor mass, isotope distribution, retention time as well as the masses and relative abundances of the top fragment ions. Such libraries can be generated specifically for each project, e.g. using peptide fractionation, or more comprehensive community libraries can be used[348,349].

Assuming a complete record of all fragment ions is present in the sample over the entire LC run time, a DIA experiments represents a nearly-complete 'digital archive' of the sample, at least within the dynamic range of the measurement, which also implies a very high reproducibility of protein identifications[346]. However, in practice, the DIA methodology could not compete with DDA in terms of proteome coverage, as exemplified by less than 1,000 identified yeast proteins in 2 hours with the first generation of instruments and software[350]. By comparison, in the same year, the Coon group published the identification of 3,977 protein groups in about 80 min single runs with DDA. Contemporary workflows have improved in terms of proteome coverage, quantitative accuracy and post-processing software[170]. On the Sciex hybrid quadrupole TOF platform, which is widely-used for SWATH, 4,000 to 5,000 proteins can now be detected in single runs of a human cell line, and good reproducibility between laboratories has been reported[351]. As all fragment ion information is recorded, DIA also holds promise to infer quantitative information about isoform and post-translational modifications[352]. Recent advances in the data acquisition speed of the Orbitrap mass analyzer, makes it now very attractive to perform DIA experiments on this instrument type as well. The performance is already on par or exceeds that of standard DDA measurements[215,353,354]. As DIA is not limited by the sequential sequencing speed, it is particularly attractive for high throughput analysis with short LC-MS runs[355,356]. However, with increasing size of the datasets, libraries and proteomic depth, more rigorous statistical controls are required and this is still subject of ongoing research[349]. Interestingly, with the advance of MS technology and bioinformatics, library-free data analysis has again received increased attention[357,358] and this approach could potentially complement generic community libraries with sample-specific protein identifications in the future.

Aims of the Thesis

In the beginning of the century, the scientific community was overwhelmed by the decoding of the human genome, and characterizing the genome on a population-wide scale alone was thought to be the key to a deep understanding of human biology and the cure of many diseases. Genomics and transcriptomics technologies kept up with these high expectations and now readily provide data for single cells as well large human cohorts. At the same time, it has become clear that genomics by itself can only provide a part of the full picture, which conversely kindled interest in proteomics. However, the inherent complexity of proteomic systems and the absence of signal amplification turns proteomics on such a scale into a great technological challenge. Even though 'complete' human proteomes now appear to be within reach, the technology still lags behind in terms of sensitivity, coverage and throughput. The overarching goal of my thesis was thus to contribute novel MS data acquisition methods that should help to eventually overcome longstanding limitations in MS-based proteomics and pave the way for comprehensive and high-throughput proteomics studies.

A straightforward approach toward increased throughput of proteomics is to measure multiple samples in a single LC-MS experiment. Isobaric labeling has shown much promise in this regard and current commercially available kits allow multiplexing of up to eleven samples. However, a major drawback of this technology is the so called 'ratio compression', which refers to the distortion caused by interfering ions on the observed abundance ratios. So far, this problem was mainly addressed with expensive instrumentation and at the cost of speed and sensitivity. In my PhD work, I contributed to the development of a novel isobaric label, termed EASI-tag, which was designed to generate 'peptide-coupled reporter ions' with excellent yields (**Article 1**). In conjunction with a tailored MS-acquisition method, this enables interference-free multiplexing on widely used MS platforms.

A core subject of my thesis is ion mobility spectrometry as an extra dimension of separation in bottom-up proteomics. This work has been performed in close collaboration with Bruker Daltonic and builds on a high-resolution quadrupole TOF platform (**Article 2**). In the course of my project, this instrument was equipped with a Trapped Ion Mobility Spectrometry (TIMS) device. TIMS accumulates and stores ions, and then releases them to the downstream mass analyzer as a function of ion mobility. We made

use of this principle to devise the 'parallel accumulation – serial fragmentation' (PASEF) method (**Article 3**), in which the quadrupole quickly switches the isolation window to capture as many precursors as possible for fragmentation. In effect, this multiplies the peptide sequencing speed many-fold over conventional MS/MS methods, however, without diminishing sensitivity. While the first iteration was implemented on a laboratory prototype and limited in several aspects, PASEF has now become the heart of a full-fledged and commercially available proteomics platform, the Bruker timsTOF Pro (**Article 4**).

Even though TOF instruments hold great promise as demonstrated in here, the vast majority of proteomics laboratories nowadays use Orbitrap mass analyzers. A strong focus during the evolution of this technology since its invention in 2007 has been to increase the speed and sensitivity of MS/MS acquisition. This development entailed ion sources that allow more ions to enter the MS and an increased ion transfer efficiency inside the MS. In proteomic practice, this leads to the paradox that less than 1% of the available ions are now mass analyzed in full scans. During my PhD, I developed the 'BoxCar' method which increases the sampling time for low-abundance ions up to 10-fold (**Article 5**). This results in a dramatically increased dynamic range for full scan mass analysis. Employing a matching strategy, this culminated in the detection of 10,000 proteins in a 100 min single run of mouse cerebellum digest. BoxCar is of particular advantage in samples with a large dynamic range. In our laboratory, first applications of BoxCar included proteomics of human heart tissue (**Article 6**) and human plasma samples from a clinical study (**Article 7**).

II.   Accurate Multiplexed Proteome Quantification

Article 1: EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification

Sebastian Virreira Winter[1,2#], **Florian Meier**[1#], Christoph Wichmann[3], Juergen Cox[3], Matthias Mann[1,4] and Felix Meissner[2]

[1] Department of Proteomics and Signal Transduction; Max Planck Institute of Biochemistry; Germany
[2] Experimental Systems Immunology; Max Planck Institute of Biochemistry; Germany
[3] Systems Biochemistry; Max Planck Institute of Biochemistry; Germany
[4] Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen; Denmark

# These authors contributed equally to the work.

Label-free quantification has, at least in our laboratory, largely replaced stable isotope labeling as the method of choice for quantitative proteomics. This is mainly due the straightforward sample preparation and ready-to-use software, which both scale very well to large sample cohorts. Even higher throughput can be achieved with isobaric labeling, which allows multiplexing several samples in one analysis. TMT is a popular example for isobaric labels, but with current methods the quantitative accuracy is imperfect due to the co-isolation of other precursor ions.

In this manuscript, we describe the development of an 'easily abstractable sulfoxide-based isobaric' (EASI) tag that overcomes this so called 'ratio compression' problem. Key to the EASI-tag method is the quantification of 'peptide-coupled reporter ions' as opposed to the low-mass reporter ions typically quantified with TMT. They retain the precursor peptide information which allows quantifying even co-isolated peptides. This possibility was exemplified with the TMT molecule by Wuhr several years ago[166]. However, that approach did not find wide application as it was limited by, first, the inefficient peptide-coupled reporter ion generation with TMT and, second, co-isolation of natural ^{13}C isotopes, which convolutes the quantitative signal and forced them to discard one multiplexing channel entirely. Here, we set out to broaden the applicability of isobaric

labeling and enable accurate multiplexed quantification on widely used benchtop mass spectrometers.

To address the inefficient yield of peptide-coupled reporter ions, inspired by CID-cleavable disucccinimidyl sulfoxide (DSSO) cross-linkers[359], Sebastian Virreira Winter and Felix Meissner devised a sulfoxide-based label that fragments at collision energies below those required for peptide-backbone fragmentation. The original symmetric sulfoxide moiety by Kao and co-workers fragments at both sides of the sulfoxide and on each side, gives rise to alkene, sulfenic acid and thiol fragments. To prevent spreading of the quantitative signal into multiple fragments, the EASI-tag features an asymmetric sulfoxide that generates only one alkene fragment. In the manuscript, we proved this concept with thousands of EASI-tag labeled peptides from a HeLa digest and, importantly, demonstrated that the EASI-tag fragments at lower collision energies than TMT and yields the expected peptide-coupled reporter ions with high yield.

Having established the chemical design of the label, we devised a tailored MS acquisition strategy, which was my major contribution to this manuscript. As mentioned above, the co-isolation of naturally-abundant ^{13}C isotopes (about 1%) yields overlapping signals with the encoded ^{13}C stable isotopes in the label. To some degree, this can be corrected computationally, which is the strategy followed by Wuhr and colleagues. However, this inevitably introduces inaccuracies in the quantification and becomes nearly impossible for high abundance ratios. Instead, we aimed to exclusively isolate the monoisotopic ^{12}C peak. To do so, we employed a narrow isolation window and further shifted the isolation window to lower m/z, which in effect suppresses the ^{13}C peak for doubly and triply charged peptide ions. With high-performance quadrupole mass filters, this strategy did not noticeably affect overall transmission efficiencies and identifications rates in our hands, while enabling a straightforward read-out of accurate ratios, even those above 1:100.

In the manuscript we report the quantification of six isotope encoded channels with EASI-tag, which proved interference-free at a depth of 10,000 human and yeast proteins in a mixed proteome experiment, and with median coefficients of variations below 10%. Ongoing software developments will further boost those numbers.

EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification

Sebastian Virreira Winter [1,2,5], Florian Meier [1,5], Christoph Wichmann[3], Juergen Cox [3], Matthias Mann [1,4]* and Felix Meissner [2]*

We developed EASI-tag (easily abstractable sulfoxide-based isobaric-tag), a new type of amine-derivatizing and sulfoxide-containing isobaric labeling reagents for highly accurate quantitative proteomics analysis using mass spectrometry. We observed that EASI-tag labels dissociate at low collision energy and generate peptide-coupled, interference-free reporter ions with high yield. Efficient isolation of ^{12}C precursors and quantification at the MS2 level allowed accurate determination of quantitative differences between up to six multiplexed samples.

Mass spectrometry (MS)-based proteomics has matured remarkably and is now being applied in a wide variety of areas[1–4]. It delivers increasingly comprehensive identification of the constituents of biological systems[5]; however, their accurate quantification across multiple conditions remains an area of active development[6]. Although label-free quantification is universally applicable and has become widely used as a result of advances in data-acquisition schemes and algorithms, it is limited in terms of sample throughput and achievable quantification accuracy. Isotopic labeling allows multiplexing and provides relative quantification in the same spectra at either the MS1 (full spectrum) or MS2 (fragment spectrum) level. In particular, isobaric labeling reagents, such as tandem mass tags (TMTs)[7] and iTRAQ[8], have become popular because they make it possible to quantify digested proteins from any sample. They generate low-molecular-mass reporter ions in which the intensity of each of the reporters is derived from the corresponding isotope-labeled peptide. The major drawback of these methods is the fact that isolation of the precursor ions inevitably coisolates other precursors, distorting the reporter patterns. This 'ratio compression' effect has been addressed with narrow isolation windows, software corrections, additional gas phase manipulation, or further fragmentation of the peptide fragment ions[9–12]. However, these methods still do not achieve accurate quantification, and they may suffer from the decreased sensitivity and acquisition speed caused by their complexity. The fragmentation event leading to low-mass reporter ions can also generate precursor ions with the remnant of the tag[13], and this has been used for quantification[14,15]. Although promising, this method is limited by inefficient generation of the peptide-coupled reporters and the fact that the isotope distributions of different channels overlap.

To broaden the applicability of isobaric labeling and to enable the determination of accurate ratios on benchtop mass spectrometers, we set out to develop a molecule that would generate peptide-coupled reporters with a high yield. Inspired by readily MS-cleavable cross-linking reagents[16], we devised a sulfoxide-based moiety that dissociates at collision energies below those required for peptide backbone fragmentation. We modified the symmetric sulfoxide cleavage site to make it asymmetric, so that fragmentation generates only one instead of three different ion species for each tagged peptide (Supplementary Fig. 1). This decreases spectral complexity and increases individual fragment ion intensities, thus improving sensitivity and computational interpretation of fragmentation spectra. Incorporation of a primary amine-reactive moiety for covalent coupling of peptides via an active succinimide ester makes the reagent compatible with standard labeling procedures (Fig. 1a). The molecule, which we term EASI-tag, can be isotope-labeled for multiplex quantification and was synthesized in a six-plex version for this study (Supplementary Note 1). In contrast with other isobaric tags, fragmentation of EASI-tag-labeled peptides produced a neutral loss, thereby retaining the charge state of the precursor[7,8,17] (Fig. 1b).

Isolation of the entire natural isotope cluster for MS/MS leads to interference of the ^{13}C peak of each channel with the ^{12}C peak of the adjacent channel, complicating quantification[14]. We exclusively focused on doubly and triply charged precursors, which comprise 96% of all multiply charged features (Supplementary Fig. 2). Furthermore, we devised an asymmetric isolation window around the ^{12}C isotope peak, with an optimized transmission width of 0.4 thomson (Th) and an offset of –0.15 Th (Fig. 1c and Methods). This entirely suppressed the ^{13}C peak for doubly charged precursors and reduced it to 0.37% for triply charged precursors, which could be further corrected computationally (Supplementary Fig. 3a,b). The asymmetric isolation window with a narrow width did not noticeably decrease transmission and identification rates of the selected precursor ion, although it did minimize the isolation of coeluting peptides (Supplementary Fig. 3c–e). Those that are still selected together can be quantified separately as a result of the high resolution employed in the EASI-tag experiment (Fig. 1d).

We investigated the fragmentation behavior of thousands of EASI-tag-coupled peptides from a HeLa digest by varying the normalized collision energies (NCEs) for each of them. This confirmed that, compared with TMT, EASI-tag fragmented efficiently at lower collision energies (Fig. 1e) and with lower median NCEs than those required for peptide backbone fragmentation (Supplementary Fig. 4). At the same time, in-source fragmentation was negligible

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]Experimental Systems Immunology, Max Planck Institute of Biochemistry, Martinsried, Germany. [3]Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany. [4]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [5]These authors contributed equally: Sebastian Virreira Winter, Florian Meier. *e-mail: mmann@biochem.mpg.de; meissner@biochem.mpg.de
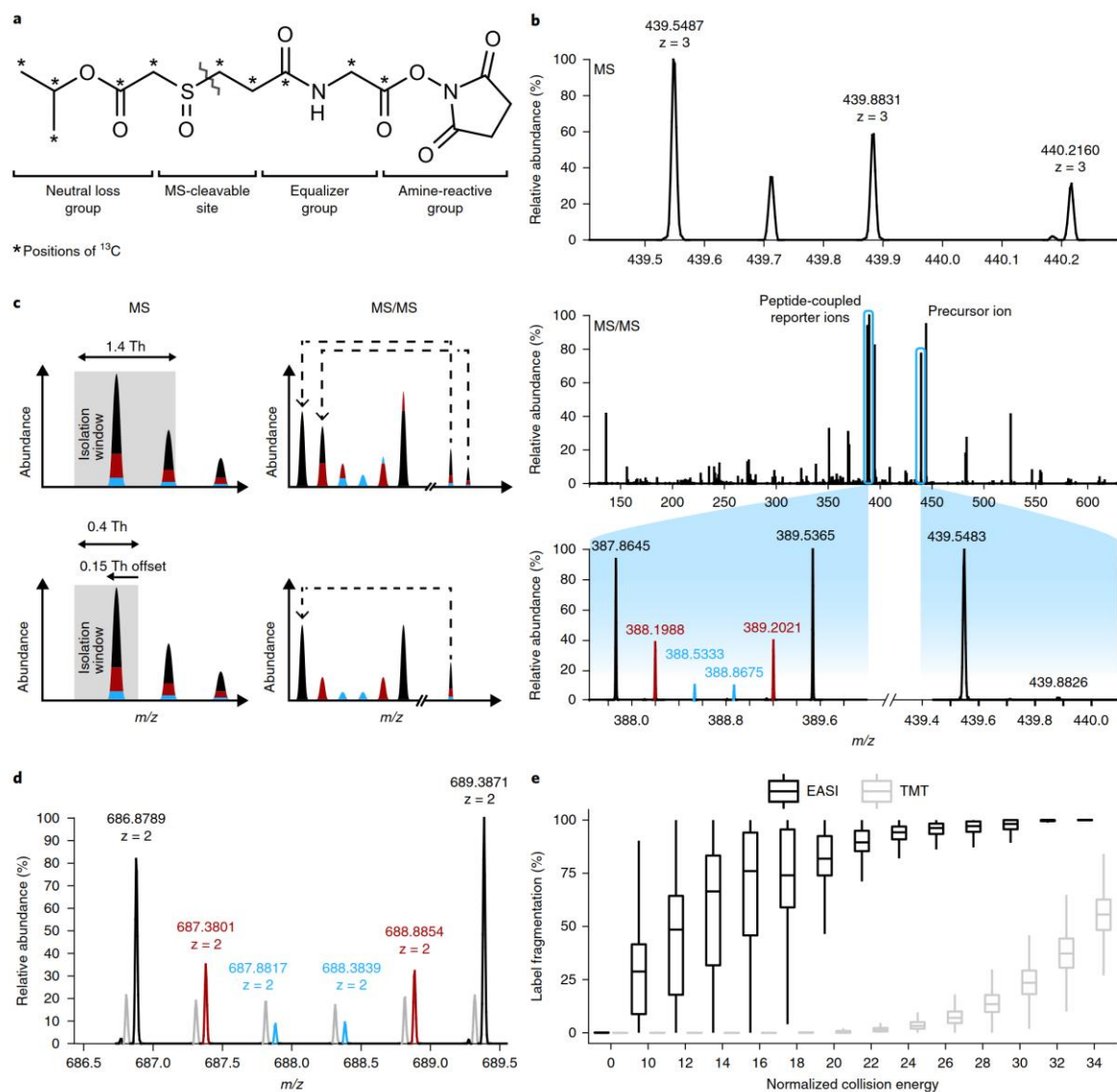
47

Fig. 1 | Design and concept of accurate and interference-free MS2-based proteome quantification with EASI-tag. a, Molecular structure of the six-plex version of EASI-tag. The isobaric labeling reagent is composed in a modular way of four functional groups and features a central sulfoxide moiety, which introduces an asymmetric, low-energy cleavage site (the zig-zag line indicates the fragmentation site). The stable-isotope-labeled positions of the neutral loss and equalizer groups for multiplexing are indicated by asterisks. Standard labeling protocols can be applied to couple peptides via the amine-reactive moiety. **b**, Mass spectra of an EASI-tag-labeled yeast peptide mixed at a ratio of 1:3:10:10:3:1. Higher-energy collisional dissociation fragmentation of the triply charged precursor ion abstracted the neutral loss group and yielded the peptide-coupled reporter ion cluster. **c**, Top, coisolation of the natural isotope cluster in a standard isolation window centered on the precursor ion convolutes the relative abundance of peptide-coupled reporter ions. Bottom, an asymmetric isolation window that suppresses the signal from adjacent isotope peaks and enabled direct quantification of reporter ions. **d**, The precursor mass information is retained in the peptide-coupled reporter ions for EASI-tag-labeled peptides. Colored peaks indicate the peptide-coupled reporter ions from an identified yeast peptide in a two-proteome experiment (mixing ratios: 1:3:10:10:3:1 for yeast and 1:1:1:1:1:1 for human). Gray peaks represent peptide-coupled reporter ions from a coisolated peptide. **e**, EASI-tag-labeled and TMT-labeled HeLa peptides were fragmented with normalized collision energies between 10 and 34 (N = 10,580, 10,568, 10,546, 10,063, 8,456, 5,513, 3,581, 2,613, 2,036, 1,656, 1,234, 669, 163, and 108 precursors for EASI-tag, and 19,098, 18,925, 18,573, 17,188, 15,959, 15,096, 14,425, 13,922, 13,422, 12,737, 11,767, 11,110, 10,131, and 6,249 precursors for TMT). Box plots show the median (center line), first and third quartiles (lower and upper box limits, respectively), and 1.5 × the interquartile range (whiskers). Spectra in **b** and **d** are from a single LC-MS/MS run, and data in **e** are derived from one experiment with eight LC-MS/MS runs per isobaric label.
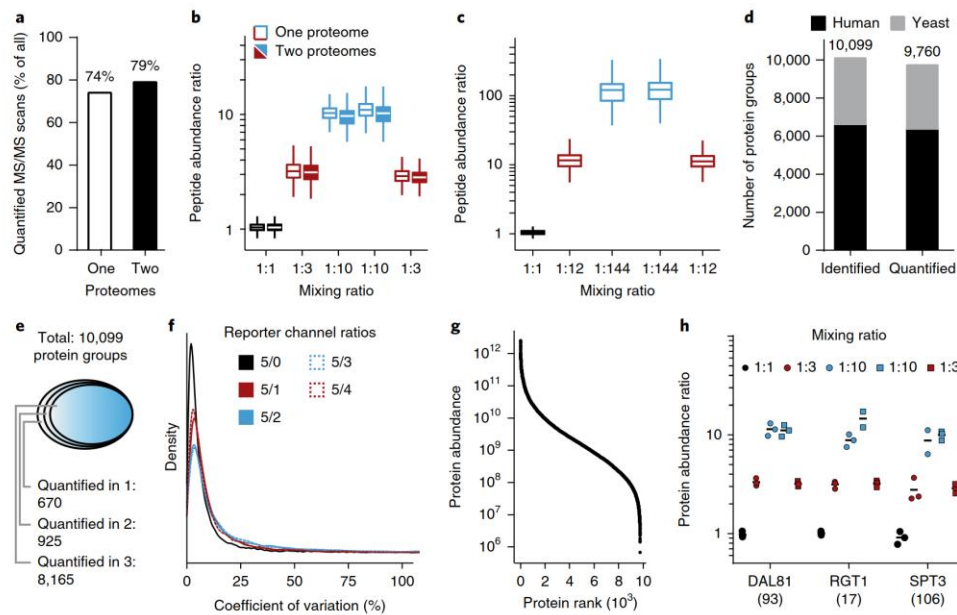
48

Fig. 2 | Proteome-wide evaluation of multiplexed quantification with EASI-tag. a, MS/MS scans in experiments of EASI-tag-labeled tryptic yeast peptides (one proteome) or of a mix of tryptic yeast and human peptides (two proteomes) with intensity signals in at least three peptide-coupled reporter channels. **b**, Quantified peptide-coupled reporter ion ratios (log$_{10}$) of yeast peptides mixed in a 1:3:10:10:3:1 ratio in the absence or presence of a HeLa background proteome (mixed 1:1:1:1:1:1) in a single-run LC-MS/MS experiment ($N = 11,906$ and 4,689 scans for 1:1, 11,860 and 4,666 scans for first 1:3, 8,463 and 3,298 scans for first 1:10, 8,601 and 3,366 scans for second 1:10, and 11,874 and 4,682 scans for second 1:3 mixing ratio). **c**, Quantified peptide-coupled reporter ion ratios (log$_{10}$) of yeast peptides mixed in a 1:12:144:144:12:1 ratio in a single-run LC-MS/MS experiment ($N = 9,540$ scans for 1:1, 9,515 scans for first 1:12, 1,487 scans for first 1:144, 1,778 scans for second 1:144, and 9,519 scans for second 1:12 mixing ratio). Box plots in **b,c** show the median (center line), first and third quartiles (lower and upper box limits, respectively), and 1.5 × the interquartile range (whiskers). **d**, A total of 10,099 protein groups were identified in 24 high-pH fractionations of the mixed yeast/HeLa two-proteome model in triplicate LC-MS/MS runs. Of these, 9,760 protein groups were quantified with intensity signals in at least three peptide-coupled reporter channels. **e**, 90% (9,090) of the identified protein groups were quantified in at least two replicate experiments. **f**, Coefficients of variation for relative quantification of human and yeast protein groups between triplicate LC-MS/MS runs. **g**, Yeast and human protein groups quantified with EASI-tag in the triplicate LC-MS/MS runs were ranked according to their abundance, which covered more than six orders of magnitude. **h**, Accurate quantification of low-abundance proteins such as yeast transcription factors with EASI-tag in the triplicate LC-MS/MS runs. Estimated copy numbers per cell[20] are indicated in parentheses; bars indicate the mean value for each ratio.

and was observed for less than 1% of the features detected by MaxQuant (Supplementary Fig. 5). Conceptually, this opens up the possibility of separately optimizing for peptide quantification and identification, which we then implemented. In Orbitrap mass spectrometers, fragmentation products from these two collision energies can be combined and read out in a single spectrum.

With the EASI-tag and a tailored acquisition strategy in place, we evaluated complex proteome quantification using the isobaric six-plex version of our molecule. We labeled tryptic yeast peptides with the EASI-tag, mixed them 1:3:10:10:3:1 with a 1:1:1:1:1:1 background of human peptides from the HeLa cancer cell line according to a two-proteome model[11], and analyzed the mixtures by single-run LC-MS/MS on an Orbitrap HF-X instrument (Methods).

We implemented the analysis of EASI-tagged peptides in MaxQuant[18], which resulted in 74–79% quantifiable MS2 scans in single runs (Fig. 2a). The peptide ratio distributions centered on the values expected from the mixing ratios and were equal in the two-proteome and one-proteome experiments, indicating the absence of any compression effects (Fig. 2b). It is very difficult to measure large ratios in isobaric labeling experiments, particularly when they are performed at the MS2 level[4,19]. Encouraged by our results, we next tested very large ratios by mixing digested yeast proteomes in 1:12:144:144:12:1 proportions. Again, the measured

ratios reflected the mixing ratios, without any evidence of compression (Fig. 2c). This capability could be crucial in biological situations in which very large fold changes are expected, such as in the regulated secretome[20], or when admixing a reference channel in a quantity that makes it likely to be picked for sequencing in the MS1 spectrum.

To investigate the EASI-tag method in a global proteome analysis, we separated the peptides from the two-proteome experiment into 24 fractions and analyzed the fractions in triplicate 120-min gradients. We obtained accurate quantification from more than 500,000 MS/MS scans (Supplementary Fig. 6). We identified 10,099 protein groups (3,485 yeast) with a mean number of 9.7 peptides (Fig. 2d). Even at this depth, the ratios closely matched the expected values, and 90% of all identified protein groups were quantified in at least two of the triplicates (Fig. 2e). Quantification accuracy of these proteins reached median coefficients of variation of 4.2%, 6.7%, 7.6%, 8.2%, and 6.2% for the reporter ion channel ratios over the entire abundance range (2.9%, 7.2%, 16.0%, 17.0%, and 7.6% for the 1:1, 1:3, 1:10, 1:10, and 1:3 yeast mixing ratios; Fig. 2f). EASI-tag-based proteome quantification covered a very large dynamic range, and even yeast transcription factors present at fewer than 100 copies per cell[21] were quantified correctly (Fig. 2g,h). Relative MaxQuant output tables can be found in Supplementary Data 1–6.

49

The EASI-tag strategy maintains the advantages of existing isobaric labeling methods and confers additional unique benefits. The tag is designed to fragment at lower collision energies than the peptide backbone, efficiently generating peptide-coupled reporter ions. In combination with the asymmetric ^{12}C-specific isolation window, EASI-tag allows accurate and precise quantification without any ratio compression. Such a capability is desired in many biomedical areas, particularly when absolute quantification of expression levels or concentrations of proteins is desired. As with alternative isobaric labeling techniques, we also anticipate good performance for post-translational modifications. Although we established the concept of accurate and interference-free MS2-based proteome quantification with the six-plex version of EASI-tag, its multiplexing capability could be increased further in the future, such as to a ten-plex, insofar as the reduced signal intensity of all channels does not affect EASI-tag performance. The previously applied strategy to increase the number of channels by neutron encoding[22] is tantalizing but would require a mass resolving power of $>250{,}000$ to distinguish between ^{13}C and ^{15}N isotopes in peptide-coupled reporter ions. In combination with a streamlined analysis workflow integrated in MaxQuant, our strategy is universally applicable for any shotgun proteomics workflow on widely used MS/MS instruments and paves the way for widespread use of isobaric labeling.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41592-018-0037-8.

References

1. Larance, M. & Lamond, A. I. *Nat. Rev. Mol. Cell Biol.* **16**, 269–280 (2015).
2. Lössl, P., van de Waterbeemd, M. & Heck, A. J. *EMBO J.* **35**, 2634–2657 (2016).
3. Aebersold, R. & Mann, M. *Nature* **537**, 347–355 (2016).
4. Rauniyar, N. & Yates, J. R.III. *J. Proteome Res.* **13**, 5293–5309 (2014).
5. Richards, A. L., Merrill, A. E. & Coon, J. J. *Curr. Opin. Chem. Biol.* **24**, 11–17 (2015).
6. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).
7. Thompson, A. et al. *Anal. Chem.* **75**, 1895–1904 (2003).
8. Ross, P. L. et al. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).
9. Bantscheff, M. et al. *Mol. Cell. Proteomics* **7**, 1702–1713 (2008).
10. Karp, N. A. et al. *Mol. Cell. Proteomics* **9**, 1885–1897 (2010).
11. Ting, L., Rad, R., Gygi, S. P. & Haas, W. *Nat. Methods* **8**, 937–940 (2011).
12. Wenger, C. D. et al. *Nat. Methods* **8**, 933–935 (2011).
13. Pichler, P. et al. *Anal. Chem.* **82**, 6549–6558 (2010).
14. Wühr, M. et al. *Anal. Chem.* **84**, 9214–9221 (2012).
15. Sonnett, M., Yeung, E. & Wühr, M. *Anal. Chem.* **90**, 5032–5039 (2018).
16. Kao, A. et al. *Mol. Cell. Proteomics* **10**, M110.002212 (2011).
17. Stadlmeier, M., Bogena, J., Wallner, M., Wühr, M. & Carell, T. *Angew. Chem. Int. Edn. Engl.* **57**, 2958–2962 (2018).
18. Cox, J. & Mann, M. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
19. Erickson, B. K. et al. *Mol. Cell* **65**, 361–370 (2017).
20. Meissner, F., Scheltema, R. A., Mollenkopf, H. J. & Mann, M. *Science* **340**, 475–478 (2013).
21. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. *Nat. Methods* **11**, 319–324 (2014).
22. Werner, T. et al. *Anal. Chem.* **86**, 3594–3601 (2014).

Acknowledgements

Author contributions

F. Meissner and M.M. conceived the study. F. Meissner and S.V.W. conceptualized and designed the EASI-tag. F. Meier, F. Meissner, and M.M. conceptualized the acquisition method, and C.W. contributed to its implementation. S.V.W., F. Meier, and C.W. performed the experiments. J.C. developed and implemented data analysis software. S.V.W., F. Meier, F. Meissner, and M.M. analyzed the data. S.V.W., F. Meier, F. Meissner, and M.M. wrote the manuscript with input from all other authors.

Competing interests

Additional information

50

Methods

Labeling reagents. TMTzero was purchased from Thermo Fisher Scientific. The EASI-tag labeling technology was conceptualized by the authors, and custom chemical synthesis of the EASI-tag reagents was performed by Taros Chemicals GmbH, Germany. The synthesis route is shown in Supplementary Note 1. For proof-of-principle experiments, we used a non-isotopically coded version (EASI-tag-zero), and for MS2-based proteome quantification, we used six isotopomers (six-plex EASI-tag) containing five ^{13}C-labeled carbons at the positions indicated in Fig. 1a. Purity was > 95% by ^{1}H and ^{13}C NMR spectroscopy (Taros Chemicals).

Cell culture. The human cervical cancer cell line HeLa (S3 clone; ATCC) was cultured in Dulbecco's modified Eagle's medium supplemented with 10% FBS, 20 mM glutamine, and 1% penicillin–streptomycin (all from PAA Laboratories). The cells were collected by centrifugation and washed once with cold phosphate-buffered saline before recentrifugation. *Saccharomyces cerevisiae* strain BY4741 (EUROSCARF) was grown to mid-logarithmic phase (OD_{600} 0.7) at 30 °C in yeast extract peptone dextrose medium (10 g/l BactoYeast extract, 20 g/l peptone, and 2% (w/v) glucose) and collected by centrifugation. The cells were washed with cold Milli-Q water and pelleted by centrifugation. All cell pellets were flash-frozen in liquid nitrogen and stored at –80 °C.

Lysis and digestion. Cell lysis, protein reduction, and alkylation were performed in a one-pot SDC buffer system containing chloroacetamide[21] (PreOmics GmbH). Briefly, HeLa and yeast cell pellets were resuspended, boiled at 95 °C for 10 min, and sonicated for 15 cycles for 30 s with a Bioruptor Plus sonication device (Diagenode) to enhance cell disruption before the addition of equal amounts of the proteases LysC and trypsin in a 1:100 (w/w) ratio. The enzymatic digestion was carried out overnight and was stopped with trifluoroacetic acid (TFA) at a final concentration of 1% (v/v). The acidified protein digests were desalted on C18 StageTips[23] and eluted with 80% acetonitrile/0.1% TFA (v/v).

Isobaric labeling. Approximately 20 μg of desalted peptide mixtures were reconstituted in 50 mM HEPES buffer, pH 8.5. The EASI-tag or TMTzero isobaric label was reconstituted in water-free acetonitrile and added in fourfold excess (w/w) to couple primary and N-terminal amino groups to the N-hydroxdysuccinimide (NHS)-activated ester. Acetonitrile was added to a final concentration of 33% (v/v), and peptides were labeled for 1 h at 25 °C. Unreacted NHS ester was hydrolyzed by the addition of 10 volumes of 1% TFA. The labeled peptides were mixed at defined ratios, purified on C18 StageTips with 2% acetonitrile/0.1% TFA, and eluted as above. The eluate was evaporated to dryness in a vacuum centrifuge, and the peptides were reconstituted in 2% acetonitrile/0.1% TFA for single-run LC-MS/MS analysis or high-pH reverse-phase fractionation. The labeling efficiency was above 98% for all experiments.

High-pH reversed-phase fractionation. Purified protein digests were fractionated on a 30-cm reversed-phase column (250-μm inner diameter) with a 'spider fractionator'[24] (PreOmics GmbH) system coupled online to an EASI-nLC 1000 chromatograph (Thermo Fisher). For EASI-tag-labeled peptides, separation was performed at a flow rate of 2 μl/min with a binary buffer at pH 10 (PreOmics GmbH) starting from 10% B, followed by a stepwise increase to 45% B within 75 min and 95% B within 7 min. For TMT-labeled peptides, the gradient started at 3% B and increased stepwise to 30% B within 45 min, 40% B within 12 min, 60% B within 5 min, and finally 95% B within 10 min.

In the deep two-proteome experiment, we concatenated 16 fractions automatically by shifting the rotor valve every 120 s and collecting into 0.2-ml tubes. To collect eight fractions (for breakdown curves; see below), we shifted the rotor valve every 100 s. Collected fractions were vacuum-centrifuged to dryness and reconstituted in 2% acetonitrile/0.1% TFA for LC-MS/MS analysis.

Liquid chromatography and mass spectrometry. LC-MS/MS analysis was performed on an EASY-nLC 1200 ultra-high-pressure system coupled online to a Q Exactive HF or Q Exactive HF-X mass spectrometer via a nano-electrospray ion source (all from Thermo Fisher Scientific). Approximately 0.5–2 μg of peptides were separated at a constant flow rate of 300 nl/min on a 45-cm reverse-phase column (75-μm inner diameter) packed with ReproSilPur C18-AQ 1.9-μm resin (Dr. Maisch GmbH). Mobile phases A and B were 100/0.1% water/formic acid (v/v) and 80/20/0.1% acetonitrile/water/formic acid (v/v/v). For EASI-tag-labeled samples, the initial concentration of 10% B was linearly increased to 40% B within 95–100 min, and then increased further to 60% within 5–6 min and 95% B within 3–5 min, with a 3–5-min plateau before re-equilibration. TMTzero-labeled and unlabeled peptides were separated as above but with initial gradients from 10% to 35% B and from 5% to 30% B, respectively.

Proteomics experiments were acquired with a data-dependent top10 or top12 method. Full MS scans were acquired in the range of m/z 300–1,650 at a resolution of 60,000 at m/z 200. The automatic gain control (AGC) target was set to 3e6. Higher-energy collisional dissociation MS/MS scans were acquired with a stepped normalized collision energy of 11 and 21 (50% each) at a resolution of 30,000 or 60,000 at m/z 200. Precursor ions were isolated in an asymmetric 0.4-Th window (offset –0.15 Th) and accumulated to reach an AGC target value of 1e5–5e5 or for

a maximum of 110 ms. Only doubly and triply charged precursors were selected for fragmentation, and the monoisotopic peak was preferably isolated. The acquisition software was adapted via the Thermo Fisher application programming interface to implement the latter functionality. Precursors were dynamically excluded for 30 s after the first fragmentation event.

To investigate the fragmentation behavior of EASI-tag-zero- and TMTzero-labeled peptides, we acquired data-dependent breakdown curves with 'Full MS' parameters set as above. Higher-energy collisional dissociation MS/MS scans were acquired at a resolution of 30,000 at m/z 200. The AGC target value and maximum ion injection time were set to 2e5 and 60 ms, respectively. Suitable precursor ions were repeatedly isolated with a symmetric isolation window of 1.5 Th or an asymmetric 0.4-Th window (offset –0.15 Th) 14 times per acquisition cycle with increasing collision energy. The normalized collision energy was set to 0 for the first scan and subsequently increased from 10 to 34 with a step size of 2.

To evaluate the monoisotopic isolation of the ^{12}C peak, we acquired a similar dataset with an unlabeled HeLa digest. Here, we fixed the collision energy at 0 (SIM scans) for all scans and shifted the center of the 0.4-Th quadrupole isolation window incrementally to lower m/z values within one acquisition cycle, starting from the precursor-centered position. The isolation offset was 0 for the first scan and increased from –0.1 to –0.2 Th with 0.01-Th increments for the subsequent scans.

To further evaluate the ion-transmission characteristics of the analytical quadrupole for narrow isolation windows, we acquired SIM scans from an unlabeled HeLa digest and decreased the window widths from 1.4 Th to 0.4 Th (0.2-Th steps), with windows either symmetric or asymmetric with offsets ranging from –0.65 Th to –0.15 Th (0.1-Th steps). SIM scans were acquired at a resolution of 60,000 at m/z 200. The injection times for each precursor ion were fixed at 5 ms.

To evaluate the sensitivity and ion transmission with an asymmetric isolation window, we acquired an unlabeled HeLa digest with a data-dependent top10 method in triplicate LC-MS/MS runs. We used either symmetric isolation windows of 1.4-Th and 0.4-Th widths or an asymmetric isolation window of 0.4-Th (–0.15-Th offset) width. Full MS parameters were set as above and MS/MS scans were acquired at a resolution of 60,000 at m/z 200. The AGC target value and maximum ion injection time were set to 5e5 and 110 ms, respectively.

Data analysis. Raw spectral data were extracted with the MSFileReader (v3.0.31, Thermo Fisher Scientific) or MSConvert (v3.0.11537, ProteoWizard), and proteomics MS Raw files were processed with MaxQuant[18] (version 1.6.1.1) for identification and quantification. The MS/MS spectra were searched against tryptic peptides with less than three missed cleavages derived from human and yeast reference proteomes (UniProt; accessed February 2017), as well as a list of potential contaminants with the Andromeda search engine. The search included cysteine carbamidomethylation as a fixed modification and methionine oxidation as a variable modification. For quantitative proteomics experiments, EASI-tag was defined as a modification of the type 'isobaric label' at any lysine residue or the peptide N terminus with a delta mass of 266.0839 Da ([12]C$_5$[13]C$_5$H$_{15}$O$_5$NS). For analysis of the breakdown curves, EASI-tag-zero and TMTzero were defined as modifications of the type 'isobaric label' with delta masses of 261.0671 Da ([12]C$_{10}$H$_{15}$O$_5$NS) and 224.1525 Da ([12]C$_{12}$H$_{20}$O$_2$N$_2$), respectively. To avoid analysis of SIM scans (NCE 0 scans) as full scans by MaxQuant, we removed these before MQ analysis. The minimum peptide length was set to 6 amino acid residues, and the maximum peptide mass was limited to 4,600 Da. The maximum allowed initial mass deviation was 4.5 p.p.m. for precursor ions after nonlinear recalibration and 20 p.p.m. for fragment ions. The initial filtering criteria for MS/MS scans was set to 20 peaks per 100-Da interval, and the 'second peptide' option was disabled. Where applicable, MS/MS precursor mass values were corrected for the isolation window offset. Identifications were controlled at a maximum false discovery rate of 1% for both peptide spectrum match and protein level.

For the quantification of EASI-tag-coupled reporter ions, a mass tolerance of 6 p.p.m. was set in MaxQuant. Neutral loss masses were defined as 150.0351 Da ([12]C$_5$[13]C$_0$H$_{10}$O$_3$S), 151.0384 Da ([12]C$_4$[13]C$_1$H$_{10}$O$_3$S), 152.0418 Da ([12]C$_3$[13]C$_2$H$_{10}$O$_3$S), 153.0451 Da ([12]C$_2$[13]C$_3$H$_{10}$O$_3$S), 154.0485 Da ([12]C$_1$[13]C$_4$H$_{10}$O$_3$S), and 155.0518 Da ([12]C$_0$[13]C$_5$H$_{10}$O$_3$S) for reporter ion channels 0, 1, 2, 3, 4, and 5, respectively.

Bioinformatic analysis. Bioinformatic analysis was performed in Perseus[25] and the R[26] statistical computing environment. Decoy database hits were strictly excluded from the analysis, as were potential contaminants and proteins that were identified exclusively by one site modification.

For EASI-tag-based quantification, we required reporter intensities in at least three channels and quantified peptides only if the ^{12}C peak was isolated and the charge state was 2 or 3. After filtering, peptide-coupled reporter ion intensities for each channel were calculated from the precursor ion with fragmented tag and all detectable b- and y-ion fragments with fragmented tag. To obtain ratios, we divided the reporter intensity in channel 5 by reporter intensities in channels 0, 1, 2, 3, and 4. Ratios were corrected for isotope impurities of the isobaric tags (Supplementary Table 1).

To determine the fragmentation efficiencies of EASI-tag-zero and TMTzero, we first processed breakdown curves with MaxQuant for peptide identification. Scan cycles with any MS/MS scan with an AGC fill below 50% were excluded from the

analysis. We calculated label fragmentation efficiencies by dividing the intensity of the peptide-coupled reporter ion by the summed intensities of the precursor ions with intact and fragmented label. Water losses were included in the analysis for both peptide-coupled reporter ions and precursor ions. Only scan cycles with identified, labeled peptides were included in the analysis, and if either the precursor ion or the peptide-coupled reporter ion was not detected, its intensity was set to 0. If both the precursor and peptide-coupled reporter ions were not detected or their summed intensity was below 5% of the precursor intensity in the SIM scan, the scan was excluded from the calculation.

For Supplementary Fig. 4, the NCE of the label fragmentation was defined as the minimal NCE at label fragmentation (see above) of 25%. If label fragmentation did not reach 25%, the NCE of the label fragmentation was defined as the NCE of the maximal label fragmentation. The NCE of peptide backbone fragmentation was defined as the NCE at which the sum of the precursor and peptide-coupled reporter intensities was below 20% of the precursor intensity in the SIM scan (NCE = 0). If the sum of the precursor and peptide-coupled reporter intensities did not drop below 20% of the precursor intensity in the SIM scan, the NCE of peptide backbone fragmentation was defined as the NCE at which the sum of the precursor and peptide-coupled reporter intensities reached its maximum.

To evaluate the quadrupole transmission, we analyzed precursor ion intensities and intensities of the M + 1 ion in SIM scans from doubly or triply charged precursors. Zero values were removed before mean calculation.

In-source fragmentation was determined for HeLa peptides labeled with the EASI-tag-zero reagent and fractioned into eight fractions by high-pH reversed-phase fractionation. Multiply charged features from the allpeptides.txt output table of MaxQuant with an $m/z > 450$ (otherwise the in-source-fragmented ion would be outside the scan range) were analyzed for in-source-fragmented features with a retention time of ± 2 s around their retention time. The mass tolerance was set to 6 p.p.m.

Data presentation. Box plot elements of all figures are defined as follows: center line, median; lower and upper box limits, first and third quartiles; whiskers, maximum $1.5 \times$ interquartile range; outliers not displayed.

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

Data availability. Mass spectrometric data and MaxQuant analysis files have been deposited in the ProteomeXchange Consortium (http://proteomecentral. proteomexchange.org) via the PRIDE partner repository[27] with the dataset identifier PXD007165. All proteomics data are summarized in the Supplementary Data. Source data for Figs. 1 and 2 are available online.

References

23. Rappsilber, J., Mann, M. & Ishihama, Y. *Nat. Protoc.* **2**, 1896–1906 (2007).
24. Kulak, N. A., Geyer, P. E. & Mann, M. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
25. Tyanova, S. et al. *Nat. Methods* **13**, 731–740 (2016).
26. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (The R Foundation, Vienna, Austria, 2008).
27. Vizcaíno, J. A. et al. *Nucleic Acids Res.* **44**, D447–D456 (2016).

III. (Trapped) Ion Mobility Spectrometry for Proteomics

Article 2: The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics

Scarlet Beck[‡], Annette Michalski[§], Oliver Raether[§], Markus Lubeck[§], Stephanie Kaspar[§], Niels Goedecke[§], Carsten Baessmann[§], Daniel Hornburg[‡], **Florian Meier**[‡], Igor Paron[‡], Nils A. Kulak[‡], Juergen Cox[¶] and Matthias Mann[‡]

[‡] Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany;
[§] Bruker Daltonik GmbH, Fahrenheitstr. 4, 28359 Bremen, Germany;
[¶] Computational Systems Biochemistry, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

When I joined the laboratory, Scarlet Beck was evaluating the proteomics performance of a novel high-resolution QTOF instrument, the Bruker *impact II*. The instrument had undergone several design iterations, which together improved the mass resolution to over 35,000 and ensured a highly efficient transfer of ions for high-sensitivity measurements. This included a re-designed collision cell, a new reflectron and careful optimization of the entire ion transfer path. These advances were complemented with adaptations of the MaxQuant software, which yielded average absolute mass deviations better than 1.5 ppm in proteomics experiments and enabled the MS/MS-based identification of 4,800 proteins in 2 h single runs of a HeLa digest. Protein quantification was very reproducible in technical replicates, and we were able to accurately reflect biological differences in model systems such as yeast and mouse cell lines. In conjunction with high pH reversed-phase fractionation, over 11,000 proteins were identified in mouse cerebellum, which led us to the conclusion that this QTOF platform is well equipped for demanding proteomics applications.

I contributed to the optimization of the MS acquisition method, analyzed data and performed experiments in the revision of the manuscript. This work lays the ground for the following studies in my thesis which would later on justify the preliminary remark "*it appears that QTOFs are set to make a comeback in proteomics*".

The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics*⑤

Scarlet Beck‡, Annette Michalski§, Oliver Raether§, Markus Lubeck§,
Stephanie Kaspar§, Niels Goedecke§, Carsten Baessmann§, Daniel Hornburg‡,
Florian Meier‡, Igor Paron‡, Nils A. Kulak‡, Juergen Cox¶, and Matthias Mann‡‖

Hybrid quadrupole time-of-flight (QTOF) mass spectrometry is one of the two major principles used in proteomics. Although based on simple fundamentals, it has over the last decades greatly evolved in terms of achievable resolution, mass accuracy, and dynamic range. The Bruker impact platform of QTOF instruments takes advantage of these developments and here we develop and evaluate the impact II for shotgun proteomics applications. Adaption of our heated liquid chromatography system achieved very narrow peptide elution peaks. The impact II is equipped with a new collision cell with both axial and radial ion ejection, more than doubling ion extraction at high tandem MS frequencies. The new reflectron and detector improve resolving power compared with the previous model up to 80%, *i.e.* to 40,000 at *m/z* 1222. We analyzed the ion current from the inlet capillary and found very high transmission (>80%) up to the collision cell. Simulation and measurement indicated 60% transfer into the flight tube. We adapted MaxQuant for QTOF data, improving absolute average mass deviations to better than 1.45 ppm. More than 4800 proteins can be identified in a single run of HeLa digest in a 90 min gradient. The workflow achieved high technical reproducibility (R2 > 0.99) and accurate fold change determination in spike-in experiments in complex mixtures. Using label-free quantification we rapidly quantified haploid against diploid yeast and characterized overall proteome differences in mouse cell lines originating from different tissues. Finally, after high pH reversed-phase fractionation we identified 9515 proteins in a triplicate measurement of HeLa peptide mixture and 11,257 proteins in single measurements of cerebellum—the highest proteome coverage reported with a QTOF instrument so far. *Molecular & Cellular Proteomics 14: 10.1074/mcp.M114.047407, 2014–2029, 2015.*

Building on the fundamental advance of the soft ionization techniques electrospray ionization and matrix-assisted laser desorption/ionization (1, 2), MS-based proteomics has advanced tremendously over the last two decades (3–6). Bottom-up, shotgun proteomics is usually performed in a liquid chromatography-tandem MS (LC-MS/MS)[1] format, where nanoscale liquid chromatography is coupled through electrospray ionization to an instrument capable of measuring a mass spectrum and fragmenting the recognized precursor peaks on the chromatographic time scale. Fundamental challenges of shotgun proteomics include the very large numbers of peptides that elute over relatively short periods and peptide abundances that vary by many orders of magnitude. Developments in mass spectrometers toward higher sensitivity, sequencing speed, and resolution were needed and helped to address these critical challenges (7, 8). Especially the introduction of the Orbitrap mass analyzers has advanced the state of the art of the field because of their very high resolution and mass accuracy (9, 10). A popular configuration couples a quadrupole mass filter for precursor selection to the Orbitrap analyzer in a compact benchtop format (11–13).

In addition to the improvements in MS instrumentation, there have been key advances in the entire proteomics workflow, from sample preparation through improved LC systems and in computational proteomics (14–16). Together, such

[1] The abbreviations used are: LC-MS/MS, liquid chromatography-tandem MS; CAA, chloroacetamide; ES, electrospray; FDR, false discovery rate; Hepa 1–6, mouse hepatoma; ID, inner diameter; LFQ, label-free quantification; MCP, multichannel plate; MEFs, mouse embryonic fibroblasts; MS/MS, tandem mass spectrometry; NSC-34, spinal cord neuron-neuroblastoma; QTOF, quadrupole time-of-flight; TOF, time-of-flight; UPS, universal protein standard.

advances are making shotgun proteomics increasingly comprehensive and deep analyses can now be performed in a reasonable time (13, 17–19). Nevertheless, complete analysis of all expressed proteins in a complex system remains extremely challenging and complete measurement of all the peptides produced in shotgun proteomics may not even be possible in principle (20, 21). Therefore, an urgent need for continued improvements in proteomics technology remains.

Besides the Orbitrap analyzer and other ion trap technologies, the main alternative MS technology is time-of-flight, a technology that has been used for many decades in diverse fields. The configuration employed in proteomics laboratories combines a quadrupole mass filter via a collision cell and orthogonal acceleration unit to a reflectron and a multichannel plate (MCP) detector (22). TOF scans are generated in much less than a millisecond (ms), and a number of these "pulses" are added to obtain an MS or MS/MS spectrum with the desired signal to noise ratio. Our own laboratory has used such a quadrupole time-of-flight (QTOF) instrument as the main workhorse in proteomics for many years, but then switched to high-resolution trapping instruments because of their superior resolution and mass accuracy. However, TOF technology has fundamental attractions, such as the extremely high scan speed and the absence of space charge, which limits the number of usable ions in all trapping instruments. In principle, the high spectra rate makes TOF instruments capable of making use of the majority of ions, thus promising optimal sensitivity, dynamic range and hence quantification. It also means that TOF can naturally be interfaced with ion mobility devices, which typically separate ions on the ms time scale. Data independent analysis strategies such as MS^E, in which all precursors are fragmented simultaneously (23, 24) or SWATH, in which the precursor ion window is rapidly cycled through the entire mass range (25), also make use of the high scanning speed offered by QTOF instruments. It appears that QTOFs are set to make a comeback in proteomics with recent examples showing impressive depth of coverage of complex proteomes. For instance, using a variant of the MS^E method, identification of 5468 proteins was reported in HeLa cells in single shots and small sample amounts (26). In another report, employing ion mobility for better transmission of fragment ions to the detector led to the identification of up to 7548 proteins in human ovary tissue (27).

In this paper, we describe the impact II™, a benchtop QTOF instrument from Bruker Daltonics, and its use in shotgun proteomics. This QTOF instrument is a member of an instrument family first introduced in 2008, which consists of the compact, the impact, and the maXis. The original impact was introduced in 2011 and was followed by the impact HD, which was equipped with a better digitizer, expanding the dynamic range of the detector. With the impact II, which became commercially available in 2014, we aimed to achieve a resolution and sequencing speed adequate for demanding shotgun proteomics experiments. To achieve this we developed an improved collision cell, orthogonal accelerator scheme, reflectron, and detector. Here we measure ion transmission characteristics of this instrument and the actually realized resolution and mass accuracy in typical proteomics experiments. Furthermore, we investigated the attainable proteome coverage in single shot analysis and we ask if QTOF performance is now sufficient for very deep characterization of complex cell line and tissue proteomes.

EXPERIMENTAL PROCEDURES

Preparation of HeLa Lysates—HeLa cells (ATCC, S3 subclone) were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal bovine serum, 20 mM glutamine and 1% penicillin-streptomycin (all from PAA Laboratories, Freiburg, Germany). Cells were collected by centrifugation at $200 \times g$ for 10 min, washed once with cold phosphate buffered saline (PBS) and centrifuged again. Supernatant was carefully discarded and the cell pellet shock frozen in liquid nitrogen and stored at −80 °C until further use. A pellet containing 5×10^7 cells was resuspended in 1.5 ml of ice cold Milli-Q water, then an equal volume of trifluoroethanol (Sigma-Aldrich, Taufkirchen, Germany) was added. The cell suspension was kept on ice for 10 min, vortexed for 1 min and sonicated for 2 min at 20% duty cycle and output control 3 (Branson Ultrasonics sonifier, Danbury, CT; model 250). After the addition of 200 μl Tris (pH 8.5, final concentration: 100 mM), 400 μl TCEP (final concentration: 10 mM) and 400 μl 2-chloroacetamide (CAA) (final concentration: 40 mM) the lysate was incubated for 10 min at 95 °C. Then the sample was diluted to 15 ml with 50 mM ammonium bicarbonate. The mixture was digested by adding LysC (Wako Chemicals GmbH, Neuss, Germany; ratio 1 μg LysC:100 μg sample protein) for 2 h at 37 °C, followed by adding trypsin (ratio 1 μg trypsin:75 μg sample protein, Promega GmbH, Mannheim, Germany) at 37 °C overnight. After a further digestion with trypsin (ratio 1:125) for 5 h at 37 °C, the digested peptides with an estimated concentration of 1 μg/μl were diluted 1:4 with water and acidified by adding formic acid (FA) (final concentration: 0.2%) and purified on Sep-Pak tC18 cartridges (Waters, Milford, MA) according to manufacturer's instructions. Peptide concentration was determined using a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE).

Preparation of Yeast Lysates—Saccharomyces cerevisiae strains BY4742 and BY4743 (EUROSCARF) were grown at 30 °C in yeast extract peptone dextrose (YPD) media (10 g/l BactoYeast extract, 20 g/l Bacto™ peptone (BD), 2% w/v glucose). Cells were grown to log phase (OD$_{600}$ of 0.6), harvested by centrifugation at $1600 \times g$ for 10 min at 4 °C, washed with cold Milli-Q water and then collected again by centrifugation at $10,000 \times g$ for 5 min at 4 °C. Cells were lysed in 1% sodium deoxycholate, 10 mM TCEP, 40 mM CAA in 100 mM Tris pH 8.5, boiled for 10 min at 95 °C and sonicated for 3 min at 30% duty cycle and output control 3 (Branson Ultrasonics sonifier; model 250). Protein concentrations were determined by tryptophan fluorescence emission assay. Cell lysates were diluted 1:2 with Milli-Q water and digested by adding LysC (Wako Chemicals GmbH, ratio 1 μg LysC:50 μg sample protein) for 4 h at 37 °C, followed by adding again LysC (ratio 1:50) overnight at 37 °C. An equal volume of ethyl acetate acidified with 1% TFA was added to the solution, samples were vortexed for 2 min and digested peptides were purified on SDB-RPS StageTips as described in Kulak *et al.* (19). Peptide concentrations were determined using a NanoDrop spectrophotometer.

Preparation of MEFs, Hepa, and NSC Cell Line Lysates—Spinal cord neuron-neuroblastoma (NSC-34) (CED-CLU140, Biozol, Eching, Germany), mouse embryonic fibroblasts (MEFs) (American Type Culture Collection, Manassas, VA), and mouse hepatoma (liver cancer, Hepa 1–6) (CRL-1830, American Type Culture Collection) cell lines

were cultured and proteins prepared as previously described (28). Briefly, the cells were lysed in lysis buffer (4% SDS, 10 mM Hepes, pH 8.0) during sonication for 15 min (level 5, Bioruptor; Diagenode, Seraing (Ougrée) - Belgium). Cell lysis was followed by reduction of disulfide bonds with 10 mM DTT for 30 min and subsequent alkylation with 55 mM IAA for 45 min. To remove the detergent, cold acetone (−20 °C) was added to 100 μg of proteins to a final concentration of 80% v/v, and proteins were precipitated for at least 2 h at −20 °C. The suspension was centrifuged for 15 min (4 °C, 16,000 × g) and the precipitate was washed with 80% acetone (−20 °C) prior to re-suspension in 50 μl of 6 M urea/2 M thiourea, 10 mM Hepes, pH 8.0. An initial digestion step (3 h) was carried out after the addition of 1 μg of LysC, followed by dilution with four volumes of 50 mM ammonium bicarbonate and the final digestion with 1 μg of trypsin overnight at room temperature. The resulting peptide mixtures were desalted on SDB-RPS StageTips (29) and subjected to single shot LC-MS/MS analysis.

Preparation of Cerebellum Lysates—Cerebellum from a single mouse (strain: C57Bl6) was homogenized in 4% SDS in 100 mM Tris pH 7.6 using a FastPrep 24 homogenizer (MP Biomedicals, Eschwege, Germany), incubated for 10 min at 95 °C and sonicated for 3 min at 30% duty cycle and output control 3 (Branson Ultrasonics sonifier; model 250). To remove the detergent, acetone (−20 °C) was added to a final concentration of 80% v/v and proteins were precipitated overnight at −20 °C. Supernatants were carefully discarded after centrifugation at 1600 × g for 20 min at 4 °C, and the pellets were washed with 80% acetone (−20 °C). The protein pellets were dissolved in 8 M Urea in 10 mM Hepes and protein concentrations were determined by the tryptophan fluorescence emission at 350 nm using an excitation wavelength of 295 nm. Proteins were reduced with 10 mM DTT for 30 min and alkylated with 55 mM iodoacetamide for 20 min. After addition of thiourea to a final concentration of 0.1 M, samples were digested by adding LysC (Wako Chemicals, ratio 1 μg LysC:100 μg sample protein) for 3 h at RT, diluted with four volumes of 50 mM ammonium bicarbonate, and further digested with trypsin (ratio 1 μg trypsin:100 μg sample protein, Promega) at RT overnight. After a further digestion with LysC and trypsin (ratio 1:100) for 8 h at RT, digested peptides were acidified by adding TFA (final concentration: 0.5%) and purified on Sep-Pak tC18 cartridges (Waters) according to manufacturer's instructions. Peptide concentrations were determined using a NanoDrop spectrophotometer.

Sample Preparation for Quantification—Universal Proteomics Standard (UPS-1, Sigma-Aldrich) and Proteomics Dynamic Range Standard (UPS-2, Sigma-Aldrich), both containing 48 human proteins, either at equimolar concentrations (UPS-1) or formulated into a dynamic range of concentrations, covering five orders of magnitude (UPS-2), were prepared according to ref (30). Predigested yeast sample (Promega) was re-suspended in 0.1% trifluoroacetic acid to a final concentration of 500 ng/μl. Digested UPS-2 sample was spiked in two different amounts of 250 fmol to 2.5 amol peptide amount for sample 1 and 500 fmol to 5 amol for sample 2 into 500 ng yeast background, thereby creating two samples with a theoretical ratio 1:1 for the yeast proteome and 1:2 for the UPS peptides. In another sample, digested UPS-1 sample (25 fmol for all components) was spiked into 500 ng yeast.

High-pH Reverse-Phase Fractionation—We performed high-pH reversed-phase peptide prefractionation with fraction concatenation on 175 μg HeLa or cerebellum peptides on a 2.1 × 300 mm Acquity UPLC Peptide BEH column packed with 130 Å pore, 1.7 μm particle size C_{18} beads (Part No. 186005792, Waters). A gradient of basic reversed-phase buffers (Buffer A: 0.1% formic acid, ammonium hydroxide pH 10; Buffer B: 0.1% formic acid, 80% acetonitrile, ammonium hydroxide pH 10) was run on a Prominence HPLC system (Shimadzu, Duisburg, Germany) at a flow rate of 150 μl/min at 60 °C.

The LC run lasted for 240 min with a starting concentration of 5% buffer B increasing to 30% over the initial 120 min and a further increase in concentration to 60% over 70 min. This elution gradient was followed by a 95% wash and re-equilibration. Fraction collection started after 0.2 ml elution and fractions were collected every 140 s resulting in 72 fractions used for concatenation into 24 fractions as described previously (31).

Inlet Capillary and CaptiveSpray—In our instrument, in contrast to many other commercial ion source designs, the high voltage for the electrospray (ES) process is applied to the vacuum capillary inlet, whereas the sprayer is kept at ground, which allows for a simpler source design (supplemental Fig. S1A). To electrically decouple the ES voltage and the electrical potential of the vacuum section, we use an inlet capillary made from high resistive glass (~1GOhm). Positioning the ES voltage at the capillary entrance means that the ions are transported opposite to the electrical gradient by the gas flow (32). In this configuration, charged molecules travel somewhat slower than the surrounding gas. According to Bernoulli's law, ions are then focused toward the area of highest gas velocity along the center axis of the capillary. The set-up tends to reduce the contamination of the inner capillary walls (33).

The Bruker CaptiveSpray nanoflow ES source is directly attached to the vacuum inlet capillary via a short capillary extension that can be heated using the instrument's drying gas (Supplemental Fig. S1A). The spray tip is automatically mechanically aligned on axis with the capillary inlet without the need for any adjustments. The principle of the CaptiveSpray is a vortex gas (usually air) that sweeps around the emitter spray tip at three different stages. The first one is designed to assist spray formation, the second and third one help to focus the spray plume into the MS inlet capillary. All three flows are created solely by the vacuum of the MS system, which requires that the entire source is vacuum sealed.

The spray emitter consists of a 2 cm long, 20 μm ID fused silica capillary. Its tip is etch-tapered, thus the inner diameter remains constant to the very end of the tip making it very robust against clogging. Furthermore, it also allows using the same emitter at flow rates ranging from 50 nl/min to 5 μl/min, thereby supporting a wide range of column types. Fused silica columns, which are often used for proteomics, are typically connected to the emitter via a low dead volume union (supplemental Fig. S1A), which also provides the electrical contact for keeping the electrospray at ground potential.

Minimizing Postcolumn Dead Volume—Using the described design, the CaptiveSpray source provides very stable ionization; however, when we initially coupled it to the LC set-up used in the Munich laboratory (17, 34), we observed broader LC peak elution distributions than we normally do (supplemental Fig. S1B). Furthermore, we wished to incorporate a column oven and pulled tip columns. We therefore constructed a modified source, which keeps the back end of the CaptiveSpray but replaces the front end by the standard set-up used in our department. The modified set-up incorporating the tip column is displayed in supplemental Fig. S1C. The modified design of the column holder allows for exact aligning and fixation of the column inside the CaptiveSpray source. Electrical grounding was applied using a connecting tee at the column head. This setup produced the desired, narrow LC peak distributions (supplemental Fig. S1D) and was used for the proteomic analyses described in this article.

LC-MS/MS Analysis—We used an Easy nLC-1000 (Thermo Fisher Scientific) on-line coupled to an impact II (Bruker Daltonics) with a CaptiveSpray ion source (Bruker Daltonics). The peptide mixtures (1 μg) were loaded onto an in-house packed column (50 cm, 75 μm inner diameter) filled with C_{18} material (ReproSil-Pur C_{18} AQ 1.9 μm reversed phase resin, Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). Chromatographic separation was carried out using a linear gradient of 5–30% buffer B (80% ACN and 0.1% FA) at a flow rate of

56

250 nl/min over 90 min. Because of loading and washing steps, the total time for an LC-MS/MS run was about 40 to 50 min longer.

Generally, LC-MS/MS data were acquired using a data-dependent auto-MS/MS method selecting the 17 most abundant precursor ions in cycle for fragmentation and an MS/MS summation time adjusted to the precursor intensity (Compass 1.8 acquisition and processing software, Bruker Daltonics). For the deep proteome measurements of a cell line in combination with peptide fractionation, we used a "dynamic method," with a fixed cycle time of 3 s. The mass range of the MS scan was set to extend from m/z 150 to 1750. Dynamic exclusion duration was 0.4 min. Isolation of precursor ions was performed using an m/z dependent isolation window of 1.5–5 Th. The collision energy was adjusted between 23–65 eV as a function of the m/z value.

For the quantitative analysis of the UPS standards in yeast we used a trapping column set-up (PepMap pre-column, 2 cm x 100 μm; Thermo Scientific) and a Dionex HPLC pump (Ultimate 3000, Thermo Scientific). For this experiment, peptides were separated on a PepMap UHPLC column (50 cm x 75 μm, 2 μm particles; Thermo Scientific) using a 90 min multistep ACN gradient (buffer A: 0.1% FA; buffer B 100% ACN in 0.1% FA). The unmodified CaptiveSpray ion source (see above) was used to interface the LC system to the impact II. For quantification full scan MS spectra were acquired at a spectra rate of 1Hz followed by acquisition of 1 MS/MS spectrum. Six replicates per sample were acquired. For data acquisition of the UPS-1 in yeast sample, the 17 most intense precursor ions were selected for fragmentation, resulting in a total cycle time of 1.2 s.

Intact Protein Analysis—Adalimumab was cleaved at the hinge region with IdeS (FabRICATOR, Genovis) and reduced to obtain the Fc/2, Fd and light chain sub units as recently described in (35). The subunits were separated by chromatography (35) prior to analysis on the impact II. Data was analyzed using the SNAP algorithm to fit the theoretical pattern (36, 37).

Development of MaxQuant for QTOF Data—In general all processing steps from the standard MaxQuant computational workflow, which was optimized for the analysis of Orbitrap data, are also applied to QTOF data. The nonlinear mass recalibration algorithm experienced major adaptations. Its original form for the Orbitrap applies a recalibration function with nonlinear dependence on the two variables m/z and retention time. It has been extended to include the peak intensity as a third dimension that the mass recalibration depends on. This is necessary because of appreciable systematic nonlinear intensity dependent peak mass shifts that are typically found in time of flight data. The intensity dependence is parameterized as a polynomial in the logarithm of peak intensities. The new mass recalibration algorithm allows for high mass accuracy without the use of internal or external calibrants.

We added a new instrument type called "Bruker QTOF" in which several relevant parameters of algorithms for the processing of spectra are set to default values that are suitable for the analysis of data generated by the impact family. These parameters include mass matching windows for the assembly of 3D peaks, mass tolerances for assembling isotope patterns and labeling pairs, initial peptide mass tolerance windows for the Andromeda search and minimum required number of scans per 3D peak. Raw data can be immediately read from the proprietary Bruker binary format and no conversion to intermediate file formats is needed. Peak centroids are utilized as determined by the centroiding algorithms of in the Bruker software. The viewer module of MaxQuant is enabled for QTOF data, among other features allowing to visualize MS data in m/z-retention time maps and to annotate and export MS/MS spectra to fulfil journal requirements for reporting of spectral evidence.

Analysis of Proteomic Data—All data were analyzed with the MaxQuant software (version 1.5.2.8 or version 1.5.0.1) (38, 39) with the Andromeda search engine (38) with the adaptions and developments described above. The false discovery rate (FDR) was set to 1% for both proteins and peptides and we specified a minimum length of seven amino acids. MaxQuant scored peptides for identification based on a search with an initial allowed mass deviation of the precursor ion of up to 0.07 Da after time-dependent recalibration of the precursor masses. The allowed fragment mass deviation was 40 ppm. The Andromeda search engine was used for the MS/MS spectra search against the Uniprot human database (downloaded on June 21, 2014, containing 88,976 entries and 247 contaminants), the Uniprot Saccharomyces cerevisiae database (downloaded on June 21, 2014, containing 6643 entries), the Uniprot mouse database (downloaded on June 21, 2014, containing 51,573 entries) and UPS fasta file provided by Sigma-Aldrich (http://www.sigmaaldrich.com/life-science/proteomics/mass-spectrometry/ups1-and-ups2-proteomic.html) for quantitative study. Enzyme specificity was set as C-terminal to Arg and Lys, also allowing cleavage at proline bonds and a maximum of two missed cleavages. Carbamidomethylation of cysteine was selected as fixed modification and N-terminal protein acetylation and methionine oxidation as variable modifications.

The "match between runs" feature of MaxQuant was used to transfer identifications to other LC-MS/MS runs based on their masses and retention time (maximum deviation 0.7 min) and this was also used in quantification experiments. Quantifications were performed with the label-free algorithms described recently (39). We required a minimum peptide ratio count of two and at least one "razor peptide" for quantification. For cerebellum, the quantification was based on normalized protein intensities. Further analysis of data was performed in the MaxQuant Viewer, in the Perseus post data acquisition package that is part of MaxQuant (all freely available at www.maxquant.org) and in the R statistical computing environment (40).

Potential contaminants as well as proteins identified only by site modification were strictly excluded from further analysis.

For the quantitative analysis of the UPS standards in yeast, entries were only accepted if they had valid values in all 12 replicates. Results were then filtered for Welch-significant regulation of UPS-2 proteins.

Analysis of the yeast samples were based on label-free intensities (LFQ values). After filtering (3 valid values in at least one group), remaining missing values were imputed from a normal distribution (width: 0.3; down shift: 1.8). Two-sample t test was performed with a FDR < 0.01.

For global cell line comparison, triplicates were analyzed twice for a total of six single shot measurements per cell line (except the NSC-34 cell line, which was only measured once). For the principal component analysis (PCA) of the different cell lines, we furthermore limited the data set (LFQ intensities) to entries with a minimum of four valid values in at least one group of six replicates. Remaining missing values were imputed from a normal distribution (see above).

Protein intensity (summed peptide intensity) for cerebellum samples were divided by the molecular weight for ranking by proteins abundance. Annotations (GO molecular function, biological process, cellular component; KEGG and Uniprot Keywords) were matched to protein groups with Perseus. We performed a 1D annotation enrichment (41) on the normalized protein intensities. To evaluate the proteome coverage, we counted the occurrence of categories in our sample and compared it to the category count for the complete murine proteome in Perseus.

MS raw data and data for protein and peptide identification and quantification were submitted as supplementary tables to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD001592.
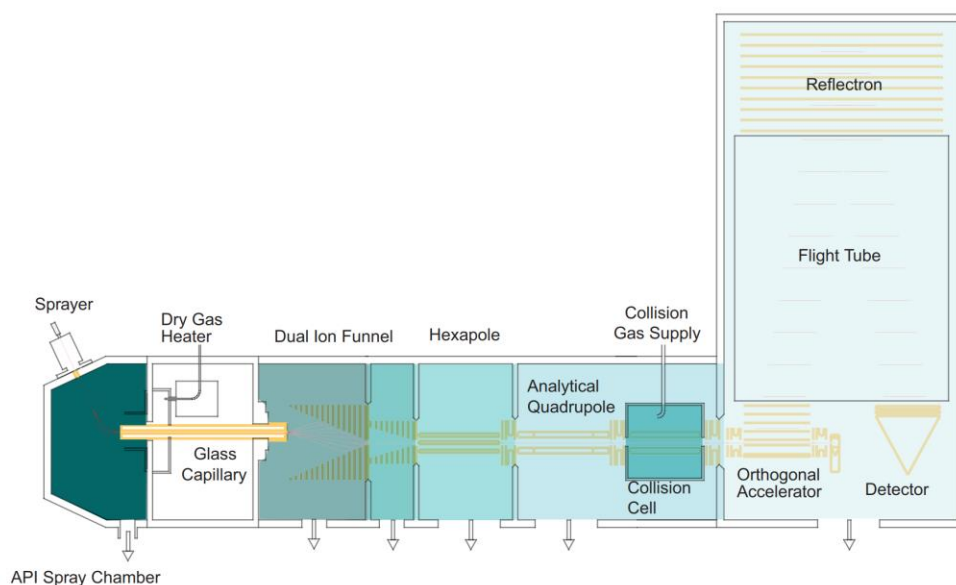
FIG. 1. **Schematic of the impact II mass spectrometer (not to scale).**

RESULTS AND DISCUSSION

Overview of the Instrument—The Bruker impact II is a QTOF in a benchtop format, featuring several improvements in its design (Fig. 1). Briefly, ions are produced in the CaptiveSpray, which is in an encased nanoelectrospray source that features a well-defined gas stream to guide the ions into the vacuum via a capillary inlet. A double ion funnel, based on principles described by Smith and co-workers (42), is positioned off axis, which prevents neutrals from further transmission along the ion path. The pressure drops by several orders of magnitude from the capillary exit to the postfunnel stage (3 mbar to 3×10^{-4} mbar), while the ion current is virtually undiminished (see below). Additionally, the funnel allows for soft transfer based on low electrical field strength independent of the mass (typically 10 V/cm, much lower than in nozzle-skimmer designs). By introducing electrical acceleration in-between the two funnels, in-source fragmentation can still be achieved intentionally. There is a hexapole ion guide between funnel and the analytical quadrupole mass filter, which has a monolithic design based on high precision glass. Precursor ions can be isolated by this quadrupole for subsequent fragmentation in the collision cell. Intact ions or fragments can be stored and extracted from the collision cell and enter the orthogonal deflection region as a very narrowly focused ion beam ($< 500 \ \mu$m). Here they are accelerated into a field-free drift region. A newly designed, two-stage reflectron further compensates the velocity distribution orthogonal to the beam direction. Finally, the ions impinge on an MCP detector coupled to a 10-bit, very high frequency (50 Gbit/s), zero noise digitizer. Data collection is coordinated by the Bruker Compass data

system and in the experiments described here, post-acquisition data processing is performed in the MaxQuant environment.

Optimization of the Collision Cell—Efficient fragmentation of precursor ions on an LC-MS/MS time scale is a key for the identification of peptides in shotgun proteomics strategies. We optimized several aspects of the collision cell (supplemental Fig. S2A): Precise geometrical alignment allows focusing of the ions along the axis of the collision cell, directly translates into well-defined starting conditions for the orthogonal accelerator and is therefore mandatory for high mass resolution. This is implemented via a quadrupolar configuration of the collision cell device providing a narrow pseudo potential well (43). We also introduced a radial ejection step between any two MS or MS/MS experiments, in order to reduce the dead time. The ion path has to be emptied to avoid crosstalk between two consecutive spectra without introducing substantial ion losses. Most importantly, we optimized the time of ion fragmentation and extraction within the fragment spectra to ensure efficient high frequency MS/MS by implementing an electrical axial field gradient. This gradient directs the ions toward the exit of the collision cell, reducing the time it takes for the first ions to reach the extraction lens after quenching the collision cell. From here, they can be rapidly released toward the orthogonal accelerator, forming packages that match the orthogonal pulser frequency. The ion densities after 3 ms (upper red traces in supplemental Fig. S2B–S2C) reveals two important aspects: Using the axial field gradient results in comparable ion density at the collision cell exit in much shorter time than without, *i.e.* after about 1 ms instead of
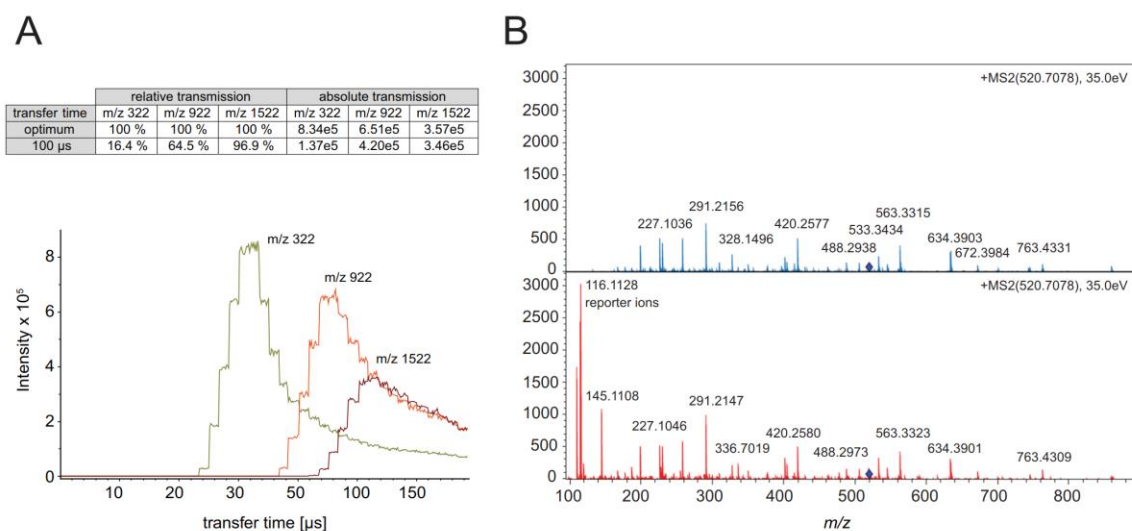
A

transfer time	relative transmission			absolute transmission		
	m/z 322	m/z 922	m/z 1522	m/z 322	m/z 922	m/z 1522
optimum	100 %	100 %	100 %	8.34e5	6.51e5	3.57e5
100 μs	16.4 %	64.5 %	96.9 %	1.37e5	4.20e5	3.46e5

B

FIG. 2. *A*, Absolute ion intensities of *m/z* 322, 922 and 1522 as function of the transfer time. The maxima of the distributions for each *m/z* value are the optimal times for efficient transfer into the flight tube. *B*, Fragment spectrum of the iTRAQ labeled peptide LFTGHPETLEK without (blue) and with (red) transfer time stepping that adds the lower *m/z* range including the reporter ions without sacrificing intensity in the higher mass range.

3 ms, respectively. This suggests a three times faster ion transfer. Moreover, the overall number of ions within the collision cell is reduced to less than 50% at the moment of quenching, which should reduce ion losses related to quenching accordingly.

The pseudo potential calculations and simulation were confirmed in different experiments comparing the performance of the optimized collision cell on the fragmentation yield of Glu-Fibrino-Peptide B at different MS/MS acquisition rates (supplemental Fig. S2D). This revealed that the reduced quench losses in fact improve the number of ions detected at a spectra rate of 16 Hz by a factor of two. We further observed that the axial field gradient improves the stability of the system even in the presence of slight contaminations on the rods.

High Transfer Efficiency to the Orthogonal Acceleration Unit—The ions travel through the flight tube and require as much time as the largest *m/z* species needs to reach the detector, before the HV pulser can send the next ion package toward the detector (typically between 100 and 150 μs). To avoid excessive loss of ions, orthogonal TOF instruments are therefore often operated in a mode in which the ions are stored in the collision cell during the TOF scan and released in time for the next extraction pulse of the orthogonal accelerator. This would allow for 100% duty cycle if all ions were indeed transferred such that they arrive in the orthogonal accelerator at the same time and with the same kinetic energy. In practice, however, the extraction time from the collision cell toward the gate lens is a function of ion mobility. We have analyzed these combined effects by varying the time from opening the gate lens to the extraction pulse of the

orthogonal accelerator (transfer time) for different *m/z* ratios (Fig. 2A). Simulations of ion trajectories and extraction times reveal that about 80% of a single ion species can be accelerated into the drift tube of the TOF analyzer under optimal transfer time conditions. In the impact II the high transfer efficiency is further optimized by reducing the distance between the trapping region and the orthogonal accelerator—it is about four times higher compared with conventional QTOF systems operated in continuous operation mode. Figure 2A summarizes the relative transmission efficiency of selected precursors (*m/z* = 322 Th, 922 Th and 1522 Th) at a transfer time of 100 μs, which is a standard setting to cover the mass range relevant in shotgun proteomics. The analysis highlights that ions with low *m/z* are compromised most, which we have previously counteracted in proteomics experiments by adding special extraction conditions (44). To tackle this problem in a more general way, Bruker introduced the "transfer time stepping" operation mode, where first high mobility species are extracted followed by species with a lower mobility. During the initially short opening times of the gate (typically 50% of the total scan) only the higher mobility, low *m/z* ions pass the gate lens, while the lower mobility ions are still accumulated in the collision cell. In the second transfer time steps, extraction times are increased to allow the low mobility, higher *m/z* ions to pass the gate lens. Spectra with and without transfer time stepping, reveal the beneficial effect on low mass ions without appreciable loss in the standard mass range (Fig. 2B). This is particularly beneficial for labeling experiments that rely on quantification of low mass reporter ions as shown in the figure. These ions are transferred with greater than 60% effi-
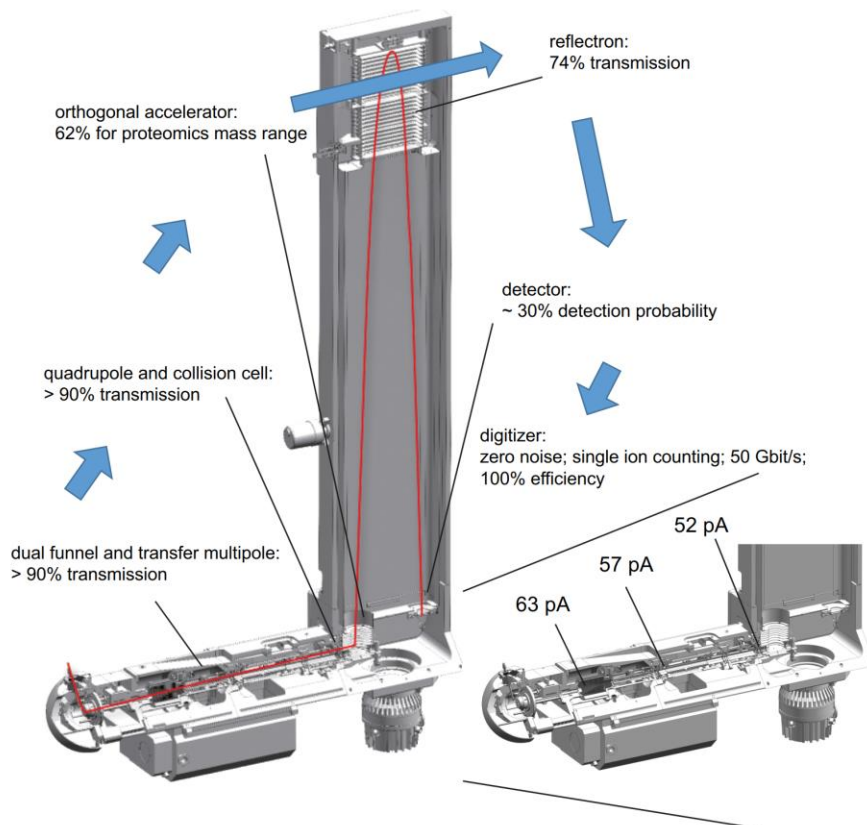
FIG. 3. **Ion transfer efficiency of mass range most relevant in shotgun-proteomics experiments (*m/z* 100–1500).** Inset shows net analyte ion currents at the indicated measurement points and transfer efficiencies at the orthogonal accelerator at various stages along the flight path of the impact II instrument.

ciency during the low *m/z* extraction phase and the total amount of ions in the analyzed fragment spectrum is increased by 58%.

Together, our developments led to an overall transmission efficiency of > 60% into the orthogonal acceleration unit. This compares favorably to a recent report, in which ion mobility was performed on fragment ions and their arrival times were synchronized with orthogonal extraction, which lead to an up to 10-fold improvement of detection sensitivities to standard operation (27).

Sensitivity and Ion Transfer—The number of ions that successfully pass through the instrument and are finally recorded determine a mass spectrometer's sensitivity. We were interested in the transfer efficiencies along the ion path, from entering the vacuum system to the detector. To experimentally determine this, we infused a 1 pmol/μl BSA solution or blank solution and measured the difference in the ion current between these conditions. When operating the outlet of the capillary and the funnel region as a Faraday cage, we measured a net ion current of 63 pA, which we defined as the

starting value (100%) (Fig. 3). The large ion acceptance aperture of the first funnel efficiently captures in the ion flux leaving the capillary. It transfers the ions to the second funnel, which also passes the ions in an almost lossless manner through the next stage as evidenced by a net current reading of 57 pA (>90%) after the hexapole. Likewise, more than 90% of the ions are transferred through the quadrupole (in nonmass selecting operation) and the collision cell. More than 60% of these ions are transmitted into the flight tube by the orthogonal accelerator (see above). The reflectron contains two grids, which are each passed twice, leading to a geometrically defined overall transmission of 74%. Although all these ions hit the MCP, not all of them enter the channels and not all of them result in secondary electrons (detector quantum yield). However, when secondary ions are generated they are greatly amplified (>10^6 fold), and can be efficiently discriminated from electronic noise (zero noise detection). Together, this leads to an estimated detection probability of 30% for our MCP detector. Although not all of these measurements and estimates are very precise, they suggest an overall detection

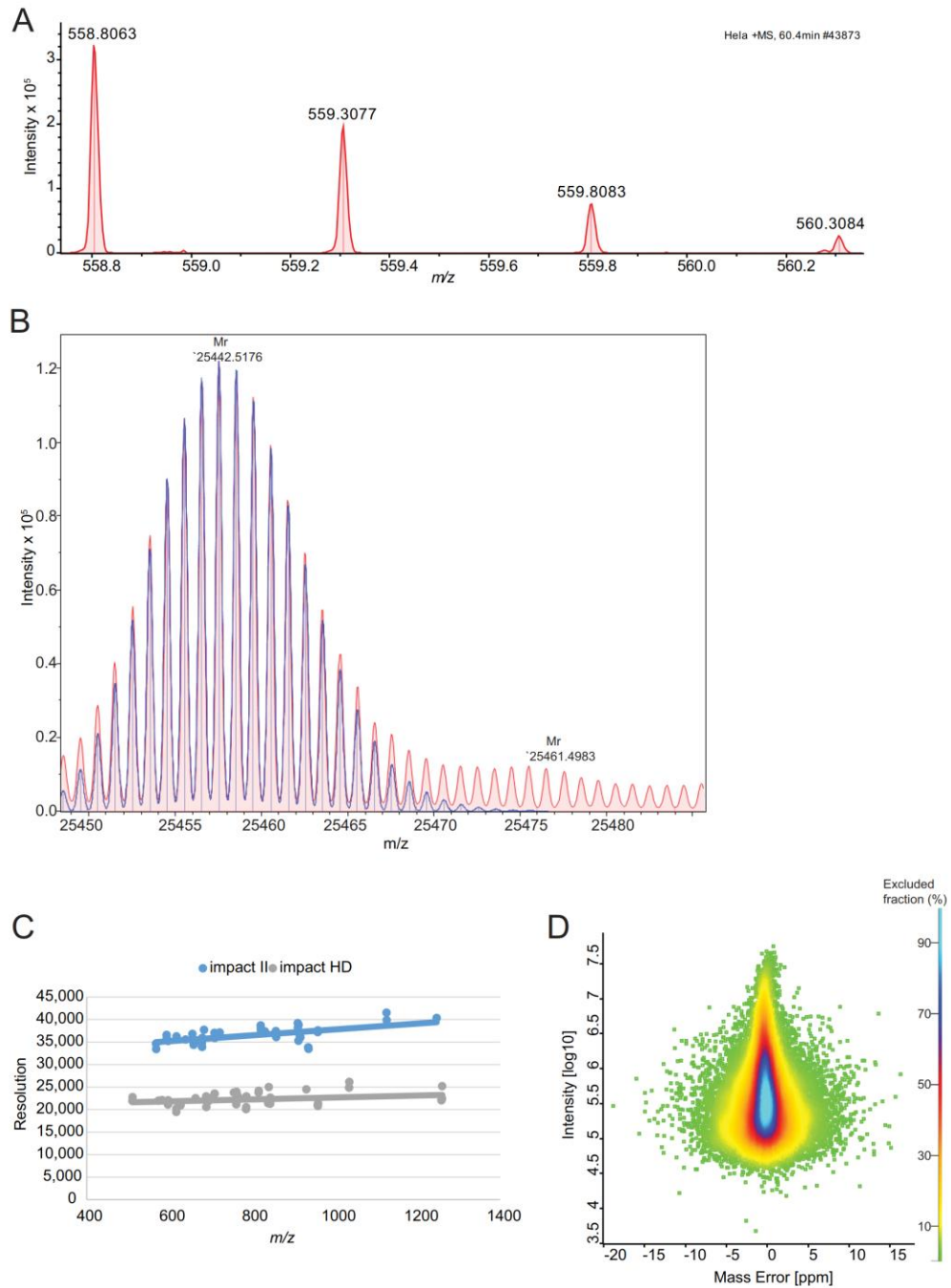FIG. 4. **Resolution and mass accuracy of *A*, a peptide isotope cluster (*m/z* 558.8063, *r* = 33k) and *B*, Fd unit of Adalimumab (*m/z* 25442.5157, *r* = 63k, 0.26 ppm).** Overall improvement of the resolution with the improved detector, *C*, and the achieved mass accuracy dependent on the summed peptide intensity, *D*, in a shotgun proteomics experiment using the QTOF optimized version of MaxQuant.

TABLE I

Identification from HeLa and yeast lysate triplicate analysis using a standard 90 min gradient

	MS scans	Isotope pattern	MS/MS scans	Identification rate [%]	Peptide sequences identified	Proteins identified
HeLa_1	7002	759,774	79,704	47.09	35,547	4870
HeLa_2	7181	769,355	79,876	47.22	35,572	4864
HeLa_3	7272	796,086	81,389	46.57	35,621	4828
Total				46.96	48,172	5210
yeast_1	4873	528,682	63,682	31.08	17,066	3361
yeast_2	4732	541,194	66,441	30.88	16,921	3325
yeast_3	4691	556,675	66,978	31.98	17,494	3369
Total				31.32	24,131	3627

probability of ions transmitted into the vacuum system of about 10%. This excellent number is because of the fact that the continuous beam generated by the ES source can be utilized and that all the ion guiding elements have been successfully optimized for high transmission.

Resolution and Mass Accuracy—For the impact II several improvements were implemented: symmetrical shielding for better ion focusing; line grids to increase the transmission; low temperature coefficient ceramic spacers to decrease temperature related mass drift and improved axel bearings for precise alignment. Together this led to about 35% increased resolution over the full proteomics mass range. As this improvement is reached solely by better focusing of the ions, mass accuracy and signal to noise are expected to increase accordingly. Improvements to the MCP detector include an increased entrance aperture, higher electron accelerating fields and optimized shielding. Overall, these measures lead to 2-fold faster ion impact transients and 30% higher detection efficiency of the MCP.

In summary, the resolving power of the TOF analyzer is expected to increase by about 70 to 80% by the introduction of the new collision cell, reflectron and detector. To test this experimentally on a standard proteomic sample, we analyzed data from a HeLa digest. Resolution for typical peptides is in excess of 33,000 as illustrated by an example in Fig. 4A. Further increase in the resolution can be obtained by the "Focus mode," which involves real time processing and alignment of successive pulses and increases accuracy of flight time determination, when multiple ions of one species reach the detector at the same time (45). This can be helpful to resolve the isotope distributions of proteins, as shown in Fig. 4B, which features a resolution in excess of 60,000 for an antibody subunit (> 25 kDa). In TOF measurements, resolution tends to increase with m/z (Fig. 4C), and reached more than 40,000 for the TuneMix component at m/z 1222. This constitutes a 70 to 80% improvement over the previous impact model, the impact HD.

This increase in resolving power should also imply better mass accuracy in proteomics samples, which we tested using the software package MaxQuant, which we adapted to QTOF data as described under "Experimental Proce-

dures." A special feature of MaxQuant is the extraction of individual mass accuracy values (46, 47), which allows to make efficient use of high mass accuracy in the identification of peptides. MaxQuant was originally developed on the basis of data from hybrid Orbitrap instruments. Here we developed MaxQuant further in order to analyze QTOF data and also in this context profit from the high mass accuracy provided by nonlinear mass recalibration algorithms. A special challenge in QTOF data is the drift in the mass scale because of thermal expansion caused by ambient temperature drift. MaxQuant employs a double search strategy, in effect supplying hundreds of reference masses internal to each proteomic sample. This feature efficiently removes any effect of the temperature related mass drift. This "software lock mass" feature makes it unnecessary to use dedicated molecular species for the calibration of spectra (48). With these developments in MaxQuant, we analyzed the peptide mass error distributions over a 90 min gradient run (Fig. 4D). This showed an average absolute mass deviation of around 1.45 ppm, which is excellent for a QTOF instrument.

Impact II Performance for Single Shot Analysis—To investigate the performance of the impact II for shotgun proteomics, we first analyzed a complex peptides mixture derived from a mammalian cell line in the single-run format (Experimental Procedures). We separated 1 μg of peptide digest by on-line HPLC with the standard 90 min gradient employed in our laboratory and performed triplicate analysis. The typical data dependent acquisition scheme in bottom-up proteomics consists of an MS scan followed by N MS/MS fragment scans of the most intense precursors (topN method). It is desirable to choose N such that the total cycle is less than a few seconds. For our measurements we aimed at a duty cycle of around 1.3 s and designed a top17 method, consisting of 200 ms for MS acquisition and a MS/MS integration time adapted to the precursor intensity. We found this to be a good balance between acquiring high S/N in the MS and achieving optimal ion intensities in the MS/MS spectra for the HeLa digest. This method reached more than 7000 MS and 79,700 MS/MS scans. In each run MaxQuant identified on average 35,580 unique peptide sequences, which results in total of 48,172 unique peptide sequences in the triplicate analysis
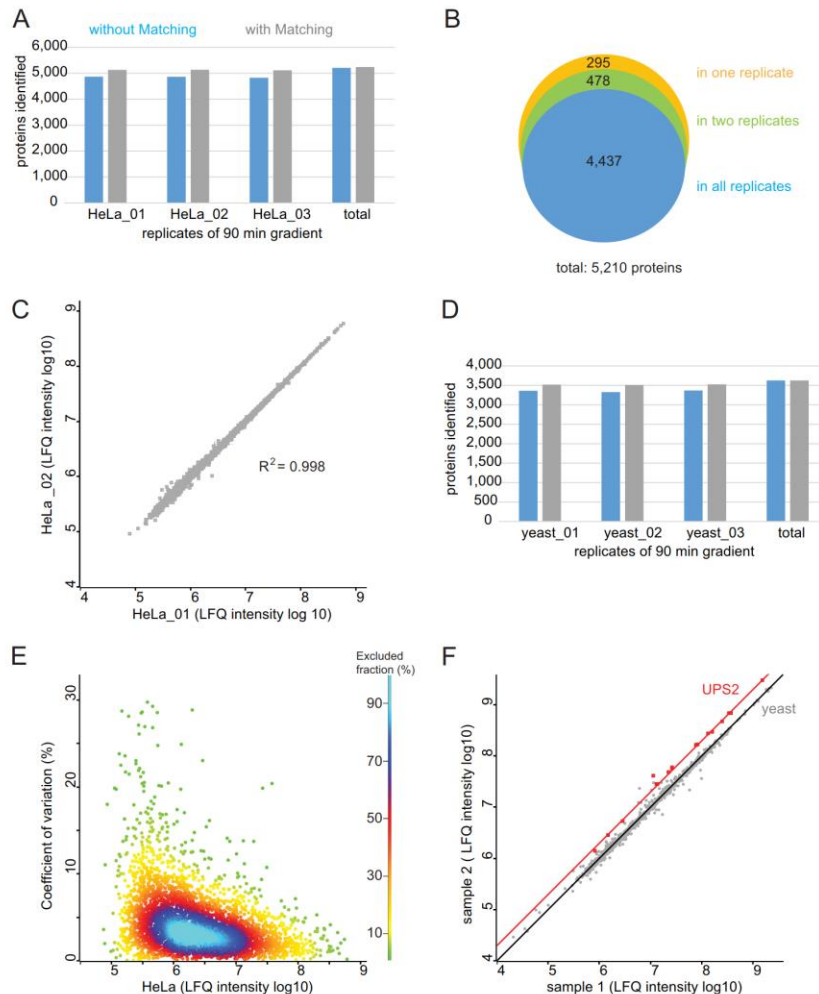
FIG. 5. **Triplicate analysis of HeLa and yeast digest using a 90 min gradient for single shot analysis.** *A*, Protein identification numbers of each replicate of 1 μg HeLa digest and *B*, the overlap of protein identities. *C*, Correlation between the rank ordered label-free quantification values of each identified protein in replicate1 (HeLa_01) and replicate2 (HeLa_02) (log10 LFQ intensities). *D*, Protein identification number of 1 μg yeast digest. *E*, Reproducibility across triplicate analysis of HeLa digest. CVs of all three replicates, representing 99% of the data points. *F*, Accuracy in quantification. Results of spike-in experiment showing the UPS-2 standard (orange) and in a yeast proteome background (gray). Yeast lysate was present in equal amounts in sample 1 and sample 2 and the UPS2 protein standard was present in twice the amount in sample 2.

(Table I). These peptides mapped to an average of 4854 proteins per run, and a total of 5210 proteins of the HeLa proteome with the three 90 min gradients (Fig. 5*A*), indicating that a deep coverage can be achieved using relatively short, single shot analysis. Transferring identification between the runs based on their mass precision and retention time ("match between runs" feature in MaxQuant) led to around 5100 proteins identifications per single run. Comparing protein identities between the triplicate analyses (without "matching between runs"), we observed that more than 90% of proteins were identified in each of them (Fig. 5*B*). This indicates high

reproducibility and a minimal 'missing value' problem. This conclusion is further supported by excellent reproducibility ($R^2 = 0.998$) in the label-free intensities determined in pairwise comparison between runs (MaxQuant LFQ values (39, 49)) (Fig. 5*C*). We also determined the number of proteins identified in a single shot run of 1 μg yeast digest. On average, we identified 3352 proteins per single, 90 min gradient, and a total of 3627 proteins when combining the three single shot measurements (Fig. 5*D*).

Reproducibility and Accuracy of Quantification—To evaluate the reproducibility of the method for label-free quantifica-
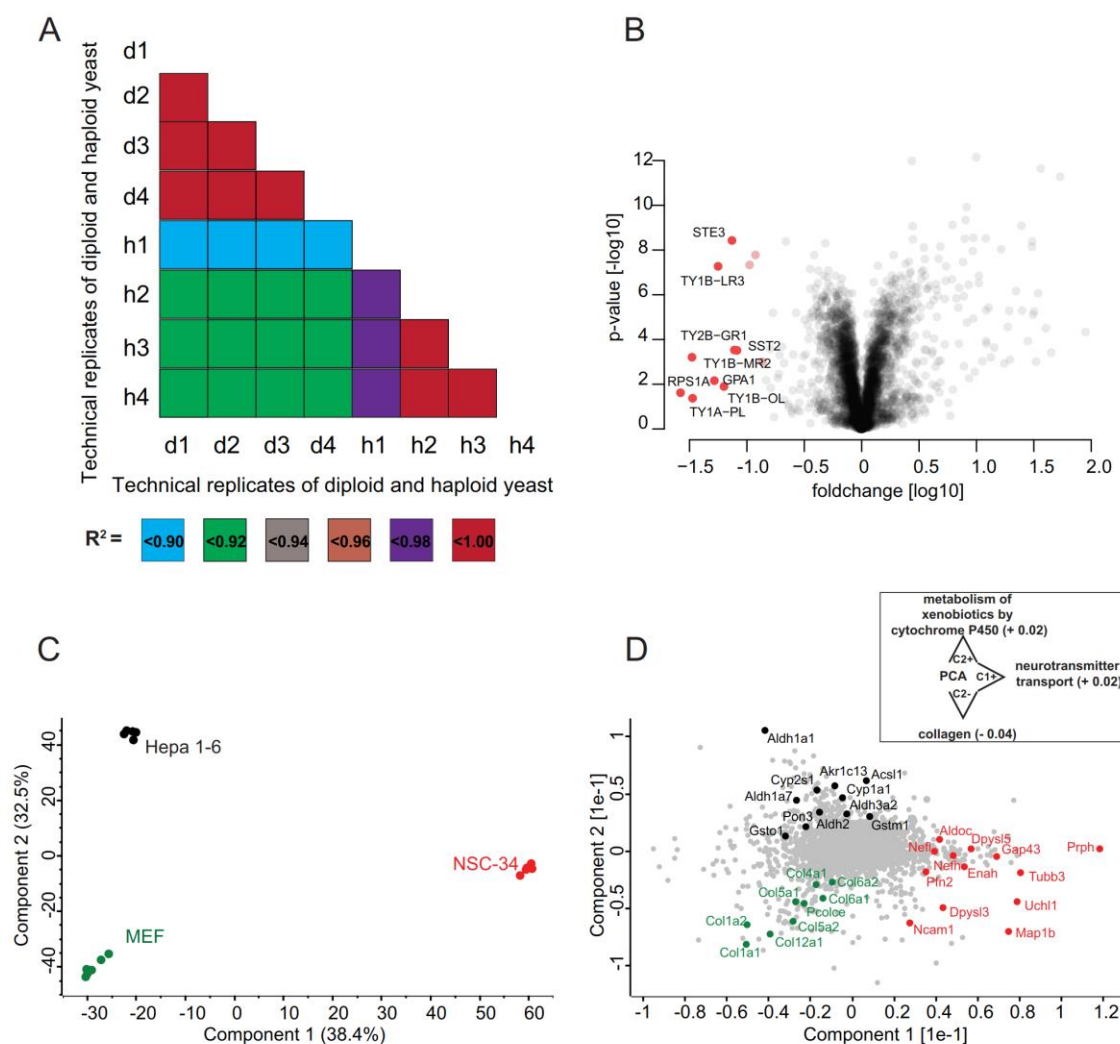
FIG. 6. *A*, Correlation of single shot LC-MS/MS measurements of haploid (h) and diploid (d) yeast samples. All technical replicate correlation values were > 0.98. *B*, Quantitative differences between the haploid and diploid yeast proteome. Proteins marked in red are significantly more abundant in haploid cells. *C*, Principal component analysis (PCA) of LFQ protein expression levels in mouse hepatoma (Hepa 1–6), mouse embryonic fibroblasts (MEF) and motoneuronal cell lines (NSC-34). *D*, Loading of the first two principal components as obtained from the PCA plotted in *C*. The inset indicates significantly enriched annotations along the PC axes (FDR < 0.05). The respective enrichment scores are specified in parentheses.

tion, we determined the coefficients of variation (CV) of the label-free intensities, determined in pair-wise comparison between three technical HeLa replicates (see above). For more than 90% of the quantified proteins the CV was smaller than 10% (Fig. 5*E*). For the lowest intensity quantile the median was 0.05 and for the highest one 0.03 (Fig. 5*E*, supplemental Fig. S3). This shows reproducible quantification over four orders of magnitude.

Accurate quantification of differently expressed proteins remains challenging over a wide concentration range and

benefits from a very stable analytical platform. QTOF instrumentation has been used widely for label-free quantification, which in principle allows convenient analysis and comparison of an arbitrary number of samples. For evaluation of the label-free quantitative capabilities of the impact II platform in complex mixtures we wished to use a reference sample set with known ratios for a small subset of proteins. We spiked the Universal Protein Standard 2 (UPS-2), consisting of 48 proteins covering a dynamic range of five orders of magnitude, in two different concentrations into the yeast proteome.

This generated two samples, in which the yeast peptides should be 1:1 whereas ratios for UPS-2 peptides should be at 1:2. To increase the number of identified peptides we then used the equimolar UPS-1, which we also spiked in a yeast background. This allowed transfer of peptide identifications to unsequenced peptides in the UPS-2 in yeast runs via the 'match between runs' algorithm in MaxQuant. We identified all of the 48 human UPS proteins in the sample containing yeast with the equimolar UPS-1 standard. Of these proteins, 23 were identified in each one of the 12 single shot yeast measurements with the dynamic UPS-2 standard either directly via MS/MS or via match between runs and 18 of these proteins showed Welch-significance (Fig. 5F). We identified and quantified UPS-2 proteins over more than three orders of magnitude in these relatively fast measurements (90 min gradients). The UPS-2 proteins have an average fold change of 0.49 ($\pm$ 0.06), which is close to the theoretical ratio of 0.5.

Quantification of Changes in the Yeast Proteome—To test the workflow in a systems biology context, we analyzed proteome changes of diploid and haploid (Matα cell) *S. cerevisiae*. We analyzed 2 μg of yeast digest from haploid and diploid cells in technical quadruplicates with our standard 90 min gradient. This identified 3769 proteins using "match between runs." For statistical analysis, we only considered LFQ intensities that were detected in at least three replicates of the haploid or diploid groups. After filtering, 3222 proteins remained for further analysis (Experimental Procedures). Remaining missing values were imputed from a normal distribution. The technical replicates correlated much more with each other than they correlated to the other genotype (R^2 greater 0.98 *versus* R^2 about 0.92; Fig. 6A). As in our previous large-scale analysis on SILAC labeled haploid and diploid yeast (50), we found transposons more abundant in haploid cells than in diploid cells (Fig. 6B). Ste3, the pheromone a factor receptor, was specific to haploid yeast, as expected from its mating status. Also absent in diploid but present in haploid cells were Sst2, a GTPase-activating protein for Gpa1 (51), which, consistently, showed higher expression in haploid cells. Conversely, Sps100, which is a sporulation-specific wall maturation protein turned out to be specific for diploid cells. Doing such systems-wide comparisons by traditional methods would have required thousands of individual Western blots. Even compared with our previous large-scale study performed by quantitative MS (50), we here used less than 1% of yeast input material and measurement time. This illustrates the rapidity by which MS-based proteomics is becoming a viable method for answering biological questions.

Global Proteomic Comparison of Different Cell Lines—As a second example of typical proteomics experiments, we applied the QTOF-based workflow to the characterization of common cellular disease model systems. For this purpose, we compared the proteomes of spinal cord neuron-neuroblastoma (NSC-34), mouse hepatoma (Hepa 1–6) and mouse embryonic fibroblast (MEF) cell lines in a quantitative manner.

All cell lysates were analyzed in single shots using 90 min gradients and subsequently quantified using the MaxQuant label-free quantification algorithm. The observed LFQ intensities were highly reproducible between biological and technical replicates as indicated by Pearson correlations coefficients > 0.97 (supplemental Fig. S4).

After stringent filtering (Experimental Procedures) we performed a principal component analysis (PCA) to evaluate the similarities and dissimilarities of the cell lines on a global scale. Replicates from a single cell line clustered very tightly in the PCA space and the first two principal components accounted for 71% of the cumulative variance within our data set (Fig. 6C). Interestingly, the motoneuronal cell line NSC-34 was clearly separated from Hepa 1–6 and MEF cells in the first principle component, whereas the variance between both non-neuronal cell lines was described by the second, orthogonal, principal component. In the latter component, NSC-34 is positioned half-way between Hepa 1–6 and MEF.

To assess individual proteins that are the main drivers for the separation between the three cell lines, we plotted the loadings of the first two principal components (Fig. 6D). Hepa 1–6 cells were characterized by gene products involved in regulatory and metabolic processes. As we had found before (52), proteins related to the glycolysis pathway, such as the aldehyde dehydrogenase (Aldh) family, were highly represented in Hepa 1–6. Not surprisingly, proteins driving the separation of MEF were predominantly linked to collagen synthesis, such as the COL gene family and the precollagen C-endopeptidase enhancer 1 (Pcolce). Proteins that differentiate along principal component 1 encompassed various components of the cytoskeleton axons, including Prph and the heavy and light chains of neurofilaments (Nefh and Nefl). In addition, gene products involved in axon guidance (Enah, Dypsl5 and Tubb3) and neuron projection (Uchl1, Gap43) were highly distinctive for NSC-34. Proteins separating NSC-34 from the other cell lines were significantly enriched for neurotransmitter transport while we observed enrichment for collagens and xenobiotic processes for the MEF and Hepa 1–6 cells on component 2, respectively, further showing how proteomics can highlight biological function. This is even more remarkable, given that we previously found that the proteomes of motoneuronal cell lines, including NSC-34, lack distinctive neuronal characteristics, as several key actors in axon growth and guidance were either depleted or low abundant (28). As a result, we had placed motoneuronal cell lines only halfway between *in vivo* motoneurons and non-neuronal controls. Nevertheless, our label-free QTOF-based workflow is very well suited to differentiate subtle alterations in biological systems in a short time of analysis.

Impact II Performance for Deep Proteome Analysis of a Cell Line—To evaluate the impact II for deep proteome coverage we performed high pH reverse-phase pre-fractionation with fraction concatenation as described in ref (31). We loaded 175 μg of a HeLa peptide mixture, collected 72 fractions and
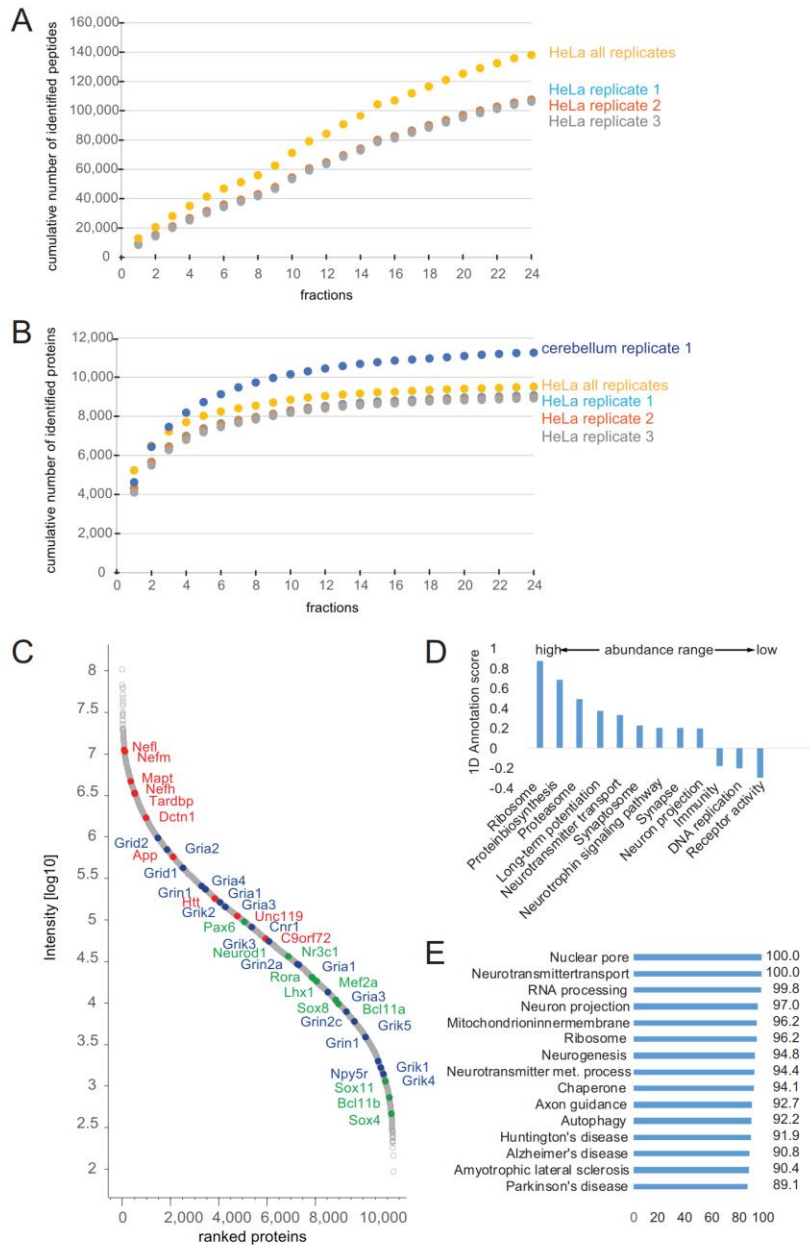
Fig. 7. **Triplicate cell line and singlet tissue analysis in 24 high-pH fractions.** *A*, Cumulative number of identified peptides of triplicate analysis of HeLa. *B*, Cumulative number of identified proteins of HeLa replicates and of the singlet analysis of mouse cerebellum (48 h total measurement time). *C*, Rank ordered intensities of each identified protein (log10 intensities) of cerebellum. Proteins involved in neurodegenerative disorders are marked in red; cerebellum specific expressed proteins are marked in blue and a selection of receptors in green. *D*, Protein abundance distribution and *E*, proteome coverage (expressed as a percentage) of different neuronal, metabolic and disease relevant processes in cerebellum.

combined them into 24 fractions (Experimental Procedures). These fractions were analyzed in technical triplicates using a standard 90 min LC-MS/MS gradient (total measurement time of 48 h per replicate). Instead of the simple Top17 method described above, we here used a so-called "dynamic method" with a fixed cycle time and a MS/MS integration time adapted to the precursor intensity. We found that this method helped to generate high quality MS/MS spectra also for low abundant peptides.

Figure 7A depicts the cumulative number of unique peptides identified as a function of the number of fractions analyzed. The increase is nearly linear, indicating a small overlap of peptide identification between fractions and the good orthogonal separation power as also observed by others using high-pH fractionation (49). The figure also indicates that the reproducibility was very high between the technical replicates. On average we identified 107,038 unique peptides, mapping to around 8995 proteins (protein FDR 1%; Fig. 7B). In total we identified 138,086 unique peptides, resulting in 9515 different protein groups. This is to our knowledge the deepest proteome coverage of a human cell line measured with a QTOF instrument.

Impact II Performance for Ultra-Deep Proteome Analysis of Tissue—Tissues are more challenging to analyze by proteomics than cell lines, because they are made up of different cell types, the extracellular matrix and other structural and connective elements. Among the different tissues, the brain is thought to be the most complex one. To evaluate the Impact II in this context, we homogenized the cerebellum of a single mouse, digested it using our standard workflow and separated part of the resulting tryptic peptides using high pH fractionation (Experimental Procedures). In total, we identified 11,257 proteins from a single analysis of 24 fractions (2 days of measurement time) (Fig. 7B, supplemental Table S1). To our knowledge, this is the deepest proteome measurement of any tissue reported by TOF instrumentation so far. Protein abundance, as indicated by the summed and normalized peptide signal varied by more than five orders of magnitudes (Fig. 7C). We identified many high to medium abundant proteins involved in neurodegenerative disorders (Fig. 7C marked in red). The transcription factors and DNA binders Pax6, Lhx1/5, Otx1/2, and Neurod1 (Fig. 7C, marked in blue) are examples of proteins that have been reported to be specifically expressed in cerebellum (53). Like various neuronal receptors, transcription factors populate the medium to low abundance range (Fig. 7C, marked in green). The distribution of molecular functions throughout the abundance range is similar to that observed in previous studies (Fig. 7D, supplemental Table S2) (54, 55). The very high depth of this cerebellum proteome is shown by almost complete coverage for neuronal, general metabolic and disease relevant processes (Fig. 7E, supplemental Table S2). Rapid estimation of protein abundances in the brain regions of a single mouse could be useful for studying tissue characteristics and disease specific alterations. For instance, knowledge about changes in the complete proteome would provide an additional layer of information on the pathological processes in neurodegeneration.

CONCLUSION

Here we have described the construction and performance of a state of the art QTOF instrument, the impact II. We documented significant improvements in the ion path, collision cell performance and, in particular, in the performance of the reflectron and detector. The latter allow a mass resolution and mass accuracy that is compatible with the high demands of shotgun proteomics of complex mixtures. For the first time, we measured and modeled the ion transmission from capillary to the detector, which revealed an excellent efficiency of about 10%. The new features of the impact II allow deep characterization in single shots, where we identified more than 5200 proteins in HeLa cells and 3600 proteins in yeast. Using off-line high pH reversed-phase fractionation we identified more than 9500 proteins in HeLa cells and 11,250 proteins in a single cerebellum tissue analysis. These are extremely high numbers for any platform and additional method developments should further improve these results. We also document excellent quantitative reproducibility and accuracy in a label-free format. In concordance with others (26, 27), we conclude that the improvements in QTOF technologies in recent years now clearly enable demanding, in-depth analysis of very complex proteomes.

REFERENCES

1. Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301
2. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71
3. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
4. Cravatt, B. F., Simon, G. M., and Yates, J. R., 3rd (2007) The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000
5. Altelaar, A. F., Munoz, J., and Heck, A. J. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48
6. Richards, A. L., Merrill, A. E., and Coon, J. J. (2014) Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24C,** 11–17
7. Domon, B., and Aebersold, R. (2006) Challenges and opportunities in

proteomics data analysis. *Mol. Cell. Proteomics* **5,** 1921–1926

8. Mann, M., and Kelleher, N. L. (2008) Precision proteomics: the case for high-resolution and high mass-accuracy. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 18132–18138

9. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162

10. Zubarev, R. A., and Makarov, A. (2013) Orbitrap mass spectrometry. *Anal. Chem.* **85,** 5288–5296

11. 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

12. Scheltema, R. A., Hauschild, J. P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014) 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

13. Kelstrup, C. D., Jersie-Christensen, R. R., Batth, T. S., Arrey, T. N., Kuehn, A., Kellmann, M., and Olsen, J. V. (2014) Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J. Proteome Res.* **13(12),** 6187–6195

14. Zhou, H., Ning, Z., Wang, F., Seebun, D., and Figeys, D. (2011) Proteomic reactors and their applications in biology. *FEBS J.* **278,** 3796–3806

15. Kocher, T., Swart, R., and Mechtler, K. (2011) Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Anal. Chem.* **83,** 2699–2704

16. Cox, J., and Mann, M. (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80,** 273–299

17. Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111 013722

18. 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

19. 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

20. 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

21. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006) 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

22. Morris, H. R., Paxton, T., Dell, A., Langhorne, J., Berg, M., Bordoli, R. S., Hoyes, J., and Bateman, R. H. (1996) High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun. Mass Spectr.* **10,** 889–896

23. Silva, J. C., Denny, R., Dorschel, C., Gorenstein, M. V., Li, G. Z., Richardson, K., Wall, D., and Geromanos, S. J. (2006) Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. *Mol. Cell. Proteomics* **5,** 589–607

24. Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., and Geromanos, S. J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144–156

25. Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012) 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

26. 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

27. Helm, D., Vissers, J. P., 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

28. Hornburg, D., Drepper, C., Butter, F., Meissner, F., Sendtner, M., and Mann, M. (2014) Deep proteomic evaluation of primary and cell line motoneuron disease models delineates major differences in neuronal characteristics. *Mol. Cell. Proteomics* **13,** 3410–3420

29. Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, prefractionation, and storage of peptides for proteomics using StageTips. *Nature Protoc.* **2,** 1896–1906

30. Wang, H., Qian, W. J., Mottaz, H. M., Clauss, T. R., Anderson, D. J., Moore, R. J., Camp, D. G., 2nd, Khan, A. H., Sforza, D. M., Pallavicini, M., Smith, D. J., and Smith, R. D. (2005) Development and evaluation of a micro- and nanoscale proteomic sample preparation method. *J. Proteome Res.* **4,** 2397–2403

31. Dwivedi, R. C., Spicer, V., Harder, M., Antonovici, M., Ens, W., Standing, K. G., Wilkins, J. A., and Krokhin, O. V. (2008) Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal. Chem.* **80,** 7036–7042

32. Whitehouse, C. M., Dreyer, R. N., Yamashita, M., and Fenn, J. B. (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57,** 675–679

33. Franzen, J. (1998) Method and device for transport of ions in gas through a capillary. Google Patents

34. Ishihama, Y., Rappsilber, J., Andersen, J. S., and Mann, M. (2002) Micro-columns with self-assembled particle frits for proteomics. *J. Chromatogr. A* **979,** 233–239

35. Ayoub, D., Jabs, W., Resemann, A., Evers, W., Evans, C., Main, L., Baessmann, C., Wagner-Rousset, E., Suckau, D., and Beck, A. (2013) Correct primary structure assessment and extensive glyco-profiling of cetuximab by a combination of intact, middle-up, middle-down, and bottom-up ESI and MALDI mass spectrometry techniques. *mAbs* **5,** 699–710

36. Senko, M. W., Beu, S. C., and McLaffertycor, F. W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectr.* **6,** 229–233

37. Tsybin, Y. O., Fornelli, L., Stoermer, C., Luebeck, M., Parra, J., Nallet, S., Wurm, F. M., and Hartmer, R. (2011) Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. *Anal. Chem.* **83,** 8919–8927

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

40. R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

41. Cox, J., and Mann, M. (2012) 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13,** S12

42. Shaffer, S. A., Prior, D. C., Anderson, G. A., Udseth, H. R., and Smith, R. D. (1998) An ion funnel interface for improved ion focusing and sensitivity using electrospray ionization mass spectrometry. *Anal. Chem.* **70,** 4111–4119

43. Gerlich, D. (1992) *Inhomogeneous RF fields a versatile tool for the study of processes with slow ions*., John Wiley & Sons, Inc, Hoboken, NJ, U.S.A.

44. Steen, H., Kuster, B., Fernandez, M., Pandey, A., and Mann, M. (2001) Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Anal. Chem.* **73,** 1440–1448

45. Räther, O. (2005) High resolution detection for time-of-flight mass spectrometers. US6870156 B2 patent

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

47. Cox, J., Hubner, N. C., and Mann, M. (2008) How much peptide sequence information is contained in ion trap tandem mass spectra? *J. Am. Soc.*

68

Mass Spectr. **19,** 1813–1820

48. Cox, J., Michalski, A., and Mann, M. (2011) Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectr.* **22,** 1373–1380

49. Wang, Y., Yang, F., Gritsenko, M. A., Wang, Y., Clauss, T., Liu, T., Shen, Y., Monroe, M. E., Lopez-Ferrer, D., Reno, T., Moore, R. J., Klemke, R. L., Camp, D. G., 2nd, and Smith, R. D. (2011) Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11,** 2019–2026

50. de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–1254

51. Apanovitch, D. M., Slep, K. C., Sigler, P. B., and Dohlman, H. G. (1998) Sst2 is a GTPase-activating protein for Gpa1: purification and characterization of a cognate RGS-Galpha protein pair in yeast. *Biochemistry* **37,** 4815–4822

52. Pan, C., Kumar, C., Bohl, S., Klingmueller, U., and Mann, M. (2009) Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Mol. Cell. Proteomics* **8,** 443–450

53. Suzuki, H., Okunishi, R., Hashizume, W., Katayama, S., Ninomiya, N., Osato, N., Sato, K., Nakamura, M., Iida, J., Kanamori, M., and Hayashizaki, Y. (2004) Identification of region-specific transcription factor genes in the adult mouse brain by medium-scale real-time RT-PCR. *FEBS Letters* **573,** 214–218

54. 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

55. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342

69

Article 3: Parallel Accumulation−Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device

Florian Meier[†,#], Scarlet Beck[†,#], Niklas Grassl[†], Markus Lubeck[‡], Melvin A. Park[§], Oliver Raether[‡], and Matthias Mann[†]

[†] Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[‡] Bruker Daltonik GmbH, Fahrenheitstrasse 4, 28359 Bremen, Germany
[§] Bruker Daltonics Inc., 40 Manning Road, Billerica, Massachusetts 01821, United States

[#] These authors contributed equally to the work.

Parallelization has greatly contributed to advances in diverse field, for example computer science. Conversely, data-dependent acquisition methods in mass spectrometry still sequence one precursor after another. This has two important implications for MS-based proteomics: First, the overall depth of the experiment is directly linked to the sequencing speed. Second, faster sequencing rates imply a lower acquisition time and therefore a lower ion count in the MS/MS spectrum. To resolve this dilemma, we introduced the concept of 'parallel accumulation – serial fragmentation' (PASEF).

The PASEF scan mode employs the Trapped Ion Mobility Spectrometry (TIMS)[281] device developed by Mel Park and colleagues at Bruker. In the course of this study, our collaboration partners at Bruker incorporated the TIMS device in the first vacuum stage of a laboratory prototype, which was based on the impact II platform. In PASEF, all precursor ions are accumulated in parallel in the TIMS device (restricted only by the charge capacity) and sequentially released from the TIMS in narrow ion mobility peaks. As the precursor ions elute, the quadrupole isolation window switches rapidly to fragment multiple peptide precursors from a single TIMS scan. In contrast to DIA, PASEF still operates with narrow isolation windows and therefore generates readily interpretable and not convoluted MS/MS spectra.

At the time, the prototype instrument required manual programming of each precursor, was limited by inefficient switching of the electronics, and was therefore not compatible with online LC separation. To still prove the PASEF concept, we directly electrosprayed

70

mixtures of four digested proteins and used a TIMS analysis time of 50 ms, which yielded average ion mobility peak half widths of about 1 ms. We demonstrated that the quadrupole is capable of switching its position within less than one millisecond, giving rise to multiple MS/MS spectra per TIMS scan. A comparison of the total ion currents with and without PASEF indicated that the full ion signal is preserved with PASEF, which means that PASEF, despite being several-fold faster, does not compromise sensitivity.

We further outlined the pre-requisite hardware improvements for a successful application of PASEF in proteomics and possible modes of operation. Extrapolating our results, we predicted an up to 10-fold increase in sequencing speed and sensitivity as compared with the sequential acquisition of precursors.

Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device

Florian Meier,[†] Scarlet Beck,[†] Niklas Grassl,[†] Markus Lubeck,[‡] Melvin A. Park,[§] Oliver Raether,[‡] and Matthias Mann*,[†]

[†]Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[‡]Bruker Daltonik GmbH, Fahrenheitstrasse 4, 28359 Bremen, Germany
[§]Bruker Daltonics Inc., 40 Manning Road, Billerica, Massachusetts 01821, United States

Ⓢ Supporting Information

ABSTRACT: In liquid chromatography-mass spectrometry (LC-MS)-based proteomics, many precursors elute from the column simultaneously. In data-dependent analyses, these precursors are fragmented one at a time, whereas the others are discarded entirely. Here we employ trapped ion mobility spectrometry (TIMS) on an orthogonal quadrupole time-of-flight (QTOF) mass spectrometer to remove this limitation. In TIMS, all precursor ions are accumulated in parallel and released sequentially as a function of their ion mobility. Instead of selecting a single precursor mass with the quadrupole mass filter, we here implement synchronized scans in which the quadrupole is mass positioned with sub-millisecond switching times at the m/z values of appropriate precursors, such as those derived from a topN precursor list. We demonstrate serial selection and fragmentation of multiple precursors in single 50 ms TIMS scans. Parallel accumulation–serial fragmentation (PASEF) enables hundreds of MS/MS events per second at full sensitivity. Modeling the effect of such synchronized scans for shotgun proteomics, we estimate that about a 10-fold gain in sequencing speed should be achievable by PASEF without a decrease in sensitivity.

KEYWORDS: proteomics, MS/MS, ion mobility, TIMS, peptide sequencing, multiplexing, time-of-flight, high resolution
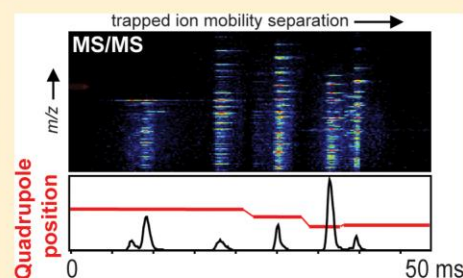
■ INTRODUCTION

High-resolution mass spectrometry (MS)-based proteomics has emerged as a powerful technique for large-scale profiling of thousands of proteins with many applications in molecular and cellular biology.[1−3] A typical bottom-up shotgun proteomics workflow starts with the extraction and solubilization of the protein material prior to enzymatic digestion. The peptide mixture is subsequently separated via liquid chromatography (LC) and electrosprayed into a mass spectrometer. To derive sequence information, suitable precursor ions are isolated by their mass-to-charge ratio (m/z) and subjected to collision-induced fragmentation, followed by database identification. Precursor scan and fragmentation are commonly performed with a data-dependent topN method, in which a MS survey spectrum is followed by fragmentation spectra of the N most abundant precursors. In complex mixtures, the depth of the analysis is thus foremost limited by the sequencing speed and sensitivity of the mass spectrometer. Previously we demonstrated that only 16% out of 100,000 peptide features eluting during a 90 min LC gradient were targeted for MS/MS.[4] State-of-the-art proteomics MS instruments partly address these challenges by increasing sequencing speed and resolving

power.[5,6] However, as more and more MS/MS spectra are acquired per second, less and less acquisition time is available for each precursor—an inherent consequence of the serial nature of the MS/MS process. Parallel fragmentation of unselected precursors in methods such as MS^E[7] or SWATH[8] addresses this problem at the expense of multiplexing the MS/MS spectra, making peptide identification more challenging and precluding iTRAQ[9] and TMT[10] based multiplexing.[10,11]

Time-of-flight (TOF) instruments acquire spectra at very high frequency. This makes them capable of using the majority of the precursor and fragment ions, thus promising optimal sensitivity and sequencing rates. However, although MS/MS rates of up to 100 Hz[12] can be readily achieved, in proteomic practice, sensitivity is generally not sufficient to detect an adequate number of fragment ions for peptide identification at such high scan speeds. That is, because the quadrupole filter transmits a given type of precursor ion for only a small fraction of the time in which that species is eluting from the LC column, only a small fraction of available precursor ions are utilized

when operating at high MS/MS spectral rates. Nevertheless, quadrupole TOF instruments have improved their performance over the last few years, and appear ready to become a viable alternative to the prevalent Orbitrap analyzer-based technology.[13−16] We recently described a state of the art QTOF instrument, with high resolution and ion transmission, which enabled rapid and in-depth analysis of complex proteomes, including the identification of more than 11,000 different proteins in brain tissue.[17]

Due to their high scanning speeds (about 0.1 ms/spectrum), TOF instruments are compatible with ion mobility spectrometry (IMS), which happens on the time scale of tens of milliseconds.[18−21] While IMS-MS can increase speed, selectivity, and sensitivity,[22−24] available platforms have entailed a considerable increase in instrumental complexity and reduced ion transmission.[25] A particular form of ion mobility, termed trapped ion mobility spectrometry (TIMS), features a particularly compact construction, without compromising resolution or transmission.[26−29] In the TIMS device, ions are accumulated in an RF-only tunnel at a position where the force of a gas flow equals the opposing force of an electric field. Ion mobility separated species are released from the device as a function of their collisional cross section. Here we asked if this sequential release after parallel accumulation can be exploited to drastically increase the speed and sensitivity of MS/MS experiments.

■ MATERIALS AND METHODS

Sample Preparation

Purified and predigested standards of enolase, phosphorylase b, alcohol dehydrogenase (ADH), and bovine serum albumin (BSA) were purchased from Waters GmbH (Eschborn, Germany) and resuspended in 0.1% formic acid to prepare stock solutions at a concentration of 10 pmol/μL each. These stock solutions were combined in an equimolar ratio and diluted in 50% water/50% acetonitrile/0.1% formic acid (v/v/v) to a final concentration of 100 fmol/μL.

Trapped ion mobility spectrometry−mass spectrometry

To mimic conditions during the analysis of complex proteomics samples, we directly infused the digested four protein mixture into a prototype, high-resolution QTOF mass spectrometer equipped with a TIMS device (flow rate 3 μL/min). For an overview of the instrument, see the Results and Discussion below. A detailed description of the construction and operation of the TIMS analyzer employed here has been published elsewhere.[26,30] Briefly, the TIMS device is composed of stacked ring electrodes, which form three distinct sections: an entrance funnel, the TIMS tunnel, and an exit funnel. The experiments were performed using nitrogen as a bath gas at room temperature, and the gas flow velocity was kept constant by regulating the pressure at the inlet and the outlet of the TIMS cartridge. Ions were accumulated for 50 ms, and mobility separation was achieved by ramping the entrance potential from −180 V to −40 V within 50 ms (435 TOF scans of 115 μs each). TIMS and MS operation were controlled and synchronized using the instrument control software otofControl (Bruker Daltonik, Bremen, Germany).

Synchronized TIMS and quadrupole operation

To enable the PASEF method, precursor m/z and mobility information was first derived from full scan TIMS-MS experiments (with a mass range of m/z 150−1850). Resulting

quadrupole mass, collision energy, and switching times were manually transferred to the instrument controller as a function of the total cycle time via direct firmware commands. The quadrupole isolation width was set to 3 Th and, for fragmentation, the collision energies were varied between 30 and 60 eV depending on precursor mass and charge.

Data analysis and modeling of increased sequencing rates

Ion mobility resolved mass spectra, nested ion mobility vs m/z distributions, as well as summed fragment ion intensities were extracted from the raw data files with a prototype version of DataAnalysis (Bruker Daltonik). S/N ratios were increased by summations of individual TIMS scans. Mobility peak positions and peak half-widths were determined based on extracted ion mobilograms (±0.05 Da) with an in-house script written in Python, using the peak detection algorithm implemented in the DataAnalysis software.

To model the anticipated benefit from PASEF scans for shotgun proteomics, we reanalyzed a single shot HeLa experiment from our recently published data set acquired on an impact II QTOF instrument.[17] Feature detection and peptide identification were performed with MaxQuant version 1.5.2.8 applying the previously described search parameters. Further analysis of the MaxQuant output was performed in the R statistical computational environment.[31] For each isotope pattern ("MS feature"), MaxQuant reports the minimum and maximum scan number where it was detected.[32] To simulate a topN method, each detected feature was assigned to a single scan (at the minimum scan number) and the N most abundant features of each bin were selected according to the original MS peak picking criteria (m/z > 300 and charge >1). This procedure reproduced the actual peak picking procedure performed by the Bruker acquisition software.

To restrict the simulations to likely peptide precursors, the MS features present in the MaxQuant output, as well as the target species assigned as described above, were further filtered for retention times between 10−100 min, charge state 2−5 and m/z values >450, which yielded a log-normal intensity distribution.

■ RESULTS AND DISCUSSION

Trapped ion mobility spectrometry (TIMS) − Mass spectrometry

To investigate the PASEF method that is the subject of this study, we made use of a prototype TIMS-QTOF mass spectrometer (Figure 1). In this instrument, a TIMS tunnel was incorporated into the first pumping stage, upstream from the transfer and selection quadrupoles and the high resolution mass analyzer.

Briefly, ions are generated in an electrospray source, transferred into the vacuum system through a glass capillary, deflected by 90°, and focused by the entrance funnel into the TIMS tunnel. This tunnel consists of pairs of stacked electrodes to which an RF field is applied. A DC field in the longitudinal direction is superimposed on this RF field (Figure 1B). Gas originating from the capillary flows through the sealed tunnel at a pressure of about 2−3 mbar. Ions entering the tunnel experience a drag due to the gas flow and a counteracting electrical field. They come to rest at a position where these two forces are equal and are radially confined by the pseudopotential induced by the RF field. Since the drag is proportional to the collisional cross section, ions of different mobility are trapped at different positions along the longitudinal axis, with
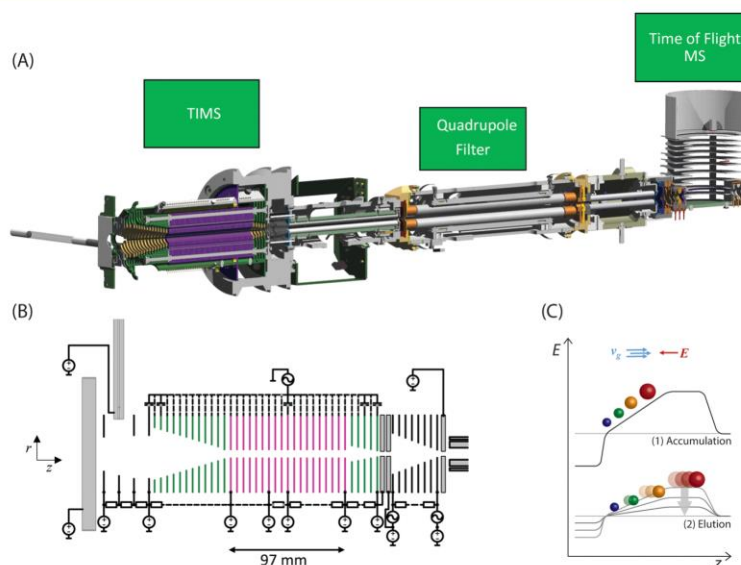
Figure 1. Trapped ion mobility spectrometry coupled to a QTOF mass spectrometer. (A) Instrument schematic of the prototype TIMS-QTOF instrument used in this study. (B) Detailed schematic of the TIMS tunnel (purple), enclosed by the entrance and exit funnels (green). (C) General mode of TIMS operation, including ion accumulation (1) and serial elution (2) of ion mobility separated ions from the TIMS device by decreasing the electrical field. The directed forces of the gas flow and electrical field are indicated by v_g and E.

high mobility ions at the entrance and low mobility ions near the exit of the tunnel (Figure 1C). After a desired ion accumulation time, further ions are prevented from entering the tunnel by a change of the potential on the deflection plate. Stored ions are then "eluted" by decreasing the electric field strength. Eluted ions are focused by the exit funnel and pass through a second funnel and transfer quadrupole. Precursor ions can be isolated by the analytical quadrupole mass filter for optional subsequent fragmentation in the collision cell. Afterward narrow ion packages are accelerated into a field-free drift region by the orthogonal deflection unit. A two-stage reflectron compensates differences in the kinetic energy of the ions and they are detected on an MCP detector coupled to a 10-bit digitizer, similar to the Bruker impact II instrument.[17]

The total length of the TIMS device is only about 10 cm (Figure 1B), and it is operated with modest potentials of less than 300 V. Since the resolution is mainly determined by the rate at which the electrical field is decreased, the user is free to adjust it based on experimental needs. For slow ramp times, ion mobility resolutions (expressed as R = $\Omega/\Delta\Omega$, where Ω is the collisional cross-section) of more than 200 have been demonstrated[28,29] and faster scan out times of about 50 ms still allow R > 40.[30]

Parallel accumulation—serial fragmentation (PASEF)

Ion mobility—mass spectrometry adds an additional dimension of separation to the standard MS scans. The elution ramp described above defines a complete cycle of TIMS-MS, with the ions that have the lowest mobility (largest collisional cross sections) in relation to their charge state passing through the instrument first. During the elution ramp, TOF spectra are recorded at high frequency (~8.7 kHz). We term the data structure generated during a TIMS-MS cycle a "TIMS scan". Several TIMS scans can be added to obtain a TIMS-MS spectrum of a desired signal-to-noise.

The TOF spectra from one scan can be summed (i.e., projected onto the m/z axis) to determine the m/z and intensity of all the ions present. This information can then be used to create a topN target list for fragmentation. MS/MS is performed in the usual way by setting the quadrupole transmission window to the values in the topN list and applying the collision energy that is optimal for the precursor mass and charge state. The MS/MS spectra produced when incorporating ion mobility are similar in most respects to those obtained by MS/MS only, with two principal differences: First, compared to operation without IMS, in which the signal is recorded continuously as long as the precursor is selected, the signal is compressed into a short time—i.e. the duration of the mobility peak. This leads to better signal-to-noise as signal is concentrated whereas noise is distributed. Second, the different precursors present in the quadrupole selection window are separated from each other by their different ion mobilities even at the same m/z. This alleviates the 'Precursor Ion Fraction (PIF) problem', which comes about because PIF values are much smaller than one for a majority of precursors in proteomics experiments.[4,33]

Despite these two advantages, MS/MS experiments performed in the above-described way still only make use of one selection window per scan, in common with all methods that use a mass selecting quadrupole to isolate precursors (Figure 2, upper panel). Inspired by the controllable nature of the release of ions from the TIMS device and the fact that cross sections roughly correlate with precursor mass,[34] we wondered if it would be possible to utilize several precursors in one scan, without giving up the mass selection of each of them by the quadrupole. In the PASEF method described and demonstrated here, the quadruple is positioned at the m/z of a precursor eluting from the TIMS analyzer and is then rapidly moved to the next one as soon as it has eluted (Figure 2, lower panel). In
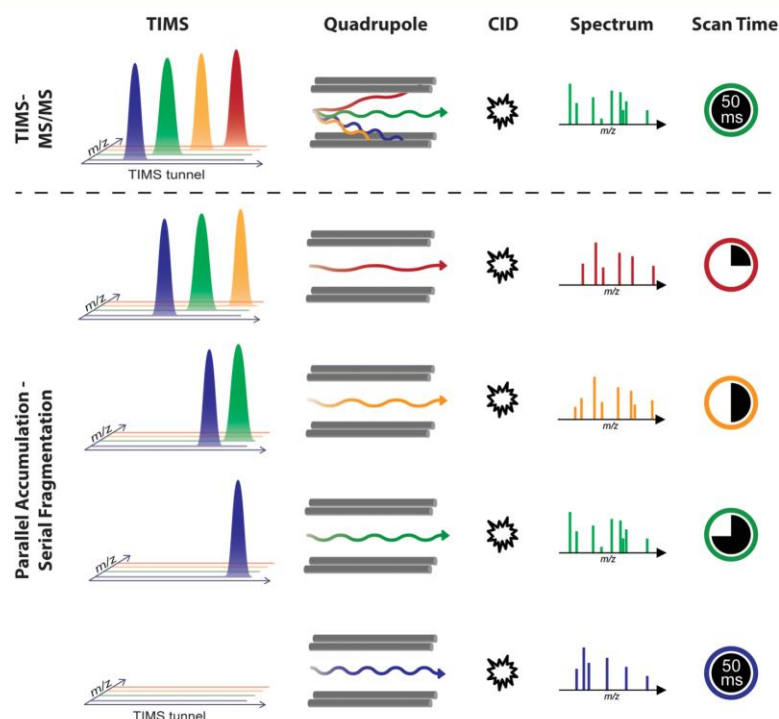
74

Figure 2. Illustration of the PASEF method in comparison with the standard TIMS-MS/MS operation mode. The top panel shows the selection of one precursor from a single TIMS scan, while all others are discarded. Conversely, the PASEF method (bottom panel) involves rapid switching of the quadrupole mass position to select multiple precursors at different m/z on the very same time scale. In this case, all targeted ions are fully used for fragmentation.

this way, the full intensity of the precursors that have been accumulated together can be utilized in one TIMS scan. This increases the speed of MS/MS by the number of precursors that are targeted. Alternatively, the same precursors can be selected in different TIMS scans to gain sensitivity; again by a factor up to the number of selected precursors per TIMS scan, or speed and sensitivity advantages can be combined as desired.

Time scales and required switching times

Key to realizing the concept of PASEF are the efficient storage of ions of the intended precursor range, high ion mobility resolution, and the extremely rapid switching of the quadrupole between precursors on the time scale of a single TIMS scan. To establish the hardware requirements for this, it is worthwhile to first consider the time scales of the individual processes involved in the LC-MS/MS experiment. With the QTOF setup in our laboratory,[17] we typically achieve average chromatographic peak widths of about 7 s (FWHM) during a 90 min LC gradient. The MS acquisition cycle for a data-dependent top17 method is completed within 1.2 to 1.4 s. Given a 200 ms interval for summing MS survey scans of 110 μs each, up to 17 precursors per second are selected by the quadrupole mass filter and fragmented in the collision cell. Therefore, the quadrupole mass position, as well as the collision energy, are switched every 60 ms. During this time, the TOF mass analyzer performs about 550 scans.

When incorporating ion mobility separation into this workflow, it becomes evident that a typical TIMS scan of 20 to 60 ms is orders of magnitude shorter than either the chromatographic peak widths or the topN cycle times (Figure 3). Conversely, the TIMS scan is about 100-fold longer than the time required for a single TOF scan and thus readily fits in between the chromatographic time scale and the TOF scan time. Individual ion mobility peaks have half widths of one millisecond or less (see below). Since the PASEF method aims to select multiple precursors from a single TIMS scan, precursor isolation should also happen within one millisecond. Note that this requires a much faster precursor isolation than the 60 ms of the normal top17 method, ideally by about 100-fold. To investigate the feasibility of quadrupole switching at these time scales, we first evaluated the rise times of the power supplies and lag times that potentially affect ion transmission. This revealed that the quadrupole is in principle capable of stepping its mass position by 3000 Th within 1.5 ms or less. Considering the much narrower m/z range relevant for proteomics, this should make even sub-millisecond switching times possible. Adding the width of a typical ion mobility peak results in a quadrupole isolation time per precursor of currently 2.5 to 4 ms. This translates into a maximum of 12−20 precursors per 50 ms TIMS scan, or 240−400 precursors per second. Importantly, the signal per precursor still corresponds to the full accumulation time. Thus, PASEF should be able to measure MS/MS spectra up to 12−20 times faster without losing sensitivity. Clearly, this number exceeds the number of precursors that can usefully be selected for fragmentation in contemporary shotgun proteomics experiments. Therefore,
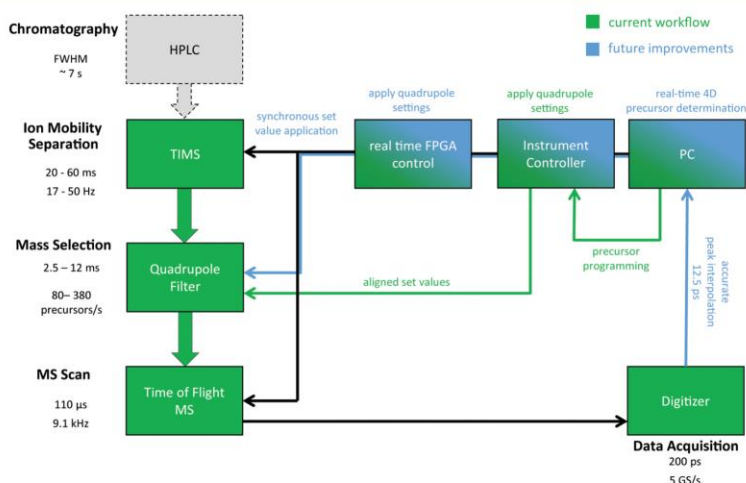
Figure 3. Investigation of hardware requirements for the PASEF method, including time scales and switching times of a typical proteomics LC-TIMS-MS/MS experiment. Arrows and lines visualize the hardware connections in the currently used setup (green) and planned future improvements (blue). GS/s, Giga-samples per second.

excess sequencing speed can be expended on improving identification scores or MS/MS-based quantification as explained in more detail below.

Precursor selection at this frequency can also be limited by the rate at which the quadrupole set values are calculated and applied by the instrument controller. In fact, we found this to be one of the major bottlenecks of our current setup since the set values are transferred via a comparably slow serial interface, precluding a fully synchronized operation of the quadrupole and TIMS. For our proof of principle experiments, we here circumvent this limitation by determining the elution times from TIMS precursor scans and applying the mass position and switching times for MS/MS scans to the quadrupole via the instrument controller in an asynchronous manner. This enabled us to select four precursors for a single TIMS scan. In the future, we will overcome this limitation by moving the calculations and set value application to the real time field-programmable gate array (FPGA), which already synchronizes TIMS and TOF analyzer. An additional direct interface to the quadrupole driver will allow set value applications in less than 50 μs, thus no longer compromising the quadrupole switching times (depicted in blue in Figure 3).

Note that the PASEF operation performed here will not change due to these improvements, except that we will be able to make use of the maximum mass selection rate of the quadrupole. Moreover, advanced peak detection and compression algorithms executed by the digitizer, capable of handling several million peaks per second, will enable real-time precursor determination in four dimensions (LC retention time, elution time from the TIMS device, m/z value and intensity) as required for data-dependent topN methods. To not compromise the proteome coverage in a shotgun experiment, this will also involve dynamic exclusion of already sequenced peptides. However, as the precursor determination itself can take up to a few milliseconds, we plan to parallelize precursor search and data acquisition. Notably, this strategy would also support more sophisticated precursor search algorithms without reducing the duty cycle for data acquisition. From

the discussion above, we conclude that current limitations are not of a fundamental nature, but tasks for engineering and data-handling which are likely to be accomplished in the near future.

Application of PASEF to complex protein digests

Having established sub-millisecond switching of the quadrupole isolation window, we next aimed to examine the novel method under circumstances that mimic the simultaneous elution of many peptides from the chromatographic column in shotgun proteomics. We thus directly infused an unseparated mixture of digested ADH, BSA, enolase and phosphorylase b to generate high peptide complexity (Materials and Methods).

Figure 4A shows the result of an ESI-TIMS-MS analysis of the four protein mixture. Without ion mobility separation, this experiment would have yielded a very complex mass spectrum with multiple overlapping signals as indicated by the projection on the m/z axis on the right. TIMS separates these overlapping ion species by their mobility, resulting in much less complex mass spectra per TOF scan. In accordance with others,[19,35] we observe a significant correlation between m/z and mobility, whereas higher m/z species are less mobile and vice versa. There are two main populations that can be assigned to different charge states, with singly charged species being less mobile than their multiply charged counterparts at similar m/z. Charge 2−5 species—potential tryptic peptides—make up a more dense population. Nevertheless, close inspection of the mobility peak widths in this regime revealed that the elution time chosen (ramping down the TIMS tunnel gradient in 50 ms) was sufficient to achieve a mobility resolution above 40 and to separate these species.

From the heat map created by the multiply charged tryptic peptides of the protein mixture, we first selected four mobility separated precursors. Corresponding mass positions at m/z 810.3 (1), 714.3 (2), 559.3 (3) and 560.6 (4) and the appropriate switching times were uploaded to the instrument controller. As apparent from Figure 4B, the quadrupole correctly isolated these precursors on the TIMS time scale. In each case, the entire peak was quantitatively captured. Projection onto the m/z scale also shows successful isolation,
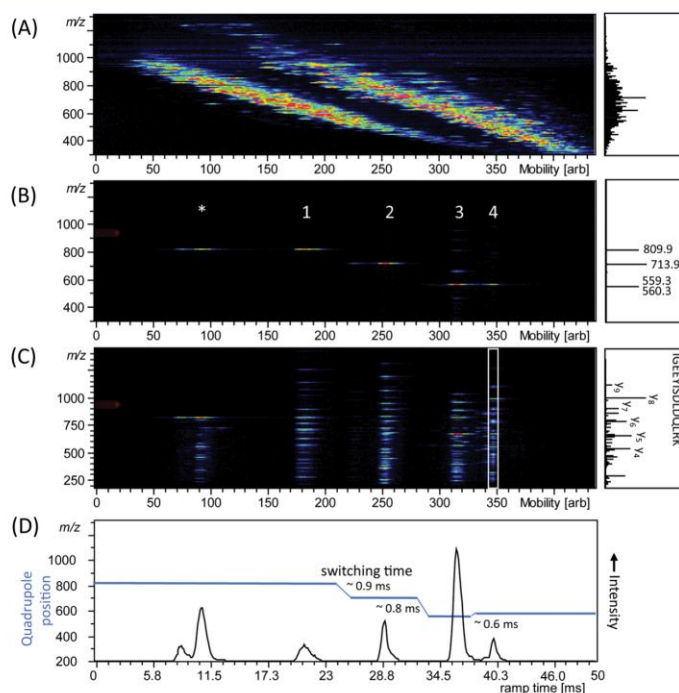
76

Figure 4. TIMS-QTOF analysis of electrosprayed peptides from a combined tryptic digest of ADH, BSA, phosphorylase b, and enolase. (A) Nested m/z and ion mobility distribution as detected in full scan MS. (B) Sequential isolation of four ions at different m/z that are separated by mobility after parallel accumulation. (C) Parallel accumulation−serial fragmentation (PASEF) of the precursors as isolated in (B). (D) Arrival time distribution of the summed fragment ions as observed by the PASEF method together with the quadrupole isolation mass as a function of the TIMS ramp time. Projected mass spectra are shown on the right of the panels, in (C) the fragments of the white box are projected.

and comparison to the spectrum without quadruple isolation demonstrates a drastic simplification of the precursor population. Precursors (3) and (4) only differ in one m/z unit, representing the classical PIF problem, but are clearly separated by the combination of ion mobility and quadrupole isolation. Another notable feature is the peak isolated before peptide (1) and marked with an asterisk. This is caused by the fact that in the current configuration the quadruple immediately started selection from the first m/z value. Consequently, it isolated a precursor from the singly charged population, which is widely separated from the actual target and upon inspection turns out to consists of two species that are distinct in mass and ion mobility.

In the next step, we applied the PASEF method by performing MS/MS on the isolated precursors. This led to a characteristic ladder of fragment ions at each precursor (Figure 4C, Supplementary Figure S1). Projection of the ladder generated by peptide (4) shows a standard MS/MS spectrum, corresponding to the sequence IGEEISDLDQLRK from phosphorylase B. The other peptides corresponded to VLGIDGGEGKEELFR (1) from enolase, HLQIIYEINQR (2) from phosphorylase B and VAAAFPGDVDR (3), also from phosphorylase B. Interestingly, we were also able to assign one of the singly charged species to ADH based on its peptide fragment spectrum (YVVDTSK). Projection of all fragment ions onto the ion mobility axis shows coherence in arrival times, with fragment ion distributions very similar to their precursors (Figure 4D). These peaks were less than 3 ms wide, with typical

half-widths of around 1 ms. In contrast, the chosen quadrupole isolation times were at least three times longer and could therefore in principle have been shortened considerably. Combined with the fast switching times apparent in the figure, this shows that at least ten precursors could have been targeted in this 50 ms ramp.

To demonstrate PASEF on a larger scale, we extended the previously described experiment to a total of ten sets of four precursors each. The results on the 40 precursors were very similar to the example discussed above and are summarized in Table 1. The selected precursors span a mass range from m/z 418 to 956 and exhibited ion mobilities corresponding to TIMS elution times between 17 and 44 ms with an average half width of 0.8 ± 0.2 ms. In this experiment we also quantified the gain due to PASEF by comparing it to standard fragmentation of the 40 precursors with TIMS. With full realization of the PASEF concept, we would expect that we would preserve full signal intensity despite being 4-fold faster (or N-fold in general, where N is the number of selected precursors). This was fully validated by the results, which showed a ratio indistinguishable from one between the two cases (0.94 ± 0.12, Table 1).

Modeling of coverage improvements in shotgun proteomics

Having demonstrated that PASEF is capable of a four to more than 10-fold increase in MS/MS speed at full sensitivity, we next modeled the effect of such an improvement on typical shotgun proteomics data. We built on our previous analysis, in

77

Table 1. PASEF Analysis of 40 Precursors from a Complex Mixture of Four Digested Proteins

Scan	Protein	Peptide Sequence Targeted	m/z	Charge	Mobility [ms]	FWHM [ms]	Ratio PASEF/TIMS[a]
1	Enolase	VLGIDGGEGKEELFR	809.956	2	22.6	1.3	0.87
	Phos b	HLQIIYEINQR	713.897	2	30.4	0.7	0.85
	Phos b	VAAAFPGDVDR	559.253	2	37.4	0.9	0.97
	Phos b	IGEEYISDLDQLRK	560.261	3	40.9	0.5	0.96
2	Enolase	SIVPSGASTGVHEALEMR	921.033	2	17.4	1.3	0.95
	Phos b	IGEEYISDLDQLR	775.914	2	28.2	0.7	0.97
	BSA	LVNELTEFAK	582.292	2	36.2	0.7	0.77
	BSA	LFTFHADICTLPDTEK	636.633	3	39.6	1.1	0.90
3	Phos b	TCAYTNHTVLPEALER	938.031	2	22.0	1.5	0.94
	Phos b	VLYPNDNFFEGK	721.862	2	28.7	0.7	1.00
		no match	658.303	4	35.1	0.6	1.12
	Phos b	LITAIGDVVNHDPVVGDR	630.659	3	39.7	0.9	1.00
4	Enolase	TAGIQIVADDLTVTNPK	878.527	2	23.2	0.7	1.07
		no match	767.925	2	29.8	0.9	0.93
	BSA	DAIPENLPPLTADFAEDKDVCK	820.099	3	34.2	0.7	1.10
	Phos b	TNFDAFPDK	527.705	2	40.3	0.6	0.82
5	Phos b	WLVLCNPGLAEIIAER	927.568	2	22.4	0.8	0.95
	Enolase	AVDDFLISLDGTANK	789.928	2	29.3	0.8	0.92
		no match	594.311	2	35.5	0.7	1.01
	Phos b	TCAYTNHTVLPEALER	625.624	3	39.9	0.9	0.82
6		no match	955.970	2	17.3	0.7	1.09
	Phos b	DFNVGGYIQAVLDR	783.927	2	27.6	0.8	0.91
	ADH	SISIVGSYVGNR	626.323	2	34.5	0.6	0.92
	Enolase	VLGIDGGEGKEELFR	540.249	3	39.0	0.8	0.92
7	BSA	YNGVFQECCQAEDK	874.403	2	27.1	0.9	0.93
	Phos b	VFADYEEYVK	631.784	2	33.6	0.6	(0.37)[b]
	Phos b	QRLPAPDEK	527.248	2	37.5	0.8	0.90
	BSA	DDPHACYSTVFDK	518.847	3	43.4	0.6	0.72
8	Phos b	IGEEYISDLDQLRK	839.973	2	24.2	0.8	1.05
	ADH	GLAGVENVTELKK	679.383	2	31.4	0.8	1.04
	Phos b	ARPEFTLPVHFYGR	563.934	3	37.7	0.6	0.78
	ADH	IGDYAGIK	418.668	2	43.6	0.6	1.03
9	BSA	KVPQVSTPTLVEVSR	820.502	2	26.4	0.8	0.98
	Phos b	LLSYVDDEAFIR	720.877	2	30.2	1.0	0.88
	Phos b	VAIQLNDTHPSLAIPELMR	706.720	3	34.9	0.8	1.05
	Phos b	APNDFNLK	459.682	2	41.2	0.6	(0.28)[b]
10		no match	782.429	2	23.8	0.7	1.16
	BSA	HLVDEPQNLIK	653.351	2	32.6	0.8	1.13
	Enolase	IGSEVYHNLK	580.281	2	35.3	0.7	0.67
	Phos b	NLAENISR	458.691	2	39.7	0.6	0.70

[a]Median summed fragment ion intensities were extracted for each precursor from PASEF (N = 331) and TIMS-MS/MS (N = 9) scans with identical quadrupole isolation settings. [b]As an artifact resulting from the asynchronous operation of TIMS and quadrupole, these two precursors were not isolated in each PASEF scan.

which we used MaxQuant to determine the number of isotope patterns ("peptide features") as a function of retention time in a 90 min analysis of a HeLa digest and used this to evaluate speed, sensitivity and separation power achievable in shotgun proteomics.[4] We repeated this analysis using our data from the impact II QTOF instrument,[17] on which our prototype is based (Materials and Methods). From the first to last eluting peptides, about 50 to 70 unique features/s are detected by MaxQuant, of which about 12 had been fragmented per second (Figure 5A). The 4-fold improvement in sequencing speed already demonstrated above would allow targeting 80% of all detected features. We noted above that a 10-fold improvement is entirely consistent with the actually achieved switching times and ion mobility resolution. This would result in a targeting rate of about 125 features/s, far above the number of features

detected in the previous data set. That said, adding ion mobility could be expected to increase the number of resolvable peptide features in shotgun proteomics runs, in which case even a 10-fold increased sequencing speed could in principle be used entirely on unique peptide features.

However, not all detected features are suitable for MS/MS fragmentation because their peptide intensity may be too low to result in useful MS/MS spectra. Redoing the peptide histogram analysis that we had done on the original Orbitrap data[4] revealed 3-fold larger numbers for the impact II data set but the same proportions: About 250,000 peptide features can be detected in the 90 min gradient, of which about 45,000 were fragmented, and about 30,000 with attendant peptide identification (Figure 5B).
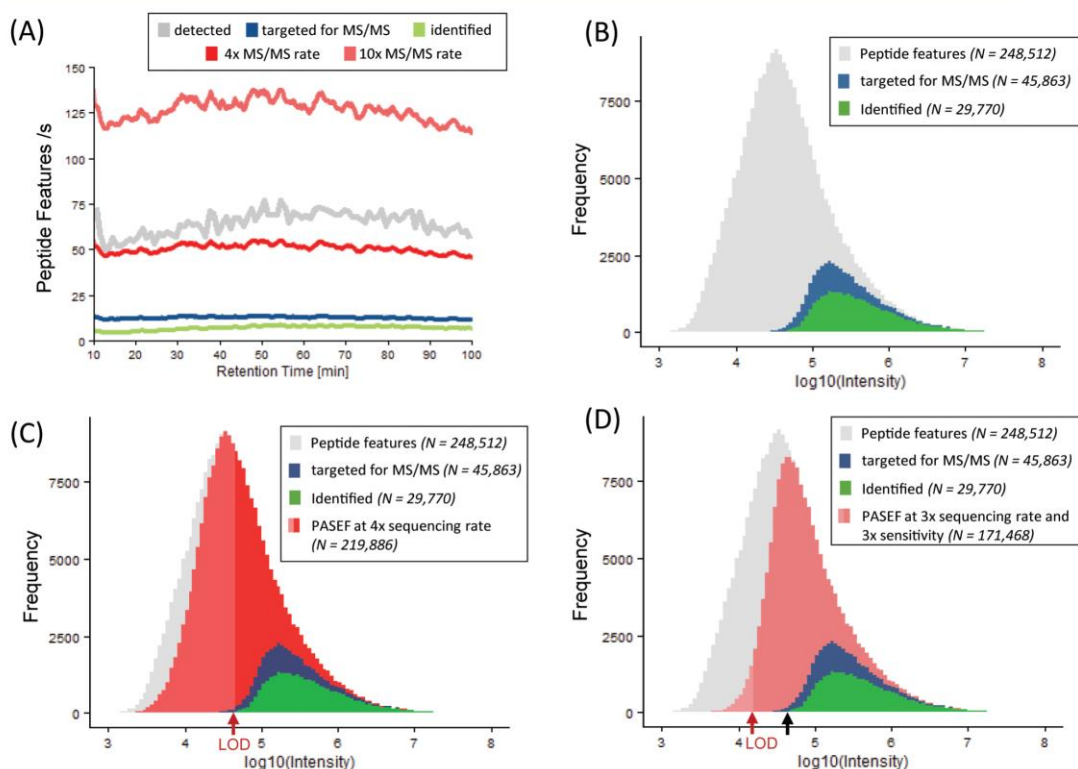
78

Figure 5. Modeling the benefits of the PASEF method in shotgun proteomics. (A) Peptide features per second eluting during a 90 min gradient. The gray trace indicates detected peptide features during conventional LC-MS/MS analysis of a tryptic HeLa digest on a QTOF instrument; blue and green traces indicate features that were targeted for MS/MS and successfully identified in the impact II data set.[17] Red traces represent the simulated sequencing speed of PASEF with four and ten precursor ions per TIMS scan, respectively. (B) Intensity distribution of detected (gray), targeted (blue), and identified (green) peptide features extracted by MaxQuant from the impact II data set. (C) Same as (B) with a simulation of targeted peptides at 4-fold sequencing speed with the PASEF method (red). (D) Same as (B) with a simulation of a PASEF experiment with three precursors per scan and 3-fold increased sensitivity due to retargeting (red). MS features below the anticipated limit of detection (LOD) are indicated by reduced opacity.

With 4-fold higher sequencing speed, essentially the entire population of detected peptide features can now be targeted, but lower intensity precursors would diminish identification success (Figure 5C). The sensitivity of the impact II instrument for peptide identification is given by the lower limit of the green population in the peptide histogram. If we assume that the current instrument reaches at least the same sensitivity, then a 4-fold improvement in sequencing at full sensitivity, allows filling in about 40% of the entire peptide precursor population, a 250% improvement from before.

As mentioned above, the PASEF advantage could be used for increasing sequencing speed or sensitivity—through targeting the same feature repeatedly. In practice, we imagine that one would use a combination of both, for instance by fragmenting sufficiently abundant precursors once but summing MS/MS spectra for lower abundant ones. To model this in a simple way, we equally assigned the expected 10-fold improvement due to PASEF to sequencing speed and sensitivity gain (Figure 5D). This leads to a similarly shaped distribution of targeted precursor as in the experimental distribution, with the difference that the target and likely identified populations are increased by 300%. As a result, 70% of the overall precursor population can now be targeted and potentially identified by MS/MS. Future experiments, using a full LC-TIMS-MS/MS set up, will reveal optimal combinations of refragmentation and precursors per TIMS scan. We also note that more sophisticated selection strategies for MS/MS could further improve the proportion of successfully identified peptides and that improvements in ion beam intensity would increase both the total number of detected features as well as the number of targetable and identifiable peptides.

■ **CONCLUSIONS**

Coupling ion mobility spectrometry to MS comprises several advantages, such as separation of ions from protein mixtures according to their size-to-charge ratio. Using the compact TIMS analyzer implemented into a state of the art QTOF mass spectrometer, we have here introduced the concept of PASEF. Sub-millisecond quadrupole switching times allowed us to select multiple precursors for fragmentation during a 50 ms ion mobility scan instead of only one. Importantly, we demonstrated that synchronized quadrupole and TIMS operation is fully quantitative in that the signal is not diminished compared to single precursor selection. An important advantage of PASEF

is that the resulting spectra—in addition to the ion mobility dimension—are fully precursor mass resolved, unlike recently proposed data independent strategies. This also makes PASEF compatible with reporter ion based chemical multiplexing strategies, such as iTRAQ or TMT. The about 10-fold gain that should be achievable by PASEF in shotgun proteomics experiments can be employed as increased sequencing speed without a decrease in sensitivity. However, modeling suggests that a combination of targeting more precursors and targeting weak precursors repeatedly, will be most effective. While demonstrated here for the TIMS-QTOF combination, the PASEF principle could be applied to any ion mobility—mass spectrometer configuration with the required sub-millisecond scan speed in the MS read out.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00932.

Supplementary Figure S-1: Annotated PASEF-MS/MS spectra of the peptides shown in Figure 4 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49-89-8578-2557. E-mail: mmann@biochem.mpg.de.

Author Contributions

All authors contributed to the research design; F.M., S.B., and M.L. performed research; F.M., S.B., M.L., O.R., and M.M. analyzed data; F.M., S.B., O.R., and M.M. wrote the manuscript. All authors have given approval to the final version of the manuscript. F.M. and S.B. contributed equally.

Notes

The authors declare the following competing financial interest(s): Several authors are employees of Bruker.

■ ABBREVIATIONS

ADH, alcohol dehydrogenase; BSA, bovine serum albumin; FPGA, field-programmable gate array; FWHM, full width at half-maximum; GS/s, Giga-samples per second; IM, ion mobility; LC, liquid chromatography; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PASEF, parallel accumulation—serial fragmentation; PIF, precursor ion fraction; QTOF, quadrupole time-of-flight; RF, radiofrequency; TIMS, trapped ion mobility spectrometry; TOF, time-of-flight

■ REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(2) Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd. The biological impact of mass-spectrometry-based proteomics. *Nature* **2007**, *450* (7172), 991−1000.

(3) Altelaar, A. F.; Munoz, J.; Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **2013**, *14* (1), 35−48.

(4) 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.* **2011**, *10* (4), 1785−93.

(5) Kelstrup, C. D.; Jersie-Christensen, R. R.; Batth, T. S.; Arrey, T. N.; Kuehn, A.; Kellmann, M.; Olsen, J. V. Rapid and Deep Proteomes by Faster Sequencing on a Benchtop Quadrupole Ultra-High-Field Orbitrap Mass Spectrometer. *J. Proteome Res.* **2014**, *13*, 6187.

(6) Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. 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* **2014**, *13* (12), 3698−708.

(7) Li, G. Z.; Vissers, J. P.; Silva, J. C.; Golick, D.; Gorenstein, M. V.; Geromanos, S. J. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* **2009**, *9* (6), 1696−719.

(8) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. 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* **2012**.

(9) 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.; Pappin, D. J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3* (12), 1154−69.

(10) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895−904.

(11) Wiese, S.; Reidegeld, K. A.; Meyer, H. E.; Warscheid, B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* **2007**, *7* (3), 340−50.

(12) Andrews, G. L.; Simons, B. L.; Young, J. B.; Hawkridge, A. M.; Muddiman, D. C. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Anal. Chem.* **2011**, *83* (13), 5442−6.

(13) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77* (7), 2187−200.

(14) Silva, J. C.; Denny, R.; Dorschel, C.; Gorenstein, M. V.; Li, G. Z.; Richardson, K.; Wall, D.; Geromanos, S. J. Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. *Mol. Cell. Proteomics* **2006**, *5* (4), 589−607.

(15) Distler, U.; Kuharev, J.; Navarro, P.; Levin, Y.; Schild, H.; Tenzer, S. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **2014**, *11* (2), 167−70.

(16) Helm, D.; Vissers, J. P.; 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.; Kuster, B. Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell. Proteomics* **2014**, *13*, 3709.

(17) 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.; Mann, M. The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics. *Mol. Cell. Proteomics* **2015**, *14* (7), 2014−29.

(18) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. Three-dimensional ion mobility/TOFMS analysis of electrosprayed biomolecules. *Anal. Chem.* **1998**, *70* (11), 2236−42.

(19) 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.* **1998**, *9* (11), 1213–6.

(20) Kanu, A. B.; Dwivedi, P.; Tam, M.; Matz, L.; Hill, H. H., Jr. Ion mobility-mass spectrometry. *J. Mass Spectrom.* **2008**, *43* (1), 1–22.

(21) May, J. C.; McLean, J. A. Ion mobility-mass spectrometry: time-dispersive instrumentation. *Anal. Chem.* **2015**, *87* (3), 1422–36.

(22) Hoaglund-Hyzer, C. S.; Clemmer, D. E. Ion trap/ion mobility/quadrupole/time-of-flight mass spectrometry for peptide mixture analysis. *Anal. Chem.* **2001**, *73* (2), 177–84.

(23) 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.* **2013**, *12* (6), 2323–39.

(24) 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.* **2014**, *6* (4), 281–94.

(25) Cumeras, R.; Figueras, E.; Davis, C. E.; Baumbach, J. I.; Gracia, I. Review on ion mobility spectrometry. Part 1: current instrumentation. *Analyst* **2015**, *140* (5), 1376–90.

(26) Fernandez-Lima, F. A.; Kaplan, D. A.; Park, M. A. Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **2011**, *82* (12), 126106.

(27) Fernandez-Lima, F.; Kaplan, D. A.; Suetering, J.; Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobility Spectrom.* **2011**, *14* (2–3), 93.

(28) Silveira, J. A.; Ridgeway, M. E.; Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* **2014**, *86* (12), 5624–7.

(29) Ridgeway, M. E.; Silveira, J. A.; Meier, J. E.; Park, M. A. Microheterogeneity within conformational states of ubiquitin revealed by high resolution trapped ion mobility spectrometry. *Analyst* **2015**, *140* (20), 6964–72.

(30) Michelmann, K.; Silveira, J. A.; Ridgeway, M. E.; Park, M. A. Fundamentals of trapped ion mobility spectrometry. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (1), 14–24.

(31) R Development Core Team. *R: A Language and Environment for Statistical Computing*; 2014.

(32) Cox, J.; Matic, I.; Hilger, M.; Nagaraj, N.; Selbach, M.; Olsen, J. V.; Mann, M. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.* **2009**, *4* (5), 698–705.

(33) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152–60.

(34) Revercomb, H. E.; Mason, E. A. Theory of Plasma Chromatography Gaseous Electrophoresis - Review. *Anal. Chem.* **1975**, *47* (7), 970–983.

(35) Taraszka, J. A.; Counterman, A. E.; Clemmer, D. E. Gas-phase separations of complex tryptic peptide mixtures. *Fresenius' J. Anal. Chem.* **2001**, *369* (3–4), 234–45.

81

Article 4: 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[§], Matthias Mann[‡,‡‡,§§]

[‡] *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, USA*
[‖] *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*

Having established the PASEF concept in the end of 2015, it took us over two years to arrive at a proteomics platform that fully integrates PASEF for online data-dependent acquisition – the Bruker *timsTOF Pro*. Major hurdles on the way were the implementation of an online precursor scheduling algorithm and the development of post-processing software for the four-dimensional LC-TIMS-MS data, which was only recently accomplished by Juergen Cox and his team.

In line with our expectations from the 2015 paper, the full implementation of PASEF in the hard- and firmware indeed achieved sequencing rates above 100 Hz. In single runs a whole-cell HeLa digest, we sequenced over 20 precursors per TIMS scan depending on the ramp times, which is fast enough to use parts of the acquisition speed for re-sequencing of low-abundance precursors. Optimization of the scan parameters resulted in the identification of about 6,000 protein groups in 2 h single runs, which outperforms the predecessor instrument, the impact II, by a large margin and is very competitive in the field. The instrument provides a high quantitative reproducibility in technical replicates and accurately quantified HeLa and E.coli proteins in a mixed proteome experiment. The performance characteristic of the instrument are particularly advantageous for low sample amounts and very fast LC measurements. This is exemplified by the identification of 2,500 proteins in single runs from 10 ng HeLa digest and over 1,000 human proteins in only about 5 min.

An exciting and yet underexplored feature of the timsTOF Pro is the possibility to measure collisional cross sections (CCS) on a large scale. In the manuscript, we demonstrate that TIMS achieves a very high precision as evident from a median 0.1% absolute deviation of the CCS values in five replicates. In combination with the proteomic depth provided with this workflow, this opens up new perspectives to study gas phase properties of peptides on a scale that goes much beyond past reports[360].

Even though the manuscript marks the end-point of the PASEF development in my thesis, it appears to be only the starting point of many more advances on the basis of TIMS and PASEF, which will lead to powerful and widely spread applications in the field of MS-based proteomics and beyond.

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

Florian Meier[1], Andreas-David Brunner[1], Scarlet Koch[2], Heiner Koch[2], Markus Lubeck[2], 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*]

[1] *Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany*
[2] *Bruker Daltonik GmbH, Fahrenheitstr. 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, Blegdamsvej 3B, 2200 Copenhagen, Denmark*

*To whom correspondence may be addressed: mmann@biochem.mpg.de

ABSTRACT

In bottom-up proteomics, peptides are separated by liquid chromatography with elution peak widths in the range of seconds, while mass spectra are acquired in about 100 microseconds with time-of-fight (TOF) instruments. This allows adding ion mobility as a third dimension of separation. Among several formats, trapped ion mobility spectrometry (TIMS) is attractive due to 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 800,000 fragmentation spectra in standard 120 min LC runs are easily achievable, which can be used for near exhaustive precursor selection in complex mixtures or re-sequencing weak precursors. MaxQuant identified more than 6,400 proteins in single run HeLa analyses without matching to a library, and with high quantitative reproducibility (R > 0.97). Online PASEF achieves a remarkable sensitivity with more than 2,900 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.

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 to genomics and transcriptomics technologies, proteomics still lags behind 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 non-specialized laboratories.

In bottom-up workflows, proteins are extracted from a biological sample of interest and enzymatically cleaved, which makes them more amenable to mass spectrometric (MS) analysis. The resulting complex peptide mixtures are typically separated via nano-flow liquid chromatography (LC), 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].

Time-of-flight (TOF) instruments have a number of 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 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[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 they emerge from the ion mobility device, they can be efficiently sampled in the ms or sub-ms time frame. Nested between LC and MS, the technology provides an additional dimension of separation[15–17] and can increase analysis speed and selectivity[18], in particular with highly complex proteomics samples[19–23]. However, many implementations of IMS, such as drift tubes, are challenging to implement due to 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 a sufficient

number of 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 IMS analyzer dimensions to about 10 cm 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. We demonstrated that PASEF increases the sequencing speed several-fold 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, USA) were grown in Dulbecco's modified Eagle's medium with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin-streptomycin (all PAA Laboratories, Germany). 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, 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, 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 reverse-phase fractionation.

Peptide fractionation - High pH reversed-phase fractionation was performed on an EASY-nLC 1000 (Thermo Fisher

Scientific, Germany) coupled to a 'spider fractionator' (PreOmics GmbH, Martinsried, Germany) as detailed in ref [33]. Purified peptides were separated on a 30 cm x 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 24 fractions in 90 s time intervals. The fractions were dried in a vacuum-centrifuge and re-constituted in water with 2% ACN and 0.1% TFA for LC-MS analysis.

Liquid Chromatography - An EASY-nLC 1200 (Thermo Fisher Scientific, Germany) 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) 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 x 75 μm i.d.) with a pulled emitter tip, packed with 1.9 μm C_{18}-coated porous silica beads (Dr. Maisch, Germany). Mobile phases A and B were water with 0.1% formic acid (vol/vol) and 80/20/0.1% ACN/water/formic acid (vol/vol/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.

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 1.5 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 to which it features an additional ion mobility region. 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 changed 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

4

form a stacked multipole in the direction of ion transfer. 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 stores and pre-separates ions according to their mobility, and the second part performs trapped ion mobility analysis in parallel. Note that equal accumulation and analysis times in both TIMS regions enable operation at duty cycles up 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 50 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. Depending on their collisional cross sections and charge states, they 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 systematically varied the ramp times from 50, 100, 150, to 200 ms while keeping the duty cycle fixed at 100%. The quantification benchmark experiment, the 60 min dilution series and the high pH reverse-phase fractions were each acquired with a 100 ms ramp and 10 PASEF MS/MS scans per topN acquisition cycle; the 30 min dilution series was acquired with a 50 ms ramp and 10 PASEF MS/MS scans per

cycle; experiments on the Evosep One were performed with a 100 ms ramp and four PASEF MS/MS scans per cycle.

MS and MS/MS spectra were recorded from m/z 100 to 1,700. 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).

Data analysis - Mass spectrometry raw files were processed with MaxQuant[37] version 1.6.1.12, 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

5

in detail separately (Cox and co-workers, *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 ions 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. The 'TIMS half width' parameter was set to 4 TOF triggers, the 'TIMS step width' to 3, the 'TIMS mass resolution' 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: 4,313 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 4,600 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. 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).

Experimental Design and Statistical Rationale - 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. 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.

6

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. 1b**). 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 50 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 in a linear manner within 50 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
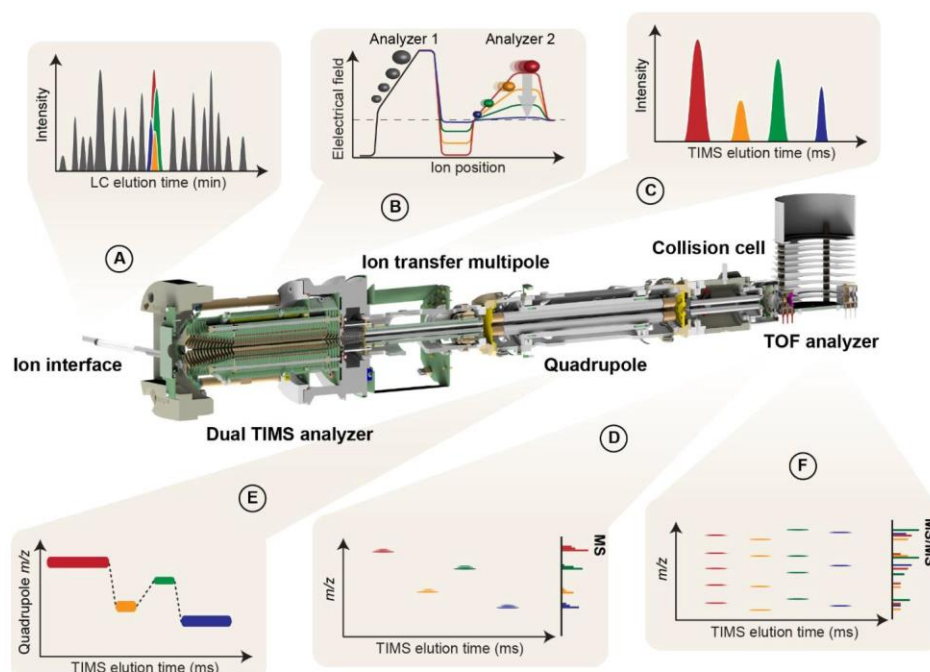

Figure 1 | Online Parallel Accumulation - Serial Fragmentation (PASEF) with the timsTOF Pro. (**A**) Peptides eluting from the chromatographic column are ionized and enter the mass spectrometer through a glass capillary. (**B**) In the dual TIMS analyzer, the first TIMS section traps and stores ion packets, 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 D and F for comparison.

7

90

less (**Fig. 1c**). 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 and the quadrupole mass filter and are 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. 1d**). The ions enter a V-shaped flight path through a two-stage reflectron and finally impinge on a multi-channel plate (MCP) ion detector coupled to a 10-bit digitizer with a sampling rate of 5 Gigasamples(GS)/s, enabling high-resolution mass analysis (R > 35,000 throughout the entire mass range). 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 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. 1e**). The *timsTOF Pro* electronics have been designed to meet these requirements and RF and DC voltages for mass selection are now calculated and set by a real-time field-programmable array (FPGA), 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 m/z windows, PASEF yields N ion-mobility-resolved MS/MS spectra for a single TIMS scan (**Fig. 1f**). 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 selection even with the ten-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, similar to the 'travelling salesman problem' in computer science. Even though exact solutions exist, for example by a brute-force method that
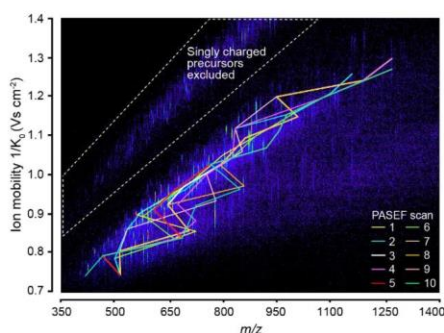
8

Figure 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).

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. An alternative strategy is to deliberately re-sequence selected low-abundance precursor ions in subsequent PASEF scans to obtain summed spectra with increased signal-to-noise. This 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, 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* vs. ion mobility plane. The flow chart in **Supplementary Fig. 1** 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. **Figure 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
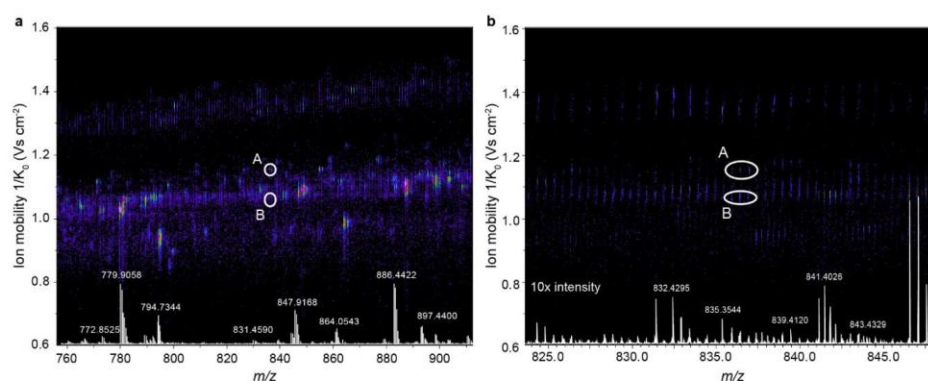
92

Figure 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 Suppl. Fig. 1.

separation of precursors along the additional ion mobility dimension was crucial as illustrated in **Figure 3**. In a projection of the data onto the *m/z* axis, no obvious precursor signals were present, even when enlarging the signal ten-fold relatively to the more abundant peaks. However, the precursor selection algorithm had found and fragmented two distinct isotope clusters in ion mobility – *m/z* space, which were separately fragmented by PASEF and clearly identified (**Supplementary Fig. 2**).

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 time allows 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 50 to 200 ms and kept the 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, an average of 23.3 precursors were sequenced per scan (**Fig. 4a**). Faster ramp times resulted in nearly proportionately less precursors per PASEF scan, but due to the higher number of scans per analysis, faster scans generated more MS/MS events in total - up to a remarkable 840,000 spectra in two hours (**Fig. 4b**). For comparison, acquiring the same number of MS/MS spectra without PASEF at the same sensitivity would have taken 12 times longer - about one day. For all ramp times, the instrument was sequencing at rates above 100 Hz during the time that peptides were eluting. We decided to use this extreme speed in part on re-sequencing low-abundance peptides to generate higher-quality summed spectra (**Fig. 4c**). On average, a given precursor ion
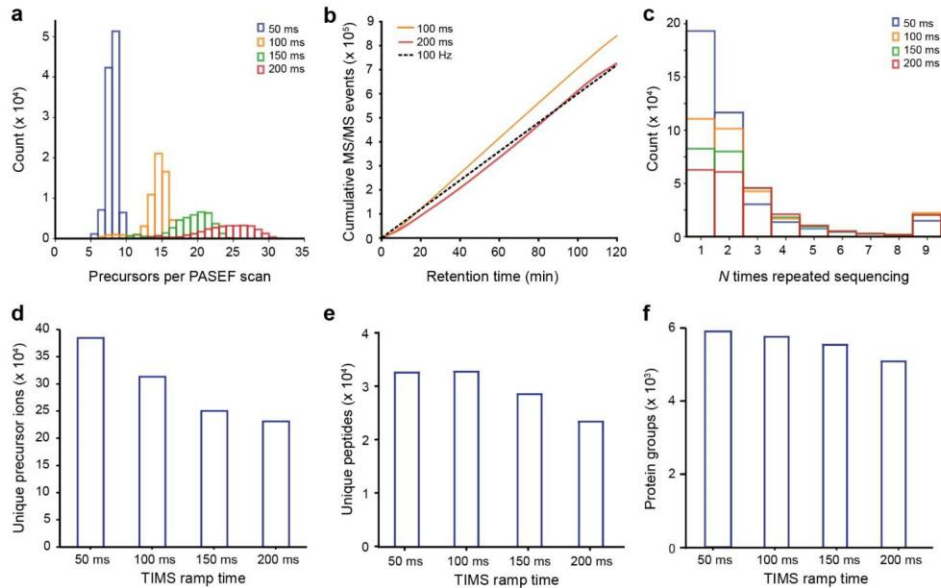
Figure 4 | Single run analyses of a HeLa digest. (**a**) Number of selected precursor ions per PASEF scan with different TIMS ramp times in 120 min runs of 200 ng HeLa digests. (**b**) Cumulative number of PASEF MS/MS spectra as a function of retention time for 100 ms and 200 ms TIMS ramps. The dashed line indicates the theoretical number of MS/MS spectra for a constant acquisition rate of 100 Hz (**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 a single run (N=4) with different TIMS settings. (**f**) Average number of protein group identifications in a single run (N=4) with different TIMS settings.

was fragmented 2.1 times in 50 ms ramps and 3.1 times with 200 ms ramps. Overall, this resulted in up to 380,000 MS/MS spectra of unique precursor ions in a single run as detected by the real-time PASEF scheduling algorithm, although post-processing in MaxQuant combined many of these (**Fig. 4d**).

From 200 ng whole-cell HeLa digest per run, we identified on average 23,696 sequence-unique peptides in quadruplicate single runs with the 200 ms method, and about 33,000 with the faster 50 ms and 100 ms methods (**Fig. 4e**). Average peptide length was 15 amino acids, similar to that expected from in silico digests of the UniProt database given our minimum peptide length of seven. The number of inferred protein groups at a false discovery rate (FDR) below 1% increased to an average of 5,970 protein groups per run with decreasing TIMS ramp times from 200 to 50 ms (**Fig. 4f**). With the 50 ms ramps, we identified in total 6,491 protein groups (5,753 with two or more peptides) with a median sequence coverage of 19.9%. This is an excellent value given the very low starting amount and the absence of fractionation or a matching library.
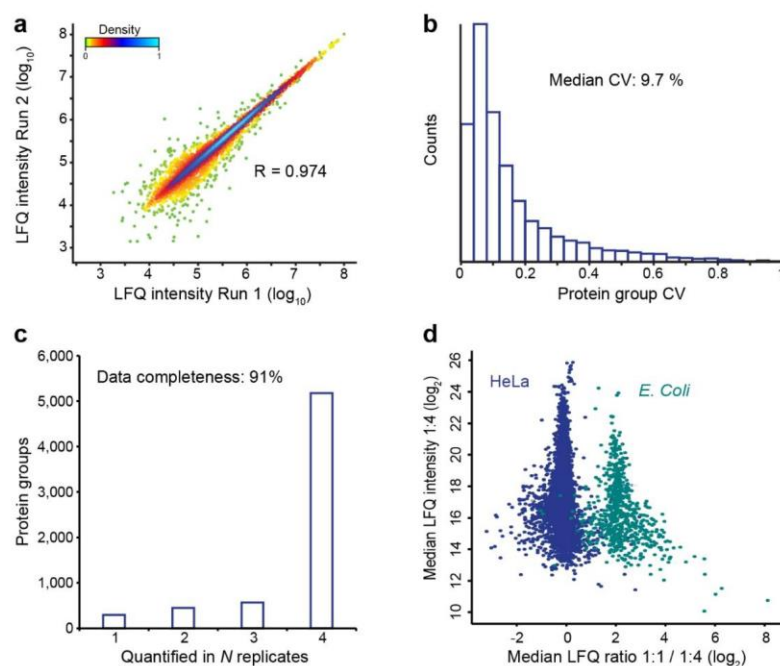
11

Figure 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. (**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 1:4 ratios (wt:wt). The scatterplot shows the median fold-change of human and E.coli proteins in quintuplicate single runs.

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 for this due to its simplicity, and it scales well to larger sample cohorts. Using the optimized 50 ms TIMS method we quantified on average 5,903 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 0.97 between the four runs, with excellent linearity over 4.5 orders of magnitude in protein abundance (**Fig. 5a**). The median coefficients of variation were 19.7% for the non-normalized peptide intensities and 9.7% at

the protein level after MaxLFQ normalization[39] (**Fig. 5b**).

Quantitative accuracy in proteomics may be limited if proteins are inconsistently measured across the samples. In data-dependent acquisition schemes, this is partially due to semi-stochastic precursor selection – a consequence of the large number of co-eluting precursor candidates and the finite sequencing speed. We asked if the several-fold 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 5,177

12

proteins in four out of four runs (**Fig. 5c**). Only 294 low-abundance proteins were exclusively quantified in a single replicate. This translated into a data completeness of 91%, 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 run' 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 1:4 ratios and measured each sample in quintuplicate 120 min single runs. This quantified 5,268 protein groups (4,565 HeLa; 703 E.coli) in at least one out of five replicates in both experimental conditions. Plotting the median fold-changes yielded two distinct clouds for HeLa and *E.coli* proteins, which were 4.6-fold separated in abundance, slightly more than the intended 4-fold mixing ratio (**Fig. 5d**). Both populations were relatively narrow (σ(HeLa) = 0.44; σ(E.coli) = 0.81) and they had minimal overlap. Without imputation, a one-sided Student's t-test returned 588 significantly changing *E.coli* proteins with at least two valid values in each group (of 621) at a permutation-based FDR below 0.05. This represents an excellent sensitivity of ~95% and at the same time, only 64 human proteins (1.5%) were false 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. 6a**). With 100 ng on column and a 1 h gradient, we reproducibly identified 4,515 protein groups, 76% of the proteome coverage with 200 ng in half the measurement time (**Fig. 6b**). Out of these, 3,346 protein groups were quantified with a CV below 20%. At 50 ng, we identified over 4,000 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 2,741 protein groups on average and 3,160 in total (2,322 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 9.2% and 1,890 proteins quantified at a CV < 20%.

To investigate achievable throughput, we repeated our sensitivity experiments with a 30 min gradient (**Fig. 6c,d**). Because of the very high sequencing speed of PASEF, reducing the measurement time had only limited effect on proteome coverage. From 100 ng HeLa digest we identified on average 3,649 protein groups in quadruplicate single runs, whereas 10 ng yielded 2,536 protein groups, all with median CVs below 12%. For the 10 ng runs, this represents 93% of the proteome

13

96

coverage of the 60 min single runs in half the time.

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/day' method on the Evosep One, which consists

of a 5.6 min gradient with 7.2 min total time between injections. Remarkably, in ten replicates, more than 1,100 proteins were identified on average without any identification transfer from libraries and with only 50 ng of injected cell lysate (**Fig. 6e,f**). 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.
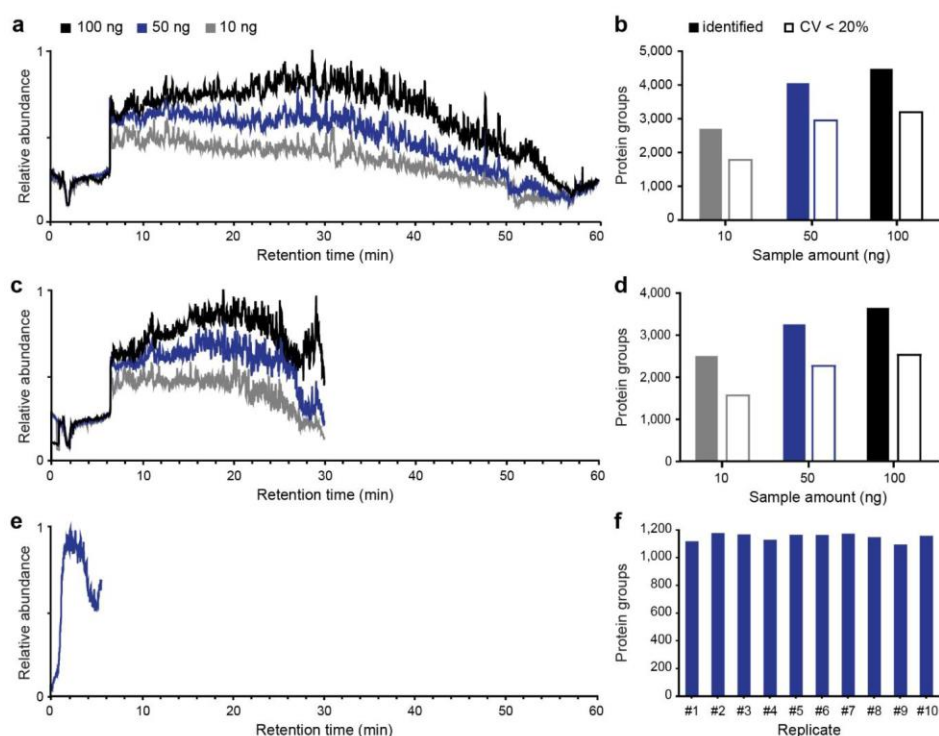
Figure 6 | Rapid and sensitive HeLa proteome measurements. (**a**) Total ion chromatograms of the 60 min gradient and three different sample amounts on column. (**b**) Average number of protein groups identified and quantified with a CV <20% in 60 min single runs (N=3). (**c**) Total ion chromatograms of the 30 min gradient and three different sample amounts on column. (**d**) Average number of protein groups identified and quantified with a CV <20% in 30 min single runs (N=3). (**e**) Total ion chromatogram of a 5.6 min gradient with 50 ng HeLa digest on column. (**f**) Number of protein groups identified in ten replicate injections with the 5.6 min gradient.

14

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. Since 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 travelling-wave ion mobility[46] and differential ion mobility[47], TIMS can
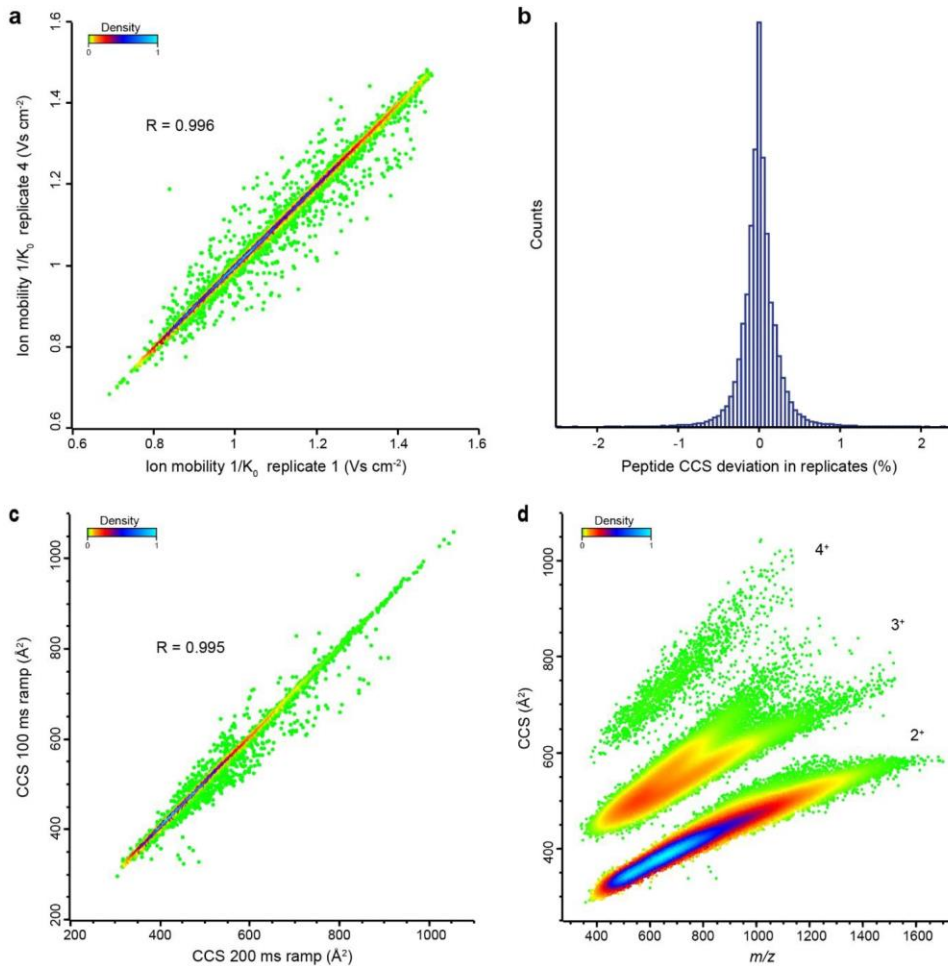
Figure 7 | Large-scale and high-precision CCS measurements. (a) Pearson correlation of peptide ion mobilities in two replicate injections of a HeLa digest (100 ms TIMS ramps). (b) Relative deviations of CCS values of all individual peptides from their mean of quadruplicate LC-MS runs. (c) Pearson correlation of measured CCS values in two injections of a HeLa digest with different TIMS ramp times (100 and 200 ms TIMS ramps). (d) Density distribution of over 100,000 CCS values from human tryptic peptide ions as a function of *m/z*. The main populations are annotated with their respective charge states.

15

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 ($1/K_0$; Vs cm^{-2}) of phosphazine derivatives from the literature[35], which can be converted to 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 21,673 $1/K_0$ values of commonly identified peptide ions in all runs with a median CV much smaller than 1% and a median pairwise correlation coefficient > 0.99 (**Fig. 7a**). Slight alterations in the gas flow can cause linear shifts in the measured mobility measurements. These can be readily taken into account by aligning the median values of all runs to the first replicate, which resulted in a median 0.1% absolute deviation of CCS values across replicates (**Fig. 7b**). 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 CCS measurements were also highly transferable across different TIMS ramp times (50 ms and 150 ms) as evident from a Pearson correlation coefficient of > 0.99 between them (**Fig. 7c**).

Having established precise CCS measurements in single runs, we next used loss-less high pH fractionation[33] to extend the scale of our data set. Measuring 24 fractions with 2 h gradients each resulted in 113,478 CCS values from 89,939 unique peptide sequences and about 9,000 protein groups. In the m/z vs. CCS plot, doubly, triply and higher-charged populations are clearly separated (**Fig. 7d**). Within each charge 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 CCS determination is still more than ten-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

16

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 over 6,000 protein groups in single runs from a human cancer cell line with minimal input material, and with high quantitative accuracy.

While 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 artefacts 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, culminated in the identification of about 2,500 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 particular 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 up 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.

Conflict of interest

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.

100

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. 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,** 1785–93 (2011).

6. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).

7. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8,** 61–80 (2015).

8. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312,** 212–7 (2006).

9. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12,** 483–490 (2008).

10. 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).

11. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H. Ion mobility-mass spectrometry. *J. Mass Spectrom.* **43,** 1–22 (2008).

12. Cumeras, R., Figueras, E., Davis, C. E., Baumbach, J. I. & Gràcia, I. Review on Ion Mobility Spectrometry. Part 2: hyphenated methods and effects of experimental parameters. *Analyst* (2014). doi:10.1039/c4an01101e

13. May, J. C. & McLean, J. A. Ion Mobility-Mass Spectrometry: Time-Dispersive Instrumentation. *Anal. Chem.* **87,** 1422–1436 (2015).

14. Eiceman, G. A., Karpas, Z. & Hill, H. H. J. *Ion Mobility Spectrometry.* (CRC Press, 2013).

15. 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,** 1213–6 (1998).

16. 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).

17. Ewing, M. A., Glover, M. S. & Clemmer, D. E. Hybrid Ion Mobility and Mass Spectrometry as a Separation Tool. *J. Chromatogr. A* 27–29 (2015). doi:10.1016/j.chroma.2015.10.080

18. 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.

18

Nat. Chem. **6,** 281–94 (2014).

19. Valentine, S. J. *et al.* Toward plasma proteome profiling with ion mobility-mass spectrometry. *J. Proteome Res.* **5,** 2977–84 (2006).

20. 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).

21. Geromanos, S. J., Hughes, C., Ciavarini, S., Vissers, J. P. C. & Langridge, J. I. Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Anal. Bioanal. Chem.* **404,** 1127–39 (2012).

22. Helm, D. *et al.* Ion Mobility Tandem Mass Spectrometry Enhances Performance of Bottom-up Proteomics. *Mol. Cell. Proteomics* **13,** 3709–15 (2014).

23. Distler, U. *et al.* Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **11,** 167–70 (2014).

24. 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).

25. 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).

26. 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).

27. 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).

28. 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).

29. Silveira, J. A, Ridgeway, M. E. & Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* **86,** 5624–7 (2014).

30. Ridgeway, M. E., Silveira, J. A., Meier, J. E. & Park, M. A. Microheterogeneity within conformational states of ubiquitin revealed by high resolution trapped ion mobility spectrometry. *Analyst* **140,** 6964–6972 (2015).

31. 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).

32. 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,** 319–24 (2014).

33. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* mcp.O116.065136 (2017). doi:10.1074/mcp.O116.065136

34. Bache, N. *et al.* A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *bioRxiv* (2018). doi:10.1101/323048
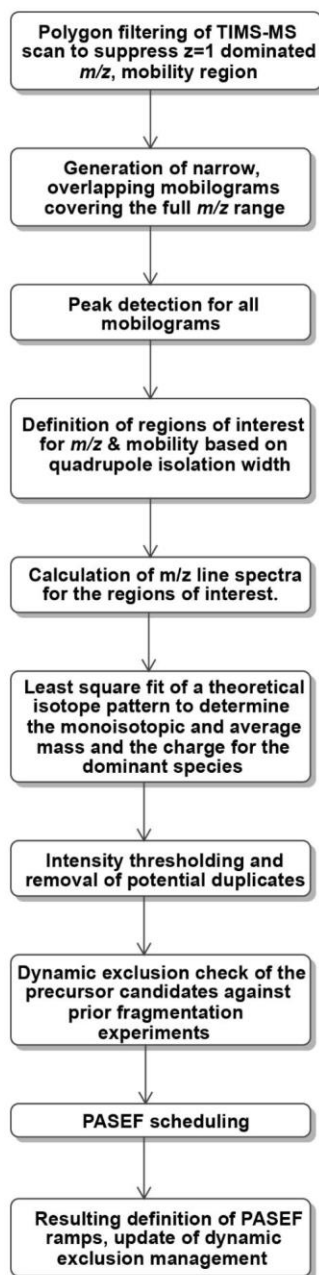
19

35. 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).

36. Mason, E. A. & McDaniel, E. W. *Transport Properties of Ions in Gases*. (John Wiley & Sons, Inc., 1988).

37. 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**, 1367–72 (2008).

38. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

39. Cox, J. *et al.* MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction. *Mol. Cell. Proteomics* M113.031591- (2014). doi:10.1074/mcp.M113.031591

40. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–40 (2016).

41. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008). at <http://www.r-project.org>

42. 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).

43. 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).

44. Hernandez, D. R. *et al.* Ion dynamics in a trapped ion mobility spectrometer. *Analyst* **139**, 1913–21 (2014).

45. Volpe P & Eremenko-Volpe T. Quantitative studies on cell proteins in suspension cultures. *Eur J Biochem* **12**, 195–200 (1970).

46. Shvartsburg, A. A. & Smith, R. D. Fundamentals of traveling wave ion mobility spectrometry. *Anal. Chem.* **80**, 9689–99 (2008).

47. Buryakov, I. A., Krylov, E. V., Nazarov, E. G. & Rasulev, U. K. 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 (1993).

48. Counterman, A. E. & Clemmer, D. E. Cis−Trans Signatures of Proline-Containing Tryptic Peptides in the Gas Phase. *Anal. Chem.* **74**, 1946–1951 (2002).

49. 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).

50. 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).

51. May, J. C., Morris, C. B. & McLean,

20

J. A. Ion mobility collision cross section compendium. *Anal. Chem.* **89,** 1032–1044 (2017).

52. 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).

53. Ross, P. L. *et al.* Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **3,** 1154–1169 (2004).

54. Virreira Winter, S. *et al.* EASI-tag enables accurate multiplexed and interference-free MS2- based proteome quantification. *bioRxiv* (2017). doi:10.1101/225649

21

104

SUPPLEMENTARY FIGURES

Polygon filtering of TIMS-MS scan to suppress z=1 dominated *m/z*, mobility region

↓

Generation of narrow, overlapping mobilograms covering the full *m/z* range

↓

Peak detection for all mobilograms

↓

Definition of regions of interest for *m/z* & mobility based on quadrupole isolation width

↓

Calculation of m/z line spectra for the regions of interest.

↓

Least square fit of a theoretical isotope pattern to determine the monoisotopic and average mass and the charge for the dominant species

↓

Intensity thresholding and removal of potential duplicates

↓

Dynamic exclusion check of the precursor candidates against prior fragmentation experiments

↓

PASEF scheduling

↓

Resulting definition of PASEF ramps, update of dynamic exclusion management

Supplementary Figure 1 | Schematic of the PASEF precursor ion selection scheme.
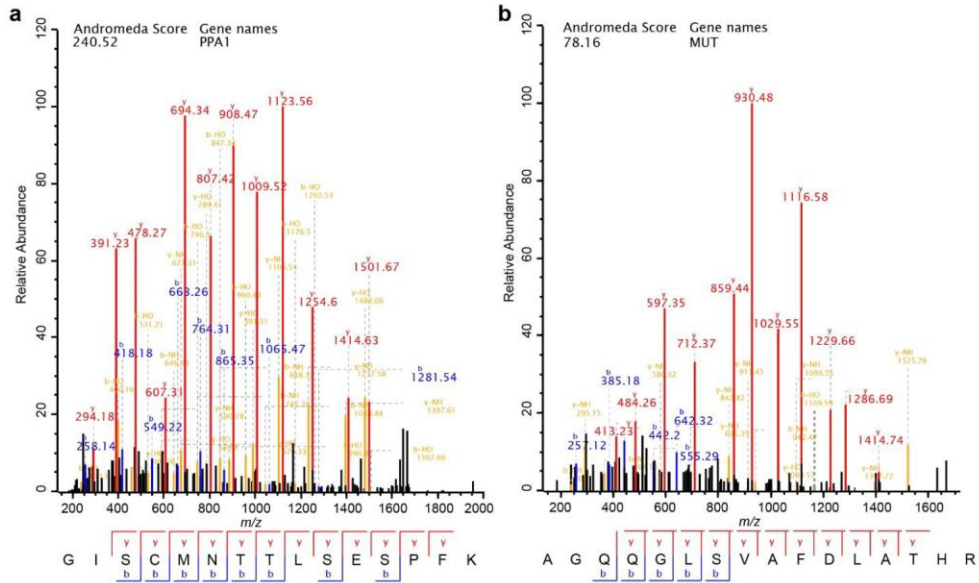
Supplementary Figure 2 | MaxQuant identification of co-eluting peptides of very similar mass, which would have been co-fragmented without TIMS.

23

106

IV.   Increasing the Dynamic Range of Orbitrap Mass Analysis

Article 5: BoxCar method enables single shot proteomics at a depth of 10,000 proteins in 100 minutes

Florian Meier[1], Philipp E. Geyer[1,2], Sebastian Virreira Winter[1], Juergen Cox[3] and Matthias Mann[1,2]

[1] Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany.
[2] NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.
[3] Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany.

Technological developments in MS-based proteomics over the last many years have always aimed to increase the depth of the proteome coverage – ideally with small starting amounts and in short time. Advances in MS instruments have almost entirely focused on the speed and sensitivity of peptide fragmentation, which led to a steep increase in the number of identified peptides and proteins per time unit. However, in most workflows, the quantitative information is derived from the full scan, which is therefore at least as important as MS/MS scans. Unfortunately, apart from increased resolution, this level has been largely abandoned. In particular, on the Orbitrap mass analyzer, which is by far the most widely used instrument in the field, the charge capacity has not changed since its introduction over ten years ago. In proteomic practice, we typically observe filling times below 1 ms, which is in stark mismatch with the scan time of over 100 ms for a high-resolution mass spectrum. This implies, that less than 1% of the available ions are used for mass analysis in full scans.

In the present manuscript, we break through this limitation by a novel scan mode termed BoxCar. Instead of filling all ions at once, BoxCar employs a serial filling scheme of multiple narrow isolation windows. This allows distributing the full charge capacity evenly across the entire mass range and filling each window separately to achieve the target number of ions. This resulted in over ten-fold increased filling times for low-abundance ions, thereby dramatically improving signal-to-noise ratios and dynamic range. To preserve quantitative accuracy, we devised a filling pattern that resembles the on/off shape of the mathematical boxcar function and assigned the full mass range to two

or three consecutive scans. We successfully benchmarked the quantitative accuracy in label-free as well as SILAC experiments.

A particularly challenging application of proteomics are clinical samples, such as human blood plasma samples. Their protein abundances typically span a huge dynamic range from overly abundant transporter proteins to extremely low-abundance regulatory proteins. Using BoxCar, we detected four-fold more features on the MS-level spanning an overall ten-fold larger dynamic range as compared with standard full scans.

To circumvent limitations implied by the finite sequencing speed and sensitivity at the MS/MS level, we combined BoxCar single runs with a peptide library and transferred identifications based on sub-ppm mass accuracies and within narrow retention time alignment windows. As a result of the increased spectral quality and dynamic range, this greatly increased the number of identified proteins in single runs of HeLa digest, to a level of coverage which was sufficient to derive copy numbers estimates and accurate complex stoichiometries. Of note, BoxCar also greatly alleviates the missing value problem and reproducibly quantified about 7,500 proteins with less than 5% missing values in 45 min HeLa single runs.

The excellent depth and reproducibility encouraged us to apply this strategy to a highly complex tissue sample of mouse cerebellum. In 100 min single runs, we found MS-level evidence for over 10,000 mouse proteins, broadly covering important biological functions and neural compartments. This highlights the performance of BoxCar and let us conclude that BoxCar essentially delivers the results of fractionation experiments, however, with minimal starting material and in single runs.

Since its publication, the method has gained much attention in the mass spectrometry field and we have already received countless requests from interested researchers. While demonstrated in the context of DDA, many other scan modes based on BoxCar are yet to be explored. Furthermore, the BoxCar method is generic and can readily be applied in other fields with similar dynamic range challenges. To facilitate dissemination, Christoph Wichmann implemented BoxCar in the *MaxQuant.Live* software which allows users to seamlessly integrate BoxCar in their workflows.

BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes

Florian Meier [1], Philipp E. Geyer[1,2], Sebastian Virreira Winter [1], Juergen Cox[3] and Matthias Mann [1,2]*

Great advances have been made in sensitivity and acquisition speed on the Orbitrap mass analyzer, enabling increasingly deep proteome coverage. However, these advances have been mainly limited to the MS2 level, whereas ion beam sampling for the MS1 scans remains extremely inefficient. Here we report a data-acquisition method, termed BoxCar, in which filling multiple narrow mass-to-charge segments increases the mean ion injection time more than tenfold as compared to that of a standard full scan. In 1-h analyses, the method provided MS1-level evidence for more than 90% of the proteome of a human cancer cell line that had previously been identified in 24 fractions, and it quantified more than 6,200 proteins in ten of ten replicates. In mouse brain tissue, we detected more than 10,000 proteins in only 100 min, and sensitivity extended into the low-attomolar range.

The proteome of a mammalian cell comprises expression products from > 10,000 protein-coding genes, with copy numbers that range from tens to over 1 million per cell, and body fluids (such as plasma) have an even larger dynamic range[1–3]. Mass spectrometry (MS)-based proteomics has become very powerful in elucidating this highly complex structural and functional entity[4,5]. However, further improvements in MS technology are needed to more fully capture the proteome in its entire complexity.

In shotgun proteomics, full scans (MS1) measure the mass-to-charge ratio (*m/z*) and signal intensity of peptides, whereas the function of MS2 scans is to determine fragmentation patterns for peptide identification in databases[6]. Innovations over the last decade have almost exclusively focused on the sensitivity and speed of peptide fragmentation by tandem mass spectrometry (MS2)[7,8], including the development of data-independent acquisition (DIA)[9,10] and peptide-centric analysis[11]. This is particularly the case for Orbitrap mass analyzers, which have become much faster but whose major bottleneck remains the charge capacity of the upstream ion storage device (C-trap) of only about 1 million charges[12,13]. With modern high-transmission ion sources and typical peptide amounts on column, this value is often reached in < 1 ms. This is < 1% of the required transient time for generating a high-resolution mass spectrum (128 ms), which implies that > 99% of the generated ions are not used for mass analysis. In contrast, when only a small proportion of the total ion current is admitted, high ion currents and high resolution can be directly used to increase proteome sampling depth[14–16].

The ion current in MS-based proteomics is generated by electrospraying hundreds of co-eluting peptides into the mass spectrometer, and in MS1 mode on the Orbitrap mass analyzer, ions from the full mass range are trapped together. However, as a consequence of the high dynamic range of the proteome and the different ionization properties of peptides, the total signal is concentrated within a few peaks rather than being evenly distributed across the full mass range. We thus reasoned that distributing the maximal charge capacity of 1 million evenly over multiple narrow *m/z* segments would limit

the proportion of highly abundant species in the C-trap and greatly increase ion injection (or 'filling') times for less abundant species. A similar benefit has been observed in DIA methods, in which the instrument cycles through *m/z* segments to acquire fragment ion spectra of all precursors in each segment. Here we aimed to improve the detection of intact precursor ions. Our strategy was built on the ability of the quadrupole–Orbitrap mass analyzer to be filled sequentially with different mass windows, which are then analyzed together in a single scan[17,18]. We termed this acquisition method 'BoxCar' and show that it drastically improves performance on the MS1 level.

Results

Development of the method. To balance the absolute ion count across the full mass spectrum, we decomposed it into narrow *m/z* segments, which were each allotted the same proportion of the total incoming ion current—for instance, 100,000 charges for each of ten segments. In the mass spectrometer, each segment was sequentially selected by the analytical quadrupole, its ions were combined in the C-trap, and the ions from all of the segments were mass-analyzed together in the Orbitrap (Fig. 1a). However, combining adjacent segments in a single scan causes artifacts at the respective edges and precludes accurate quantification. Instead, we allotted the full mass range to multiple, subsequent scans (three in Fig. 1b), each of which was defined by interspaced, rectangular BoxCar patterns for mass selection (Methods and Supplementary Protocol). The term BoxCar refers to the mathematical on–off shape of the transmission function. The superimposition of the BoxCar scans in an acquisition cycle reconstructs a full scan, with the difference being that the boxes with low ion signal benefit from a dramatic increase in fill time, whereas the ion proportion from the few boxes with very high ion current is restricted. In Fig. 1c,d, the summed ion injection time of the BoxCar scan (177 ms) dwarfed the ion injection time of the standard full scan (1 ms), and the average ion injection times of individual boxes were 22-fold higher than in the standard full scan. As a result, the low-abundance species with poor signal-to-noise

[1]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark. [3]Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany. *e-mail: mmann@biochem.mpg.de
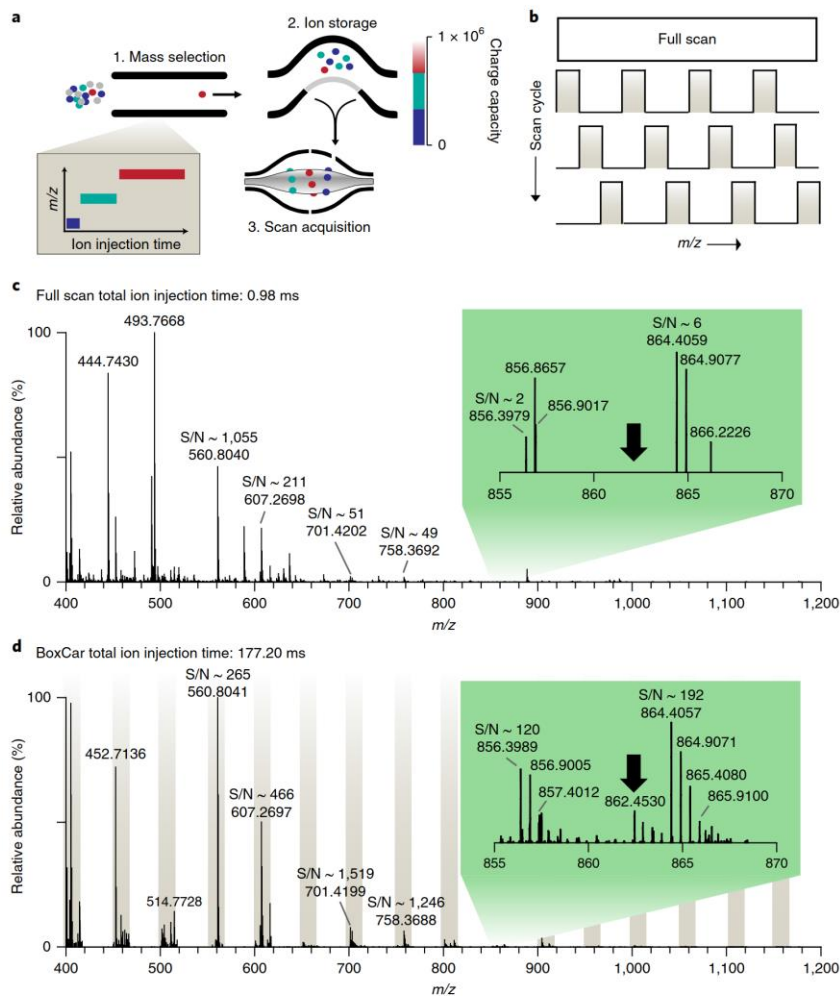
Fig. 1 | The BoxCar acquisition method. a, The high dynamic range of the incoming ion population is normalized by (1) sequentially filtering narrow m/z windows in the analytical quadrupole, (2) accumulating the ions in the C-trap and (3) analyzing the stored ion population in a single Orbitrap scan. **b**, The full mass range is covered by an interspaced, rectangular boxcar function for mass selection in subsequent scans. **c**, Representative example of a standard full scan in shotgun proteomics. **d**, The adjacent BoxCar scan composed of 16 narrow isolation windows ('boxes'). In **c** and **d**, the green insets highlight an apparently empty m/z region in the full scan that was filled for 16 ms in the BoxCar scan. Signal-to-noise ratios (S/N) are provided for comparison. The spectra were selected from a single LC–MS run of HeLa digest.

ratios (S/N) were now detected at 30- to 60-fold increased S/N. Furthermore, we observed an additional isotope pattern at nominal m/z 862, which had completely escaped detection in the full scan (Fig. 1c,d).

Next, we systematically investigated the effect of the acquisition parameters on the number of quantified peptide features in a single-shot experiment using a 'statistical design-of-experiment' (DoE) approach (Supplementary Fig. 1 and Supplementary Table 1). Small boxes have longer injection times, greater S/N, increased dynamic range and improved 'isolation' of very-high-abundance peaks. However, this needs to be balanced by the total available fill time, as well as the overall cycle time. To achieve cycle times around 1 s for the full acquisition cycle of one full scan, BoxCar scans and data-dependent MS2 scans, here we limited the number of BoxCar scans to two or three per acquisition cycle. For the same reason,

we limited the number of MS2 scans to five—with precursors selected from full scans. This yielded an average of eight data points per chromatographic peak with two BoxCar scans for the 45-min gradient. We further scaled box widths to be inversely proportional to the m/z density of tryptic peptides (Supplementary Table 2).

Having established sensible parameters for BoxCar, we did a first comparison to the established shotgun method with standard full scans. In analyses using a HeLa cell digest, we found that BoxCar doubled the number of detected multiply charged isotope patterns and increased the dynamic range by nearly one order of magnitude (Supplementary Fig. 2 and Supplementary Table 3).

Reproducible and accurate protein quantification. Ion transmission through the mass-selecting quadrupole inevitably deviates from an ideal rectangular shape, particularly at the edges of the
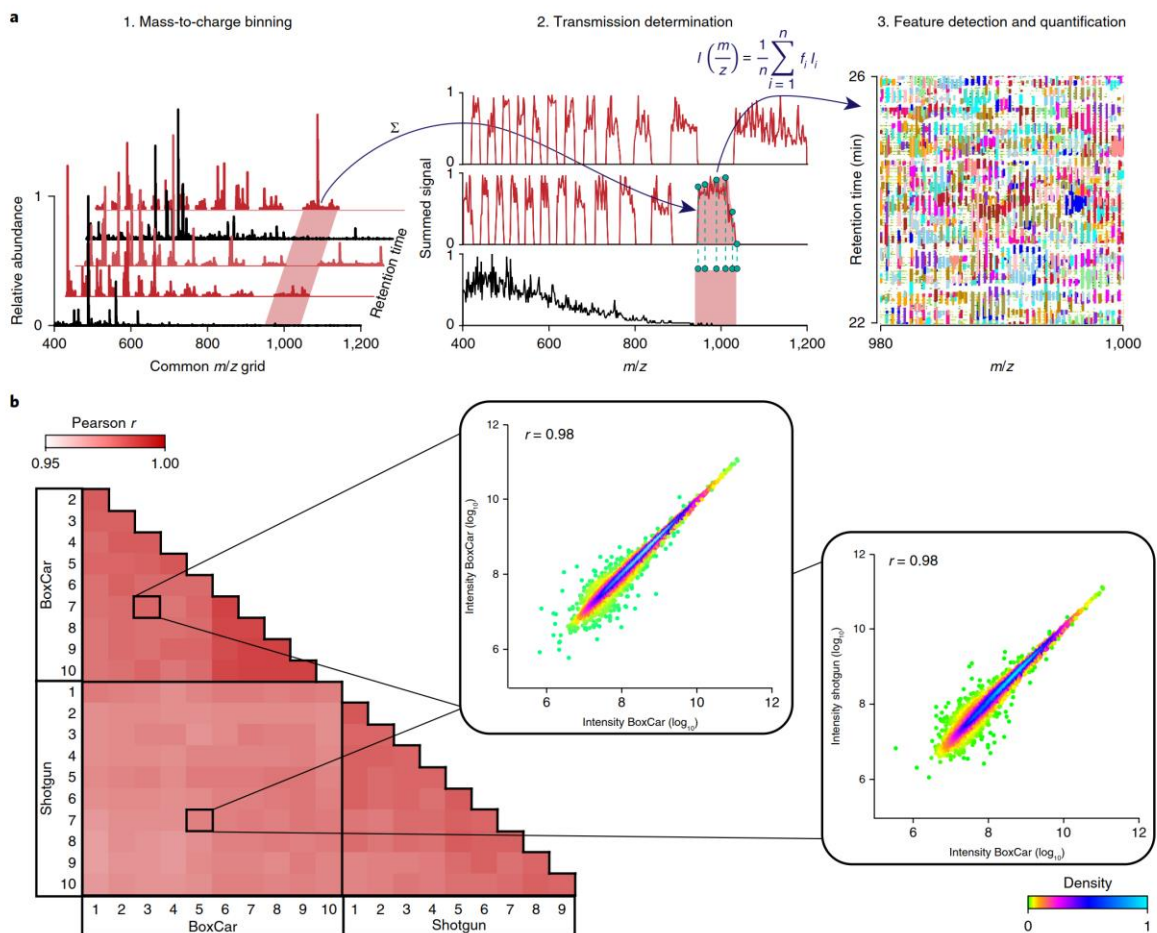
110

Fig. 2 | Label-free protein quantification with BoxCar scans. a, MaxQuant assembles the full scan (black traces) and BoxCar scans (red traces) of an acquisition cycle to a single high-dynamic-range scan. First, all spectra are transformed to a common high-resolution m/z grid, and the signals from each scan (here one full scan and two BoxCar scans) are integrated over the entire LC elution time (step 1). From the integrated signals, the shape of the quadrupole transmission function for each BoxCar scan is globally determined by a pointwise comparison to the full scan (step 2). The resulting relative transmission factors for each m/z bin in each BoxCar scan are used as weights for calculating the average signal intensity from the full scan and the BoxCar scans. These hybrid spectra are taken as a replacement for standard full scans in all subsequent processing steps without further adjustments (step 3). See the main text and Methods for further details. **b**, Pearson correlation analysis of label-free protein intensities in ten replicate 45-min runs for HeLa samples using BoxCar or a standard shotgun method ($N = 4,292$ proteins). The median correlation coefficients were 0.98 for BoxCar and the shotgun replicates, and 0.98 for the cross-correlated protein intensities. Density plots illustrate the protein intensity correlation of two representative pairs.

isolation window. To alleviate these 'edge effects', we incorporated a small overlap of neighboring boxes on the m/z axis. Further correction for non-uniform ion transmission efficiencies was crucial for the detection and accurate quantification of isotope patterns, in particular those that reached across the borders of a single box. To achieve this, we adapted the post-processing workflow in MaxQuant[19] (Fig. 2a and Methods). Our algorithm first assembled all BoxCar scans and the preceding full scans on a common high-resolution m/z grid. Next, all of the ion signals from these scans were integrated over the entire liquid chromatography (LC) run time to determine the empirical ion transmission function by comparison to the full scan. The resulting relative ion transmission factors for each m/z bin were then applied to correct the relative ion abundances in each individual BoxCar scan. Finally, for each acquisition cycle, a single high-dynamic-range spectrum was calculated

as the weighted average of the standard full scan and the intensity-corrected BoxCar scans.

Using the full information over the entire elution time makes our algorithm robust against interscan and time-dependent fluctuations in the ion current. Further downstream analysis in MaxQuant remained unchanged, while greatly benefiting from the increased spectral quality and dynamic range.

To test the quantitative accuracy of the resulting hybrid spectra, we first analyzed a mixture of heavy and light stable isotope-labeled HeLa lysate, in which BoxCar accurately quantified the pipetting ratio of 1:3 throughout the entire m/z range (Supplementary Fig. 3). To further provide a benchmark for the BoxCar method for label-free quantification, we assessed the quantitative reproducibility and accuracy between ten technical replicate injections of a whole HeLa proteome digest and compared them to the data
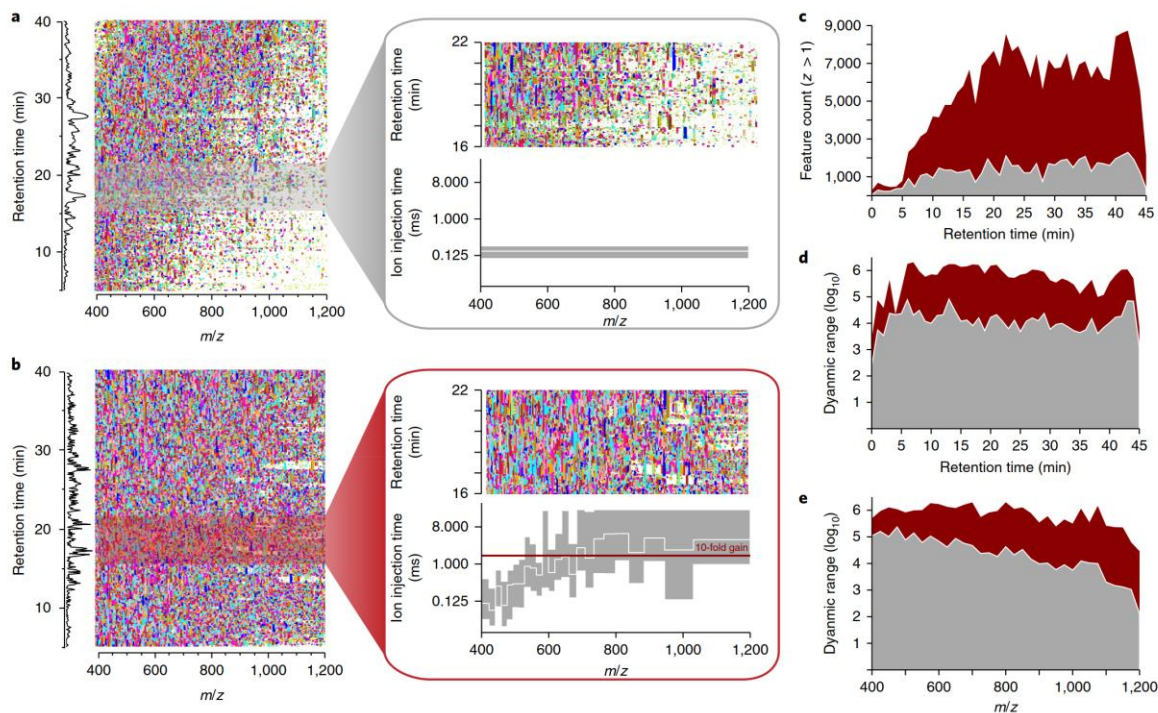
111

Fig. 3 | High-dynamic-range analysis of a human plasma sample in single runs. a,b, Isotope patterns detected in the m/z-retention time plane are color-coded for standard full scans (**a**) and BoxCar scans (**b**). Median ion injection times (full scans, $N=604$; BoxCar scans, $N=347$) are illustrated in the insets as a function of m/z (white line). The top and bottom of the gray boxes represent the 75th and 25th percentiles, respectively. **c,** Detected multiply charged features per minute over the entire gradient in a single run with the standard (gray) and BoxCar (red) method. **d,e,** Dynamic range of the detected features as a function of retention time (**d**) and m/z (**e**). Full scans, $N=59,491$ features; BoxCar scans, $N=249,162$ features.

obtained from our standard method (Fig. 2b). In total, we identified 4,302 protein groups, of which 4,156 (on average) were quantified per 45-min run. Median coefficients of variation (CVs) were 17% and 23% for the non-normalized peptide intensities with BoxCar and the standard method, respectively, and 8% on the protein level after label-free normalization for both. Pairwise comparisons of all of the replicates revealed excellent quantitative reproducibility of the BoxCar method with a median Pearson correlation of 0.98 on the protein level, which was virtually identical to the overall correlation of standard full scan quantifications to each other ($r=0.98$). Similarly, BoxCar–BoxCar correlations were very similar to BoxCar–full scan correlations, both at the peptide and protein levels, further confirming the absence of any systematic bias (Fig. 2b, insets and Supplementary Fig. 4). In a two-proteome experiment (HeLa and *Escherichia coli* lysates mixed in 1:2 and 1:12 ratios), BoxCar performed particularly well for the low-abundance *E. coli* proteins that were quantified either with low accuracy or had missing values with the standard method. This indicated an approximately tenfold-increased dynamic range (Supplementary Fig. 5).

Extended dynamic range in body fluids. Analysis of proteomes from body fluids is an important clinical application of proteomics; however, it can also be very challenging owing to their extreme dynamic range, which exceeds ten orders of magnitude for plasma[20]. The finite intrascan dynamic range severely limits the ability to detect low-abundance proteins in standard single-run measurements[21,22]. In trapping instruments, including the Orbitrap analyzer,

this is mainly due to peptides from highly abundant proteins, such as serum albumin. Figure 3a illustrates the detected peptide features in standard full scans and the total ion chromatogram, which was dominated by individual abundant peptides for much of the elution time. While these ions eluted, the mass spectrometer was essentially blind to all other peptides, which resulted in sparse regions in the m/z–retention-time plane. In contrast, the BoxCar method effectively limited the proportion of abundant species by confining them to narrow mass windows with an equalized number of charges (Fig. 3b).

In the 16- to 22-min elution window, for instance, injection times for the standard MS spectrum were only 0.15 ms, which corresponded to a duty cycle of 0.12% for a 128-ms transient. In the BoxCar measurement, ion injection time scaled with the expected ion current in each m/z window up to the defined maximum of 21 ms. This 140-fold increased ion fill time revealed previously hidden, co-eluting isotope features and thus potential peptides (Fig. 3a,b). Gains were particularly high in less-populated m/z ranges; however, the median gain across the full mass range was more than tenfold (Fig. 3b).

In total, we detected 249,162 multiply charged isotope patterns with BoxCar, which was fourfold more than the 59,491 observed by the standard method (Fig. 3c). Patterns identified in both methods exhibited more isotope peaks and were detected for longer elution times (Supplementary Fig. 6). The dynamic range improvement was uniform over the retention time but increased with m/z, for an overall tenfold gain, which was consistent with the gain in ion fill time (Fig. 3d,e).
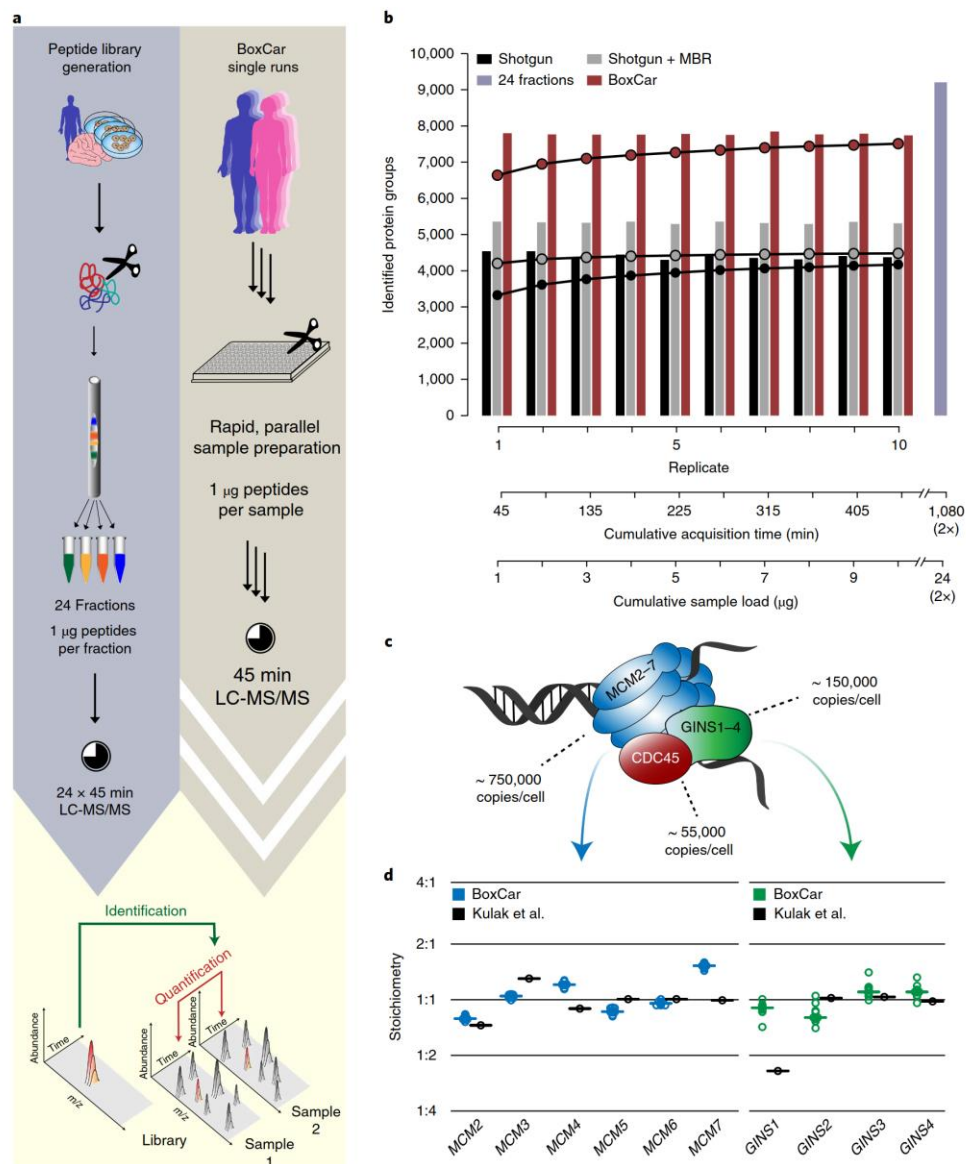
112

Fig. 4 | Library approach for deep proteome coverage in BoxCar single runs. a, Schematic workflow for peptide library generation from fractionated protein digests (left) and high-throughput proteome quantification in BoxCar single runs (right). **b**, Comparison of protein identifications in ten 45-min analyses of a HeLa digest with a standard shotgun method, with 'matching between runs' (MBR), and with the BoxCar approach, using a library. Line plots indicate the cumulative number of proteins identified with two or more peptides in at least one replicate. Cumulative peptide load and measuring times (bottom axes) show that BoxCar minimizes the requirements for acquisition time and starting material, while proteome coverage reaches the level of 24 high-pH reverse-phase fractions. **c**, Representation of the CMG complex during DNA replication together with copy numbers estimated from BoxCar single runs. **d**, Stoichiometry of the MCM and GINS complexes reconstructed from relative copy numbers in BoxCar single runs and in six strong cation exchange fractions[30]. Horizontal bars indicate medians ($N = 10$ replicate injections).

Deep proteome quantification with BoxCar. For very deep proteome coverage, peptide prefractionation is generally used, as it increases the overall amount of material that can be loaded and the time available for peptide sequencing[8,23]. Given the greatly improved dynamic range and number of detected features of the BoxCar method, we asked whether comprehensive proteome character-

ization could be reached by combining single BoxCar runs with a peptide library[24,25]. In such an approach, peptide identifications would be transferred from the library using the 'match between runs' feature of MaxQuant[26], whereas the quantitative information was provided by BoxCar single runs (Fig. 4a and Supplementary Note 1). Unlike a standard shotgun approach, the library strategy is
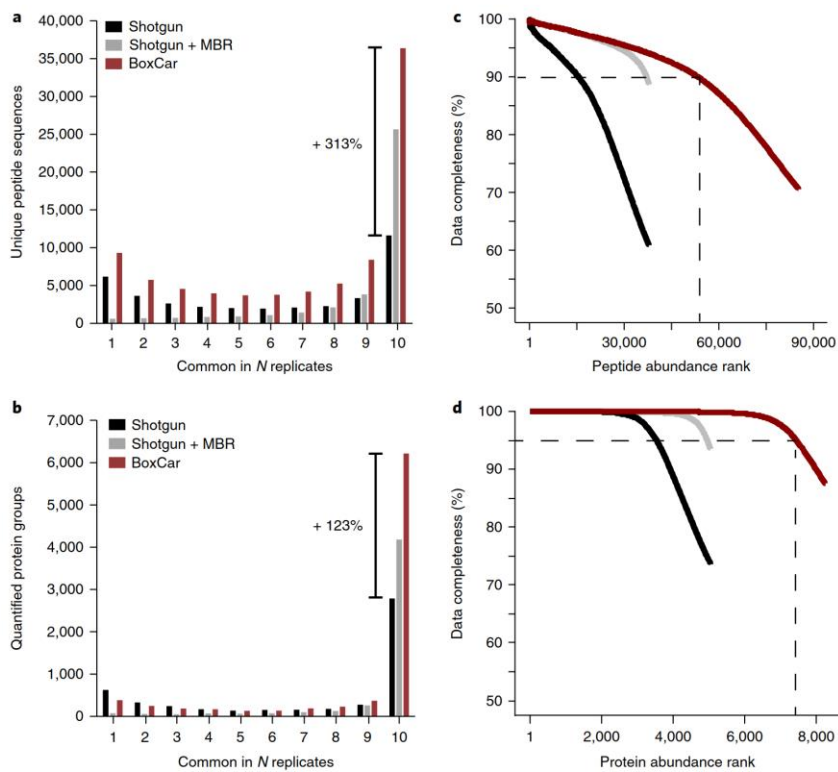
113

Fig. 5 | Assessment of missing value rates in ten replicate 45-min analyses of HeLa digest. a,b, Number of unique peptide sequences (**a**) and protein groups (**b**) quantified in specific numbers of replicates (*N*) with our standard shotgun method, using MBR, and with BoxCar in conjunction with a matching library. **c,d,** Completeness of the peptide (**c**) and protein group (**d**) quantification matrix as a function of descending peptide or protein abundance in all three methods (*N* = 10 replicate injections for each).

not limited by the MS2 scan speed of the mass spectrometer, and, with BoxCar, a much larger proportion of the incoming ion current can be used for quantification.

To test this concept, we acquired a library of HeLa peptides from 24 high-pH reverse-phase fractions[27], which comprised 128,720 sequence-unique peptides from 9,210 protein groups. Matching was facilitated by sub-parts per million absolute average mass accuracies, as well as by median absolute elution time differences of 1.7 s after alignment (as compared to median peak widths of 11.3 s). On average, we assigned 80,714 isotope patterns to peptide sequences in 45-min BoxCar runs (Supplementary Table 3), which corresponded to a maximum of 50 peptide features per second (Supplementary Fig. 7). At typical MS2 identification rates, this would otherwise have required a sequencing rate of more than 60 Hz. Similarly, matching the library to standard shotgun runs was previously limited by the dynamic range of the full scan, which is expanded up to one order of magnitude in BoxCar scans throughout the LC gradient and mass range (Supplementary Fig. 7).

Each BoxCar run identified (on average) 60,228 unique peptide sequences and 7,775 protein groups from 1 µg of peptides (Fig. 4b). Combining the ten BoxCar single runs retrieved two-thirds of all unique peptide sequences and 93% of all protein groups present in the library, with 7,508 proteins identified by two or more peptides in at least one replicate. We reasoned that this level of coverage would be sufficient to estimate the copy numbers and stoichiometries of most of the protein complexes in the cell. As an example, we studied the six-membered minichromosome maintenance (MCM)

complex, a key factor in DNA replication[28] (Fig. 4c). In eukaryotes, MCM associates with the four-membered GINS complex and the initiation factor CDC45, also known as the CMG complex[29]. From our 45-min BoxCar single runs, we estimated an abundance of ~750,000, 150,000 and 55,000 copies per cell for MCM, GINS and CDC45, respectively, in agreement with our previous results from fractionated samples[30]. Moreover, all members of the MCM and GINS complexes were quantified with the expected 1:1 stoichiometry (Fig. 4d).

Alleviating the 'missing value' problem. The semistochastic nature of precursor selection in shotgun proteomics results in not every peptide being quantified in every run; this is often referred to as the 'missing value' problem[31,32]. The complete quantification matrix of ten HeLa runs analyzed with a standard shotgun method comprised 37,869 unique peptide sequences and 5,050 protein groups, of which only 11,626 peptides and 2,789 proteins were quantified in each of the 45-min runs (Fig. 5a,b). Using the 'matching between runs' function in MaxQuant greatly mitigated this problem by transferring peptide identifications between experiments, and this led to complete quantification of 25,653 peptides and 4,180 protein groups in ten replicates. BoxCar in conjunction with the peptide library achieved even more consistent profiles across all of the replicates, providing quantification of ~70% of all peptides and > 95% of all proteins from the entire shotgun experiment in every replicate. In total, BoxCar quantified 7,222 proteins per 45-min run, with a median CV of 10%, and it more than doubled the proportion of
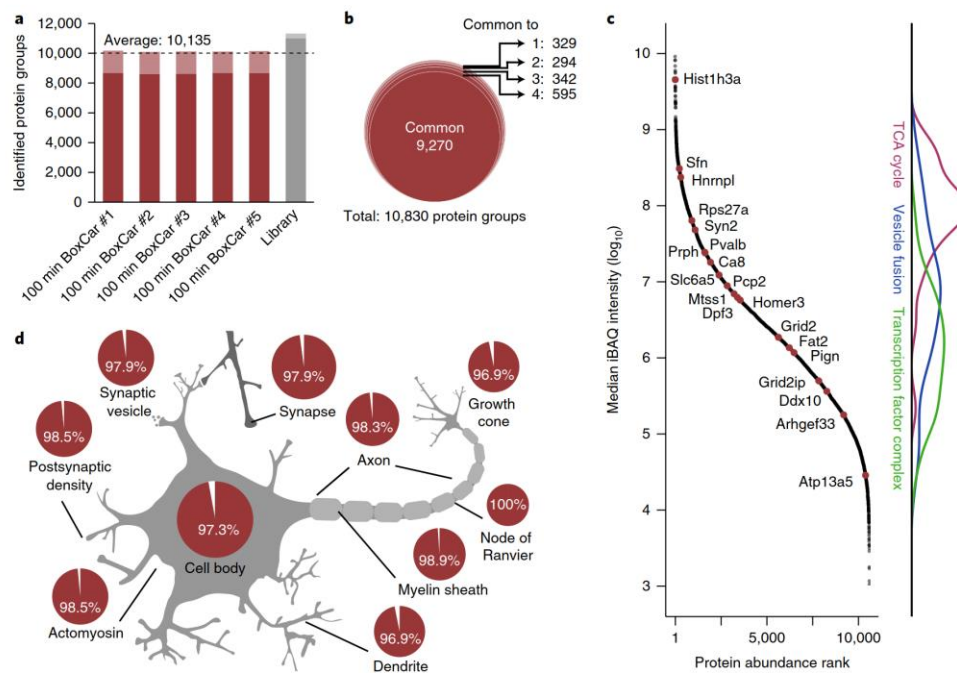
114

Fig. 6 | In-depth proteome analysis of rodent cerebellum in 100-min BoxCar single runs. a, Number of proteins identified in each of five BoxCar replicates and in a high-pH fractionation experiment with 24 fractions (library column). Light bars indicate proteins identified by a single peptide. **b,** A total of 10,850 protein groups were identified in 100-min BoxCar single runs, and 97% of the proteins were identified in at least two of five replicates. **c,** Ranking of cerebellum proteins by their median, normalized intensity in BoxCar single runs (N = 5). Differentially expressed brain proteins[36] are highlighted in red, and the density distribution of key metabolic and neuronal Gene Ontology (GO) annotations is projected onto the right axis. **d,** Relative coverage of GO cellular compartments in BoxCar single runs (N = 5) as compared to the full library. iBAQ, intensity-based absolute quantification.

peptides and proteins quantified in ten of ten runs to 36,377 and 6,216, respectively.

Next, we calculated the completeness of the data matrices as a function of the peptide and protein abundance rank (Fig. 5c,d). For the standard shotgun method, both curves declined fast and reached 60% and 75% at the full depth. With matching between runs the curve was much shallower, and only 11% of the peptide and 6% of the protein data points were missing. The BoxCar workflow outperformed both methods in terms of proteome coverage and reproducibility. From the 54,000 most abundant peptides, only 10% of the values were missing, and the protein data completeness did not drop < 95% until a depth of 7,450 proteins. Even at the full depth of 8,253 protein groups, the fraction of missing data points was as low as 12.5%. Our results compared favorably to reported missing value rates in DIA methods, which are furthermore reached at lower proteome coverage[33–35].

Ten thousand protein groups in 100 min. The excellent depth and reproducibility in single-shot analyses of cell lines encouraged us to investigate the performance of BoxCar in a highly complex tissue sample from mammalian brain[36]. First, we used high-pH reverse-phase fractionation to generate a deep peptide library from fresh-frozen mouse cerebellum, comprising 154,305 unique peptide sequences from 11,328 protein groups, with a median sequence coverage of 30%. We then analyzed quintuplicates of unfractionated cerebellum digest in 100-min gradients. BoxCar yielded > 100,000 identified features per single run, 60% more than the analog shotgun–library approach and spanning more than one additional order

of magnitude (Supplementary Fig. 7). Overall, the BoxCar-library-based workflow identified (on average) 72,896 unique peptide sequences and > 10,000 protein groups in each replicate, which was equivalent to the identification of 100 proteins/min over the entire gradient (Fig. 6a). A total of 9,270 protein groups were detected in all five replicates and 10,207 were detected in at least three of the five replicates (Fig. 6b). We found evidence for 65% of all unique peptide sequences and 92% of all library proteins in our 100-min BoxCar runs, with 9,583 proteins being identified with two or more peptides in at least one replicate. Protein abundances spanned more than six orders of magnitude, and we estimated the amount of the 1,200 least abundant proteins to be < 10 attomoles (Fig. 6c and Supplementary Fig. 8). Label-free protein quantification was highly reproducible with a median CV of 9.9% (20% for non-normalized peptide and summed peptide intensities).

Coverage of biological functions and neural compartments in the single-run measurements, in comparison with the cerebellum library, was almost complete for structural proteins, for instance, 100% of all of the proteins that constitute the node of Ranvier and 98.5% of those associated with actomyosin (Fig. 6d). Similarly, proteins involved in neuronal activity were broadly covered, including 97.9% and 98.5% of those annotated as 'synaptic vesicles' and 'postsynaptic density', respectively.

Bioinformatics analysis indicated that high- and medium-abundance proteins were associated with key metabolic functions, such as the tricarboxylic acid (TCA) cycle, and important neuronal functions, such as vesicle fusion (Fig. 6c). Members of transcription factor complexes were among the lower-abundance proteins, often

with estimated copy numbers of < 10,000 per cell (Supplementary Fig. 8). An array of proteins that we had previously reported to be specifically enriched in cerebellum as compared with other brain regions[36] were distributed across the entire abundance range. These included regulatory proteins (Syn2 and Sfn) and scaffolding proteins (Grid2ip and Homer3), as well as proteins that are crucial for cerebellar development (Ca8 and Fat2).

Discussion

A key limitation of trapping-based mass spectrometers—including the Orbitrap—is the limited charge capacity of the ion trap[13], which excludes the majority of ions from MS1-level analysis. Our BoxCar acquisition method overcomes this limitation by using the incoming ion current far more efficiently, by redistributing the proportion of ions allocated across the mass range. The BoxCar method increases the MS1-level dynamic range by one order of magnitude on average and even more in the case of very high dynamic range proteomes, such as body fluids. BoxCar does not require any changes to the instrument hardware, and the post-processing algorithms are fully integrated in MaxQuant. Insertion of BoxCar scans does lengthen the acquisition cycle time, but we were able to balance this by reducing the number of MS2 scans and transferring identifications from a library. Ongoing improvements in resolving power[37] will allow BoxCar scans to be inserted into an acquisition cycle with less effect on its duration.

Library-based workflows are well established for DIA[11,38], and the increased spectral quality and extended dynamic range of the BoxCar method make this strategy also very attractive on the MS1 level, even though stringent control for false matches remains challenging and currently relies on reproducible chromatography, as well as on project-specific libraries. In a human cancer cell line, we consistently quantified almost the entire contents of the protein library in only 1 h, using 4% of the sample amount used for library generation. Likewise, the 10,000 protein groups detected in 100-min gradients of the cerebellum proteome broadly covered the relevant biological functions from minimal input material. Thus, the combination of BoxCar scans with a peptide library essentially delivers the results of a pre-fractionation approach without the attendant consumption of sample and measurement time.

Although demonstrated here in the context of proteomics, the BoxCar method is generic and can seamlessly be applied in other fields with similar dynamic range challenges, for instance metabolomics. Note that all of the advantages of BoxCar follow from its greater dynamic range at the MS1 level. Therefore, they are not limited to shotgun proteomics but should also benefit any method that relies on accurate and repeatable picking of peptide precursors, such as targeting and isobaric tagging[39,40]. Furthermore, MS2-level DIA methods would also profit from BoxCar by much improved precursor information.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41592-018-0003-5.

References

1. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537**, 328–338 (2016).
2. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
3. Schwenk, J. M. et al. The Human Plasma Proteome draft of 2017: building on the Human Plasma PeptideAtlas from mass spectrometry and complementary assays. *J. Proteome Res.* **16**, 4299–4310 (2017).
4. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
5. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The coming age of complete, accurate and ubiquitous proteomes. *Mol. Cell* **49**, 583–590 (2013).
6. Link, A. J. et al. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
7. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: toward an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).
8. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24**, 11–17 (2015).
9. 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).
10. 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).
11. Ting, Y. S. et al. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol. Cell. Proteomics* **14**, 2301–2307 (2015).
12. Eliuk, S. & Makarov, A. Evolution of Orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem.* **8**, 61–80 (2015).
13. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **85**, 5288–5296 (2013).
14. Yi, E. C. et al. Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23**, 3205–3216 (2002).
15. 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).
16. Kelstrup, C. D. et al. Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
17. Michalski, A. et al. Mass-spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, M111.011015 (2011).
18. Egertson, J. D. et al. Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* **10**, 744–746 (2013).
19. 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).
20. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13**, 942 (2017).
21. Geyer, P. E. et al. Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2**, 185–195 (2016).
22. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786 (2015).
23. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599 (2017).
24. Pasa-Tolić, L., Masselon, C., Barry, R. C., Shen, Y. & Smith, R. D. Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* **37**, 621–624 (2004).
25. Zhang, B., Käll, L. & Zubarev, R. A. DeMix-Q: quantification-centered data processing workflow. *Mol. Cell. Proteomics* **15**, 1467–1478 (2016).
26. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of 11 common cell lines reveals ubiquitous, but varying, expression of most proteins. *Mol. Cell. Proteomics* **11**, M111.014050 (2012).
27. 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).
28. Tye, B. K. MCM proteins in DNA replication. *Annu. Rev. Biochem.* **68**, 649–686 (1999).
29. Moyer, S. E., Lewis, P. W. & Botchan, M. R. Isolation of the Cdc45–Mcm2–7–GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc. Natl. Acad. Sci. USA* **103**, 10236–10241 (2006).
30. 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).
31. 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).
32. Röst, H. L., Malmström, L. & Aebersold, R. Reproducible quantitative proteotype data matrices for systems biology. *Mol. Biol. Cell* **26**, 3926–3931 (2015).
33. Röst, H. L. et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13**, 777–783 (2016).
34. Navarro, P. et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).

35. Collins, B. C. et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH–mass spectrometry. *Nat. Commun.* **8**, 291 (2017).

36. Sharma, K. et al. Cell-type- and brain-region-resolved mouse brain proteome. *Nat. Neurosci.* **18**, 1819–1831 (2015).

37. 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).

38. 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).

39. Gallien, S., Kim, S. Y. & Domon, B. Large-scale targeted proteomics using internal standard triggered–parallel reaction monitoring (IS-PRM). *Mol. Cell. Proteomics* **14**, 1630–1644 (2015).

40. Erickson, B. K. et al. A strategy to combine sample multiplexing with targeted proteomics assays for high-throughput protein signature characterization. *Mol. Cell* **65**, 361–370 (2017).

Acknowledgements

Author contributions

F.M. and M.M. conceptualized the method; P.E.G. contributed to the development of the post-processing algorithms; J.C. designed and implemented the post-processing algorithms; F.M. and P.E.G. performed the experiments; F.M., P.E.G., S.V.W., J.C. and M.M. analyzed the data; and F.M. and M.M. wrote the manuscript with input from all of the authors.

Competing interests

F.M., J.C. and M.M. are inventors on a patent application covering the method described herein (applicant: Max-Planck-Gesellschaft zur Förderung der Wissenschaften; application number: PCT/EP2018/051290).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0003-5.

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

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

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

117

Methods

Cell culture. Human malignant epithelial cells (HeLa S3, ATCC) were cultured in Dulbecco's modified Eagle's medium (DMEM), fortified with 10% fetal bovine serum, 20 mM glutamine and 1% penicillin–streptomycin (all from PAA Laboratories, Germany). For the 'stable isotope labeling with amino acids in cell culture' (SILAC) experiment, cells were cultured in arginine- and lysine-free DMEM, and the medium was fortified with equal concentrations of either natural L-arginine and L-lysine (light) or their stable-isotope-labeled counterparts (Arg10, [$^{13}C_6{}^{15}N_4$]arginine; Lys8, [$^{13}C_6{}^{15}N_2$]lysine; both from Cambridge Isotope Laboratories, USA) following our standard protocol[41]. *E. coli* (strain: XL1 blue) was grown in LB medium at 37 °C until log phase and harvested at OD 0.5 ($\lambda = 600$ nm). Cells were collected by centrifugation at 200 *g* for 10 min and washed once with cold phosphate-buffered saline (PBS) before re-centrifugation. Cell pellets were flash-frozen in liquid nitrogen and stored at –80 °C until needed.

HeLa and *E. coli* sample preparation. HeLa cells were lysed, reduced and alkylated in SDC buffer with chloroacetamide (PreOmics GmbH, Germany) as previously described[30]. Briefly, the samples were boiled at 95 °C for 10 min, followed by 15 min of sonication at maximum power (Bioruptor, Diagenode, Belgium) to disrupt cells and shear nucleic acids. Protein digestion was performed overnight at 37 °C by adding equal amounts of the proteolytic enzymes LysC and trypsin in a 1:100 (wt/wt) ratio to the sample. The reaction was stopped with five volumes of isopropanol containing 1% trifluoroacetic acid (TFA).

Mouse cerebellum sample preparation. All mouse experiments were performed according to the institutional regulations of the Max Planck Institute of Biochemistry and approved by the government agencies of Upper Bavaria. The cerebellum was dissected from an individual male mouse (strain: C57BL/6) at postnatal day 5 and snap-frozen in liquid nitrogen. The frozen cerebellum was ground using a mortar and pestle and suspended in 100 μl SDC reduction and alkylation buffer (PreOmics GmbH). Further sample processing and enzymatic digestion was performed as described above for HeLa and *E. coli* samples.

Preparation of human plasma samples. Blood sampling complied with all relevant ethical regulations and was approved by the ethics committee of the Max Planck Society (24th February 2015, Heidelberg, Germany). Donors provided written informed consent. The human plasma samples from a single healthy donor was prepared as previously reported[21]. In brief, SDC reduction and alkylation buffer was added to blood plasma in 25-fold excess (vol/vol), and the mixture was boiled at 95 °C for 10 min. The samples were digested for 1 h at 37 °C with LysC and trypsin in a 1:100 ratio (wt/wt) before quenching the digestion reaction with isopropanol containing 1% TFA.

Peptide purification for mass spectrometry. Following our previously published standard protocol[30], acidified protein digests were loaded onto a styrenedivinylbenzene reversed-phase sulfonate (SDB-RPS) sorbent. Following two consecutive purification and desalting steps with isopropanol containing 1% TFA and double-distilled water containing 0.1% TFA, the peptides were eluted with 80% acetonitrile (ACN) containing 1% ammonia. The eluate was vacuum-centrifuged to dryness (SpeedVac, Eppendorf, Germany) and reconstituted in double-distilled water with 2% ACN and 0.1% TFA for single-run analyses or reversed-phase chromatography at basic pH.

High-pH reversed-phase fractionation. Peptide pre-fractionation was performed on a fully automated 'spider fractionator' (PreOmics GmbH, Germany) coupled online to an EASY-nLC1000 chromatography system (Thermo Fisher Scientific, Germany). A detailed description of the setup is available in ref. [27]. Approximately 60 μg of purified peptides was loaded onto a 30-cm reversed-phase column with an inner diameter of 250 μm and separated at a flow rate of 2 μl/min with a binary buffer system (buffers A and B, Article No. P.O. 00009) at pH 10 (all from PreOmics GmbH). The gradient started from 3% buffer B and was increased stepwise to 30% buffer B within 45 min and to 60% within 17 min; this was then further increased to 95% buffer B within 5 min, held constant for 3 min and then decreased to 3% buffer B within 10 min for re-equilibration. For each sample, 24 fractions were automatically concatenated and collected into 0.2-ml tubes by shifting the rotor valve every 90 s. The fractions were vacuum-centrifuged to dryness and subjected to LC-MS/MS analysis after re-constitution in double-distilled water containing 2% ACN and 0.1% TFA without any further purification.

Liquid chromatography and mass spectrometry (LC-MS). Nanoflow reversed-phase LC was performed on an EASY-nLC 1000 or EASY-nLC 1200 ultra-high-pressure system coupled online to a Q Exactive HF mass spectrometer equipped with a nano-electrospray ion source (all from Thermo Fisher Scientific). Approximately 1–2 μg purified peptides was loaded on a 40-cm column (75 μm inner diameter) with a pulled emitter, which was packed in-house with ReproSil-Pur C_{18}-AQ 1.9-μm beads (Dr. Maisch GmbH, Germany). To reduce the backpressure at high flow rates and to enhance separation efficiency, the column compartment was kept at 60 °C. For 45-min gradients, mobile phases A and B were water with 0.1% formic acid (vol/vol) and 60/40/0.1% ACN/water/formic

acid (vol/vol/vol). Peptides were separated at a constant flow rate of 450 nl/min with a linear gradient of 3–40% mobile phase B within 35 min, followed by a linear increase from 40–90% mobile phase B within 8 min and a 2-min plateau before re-equilibration. For 100-min gradients, mobile phase A was as described above, and mobile phase B was 80/20/0.1% ACN/water/formic acid (vol/vol/vol). Peptide separation was performed at a constant flow rate of 350 nl/min with a binary gradient of 2–5% mobile phase B for the first 2 min, followed by a linear increase to 35% mobile phase B over 93 min and column washing for a total of 5 min with a maximum of 98% mobile phase B before re-equilibration.

Standard shotgun LC-MS experiments were performed with a data-dependent top15 method. Full MS scans were acquired from *m/z* 300–1,650 at a resolution of 60,000 at *m/z* 200 with a target of 3×10^6 charges for the automated gain control (AGC). For higher-energy collisional-dissociation MS/MS scans, the normalized collision energy was set to 27 and the resolution to 15,000 at *m/z* 200. Precursor ions were isolated in a 1.4 Thomson (Th) or 1.5 Th window and accumulated for a maximum of 25 ms or until the AGC target of 1×10^5 ions was reached. Precursors with unassigned charge states, a charge of 1+, or a charge of 6+ and higher were excluded from sequencing. Previously targeted precursors were dynamically excluded from re-sequencing for 20 s.

Implementation of BoxCar scans on a quadrupole Orbitrap mass spectrometer. The Q Exactive HF mass spectrometer was operated with the instrument control software Tune v2.6 (build 2640, Thermo Fisher Scientific). Customized MS acquisition methods were built in the 'text method interface' of Xcalibur (v3.1.66.10, Thermo Fisher Scientific), which was accessed through 'XmlMode' as provided by Thermo Fisher Scientific. A detailed step-by-step procedure for implementing BoxCar scans is available in the Supplementary Protocol.

We implemented BoxCar scans using the built-in multiple-ion injection method, with the normalized collision energy set to 0. The scan sequence for a single BoxCar scan was as follows: (i) the mass range of the first 'box' was selected in the analytical quadrupole, and the ions were accumulated in the C-trap until either the AGC target value or the maximum ion injection time was reached; (ii) the selection window of the analytical quadrupole was switched to the next 'box', and the ions were accumulated as before; (iii) step ii was repeated as often as required; and (iv) the combined ion population was injected into the Orbitrap cell and mass-analyzed in a single transient. Within one LC-MS experiment, the number of BoxCar scans, as well as the number and dimensions of all boxes per BoxCar scan, were fixed. For each BoxCar scan, the total AGC target value was adjusted to less than 1×10^6, to not exceed the space charge limit of the C-trap, and evenly distributed across all boxes. Likewise, the maximum total ion injection time given by the Orbitrap transient time was evenly allotted to all of the boxes. The overall acquisition cycle comprised a standard full scan, two or more BoxCar scans covering a range of *m/z* 400–1,200 and five data-dependent MS/MS scans. Suitable precursor ions were selected from the preceding full scan, applying the same criteria and settings as in the standard shotgun experiments.

Investigation of BoxCar scan parameters. To assess the effect of different BoxCar acquisition parameters on the number of detectable MS1 features, we used a statistical DoE approach. The DoE was designed and analyzed with the MODDE software package version 10.0.0 (Umetrics, Sweden). We varied the number of Boxcar scans per acquisition cycle, the number of 'boxes' per Boxcar scan, the maximum ion injection time per scan and the mass resolution. We chose a D-optimal design with 33 design runs and five center points (Supplementary Table 1). All experiments were run with 45-min gradients of 1 μg peptides from a HeLa digest in randomized order, and the number of detected isotope patterns (*m/z* 400–1,200, *z* > 1) was extracted from the MaxQuant output tables. In total, we included *N* = 37 data points to fit a quadratic model with multiple linear regression. R^2 and Q^2 values > 0.978 indicated a valid model with excellent predictive power and very high reproducibility (0.995). The results are represented as four-dimensional (4D) response contour curves (Supplementary Fig. 1).

Analysis of HeLa proteomes in single runs. Single runs of 1 μg HeLa digest were measured with the 45-min gradient using both the standard shotgun method and BoxCar. BoxCar runs were acquired at a MS1 resolution of 120,000 at *m/z* 200, and the acquisition cycle included two BoxCar scans at 12 boxes, each with a 1 Th overlap at the edges. The maximum ion injection time for each box was set to 21.3 ms to reach an AGC target of 16,666 ions. The width of the individual boxes scaled inversely proportionally to the *m/z* density of the tryptic peptides (Supplementary Table 2). To assess the quantitative accuracy of BoxCar, we mixed HeLa and *E. coli* digests, as well as heavy and light SILAC HeLa digests, as specified in the main text, and analyzed (in each case) 1 μg of the peptide mixtures with the 100-min gradient. The BoxCar parameters were set as described above, but with three BoxCar scans at 12 boxes per acquisition cycle, and the AGC target was set to 83,333 ions per box.

Analysis of human plasma digests in single runs. Purified peptides from human plasma were acquired with a standard shotgun method and BoxCar for comparison, using the 45-min gradient. The MS1 resolution was set to 120,000 at *m/z* 200 for BoxCar, and the acquisition cycle comprised two BoxCar scans at

12 boxes (scaled width, 1 Th overlap) with a maximum ion injection time of 21.3 per box with the individual AGC target set to 25,000.

Library generation and deep proteome coverage in single runs. To increase the proteome coverage in single runs of HeLa digest, we first generated a peptide library from two separate high-pH reversed-phase fractionations as described above and measured each fraction with the 45-min gradient using our standard shotgun acquisition method. The library raw files were analyzed together with the single-run HeLa dataset from above.

To generate a deep library of cerebellum peptides, we fractionated the purified peptides five times and measured each fraction with the 100-min gradient using standard acquisition parameters. Then we acquired quintuplicate single runs of 1 μg digested proteins with our standard shotgun method, as well as with BoxCar. The BoxCar parameters were set as described above for the 100-min gradient.

Adaptation of MaxQuant for BoxCar scans. MaxQuant contains a new data analysis mode for processing BoxCar scans, which assembles the separate BoxCar scans and the preceding full scan into a single high-dynamic-range scan. For this purpose, the BoxCar scans, as well as the full scans, were first transformed to a common m/z grid. This was facilitated by having the same mass resolution in each of these scans. The m/z-dependent bin size of the common m/z grid was chosen to be the local mean of the bin sizes of the full scan and the BoxCar scans in the raw data. Note that the typical spacing of the raw data points in 'profile mode' is about $10\times$ more finely grained than the mass resolution. The intensities on the new m/z grid with common binning were calculated from the original scans as linear interpolations from the closest m/z values greater than and less than the m/z value on the new grid. Next, BoxCar scans for each range combination, as well as the full scans, were summed up over the whole retention time range. The summed full scan, as well as the summed BoxCar scans, allowed us to calculate a transmission function for each BoxCar range as follows.

Because the full scan summed over the whole LC-MS run, as well as the summed BoxCar scans, had the same common m/z grid, we were able to divide intensity values point wise. The transmission function for BoxCar scan j is the point-wise ratio between the intensities of the full scan and the BoxCar scan:

$$\text{transmission}_j(m/z) = \text{BoxCar intensity}_j(m/z)\,/\,\text{full} - \text{scan intensity}(m/z)$$

These transmission functions were used to calculate a single high-dynamic-range scan by using the transmission function as a weight for a weighted average of the full scan and all BoxCar scans from one acquisition cycle. Notably, the algorithm does not require any user input but rather adapts to the experimental design, preserving full flexibility in choosing the scan range for full scans and BoxCar scans, the number of BoxCar scans, as well as the number of boxes per scan. All subsequent parts of the MaxQuant computational workflow took these scans as input as a replacement for the commonly used MS1 survey scans. The BoxCar analysis mode as described here is fully integrated in the latest public release of MaxQuant (version 1.6.1.0), which is permanently available free-of-charge from http://www.maxquant.org/.

Data analysis. MS Raw files were processed with MaxQuant version 1.5.3.55 or 1.6.0.15, and the extracted MS/MS spectra were matched by the Andromeda search engine against tryptic peptides (maximum of two missed cleavages) derived from human or mouse reference proteomes (Uniprot, version 2016/05) and a list of common contaminants at a false discovery rate (FDR) of 1% on peptide spectrum match and protein levels. The search included cysteine carbamidomethylation as a fixed modification and methionine oxidation and acetylation of the protein N terminus as variable modifications. The required minimum peptide length was seven amino acids, and the maximum peptide mass was limited to 4,600 Da. Maximum mass tolerances were 4.5 p.p.m. for precursor ions after nonlinear recalibration and 20 p.p.m. for fragment ions. MaxLFQ[42] was performed separately in parameter groups with a minimum ratio count of 1. If applicable, peptide

identifications were transferred between samples by 'match between runs' within a 0.3-min (45-min gradients) or a 0.5-min (100-min gradients) window after retention time alignment. The matching type was set to 'matching from' for library runs and 'matching from and to' for single runs. To minimize the potential for false matching (Supplementary Note 1), peptide libraries were generated from the very same cell culture batch or specimen and, if applicable, processed in parallel with the samples for single-run experiments and analyzed using the same chromatography method. Identifications in the library were stringently filtered for a FDR < 1% on the peptide spectrum match and protein group level, which implies an even lower FDR on the peptide level. To estimate the rate of potentially false matches in the current study, we devised a decoy model on the level of MS features by offsetting masses by 40 p.p.m. and permuting all retention times with a minimal shift of 180 s. This resulted in a conservative FDR estimate of < 5% at the matched feature level.

MS metadata, such as ion injection times, were accessed and extracted from raw files with the MSFileReader (v3.0.31, Thermo Fisher Scientific). Further downstream analysis of the MaxQuant output tables and metadata was performed in either Perseus[43] (v1.6.0.8) or the R statistical computing environment[44] (v3.2.1). Copy numbers and absolute protein abundances were estimated with the 'proteomic ruler' as previously described[45]. Decoy database hits were strictly excluded from the analysis, as well as proteins that were identified as potential contaminants or identified exclusively by one site modification.

Statistical analysis. To assess technical differences between BoxCar and shotgun methods, purified peptides were pooled before LC-MS analysis, and the methods were run in alternating order (ABAB). Replicates refer to repeated LC-MS analyses (technical replicates). The exact N numbers for statistical tests are reported in the figure legends where applicable. To calculate coefficients of variation, missing values (displayed as 0 or 'NaN' in the MaxQuant output) were removed before the analysis. Pearson correlation coefficients were used as a measure for linear correlation between logarithmized intensity values from two experiments. In Supplementary Fig. 5, separately for each acquisition method, the data were filtered for protein groups with no missing values in both conditions, and a one-sided two-sample t-test was performed ($N = 3$ replicates per group). To correct for multiple-hypothesis testing, significant hits by using a t-test were truncated at a permutation-based FDR threshold of 0.05 (250 randomizations).

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

Data availability. The datasets generated during and analyzed during the current study have been deposited to the ProteomeXchange Consortium via the PRIDE[46] partner repository with the dataset identifier PXD006109. Source data are available for Figs. 1–6.

References

41. Ong, S.-E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1**, 2650–2660 (2006).
42. 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).
43. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
44. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, Vienna, 2008).
45. 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).
46. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**(D1), D447–D456 (2016).

Applications of the BoxCar Acquisition Method

The translation of proteomics from basic research to clinical applications is a central goal of many laboratories, including ours. On this way, many challenges are yet to be overcome – some of them on the data acquisition level. In hospitals, blood samples and tissue biopsies are the two most common samples collected from patients. As mentioned above, these samples comprise proteins spanning a very high dynamic range. We reasoned that the BoxCar acquisition method could be particularly advantageous in this areas and 'battle-tested' it in two real-life studies with human samples.

The full text of the published paper on the human heart atlas and the manuscript of the plasma proteome profiling study can be found in the appendix.

Article 6: Region and cell-type resolved quantitative proteomic map of the human heart

Sophia Doll[1,2], Martina Dreßen[3], Philipp E. Geyer[1,2], Daniel N. Itzhak[1], Christian Braun[4], Stefanie A. Doppler[3], **Florian Meier[1]**, Marcus-Andre Deutsch[3,5], Harald Lahm[3], Rüdiger Lange[3,5], Markus Krane[3,5] & Matthias Mann[1,2]

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
2 Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
3 Department of Cardiovascular Surgery, German Heart Center Munich at the Technische Universität München, Munich, Germany
4 Forensic Institute, Ludwig-Maximilians-University, Munich, Germany
5 DZHK (German Center for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany

In this study, Sophia Doll from our group and collaboration partners from the German Heart Center set out to map the protein constituents of the healthy human heart by measuring 16 anatomical regions and three major cardiac cell types in depth. Collectively, this resulted in over 10,700 quantified proteins which were assembled to a quantitative and multi-level model of the human heart proteome – a human heart atlas. Having curated this in-depth resource, we applied it as a project-specific peptide library for BoxCar measurements of tissue samples from patients suffering from atrial fibrillation. Note that in this case, maximizing the proteomic depth while working with minimal starting material were key. With BoxCar, we quantified about 3,600 to 4,100 proteins in 100 min single runs of the patient samples with excellent reproducibility in technical replicates, and also high correlation in-between patients as expected from the assumption that only a small proportion of the proteome is altered in disease states. The comparison of patient samples to reference values from the human heart atlas revealed many alterations in the proteome, which appear to be linked to the disease. While the study was too small to draw clinical conclusions, it still highlights the potential of BoxCar in this area.

Article 7: Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery

Nicolai J. Wewer Albrechtsen[1,2,3,4,5,*], Philipp E. Geyer[1,5,*], Sophia Doll[1,5], Peter V. Treit[5], Kirstine N. Bojsen-Møller[3,6], Christoffer Martinussen[3,6], Nils B. Jørgensen[6], Signe S. Torekov[2,3], **Florian Meier**[1,5], Lili Niu[1], Alberto Santos[1], Eva C. Keilhauer[5], Jens J. Holst[2,3], Sten Madsbad[3,6], Matthias Mann[1,5,]

[1] NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
[2] Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[3] NNF Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[4] Department of Clinical Biochemistry, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark.
[5] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
[6] Department of Endocrinology, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark

*These authors contributed equally

Over the last years, Philipp Geyer has been driving efforts in our laboratory to establish a robust MS-based pipeline for 'plasma proteome profiling'[30,67]. In this study, the pipeline was applied to two longitudinal cohorts of 47 morbidly obese patients undergoing bariatric surgery, which is generally recognized as the most effective treatment for obesity and type 2 diabetes. Application of BoxCar increased the number of identified plasma proteins to over 1,000 per sample, which is a dramatic improvement in proteome coverage as compared to previous studies. This depth and scale revealed distinct rearrangements of the human plasma proteome after surgery, mostly in proteins related to lipid metabolism and inflammation response. In the manuscript, data from this study were correlated with a previous study on sustained weight loss[361] as well as clinical parameters, resulting in a panel of potential marker proteins for insulin resistance. Furthermore, the 'two study' design confirmed the value of the so called 'rectangular' stragegy[30] for biomarker research. This strategy aims at large-scale profiling of many patients in many diseases and therefore ultimately requires a very high throughput, however, without compromising proteomic depth. BoxCar holds promise to be a valuable addition to this workflow.

V. Discussion and Outlook

We are entering an era of ubiquitous and complete proteome measurements, which will undoubtedly provide novel insight in human biology and unravel molecular mechanisms of health and disease[110]. A particular focus in our group is the clinical application of proteomics[30]. Routine and large-scale screenings of 1,000s of human plasma proteomes (**Article 6**) or diseased tissues (**Article 7**) hold great promise for revolutionizing medical diagnostics and could become an important pillar of personalized medicine. At the other end of the scale, there are heralds of proteomics on the level of single cells[362]. However, to turn these great adventures into successful journeys, many technological obstacles are yet to overcome. **Figure 16** provides an overview of the contemporary MS-based proteomics workflow and areas for future improvements.

Any proteomics experiment starts with the extraction of proteins from a biological sample of interest and, in case of bottom-up proteomics, proteolytic digestion. This step has tremendously progressed toward miniaturization and automation in the past years and by now, robot-assisted sample preparation enables the parallelized processing of hundreds of samples per day[67]. Excitingly, there are now also platforms operating on the nano-scale to process samples in volumes of less than 200 nL[363]. From this, one can conclude that sample preparation is primed for single cells well as large clinical studies.

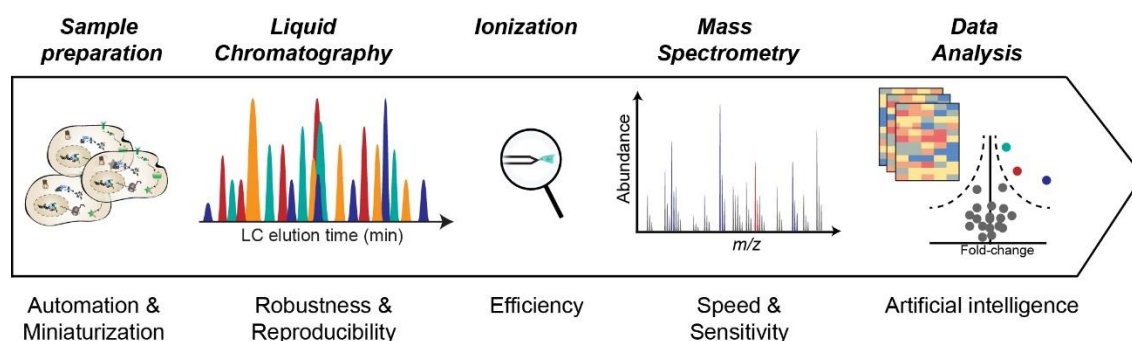**Figure 16 | The future of MS-based proteomics.** The figure highlight areas of active research in the current bottom-up proteomics workflow. On the way toward robust and comprehensive proteomics workflows that scale to population-wide sample cohorts as well as single cell sensitivity multiple improvements are needed. Particularly promising research topics are indicated underneath the workflow.

The next challenge is reproducible and robust high-resolution chromatography[364]. Nano-flow LC provides excellent sensitivity and resolution, however, it remains a finicky technology that requires frequent interventions by the user. In view of larger sample cohorts, there has been a trend toward higher flow rates in the lower µL/min range which tends to be more robust[355]. A novel setup that promises to improve the robustness of LC systems so they are ready for clinical use is incorporated into the Evosep One[356]. Instead of two high-pressure pumps, which are typically the most error prone components, this system employs only one high-pressure pump and stores the sample in a pre-formed gradient. In this way, the system also minimizes overhead times for sample loading and column equilibration, thus allowing analysis of more than 200 samples per day. The column technology itself has remained largely unchanged in recent years, and 50 cm columns packed with sub-2 µm particles are routinely used in our laboratory since 2011[365]. However, also in this area there are new inventions just around the corner and promise more robust and reproducible chromatography, for example new generations of monolithic[366] or chip-based columns[367]. Reproducible and standardized chromatography will be key to future proteomics workflows, and in particular those employing peptide-libraries. This includes the BoxCar method described here, MS/MS-level DIA methods, as well as targeted proteomics, which all dramatically benefit from predictable retention times with narrow uncertainty intervals.

At the interface of LC and MS there is currently no viable alternative to ESI in sight. However, the efficiency of the ionization process and the ion transfer decreases with higher flow rates[194], which is why multi-nozzle emitters could provide a feasible way forward[368]. For proteomics of low cell counts or miniscule tissue biopsies, sub-ambient pressure ionization also appears promising to increase the number of ions entering the MS and thus improve sensitivity[369]. However, although in some cases first described over ten years ago, these technologies still have to prove their applicability in proteomic practice.

Multiplexing can be a powerful answer to the demands of clinical samples and other large scale studies which can encompass 100s and 1,000s of samples. However, while this approach provides a complete quantification matrix within a set of multiplexed samples, it may not at all be complete in between sets of multiplexed samples due to inconsistent precursor selection. This problem is only partially addressed by peptide fractionation as this negates most of the advantages in throughput. Furthermore, to not compromise quantitative accuracy, the most popular method so far requires very expensive

124

instrumentation with MS^3 capabilities and is, even conceptually, not free from artefacts as exemplified by low-abundance precursors[165]. In contrast, the EASI-tag technology introduced here provides interference-free and accurate quantification at the MS/MS level (**Article 1**). Considering that this is only the first generation of the tag, there is plenty of room for improvements, e.g. by increasing the number of channels. An apparent limitation derives from the complexity of the MS/MS spectrum as serial fragmentation events lead to *b* and *y* ions with the full reporter ion pattern. However, computationally, this could even be turned into an advantage, given that the exact position of the expected reporter ions can be calculated and used in scoring algorithms. Additionally, the signals with the correct delta masses could be stacked to clean up the spectra prior to the database search. Given that the noise level in MS/MS spectra with the very narrow isolation windows is generally very low, such strategies appear feasible and promising.

In an ideal mass spectrometer, all ions entering the MS are ultimately used for mass analysis. The BoxCar method developed in this thesis highlights the dramatic improvements that can still be achieved in this area (**Article 5**). While largely abandoned over the last decade, full scans are now receiving more attention and BoxCar has great potential to fill this gap. In DDA methods, this is supported by the fact that very fast instruments eventually run out of suitable precursor ions[297]. In targeted proteomics, a high-quality MS signal can be used to trigger quantification events more precisely, thereby increasing throughput and sensitivity[165,324]. The improved precursor information with BoxCar also opens up new possibilities for 'intelligent' data acquisition schemes that go beyond standard topN picking schemes and consider further parameters to predict the potential information content of a precursor candidate. This should allow targeted proteomics on a proteome-wide scale and in a highly automated manner, avoiding manual optimization and validation of target peptides. Having established such a strategy, it could readily be extended to the EASI-tag technology for reproducible and accurate multiplexed quantification. In DIA methods, the growing interest in full scans is connected to re-emerging library-free approaches, which inevitably require accurate and high-quality precursor ion information[357,358]. To a lesser extent, this is also true for library-based strategies, in which improved precursor information increases matching scores and allows better stratification of true and false hits[354].

Conceptually, TIMS-PASEF comes closest to the ideal mass spectrometer (**Articles 2-4**). Within a PASEF set, all precursor ions are stored in parallel and focused into narrow ion

mobility peaks. By the time precursors elute from the ion mobility device, the synchronized quadrupole ideally transmits all precursor ions into the collision cell. In the present thesis, I used this principle to multiply the peptide sequencing rate up to the sensitivity limit of the mass spectrometer. Reaching this limit, the tremendous speed can also be used in part on re-sequencing low-abundance precursors, yielding cumulative MS/MS spectra with increased signal-to-noise ratios. Note that this resembles targeted proteomics and also DIA, in which precursors of interest are sampled repeatedly over the entire chromatographic elution peak. In targeted proteomics, narrow isolation windows are employed to achieve maximum selectivity and sensitivity – at the cost of proteome coverage. In DIA, wider isolation windows (yielding highly convoluted MS/MS spectra) are used to maximize the proteome coverage and ideally acquire a digital record of fragment ions for all eluting precursor ions.

Even the most sophisticated data acquisition strategy is worthless without post-processing software that makes good use of the data record. In this regard, it is surprising that the groundbreaking developments of artificial intelligence in conjunction with deep neural networks[370] have not yet been widely translated into proteomics research. In combination with the improvements in data acquisition outlined above and the availability of high-quality data for virtually all (tryptic) peptides[330], it appears clear that the application of neural networks will render labor-intensive peptide library generation superfluous and revolutionize data analysis as we know it today.

In summary, in this thesis I have presented three novel data acquisition strategies – EASI-tag, PASEF and BoxCar – which help pave the way for the next generation of MS-based proteomics in its quest for comprehensive large-scale analysis and single cell sensitivity. These developments will go hand in hand with upcoming advances in data analysis and open up exciting perspectives for proteome research. Importantly, while developed in the context of proteomics, the strategies outlined in this thesis are generic and readily transferrable to other MS-based omics technologies, for example metabolomics and lipidomics, which face similar analytical challenges in terms of dynamic range and sensitivity, and perhaps even greater challenges in terms of selectivity.

VI. References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

2. Venter, J. C. The Sequence of the Human Genome. *Science* **291,** 1304–1351 (2001).

3. Collins, F. S., Lander, E. S., Rogers, J. & Waterson, R. H. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

4. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470,** 187–197 (2011).

5. Goffeau, A. *et al.* Life with 6000 Genes conveniently among the different interna- Old Questions and New Answers The genome . At the beginning of the se- of its more complex relatives in the eukary- cerevisiae has been completely sequenced Schizosaccharomyces pombe indicate. *Science* **274,** 546–567 (1996).

6. Waterston, R., Sulstont, J. & Sulston, J. The genome of Caenorhabditis elegans. *Pnas* **92,** 10836–10840 (1995).

7. Adams, M. D. *et al.* The Genome Sequence of Drosophila melanogaster. *Science* **287,** 2185–2195 (2000).

8. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74,** 5463–5467 (1977).

9. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Mol. Cell* **58,** 586–597 (2015).

10. 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).

11. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

12. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the science. *Nat. Rev. Genet.* **17,** 175–188 (2016).

13. Biesecker, L. G. & Green, R. C. Diagnostic Clinical Genome and Exome Sequencing. *N. Engl. J. Med.* **370,** 2418–2425 (2014).

14. Tabarrok, A. The rise, fall and rise again of 23ANDME. *Nature* **11,** 565–577 (2007).

15. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

16. Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88,** 243–251 (1997).

17. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7,** 862–872 (2006).

18. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

19. McGettigan, P. A. Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* **17,** 4–11 (2013).

20. Wilkins, M. Proteomics data mining. *Expert Rev Proteomics.* **6,** 599–603 (2009).

21. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165,** 535–550 (2016).

22. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534,** 55–62 (2016).

23. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387 (2014).

24. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547,** E19–E20 (2017).

25. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).

26. Volpe P & Eremenko-Volpe T. Quantitative studies on cell proteins in suspension cultures. *Eur J Biochem* **12,** 195–200 (1970).

27. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549 (2011).

28. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* **18,** 1–16 (2015).

29. Anderson, N. L. & Anderson, N. G. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. *Mol. Cell. Proteomics* **1,** 845–867 (2002).

30. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13,** 942 (2017).

31. Uhlen, M. & Ponten, F. Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* **4,** 384–93 (2005).

32. The Human Protein Atlas. at <www.proteinatlas.org>. Date accessed: 2018/06/17.

33.   Smith, L. M. & Kelleher, N. L. Proteoform: A single term describing protein complexity. *Nat. Methods* **10,** 186–187 (2013).

34.   Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14,** 206–214 (2018).

35.   Prabakaran, S., Lippens, G., Steen, H. & Gunawardena, J. Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **4,** 565–583 (2012).

36.   Olsen, J. V. *et al.* Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127,** 635–648 (2006).

37.   Choudhary, C. & Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11,** 427–39 (2010).

38.   Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358,** (2017).

39.   Barrios-Rodiles, M. *et al.* High-Throughput Mapping of a Dynamic Signaling Network in Mammalian Cells High-Throughput Mapping of a Dynamic Signaling Network in Mammalian Cells. *Science* **307,** 1621–1625 (2005).

40.   Wild, D. *The Immunoassay Handbook*. (Elsevier Science, 2013).

41.   Goldberg. A Theory of Antibody Antigen Reactions .1. Theory for Reactions of Multivalent Antigen with Bivalent and Univalent Antibody. *J. Am. Chem. Soc.* 5715–5725 (1952). doi:10.1021/ja01142a045

42.   Coons, A. H., Creech, H. J. & Jones, R. N. Immunological Properties of an Antibody Containing a Fluorescent Group. *Exp. Biol. Med.* **47,** 200–202 (1941).

43.   Uhlén, M. *et al.* A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. *Mol. Cell. Proteomics* **4,** 1920–1932 (2005).

44.   Lequin, R. M. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clin. Chem.* **51,** 2415–2418 (2005).

45.   Hoofnagle, A. N. & Wener, M. H. The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. *J. Immunol. Methods* **347,** 3–11 (2009).

46.   Baker, M. Blame it on the antibodies. *Nature* **521,** 274–276 (2015).

47.   Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).

48.   Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).

49.   Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **68,** 850–858 (1996).

50.   Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682 (1999).

51.   Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247 (2001).

52.   Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **4,** 1419–1440 (2005).

53.   Heck, A. J. R. Native mass spectrometry: A bridge between interactomics and structural biology. *Nat. Methods* **5,** 927–933 (2008).

54.   Kelleher, N. L. Peer Reviewed: Top-Down Proteomics. *Anal. Chem.* **76,** 196 A-203 A (2004).

55.   Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* **445,** 683–693 (2014).

56.   Forbes, A. J., Mazur, M. T., Patel, H. M., Walsh, C. T. & Kelleher, N. L. Toward efficient analysis of <70 kDa proteins with 100% sequence coverage. *Proteomics* **1,** 927–933 (2001).

57.   Wu, C. *et al.* A protease for 'middle-down' proteomics. *Nat. Methods* **9,** 822–4 (2012).

58.   Hein, M. Y., Sharma, K., Cox, J. & Mann, M. in *Handbook of Systems Biology* 3–25 (Elsevier, 2013). doi:10.1016/B978-0-12-385944-0.00001-0

59.   Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–62 (2009).

60.   Olsen, J. V, Ong, S.-E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–14 (2004).

61.   Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9,** 1323–1329 (2010).

62.   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).

63.   Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–62 (2009).

64.   Wiśniewski, J. R., Zougman, A. & Mann, M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* **8,** 5674–8 (2009).

65.   Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2,** 1896–906 (2007).

66.   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,** 319–24 (2014).

67.   Geyer, P. E. *et al.* Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2,** 185–195 (2016).

68.   Doll, S. & Burlingame, A. L. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem. Biol.* **10,** 63–71 (2015).

69.   Olsen, J. V. & Mann, M. Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry. *Mol. Cell. Proteomics* **12,** 3444–3452 (2013).

70.   Riley, N. M. & Coon, J. J. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Anal. Chem.* **88,** 74–94 (2016).

71.   Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell* **44,** 325–340 (2011).

72.   Wagner, S. A. *et al.* A Proteome-wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles. *Mol. Cell. Proteomics* **10,** M111.013284 (2011).

73.   Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23,** 94–101 (2005).

74.   Gingras, A.-C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8,** 645–654 (2007).

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

76.   Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440 (2015).

77.   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.* **113,** E3501–E3508 (2016).

78.   Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biol. Mass Spectrom.* **16,** 99–111 (1988).

79.   Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5,** 699–711 (2004).

80.   Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–12 (2007).

81.   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,** 9528–33 (2004).

82.   Reilly, J. P. Ultraviolet photofragmentation of biomolecular ions. *Mass Spectrom. Rev.* **28,** 425–447 (2009).

83.   Brodbelt, J. S. Photodissociation mass spectrometry: New tools for characterization of biological molecules. *Chem. Soc. Rev.* **43,** 2757–2783 (2014).

84.   Fournier, M. L., Gilmore, J. M., Martin-Brown, S. A. & Washburn, M. P. Multidimensional separations-based shotgun proteomics. *Chem. Rev.* **107,** 3654–3686 (2007).

85.   Dowell, J. A., Frost, D. C., Zhang, J. & Li, N. Comparison of two-dimensional fractionation techniques for shotgun proteomics. *Anal. Chem.* **80,** 6715–6723 (2008).

86.   Mallick, P. & Kuster, B. Proteomics: A pragmatic perspective. *Nat. Biotechnol.* **28,** 695–709 (2010).

87.   Motoyama, A. & Yates, J. R. Multidimensional LC separations in shotgun proteomics. *Anal. Chem.* **80,** 7187–7193 (2008).

88.   Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690 (2001).

89.   Gilar, M., Olivova, P., Daly, A. E. & Gebler, J. C. Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J. Sep. Sci.* **28,** 1694–1703 (2005).

90.   Toll, H., Oberacher, H., Swart, R. & Huber, C. G. Separation, detection, and identification of peptides by ion-pair reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry at high and low pH. *J. Chromatogr. A* **1079,** 274–286 (2005).

91.   Gilar, M., Olivova, P., Daly, A. E. & Gebler, J. C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.* **77,** 6426–6434 (2005).

92.   Dwivedi, R. C. *et al.* Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both

dimensions for high-throughput bottom-up proteomics. *Anal. Chem.* **80,** 7036–7042 (2008).

93. Wang, Y. *et al.* Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11,** 2019–2026 (2011).

94. Song, C. *et al.* Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides. *Anal. Chem.* **82,** 53–56 (2010).

95. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* mcp.O116.065136 (2017). doi:10.1074/mcp.O116.065136

96. de Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–4 (2008).

97. Hebert, A. S. *et al.* The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **13,** 339–347 (2013).

98. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24,** 11–17 (2015).

99. Feng, Y., Cappelletti, V. & Picotti, P. Quantitative proteomics of model organisms. *Curr. Opin. Syst. Biol.* **6,** 58–66 (2017).

100. Walther, D.M. *et al.* Widespread Proteome Remodeling and Aggregation in Aging C . elegans. *Cell* **161,** 919–932 (2015).

101. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol. Cell. Proteomics* **11,** M111.014050-M111.014050 (2012).

102. Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell. Proteomics* **12,** 1709–22 (2013).

103. Azimifar, S. B., Nagaraj, N., Cox, J. & Mann, M. Cell-type-resolved quantitative proteomics of murine liver. *Cell Metab.* **20,** 1076–1087 (2014).

104. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–81 (2014).

105. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–7 (2014).

106. Schmidt, T. *et al.* ProteomicsDB. *Nucleic Acids Res.* **46,** D1271–D1281 (2018).

107. Schirle, M., Heurtier, M.-A. & Kuster, B. Profiling Core Proteomes of Human Cell Lines by One-dimensional PAGE and Liquid Chromatography-Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **2,** 1297–1305 (2003).

108. Nesvizhskii, A. I. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11,** 1114–1125 (2014).

109. 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).

110. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Mol. Cell* **49,** 583–590 (2013).

111. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166,** 755–765 (2016).

112. Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11,** 786 (2015).

113. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4,** 787–797 (2007).

114. 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,** 1367–72 (2008).

115. P. Sinitcyn, J. Rudolph, J. C. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1,** 207–234 (2018).

116. 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).

117. Mann, M. & Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18132–8 (2008).

118. Muddiman, D. C. & Oberg, A. L. Statistical evaluation of internal and external mass calibration laws utilized in Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **77,** 2406–2414 (2005).

119. 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,** 2010–21 (2005).

120. Cox, J., Michalski, A. & Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* **22,** 1373–80 (2011).

130

121.   Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).

122.   Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–14 (2007).

123.   Brosch, M. Sensitive and accurate peptide identification with Mascot Percolator Terminology : FPR ,. *J. Proteome Res.* 3176–3181 (2009).

124.   Johnson, R. S. & Taylor, J. A. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Appl. Biochem. Biotechnol. - Part B Mol. Biotechnol.* **22,** 301–315 (2002).

125.   Muth, T. & Renard, B. Y. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief. Bioinform.* 1–17 (2017). doi:10.1093/bib/bbx033

126.   Mann, M. & Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **66,** 4390–4399 (1994).

127.   Tabb, D. L., Saraf, A. & Yates, J. R. GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Anal. Chem.* **75,** 6415–6421 (2003).

128.   Nesvizhskii, A. I. *et al.* Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **5,** 652–670 (2006).

129.   Zhou, X. X. *et al.* PDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **89,** 12690–12697 (2017).

130.   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,** (2015).

131.   Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **13,** 651–656 (2016).

132.   Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

133.   Reiter, L. *et al.* Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **8,** 2405–2417 (2009).

134.   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).

135.   Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry – based shotgun proteomics. *Nat. Protoc.* **11,** 2301–2319 (2016).

136.   Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–40 (2016).

137.   Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

138.   Ong, S.-E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–62 (2005).

139.   Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389,** 1017–31 (2007).

140.   Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404,** 939–65 (2012).

141.   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).

142.   Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7,** 952–8 (2006).

143.   Schwanhäusser, B., Gossen, M., Dittmar, G. & Selbach, M. Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* **9,** 205–209 (2009).

144.   Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

145.   McShane, E. *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167,** 803–815.e21 (2016).

146.   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).

147.   Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7,** 383–5 (2010).

148.   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).

149.   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).

150.   Hebert, A. S. *et al.* Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **10,** 332–334 (2013).

151.   Merrill, A. E. *et al.* NeuCode Labels for Relative Protein Quantification. *Mol. Cell. Proteomics* **13,** 2503–2512 (2014).

152.   Hsu, J. L., Huang, S. Y., Chow, N. H. & Chen, S. H. Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Anal. Chem.* **75,** 6843–6852 (2003).

153.   Hsu, J. L., Huang, S. Y. & Chen, S. H. Dimethyl multiplexed labeling combined with microcolumn separation and MS analysis for time course study in proteomics. *Electrophoresis* **27,** 3652–3660 (2006).

154.   Ross, P. L. *et al.* Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **3,** 1154–1169 (2004).

155.   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).

156.   Werner, T. *et al.* High-resolution enabled TMT 8-plexing. *Anal. Chem.* **84,** 7188–7194 (2012).

157.   Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* **12,** 3586–3598 (2013).

158.   Bantscheff, M. *et al.* Robust and Sensitive iTRAQ Quantification on an LTQ Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **7,** 1702–1713 (2008).

159.   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).

160.   Karp, N. A. *et al.* Addressing Accuracy and Precision Issues in iTRAQ Quantitation. *Mol. Cell. Proteomics* **9,** 1885–1897 (2010).

161.   Sturm, R. M., Lietz, C. B. & Li, L. Improved isobaric tandem mass tag quantification by ion mobility mass spectrometry. *Rapid Commun. Mass Spectrom.* **28,** 1051–1060 (2014).

162.   Wenger, C. D. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* **8,** 933–935 (2011).

163.   Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8,** 937–940 (2011).

164.   McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86,** 7150–7158 (2014).

165.   Erickson, B. K. *et al.* A Strategy to Combine Sample Multiplexing with Targeted Proteomics Assays for High-Throughput Protein Signature Characterization. *Mol. Cell* **65,** 361–370 (2017).

166.   Wühr, M. *et al.* Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* **84,** 9214–9221 (2012).

167.   Sonnett, M., Yeung, E. & Wühr, M. Accurate, Sensitive, and Precise Multiplexed Proteomics using the Complement Reporter Ion Cluster. *Anal. Chem.* acs.analchem.7b04713 (2018). doi:10.1021/acs.analchem.7b04713

168.   Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **28,** 710–721 (2010).

169.   Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for Label-free Peptide Quantification. *Mol. Cell. Proteomics* **12,** 549–556 (2013).

170.   Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3685

171.   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).

172.   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).

173.   Ning, K., Fermin, D. & Nesvizhskii, A. I. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res.* **11,** 2261–2271 (2012).

174.   Clough, T. *et al.* Protein quantification in label-free LC-MS experiments. *J. Proteome Res.* **8,** 5275–5284 (2009).

175.   Chelius, D. & Bondarenko, P. V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome Res.* **1,** 317–323 (2002).

176.   Bondarenko, P. V., Chelius, D. & Shaler, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography - Tandem mass spectrometry. *Anal. Chem.* **74,** 4741–4749 (2002).

177.   Old, W. M. *et al.* Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol. Cell. Proteomics* **4,** 1487–1502 (2005).

178.   Silva, J. C., Gorenstein, M. V, Li, G.-Z., Vissers, J. P. C. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144–56 (2006).

179.   Cox, J. *et al.* MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction. *Mol. Cell. Proteomics* M113.031591- (2014). doi:10.1074/mcp.M113.031591

180.   Rost, H. L., Malmstrom, L. & Aebersold, R. Reproducible quantitative prototype data matrices for systems biology. *Mol. Biol. Cell* **26,** 3926–3931 (2015).

181.   Shen, X. *et al.* IonStar enables high-precision, low-missing-data proteomics quantification in large biological cohorts. *Proc. Natl. Acad. Sci.* 201800541 (2018). doi:10.1073/PNAS.1800541115

182.   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).

183.   Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Anal. Chem.* **60,** 2299–2301 (1988).

184.   Tanaka, K. *et al.* Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2,** 151–153 (1988).

185.   Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* **88,** 4451–4459 (1984).

186.   Meng, C. K., Mann, M. & Fenn, J. B. Of protons or proteins. *Zeitschrift fÜr Phys. D Atoms, Mol. Clust.* **10,** 361–368 (1988).

187.   Fenn, J., Mann, M., Meng, C., Wong, S. & Whitehouse, C. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).

188.   Mann, M., Meng, C. K. & Fenn, J. B. Interpreting Mass Spectra of Multiply Charged Ions. *Anal. Chem.* **61,** 1702–1708 (1989).

189.   Fenn, J. B. Electrospray wings for molecular elephants (Nobel lecture). *Angew. Chemie - Int. Ed.* **42,** 3871–3894 (2003).

190.   Knochenmuss, R. The Coupled Chemical and Physical Dynamics Model of MALDI. *Annu. Rev. Anal. Chem.* **9,** 365–385 (2016).

191.   McDonnell, L. A. & Heeren, R. M. A. Imaging mass spectrometry. *Mass Spectrom. Rev.* **26,** 606–43 (2007).

192.   Spengler, B. Mass spectrometry imaging of biomolecular information. *Anal. Chem.* **87,** 64–82 (2015).

193.   Taylor, G. Disintegration of Water Drops in an Electric Field. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **280,** 383–397 (1964).

194.   Wilm, M. Principles of electrospray ionization. *Mol. Cell. Proteomics* **10,** M111.009407 (2011).

195.   Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379,** 466–469 (1996).

196.   Belov, M. E., Gorshkov, M. V, Udseth, H. R., Anderson, G. A. & Smith, R. D. Zeptomole-Sensitivity Electrospray Ionization−Fourier Transform Ion Cyclotron Resonance Mass Spectrometry of Proteins. *Anal. Chem.* **72,** 2271–2279 (2000).

197.   Brunnée, C. The ideal mass analyzer: fact or fiction? *Int. J. Mass Spectrom. Ion Process.* **76,** 125–237 (1987).

198.   Gross, J. H. *Mass Spectrometry - A Textbook.* (Springer, 2017).

199.   Paul, W. & Steinwedel, H. Notizen: Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift für Naturforsch. A* **8,** (1953).

200.   Paul, W. & Raether, M. Das elektrische Massenfilter. *Zeitschrift f??r Phys.* **140,** 262–273 (1955).

201.   Dawson, P. Quadrupole Mass Analysers: Performance, design and some recent applications. *Mass Spectrom. Rev.* **5,** 1–37 (1986).

202.   Douglas, D. J. Linear quadrupoles in mass spectrometry. *Mass Spectrom. Rev.* **28,** 937–960 (2009).

203.   Paul, W. Electromagnetic Traps for Charged and Neutral Particles(Nobel Lecture). *Angew. Chemie Int. Ed. English* **29,** 739–748 (1990).

204.   Yost, R. A. & Enke, C. G. Selected Ion Fragmentation with a Tandem Quadrupole Mass Spectrometer. *J. Am. Chem. Soc.* **100,** 2274–2275 (1978).

205.   Yost, R. A. & Enke, C. G. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Anal. Chem.* **51,** 1251–1264 (1979).

206.   Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72,** 1156–1162 (2000).

207. 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).

208. Scigelova, M. & Makarov, A. in *Encyclopedia of Analytical Chemistry* 1–36 (John Wiley & Sons, Ltd, 2013). doi:10.1002/9780470027318.a9309

209. Perry, R. H., Cooks, R. G. & Noll, R. J. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev.* **27,** 661–99 (2008).

210. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8,** 61–80 (2015).

211. Hardman, M. & Makarov, A. a. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal. Chem.* **75,** 1699–705 (2003).

212. Makarov, A. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78,** 2113–2120 (2006).

213. 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* M114.043489- (2014). doi:10.1074/mcp.M114.043489

214. Kelstrup, C. D. *et al.* Rapid and Deep Proteomes by Faster Sequencing on a Benchtop Quadrupole Ultra-High-Field Orbitrap Mass Spectrometer. *J. Proteome Res.* (2014). doi:10.1021/pr500985w

215. Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* acs.jproteome.7b00602 (2017). doi:10.1021/acs.jproteome.7b00602

216. Grinfeld, D., Aizikov, K., Kreutzmann, A., Damoc, E. & Makarov, A. A. Phase-Constrained Spectrum Deconvolution for Fourier Transform Mass Spectrometry. *Anal. Chem.* acs.analchem.6b03636 (2016). doi:10.1021/acs.analchem.6b03636

217. Cameron, A. E. & Eggers, D. F. An Ion 'Velocitron''''. *Rev. Sci. Instrum.* **19,** 605–607 (1948).

218. Wolff, M. M. & Stephens, W. E. A pulsed mass spectrometer with time dispersion. *Rev. Sci. Instrum.* **24,** 616–617 (1953).

219. 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).

220. Dawson, J. H. J. & Guilhaus, M. Orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **3,** 155–159 (1989).

221. Mirgorodskaya, O. A., Shevchenko, A. A., Chernushevich, I. V., Dodonov, A. F. & Miroshnikov, A. I. Electrospray-ionization time-of-flight mass spectrometry in protein chemistry. *Anal. Chem.* **66,** 99–107 (1994).

222. Andrews, G. L., Simons, B. L., Young, J. B., Hawkridge, A. M. & Muddiman, D. C. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Anal. Chem.* **83,** 5442–5446 (2011).

223. 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).

224. Mamyrin, B. A. Laser assisted reflectron time-of-flight mass spectrometry. *Int. J. Mass Spectrom. Ion Process.* **131,** 1–19 (1994).

225. Casares, A., Kholomeev, A. & Wollnik, H. Multipass time-of-flight mass spectrometers with high resolving powers. *Int. J. Mass Spectrom.* **206,** 267–273 (2001).

226. Wollnik, H. & Casares, A. An energy-isochronous multi-pass time-of-flight mass spectrometer consisting of two coaxial electrostatic mirrors. *Int. J. Mass Spectrom.* **227,** 217–222 (2003).

227. 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).

228. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312,** 212–7 (2006).

229. 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).

230. Eiceman, G. A., Karpas, Z. & Hill, H. H. J. *Ion Mobility Spectrometry*. (CRC Press, 2013).

231. Mason, E. A. & McDaniel, E. W. *Transport Properties of Ions in Gases*. (John Wiley & Sons, Inc., 1988).

232. Revercomb, H. E. & Mason, E. A. Theory of plasma chromatography/gaseous electrophoresis. Review. *Anal. Chem.* **47,** 970–983 (1975).

233. Asbury, G. & Hill, H. Using different drift gases to change separation factors (alpha) in ion mobility spectrometry. *Anal. Chem.* **72,** 580–4 (2000).

234. Ruotolo, B. T., McLean, J. A., Gillig, K. J. & Russell, D. H. The influence and utility of varying field strength for the separation of tryptic peptides by ion mobility-mass spectrometry. *J. Am. Soc. Mass Spectrom.* **16,** 158–165 (2005).

235. Eiceman, G. A. Advances in Ion Mobility Spectrometry: 1980—1990. *Crit. Rev. Anal. Chem.* **22,** 471–490 (1991).

236. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H. Ion mobility-mass spectrometry. *J. Mass Spectrom.* **43,** 1–22 (2008).

237. May, J. C. & McLean, J. A. Ion Mobility-Mass Spectrometry: Time-Dispersive Instrumentation. *Anal. Chem.* **87,** 1422–1436 (2015).

238. 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).

239. Hill, H. H., Siems, W. F. & St. Louis, R. H. Ion mobility spectrometry. *Anal. Chem.* **62,** 1201A–1209A (1990).

240. Thomson, J. J. & Rutherford, E. On the passage of electricity through gases exposed to Röntgen rays. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **42,** 392–407 (1896).

241. Zeleny, J. On Air Electrified by the Discharging Action of Ultra-violet Light. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **46,** 272–273 (1898).

242. Langevin, P. L'Ionisation des Gaz. *Ann. Chim. Phys.* **28,** 289–384 (1903).

243. Tyndall, A. M. & Grindley, G. C. The Mobility of Ions in Air. Part I. Negative Ions in Moist Air. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **110,** 341–358 (1926).

244. Tyndall, A. M. & Powell, C. F. The Mobility of Positive Ions in Helium. Part I. Helium Ions. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **134,** 125–136 (1931).

245. Bradbury, N. E. The Absolute Values of the Mobility of Gaseous Ions in Pure Gases. *Phys. Rev.* **40,** 508–523 (1932).

246. Barnes, W. S., Martin, D. W. & McDaniel, E. W. Mass Spectrographic Identification of the Ion Observed in Hydrogen Mobility Experiments. *Phys. Rev. Lett.* **6,** 110–111 (1961).

247. McAfee, K. B. & Edelson, D. Identification and Mobility of Ions in a Townsend Discharge by Time-resolved Mass Spectrometry. *Proc. Phys. Soc.* **81,** 382–384 (1963).

248. Zheng, X. *et al.* Coupling Front-End Separations, Ion Mobility Spectrometry, and Mass Spectrometry For Enhanced Multidimensional Biological and Environmental Analyses. *Annu. Rev. Anal. Chem.* **10,** 71–92 (2017).

249. Kemper, P. R., Dupuis, N. F. & Bowers, M. T. A new, higher resolution, ion mobility mass spectrometer. *Int. J. Mass Spectrom.* **287,** 46–57 (2009).

250. Dugourd, P., Hudgins, R. R., Clemmer, D. E. & Jarrold, M. F. High-resolution ion mobility measurements. *Rev. Sci. Instrum.* **68,** 1122–1129 (1997).

251. 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).

252. 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).

253. Tang, K. *et al.* High-sensitivity ion mobility spectrometry/mass spectrometry using electrodynamic ion funnel interfaces. *Anal. Chem.* **77,** 3330–3339 (2005).

254. Allen, S. J. & Bush, M. F. Radio-Frequency (rf) Confinement in Ion Mobility Spectrometry: Apparent Mobilities and Effective Temperatures. *J. Am. Soc. Mass Spectrom.* **27,** 2054–2063 (2016).

255. 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,** 2451–62 (2007).

256. Ibrahim, Y. M. *et al.* Development of a new ion mobility (quadrupole) time-of-flight mass spectrometer. *Int. J. Mass Spectrom.* **377,** 655–662 (2015).

257. Baker, E. S. *et al.* Enhancing bottom-up and top-down proteomic measurements with ion mobility separations. *Proteomics* 2766–2776 (2015). doi:10.1002/pmic.201500048

258. Giles, K. *et al.* Applications of a travelling wave-based radio-frequency-only stacked ring ion guide. *Rapid Commun. Mass Spectrom.* **18,** 2401–14 (2004).

259. 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).

260. Giles, K., Williams, J. P. & Campuzano, I. Enhancements in travelling wave ion mobility resolution. *Rapid Commun. Mass Spectrom.* **25,** 1559–1566 (2011).

261. Zhong, Y., Hyung, S.-J. & Ruotolo, B. T. Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions. *Analyst* **136,** 3534–3541 (2011).

262. Shvartsburg, A. A. & Smith, R. D. Fundamentals of traveling wave ion mobility spectrometry. *Anal. Chem.* **80,** 9689–99 (2008).

263. Dodds, J. N., May, J. C. & McLean, J. A. Correlating Resolving Power, Resolution, and Collision Cross Section: Unifying

Cross-Platform Assessment of Separation Efficiency in Ion Mobility Spectrometry. *Anal. Chem.* **89,** 12176–12184 (2017).

264. Bush, M. F. *et al.* Collision cross sections of proteins and their complexes: A calibration framework and database for gas-phase structural biology. *Anal. Chem.* **82,** 9557–9565 (2010).

265. 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).

266. 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).

267. Helm, D. *et al.* Ion Mobility Tandem Mass Spectrometry Enhances Performance of Bottom-up Proteomics. *Mol. Cell. Proteomics* **13,** 3709–15 (2014).

268. Distler, U. *et al.* Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **11,** 167–70 (2014).

269. 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,** 2323–39 (2013).

270. Buryakov, I. A., Krylov, E. V., Nazarov, E. G. & Rasulev, U. K. 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 (1993).

271. Bennett, R. V, Gamage, C. M., Galhena, A. S. & Fernández, F. M. Contrast-enhanced differential mobility-desorption electrospray ionization-mass spectrometry imaging of biological tissues. *Anal. Chem.* **86,** 3756–63 (2014).

272. 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).

273. Santiago, B. G., Harris, R. A., Isenberg, S. L. & Glish, G. L. Resolving powers of >7900 using linked scans: how well does resolving power describe the separation capability of differential ion mobility spectrometry. *Analyst* **140,** 6871–6878 (2015).

274. Venne, K., Bonneil, E., Eng, K. & Thibault, P. Improvement in peptide detection for proteomics analyses using nanoLC-MS and high-field asymmetry waveform ion mobility mass spectrometry. *Anal. Chem.* **77,** 2176–2186 (2005).

275. Canterbury, J. D., Yi, X., Hoopmann, M. R. & MacCoss, M. J. Assessing the Dynamic Range and Peak Capacity of Nanoflow LC-FAIM-?MS on an Ion Trap Mass Spectrometer for Proteomics. *Anal. Chem.* **80,** 6888–6897 (2008).

276. Saba, J., Bonneil, E., Pomiès, C., Eng, K. & Thibault, P. Enhanced sensitivity in proteomics experiments using FAIMS coupled with a hybrid linear ion trap/orbitrap mass spectrometer. *J. Proteome Res.* **8,** 3355–3366 (2009).

277. 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).

278. Shvartsburg, A. A. *Differential Ion Mobility Spectrometry: Nonlinear Ion Transport and Fundamentals of FAIMS*. (CRC Press, 2008).

279. 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).

280. 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).

281. 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).

282. Benigni, P., Thompson, C. J., Ridgeway, M. E., Park, M. A. & Fernandez-Lima, F. Targeted high-resolution ion mobility separation coupled to ultrahigh-resolution mass spectrometry of endocrine disruptors in complex mixtures. *Anal. Chem.* **87,** 4321–4325 (2015).

283. Pu, Y. *et al.* Separation and Identification of Isomeric Glycans by Selected Accumulation-Trapped Ion Mobility Spectrometry-Electron Activated Dissociation Tandem Mass Spectrometry. *Anal. Chem.* **88,** 3440–3443 (2016).

284. Benigni, P. & Fernandez-Lima, F. Oversampling Selective Accumulation Trapped Ion Mobility Spectrometry Coupled to FT-ICR MS: Fundamentals and Applications. *Anal. Chem.* **88,** 7404–7412 (2016).

285. 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).

286. 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).

287. Hernandez, D. R. *et al.* Ion dynamics in a trapped ion mobility spectrometer. *Analyst* **139,** 1913–21 (2014).

288. Silveira, J. A., Ridgeway, M. E. & Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* **86,** 5624–5627 (2014).

136

289. Ridgeway, M. E., Silveira, J. A., Meier, J. E. & Park, M. A. Microheterogeneity within conformational states of ubiquitin revealed by high resolution trapped ion mobility spectrometry. *Analyst* **140,** 6964–6972 (2015).

290. Silveira, J. A., Danielson, W., Ridgeway, M. E. & Park, M. A. Altering the mobility-time continuum: nonlinear scan functions for targeted high resolution trapped ion mobility-mass spectrometry. *Int. J. Ion Mobil. Spectrom.* (2016). doi:10.1007/s12127-016-0196-1

291. 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).

292. 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,** 1785–93 (2011).

293. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 1–8 (2011).

294. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537,** 328–338 (2016).

295. Schwenk, J. M. *et al.* The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J. Proteome Res.* **16,** 4299–4310 (2017).

296. 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).

297. Hebert, A. S. *et al.* Improved Precursor Characterization for Data-Dependent Mass Spectrometry. *Anal. Chem.* **90,** 2333–2340 (2018).

298. Spahr, C. *et al.* Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I. Profiling an unfractionated tryptic digest. *Proteomics* **1,** 93–107 (2001).

299. 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).

300. 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,** M111.013185 (2012).

301. Broeckling, C. D., Hoyes, E., Richardson, K., Brown, J. M. & Prenni, J. E. DataSet-Dependent Acquisition enables comprehensive tandem mass spectrometry coverage of complex samples. *Anal. Chem.* acs.analchem.8b00929 (2018). doi:10.1021/acs.analchem.8b00929

302. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12,** 483–490 (2008).

303. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **2,** 43–50 (2003).

304. 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,** O111.013698 (2012).

305. Yi, E. C. *et al.* Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23,** 3205–3216 (2002).

306. Vincent, C. E. *et al.* Segmentation of precursor mass range using 'tiling' approach increases peptide identifications for MS1-based label-free quantification. *Anal. Chem.* **85,** 2825–2832 (2013).

307. Sabidó, E., Selevsek, N. & Aebersold, R. Mass spectrometry-based proteomics for systems biology. *Curr. Opin. Biotechnol.* **23,** 591–597 (2012).

308. Zhang, B., Kall, L. & Zubarev, R. A. DeMix-Q: Quantification-centered Data Processing Workflow. *Mol. Cell. Proteomics* mcp.O115.055475 (2016). doi:10.1074/mcp.O115.055475

309. Gillette, M. A. & Carr, S. A. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nat. Methods* **10,** 28–34 (2013).

310. Whiteaker, J. R. *et al.* A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat. Biotechnol.* **29,** 625–634 (2011).

311. Huttenhain, R. *et al.* Reproducible Quantification of Cancer-Associated Proteins in Body Fluids Using Targeted Proteomics. *Sci. Transl. Med.* **4,** 142ra94-142ra94 (2012).

312. Soste, M. *et al.* A sentinel protein assay for simultaneously quantifying cellular processes. *Nat. Methods* **11,** 1045–1048 (2014).

313. Yost, R. A. & Enke, C. G. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Anal. Chem.* **51,** 1251–1264 (1979).

314. Kondrat, R. W., McClusky, G. A. & Cooks, R. G. Multiple Reaction Monitoring in Mass Spectrometry/Mass Spectrometry for Direct Analysis of Complex Mixtures. *Anal. Chem.* **50,** 2017–2021 (1978).

315. Charles, T., Neuroscience, B. S. & Spectrometry, M. Peptide Quantification By Mass Spectrometry. *Mass Spectrom. Rev.* **15,** 213–240 (1996).

316. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nat. Methods* **9,** 555–566 (2012).

317. Kennedy, J. J. *et al.* Demonstrating the feasibility of large-scale development of standardized assays to quantify human proteins. *Nat. Methods* **11,** 149–155 (2014).

318. Anderson, L. & Hunter, C. L. Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. *Mol. Cell. Proteomics* **5,** 573–588 (2006).

319. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. & Aebersold, R. Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics. *Cell* **138,** 795–806 (2009).

320. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494,** 266–70 (2013).

321. 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).

322. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** M111.011015 (2011).

323. Gallien, S. *et al.* Targeted Proteomic Quantification on Quadrupole-Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **11,** 1709–1723 (2012).

324. Gallien, S., Kim, S. Y. & Domon, B. Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction. *Mol. Cell. Proteomics* 1630–1644 (2015). doi:10.1074/mcp.O114.043968

325. Abbatiello, S. E., Mani, D. R., Keshishian, H. & Carr, S. A. Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. *Clin. Chem.* **56,** 291–305 (2010).

326. Reiter, L. *et al.* MProphet: Automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **8,** 430–435 (2011).

327. Picotti, P. *et al.* High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* **7,** 43–46 (2010).

328. Kusebauch, U. *et al.* Human SRMAtlas : A Resource of Targeted Assays to Quantify the Complete Human Proteome Resource Human SRMAtlas : A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **166,** 766–778 (2016).

329. Matsumoto, M. *et al.* A large-scale targeted proteomics assay resource based on an in vitro human proteome. *Nat. Methods* **14,** 251–258 (2017).

330. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14,** 259–262 (2017).

331. Zauber, H., Kirchner, M. & Selbach, M. Picky: A simple online PRM and SRM method designer for targeted proteomics. *Nat. Methods* **15,** 156–157 (2018).

332. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12,** 35 (2014).

333. Masselon, C. *et al.* Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures. *Anal. Chem.* **72,** 1918–1924 (2000).

334. Zhang, N. *et al.* ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **5,** 4096–4106 (2005).

335. Bern, M. *et al.* Deconvolution of mixture spectra from ion-trap dat-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82,** 833–841 (2010).

336. Houel, S. *et al.* Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomic studies. *J. Proteome Res.* **9,** 4152–4160 (2010).

337. Wang, J., Pérez-Santiago, J., Katz, J. E., Mallick, P. & Bandeira, N. Peptide Identification from Mixture Tandem Mass Spectra. *Mol. Cell. Proteomics* **9,** 1476–1485 (2010).

338. Purvine, S., Eppel, J.-T., Yi, E. C. & Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3,** 847–850 (2003).

339. Silva, J. C. *et al.* Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **77,** 2187–2200 (2005).

340. Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **9,** 2252–2261 (2010).

341. Panchaud, A. *et al.* Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81,** 6481–8 (2009).

138

342.   Panchaud, A., Jung, S., Shaffer, S. A., Aitchison, J. D. & Goodlett, D. R. Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal. Chem.* **83,** 2250–2257 (2011).

343.   Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* **10,** 744–746 (2013).

344.   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).

345.   Bilbao, A. *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15,** 964–980 (2015).

346.   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-O111.016717 (2012).

347.   Ting, Y. S. *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol. Cell. Proteomics* **14,** 2301–2307 (2015).

348.   Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1,** 140031 (2014).

349.   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).

350.   Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32,** 219–223 (2014).

351.   Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8,** 291 (2017).

352.   Rosenberger, G. *et al.* Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat. Biotechnol.* **35,** 781–788 (2017).

353.   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).

354.   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).

355.   Vowinckel, J. *et al.* Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. *Sci. Rep.* **8,** 1–10 (2018).

356.   Bache, N. *et al.* A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *bioRxiv* (2018). doi:10.1101/323048

357.   Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12,** 258–264 (2015).

358.   Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14,** (2017).

359.   Kao, A. *et al.* Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **10,** M110.002212 (2011).

360.   May, J. C., Morris, C. B. & McLean, J. A. Ion mobility collision cross section compendium. *Anal. Chem.* **89,** 1032–1044 (2017).

361.   Geyer, P. E. *et al.* Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12,** 901 (2016).

362.   Specht, H. & Slavov, N. Routinely quantifying single cell proteomes: A new age in quantitative biology and medicine. 1–11 (2018). doi:10.7287/peerj.preprints.26821v1

363.   Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat. Commun.* **9,** 1–10 (2018).

364.   Shishkova, E., Hebert, A. S. & Coon, J. J. Now, More Than Ever, Proteomics Needs Better Chromatography. *Cell Syst.* **3,** 321–324 (2016).

365.   Thakur, S. S. *et al.* Deep and Highly Sensitive Proteome Coverage by LC-MS/MS Without Prefractionation. *Mol. Cell. Proteomics* **10,** M110.003699 (2011).

366.   Ikegami, T. & Tanaka, N. Recent Progress in Monolithic Silica Columns for High-Speed and High-Selectivity Separations. *Annu. Rev. Anal. Chem.* **9,** 317–342 (2016).

367.   Op de Beeck, J. *et al.* A novel nanoflow LCMS limited sample proteomics approach using micro pillar array columns (μPAC). in *16th Human Proteome Organisation World Congress* (2017).

368.   Kim, W., Guo, M., Yang, P. & Wang, D. Microfabricated monolithic multinozzle emitters for nanoelectrospray mass

spectrometry. *Anal. Chem.* **79,** 3703–3707 (2007).

369.   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).

370.   Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* **550,** 354–359 (2017).

Appendix

1) *Article 6:* Region and cell-type resolved quantitative proteomic map of the human heart
2) *Article 7:* Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery

Region and cell-type resolved quantitative proteomic map of the human heart

Sophia Doll[1,2], Martina Dreßen[3], Philipp E. Geyer[1,2], Daniel N. Itzhak[1], Christian Braun[4], Stefanie A. Doppler [3],
Florian Meier [1], Marcus-Andre Deutsch[3,5], Harald Lahm[3], Rüdiger Lange[3,5],
Markus Krane[3,5] & Matthias Mann [1,2]

The heart is a central human organ and its diseases are the leading cause of death worldwide, but an in-depth knowledge of the identity and quantity of its constituent proteins is still lacking. Here, we determine the healthy human heart proteome by measuring 16 anatomical regions and three major cardiac cell types by high-resolution mass spectrometry-based proteomics. From low microgram sample amounts, we quantify over 10,700 proteins in this high dynamic range tissue. We combine copy numbers per cell with protein organellar assignments to build a model of the heart proteome at the subcellular level. Analysis of cardiac fibroblasts identifies cellular receptors as potential cell surface markers. Application of our heart map to atrial fibrillation reveals individually distinct mitochondrial dysfunctions. The heart map is available at maxqb.biochem.mpg.de as a resource for future analyses of normal heart function and disease.

[1] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany. [2] Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen 2200, Denmark. [3] Department of Cardiovascular Surgery, German Heart Center Munich at the Technische Universität München, Munich 80636, Germany. [4] Forensic Institute, Ludwig-Maximilians-University, Munich 80336, Germany. [5] DZHK (German Center for Cardiovascular Research), Partner Site Munich Heart Alliance, Munich 80802, Germany. Correspondence and requests for materials should be addressed to M.K. (email: krane@dhm.mhn.de) or to M.M. (email: mmann@biochem.mpg.de)

142

The human heart beats more than two billion times in an average life span and each contraction is precisely controlled by an intricate interplay between electrical signals and mechanical forces. At the anatomical level, it is composed of four cavities, four valves, large arteries, and veins, which act in concert to achieve proper filling, ejection, contraction, and overall pump function. The heart's own blood supply is ensured by two coronary arteries. The human heart is composed of four major cell types—cardiac fibroblasts (CFs), cardiomyocytes, smooth muscle cells (SMCs), and endothelial cells (ECs)[1]. Their proportion with respect to number and volume, however, remains controversial. CFs are mesenchymal cells, which produce the extracellular matrix (ECM) scaffold of the heart and are thought to constitute more than half of all heart cells[2]. Cardiomyocytes are estimated to provide about 30% of the total cell number but account for over 70% of the total cardiac mass because of their large volume. In contrast, SMCs, which support the vascular system, and ECs, which form the interior lining of the heart, blood vessels, and cardiac valves, are generally believed to be much less abundant. However, these estimates have been challenged and a recent report claims that ECs are the largest cellular population within the heart[3].

In common with other muscle tissues, the heart is dominated by a small number of proteins involved in the contractile apparatus. It employs tissue-specific isoforms such as cardiac troponins, which are used in the diagnosis of myocardial infarction. From a physiological and pathophysiological perspective, it would be desirable to gain deeper insights into the molecular characteristics of the heart at the spatial and cellular levels. In particular, characterization of the healthy state of the human heart would be an important starting point to investigate heart disease, which—despite major progress remains the leading cause of death in developed countries and is rapidly increasing in developing ones[4].

Relatively little is known about the protein composition of the different regions and cell types of the heart. Previous studies have focused on defining differences between specific regions of the heart, or single-diseased heart compartments[5], or from nonhuman, or subcellular material[6,7]. Phosphoproteomic studies have also been applied for the analysis of mammalian hearts[8,9]. Moreover, other studies use transcriptomic approaches[10–12], which is an imperfect proxy for protein levels and their dynamics. However, proteins are the driving forces of the cellular machinery and they are involved in the control of virtually all physiologic
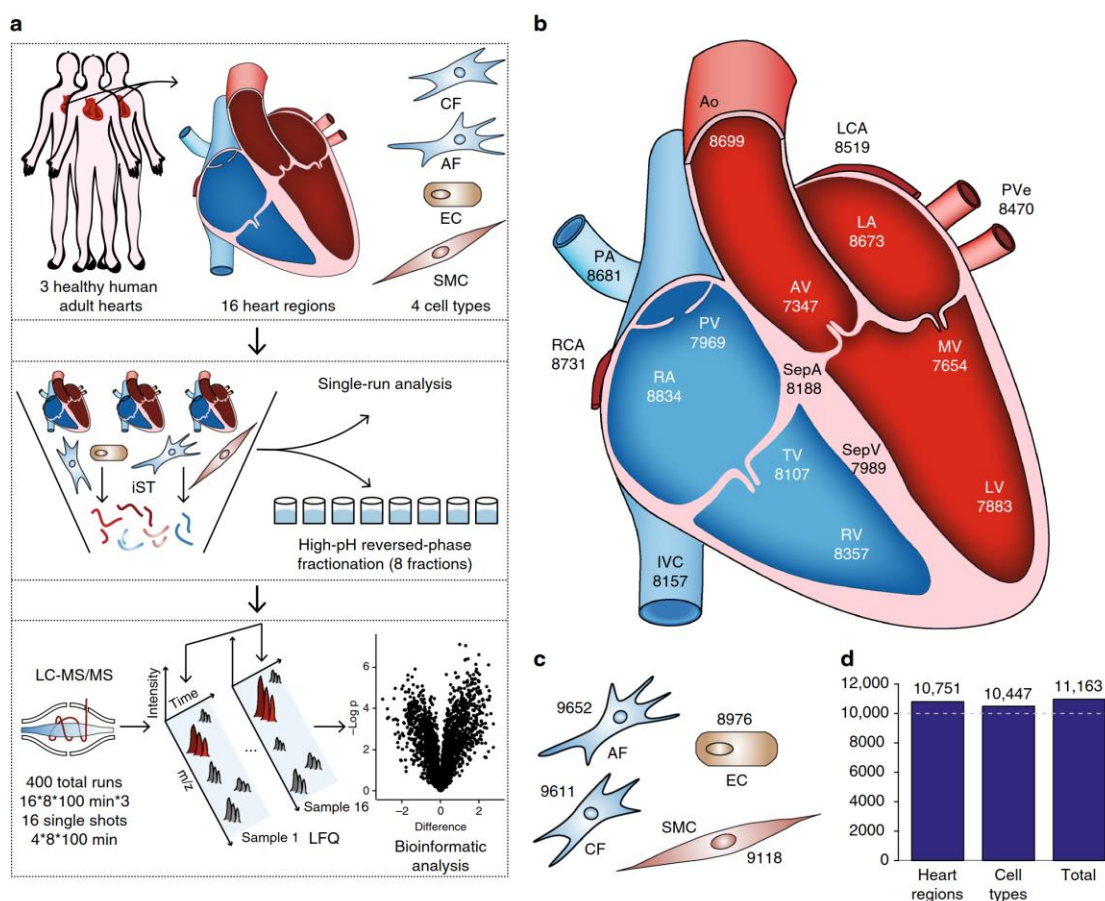

Fig. 1 The quantitative landscape of the human heart proteome. **a** Experimental design, including the source of material (upper panel), in-depth vs. single-run analyses (middle panel), and schematic depiction of the analytical workflow (lower panel). **b** Graphical illustration of the human heart showing the total number of quantified proteins in each region. **c** Quantified proteins in three cardiac cell types and adipose fibroblasts. **d** Bar plot of the total number of quantified proteins in all heart regions, cell types, and the entire data set
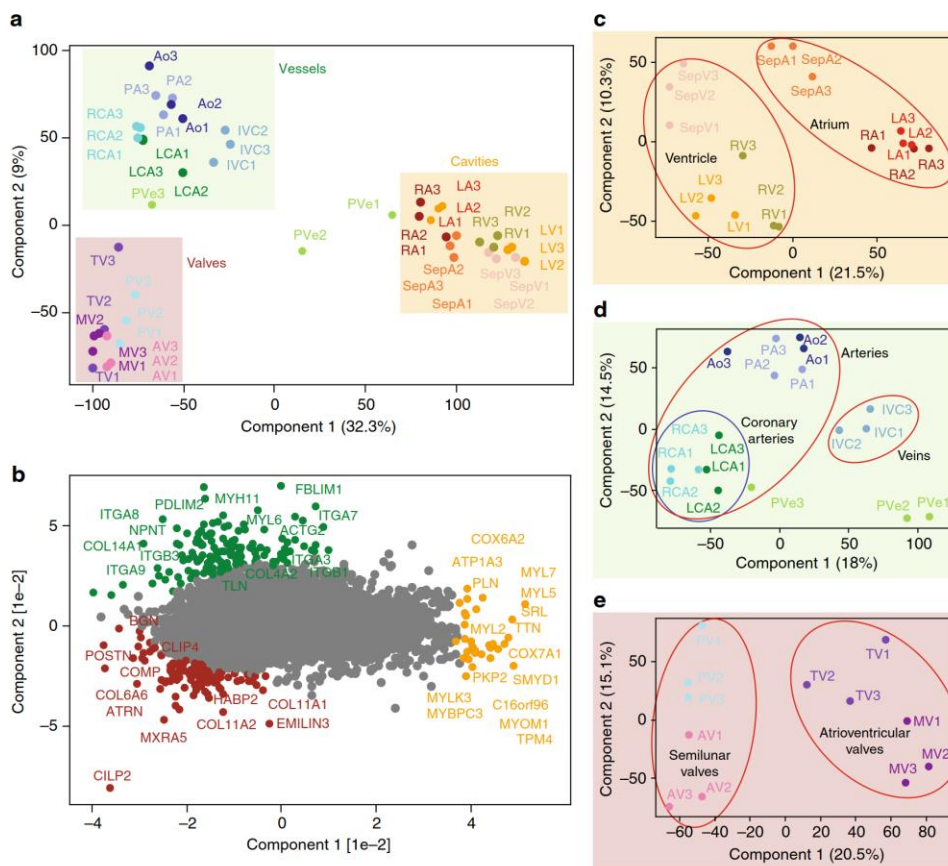
Fig. 2 Principal component analysis (PCA) of the 16 heart regions based on their proteomic expression profiles. **a** The proteomes of the cavities (RA, LA, RV, LV, SepA, and SepV), vessels (Ao, PA, RCA, LCA, IVC, and PVe), and valves (TV, MV, AV, and PV) depicted by replicate number (individuals 1, 2, and 3). The first and second component segregate the heart areas and account for 32.3 and 9% of the variability, respectively. **b** Proteins driving the segregation between the three heart areas. **c** Cavities segregate into the ventricular and atrial part, **d** vessels into coronary arteries (RCA, LCA) and outgoing vessels (Ao, PA), and **e** valves into ventricular (MV, TV) and semilunar valves (AV, PV)

events. The high dynamic range of the muscle proteome presents a formidable challenge to the comprehensive analysis of the heart at the level of expressed proteins. This is because very abundant proteins make it difficult to detect low abundant regulatory proteins in the same sample. The majority of studies only identified a few thousand proteins, and there is a paucity of studies of the human, nondiseased heart, because of the difficulty in obtaining the relevant tissue.

A global protein expression "footprint" of the healthy heart can be used as a reference library to compare against footprints of malfunctioning hearts in the search for biomarkers, therapeutic targets, or disease signatures. Recent advances in MS-based proteomics technology now allow the identification of very deep proteomes[13,14]. Our group has already established proteomics maps of the mouse liver and brain[15,16] and analyzed skeletal muscle in considerable depth and sensitivity[17,18]. Here, we set out to generate a spatial and cell-type-resolved proteomic map of the healthy human heart. To this end, we measured 16 regions of three human hearts, as well as primary cell types. We employed high-sensitivity sample preparation, peptide fractionation, and an advanced label-free LC–MS workflow to quantify a total of more than 11,000 proteins. Our results establish proteomic differences

between heart regions, suggest functional differences, and pinpoint potential cell-type markers. To illustrate the usefulness of the heart proteomic map, we apply it to define molecular changes in patients suffering from atrial fibrillation (AFib).

Results

Establishing a proteomic map of the human heart. Three adult hearts were obtained from male trauma victims aged 21–47 years with no apparent adverse heart condition (Supplementary Table 1). We selected a total of 16 anatomically defined regions from each heart for MS analysis (Fig. 1a, b): the atrial and ventricular septa (SepA and SepV) separating the atria and ventricles, respectively; the right atrium (RA) and right ventricle (RV) connected via the tricuspid valve (TV); the left atrium (LA) and left ventricle (LV) linked via the mitral valve (MV); the right and left ventricles connected to the pulmonary artery (PA) and aorta (Ao) via the pulmonary and aortic valves (PV and AV); the inferior vena cava (IVC) collecting deoxygenated blood; the pulmonary vein (PVe) carrying oxygenated blood; and the main right and left coronary arteries (RCA and LCA) supplying the heart with oxygen-rich blood. In addition, we isolated CFs, ECs,
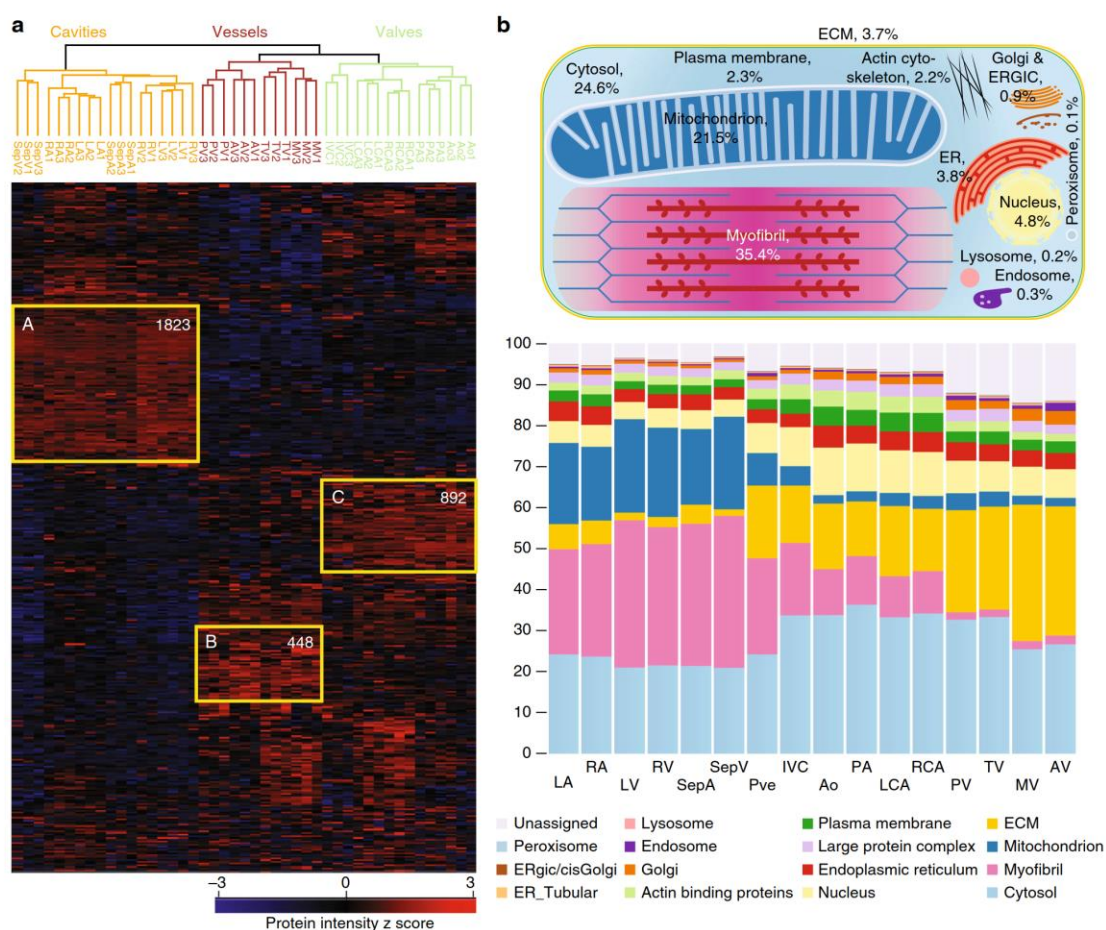
144

Fig. 3 Proteins differentially expressed across the different heart areas. **a** Heat map of z-scored protein abundances (LFQ intensities) of the differentially expressed proteins (ANOVA, FDR < 0.05) after unsupervised hierarchical clustering reveals proteins significantly upregulated in the cavities, vessels, or valves (highlighted in yellow: A, B, and C). **b** The upper panel shows a schematic of an average heart cavity cell, where organelles are sized according to their contribution to total protein mass. Percentages are taken from the median of all cavities and scaled to account for unassigned proteins. The lower panel shows the contribution of each organelle to cellular protein mass, as a percentage of the total, in each heart region

and SMCs from patients undergoing cardiovascular surgery (Fig. 1c).

After tissue homogenization in liquid nitrogen, we performed all sample preparation using the "in-StageTip (iST) method" (see "Methods" section), reducing sample contamination, loss, preparation time, and increasing quantification accuracy[19]. The recently described "loss-less" nano-fractionator enabled efficient fractionation of a total of only 30 μg of peptides into eight fractions, of which a third of each fraction was loaded in the subsequent LC–MS step[20]. The resulting 400 samples were analyzed with a state-of-the-art label-free workflow on a quadrupole– Orbitrap mass spectrometer (Fig. 1a).

Analysis in the MaxQuant environment using a false-discovery rate (FDR) of less than 1% at the peptide and protein levels[21], identified a total of 181,814 sequence-unique peptides. These assembled into 11,236 protein groups. Many high-abundance proteins had very high sequence coverage—such as 100% for myosin regulatory light chain 2 (MYL7)—whereas median coverage of all proteins was ~38%. The MaxLFQ algorithm[22]

quantified 11,163 proteins, 10,751 in the 16 heart regions, and 10,447 in the noncardiomyocyte cell types, including AFs (Fig. 1d). Proteomic depth was high in all regions, including the four cardiac valves, in which we identified a mean of about 7800 proteins despite the fact that it mainly consists of ECM. To put this number in perspective, reanalysis of the "human draft proteome" heart data[23] with the settings used here revealed that our study identified more than three times as many proteins, most of which were of low abundance (Supplementary Fig. 1). For further analysis, we only considered a subset of 8908 proteins with quantitative values in all biological triplicates of at least one heart region.

Signal intensities for the quantified proteins spanned more than six orders of magnitude, while only six proteins—myosin 7 (MYH7), titin (TTN), cardiac muscle-specific actin (ACTC1), alpha-actinin-2 (ACTN2), and hemoglobin (HBA1 and HBB) represented 25% of the total protein molecules in cavities, with similar values in vessels and valves (Supplementary Fig. 2a–f). A large amount of hemoglobin remained, despite extensive washing
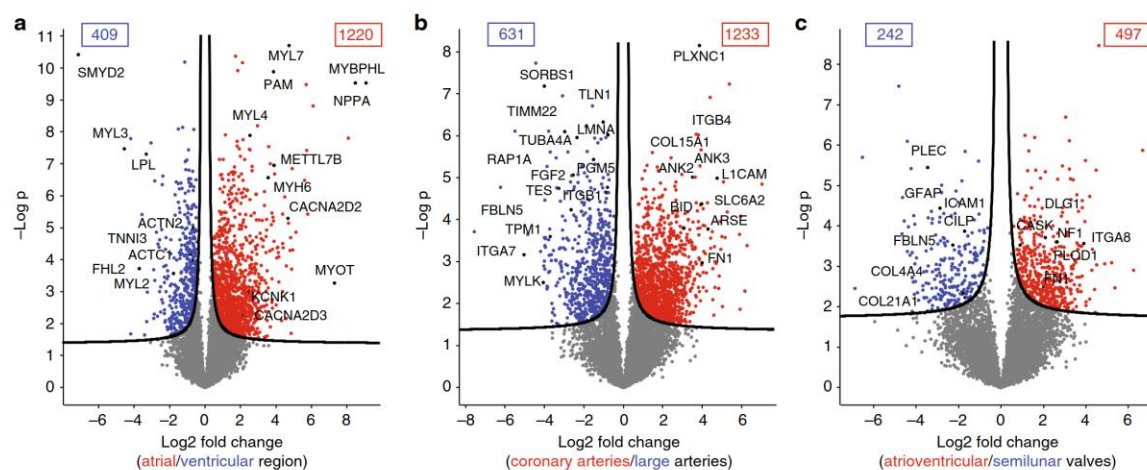
145

Fig. 4 Proteins differentially expressed in human heart regions. Volcano plots of the p values vs. the log2 protein abundance differences between regions, with proteins outside the significance lines colored in red or blue (FDR < 0.05). p values are calculated from the data of three healthy hearts. **a** Ventricular (LV, RV, and SepV) compared to the atrial (LA, RA, and SepA) regions, **b** coronary arteries (LCA, RCA) compared to arteries (Ao, PA), and **c** semilunar (AV, PV) compared to atrioventricular (TV, MV) valves

of the samples with PBS, since hearts of trauma victims cannot be perfused. Due to efficient peptide fractionation, our measurement covered regulatory proteins such as transcription factors GATA4, GATA6, TBX20, TBX3, and TBX5 controlling cardiac-specific gene expression (Supplementary Fig. 3).

To assess quantitative reproducibility, we analyzed several samples in technical triplicates. Pearson correlation coefficients (0.97–0.99) were on par with, or exceeded the values previously achieved in cell line systems[24] (Supplementary Fig. 4). Likewise, we observed high correlation values between biological replicates; ranging from 0.83 (PVe) to 0.95 (LA) (Supplementary Fig. 5). As these values incorporate any differences due to postmortem sample treatment, we conclude that our results from three individuals can likely be generalized to the adult male population at large. Raw data and MaxQuant results are provided online and the human cardiac proteome resource is available in our online database MaxQB[25] (see below).

Comparative analysis between anatomical areas of the heart. For an overall assessment of proteomics similarities and differences of the 16 heart regions, we employed principal component analysis (PCA). Cavities, vessels, and valves clearly clustered separately with the samples from different individuals tightly grouped together (Fig. 2a). The only exception was the PVe from patient 1 and 2, which clustered closer to the atrium than other vessels. This is readily explained by the difficulty of resecting PVe without contamination from LA; thus, PVe samples were excluded from the subsequent analyses.

The segregation of the three groups was mainly driven by MYL7, MYL5, cytochrome c oxidase subunit 7A1 (COX7A1), sarcalumenin (SRL) and TTN (highlighted in red, driving segregation of the cavities), and collagen proteins, such as COL4A2, COL14A1, and integrins, including ITGA7, ITGA8, and ITGB1 (highlighted in blue, segregating vessels), as well as biglycan (BGN), COL11A1, COL11A2, and COL6A6 (highlighted in purple, segregating valves) (Fig. 2b). As these proteins reflect known biological differences between the cardiac cavities and the vessels and valves that are rich in ECM components, they serve as positive controls of our proteomic analysis. Furthermore, the PCA analysis highlighted several interesting candidates, such as

cartilage intermediate layer protein 2 (CILP2) and MXRA5 (valves), nephronectin (NPNT), a functional ligand of ITGA8 and ITGB1 (vessels), and uncharacterized proteins such as C16orf96 (cavities).

Each of the three main clusters exhibited further subgroupings. Heart cavities were divided into atrial (RA, LA, and SepA) and ventricular (RV, LV, and SepV) parts and within them atrial and ventricular septa were separated from atria and ventricles, respectively. Furthermore, there was a moderate but clear distinction of the left and right side of the heart (Fig. 2c). The vessel group subdivided into large arteries (Ao and PA) and large veins (IVC) (Fig. 2d). Within the arteries, the RCA and LCA formed a subcluster, demonstrating differences between coronary and large arteries at the proteomic level. Finally, both atrioventricular valves (MV and TV) clustered together, whereas semilunar valves (AV and PV) formed a separate group (Fig. 2e). The main drivers of the PCA separation are highlighted in Supplementary Fig. 6.

For a functional view of the proteomic differences in the human heart, we performed unsupervised hierarchical clustering of the 6807 proteins with statistically different expression across the heart regions (FDR < 0.05) (Supplementary Data 1). This again clustered individuals in all but one case (RV of one individual), followed by cavities, vessels, and valves with their subdivisions (Fig. 3a). The heat map shows one major cluster of highly and coexpressed proteins for each of the three anatomical areas. Gene ontology and GSEA[26] revealed that proteins in cluster A (high expression in the cavities) were enriched ($p < 10^{-12}$) in terms of cardiac muscle contraction, Z disc, and sarcomere organization compared to clusters B and C (high expression in the vessels and valves, respectively). The terms mitochondria and respiratory electron transport chain were also enriched ($p < 10^{-55}$) in this cluster, concordant with the large number of mitochondria to ensure sufficient amounts of ATP for continuous muscle contraction (Supplementary Data 2). Thus, our proteomic data provide a global protein expression basis for the functional specialization of cardiac muscle tissue.

To provide insights into the organelle sizes in the heart proteome at a quantitative level, we used the proteomic ruler approach to estimate copy numbers per cell[27] together with

146

subcellular localization annotations[28,29]. We calculated an approximate total protein content of 1 ng per diploid nucleus and found that a heart cavity muscle cell has an approximate volume of 5 pL per nucleus (note that about 30% of all cardiomyocytes have two or more nuclei[2]). These values are roughly double that of mouse heart muscle cells[30]. Estimated protein copy numbers per diploid nucleus and protein concentration[31] across our samples ranged from ~10 to 10^9 and <0.1 nM to 200 µM, respectively (Supplementary Data 3). We found that mitochondria constituted 21% of protein mass in the cavities (Fig. 3b, "Methods" section, and Supplementary Data 3). This compares to 7% of mitochondrial protein mass in HeLa cells[28] and 3% in valves and vessels, demonstrating the immense aerobic respiration in cardiac muscle cells localized in the atrial and ventricular part of the heart at a quantitative level.

To further mine our quantitative and in-depth proteome resource, we used volcano plots[32] to compare expression differences within the three anatomical areas of the heart. We specifically focused on proteins that were in the top 75% in abundance and only identified in one of the two regions that we compared (Supplementary Data 4).

The atrial vs. ventricular proteome. Although they are both heart muscles, the main role of atria is to collect and transfer pulmonary and systemic blood, whereas ventricles need to pump the blood throughout the entire body. Consistent with these different functions, we found drastic differences in their proteomes, with 1220 (13.7%) proteins showing significantly higher expression in the atria and 409 (4.6%) displaying higher expression in the ventricles (Fig. 4a). As expected, mitochondrial proteins were more abundant in ventricles ($p < 10^{-150}$ by GSEA analysis, Supplementary Data 5). Cardiomyocytes use fatty acids as their main energy source and ventricular myocytes have higher energy demands due to the greater force of contraction. Accordingly, lipid metabolic processes were overrepresented in the ventricular region ($p\ 7 \times 10^{-26}$), exemplified by the ~tenfold increased expression of lipoprotein lipase (LPL). Likewise, it was enriched in muscle contraction ($p\ 9 \times 10^{-27}$), due to increased cardiomyocyte size, as estimated by the proteomic ruler ("Methods" section). The known markers for ventricles vs. atria, such as MYL2, MYL3, and LPL were clearly recovered as such, and our data set contains many additional ones, including the lysine methyltransferase SMYD2, which is thought to have a role in myocyte function[33]. Several interesting candidates were only identified in the ventricles but not in the atria, such as the probable histone demethylase JMJD1C, the ubiquitin ligase TRIM38, the tumor suppressor RASSF8, and the uncharacterized KIAA1324L protein. These proteins have not been associated with ventricular functions before and suggest starting points for exploring their role in heart physiology. Proteins previously reported as atrium specific[34–36], including myosin 6 (MYH6), peptidyl-glycine alpha-amidating monooxygenase (PAM), and natriuretic peptides A (NPPA) displayed ten to several hundred-fold higher abundance in the atrium. Interestingly, these proteins are only highly expressed in the ventricular regions under pathological conditions—for example, cardiac hypertrophy leads to elevated PAM levels in ventricles[37]. We found potassium ion channels predominantly in the atrial part, such as KCNK1, which induces background currents[38], and calcium-dependent ion channels, including CACNA2D2 and 3 (>fourfold), as well as gap junction GJA5 (>ninefold), reflecting the presence of the sinus and AV node, which generate the electrical impulse for heart contraction. Others, such as CACNA1C, which play an important role in excitation–contraction coupling in the heart, were equally expressed in the atria and ventricles (see also Supplementary Fig. 7).

The high protein sequence coverage encouraged us to investigate isoform-specific expression patterns of sarcomeric proteins (Supplementary Data 6). These isoforms are of particular importance because their altered expression has been associated with diverse cardiac dysfunctions[39] and because troponins are routinely used biomarkers for myocardial infarction[40]. The myosin isoform families MYH6, MYH7, MYL2-7, and MYL8 localized highly specifically to atria or ventricles, confirming the regional specificity of our data set despite high sequence identity, for instance, over 80% between MYHs[41].

Myocardial infarction remains one of the largest causes of death and although rapid ELISA tests against cardiac troponins TNNT2 and TNNI3 play a crucial role, further improvements in diagnosis would be of great clinical benefit[42,43]. Here, we found that TNNT2 was more abundant (>twofold) than TNNI3 in all cavities and that their expression largely correlated across regions (Pearson correlation: 0.99). Interestingly, the cardiac isoform myosin-binding protein C3 (MYBPC3), which participates in stabilizing sarcomere structures, displayed a strikingly similar protein expression profile to cardiac troponins and was similarly abundant to TNNT2 (about 1.5-fold higher abundant) (Supplementary Fig. 8). Moreover, it can be detected by antibody- and MS-based approaches in human plasma after myocardial injury[44,45], showing that it can be used as an useful additional parameter to monitor myocardial infarct.

Finally, while the RV and LV did not show any significantly altered protein expression, the protein myotilin, which stabilizes thin filaments during muscle contraction was much more abundant in LA compared to RA (>100-fold; Supplementary Fig. 9).

Large vs. small arteries proteomes. Coronary arteries are relatively small as they supply the heart itself with blood; however, their malfunction is responsible for the high prevalence of coronary artery diseases (CADs), affecting more than 16 million individuals in the United States alone[46]. Overall, 1233 (13.8%) of quantified proteins were significantly more abundant in coronary arteries (RCA and LCA) and 631 (7.1%) in large arteries (Ao and PA) (Fig. 4b). Proteins involved in mitochondrial functions, collagen proteins, and integrins, such as COL15A1 and ITGB4 were highly enriched in the coronary vs. large arteries. Fibronectin (FN1) showed a 12-fold increase and has been previously associated with CADs[47], although it would need to be investigated whether or not it is a specific marker for CAD patients. Arylsulfatase E (ARSE) is a constituent of artery walls where it regulates the composition of cartilage, and we found it to be >20-fold more abundant in coronary arteries.

Large arteries showed significant ($p < 3 \times 10^{-42}$) enrichment for cytoskeleton proteins and proteins involved in cell junction, consistent with the higher structural demands on them. For instance, fibulin 5 (FBLN5) was six- to tenfold higher expressed in PA and Ao than the other heart compartments (Supplementary Fig. 10). It is required to form the elastic lamina, has a protective role against vascular injury, and its downregulation has been associated with aortic aneurysm[48]. Interestingly, fibroblast growth factor 2 (FGF2) and ras-related protein (RAP1A), described as a key regulator of FGF2-induced angiogenesis[49], were three- to sixfold more abundant in the large arteries. As we had not identified FGF2 in a deep plasma proteome previously[50], it is unlikely to derive from blood remnants and may instead represent an ECM-bound form[51]. A total of 92 medium- to high-abundance proteins were exclusively quantified in the large arteries (Supplementary Data 4). This included the key focal
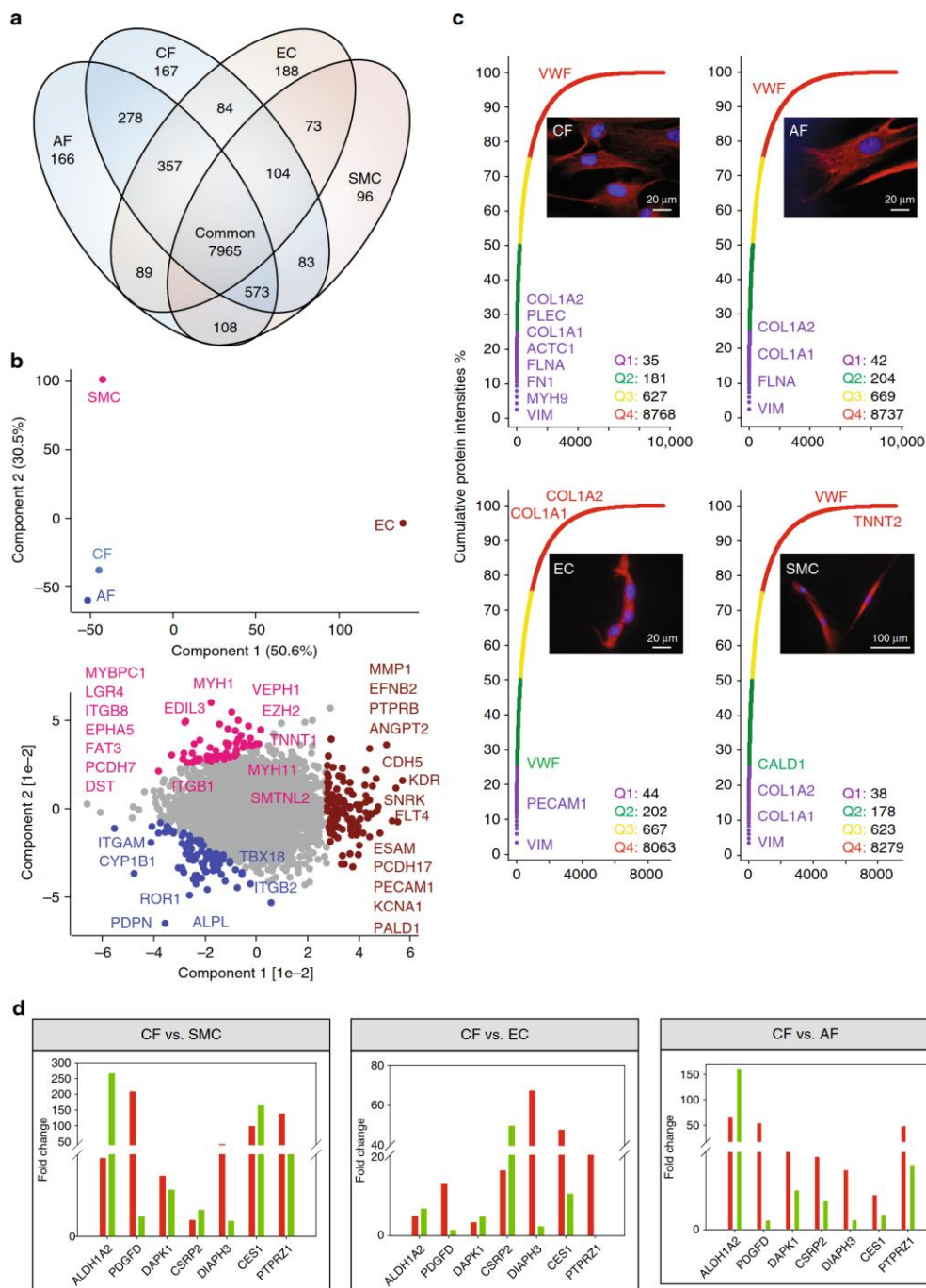
Fig. 5 Comparative analysis of the cell-type proteomes. **a** Commonly and exclusively quantified proteins in three cardiac cell types and adipose fibroblasts. **b** PCA comparing the four cell types based on component 1 and 2, which accounted for 50.6 and 30.5% of the variability, respectively. **c** Cumulative protein abundances for each cell type and total number of proteins constituting the quantiles (Q1–Q4). The corresponding cell types are illustrated with immunofluorescence pictures at ×100 magnification. **d** RT-qPCR (green) and proteomic (red) fold-changes of the indicated genes in CF compared to all other cell types

148

adhesion protein SORBS1, COL26A1, and MYLK and MYH11, which are both involved in smooth muscle contraction, reflecting the higher proportion of SMCs in the large arteries' wall compared to coronary arteries.

The atrioventricular vs. semilunar valves proteome. The atrioventricular valves separating atria from ventricles (TV and MV) are morphologically quite different from the semilunar valves (AV and PV) preventing backflow of blood from aorta or pulmonary artery to ventricles. We found that only 497 (5.6%) proteins were significantly more abundant in the atrioventricular valves and 242 (2.7%) were more abundant in the semilunar valves (Fig. 4c). Valves are composed of highly organized ECM proteins and changes in its composition and their possible release are expected during valve deterioration, leading to dysfunction and failing heart valves[52–54]. Although the overall changes were limited, we found that among the ECM proteins, plectin isoform 3 and GFAP were highly expressed in the semilunar valves, whereas peripheral plasma membrane CASK, collagen enzyme P4HA1, integrin ITGA8, and neurofibromin (NF1) were significantly higher in atrioventricular valves.

Cell-type-resolved proteome of the human heart. Our region-resolved proteome achieved great depths, but as we used homogenized tissue, we do not have cell-type-specific information about the origin of the proteins. To address this, we isolated CFs, ECs, and SMCs from tissue samples harvested during cardiac surgery (Supplementary Table 2, "Methods" section). AFs were included to help in defining the CF-specific proteome. Cardiomyocytes were not investigated because of the impossibility to culture these cells from surgical biopsies. We achieved highly purified cell populations with values for CFs, AFs, ECs, and SMCs of 96%, 97%, 96%, and 92%, respectively (Supplementary Fig. 11a–d). Of a total of 11,236 different proteins, 7965 were identified in all four cell types, indicating that the majority of the cardiac cell proteome is expressed in its major cell types (Fig. 5a). We found high correlation (0.92) in protein expression between the fibroblast cell types (CF and AF), whereas SMC and EC were somewhat less related (0.81) and this is also reflected in the PCA (Supplementary Fig. 12 and Fig. 5b). On average, the 40 most abundant proteins accounted for 25% of the total protein mass in all four cell types (Fig. 5c). Consistent with the mesodermal origin of these cell types, vimentin (VIM) was the most abundant protein, accounting for 3% of the total protein mass. In conjunction with LARP6, VIM stabilizes type I collagen mRNAs, leading to upregulation of the collagens COL1A1 and COL1A2[55]. We found the collagens in the top quartile (Q1) of expression in CFs, AFs, and SMCs, whereas they were among the least abundant proteins (Q4) in ECs. Cell-type-enriched proteins—those with at least twofold higher expression in one of the cell types compared to all others—are listed in Supplementary Data 7.

CF-enriched cell surface markers. Over the past decade, CFs have been shown to hold great promise as a potential target population for cardiac regenerative therapies[1,56,57]. Selection and targeting of CFs, however, remains challenging and currently relies on unspecific CF markers, including VIM, discoidin domain-containing receptor 2 (DDR2), periostin (POSTN), protein S100A4, ACTA2, platelet-derived growth factor receptors PDGFRα and β, T-box transcription factor TBX18, and the THY1 membrane glycoprotein[58]. Among these, only PDGFRB, S100A4, and ACTA2 showed at least twofold enrichment in CFs compared to ECs and SMCs, whereas all other currently employed CF markers were not enriched in CFs. Remarkably, compared to another fibroblast cell type (AFs), none of these markers were

even twofold enriched. GATA4 and TBX20 have been reported as specific CF markers[59]. GATA4 was indeed only identified and quantified in CFs; however, TBX20 was fourfold more abundant in SMCs compared to CFs.

Globally, 609 (5.8%) proteins were specifically enriched in CFs compared to AFs, ECs, and SMCs. These encompassed 25 cell membrane receptors (Supplementary Data 8). The presence of the tyrosine kinase ROR1 in CFs has not been reported before but we found it to be 200-fold more abundant than in the other cardiac cell types. Activin receptor ACVR1 is required for normal heart development[60] and it was also one of the most highest expressed proteins in CFs. Natriuretic peptide receptor NPR3 has a central role in vasodilatation, is known to be present in CFs,[61] and our data showed sixfold higher expression in CFs. The drug target hepatocyte growth factor receptor MET, was also increased in CFs (fourfold higher than ECs and SMCs). BDKRB2, the receptor for bradykinin plays a pivotal role in the cardiovascular system by regulating blood pressure. Interestingly, we exclusively identified it in fibroblasts, with 11-fold higher expression in CFs compared to AFs. Likewise, protein levels of the cell membrane phosphatase PTPRZ1 were more than 40-fold higher in CFs compared to all other investigated cell types, a finding supported by qPCR. Interestingly, the direction of expression changes was concordant between the mRNA and protein levels. Importantly, however, for this and six other genes, the fold-changes indicated by qPCR were not predictive of the actual protein-level changes (Fig. 5d). Our data provide a catalog of CF-enriched marker candidates that hold promise for better definition and targeting of human CFs.

EC and SMC proteome. Two further major cell types of the human heart include ECs, which form the inner lining of heart blood vessels and SMCs, the major constituents of the heart vasculature. ECs contain numerous storage granules filled with von Willebrand factor (VWF), which is involved in hemostasis. VWF was one of the most abundant proteins in ECs, whereas it was among the least abundant in CFs, AFs, and SMCs (Fig. 5c). The platelet endothelial cell adhesion molecule (PECAM1/CD31) was among the most abundant (Q1) proteins in ECs. Furthermore, EC-specific proteins ESAM (>22-fold) and ESM1 (exclusively) were overrepresented. Proteins involved in blood vessel morphogenesis such as the VEGF receptors FLT1, FLT4, KDR, as well as EPHB4 and its ligand Ephrin B2 showed 3- to 250-fold higher expressions compared to the other cell types. Comparing our data with single-cell transcriptomic data[62] revealed that all genes identified as EC specific compared to CF (except CAV2, which only displayed a moderate increase) also showed at least several 10-fold upregulation in ECs at the proteomic level (Supplementary Data 8).SMCs contain the same muscle-contracting proteins as cardiac cells but do not have troponin. The low levels of cardiac TNNT2 detected (>100-fold less than in cavities), are likely due to the 8% impurity of isolated SMCs. In place of troponins, caldesmon (CALD1), which was among the highest abundant proteins in SMCs, blocks the myosin-head binding site on actin filaments[63]. Proteins segregating the SMC group (Fig. 5b, highlighted in pink) from EC, AF, and CF included typical smooth muscle proteins such as SMTNL2 and MYH11 of which we quantified three splice variants and two were exclusively identified in SMCs. Our analysis also opens up for the investigation of new epigenetic mechanisms, for instance, based on the very significant enrichment for EZH2-regulated proteins ($p < 10^{-34}$).

Clinical application of the heart map to atrial fibrillation. Having generated a map of the healthy human heart, we next investigated if it could serve as a reference to pinpoint molecular
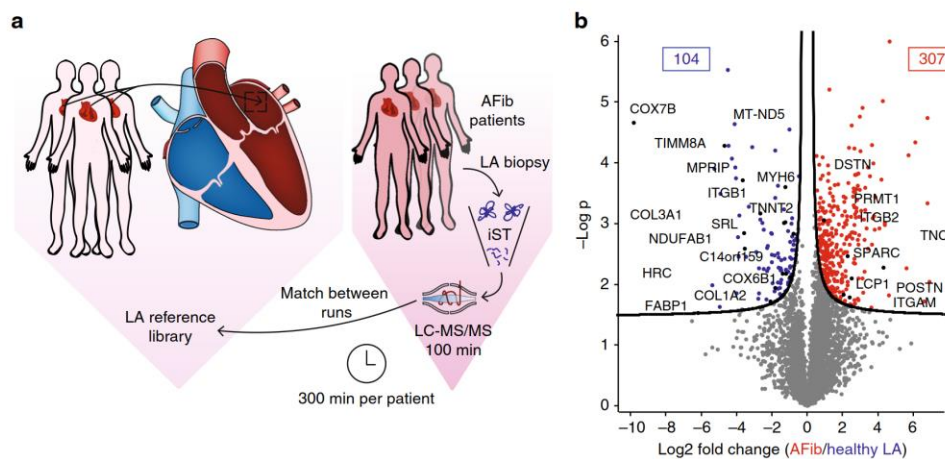
Fig. 6 Clinical application of the healthy human heart atlas to atrial fibrillation. **a** Experimental workflow: LA tissues from three atrial fibrillation patients (AFib) were single-runs of technical triplicates. Data were matched against the healthy human LA library. **b** Volcano plot of the p values vs. the log2 protein abundance differences in AFib compared to healthy LA. Significantly up- and downregulated proteins are highlighted in red and blue, respectively (FDR 0.05)

differences between healthy and diseased tissue. To this end, we collected LA samples from three patients suffering from AFib (Supplementary Table 3), the most common heart arrhythmia and a major cause of mortality[64]. We applied a single-run method, in which an in-depth measurement of the proteomic system in question serves as a reference set of identified peptides for deep and high-throughput single-run measurements[65] to assess the proteomic changes in AFib patients from minimal material and in a timely manner. We found that combining the iST sample preparation, our established healthy reference heart "library," and single-run, triplicate measurements, any cardiac sample can be profiled in less than two days, of which only 300 min are MS-measuring time (Fig. 6a). In this way, we quantified an average of 3681 proteins for the healthy LAs and 4147 proteins in the AFib group, with excellent average Pearson correlations for technical and "biological replicates" of 0.97 and 0.93, respectively (Supplementary Fig. 13). In the AFib group compared to healthy samples, 104 proteins were significantly downregulated and 307 were upregulated (Fig. 6b and Supplementary Data 9). Proteins with increased expression in AFib are involved in ribonucleoprotein complexes and transcription. Some ECM proteins, such as DSTN, ITGB2, ITGAM, and FLNB, were significantly upregulated in the AFib group, whereas others, including COL1A2, COL3A1, and ITGB1 were downregulated. These results point to a reorganization of the ECM in AFib, explaining previous observations at the level of expressed proteins[66]. There is also evidence of significant contractile remodeling in AFib with several-fold lower expression of TNNT2, HRC, MYH6, SCN5A, and SRL, suggesting disruption of the cardiac tissue. Furthermore, MPRIP, a protein that has been previously associated with an increased number of stress fibers when downregulated[67] showed 12-fold lower expression. The most significantly downregulated proteins in the AFib group (Fig. 6b and lower yellow box in Supplementary Fig. 14) were enriched for "mitochondrion" ($p < 10^{-100}$). This included the most significantly downregulated protein (COX7B) and two other key mitochondrial proteins—IMMT and TIMM8A, all of which were >25-fold less abundant than in healthy tissue. Mitochondrial dysfunction has already been reported in AFib[68–71]; however, the broader molecular nature of their defects and whether they are different between patients are not fully understood.

Interestingly, inspection of the hierarchical clustering plot within the AFib group revealed distinct and nonoverlapping clusters of up- and downregulated proteins for each of the three patients (orange and green clusters, Supplementary Fig. 14), whereas nondiseased biological controls showed similar expression to each other. Although the number of AFib patients is much too small to derive a general signature, our data clearly show that the mitochondrial defects reflected in the proteomes are very different between individuals. These proteomic patterns point to a potential molecular subclassification of AFib patients. These observations, however, will require a more thorough analysis including a larger patient cohort.

Discussion

Creating anatomical and cellular maps increases our understanding of human biology and diseases[15,16,72]. Here, we used "loss-less" high-pH reversed-phase fractionation and high resolution, quantitative MS to generate a heart region, and cell-type-resolved human heart map. Starting from low microgram sample amounts, we quantified over 11,000 proteins, representing by far the deepest proteome of the healthy human heart, which is available in the online database MaxQB[25]. The ability to work with minimal starting material enables in-depth proteomics analyses from heart biopsies that can be obtained during surgery. Furthermore, our streamlined proteomics workflow enables the profiling of any cardiac sample in less than two days, a realistic time frame for future clinical application. At the anatomic level, we found that the 16 heart regions clustered into the expected three main areas (cavities, vessels, and valves). Binary comparison of subgroups, such as the atria and ventricles, provides crucial information to understand the basis for atrioventricular differences in healthy as well as diseased human hearts, a precondition to identify more specific and reliable biomarkers. To complement the region-resolved heat map, we also established a comprehensive proteomic map of three noncardiomyocyte cell types. This should be particularly useful in future studies to better define and target the CF population. CFs are activated into myofibroblasts after acute myocardial infarction, leading to increased ECM production and wound healing by scar formation within the infarction area. The direct reprogramming of resident CFs after myocardial infarction into induced cardiomyocytes or cardiac

150

progenitor cells is currently a promising strategy for cardiac regeneration. Our quantitative proteomic data question the specificity of currently used CF markers while providing a promising panel of enriched cell surface markers for therapy. There is a lack of reliable biomarkers for aortic aneurysms (enlargement of the Ao) or more importantly aortic dissection (tear in the wall of the Ao). Currently, both diagnoses rely on clinical examinations and laborious imaging techniques. Our deep quantitative proteome of large arteries, in particular the Ao, could help in establishing a healthy baseline in future studies aiming to define protein expression indicative of these conditions. Likewise, our human cardiac valve proteome can be used as a background for future studies aiming to uncover biomarkers indicative of cardiac valve deterioration. Finally, we show that patients suffering from AFib present both common and distinct proteome profiles, potentially pointing to individual-specific disease manifestation. While our investigation is only a first step, it opens up as a yet-unexplored molecular classification at the level of expressed proteins. Further directions in human heart proteomics could include region- and cell-type-specific mapping at higher resolution, investigation of PTMs, and the combination of proteomics with detailed mechanistic investigations of disease etiology.

Methods

Tissue preparation. In total, 16 healthy heart regions from three adult male individuals (Supplementary Table 1) were collected less than 72 h postmortem during an autopsy after a court order. The hearts of the subjects did not present any relevant injury or signs of cardiac malfunction and were therefore defined as healthy. The 16 heart regions included four main vessels (aorta (Ao), pulmonary artery (PA), vena cava inferior (IVC), and pulmonary vein (PVe)), four heart cavities (right atrium (RA), left atrium (LA), right ventricle (RV), and left ventricle (LV)), four heart valves (tricuspid valve (TV), pulmonary valve (PV), aortic valve (AV), and mitral valve (MV)), the ventricular septum (SepV), the atrial septum (SepA), the left coronary artery (LCA), and the right coronary artery (RCA) were explanted by an official medicolegal expert. Samples were stored at −80 °C after collection. The investigation was approved by the local ethical committee of the Medical Faculty of the Technical University of Munich (project no. 247/16s). The ethical committee explicitly approved the use of human samples in the context of trauma.

Cell isolation. Atrial samples from patients undergoing cardiovascular surgery were cut into 1–2-mm² fragments and digested with 2 mg per ml of collagenase type II (Life Technologies, Cat. No. 17101-015, Carlsbad, CA) in PBS (1 h, 37 °C). After filtration (70-μm cell strainer (Greiner Bio-One, Cat. No. 542070, Frickenhausen, Germany) and red cell lysis (Red Blood Cell Lysis Solution, Miltenyi Biotec, Cat. No. 130-094-183, Bergisch Gladbach, Germany), the remaining cells were resuspended in 1 ml of auto-running MACS buffer (Miltenyi Biotec, Cat. No. 130-090-221). After preseparation (30-μm filter, Miltenyi Biotec, Cat. No. 130-041-407), CD31-positive endothelial cells (ECs) were isolated using the CD31 MicroBead Kit (Miltenyi Biotec, Cat. No. 130-091-935) and the human FcR-Blocking Regent (Miltenyi Biotec, Cat.No. 130-059-901) in the MACS system (Miltenyi Biotec) according to the manufacturer's instruction. Isolated ECs were cultured in Endothelial Cell Growth Medium 2 (PromoCell, Cat. No. C-22011, Heidelberg, Germany) until confluence. Adipose fibroblasts (AF) were isolated from subcutaneous fat tissue and cardiac fibroblasts (CF) were isolated from atrial samples. Tissues were cut and digested using collagenase solution type II (2.5 h, 37 °C), and resuspended in DMEM high glucose (Biochrom-Millipore, Cat. No. FG 0435; Berlin, Germany) containing 10% fetal calf serum (FCS, Fisher Scientific, Schwerte, Germany), penicillin (100 U per ml), streptomycin (100 μg per ml), both from PanReacAppliChem (AppliChem, Darmstadt, Germany), and sodium-pyruvate (1 mM, Gibco, Karlsbaden, CA) and grown to confluence. Smooth muscle cells (SMCs) were isolated from the arteria mammaria interna (ITA). Vessels were cut longitudinally. With a scalpel, the tissue was cut into square pieces. The pieces were put on BD Primaria™ 6-well plates (Greiner Bio-One, Frickenhausen, Germany) and dried (2–3 h, 37 °C, 5% CO₂). SMC growth medium 2 (PromoCell GmbH, Heidelberg, Germany) was added. When the first SMCs migrated from the tissue, the pieces were removed and cells were grown to confluence. All biopsies from patients undergoing cardiovascular surgery were transferred within 10 min from the operation room in PBS. To passage the ECs, SMCs, and fibroblasts, cells were incubated with trypsin (Life Technologies) diluted 1 in 4 in PBS (5 min, 37 °C, 5% CO₂). Trypsin solution was neutralized using 2 times of media. ECs and SMCs were frozen in liquid nitrogen in 10% DMSO (Roth, Karlsruhe, Germany), 40% FCS, and 50% cell-type-specific medium. CFs and AFs were frozen in liquid nitrogen using 10% DMSO and 10% FCS in fibroblast-specific medium as described above. The generation of cell cultures from patient biopsies

was approved by the local ethical committee of the Medical Faculty of the Technical University of Munich (project no. 1588/06 (amendment) and 2919/10).

Flow cytometry-based quantification of the cell types. Frozen SMCs isolated from the ITA (passage p2) and cultured as described above were thawed and used for flow cytometry analysis. SMCs were fixed in 1% formaldehyde (20 min) and blocked in PBS containing 10% FCS (v/v) and 0.05% sodium azide (w/v). The first antibody anti-alpha smooth muscle actin (ab5694, Abcam, UK) was added at a ratio of 1:20 in wash/permeabilization buffer (PBS containing 5% FCS (v/v), saponin 0.5% (w/v), and sodium azide (0.05%, w/v)) for 30 min at 4 °C. The secondary antibody goat-anti-rabbit Alexa Fluor 488® (ab150077, Abcam, UK) was added at a ratio of 1:2000 in buffer containing saponin (30 min, ice, dark). ECs from atrial biopsies were isolated and cultured as described above. After preparation of a single-cell suspension, one sample of the total biopsy and the positive and negative fraction obtained after MACS were analyzed. All cell fractions were resuspended in ice-cold PBS/0.5% BSA/2 mM EDTA (FACS buffer) containing 5% anti-human CD31 PE-Cy7 (25-0319, eBioscience, Frankfurt, Germany) and incubated for 30 min in the dark. CFs and AFs were isolated as described before. Cells were cultured on gelatin-coated plates and analyzed at passage 1 or 2. Cells were detached using 0.25% trypsin (Life Technologies) diluted at a ratio of 1:4 in PBS. Monoclonal anti-human CD90 PE-Cy5 (eBioscience, 15-0909), anti-human CD105 APC (eBioscience, 17-1057), and anti-human CD45 FITC (eBioscience, 11-9459) antibodies were resuspended at a ratio of 1:20 in ice-cold FACS buffer and incubated for 30 min on ice in the dark. After staining, all cells undergoing flow cytometry analyses were resuspended in ice-cold PBS/0.5% BSA/4 mM EDTA and kept on ice until flow cytometric analysis was performed with a BD LSRFortessa (BD, San Jose, CA). Negative controls were unstained or stained with the secondary antibody alone. Cytometry data were analyzed with the FlowJo software version 7.6.5 (flowjo@treestar.com).

Immunocytochemistry. SMCs were fixed (4% PFA, 20 min) and permeabilized in PBS-T (0.1% Triton-X-100 in PBS, 10 min). Unspecific binding was blocked with 5% normal goat serum (Abcam, ab7841, Cambridge, UK) in PBST for 30 min. Polyclonal rabbit anti-alpha smooth muscle actin (Abcam, ab5694) was diluted at a ratio of 1:20 in PBS-T and SMCs were incubated with the first antibody overnight at 4 °C. The secondary antibody goat-anti-rabbit IgG (H&L) Alexa Fluor 555® (Abcam, ab150078) was diluted at a ratio of 1:200 in PBS-T. Cells were incubated in the dark for 60 min. ECs isolated from LA and the positive and the negative fraction after MACS sort were stained for CD31. Cells were isolated and processed using the MACS system as described above and plated on cover slides until confluence. Cells were fixed using 4% PFA (10 min) and blocked (5% goat serum in PBS). Polyclonal rabbit anti-CD31 (Abcam, ab28364) was diluted at a ratio of 1:25 and incubated overnight at 4 °C. Secondary antibody goat-anti-rabbit IgG (H&L) Alexa Fluor 555® (Abcam, ab150078) was diluted at a ratio of 1:500 and incubated (60 min, dark). Both antibodies were diluted in PBS. CFs and AFs were fixed using 4% PFA/sucrose in PBS (15 min). For VIM staining, cells were permeabilized (0.25% Triton-X-100 in PBS) and unspecific binding sites were blocked (5% goat serum in PBS-T, 1 h). Polyclonal rabbit anti-VIM (Abcam, ab45939) was used as cytoskeleton marker at 1 μg per ml of final concentration in PBS-T. For DDR2, staining cells were washed with PBS after fixation and blocked with 5% goat-serum in PBS. Polyclonal rabbit anti-DDR2 (LSBio, LS-C99151, Seattle, WA) was diluted at a ratio of 1:20 in PBS. Both antibodies were incubated for 1 h. Secondary antibody goat-anti-rabbit IgG (H&L) Alexa Fluor 555® (Abcam, ab150078) was diluted at a ratio of 1:500 in either PBS-T or PBS and incubated for 1 h in the dark. For immunocytochemistry, SMCs, ECs, CFs, and AFs were grown to approximately 80% confluence on 4-well chamber cover slides (Millizell EZ slides, Millipore, Darmstadt, Germany). All incubations were performed at room temperature, except overnight incubations. After the last wash, slides were air-dried, mounted in Abcam-mounting medium containing DAPI (Abcam, ab104139), sealed with coverslips, and evaluated under a fluorescent microscope (Axiovert 200 M, Zeiss, D-73447 Oberkochen).

RT-qPCR analysis. Expression of CF-specific proteins was confirmed on transcriptional level by RT-qPCR analysis. ECs, SMCs, CFs, and AFs (p0-4) were lysed with RNA lysis buffer (Peqlab, Erlangen, Germany). Total RNA was purified using the peqGOLD total RNA kit (Peqlab) and reverse-transcribed into cDNA with M-MLV reverse transcriptase (Invitrogen, Darmstadt, Germany) according to the manufacturer's recommendation. Expression of ALDH1A2, PDFD, DAPK1, CSRP2, DIAPH3, CES1, and PTPRZ1 was evaluated on a QuantStudio3 (Applied Biosystems, Foster City, CA) using Power SYBR Green Master Mix (Applied Biosystems) and the following conditions: activation of *Taq* DNA polymerase (15 min at 95 °C) followed by 40 cycles with 15 s at 95 °C, 60 s at 60 °C. The sequences of the used primers are noted in Supplementary Table 4. Quantification was performed using the relative expression software tool REST©. Data were normalized to β-actin.

Sample preparation for MS analysis. All 16 heart regions dissected from three trauma victims were washed three times with cold PBS before being crushed in liquid nitrogen using a mortar and pestle. Powdered samples were then resuspended

in 500 µl of SDC reduction and alkylation buffer and boiled for 10 min to denature proteins[19]. Samples were further mixed (six times for 30 s and cooled on ice in-between) using a FastPrep®-24 Instrument (MP Biomedicals). Protein concentration was measured using the Tryptophan assay and 300 µg were further processed for overnight digestion by adding Lys-C and trypsin in a 1:50 ratio (µg of enzyme to µg of protein) at 37 °C and 1700 rpm. On the following day, samples were sonicated using a bioruptor (15 cycles of 30 s) and further digested for 3 h with Lys-C and trypsin (1:100 ratio). Peptides were acidified to a final concentration of 0.1% trifluoroacetic acid (TFA) for SDB-RPS binding and 40 µg of peptides were loaded on four 14-gauge Stage-Tip plugs. Peptides were washed first with isopropanol/1% TFA (200 µl) and then 0.2% TFA (200 µl) using an in-house-made Stage-Tip centrifuge at 2000×g. Peptides were eluted with 60 µl of elution buffer (80% acetonitrile/1% ammonia) into auto sampler vials and dried at 60 °C using a SpeedVac centrifuge (Eppendorf, Concentrator plus). Peptides were resuspended in 2% acetonitrile/0.1% TFA and sonicated (Branson Ultrasonics, Ultrasonics Cleaner Model 2510) before peptide concentration estimation using the Nanodrop. About 30 µg of peptides of each sample were further fractionated into 54 fractions and concatenated into 8 fractions by high-pH reversed-phase fractionation using the recently described[20] "loss-less" nano-fractionator. CFs, AFs, ECs, and SMCs were processed similarly to the heart tissue samples without liquid nitrogen crushing and FastPrep®-24 Instrument.

Liquid chromatography–MS analysis. Nanoflow LC–MS/MS analysis of tryptic peptides was conducted on a quadrupole Orbitrap mass spectrometer[73] (Q Exactive HF, Thermo Fisher Scientific, Rockford, IL, USA) coupled to an EASYnLC 1200 ultra-high-pressure system (Thermo Fisher Scientific) via a nano-electrospray ion source. About 1 µg of peptides were loaded on a 40-cm HPLC-column (75-µm inner diameter; in-house packed using ReproSil-Pur C18-AQ 1.9-µm silica beads; Dr Maisch GmbH, Germany). Peptides were separated using a linear gradient from 2 to 20% B in 55 min and stepped up to 40% in 40 min followed by a 5 min wash at 98% B at 350 nl per min where solvent A was 0.1% formic acid and 5% DMSO in water and solvent B was 80% acetonitrile, 5% DMSO, and 0.1% formic acid in water. The total duration of the run was 100 min. Column temperature was kept at 60 °C by a peltier element-containing, in-house-developed oven. The mass spectrometer was operated in "top-15" data-dependent mode, collecting MS spectra in the Orbitrap mass analyzer (60,000 resolution, 300–1650 m/z range) with an automatic gain control (AGC) target of 3E6 and a maximum ion injection time of 15 ms. The most intense ions from the full scan were isolated with an isolation width of 1.5 m/z. Following higher-energy collisional dissociation (HCD) with a normalized collision energy (NCE) of 27%, MS/MS spectra were collected in the Orbitrap (15,000 resolution) with an AGC target of 5E4 and a maximum ion injection time of 25 ms. Precursor dynamic exclusion was enabled with a duration of 30 s. For clinical AFib samples, a "top-5" data-dependent acquisition method as described above was modified to increase the dynamic range on the MS1 level by including three segmented MS scans (12 segments each; total AGC target 1E6), covering a m/z range of 400–1200. MS1 resolution was set to 120,000 at m/z 200 throughout.

MS data analysis. Tandem mass spectra were searched against the 2015 Uniprot human databases (UP000005640_9606 and UP000005640_9606_additional) using MaxQuant[21] version 1.5.5.6 with a 1% FDR at the peptide and protein level, peptides with a minimum length of seven amino acids with carbamidomethylation as a fixed modification, and N-terminal acetylation and methionine oxidations as variable modifications. Enzyme specificity was set as C-terminal to arginine and lysine using trypsin as protease and a maximum of two missed cleavages were allowed in the database search. The maximum mass tolerance for precursor and fragment ions was 4.5 ppm and 20 ppm, respectively. If applicable, peptide identifications by MS/MS were transferred between runs to minimize missing values for quantification with a 0.7-min window after retention time alignment. Label-free quantification was performed with the MaxLFQ algorithm using a minimum ratio count of 1. For clinical AFib samples, the identification transfer was restricted to the healthy LA library only, and we set a minimum ratio count of 2 for label-free quantification.

Statistical analysis. Statistical and bioinformatics analysis was performed with the Perseus software[32] (version 1.5.5.0), Microsoft Excel, and R statistical software. Proteins that were identified in the decoy reverse database or only by site modification were not considered for data analysis. We also excluded potential contaminants. Data were further filtered to make sure that identified proteins showed expression in all biological triplicates of at least one heart region and missing values were imputed on the basis of normal distribution (down shift = 1.8, width = 0.15). PCA analysis of the heart region and cell types relied on singular value decomposition and the original feature (protein) space was orthogonally transformed into a set of linearly uncorrelated variables (principal components). These account for distinct types of variability in the data. For hierarchical clustering, LFQ intensities were first z-scored and clustered using Euclidean as a distance measure for column and row clustering. Gene set enrichment analysis (GSEA) was performed using gene set collections from the MSigDB[26]. Mean log2 ratios of biological triplicates and the corresponding p values were visualized with volcano plots. We used t-test

for binary comparisons and SAM with $s_0 = 0.1$ and FDR < 0.05 for the assessment of t-test results in volcano plots[32].

Copy number calculation and subcellular heart proteome model. Conversion of LFQ intensities to copy number estimations was achieved using the proteomic ruler[27]. The proteomic ruler plug-in v.0.1.6 was downloaded from the Perseus plug-in store, for use with Perseus version 1.5.5.0. Protein intensities were filtered for three valid values in at least one heart region. Proteins belonging to the GO term "blood microparticle" were removed from the analysis (see Supplementary Data 3 for a full list of removed proteins). Protein groups (proteins that can be distinguished based on the available peptide information) were annotated with amino acid sequence and tryptic peptide information for the leading protein ID, using the .FASTA file used for processing in MaxQuant. Copy numbers per diploid nucleus were estimated using 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," and total cellular protein concentration—"200 g per l." To build a subcellular model of the heart atlas proteome, subcellular localization predictions from spatial proteomics data[28,29] were matched to the protein groups using the leading canonical protein ID. The median copy number of the three replicates was multiplied by the protein molecular weight to calculate protein mass. The mass of each protein was attributed to the nucleus, cytosol, or a specific organelle according to its distribution in HeLa cells. Since many highly abundant heart-specific proteins were not present in this spatial proteomics database, the top 100 proteins in each heart region were completed for subcellular localization using annotation from UniProt. These manually annotated protein masses were assigned entirely to the respective organelle. This led to a median of 94% of total protein mass being assigned to a specific location.

References

1. Xin, M., Olson, E. N. & Bassel-Duby, R. Mending broken hearts: cardiac development as a basis for adult heart regeneration and repair. *Nat. Rev. Mol. Cell Biol.* **14**, 529–541 (2013).
2. Bergmann, O. et al. Dynamics of cell generation and turnover in the human heart. *Cell* **161**, 1566–1575 (2015).
3. Pinto, A. R. et al. Revisiting cardiac cellular composition novelty and significance. *Circ. Res.* **118**, 400–409 (2016).
4. Celermajer, D. S., Chow, C. K., Marijon, E., Anstey, N. M. & Woo, K. S. Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection. *J. Am. Coll. Cardiol.* **60**, 1207–1216 (2012).
5. Aye, T. T. et al. Proteome-wide protein concentrations in the human heart. *Mol. BioSyst.* **6**, 1917–1927 (2010).
6. Lau, E. et al. A large dataset of protein dynamics in the mammalian heart proteome. *Sci. Data* **3**, 160015 (2016).
7. Zhang, J. et al. Systematic Characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria. *Proteomics* **8**, 1564–1575 (2008).
8. Peng, Y. et al. Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction. *Mol. Cell. Proteomics* **13**, 2752–2764 (2014).
9. Lundby, A. & Olsen, J. V. Phosphoproteomics taken to heart. *Cell Cycle* **12**, 2707–2708 (2013).
10. Ounzain, S. et al. Genome-wide profiling of the cardiac transcriptome after myocardial infarction identifies novel heart-specific long non-coding RNAs. *Eur. Heart J.* **36**, 353–368 (2015).
11. Lindskog, C. et al. The human cardiac and skeletal muscle proteomes defined by transcriptomics and antibody-based profiling. *BMC Genomics* **16**, 475 (2015).
12. Kääb, S. et al. Global gene expression in human myocardium—oligonucleotide microarray analysis of regional diversity and transcriptional regulation in heart failure. *J. Mol. Med.* **82**, 308–316 (2004).
13. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
14. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* https://doi.org/10.1016/j.cels.2017.05.009 (2017).

15. Azimifar, S. B., Nagaraj, N., Cox, J. & Mann, M. Cell-type-resolved quantitative proteomics of murine liver. *Cell Metab.* **20**, 1076–1087 (2014).

16. Sharma, K. et al. Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* **18**, 1819–1831 (2015).

17. Deshmukh, A. S. et al. Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. *Mol. Cell. Proteomics* **14**, 841–853 (2015).

18. Murgia, M. et al. Single muscle fiber proteomics reveals fiber-type-specific features of human muscle aging. *Cell Rep.* **19**, 2396–2409 (2017).

19. 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).

20. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 065136 (2017).

21. 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).

22. 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).

23. Kim, M.-S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).

24. Coscia, F. et al. Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. *Nat. Commun.* **7**, 12645 (2016).

25. Schaab, C., Geiger, T., Stoehr, G., Cox, J. & Mann, M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell. Proteomics* **11** (M111), 014068 (2012).

26. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).

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. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, e16950 (2016).

29. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* **44**, D1251–D1257 (2016).

30. Rakus, D., Gizak, A. & Wiśniewski, J. R. Proteomics unveils fibroblast-cardiomyocyte lactate shuttle and hexokinase paradox in mouse muscles. *J. Proteome Res.* **15**, 2479–2490 (2016).

31. Wiśniewski, J. R. et al. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8**, 611 (2012).

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

33. Voelkel, T. et al. Lysine methyltransferase Smyd2 regulates Hsp90-mediated protection of the sarcomeric titin springs and cardiac function. *Biochim. Biophys. Acta* **1833**, 812–822 (2013).

34. Ouafik, L., May, V., Keutmann, H. T. & Eipper, B. A. Developmental regulation of peptidylglycine alpha-amidating monooxygenase (PAM) in rat heart atrium and ventricle. Tissue-specific changes in distribution of PAM activity, mRNA levels, and protein forms. *J. Biol. Chem.* **264**, 5839–5845 (1989).

35. England, J. & Loughna, S. Heavy and light roles: myosin in the morphogenesis of the heart. *Cell. Mol. Life Sci.* **70**, 1221–1239 (2013).

36. Lu, Z. Q., Sinha, A., Sharma, P., Kislinger, T. & Gramolini, A. O. Proteomic analysis of human fetal atria and ventricle. *J. Proteome Res.* **13**, 5869–5878 (2014).

37. Kumar, C., Saidapet, C., Delaney, P., Mendola, C. & Siddiqui, M. A. Expression of ventricular-type myosin light chain messenger RNA in spontaneously hypertensive rat atria. *Circ. Res.* **62**, 1093–1097 (1988).

38. Lesage, F. & Lazdunski, M. Molecular and functional properties of two-pore-domain potassium channels. *Am. J. Physiol.* **279**, F793–F801 (2000).

39. Yin, Z., Ren, J. & Guo, W. Sarcomeric protein isoform transitions in cardiac muscle: a journey to heart failure. *Biochim. Biophys. Acta* **1852**, 47–52 (2015).

40. Hamm, C. W. et al. The prognostic value of serum troponin T in unstable angina. *N. Engl. J. Med.* **327**, 146–150 (1992).

41. Weiss, A., Schiaffino, S. & Leinwand, L. A. Comparative sequence analysis of the complete human sarcomeric myosin heavy chain family: implications for functional diversity11Edited by J. Karn. *J. Mol. Biol.* **290**, (61–75 (1999).

42. Eggers, K. M., Johnston, N., James, S., Lindahl, B. & Venge, P. Cardiac troponin I levels in patients with non–ST-elevation acute coronary syndrome—The importance of gender. *Am. Heart J.* **168**, 317–324.e1 (2014).

43. Shave, R. et al. Exercise-induced cardiac troponin elevation: evidence, mechanisms, and implications. *J. Am. Coll. Cardiol.* **56**, 169–176 (2010).

44. Keshishian, H. et al. Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. *Mol. Cell. Proteomics* **14**, 2375–2393 (2015).

45. Govindan, S. et al. Increase in cardiac myosin binding protein-C plasma levels is a sensitive and cardiac-specific biomarker of myocardial infarction. *Am. J. Cardiovasc. Dis.* **3**, 60–70 (2013).

46. Members, W. G. et al. Heart disease and stroke statistics—2009 update. *Circulation* **119**, 480–486 (2009).

47. Ozcelik, F. et al. Diagnostic value of plasma fibronectin level in predicting the presence and severity of coronary artery disease. *Ann. Hematol.* **88**, 249–253 (2009).

48. Orriols, M. et al. Down-regulation of Fibulin-5 is associated with aortic dilation: role of inflammation and epigenetics. *Cardiovasc. Res.* **110**, 431–442 (2016).

49. Yan, J., Li, F., Ingram, D. A. & Quilliam, L. A. Rap1a is a key regulator of fibroblast growth factor 2-induced angiogenesis and together with Rap1b controls human endothelial cell functions. *Mol. Cell. Biol.* **28**, 5803–5810 (2008).

50. Geyer, P. E. et al. Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12**, 901 (2016).

51. Schiller, H. B. et al. Time- and compartment-resolved proteome profiling of the extracellular niche in lung injury and repair. *Mol. Syst. Biol.* **11**, 819 (2015).

52. Levine, R. A. et al. Mitral valve disease—morphology and mechanisms. *Nat. Rev. Cardiol.* **12**, 689–710 (2015).

53. Martin-Rojas, T. et al. iTRAQ proteomic analysis of extracellular matrix remodeling in aortic valve disease. *Sci. Rep.* **5**, 17290 (2015).

54. Tan, H. T. et al. Unravelling the proteome of degenerative human mitral valves. *Proteomics* **15**, 2934–2944 (2015).

55. Challa, A. A. & Stefanovic, B. A novel role of vimentin filaments: binding and stabilization of collagen mRNAs ▽. *Mol. Cell. Biol.* **31**, 3773–3789 (2011).

56. Nam, Y.-J. et al. Reprogramming of human fibroblasts toward a cardiac fate. *Proc. Natl Acad. Sci. USA* **110**, 5588–5593 (2013).

57. Ieda, M. et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**, 375–386 (2010).

58. Souders, C. A., Bowers, S. L. K. & Baudino, T. A. Cardiac fibroblast. *Circ. Res.* **105**, 1164–1176 (2009).

59. Furtado, M. B. et al. Cardiogenic genes expressed in cardiac fibroblasts contribute to heart development and repairnovelty and significance. *Circ. Res.* **114**, 1422–1434 (2014).

60. Thomas, P. S., Rajderkar, S., Lane, J., Mishina, Y. & Kaartinen, V. AcvR1-mediated BMP signaling in second heart field is required for arterial pole development: Implications for myocardial differentiation and regional identity. *Dev. Biol.* **390**, 191–207 (2014).

61. Rose, R. A. & Giles, W. R. Natriuretic peptide C receptor signalling in the heart and vasculature. *J. Physiol.* **586**, 353–366 (2008).

62. DeLaughter, D. M. et al. Single-cell resolution of temporal gene expression during heart development. *Dev. Cell* **39**, 480–490 (2016).

63. Hemric, M. E. & Chalovich, J. M. Effect of caldesmon on the ATPase activity and the binding of smooth and skeletal myosin subfragments to actin. *J. Biol. Chem.* **263**, 1878–1885 (1988).

64. Kirchhof, P. et al. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur. Heart J.* **37**, 2893–2962 (2016).

65. Nagaraj, N. et al. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. *Mol. Cell. Proteomics* **11**, M111.013722 (2012).

66. Burstein, B. & Nattel, S. Atrial fibrosis: mechanisms and clinical relevance in atrial fibrillation. *J. Am. Coll. Cardiol.* **51**, 802–809 (2008).

67. Surks, H. K., Riddick, N. & Ohtani, K.-I. M-RIP targets myosin phosphatase to stress fibers to regulate myosin light chain phosphorylation in vascular smooth muscle cells. *J. Biol. Chem.* **280**, 42543–42551 (2005).

68. Ad, N., Schneider, A., Khaliulin, I., Borman, J. B. & Schwalb, H. Impaired mitochondrial response to simulated ischemic injury as a predictor of the development of atrial fibrillation after cardiac surgery: in vitro study in human myocardium. *J. Thorac. Cardiovasc. Surg.* **129**, 41–45 (2005).

69. Xie, W. et al. Mitochondrial oxidative stress promotes atrial fibrillation. *Sci. Rep.* **5**, 11427 (2015).

70. Chugh, S. S., Blackshear, J. L., Shen, W.-K., Hammill, S. C. & Gersh, B. J. Epidemiology and natural history of atrial fibrillation: clinical implications. *J. Am. Coll. Cardiol.* **37**, 371–378 (2001).

71. Goudarzi, M. et al. Development of a novel proteomic approach for mitochondrial proteomics from cardiac tissue from patients with atrial fibrillation. *J. Proteome Res.* **10**, 3484–3492 (2011).

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

73. Kelstrup, C. D. et al. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field orbitrap mass spectrometer. *J. Proteome Res.* **13**, 6187–6195 (2014).

Author contributions

S.D. acquired and interpreted the proteomics data, developed the concept, and wrote the manuscript. M.D. performed sample collection, FACS sorting, cell culture experiments, gene expression analysis, immunocytochemistry, and edited the paper. P.G. helped with the sample preparation. D.I. analyzed the subcellular proteome. C.B. collected samples from healthy probands. S.D. performed sample collection, immunocytochemistry, and FACS analysis. F.M. helped with the MS measurements. M.A.D. helped with the study design and provided clinical data of included patients. H.L. did the biobanking of selected patient materials, cell culture experiments, and edited the paper. R.L. designed and supervised the study and edited the paper. M.K. developed the concept, supervised the study, and edited the paper. M.M. designed and supervised the study and edited the paper. All the authors have approved the final version.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-01747-2.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

154

Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery

Authors: Nicolai J. Wewer Albrechtsen[1,2,3,4,5,*], Philipp E. Geyer[1,5,*], Sophia Doll[1,5], Peter V. Treit[5], Kirstine N. Bojsen-Møller[3,6], Christoffer Martinussen[3,6], Nils B. Jørgensen[6], Signe S. Torekov[2,3], Florian Meier[1,5], Lili Niu[1], Alberto Santos[1], Eva C. Keilhauer[5], Jens J. Holst[2,3], Sten Madsbad[3,6], Matthias Mann[1,5,†]

Affiliations:

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

[2]Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[3]NNF Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[4]Department of Clinical Biochemistry, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark.

[5]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[6]Department of Endocrinology, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark

*These authors contributed equally

†Lead contact and to whom correspondence should be addressed: Matthias Mann: mmann@biochem.mpg.de

HIGLIGHTs:
- **Plasma proteome profiling of independent Roux-en-Y gastric bypass cohorts**
- **Global correlation maps of the plasma proteome reveal functional networks**
- **Systemic inflammation and lipid transport are the major remodeled processes**
- **Gastric bypass has specific and common effects to other weight loss interventions**

1

SUMMARY

Obesity-related diseases affect half of the global population, and bariatric surgery is one of the few interventions with long-lasting weight loss and cardio-metabolic effects. Here, we investigated the effect of Roux-en-Y gastric bypass surgery on the plasma proteome, hypothesizing that specific proteins or protein patterns may serve as key mediators and markers of the metabolic response. We performed mass spectrometry (MS)-based proteomics on two longitudinal studies encompassing 47 morbidly obese patients, generating quantitative information on more than 1700 proteins. A global correlation matrix incorporating about 200,000 relationships revealed functional connections between proteins and assigned them to physiological processes. The main classes of significantly altered proteins were markers of systemic inflammation and those involved in lipid metabolism. Our data highlight robust correlative and anti-correlative behaviors of circulating proteins to each other and to clinical parameters. A group of inflammation-related proteins showed distinct inverse relationships to proteins consistently associated with insulin sensitivity.

Keywords: Biomarker, Diabetes, Gastric bypass surgery, Global correlation profiles, Insulin resistance, Mass spectrometry, Obesity, Plasma Proteome Profiling, Serum, Shot gun proteomics

156

INTRODUCTION

Over 600 million people are obese (BMI > 30) and a further 1.3 billion overweight ((WHO), 2016). The proportion of the latter subgroup has doubled since 1980, representing an unprecedented number of people predisposed to or already affected by a broad spectrum of co-morbidities (Afshin et al., 2017). These include type 2 diabetes (T2D), cardiovascular disease, endocrine disturbances, neurodegenerative diseases and several types of cancers.

In response to this global and dramatically increasing burden, the scientific community has devoted substantial efforts to obesity and its clinical consequences, aiming to gain a deeper understanding of the underlying disease mechanisms while developing new treatment possibilities. From a vast array of dietary, behavioral and pharmacological alternatives, bariatric surgery has emerged as the most effective intervention for treating obesity and type 2 diabetes. There are different bariatric procedures, with Roux-en-Y gastric bypass (RYGB) among the most established ones, where the creation of a small gastric pouch and bypass of the stomach and upper small intestine, causes accelerated entry of nutrients into the small intestine (Griffen et al., 1977; Mason and Ito, 1967; Mason et al., 1975; Wittgrove et al., 1994). Besides major weight loss (on average 40 kg) the physiological responses to RYGB include broad improvements in cardiovascular status and other measures of metabolic health (Aab et al., 2016; Adams et al., 2017; Cardoso et al., 2017; Madsbad and Holst, 2014; Sjostrom et al., 2007). Interestingly, RYGB is curative for many individuals suffering from T2D. Hepatic insulin sensitivity tends to improve markedly within a week after surgery, even before substantial weight loss, followed by increased sensitivity in peripheral tissues in weeks (Bojsen-Moller et al., 2014). While gastrointestinal peptides, bile acids and even the microbiome have been implicated in the rapid and drastic physiological changes after RYGB, the molecular mechanisms are poorly understood (Cummings and Rubino, 2018; Miras and le Roux, 2013).

Metabolic processes are regulated by proteins and the technology of choice to study changes in the proteome in an unbiased manner is mass spectrometry (MS)-based proteomics (Aebersold and Mann, 2016). Proteins circulating in the blood can be both mediators of organ cross talk and markers of whole body states. However, a system-wide approach has only recently become possible due to technological improvements of the proteomic pipeline (Geyer et al., 2017). We recently developed 'Plasma Proteome Profiling' as a robust and automated platform for the reproducible analysis of hundreds of clinical samples. The quantified proteome included the main functional plasma proteins such as the complement system, apolipoproteins and other transporters and allowed assessment of sample quality (Geyer et al., 2016a). In a subsequent study, we investigated the longitudinal effects of caloric restriction-induced weight loss on the plasma proteome of 52 individuals over more than a year. Analysis of more than 1000

individual plasma proteomes revealed that protein levels in general were much more stable within individuals over time than between them. We extracted panels indicative of systemic inflammatory status, allowing stratification of the cohorts. The data also highlighted specific proteins correlated to relevant clinical parameters such as BMI, HDL, LDL and insulin sensitivity (Geyer et al., 2016b).

Here we wished to gain further insights into the global metabolic changes caused by RYGB surgery as reflected in the plasma proteome. To obtain robust and generalizable results, we selected two studies with similar design to investigate improvements in hepatic insulin sensitivity and beta-cell function (Bojsen-Moller et al., 2014; Martinussen et al., 2015). We applied an improved version of our plasma proteome profiling pipeline, which doubled our coverage of the plasma proteome. Furthermore, for a more comprehensive understanding of the dynamics of the plasma proteome, we implemented global correlation maps. These consist of hundreds of thousands of protein-protein connections and protein – clinical parameter associations, highlighting co-regulated proteins and revealing underlying physiological mechanisms.

RESULTS

In-depth dataset of the human plasma proteome in RYGB

Study 1 consisted of 20 and study 2 of 27 morbidly obese individuals, BMI 40±4 kg/m^2 (Bojsen-Moller et al., 2014) and 42±4 kg/m^2 (Martinussen et al., 2015), respectively (Figure 1A). Blood was drawn at four time points, once directly before and three times after surgery (0, 1, 12 and 52 weeks). From 19 individuals, we also obtained blood samples at two or four years post-operation (>104 weeks).

Weight declined steadily in the cohorts, stabilizing at an average loss of 28% of total body mass one year after surgery (Figure 1A). Insulin resistance of the liver as estimated by HOMA-IR, decreased from 3.3±0.7 to a near normal value of 2.0±0.9 after 12 weeks.

In this study, we further improved the Plasma Proteome Profiling pipeline by first generating deep plasma libraries by consecutive depletion of pooled plasma from the top6 and the top14 highest abundant proteins. Following depletion, we used the recently developed high-pH reversed-phase 'Spider Fractionator' (Kulak et al., 2017) to generate a deep library of 1928 plasma proteins and 14,588 peptides (excluding contaminants) (Figure 1B). Cohort samples were automatically prepared without any depletion as described previously (Geyer et al., 2016a) (Figure 1C). The second improvement of our workflow was a new acquisition method termed BoxCar, which results in about a ten-fold increase in dynamic range of peptide signals via an equalized filling pattern of the ion trap. This results in substantial coverage of undepleted plasma samples even without fractionation (Meier et al., 2018).

4

158

We generated a large plasma proteome dataset, comprising a total of 175 human plasma proteomes. In triplicate analysis of the samples, an average of 1025 proteins were identified per study participant, more than twice the number in our previous plasma proteome study (Geyer et al., 2016b) (Figure 1D; Table S1). In total, we quantified 1700 plasma proteins with at least two peptides in the study samples (Figure 1D and E). We also evaluated the cohorts based on our previously defined sets of quality markers. None of the samples suffered from extensive erythrocyte lysis or partial coagulation, but we observed a slight trend toward higher erythrocyte lysis in study 2 (Figure S1).

As this is the largest proteome dataset generated by unbiased MS-based proteomics on non-depleted plasma, we provide the data in an easily accessible and minable format for use by the community (Table S1). Ranking the proteins according to their abundance and annotating them according to their function using the Gene Ontology Biological Process terms (GOBP) and performing a one-dimensional annotation enrichment analysis resulted in 117 statistically significant terms (Figure 1F). At this depth of coverage, processes connected to lipid transport, coagulation and inflammation are very well reflected in the higher abundance ranges, whereas intracellular processes - presumably as a result of tissue leakage - are enriched in the low abundance ranges (Table S2).

Plasma proteome rearranges after RYGB

Our primary aim was to investigate plasma proteome dynamics after RYGB and to relate these to the pathophysiological changes stemming from the procedure. Over the entire time course the levels of 114 proteins changed significantly (Table S3). We expected the surgery intervention itself to have a strong effect on the plasma proteome. Indeed, most of the significantly altered proteins – 88 in total – changed one week after surgery, 39 after 12 and 52 weeks and 33 after more than two years. The two studies differed only minimally in their design, allowing us to initially use study 1 as the 'discovery cohort' and study 2 as the 'validation cohort' for plasma proteome profiling. For the first two time points, where the number of samples was most comparable, 90% of the significantly changed proteins in study 1 were also significant in study 2 (Table S3). Due to the high overlap of proteins regulated in the two independent RYGB studies and the fact that we did not detect any significant differences between them, we henceforth integrated the data from both of them.

To investigate plasma proteome dynamics after RYGB, we calculated the median, Z-scored intensities of all significantly altered proteins at each time point. A hierarchical clustering analysis stratified proteins into four main groups based on their response to RYGB (Figure 2A and B). Nearly half of these proteins showed a long-term decrease (group 1, 10 proteins; group 2, 40 proteins). The main difference between the two patterns was that group 2 increased at the first post-operative time point before dropping below baseline levels for the entire period of up to four years. The 39 proteins in group 3 showed a long-term increase, with many of them initially exhibiting a decrease after surgery. This was also the pattern of group 4, although here levels only recovered to baseline. The remaining proteins had a more heterogeneous behavior with increased or decreased levels between the different time points.

5

We annotated each of the 114 significantly changed proteins with Gene Ontology Biological Process (GOBP), Molecular Function (GOMF), and Cellular Compartment (GOCC) as well as UniProt-KB Keyword terms, to reveal the biological processes that were influenced by RYGB. Fisher exact test on these keyword annotations yielded 322 connections based on 20 central keywords, drawing a broad picture of biological functions affected by RYGB (STAR Methods; Table S4). These range from inflammation, over lipid homeostasis and protease activity to changes in metal homeostasis.

Next, we performed hierarchical clustering on a Boolean table of the proteins annotated by the 20 keywords, to investigate if proteins belonging to the same cluster also had a functional association (Figure 2C). Terms related to inflammation were most strongly enriched (cluster 2). Manual inspection revealed even more inflammatory proteins, including frequently clinically used marker of inflammation, CRP (C-reactive protein), the acute phase proteins SAA1, SAA4, LBP and the complement factors, C1S, CFHR3 and CFHR5. The second most keyword-enriched group belonged to the lipid homeostasis system (cluster 3), as expected upon weight loss, and their individual dynamics will be described below.

`Metal-binding` was also significantly enriched, but spread over clusters 1, 2 and 3, indicating that RYGB differently affected their functional roles (Table S4). The absorption of iron is compromised after gastric bypass surgery since the antrum, duodenum and the proximal part of the jejunum, where absorption takes place is separated from the contact with the ingested nutrients and results in anemia, due to malabsorption of micronutrients (Worm et al., 2015). Interestingly, levels of transferrin (TF), the main iron transporter in plasma increased by 20% (cluster 3), indicating a sensing mechanism for low iron levels. In contrast, ceruloplasmin (CP), which oxidizes Fe^{2+} to Fe^{3+} for effective transport by transferrin, was down-regulated and present in cluster 2 that anti-correlates with cluster 3, perhaps because less iron needs to be oxidized.

We traced the significantly regulated proteins connected to protease keyword annotations to members of the complement cascade system and different serine protease inhibitors (SERPINs). SERPINs are a functionally heterogeneous group of proteins, including the inflammation associated SERPINA1 and SERPINA3 (cluster 2) and also the adipocyte secreted SERPINF1, whose decrease after operation indicates the loss of body fat (cluster 1).

Encouraged by the fact that plasma proteome profiling revealed specific physiologic changes upon RYGB, we set out to systematically exploit the observed regulation to construct a global network of co-regulated plasma proteins and functional modules.

Global correlation maps reveal co-regulation of hundreds of plasma proteins

The levels of plasma proteins have traditionally been measured and considered independently, however proteins often work together as complexes or groups. Treating RYGB as a generic perturbation, we next investigated the co-regulation of the plasma proteome at a systems level to identify protein networks and search for new biomarker candidates. We constructed a global correlation map containing pairwise relationships between all proteins and further included clinical parameters. In our case, there are up to

175 abundance values for each plasma protein (47 individuals, 5 time points). For instance, the Pearson correlation coefficient between the lipoprotein proteins APOA1 and APOC1 was 0.57 over all data points (Figure 3A).

We considered only the 580 proteins that were quantified in at least 25% of the samples and added 26 clinical parameters. As a third parameter group, we included the levels of glycated peptides, which are related to blood sugar control and can be determined by proteomics as previously described (Keilhauer et al., 2016). Here, we quantified 203 glycated peptides and combined them to a single 'peptide glycation' factor (STAR Methods). The protein abundance levels, clinical parameters and the peptide glycation factor constituted 607 items in total, which we cross-correlated to generate a matrix of 183,921 correlation coefficients. Unsupervised hierarchical clustering and color-coding the correlation coefficients then generated the global correlation map. We hypothesized that such a map would group the plasma proteome in such a way that proteins and clinical parameters associated with the same underlying regulatory mechanism would cluster in the same area. Thus, the global correlation map would both capture large-scale organizational features of the plasma proteome as well as individual associations between proteins and connect them to measures of disease state represented by clinical parameters.

A wealth of information can be found in the correlation map (Figure 3A), with co-regulated areas as large as 76x76 items and as small as binary associations. For functional interpretation, we employed bioinformatic keyword annotation and subsequent Fisher exact tests. The largest structure belonged to the innate immune system and was further dominated by a sub-cluster with very strong correlations. A cluster consisting of immunoglobulins was nearly as large (68x68). The global correlation analysis also identified a cluster dominated by HDL, but also containing adiponectin (the protein and the clinical parameter), which is positively associated with peripheral insulin sensitivity (Lihn et al., 2005). Out of more than 20 clusters of similar size, one is associated with low density lipoprotein particles (LDL), another with proteins participating in the coagulation cascade such as the von Willebrand factor (vWF) and fibrinogens, and one contains three steroid binding proteins (PZP, SHBG, SERPINA6).

The strongly correlated sub-cluster of the inflammation system includes proteins like CRP, SAA1 and members of the complement cascade. Furthermore, it contains diabetes-associated clinical parameters like insulin levels, C-peptide levels and HOMA-IR, which reflects insulin resistance (Figure 3A). The fact that these disease markers correlate directly within the group of inflammation proteins, reflects a link between insulin resistance and systemic inflammation, which is the subject of extensive metabolic research (Moran et al., 2005; Ndumele et al., 2006). Notably, there are also anti-correlating clusters, characterized by, off-diagonal areas in the global correlation map (Figure 3A, blue patterns). The main clusters contain inflammation-associated proteins, apolipoproteins and insulin-sensitivity parameters. For instance, there is a negative correlation between CRP and adiponectin of -0.35, highlighting direct protein-based connections between decreased inflammatory status and increased insulin sensitivity.

Clearly, the global map contains a treasure trove of information about the plasma proteome, encoded in its almost 200,000 correlations. Co-regulated proteins identify proteins associated with diverse physiological processes. Apart from the most apparent

7

functional associations, there are many clusters with no significant keywords as well as many novel connections within the clusters.

Interestingly, some clusters contained information about the sample analysis procedure. These included a tightly co-regulated group of platelet proteins. Although they accurately report on sample quality, they do not reflect intrinsic plasma proteome properties and were therefore removed from downstream analysis (STAR methods).

To explore the connections between plasma proteins in a complementary way, we generated a network in which the connections are given by their correlation coefficients. For all proteins and clinical parameters in the 20 clusters with the highest number of nodes, we only selected proteins with correlation coefficients of at least 0.4 to display their relationship within and between these clusters (Figure 3B). This highlights the connections within main physiological processes. In the inflammation cluster, for instance, there are strong relationships of the complement system to other inflammatory proteins like lipoprotein binding protein (LBP). The network clearly captures the relationship of inflammation, insulin resistance and lipid metabolism.

RYGB and the response to low-grade inflammation

Inflammation associated proteins were found in the largest co-regulated cluster and they also contained the most highly regulated ones after RYGB (Figures 2, 4). To define an inflammatory panel in an unbiased manner, we filtered for relevant keywords (STAR Methods). This resulted in 51 proteins that covered the major pathways, such as complement activation and acute phase reaction. This inflammatory panel was quantified with high consistency in the dataset (90% valid values). We combined the changes of these proteins, normalized to their starting level, into a single value by taking the median, which represents a robust and accurate reflection of the global inflammation status of each individual at each time point. Then, we aggregated the inflammatory status of all participants over time. The longitudinal trajectory of this inflammation status had a maximum at the first post-operative time point with an average increase of 10% (Figure 4A), presumably as a direct consequence of the surgery (Finnerty et al., 2013). Following this brief increase, the global inflammation panel decreased gradually, reaching baseline level 12 weeks after RYGB surgery and the minimum of -11% at the final time point (>104 weeks). This dynamic regulation was quite consistent across almost all participants in the study cohort (Figure S2A).

Proteins included in our inflammation panel were selected by keyword annotation, which could be spurious, including proteins with negative correlation to inflammation. To assess their suitability as markers, we inspected their dynamics and found that the large majority went up at the immediate post-operative time point and stabilized at lower levels at 52 and greater than 104 weeks (Figure 4A). Next, we compared our data to a plasma proteomic profiling study where weight loss was induced by caloric restriction (Cambridge diet), in which 43 participants lost on average 12% of their body mass during an eight-week period and maintained their weight for one year (Iepsen et al., 2015). From that study, we had already defined a ten protein inflammation panel (Geyer et al., 2016b), which was completely contained in the current one. After exclusion of the first post-operative time point in the RYGB study, we calculated the slope of the inflammation

8

162

proteins in both studies across all time points (Table S5). Between caloric restriction achieved through the Cambridge diet and by RYGB surgery, there was an overlap in 31 increased inflammation proteins. Thus the direction of changes in the full 51 panel in the previous study generally agreed with the current one (Figure 4B). C-reactive protein (CRP), a prominent risk factor for cardiovascular disease, showed the strongest response and its levels increased 5-fold at the one-week time point, and decreased 14-fold afterwards.

For further insights into the connection of the plasma proteome to weight loss, we correlated all 1700 quantified plasma proteins directly to BMI. In total, 51 proteins had a significant correlation of which 18 were part of our inflammation panel (Figure S2B, Table S6). CRP, complement factor C3 and S100A8 were the proteins with the strongest correlation (Pearson correlation of 0.63, 0.61 and 0.45, respectively). Together with the fact that the BMI clustered with inflammation annotated proteins in the global correlation map, this underlines the strong connection between inflammation and obesity.

Core proteins in lipid metabolism are significantly altered by RYGB (407 words)

Lipid homeostasis was the second most strongly affected physiological system by RYGB and it clustered into two distinct regions of the global correlation map (Figure 2). Many of the proteins found in this cluster are well-known constituents of different lipoprotein particles like APOB or receptors of these particles such as the LDL receptor (LDLR). To elucidate their behavior, we filtered the identified plasma proteome for the gene ontology cellular compartment (GOCC) terms 'lipid transport', 'chylomicron', 'high density lipoprotein (HDL)', 'low density lipoprotein (LDL)', 'intermediate density lipoprotein (IDL)' and 'very low density lipoprotein (vLDL)' particle. To this set, we added non-annotated proteins well known to be associated with lipid metabolism like apolipoprotein D (APOD), apolipoprotein(a) (LPA), Prolow-density lipoprotein receptor-related protein 1 (LRP1) and phospholipid transfer protein (PLTP). Excluding the keyword annotated inflammation proteins SAA1, SAA2 and SAA4, resulted in 25 proteins that reflect the lipid homeostasis system in blood. They were consistently quantified across all patients and time points. Furthermore, their correlation map reveals a large degree of similarity between the caloric restriction and RYGB studies (Figure 5A, B). For instance, the highly correlated clusters for apolipoproteins APOC1, APOC2, APOC3, APOC4 and APOE, which are LDL constituents, are almost superimposable. Likewise, the HDL related proteins – APOA1, APOA2, APOA4 and PON1 – correlated highly.

Inspecting the direction of changes in the expression levels of proteins, we found that the HDL constituents APOA1, APOA2, APOA4 and PON1 increased in RYGB, but decreased in caloric restriction, reflecting the HDL laboratory values that were measured independently in these studies (Figure S3). Furthermore, HDL-associated proteins like APOC1, APOC2, APOC4, APOM, PLTP, LPA and several others that may be associated, were regulated in opposite directions. One of those - APOC1 - has broad physiological functions including inhibition of fatty acid uptake from adipocytes and inhibition of lipoprotein binding to LDL. Interestingly, APOC1 was a member of both the LDL and HDL correlation clusters. In caloric restriction, its levels decreased by 12% whereas in RYGB they increased by 7%. This could be due to a lower need for transport of fatty acids after caloric restriction but not after RYGB and may reflect the mechanistic differences of these

9

two weight loss interventions. Interestingly, mice overexpressing APOC1 seem protected for obesity and insulin resistance due to reduced fatty acid uptake in adipocytes (Jong et al., 2001).

After RYGB, insulin sensitivity of the liver rapidly improves (Bojsen-Moller et al., 2014). In line with this, we observed increases in APOA4, another surrogate marker of hepatic insulin sensitivity (VerHague et al., 2013).

Improvement in insulin sensitivity after RYGB is reflected in the plasma proteome

The immediate positive effect of RYGB on insulin sensitivity has been widely discussed, but any reflections of this in the plasma proteome have not been described. In both RYGB cohorts we determined hepatic insulin resistance by the `homeostatic model assessment of insulin resistance` (HOMA-IR based on C-peptide) and in study 1, we further measured `hepatic insulin sensitivity` by tracer estimation of hepatic glucose production, `peripheral insulin sensitivity` by the hyperinsulinemic euglycemic clamp and ´oral glucose insulin sensitivity` from oral glucose tolerance test, resulting in a system-wide assessment of insulin sensitivity.

Connecting the insulin sensitivity measures to the plasma proteome profile retrieved 75 significantly correlating proteins (Figure 6A, B; Table S7). Adiponectin, sex hormone-binding globulin, vitamin K-dependent protein S, cholinesterase, the complement factors C3, CFB and C4-binding protein alpha chain correlated with all four assessments. We calculated a summed insulin sensitivity correlation, taking the strength and number of correlations into account and followed this measure over time since surgery (Figure 6C, Table S8). This factor reached its maximum one year after the gastric bypass operation. Interestingly, the summed insulin sensitivity correlation was almost superimposed in both studies (Figure 6B), validating the individual correlating proteins.

Reassuringly, adiponectin, a classical marker for insulin sensitivity was the protein with the highest score (ADIPOQ; S:16) and sex hormone-binding globulin (SHBG; S:8) and complement factor C3 (C3; S:8) were likewise in the top group. Strikingly, another 25 proteins and the 'peptide glycation' factor correlated or anti-correlated highly with insulin sensitivity (Figure 6A). Other proteins in this heterogeneous group included pigment epithelium-derived factor (SERPINF1; S:11), afamin (AFM; S:10), antithrombin-III (SERPINC1; S:9), apolipoprotein A1 (APOA1; S:8). Note that some of these correlations may simply reflect the weight loss, rather than a more specific effect on insulin sensitivity. For instance, SERPINF1, a protein secreted by adipocytes, is clearly and highly significantly anti-correlated to insulin sensitivity – especially to hepatic insulin sensitivity (Pearson's R -0.66; p < 10^{-9}). Of note, we had previously found a strong association of SERPINF1 and weight loss in the caloric restriction study as well as a correlation to insulin resistance based on HOMA-IR (Geyer et al., 2016b). It would be interesting to determine if SERPINF1 has a functional involvement in insulin sensitivity or if it mainly reflects total fat mass.

Conversely, a group of high-scoring proteins reflecting the inflammatory state and lipid transport likely represent functional connections: both systemic inflammation and dysregulated lipid homeostasis. These included several components of the complement

10

164

system and a number of apolipoproteins, which also correlated or anti-correlated with insulin sensitivity measures in our global correlation map (Figure 3, Table S8). Clearly, our protein correlation profiles reflect the complex relationships between insulin sensitivity, lipid homeostasis and inflammation. However, the majority of proteins correlated with insulin sensitivity had no reported connection to these two physiological processes. As insulin has a broad effect on different organs, it is not surprising that proteins of different physiological processes correlated to insulin sensitivity, which makes them interesting candidates to study as cause or consequence of insulin resistance.

DISCUSSION

Bariatric surgery fundamentally alters human metabolism, and much remains to be learned about the molecular mechanisms underlying surgical therapies such as RYGB. The majority of research has focused on anthropometric as well as clinical and physiological data (Puzziferri et al., 2014), whereas there are few hypothesis-free approaches (Arora et al., 2015; Luo et al., 2016). In this work, we aimed to describe global rearrangements of the plasma proteome in a systems-wide view and discover protein markers of gastric bypass surgery with possible functional relevance. This was enabled by a recently developed robust and highly reproducible pipeline for mass spectrometry-based investigation of plasma, which we termed Plasma Proteome Profiling (Geyer et al., 2016a). In the current study, we integrated several technological advances into the workflow including the ´Spider fractionator` to generate deep peptide libraries and combined it for the first time with the `BoxCar` scan method to increase the dynamic range of our measurements (Meier et al., 2018). Together, this achieved the most extensive quantitative plasma proteome so far, providing the community with concentration estimates of over 1700 proteins and their bioinformatic keyword annotation.

We described longitudinal trajectories of 114 significantly altered proteins in response to RYGB and applied functional annotation analysis to reveal the underlying physiological mechanisms. This identified two main groups, belonging to the inflammation and the lipid homeostasis systems. We found a strong overlap of inflammation proteins in RYGB with our previous study on caloric restriction induced weight loss (Geyer et al., 2016b) with the same regulation and tight correlation to BMI. We conclude that body mass is a major determinant of systemic low-grade inflammation levels and that weight loss – regardless of the mode by which it is achieved – is its most important driver. These dramatic changes are illustrated by regulation of 51 inflammation-associated proteins, the most prominent of which is CRP that decreased 14-fold during the average weight reduction of 28%. Note that a few of these proteins were up-regulated, for instance, the complement factor C7. Interestingly, our unbiased analysis revealed that the weight loss driven decrease over the subsequent year was as strong as the immediate surgery-induced increase. Such dramatic changes in systematic inflammation could affect the risk for cardiovascular events (Danesh et al., 2000).

In contrast to the inflammation system, the lipid homeostasis system was differently regulated between caloric restriction and RYGB. While LDL associated proteins decreased in both interventions, the HDL associated cluster was only increased after RYGB. This is interesting given the link between high HDL levels and lower risk of

coronary heart disease in associative studies, but not when increased by pharmacological intervention (Rader and Hovingh, 2014). At this point, the majority of differently regulated proteins between both types of weight loss cannot be directly linked to clinically measured HDL levels alone.

Lipoprotein particles can be sub-classified into several types that interact or are converted into each other (Kontush et al., 2013). Proteins are the defining elements of all of them and could be the basis for more rational classification of the lipid homeostasis system. In our study, several cardiovascular risk markers from the lipid homeostasis system like the LDL-associated APOB and LPA dropped in both caloric restriction and RYGB, whereas APOA1, whose levels positively correlate with decreased cardiovascular risk (Emerging Risk Factors et al., 2012), significantly increased only in RYGB. A number of such factors emerged from our analyses (Table S9), and it would be interesting to study them in-depth for a possible differential involvement in cardiovascular diseases. More generally, it has been widely discussed which types of cholesterol particles are the most relevant to disease risk (Krauss, 2010) and our data now suggest that it may be promising to focus on the different lipoprotein-associated proteins themselves (Hegele, 2009). For the future, we imagine that plasma proteome profiling of cardiovascular risk studies may help to unravel the role and biomarker potential of individual proteins.

The plasma proteome profiles of a well-characterized longitudinal cohort with comprehensive clinical characterization provided the first opportunity to systematically study the dynamics of the human plasma proteome. For this purpose, we here employed global correlation maps, in which we correlated the quantitative levels of all proteins and clinical parameters to each other. We sorted the resulting matrix by hierarchical clustering and applied bioinformatic enrichment analysis. This disentangled complex processes in the plasma proteome and associated proteins to common physiological functions and clinical risk measures.

Standard clinical parameters for insulin resistance include insulin levels themselves, C-peptide levels and HOMA-IR, clustered with proteins of the inflammation system. Hepatic insulin sensitivity, peripheral insulin sensitivity and oral glucose insulin sensitivity all tended to associate with proteins of the lipid homeostasis system. Network analysis highlighted an anti-correlation between those two systems (Figure 3B). This is consistent with the detrimental effect of systemic inflammation on the risk of metabolic disease and the positive and negative effects of the levels of different cholesterol lipoprotein particles (Dali-Youcef et al., 2013; Glass and Olefsky, 2012). Thus, we here recapitulated the relationships between insulin sensitivity, the lipid homeostasis system and systemic inflammation purely from an unbiased assessment of the plasma proteome in a human intervention study for diabetes. Interestingly, our analysis revealed that known single protein markers of insulin sensitivity, such as adiponectin (Li et al., 2009), were indeed the highest scoring ones, and that there are many others that have not yet been connected to insulin sensitivity (Figure 6, Table S8).

In conclusion, plasma proteome profiling is a powerful technology for studying the effects of physiological interventions such as RYGB in humans, which may provide new molecular markers and targets for the treatment of metabolic syndrome and diabetes. The approach can assess changes in circulating proteins resulting from any metabolic perturbation. Our bioinformatic analysis integrates these data with existing knowledge,

thereby providing a functional and clinical context for interpretation. Further streamlining of the analytical methods developed here will make it possible to measure many clinical studies, building up a comprehensive 'knowledge base' of the human plasma proteome in health and disease states (Geyer et al., 2017).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

NJWA and PEG designed, performed and interpreted the MS-based proteomic analysis of patient plasma and wrote the paper and generated the figures. SD, PVT, FM, LN, AS and EK interpreted proteomic data and revised the manuscript. KNBM, CM, SM, JJH, and SST provided patient material and clinical data and revised the manuscript. MM designed, and interpreted the MS-based proteomic analysis of patient plasma, supervised and guided the project and wrote the paper.

FUNDING

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

(WHO), W.H.O. (2016). Global Health Observatory (GHO) data. Overweight and obesity. http://www.who.int/gho/ncd/risk_factors/overweight/en/.

Aab, A., Abreu, P., Aglietta, M., Ahn, E.J., Al Samarai, I., Albuquerque, I.F., Allekotte, I., Allison, P., Almela, A., Alvarez Castillo, J., *et al.* (2016). Measurement of the Radiation Energy in the Radio Signal of Extensive Air Showers as a Universal Estimator of Cosmic-Ray Energy. Phys Rev Lett *116*, 241101.

Adams, T.D., Davidson, L.E., Litwin, S.E., Kim, J., Kolotkin, R.L., Nanjee, M.N., Gutierrez, J.M., Frogley, S.J., Ibele, A.R., Brinton, E.A., *et al.* (2017). Weight and Metabolic Outcomes 12 Years after Gastric Bypass. N Engl J Med *377*, 1143-1155.

13

Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature *537*, 347-355.

Afshin, A., Forouzanfar, M.H., Reitsma, M.B., Sur, P., Estep, K., Lee, A., Marczak, L., Mokdad, A.H., Moradi-Lakeh, M., Naghavi, M.*, et al.* (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N Engl J Med *377*, 13-27.

Arora, T., Velagapudi, V., Pournaras, D.J., Welbourn, R., le Roux, C.W., Oresic, M., and Backhed, F. (2015). Roux-en-Y Gastric Bypass Surgery Induces Early Plasma Metabolomic and Lipidomic Alterations in Humans Associated with Diabetes Remission. PLoS One *10*, e0126401.

Bojsen-Moller, K.N., Dirksen, C., Jorgensen, N.B., Jacobsen, S.H., Serup, A.K., Albers, P.H., Hansen, D.L., Worm, D., Naver, L., Kristiansen, V.B.*, et al.* (2014). Early enhancements of hepatic and later of peripheral insulin sensitivity combined with increased postprandial insulin secretion contribute to improved glycemic control after Roux-en-Y gastric bypass. Diabetes *63*, 1725-1737.

Bojsen-Moller, K.N., Dirksen, C., Svane, M.S., Jorgensen, N.B., Holst, J.J., Richter, E.A., and Madsbad, S. (2017). Variable reliability of surrogate measures of insulin sensitivity after Roux-en-Y gastric bypass. Am J Physiol Regul Integr Comp Physiol *312*, R797-R805.

Cardoso, L., Rodrigues, D., Gomes, L., and Carrilho, F. (2017). Short- and long-term mortality after bariatric surgery: a systematic review and meta-analysis. Diabetes Obes Metab.

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.

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., 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.

Cummings, D.E., and Rubino, F. (2018). Metabolic surgery for the treatment of type 2 diabetes in obese individuals. Diabetologia *61*, 257-264.

Dali-Youcef, N., Mecili, M., Ricci, R., and Andres, E. (2013). Metabolic inflammation: connecting obesity and insulin resistance. Ann Med *45*, 242-253.

Emerging Risk Factors, C., Di Angelantonio, E., Gao, P., Pennells, L., Kaptoge, S., Caslake, M., Thompson, A., Butterworth, A.S., Sarwar, N., Wormser, D.*, et al.* (2012). Lipid-related markers and cardiovascular disease prediction. JAMA *307*, 2499-2506.

Finnerty, C.C., Mabvuure, N.T., Ali, A., Kozar, R.A., and Herndon, D.N. (2013). The surgically induced stress response. JPEN J Parenter Enteral Nutr *37*, 21S-29S.

Geyer, P.E., Holdt, L.M., Teupser, D., and Mann, M. (2017). Revisiting Biomarker Discovery by Plasma Proteomics. Molecular Systems Biology *Submitted*.

Geyer, P.E., Kulak, N.A., Pichler, G., Holdt, L.M., Teupser, D., and Mann, M. (2016a). Plasma Proteome Profiling to Assess Human Health and Disease. Cell Syst *2*, 185-195.

Geyer, P.E., Wewer Albrechtsen, N.J., Tyanova, S., Grassl, N., Iepsen, E.W., Lundgren, J., Madsbad, S., Holst, J.J., Torekov, S.S., and Mann, M. (2016b). Proteomics reveals the effects of sustained weight loss on the human plasma proteome. Mol Syst Biol *12*, 901.

Glass, C.K., and Olefsky, J.M. (2012). Inflammation and lipid signaling in the etiology of insulin resistance. Cell Metab *15*, 635-645.

Griffen, W.O., Jr., Young, V.L., and Stevenson, C.C. (1977). A prospective comparison of gastric and jejunoileal bypass procedures for morbid obesity. Ann Surg *186*, 500-509.

Hegele, R.A. (2009). Plasma lipoproteins: genetic influences and clinical implications. Nat Rev Genet *10*, 109-121.

Iepsen, E.W., Lundgren, J., Dirksen, C., Jensen, J.E., Pedersen, O., Hansen, T., Madsbad, S., Holst, J.J., and Torekov, S.S. (2015). Treatment with a GLP-1 receptor agonist diminishes the decrease in free plasma leptin during maintenance of weight loss. Int J Obes (Lond) *39*, 834-841.

Jong, M.C., Voshol, P.J., Muurling, M., Dahlmans, V.E., Romijn, J.A., Pijl, H., and Havekes, L.M. (2001). Protection from obesity and insulin resistance in mice overexpressing human apolipoprotein C1. Diabetes *50*, 2779-2785.

Keilhauer, E.C., Geyer, P.E., and Mann, M. (2016). HCD Fragmentation of Glycated Peptides. J Proteome Res *15*, 2881-2890.

Kontush, A., Lhomme, M., and Chapman, M.J. (2013). Unraveling the complexities of the HDL lipidome. J Lipid Res *54*, 2950-2963.

Krauss, R.M. (2010). Lipoprotein subfractions and cardiovascular disease risk. Curr Opin Lipidol *21*, 305-311.

Kulak, N.A., Geyer, P.E., and M., M. (2017). Loss-less nano-fractionator for high sensitivity, high coverage proteomics. Molecular Cellular Proteomics; Manuscript in Press.

Li, S., Shin, H.J., Ding, E.L., and van Dam, R.M. (2009). Adiponectin levels and risk of type 2 diabetes: a systematic review and meta-analysis. JAMA *302*, 179-188.

Lihn, A.S., Pedersen, S.B., and Richelsen, B. (2005). Adiponectin: action, regulation and association to insulin sensitivity. Obes Rev *6*, 13-21.

Luo, P., Yu, H., Zhao, X., Bao, Y., Hong, C.S., Zhang, P., Tu, Y., Yin, P., Gao, P., Wei, L._, et al._ (2016). Metabolomics Study of Roux-en-Y Gastric Bypass Surgery (RYGB) to Treat Type 2 Diabetes Patients Based on Ultraperformance Liquid Chromatography-Mass Spectrometry. J Proteome Res *15*, 1288-1299.

Madsbad, S., and Holst, J.J. (2014). GLP-1 as a mediator in the remission of type 2 diabetes after gastric bypass and sleeve gastrectomy surgery. Diabetes *63*, 3172-3174.

Martinussen, C., Bojsen-Moller, K.N., Dirksen, C., Jacobsen, S.H., Jorgensen, N.B., Kristiansen, V.B., Holst, J.J., and Madsbad, S. (2015). Immediate enhancement of first-phase insulin secretion and unchanged glucose effectiveness in patients with type 2 diabetes after Roux-en-Y gastric bypass. Am J Physiol Endocrinol Metab *308*, E535-544.

Mason, E.E., and Ito, C. (1967). Gastric bypass in obesity. Surg Clin North Am *47*, 1345-1351.

Mason, E.E., Printen, K.J., Hartford, C.E., and Boyd, W.C. (1975). Optimizing results of gastric bypass. Ann Surg *182*, 405-414.

Meier, F., Geyer, P.E., Cox, J., and Mann, M. (2018). BoxCar method enables single shot proteomics at a depth of 10,000 proteins in 100 minutes. Under revision.

Miras, A.D., and le Roux, C.W. (2013). Mechanisms underlying weight loss after bariatric surgery. Nat Rev Gastroenterol Hepatol *10*, 575-584.

Moran, A., Steffen, L.M., Jacobs, D.R., Jr., Steinberger, J., Pankow, J.S., Hong, C.P., Tracy, R.P., and Sinaiko, A.R. (2005). Relation of C-reactive protein to insulin resistance and cardiovascular risk factors in youth. Diabetes Care *28*, 1763-1768.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol Cell Proteomics *11*, M111 013722.

Ndumele, C.E., Pradhan, A.D., and Ridker, P.M. (2006). Interrelationships between inflammation, C-reactive protein, and insulin resistance. J Cardiometab Syndr *1*, 190-196.

Puzziferri, N., Roshek, T.B., 3rd, Mayo, H.G., Gallagher, R., Belle, S.H., and Livingston, E.H. (2014). Long-term follow-up after bariatric surgery: a systematic review. JAMA *312*, 934-942.

Rader, D.J., and Hovingh, G.K. (2014). HDL and cardiovascular disease. Lancet *384*, 618-625.

Sjostrom, L., Narbro, K., Sjostrom, C.D., Karason, K., Larsson, B., Wedel, H., Lystig, T., Sullivan, M., Bouchard, C., Carlsson, B._, et al._ (2007). Effects of bariatric surgery on mortality in Swedish obese subjects. N Engl J Med *357*, 741-752.

15

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M., Geiger, T., Mann, M., Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods.

VerHague, M.A., Cheng, D., Weinberg, R.B., and Shelness, G.S. (2013). Apolipoprotein A-IV expression in mouse liver enhances triglyceride secretion and reduces hepatic lipid content by promoting very low density lipoprotein particle expansion. Arterioscler Thromb Vasc Biol *33*, 2501-2508.

Wittgrove, A.C., Clark, G.W., and Tremblay, L.J. (1994). Laparoscopic Gastric Bypass, Roux-en-Y: Preliminary Report of Five Cases. Obes Surg *4*, 353-357.

Worm, D., Madsbad, S., Kristiansen, V.B., Naver, L., and Hansen, D.L. (2015). Changes in Hematology and Calcium Metabolism After Gastric Bypass Surgery--a 2-Year Follow-Up Study. Obes Surg *25*, 1647-1652.

FIGURE TITLES AND LEGENDS

Figure 1: Study design and analytical performance.
(A) The two studies involved 47 patients who have undergone Roux-en-Y gastric bypass. The boxplots for the BMI distribution indicate a 10-90 percentile and the median is highlighted with a black bar.
(B) Highly abundant proteins were only depleted for the library generation in seven pools of plasma. After digestion peptides were separated at high pH on a reversed phase material into 24 fractions.
(C) Plasma Proteome Profiling pipeline applied to 175 plasma proteomes in triplicate. Sample preparation took a total of 4h and 45 min gradients were used, resulting in a 3D data matrix of quantified proteins as a function of individuals over time.
(D) Number of identified proteins in each of the seven library pools (red dots) and the 175 plasma samples (blue dots).
(E) Total number of proteins identified in the library and in the study cohort.
(F) Quantitative values of 1700 plasma proteins ranked according to their abundance. Several proteins are exemplified over the abundance range (blue dots). Functional annotation and 1D enrichment resulted in 117 significantly enriched annotation terms (right graph, grey dots). Eight of these categories are highlighted as boxplots with a 10-90 percentile.

Figure 2: Longitudinal trajectories and functional interpretation of significantly influenced proteins.
(A) Proteins clustered into four main groups according to their patterns of adaption after RYGB. Z-scores are plotted over five time points.
(B) Hierarchical clustering of the plasma proteins in the four groups.
(C) A Fischer exact test identified 20 main keywords, which were used in a second hierarchical clustering analysis for functional interpretation of the longitudinal trajectories. The dominant keywords – related to inflammation and the lipid homeostasis system - are highlighted by red and blue rectangles, respectively.

Figure 3: Global correlation map of the plasma proteome.
(A) Pairwise correlation of proteins using their intensity levels over the 175 study samples results in a matrix of correlation coefficients where one protein is compared to all other proteins and clinical parameters. The left panel illustrates the calculation of the Pearson correlation coefficient between two proteins with high and low correlation, respectively. The right hand panel shows cross-correlations of all quantified proteins and clinical parameters (25% valid values) after subsequent hierarchical clustering, revealing an extensive map of co-regulated items (proteins,

170

clinical parameters, peptide glycation). In the map, items with a strong correlation or anti-correlation to each other cluster together in red or blue areas, respectively. The main clusters are functionally annotated with keywords.

(B) Network analysis of the main clusters in (A) reveals connections within and between them. The thickness of the lines corresponds to the Pearson correlation coefficient (Red lines: positive correlations; blue lines: anti-correlations; green cycles: clinical parameters).

Figure 4: Connection between global inflammation status and body mass.

(A) Changes of the global inflammation status of all study participants. The global inflammation status is calculated as the median of 51 inflammation proteins, selected by keyword annotation. Pie charts over each of the post-operation time points show the number of proteins, which were up or down-regulated.

(B) Overlap of inflammation proteins in the caloric restriction (CR) and the RYGB studies. Plus and minus indicates if a protein was decreased in both or in one of the two studies. The color code shows how many of the proteins were significantly regulated.

(C) Correlation plot of protein intensities with the BMI in the RYGB study. Inflammation proteins are highlighted in red or orange if they were significantly or not significantly correlated to the BMI, respectively. All proteins above the dashed line show significant correlations.

Figure 5: Response of lipid homeostasis regulating proteins to RYGB and caloric restriction.

(A) Correlation maps of lipid homeostasis associated proteins in the RYGB study.

(B) Correlation maps of lipid homeostasis associated proteins in the caloric restriction study.

(C) Median of protein intensities showing increased or decreased expression after RYGB and caloric restriction (CR). Significantly changed proteins are marked with an asterisk.

Figure 6: Correlation to insulin sensitivity assessments.

(A) Correlation of the plasma proteome profile to the four insulin sensitivity assessments: hepatic insulin sensitivity, homeostatic model assessment of insulin resistance based on C-peptide (HOMA-IR); peripheral insulin sensitivity; oral glucose insulin sensitivity. Pearson correlations were rank-ordered and proteins grouped, scored and color-coded from 1-4.

(B) Example correlations of four proteins to the insulin sensitivity assessments.

(C) Trajectories of the summed insulin sensitivity correlations defined in (A) over time. Blue RYGB study 1 and Grey, RYGB study 2.

STAR METHODS

Study design

RYGB study 1: Ten obese patients with type 2 diabetes (age 43.6 ± 3.4 years, male/female 4/6, BMI 38.9 ± 1.6 kg/m², HbA1c 7.0 ± 0.3%, diabetes duration 3.3 ± 1.0 years) and ten obese patients with normal glucose tolerance (age 40.1 ± 2.8 years, male/female 3/7, BMI 40.2 ± 0.8 kg/m², HbA1c 5.4 ± 0.1%) scheduled for laparoscopic Roux-en-Y gastric bypass (RYGB) were investigated before (n=20), 1 week (n=16), 3 months (n=20) and 1 year (n=18) postoperatively (Clinical trial NCT 01202526), as described previously (Bojsen-Moller et al., 2014). At 4 years post-RYGB, an extension study (NCT 03046147) was initiated and 16 patients of the initial cohort accepted to be

17

included. Written informed consent was obtained from all participants and the study was approved by the Municipal Ethical Committee of Copenhagen in accordance with the Helsinki declaration and by the Danish Data Protection Agency. Triple sampling at fasting was performed at all study visits, while hyperinsulinemic-euglycemic clamps (HEC) including basal glucose tracer infusions were performed at all except the 4 year's visit as previously described in detail. Oral glucose tolerance tests (OGTTs) were done on separate study days before, 3 months, 1 year and 4 years after RYGB. Antidiabetic agents were discontinued ≥3 days before each study day, and participants were instructed to refrain from strenuous physical activity and alcohol for 3 days and to fast overnight (10-12 h) prior to all experiments. Calculations: HOMA2-IR was calculated from all study visits based on a triple sampling of fasting glucose and C-peptide concentrations using the HOMA2. Hepatic insulin sensitivity and peripheral insulin sensitivity were estimated from HEC experimental days as the tracer determined basal rate of appearance (Ra) of glucose (in mg per minute) with correction for C-peptide concentrations and the tracer determined rate of disappearance (Rd) of glucose during the HEC expressed as mg per min per kg fat free mass (ffm), respectively. The Oral Glucose Insulin Sensitivity index was estimated from all OGTTs and has recently been validated against the HEC in this particular cohort (Bojsen-Moller et al., 2017).

RYGB study 2: Participants: Ten obese patients with type 2 diabetes (age 46.1 ± 2.8 years, male/female 4/6, BMI 41.2 ± 1.3 kg/m^2, HbA1c 45.1 ± 2.2 mmol/mol, diabetes duration 3.6 ± 1.3 years), seven obese patients with impaired glucose tolerance (age 41.8 ± 3.7 years, male/female 2/5, BMI 40.3 ± 1.6 kg/m^2, HbA1c 40.0 ± 1.0 mmol/mol) and ten obese patients with normal glucose tolerance (age 41.6 ± 3.0 years, male/female 4/6, BMI 43.8 ± 1.5 kg/m^2, HbA1c 35.0 ± 1.3 mmol/mol) were examined before (n=27) and 1 week (n=25) and 3 months (n=25) after RYGB (NCT01993511) as previously described (Martinussen et al., 2015). A few of the patients were also investigated after 1 year (n=6) and 2 years (n=3) postoperatively. Written informed consent was obtained from all participants and the study was approved by the Municipal Ethical Committee of Copenhagen in accordance with the Helsinki declaration and by the Danish Data Protection Agency. Triple sampling at fasting was performed at all study visits. Oral glucose tolerance testing was done on separate study days before (n=23) and 3 months (n=23) and 1 year (n=5) after RYGB. Antidiabetic medication was discontinued >3 days (>10 days for Liraglutide) prior to surgery and patients were instructed to fast overnight (10-12 hours) prior to testing. Calculations: HOMA2-IR and oral glucose insulin sensitivity was obtained as described for study 1.

High abundant protein depletion for building a matching library

To construct a library of peptide identifications, we used two commercial depletion kits that together remove the top14 highest abundant proteins in plasma (Geyer et al., 2016a). Following depletion, we separated our samples into 24 fractions, using our recently developed high-pH reversed-phase 'Spider fractionator' (Figure 1B) (Kulak et al., 2017).

Sample preparation for study samples

Plasma samples for the quantitative analysis of the two bariatric surgery cohorts were prepared for all samples according to the previously published Plasma Proteome Profiling pipeline (Geyer et al., 2016a). In brief, this involved optimized conditions for denaturation, alkylation, digestion and peptide purification, and liquid chromatography and MS settings. Sample preparation was carried out in an automated liquid handling platform (Agilent Bravo) in a 96 well format. For sample preparation, we used the 'iST' Kit for proteomic sample preparation (P.O. 00001, PreOmics GmbH). Depletion was only used for the library and not for the study samples.

High pressure liquid chromatography and mass spectrometry

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1200 system (Thermo Fisher Scientific), which was combined with a Q Exactive HF Orbitrap (Thermo Fisher Scientific) and a nano-electrospray ion source (Thermo Fisher Scientific). Purified peptides were separated on 40 cm HPLC-columns (ID: 75 μm; in-house packed into the tip with ReproSil-Pur C18-AQ 1.9 μm resin (Dr. Maisch GmbH)). For each LC-MS/MS analysis, around 1 μg peptides were used for the 45 min gradients and for the fractions of the deep plasma data set.

Peptides were loaded in buffer A (0.1% formic acid, 5% DMSO (v/v)) and eluted with a linear 35 min gradient of 3-30% of buffer B (0.1% formic acid, 5% DMSO, 80% (v/v) acetonitrile), followed by a 7 min increase to 75% of buffer B and a 1 min increase to 98% of buffer B, and a 2 min wash of 98% buffer B at a flow rate of 450 nl/min. Column temperature was kept at 60 °C by a Peltier element containing in-house-developed oven. MS data were acquired with a Top15 data-dependent MS/MS scan method (topN method) for the library and the BoxCar scan method (Meier et al., 2018) for the study samples. Target values for the full scan MS spectra was 3×10^6 charges in the 300-1,650 m/z range with a maximum injection time of 55 ms and a resolution of 120,000 at m/z 200. Fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27 eV. MS/MS scans were performed at a resolution of 15,000 at m/z 200 with an ion target value of 5×10^4 and a maximum injection time of 25 ms.

Data analysis

MS raw files were analyzed using the MaxQuant software (Cox and Mann, 2008) and peptide lists were searched against the human Uniprot FASTA database with the Andromeda search engine (Cox et al., 2011). A contaminants database was employed and cysteine carbamidomethylation was set as a fixed modification and N-terminal acetylation and methionine oxidations as variable modifications. False discovery rate (FDR) was 0.01 for both the protein and peptide level with a minimum length of 7 amino acids for peptides and this FDR was determined by searching a reverse sequence database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin and LysC as proteases, and a maximum of two missed cleavages were allowed. Peptides were identified with an initial precursor mass deviation of up to 7 ppm and a fragment mass deviation of 20 ppm. 'Match between run algorithm' in MaxQuant

(Nagaraj et al., 2012) was performed after constructing a matching library consisting of depleted plasma samples. All proteins and peptides matching to the reversed database were filtered out. For label-free protein quantitation (LFQ) we required a minimum ratio count of 2 (Cox et al., 2014). Glycated peptide were identified as previously reported (Keilhauer et al., 2016).

Bioinformatic analysis

All bioinformatic analyses were done with the Perseus software of the MaxQuant computational platform (Cox and Mann, 2008; Tyanova, 2016). For statistical analysis of significantly changed proteins before and after RYGB, we used a one-sample t-test with a false discovery rate of <0.05 after Benjamini-Hochberg correction. Correlation to clinical parameters like BMI and insulin resistance data to MS-based proteomic data were also done within the Perseus environment.

Global correlation analysis: Proteins were only considered for the analysis, if they were present in at least 25% of all samples, which is equal to 44 out of the 175 measured samples. After removal of contaminants like platelet proteins, this left 580 proteins. Peptide glycation: Quantified glycation sites were also filtered for at least 25% valid values and Z-scored over all individuals and time points. The median of this normalized index over 203 glycation sites is reported as the 'peptide glycation' factor.
Inflammation system: We filtered for proteins of the inflammation system by including keywords enriched in the analysis of Figure 2 (complement pathway, immunity, innate immunity, membrane attack complex, cytolysis, complement alternate pathway, inflammatory response, antimicrobial) and further included the term 'acute phase' as it added well-known inflammation proteins like the CRP.
Lipid homeostasis system: We filtered the plasma proteome for the gene ontology cellular compartment (GOCC) keywords lipid transport, chylomicron, high density lipoprotein (HDL), low density lipoprotein (LDL), intermediate density lipoprotein (IDL) and very low density lipoprotein (vLDL) particles. We further included non-annotated proteins well known to associate with lipid metabolism: apolipoprotein D (APOD), apolipoprotein(a) (LPA), Prolow-density lipoprotein receptor-related protein 1 (LRP1) and phospholipid transfer protein (PLTP), and excluded the keyword annotated inflammation proteins SAA1, SAA2 and SAA4, which are not associated with lipoprotein particles. Filtering for at least 10% valid values, resulted in a set of 25 proteins (Table S9). The median changes of the proteins were calculated across all time points.
Scoring system for insulin sensitivity: We ranked proteins correlating to the four insulin sensitivity parameters ('homeostatic model assessment of insulin resistance' (HOMA-IR), 'hepatic insulin sensitivity', 'peripheral insulin sensitivity', 'oral glucose insulin sensitivity') according to their Pearson correlation coefficient. The five proteins with the highest correlation to each of the assessments were assigned the score 4, the next five proteins 3, the following five 2 and the rest 1 (Table S8).
Insulin sensitivity score: We defined an overall score for each protein adding the individual scores in each of the four insulin sensitivity assays. Next, we normalized the protein trajectories over time, using baseline level before surgery for each individual as 100%. The score weighted the impact of each of the proteins and the median was plotted for both RYGB studies over time.
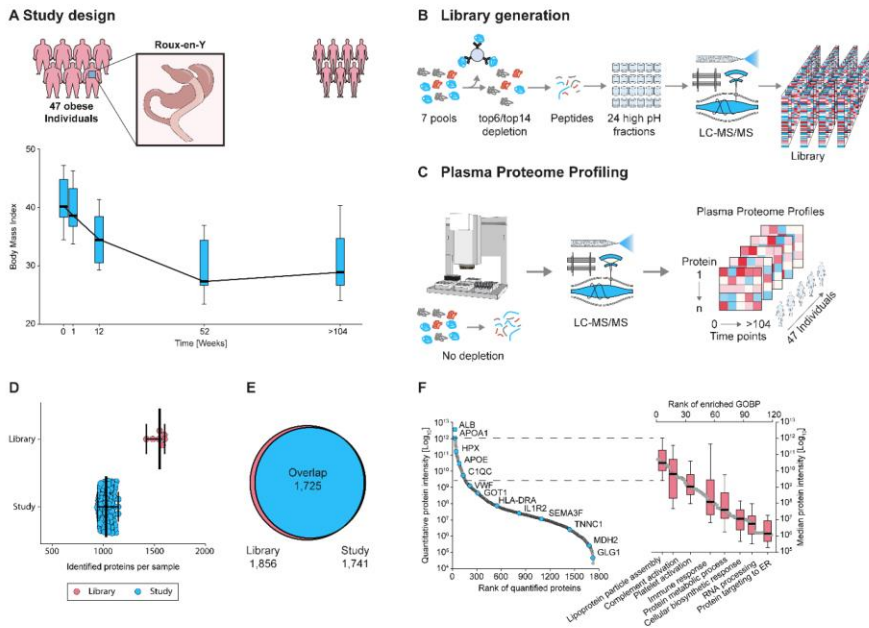
20

174

FIGURES

Figure 1: Study design and analytical performance.
(A) The two studies involved 47 patients who have undergone Roux-en-Y gastric bypass. The boxplots for the BMI distribution indicate a 10-90 percentile and the median is highlighted with a black bar.
(B) Highly abundant proteins were only depleted for the library generation in seven pools of plasma. After digestion peptides were separated at high pH on a reversed phase material into 24 fractions.
(C) Plasma Proteome Profiling pipeline applied to 175 plasma proteomes in triplicate. Sample preparation took a total of 4h and 45 min gradients were used, resulting in a 3D data matrix of quantified proteins as a function of individuals over time.
(D) Number of identified proteins in each of the seven library pools (red dots) and the 175 plasma samples (blue dots).
(E) Total number of proteins identified in the library and in the study cohort.
(F) Quantitative values of 1700 plasma proteins ranked according to their abundance. Several proteins are exemplified over the abundance range (blue dots). Functional annotation and 1D enrichment resulted in 117 significantly enriched annotation terms (right graph, grey dots). Eight of these categories are highlighted as boxplots with a 10-90 percentile.

1

Figure 2: Longitudinal trajectories and functional interpretation of significantly influenced proteins.
(A) Proteins clustered into four main groups according to their patterns of adaption after RYGB. Z-scores are plotted over five time points.
(B) Hierarchical clustering of the plasma proteins in the four groups.

(C) A Fischer exact test identified 20 main keywords, which were used in a second hierarchical clustering analysis for functional interpretation of the longitudinal trajectories. The dominant keyword – related to inflammation and the lipid homeostasis system - are highlighted by red and blue rectangles, respectively.

2

176

Figure 3: Global correlation map of the plasma proteome.

(A) Pairwise correlation of proteins using their intensity levels over the 175 study samples results in a matrix of correlation coefficients where one protein is compared to all other proteins. The left panel illustrates the calculation of the Pearson correlation coefficient between two proteins with high and low correlation, respectively. The right hand panel shows cross-correlations of all quantified proteins and clinical parameters (25% valid values) after subsequent hierarchical clustering, revealing an extensive map of co-regulated items (proteins, clinical parameters, peptide glycation factor). Items with a strong correlation to each other cluster together in red or blue areas in the map. The main clusters are functionally annotated with keywords.

(B) Network analysis of the main clusters in (A) reveals connections within and between them. The thickness of the lines corresponds to the Pearson correlation coefficient (Red lines: positive correlations; blue lines: anti-correlations; green cycles: clinical parameters).

Figure 4: Connection between global inflammation status and body mass.

(A) Changes of the global inflammation status of all study participants. The global inflammation status is calculated as the median of 51 inflammation proteins selected by keyword annotation. Pie charts over each of the post-operation time points show the number of proteins, which were up or down-regulated.

(B) Overlap of inflammation proteins in the caloric restriction (CR) and the RYGB studies. Plus and minus indicates if a protein was decreased in both or in one of the two studies. The color code shows how many of the proteins were significantly regulated.

(C) Correlation plot of protein intensities with the BMI in the RYGB study. Inflammation proteins are highlighted in red or orange if they were significantly or not significantly correlated to the BMI. All proteins above the dashed line show significant correlations.

Figure 5: Response of lipid homeostasis regulating proteins to RYGB and caloric restriction.
(A) Correlation maps of lipid homeostasis associated proteins in the RYGB study.
(B) Correlation maps of lipid homeostasis associated proteins in the caloric restriction study.
(C) Median of protein intensities showing increased or decreased expression after RYGB and caloric restriction (CR). Significantly changed proteins are marked with an asterisk.
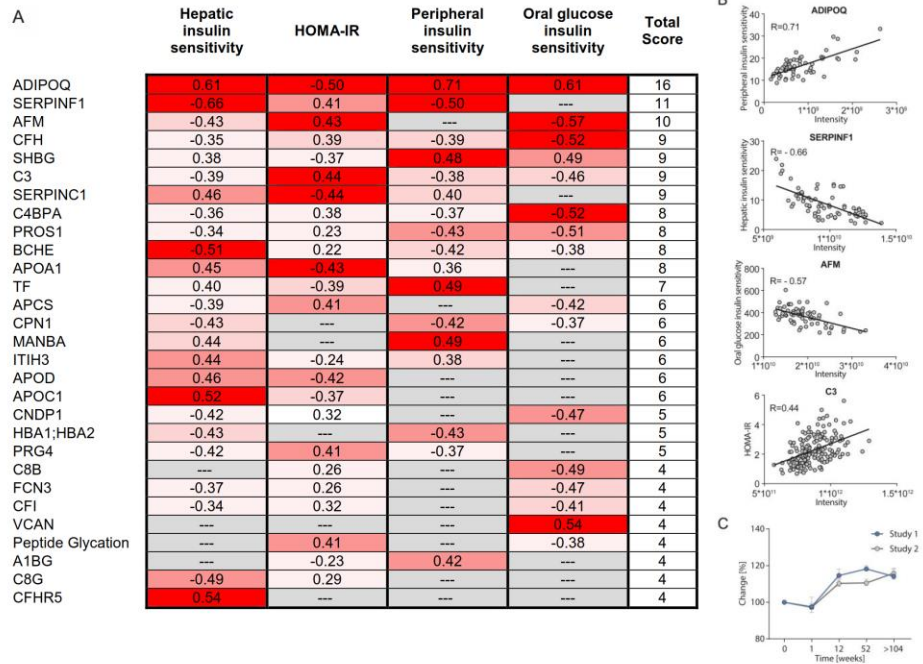
180

A

	Hepatic insulin sensitivity	HOMA-IR	Peripheral insulin sensitivity	Oral glucose insulin sensitivity	Total Score
ADIPOQ	0.61	-0.50	0.71	0.61	16
SERPINF1	-0.66	0.41	-0.50	---	11
AFM	-0.43	0.43	---	-0.57	10
CFH	-0.35	0.39	-0.39	-0.52	9
SHBG	0.38	-0.37	0.48	0.49	9
C3	-0.39	0.44	-0.38	-0.46	9
SERPINC1	0.46	-0.44	0.40	---	9
C4BPA	-0.36	0.38	-0.37	-0.52	8
PROS1	-0.34	0.23	-0.43	-0.51	8
BCHE	-0.51	0.22	-0.42	-0.38	8
APOA1	0.45	-0.43	0.36	---	8
TF	0.40	-0.39	0.49	---	7
APCS	-0.39	0.41	---	-0.42	6
CPN1	-0.43	---	-0.42	-0.37	6
MANBA	0.44	---	0.49	---	6
ITIH3	0.44	-0.24	0.38	---	6
APOD	0.46	-0.42	---	---	6
APOC1	0.52	-0.37	---	---	6
CNDP1	-0.42	0.32	---	-0.47	5
HBA1;HBA2	-0.43	---	-0.43	---	5
PRG4	-0.42	0.41	-0.37	---	5
C8B	---	0.26	---	-0.49	4
FCN3	-0.37	0.26	---	-0.47	4
CFI	-0.34	0.32	---	-0.41	4
VCAN	---	---	---	0.54	4
Peptide Glycation	---	0.41	---	-0.38	4
A1BG	---	-0.23	0.42	---	4
C8G	-0.49	0.29	---	---	4
CFHR5	0.54	---	---	---	4

B

C

Figure 6: Correlation to insulin sensitivity assessments.
(A) Correlation of the plasma proteome profile to the four insulin sensitivity assessments: hepatic insulin sensitivity, homeostatic model assessment of insulin resistance based on C-peptide (HOMA-IR); peripheral insulin sensitivity; oral glucose insulin sensitivity. Pearson correlations were rank-ordered and proteins grouped, scored and color-coded from 1-4.
(B) Example correlations of four proteins to the insulin sensitivity assessments.
(C) Change of the summed insulin sensitivity correlation defined in (A) over time. Blue RYGB study 1 and Grey, RYGB study 2.

7

181

Acknowledgements