Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Mass spectrometry-based proteomics: from deep proteomes to clinical application

Johannes Bruno Müller-Reif,
geb. Müller

aus
Nürnberg, Deutschland

2021

Erklärung

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

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 02.05.2021

_____

Johannes Bruno Müller-Reif

Dissertation eingereicht am 06.05.2021

1. Gutachter: Hon.-Prof. Dr. Matthias Mann _____

2. Gutachter: Apl. Prof. Dr. Henrik Daub _____

Mündliche Prüfung am 07.06.2021

Abstract

Proteins perform the vast majority of molecular functions in biological life and the quantitative investigation of the proteome, the sum of all proteins in a system, has been an interest in the life sciences for the last decades. Technological developments that have shaped the field are advancing unabated, playing a crucial role in state-of-the-art proteomics research. With mass spectrometry (MS)-based methods and ultra-high-performance liquid chromatography (UHPLC), thousands of proteins can be quantitatively measured in a cell culture digest, in patient tissue or most other biological specimens in under an hour of analysis time. Liquid chromatography techniques play an especially important role because they distribute the overwhelming bulk of analytes into a time-ordered landscape of eluting peaks.

One focus of this PhD thesis was the development and application of advanced chromatographic methods for MS-based proteomics. This encompasses the foundational publication of the Evosep One system, which has become a valuable instrument for reproducible high throughput proteomics studies as well as a high-profile application of the novel µPAC column, a chip based nano-flow separation device which I used for the proteomics measurements of 100 taxonomically diverse organisms. Furthermore, to enhance throughput and quality, I designed a high-pressure packing station for the production of capillary columns, which are still the mainstay in the proteomics field, making their production more than hundred times more time efficient.

The main project of my PhD is the ‘proteome landscape of the kingdoms of life’, an unprecedented investigation of 100 organisms’ proteomes from across all of known biological life. For the first time, this makes it possible to compare proteomes from organisms of all domains. We hope to have showcased the universal application of proteomics to facilitate model organism independent and unbiased research in the future.

Finally, I also addressed the use of proteomics techniques for clinical studies. The investigation of contamination markers in common blood-based searches lays a foundation for biomarker studies, which can be extended to other matrices like urine or cerebrospinal fluid (CSF). In the latter we uncover potential new biomarkers for Alzheimer’s Disease in a multicentric study.

In summary, this thesis is a cross-section of state-of the art MS-based proteomics from technological developments through deep organism proteomes across the tree of life to clinical applications.

Table of contents

1. Introduction 1

1.1. Liquid chromatographic techniques in mass spectrometry-based proteomics 7

1.1.1. Particle packed columns 8

1.1.2. Chip-based stationary phases 9

1.1.3. Benefits and drawbacks of nano- and microflow liquid chromatography 12

1.2. Mass spectrometry-based proteomics 13

1.3. Scan modes for MS-based proteomics 19

1.4. MS-based proteomics as a tool for multi-organism studies 22

1.5. Organism case study - bear proteomics 26

1.6. Clinical proteomics 27

1.7. Clinical study design for proteomics 30

2. Aims of the thesis 33

3. Publications 35

3.1. Article 1: The proteome landscape of the kingdoms of life 35

3.2. Article 2: A new high-pressure packing system enables rapid multiplexed packing of capillary columns 44

3.3. Article 3: A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics 60

3.4. Article 4: Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies 75

3.5. Article 5: Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease 88

3.6. Article 6: Molecular origin of blood-based infrared fingerprints 107

3.7. Article 7: Cohort Profile: MUNICH Preterm and Term Clinical Study (MUNICH-PreTCl) 131

3.8. Article 8: Cotranslational N-degron masking by acetylation promotes proteome stability in plant 154

4. Discussion 172

5. References 175

6. Acknowledgements 190

Abbreviations

AA	Amino acid
C-18	Carbon-18
CID	Collision induced dissociation
CRISPR/Cas	Clustered Regularly Interspaced Short Palindromic Repeats / Caspase method
CSF	Cerebrospinal fluid
DARTS	Drug affinity responsive target stability
DBS	Dried blood spot
DC	Direct current
DDA	Data dependent acquisition
DIA	Data independent acquisition
diGly	di-Glycin
DNA	Desoxyribonucleic acid
DTT	Dithiothreitol
EggNOG	Evolutionary genealogy of genes: Non-supervised Orthologous Groups
ELISA	Enzyme-linked immunosorbent assay
ES	Electro spray
ETD	electron transfer dissociation
FAIMS	Field asymmetric waveform ion mobility spectrometry
FASP	Filter assisted sample preparation
FTIR	Fourier transform infrared
GO	Gene ontology
HCD	higher-energy C-trap collisional dissociation
HPLC	High performance liquid chromatography
ID	Inner diameter
IMF	infrared molecular fingerprints
LC	Liquid chromatography
LSTM	Long short term memory
m/z	Mass to charge ratio
MALDI	Matrix-assisted laser desorpion/ionisation
mRNA	Messenger ribo nucleic acid

MS	Mass spectrometry
MS1	Precursor mass scan
MS2	Precursor fragment mass scan
MS3	2nd fragmentation scan
NTA	Nitrilotriacetic acid
PASEF	Parallel accumulation parallel fragmentation
PCR	Polymerase chain reaction
POI	Protein of interest
PPP	Plasma protein profiling
PRM	parallel reaction monitoring
PTM	Post translational modification
RF	Radio frequency
RNA	Ribonucleic acid
SDC	Sodium deoxycholate
SDS	Sodium dodecyl sulfate
SIM	selected ion monitoring
SRM	selected reaction monitoring
TCEP	Tris-(2-carboxyethyl)-phosphin
Th	Thompson
Tims	Trapped ion mobility spectrometry
TOF	Time of flight
UVPD	ultra-violet photodissociation

1. Introduction

The central dogma of cell biology defines proteins, the products of gene expression via the transcriptional and translational cellular machinery, as the main effector molecules of biological life. While the one-dimensionality of the central dogma has long been revised, the fact that proteins carry out the vast majority of functions in living cells remains. Long before this had been widely acknowledged, protein analysis was recognized as a major field of medical, biological, chemical and biochemical research. Already in the 18th century proteins were isolated and described qualitatively for scientific reasons (J. B. Beccari, 1731). The name 'Protein' originated in 1839, when Gerrit J. Mulder, who employed element analyses to describe the molecular composition of various 'albuminous materials' found that the majority of the compounds he analyzed had the same molecular formula (Mulder, 1839). For these compounds he borrowed the Greek word *proteios*, and coined the presently still used denomination.

The following century in protein research was mainly spent in exploration of the primary structure of proteins. This was only achieved in 1902 by Franz Hofmeister who postulated that proteins are a condensate from amino acids (Hofmeister, 1889) and later confirmed by Emil Fischer, who found that dipeptides are the smallest macromolecular form of proteins which can be gained by hydrolysis of proteins. He introduced the term peptides and founded the field of peptide synthesis (Fischer, 1906).

Protein analysis in the following century focused on describing their context in biological systems (central dogma of cell biology) or the overall structure (starting with crystal structures) but most analyses relied on size, chemical properties and affinity-based methods to identify proteins. The first step towards sequence-based identification of proteins was done by Paul Schlack and W. Kumpf in 1926 by developing a process to sequentially cleave peptides from a protein's C-terminus (Schlack & Kumpf, 1926). The nowadays better-known Edman degradation was invented by Pehr Edman in 1949 and can be used to sequentially analyze peptides up to 50 residues from the N-terminus (Edman, 1949).

While these methods seem anachronistic compared to modern protein analysis tools, they established the fact that peptides of proteins can further be sequentially cleaved between their amino acids to identify the primary structure of the peptide and subsequently the protein. This still forms the basis for modern mass spectrometric (MS) methods for proteomics analyses.

The special case of proteomics

Proteomics mainly differs from the biologically upstream omics techniques genomics and transcriptomics for a single reason: Genomics and transcriptomics rely on sequence amplification which is biotechnologically enabled by enzymes 'high-jacked' from natural DNA replication, transcription and reverse-transcription machineries found in various organisms (Saiki et al., 1988; Weis et al., 1992). There is no known biological function or process engineered which can do the same for proteins. In contrast to biologically downstream omics techniques like metabolomics or lipidomics, proteomics shares the sequence type analysis with the biologically upstream techniques. DNA, RNA and proteins consist of a number of fixed building blocks (generally) in linear combination and characterization simply means to sequence the building blocks and assemble them in the right order. The same is not true for small molecules or the broad variety of lipid structures, where bulk must be done with a fingerprint type of information because structures do not give rise to a systematic sequential approach as for proteins (Peake, 2018; Wishart, 2011; Yang & Han, 2016).

Proteins are diverse molecules and very variable in size and chemical properties. This is another difference from the upstream omics techniques, where the restriction to four initial bases as building blocks for DNA and RNA limits their variability. For instance, mRNA for transcriptomics analysis can always be accessed via its polyA tail. Proteins in general do not share such patterns and whole proteome analysis is difficult because few techniques enable access to all chemically diverse proteins. The method to make proteins accessible for molecular omics techniques is to break them down into peptides of a length of manageable size, whose properties are much more similar to each other than the properties of different proteins. This approach is called bottom-up proteomics and the method of choice to break the proteins down in peptides is enzymatic digestion. The most common enzyme employed is trypsin, with a cleavage specificity at the C-terminal side of lysine and arginine residues. This is beneficial for the downstream MS analysis because both amino acids carry a positive C-terminal charge at acidic pH, but other peptidases like GluC or AspN are also used e.g. to increase the sequence coverage of proteins (**Figure 1**).

Most workflows for bottom-up proteomics share the extraction, homogenization and chemical modification steps. Samples are commonly boiled for a short time in a lysis buffer containing SDS, SDC or other detergents or denatured and unfolded in urea. After reduction with reagents like TCEP, DTT or ß-mercaptoethanol, the reduced disulfide bonds are quenched by acetylation or carbamidomethylation e.g. with chloro- or iodoacetamide. In the case of samples with cellular content, this is followed by a DNA

shearing step by sonication with ultrasound. For fibrous material this technique can also be applied before or between the previous steps, to disrupt the material and extract proteins. Recently, these procedures have been combined in a single pot reaction by making the reagents cross-compatible (Kulak et al., 2014). This decreases hands on time and makes the process more reproducible. Additionally, by avoiding the previously popular precipitation step (Jiang et al., 2004), or filter steps, this process can be readily automated on liquid handling robotic platforms (Geyer, Kulak, et al., 2016).

Following digestion, the peptides must be isolated from cellular debris, chemicals and buffer salts which can be done by desalting. Reversed phase material like C-18 or weak cation exchange functionalized surfaces can be applied. Other strategies are the filter aided sample preparation (FASP) technique in which all molecules smaller than a cutoff below the protein level are washed away before digestion, and the clean peptides are eluted from the filter afterwards (Wiśniewski, 2017). A recent development is the precipitation of proteins on magnetic particles and peptide elution by digestion from the beads after several washing steps (Batth et al., 2019; Hughes et al., 2019). All protocols yield sufficiently clean peptides and may somewhat prefer certain peptide species, as studies show that the resulting peptides detectable by LCMS only partly overlap (Sielaff et al., 2017). Additionally, the particle-based methods can be engineered to enrich for certain protein subsets (Blume et al., 2020). The arguments for a specific sample preparation procedure are often justified by up- and downstream compatibility to sample type and LC instrumentation and preference of the user.

While sample preparation for whole proteome analysis is comparably straightforward, strategies to enrich for certain protein or peptide classes add a new layer of complexity. The most commonly employed techniques are pulldown analyses for the study of protein interactions and phospho-peptide enrichment to study cellular signaling, but other enrichment processes, like acetylation- or glycosylation enrichment are possible as well. Phosphopeptide enrichment is done by positively charged metal ion raisins like Fe(III)-NTA (Andersson & Porath, 1986) or TiO2 (Pinkse et al., 2004). Being employed after protein digestion, the peptides with the negatively charged phosphorylations on serine, threonine or tyrosine residues bind to the raisin while other peptides are washed off. State of the art protocols for highest yields employ this technique instead of desalting and directly injecting the eluted peptides into the LCMS (Humphrey et al., 2018). Automated high throughput procedures e.g. on the Agilent *Assaymap* platform require relatively large amounts of purified peptides, but enable 96 well processing for the enrichment step with high yield (Russell & Murphy, 2016). Protein interaction studies like pulldown analyses are done with an enrichment step before the start of the sample

preparation. This can involve immunoprecipitation and genetically encoded tags like GFP linked to the protein of interest (POI). Interacting proteins bind to the POI which is 'pulled down' and other proteins are washed away under specific buffer conditions. Sample preparation as described above is applied and binders to the POI can be identified by comparison to control experiments with unspecific binders (Keilhauer et al., 2015; Wierer & Mann, 2016).

To illustrate how peptides are identified in the mass spectrometer the Edman degradation mentioned earlier can be used as a contrast. Peptides of interest in a bottom up proteomics experiment are commonly between 7 and 30 amino acids (AA) long and can be visualized as pearls on a string. The Edman degradation identifies the sequence by freeing and identifying one amino acid after another, beginning from the N-terminus. A mass spectrometer enables a similar procedure, by breaking multiple copies of a peptide ion semi-randomly between AAs. When lining up all fragments by size from both peptide ends in an ms/ms spectrum, the mass differences between fragments yield part of the peptide sequence (Ruedi Aebersold & Mann, 2003; Sinha & Mann, 2020). This clarifies why mass spectrometers are used for proteomics analysis: Unlike genomics and transcriptomics, where light emission readout methods can be used coupled to signal amplification with polymerase chain reaction (PCR) (Bustin et al., 2005), in proteomics the mass and fragment mass differences of a peptide are the primary readout (R. Aebersold & Goodlett, 2001). In theory, this enables de novo sequencing of peptides but in practice with possibly missing fragments in the spectrum and thousands of peptide sequences from a whole organism proteome to compare for a match, the need for automated statistically solid peptide spectrum match algorithms becomes apparent (Cox & Mann, 2008; Johnson & Biemann, 1989; Mann & Wilm, 1994). This is nowadays implemented in ready to use software packages whose input are MS raw files, sequence files for comparison and some parameter settings to yield peptide and protein identifications and quantifications with statistical significance.

To unfold its true power, the MS must be coupled to a continuous separation system. Most state-of-the-art systems employ liquid chromatography and in particular reversed phase liquid chromatography (Horváth et al., 1976), making use of the differential hydrophobic specificity of peptides assembled from different amino acids. In this manner, the tens of thousands of peptides from e.g. a cell culture tryptic digest are separated in two dimensions, a retention time domain, which follows peptide length and hydrophobicity and a m/z domain of the eluting molecules ionized as they elute from the end of the chromatographic column. For peptide identification the mass spectrometer switches between precursor scans (MS1) and fragmentation scans of selected ions

(DDA) or ion mass ranges (DIA) (MS2) for subsequent sequence-based identification of peptides from a database. With this method, proteomes of biological systems can be quantitatively measured with increasing depth in single shots. Despite the differing scan modes described below, the MS1 and MS2 scans are a shared feature of all MS-based proteomics experiments.

Improvements in instrumentation, scanning methods and analysis programs have pushed the limits of MS-based proteomics in the last decades (Hebert et al., 2018; Kelstrup et al., 2018; Makarov, 2000; Meier, Brunner, et al., 2018; Meier, Geyer, et al., 2018). Only in 2008, the first quasi complete proteome map of a complex organism, budding yeast, was achieved by extensive fractionation and instrument time effort (De Godoy et al., 2008), a task, which only six years later could be completed in an hour and by now within minutes with comparable completeness (Hebert et al., 2014; Nagaraj et al., 2012). In 2014, the first draft maps of the human proteome were presented – essentially compendia of protein identifications covering more than 70% of known human protein coding genes (M. S. Kim et al., 2014; Wilhelm et al., 2014). Nowadays, proteomes of mammals are routinely measured to a depth of 10,000 proteins. While whole proteome measurement is still one of its main applications and obtaining deep proteomes on a routine basis is still a challenging task, MS-based proteomics has become established in combination with several other experimental upfront methods. For example, proteomics can be used to build interaction networks with genetic tagged libraries in combination with pulldown screens (Hein et al., 2015). Proximity labeling analyses combined with pulldown of modified protein residues has the same aim (Roux et al., 2012). Enrichment strategies for post-translational modifications (PTMs) give insights into cell signaling, trafficking and protein turnover (Bard & Chia, 2016; Hansen et al., 2020; Robles et al., 2017; Tanzer et al., 2020). With spatial proteomics methods involving gradient centrifugation, the localization of proteins within the cellular organelles can be mapped on a global scale (Andersen et al., 2005; Itzhak et al., 2016; Krahmer et al., 2018). Proteomics can also be used for drug target identifications with tools such as limited proteolysis or drug affinity responsive target stability (DARTS) (Pai et al., 2015; Pepelnjak et al., 2020).

This brief summary represents just the tip of the iceberg of MS-based proteomics methods and applications. The following chapters will give a detailed introduction and focused view on the parts of the proteomics workflow that are of special importance in my PhD thesis.

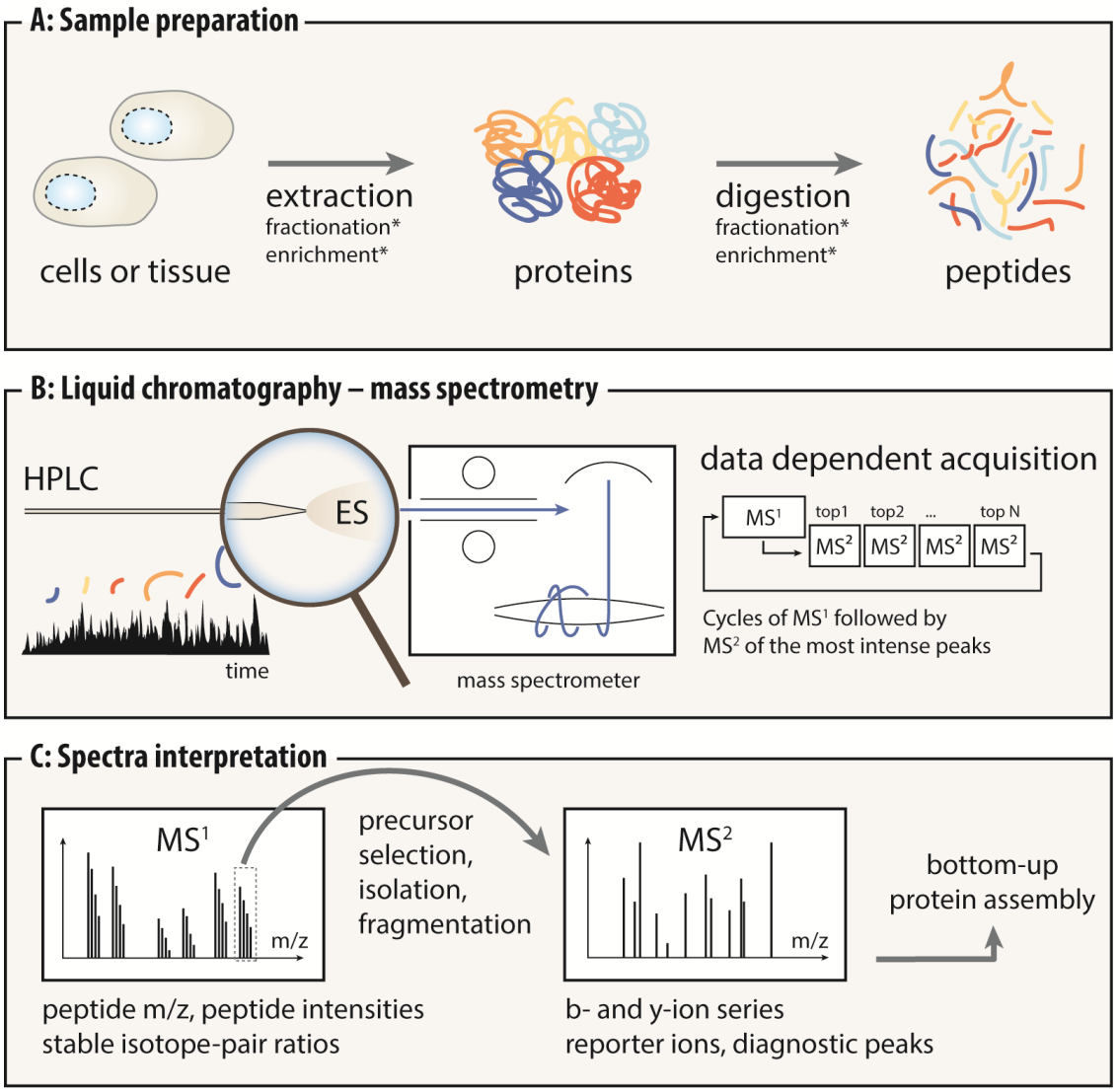


Figure 1: Shotgun proteomics workflow. a) In sample preparation for bottom-up proteomics, proteins are extracted from biological material and digested into peptides by enzymatic cleavage. **b)** The purified peptides are separated via HPLC in a time domain and brought into the mass spectrometer by ES ionization. The mass to charge ratio of the entering ions are detected by the MS and dependent on the scan mode of the MS, single or multiple ions are subjected to fragmentation and detection of the fragment ions. The figure exemplifies the scan mode cycle of a data dependent acquisition scheme, where N single MS^1 precursors are isolated and subjected to fragmentation and detection of MS^2 ions. **c)** The scanned spectra information is saved to a raw file and interpretation is done by specialized software which assembles or recognizes peptides from the MS^1 precursor mass and fragment ion information. Proteins information is assembled from the LCMS peptide information by software. Adapted from (Hein et al., 2013).

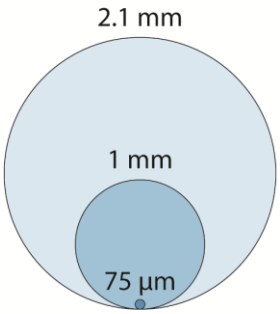
1.1. Liquid chromatographic techniques in mass spectrometry-based proteomics

Separation techniques like liquid chromatography upfront to the MS are not a precondition for MS-based proteomics. Bottom up proteomics matrix-assisted laser desorption/ionization (MALDI)-imaging has been around since 2001 (Stoeckli et al., 2001) and a recent publication shows that directly injecting ions into the MS via ES and simple gas-phase separation can analyze complex peptide mixtures (Meyer et al., 2020). These approaches, however, are attractive mainly for high throughput but shallow screening efforts because they lack the analysis depth as they give up the retention time dimension. The mass spectrometer is simply not sensitive or fast enough to deal with all incoming ions of a complex biological system, which results in an overall reduced dynamic range and only the most abundant peptides being sequenced and identified. By ‘stretching out’ the peptide mixtures in an additional domain orthogonal to the m/z domain – while concentrating them into narrow elution peaks, the ion mixtures entering the MS are reduced in complexity so that the MS has more time to sequence specific precursors selectively. Furthermore, the complexity of co-fragmented precursors is low enough to statistically assign them to a peptide of the proteome that is analyzed.

The time dimension of the chromatographic separation has changed over the last decade hand-in-hand with the increase in scanning speed of the MS-instrumentation. While four hour runs for single shots where standard not long ago (Kulak et al., 2014; Nagaraj et al., 2012), Geyer et al. showed that especially for samples in which only low complexity can be resolved by LCMS (e.g. blood plasma) 20-minute gradients can achieve similar proteome depth at increased throughput. Partly as a result, the community is shifting to shorter gradients and higher throughput (Geyer, Kulak, et al., 2016; Riley et al., 2016). This change in paradigm was adopted for several developments, e.g. the Evosep One HPLC or the idea to employ peptide prefractionation with short LCMS gradients for deep proteomes (Bekker-Jensen et al., 2017).

It has recently been demonstrated that very short LC-gradients also work for complex proteomes (Messner et al., 2019), but for deep proteome profiling the gradient is still commonly stretched up to several hours (Wang et al., 2019). LC-systems coupled to MS can vary in flow rates and therefore chromatographic column shape from milliliter (ml) to the nanoliter (nl) range, which is reflected in the naming of the corresponding technique – high-flow, microflow or nano-flow. Between them, the linear velocity of the mobile phase over the stationary phase is roughly the same, as the column diameter is adapted to the flow conditions to reach reasonable pressure conditions to run the chromatography.

a



b

	Ralser	Küster	nanoflow
d [mm]	2.1	1	0.075
area [mm²]	3.463606	0.785398	0.004418
length [mm]	50	150	500
bed volumne [mm³]	173.1803	117.8097	2.208932
flow [mm³/min]	800	50	0.3
linear velocity [mm/min]	230.9732	63.66198	67.90611

Figure 2: Column diameters for micro and nanoflow proteomics applications. a) Column diameters used in published MS-based proteomics setups compared in the scale of 20:1. **b)** Technical parameters for the columns exemplified in (A). The 2.1 mm diameter columns are used for 30 seconds gradients whereas the 1 mm columns are applied in 30-minute gradients. The column length is scaled down to enable a higher flowrate resulting in a somewhat higher linear velocity for the 2.1 mm column. The nanoflow column condition is simply a down scale of the 1 mm microflow conditions with nearly the same linear velocity of mobile phase over stationary phase. Adapted from (Messner et al., 2019) and (Bian et al., 2020).

Most cutting-edge systems for LCMS based proteomics employ nano-flow from 100 to 1000 nl/min and capillaries with 50 to 150 µm ID as columns, whereas 'industrial scale' and high throughput systems try to make use of microflow with mm sized columns for increased robustness and higher throughput. Column length in both cases is scaled to fit the pressure conditions of the system, with the maximum length as possible to achieve maximum performance (**Figure 2**).

1.1.1. Particle packed columns

As described above, apart from specialized developments (Bian et al., 2020; Messner et al., 2019), nanoflow conditions and µm sized capillaries are primarily employed for cutting-edge LCMS-based proteomics. Like columns for higher flow applications they can be purchased as particle packed tubes ready for use for similar prices, but in the case of ultra-high-pressure applications the lifetime of capillary columns tends to be only in the range of weeks. This is one of the reasons why laboratories with high throughput needs tend to prepare their own columns to save costs. Descriptions on how to prepare columns are publicly available (https://proteomicsresource.washington.edu/docs/protocols05/Packing_Capillary_Columns.pdf) and equipment to do so can be purchased from several vendors.

Packed capillary columns come in two different types, either as packed emitters, which directly function as ES-emitter or, more similar to regular commercial HPLC columns, as particle packed tubes with a porous frit and HPLC-ready connection at the end. Both systems have their drawbacks and benefits. While fritted columns usually have lower backpressures and therefore increased lifetimes, the porous frit and post-column dead volume of a downstream connected emitter typically lead to peak broadening compared to the chromatography to spray design in packed emitters (Gritti & Gilar, 2019), where

the bead bed ends with the opening of the electrospray (ES) tip (Emmett & Caprioli, 1994; Ishihama et al., 2002; Kennedy & Jorgenson, 1989).

With the aim of highest chromatographic performance, the columns in our laboratory in the past years were of the packed emitter type, which would be expensive to buy but which we manufacture with reasonable effort at low costs. While traditional approaches for the packing of capillary columns are time consuming and have low overall yield especially for columns with a bead bed of more than 30 cm in 75 µm ID capillaries, the procedure has been overhauled in the last years (Kovalchuk et al., 2019; Shishkova et al., 2018). I combined these published principles of high pressure packing and high-density slurry packing into a conceptionally novel approach for multiplexed packing and increased column production efficiency many-fold compared to previous standards, at undiminished chromatographic performance. With the new approach a single 50 cm with 75 µm ID column can be packed with sub 2 µm particles in under 2 minutes. Furthermore, with the construction of a packing station this process can be multiplexed to pack 10 columns in under 2 minutes.

1.1.2. Chip-based stationary phases

An alternative to packed capillaries for reversed phase nano-flow LCMS-systems, the µPAC column, was recently developed and commercialized by the company Pharmafluidics from Gent (De Beeck et al., 2018). They completely avoid the reproducibility issues between capillary columns with spherical particles by performing the chromatographic separation in a flow path etched into silica. This results in a micrometer-sized pillar structure, which is made porous and coated with C-18 molecules for reversed phase conditions (**Figure 3**).

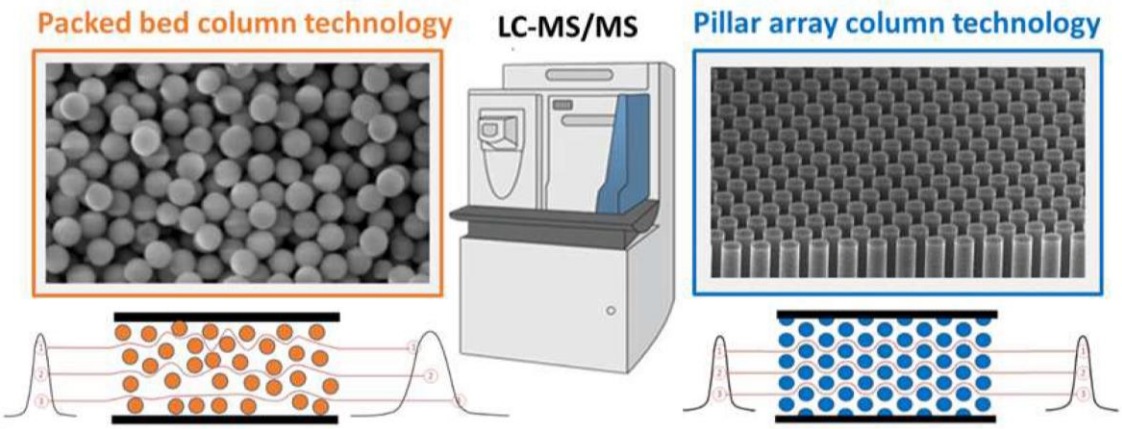


Figure 3: Spherical particles vs. etched pillar structure as stationary phase for HPLC. Electron microscopy pictures of packed spherical beads and the uniformly etched pillar array structure of the µPAC column are depicted. The peak

broadening from un-uniform flow paths in the packed bead bed vs. the regularly spaced pillar array is visualized. From (De Beeck et al., 2018).

Drawbacks of these columns are their higher costs and their reduced flexibility compared to in-house packed capillaries, which can be manufactured in all sizes and shapes. Furthermore, the first generation of commercially available columns has a relatively large dead-volume and a low pressure resistance (350 bar max), making it unsuitable for short LC gradients and fragile when not used with care. The chromatographic performance did not match the cutting-edge capillary columns with respect to effective use of LCMS acquisition time, ionization efficiency and peak capacity, but is appropriate for standard LCMS-based proteomics. However, the µPAC-column has a large benefit, which is the reproducibility of retention times of each peptide between runs and columns and laboratories. The chip structure is identical by virtue of the production process and therefore less variable as a packed bead bed. Additionally, the pillar structure is less affected by pressure changes as they occur between LCMS runs and therefore they have a longer life-time and less variability between runs (**Figure 4**).

I successfully employed the µPAC column in a large MS-based proteomics study and optimized a gradient for maximum sampling time and best ionization efficiency. This was done with a flow rate gradient together with a mobile phase gradient. With this combination, peptides are brought to an early elution from the column despite the high column volume while maintaining a high ionization efficiency due to a low flow rate (300 nl/min) during the peptide elution from the column. We demonstrate the superior performance compared to particle packed columns in terms of cross-run reproducibility and also the cross-laboratory reproducibility by comparison of peptide retention times from measurements in our laboratories in Munich and Copenhagen (**Figure 5**a).

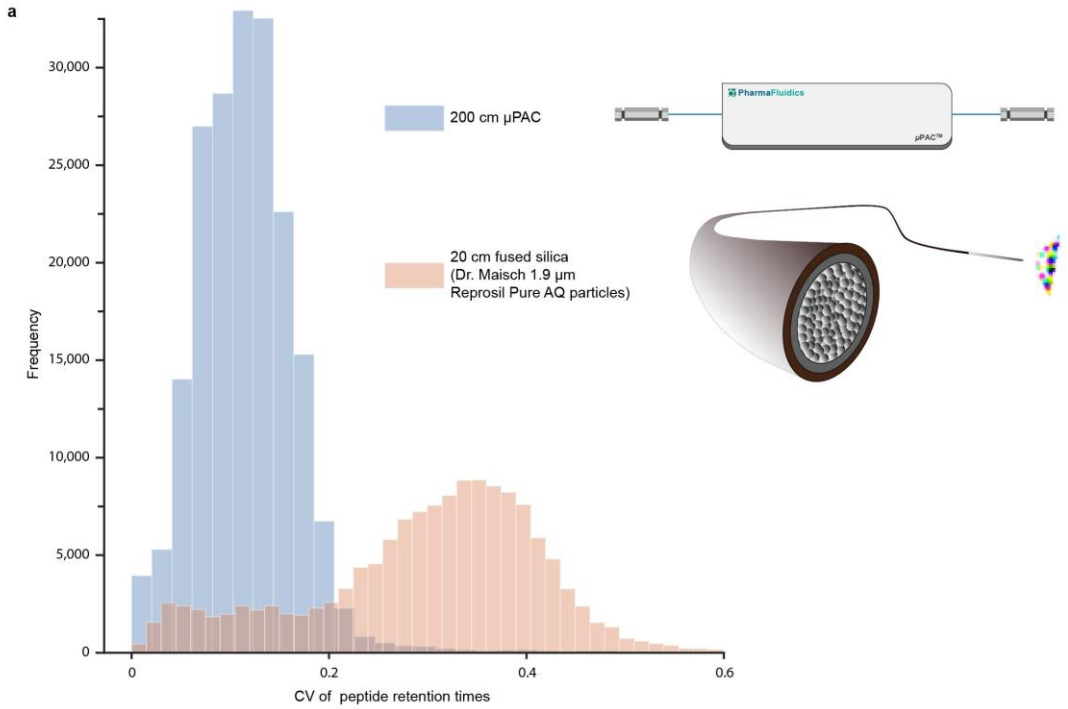


Figure 4: Peptide retention time reproducibility of capillary and µPAC column. The coefficient of variation (CV) of peptide retention times is showed as a histogram for eight measurements of HeLa peptides with a capillary column packed with 1.9 µm particles (Reprosil-Pur AQ, Dr. Maisch) and a 200 cm µPAC column. CVs are several-fold better for the µPAC column.

Furthermore, we demonstrated the highly accurate chromatography performance by application of machine learning to predict the peptide retention times of previously not analyzed peptides and successfully retrieving those in a global targeting experiment (**Figure 5**b).

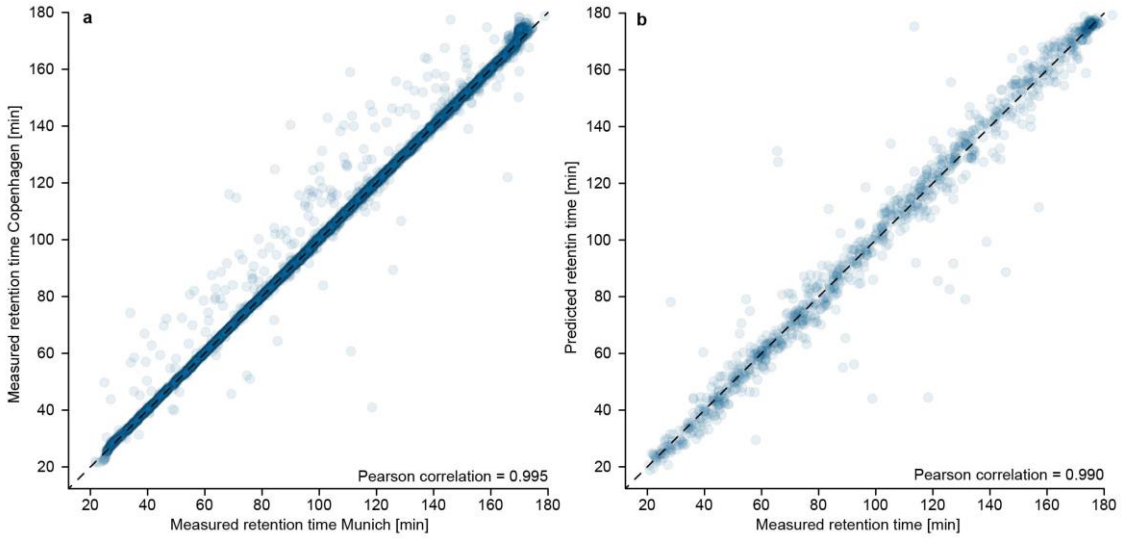


Figure 5: Peptide retention time reproducibility and prediction of peptide retention times from their amino acid sequences. a) HeLa peptides from different digests were measured on a µPAC column in our laboratories in Munich and Copenhagen. The retention times of the overlapping identified peptides have a high similarity score with a Pearson correlation of 0.995. **b)** Peptides from a holdout set not employed for training of a bidirectional LSTM model (see below) for peptide retention time prediction from peptide sequence, are displayed with their predicted retention time from the trained model and the experimentally determined retention time. The Pearson correlation is 0.990.

1.1.3. Benefits and drawbacks of nano- and microflow liquid chromatography

Nano- and microflow chromatography both have their benefits and drawbacks for the use with ES-MS. Microflow is robust due to lower pressure requirements and this is also reflected in longer lifetimes of columns. On the downside, it is less sensitive because of larger column diameters, which results in higher material loading requirements. The higher flow also implies a larger dilution factor and subsequently lower concentration of the sample at the point of ES, which also lowers the sensitivity. When comparing different flow and column conditions this must be adjusted for by correcting the flow rate and loading amount for linear velocity (meaning the column ID).

Nano-flow LCMS has its drawbacks on the robustness and throughput side: Higher flow rates require high pressures (up to >1000 bar) under which a solvent gradient must be formed, resulting in material problems like leaks and fast column decay (Richards et al., 2015; Shishkova et al., 2016). When running on the lower end of nano-flow range (e.g. sub 100 nl/min), gradients tend to become unstable and suffer in reproducibility between runs. With the low flow rates allowed by capillary columns, overhead times for LCMS analyses, namely column equilibration and sample loading onto the column, tend to be long and therefore short gradient times quickly result in idle times of the MS up to 50%. Solutions to this problem like the use of two columns with a single chromatographic system are technically challenging and have rarely been employed routinely (Hosp et al., 2015). However, the great advantage of nano-flow is increased sensitivity allowing low sample input amounts which are especially important for PTM enriched samples, resulting from a high analyte to solvent molecule ratio at the ES and the smaller column IDs.

As a result of the previous considerations and requirements for chromatography for LCMS setups, employing the preferred nano-flow setup for higher throughput analyses is challenging. The ideal system would combine the throughput and reproducibility of a micro-flow system with the sensitivity of a nano-flow system.

In cooperation with the company Evosep from Odense (Denmark), we established a new chromatographic system for LCMS application, the Evosep One. This LC is especially suited to high throughput analyses because it reduces overhead time to a minimum (several minutes) and therefore enables schedules of up to 300 samples per day without sacrificing much MS time to overhead. This is done by several innovations: A disposable precolumn system is employed, where peptides are supplied, pre-loaded on C-18 material and eluted into a sample loop by a low-pressure system consisting of four pumps. To avoid forming a gradient under pressure, this is done by the low-pressure

system when eluting the sample from the C-18 and diluting it with aqueous buffer before storing the peptides lined up in the sample loop in a preformed gradient. The aqueous dilution is necessary to ensure that the peptide peaks are stopped and re-sharpened at the begin of the analytical column. After gradient formation is completed, the sample-loop is switched in-line with the high-pressure pump where the preformed gradient is release to the analytical column and the proper LCMS analysis starts (**Figure 6**).

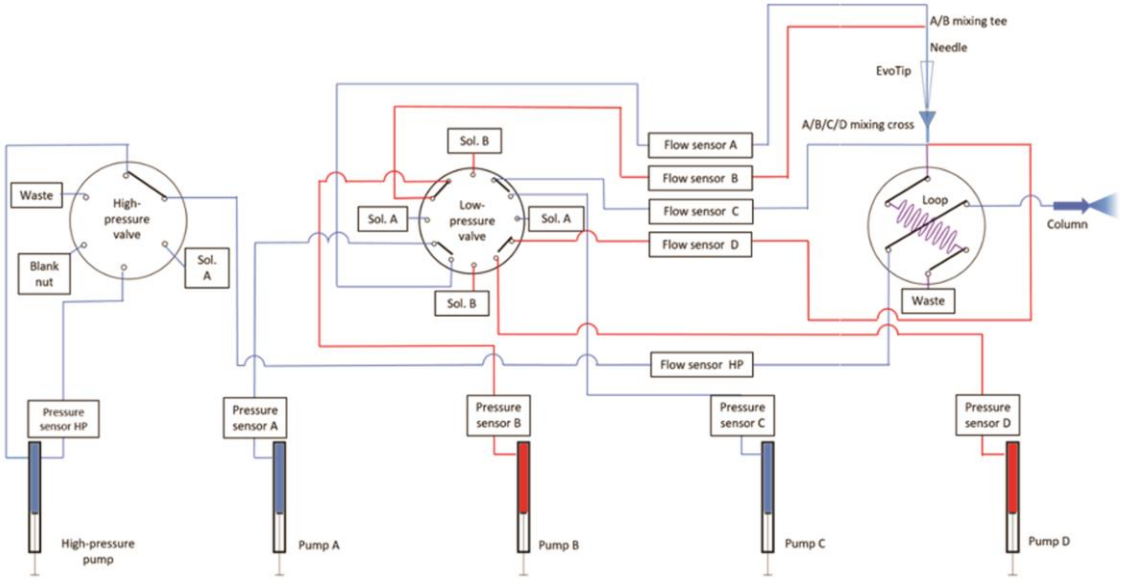


Figure 6: Flow-chart of the Evosep One HPLC from the paper 'A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics'.

1.2. Mass spectrometry-based proteomics

Mass spectrometers were invented in 1918 and have since been employed in various scientific and industrial use cases (Dempster, 1918). In general, mass spectrometers are devices for the detection of ion mass to charge ratios. Their broad application for biological sciences was initiated by the invention of the soft ionization techniques MALDI and ES (Fenn et al., 1989; Karas & Hillenkamp, 1988). Especially ES, whose inventor John Fenn was awarded the Nobel Prize in chemistry in 2002, is of interest for proteomics analysis because of its effective coupling to liquid flow techniques (Whitehouse et al., 1985).

To enter the electrodynamic flow path of the MS instrument, molecules must be ionized and introduced into the instrument's vacuum system. In ES, liquid eluting from a etched or pulled capillary or needle forms a Taylor cone that breaks up into charged droplets at the end of the ES emitter (Taylor, 1964; Zeleny, 1914). This is accomplished by an electric field in the kV range between the emitter and the MS entrance. By the use of electrostatic force and electric fields, ions can be focused or stored in near vacuum

conditions. This is the fundamental physical basis for ion guides, ion traps and mass analyzers which are combined within modern hybrid mass spectrometers to enable state of the art proteome measurements.

In simple terms, a hybrid mass spectrometer for proteomics needs a number of coupled devices typically operated in the following order: A mass filter (normally a quadrupole) to filter ions for precursor selection, a collision cell for optional fragmentation of precursors, and a mass analyzer for the detection of the analyte's mass to charge ratio (**Figure 7**a) (e.g. a quadrupole, a time-of-flight detector or an Orbitrap analyzer)(Ruedi Aebersold & Mann, 2003). All additional parts are implemented for further improvements, to enhance the number of ions that reach the analyzer (higher sensitivity), increase the resolution of mass spectra, the accuracy and speed of the analysis, expand the dynamic range of the MS and find more efficient ways to make use of the existing ions via new scan modes.

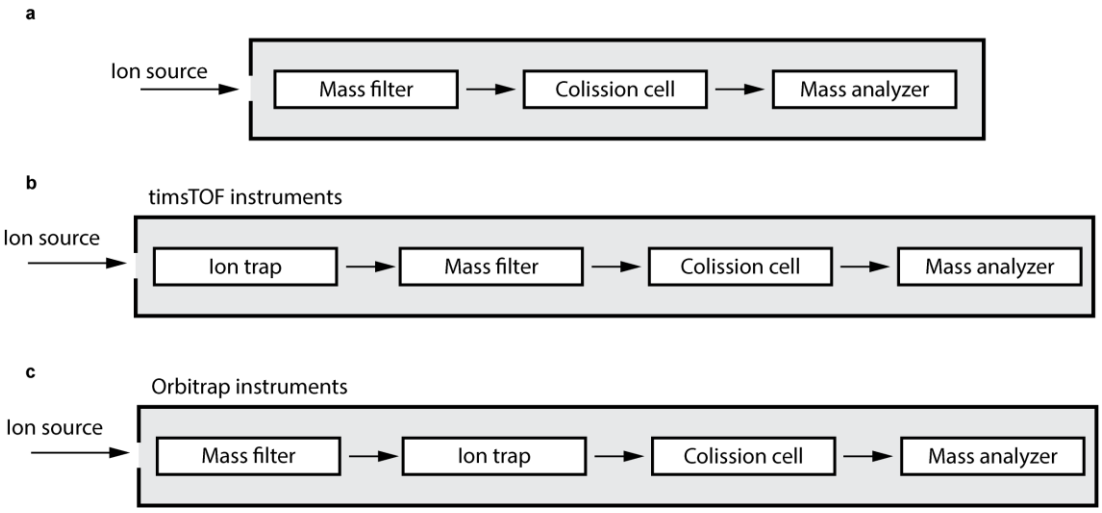


Figure 7: Basic components of modern mass spectrometers for proteomics analyses. a) To be used in bottom up proteomics experiments, instruments must have a mass filter, a collision cell or other ion fragmentation device and a mass analyzer. By the introduction of ion traps, high performance MS-based proteomics is facilitated because lower abundant ion species can be accumulated to increase sensitivity and signal to noise ratio. This is the case e.g. in timsTOF instruments, where the ion trap is located in front of the mass filter and is also used as a device for ion mobility separation (b) or in Orbitrap instruments, where the ion trap is located downstream of the mass filter, so that specific ion species can be accumulated (c).

The quadrupole is a device used in virtually all hybrid mass spectrometers. It consists of four metal rods arranged in parallel with the opposite rods being connected electrically (**Figure 8**a). The optimal quadrupole shape would be hyperbolical towards the center, but most devices have circular rods, which is a well working approximation. When operated in radio frequency (RF) only mode, the quadrupole functions as an ion transmitter with a low m/z cutoff by employing a radio frequency between the rod pairs. This can also be used for ion trapping, by adding electric field lenses at both ends, which creates a

potential well along the quadrupole rods. When applying a direct current (DC) field additionally to the RF on the opposing rods, the quadrupole functions as a mass filter, because only specific m/z ranges have a stable oscillating path through the device whereas the other ions collide with the rods. The mass filter can be scanned across the whole m/z range and narrowed to windows below 1 Thompson (Th), at the cost of low transmission efficiency at narrow isolation windows. Together with an ion detector, the quadrupole can be used as a mass analyzer by going through the m/z range with a narrow isolation window while detecting the transmitted ions. These features make the quadrupole one of the most flexible and widely applicable devices in mass spectrometry (Dawson, 1986; Douglas, 2009; W. Paul & Raether, 1955; Wolfgang Paul & Steinwedel, 1953). This is exemplified by triple quadrupole devices, where three quadrupoles are used in a linear arrangements as mass filter, collision cell and mass analyzer (Yost & Enke, 1978).

In line with the quadrupole, hexa- and octopoles are commonly used in hybrid mass spectrometers. While lacking the excellent mass filter properties of the quadrupole, the pseudopotential well formed by the RF field in the devices offers advantages. While the quadrupolar device forms a narrow potential well, this is broader for the hexa- and octopolar device. This favors use for ion trapping and transmission, with the hexa- and octopoles having more volume for ions to fill and the quadrupole being better suited to focus ion beams (Dawson, 1986; Douglas, 2009).

The ion funnel is another device that is commonly used in hybrid MS instruments for ion beam focusing. The ion funnel operates at higher pressures (mbar range) and is a special case of the stacked ring ion guide, which consists of stacked conductor ring plates with an RF-potential applied to every second of the plates (Guan & Marshall, 1996a). By narrowing down the diameter of the rings towards an end plate and applying a DC gradient over the plates, ions are focused into an ion beam and accelerated towards the end plate (**Figure 8**b). These devices are often placed at instrument entrances and after ion trapping regions. In those regions, ion beam focusing is needed for downstream parts of an instrument such as a quadrupole mass filter, where ions not focused to the center of the quadrupole would be instantly lost (T. Kim et al., 2000; Shaffer et al., 1998).

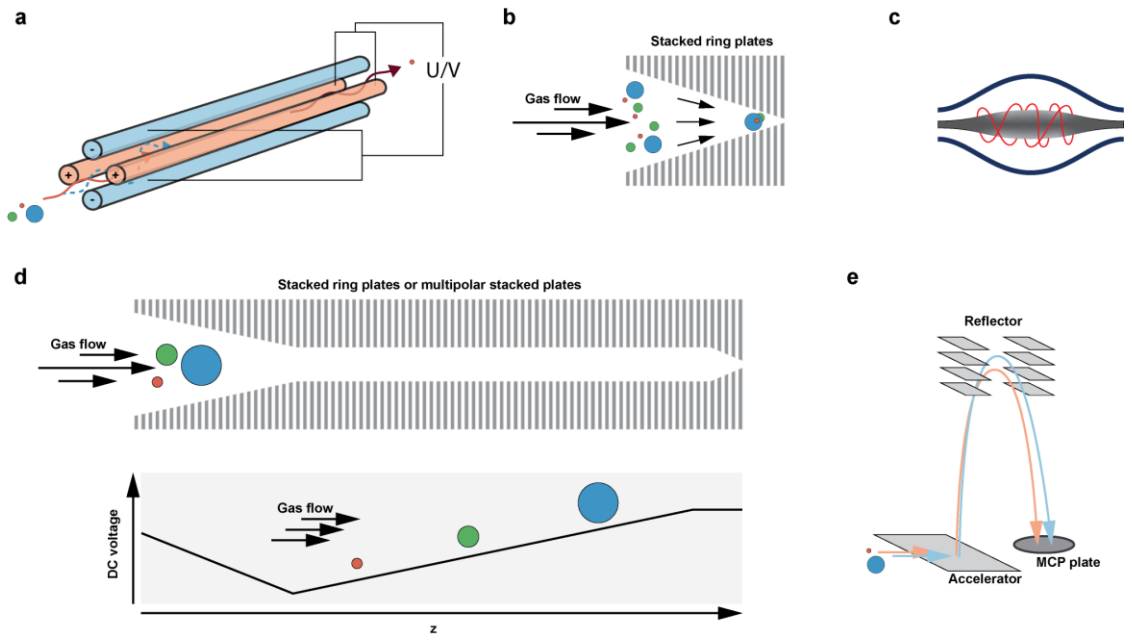


Figure 8: Schematic view of functional mass spectrometer parts. a) A quadrupole consists of four parallel rods, where the opposing rods are connected to the same RF and DC. By altering their parameters, ions can be transmitted or specific ions can be isolated. b) Ion funnels are mainly used for ion focusing, e.g. after transfer tubes, and consist of stacked ring plates with a RF frequency and DC gradient over the field applied. The RF facilitates a pseudopotential well, which pushes the ions to the funnel center and the DC gradient accelerates the ions towards a narrow exit electrode lens. c) The orbitrap revolutionized the MS-based proteomics field in the early 2000s because of its high mass resolution. It consists of a center and ring electrodes where the entering ions are stabilized on an oscillating orbit around the center electrode and the oscillation frequency can be measured and transformed to m/z information. The transient time in the analyzer defines the mass resolution. d) The tims device is capable of trapping ions and separating them by their mobility at the same time. This happens by a gas flow through the tims tunnel and a DC gradient decreasing (for positive ions) in the opposite direction. Ions with large collisional cross-section at similar charge compared to smaller ions are more affected by the gas flow and find an equilibrium at a position further in the tunnel. By raising the lower end of the tims ramp the ions can be released in their mobility order. e) The TOF detector consists of two important parts: The accelerator pushes the analyte ions towards detector plate, commonly an MCP plate. The reflector enhances the resolution and at the same time decreases the instrument dimensions. Adapted from Gross, 2017.

Like quadrupoles, stacked ring ion guides can be used for ion trapping. The dipolar RF device creates a pseudopotential well towards the center of the ring channel in which ions can be held. By adding lenses with higher DC potential than the static potential of the ring plates at the entrance and exit, a potential well is formed in which ions can be accumulated. This setup can be used to separate ions by mobility which measures the ratio of ion charge to collisional cross section of the ion. By employing a steady gas flow through the stacked ring ion guide and adding a DC gradient over the plates, ions are lined up by mobility in the device (**Figure 8**d). This basic principle was discovered by Mel Park and is used to introduce the ion mobility dimension in trapped ion mobility spectrometry time of flight (timsTOF) instruments by Bruker (Meier et al., 2015; Silveira et al., 2014).

Other concepts of ion mobility separation are also used in commercial instruments. These encompass drift tubes (Tyndall et al., 1926), traveling wave separation (Giles et

al., 2004) or high field asymmetric waveform ion mobility spectrometry (FAIMS) (Kolakowski & Mester, 2007; Swearingen & Moritz, 2012).

Ion fragmentation at specific chemical bonds is a prerequisite for hybrid mass spectrometers and is typically done in a collision cell by imparting energy in different ways. The most popular are collision induced fragmentation (CID) (Levsen & Schwarz, 1976; McLafferty et al., 1973) and higher-energy C-trap collisional dissociation (HCD) (Olsen et al., 2007). These methods primarily yield b and y ion series by breaking peptide bonds. Other fragmentation methods are especially helpful for the fragmentation of certain PTM by dissociation of specific chemical bonds or for yielding different ion series from peptides. Examples are electron transfer dissociation (ETD) (Syka et al., 2004) and ultra-violet photodissociation (UVPD) (Zubarev et al., 1998).

Common mass analyzers in hybrid instruments for MS-based proteomics are the Orbitrap (**Figure 8**c) (Hu et al., 2005) and TOF instruments (**Figure 8**e) (Wolff & Stephens, 1953).

With proteomics applications being adopted in areas from biological to clinical research, a common question is which developments will shape the future of this technology. One aim is to minimize input material to enable detection of proteomes from low amounts of biological material like single cells. This is generally done by scaling down sample preparation workflows and developing LC and MS instrumentation for maximum sensitivity. This topic towards single cell proteomics has got much attention in the last years (Ctortecka & Mechtler, 2021; Marx, 2019).

Another possibility to increase the performance of MS-based proteomics is the expansion of the detectable dynamic range. This is for example done manually and off-line by prefractionation approaches to specifically cut out certain high abundant species like K48-peptides in diGly enriched samples (Hansen et al., 2020) or in a non-specific way, in a LC step orthogonal to the low pH reversed phase chromatography online with the MS instrument (Bekker-Jensen et al., 2017; Kulak et al., 2017). The latter approach results in several fractions which have to be measured separately by LCMS. It obtains its power by giving the MS more time for peptide sequencing at a certain retention time, enabling the loading of more input material and simplifying the interpretation of MS2 fragment spectra for data independent identification approaches from co-fragmentation spectra.

The above fractionation approaches expand the dynamic range of the proteomic experiment by enabling detection of low abundant peptide species otherwise covered up by higher abundant, coeluting peptides in a single shot experiment. Untargeted MS-

based proteomics - unlike affinity-based approaches - is a technique which identifies proteins in the dynamic range roughly in order of abundance, with a statistically higher chance of detection and identification for higher abundant species. Therefore, improving the dynamic range directly enhances the proteomic depth in experiments. The endless race for the deepest proteomes (e.g. in a simple cell culture digest) is currently dominated by new software approaches coupled to intelligent scan modes of the MS. In the end however, the physical limits of MS-based proteomics can only be overcome by new developments in instrumentation.

A shared principle of different high-end mass spectrometers is the use of a collection device for ions to enhance the signals that reach the analyzer. The dynamic range of an MS is mainly determined by the capacity of this ion trap, like the tims device for timsTOF instruments (**Figure 7**b) or the C-trap for Orbitrap instruments (**Figure 7**c). When filled with ions of different abundances at the same time, there is a ratio threshold of total ions to an individual ion species at which it cannot be separated from noise, which specifies the lower detection threshold. By enhancing the total amount of ions in the trap this ratio threshold is shifted and the detection limit is lowered, resulting in higher dynamic range.

In cooperation with Bruker Daltonics, I worked on prototype parts for the storage of large ion populations. Devices for ion storage are usually multipolar systems. Examples are quadrupolar traps like the C-trap in Thermo Fisher instruments or the formerly used PC-board cartridges for ion trapping and simultaneous ion mobility separation in the tims device in Bruker instruments, which consisted of a quadrupolar and an octopolar region. By generating an oscillating field between the rods of the device in RF-only mode, a pseudopotential well is created in the center between the rods which can hold ions (Miller' & Denton, n.d.; Wolfgang Paul, 1990). Stacked ring ion guide dipolar devices are a relatively new development that matches the capacity of quadrupolar ion traps (Guan & Marshall, 1996b). Briefly, these are built quite similar to ion funnels, which are stacked conductor ring plates with a dynamic electric potential to focus and accelerate ion beams (Kelly et al., 2010). To be able to trap ions, a frequency is applied between every second of the stacked electrode plates without having a DC gradient over the rings, which again creates a pseudopotential well in the middle of the rings.

I compared a traditional quadrupole and different forms of stacked ring ion guide type multipolar devices to characterize physical parameters of these different parts. Of interest were the charge capacity, the ion decay times, the exit times under isobaric conditions and the behavior of those parameters under different pressure conditions, meaning the gas flow through the system when ions were accumulated. My experiments show that a novel device has the ability to store more than 100 million ions, can hold

these ions almost without decay and is able to eject the ions within milliseconds of time by applying an axial DC field. With this development, it would be possible to rethink the current construction of timsTOF devices with their parallel accumulation serial fragmentation (PASEF) scan mode in a single tims device (Meier, Brunner, et al., 2018). By changing the localization of ion accumulation to a place prior to the current tims device, and therefore enabling the use of the full length tims for mobility separation, the capacity of the entire system can be scaled up. This approach is under development and currently called orthogonal PASEF due to the orthogonal placement of the ion accumulation region to the tims device in current constructions.

These developments will most likely impact all kinds of proteomics experiments with the timsTOF instrument in the future. It may become possible to acquire deep proteomes in short measurement times and without prior fractionation. The impact on body fluid proteomics might even be higher: For the last decade, only small improvements have been achieved on the way to deeper proteomes for plasma and therefore efforts have been concentrated on throughput and reproducibility. With the dynamic range of these sample types being the main cause for this issue, MS instrumentation specifically tailored to approach this problem might yield not only a small increase, but rather an order of magnitude improvement, something which is unprecedented in the field of plasma proteomics.

1.3. Scan modes for MS-based proteomics

New scan modes for MS-based proteomics have been developed hand-in-hand with the hardware instrumentation within the last decade. The possibility to do certain modes of MS1 and MS2 or even MS3 detection for advanced precursor identification arises from the architecture of instruments and is therefore mostly specific for each one of those. Two principles form the basis for recent developments: Data dependent acquisition (DDA) scan modes (Link et al., 1999; Venable et al., 2004) and targeted scan modes such as selected ion monitoring, multiple ion monitoring or parallel reaction monitoring (SIM, SRM, PRM) (Price, 1991; Yost & Enke, 1978). DDA starts with MS1 scans, which are scans of all ions at a timepoint in the chromatographic gradient in an m/z range. These scans detect intact peptides which are called precursors. Two strategies can be applied to fragment precursors for detection of b and y ion series (the main backbone fragments from the N- and C-terminus, respectively) and subsequent peptide identification: In DDA, after each MS1 scan, the top N most abundant peptides are chosen for isolation and fragmentation in a serial manner. Exclusion lists or dynamic exclusions prevent picking a precursor again before a specified number of seconds.

Typically, only precursors with a charge of at least two are considered for fragmentation, because they have a higher likelihood of being peptides and of yielding high quality fragment spectra. In SIM, m/z windows are chosen for isolation and fragmentation of a chosen precursor with known retention time and fragmentation behavior. Gradients are normally short, given that only a few peptides are monitored. Specificity and quantification can suffer as SIM is usually performed on low resolution instruments like triple quadrupoles. By design targeted proteomics misses out on the discovery properties of the DDA mode. The SIM concept can be extended to multiple reaction monitoring (Kondrat et al., 1978) or parallel reaction monitoring (PRM) (Rauniyar, 2015), where multiple ions are selected for isolation but still fragmented sequentially over the whole elution time as in SIM. These methods are frequently applied to detect low abundant small molecules but are also for specific peptide identification. Where the targeted methods suffer in unbiased detection for discovery of unknown features, the DDA mode has its drawbacks in data completeness. The top N sampling method is semi-stochastic and in repeated runs different precursors get selected for fragmentation and identification, resulting in data with a larger proportion of missing data.

There are multiple approaches to combine these techniques and therefore generate complete data which also allows the discovery of peptides in an unbiased way. The data independent acquisition (DIA) strategy (Doerr, 2014), is most common, where MS2 scans are done with broad isolation windows, stepping through the whole range of precursor m/z. When scanning in this fashion, no precursor stays unfragmented but the MS2 spectra consist of the ion series from multiple peptides which makes it more challenging to identify peptides with statistical significance. Only recently has the community set guidelines for the stringent and confident identification of peptides from DIA experiments (Chalkley et al., 2019) and software solutions are beginning to enable the universal application of DIA scan modes (Bernhardt et al., 2014; Demichev et al., 2020). Recent developments have shown that this scan mode can be applied to many different types of experiments and also additional dimensions, like ion mobility coupled to TOF scans for eluting ions to make the identifications more specific (Meier et al., 2020).

Another approach to overcome the above-mentioned limitations of targeted MS is a ‘global targeting’ strategy which takes the PRM concept to the next level. By measuring peptide libraries of a sample or generating those with deep learning for retention time and fragment ion parameters, a multi-dimensional scan map of peptides in the RT, m/z and MS2 directions can be generated. By targeting a large number of those precursors in narrow RT and m/z windows, the data completeness of whole proteome

measurements can be kept high. Discovery experiments are also possible by focusing, for instance, on thousands of peptides that changed between experimental conditions (Wichmann et al., 2019).

In sample types with large dynamic range e.g. blood-plasma, a key limitation for DDA methods is the dynamic range of MS1 scans. When a single ion type fills the ion trap in a short time, lower abundant peptides cannot be separated from the noise and are not picked for MS2 identification. DIA scan modes partially address this by collecting ions from those low abundant regions for fragmentation in an unbiased manner. Alternatively, this can also be addressed by additional MS1 scans, avoiding the high abundant precursors. This has been implemented in the BoxCar scan mode, where whole MS1 scans are complemented with MS1 scans that only collect ions from boxed windows over the m/z range. In this way, the MS1 filling time can be allocated to highlight precursors from formerly underrepresented areas (Meier, Geyer, et al., 2018).

An alternative way to solve the missing value problem from semi-stochastic sampling in DDA is speed. Arguably, if the instrument is fast enough to sequence all or nearly all multiple charged MS1 features in every run, the data should be complete in this regard, at least for all precursors that are still detectable in MS1 scans. This principle can be implemented on instruments with certain hardware, that features rapid detection of ion m/z like TOF analyzers. The orbitrap mass analyzer which revolutionized MS-based proteomics with its high-resolution mass spectra in the early 2000s has a speed drawback, because high resolution can only be obtained from ion transient times in the tens of milliseconds range. The TOF analyzer scans the entire mass range in every single pulse, which only lasts about 100 µs, and is therefore suited extremely well for rapid detection. This is made used of in the PASEF scan mode, where precursors are eluted from a mobility device, further isolated in a narrow m/z window and fragmented. Because of the high TOF scanning speed, the limiting factor is the quadrupole switching time and the timed release of ions from the mobility device (Meier et al., 2015; Meier, Brunner, et al., 2018).

In summary, the scan mode to choose for an experiment depends on several factors, the aim of the study, the sample type and the MS instrumentation at hand. One might argue that DIA is now the most powerful scan mode, delivering proteome coverage as well as depth. Still, for a long time and even now, the community was suspicious of the quality of peptide spectrum matches from multiplexed fragmentation which is always the case in DIA. By now, this is demonstrated to work in well-defined search spaces like common model organisms and especially human samples. This is one of the reasons

we still used DDA scan modes for the multi organism study described just below, which brought proteomics to previously never investigated organisms.

1.4. MS-based proteomics as a tool for multi-organism studies

The analysis of biological samples from different organisms is a topic of interest not only for evolutionary biologists but also for newer scientific disciplines like evolutionary medicine or life science in general. Traditionally, organisms' taxonomic relations have been determined by phylogenic features and only the revolutionary technique of genome sequencing has brought comparisons and classification to a molecular level. On the protein level, human proteins have been the scientific focus, mainly because of their medical relevance. Apart from this, proteins have often been selected for study because of their accessibility and in certain areas for their unique functions. The first crystals of proteins for example have been described for various organisms' hemoglobin (Giegé, 2013). Comparing different evolutionary solutions to a process performed by proteins can greatly contribute to the understanding of proteins as molecular machineries (Sikosek & Chan, 2014).

However, studies comparing the proteome of organisms rather than the genome or transcriptome had rarely been done. This may be because high performance MS-based proteomics, the only technique capable of this task, is still young and investigations of model organisms have been the main focus of the field. If cross organism studies are done, comparisons are mainly between a limited number of model organisms which can be cultured and genetically modified in scientific investigations. This is because most traditional biochemical analyses rely on affinity-based methods like antibodies for protein identification and interaction studies and these antibodies are predominantly available for a few model organisms. An alternative to affinity-based methods is genetic tagging to selectively trace or purify a certain protein in a single organism. Traditionally this was limited only to organisms easily accessible for genetic engineering, but the discovery of the CRISPR/Cas gene editing makes this possible for a wider spectrum of organisms (Jinek et al., 2012). Still, these methods are time consuming and expensive, especially if a large number of proteins are the aim. MS-based proteomics does not suffer from those drawbacks. Due to its purely sequence based identification of proteins, it can be directly applied to any organism for de novo protein analyses. However, de novo sequencing is challenging and to give statistically solid evidence of proteins, a draft proteome from genome analyses is needed for reliable bioinformatic identification of

proteins. With these precondition, MS-based proteomics can be applied to any biological sample, wherefrom proteins can be extracted.

In our publication, 'The proteome landscape of the kingdoms of life' we quantitatively measure the proteomes of 100 organisms across the tree of life. We apply a standardized workflow to 19 archaea, 49 bacteria and 32 eukaryote species and achieve good sequence coverage for the archaea and bacteria in single shot measurements as well as excellent coverage for the eukaryotes after high-pH prefractionation into 8 samples each. Using publicly available sequence databases (UniProt) ("UniProt: A Worldwide Hub of Protein Knowledge," 2019) we identify and quantify thousands of proteins which were predicted from genetic material but for which there had never been any experimental proof of existence. We identified more than 340,000 protein groups from more than 2 million unique peptide sequences, from 1.1 million protein entries in the UniProt database. From the ~560,000 entries with verified existence in the SwissProt part of Uniprot, we identified close to 80,000 and additionally give experimental evidence to more than 1 million entries of the TrEMBL part of the database which mainly consists of predicted protein sequences from genome sequencing. After subtracting sequences which were found in previous studies in a literature research from the PRIDE database, we more than double the number of proteins with experimental evidence with the identification of more than 800,000 entries.

Proteome comparison is normally done on the protein and peptide level, but this is not directly possible in the case of cross organism comparisons. In single organism experiments, like cell culture treatments or clinical cohorts, every protein quantified in different conditions or cohort arms and split by a categorical parameter can be compared statistically. As there are no overlapping proteins between organisms this is not applicable here.

Looking at the proteomes as a whole it is remarkable to note that all proteomes follow a power law distribution in abundance (after examining this in more detail in our data, it turns out to be closer to a beta distribution, which is among the family of exponential distributions). This implies that the biology of all species in this cross section is similar in terms of proteome organization. Without exception, a hand full of extremely abundant proteins dominate the total protein mass. The dynamic range of our measurements is somewhat limited in the single-run experiments, whereas the pre-fractionation used on eukaryotic samples resulted in deeper proteomes and better isoform coverage.

To comprehensively compare the proteomes of different organisms, one must make use of links between the proteomes' proteins to group them into comparable parameters at

a cross organism level. This is possible on an evolutionary basis by homology information available from databases like EggNOG (*evolutionary genealogy of genes: Non-supervised Orthologous Groups*) (Huerta-Cepas et al., 2019). This information links proteins as gene products by their biological evolutionary ancestry on different taxonomical levels. For example, proteins from different *ophistokonts* (including animal and fungi kingdom) can be defined as a homology cluster specific to the taxonomical level for a common ancestry. This can reveal structure or functionality and therefore is a good way to see which specific proteins have importance in which organisms of taxonomical sub clusters. However, there are also clusters including the taxonomical root level, which tie together the entire known biological life. One of the most abundant homological clustered proteins turned out to be the chaperon GroEL/HSPD1/Hsp60 which seems to have a fundamental role from distant bacteria species to higher eukaryotes. Our data makes the expression levels of such cross species related and relevant proteins directly comparable and enables screening for proteins with important evolutionary conserved functions across different taxonomic levels.

More database resources are required to make the cross-species data interpretable beyond the pure protein amino acid sequence information. Sequence predicted structural features (secondary and domain structures) from the PFAM database are bridging this gap (El-Gebali et al., 2019). Specific protein subdomains tie together distinct proteins which are evolutionary conserved for important functional reasons. The most abundant of those domains captured in our dataset are multiple *Elongation factor Tu* domains, but also the Hsp70 as a whole protein and GAPDH C-terminal domain are amongst the universally expressed and catalogued protein substructures. A common paradigm in structural biology is that ‘structure defines function’. Therefore, for a large number of proteins that only have predictive sequence information from genome sequencing there can still be a suggestion for molecular functions if they are involved in biological processes annotated in the Gene Ontology Resource (“The Gene Ontology Resource: 20 Years and Still GOing Strong,” 2019). We focused on the biological processes and report the organism that do not have extensive information here. This correlates well with the organisms which have the worst proteome coverage in our experiments, a fact that we trace back to poor genome annotation. The overall most abundant biological processes across the tree of life are oxidation-reduction-processes, translation and protein folding. From more than 13,500 biological process annotations in the dataset, only a minority has proteins with a function in all organisms measured here (**Figure 9**). Some functions are expectedly exclusive to certain taxonomic organism

subsets, like histone related processes to eukaryotes or photo complex reactions to photosynthesis employing species. With translation, protein folding and proteolysis being among the most abundant biological processes executed across the tree of life, our data supports the maintenance of the proteome itself as one of the most important features of biological life as we know it. Especially the high abundance and therefore importance of proteins that help and keep other proteins folded correctly is noteworthy.

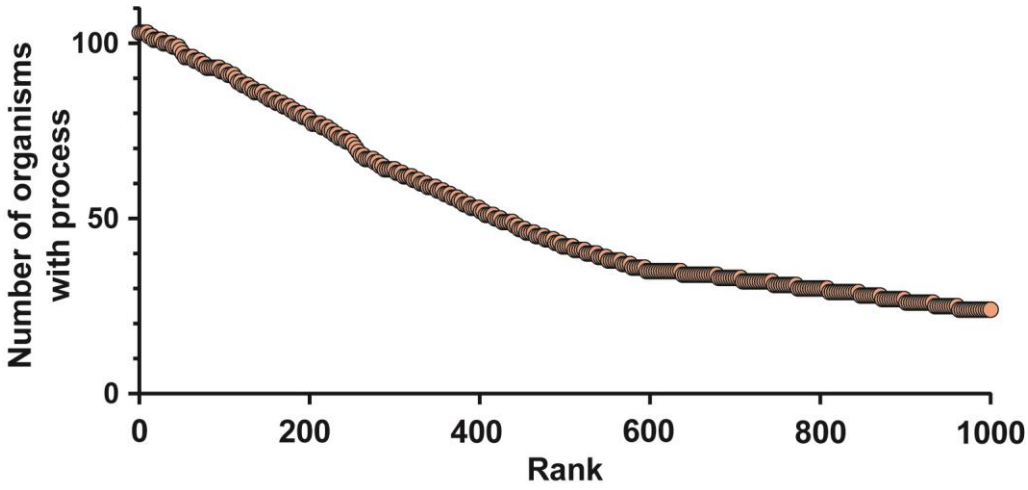


Figure 9: Data completeness of the biological process information for proteins from our 100-organism dataset. The number of organisms for which a specific biological process is annotated for at least one quantified protein (y-axis) is sorted by the rank (x-axis)

With the proteomes of 100 organisms and the added layers of annotation data, we have provided a complex and unprecedented dataset that we hope to be useful for years to come. Simple take home messages include the large number of highly abundant proteins with little to no database information pointing to any known function, which should be of interest to biologists and biotechnologists. These proteins clearly play important roles within the specific organisms because of the high expression levels we report, but have mainly only been described as potential proteins from genome sequence with no expression information at all. With a virtually endless number of those proteins being present in the known living organisms, our methodology clearly points out a manageable number of candidates that may be worth looking at in detail.

More generally, we deliver qualitative and quantitative information for many proteins which have previously only been predicted by sequencing data. Such information might be valuable for proteins with homologs in different organisms. Researchers working on uncommon proteins can look up those or the respective homologs in our dataset and find organisms where closely related proteins are expressed in higher levels, suggesting important biological functions, or find organisms which are evolutionary close but lack the specific protein or protein class in our dataset, to study how these get along without the protein.

These are just examples for the applicability of our dataset as a unique resource for future research. To have all information readily pooled and available we decided to store all organism, protein, peptide and annotation information in a graph database. The different layers of information are established as nodes with every organism, protein, peptide or GO annotation term being a single node of the super class. These nodes are then connected by relationships e.g. proteins found in an organism, peptides identified of a protein or GO terms annotated to a protein. This builds up an immense data structure with all the information and analysis illustrated above in a single database with 8 million datapoints and 53 million relationships between them, all ready to be queried.

The network graph database is publicly available in our program, which uses the neo4j graph platform that enables the extraction of information with the industry standard programming language Cypher. For easy exploration of the dataset we exemplified some functional analyses in interactive applications on our webpage proteomesoflife.com.

1.5. Organism case study - bear proteomics

The above-described advantages of MS-based proteomics for the analysis of various organisms are still very rarely made use of. During the time of my PhD I worked with several organisms for standalone studies of which two should be mentioned here.

In an approach to analyze the functions of a ribosomal associated protein I measured the total proteome of *Arabidopsis* wild-type and POI depleted samples. *Arabidopsis* is a widely used model organism which we had included in our multiorganism study as well and the proteome has previously been studied extensively (Mergner et al., 2020). Our data helped to identify a new regulator of protein turnover which acts co-translationally by detection of N-degrons and enhancing protein lifetime by N-terminal acetylation.

On a broader scope, a scientific field that could directly benefit from the universal applicability of MS-based proteomics is evolutionary medicine. Among others, this scientific approach tries to gain biological understanding through the observation of evolutionary processes. In more detail, if an organism has adapted to a specific niche and therefore solved a biological problem on a molecular level, it might be possible to first understand the underlying biological principles on a molecular mechanistical level and second make use of this biological knowledge to target the same mechanisms for medical treatments.

This is exemplified by our so far unpublished study of active and hibernating brown bears. Unlike other animals such as rodents, hibernating brown bears have a relatively

high body temperature (~30°C) and properly hibernate – without frequent active phases (Hellgren, 1998). There is medical interest in their unique body conditions for multiple reasons, one being the absence of venous thrombosis and their subsequent effects. This suggests that coagulation is blocked or coagulation stimuli decreased in hibernating bears and these changes and the reasons for them should be detectable on a functional protein level (Welinder et al., 2016). With venous thrombosis being a frequent medical condition, this would be valuable information for identification of new medical targets and understanding of the disease.

Affinity based methods are not available for bear samples because there are no commercial antibodies against bear proteins specifically, and production of a range of antibodies from immunization techniques e.g. in rabbits would exceed the resources of most single scientific studies. This makes MS-based proteomics the only method for the comparative investigation of these samples and it also holds the promise of new discoveries because of its untargeted way of detecting proteins.

In our comparative analyses of plasma and thrombocytes from active and hibernating brown bears we find several metabolic marker proteins and suggest a range of described and previously undescribed proteins to play a role in suppression of venous thrombosis. Overall, we observe drastic changes in both examined specimen.

1.6. Clinical proteomics

The case of proteomics in evolutionary medical studies in hibernating brown bears described above already indicates the enormous potential of MS-based proteomics for medical research. The bear example deals with extreme body conditions (month long fasting and rest time) which results in the drastic changes manifested in plasma and thrombocyte proteins. When conducting studies with human individuals, effect sizes tend to be lower. Compared to cell culture experiments with gene knock out conditions or strong perturbations and even model organisms with genetically homogeneous individuals, studies with patients prove to be more challenging. Nearly any patient cohort comes with a high ‘genetic noise’, resulting in much higher naturally occurring variation in protein levels compared to the above described ‘laboratory conditions’. Additionally, recruiting patients for studies is always difficult and often happens over longer time periods, which can entail changes in sample taking and processing. This increases the variation on samples especially compared to a single researcher handling organisms or samples in a controlled environment and short timeframe.

This may seem like a disadvantage for discovery-based proteomics studies, but I argue that one can look at it another way. In fact, any routine clinical laboratory faces the same difficulties but is not able to adjust for it. By genetic and biological variation, parameters used as biomarkers can and do vary between patients. Today those parameters are normally measured in single readout technique assays like ELISA, or enzymatic activity, which are commonly automated in current commercial laboratories. When a measurement at a single time point is taken from a patient, the assessment of the measured value relative to the population-based reference can be misleading and a time course of measurements to find the individual specific thresholds would be more appropriate. This would come with regular invasive blood taking procedure which is elaborate and uncomfortable for most patients, and is therefore rarely done. Proteomics offers a far more universal readout than a single parameter and the proteomic signature can be used to tailor the reference values to the individual patient. Additionally, proteomics only requires a minimal amount of sample (1µl of plasma is sufficient for state-of-the-art techniques) and the sample taking can therefore be done by minimal invasive procedures like finger pricks.

In general, clinical decision making is often based on laboratory testing, and quantitative analyses of molecules in body fluids constitute the large majority. In current industrial practice these tests are mainly enzymatic or affinity based and performed as single analyses on automatic test handling platforms (Roche COBAS etc.). MS-based methods are currently mainly employed for the quantification of small molecules in the clinical laboratory or in discovery phase projects to find new biomarkers. Proteomics techniques can be useful here for multiple reasons. One of the main ones is that MS-based proteomics experiments are designed to detect proteins in an untargeted manner and are therefore suited to be employed in discovery phase experiments.

This has been done especially for plasma in various studies of our laboratory in the last years (Geyer, Wewer Albrechtsen, et al., 2016; Niu et al., 2019; Wewer Albrechtsen et al., 2018), but can also be applied to other body fluids like urine (Virreira Winter et al., 2020), saliva (Grassl et al., 2016), stool (Zhang et al., 2018) or tears (Nättinen et al., 2019). Tissue proteomics is also widely used in model organisms (Geiger et al., 2013), patient samples for tissue atlases (Aizarani et al., 2019; Angelidis et al., 2019; Doll et al., 2017; Dyring-Andersen et al., 2020) or characterization of tumor expression patterns (Doll et al., 2018). In the case of tissues, the tissue slides commonly used in pathological applications like staining and immunohistological chemistry are especially interesting. These tissue slides can be assessed by pathology experts under the microscope and specific areas cut out macro- or microscopically at the cellular level for downstream

proteomics analysis. This field is still young and new revolutionary developments in specific automated analysis in a cell type resolved manner are to be expected in the future.

Proteomics offers multiple advantages compared to other, targeted approaches of quantitative protein detection methods. Quality control of study parameters can be done simultaneously because of the untargeted nature of many MS-based proteomics workflows from intrinsic readouts. We exemplified the use of quality markers in plasma in our publication 'Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies'. The implications from this study lead in two directions: When used as a fast and high throughput tool, proteomics can be applied in addition to the other measurements to complete the dataset in an unbiased manner while also uncovering any quality issues within the cohort. This helps to better understand the problems and results of newly developed analytic tools and to evaluate if study results from a single readout technique can be trusted. Additionally, proteomics is a valuable tool in the classification of patients into expression-based population clusters and therefore can be employed as a tool for stratifying cohorts for outcome prediction from single readout biomarker tests. These use cases also apply intrinsically to proteomics studies in general.

In a cooperative effort with a research group aiming to use Fourier-transform infrared spectroscopy of liquid biopsies for diagnostic purposes, we exemplify the use of MS-based proteomics for quantitative sample assessment in the manuscript 'Molecular origin of blood-based infrared fingerprints'. The aim of our collaborators was to detect disease related changes or patterns within the infrared molecular fingerprints of serum samples and a cohort of lung cancer patients and controls were recruited. For this groups, specific signatures could be detected employing their method and with the aim to identify the cause for the changes in the fingerprints we set out to do MS-based proteomics analysis of the samples. Serum contains diverse molecular groups, but the most likely ones effecting the infrared spectra were the serum proteins. However, the actual cause for changes in the fingerprint spectra detained with infrared laser was not clear at all, and multiple theories from concentration changes, small molecules interacting with parts of the proteins or changes in secondary structure of proteins were discussed. With the MS-based proteomics information about high and similar sample quality, biases from study arm differences could be excluded and the quantitative information from our measurements helped to reassemble fingerprint spectra from molecular origin. Additionally, we found the acute phase proteins (SAA1, SAA2 and CRP) to be changed between cases and controls which in combination with the changed

albumin concentrations in the lung cancer part of the cohort explain the changes in fingerprint spectra.

Looking toward future personalized precision medicine, the progress of MS-based proteomics will not be in clinical studies alone. Population wide studies are necessary to build the foundations for a future application of proteomics as a routine medical analysis tool. For instance, we envision frequent Plasma Proteome Profiling (PPP) of patients (Geyer, Kulak, et al., 2016). With a knowledge base from millions of patients and samples as a background, this could revolutionize personalized medicine by the application of untargeted, non-invasive tests for the improvement of public health and helping to change lifestyle choices.

1.7. Clinical study design for proteomics

Clinical studies have to be designed carefully for MS-based proteomics biomarker discovery. A range of parameters has to be considered before any sample is collected. Cohorts must be well defined and one must ensure that no bias is introduced at the cohort level. This includes case and control collection at different sites or locations, by different personnel, with alternating material or on different times of the day, which all can introduce statistical noise if variation it is not evenly distributed across cases and control cohorts. Effect sizes must be considered when planning a cohort. The basic underlying principle is that the cohort size must be adapted to detect a statistically significant difference in the measured parameters at a certain technical and biological variance. This requires preliminary data collection before a study can be planned. The technical variance of the whole workflow from sample collection to data acquisition should be known for the desired sample type and it can be determined by technical workflow replicates quite easily. The accuracy of protein quantification by MS-based proteomics is protein dependent, and – for a given work-flow – can entail coefficients of variation of more than 100% for some. The second step is to determine the biological variance in an example population cohort. With this information and an estimate of the effect size, meaning the average expected change of a certain protein between cohort and control group, it is possible to tailor a study to the biological and technical needs.

Most other considerations are based on infrastructure at the study site. Proteomics analyses greatly benefit from additional information like anthropometric data, clinical laboratory results or patient questionnaires. To collect such information in a systematic way is not necessarily part of the daily schedule of a clinician, so new workflows, infrastructure and often personnel must be involved.

These challenges are not unique to proteomics studies, but certain features have more impact on proteomics than on ELISA based studies. In proteomics, quality issues derived from sample collection bias are detectable based on the proteomics background. Cohort differences can be discovered by metadata correlation to known markers in the analyzed sample type, which helps to prevent misinterpretation of the outcomes.

In our manuscript 'Cohort Profile:MUNICH-PretCl study Preterm and Term Clinical Study (MUNICH-PreTCl)', we describe a cohort of term and preterm born infants, from which dried-blood-spot (DBS) and plasma samples have been collected for a proteomic study. The cohort was collected in the Frauenklinik of the Ludwig Maximilian University Munich (LMU) over multiple phases including the report of clinical metadata and questionnaires. This information is laid out in the manuscript and builds a basis for the interpretation of the subsequent proteomics study. With correlation analyses between the meta- and the proteomic data, additional insights can be gained and overinterpretation avoided.

The importance of clinical metadata and study design is also worth noting for our publication 'Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease'. Here, we employ our previously proposed 'rectangular study' principle for biomarker discovery (Geyer et al., 2017) (**Figure 10: The triangular and rectangular study approach for biomarker discovery. (A)** In the triangular study approach, a small subset of patients is screened in depth for potential biomarkers. In follow-up studies the number of participants is enlarged whereas the number of screened parameters decreased. **(B)** The rectangular study approach is characterized by two (or more) large study cohorts. The overlapping candidates from both study arms are high confidence potential biomarkers, whereas the study specific hits can be ignored. From (Geyer et al., 2017). In the 'triangular study design', that had been the standard before, a small number of samples is analyzed in depth for biomarkers and only a few candidates are validated in follow up studies with larger cohorts. This yielded few or no new biomarkers because promising targets from the discovery phase first stage of the study nearly always turned out to be cohort specific. Furthermore, protein patterns are more predictive than a few individual proteins. With the rectangular approach this is less likely to happen, as two or more large cohorts are screened in a discovery phase and the overlapping significant hits are subjected to follow up work.

Our three different cohorts of Alzheimer´s disease patients and corresponding control samples separate to different degrees between cases and controls on the basis of the

CSF proteome. By making use of the clinical data we are able to show that despite the reduced power in separation of case and control samples in one of the cohorts (presumably due to less stringent classification), the biological context and resulting stratification of CSF proteins is identical to the two additional cohorts.

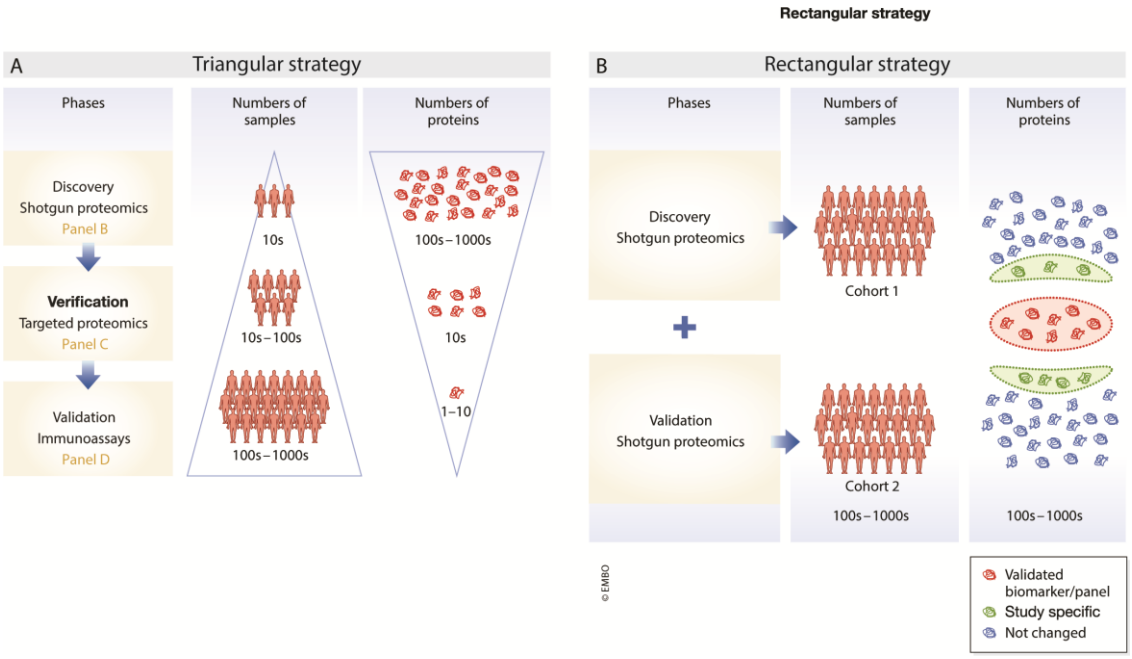


Figure 10: The triangular and rectangular study approach for biomarker discovery. (A) In the triangular study approach, a small subset of patients is screened in depth for potential biomarkers. In follow-up studies the number of participants is enlarged whereas the number of screened parameters decreased. **(B)** The rectangular study approach is characterized by two (or more) large study cohorts. The overlapping candidates from both study arms are high confidence potential biomarkers, whereas the study specific hits can be ignored. From (Geyer et al., 2017).

2. Aims of the thesis

In my PhD thesis I had two main goals: The development of tools for use in MS-based proteomics and the general application of the method to biological and clinical studies. These two aims were combined in some projects, as can be seen in the publication ‘The proteome landscape of the kingdoms of life’, where the investigation of a new, highly reproducible column technology led to the measurement of a peptide library for the prediction of physical peptide probabilities. This rather technical project was transformed entirely, when it became clear that the underlying dataset proved to be highly valuable in itself when extended to a wide range of species. It contributes to the scientific community by giving proof of existence for thousands of previously merely predicted proteins and enabling quantitative comparisons between organisms on a functional level. Along the same lines, my aim to reproduce published technical data and increase chromatographic performance by the introduction of high-pressure packing of capillary columns resulted in the development of a new concept for multiplexed column production. This will save hours of hands-on time during the preparation of in-house packed capillary columns for use with state of the art HPLC-MS applications. These cutting edge HPLC methods might be supplemented in the future by streamlined systems like the Evosep One described in the publication ‘A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics’.

The advantages of higher throughput and reproducibility will be important especially for clinical applications where hundreds and even thousands of samples have to be measured with constant quality. I worked on two examples for such studies during my PhD: The publication ‘Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer’s disease’ is a paradigm of body fluid proteomics for the discovery of new biomarkers and the ‘Cohort Profile: MUNICH Preterm and Term Clinical Study (MUNICH-PreTCl)’ illustrates the planning needed to recruit a cohort for a successful study. During my PhD I learned that successful clinical proteomics studies mainly depend on accurate planning of the cohort and problems on the cohort side are more serious than problems on the MS-based proteomics side. This is made clear in the publication ‘Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies’. Here possible sample related biases in plasma proteomics studies are dissected and their origins investigated, especially with regards to sample taking and processing.

3. Publications

3.1. Article 1: The proteome landscape of the kingdoms of life

Authors: Johannes B. Müller[1,7], Philipp E. Geyer[1,2,7], Ana R. Colaço[3], Peter V. Treit[1], Maximilian T. Strauss[1,2], Mario Oroshi[1], Sophia Doll[1,2], Sebastian Virreira Winter[1,2], Jakob M. Bader[1], Niklas Köhler[4], Fabian Theis[4,5], Alberto Santos[3,6] & Matthias Mann[1,3]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]OmicEra Diagnostics GmbH, Planegg, Germany. [3]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany. [5]Technical University of Munich, Department of Mathematics, Garching, Germany. [6]Li-Ka Shing Big Data Institute, University of Oxford, Oxford, UK. [7]These authors contributed equally: Johannes B. Müller, Philipp E. Geyer.

In our paper ‘The proteome landscape of the kingdoms of life’, we apply MS-based proteomics as a tool for cross organism studies. By quantitatively measuring the proteomes of 100 organisms (29 archaea, 49 bacteria and 32 eukaryotes) across the tree of life, we present an unprecedented dataset which enables the study of differences between organisms’ functional machineries.

The set of organisms was chosen with three criteria in mind: The genome of all organisms had to be sequenced and, ideally, protein sequences for predicted genes should be available in online databases. Organisms had to be available from a collection or similar source and the price per organism had to be sufficiently low. We employed a universally applicable proteomics workflow from sample preparation to data analysis for all 100 organisms.

We made use of a chip-based column for HPLC separation of peptides prior to MS and demonstrated their very high reproducibility. With high-quality data for a set of more than two million unique peptides in hand, we employed machine learning to predict the retention time of peptides. The bidirectional Long Short Term Memory (LSTM) neural network model that our cooperation partners developed is able to predict peptide retention times from peptide sequence alone in a highly accurate manner. We devised an experiment to test the accuracy of the predictions as follows. By programming the MS instrument to only sequence peptides if they elute in a narrow window at the predicted retention times of peptides in previously unexamined organisms, we show that the predicted RTs are accurate enough to identify almost the same number of proteins as in an untargeted experiment.

Comparing the organisms' proteomes is only possible after linking proteins between organisms from annotation databases. We make use of biological resources such as GO, PFAM and EggNOG annotations for protein functions, protein domains and protein homologies predicted from sequence to enable the comparison of organisms based on our proteomics data. On a functional level we demonstrate that the most common biological processes are equally important in all organisms and that the most abundant protein domains and homology clusters are distributed similarly.

This study is only a beginning of the topic of cross-organism proteomics studies. With 100 organisms widely spread across the tree of life we provide a deep but sparse view. In the future, tissue or cell type resolved studies of multicellular organisms and perturbations for single cellular organisms would provide a more detailed view. We hope that our publication encourages more laboratories to explore this field. Such initiatives will make use of the potential of MS-based proteomics to overcome the limitation of studying only model organisms.

Article

The proteome landscape of the kingdoms of life

https://doi.org/10.1038/s41586-020-2402-x

Received: 2 August 2019

Accepted: 27 April 2020

Published online: 17 June 2020

Check for updates

Johannes B. Müller[1,7], Philipp E. Geyer[1,2,7], Ana R. Colaço[3], Peter V. Treit[1], Maximilian T. Strauss[1,2], Mario Oroshi[1], Sophia Doll[1,2], Sebastian Virreira Winter[1,2], Jakob M. Bader[1], Niklas Köhler[4], Fabian Theis[4,5], Alberto Santos[3,6] & Matthias Mann[1,3]✉

Proteins carry out the vast majority of functions in all biological domains, but for technological reasons their large-scale investigation has lagged behind the study of genomes. Since the first essentially complete eukaryotic proteome was reported[1], advances in mass-spectrometry-based proteomics[2] have enabled increasingly comprehensive identification and quantification of the human proteome[3–6]. However, there have been few comparisons across species[7,8], in stark contrast with genomics initiatives[9]. Here we use an advanced proteomics workflow—in which the peptide separation step is performed by a microstructured and extremely reproducible chromatographic system—for the in-depth study of 100 taxonomically diverse organisms. With two million peptide and 340,000 stringent protein identifications obtained in a standardized manner, we double the number of proteins with solid experimental evidence known to the scientific community. The data also provide a large-scale case study for sequence-based machine learning, as we demonstrate by experimentally confirming the predicted properties of peptides from *Bacteroides uniformis*. Our results offer a comparative view of the functional organization of organisms across the entire evolutionary range. A remarkably high fraction of the total proteome mass in all kingdoms is dedicated to protein homeostasis and folding, highlighting the biological challenge of maintaining protein structure in all branches of life. Likewise, a universally high fraction is involved in supplying energy resources, although these pathways range from photosynthesis through iron sulfur metabolism to carbohydrate metabolism. Generally, however, proteins and proteomes are remarkably diverse between organisms, and they can readily be explored and functionally compared at www.proteomesoflife.org.

To collect a diverse set of representative organisms across the tree of life, we considered the availability of assembled genome sequences and the accessibility of cultured or tissue material, and included common model organisms for comparison. This resulted in 19 archaea, 49 bacteria and 32 eukaryotes—a total of 100 different species (Fig. 1a, b). We also added 14 viruses (Supplementary Table 1).

To obtain the proteomes of these extremely different biomaterials, we tested a number of extraction protocols and found that the in-StageTip (iST) protocol[10] was most universally applicable and allowed automated and highly reproducible sample preparation. We incorporated the latest advances into our workflow for high-resolution bottom-up proteomics, and implemented a recently developed chip-based method[11] (Fig. 1c–e). C_{18}-covered beads are replaced by a uniformly ordered and statically fixed micrometre-sized pillar structure[12] (Fig. 1d), leading to 2.5-fold improvements in coefficients of variation for peptide retention times and high interlaboratory reproducibility (Extended Data Figs. 1, 2a). For all prokaryotes we performed single-run mass spectrometry (MS) analyses, whereas we used a loss-less prefractionator[13] for the more complex eukaryotic samples.

We reasoned that our chip-based chromatographic method, combined with the very large data set of more than two million unique peptides, should be well suited to deep learning algorithms, which have recently been shown to be applicable to MS-based proteomics[14–16] (Extended Data Fig. 3). We developed a long short-term memory (LSTM) deep learning model with an interpretable attention layer to precisely predict chromatographic retention times, achieving a Pearson correlation of 0.990 (Extended Data Figs. 2b, 4). To test the model on a completely unknown proteome, we instructed the mass spectrometer to sequence peptides from *B. uniformis*, *Bacillus megaterium* or *Enterobacter aerogenes* only if they eluted in a narrow band around the retention times predicted by deep learning. This resulted in only slightly diminished proteome depths (at least 88% on the protein level), showing that these peptide properties were successfully modelled in silico (Fig. 2).

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]OmicEra Diagnostics GmbH, Planegg, Germany. [3]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany. [5]Technical University of Munich, Department of Mathematics, Garching, Germany. [6]Li-Ka Shing Big Data Institute, University of Oxford, Oxford, UK. [7]These authors contributed equally: Johannes B. Müller, Philipp E. Geyer. ✉e-mail: mmann@biochem.mpg.de

Article

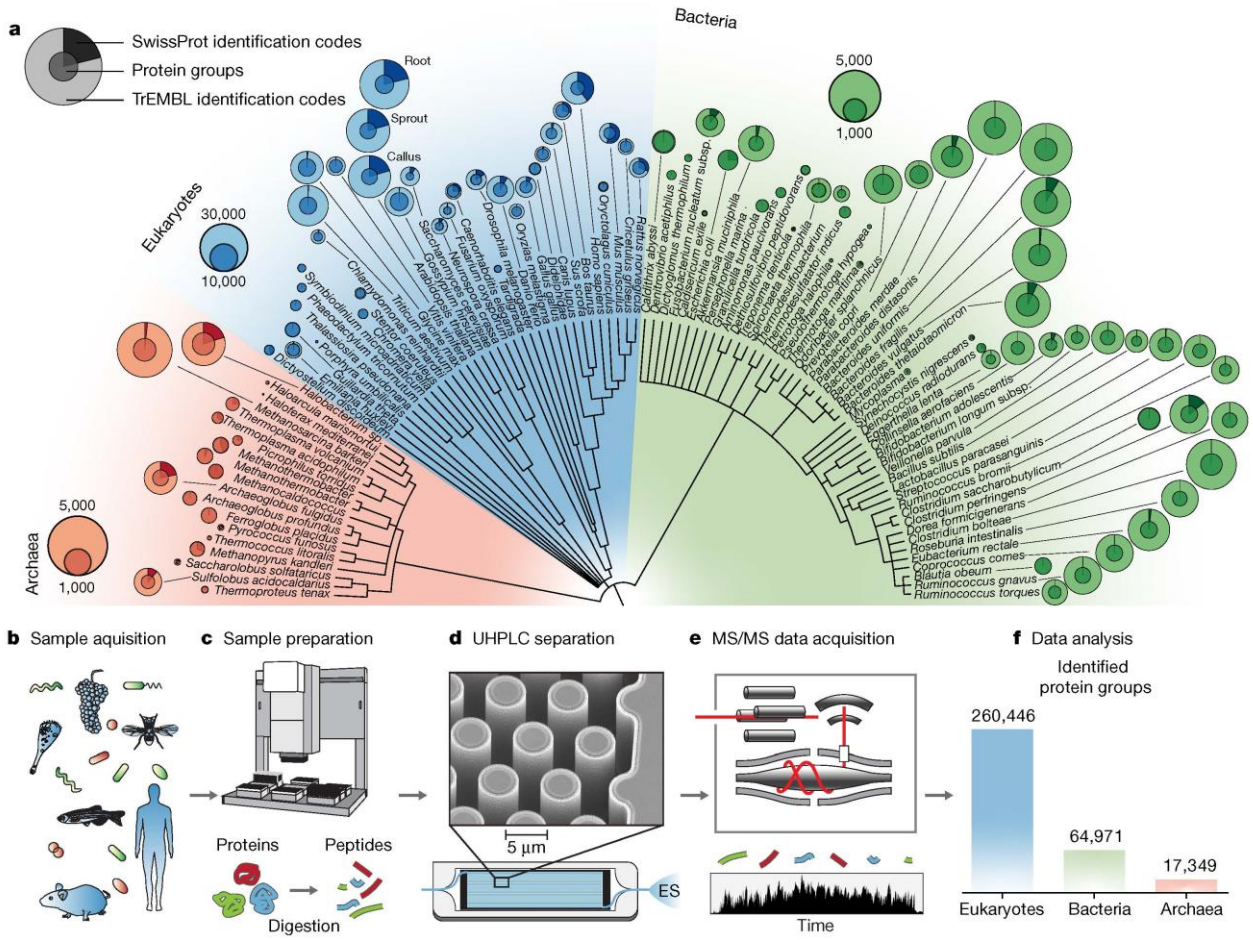


Fig. 1 | Collection of organism samples across the tree of life, and integration of the proteomic workflow. a, All organisms used herein were ordered and ranked on the basis of National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) taxonomy. Pie charts refer to the numbers of protein groups (proteins distinguishable by their identified peptides) and to database protein entries found here. **b**, **c**, The acquired samples were subjected to protein extraction and digestion into peptides for sample preparation. **d**, Peptides were separated using a silica-chip-based micropillar array column (µPAC) with etched pillar structures that are coated with C_{18}. UHPLC, ultra-high performance liquid chromatography. The magnification shows a scanning electron microscopy image of the pillar structures (adapted with permission from PharmaFluidics). **e**, Peptides were ionized by electrospray (ES) and analysed in a high-resolution mass spectrometer. **f**, Numbers of identified proteins across the three superkingdoms.

Across the 100 organisms, we identified 349,164 proteins that were distinguishable by their identified peptides (Supplementary Table 2). These protein groups covered 1,136,558 entries, 93% of which were from TrEMBL—the section of the UniProt database (https://www.uniprot.org) that contains protein sequences predicted from genomes[17] (Fig. 1 and Extended Data Fig. 5). Because we have statistically significant evidence for the existence and correctness of our MS-derived peptide sequences, our data greatly increase the number of experimentally verified proteins, especially in bacteria and archaea. Contrary to our expectations, even well-studied model organisms still contributed many previously unknown proteins. The current Swiss-Prot database (version 2019_03, reviewed section of UniProt; see Methods) encompasses 559,634 experimentally verified proteins from all species. After taking into account proteins that have been described previously in the PRIDE/ProteomeXchange repository (https://www.ebi.ac.uk/pride/archive/), our additional 803,686 proteins more than double the number of proteins with experimental evidence.

To check the depth of proteome coverage, we inspected identifications for model organisms. With more than 5,000 identified protein groups in the yeast *Saccharomyces cerevisiae*, 9,000 in the zebrafish *Danio rerio* and 11,000 in the cotton plant *Gossypium hirsutum*, we obtained an even higher depth in comparison to previous large-scale efforts that focused on individual organisms. In prokaryotes we identified about half of all predicted genes at the protein level, representing a large fraction of the total proteome expressed in a single condition. However, this is less than the coverage obtained in several dedicated studies that used fractionation in these organisms and investigated different conditions. Eukaryotes generally have larger genomes and we identified correspondingly higher numbers of proteins (Fig. 1a). For instance, in a single human cell line, we identified 9,500 protein groups in our standardized workflow—a large proportion of the expressed proteome[6]—whereas 14 cell lines yielded 12,005 protein groups (Supplementary Table 4). Several species had very low proteome coverages. As the MS data were of similar quality in most of these cases (Supplementary Table 5), but the identification rates were low, we attribute

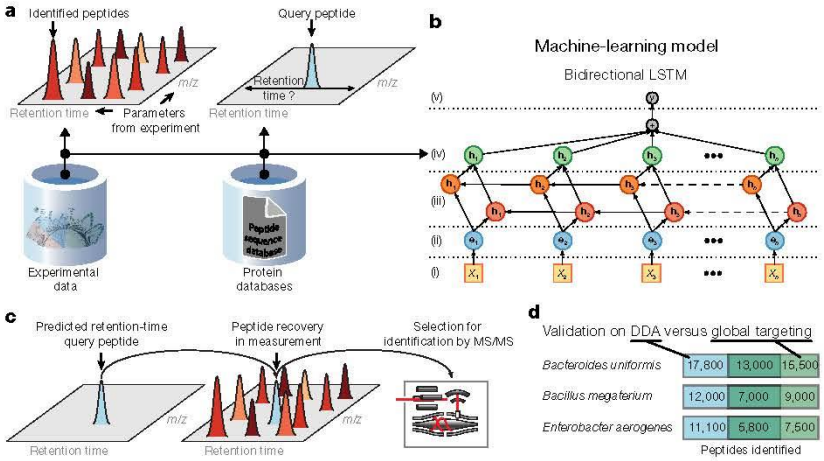


Fig. 2 | Application of a deep learning model to predict peptide retention times for liquid chromatography with tandem mass spectrometry (LC-MS/MS) measurements. a, The data used as inputs for retention time predictions are: left, our experimental data (from Fig. 1a), yielding retention time information on 2 million sequence-unique peptides from 100 organisms; and right, a list of query peptides with unknown retention times derived from a protein database. **b**, Bidirectional LSTM model with attention layer: (i), amino-acid sequence input (x_n); (ii), vectorization of amino-acid information for processing (yielding $\mathbf{e}_n$); (iii), generation of bidirectional LSTM layers ($\mathbf{h}_n$); (iv), attention-based reduction to fixed-length peptide-feature vector ($\mathbf{h}_n$); (v), prediction of retention time (y). **c**, Principle of the global targeting approach displayed for a single peptide: the instrument is set to select the peptide m/z peak for MS/MS identification if it is observed in a narrow retention time window predicted by deep learning. **d**, Application of the 'blind global targeting procedure' to all peptides of three previously unanalysed organisms resulted in the successful detection of predicted peptides in the organism samples. DDA, data-dependent acquisition.

the low proteome coverage to poor genome annotation or proteome prediction, which our data could help to improve through proteogenomics approaches.

In contrast to genomics and transcriptomics, proteomics data allow the direct estimation of the end product of gene expression[18]. We used label-free quantification in MaxQuant to estimate fractional protein intensities across multiple species[19]. Next, we asked how the proteins are distributed across the abundance range of the different organisms, and calculated the number of proteins that contribute to 90% of the total protein amount. The average was 1,546 proteins in eukaryotes, 306 in bacteria and 262 in archaea (Fig. 3a and Extended Data Figs. 6, 7). We used protein homology to enable the quantitative comparison of protein levels between the different organisms. Homology inference is a challenging bioinformatics problem, especially in poorly annotated organisms[20]. To perform the comparison across the studied species, we used high-quality homology prediction from Evolutionary Genealogy of Genes: Non-Supervised Orthologous Groups (EggNOG 5.0)[21]—a database of orthologous groups and functional annotations. We connected our quantitatively determined proteins and corresponding peptides with annotation and structural information data from various sources[17,22–24] in a graph database[25] yielding an explorable network structure with more than 8 million nodes (from proteins, peptides, gene ontology terms, and so on) and more than 53.8 million relationships between them (from homologies, associations, and so on) (Fig. 3b). The graph can be easily queried for any relationship between all of these nodes, as visualized for MS-identified homologues of two species (Fig. 3b). Here an abundant but uncharacterized protein from soybean (*Glycine max*) is linked to its counterpart in wine (*Vitis vinifera*), allowing direct comparison of MS identification, quantification and functional annotations. Similar queries can be performed for entire MS-characterized pathways, organelles or cell compartments. Co-varying pathways or gene ontology terms can also be explored, as well as their relationships to uncharacterized proteins (see www.proteomesoflife.org).

For instance, in soybean, the 11,208 quantified proteins covered more than five orders of magnitude (Fig. 3c) and had 1,763 annotated gene ontology terms. Applying a one-dimensional enrichment analysis to the annotated proteins[26] resulted in 734 statistically significantly enriched terms ($P < 0.05$) (Fig. 3d). Proteins linked to oxidation and reduction processes were the most abundant, reflecting the dominant roles of redox chemistry as a foundation for biochemical reactions such as glycolytic and carbohydrate metabolic processes (among the next most abundant categories). Apart from 'translation process', the most abundant gene ontology term of a biological process was 'protein folding', with an entire 3% of the protein mass. Altogether, functions dedicated to the life cycle of the proteome (translation, elongation, folding and proteolysis) made up a remarkable 10% of proteome mass in living organisms.

Conversely, certain classes of proteins were predominant only in specific branches of life (Extended Data Fig. 8). As expected, photosynthesis-related proteins were present only in photoautotrophic organisms such as plants, algae, protozoa or cyanobacteria (13 out of the 100 organisms) (Fig. 4 and Extended Data Fig. 9). Likewise, numerous functional associations can only be found within Bilateria or even Amniota. These mainly concern proteins associated with differentiation and tissue formation, higher intracellular spatial organization and well-described but subtaxonomy-specific signalling cascades. As expected, protein phosphorylation is predominantly but not exclusively present in eukaryotes. The bacteria and archaea both encompass organisms using this process (for instance in phosphorelay signalling), yet the proportion of the proteome mass involved in it is an order of magnitude lower in these organisms than in eukaryotes.

Much of proteome regulation is accomplished by post-translational modifications, which are typically investigated using specific enrichment protocols followed by MS analysis. However, even our nonenriched workflow in combination with the pFind tool[27] yielded a very large number of peptides with post-translational modifications for which the numbers of modified peptides were proportional to the size of the identified proteome (Extended Data Fig. 10). For instance, we found 29,426 serine phosphorylation sites, almost exclusively in eukaryotes, and 2,862 phosphotyrosine sites were largely restricted to ophistokonts (Supplementary Table 3).

Article

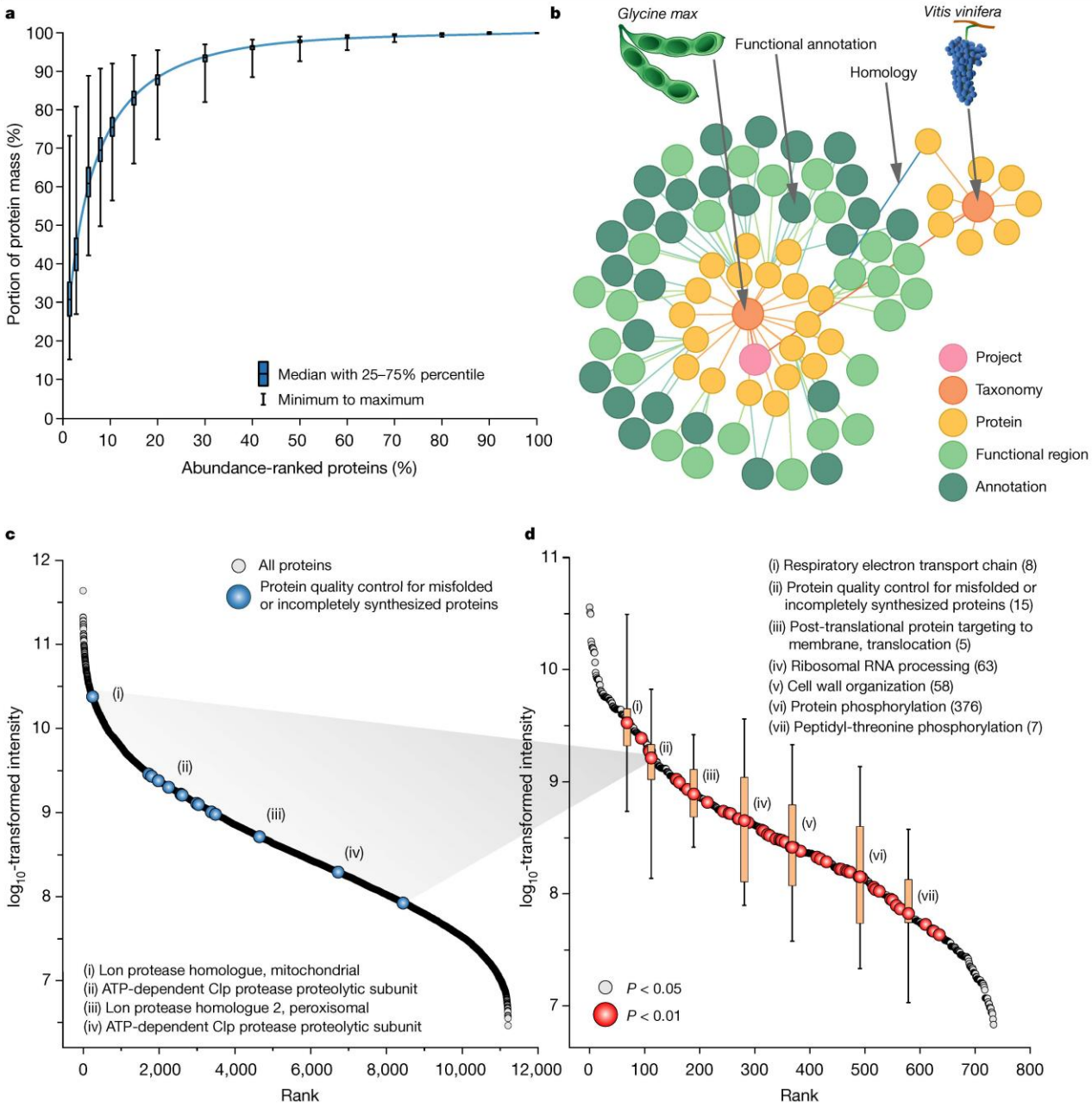


Fig. 3 | Organism-resolved integration of proteome data into a global analysis. a, Cumulative protein intensities (ranked by abundance; *x* axis) and their contribution to total protein mass (*y* axis) across all organisms ($n = 100$ organisms). **b**, Exemplified structure from the data model of the graph database, illustrating the connection between two homologous proteins of *G. max* and *V. vinifera*, and related annotations. **c**, All quantified proteins from *G. max* are displayed, plotting their intensities against their rank in the dynamic range. All proteins for which the functions are associated with 'protein quality control for misfolded or incomplete synthesized proteins' are highlighted. **d**, Significantly enriched functions (grey circles, $P < 0.05$; red circles, $P < 0.01$) within the proteome of *G. max* (with seven specific examples) and their distribution across the dynamic range (sample sizes in parentheses; one-sided Mann–Whitney *U*-test to the mean functional expression level). Error bars represent minimum to maximum values, and boxes show 10–90% percentiles.

Overall, 38.4% of the identified proteins did not have any functional annotation for the biological processes, and interestingly this was true even for 22.9% of the 100 most highly abundant proteins of each species at the biological-process level, and for 10% when considering protein functional domains (Extended Data Fig. 7 and Supplementary Table 6). Thus, our data point to a very large number of highly expressed proteins without any functional annotation or sequence homology to proteins with known gene ontology terms. Exploration of this part of the 'dark proteome' would be attractive: these proteins may indicate essential but unique features in the evolutionary development of these organisms that may be of biological or biotechnological interest.

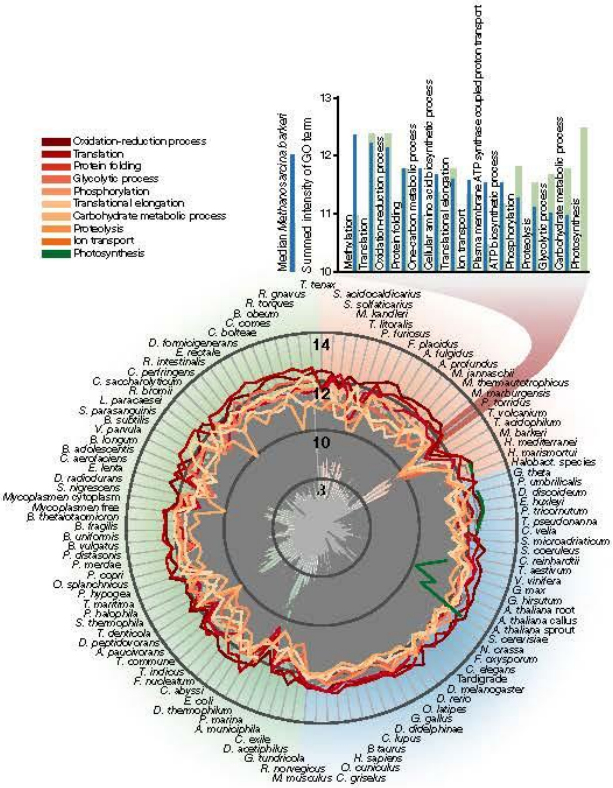


Fig. 4 | Global view of the expression levels of functional groups across the 100 organisms. The main diagram shows summed intensities for functional terms (grey lines), with the ten most abundant terms in all organisms colour-coded according to the key in the top left. The inset in the top right shows the most abundant gene ontology (GO) terms for the archaea *Methanosarcina barkeri* (blue lines), together with the median abundance of all 100 organisms for the displayed terms (green lines).

Advances in sequencing technology are now delivering the genome sequences of an exponentially increasing number of organisms, and we here made a first step towards a parallel scale-up of the characterization of proteomes. Sampling across the taxonomy of life, we created a large set of proteomes with high coverage of their expressed proteins. Label-free quantification values allow us to infer common and specialized biological functions and to compare them to close and distant relatives from all taxonomic levels. The data can be interactively explored at www.proteomesoflife.org.

Limitations of this study include the fact that we measured only selected cell types, tissues and biological states, and that the depth of proteome coverage is not yet comprehensive. Likewise, we have hardly touched upon the post-translational modification of proteins and their evolutionary diversity[28]. Ongoing improvements in MS-based proteomics—including more-refined abundance estimates[29], as well as entire streamlined workflows as described here—will substantially increase throughput in the future[2]. Given the cost effectiveness of proteomic measurements (marginal costs of less than $1,000 per species if its genome is available) and considering the wealth of novel data generated, we propose a community effort to explore many more organisms in different functional states. Integration with genomic, metabolomic and other data, together with incorporation of machine learning methods for species-specific libraries, would expand the systems-biological perspective beyond model organisms to the entire tree of life.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2402-x.

1. de Godoy, L. M. F. et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
2. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
3. 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).
4. Kim, M.-S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
5. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
6. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599 (2017).
7. Weiss, M., Schrimpf, S., Hengartner, M. O., Lercher, M. J. & von Mering, C. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* **10**, 1297–1306 (2010).
8. Marx, H. et al. A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nat. Biotechnol.* **34**, 1198–1205 (2016).
9. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017); correction *Nature* **568**, E11 (2019).
10. 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).
11. Geyer, P. E. et al. Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2**, 185–195 (2016).
12. De Beeck, J. O. et al. Digging deeper into the human proteome: a novel nanoflow LCMS setup using micro pillar array columns (µPAC™). Preprint at *bioRxiv* https://doi.org/10.1101/472134 (2018).
13. 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).
14. Zhou, X.-X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
15. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
16. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
17. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47** (D1), D506–D515 (2019).
18. Muñoz, J. & Heck, A. J. R. From the human genome to the human proteome. *Angew. Chem. Int. Edn* **53**, 10864–10866 (2014).
19. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
20. Altenhoff, A. M. et al. Standardized benchmarking in the quest for orthologs. *Nat. Methods* **13**, 425–430 (2016).
21. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47** (D1), D309–D314 (2019).
22. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47** (D1), D330–D338 (2019).
23. Geer, L. Y. et al. The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492–D496 (2010).
24. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47** (D1), D427–D432 (2019).
25. Santos, A. et al. Clinical knowledge graph integrates proteomics data into clinical decision-making. Preprint at bioRxiv https://doi.org/10.1101/2020.05.09.084897 (2020).
26. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13** (Suppl 16), S12 (2012).
27. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
28. Zielinska, D. F., Gnad, F., Schropp, K., Wiśniewski, J. R. & Mann, M. Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Mol. Cell* **46**, 542–548 (2012).
29. Wiśniewski, J. R., Wegler, C. & Artursson, P. Multiple-enzyme-digestion strategy improves accuracy and sensitivity of label- and standard-free absolute quantification to a level that is achievable by analysis with stable isotope-labeled standard spiking. *J. Proteome Res.* **18**, 217–224 (2019).

Article

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample preparation

Organisms were obtained as stated in Supplementary Table 1. Cell lines were implicitly authenticated by MS and tested for mycoplasma contamination. The LLC-PK1 cell line was contaminated and mycoplasma contamination was harvested for analysis.

We carried out sample preparation according to the in-StageTip protocol[10] with an automated set-up on an Agilent Bravo liquid-handling platform as described[11]. In brief, samples were incubated in PreOmics lysis buffer (catalogue number P.O. 00001, PreOmics) for reduction of disulfide bridges, cysteine alkylation and protein denaturation at 95 °C for 10 min. Root and sprout parts of *Arabidopsis thaliana*, whole *Drosophila melanogaster* and leaves of *Porphyra umbilicalis* were ground in liquid nitrogen with a mortar and pestle beforehand. Samples were sonicated using a Bioruptor Plus from Diagenode (15 cycles, each of 30 s), and the protein concentration was measured using a tryptophan assay. In total, 200 µg of protein from each organism were further processed on the Agilent Bravo liquid-handling system by adding trypsin and LysC (at a 1:100 ratio of enzyme to sample protein, both in micrograms), mixing and incubating at 37 °C for 4 h.

We purified the peptides in consecutive steps according to the PreOmics iST protocol (www.preomics.com). After elution from the solid-phase extraction material, the peptides were completely dried using a SpeedVac centrifuge at 60 °C (Eppendorf, Concentrator Plus). Peptides were suspended in buffer A* (2% acetonitrile (v/v), 0.1% trifluoroacetic acid (v/v)) and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). Eukaryotes generally have larger numbers of genes than bacteria and archaea, resulting in a larger number of proteins and consequently of peptides. To reduce the complexity in the MS measurements, we separated eukaryotic peptide mixtures into eight fractions using the high-pH reversed-phase 'spider fractionator' as described[13].

UHPLC and mass spectrometry

We analysed the samples by applying LC-MS instrumentation, comprising an EASY-nLC 1200 ultrahigh-pressure system (Thermo Fisher Scientific) coupled to a Q Exactive HFX Orbitrap instrument[30] (Thermo Fisher Scientific) with a nano-electrospray ion source (Thermo Fisher Scientific).

For each analysis, 500 ng of purified peptides were separated on a 200 cm µPAC C_{18} microchip nano-LC column (PharmaFluidics). Peptides were loaded in buffer A*. To overcome the void volume of 10 µl, we applied a concentration gradient from 5% buffer B (0.1% formic acid (v/v), 80% acetonitrile (v/v)) to 10% buffer B coupled with a flow gradient from 750 nl min^{-1} to 300 nl min^{-1} for the first 15 min. Subsequently peptides were eluted with a linear gradient from 10% to 30% buffer B in 125 min at a constant flow rate of 300 nl min^{-1}. This was followed by a stepwise increase of buffer B to 60% in 5 min and to 95% buffer B in 5 min. Afterwards we applied a 5 min wash with 95% buffer B, followed by a 5 min decrease to 1% buffer B and a 20 min wash. We kept the column temperature constant at 50 °C by using an oven from Phoenix S&T (catalogue number PST-BPH-15). To avoid interference between the electrospray voltage and the µPAC chip column, we grounded the post-column connection, which was connected by a 20 cm long, 20 µm inner diameter fused silica post-column line to a New Objective Pico-Tip Emitter. This setup is further detailed in Extended Data Fig. 1b. The electrospray voltage was applied by connecting the mass spectrometer source output to the metal connection between the post-column sample line with an in-house-made clamp connection.

HPLC parameters were monitored in real time using SprayQC software[31]. MS data were acquired with a Top15 data-dependent MS/MS method. Target values for the full-scan MS spectra were 3×10^6 charges in the m/z range 300–1,650, with a maximum injection time of 20 ms and a resolution of 60,000 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 a target value of 1×10^5 and a maximum injection time of 28 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides.

Data analysis

MS raw files were analysed using MaxQuant software, version 1.6.1.13 (ref. [32]), and peptide lists were searched against their species-level UniProt FASTA databases. A contaminant database generated by the Andromeda search engine[33] was configured with cysteine carbamidomethylation as a fixed modification and amino-terminal acetylation and methionine oxidation as variable modifications. We set the false discovery rate (FDR) to 0.01 for protein and peptide levels, with a minimum length of seven amino acids for peptides. The FDR was determined by searching a reverse database. Enzyme specificity was set as carboxy-terminal to arginine and lysine as expected, using trypsin and LysC as proteases. A maximum of two missed cleavages was allowed. Peptide identification was performed in Andromeda with an initial precursor mass deviation of up to 7 ppm and a fragment mass deviation of 20 ppm. All proteins and peptides matching the reversed database were filtered out. All bioinformatics analyses were performed using Perseus[34] as well as standard analysis in Python version 3.6.4.

Machine learning model to predict retention times

To predict the retention times of peptides by machine learning, we isolated all detected peptide sequences, including modified peptides. For solvent-induced microshifts between runs, we corrected the detected retention times per peptide by the median shift of all peptides from one run to the median peptide retention time. This resulted in a total of 5,168,800 peptide sequences corresponding to 2,196,869 unique peptide sequences with a median retention time value for retention time prediction.

Our neural network architecture model takes a raw peptide sequence as input. Each amino acid was encoded into a 26-dimensional vector representation for processing using a one-hot encoding scheme, resulting in an $Lx26$ feature vector for a peptide with length L. This vector was connected to a two-layer bidirectional recurrent network with LSTM units with 500 hidden nodes each, which extract context-based features for each individual amino acid. This amino-acid-based feature embedding was reduced to a global 128-dimensional peptide-feature vector by an attention layer, which predicts the contribution of each individual amino-acid feature vector to the regression task. This peptide-feature vector was the input to a logistic regression layer, which regresses the expected retention time for the peptide sequence. The combination of recurrent layers with the attention layer allowed the model architecture to process peptide sequences with arbitrary lengths, but at the same time allow interpretability. The model was end-to-end trained on 2,125,113 peptides and validated on 54,490 holdout peptides. To validate the retention time prediction in vitro, we used the trained model to predict the peptide retention times of all tryptic peptides from *B. uniformis*, which were not included in the training set. We set the mass spectrometer to sequence only if the peptide eluted in a window of 1.4 s around the predicted retention time. This 'global targeting' was done using MaxQuant.live software (version 0.15)[35].

Graph database and cloud data-analysis notebook

To allow exploration of the MS experimental results, we developed a graph database (Neo4j: http://neo4j.com/, version 3.5.8, community edition) that collects all of the experimental data as well as homology and

functional annotations from different publicly available resources[17,21–24,36]. The implemented data model contains 11 different types of node and 14 types of link among the nodes; the data amount to 7,410,594 nodes and 35,517,979 relationships (5.02 GB). To populate the graph, flat files from source databases were downloaded and parsed to generate tab-delimited files comprising nodes and relationships, and standardized using selected terminologies and ontologies. The relationships collected in the database describe ontology structures (Directed Acyclic Graph relationships) and homology (orthology or paralogy) or functional associations (biological processes, functional regions, and so on). A version of the database is accessible at http://www.proteomesoflife.org.

The website gives access to interactive analyses implemented in Python (version 3.6), and uses Cypher as the query language (https://neo4j.com/developer/cypher-query-language/) (see also ref. [37]).

Data integration and comparison

We compared data in online proteomics repositories (PRIDE (https://www.ebi.ac.uk/pride/) and ProteomeXchange (http://www.proteomexchange.org)) with our data from 100 organisms, and downloaded either the provided protein tables or the raw files (Supplementary Table 6). We analysed the raw files with the same MaxQuant version and sequence files as used in our study. If identifiers other than UniProt identifiers were used, we applied the UniProt database to find the corresponding entries and to determine those proteins for which there was previous MS evidence.

Reporting summary

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

Data availability

The MS-based proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository and are available via ProteomeXchange with identifier PXD014877 and PXD019483.

Code availability

Custom computer code is available at https://github.com/MannLabs/proteomesoflife.

30. Kelstrup, C. D. et al. Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
31. Scheltema, R. A. & Mann, M. SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11**, 3458–3466 (2012).
32. 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).
33. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
34. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
35. Wichmann, C. et al. MaxQuant.Live enables global targeting of more than 25,000 peptides. *Mol. Cell. Proteomics* **18**, 982–994 (2019).
36. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47** (D1), D442–D450 (2019).
37. Perkel, J. M. Why Jupyter is data scientists' computational notebook of choice. *Nature* **563**, 145–146 (2018).

Acknowledgements We thank all members of the Proteomics and Signal Transduction Group and the Clinical Proteomics Group at the Max Planck Institute of Biochemistry, Martinsried, for help and discussions, and in particular I. Paron, C. Deiml, A. Strasser and B. Splettstoesser for technical assistance. We further thank the P. Bork group for supplying bacteria, the A. Pichlmair group for virus samples, F. Hosp for *A. thaliana*, I. Sinning for *Neurospora crassa* and the K.-P. Janssen group for cell line samples. Our work was partially supported by the Max Planck Society for the Advancement of Science, by the European Union's Horizon 2020 research and innovation program with the Microb-Predict project (grant 825694), by grants from the Novo Nordisk Foundation (NNF15CC0001 and NNF15OC0016692), and by the Deutsche Forschungsgemeinschaft (DFG) project 'Chemical proteomics inside us' (grant 412136960).

Author contributions J.B.M. and P.E.G. designed the experiments, performed and interpreted the MS-based proteomic analyses, carried out bioinformatics analyses and generated text and figures for the manuscript. P.V.T., S.D., S.V.W. and J.M.B. designed experiments and performed MS-based proteomics analyses. A.R.C. and A.S. integrated annotation data with proteomics data and implemented the Python code as well as graph-based structures. A.S. and M.O. implemented the web-accessible analyses. N.K., F.T. and M.T.S. carried out the machine learning analysis. M.M. supervised and guided the project, designed the experiments, interpreted MS-based proteomics data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2402-x.
Correspondence and requests for materials should be addressed to M.M.
Peer review information *Nature* thanks Joshua Coon, Vera van Noort and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

3.2. Article 2: A new high-pressure packing system enables rapid multiplexed packing of capillary columns

Authors: Johannes B. Müller-Reif[‡ǁ], Fynn M. Hansen[‡], Lisa Schweizer[‡], Peter V. Treit[‡], Philipp E. Geyer[‡,ǁ,¶], Matthias Mann[‡,¶,*]

‡Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
¶NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
ǁNew address: OmicEra Diagnostics GmbH, Planegg, Germany
*Corresponding author. Tel: +49 89 8578 2557; E-mail: mmann@biochem.mpg.de

By bringing different principles for column packing together, I developed the new packing station for capillary columns presented in the manuscript ‘A new high-pressure packing system enables rapid multiplexed packing of capillary columns’. Previous publications reported that columns can be packed with pressures up to 2000 bar and that this increases the performance of the chromatography. Packing with dense bead slurries has also been linked to increased chromatographic performance and to rapid column packing. By combining both principles, I speeded up the packing process of capillary columns many-fold.

Despite alternative systems like commercial columns combined with high flow LC or chip-based columns, the vast majority of LCMS-based proteomics applications is performed with capillary columns. Those can be purchased in different forms, as packed emitters, ready to be used for ES (Ion opticks) or fritted columns for use with different types of ES emitters (PepSep). These options are relatively expensive and therefore laboratories with high throughput traditionally pack their own capillaries. Empty pulled emitters can be purchased or produced from fused silica with a laser puller to form an ES emitter tip ready to be packed with particles all the way into the end of a column. To pack sub 5 µm particles into the capillaries from 50 to 150 µm ID, gas pressure bombs are most commonly used, but this process is slow and limited to pressures in the range of 100 bar or 300 bar for most systems.

Our new packing station overcomes these limitations by enabling packing of high-density bead slurry into the capillaries under pressures up to 3000 bar. With this system a large number of beads can enter the column at the same time and the flow rate during packing is kept high. By the introduction of multiplexed packing of up to ten columns at the same time we increase productivity even more.

We find similar performance metrics for columns produced increasing packing pressure, so that the new packing station is now in routine use in the column production process of our group.

A new parallel high-pressure packing system enables rapid multiplexed production of capillary columns

Johannes B. Müller-Reif[‡,ǁ], Fynn M. Hansen[‡], Lisa Schweizer[‡], Peter V. Treit[‡], Philipp E. Geyer[‡,ǁ,¶], Matthias Mann[‡,¶,*]

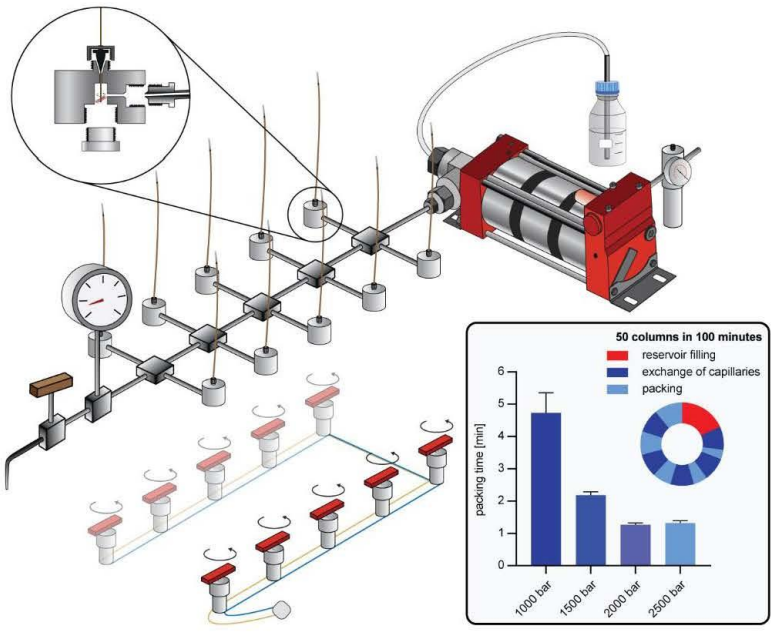
Affiliations

[‡]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

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

[ǁ]New address: OmicEra Diagnostics GmbH, Planegg, Germany

*Corresponding author. Tel: +49 89 8578 2557; E-mail: mmann@biochem.mpg.de



Abstract

Reversed-phase high performance liquid chromatography (HPLC) is the most commonly applied peptide separation technique in mass spectrometry (MS)-based proteomics. Particle-packed capillary columns are predominantly used in nano-flow HPLC systems. Despite being the broadly applied standard for many years capillary columns are still expensive and suffer from short lifetimes, particularly in combination with ultra-high-pressure chromatography systems. For this reason, and to achieve maximum performance, many laboratories produce their own in-house packed columns. This typically requires a considerable amount of time and trained personnel. Here, we present a new packing system for capillary columns enabling rapid, multiplexed column production with pressures reaching up to 3000 bar. Requiring only a conventional gas pressure supply and methanol as driving fluid, our system replaces the traditional setup of helium pressured packing bombs. By using 10x multiplexing, we have reduced the production time to just under 2 minutes for several 50 cm columns with 1.9 µm particle size, speeding up the process of column production 40 to 800 times. We compare capillary columns with various inner diameters (ID) and length packed under different pressure conditions with our newly designed, broadly accessible high-pressure packing station.

One sentence summary: A newly constructed parallel high-pressure packing system enables the rapid multiplexed production of capillary columns.

Introduction

State-of-the-art mass spectrometry (MS)-based proteomic pipelines typically consist of a sample preparation workflow to digest proteins and harvest pure peptides, a liquid chromatography (LC) system for peptide separation, a mass spectrometer and a sophisticated bioinformatics pipeline for raw data interpretation and subsequent statistical analysis (1,2). The LC system plays a central role by separating the complex mixture of tens of thousands of peptides in a time-resolved manner according to their biochemical properties, making them ultimately manageable for the MS system over the course of a gradient (3,4). The most widely applied technique for high-performance applications is reversed-phase separation, originally introduced in the 1970s (5). In essence, chromatographic systems are made of programmable pumps with the ability to form a gradient of a mixture of different agents. In the case of reversed-phase LC, the stationary phase is nonpolar, separating analytes by hydrophobicity over the course of a gradient of increasing nonpolar mobile phase. The LC system is coupled to the mass spectrometer by electrospray (ES) ionization via an emitter (6). Glass or steel needles are commonly connected to the column. Particle packed capillaries for chromatography can also be used for ES without being coupled to an additional emitter (7–9). These basic attributes are shared by most LC-MS systems and differences are mainly defined by operational flow. Nano-flow LC operates at flow rates of several hundred nano-liters per minute and is the standard in proteomics due to the high sensitivity obtainable.

High flow rates in the µl to ml range, applied to columns with large inner diameters, are typically used in high-throughput or industrial-scale analysis as well as analytical MS application areas. Although these micro-flow systems limit sensitivity, recent work has demonstrated robust and

reproducible performance (10,11). Reproducibility and stability of those systems are high, but drawbacks are lowered sensitivity and a need for high sample amounts. Compared to developments in sample preparation, MS instrumentation, scan modes and software, the LC apparatus has been largely unchanged in cutting edge MS-based proteomics. While identifications in proteomics experiments have doubled in single-shot experiments this can mainly be traced to improvement on the MS instrumentation and software (12–17). Current trends in LC developments aim rather towards systems for higher throughput and increasing robustness required for clinical applications (18), whereas the race for better separation in single-shot high performance runs with increasingly higher pump pressures has been comparatively abandoned. Consequently, a typically used setup for maximum sensitivity and performance for most experiments still consists of columns around 75 µm ID with a length of 20 to 50 cm, packed with sub 2 µm particles. Although, better performance could be reached by longer columns or smaller particles, both conditions would result in higher operational pressures which tend to make the LC systems unstable (4,19). For example, very high pressures can lead to leaks in the LC flow paths, resulting in poor reproducibility and subsequently a loss of measurement time.

Commercially available capillary columns in the aforementioned dimensions are expensive, especially considering how frequently they must be replaced. Therefore, many high-throughput laboratories produce packed capillaries in-house. Empty glass capillaries, ready to be packed and used, can either be purchased or produced from cheap polyimide coated capillaries using a laser puller. Typically, a gas pressure system is deployed to pack such columns with particles in the low µm range and instructions on the manufacturing process can be found online with open access (https://proteomicsresource.washington.edu/docs/protocols05/Packing_Capillary_Columns.pdf). However, this process is inherently slow and interesting methods have recently been established with the aim of speeding up the packing process with high-pressure (20) or dense bead-slurry, as in the FlashPack method (21).

Combining these principles, we here present a high-pressure packing system for capillary columns using a high-concentration bead slurry that has previously been described as beneficial for column performance (22). These high slurry concentrations and packing pressures of 1000 - 3000 bar allow us to achieve packing times for 50 cm columns in the minute range with our system, compared to hours for traditional procedures. Deploying a manifold system and a pump capable of high flow rates further multiplexes packing to up to 10 columns, making column production 40 to 800 times more time efficient compared to previous systems. We observe consistently good column performance for packing pressures at over 1000 bar with no adverse effects on the column backpressure and lifetime, while packing times continued to decrease at higher pressures. We provide a detailed blueprint of the system so it can readily be set up in interested laboratories (Suppl. Table 1).

Experimental Procedures

Preparation of fused silica

Fused silica from Polymicro® (TSP075365 for 75 µm ID, TSP100365 for 100 µm ID or TSP150365 for 150 µm ID) was cut to 140 cm. Polyimide coating was removed by a Bunsen burner and polishing with an ethanol-soaked tissue in the middle of the cut capillary at a width of 2 cm. An electro spray emitter tip was pulled with a laser puller (Sutter P2000) at the polished part of the capillary resulting in two empty capillary columns ready to be packed.

Sample preparation: Protein digestion and in-StageTip purification

HeLa cells were cultured in high glucose DMEM with 10% fetal bovine serum and 1% penicillin/streptomycin (all from Life Technologies, Inc.). Cells were counted using a countess cell counter (Invitrogen), and aliquots of $1x10^6$ cells were washed twice with PBS (Life Technologies, Inc.), snap-frozen and stored at −80°C. Sample preparation was carried out with the PreOmics iST kit (www.preomics.de). We used one HeLa pellet with 1 million cells per cartridge, determinate the peptide concentration after peptide cleanup via NanoDrop and adjusted the peptide concentration to 0.2 mg/ml.

Ultra-high-pressure liquid chromatography and mass spectrometry

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1200 ultra-high-pressure system (Thermo Fisher Scientific), coupled to an Orbitrap Exploris 480 instrument (Thermo Fisher Scientific) using a nano-electrospray ion source (Thermo Fisher Scientific). Purified peptides were separated on high-pressure packed columns as described in the results section. For each LC-MS/MS analysis with 75 µm ID columns, 500 ng peptides were used. For 100 µm ID columns, 888 ng peptides were used and for 150 µm ID columns 2000 ng peptides were used to adjust for the higher column volume.

Peptides were loaded in buffer A* (2% acetonitrile (v/v), 0.1% trifluoroacetic acid (v/v)) and eluted with a linear 105 min gradient of 5-30% of buffer B (0.1% formic acid, 80% (v/v) acetonitrile), followed by a 10 min increase to 95% of buffer B, followed by a 5 min wash of 95% buffer B. For the 75 µm ID columns flow rate was 300 nl/min, 535 nl/min for 100 µm ID columns and 1200 nl/min for 150 µm ID columns to adjust for the linear flow velocity. Column temperature was kept at 60 °C by an in-house-developed oven containing a Peltier element, and parameters were monitored in real time by the SprayQC software. MS data was acquired with a Top15 data-dependent MS/MS scan method. MS1 AGC Target was set to 300% in the 300-1650 m/z range with a maximum injection time of 25 ms and a resolution of 60,000 at m/z 200. Fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 30 eV. MS/MS scans were performed at a resolution of 15,000 at m/z 200 with an AGC Target of 100% and a maximum injection time of 28 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides. Each column was equilibrated with two 120 min HeLa runs before the representative run for column cross-comparison.

Data analysis

MS raw files were analyzed by MaxQuant software, version 1.6.11.0, and peptide lists were searched against the human Uniprot FASTA database. A contaminant database generated by the Andromeda search engine was configured with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. We set the false discovery rate (FDR) to 0.01 for protein and peptide levels with a minimum length of 7 amino acids for peptides and the FDR was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin and LysC as proteases. A maximum of two missed cleavages were allowed. Peptide identification was performed with an initial precursor mass deviation up to 7 ppm and a fragment mass deviation of 20 ppm. All proteins and peptides matching to the reversed database were filtered out.

Bioinformatics analysis

Bioinformatics analyses were performed in Python (version 3.6.4.) using Numpy (1.19.2), Pandas (1.1.4), Matplotlib (3.3.2), Seaborn (0.11.0) and Scipy (1.5.2) packages.

Experimental design and statistical rationale

The overall experimental design was focused on making different capillary columns for proteomics experiments as comparable as possible. To achieve this, statistical analysis was done from triplicate experiments for the packing time and pressure performance experiments. Experimental conditions for column cross-comparisons were chosen to eliminate outer influences, including measurements on the similar LC and MS system and equilibration procedures.

Results and discussion

A high-pressure packing chamber for high density bead-slurries

A central challenge of nano-flow chromatography in proteomics laboratories is the constant demand for new capillary columns. Due to their costs, commercial columns cannot be treated as a disposable item. However, in our hands, we frequently observe highest performance only for a short lifespan for ultra-high-performance applications. Therefore, to reach the needed quantity and cost requirements, we and many other laboratories produce their own capillary columns. However, the throughput of production is limited, especially for columns with small inner diameter and extended length such as the 50 cm 75 µm inner diameter columns used in most applications in our laboratories. We produce pulled or fritted capillaries and pack them with solid phase material, typically sub 2 µm C18 beads. A skilled person can pull hundreds of empty columns within a day and fritted columns are also easy to produce. However, the packing process is an inherently low-throughput and error-prone process, which makes high-performance columns prized items in mass spectrometry laboratories. Particularly achieving longer columns length is – in our experience – a precondition for ultra-high-performance.

We hypothesized that high-throughput packing of capillary columns could be achieved by highly concentrated bead-slurries (21) in combination with very high-pressure packing (>1000 bar) (20). However, increased packing pressure and bead-slurry concentration can lead to column blocking, slowing down and eventually halting the packing procedure. Chloroform as a bead-solvent was reported as an approach to avoid this issue, because it can solvate higher bead concentrations. However, in combination with our bead-particles, we observed poor chromatographic performance during proteomic experiments. Instead, we combined elevated packing pressure with the FlashPack system (21), which prohibited bead aggregation at the column entrance via stirring.

To test our concept, we constructed a custom–made chamber for high-pressure packing, where the pressure derives from a conventional HPLC system (EASY-LC 1000 in our case). The device consists of a central chamber, containing the bead slurry and magnetic stirring bar, and has three openings. A large-bore access allows filling the chamber with the bead-slurry, a micro-bore fitting holds the capillary entrance into the chamber and a nano-viper connection is used as an inlet for the pressure from the HPLC system (Suppl. Fig. 1). The prototype packing chamber enabled us to fill single capillaries within minutes using the HPLC high-pressure pumps (950 bar). However, this system was not suited for high-throughput column production and the low pump volume of the HPLC system resulted in a non-continuous packing as the pump had to be refilled several times until a column was filled with beads.

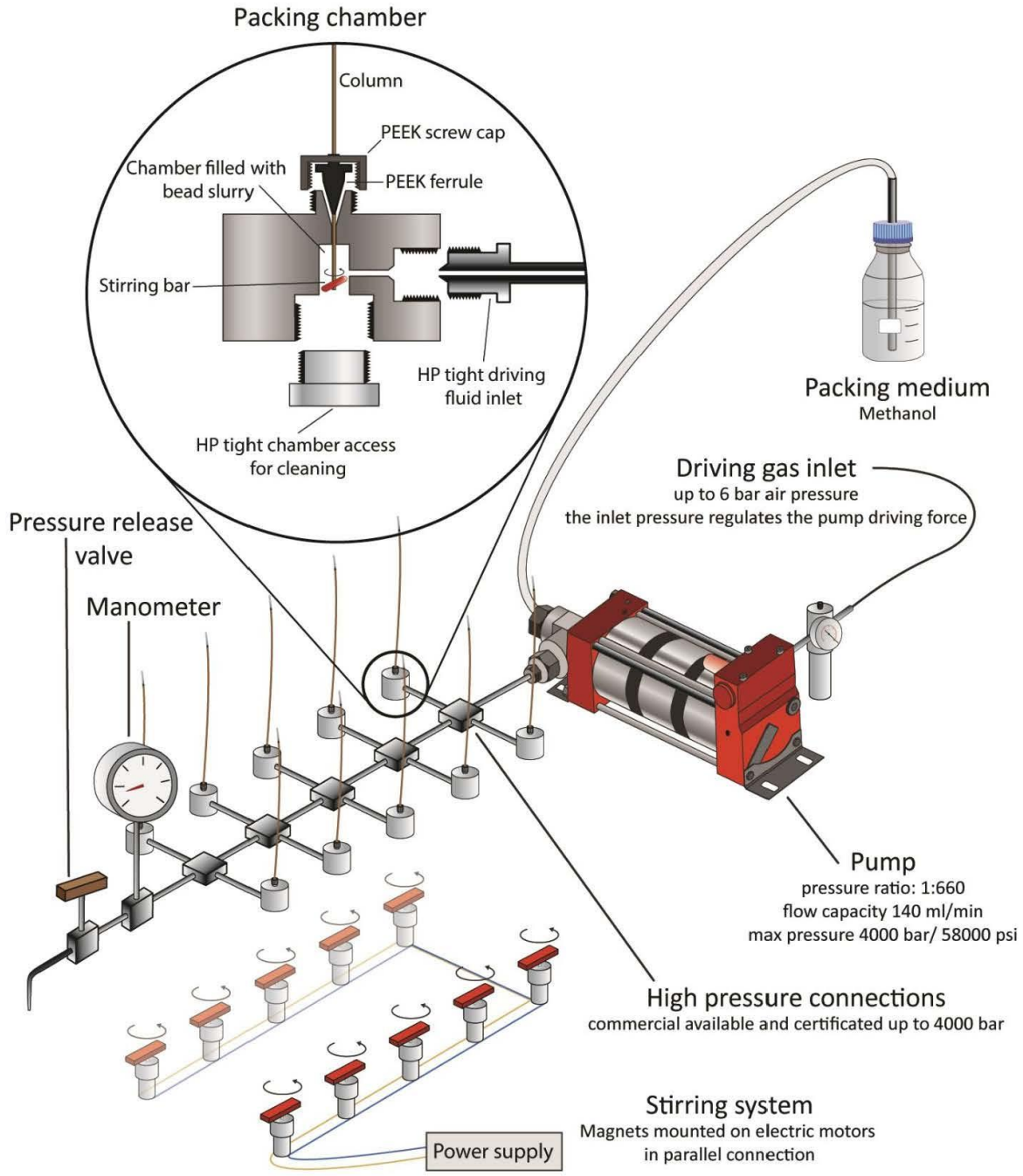


Fig. 1: High-pressure packing station. Scheme of the high-pressure packing station with detailed description of the crucial parts. The high-pressure pump is powered by a driving gas inlet and increases the pressure of a packing medium that is provided in a large volume flask by 660-fold. The compressed packing medium is channeled to ten packing chambers, placed on top of a magnetic stirring rack. A manometer is installed to monitor the system pressure as well as a pressure release valve to facilitate time efficient system depressurization. The inset depicts a packing chamber in detail, including high-pressure fittings, stirring bar and capillary column.

Encouraged by aspects of our newly devised packing system, we set out to further streamline column production. We replaced the small volume HPLC pump with a Maximator HD-pump (Experimental Methods). This high-flow continuous system converts driving gas from a standard laboratory gas supply line at a pressure ratio of 1:660 to a fluid outlet with a maximal pressure rating of 4000 bar and maximal flow capacity of 140 ml/min (Fig 1). To use the FlashPack principle we used methanol as the packing medium, which settles C18 beads at the chamber bottom. The high flow capacity allowed us to implement multiple pump outlets for multiplex packing of up to ten columns with our station. We redesigned the original packing chamber to fit high-pressure connections (Suppl. Fig. 2). For optimal stirring, we further created a rack system with magnets mounted on electric motors via 3D printed components to fit directly underneath the packing stations (Detailed in Experimental Methods and Suppl. Fig. 3). Moreover, we connected a high-pressure range manometer to monitor packing pressure and added a pressure relief valve for efficient and controlled depressurization of the system, a notoriously time-consuming process. Even though the system is typically running at 1500 bar in our laboratory, the relief of pressure takes only 60 seconds, without flow-back from the running beads from the capillary. Additionally, the system is secured from capacity exceeding driving gas pressure by a control valve, which prevents the pump to be exposed to higher input than 6 bar. As with conventional packing systems, the weakest connection is the sealing of the capillary to the high-pressure chamber. We used a standard polyether-ether-ketone (PEEK) ferrule employed in HPLC applications in combination with a newly designed, reinforced PEEK screw cap (Suppl. Fig. 2D) to pin the column under very high pressure. Nevertheless, if the system pressure exceeds the durability of the material, the column is ejected. Due to the low compression capabilities of methanol, this is dangerous if one has body parts directly above the fitting when a rupture occurs and this must be prevented. Compared to gas, which can compress much more than liquid, no explosion risk should arise from our new packing station. Nevertheless, our recommendation is to use this device only within a secured area.

Ultra-fast column packing

The time required to fill a capillary column with beads depends on two variables, the bead concentration of the packing slurry and the flow rate through the capillary. Empty capillaries with pulled electrospray emitter have high flow rates in the µl/min range even for conventional gas-based packing bombs with lower pressure (<100 bar). However, as the bead bed grows, the flow rate through the column decreases drastically. Hence, the high-density bead slurry of FlashPack enables short packing times especially for shorter columns (21). We anticipated that combining this principle with the potentially high flow rates of our extremely high-pressure system would significantly reduce packing times.

To quantify the production throughput of our system, we consecutively packed 50 cm capillaries with 75 µm ID at different pressures (1000-2500 bar) and measured the time required. With a freshly filled bead reservoir, packing at the lowest tested pressure took on average 4.7 min. Increasing pressure to 2000 bar results in packing times just over a minute. Even higher pressure did not result in faster packing. Overall our system decreased the time for making a single column 10- to 100-fold compared to previous packing procedures (20,21) (Fig. 2A-B). Of note, the total production throughput is even higher due to multiplex-packing and the option to quickly exchange capillaries and bead slurries. This results in a 40-800 times faster column production (Fig. 2C). Once filled with bead-slurry and mounted on the high-pressure system, the packing

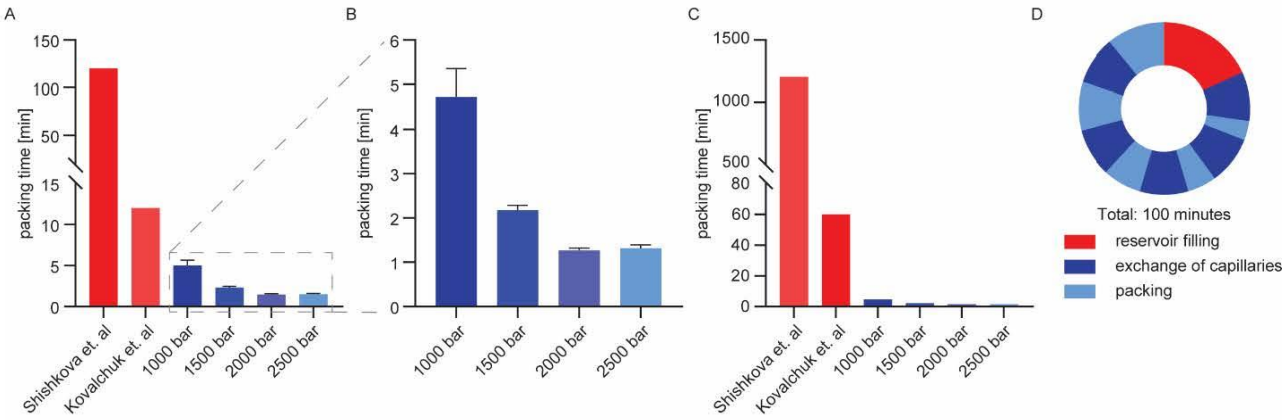


Fig. 2: Comparison of packing times. A Packing times of single columns as described in previous efforts and for different packing pressures (data collected in triplicates, displayed with standard deviation) with a detailed view of the tested pressure conditions (B). C Production time for 10 columns considering multiplexing (2x multiplexing for Kovalchuk *et al.* and 10x for the system presented here) (20,21). D Times of a packing cycle of 10 x 5 columns, taking a total of 100 minutes with filling of the reservoir and changing of capillaries between the actual packing steps.

chambers can be used to pack several columns consecutively. This merely requires depressurizing the system via the pressure relief valve and exchange the filled columns with empty capillaries. Consecutive packing of several columns from the same reservoir will decrease the packing speed due to the removal of beads from the reservoir. To fully restore packing speed, the bead chamber has to be opened and refilled, which takes about 10 min for all ten chambers together. Typically, we refilled the reservoir after five capillary exchanges. The average turn-around cycle for producing ten columns is thus 20 minutes, allowing the production of hundreds of columns in a working day (Fig. 2D). An additional advantage of the high-throughput system is that it allows us to discard non-properly packed columns, which occur in approximately 10% of cases.

The high-pressure system faces the same two main challenges as other packing stations, which are particle clogging within the capillary and bead aggregation at the column entrance. Particle clogging can only be avoided by clean working conditions. This means dust free storage and clean cutting of fused silica and the use of filtered fluids and dust free particles for bead slurry preparation. Bead aggregation from dense slurry can be circumvented by optimized stirring conditions according to the FlashPack principle (21).

Influence of packing pressure on column performance

To evaluate the effect of packing pressure on column performance on realistic samples, we analyzed three of our laboratory standard HeLa digests on each column. Across all packing conditions, we observed no significant variation in the number of identified peptides and protein groups (Fig. 3A/B). Moreover, the median peak widths of identified peptides were comparable for all conditions (Fig. 3C). Correlation between the non-corrected retention times of peptides analyzed using columns produced at varying pressures was remarkably high (Pearson correlation coefficients > 0.996) and not significantly altered from replicates packed with similar pressure conditions (Fig. 3D).

Multiplexed high-pressure column packing

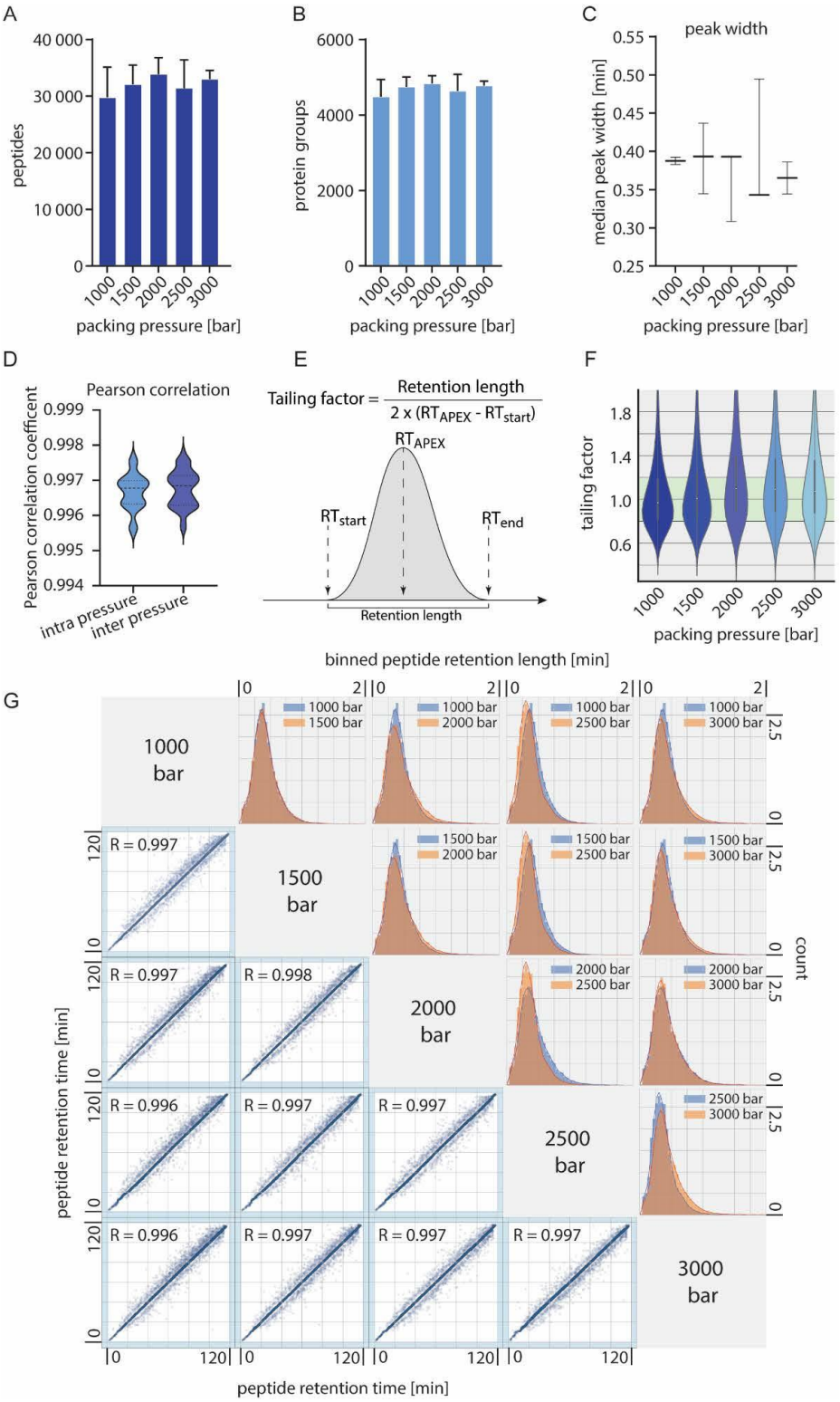


Fig. 3: Comparison of capillary columns packed at different pressures. A, Numbers of identified peptides of triplicate measurements of 500 ng HeLa digests on columns filled at the indicated packing pressures. Peptides were separated on 50 cm, 75 µm ID columns packed with 1.9 µm Reprosil AQ Beads (Dr. Maisch) with a 2-hour gradient. B, Numbers of identified protein groups of the same conditions as in (A). Error bars indicate the standard deviation from triplicate measurements. C, Median peak widths of identified peptides. D, Distribution of Pearson correlation coefficients calculated on peptide retention times between columns packed at the same pressure and columns packed at different pressures (p-value of unpaired t-test for difference: 0.6). E, Visualization of the tailing factor calculation. F, Tailing factors for all identified peptides from runs with 75 µm ID columns and different packing pressures. G, Correlation of peptide retention times across packing conditions. The density of peptides is color-coded. The histograms show the peak width distribution of four representative runs.

Another factor often used to characterize column performance is the tailing factor which can be calculated as depicted in figure 3E (23). Usually, the peak width at 5% peak height is used for peak width calculation but in proteomics experiments where tens of thousands of peaks are investigated, the base-to-base peak width is typically calculated, although full width at half maximum (FWHM) is also often reported. In general, the distribution of peak shapes was wider than what would be expected from an analysis run of few analytes, but the median typically centered around the optimum of 1. The median of the tailing factor was below 1.0 for the lower and shifts above 1.0 for higher packing pressures up to a median of 1.2 (Fig 3F). In the literature tailing factors in the range between 1 – 1.2 are often described (24). The shift towards this range with the higher packing pressures could result from denser compressed bead bed. As described above the general performance was not altered for the proteomics metrics,

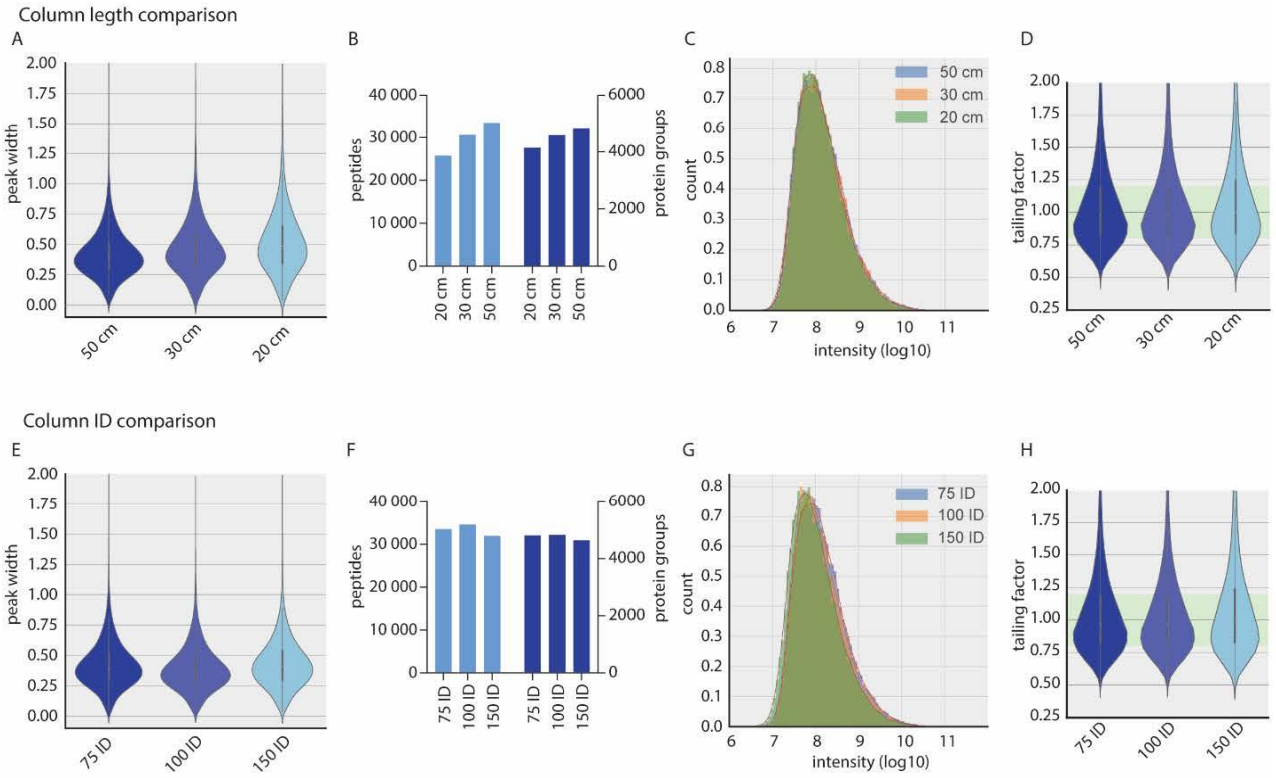


Fig. 4: Length and inner diameter comparison. All columns were packed with 1000 bar packing pressure. A, Peak width distribution from HeLa runs with different column length with the respective number of peptide and protein identifications (B), peptide intensity distribution (log10) (C) and tailing factor distribution (D). E, Peak width distribution from HeLa runs with different column IDs with the respective number of peptide and protein identifications (F), peptide intensity distribution (log10) (G) and tailing factor distribution (H).

which leads us to the conclusion that the minor change in tailing factors with higher packing pressures is not changing the LC-MS performance. This manifests in an only slightly altered distribution of peak widths between representative experiments of columns packed at different pressures (Fig. 3G). From the correlation of peptide retention times, it is visible that for all representative comparisons, the peptides elute in a narrow and reproducible time window that is not influenced by the applied packing pressure. This retention time stability is accompanied by similar separation properties of the different columns, which can be visualized directly by the retention length of analyzed molecules. Figure 3G shows bulk analysis of all identified peptides with nearly overlapping retention length distributions whereas the minor differences do not constitute a significant trend towards a better performance for lower or higher packing pressures of capillary columns. Based on these results it seems that the packing pressure has no or only minimal effect on the column performance.

LC-MS performance of columns with different length and inner diameter

Length and inner diameter of capillary columns allow their adaptation to a plethora of sample materials and LC systems, specifically regarding separation power and backpressure. In MS-based proteomics, 75 µm ID columns in combination with flow rates in the range of 200-400 nanoliter per minute are typical. Hence, we packed such capillary columns with different lengths (20, 30, 50 cm) with our high-pressure system and compared their performance. Packing time for the shorter columns was even faster and in the range of 30 sec. The longest columns produced the smallest peak widths and subsequently resulted in the highest numbers of identified peptides and proteins (Fig. 4A-B). Interestingly, the distribution of peptide intensities did not change significantly, and the tailing factor also remained unaffected (Fig 4C-D). Over the last years the demand for high-throughput analysis has become apparent for the analysis of clinical samples, especially blood plasma as we have described before (25). This has been addressed by a novel HPLC principle with pre-formed gradients and slightly higher flow rates (18) and by higher-flow systems operating in the upper microliter per minute range (10,26). As these strategies require columns with higher inner diameter to maintain acceptable pressure during analysis, we produced columns with 75 µm, 100 µm and 150 µm ID and tested their performance.

When comparing column inner diameters, the experimental setup has to be adapted to the conditions. To enable direct comparison of capillaries with different ID, we scaled the flow rates to reach the same linear velocities and the amount of input material to the column volume (Experimental Procedures). For the 100 µm ID columns this results in a flow rate of 535 nl/min and 888 ng of peptides for loading, whereas for the 150 µm ID column 1200 nl/min and 2 µg of peptide material was loaded to be comparable to the 300 nl/min and 500 ng employed for the 75 µm ID columns. This requirement of higher sample amount already limits the applicability of larger column diameters for samples with limited accessibility. The 1400 µl pump volume of the Easy-LC 1200 used for the experiment were sufficient to run a 2-hour gradient with the 150 µm ID column, but longer gradients or higher flow rates would exceed the capabilities of the LC-system and require higher flow rates. The larger column IDs led to slightly broader peak widths, but peptide and protein identifications were not affected. Due to the correction of the sample input amount, we have not seen a difference in peptide intensity distributions, and the peak tailing has not been affected by the column ID (Fig 4E-H).

Conclusion

Here, we aimed to increase the throughput and to streamline the production of capillary columns for MS-based proteomics. We provide a detailed list for commercial parts and blueprints describing the construction of our high-pressure packing station. The setup can be built at relatively low costs (<$10,000), compared to the cumulative expenses for high performing commercial columns. We designed this new station to fill multiple columns simultaneously within a few minutes, which accelerates the packing process of capillary columns more than 100-fold compared to traditional gas pressure driven stations. In this way, we hope our system helps researchers streamlining the often work-intensive and fragile column production process. In addition, the extreme high pressures enable the packing of long, high-performing columns (> 50 cm). The ability to produce high-performing columns at high-throughput opens up the possibility of using capillary columns always at the peak of their performance, replacing them as soon as peak-broadening or decreased ionization is observed. Reassuring in terms of the robustness of the packing process itself and the stability achieved at exceedingly high pressures, we have not observed variation in the performance characteristics over a wide range of packing pressure from 1000 to 3000 bar. We hope the technology described here will enable laboratories of any scale to mass-produce high performance long capillary columns.

Literature

1. Aebersold R, Mann M. Mass spectrometry-based proteomics [Internet]. Vol. 422, Nature. Nature Publishing Group; 2003 [cited 2020 Dec 15]. p. 198–207. Available from: www.nature.com/nature
2. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature [Internet]. 2016 Sep 15 [cited 2019 Apr 21];537(7620):347–55. Available from: http://www.nature.com/articles/nature19949
3. 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 [Internet]. 2011 Apr 1 [cited 2021 Feb 20];10(4):1785–93. Available from: https://pubs.acs.org/sharingguidelines
4. Shishkova E, Hebert AS, Coon JJ. Now, More Than Ever, Proteomics Needs Better Chromatography. Vol. 3, Cell Systems. Cell Press; 2016. p. 321–4.
5. Horváth C, Melander W, Molnár I. Solvophobic interactions in liquid chromatography with nonpolar stationary phases. J Chromatogr A. 1976 Sep 29;125(1):129–56.
6. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989.
7. Kennedy RT, Jorgenson JW. Preparation and Evaluation of Packed Capillary Liquid Chromatography Columns with Inner Diameters from 20 to 50 µm. Anal Chem [Internet]. 1989 [cited 2021 Feb 20];61(10):1128–35. Available from: https://pubs.acs.org/sharingguidelines
8. Emmett MR, Caprioli RM. Micro-Electrospray Mass Spectrometry: Ultra-High-Sensitivity Analysis of Peptides and Proteins [Internet]. 1994 [cited 2021 Feb 20]. Available from: https://pubs.acs.org/sharingguidelines
9. Ishihama Y, Rappsilber J, Andersen JS, Mann M. Microcolumns with self-assembled particle frits for proteomics. J Chromatogr A [Internet]. 2002 Dec 6 [cited 2020 Oct 29];979(1–2):233–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021967302014024
10. Bian Y, Zheng R, Bayer FP, Wong C, Chang YC, Meng C, et al. Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. Nat Commun. 2020 Dec 1;11(1).
11. Bian Y, Bayer FP, Chang Y-C, Meng C, Hoefer S, Deng N, et al. Robust Microflow LC-MS/MS for Proteome Analysis: 38 000 Runs and Counting. Anal Chem [Internet]. 2021 Feb 17 [cited 2021 Feb

22];acs.analchem.1c00257. Available from: https://pubs.acs.org/doi/10.1021/acs.analchem.1c00257

12. Bernhardt O, Selevsek N, Gillet L, Rinner O, Picotti P, Aebersold R, et al. Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data . F1000Research. 2014 Aug 14;5.

13. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2019 Apr 21];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511

14. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Cooks RG. The Orbitrap: A new mass spectrometer [Internet]. Vol. 40, Journal of Mass Spectrometry. John Wiley & Sons, Ltd; 2005 [cited 2021 Jan 20]. p. 430–43. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/jms.856

15. Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen J V. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. J Proteome Res [Internet]. 2018 Jan 5 [cited 2020 Dec 15];17(1):727–38. Available from: https://pubs.acs.org/sharingguidelines

16. Meier F, Beck S, Grassl N, Lubeck M, Park MA, Raether O, et al. Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. J Proteome Res [Internet]. 2015 [cited 2021 Jan 20];14(12):5378–87. Available from: https://pubmed.ncbi.nlm.nih.gov/26538118/

17. Meier F, Brunner AD, Frank M, Ha A, Bludau I, Voytik E, et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. Nat Methods [Internet]. 2020 Dec 1 [cited 2021 Jan 20];17(12):1229–36. Available from: https://doi.org/10.1038/s41592-020-00998-0

18. Bache N, Geyer PE, Bekker-Jensen DB, Hoerning O, Falkenby L, Treit P V., et al. A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. Mol Cell Proteomics. 2018 Nov 1;17(11):2284–96.

19. Richards AL, Merrill AE, Coon JJ. Proteome sequencing goes deep. Vol. 24, Current Opinion in Chemical Biology. Elsevier Ltd; 2015. p. 11–7.

20. Shishkova E, Hebert AS, Westphall MS, Coon JJ. Ultra-High Pressure (>30,000 psi) Packing of Capillary Columns Enhancing Depth of Shotgun Proteomic Analyses. Anal Chem [Internet]. 2018 Oct 2 [cited 2020 Oct 29];90(19):11503–8. Available from: https://pubs.acs.org/sharingguidelines

21. Kovalchuk SI, Jensen ON, Rogowska-Wrzesinska A. FlashPack: Fast and simple preparation of ultrahigh-performance capillary columns for LC-MS. Mol Cell Proteomics [Internet]. 2019 Feb 1 [cited 2020 Dec 15];18(2):383–90. Available from: https://www.mcponline.org

22. Bruns S, Franklin EG, Grinias JP, Godinho JM, Jorgenson JW, Tallarek U. Slurry concentration effects on the bed morphology and separation efficiency of capillaries packed with sub-2μm particles. J Chromatogr A [Internet]. 2013 Nov 29 [cited 2020 Oct 30];1318:189–97. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021967313016178

23. Analytical Separation Science. Analytical Separation Science. Wiley-VCH Verlag GmbH & Co. KGaA; 2015.

24. Birdsall RE, Kellett J, Yu YQ, Chen W. Application of mobile phase additives to reduce metal-ion mediated adsorption of non-phosphorylated peptides in RPLC/MS-based assays. J Chromatogr B Anal Technol Biomed Life Sci. 2019 Sep 15;1126–1127:121773.

25. Geyer PE, Holdt LM, Teupser D, Mann M. Revisiting biomarker discovery by plasma proteomics. Mol Syst Biol [Internet]. 2017 Sep [cited 2020 Dec 15];13(9):942. Available from: https://pubmed.ncbi.nlm.nih.gov/28951502/

26. Messner C, Demichev V, Bloomfield N, Ivosev G, Wasim F, Zelezniak A, et al. ScanningSWATH enables ultra-fast proteomics using high-flow chromatography and minute-scale gradients [Internet]. bioRxiv. bioRxiv; 2019 [cited 2020 Dec 15]. p. 656793. Available from: https://doi.org/10.1101/656793

Acknowledgments

We thank all members of the Proteomics and Signal Transduction Group at the Max Planck Institute of Biochemistry and the Clinical Proteomics Group of the NNF Center for Protein Research for help and discussions and in particular Igor Paron, Christian Deiml for technical assistance, Mario Oroshi for help with the online resource. We thank the mechanical workshop and the educational workshop especially Martin Wied, Andreas Kucher and Harry Spangenberg of the Max Planck Institute of Biochemistry for the fabrication and iterative optimization of all self-constructed parts.

Funding

The work carried out in this project was partially supported by the Max Planck Society for the Advancement of Science.

Author contributions

J.B.M-R. designed and assembled the packing station parts and carried out the bioinformatics analyses. J.B.M-R., L.S., F.H., P.G., M.M. and P.V.T. designed the experiments, performed and interpreted the MS-based proteomic analyses, generated text and figures and wrote the manuscript. M.M. supervised and guided the project, designed the experiments, interpreted MS-based proteomics data.

Competing interests

The authors declare no competing interests.

3.3. Article 3: A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics

Authors: Nicolai Bache‡,§,*, Philipp E. Geyer¶,ǁ,*, Dorte B. Bekker-Jensenǁ, Ole Hoerning‡, Lasse Falkenby‡, Peter V. Treit¶, Sophia Dollǁ, Igor Paron¶, Johannes B. Müller¶, Florian Meier¶, Jesper V. Olsenǁ, Ole Vorm‡ and Matthias Mann¶,ǁ,§

‡ Evosep Biosystems, Odense, Denmark

¶ Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

ǁ Novo Nordisk Foundation Center for Protein Research, Proteomics Program, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

* These authors contributed equally

§ To whom correspondence should be addressed

This publication marks the dawn of a new era of reproducibility and throughput at the intersection of low micro to nano-flow LCMS. With the Evosep One, we present a chromatography apparatus which combines the benefits of low flow rates for high ionization efficiency and sensitivity of capillary column dimensions with the throughput and robustness of high-flow systems. This is achieved by decoupling the gradient formation from the actual analysis. The former happens in a low-pressure region and results in a preformed gradient in a sample loop before it is pushed over the analytical column under high pressure. This avoids the challenges of gradient formation under high pressure with two pumps as in previous systems and puts less stress on switching valves. Additionally, there is a special sample loading system in the form of a StageTip-like C-18 material on which peptides (or other analytes) are concentrated prior to injection to the LC. This acts like a precolumn or trapping column and reduces the amount of contamination injected into the MS.

These developments make higher throughput analyses possible because overhead times between runs are cut down to a minimum. A throughput of 300 samples a day is possible with a gradient to overhead ratio of less than 2:1. In our hands the most common gradients are the 30 and 60 samples per day methods with gradient to overhead ratios of 11:1 and 7:1.

We demonstrate that the principle of gradient preformation is practically useful and only leads to minimal if any mixing of different phases like water and acetonitrile in the sample loop. The elution of the peptides from the disposable pre-column requires another precaution because the gradient with which the peptides are eluted from the precolumn must be diluted with the aqueous phase before forming the final gradient for the analytical column. In this way peptides are retained at the head of the analytical column and peptide peaks are re-sharpened when entering the analytical column because the

surrounding gradient ratio of aqueous to hydrophobic mobile phase is lower than the ratio required for peptide elution. We exemplified the cause and fixes for these issues in the testing phase of several thousand HeLa runs and demonstrate high retention time stability and low technical variability on 96 plasma proteomic experiments.

In summary, the Evosep One HPLC has its ideal use case in proteomic experiments that need short gradient times and higher reproducibility such as clinical studies with body fluids or measurement of fractionated samples for deep proteome profiling.

A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-Robust Proteomics

Authors

Nicolai Bache, Philipp E. Geyer, Dorte B. Bekker-Jensen, Ole Hoerning, Lasse Falkenby, Peter V. Treit, Sophia Doll, Igor Paron, Johannes B. Müller, Florian Meier, Jesper V. Olsen, Ole Vorm, and Matthias Mann

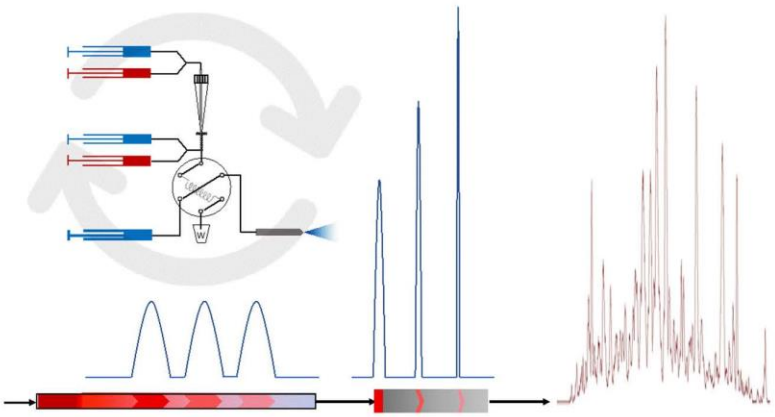
Correspondence

mmann@biochem.mpg.de

In Brief

Because of low throughput and limited robustness, nano-scale liquid chromatography has been a bottleneck for advancing proteomics in biomedical research. Here, we developed and evaluated two new LC concepts—"pre-formed gradients" and "offset gradients for peptide refocusing"—that are both implemented in the Evosep One instrument. We evaluated robustness with more than 2000 HeLa runs, demonstrated absence of cross-contamination with crude plasma samples, high proteome coverage by fractionated HeLa and routinely measuring more than 5000 proteins/sample in just 21 minutes.

Graphical Abstract



Highlights

- Pre-formed and offset gradients for high throughput, robustness and peptide re-focusing.
- Minimal cross-contamination by disposable trap columns and partial elution.
- Single shot DIA measurements achieve >5000 proteins in 21 min.

Technological Innovation and Resources

Author's Choice

A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics*

Nicolai Bache‡‖, Philipp E. Geyer§¶‖, Dorte B. Bekker-Jensen¶, Ole Hoerning‡, Lasse Falkenby‡, Peter V. Treit§, Sophia Doll§, Igor Paron§, Johannes B. Müller§, Florian Meier§, Jesper V. Olsen¶, Ole Vorm‡, and Matthias Mann§¶**

To further integrate mass spectrometry (MS)-based proteomics into biomedical research and especially into clinical settings, high throughput and robustness are essential requirements. They are largely met in high-flow rate chromatographic systems for small molecules but these are not sufficiently sensitive for proteomics applications. Here we describe a new concept that delivers on these requirements while maintaining the sensitivity of current nano-flow LC systems. Low-pressure pumps elute the sample from a disposable trap column, simultaneously forming a chromatographic gradient that is stored in a long storage loop. An auxiliary gradient creates an offset, ensuring the re-focusing of the peptides before the separation on the analytical column by a single high-pressure pump. This simplified design enables robust operation over thousands of sample injections. Furthermore, the steps between injections are performed in parallel, reducing overhead time to a few minutes and allowing analysis of more than 200 samples per day. From fractionated HeLa cell lysates, deep proteomes covering more than 130,000 sequence unique peptides and close to 10,000 proteins were rapidly acquired. Using this data as a library, we demonstrate quantitation of 5200 proteins in only 21 min. Thus, the new system - termed Evosep One - analyzes samples in an extremely robust and high throughput manner, without sacrificing in depth proteomics coverage. ***Molecular & Cellular Proteomics 17: 10.1074/mcp.TIR118.000853, 2284–2296, 2018.***

Bottom-up proteomics is a highly successful and generic technology, which now allows the analysis of complex samples ranging from bacteria through cell line systems and even human tissue samples (1). State-of-the-art workflows begin with a robust sample preparation to digest proteins and harvest purified peptides (2), which are separated by a liquid chromatography (LC)[1] system before they are analyzed by a mass spectrometer (MS). Established software solutions automatically interpret the acquired spectra, generating lists of thousands of quantified proteins (3–8).

The current performance level is a result of improvements not only in the mass spectrometric components but also the chromatographic part of the LC-MS workflow. In the quest for ever increasing chromatographic separation power, columns have become longer and particle sizes smaller - now reaching the sub 2 μm range. This may require pump pressures more than 1000 bar, presenting great engineering challenges for both the pumps and the entire LC system, often limiting robustness in routine operation. Thus, chromatography remains a weak link in MS-based proteomics workflows, leading to calls for new approaches (9). Furthermore, irreproducibility of retention times within and between laboratories severely limits strategies that rely on the transfer of accurate retention times, especially targeted proteomics (10), data independent acquisition (11) and "match between runs" at the MS level (12, 13).

There is great interest in applying the increasing power of MS-based proteomics to diagnostic and clinical questions (14). "Clinical proteomics", however, requires far more stability and reproducibility than that available even in the most advanced MS-based proteomics laboratories. Note that irreproducibility and robustness issues are not features of LC-MS *per se*, as the measurement of small molecules is firmly established in clinical laboratories around the world, which routinely measure hundreds of samples per day. The two key differences of these LC systems to the one applied in proteomics are their much larger column diameters (20-fold) and flow rates (1000-fold), making them much easier to control and less error-prone. Increasing the flow rates to achieve greater robustness has already been advocated in the context

From the ‡Evosep Biosystems, Odense, Denmark; §Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany; ¶Novo Nordisk Foundation Center for Protein Research, Proteomics Program, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Received May 11, 2018, and in revised form, July 13, 2018
Published, MCP Papers in Press, August 13, 2018, DOI 10.1074/mcp.TIR118.000853

of cancer proteomics (15). However, the signal intensity in electrospray ionization is concentration dependent and reducing sensitivity at higher flow rates, which limits these approaches to a few μl/min. Apart from high robustness, throughput is the other central requirement for MS-based proteomics, if it is to enter routine clinical use. Instead, current proteomics workflows generally employ fractionation—multiplying measurement time—or use relatively long gradient times.

In a recent large-scale plasma proteomics study measured in our laboratory, involving more than a thousand samples, 80% of the overall down time was attributable to the HPLC system rather than the MS. At the same time, column equilibration, loading and washing steps between runs limited the attractiveness of very short gradients (16, 17).

Several years ago, some of the current authors devised a very different sample loading and injection approach. Termed speLC, for solid-phase-extraction (nano) liquid chromatography, it was intended for very high sample throughput needed for clinical application (18). The speLC made use of the same StageTips that are commonly employed in proteomics for micro-scale purification of peptides and crude manual fractionation (19–21). Instead of eluting into the autosampler vial of the HPLC system, a low-pressure pump passed a 5–10 min gradient through the StageTip itself and directly toward the MS. The speLC system can analyze 192 *E. coli* samples in only 30 h, as well as identifying more than 500 proteins from a HeLa cell lysate in less than 10 min (18). In subsequent work, speLC was combined with pre-fractionation such as 1D gel electrophoresis or strong cation exchange (SCX), capitalizing on its ability to analyze each of the fractions in 10 min or less (22). Although useful for simple protein mixtures, the low-pressure elution from StageTips and use of only very short analytical columns inherently limited chromatographic separation power of this system.

In the work reported here, we aimed to preserve the benefits of the original speLC device while also achieving the desirable features of modern HPLC instruments. We realized this goal by coupling elution through the StageTips to a novel downstream workflow. In the Evosep One design, peptides are eluted at low pressure and flow rates of tens of μl/min from a special StageTip - termed Evotip™. Notably, the gradient along with the eluted analytes are captured in a long capillary loop. A single high-pressure pump then applies the stored gradient to an analytical nano-scale column. This results in undiminished chromatographic separation performance while eliminating the need to form a gradient at high pressure. Thus, this layout marries the convenience and robustness of large columns, high-flow systems with the sensitivity of narrow column diameters and low-flow rates of nano-LC systems. We further detail the principle of operation and development of the Evosep instrument in detail and investigate its robustness, throughput, and reproducibility in typical applications encountered in MS-based proteomics.

[1] The abbreviation used is: LC, liquid chromatography.

EXPERIMENTAL PROCEDURES

Description of the Liquid Chromatography System—The Evosep One incorporates four low-pressure single stroke piston pumps (A, B, C, and D) and one high-pressure single stroke piston pump (HP) (Fig. 1*A*; supplemental Fig. S1*A*, S1*B*). Together they create a separate low- and high-pressure sub-system. Each pump is equipped with a pressure and flow sensor to monitor and precisely control the flow of the individual solvent. A custom 12-port valve (operating at low-pressure) diverts the flow of the low-pressure pumps either toward the solvent bottles (sol A, sol B) for refilling or toward the system for analysis. The high-pressure pump has a separate 6-port valve (operating at high-pressure) for refilling.

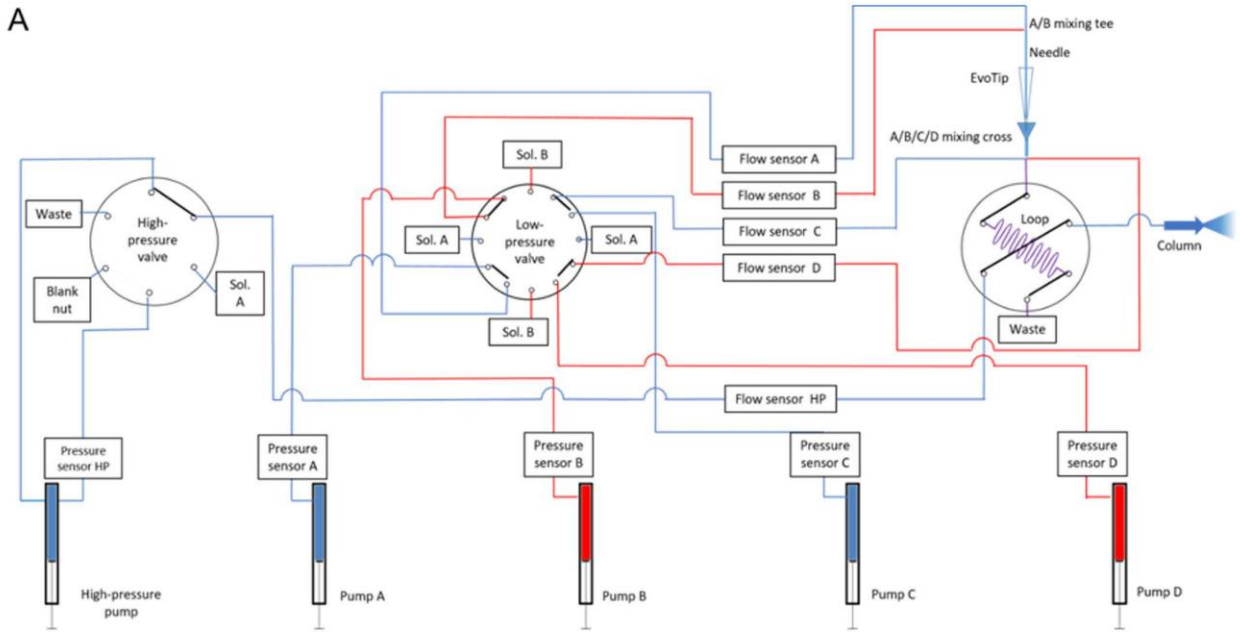
The only common flow path is a storage loop, which is either connected to the low- or high-pressure sub-system and is controlled by a 6-port rotary valve (Fig. 1*A*). In this way, the high-pressure sub-system is always connected to the analytical or separation column but is either in-line or bypasses the storage loop. In contrast, the low-pressure sub-system is always connected to waste but either in-line or bypassing the storage loop. Thus, the storage loop becomes the bridge between the low- and high-pressure sub-systems.

The separate steps are illustrated in the timetable and in the flow path diagrams, highlighting the individual stages of an LC-MS run (Fig. 1*B*, supplemental Fig. S2–S9). At the beginning of a new LC-MS run, the XYZ-axis manipulator of the Evosep One picks up an individual disposable trap column (Evotip) with its ceramic needle and positions it in-line with the solvent flow path at the A/B/C/D mixing cross (Fig. 1*A*, supplemental Fig. S1*A*, S1*B*).

In the second step, pumps A and B then form a primary gradient at the A/B mixing tee that flows through the disposable trap column, eluting the analytes of interest (Fig. 1*B*, supplemental Fig. S2). The organic content of this initial gradient is limited to less than 35% to ensure that only peptides of interest are eluted off the tips while unwanted compounds such as polymers, lipids, and other highly hydrophobic compounds remain bound to the single-use, disposable tips along with any particulate matter from the loaded samples. Furthermore, the final elution volume of this initial gradient is limited to few μl to ensure very precise elution and minimize bleeding of the more hydrophobic molecules. This "partial elution" concept will be further described in RESULTS AND DISCUSSION.

The two additional low-pressure pumps, C and D then modify the eluent at the mixing cross A/B/C/D to create an "offset" to the initial gradient (supplemental Fig. S2). This has the purpose of lowering the organic contents, such that the analytes are initially retained on the analytical column. The offset gradient with the embedded analytes is moved into the storage loop before being switched in-line with the high-pressure pump. In parallel to the first two steps, the high-pressure pump is filled (supplemental Fig. S3) and the analytical column is equilibrated (supplemental Fig. S4). Subsequently, the Evosep One switches the storage loop in-line with the high-pressure pump and the preformed gradient with the embedded analytes is pushed toward the analytical column for high performance separation (supplemental Fig. S5). In parallel to the LC-MS run, the Evosep One is prepared for the next sample by ejecting the disposable trap column, washing the mixing cross A/B/C/D and the ceramic needle, refilling the low-pressure pumps and aligning the solvents of the low-pressure pumps (Fig. 1*B*, supplemental Fig. S6–S9).

The instrument contains procedures to monitor its state during an LC run and can detect high pressure in different parts of the system, warns of potential leaks or the lack of an Evotips in the designated autosampler position. It also has built in trouble shooting procedures



B

Time [min]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Tip pickup and seal																								
Tip elute and gradient formation																								
Refill high-pressure pump																								
Equilibrate analytical column																								
Mixing cross wash																								
Run gradient to analytical column																								
Tip eject																								
Needle wash																								
Refill pump A/B/C/D																								
Align solvent pumps A/B/C/D																								

FIG. 1. **Evosep One flow diagram and time schedule.** *A,* Almost all of the system runs at low pressure (10–20 bar), increasing the lifetime and robustness of the LC. Only a single pump and flow path operates at high pressure and this does not involve any solvent mixing. *B,* Stepwise timetable including all steps that the Evosep One is performing during a LC-MS run for the 60 samples/day method. The activities for the autosampler, the low-pressure pumps and the high-pressure pump are color-coded in green, yellow and blue, respectively. For a detailed flow diagram with highlighted flow path states see supplemental Fig. S2–S9.

through the Chronos software user interface. Moreover, the Evosep One carries out preparatory actions before a sample run.

Cell Culture—HeLa cells were cultured in high glucose DMEM with 10% fetal bovine serum and 1% penicillin-streptomycin (Life Technologies, Inc.). Cells were counted using an Invitrogen countess cell counter and stored after snap freezing at −80 °C.

Tryptophan Fluorescence Emission Assay for Protein Quantification—Protein concentrations were determined in 8 M urea by tryptophan fluorescence emission at 350 nm, using an excitation wavelength of 295 nm. Tryptophan at a concentration of 0.1 μg/μl in 8 M urea was used to establish a standard calibration curve (0–4 μl). We estimated that 0.1 μg/μl tryptophan are equivalent to the emission of 7 μg/μl of human protein extract, if tryptophan on average accounts for 1.3% of human protein amino acid composition.

Protein Digestion—For sample preparation we used the iST kit for proteomic samples (2), starting with 10^6 HeLa cells according to the manufacturer's instructions (P.O. 00001, PreOmics GmbH).

Robustness Optimization—To test and optimize robustness, we injected and analyzed over 2000 times tryptic peptides of HeLa cells, initially in exploratory batches. For this experiment, we used the breadboard model of the Evosep One coupled to a LTQ Orbitrap instrument. All issues were protocolled, and the system was optimized during the test and in the exploratory phase, the instrument was only stopped for the optimization of hardware and software components. The last 1500 HeLa samples were analyzed on a single column to analyze variation in the system and the wear of the column.

Plasma Proteomics—Blood was taken by venipuncture using a commercially available winged infusion set and collection tubes containing EDTA and centrifuged for 15 min at 2000 × *g* to harvest plasma. Blood was sampled from a healthy donor, who provided written informed consent, with prior approval of the ethics committee of the Max Planck Society. The plasma was distributed into a 96-well plate and subsequently processed with an automated sam-

Pre-formed Gradient Liquid Chromatography

ple preparation for Plasma Proteome Profiling as described previously (16).

High Throughput of Low Complexity Samples—the "UPS1 Proteomic Standard" (Sigma-Aldrich) was digested as indicated above using the PreOmics iST kit and the peptides were analyzed with the 200 samples/day method (5.6 min gradient) with a 2 μl/min flow on a 5 cm C18 column (3 μm particle size).

Prefractionation—Peptides for deep proteome analysis were fractionated using a reversed-phase Acquity CSH C18 1.7 μm 1 $\times$ 150 mm column (Waters, Milford, MA) on an Ultimate 3000 high-pressure liquid chromatography (HPLC) system (Dionex, Sunnyvale, CA) operating at 30 μl/min. Buffer A (5 mM ammonium bicarbonate) and buffer B (100% ACN) were used. Peptides were separated by a linear gradient from 5% B to 35% B in 55 min, followed by a linear increase to 70% B in 8 min. In total, 46 fractions were collected without concatenation. For nano-flow LC-MS/MS, the loading amount was kept constant at 500 ng per injection for the Easy-nLC 1200, while 500 ng from each fraction was loaded on an Evotip.

UV Gradient Storage Experiment—To assess the effect of diffusion as a function of storage time in a storage loop, we built a test rig to mimic Evosep One operation as illustrated in figure 2. A set of Zirconium nano pumps (Prolab Instruments, GmbH, Switzerland, pump A: 0.1% formic acid (FA) in H_2O, Pump B: 0.1% FA, 1% acetone in acetonitrile) were programed to create the following composition profile: 0–5 min 5% B, 5–10 min 5–95% B, 10–13 min 95% B, 13–15 min 95–5%, 15–18 min 5%, 18–23 min 5–95% B, 23–25 min 95% B, 25–27 min 95–5% B, 27–30 min 5% B. This was delivered into a coiled (diameter 10 cm) fused silica storage loop (length 7 m, i.d. 100 μm, OD375, Polymicro Technologies). After a specified storage time had passed, a third Zirconium pump pushed the content out of the loop at a flow rate of 2 μl/min toward a UV detector (SpectraFlow 501, SunChrom) equipped with a nano-flow cell (5 nl) set to record the absorption at 265 nm. The storage loop and the three pumps were all connected to a standard 6-port Vici valve (Valco Instruments Co. Inc.) to control the flow path using a script.

Evaluation of Chromatographic Performance—We loaded 250 ng of a commercial HeLa digest (Pierce, no. 1862824) spiked in with 100 fmol of PicoSure Test Standard (eight synthetic peptide mix, New Objective PS-STDN) and loaded the mix on Evotips. For each of the five gradient methods, four replicates were analyzed using a Thermo Q Exactive set to acquire full scans (resolution 35k) and targeted MS2 (resolution 17.5k) of the eight synthetic peptides in a scheduled table (Fig. 6*B*). Skyline was used to extract between 4 and 6 MS2 ions (parallel reaction monitoring) for each of the 8 peptides (23). Chromatographic profiles were exported from Skyline and peak characteristics for each peak was extracted using a script.

Loading of Evotips—Tips were activated with consecutive 100 μl wash steps of 100% ACN, 50% ACN in 0.5% formic acid in H_2O followed by two times 0.5% formic acid in H_2O. BSA or HeLa peptides were loaded in 0.5% formic acid in H_2O. The tip activation protocol was later optimized to use 1-propanol for wetting the C18 material prior to equilibration.

High-pressure Liquid Chromatography and Mass Spectrometry—LC-MS instrumentation consisted of a breadboard Evosep One coupled to an LTQ Orbitrap for the more than 2000 HeLa injection experiment, and the Evosep One production version coupled to an Q Exactive HF-X Orbitrap (Thermo Fisher Scientific) for all other experiments. Purified peptides were separated on the HPLC columns with 3 μm Reprosil-Pur C18 beads (Dr. Maisch, Ammerbuch, Germany) and dimensions indicated below in Fig. 6*B*. On the LTQ Orbitrap MS, data were acquired with a Top6 data dependent shotgun method and with a Top12 method for the Q Exactive HF-X instrument. On the Q Exactive HF-X Orbitrap, the target value for the full scan MS spectra was 3 $\times$ 10^6 charges in the 300–1650 *m/z* range with a maximum injection time of 50 ms and a resolution of 60,000 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 (24). MS/MS scans were performed at a resolution of 15,000 at *m/z* 200 with an ion target value of 5 $\times$ 10^4 and a maximum injection time of 25 ms. Dynamic exclusion was set to 15 s to avoid repeated sequencing of identical peptides.

Deep Proteome and DIA Experiments—HeLa cells were harvested at ~80% confluence by washing twice with PBS and subsequently adding boiling lysis buffer (6 M guanidinium hydrochloride (GndCl), 5 mM tris(2-carboxyethyl)phosphine, 10 mM chloroacetamide, 100 mM Tris pH 8.5) directly to the plate. The cell lysate was collected by scraping the plate and boiled for an additional 10 min, followed by micro tip probe sonication (Vibra-Cell VCX130, Sonics, Newton, CT) for 2 min with pulses of 1 s on and 1 s off at 80% amplitude. Protein concentration was estimated by Bradford assay, and the lysate was digested with LysC (Wako) in an enzyme/protein ratio of 1:100 (w/w) for 1 h, followed by dilution with 25 mM Tris, pH 8.5, to 2 M GndCl and further digested overnight with trypsin (1:100 w/w). Protease activity was quenched by acidification with trifluoroacetic acid (TFA) to a final concentration of ~1%, and the resulting peptide mixture was concentrated on Sep-Pak (C18 Classic Cartridge, Waters, Milford, MA). Elution was done with 2 ml of 40% acetonitrile (ACN), followed by 2 ml of 60% ACN. The eluates were combined and volume reduced by SpeedVac (Eppendorf, Germany), and the final peptide concentration was estimated by measuring absorbance at 280 nm on a NanoDrop spectrophotometer (NanoDrop 2000C, Thermo Fisher Scientific, Germany). For DIA samples, iRT peptides (Biognosys AB, Schlieren, Switzerland) were added prior to MS analysis according to the manufacturer's protocol. For samples analyzed on the Evosep One, an in-house packed 12 cm, 150 μm i.d. capillary column with 1.9 μm Reprosil-Pur C18 beads (Dr. Maisch, Ammerbuch, Germany) was used, while samples analyzed on the Easy-nLC 1200 were separated in an in-house packed 15 cm, 75 μm i.d. capillary column with the specifications as described above. The column temperature was maintained at 40 °C using an integrated column oven (PRSO-V1, Sonation, Biberach, Germany) and interfaced online with the mass spectrometer.

Data Analysis—MS raw files were analyzed by the MaxQuant software (version 1.5.6.8) (3) and fragments lists were searched against the human Uniprot Reference Proteome without isoforms (April 2017 release with 21,042 protein sequences) by the Andromeda search engine (25) with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidations as variable modifications. The experiment for the 200 samples/day method was analyzed with the UPS1 FASTA file, downloaded from the homepage of Sigma-Aldrich (April 2018). We set the false discovery rate (FDR) to 0.01 at the peptide and protein levels and specified a minimum length of 7 amino acids for peptides. 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. An initial precursor mass deviation up to 7 ppm and a fragment mass deviation of 20 ppm were specified.

Data independent analysis (DIA) results were processed with Spectronaut version 11.0.15038.19.19667, using default settings (Biognosys, Zurich, Switzerland). A project specific spectral library was imported from the separate MaxQuant analysis of the combined analysis of the 46 pre-fractionated HeLa fractions, and DIA files were analyzed using default settings. Information about precursors, peptides and proteins identified by the Spectronaut software are available in Supplemental Table S1 and S2.

All bioinformatics analyses were done with the Perseus software (26) of the MaxQuant computational platform.

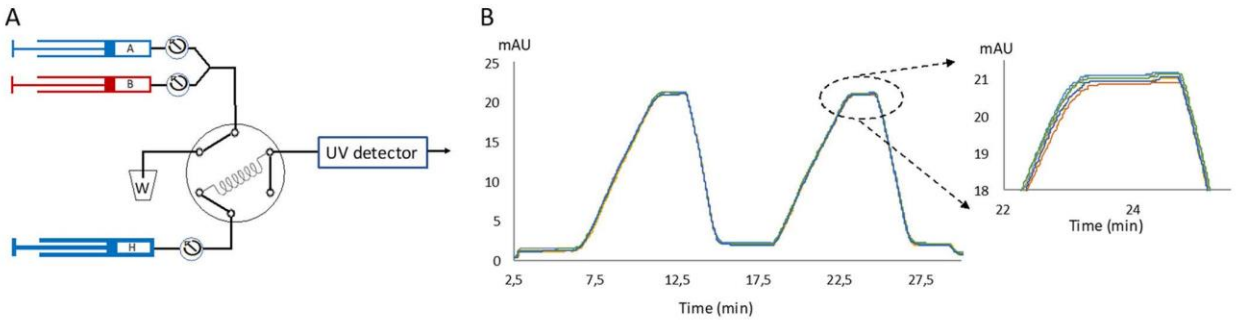


FIG. 2. **UV set up to test gradient storage.** *A,* Flow diagram for testing potential gradient mixing during storage in the capillary loop. *B,* Profiles of the acetonitrile and water plugs that were recorded by the UV detector for different storage times. Profiles were almost completely superimposed, consistent with minimal mixing of the two phases during storage.

RESULTS AND DISCUSSION

Principle of Analyte Embedding in Pre-formed Gradients—Our key idea in making nano-LC as robust as high-flow LC was to decouple gradient formation from the high resolution, high-pressure separation on an analytical column. As in established peptide purification strategies, the peptides are first loaded on Evotips (a form of solid phase extraction tips like StageTips (20)). However, instead of eluting the peptides from the tips, drying them to remove the organic content and re-suspending them in injection buffer, we directly elute from the Evotip into the capillary loop. This is accomplished at pressures of only a few bar by two syringe pumps A and B at flow rates of 10 to 20 μl/min (Fig. 1). Note that an entire gradient can be stored in a several meters long fused silica capillary - already containing the individual peptides at the organic content where they elute from the C18 material. For instance, a 4 m long capillary of 100 μm inner diameter (i.d.) has a volume of 31.5 μl, enough for a subsequent analytical column separation of 31.5 min at 1 μl/min or 90 min at 350 nl/min.

We first asked if the gradient would be affected over time in the storage loop due to diffusion (27). Considering the very high aspect ratio of column length compared with i.d. (40,000 in the example above), this appears to be unlikely. Further, in a similar capillary storage scheme in the RePlay system we did not observe such mixing (28). To experimentally investigate this question, we placed defined plugs of ACN/1% acetone and water in the capillary loop, stored them for 0 or 60 min and monitored them with a UV detector (Fig. 2*A*, EXPERIMENTAL PROCEDURES). This did not lead to detectable mixing (Fig. 2*B*), confirming that storage of pre-formed gradients in a capillary loop is suitable for our purposes.

Having established that an analyte-containing gradient can be formed easily and stored in a loop, the next challenge was to obtain high chromatographic resolution with the help of an analytical column. A common issue in pre-column setups is peak broadening because peptides eluting from the pre-column are not sufficiently retained on the analytical column. To solve this issue, and to take account of the relatively large elution volume from the Evotip, we designed a gradient offset strategy. Once the Evotip is sealed in-line with the solvent system, a gradient from pumps A and B subsequently elutes the peptides from the tip. Directly after the Evotip, a secondary gradient from pumps C and D modifies the composition of the initial gradient and thus, reduces the effective organic content (Fig. 3*A*, 3*B*). With the offset gradient, peptides eluting from the loop are shortly retained at the head of the column and thereby focused (Fig. 3*C*). After separating on the analytical column, this results in the highest possible peak capacity. Note that due to the pre-formed and offset gradient the analytes are effectively loaded on the column in a sequential manner. Consequently, only a few percent of total peptide load is on the column at any given time (for instance, with a loop of 30 μl, a maximum of 3% for a 12 cm, 75 μm i.d. column which has a bed volume of less than one 1 μl).

After generation of the gradient, the loop-valve switches the storage loop in-line with the high-pressure pump and the analytical column (Fig. 3*A*). The high-pressure pump then pushes the pre-formed and offset gradient with embedded, pre-separated peptides over the analytical column. The fact that almost all the system's functionality is contained in the low-pressure sub-system (Fig. 1), should ensure long lifetime of the mechanical components, and opens for ultra-precise flow manipulation, at a low risk of critical leaks and malfunction.

To test the Evosep One separation scheme, we loaded a BSA digest on an Evotip and eluted it in a 21 min gradient from an 8 cm analytical column (100 μm i.d., 3 μm C18 beads). This resulted in low peak widths (4.8 s median FWHM) and corresponding column capacities. Multiple injections illustrate that the chromatograms are virtually superimposable (Fig. 3*D*). An interesting consequence of our design is that it almost eliminates the loading and washing steps that are otherwise necessary between injections. Instead, the washing step is also encoded in the loop composition, and all remaining procedures take less than 3 min. This brings the total analysis time (injection to injection) very close to actual gradient time (21 min + 3 min). (Note that the instrument further-

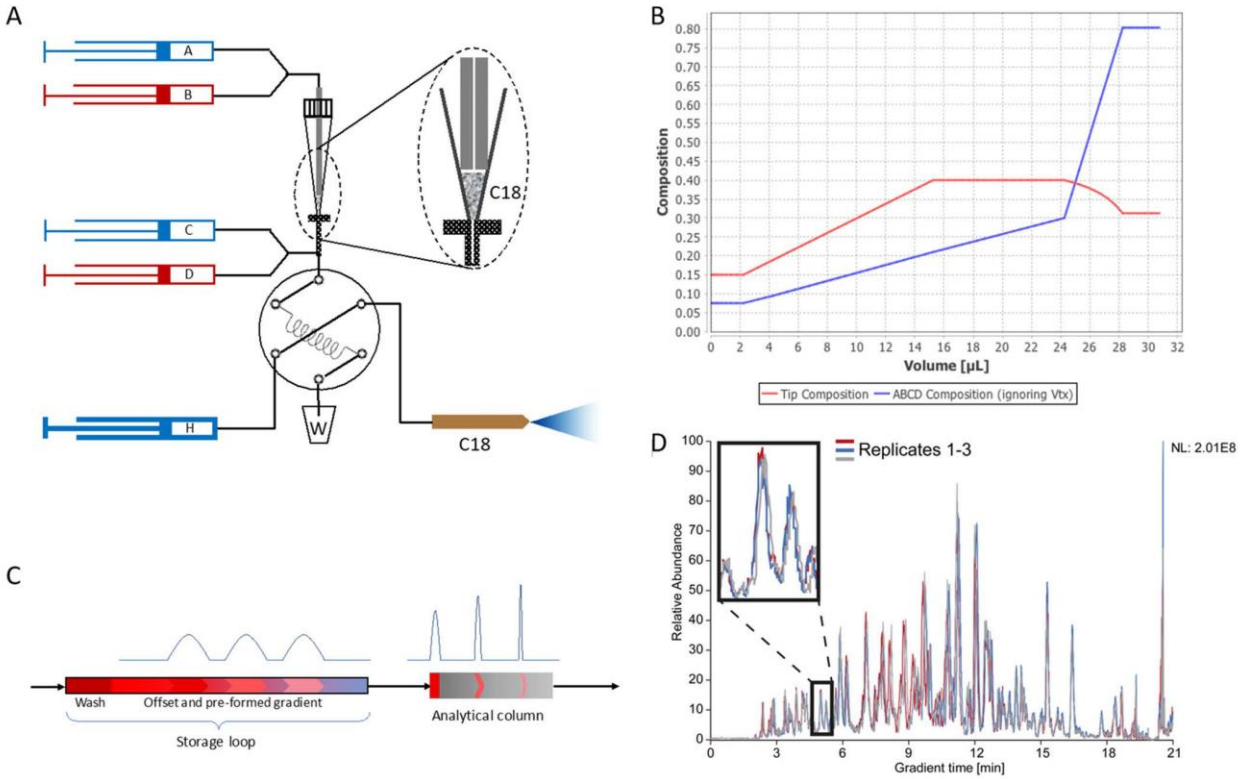


FIG. 3. **Pre-formed gradient.** *A,* Peptides are eluted from the C18 containing Evotip by pumps A and B. Low pressure pumps C and D form the final gradient, which is stored in the capillary loop together with the analytes. Subsequently, the valve switches and the high-pressure pump (H) simply pushes the gradient with its peptides over the analytical column. *B,* Composition of the gradient resulting from the confluence of the flows from pumps A, B and pumps C, D (*x* axis designates the volume entering the storage loop). The proportion of acetonitrile is indicated on the *y* axis. *C,* Analytes embedded in the storage loop are represented in red and as peak intensities. Because of the offset provided by pumps C and D, peptides are shortly retained at the head of the analytical column and elute with narrow peak widths. *D,* Comparison of three base peak chromatograms from a HeLa digest, demonstrating the reproducibility.

more allows a higher flowrate at the beginning of the gradient, which would further compress the time to appearance of the first peptides in the gradient.) Compared with conventional designs, this dramatically increases throughput, especially for short gradients, while avoiding the complexity and reproducibility issues of double column designs (29).

Robustness Development and Stress Test—Having established the basic principles of operation, we constructed a breadboard model that incorporates all functional components. As far as possible, we chose industry leading standard components, such as the CTC Analytics auto sampler and Vici rotary valves, whereas other components were custom designed for our throughput and robustness requirements (EXPERIMENTAL PROCEDURES). Pump firmware development was done in house but for other software development, we used the Chronos environment, an industry standard and widely used platform, with a view to integrate our instrument with the different MS manufacturers.

To fine-tune operation and optimize robustness, we injected 1 μg of a tryptic HeLa cell digest over 2000 times in a consecutive manner. We logged all issues over time and stopped the test only to optimize hardware or software components. In total, 35% of the measurements within the first 250 samples suffered from sample loss due to an imperfect seal of the autosampler needle and the tip. In a first step, we optimized the needle, which resulted in an immediate reduction of errors. After changing the seal between the tip and the entrance of the flow path as well, these issues were eliminated (Figs. 3*A*, 4*A*). From injection 513 on, all instrument related issues appeared to be resolved. We then mounted a new column to test the "partial elution" concept (as described in EXPERIMENTAL PROCEDURES) in subsequent injections. Over these 1500 samples the total ion current remained unchanged until the end of the experiment (Fig. 4*B*). A few LC-MS runs were blank, but this turned out to be due to incorrect manual loading of the corresponding Evotips.

We also recorded the pressure profiles for all runs. Validating the partial elution concept, there was only a very slight increase in backpressure, indicating that the column had remained free of deposits and as further evidence of the effect,

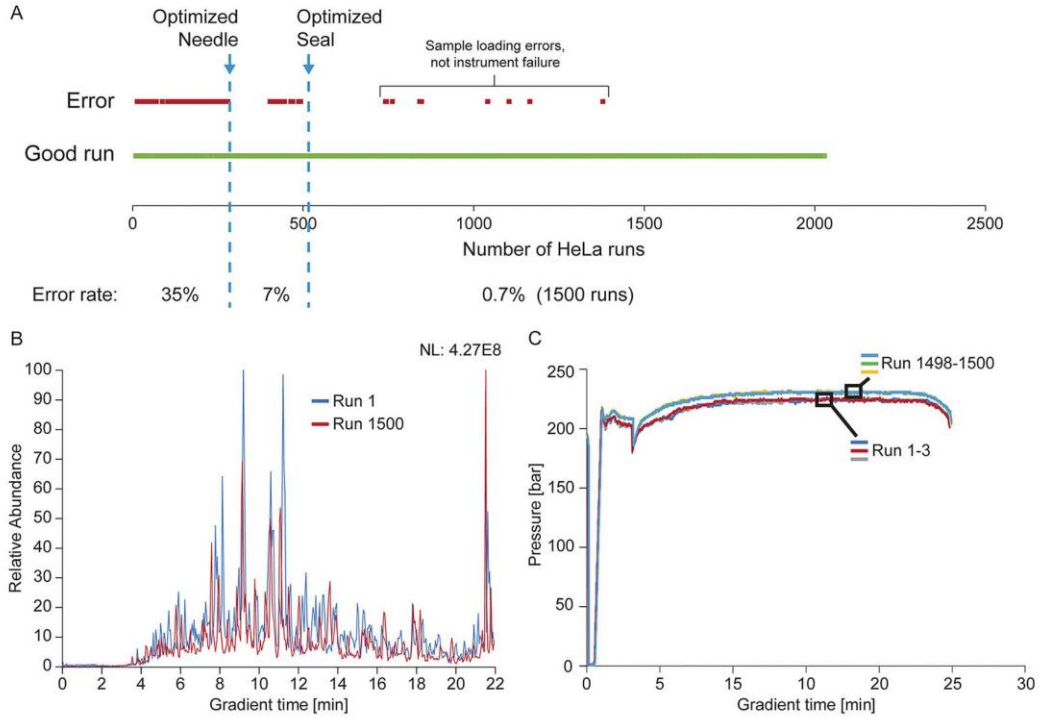


FIG. 4. **Robustness evaluation.** *A,* Error frequency during the development phase of the system assessed by consecutive measurements of HeLa digests. *B,* The first and last base peak chromatogram of a HeLa digest in a series of 1500 measurements using a 22 min gradient. *C*, Pressure profiles over the gradient for the first and last three HeLa digests of the same experiment.

the TICs of runs 1 and 1500 were indeed highly similar and showing no decay in separation performance of the column. Pressure profiles of adjacent runs were virtually indistinguishable (Fig. 4*C*).

The Evosep One was intended and constructed for high throughput applications, with a focus on clinical analysis. Blood plasma is the most widely analyzed clinical matrix, with millions of samples drawn daily. Yet it is difficult to analyze plasma robustly by nano-LC/MS, mainly because of the large number of non-protein blood components. To demonstrate clinical applicability of the system, we employed our automated sample preparation pipeline - termed Plasma Proteome Profiling (Geyer *et al.* 2016a). Plasma samples were prepared and loaded on the Evotips in a 96 well format, using a robotic platform. The total measurement time for the 96 samples on the Evosep One was less than 2 days, corresponding to a throughput of 60 samples per day. Reproducibility over all 96 independent, parallel sample preparations and injections of the same original plasma was excellent (median Pearson correlation coefficient of 0.98) over all runs (Fig. 5*B*). For clinical decision making based on the concentration of biomarkers, it is crucial to ensure low carry-over from one analysis to the next. Therefore, we performed a cross contamination experiment with six alternating injections of plasma and blanks (Fig. 5*B*). The average carry-over was as low as 0.07% and 80% of this can be traced back to just 20 peptides (Fig. 5*C*).

Design of Methods for Desired Throughput and Depth—Based on the principles explained above and the experiences from the robustness testing on the breadboard model, we then constructed the production unit. We devised a number of standard gradients and column combinations tailored to different applications, ranging from high throughput quality control of low complexity samples, through comprehensive proteomics using fractionation, to the in depth single run characterization of complex proteomes. The short gradients made possible by the Evosep system can be used for low complex samples and the somewhat longer ones for more complex samples.

Note that the design choices embodied in the Evosep One also imply certain limitations, at least in the current version. In common with previous efforts in “industrialized proteomics”, we chose to prioritize reliability, robustness and throughput over certain other parameters. The choice of relatively short and somewhat larger i.d. columns together with a flowrate of 1 μl/min, does not maximize sensitivity (however, this can easily be adjusted by the user). Likewise, sample introduction through the Evotip currently results in an elution volume optimized for gradients up to 44 min, whereas longer gradients would lead to broader peaks.

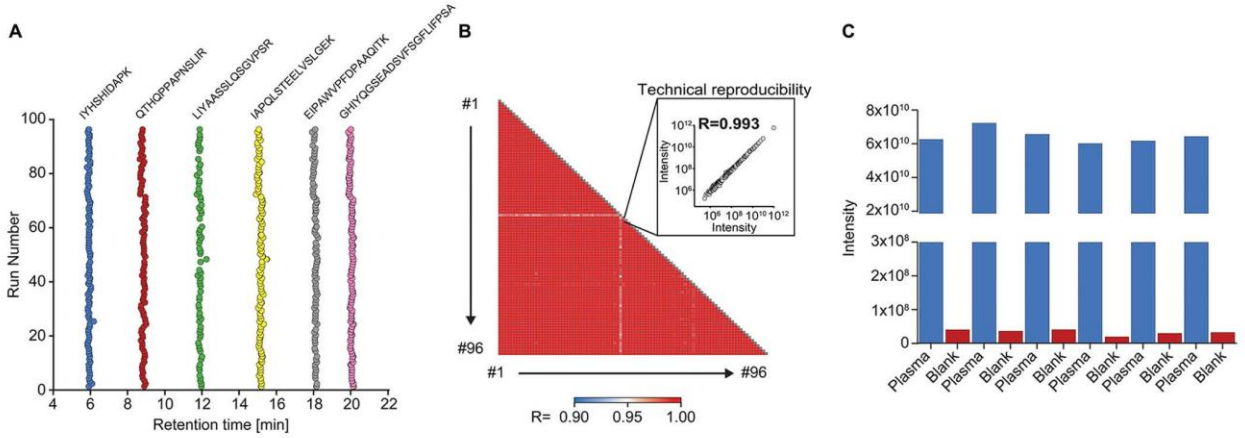


FIG. 5. **Clinical applicability to the plasma proteome.** *A,* Retention time stability of selected peptides spanning a range of elution times over 96 plasma proteome runs. *B,* Pearson correlation matrix comparing all 96 plasma runs to each other. A single correlation graph with the median Pearson value is shown in the inset. *C,* Summed total peptide intensities in alternating plasma and blank runs.

We characterized the chromatographic performance of each method using a synthetic peptide mix, spiked into the complex background of a HeLa digest. Parallel reaction monitoring targeting only the synthetic peptides extracted a detailed elution profile representative of typical proteomic measurements. From this data we calculated peak and retention time properties for the eluting peptides (Fig. 6*A*; supplemental Fig. S10–S14). Fig. 6*B* shows these data in tabular form for the optimized gradients and column dimensions for the standard use cases and sample types.

We first wished to demonstrate the possible throughput on low complexity samples. We digested the "UPS1 Proteomic Standard" (EXPERIMENTAL PROCEDURES) and used the 5.6 min gradient with the 2 μl/min flow on the 5 cm column (200 samples/day method). In a single day, this resulted in 200 data sets with very consistent protein coverage (Fig. 6*C*). The UPS1 should contain 48 different proteins but curiously four of them were never identified. As this standard is equimolar this is not an issue of dynamic range. Furthermore, the remaining 44 proteins were quantified essentially completely in all runs (average of 43.5 $\pm$ 1) (Fig. 6*C*). We conclude that the remaining proteins were likely missing from the kit. The high throughput for low complexity samples would be very interesting for single protein identification experiments in gel bands, for instance, or for contaminant analysis in recombinant protein expression in biotechnology. In many cases, it could also be enough for somewhat more complex mixtures such as those resulting from pull-down experiments.

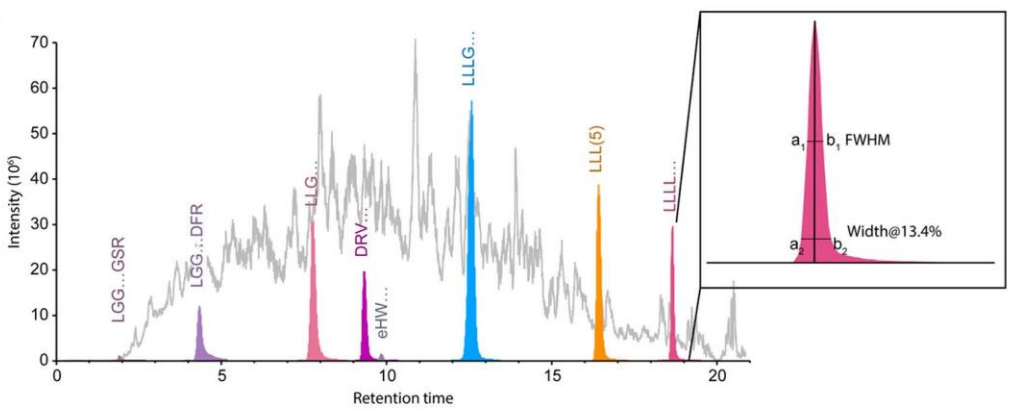
Rapid Generation of In-depth Mammalian Cell Line Proteomes—Having shown the applicability of the system for low complexity samples in high throughput, we next investigated the rapid characterization of fractionated, high complexity proteomics samples. A fractionation step is very common in the analysis of cell line or tissue proteomes, but usually comes with the caveat of a drastic increase in measuring time as the number of factions increases.

We built on a recently described strategy that combined extensive high pH reversed-phase peptide pre-separation in a first HPLC dimension without "concatenation" of the resulting fractions and relatively short gradients (4). Up to 70 such fractions were analyzed in gradients of 30 min, allowing for overall high peptide loading and high combined peak capacity and making optimal use of the high acquisition speed of state-of-the-art mass spectrometers (30). This resulted in a very deep coverage of cell line and tissue proteomes, on par with RNA-seq results (4). A bottleneck of the workflow was the low utilization of the mass spectrometer, due to the washing, equilibration and loading times of the HPLC, which are minimized with the Evosep system.

To characterize the efficiency for fractionated proteomes and to compare this to the Easy-nLC 1200 used as a standard in our laboratories as well as in the study described above, we performed an analysis of 46 HeLa fractions on both systems. Each of the fractions was divided and separately measured on the Easy-nLC and the Evosep One on the same MS instrument, recording total instrument time, the time utilized for gradients and the numbers of peptides and proteins identified. The Easy-nLC 1200 was run with our previously optimized 15 min gradients, whereas we used the 21 min gradient of the 60 proteomes/day method for the Evosep One.

As expected because of the short overhead time between runs, the Evosep One was significantly more efficient in terms of utilization of the mass spectrometer. A full 88% of the total analysis time of 18.4 h was spent on data acquisition (Fig. 7*A*). In contrast, the Easy-nLC 1200 occupied the mass spectrometer for 28.3 h, but only 14.6 h (52%) were productively used. This difference did not come at the expense of the numbers of identified peptides and proteins, which was very comparable with 132,850 peptides (9918 proteins) and 130,450 peptides (9603 proteins) for the Evosep One and the Easy-nLC 1200, respectively. A detailed view of peptides identified in each

A



B

Throughput	Cycle time	Gradient length	Overhead	Flow rate	Column length / ID	FWHM	Peak capacity
Samples / day	[min]	[min]	[min]	[µl / min]	[cm / µm]	[sec]	4σ / FWHM
300	4.8	3.2	1.6	3.0	4 / 150	1.5	35 / 72
200	7.2	5.6	1.6	2.0	4 / 150	2.2	41 / 97
100	14.4	11.5	2.9	1.5	8 / 100	4.0	63 / 129
60	24.0	21.0	3.0	1.0	8 / 100	6.8	79 / 161
30	48.0	44.0	4.0	0.5	15 / 100	15.2	77 / 151

C

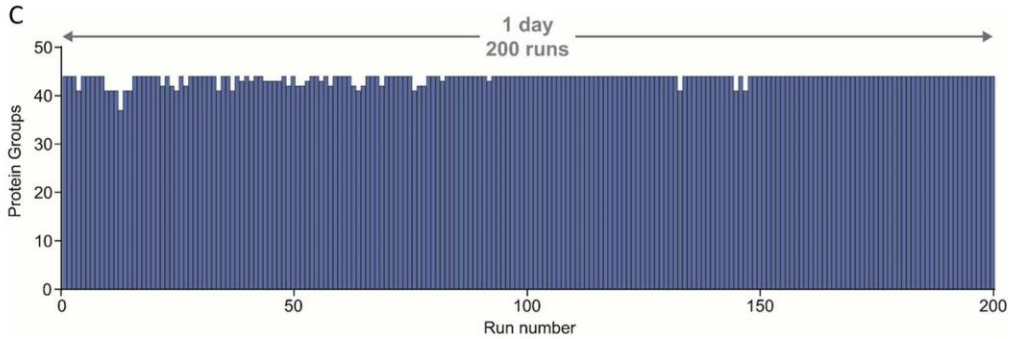


FIG. 6. **Evosep One methods and chromatographic performance.** *A,* Extracted peaks of synthetic peptides (colored) in a HeLa background (gray). The inset illustrates the extracted peak properties. *B,* For ease of use, five optimized methods have been pre-set to provide the best performance to time compromise. They are defined by the total number of samples that can be run per day rather than referring to the length of the gradient. The peak width and peak capacity values are averages on a HeLa digest with spiked in synthetic peptides (for details see supplemental Fig. S10–S14). *C,* Technical replicates of a digest of the UPS1 Proteomic Standard were injected 200 times with the 200 samples/day method. The number of identified proteins for each sample is shown as a bar graph in chronological order.

fraction separately or cumulatively, showed that they are very similar (Fig. 7*B*, 7*C*). This confirms our conclusion that the design principle of the Evosep One resulted in saving substantial measurement time (35% in this case), at undiminished performance. For longer gradients, the proportional time savings would be lower, however, given the high price of modern mass spectrometers, they would still be economically attractive. The above experiments show that the Evosep is well suited for the in-depth characterization of proteomes via the rapid analysis of the high pH or other fractions that are commonly used in proteomics. While we employed label-free quantitation here, the results should equally apply to isobaric labeling strategies. We also note that an average of 2700 proteins were identified in these fractions. There are several proteomics strategies that produce many fractions, such as thermal shift assays (31) or organellar proteomics (32), and our approach opens up for strategies to rapidly and robustly measure these.

Single Shot, High Throughput HeLa Proteomes Using DIA—The experiments described so far used data dependent ac-

A

LC system	Total time	Gradient time	Peptides	Proteins
Evosep One	18.4 h	16.1 h	132,850	9918
Easy-nLC 1200	28.3 h	14.6 h	130,450	9603

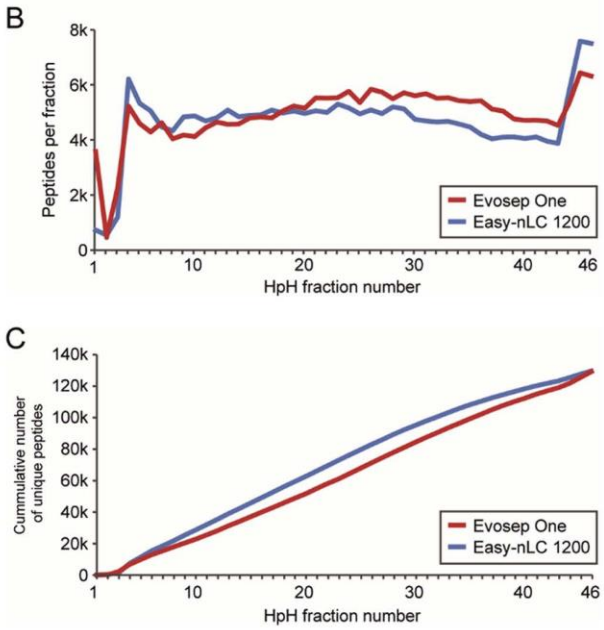


FIG. 7. **Rapid generation of mammalian cell line proteomes.** *A,* Table for the comparison of the Evosep One with the Easy-nLC 1200, including total measurement time, gradient time and the numbers for identified proteins and peptides. *B,* Numbers of identified peptides per fraction over the 46 high pH reversed-phase fractions for both LC systems. *C,* Cumulative numbers of unique peptides across the fractions.

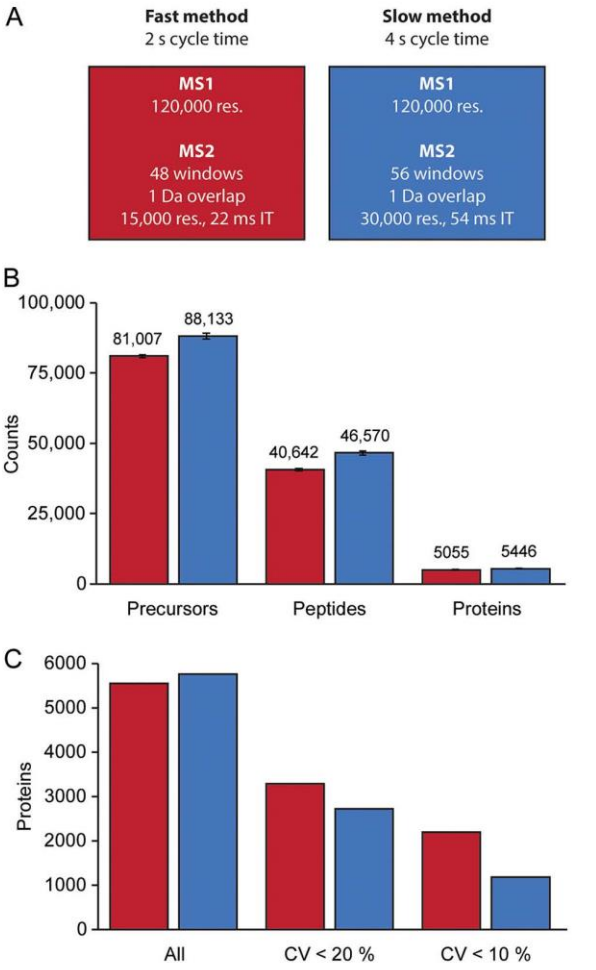


FIG. 8. **Rapid generation of mammalian cell line proteomes.** *A,* Two scan modes for the acquisition of DIA data were devised and tested. *B,* Average number of precursors, identified peptides and protein groups for five HeLa measurements with 21 min gradients on the Evosep One. *C,* Number of proteins quantified with a coefficient of variation (CV) below 20 and 10%.

quisition (DDA). However, data independent acquisition (DIA) is becoming increasingly popular and competitive (7). In our hands, we have found DIA to perform particularly well with relatively short gradients on fast and high resolution Orbitrap analyzers (5). The Evosep One with its fast turn-around between runs appeared to be a good addition to this strategy and we were curious to see how deep the proteome could be covered with such a combination. For this purpose, we made use of the very extensive peptide library generated in our previous experiments of the 46 fractions of HeLa digests using the Spectronaut software with one percent FDR at both precursor and protein levels.

Especially in short gradients, there is a trade-off between the number of peptide identifications and the quantification accuracy because of the finite time for a DIA cycle. To investigate this, we designed a faster (2 s cycle time, 15k MS/MS resolution) and a slower scanning method (4 s cycle time, 30k MS/MS resolution) as visualized in Fig. 8*A*. Given the short 21 min gradients (60 samples per day) the proteome coverage was very high for both methods with more than 5000 quantified proteins from more than 40,000 matched peptides. This equates to 250 unique proteins per gradient minute throughout the gradient. As expected, the slower method was somewhat superior in terms of identifications with a higher number of precursors (88,133 *versus* 81,007), peptide identifications (46,570 *versus* 40,642) and protein groups (5446 *versus* 5055) (Fig. 8*B*). For the fast and the slow method, the overlap of proteins between replicates was 81% and 85% with 4491 and 4904 proteins found in all five measurements, respectively (supplemental Fig. S15*A*). The faster method performed better with regard to protein quantification with 3286 proteins with a CV less than 20% in the slower *versus* 2724 in the faster method, respectively (Fig. 8*C*, supplemental Fig. S15*B*). For the top 70% of the proteome by abundance, data completeness was close to 100% (supplemental Fig. S15*C*). These results indicate that the short gradients enabled by

the Evosep One can very efficiently be combined with DIA for high-throughput and in-depth acquisition of proteomic data.

CONCLUSION

Despite the great technological advances in high sensitivity nano-flow MS-based proteomics, the robustness and throughput have been weak links even in state of the art MS-based proteomic workflows. This has led to a move toward microflow systems—especially with a view toward clinical applications—however, at the cost of sensitivity (33). Here, we have introduced an entirely novel concept based on the pre-formation of gradients at relatively high-flow and low-pressure. This pre-stored gradient already has the analytes embedded and is moved across a high-resolution column by a single, high-pressure pump. Based on these principles, we first designed a breadboard system that was progressively developed into a commercial HPLC system - the Evosep One. We established that pre-storing of the gradient, followed by "re-focusing" of the peaks at the head of the analytical column, assures full chromatographic peak capacity of the overall system. Together with the Evotip as a disposable sample clean up cartridge, the system is designed for sensitivity, throughput, and robustness - tailor made for large clinical studies. To test this, we performed thousands of runs with cell lysates as well as complex clinical samples such as blood plasma. We found that the decoupling of gradient formation with a low-pressure system and the high-pressure peptide separation ensured stable and uninterrupted operation without instrument related issues or deterioration in chromatographic performance. As expected from its design, the Evosep One proved to have minimal or absent cross contamination and very high consistency of label-free quantitation results across injections.

The time required for the formation of the pre-stored gradient, including the washing step, happens within 2–3 min, reducing the idle time of the mass spectrometer between injections. This opens up for the rapid analysis of samples of medium complexity, as we demonstrated with the measurement of 200 standard mixtures (UPS1) in a single day. The short gradients on the Evosep One are especially attractive in combination with time-of-flight (TOF) instruments because of their very high scanning speed. This was recently demonstrated by the identification of more than 1000 HeLa proteins in only 5.6 min (200 samples/day method) (34). Deep proteomes are typically achieved after extensive fractionation. In this context, the fast turn-around of the Evosep One ensures very high utilization of the MS instrumentation as we show by the analysis of 46 HeLa fractions in 18 h. Finally, we used a state-of-the-art data independent workflow that enabled a remarkable proteome depth of 5200 proteins in only 21 min (60 samples/day method). With ongoing developments on the mass spectrometric side, the proteome coverage is likely to improve further.

The minimal run-to-run times make even very short gradients efficient and attractive, opening up for high-throughput proteomics in areas like screening of host-cell proteins in pharma research, protein interaction studies and in particular clinical proteomics. We further imagine applications in top-down proteomics and in small molecule analysis, in particular metabolomes.

Acknowledgments—We thank all members of the department of Proteomics and Signal Transduction at the Max Planck Institute of Biochemistry in Martinsried for help and discussions, and Gaby Sowa for technical assistance.

DATA AND MATERIAL AVAILABILITY

The MS-based proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository and are available via ProteomeXchange with the identifier PXD010393. Tables containing all identified proteins are available with accession number, sequence coverage, number of identified peptides and quantitative values (Supplemental Tables S1–6).

* This work was partially supported by the Max Planck Society for the Advancement of Science, the European Union's Horizon 2020 research and innovation program (grant agreement no. 686547; MSmed project) and by the Novo Nordisk Foundation (grant NNF14CC0001).

[S] This article contains supplemental Figures and Tables. The authors state that they have potential conflicts of interest regarding this work: NB, OH, LF, OV, are employees of Evosep and MM is an indirect investor in Evosep.

** To whom correspondence should be addressed: Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry Martinsried, Germany. E-mail: mmann@biochem.mpg.de.

‖ These authors contributed equally.

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

REFERENCES

1. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355
2. 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
3. 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
4. Bekker-Jensen, D.B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen, C. L., Nielsen, M. L., and Olsen, J. V. (2017) An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Systems* **4,** 587–599
5. Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hogrebe, A., Harder, A., and Olsen, J. V. (2018) Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J Proteome Res,* **17,** 727–738
6. Kulak, N.A., P.E. Geyer, and Mann, M. (2017) Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* Manuscript in Press, 2017

7. Bruderer, R., Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, Gomez-Varela D, Reiter L. (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics* **16,** 2296–2309,
8. Meier, F., Geyer, P. E., Virreira, Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* doi:10.1038/s41592-018-0003-5
9. Riley, N. M., Hebert, A. S., and Coon, J. J. (2016) Proteomics Moves into the Fast Lane. *Cell Syst,* **2,** 142–143
10. Picotti, P., and Aebersold, R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9,** 555–566
11. Gillet, L. C., Leitner, A., and Aebersold, R. (2016) Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annu. Rev. Anal. Chem.* **9,** 449–472
12. 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
13. Geiger, T., Wehner, A, Schaab, C, Cox, J, 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
14. Geyer, P. E., Holdt, L.M., Teupser, D., and Mann, M. (2017) Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol,* **13,** 942
15. Liu, Y., Hüttenhain, R., Collins, B., and Aebersold, R. (2013) Mass spectrometric protein maps for biomarker discovery and clinical research. *Expert Rev. Mol. Diagn.* **13,** 811–825
16. Geyer, P. E., Kulak, N. A., Pichler, G., Holdt, L. M., Teupser, D., and Mann, M. (2016) Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2,** 185–195
17. 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. (2016) Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12,** 901
18. Falkenby, L. G., Such-Sanmartín, G., Larsen, M. R., Vorm, O., Bache, N., and Jensen, O. N. (2014) Integrated solid-phase extraction-capillary liquid chromatography (speLC) interfaced to ESI-MS/MS for fast characterization and quantification of protein and proteomes. *J. Proteome Res.* **13,** 6169–6175
19. Ishihama, Y., Rappsilber, J., and Mann, M. (2006) Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics. *J. Proteome Res.* **5,** 988–994
20. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
21. Wisniewski, J. R., Dus, K., and Mann, M. (2013) Proteomic workflow for analysis of archival formalin-fixed and paraffin-embedded clinical samples to a depth of 10 000 proteins. *Proteomics Clin. Appl.* **7,** 225–233
22. Binai, N. A., Marino, F., Soendergaard, P., Bache, N., Mohammed, S., and Heck, A. J. (2015) Rapid analyses of proteomes and interactomes using an integrated solid-phase extraction-liquid chromatography-MS/MS system. *J. Proteome Res.* **14,** 977–985
23. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* **26,** 966–968
24. Olsen, J. V., et al. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712
25. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) *Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res,* **10,** 1794–1805
26. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M., Geiger, T., Mann, M., and Cox, J., (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Method* **13,** 731–740
27. Davis, M. T., Stahl, D. C., and Lee, T. D. (1995) Low flow high-performance liquid chromatography solvent delivery system designed for tandem capillary liquid chromatography-mass spectrometry. *J. Am. Soc. Mass Spectrom.* **6,** 571–577
28. Waanders, L. F., Almeida, R., Prosser, S., Cox, J., Eikel, D., Allen, M. H., Schultz, G. A., Mann, M. (2008) A novel chromatographic method allows on-line reanalysis of the proteome. *Mol. Cell. Proteomics* **7,** 1452–1459
29. Hosp, F., Scheltema, R. A., Eberl, H. C., Kulak, N. A., Keilhauer, E. C., Mayr, K., and Mann, M. (2015) A double-barrel liquid chromatography-tandem mass spectrometry (LC-MS/MS) system to quantify 96 interactomes per day. *Mol. Cell. Proteomics* **14,** 2030–2041
30. 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,** 6187–6195
31. Savitski, M. M., Reinhard, F.B., Franken, H., Werner, T., Savitski, M.F., Eberhard, D., Martinez Molina, D., Jafari, R., Dovega, R.B., Klaeger, S., Kuster, B., Nordlund, P., Bantscheff, M., and Drewes, G. (2014) Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346,** 1255784
32. Itzhak, D.N., Tyanova, S., Cox, J., and Borner, G.H. (2016) Global, quantitative and dynamic mapping of protein subcellular localization. Elife, **5**
33. Fu, Q., Kowalski, M. P., Mastali, M., Parker, S. J., Sobhani, K., van den Broek, I., Hunter, C. L., and Van Eyk, J. E. (2018) Highly reproducible automated proteomics sample preparation workflow for quantitative mass spectrometry. *J. Proteome Res.* **17,** 420–428
34. Meier, F., et al. (2018) Online parallel accumulation – serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer bioRxiv doi: https://doi.org/10.1101/336743

3.4. Article 4: Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies

Authors: Philipp E Geyer[1,2], Eugenia Voytik[1], Peter V Treit[1], Sophia Doll[1,2], Alisa Kleinhempel[3], Lili Niu[2], Johannes B Müller[1], Marie-Luise Buchholtz[3], Jakob M Bader[1], Daniel Teupser[3], Lesca M Holdt3&Matthias Mann[1,2,*]

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

2 NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

3 Institute of Laboratory Medicine, University Hospital, LMU Munich, Munich, Germany

*Corresponding author. Tel: +49 89 8578 2557; E-mail: mmann@biochem.mpg.de

Our publication ‘Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies’ is directed towards the resurgent trend of plasma proteomics for biomarker screening. As the literature reveals, a large proportion of studies tend to report proteins as biomarkers which our publication terms ‘quality markers’ but really indicating either erythrocyte, thrombocyte or coagulation contamination. To pinpoint the specific proteins that must be treated with caution, we acquire deep proteomes of erythrocytes, thrombocytes and pure plasma and compare the proteomes of plasma and serum to identify markers for coagulation in plasma samples.

These proteomes for the first time yield a list of bias specific proteins that can be applied for convolution analysis. For all samples of a plasma proteome study, a quality assessment can now be done to flag the samples which have problems in the sample taking process and potentially should be excluded. Additionally, when performing correlation analysis of the quantified proteins within a study, proteins can be tested for ‘origin bias’. For instance, if a protein of interest clusters together with known quality marker for thrombocytes this is a sign that this candidate should be treated with caution as it is most likely a thrombocyte derived protein and only appears to be regulated in the disease process.

We supply these analyses to the community in an online tool to check uploaded datasets at www.plasmaproteomeprofiling.org.

Report

TRANSPARENT PROCESS OPEN ACCESS

EMBO Molecular Medicine

Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies

Philipp E Geyer[1,2], Eugenia Voytik[1], Peter V Treit[1], Sophia Doll[1,2], Alisa Kleinhempel[3], Lili Niu[2], Johannes B Müller[1], Marie-Luise Buchholtz[3], Jakob M Bader[1], Daniel Teupser[3], Lesca M Holdt[3] & Matthias Mann[1,2,*]

Abstract

Plasma and serum are rich sources of information regarding an individual's health state, and protein tests inform medical decision making. Despite major investments, few new biomarkers have reached the clinic. Mass spectrometry (MS)-based proteomics now allows highly specific and quantitative readout of the plasma proteome. Here, we employ Plasma Proteome Profiling to define quality marker panels to assess plasma samples and the likelihood that suggested biomarkers are instead artifacts related to sample handling and processing. We acquire deep reference proteomes of erythrocytes, platelets, plasma, and whole blood of 20 individuals (> 6,000 proteins), and compare serum and plasma proteomes. Based on spike-in experiments, we determine sample quality-associated proteins, many of which have been reported as biomarker candidates as revealed by a comprehensive literature survey. We provide sample preparation guidelines and an online resource (www.plasmaproteomeprofiling.org) to assess overall sample-related bias in clinical studies and to prevent costly miss-assignment of biomarker candidates.

Keywords biomarker discovery; mass spectrometry; plasma proteomics; sample quality; study design
Subject Categories Biomarkers; Proteomics
DOI 10.15252/emmm.201910427 | Received 4 February 2019 | Revised 26 August 2019 | Accepted 3 September 2019 | Published online 30 September 2019
EMBO Mol Med (2019) 11: e10427

Introduction

Protein levels determined in blood-based laboratory tests can be useful proxies of diseases. These biomarkers assess normal physiological status, pathogenic processes, or a response to an exposure or intervention (FDA-NIH:Biomarker-Working-Group, 2016). Proteins and enzymes constitute the largest proportion of laboratory tests, reflecting the importance of the plasma proteome in clinical diagnostics (Geyer *et al*, 2017). Typical protein biomarkers such as the enzymes aspartate aminotransferase (ASAT) and alanine aminotransferase (ALAT) for the diagnosis of liver diseases or cardiac troponins indicating myocardial necrosis are used routinely in clinical decision making. Enzymatic activity or antibody-based laboratory tests are performed in high-throughput and at relatively low costs, as the standard of health care. However, specific biomarkers are only available for a very limited number of conditions and most have been introduced decades ago (Anderson *et al*, 2013). There is thus a critical need to make the biomarker discovery process more efficient.

Protein-binder assays quantifying many plasma proteins in parallel have become available (Gold *et al*, 2010; Assarsson *et al*, 2014), resulting in large-scale biomarker mining efforts (Ganz *et al*, 2016; Herder *et al*, 2018; Sun *et al*, 2018). Orthogonal to those technologies, mass spectrometry (MS)-based proteomics has become increasingly powerful in all domains of protein research (Aebersold & Mann, 2003, 2016; Munoz & Heck, 2014). MS measures the mass and fragmentation spectra of tryptic peptides derived from the sample with very high accuracy. Because these peptide and fragment masses are unique, MS-based proteomics is inherently specific, which can be an advantage over enzyme tests and immunoassays (Wild, 2013). Within its limit of detection, MS-based proteomics can analyze all proteins in a system and is unbiased and hypothesis-free in this sense.

The proteomic community has developed guidelines for the development, specificity, and potential clinical application of biomarkers. These discuss quality standards and emphasize the importance of selecting cohorts that are appropriate in size, thus ensuring the statistical significance of potential findings (Mischak *et al*, 2010; Surinova *et al*, 2011; Skates *et al*, 2013; Hoofnagle *et al*, 2016; Geyer *et al*, 2017). That being said, there are no systematic procedures in place to assess the proteome-wide effects of pre-analytical handling of blood-based samples. Considering that plasma samples are often collected during daily clinical routine and variably processed, sample collection and processing clearly have the potential to negatively influence clinical studies, making it difficult to uncover true biomarkers, while potentially contributing incorrect ones. Especially in case–control studies, any difference in the

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
2 NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
3 Institute of Laboratory Medicine, University Hospital, LMU Munich, Munich, Germany
*Corresponding author. Tel: +49 89 8578 2557; E-mail: mmann@biochem.mpg.de

collection and processing of samples may result in systematic bias. So far, relatively little attention has been paid to this crucial aspect on a proteome-wide scale and these studies mainly investigate pre-analytical effects (Rai *et al*, 2005; Timms *et al*, 2007; Schrohl *et al*, 2008; Qundos *et al*, 2013; Hassis *et al*, 2015).

Recently, we developed "Plasma Proteome Profiling", an automated MS-based pipeline for high-throughput screening of plasma samples (Geyer *et al*, 2016a). In this article, we apply this technology to systematically assess the quality of individual samples and clinical studies with the aim to identify generally applicable quality marker panels. Blood collection and subsequent errors in preparation are likely sources of plasma contamination. To address this issue, we construct proteomic catalogs of contaminating cell types as well as proteomic changes that may be induced during processing. This results in three panels of contaminating proteins, recommendations for assessing the quality of plasma samples and for consistent sample processing. We develop an online tool for biomarker studies and test the applicability of the panels on a recent investigation on the effects of weight loss on the plasma proteome (Geyer *et al*, 2016b). A comprehensive literature review of plasma proteome studies highlights that about half of them potentially suffer from limitations related to sample processing.

Results

Erythrocyte and platelet proteins in the plasma proteome

During the development of our Plasma Proteome Profiling pipeline and its optimization for high-throughput screening of human cohorts (Geyer *et al*, 2016a), we repeatedly observed proteins that tended to emerge as groups of statistically significant outliers but appeared to be independent of the particular study. We hypothesized that they reflected sample quality issues. Manual and bioinformatic inspection revealed three classes of origin: erythrocytes, platelets, and the blood coagulation system. Consequently, we designed experiments to systematically characterize these main quality issues of the plasma proteome.

First, we acquired reference proteomes of erythrocytes and platelets, which are by far the most abundant cellular components (5×10^6 and 3×10^5 cells per µl). We harvested these cellular components from 10 healthy females and 10 males to obtain representative erythrocytes, platelets, and pure (platelet-free) plasma and further collected platelet-rich plasma and whole blood (Fig 1A; see Materials and Methods). Cell counting confirmed the purity of the samples (Table EV1). All five blood fractions were separately prepared for each individual by our automated proteomic sample preparation pipeline, followed by liquid chromatography coupled to high-resolution mass spectrometry (LC-MS/MS). To create reference proteomes, we generated a very deep library from pooled samples by analyzing extensively pre-fractionated peptides (Kulak *et al*, 2017; see Materials and Methods). A total of 6,130 different proteins were identified from 61,654 sequence-unique peptides (Fig 1B and C). The platelet proteome was the most extensive (5,793 proteins), whereas we detected 2,069 proteins in erythrocytes, 1,682 in platelet-rich plasma, and 912 in platelet-free plasma. The comparison of platelet-rich plasma to platelet-free plasma (84% additional proteins) demonstrates the extent of proteins that can be introduced by platelets.

Next, we investigated purified samples for all 20 study participants individually. The average numbers of identified proteins and peptides were very consistent in all individuals (Appendix Fig S1). To construct panels of easily detectable and robust quality markers, we calculated the average protein intensities and the coefficient of variation (CV) across the study participants. As a prerequisite, we required that the proteins should be substantially more abundant in erythrocytes as well as platelets rather than in plasma. According to these criteria, we selected the 30 most abundant proteins with CVs below 30% and at least a 10-fold higher expression level in the contaminating cell type than in plasma (Fig 1D and E). NIF3-like protein 1 (NIF3L1), a low-abundance erythrocyte-specific protein, was excluded, because it was inconsistently identified as was the platelet-bound coagulation factor F13A1, whose function makes it an unsuitable platelet marker. The remaining proteins represent our cellular quality marker panels (Table EV2). They overlap by just two proteins (actin/ACTB and glyceraldehyde-3-phosphate dehydrogenase/GAPDH), and their quantities were not correlated with each other (Appendix Fig S2). Thus, they are specific and independent indicators for the origin of plasma quality.

Comparing median expression values of proteins shared between the blood components revealed that plasma proteins do correlate with whole blood (Pearson's correlation coefficient $R = 0.43$), as expected. In contrast, there was no correlation between the platelet, erythrocyte, and plasma proteomes (Appendix Fig S2). This indicates that the levels of cellular proteins in plasma are not a constant fraction of those in the cellular proteomes. The platelet panel was enriched in platelet-rich plasma compared to normal (platelet-free) plasma. Both panels are de-enriched in pure plasma compared to whole blood, however, this effected the erythrocyte panel even more strongly, because centrifugation removes erythrocytes more efficiently than platelets. A histogram of both panels over the abundance range visualizes their distribution in the different blood compartments (Appendix Fig S2). Erythrocytes are 10-fold more abundant and fourfold larger than platelets, and indeed, the corresponding panel proteins have a 42-fold difference in whole blood. In plasma, however, their ratio was nearly one to one, again pinpointing a more efficient removal of erythrocytes than of platelets in standard sample preparation. The fact that several proteins of both panels were still detectable in pure plasma indicates a baseline level of contaminants due to imperfect de-enrichment or the life cycle of these cells. The four most abundant erythrocyte proteins, HBA1, HBB, CA1, and HBD, were present in pure plasma of almost all individuals, whereas lower abundant proteins were only sporadically identified. In contrast, platelet proteins were quantified over a larger abundance range and some of them were found in every individual.

In addition to the sum of panel protein abundances, we calculated their correlation to the standard reference panel defined by the 20 participants to several hundred plasma samples of a previous study (Geyer *et al*, 2016b). A distinct contamination of erythrocyte proteins seems to be a part of the plasma proteome as the erythrocyte panel has in general a relatively high correlation between the reference cohort erythrocyte levels and the plasma samples in the above-mentioned study. In contrast, in many plasma samples there

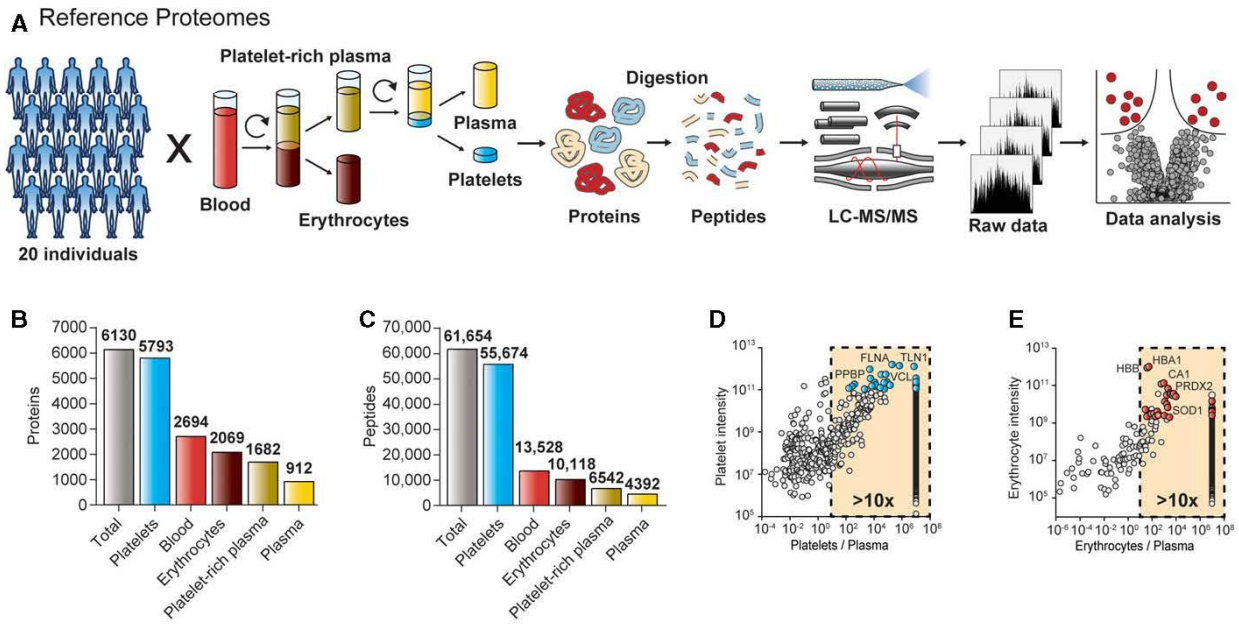


Figure 1. Identification of blood cell markers.

A Study outline and proteomic workflow. Erythrocytes, thrombocytes, platelet-rich, and platelet-free plasma were generated from 10 healthy female and male individuals by differential centrifugation and successive purification steps. To generate reference proteomes for each of the blood compartments, the respective protein samples of the 20 study participates were digested to peptides.

B, C Proteins (B) and peptides (C) identified for platelets, erythrocytes, platelet-rich, and platelet-free plasma.

D, E Selection of the most suitable quality marker proteins for (D) platelet contamination (blue dots) and (E) erythrocyte contamination (red dots) based on their abundance, the platelet/erythrocyte-to-plasma ratio, and the coefficient of variation. Proteins that were only detected in platelets or erythrocytes, but not in plasma are aligned on the right side of the graph.

was no correlation detectable between the reference cohort platelet levels and the plasma samples in the study. In practice, a correlation > 0.5 indicated that the proteins are present as a result of contamination (Appendix Fig S3A–C). Note that an apparent contaminant protein could still be applied as a biomarker—however, in this case its abundance value should be different from the pattern in the reference quality panel.

Serial dilution experiments validate the erythrocyte and platelet quality marker panels

To determine whether the two protein panels correctly quantify contamination in plasma, we generated four pools of erythrocytes and platelets from five study participants at a time. These pools were diluted in nine steps into platelet-free plasma for a total range of 10^7, followed by cell counting and proteomic analysis (Fig 2A). This resulted in an expected decrease in the cellular proteome ratio to plasma (Fig 2B and C). All but two of the panel proteins were consistently quantified over the dilution range. As the protein within each panel has the same origin, we defined a single variable for each cell type by summing their intensities and dividing by the summed intensities of all quantified plasma proteins. This yielded two remarkably robust "contamination indices" that turned out to be linear with respect to the cell numbers determined by cell cytometry (Table EV3; $R = 0.98$ and 0.99, Fig 2D and E). Spiked-in contaminations of 1:100 could readily be detected, which corresponds to a concentration of 70,000 erythrocytes or 30,000 platelets per µl plasma.

Quality marker panel for blood coagulation

In addition to contamination due to cellular constituents, partial and variable coagulation could contribute to systematic bias in biomarker studies. Indeed, we had found coagulation-related proteins to be connected to sample handling from finger pricks while developing our plasma proteomics pipeline (Geyer *et al*, 2016a). In clinical practice, an anticoagulant is pre-added to commercially available containers so that it is combined with blood upon withdrawal. Prompt inversion mixes the anticoagulant with the blood, yielding pure plasma after centrifugation (Fig 3A). Any delay in adding or mixing could cause partial coagulation—in the extreme case of missing anticoagulant and waiting for 30 min, one would obtain serum instead of plasma.

To generate a panel for assessing blood coagulation, we systematically compared 72 plasma vs. 72 serum samples (four individuals, 18 aliquots). From a total of 2,099 quantified proteins, 299 were significantly altered (Fig 3B). The most significantly de-enriched proteins after clotting were typical constituents of the coagulation cascade such as fibrinogen chains alpha (FGA), beta (FGB), and gamma (FGG) ($P < 10^{-130}$, > 40-fold), whereas the platelet-associated

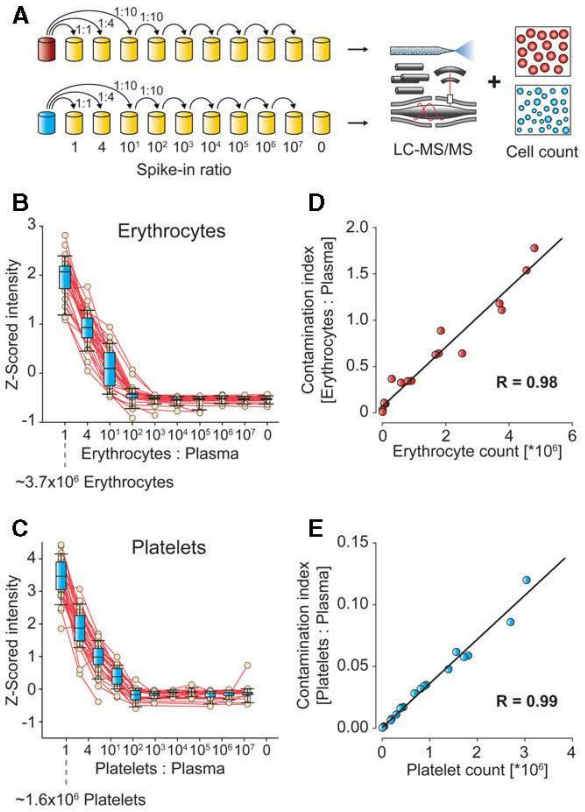


Figure 2. Spike-in of erythrocyte and platelet fractions into pure plasma.

A Dilution and analysis scheme.

B, C Protein intensities were Z-scored across the dilution series (B) for the 29 quality markers of the erythrocyte panel and (C) for the 29 markers of the platelet panel as a function of their spike-in proportion to plasma. Whiskers indicate 10–90 percentiles, and horizontal lines denote the mean.

D Correlation of erythrocyte count to the "contamination index" for the erythrocyte marker panel.

E Correlation of platelet count to contamination index for the platelet marker panel.

coagulation factor F13A1 and antithrombin-III (SERPINC1) decreased by more than half. Interestingly, the strongest elevated proteins in serum were highly abundant platelet proteins: platelet basic protein (PPBP), platelet glycoprotein Ib alpha chain (GP1BA), thrombospondin 1 (THBS1), and platelet glycoprotein V (GP5) ($P < 10^{-10}$; twofold to fivefold increase). In total, 208 proteins increased and 91 decreased due to coagulation. The former set of proteins, which have higher levels in serum than in plasma, were also quantitatively enriched with high-abundant platelet proteins ($P < 10^{-5}$; median rank 699 of 3,150 proteins), indicating coagulation-induced activation of platelets.

To define a robust panel of quality markers for the extent of coagulation, we first selected the 30 most significantly altered proteins between serum and plasma. Although not among the top 30, we added the platelet factor 4 variant 1 (PF4v1; $P < 10^{-11}$, 2.2-fold up in serum), because it was an excellent indicator of coagulation in our studies and has already been reported in the context of pre-analytical variation (Timms *et al*, 2007).

In contrast to the erythrocyte and platelet panels, proteins of the coagulation panel increase or decrease due to blood clotting and the fold changes vary strongly between them. Because fold changes are greatest for the decreasing proteins, we calculated the coagulation marker ratio only from them (sum of all plasma proteins divided by sum of plasma-elevated coagulation proteins). This ratio was very robust when comparing serum and plasma, clearly separating them with median ratios of 9 and 120 for these distinct sample types (Fig 3C). Of the coagulation marker panel, only F13A1, PPBP, and THBS1 were in common with the platelet panel and none with the erythrocyte panels (Fig 3D). The low overlap observed for the three quality marker panels should make them highly specific tools to elucidate the presence and origin of sample-related bias.

Application of the quality marker panels to a biomarker study

The above-defined marker panels can assess sample-related issues at three levels: the quality of each sample in a clinical cohort, potential systematic bias in the entire study, and the likelihood that individual biomarker candidates belong to the contaminant proteomes.

We recently investigated changes in the plasma proteome upon weight loss (Geyer *et al*, 2016a,b). Briefly, caloric restriction in 52 individuals for 2 months was followed by weight maintenance for 1 year. Plasma Proteome Profiling of seven longitudinal samples revealed significant changes in the profile of apolipoproteins, a decrease in inflammatory proteins and markers correlating with insulin sensitivity. Given that protein abundance changes of < 20% were often highly significant, we expected that overall sample quality was high, making this study suitable for testing the practical applicability of the quality marker panels.

First, we assessed the quality of each sample separately by calculating the three contamination indices and plotting their distribution in the total of 318 measurements. For each index, we initially defined potentially contaminated samples as those with a value more than two standard deviations above the mean (red lines in Fig 4A). This flagged 12 samples, six with platelet contamination, one with increased erythrocyte levels, and five with signs of partial coagulation. Resolving the three quality marker panels to the levels of individual proteins resulted in almost perfectly parallel trajectories (Appendix Fig S4A–C). Accordingly, the correlations to the reference quality marker panels were substantial ($R > 0.77$). Overall, the variation of the contamination indices was highest for the platelets also visible by a contamination index difference (max/min ratio) of a factor 182 between the least and the most contaminated sample, followed by erythrocytes (max/min 23), and lowest for coagulation (max/min 5). The platelet proteins talin-1 (TLN1), myosin-9 (MYH9), and alpha-actinin-1 (ACTN1) had the largest variations, all with maximal changes > 5,000-fold. Catalase (CAT), carbonic anhydrase 1 and 2 (CA1, CA2) from the erythrocyte index varied maximally by more than 500-fold. The three fibrinogens in the coagulation panel changed by up to 20-fold, indicating that only partial coagulation events took place (Fig 4A).

Note that evaluating individual sample quality based on the standard deviation of all samples, as done here, has the benefit of being independent of the specific proteomic method used to measure protein amounts. However, this requires that most samples have

low levels of contamination, so that outliers of the statistical distribution are clearly apparent. If this is not the case, we propose using general, study-independent cutoff values to differentiate between samples of high and poor quality in such studies.

To assess potential systematic bias for groups of samples such as cases and controls or different time points, we applied a *t*-test based volcano plot. Most of the significantly upregulated proteins at time point 4 were members of the platelet panel (Fig 4B). With this information in hand, we contacted our collaboration partners, who tracked down the platelet contamination to a switch of the blood-taking equipment due to low supplies.

In practice, such sample issues will occasionally happen in a clinical study, and our quality marker panels would allow elimination of the affected samples. However, if contaminating proteins can reliably be distinguished from relevant biomarker candidates, the data could still be used. In our example, six of the eight significant outliers were from the platelet panel, and the other two proteins—GP1BA and NRP1—could still be of interest. To investigate this further, we inspected the global correlation map of all proteins, time points, and participants (Albrechtsen *et al*, 2018). In this hierarchical clustering analysis, proteins that are co-regulated have a high correlation to each other and appear in groups, visualized as red patches (Fig 4C). Here, the platelet cluster was the second largest one with 38 proteins ($R = 0.69$). All quantified platelet panel proteins were in this cluster, as was GP1BA, flagging them as likely contaminants (Fig 4C and inset). Interestingly, NRP1, a receptor involved in angiogenesis, did not group with the platelet proteins, suggesting a potential biological role. This is supported by the fact that NRP1 was significantly regulated over all time points compared to the baseline, in contrast to the platelet cluster proteins.

The other two quality marker panels are also readily apparent in the global correlation map. Ten members of the erythrocyte panel cluster tightly as do the three fibrinogen chains (Appendix Fig S5). However, in this study the fibrinogens group with proteins involved in low-grade inflammation, reduction of which was one of the main findings of our study (Appendix Fig S5). In contrast, the coagulation marker PF4v1, which is also a highly abundant protein in platelets, clustered in the platelet group in this analysis, indicating that it varied as a result of sample preparation.

To make the above-described analysis readily available, we created an online platform at www.plasmaproteomeprofiling.org. It provides a toolbox for the interactive assessment of the quality of plasma proteomic data. Lists of protein abundances from MaxQuant search result tables or the template (Table EV4) can be uploaded by a simple drag and drop system. The system automatically generates the three contamination index values as shown in Fig 4A. If the user indicates cases and controls, the data set will be analyzed for systematic bias as visualized in a volcano plot (Fig 4B). The global correlation map is also displayed with the clusters of the quality marker panels (Fig 4C). The website is designed in the Dash data visualization framework, which allows further interactive analysis of the data (see Materials and Methods). Potential biomarker candidates in the volcano plot can be selected and displayed in the global correlation map to check whether the protein falls into or near one of the quality marker clusters.

Revisiting results of published biomarker studies

Having examined one study in detail, we set out to survey the extent to which quality marker proteins are reported as biomarker candidates in the literature. To this end, we performed a comprehensive PubMed search requiring the terms 'proteomics', 'proteome', 'plasma OR serum', 'biomarker' and 'mass spectrometry' spanning the time frame from 2002 to April 2018. We excluded review papers, purely technological publications without biomarker candidates, animal studies, and publications without proteins as qualitative or quantitative variables. From the resulting 210 publications, we manually extracted the lists of the biomarker candidates that were reported as "significantly altered proteins" by the authors. Gene and protein names were mapped to the corresponding protein identifiers in our reference panels and analyzed for their frequencies.

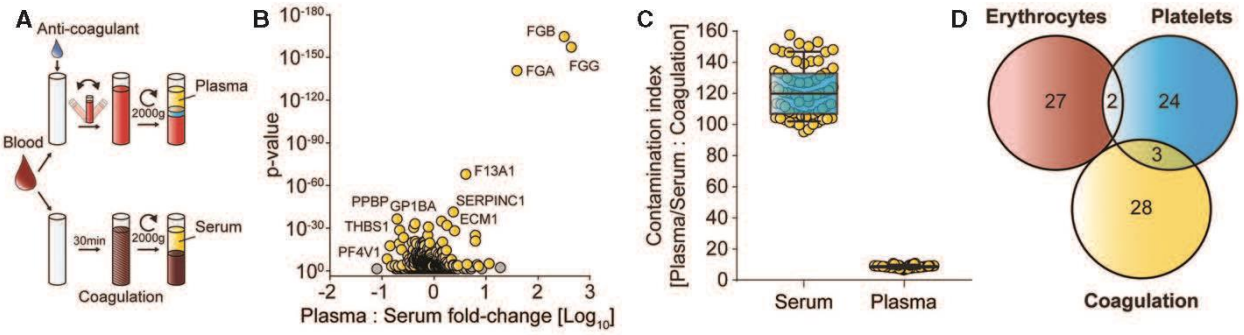


Figure 3. Quality marker panel for blood coagulation.

A Preparation of plasma and serum samples. EDTA was used as anticoagulation agent, and incubation and centrifugation values are indicated.

B Volcano plot comparing 72 plasma vs. 72 serum proteomes. Proteins highlighted in yellow were chosen according to their *P*-value as markers for coagulation. Only the plasma-enriched proteins (compared to serum) were used in the calculation of the coagulation contamination index.

C Ratio of the summed intensities of all plasma or serum proteins to the sum of the plasma-enriched panel proteins is plotted for all samples. Whiskers indicate the 10–90 percentile, and horizontal lines denote the mean.

D Overlap of the three quality marker panels.

Remarkably, 113 studies (54%) reported at least one potential quality marker as a biomarker candidate or as a statistically significant association (Fig 4D). As the total quality marker panel consists of 84 proteins and the median number of candidates per clinical study was seven, a certain overlap is not entirely unexpected. However, the candidates in question almost always were near the top of most abundant proteins of the quality marker panels, making it highly likely that they are indeed contaminants. Furthermore, while an individual protein could still be a genuine biomarker candidate, the fact that 22 studies (11%) reported two of them, and a further 23 studies (11%) three or more, again makes quality issues the likely explanation.

The majority of these studies reported proteins as potential biomarkers or as significant outliers of the coagulation panel, followed by the erythrocyte and platelet panels (Fig 4E). The most frequent one was clusterin (CLU; 27 times), followed by the fibrinogens (alpha, beta, and gamma; 22, 10, and 15 times), prothrombin (F2; 17 times), kininogen (KNG1; 15 times), antithrombin-III (SERPINC1; 13 times), and platelet basic protein (PPBP; 10 times). It is worth noting that proteins related to erythrocyte leakage may falsely be taken to indicate activation of oxidative pathways. For example, the hemoglobin subunits (e.g. HBA1, HBB, and HBD, listed 1, 6, and 1 time), carbonic anhydrases (CA1 and CA2, 6 and 6 times), fructose-bisphosphate aldolase (ALDOA, 5 times), peroxiredoxin 2 (PRDX2, 3 times), and superoxide dismutase (SOD1; 2 times) are annotated with keywords linked to oxidation. To illustrate this, a recent publication connected plasma proteome alterations in type 1 diabetes to oxidative stress. This may be a spurious link because the reported proteins were mostly members of the erythrocyte quality marker panel (Liu *et al*, 2018). Although platelet panel proteins are not prominent in the biomarker literature yet, we expect that they—along with lower abundant erythrocyte-specific proteins—will play an increasing role as technological progress enables higher plasma proteome coverage. We caution that platelet proteins already found in the biomarker literature such as PPBP, THBS1, and PF4 are often linked to coagulation events.

Recommendations for future proteomic studies

Based on our experience with the above-defined three quality marker panels (Table EV2) and analysis of thousands of plasma proteomes, we devised a general guideline for minimizing and detecting biases related to sample taking and processing (Table 1).

To further document the influence of common variables in the blood-taking process, we invited 10 healthy individuals and collected blood in 10 different blood sampling tubes. In this experiment, we systematically varied the type of plasma/serum, the blood specimen tubes (with or without gel), and the deposition of blood into the sampling tube (vacuum vs. pull system).

The most prominent differences were again between serum and plasma (Fig 3B; Appendix Fig S6). Apart from this, we found that contaminations with high-abundant erythrocyte-specific proteins appeared in several comparisons. Serum and EDTA plasma both had significantly higher levels than lithium heparin and citrate plasma (Appendix Fig S6A–F). Moreover, vacuum sampling can have an influence on erythrocyte-specific protein levels for some tubes. For instance, we found significantly increased levels of HBA1 and HBB in lithium heparin plasma tubes after vacuum sampling compared to a pull system, but not in the same comparison when using serum tubes (Appendix Fig S7A–D). Furthermore, erythrocyte-specific proteins were significantly increased in lithium heparin pull tubes (more than twofold), which contain a gel plug compared to pull tubes without a gel plug (Appendix Fig S8A–D). In contrast, there were no differences between serum tubes with and without gel. These findings illustrate how even seemingly minor changes in blood-taking equipment can result in statistically significant differences of protein levels, which could confound biomarker studies. They also highlight the value of unbiased, system-wide investigation of the blood proteome and our quality marker panels.

We also found that the procedure of sampling the plasma from the tubes has a prominent effect on platelet contamination (Appendix Figs S9 and S10). Thus, we recommend not to collect the lowest layer of the plasma above the platelet bed after centrifugation. Furthermore, any delay from centrifugation to plasma harvest has the potential to induce platelet protein contamination. These factors mainly influence the platelet rather than the erythrocyte contamination index, indicating that proteins from the platelet proteome are the most likely cause of erroneous assignment of biomarker candidates.

Discussion

Blood plasma remains the predominant biological matrix to assess health and disease in clinical settings. Around the world, every day hundreds of thousands of samples are analyzed to determine the levels of individual proteins. Likewise, blood plasma is directly or indirectly assessed in most clinical trials. Protein levels in plasma can readily be affected by cellular contamination or handling-related issues, and in clinical practice, this is partially addressed by simple tests such as those for hemoglobin contamination. However, these tests are not systematic or quantitative and they can only be used to exclude clearly contaminated samples.

Because of its high specificity and unbiased nature, MS-based proteomics is ideally suited to characterize the quality of blood plasma and it requires < 1 µl of material. So far, research on sample quality involving MS has mainly been restricted to the stability of internal standards in targeted assays and has rarely addressed overall sample quality (Schrohl *et al*, 2008; Hassis *et al*, 2015; Hoofnagle *et al*, 2016). Employing our Plasma Proteome Profiling pipeline to various clinical studies suggested that platelets, erythrocytes, and coagulation are by far the most important causes of plasma quality issues. We acquired very deep reference proteomes for these cell types and blood compartments, which we provide to the community to evaluate the possible origin of proteins emerging from biomarker studies. We defined three panels of about 30 proteins each that can serve as contamination indices (Table EV2). Using the example of a longitudinal Plasma Proteome Profiling study of weight loss and our online resource, we illustrated how the contamination indices can flag individual suspect samples and systematic biases. Furthermore, correlation analysis reveals whether potential biomarkers emerging from a given study are likely to be associated with quality-related proteome changes instead. Conversely, this procedure can "rescue" genuine biomarker candidates that are part of the quality marker proteomes. As an example, fibrinogens, a member of the coagulation quality marker panel, can also change during an inflammatory condition and might be correlated with classical inflammation markers such as CRP. In certain diseases, the entire set of proteins of a quality marker panel can be altered. For example, increased platelet

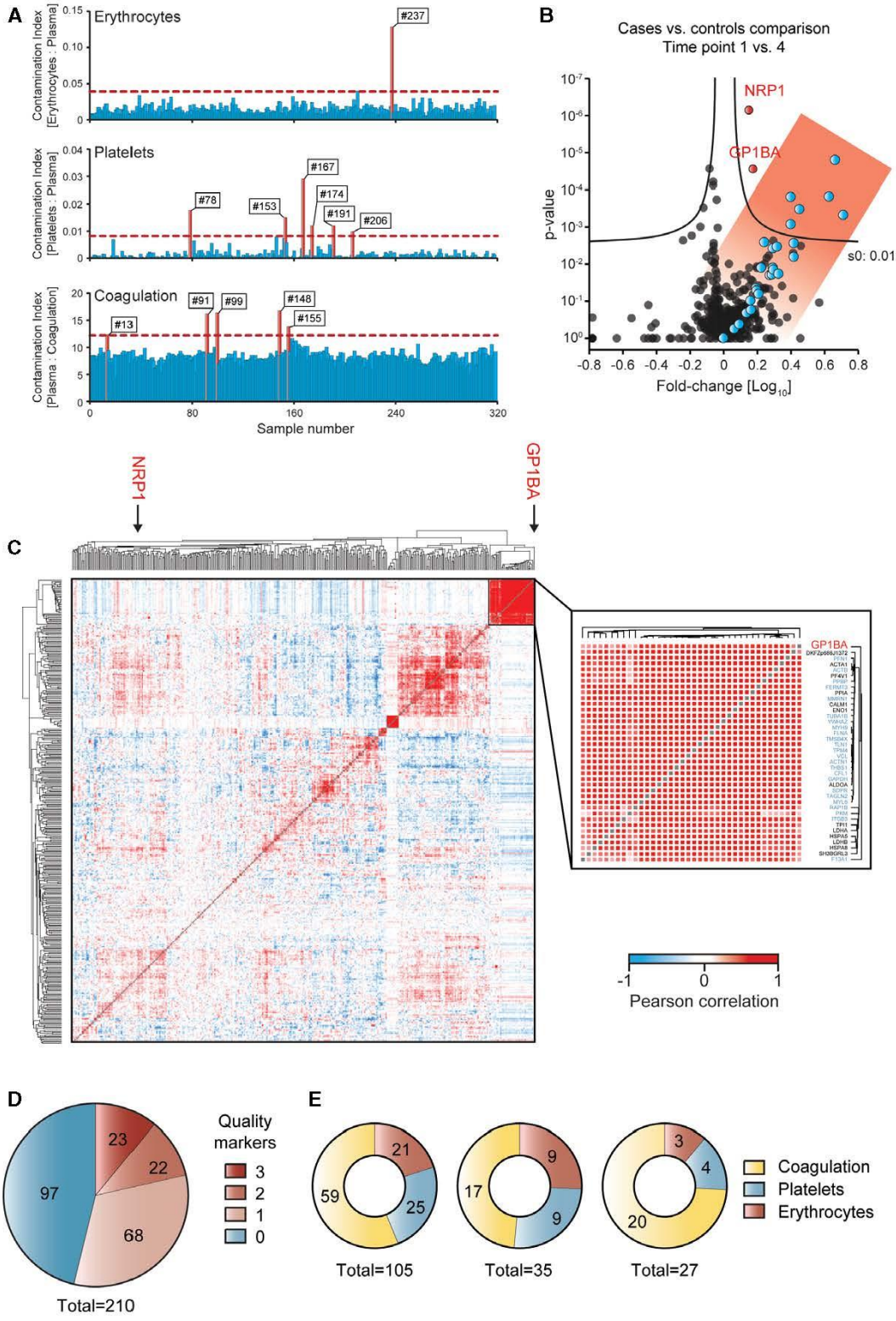

A
Erythrocytes
Contamination Index [Erythrocytes : Plasma]
#237
Platelets
Contamination Index [Platelets : Plasma]
#78
#153
#167
#174
#191
#206
Coagulation
Contamination Index [Plasma : Coagulation]
#13
#91
#99
#148
#155
Sample number
B
Cases vs. controls comparison
Time point 1 vs. 4
NRP1
GP1BA
p-value
s0: 0.01
Fold-change [Log10]
C
NRP1
GP1BA
Pearson correlation
D
Quality markers
3
2
1
0
23
22
97
68
Total=210
E
21
59
25
Total=105
9
17
9
Total=35
3
4
20
Total=27
Coagulation
Platelets
Erythrocytes

Figure 4.

Figure 4. Quality marker panels in a weight loss study and literature study.

A Assessment of individual sample quality with respect to the three contamination indices using the online tool at www.plasmaproteomeprofiling.org. Samples with indices that are more than two standard deviations from the mean (horizontal red lines) are flagged as potentially contaminated (red bars and sample numbers).

B Volcano plot of the proteome comparison of time point 1 vs. 4. Proteins of the platelet panel are highlighted in blue and two additional significantly regulated proteins in red.

C Global correlation map on the left with an inset of the platelet cluster on the right. The two significant outliers of the volcano plot in (B) are marked in red. Platelet panel proteins are highlighted in blue in the inset. Red patches in the global correlation map indicate positive and blue patches negative correlations.

D Literature analysis of 210 publications using MS-based plasma proteomics to identify new biomarkers. The number of quality markers reported as biomarker candidates in these studies is indicated.

E Distribution of the reported quality markers according to the three types of likely contaminations. The distribution is shown across studies that report one, two, or three proteins of the same quality marker panel.

levels—thrombocythemia—can have a variety of causes ranging from chronic inflammation to myeloproliferative diseases. Likewise, increased concentration of erythrocyte-specific proteins can be caused by hemolytic diseases such as in autoimmunity. While these cases are not the usual reasons why a quality marker panel is altered, they need to be considered when judging the analytical validity of a plasma measurement.

The clinical potential of the plasma proteome has long been realized and is also emphasized by the fact that more than 50 FDA-approved biomarkers can be quantified even in relatively shallow proteomic measurements of plasma (Geyer *et al*, 2016a). If there are as many new biomarkers among the less abundant proteins, there should be a diagnostic treasure trove still to be discovered (Geyer *et al*, 2017). Millions of plasma samples are stored in biobanks worldwide, representing an immense untapped resource that could be analyzed by MS-based proteomics or large-scale affinity-based methods. Despite initial enthusiasm and community efforts such as the Human Proteome Organization's plasma proteomic initiative (Omenn *et al*, 2005; Schwenk *et al*, 2017), few if any new protein biomarkers have entered the clinic in recent decades. This is probably at least partially due to technological limitations to characterize the vast dynamic range of the plasma proteome, which in turn has led to underpowered study designs (Geyer *et al*, 2017). While many of these challenges are already being addressed, we suspect that problems with sample quality represent another important reason for the paucity of new biomarkers and, even more seriously, for incorrect biomarkers being used. Examining our own data as well as the scientific literature, we here show that sample quality issues indeed have an impact on reported results. Nearly half of the reviewed studies reported at least one potential biomarker that is in our quality marker panels, and many had two or more, making sample contamination very likely. While coagulation-related issues are currently most prominent, increasing depth of plasma proteome coverage may replace platelet contamination as the most important source of error in the future. A corollary of the very large abundance variation of proteins introduced by quality issues is that it should further discourage pooling of samples. While this increases throughput, even a single contaminated sample can readily skew an entire batch.

Systematic bias introduced by imperfect sample handling or processing may lead to reporting incorrect biomarkers. Conversely, randomly distributed samples with poor quality will diminish overall statistical quality and may obscure true biomarker candidates.

The sources of quality issues are different kinds of variations in the pre-analytical processes, and we found platelet contamination during plasma harvesting to be one of the main culprits. Among the few previous studies, Hassis *et al* (2015) investigated different sample handling errors and concluded that only extreme conditions, such as delay in sample storage for 4 days, substantially changed the plasma proteome. However, proceeding with such extreme cases is rare, and quality issues are much more likely to originate from recontamination with whole blood after centrifugation during the plasma harvest or post-centrifugation times and resuspension of platelets, for instance. The comparison of 10 different blood sampling tubes showed that even seemingly minor differences in

Table 1. Practical considerations to minimize systematic bias.

General instructions
Avoid pooling of samples
Use plasma or serum exclusively, not a combination
Sample collection
Standardize blood collection and pre-analytical procedures (preferably same person collecting blood, centrifuge, sampling container, storage temperature, and time)
Centrifuge blood to generate plasma immediately
Centrifuge according to manufacturer's instruction
Harvest plasma immediately after centrifugation
Harvest the plasma starting from the top of the container and pool it before aliquotting
Discard the last 500 µl of plasma to avoid contamination with platelets or use a second centrifugation step to generate platelet-poor plasma
Freeze samples immediately after harvesting
Principal assessment of study sample quality
When working with a new batch of samples from collaborators: run at least 10 test samples of each study group by mass spectrometry
Use quality marker panels to check for any indication of contamination
Main study
Continuously assess quality during the project to detect and avoid systematic bias (pre-analytics, mass spectrometric analyses)
Overall quality: report the number of contaminated samples
Systematic bias: report potential systematic bias
Check whether biomarker candidates are contained in the quality marker panels
Identification of several quality markers as biomarker candidates may be indicative of a study vector
If a quality marker is among the biomarker candidates, thorough validation is required

the sample handling devices like a pull vs. a vacuum deposition system can have a statistically significant effect on the measured proteome. Therefore, we want to stress the importance of strictly following standard operating procedures. We here provide general considerations for minimizing sample-related issues, ranging from immediate harvest of the plasma after centrifugation to discarding the lowest layer of plasma to avoid recontamination with platelets (Table 1). These recommendations update and extend general good laboratory practices as well as HUPO guidelines (Omenn *et al*, 2005; Rai *et al*, 2005). We also advocate that plasma samples are quality-checked by MS-based proteomics, at least for a representative subset. This is especially important for clinical studies but also for targeted single-analyte measurements, which by their nature are blind to the overall composition of the sample. Although it would be possible to determine contamination indices by multiplexed affinity-based methods, we recommend MS for this purpose because of its very high specificity and its unbiased nature. Furthermore, the proteomic depth needed to assess the quality is easily achievable even in rapid and economical measurements.

The concepts and methods put forward in this study could readily be adapted to other body fluids such as urine, saliva, or cerebrospinal fluid. This would require developing the appropriate contamination indices. Furthermore, the three quality marker categories are the largest but not the only ones. For instance, we imagine that similar experiments can be performed to gauge the effect of storage duration and temperature on the plasma proteome as it influences MS-based proteomics.

In conclusion, sample-related quality issues are clearly a concern for biomarker studies. However, we show here that they can be addressed rigorously and comprehensively by MS-based proteomics. As this technology continues to improve in throughput, depth, and robustness, we envision that it will be employed in routine clinical practice. Biomarker panels instead of single markers will be measured by MS-based proteomics as this takes advantage of its inherently multiplexed nature and allows the characterization of clinical conditions more comprehensively. These biomarker panels could routinely be extended with quality marker panels as introduced here, helping to establish biomarker-guided decisions in a wide variety of clinically important areas.

Materials and Methods

Samples for defining the three quality marker panels

All participants gave written informed consent for their participation in the Munich Study on Biomarker Reference Values (MyRef), which is registered under the local ethic number 11-16. All experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.

To establish the quality marker panels, whole blood was harvested by venipuncture of 10 females and 10 males into commercial EDTA-containing sampling containers. The blood was centrifuged at 200 *g* for 10 min, and both the pellet and the supernatant were kept for further processing steps. The bottom layer of 500 µl plasma was discarded to avoid contamination of the platelet-rich plasma fraction with erythrocytes. The pellet was centrifuged at 2,000 *g* for 15 min, and the top layer containing plasma, the buffy coat, and 1 ml of erythrocytes were discarded. After adding 4 ml PBS containing 1.6 mg/ml EDTA, the suspension was centrifuged at 2,000 *g* for 15 min and the supernatant was discarded together with 500 µl of the top layer of the erythrocytes. This step was repeated, and the pure erythrocyte fraction was harvested. We centrifuged the supernatant from the first centrifugation step containing plasma and platelets a second time at 200 *g* for 10 min and harvested the supernatant, which constitutes the platelet-rich plasma. This step was repeated, and we collected the supernatant and the platelet after centrifugation at 2,000 *g* for 15 min. The supernatant was centrifuged a second time at 2,000 *g* for 15 min to harvest platelet-free plasma by sampling only top layer of the supernatant, but discarding the bottom layer of 500 µl. The platelets were washed twice by adding 4 ml PBS containing 1.6 mg/ml EDTA and centrifugation at 2,000 *g* for 15 min. The supernatant was discarded, and the pure platelet fraction was harvested.

For the serum and plasma comparison, blood samples from two females and two males were split into 18 samples each and serum and plasma were harvested after centrifugation at 2,000 *g* for 15 min.

To investigate the effects of different blood sampling devices on the blood plasma proteome, we invited 10 healthy individuals (five female and five males) and collected blood in the 10 different blood sampling devices (Table EV5). After collecting whole blood, it was incubated at room temperature for 30 min to allow coagulation in the serum tubes. The plasma tubes were also stored at room temperature for the same time, and the different tubes were centrifuged together. Afterward, 0.5 ml of plasma or serum was sampled from the top of the tubes.

To evaluate the platelet contamination in different layers of plasma after centrifugation, blood was collected in two different 9-ml S-Monovette EDTA-containing sampling containers (Sarstedt). The blood of one container was transferred to a 15-ml centrifugation tube without separation gel. Both containers were centrifuged at 2,000 *g* for 15 min. Plasma was harvested in nine volume fractions starting from the top layer in 500 µl steps to the top of the buffy coat. The buffy coat itself was not touched, and a small amount of plasma (~200 µl) remained on top.

High-abundant protein depletion for building a matching library

We created a matching library and applied a consecutive depletion strategy, in which the top 6 and top 14 most abundant plasma proteins were depleted by using a combination of two immunodepletion kits, as described in ref. Geyer *et al* (2016a). Briefly, the Agilent Multiple Affinity Removal Spin Cartridge was used for the depletion of the top six highest abundant proteins (albumin, IgG, IgA, antitrypsin, transferrin, and haptoglobin), followed by Seppro Human 14 Sigma immunodepletion for the 14 highest abundant proteins (albumin, IgG, IgA, IgM, IgD, transferrin, fibrinogen, α2-macroglobulin, α1-antitrypsin, haptoglobin, α1-acid glycoprotein, ceruloplasmin, apolipoprotein A-I, apolipoprotein A-II, apolipoprotein B, complement C1q, complement C3, complement C4, plasminogen, and prealbumin). Following depletion, we fractionated our samples using the high pH

The paper explained

Problem

New biomarkers are urgently needed in many health and disease contexts and mass spectrometry-based proteomics is a potentially powerful and promising technology for their discovery, as it can analyze the plasma proteome in a quantitative and specific manner. However, a systematic analysis of pre-analytical variations might obscure the discovery of novel biomarkers and has not been performed so far.

Results

We employ Plasma Proteome Profiling to discover three quality marker panels that report on the status of plasma samples with regards to erythrocyte lysis, platelet contamination, and partial coagulation. These panels can identify individual samples of poor quality and correct for systematic bias in biomarker studies. Moreover, they can be applied to evaluate whether a novel biomarker candidate is linked to one of the sources of contamination. We further provide sample preparation guidelines and an online resource to assess the overall sample-related bias in individual samples in clinical studies.

Impact

Quality issues due to erythrocyte lysis, platelet contamination, and partial coagulation might affect up to 50% of all biomarker studies as we showed by a literature survey of more than 200 published manuscripts. Our quality marker panels will prevent costly miss-assignment of potential biomarker candidates and support the discovery of promising biomarkers.

reversed-phase "Spider fractionator" into 24 fractions as described previously (Kulak *et al*, 2017).

Sample preparation: protein digestion and in-StageTip purification

Sample preparation was carried out according to our Plasma Proteome Profiling pipeline as described in Geyer *et al* (2016a,b) with an automated setup on an Agilent Bravo Liquid Handling Platform. In brief, plasma samples were diluted 1:10 with $_{dd}H_2O$ and 10 µl of the sample was mixed with 10 µl PreOmics lysis buffer (P.O. 00001, PreOmics GmbH) for reduction of disulfide bridges, cysteine alkylation, and protein denaturation at 95°C for 10 min (Kulak *et al*, 2014). Trypsin and LysC were added to the mixture after a 5-min cooling step at room temperature, at a ratio of 1:100 micrograms of enzyme to micrograms of protein. Digestion was performed at 37°C for 1 h. An amount of 20 µg of peptides was loaded on two 14-gauge StageTip plugs, followed by consecutive purification steps according to the PreOmics iST protocol (www.preomics.com). The StageTips were centrifuged using an in-house 3D-printed StageTip centrifugal device at 1,500 g. The collected material was completely dried using a SpeedVac centrifuge at 60°C (Eppendorf, Concentrator plus). Peptides were suspended in buffer A* [2% acetonitrile (v/v), 0.1% formic acid (v/v)] and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). Pools for each of the five sample types (whole blood, erythrocytes, platelets, plasma, and platelet-free plasma) were generated from the 20 individuals and prepared according to the procedure above. The peptides were fractionated using the high pH reversed-phase "Spider fractionator" into 24 fractions as described previously to generate deep proteomes (Kulak *et al*, 2017).

Ultra-high-pressure liquid chromatography and mass spectrometry

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1000 or 1200 ultra-high-pressure system (Thermo Fisher Scientific), which was coupled to a Q Exactive HF Orbitrap (Thermo Fisher Scientific) using 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, about 0.5 µg peptides were used for 45-min runs and for each fraction of the deep plasma data set.

Peptides were loaded in buffer A [0.1% formic acid and 5% DMSO (v/v)] and eluted with a linear 35-min gradient of 3–30% of buffer B [0.1% formic acid, 5% DMSO, and 80% (v/v) acetonitrile], followed stepwise by a 7-min increase to 75% of buffer B and a 1-min increase to 98% of buffer B, followed by a 2-min wash of 98% buffer B at a flow rate of 450 nl/min. Column temperature was kept at 60°C by an in-house-developed oven containing a Peltier element, and parameters were monitored in real time by the SprayQC software (Scheltema & Mann, 2012). MS data were acquired with a Top15 data-dependent MS/MS scan method for the construction of the library and BoxCar scans (Meier *et al*, 2018) for the study samples. Target values for the full-scan MS spectra were 3×10^6 charges in the 300–1,650 m/z range with a maximum injection time of 55 ms and a resolution of 60,000 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 30,000 at m/z 200 with an ion target value of 1×10^5 and a maximum injection time of 120 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides.

Data analysis

MS raw files were analyzed by MaxQuant software, version 1.5.6.8, (Cox & Mann, 2008), and peptide lists were searched against the human UniProt FASTA database. A contaminant database generated by the Andromeda search engine (Cox *et al*, 2011) was configured with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. We set the false discovery rate (FDR) to 0.01 for protein and peptide levels with a minimum length of 7 amino acids for peptides, and the FDR was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin and LysC as proteases. A maximum of two missed cleavages were allowed. Peptide identification was performed with an initial precursor mass deviation up to 7 ppm and a fragment mass deviation of 20 ppm. The "match between run algorithm" in the MaxQuant quantification (Nagaraj *et al*, 2012) was enabled after constructing a matching library consistent of depleted and all the undepleted plasma samples. All proteins and peptides matching to the reversed database were filtered out. Label-free protein quantitation (LFQ) was performed with a minimum ratio count of 2 (Cox *et al*, 2014).

Bioinformatic analysis

All bioinformatic analyses were performed with the Perseus software of the MaxQuant computational platform (Cox & Mann, 2008;

Tyanova *et al*, 2016). For the global correlation analysis, proteins were filtered for at least 50% valid values in the weight loss study and the hierarchical clustering was performed using Euclidean distance. The weight loss study contained in total 28 proteins of the platelet panel, but after sorting for 50% valid values only 24 were left and all of them clustered in the platelet panel.

Online platform for automated analysis of clinical studies

Our online portal is equipped with a user-friendly graphical interface that supports the most common web browsers, such as Google Chrome, Firefox, and Internet Explorer. For the front-end development, a Dash framework was used (version 0.27.0), which consists of a Flask server (1.0.2) that communicates with front-end React.js components using JSON, or JavaScript Object Notation, packets (a minimal, readable format for structuring data) over HTTP, or Hypertext Transfer Protocol, requests that work as request–response protocols between a client and server. Taking advantage of the full power of Cascading Style Sheets (CSS), every graphical element was customized: the sizing, the positioning, the colors, and the fonts.

The platform takes the results of the MS data processed by the MaxQuant software (Cox & Mann, 2008) from the proteinGroups table (to be extended to other formats). During the data uploading, the input file is verified through a combination of preliminary tests. We built a complex data structure using general Python libraries, such as NumPy, Pandas, and SciPy. Using three panels of markers for platelet contamination, erythrocyte contamination, and coagulation events in plasma samples, respectively, we identify samples affected by quality issues. Samples having at least 50% "valid values" (i.e. those with quantification results) are preprocessed by cleaning the data and prepare them for the subsequent visualization step.

Data availability

The MS-based proteomic data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository and are available via ProteomeXchange with identifier PXD011749 (https://www.ebi.ac.uk/pride/archive/projects/PXD011749).

Expanded View for this article is available online.

Acknowledgements

We thank all members of the Proteomics and Signal Transduction Group and the Clinical Proteomics Group for help and discussions and in particular Igor Paron, Christian Deiml, Alexander Strasser, and Gaby Sowa for technical assistance; Mario Oroshi for help with the online resource; Nicolai J. Wewer Albrechtsen, Nils A. Kulak, Niels Skotte, and Martin Steger for discussion; and Jürgen Cox for bioinformatic tools. The work carried out in this project was partially supported by the Max Planck Society for the Advancement of Science, the European Union's Horizon 2020 research and innovation program with the MSmed project (no. 686547), and grants from the Novo Nordisk Foundation (NNF15CC0001; NNF15OC0016692) and the BMBF grant German Biobank Alliance (BMBF 01EY1711C).

Author contributions

PEG designed, performed, and interpreted the MS-based proteomic analysis of patient plasma; wrote the paper; and generated the figures. PVT wrote the manuscript and performed together with LN, SD, JBM, AK, MLB, and JB experiments and generated article text. DT and LMH designed experiments, drafted practical considerations for sample preparation, and worked on the article text. EV designed and established the interactive online resource. MM designed and interpreted the MS-based proteomic analysis of plasma, supervised and guided the project, and wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

For more information

(i) https://www.biochem.mpg.de/en/rd/mann
(ii) https://www.cpr.ku.dk/research/proteomics/mann-group/
(iii) http://www.plasmaproteomeprofiling.org/

References

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

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

Albrechtsen NJW, Geyer PE, Doll S, Bojsen-Moller KN, Martinussen C, Torekov SS, Keilhauer E, Treit PV, Meier F, Holst JJ *et al* (2018) Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery. *Cell Syst* 7: 601 – 612 e3

Anderson NL, Ptolemy AS, Rifai N (2013) The riddle of protein diagnostics: future bleak or bright? *Clin Chem* 59: 194 – 197

Assarsson E, Lundberg M, Holmquist G, Bjorkesten J, Thorsen SB, Ekman D, Eriksson A, Rennel Dickens E, Ohlsson S, Edfeldt G *et al* (2014) Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One* 9: e95192

Cox J, 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 RA, Olsen JV, Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10: 1794 – 1805

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, 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

FDA-NIH:Biomarker-Working-Group (2016) BEST (Biomarkers, EndpointS, and other Tools) Resource. Maryland: Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US)

Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, Sterling DG, Williams SA (2016) Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* 315: 2532 – 2541

Geyer PE, Kulak NA, Pichler G, Holdt LM, Teupser D, Mann M (2016a) Plasma proteome profiling to assess human health and disease. *Cell Syst* 2: 185 – 195

Geyer PE, Wewer Albrechtsen NJ, Tyanova S, Grassl N, Iepsen EW, Lundgren J, Madsbad S, Holst JJ, Torekov SS, Mann M (2016b) Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol Syst Biol* 12: 901

Geyer PE, Holdt LM, Teupser D, Mann M (2017) Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol* 13: 942

Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, Carter J, Dalby AB, Eaton BE, Fitzwater T *et al* (2010) Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* 5: e15004

Hassis ME, Niles RK, Braten MN, Albertolle ME, Ewa Witkowska H, Hubel CA, Fisher SJ, Williams KE (2015) Evaluating the effects of preanalytical variables on the stability of the human plasma proteome. *Anal Biochem* 478: 14 22

Herder C, Kannenberg JM, Carstensen-Kirberg M, Strom A, Bonhof GJ, Rathmann W, Huth C, Koenig W, Heier M, Krumsiek J *et al* (2018) A systemic inflammatory signature reflecting cross talk between innate and adaptive immunity is associated with incident polyneuropathy: KORA F4/FF4 study. *Diabetes* 67: 2434 2442

Hoofnagle AN, Whiteaker JR, Carr SA, Kuhn E, Liu T, Massoni SA, Thomas SN, Townsend RR, Zimmerman LJ, Boja E *et al* (2016) Recommendations for the generation, quantification, storage, and handling of peptides used for mass spectrometry-based assays. *Clin Chem* 62: 48 69

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

Kulak NA, Geyer PE, Mann M (2017) Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol Cell Proteomics* 16: 694 705

Liu CW, Bramer L, Webb-Robertson BJ, Waugh K, Rewers MJ, Zhang Q (2018) Temporal expression profiling of plasma proteins reveals oxidative stress in early stages of Type 1 Diabetes progression. *J Proteomics* 172: 100 110

Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods* 15: 440 448

Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, Bennett SE, Bischoff R, Bongcam-Rudloff E, Capasso G *et al* (2010) Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* 2: 46 ps42

Munoz J, Heck AJ (2014) From the human genome to the human proteome. *Angew Chem Int Ed Engl* 53: 10864 10866

Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, Vorm O, 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

Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS *et al* (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 5: 3226 3245

Qundos U, Hong MG, Tybring G, Divers M, Odeberg J, Uhlen M, Nilsson P, Schwenk JM (2013) Profiling post-centrifugation delay of serum and plasma with antibody bead arrays. *J Proteomics* 95: 46 54

Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, Mehigh RJ, Cockrill SL, Scott GB, Tammen H *et al* (2005) HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics* 5: 3262 3277

Scheltema RA, Mann M (2012) SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J Proteome Res* 11: 3458 3466

Schrohl AS, Wurtz S, Kohn E, Banks RE, Nielsen HJ, Sweep FC, Brunner N (2008) Banking of biological fluids for studies of disease-associated protein biomarkers. *Mol Cell Proteomics* 7: 2061 2066

Schwenk JM, Omenn GS, Sun Z, Campbell DS, Baker MS, Overall CM, Aebersold R, Moritz RL, Deutsch EW (2017) 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

Skates SJ, Gillette MA, LaBaer J, Carr SA, Anderson L, Liebler DC, Ransohoff D, Rifai N, Kondratovich M, Tezak Z *et al* (2013) Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. *J Proteome Res* 12: 5383 5394

Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P *et al* (2018) Genomic atlas of the human plasma proteome. *Nature* 558: 73 79

Surinova S, Schiess R, Huttenhain R, Cerciello F, Wollscheid B, Aebersold R (2011) On the development of plasma protein biomarkers. *J Proteome Res* 10: 5 16

Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'Jampens D, Podust VN, Ford J, Fung ET, Gammerman A, Jacobs I *et al* (2007) Preanalytic influence of sample handling on SELDI-TOF serum protein profiles. *Clin Chem* 53: 645 656

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. *Nat Methods* 13: 731 740

Wild D (2013) *The immunoassay handbook: theory and applications of ligand binding, ELISA, and related techniques*, 4th edn. Oxford; Waltham, MA: Elsevier

3.5. Article 5: Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease

Authors:

Jakob M Bader[1,†], Philipp E Geyer[1,2,†], Johannes B Müller[1], Maximilian T Strauss[1], Manja Koch[3], Frank Leypoldt[4,5], Peter Koertvelyessy[6,7], Daniel Bittner[6], Carola G Schipke[8], Enise I Incesoy[9], Oliver Peters[9,10], Nikolaus Deigendesch[11], Mikael Simons[12,13], Majken K Jensen[3,14], Henrik Zetterberg[15,16,17,18] & Matthias Mann[1,2,*]

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

2 NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

3 Departments of Nutrition & Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

4 Institute of Clinical Chemistry, Faculty of Medicine, Kiel University, Kiel, Germany

5 Department of Neurology, Faculty of Medicine, Kiel University, Kiel, Germany

6 Department of Neurology, Medical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany

7 Department of Neurology, Charité Universitätsmedizin Berlin, Berlin, Germany

8 Experimental & Clinical Research Center (ECRC), Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, &

Berlin Institute of Health, Berlin, Germany

9 Department of Psychiatry, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin & Berlin Insti tute of Health, Charité Universitätsmedizin Berlin,

Berlin, Germany

10 German Center for Neurodegenerative Diseases, Berlin, Germany

11 Institute of Medical Genetics and Pathology, University Hospital Basel, Basel, Switzerland

12 German Center for Neurodegenerative Diseases (DZNE), Munich, Germany

13 Munich Cluster for Systems Neurology, Munich, Germany

14 Department of Public Health, University of Copenhagen, Copenhagen, Denmark

15 Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden

16 Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden

17 UK Dementia Research Institute at UCL, London, UK

18 Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK

*Corresponding author. Tel: +49 8985782557; E-mail: mmann@biochem.mpg.de

† These authors contributed equally to this work

The publication 'Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease' applies MS-based proteomics to the characterization of CSF from Alzheimer's Disease and control patients. Our general aim was the discovery of new biomarkers or more general disease associated signatures. This was done with the 'rectangular approach' (Geyer et al., 2017), with three cohorts from different clinical sites and overall about 200 individuals. We supply CSF proteomes with an average depth of 1000 protein groups and are able to specify functional subclusters in the proteome like metabolic factors and neuronal tissue leakage proteins. Despite the fact that one of the

three cohorts shows less separation between cases and controls, we are able to identify a 40-protein signature, 35 with elevated abundance and 5 with decrease abundance in Alzheimer's disease. The majority of those proteins highly correlate with parameters for clinical identification of Alzheimer's disease like t-tau and are enriched for glycolysis associated functions.

By comparison of our results to a study with similar aims published as a preprint in parallel, we reduce the set of 40 proteins to 26, to which we apply machine learning. We find that a set of 14 proteins can correctly classify Alzheimer's disease status with a decision tree. After tuning the model, we demonstrate that the signature of the six most important parameters for classification can split between Alzheimer's disease patients and control with a sensitivity of 82% and a specificity of 87%.

Article

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease

Jakob M Bader[1,†], Philipp E Geyer[1,2,†], Johannes B Müller[1], Maximilian T Strauss[1], Manja Koch[3], Frank Leypoldt[4,5], Peter Koertvelyessy[6,7], Daniel Bittner[6], Carola G Schipke[8], Enise I Incesoy[9], Oliver Peters[9,10], Nikolaus Deigendesch[11], Mikael Simons[12,13], Majken K Jensen[3,14], Henrik Zetterberg[15,16,17,18] & Matthias Mann[1,2,*]

Abstract

Neurodegenerative diseases are a growing burden, and there is an urgent need for better biomarkers for diagnosis, prognosis, and treatment efficacy. Structural and functional brain alterations are reflected in the protein composition of cerebrospinal fluid (CSF). Alzheimer's disease (AD) patients have higher CSF levels of tau, but we lack knowledge of systems-wide changes of CSF protein levels that accompany AD. Here, we present a highly reproducible mass spectrometry (MS)-based proteomics workflow for the in-depth analysis of CSF from minimal sample amounts. From three independent studies (197 individuals), we characterize differences in proteins by AD status (> 1,000 proteins, CV < 20%). Proteins with previous links to neurodegeneration such as tau, SOD1, and PARK7 differed most strongly by AD status, providing strong positive controls for our approach. CSF proteome changes in Alzheimer's disease prove to be widespread and often correlated with tau concentrations. Our unbiased screen also reveals a consistent glycolytic signature across our cohorts and a recent study. Machine learning suggests clinical utility of this proteomic signature.

Keywords Alzheimer's disease; cerebrospinal fluid; mass spectrometry; neurodegeneration; proteomics

Subject Categories Biomarkers; Neuroscience; Proteomics

DOI 10.15252/msb.20199356 | Received 14 November 2019 | Revised 29 April 2020 | Accepted 30 April 2020

Mol Syst Biol. (2020) 16: e9356

Introduction

Alzheimer's disease (AD) is the most common type of dementia, and its prevalence is growing rapidly in aging societies (GBD 2016 Neurology Collaborators, 2019). In 2015, almost 47 million people worldwide were estimated to be affected by dementia, and the numbers are expected to reach 75 million by 2030, and 131 million by 2050, with the greatest increase expected in low-income and middle-income countries (Winblad *et al*, 2016). Patients with AD typically present with memory impairment and difficulty performing activities of daily living (Scheltens *et al*, 2016). However, symptoms may manifest decades after the underlying pathology has initiated, including the deposition of amyloid plaques and development of neurofibrillary tangles (Jack *et al*, 2010).

Biomarkers have become important diagnostic tools to define the presence and absence of dementia before onset of memory loss.

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
2 NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
3 Departments of Nutrition & Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
4 Institute of Clinical Chemistry, Faculty of Medicine, Kiel University, Kiel, Germany
5 Department of Neurology, Faculty of Medicine, Kiel University, Kiel, Germany
6 Department of Neurology, Medical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany
7 Department of Neurology, Charité Universitätsmedizin Berlin, Berlin, Germany
8 Experimental & Clinical Research Center (ECRC), Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, & Berlin Institute of Health, Berlin, Germany
9 Department of Psychiatry, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin & Berlin Institute of Health, Charité Universitätsmedizin Berlin, Berlin, Germany
10 German Center for Neurodegenerative Diseases, Berlin, Germany
11 Institute of Medical Genetics and Pathology, University Hospital Basel, Basel, Switzerland
12 German Center for Neurodegenerative Diseases (DZNE), Munich, Germany
13 Munich Cluster for Systems Neurology, Munich, Germany
14 Department of Public Health, University of Copenhagen, Copenhagen, Denmark
15 Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden
16 Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden
17 UK Dementia Research Institute at UCL, London, UK
18 Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK
*Corresponding author. Tel: +49 8985782557; E-mail: mmann@biochem.mpg.de
†These authors contributed equally to this work
[Correction added on 28 September 2020, after first online publication: Projekt Deal funding statement has been added.]

While a research framework for defining AD based on beta amyloid (Aβ) deposition, pathologic tau, and neurodegeneration (ATN) has been proposed (Jack *et al*, 2018), clinical criteria for AD are not universally standardized and range from clinical presentation to brain imaging by MRI and PET to clinical chemistry analysis of $A\beta_{1-42}/A\beta_{1-40}$, total-tau (t-tau), and phosphorylated-tau ($p\text{-}tau_{181}$) in cerebrospinal fluid (CSF; Frisoni *et al*, 2010; McKhann *et al*, 2011; Ferreira *et al*, 2014; Rice & Bisdas, 2017). Most research currently focuses on Aβ and tau, because they are the main components of amyloid plaques and neurofibrillary tangles (Serrano-Pozo *et al*, 2011). However, the search for a disease-modifying therapy has yet to show clinically relevant results and it is becoming increasingly clear that many additional pathological changes in multiple pathways occur in dementia.

Thus, we propose an unbiased analysis of CSF proteins in participants with and without AD for a comprehensive search for novel diagnostic biomarkers. A set of reliable protein biomarkers rather than a single marker could also enable the development of highly specific tests for early disease detection in at-risk segments of the population. Ideally, such markers should identify unexpected biological pathways and new potential therapeutic targets for future development.

Mass spectrometry (MS)-based proteomics has become a very powerful technology for the analysis of protein abundance levels, modifications, and interactions, with important discoveries in biological and biochemical research, including neuroscience (Aebersold & Mann, 2016; Hosp & Mann, 2017). MS-based proteomics is unbiased in the sense that it identifies and quantifies proteins in an untargeted manner. Additionally, the identification is extremely specific through the amino acid sequence information at the peptide level. These inherent features differentiate MS-based from affinity-based methods and should make MS an ideal tool for biomarker discovery; however, in body fluids this long-standing goal has not generally been realized so far. This has been due to a variety of technological and conceptual limitations, compromising reproducibility, the number of consistently quantified proteins and throughput (Geyer *et al*, 2017). For instance, a general issue in body fluid proteomics is the presence of highly abundant proteins such as albumin that hamper efficient identification of less abundant proteins. Previous workflows were laborious, typically quantified a few hundred proteins at most per sample and required hundreds of microliters of precious CSF, thereby limiting the availability of suitable samples (Dayon *et al*, 2018). Reproducibility was low with only a minority of proteins having clinically accepted coefficients of variation (CV) of $< 20\%$. Furthermore, many proteins were not quantifiable in all study participants and validation in well-characterized study populations was lacking. Therefore, entire databases have been curated to navigate reported CSF proteome alterations across studies in the field of neurodegeneration including AD (Guldbrandsen *et al*, 2017).

Recent technological advances enable substantially higher proteome coverage and better and more comprehensive protein quantitation. These developments include automated sample preparation, technological improvements in mass spectrometers, MS data acquisition, and processing software that synergize to enhance the overall analytical performance (Bruderer *et al*, 2017; Kelstrup *et al*, 2018). Based on these advances, we here developed a streamlined and highly reproducible workflow from sample preparation to data-independent MS acquisition (Ludwig *et al*, 2018) and an integrated analysis of the results for CSF. This workflow enabled us to clearly identify the established markers as well as a large number of consistent and biologically meaningful proteome changes across several independent cohorts.

Results

Overview of study populations

We recently proposed a shift in the study design of clinical discovery proteomics termed "rectangular strategy" (Geyer *et al*, 2017). In the previous "triangular strategy" study design, selected samples were characterized with extensive workflows and a small number of candidates were then assessed in a larger number of individuals using targeted methods. However, these candidates often turned out to be specific to the discovery population and could not be validated in independent study populations. In contrast, in the "rectangular strategy", multiple studies are subjected to the same high proteome depth workflow, moving the discovery to the population-wide setting in order to discern pathological from study-specific effects.

To implement the rectangular strategy, we analyzed three separate study populations of about 30 AD patients and about 30 or 50 controls, amounting to 197 individuals in total (Fig 1A). We refer to the study populations as cohorts throughout the manuscript, because each cross-sectional study was slightly different, conducted in distinct settings and geographical regions. One cohort originated from western Sweden, another from the German cities Magdeburg and Kiel (obtained through Harvard T. H. Chan School of Public Health), and the third from Berlin. The overall median age was 70.0 ± 12.1 years (± SD) (Fig EV1A). However, the 16 non-AD control patients of the Kiel sub-cohort were younger (median 32.0 ± 17.1 years). In each of our cohorts, patients were classified as AD if the t-tau concentration was above 400 ng/l, and the $A\beta_{1-42}$ concentration below 550 ng/l or the $A\beta_{1-42}/A\beta_{1-40}$ ratio below 0.065 as determined by ELISA measurement at the clinical collection site (Materials and Methods).

The degree of separation of AD cases and controls by clinical AD CSF biomarker concentrations differed across cohorts. AD and non-AD were best separated in the Sweden cohort but the Magdeburg cohort also exhibited a good overall separation (Fig EV1B–K, Materials and Methods). In the Berlin cohort, however, AD and control groups overlapped to some degree regarding CSF $A\beta_{1-42}$ and slightly regarding t-tau.

Characterization of the CSF proteomics workflow

Previously, we developed a streamlined Plasma Proteome Profiling pipeline, in which the proteins in one microliter of plasma are digested to peptides and purified for MS analysis in an automated system (Geyer *et al*, 2016). CSF contains much less protein than plasma, with about 0.17–0.70 g/l and 60–80 g/l total protein content, respectively (Seyfert *et al*, 2002; Laub *et al*, 2010). Nevertheless, we achieved a very robust workflow with high proteome depth from only a few microliter of sample that was not depleted of highly abundant proteins (Fig 1A and C). We adopted a data-

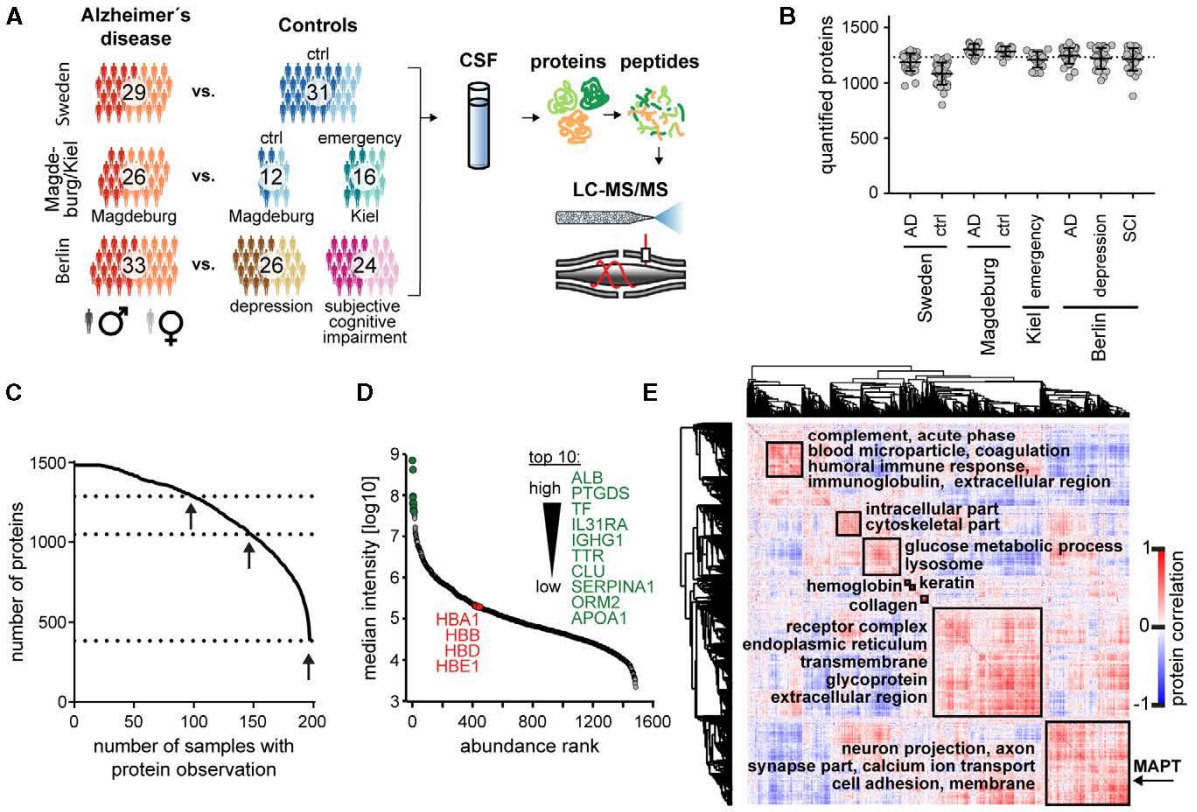


Figure 1. Study overview and CSF proteome characterization.

A Overview of the study populations (cohorts) and schematic proteomic workflow. The CSF of three cohorts comprising AD and control subjects was analyzed. The total number of subjects per cohort group is depicted. Light and dark shades represent female and male subjects, respectively. "Ctrl" refers to non-AD control subjects.

B Number of proteins identified and quantified passing the 1% FDR cutoffs in each sample. Horizontal lines show the mean and the error bars ± SD. The dashed line indicates the level of the meta-median (1,233 proteins) of the group medians of quantified proteins. Number of samples per group as shown in A).

C Data completeness curve. The number of proteins in the dataset (Y axis) depending on the minimum number of samples in which the proteins have each been quantified (X axis) is plotted. The arrows indicate 50%, 75%, and 100% data completeness.

D Median CSF protein abundance distribution as calculated from MS intensities of quantified peptides of each protein. The top ten most abundant proteins and hemoglobins are highlighted.

E Global correlation map of proteins generated by clustering the Pearson correlation coefficients of all possible protein combinations. The abundance of proteins with common regulation correlates across samples, and they therefore form a cluster. Prominent clusters are annotated with functional terms obtained from bioinformatics enrichment analysis. The position of tau (gene name MAPT) is labeled on the Y axis. The inset shows the color code for Pearson correlation coefficients.

independent acquisition strategy (DIA), both because it can achieve high data completeness (Gillet *et al*, 2012) and because it has been shown to perform excellently on the linear quadrupole-Orbitrap instruments employed here (Kelstrup *et al*, 2018). A DIA library of about 2,700 proteins was computationally merged from pooled AD and non-AD samples after separation into 24 fractions each and a direct-DIA search for all single-run samples (Materials and Methods). CSF proteomes were acquired by measuring single 100-min gradient runs for each patient.

On average, we quantified 1,233 proteins per CSF sample (Fig 1B, Datasets EV1–EV3). The data acquired with DIA had 100% completeness for 385 proteins (26%), 75% for 1,050 proteins (71%), and 50% for 1,288 proteins (87%) (Fig 1C). The quantified protein intensities spanned over six orders of magnitude, in which the top ten most abundant proteins contributed 65% of total protein intensity of the entire 1,484 proteins in our dataset (Fig 1D). To achieve such CSF proteome depth, extensive fractionation and depletion of abundant proteins often combined with isobaric labeling were previously required, with its associated disadvantages (preprint: Higginbotham *et al*, 2019; Sathe *et al*, 2019). For a single-shot CSF proteomics workflow that is amenable to high-throughput and large cohorts, this presents an unprecedented depth at high data completeness.

We investigated intra- and inter-assay variability of our automated CSF pipeline by repeated sample preparation (Materials and Methods), which revealed high reproducibility with over 1,000 proteins having inter-assay CVs below 20% (Fig EV2A and B, Datasets EV4 and EV5). This level of variability is much smaller than the proteome differences between subjects, as assessed by calculating the inter-individual variability within the cohorts. Here, only 225 proteins had a CV below 20% (Fig EV2C).

The availability of a large set of 197 CSF samples prompted us to investigate the relationship between different proteins in order to functionally interpret co-regulation of proteins that cluster with each other or with clinical parameters. The global protein correlation map (Wewer Albrechtsen *et al*, 2018) resulting from more than a million protein–protein comparisons highlighted eight main clusters of proteins which follow common functions or themes (Dataset EV6). For instance, neuronal annotation terms such as the gene ontology cellular compartments (GOCC) terms neuron projection, axon, and synapse were selectively enriched in the second largest cluster (Figs 1E and EV2D). Identification of neuronal proteins in the CSF highlights that proteins originating in the central nervous system accumulate in the CSF, thus making the CSF reflective of physiological or pathological proteome alteration in this organ.

Another cluster was enriched in blood plasma proteins relating to humoral immunity, the complement system or coagulation. Vascular proteins have been reported to be increased in AD brains while decreased in AD CSF (preprint: Higginbotham *et al*, 2019). However, apart from disease-associated effects such as a modulation of the blood–brain barrier, apparent alterations of blood protein abundances in CSF may be caused by blood contamination during CSF sampling which is hard to avoid entirely. Proteins are likely blood contaminants in CSF if they exhibit the same abundance profile across samples as known blood proteins and occur in the same abundance ratio to these blood proteins in CSF as in blood. Conversely, if a protein also found in blood does not correlate with the blood proteins, it may still be a genuine biomarker for AD. The global correlation map presents an efficient approach to distinguish biomarkers from contaminants (Geyer *et al*, 2019). Here, CSF signatures of proteins biologically relevant to AD clearly separated from protein clusters that are at higher risk to be contamination-associated (Fig 1E).

Proteomics detects differences in CSF t-tau in individuals with or without AD and neuronal and widespread novel proteome alterations

In the Sweden and the Magdeburg/Kiel cohorts, AD was associated with drastic CSF proteome alterations, with 540 and 453 proteins significantly ($P < 0.05$) differing by AD status, respectively. These changes encompassed up- and down-regulated proteins, and significant proteins had a median absolute fold change of about 1.3-fold in both studies. The extensive brain atrophy apparent upon autopsy and the widespread brain proteome alterations harmonize well with the observed substantial alterations in the CSF proteome in AD and other neurodegenerative diseases (Hosp *et al*, 2017; preprint: Higginbotham *et al*, 2019). In all three cohorts, tau (gene name MAPT) was the most significantly or among the most significantly altered proteins between individuals with or without AD, with higher levels in AD (Fig 2A–C, Appendix Fig S1A–E). The fact that tau levels are elevated in AD CSF has been known for more than two decades but this important protein is not easily quantified in large proteomics discovery cohorts. Typically, tau quantitation by mass spectrometry has required extensive fractionation and depletion of abundant proteins, limiting throughput (preprint: Higginbotham *et al*, 2019; Sathe *et al*, 2019). Alternatively, targeting instead of discovery strategies can in principle quantify proteins such as tau in larger sample numbers (Barthélemy *et al*, 2016).

Cerebrospinal fluid is expected to reflect pathological alterations in functional classes of proteins. AD is characterized by synaptic dysfunction and neuronal cell death. Proteins associated with the gene ontology (GO) term "neuron projection" were indeed enriched in AD CSF compared with non-AD CSF ($P < 0.01$ in the all three cohorts; Fig 2A–C, Appendix Fig S1F). Likewise, proteins of the GO term "synapse part" were significantly enriched in AD CSF in the Sweden and Berlin cohorts ($P < 0.01$).

In the Berlin cohort, proteome alterations between AD and non-AD CSF were smaller with only 168 proteins exhibiting significantly ($P < 0.05$) different abundances (Figs 2C and 3A, Appendix Fig S1D and E). This finding concurs with the reduced biochemical separation of the AD and non-AD groups in the Berlin cohort based on clinical AD CSF biomarkers (Fig EV1B–K).

Despite fewer significantly different proteins, the Berlin cohort exhibited the same key features of the two other cohorts such as tau being a top outlier and the enrichment of neuronal and synaptic proteins. The second dominant outlier 14-3-3γ (gene name YWHAG) in the Berlin cohort was likewise enriched in AD CSF in the other cohorts. The family of 14-3-3 proteins is very abundant in the brain and has been implicated in neurodegenerative diseases, and increased levels of 14-3-3γ have been reported in AD brain tissue and CSF (Fountoulakis *et al*, 1999; Foote & Zhou, 2012; Sathe *et al*, 2019). Together, this shows a reduced but equivalent AD-associated effect on the CSF proteome in the Berlin cohort.

Replication of AD-associated proteins across cohorts

As it had previously been challenging to establish biomarker panels that could be replicated across cohorts, we next assessed the consistency of AD-associated protein changes in this multi-cohort study. Of the significantly changed proteins described above, large proportions were consistent in their AD/non-AD association (Fig 3A and B, Dataset EV4). Comparing the Sweden and Magdeburg/Kiel cohorts, 89% (172/194 proteins) and 95% (102/107) were consistent at significance levels of $P < 0.05$ and $q < 0.05$, respectively. Likewise, comparing the Sweden and Berlin cohorts 95% (70/74) and 100% (16/16) were consistent applying the same criteria, respectively, equivalent to 93% (64/69) and 100% (14/14) comparing the Magdeburg/Kiel and Berlin cohort.

Furthermore, quantitative alterations of protein levels between AD and non-AD CSF were very consistent across the cohorts. AD/non-AD fold changes of proteins were highly correlated with Pearson's correlation coefficients at $r = 0.91$, $r = 0.80$, and $r = 0.90$ for the comparisons of Sweden and Magdeburg/Kiel, Sweden and Berlin, and Magdeburg/Kiel and Berlin, respectively (Fig 3C–E).

We assessed whether AD and non-AD samples clustered together independent of the cohort, based on either the global unfiltered CSF proteome profile, the less stringent ($P < 0.05$) intersection, or the more stringent ($q < 0.05$) intersection set of proteins significant in all three cohorts. After Z-scoring protein intensities within cohorts, unsupervised clustering clearly separated AD from non-AD groups in all three cases (global proteome, both intersection sets; Fig 3F and G, Appendix Fig S2A and B). In the $P < 0.05$ intersection set, 40 out of 43 proteins (93%) differed consistently in abundance by AD status, 35 of which had an elevated abundance in AD CSF and five an elevated abundance in non-AD CSF (Fig 3F, Appendix Fig S3A and B). We discuss these

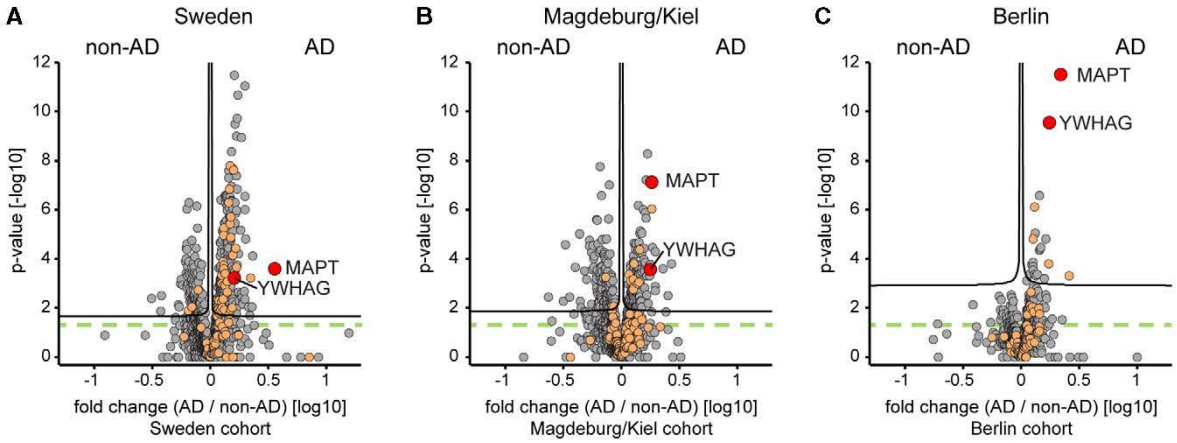


Figure 2. Differences in AD vs. non-AD CSF proteome in the three cohorts.

A–C Protein AD/non-AD fold changes plotted vs. statistical significance for Sweden (A), Magdeburg/Kiel (B), and Berlin (C) cohorts. Proteins associated with the GO annotation neuron projection labeled in orange. Proteins above the dashed green line are statistically significant ($P < 0.05$), and those above the black curves have a q-value below 0.05 (see Materials and Methods).

proteins as "the 40-protein signature" of AD in the remainder of this paper.

Next, we investigated if our results depended on the control groups in the Magdeburg/Kiel cohort and the Berlin cohort. The former controls were collected in Magdeburg or in Kiel, and in Berlin, the controls comprised subjective cognitive impairment patients and depression patients. Furthermore, the Kiel controls were younger than other cases or controls, and accordingly, their proteomes separated from the other non-AD controls (Fig EV1A, Appendix Fig S2A and B). Despite these differences, AD vs. non-AD fold changes of our 40-protein signature were independent of the specific non-AD control group subtype in these cohorts (Fig EV3A and B). To specifically investigate the effect of age and sex on the AD regulation of the 40-protein signature, we employed a linear regression model. After correction for age and sex in this way, the CSF abundance of all 40 proteins still significantly depends on AD status (Fig EV3C, Dataset EV7). Interestingly, CSF proteome alterations were of smaller magnitude in males compared to females in this study population.

Taken together, the "rectangular strategy" was able to discern AD-related alterations that reflect a small subset of the CSF proteome (< 50 proteins) from other cohort-specific effects comprising larger parts of the quantified CSF proteome (> 1,000 proteins) even in cohorts partially constrained by other biases such as age differences.

AD-associated proteins in CSF are linked to neurodegeneration

Many proteins among our 40-protein signature have known or suspected links to AD or other neurodegenerative diseases (Fig 3F). For instance, PARK7 (protein/nucleic acid deglycase DJ-1) and SOD1 (superoxide dismutase 1) are risk genes for Parkinson's disease and amyotrophic lateral sclerosis, respectively (Bonifati *et al*, 2003; Renton *et al*, 2014). Notably, the two cellular superoxide dismutases SOD1 and SOD2 were more abundant in AD CSF than in non-AD CSF, whereas the extracellular SOD3 was more abundant in non-AD CSF. Moreover, a genetic interaction of YWHAZ (14-3-3 protein ζ/δ) and BChE (buturyl cholinesterase) modulates the risk for AD (Mateo *et al*, 2008). CHI3L1 (protein YKL-40/chitinase-3-like protein 1), an astrocyte-derived protein, is elevated in AD CSF and discussed as a marker for progression from mild cognitive impairment to AD (Olsson *et al*, 2016; Baldacci *et al*, 2017). Similarly, fatty acid-binding protein 3 (FABP3) is elevated in AD CSF in our data and has been discussed as an AD CSF biomarker before (Sepe *et al*, 2018). CRYM (Ketimine-reductase mu-crystallin) has been reported as a modulator of huntingtin toxicity to striatal neurons in Huntington's disease (Francelle *et al*, 2015).

Proteins differing by AD status correlate with CSF t-tau abundance and MMSE score

As CSF composition reflects brain health, proteins in CSF may differ between AD and control subjects and additionally correlate with severity of AD pathology as reflected by classical clinical parameters such as t-tau abundance in CSF. Indeed, in the total dataset of 1,484 proteins, 124 proteins correlated significantly ($P < 0.05$) with t-tau concentration, 19 of which had a correlation q-value below 0.05 (Fig 4A–D, Appendix Fig S4A, Dataset EV4). All 124 proteins showed a consistent directionality of positive or negative correlation across the three cohorts. The abundance of tau as measured by MS correlated well with the ELISA measurements (Pearson $r = 0.82$ for Sweden, $r = 0.66$ for Magdeburg, $r = 0.68$ for Berlin).

We next asked how our 40-protein signature correlated with clinical t-tau measurements. Indeed, a large fraction—29 of 40 proteins—significantly correlated with t-tau in each of the three cohorts, and the directionality of change was also as expected for the non-significant proteins (Fig 4A–E, Appendix Fig S4A). This is a substantial enrichment over the numbers expected by chance in this dataset ($P < 0.0001$, odds ratios 37). Upon adjustment for age, sex, and cohort in a linear regression model comprising all three cohorts,

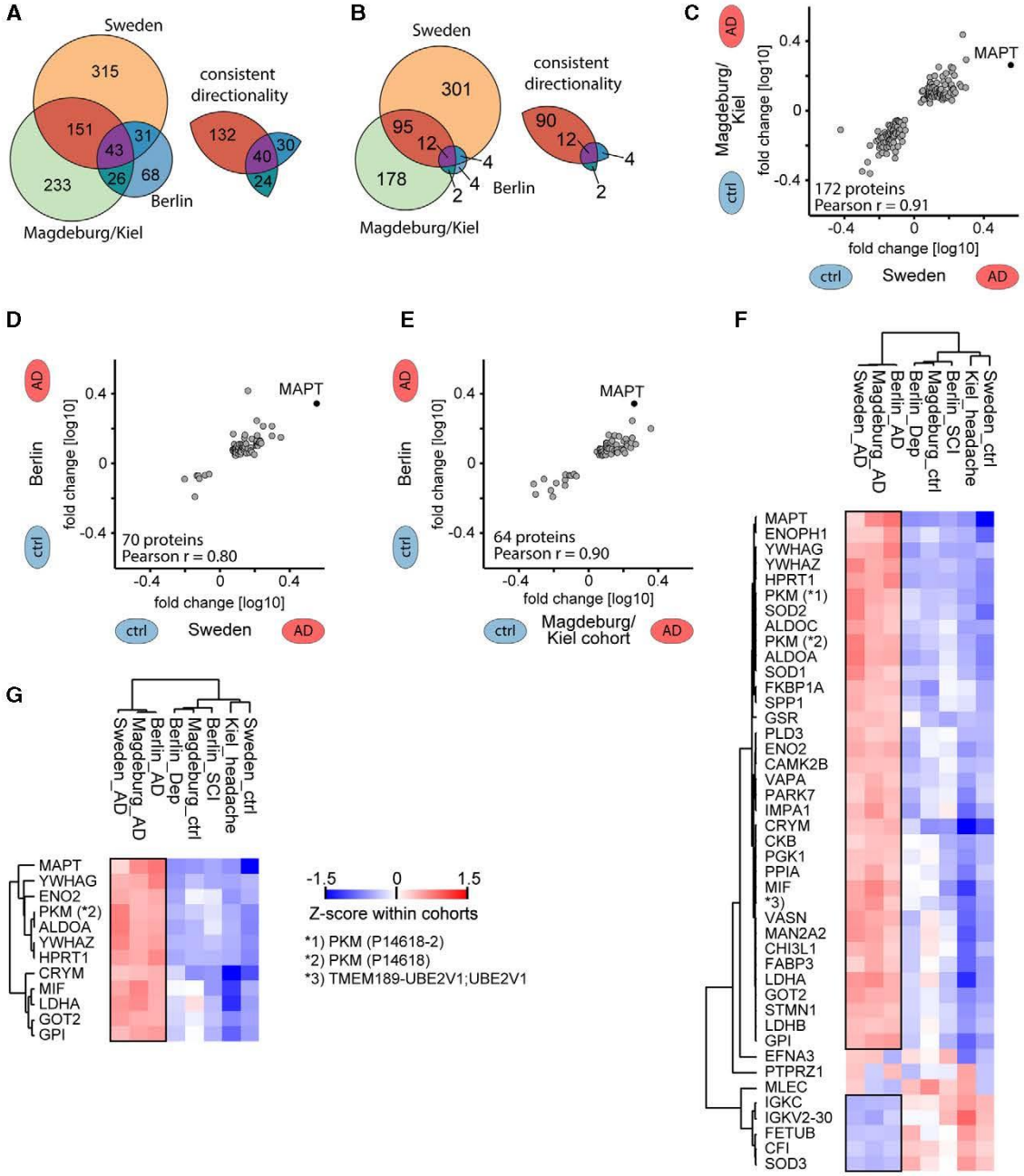


Figure 3. CSF proteome alterations across the three cohorts.

A, B On the left, number of proteins that differ significantly (*P*-value < 0.05 in A; *q*-value < 0.05 in B) in abundance by AD status within each cohort. On the right, number of proteins thereof that have a consistent directionality of either elevated or reduced abundance in AD CSF in pairwise comparisons of cohorts.

C–E Correlation of protein AD/non-AD fold changes in pairwise combinations of two cohorts each. Combinations are Sweden vs. Magdeburg/Kiel (C), Sweden vs. Berlin (D), and Magdeburg/Kiel vs. Berlin (E). Proteins included differ significantly ($P < 0.05$) and consistently in abundance by AD status in both cohorts each.

F, G Proteins that differ significantly ($P < 0.05$ in E; $q < 0.05$ in F) in abundance by AD status across all three cohorts. Z-scored abundances of proteins in the AD and non-AD groups of all cohorts shown by the heat map (see Materials and Methods). Hierarchical clustering separates AD from non-AD groups. Pyruvate kinase PKM (PKM) was quantified in two isoforms, and UniProt IDs are given in parentheses. Black frames highlight proteins with consistent AD/non-AD fold changes across cohorts.

all 40 proteins were significantly associated with t-tau (Materials and Methods; Appendix Fig S4B, Dataset EV7).

Some of these proteins, including fructose-bisphosphate aldolase A (ALDOA), superoxide dismutase 1 (SOD1), and YKL-40/chitinase-3-like protein 1 (CHI3L1), have previously been reported to correlate positively with CSF t-tau levels (Dayon *et al*, 2018).

In the clinic, AD is routinely diagnosed by biochemical parameters or by cognitive tests. We therefore investigated the relation

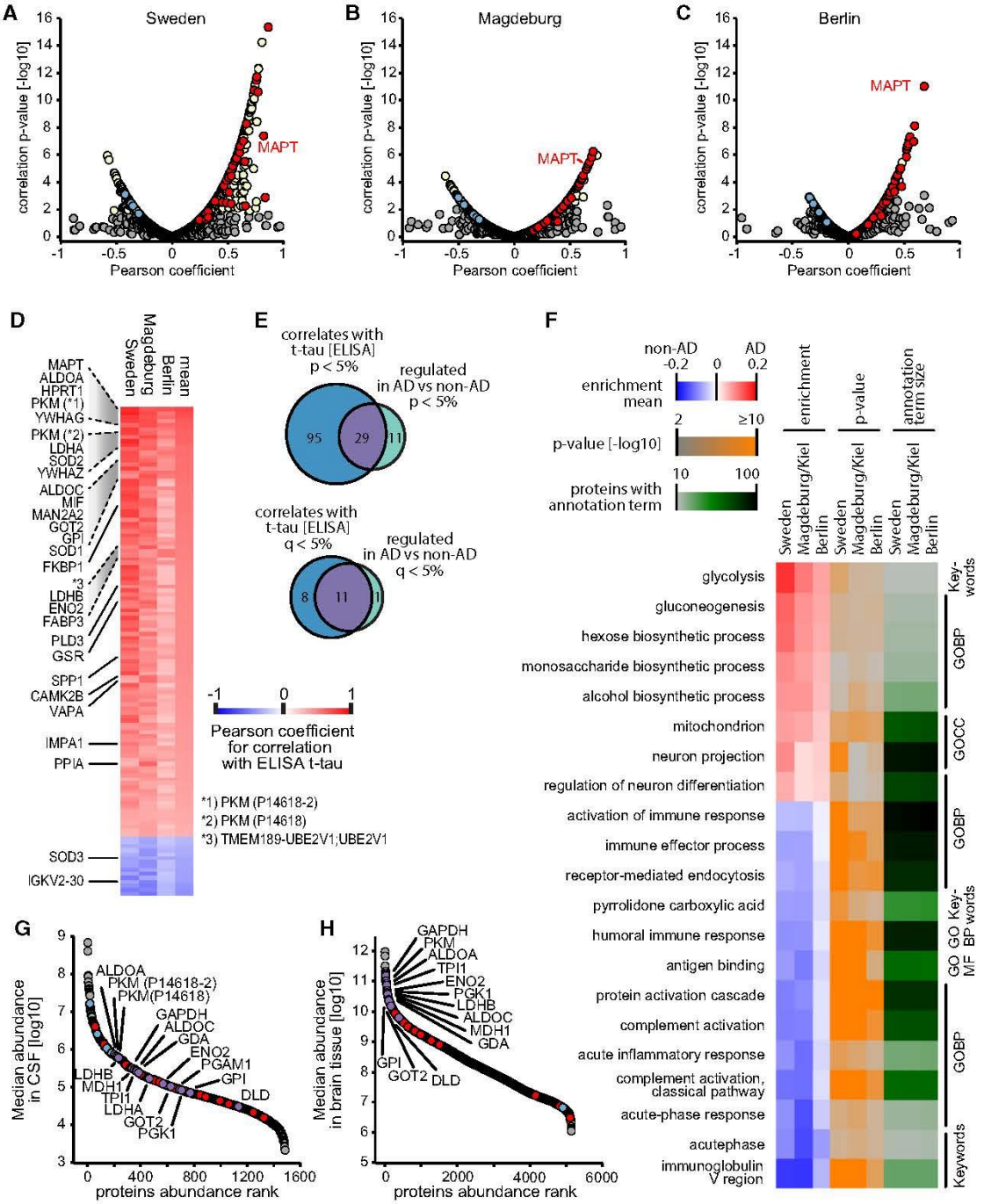


Figure 4.

Figure 4. Protein correlation with t-tau measurements and analysis of annotation term enrichment.

A–C Correlation of proteins with ELISA-measured t-tau concentration across samples within the Sweden (A), Magdeburg (B), and Berlin (C) cohorts. Proteins with a q-value below 0.05 are labeled in yellow. Proteins of the 40-protein signature are colored in red for those with higher abundance in AD CSF and in blue for those with higher abundance in non-AD CSF.

D Three-cohort summary of proteins significantly correlating with ELISA-measured t-tau. Protein names given for the 29 proteins out of the 40-protein signature with significant ($P < 0.05$) correlation in each of the three cohorts. Pyruvate kinase PKM (PKM) was quantified in two isoforms, and UniProt IDs are given in parentheses.

E Overlap of proteins significantly differing by AD status with proteins significantly correlating to ELISA-measured t-tau.

F Annotation enrichment in the AD versus non-AD fold change dimension. Terms with positive enrichment means are enriched in AD CSF over non-AD CSF. Conversely, terms with enrichment means below zero are enriched in non-AD compared with AD CSF. Annotations filtered for significance of enrichment ($P < 0.05$) and term size (10–100 proteins per term) in all three cohorts.

G, H Protein abundance distribution of CSF (G) and brain (H) showing the abundances of AD-modulated CSF proteins. Proteins of our 40-protein signature are highlighted in red (elevated abundance in AD) and blue (elevated abundance in non-AD). Proteins linked to glucose metabolism are highlighted in purple and labeled.

between our proteomics results and the mini-mental state examination (MMSE) scores as a measure of cognitive performance, which were assessed in the Berlin cohort (Fig EV1L). In the literature, reference population means of MMSE scores were 29, 27, and 20 for cognitively normal, mild cognitive impairment (MCI), and AD participants, respectively (Chapman *et al*, 2016), while the MMSE scores in the Berlin cohort were 27.7 ± 1.9 (mean ± SD) for non-AD and 22.7 ± 4.5 for AD. Tau (MAPT), osteopontin (SPP1), and 14-3-3γ (YWHAG) were the top three proteins inversely correlating with the MMSE score (Fig EV4A). Osteopontin has already been reported to inversely correlate with the MMSE score in AD (Comi *et al*, 2010). Moreover, in our 40-protein signature proteins with higher abundance in AD CSF correlated negatively with the MMSE score and vice versa.

When stratifying the Berlin cohort into "high MMSE score" and "low MMSE score" groups over a cutoff range from 29 to 21, we obtained the greatest separation at a cutoff of 25. Reassuringly, MAPT and YWHAG were the top outliers and our 40-protein signature showed the expected association with the MMSE groups at all cutoff values in spite of the limited diagnostic performance of the MMSE evaluation (Fig EV4B–F) (Perneczky *et al*, 2006; Mitchell, 2009; Arevalo-Rodriguez *et al*, 2015). Thus, CSF protein signatures linked to biochemically defined AD also associate with cognitive performance.

Neuronal and glycolytic signature in AD CSF

To identify biological signatures in the AD-associated proteome alterations, we performed an annotation enrichment analysis of functional terms (GO biological process, GO cellular compartment, UniProt Keywords) in the global proteome AD/non-AD fold changes. We obtained 21 annotation terms below a P-value of 0.05, all of which showed consistency across the three cohorts (Materials and Methods, Fig 4F). Terms including "neuron projection" and "regulation of neuron differentiation" underline the neuronal signature in the AD CSF proteome. Interestingly, glycolysis and gluconeogenesis presented as top terms with enrichment in AD CSF in this unbiased analysis. This concurs with the presence of glycolytic proteins in our 40-protein signature. These include fructose-bisphosphate aldolase A (ALDOA) and C (ALDOC), pyruvate kinase PKM (PKM), γ-enolase (ENO2), aspartate aminotransferase, mitochondrial (GOT2), phosphoglycerate kinase 1 (PGK1), L-lactate dehydrogenase A chain (LDHA), and B chain (LDHB) (Fig 3F). Moreover, other glycolytic proteins in the dataset not passing the significance cutoffs nevertheless uniformly followed the same trend of elevated abundances in AD CSF (Appendix Fig S5). Glycolytic proteins may originate from astrocytes as glycolysis in the brain is mainly performed by these cells to provide lactate for oxidative phosphorylation in neurons (Bélanger *et al*, 2011; preprint: Higginbotham *et al*, 2019). Furthermore, the GO cellular compartment annotation term "mitochondrion" was also enriched in AD CSF, and mitochondrial dysfunction is a known hallmark of AD (Querfurth & LaFerla, 2010). When we mapped the up-regulated proteins of our 40-protein signature onto a deep human brain proteome (Carlyle *et al*, 2017), their corresponding abundance in brain was generally in the more abundant range (Fig 4G and H). This observation is consistent with mechanisms in which cellular proteins are released into the CSF by tissue damage-associated loss of membrane integrity, exosome release, or others.

Further confirmation of AD-associated proteome alterations in an independent cohort

After completion of our study, a related preprint appeared (preprint: Higginbotham *et al*, 2019). Similarly to our study, the authors investigated proteomic profiles in a study of 20 AD cases and 20 controls, although they used a different experimental workflow. CSF samples were depleted, digested, chemically labeled for multiplexing by an isobaric tag, fractionated, and analyzed by mass spectrometry, achieving a remarkable depth of quantitation. A second cohort, consisting of 33 AD and 32 controls and 30 asymptomatic cases, was also measured, although with a somewhat different method and a reduced proteome depth. Many AD-associated CSF signatures observed in our study including the glycolytic signature, the neuronal signature, and the 14-3-3 protein signature are also reported in the manuscript. This provides additional evidence for these signatures to be AD-associated from independent cohorts identified by a different experimental approach.

To determine a panel of consistently AD-regulated proteins and to assess inter-study consistency in more detail, we downloaded the available data and compared them to our data. As tau was not contained in the second cohort dataset and only 31 proteins significantly differed by AD status in both cohorts of that independent study, we limited our comparison to the 20 AD cases versus 20 controls cohort by Higginbotham *et al*. This dataset contained 2,875 proteins quantified in at least half of the samples and 528 proteins thereof differed significantly ($P < 0.05$) by AD status. Notably,

despite the different proteome depth the number of proteins that differed by AD status is similar to the proteins that differed significantly by AD status in the Sweden (540) and Magdeburg/Kiel (453) cohorts. These similar numbers in three out of four cohorts suggest that both proteomic approaches cover a substantial part of the CSF proteome signature related to AD.

Out of our 40-protein signature, 38 proteins were contained in the dataset of this independent study and 26 of 38 (68%) thereof were also significant (Fig EV5A, Dataset EV4). This is a highly significant enrichment among all significant proteins in the dataset of that independent study (odds ratio 10, $P < 0.0001$, Fig EV5B). The directionality of abundance elevation in either AD or non-AD CSF was consistent across studies for these 26 core proteins (Fig EV5C). Moreover, quantitative fold change agreement was high (Pearson $r = 0.76$; Fig EV5D). Among these 26 proteins, only one protein, fetuin-B (FETUB), had an elevated abundance in non-AD CSF, while 25 proteins were elevated in AD CSF including tau, glycolysis-related proteins, 14-3-3 proteins, protein/nucleic acid deglycase DJ-1 (PARK7), superoxide dismutase 1 (SOD1), fatty acid-binding protein 3 (FABP3) and hypoxanthine-guanine phosphoribosyltransferase (HPRT1). Taken together, AD-associated protein signatures identified in our work are validated in a completely separate study using an independent cohort and different experimental strategy.

AD classification by machine learning on the CSF signature

Next, we next assessed if the MS intensities of the set of 26 core proteins which overlap between our and the Higginbotham studies could be applied to classify participants by AD status using machine learning and we explored a variety of machine learning models. First, to determine feature importance, we employed a decision tree and found that a model with a maximum depth of six levels, using the intensities of 14 proteins could correctly classify the participants in the three studies by AD status. A visualization of the decision tree revealed that levels of tau itself were at the root, followed by the glycolytic enzyme pyruvate kinase PKM (PKM), and macrophage migration inhibitory factor (MIF) at the next level (Fig 5A). As protein intensities are correlated, a decision tree could potentially rank proteins differently depending on its initial state. However, when repeatedly training the decision tree ($n = 10{,}000$) with random initial states and also shuffling the dataset, the root of the tree remained similar (MAPT at rank 1 in all cases, PKM at rank 2 or 3 in 82.8%, and MIF at rank 2 or 3 in 84.3% of all cases, respectively). This underlines the importance of these three proteins among the CSF proteome as indicators of AD.

To test models for generalizability, we considered several tree-based ensemble methods. We trained six commonly used methods (AdaBoost, Bagging, ExtraTrees, GradientBoosting, RandomForest, and XGBoost) on the intensities of the 14 proteins selected by the decision tree above such that the tree needed to completely classify the participants. The protein intensities were randomly shuffled and split using a k-folds cross-validator ($k = 6$) into six training/test splits. Accordingly, shuffling entailed mixing of patients from different cohorts but each sample was in the testing dataset exactly once. For each method, we performed cross-validation and determined a receiver operating characteristic (ROC) curve.

All classifiers reached an area under the ROC curve (AUC) of at least 0.84. XGBoost had the best performance with a mean AUC of 0.91 and was selected for further analysis. To determine the optimal number of features, we iteratively added them in them in their order of importance in the decision tree. The overall model performance increased with the number of proteins and reached a plateau at six proteins (MAPT, PKM [P14618-2 isoform], MIF, IMPA1, YWHAZ, and ALDOC), which we selected for the final model.

To assess the performance of our final model as a predictive test we again used k-fold cross-validation in six different training/test splits. The different splits exhibited good agreement with each other at AUC's ranging from 0.87 to 0.98, indicating robustness of classification (Fig 5B). We then determined the overall confusion matrix combining the six splits ("net reclassification", the number of correctly and incorrectly classified participants) (Fig 5C). In total, 72 out of 88 AD patients and 95 out of 109 non-AD patients were correctly identified, corresponding to a sensitivity of 82% and a specificity of 87%.

Discussion

We have combined advanced sample preparation, cutting-edge mass spectrometry hardware, acquisition schemes, MS data processing and bioinformatic analysis and optimized it for CSF to build a high-performance CSF proteomics workflow amenable to high-throughput and large cohorts. About 1,500 proteins can be quantified and over 1,000 with intra- and inter-assay coefficients of variation (CVs) below 20%. Using this technology, we identified known biomarkers such as tau as top candidates as well as a range of novel potential biomarkers. Harnessing this pipeline, we compared AD and non-AD CSF in three independent cohorts. This led to a 40-protein signature whose members are consistently up- or down-regulated in AD CSF vs. non-AD CSF across the three cohorts.

Cases and controls in two of our cohorts separated better on the basis of clinical AD CSF biomarker concentrations (t-tau, p-tau$_{181}$, Aβ_{1-42}, Aβ_{1-40}) than in the third one. Likewise, AD-associated differences in the CSF proteome were smaller and fewer protein alterations were statistically significant in that third cohort. The attenuated separation according to clinical CSF values suggests that this third cohort comprised milder AD cases and early-stage AD patients in the non-AD group just below the cutoff values. This would lead to the attenuated overall differences in the CSF proteome profile between the AD and the non-AD groups that we observe.

There is no universally accepted AD classification system; however, various different integration schemes of clinical AD CSF biomarkers have been explored (Bloudek *et al*, 2011; Ferreira *et al*, 2014; Ritchie *et al*, 2017). Using the Hulstaert index, a variation of the Aβ_{1-42}/t-tau ratio, for AD classification of the three cohorts we obtained largely the same, but fewer statistically significant potential marker proteins compared to our uniform AD classification (Appendix Fig S6A–D, Materials and Methods) (Hulstaert *et al*, 1999; Molinuevo *et al*, 2013; Vos *et al*, 2013). Furthermore, the mini-mental state examination (MMSE) cognitive test was performed in one of our cohorts. It was encouraging to find the proteomic outliers identified by analysis of biochemically defined AD CSF to be associated with the MMSE score performance.

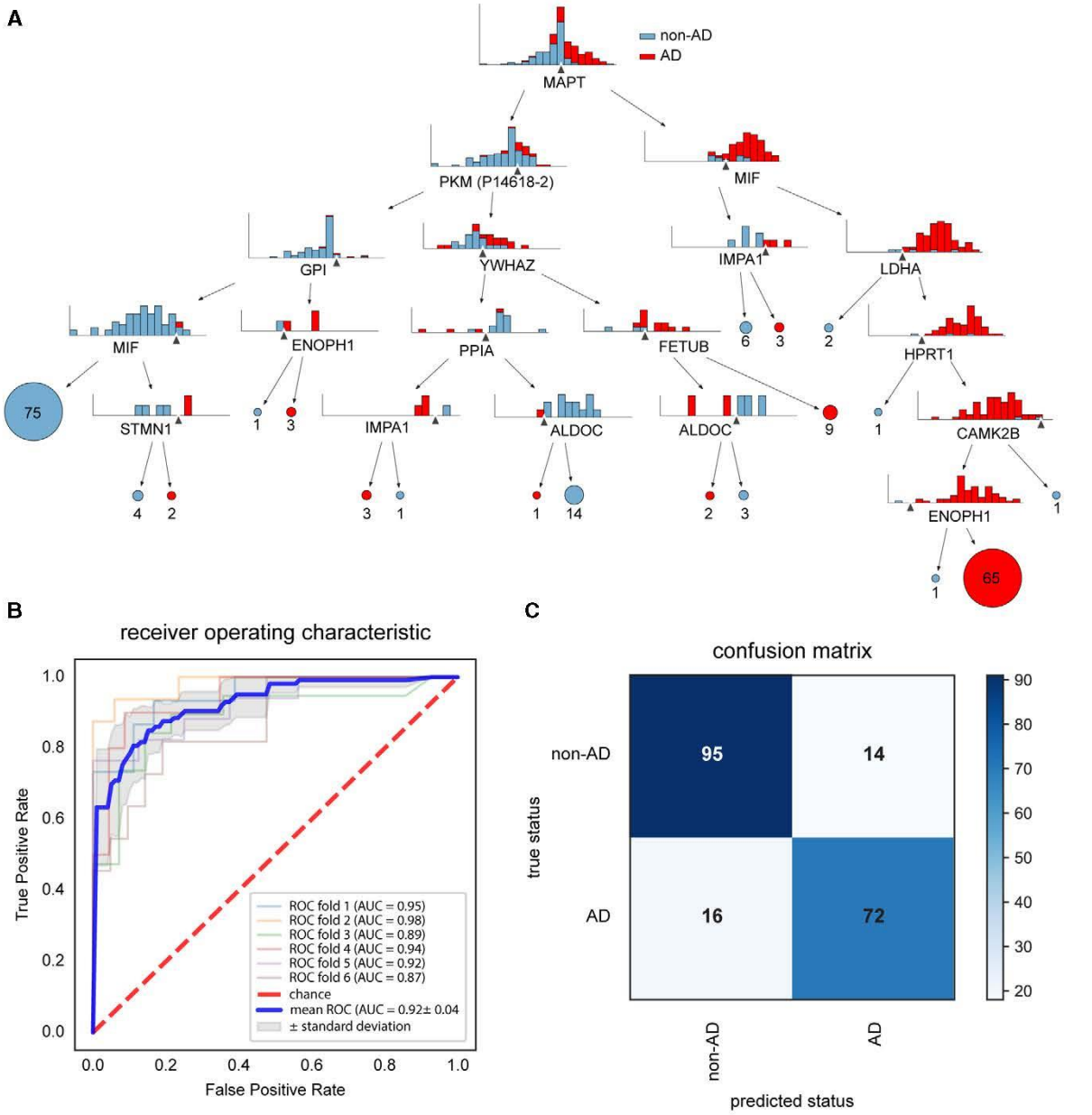


Figure 5. Machine learning separates AD from non-AD CSF at high performance.

A Decision tree to classify AD vs. non-AD participants based on the protein levels of a core 26 protein set. Splits are indicated by black triangles. A tree with a minimum depth of six can correctly classify the participants by AD status.

B Receiver operating characteristic (ROC) curve for the model based on XGBoost. The diagonal line indicates random performance. Blue line represents the mean performance of the model when trained on six stratified train—test splits (k-fold). The gray areas represent the standard deviation of ROC values.

C Confusion matrix indicating model performance when predicted on the test split of the cross-validation. Overall accuracy is 0.85.

Another general challenge in biomarker discovery studies are cohort-specific effects. This relates particularly to multi-centric studies with distinct inclusion criteria for cases and controls. Despite cohort-specific effects and attenuated AD/non-AD differences in the third cohort of our study, proteins that statistically significantly differed by AD status in multiple cohorts exhibited very good qualitative and quantitative cross-cohort agreement in their AD modulation. A signature of 40 CSF proteins was consistently associated with AD status and showed high correlation values of protein fold changes across cohorts. When further combined with a recent,

independent effort on bioRxiv (preprint: Higginbotham *et al*, 2019; Johnson *et al*, 2020), which used different MS technology, this resulted in a set of 26 core proteins consistent across four independent cohorts. This highlights the power of the "rectangular strategy" study design in discerning cohort-specific from pathological effects for biomarker discovery.

Our relatively large dataset with nearly 200 participants prompted us to explore machine learning for the purpose of assessing AD status on the basis of the levels of the 26 core proteins. We found that an ensemble method-based classifier reached high specificity (87%) and sensitivity (82%), while showing promising generalizability. Intriguingly, tau itself, one of the glycolysis-related proteins, and an immunological factor were selected by the machine learning algorithm as the most important features for classification, proving further validation of our biomarker panel and biomarker identification pipeline. The modeling also indicated that additional and more uniform training data could further improve diagnostic performance. Furthermore, additional clinical data, such as cognitive assessments, can naturally be incorporated in this framework.

In the list of the 26 core proteins, several have known links to neurodegeneration such as protein/nucleic acid deglycase DJ-1 (PARK7) and superoxide dismutase 1 (SOD1) or genetic interaction links to AD like 14-3-3 protein ζ/δ (YWHAZ) (Bonifati *et al*, 2003; Mateo *et al*, 2008; Renton *et al*, 2014). Likewise, the set also contains the tentative AD biomarker CHI3L1 (protein YKL-40) likely reflecting astrocytic activation (Olsson *et al*, 2016; Baldacci *et al*, 2017). Moreover, we identify a number of glucose metabolism-associated proteins elevated in AD CSF in line with other reports (Dayon *et al*, 2018; preprint: Higginbotham *et al*, 2019; Sathe *et al*, 2019). These glycolytic proteins and other AD-associated proteins in CSF are highly abundant in brain and could be released into CSF from brain tissue. Regardless of the mechanism of accumulation in the CSF, the utility of abundant cellular proteins as markers is generally accepted in clinical practice. In the plasma proteome, this is demonstrated by troponin levels indicative of acute myocardial infarction (Keshishian *et al*, 2015) and liver proteins indicative of fatty liver disease (Niu *et al*, 2019).

The fact that CSF proteomics is now able to detect brain-derived proteins and determine protein signatures consistent across multiple independent multi-centric cohorts sets the stage for future biomarker discovery studies in neurodegenerative diseases. Next steps should include investigating the added diagnostic value of the AD CSF protein signature when combined with established diagnostic criteria in the clinic, preferably in a machine learning framework. Further, we speculate that the workflow presented here would be highly suited for the discovery of additional clinically and etiologically relevant biomarkers. There is a great need for early diagnosis, prognosis, and treatment efficacy biomarkers (Winblad *et al*, 2016). Further studies are warranted assessing the relevance of these proteins in prospective studies of dementia-free individuals in midlife with repeated brain imaging, cognitive testing, and long-term follow-up for dementia incidence. Recent developments in MS-based proteomics now enable fast and efficient quantitative readout of relatively large panels of proteins in a targeted or "globally targeted" manner (Abbatiello *et al*, 2013; Wichmann *et al*, 2019). This may enable the use of MS-based proteomics not only for the discovery of disease-associated protein patterns but also for routine clinical tests (Geyer *et al*, 2017).

Materials and Methods

Study populations

Three cohorts of AD and non-AD control CSF samples were obtained, one from Sweden, one originating from the German cities of Magdeburg and Kiel (through the Harvard T. H. Chan School of Public Health), and one from Berlin. The CSF concentration values of the clinical AD biomarkers t-tau, p-tau$_{181}$, Aβ_{1-42}, and Aβ_{1-40} were available as follows: t-tau, p-tau$_{181}$, Aβ_{1-42} for the Sweden cohort; t-tau, p-tau$_{181}$, Aβ_{1-42}, and Aβ_{1-40} for the Magdeburg cohort; and t-tau, Aβ_{1-42}, and Aβ_{1-40} for the Berlin cohort.

Sweden CSF samples were obtained from patients with cognitive impairment at several memory clinics in western Sweden. De-identified diagnostic remnant CSF material was used in this study, which was approved by the Gothenburg ethics committee. The AD and non-AD groups as classified by the primary AD criteria of this study were well separated biochemically based on the clinical AD CSF biomarkers. CSF biomarker levels were measured using the INNOTEST assays (Fujirebio, Ghent, Belgium) in the Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden, by board-certified laboratory technicians who were blinded to clinical data. The laboratory procedures were accredited by the Swedish Board for Accreditation and Conformity Assessment (SWEDAC).

Magdeburg CSF samples originated from patients at the outpatient memory clinic at the Otto-von-Guericke University Magdeburg. CSF biomarker levels were measured at the site of collection using commercially available INNOTEST ELISA kits (Fujirebio, Ghent, Belgium). The AD and non-AD groups as defined by our primary AD classification criteria were well separated biochemically based on the clinical AD CSF biomarkers. The local ethics committee approved the use of the CSF samples. Additional control samples from Kiel were acquired from patients treated at the emergency department at the University Hospital Schleswig-Holstein. Informed consent for scientific analysis of diagnostic remnant samples collected for clinical care and ethics committee approval for use of the samples were obtained.

Berlin CSF samples were obtained from patients at the Memory Clinic of Charité Universitätsmedizin Berlin. The clinical AD biomarkers t-tau, Aβ_{1-42}, and Aβ_{1-40} were measured at the site of collection. The V-PLEX Aβ Peptide Panel 1 (6E10) Kit (Meso Scale Diagnostics, Rockville, MD, USA) was used for Aβ peptide quantitation and the INNOTEST hTAU Ag (Fujirebio Germany GmbH, Hannover, Germany) for tau. The AD and non-AD groups as defined by our primary AD classification criteria were moderately separated biochemically based on the clinical AD CSF biomarkers. CSF collection was standardized as described elsewhere (Schipke *et al*, 2011). The local ethics committee approved the use of the CSF samples. All participants provided written informed consent.

Primary AD classification

To enable uniform analysis, we standardized classification of AD and non-AD for the different cohorts uniformly based on the CSF concentrations of t-tau, Aβ_{1-42}, and Aβ_{1-40} for the Sweden, Magdeburg, and Berlin cohorts. Patients were classified as AD if the t-tau concentration was above 400 ng/l and the Aβ_{1-42} concentration

below 550 ng/l or the $A\beta_{1-42}/A\beta_{1-40}$ ratio was below 0.065. The t-tau criterion and at least one of the two Aβ criteria had to be met for a patient to be classified as having AD and patients were classified as not having AD otherwise. The classification here is derived from a classification according to the cutoffs of t-tau being higher than 400 ng/l, p-tau_{181} higher than 60 ng/l, and $A\beta_{1-42}$ lower than 550 ng/l (Sjogren *et al*, 2001; Hansson *et al*, 2006). We additionally included the CSF $A\beta_{1-42}/A\beta_{1-40}$ ratio as it has a superior diagnostic performance than the $A\beta_{1-42}$ concentration alone (Spies *et al*, 2010; Dubois *et al*, 2014; Niemantsverdriet *et al*, 2017). Participants with missing information on the CSF t-tau or $A\beta_{1-42}$ concentration in the Sweden, Magdeburg, or Berlin cohort were excluded. Kiel CSF samples originated from young patients (32.0 ± 17.1 years, median ± SD) treated at an emergency department with no indications of AD or other neurodegenerative diseases. Thus, we included these samples as non-AD controls despite the missing clinical biomarker CSF concentrations.

Hulstaert index

The Hulstaert index for AD classification is a variant of the $A\beta_{1-42}$/t-tau ratio with improved diagnostic performance (Molinuevo *et al*, 2013). It is calculated as $A\beta_{1-42}/(240 + (1.18*\text{t-tau}))$ using ng/l concentrations, and samples below a cutoff value of one are classified as AD (Hulstaert *et al*, 1999). We performed an independent analysis using the Hulstaert index instead of our uniform classification. As shown in Appendix Fig S6, the results overlap almost completely; however, the Hulstaert index, although less stringent in AD inclusion, leads to a smaller number of significantly different proteins.

Clinical AD diagnosis

Information about clinical AD status, i.e. the diagnosis of symptomatic AD according to site-specific criteria, was available for the Magdeburg, Kiel, and Berlin cohorts. At these sites, clinical AD diagnoses had been reached by assessing the clinical presentation of patients according to distinct guidelines.

In Magdeburg, the clinical AD diagnosis was based on the patient's clinical presentation using the National Institute of Neurological and Communicative Disorders and Stroke—Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria (McKhann *et al*, 2011). The clinical evaluation included the CERAD (Consortium to Establish a Registry for Alzheimer's Disease) neuropsychological test battery and magnetic resonance imaging (Morris *et al*, 1989). AD and control subjects had no clinical signs of stroke, epilepsy, or other neurodegenerative diseases. For the clinical diagnosis of AD, local concentration cutoffs for core AD biomarkers were used; however, fulfillment of the cutoff criteria was considered indicative but not sufficient for an AD diagnosis which also depended on the patient's clinical presentation. AD was considered likely if the criteria p-$tau_{181} > 80$ ng/l and t-tau > 450 ng/l were simultaneously met. Likewise, AD was considered likely if the criteria $A\beta_{1-42} < 485$ ng/l and the amyloid ratio $A\beta_{1-42}/A\beta_{1-40} < 0.06$. Non-AD control patients underwent CSF withdrawal to exclude neuroinflammation and dementia. Control subjects showed no signs of neurodegeneration and had normal CSF parameters regarding cell count, protein content, and lactate concentration. All but one of 26 biochemically defined AD cases according to our primary AD classification study also had a clinical AD diagnosis, while none of the non-AD controls had a clinical AD diagnosis.

Kiel CSF samples originated from patients presenting with acute headache. No patient had an AD diagnosis or showed clinical indications of neurodegenerative diseases. CSF sampling was performed to exclude meningitis which is not present in any subject in this study. Subjects with a history of dementia, systemic or CSF inflammatory signs, and blood–brain barrier dysfunction (CSF-to-serum albumin ratios $\geq 9 \times 10^3$) were excluded, and clinical diagnoses were diverse and predominantly migraine, headache, common cold or sinusitis or skin sensation disturbance (Koch *et al*, 2017).

In Berlin, patients were diagnosed as having AD based on the clinical presentation according to the American Psychiatric Association guidelines, the Diagnostic and Statistical Manual of Mental Disorders (DSM), version DSM-5. Diagnoses were reached at a consensus panel composed of psychiatrists, neurobiologists, and neuropsychologists according to the DSM-5. Specifically, patients' relevant medical history, standard cognitive and functional measurements (e.g., MMSE), CSF biomarker values for t-tau and amyloid peptides, and cMRI findings were examined in parallel. For the clinical diagnosis of AD, site-specific CSF concentration cutoffs for core AD biomarkers were used. Under these conditions, the following CSF biomarker values were rated as indicative of AD: $A\beta_{1-42} < 600$ ng/l or $A\beta_{1-42}/A\beta_{1-40}$ ratio ≤ 0.060 (in 2014 and before) or $A\beta_{1-42}/A\beta_{1-40}$ ratio ≤ 0.065 (from 2015 on), in addition to t-tau > 350 ng/l. Again, however, fulfillment of these cutoff criteria was considered indicative but not sufficient for an AD diagnosis which also depended on the patient's clinical presentation. Out of 33 biochemically defined AD cases according to of our primary AD classification, 24 also had a clinical AD diagnosis at the time of CSF withdrawal, while none of the non-AD controls had a clinical AD diagnosis. For three of the nine biochemically defined AD cases without a clinical AD diagnosis, the medical records included additional clinical information or information collected months to years after the CSF withdrawal as the patient returned to the clinic again. These three patients either developed clinical AD within 2 years, presented with mild cognitive deficiencies of the AD type or a "not yet specified neurodegenerative disease".

Sample preparation

The sample preparation was optimized for CSF on the basis of our Plasma Proteome Profiling workflow (Geyer *et al*, 2016). CSF was aliquoted in 96-well plates and processed with an automated set-up on an Agilent Bravo liquid handling platform. In total, 40 µl of CSF was mixed with 40 µl PreOmics lysis buffer (PreOmics GmbH) for reduction of disulfide bridges, cysteine alkylation, and protein denaturation at 95°C for 10 min. After a 10-min cooling step, 0.2 µg trypsin and 0.2 µg LysC were added to each sample and digestion was performed at 37°C for 4 h. Peptides were purified on two 14-gauge StageTip plugs according to the PreOmics iST protocol (https://preomics.com/products). The StageTips were centrifuged using an in-house 3D-printed StageTip tray at 1,500 g for washing and elution. The eluate was completely dried using a SpeedVac centrifuge at 45°C (Eppendorf, Concentrator plus), resuspended in

10 µl buffer A* (2% v/v acetonitrile, 0.1% v/v trifluoroacetic acid, and stored at −20°C. Upon thawing, samples were shaken for 1 min at 2,000 rpm (thermomixer C, Eppendorf). Peptides were then subjected to LC-MS/MS analysis.

Additionally, for library generation for the DIA measurements, peptides of the Sweden cohort were pooled into one AD sample pool and one non-AD sample pool of 75 µg each. Peptide concentration was measured spectroscopically by absorbance at 280 nm (Nanodrop 2000, Thermo Scientific). The AD sample pool and the non-AD sample pool were fractionated into 24 fractions each by high-pH reversed-phase chromatography on the "spider fractionator" (Kulak *et al*, 2017). Fractions were completely dried and resuspended in 10 µl buffer A*. To determine coefficients of variation, five aliquots of a CSF pool on one plate were subjected to sample preparation (intra-plate) and this was repeated on three different days (inter-plate).

Mass spectrometry analysis

Samples were measured using an EASY-nLC 1200 (Thermo Fisher Scientific) coupled to a Q Exactive HF-X Orbitrap mass spectrometer (Thermo Fisher Scientific) via a nano-electrospray ion source (Thermo Fisher Scientific). Purified peptides were separated on 50 cm UHPLC columns with an inner diameter of 75 µm packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH). In total, 500 ng of purified peptide in buffer A* was loaded onto the column in buffer A (0.1% v/v formic acid) and eluted at a flow rate of 300 nl/min and a temperature of 60°C by a linear 80-min gradient from 5% to 30% buffer B (0.1% v/v formic acid, 80% v/v acetonitrile), followed by a 4-min increase to 60% B, a further 4-min increase to 95% B, a 4-min plateau phase at 95% B, a 4-min decrease to 5% B, and a 4-min wash phase of 5% B. To acquire MS data, the data-independent acquisition (DIA) scan mode was used for single-shot patient samples, whereas the fractionated samples of the AD pool and non-AD pool were acquired with a top12 data-dependent acquisition (DDA) scan mode. Both acquisition schemes were combined with the same liquid chromatography gradient. The mass spectrometer was operated by the Xcalibur software (Thermo Fisher). DDA scan settings on full MS level included an ion target value of 3×10^6 charges in the 300–1,650 m/z range with a maximum injection time of 20 ms and a resolution of 60,000 at m/z 200. At the MS/MS level, the target value was 10^5 charges with a maximum injection time of 60 ms and a resolution of 15,000 at m/z 200. For MS/MS events only, precursor ions with 2–5 charges that were not on the 20 s dynamic exclusion list were isolated in a 1.4 m/z window. Fragmentation was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27 eV. DIA was performed with one full MS event followed by 33 MS/MS windows in one cycle resulting in a cycle time of 2.7 s. The full MS settings included an ion target value of 3×10^6 charges in the 300–1,650 m/z range with a maximum injection time of 60 ms and a resolution of 120,000 at m/z 200. DIA precursor windows ranged from 300.5 m/z (lower boundary of the first window) to 1649.5 m/z (upper boundary of the 33rd window). MS/MS settings included an ion target value of 3×10^6 charges for the precursor window with an Xcalibur-automated maximum injection time and a resolution of 30,000 at m/z 200.

Mass spectrometry data processing

The MS data of the fractionated pools (DDA MS data, 24 AD fractions, 24 non-AD fractions) and the single-shot subject samples (DIA MS data, all samples from all three cohorts) were used to generate a DDA-library and direct-DIA-library, respectively, which were computationally merged into a hybrid library in the SpectroMine software, version 1.0.21621.8.15296 (Biognosys AG, Schlieren, Switzerland). The hybrid library contained 33,392 precursors linked to 23,855 unique peptides considering peptide modifications or 17,301 unique peptides based on the amino acid sequence corresponding to 2,733 protein groups. The hybrid spectral library was used to search the MS data of the single-shot patient samples in the Spectronaut software, version 12.0.20491.9.26669 (Biognosys AG), for final protein identification and quantitation. All searches were performed against the human UniProt reference proteome of canonical and isoform sequences with 93,786 entries downloaded in March 2018. Searches used carbamidomethylation as fixed modification and acetylation of the protein N-terminus, oxidation of methionines and deamidation of asparagine or glutamine as variable modifications. Default settings were used for other parameters. In brief, a trypsin/P proteolytic cleavage rule was used, permitting a maximum of two miscleavages and a peptide length of 7–52 amino acids. Protein intensities were normalized using the "Local Normalization" algorithm in Spectronaut based on a local regression model (Callister *et al*, 2006). Spectral library generation stipulated a minimum of three fragments per peptide, and maximally, the six best fragments were included. A protein and precursor FDR of 1% were used and protein quantities were reported in samples only if the protein passed the filter ("Q-value sparse" mode data filtering).

Bioinformatics data analysis

Data analysis was mainly performed in the Perseus environment version 1.6.1.3 but also in version 1.6.0.9 for correlation analysis and version 1.5.2.11 for Venn diagram analysis (Tyanova *et al*, 2016). Proteins with < 20 observations across the entire dataset were excluded, reducing the dataset from 1,542 to 1,484 proteins. Protein intensities were $\log_{10}$-transformed for further analysis, apart from correlation and coefficient of variation analysis. All t-tests performed were two-sided and unpaired. False discovery rate (FDR) control to account for multiple hypothesis testing in statistical tests was performed by a permutation-based model in conjunction with a SAM-statistic with an s0-parameter of 0.001 (Tusher *et al*, 2002). Annotation term enrichment was performed with the 1D enrichment tool in Perseus separately for each cohort (Cox & Mann, 2012). Annotation terms were filtered for terms with a *P*-value cutoff of 0.5% in each cohort. Moreover, terms comprising less than 10 or more than 100 proteins in our dataset of 1,484 proteins were excluded because we found that annotation enrichment analysis is often dominated by very small or large but not meaningful terms. Hierarchical clusters were generated using the built-in tool in Perseus. When protein abundances were reported on the group level (e.g. Sweden AD), Z-scoring across samples either within the cohort or across cohorts (for all 197 samples) was performed as stated in the figure legends and the median Z-score was taken as group abundance. Sample groups (e.g. Sweden AD) were clustered based on

Pearson's correlation coefficient, while proteins were clustered based on Euclidian distance unless ranked by the three-cohort mean.

A deep human brain proteome was used for comparison to the CSF proteome, and 753 proteins were matched based on ensemble identifiers (Carlyle *et al*, 2017, supplementary table 5). For generation of the abundance distribution curves, median protein abundances across all samples within a proteome were used. For the comparison of AD CSF proteomes with the independent report (preprint: Higginbotham *et al*, 2019), data for the CSF1 dataset were downloaded from bioRxiv.org. Proteins were matched to our data based on UniProt protein identifiers, apart from of MAPT, ALDOA, and SOD2 which were matched based on gene names. Fisher's exact test in combination with the Baptista–Pike method was used in GraphPad Prism version 7.03 to assess the significance of enrichment and odds ratios in contingency table settings. This included the analysis of association of t-tau concentration-correlated proteins with proteins differing by AD status and the analysis of enrichment of AD-regulated proteins identified in this study among the proteins differing significantly ($P < 0.05$) by AD status in the Higginbotham CSF1 dataset.

We used linear regression analysis computed in RStudio version 1.2.5033 using R version 3.6.3 and assessed the association of $\log_{10}$-transformed protein intensities first with AD status (Fig EV3C) and second with the $\log_{10}$-transformed ELISA-measured CSF t-tau concentration (Appendix Fig S4B), adjusting for age, sex, and cohort (Sweden, Magdeburg/Kiel, or Berlin) in both models. To compare estimators of binary (AD status, sex) to those of continuous variables (age, t-tau concentration [$\log_{10}$]), the estimators for continuous variables (i.e. per 1 year [age] and per 1 unit in $\log_{10}$ space of t-tau concentration/[ng/l]) were multiplied with the interquartile range (IQR) of the variable for plotting. IQRs for age were eleven years for the complete dataset (Fig EV3C) and 9 years for the reduced dataset excluding the Kiel samples due to missing t-tau concentration values (Appendix Fig S4B). The t-tau concentration 75% and 25% quantiles were 802 ng/l and 275 ng/l, respectively, corresponding to an interquartile range of 527 and 0.4648 in linear and $\log_{10}$ space, respectively (Appendix Fig S4B). Regression coefficients for age and sex were displayed in the heat map if the P-value for these estimators was below 0.05. All proteins were associated with AD status or t-tau concentration at a significance below of 0.05 in each plot.

Coefficients of variation (CVs) were calculated in RStudio for all inter-plate and intra-plate combinations of three samples, the median thereof was reported as overall coefficient of variation. Combinations with only one observation in three samples of a given protein were excluded. The protein CVs of the main study were calculated likewise within cohorts individually. The median CVs were calculated within the three cohorts, and the median thereof reported as final CV.

Machine learning for participant classification

All data processing was done in Python (3.7.3). Protein intensity data were Z-scored within cohorts, saved in Excel, and imported via the pandas package (0.25.3). Except for the XGBoost classifier, missing intensities were replaced with 0. Machine learning classifiers were employed using the scikit-learn package (0.21.3) and the XGBoost package package (0.90) (Fabian *et al*, 2011). Results were plotted via matplotlib (3.1.2). Visualization of the decision tree was performed with the dtreeviz package (https://github.com/parrt/dtreeviz).

In order to estimate features important for AD prediction, we employed a decision tree (Freund & Schapire, 1997). The minimum depth of the tree was increased until a training accuracy of 1.0 was achieved. At a tree depth of 2, using the protein intensities of MAPT, PKM (protein group P14618-2), and MIF, the training accuracy had reached 0.86, highlighting the importance of these proteins for the classifier. For a tree depth of six, intensities of a total of 14 proteins were used by the algorithm.

For estimating how well our tree-based approach would generalize to new data, we tested several ensemble methods (AdaBoost, Bagging, ExtraTrees, GradientBoosting, RandomForests, XGBoost). The subset of 14 protein intensities selected by the decision tree above were randomly shuffled and split using a k-Folds cross-validator ($k = 6$). Each model was used with its default parameters. XGBoost had the best performance and was selected for further analysis. To determine the optimum set of features, we added proteins to the model iteratively according to their feature importance within the tree (Fig 5A) and compared the AUC as a measure of model performance. To control for overfitting, we employed early stopping with 10 rounds and logloss as evaluation metric for best generalizability. No further tuning of hyperparameters was performed at this stage.

To assess the sensitivity and specificity of the final method, we combined each train and test set of the cross-validation and calculated the confusion matrix. Here, a test accuracy of 0.85 was achieved (training accuracy 0.94; sensitivity 82%, specificity 87% on our AD data).

Data availability

The datasets produced in this study are available in the following databases:

- Proteomic datasets: PRIDE archive PXD016278 (Perez-Riverol *et al*, 2019; https://www.ebi.ac.uk/pride/archive/)

Expanded View for this article is available online.

Acknowledgements

We thank all members of the Proteomics and Signal Transduction Group at the Max Planck Institute of Biochemistry and the Clinical Proteomics Group at the NNF Center for Protein Research for help and discussions and in particular Igor Paron, Christian Deiml, and Alexander Strasser for technical assistance. We further thank Martin Steger for his contribution in establishing data-independent mass spectrometry for CSF and Sebastian Virreira Winter, Özge Karayel, Niels H. Skotte, Felix Meissner, and Daniel Hornburg for discussions and supplying samples. The work carried out in this project was partially supported by the Max Planck Society for the Advancement of Science, the European Union's Horizon 2020 research and innovation program with the Microb-Predict project (no. 825694), by grants from the Novo Nordisk Foundation (NNF15CC0001; NNF15OC0016692), the DFG project "Chemical proteomics inside us" (no. 412136960), the European Research Council Synergy Grant under FP7 GA number ERC-2012-SyG_318987-Toxic Protein Aggregation in Neurodegeneration (ToPAG), and funding from the Harvard T.H. Chan School of Public Health Dean's Challenge Program sponsored by the McLennan Family Fund. Open access funding enabled and organized by Projekt DEAL.

Author contributions

JMB, PEG, and JBM designed the experiments, performed, analyzed, and interpreted the MS-based proteomic data. JMB optimized the MS data processing, performed bioinformatics analysis, and generated text and figures for the manuscript, and PEG contributed to these aspects. MTS performed the machine learning analysis. MK, FL, PK, DB, CGS, EII, OP, ND, MS, MKJ, and HZ designed and established the study cohorts and contributed to writing the paper. MM supervised and guided the project, designed the experiments and interpreted MS-based proteomics data and wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abbatiello SE, Mani DR, Schilling B, Maclean B, Zimmerman LJ, Feng X, Cusack MP, Sedransk N, Hall SC, Addona T *et al* (2013) Design, implementation and multisite evaluation of a system suitability protocol for the quantitative assessment of instrument performance in liquid chromatography-multiple reaction monitoring-MS (LC-MRM-MS). *Mol Cell Proteomics* 12: 2623 2639

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

Arevalo-Rodriguez I, Smailagic N, Roque I Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, Pedraza OL, Bonfill Cosp X, Cullum S (2015) Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev* CD010783

Baldacci F, Lista S, Cavedo E, Bonuccelli U, Hampel H (2017) Diagnostic function of the neuroinflammatory biomarker YKL-40 in Alzheimer's disease and other neurodegenerative diseases. *Expert Rev Proteomics* 14: 285 299

Barthélemy NR, Fenaille F, Hirtz C, Sergeant N, Schraen-Maschke S, Vialaret J, Buée L, Gabelle A, Junot C, Lehmann S *et al* (2016) Tau protein quantification in human cerebrospinal fluid by targeted mass spectrometry at high sequence coverage provides insights into its primary structure heterogeneity. *J Proteome Res* 15: 667 676

Bélanger M, Allaman I, Magistretti PJ (2011) Brain energy metabolism: focus on Astrocyte-neuron metabolic cooperation. *Cell Metab* 14: 724 738

Bloudek LM, Spackman DE, Blankenburg M, Sullivan SD (2011) Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. *J Alzheimers Dis* 26: 627 645

Bonifati V, Rizzu P, Van Baren MJ, Schaap O, Breedveld GJ, Krieger E, Dekker MCJ, Squitieri F, Ibanez P, Joosse M *et al* (2003) Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science* 299: 256 259

Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, Gomez-Varela D, Reiter L (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol Cell Proteomics* 16: 2296 2309

Callister SJ, Barry RC, Adkins JN, Johnson ET, Qian W-J, Webb-Robertson B-JM, Smith RD, Lipton MS (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5: 277 286

Carlyle BC, Kitchen RR, Kanyo JE, Voss EZ, Pletikos M, Sousa AMM, Lam TKT, Gerstein MB, Sestan N, Nairn AC (2017) A multiregional proteomic survey of the postnatal human brain. *Nat Neurosci* 20: 1787 1795

Chapman KR, Bing-Canar H, Alosco ML, Steinberg EG, Martin B, Chaisson C, Kowall N, Tripodis Y, Stern RA (2016) Mini mental state examination and logical memory scores for entry into Alzheimer's disease trials. *Alzheimers Res Ther* 8: 9

Comi C, Carecchio M, Chiocchetti A, Nicola S, Galimberti D, Fenoglio C, Cappellano G, Monaco F, Scarpini E, Dianzani U (2010) Osteopontin is increased in the cerebrospinal fluid of patients with Alzheimer's disease and its levels correlate with cognitive decline. *J Alzheimer's Dis* 19: 1143 1148

Cox J, Mann M (2012) 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* 13(Suppl. 1): 1 11

Dayon L, Núñez Galindo A, Wojcik J, Cominetti O, Corthésy J, Oikonomidi A, Henry H, Kussmann M, Migliavacca E, Severin I *et al* (2018) Alzheimer disease pathology and the cerebrospinal fluid proteome. *Alzheimer's Res Ther* 10: 1 12

Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R *et al* (2014) Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol* 13: 614 629

Fabian P, Gaël V, Alexandre G, Vincent M, Bertrand T, Olivier G, Mathieu B, Peter P, Ron W, Vincent D *et al* (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12: 2825 2830

Ferreira D, Perestelo-Pérez L, Westman E, Wahlund LO, Sarrisa A, Serrano-Aguilar P (2014) Meta-review of CSF core biomarkers in Alzheimer's disease: the state-of-the-art after the new revised diagnostic criteria. *Front Aging Neurosci* 6: 1 24

Foote M, Zhou Y (2012) 14-3-3 proteins in neurological disorders. *Int J Biochem Mol Biol* 3: 152 164

Fountoulakis M, Cairns N, Lubec G (1999) Increased levels of 14-3-3 gamma and epsilon proteins in brain of patients with Alzheimer's disease and Down syndrome. *J Neural Transm Suppl* 57: 323 335

Francelle L, Galvan L, Gaillard MC, Guillermier M, Houitte D, Bonvento G, Petit F, Jan C, Dufour N, Hantraye P *et al* (2015) Loss of the thyroid hormone-binding protein Crym renders striatal neurons more vulnerable to mutant huntingtin in Huntington's disease. *Hum Mol Genet* 24: 1563 1573

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Lect Notes Comput Sci* 904: 23 37

Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* 6: 67 77

GBD 2016 Neurology Collaborators (2019) Global, regional, and national burden of neurological disorders, 1990 2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* 18: 459 480

Geyer PE, Kulak NA, Pichler G, Holdt LM, Teupser D, Mann M (2016) Plasma proteome profiling to assess human health and disease. *Cell Syst* 2: 185 195

Geyer PE, Holdt LM, Teupser D, Mann M (2017) Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol* 13: 942

Geyer PE, Voytik E, Treit PV, Doll S, Kleinhempel A, Niu L, Muller JB, Buchholtz M-L, Bader JM, Teupser D *et al* (2019) Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies. *EMBO Mol Med* 11: e10427

Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, 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

Guldbrandsen A, Farag Y, Kroksveen AC, Oveland E, Lereim RR, Opsahl JA, Myhr K-M, Berven FS, Barsnes H (2017) CSF-PR 2.0: an interactive literature guide to quantitative cerebrospinal fluid mass spectrometry data from neurodegenerative disorders. *Mol Cell Proteomics* 16: 300 309

Hansson O, Zetterberg H, Buchhave P, Londos E, Blennow K, Minthon L (2006) Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol* 5: 228 234

Higginbotham L, Ping L, Dammer EB, Duong DM, Zhou M, Gearing M, Johnson ECB, Hajjar I, Lah JJ, Levey AI *et al* (2019) Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease. *bioRxiv* https://doi.org/10.1101/806752 [PREPRINT]

Hosp F, Gutierrez-Angel S, Schaefer MH, Cox J, Meissner F, Hipp MS, Hartl F-U, Klein R, Dudanova I, Mann M (2017) Spatiotemporal proteomic profiling of Huntington's disease inclusions reveals widespread loss of protein function. *Cell Rep* 21: 2291 2303

Hosp F, Mann M (2017) A primer on concepts and applications of proteomics in neuroscience. *Neuron* 96: 558 571

Hulstaert F, Blennow K, Ivanoiu A, Schoonderwaldt HC, Riemenschneider M, De Deyn PP, Bancher C, Cras P, Wiltfang J, Mehta PD *et al* (1999) Improved discrimination of AD patients using beta-amyloid(1-42) and tau levels in CSF. *Neurology* 52: 1555 1562

Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ (2010) Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol* 9: 119 128

Jack CRJ, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J *et al* (2018) NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement* 14: 535 562

Johnson ECB, Dammer EB, Duong DM, Ping L, Zhou M, Yin L, Higginbotham LA, Guajardo A, White B, Troncoso JC *et al* (2020) Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat Med* 26: 769 780

Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen JV (2018) Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J Proteome Res* 17: 727 738

Keshishian H, Burgess MW, Gillette MA, Mertins P, Clauser KR, Mani DR, Kuhn EW, Farrell LA, Gerszten RE, Carr SA (2015) Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. *Mol Cell Proteomics* 14: 2375 2393

Koch M, Furtado JD, Falk K, Leypoldt F, Mukamal KJ, Jensen MK (2017) Apolipoproteins and their subspecies in human cerebrospinal fluid and plasma. *Alzheimer's Dement* 6: 182 187

Kulak NA, Geyer PE, Mann M (2017) Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol Cell Proteomics* 16: 694 705

Laub R, Baurin S, Timmerman D, Branckaert T, Strengers P (2010) Specific protein content of pools of plasma for fractionation from different sources: impact of frequency of donations. *Vox Sang* 99: 220 231

Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol* 14: e8126

Mateo I, Llorca J, Infante J, Rodríguez-Rodríguez E, Berciano J, Combarros O (2008) Gene-gene interaction between 14-3-3 zeta and butyrylcholinesterase modulates Alzheimer's disease risk. *Eur J Neurol* 15: 219 222

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R *et al* (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement* 7: 263 269

Mitchell AJ (2009) A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J Psychiatr Res* 43: 411 431

Molinuevo JL, Gispert JD, Dubois B, Heneka MT, Lleo A, Engelborghs S, Pujol J, De Souza LC, Alcolea D, Jessen F *et al* (2013) The AD-CSF-index discriminates alzheimer's disease patients from healthy controls: a validation study. *J Alzheimer's Dis* 36: 67 77

Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C (1989) The consortium to establish a registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 39: 1159 1165

Niemantsverdriet E, Ottoy J, Somers C, De Roeck E, Struyfs H, Soetewey F, Verhaeghe J, Van den Bossche T, Van Mossevelde S, Goeman J *et al* (2017) The cerebrospinal fluid Abeta1-42/Abeta1-40 ratio improves concordance with amyloid-PET for diagnosing Alzheimer's disease in a clinical setting. *J Alzheimers Dis* 60: 561 576

Niu L, Geyer PE, Wewer Albrechtsen NJ, Gluud LL, Santos A, Doll S, Treit PV, Holst JJ, Knop FK, Vilsboll T *et al* (2019) Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Mol Syst Biol* 15: e8793

Olsson B, Lautner R, Andreasson U, Ohrfelt A, Portelius E, Bjerke M, Holtta M, Rosen C, Olsson C, Strobel G *et al* (2016) CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol* 15: 673 684

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M *et al* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47: D442 D450

Perneczky R, Wagenpfeil S, Komossa K, Grimmer T, Diehl J, Kurz A (2006) Mapping scores onto stages: mini-mental state examination and clinical dementia rating. *Am J Geriatr Psychiatry* 14: 139 144

Querfurth HW, LaFerla FM (2010) Alzheimer's disease. *N Engl J Med* 362: 329 344

Renton AE, Chiò A, Traynor BJ (2014) State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* 17: 17 23

Rice L, Bisdas S (2017) The diagnostic value of FDG and amyloid PET in Alzheimer's disease A systematic review. *Eur J Radiol* 94: 16 24

Ritchie C, Smailagic N, Noel-Storr AH, Ukoumunne O, Ladds EC, Martin S (2017) CSF tau and the CSF tau/ABeta ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev* 3: CD010803

Sathe G, Na CH, Renuse S, Madugundu AK, Albert M, Moghekar A, Pandey A (2019) Quantitative proteomic profiling of cerebrospinal fluid to identify candidate biomarkers for Alzheimer's disease. *Proteomics Clin Appl* 13: e1800105

Scheltens P, Blennow K, Breteler MMB, de Strooper B, Frisoni GB, Salloway S, Van der Flier WM (2016) Alzheimer's disease. *Lancet* 388: 505 517

Schipke CG, Prokop S, Heppner FL, Heuser I, Peters O (2011) Comparison of immunosorbent assays for the quantification of biomarkers for Alzheimer's disease in human cerebrospinal fluid. *Dement Geriatr Cogn Disord* 31: 139 145

Sepe FN, Chiasserini D, Parnetti L (2018) Role of FABP3 as biomarker in Alzheimer's disease and synucleinopathies. *Future Neurol* 13: 199 207

Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT (2011) Neuropathological alterations in Alzheimer disease. *Cold Spring Harb Perspect Med* 1: a006189

Seyfert S, Kunzmann V, Schwertfeger N, Koch HC, Faulstich A (2002) Determinants of lumbar CSF protein concentration. *J Neurol* 249: 1021 1026

Sjogren M, Vanderstichele H, Agren H, Zachrisson O, Edsbagge M, Wikkelso C, Skoog I, Wallin A, Wahlund LO, Marcusson J *et al* (2001) Tau and Abeta42 in cerebrospinal fluid from healthy adults 21-93 years of age: establishment of reference values. *Clin Chem* 47: 1776 1781

Spies PE, Slats D, Sjogren JMC, Kremer BPH, Verhey FRJ, Rikkert MGMO, Verbeek MM (2010) The cerebrospinal fluid amyloid beta42/40 ratio in the differentiation of Alzheimer's disease from non-Alzheimer's dementia. *Curr Alzheimer Res* 7: 470 476

Tusher VG, Tibshirani R, Chu G (2002) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98: 5116 5121

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13: 731 740

Vos SJB, van Rossum IA, Verhey F, Knol DL, Soininen H, Wahlund L-O, Hampel H, Tsolaki M, Minthon L, Frisoni GB *et al* (2013) Prediction of Alzheimer disease in subjects with amnestic and nonamnestic MCI. *Neurology* 80: 1124 1132

Wewer Albrechtsen NJ, Geyer PE, Doll S, Treit PV, Bojsen-Moller KN, Martinussen C, Jorgensen NB, Torekov SS, Meier F, Niu L *et al* (2018) Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-En-Y gastric bypass surgery. *Cell Syst* 7: 601 612.e3

Wichmann C, Meier F, Virreira Winter S, Brunner A-D, Cox J, Mann M (2019) MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Mol Cell Proteomics* 18: 982 994

Winblad B, Amouyel P, Andrieu S, Ballard C, Brayne C, Brodaty H, Cedazo-Minguez A, Dubois B, Edvardsson D, Feldman H *et al* (2016) Defeating Alzheimer's disease and other dementias: a priority for European science and society. *Lancet Neurol* 15: 455 532

3.6. Article 6: Molecular origin of blood-based infrared fingerprints

Authors:

Liudmila Voronina[1,2,*], Cristina Leonardo[1,2], Johannes B. Mueller-Reif[3,†], Philipp E. Geyer[3,4,†], Marinus Huber[1,2], Michael Trubetskov[2], Kosmas V. Kepesidis[1], Jürgen Behr[5], Matthias Mann[3,4], Ferenc Krausz[1,2,6], Mihaela Žigman[1,2,6,*].

1. Department of Physics, Ludwig Maximilian University of Munich, Garching, 85748 Germany;

2. Max Planck Institute of Quantum Optics, Garching, 85748 Germany;

3. Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, 82152 Germany;

4. Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, 2200 Denmark;

5. Comprehensive Pneumology Center, Department of Internal Medicine V, Clinic of the Ludwig Maximilians University Munich (LMU), Member of the German Center for Lung Research;

6. Center for Molecular Fingerprinting, Budapest, 1093 Hungary.

† New Adress: OmicEra Diagnostics GmbH, Planegg, 82152 Germany

* Corresponding authors. Address: Am Coulombwall 1, 85748 Germany. Phone: +4917629522131

Emails: liudmila.voronina@mpq.mpg.de, mihaela.zigman@mpq.mpg.deJakob

Body fluids as specimens for clinical test development are of interest for molecular identifying methods like proteomics or metabolomics, but also for more complex approaches. In our manuscript ‘Molecular origin of blood-based infrared fingerprints’, we teamed up with the Department of Physics from the LMU and the MPI of quantum optics to use proteomics for the interpretation of infrared molecular fingerprints (IMF) of human serum samples. A cohort of 148 lung cancer, benign condition and non-symptomatic patients was probed and investigated with MS-based proteomics and Fourier transform infrared (FTIR) spectroscopy. It has been shown previously that FTIR spectroscopy can distinguish lung cancer patients form controls, but the understanding of the underlying molecular changes has not been addressed so far. With the information of the proteomics measurements, the molecular fingerprints were reconstructed from the high abundant proteins of serum and the underlying changes could be traced to concentration changes of the high abundant proteins and adding the metabolite fraction of serum yielded in an accurate model of the original serum spectra. Proteomics and IMF both allowed binary classification of lung cancer versus reference individuals at > 80%.

This study is a perfect showcase how proteomics can contribute not only to biomarker identification in unbiased study designs but also to establish novel screening techniques for medical conditions in liquid biopsies. The FTIR spectroscopy method to investigate

serum samples is quick, easy applicable and affordable, compared to more specific, comprehensive but also time-consuming techniques like proteomic profiling. It is reasonable to establish such methods for preselection of patients for deeper screening as a first stage of diagnosis. MS-based proteomics helps to investigate the molecular origins of changes in such fingerprint methods and ensures that findings are not only on the result of quality bias from sample taking and processing. This described in the publication ‘Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies’, which is now accepted for publication in Angewandte Chemie,

Molecular Origin of Blood-based Infrared Fingerprints

Liudmila Voronina[1,2,*], Cristina Leonardo[1,2], Johannes B. Mueller-Reif[3,†], Philipp E. Geyer[3,4,†], Marinus Huber[1,2], Michael Trubetskov[2], Kosmas V. Kepesidis[1], Jürgen Behr[5], Matthias Mann[3,4], Ferenc Krausz[1,2], Mihaela Žigman[1,2,*].

1. Department of Physics, Ludwig Maximilian University of Munich, Garching, 85748 Germany;
2. Max Planck Institute of Quantum Optics, Garching, 85748 Germany;
3. Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, 82152 Germany;
4. Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, 2200 Denmark;
5. Comprehensive Pneumology Center, Department of Internal Medicine V, Clinic of the Ludwig Maximilians University Munich (LMU), Member of the German Center for Lung Research;

† New Address: OmicEra Diagnostics GmbH, Planegg, 82152 Germany

* Corresponding authors. Address: Am Coulombwall 1, 85748 Germany. Phone: +4917629522131

Emails: liudmila.voronina@mpq.mpg.de, mihaela.zigman@mpq.mpg.de

Author contributions: L.V. and M. Ž. designed the research plan; M. Ž., F.K. and M.M. initiated and led the study plan; J.B. led the clinical study; L.V., C.L., J.M., P.G. and M.H. performed the measurements; L.V., C.L., J.M., M.T., K.K. analyzed the data; L.V., M. Ž., J.M.-R. wrote original draft; all authors contributed to editing and revising the manuscript.

Competing Interests: The authors declare no competing interests.

Pre-print information: http://arxiv.org/abs/2102.00765.

Keywords: IR spectroscopy • cancer • infrared molecular fingerprinting • proteomics • liquid biopsy

Abstract

Infrared spectroscopy of liquid biopsies is a time- and cost-effective approach that may advance biomedical diagnostics. However, molecular nature of disease-related changes of infrared molecular fingerprints (IMFs) remains poorly understood, impeding the method's applicability. Here we probe 148 human blood sera and reveal the origin of the variations in their IMFs. To that end, we supplemented infrared spectroscopy with biochemical fractionation and proteomic profiling, providing molecular information about serum composition. Using lung cancer as an example for a medical condition, we demonstrate that the disease-related differences in IMFs are dominated by contributions from twelve highly abundant proteins - that, if used as a pattern, may be instrumental for detecting malignancy. Tying proteomic to spectral information and machine learning advances our understanding of infrared spectra of liquid biopsies, a framework that could be applied to probing of any disease.

Introduction

Infrared spectroscopy is a well-established method of studying chemical substances *via* analyzing the vibrational transitions that are characteristic of their molecular structure.[1] In particular, infrared molecular fingerprinting of human biofluids has the potential to provide information about the health state of individuals when combined with appropriate machine learning algorithms.[2–14] The idea behind is to record an infrared absorption spectrum of the whole molecular ensemble composing blood serum using Fourier-transform infrared (FTIR) spectroscopy and pinpoint the deviations, associated with a given pathophysiological condition. However, the molecular origin of such changes in infrared molecular fingerprints (IMFs) is poorly understood.[15,16] The interpretation of the infrared absorption spectra is currently largely restricted to the characteristic spectral signatures of various functional groups.[17–19] However, these are contained in many different types of biomolecules, their spectral features in aqueous environment are broad and strongly overlapping, and the molecular complexity of biofluids is extremely high. Therefore, the understanding of the underlying molecular changes of the IMFs has so far been limited.[20,21]

Thorough exploration of the molecular origin of IMFs would be instrumental for successful application and verification of molecular fingerprinting in clinical settings.[3] It would allow for improved sample preparation, ensure that the spectral features used for building the computational models are indeed caused by a medical condition and not by confounding factors and help define the possible limitations of blood-based IMFs' applicability.[22] To that end, several studies measured the concentrations of a range of analytes in human blood serum using conventional biochemical methods and demonstrated that IMFs can be used to retrieve these concentrations using multivariate regression or consecutive spectral subtraction approaches.[14,23–28] However, they come up short in determining how exhaustive the list of molecular constituents is and connecting disease-related changes in the molecular composition of biofluids to the changes in the corresponding IMFs.[26]

It had been suggested that large variations in blood-based IR spectra may be caused by a varying albumin-to-globulin ratio.[29] Indeed, the spectroscopic signature of human blood serum is vastly dominated by a few highly abundant molecular components, such

as human serum albumin (HSA) and immunoglobulins.[30] To overcome the challenge of strong molecular signals that overshadow the signals from less abundant molecules, splitting complex biological samples into several fractions of different chemical nature is beneficial.[28,31,32] Previously, ultrafiltration has been used to fractionate human blood serum based on molecular weight of the components.[15,24,28,33,34] However, these methods introduce unwanted chemicals in non-reproducible fashion.[35] In this study, we chose to adapt a combination of solvent-extraction sample preparation protocols, which are typically used in metabolomics[36] and proteomics,[37] because of their robustness and speed.[38]

In order to explore the dependence of the IMFs of human blood serum on its molecular composition, spectroscopic molecular fingerprinting should be ultimately combined with a technique that is able to provide molecular-specific information over a high dynamic range.[39] Recently, a high-throughput mass spectrometry (MS)-based proteomic workflow has been established for the analysis of human blood plasma profiles.[40] We adapted this technology for human blood serum and applied it to our sample set in order to model the IMFs of hydrated biofluids as a linear combination of molecular components. Such a parallelized FTIR-MS approach for molecular annotation of disease-relevant vibrational fingerprints of human blood derivatives has been lacking this far.

With the gained understanding of the molecular composition underlying the IMFs of human blood serum, we compare the samples of lung cancer patients (TNM clinical stages II and III) with reference individuals matched in age, gender and smoking status. We focused on lung cancer as a prototypical disease for which non-invasive early detection from blood profiling would be highly beneficial.[41,42] The ability of FTIR spectroscopy of blood serum to discriminate lung cancer cases from controls has been previously shown in several studies.[43,44] Pattern recognition algorithms were used to identify non-small cell lung carcinoma and subtype the disease conditions.[43] Independently, the ratio between intensities at 1080 and 1170 cm^{-1} was put forward as the most informative for disease detection, and it was suggested that changes in the protein secondary structure might be correlated with lung cancer.[44] Other types of cancer have also been detected with various efficiencies using blood-based IMFs, with little insight into molecular changes for the reasons stated above.[10,45–49]

In this study, we obtain reproducible, cost- and time-efficient IMFs of human sera and use proteomic measurements to facilitate their understanding at a molecular-level. In particular, we reveal a pattern of changes of human blood serum composition, which correlates with the presence of lung cancer and results in an observable difference between IMFs of blood sera of lung cancer patients compared to the reference group. Both spectral and molecular information was used to build explainable classification models for lung cancer detection.[50] This paradigm can be applied to possibly any other health phenotypes in order to develop efficient and explainable diagnostic tools.

Results

1. Decomposing complexity of human blood sera using biochemical fractionation

We recorded infrared absorption spectra of liquid human blood sera in the range from 1000 to 3000 cm^{-1}. The spectra are dominated by amide bands that are attributed to the vibrations of protein backbone.[51] In particular, the most prominent feature between 1600 and 1700 cm^{-1} (Amide I band) is characteristic of the secondary structure of the proteins.[51] The region on the red side of the spectrum (1000-1200 cm^{-1}) is often referred to as “carbohydrate region”, because of the typical absorption patterns that glycans exhibit here.[18] Finally, lipids produce several absorption bands around 1735 cm^{-1}, 2852 cm^{-1} and 2926 cm^{-1}.[52]

Attributing the distinct features of the mid-infrared absorption spectrum of human blood serum to a specific molecular class is somewhat oversimplified, since absorption spectra of various biological molecules often overlap. In order to gain deeper insight into the origins of different spectral features, we built a comprehensive model of the human blood serum absorption. To this end, we used a set of 148 prospectively collected blood serum samples (Figure 1*A*).

As a first step, we recorded the IMFs of each full intact, fluid, serum sample using high-throughput automated FTIR spectrometer in transmission mode (black line in Figure 1*B*).[11] Next, we biochemically fractionated each sample into three fractions and recorded their IMFs (colored lines in Figure 1*B*) in order to assess the relative contributions of roughly defined molecular classes, i.e. proteins and metabolites. In parallel, we used

proteomic analysis of the crude sera and human serum albumin (HSA)-depleted fractions to characterize the efficiency of HSA depletion and the molecular composition of each protein fraction.

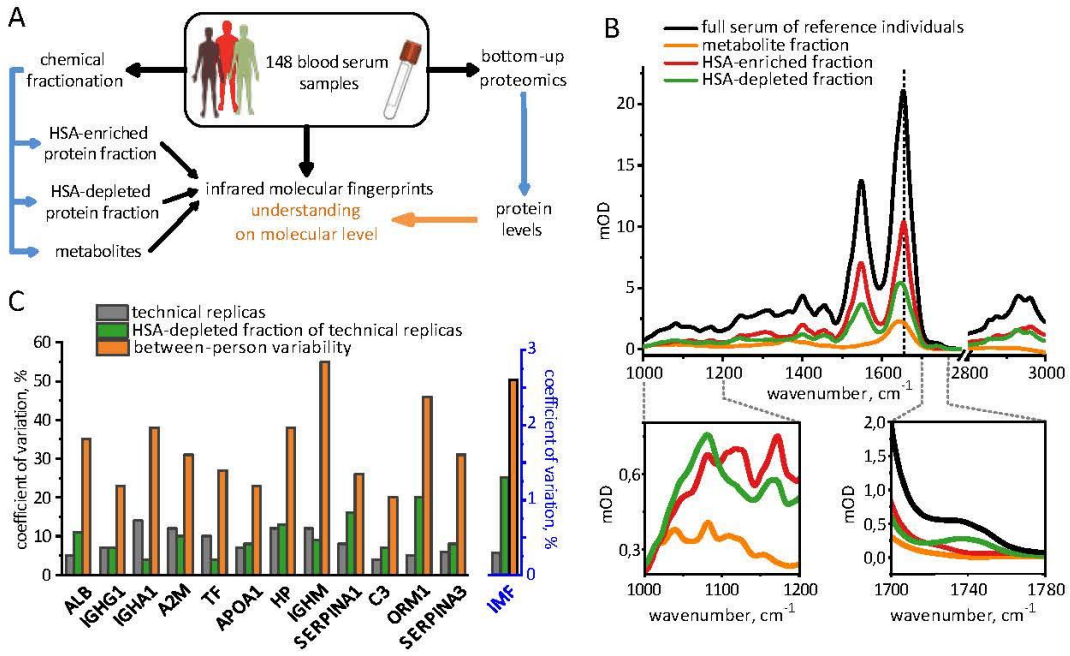


Figure 1. Decomposing complexity of human blood sera using chemical fractionation. (A) Overview of the workflow of the study. (B) Average infrared molecular fingerprint (IMF) of human blood serum of 93 reference individuals and the corresponding IMFs of 3 fractions. The dashed vertical line shows the position of the Amide I band in the HSA-enriched fraction. The two lower inserts highlight the regions with the largest relative differences between the fractions. (C) Reproducibility of the fractionation protocol assessed with proteomic and FTIR measurements. Left axis: coefficients of variation for the levels of 12 proteins considered in this study for the same 8 serum samples with and without fractionation as well as their between-person variability in 93 control individuals. Right axis: the corresponding variations in the IMFs, averaged across wavenumbers.

Human serum albumin is the most abundant serum protein and constitutes about a half of total protein mass.[30] It is helpful to separate HSA away from other proteins, because its intense absorption potentially obscures the signals from other molecules.[31] For this purpose, we first precipitated most of the proteins using cold ethanol.[37] The supernatant was enriched in HSA, which we precipitated in the next step to separate it

from metabolites.[53] All three fractions (HSA-depleted proteins, HSA-enriched proteins and metabolites) were dried in vacuum and re-dissolved in water prior to the spectroscopic measurements.

We assessed the reproducibility of our fractionation protocol both with FTIR spectroscopy and proteomic analyses (Figure 1*C*). First, we estimated the measurement uncertainty of the proteomic workflow as the coefficient of variation (CV) in repeated measurements of the same single human blood plasma sample. The average CV for the 12 proteins considered in this study (see below) in the crude plasma samples is 9 %, and it rises to 10% in the HSA-depleted fraction of the same sample, suggesting that the process of fractionation adds only minor error compared to the instrumental one. The CV measured for 93 reference individuals provides a rough estimate for the between-person variability, which is higher than the instrumental error for all considered proteins (33 % on average). The analysis based on IMFs leads to similar conclusions (Figure 1*C*, right axis).

We further compared the spectral intensities of each of the fractions (Figure 1*B*). This procedure facilitates several unexpected conclusions about the nature of the IMFs of crude blood sera: Firstly, the signals between 1000 and 1200 cm^{-1} are typically attributed to carbohydrates.[17] Indeed, we detected the metabolite fraction containing free carbohydrates, exhibiting characteristic pattern in this region of the spectra. However, the intensity of the signals from both two protein fractions combined is an order of magnitude higher than that of metabolite fraction in this spectral region. We attribute this effect to glycosylation of proteins and further demonstrate it below. Additionally, we show that around 10% of the intensity of the Amide I band (1654 cm^{-1} in crude serum), which is typically attributed to proteins,[17] is actually contributed by metabolites.

Altogether, our fractionation workflow enabled us to disentangle the quantitative contributions of metabolites and proteins to the IMF of crude blood sera. Since the absorption of proteins fractions is, as expected, significantly higher than that of metabolites, in the next step we focused on understanding and modeling the contribution of protein absorption to the overall fingerprints.

2. Towards molecular understanding of infrared fingerprints using proteomics

We demonstrated that the IR spectrum of blood serum mostly exhibits signals originating from the protein absorption. It is therefore important to understand how various proteins of blood sera contribute to the overall IR absorption spectra of this biofluid. To that end, we performed bottom-up proteomic analysis of the same samples. They were subjected to an established mass-spectrometry based proteomics pipeline.[40] In brief, proteins in the sample are denatured and disulfide bonds reduced and quenched. Proteins are then digested into tryptic peptides and desalted. The peptides are separated by reversed phase chromatography coupled online to the mass spectrometer to detect the mass to charge ratios of peptides and their fragments in a quantitative manner. This enables software-dependent peptide identification and subsequently quantitative protein assembly from detected peptides.[54,55]

The first ten proteins listed in Figure 1*C* are the ten most abundant proteins in human blood serum (Table S1). The quantitative values for each protein (so called 'label-free quantification' or LFQ values) provided by proteomic measurements are suited to characterize the differences between subjects in a study, but not directly proportional to the absolute concentrations of proteins,[56] as revealed by Table S1. To obtain the actual protein concentrations, we re-scaled the LFQ values using the average reference concentrations of these proteins in healthy subjects.

To be able to link the actual individual protein levels directly to the IMFs of blood sera, we measured IR absorption spectra of each of the 10 most abundant proteins separately, dissolved in phosphate-buffered saline (PBS). Figure 2*A* demonstrates the IR spectra of 5 highly abundant proteins (Figure S2 for all proteins). The position and shape of the Amide I band is characteristic for their secondary structure and qualitatively corresponds to the known β-sheet and α-helix content of proteins.[51] As expected, alpha-1-acid glycoprotein (ORM1 in Figure 2*A*) shows particularly high absorption in the region of 1000-1200 cm^{-1}, because about 45 % of its dry mass is comprised of carbohydrates.[57]

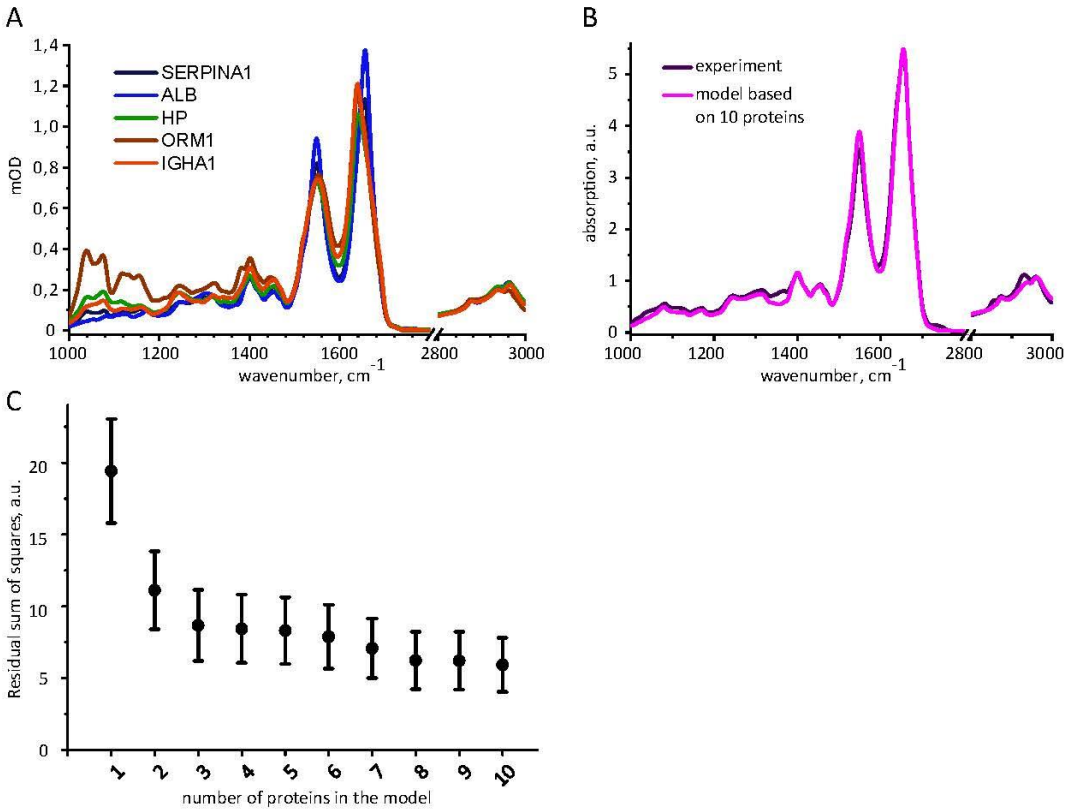


Figure 2. Molecular modeling of infrared fingerprints based on serum proteomic profiling. (A) Examples of infrared absorption spectra of human serum proteins at the same concentration, 5 mg/mL. (B) Average IMF of 148 human blood sera, each modelled as a sum of contributions of 10 proteins compared to the average experimentally measured IMF. (C) Average vector distance between the model and experimental spectra for all 148 samples depending on the number of proteins introduced into the model.

In order to estimate the contribution of each protein to the IMF of blood serum, we modeled the absorption spectra of every individual's serum as a sum of IR absorption spectra of proteins multiplied by their respective concentrations, measured by proteomics:

$$IMF(\tilde{\nu}) = \sum_i C_i * S_i(\tilde{\nu}) ,$$

where $\tilde{\nu}$ represents wavenumber, C_i – concentration of the protein *i* in mg/mL, $S_i(\tilde{\nu})$ – absorption spectrum of the protein *i* for 1 mg/mL.

We started by taking into account the spectral contribution of HSA only (*i*=1) and building complexity by adding proteins one by one, in the order as listed in Table S1. Figure 2*C* shows how the model becomes closer to the experimentally measured IMFs with every additional protein. Adding further lower abundant proteins to the model is

expected to yield only small improvements, since the total concentration of remaining proteins that are beyond the ten molecules considered here is about the same order of magnitude as the level of complement component C3.

In Figure 2*B* we compare the average modeled and experimental absorption spectra of human blood serum. Given the linear character of the model and the limited number of considered components, the matching is remarkably high. The only prominent peaks missing from the modeled spectra are the C=O (at 1735 cm^{-1}) and C-H stretches (at 2852 cm^{-1} and 2926 cm^{-1}) known to be unique for lipids.[52] Indeed, the average concentration of cholesterol in human blood serum is of the same order of magnitude as the last proteins we considered.[58] The model can, therefore, be further refined by including cholesterol and other metabolites, such as ATP, melanin, glucose and urea. In fact, adding the entire metabolite fraction to the model further reduces the RSS between the model and the experiment by 50 % (Figure S3).

3. Combining MS-based proteomics and IR fingerprinting reveals lung cancer-related molecular changes in blood serum

Having obtained a simple model of the IR absorption of human blood serum, we can address the question how this absorption changes as a consequence of a disease. In this study we focused on lung cancer, as the most common cause of cancer-related deaths worldwide.[41] We compare the IMFs of prospectively collected sera between two cohorts: 55 lung cancer patients (therapy naïve, prior to any cancer-related therapy, at TNM clinical stages 2 and 3) with 93 reference individuals. In the latter cohort we gathered non-symptomatic individuals ("healthy"), patients with chronic pulmonary obstructive disease (COPD) and individuals with lung hamartoma, to challenge our detection regime by non-cancerous lung diseases. Importantly, to avoid possible confounding bias the cohorts are gender, age and smoking-status matched (Table S2).

We find that infrared molecular fingerprints of lung cancer patients clearly differ from that of reference individuals. The black line in Figure 3*A* shows the difference between the average IMF of lung cancer patients and those of references as a function of wavenumber, which we specify as "differential fingerprint". The p-values of the most prominent spectral peaks are below 10^{-6} (Table S3), strongly suggesting that the

differences between the IMFs of two cohorts are statistically significant. To further quantify these differences, we applied support vector machine (SVM) algorithm to classify the samples into two classes – cancer cases and reference individuals. To that end, the data were split into train and test sets, employing 10-times repeated 10-fold cross-validation. The area under the curve (AUC) of the receiver operating characteristics (ROC) curve was used as a measure of classification efficiency. For the classification of lung cancer patients versus references, the model reveals an AUC of 0.85±0.1, implying that the SVM model can, in principle, be trained to distinguish between the two cohorts.

We find that the differential fingerprint of lung cancer has a specific shape, with prominent features around 1000-1200 cm^{-1}, as well as in the Amide I and Amide II regions. Such shape could result from alternations in the proteins secondary structure, as previously suggested[44] or, alternatively, from the changes in their concentration.[22] The distinction between the two possibilities can only be obtained by comparison of two sample sets with a technique that provides information about molecular concentrations.

The HSA-enriched and HSA-depleted fractions reflect the largest differences between lung cancer and reference samples with p-values below 10^{-6} (Table S3), while the metabolite fraction is not significantly different in the samples from reference individuals versus these of the lung cancer patients. This finding is confirmed by the AUC values: for the metabolite fraction the AUC is 0.62±0.2, while for the HSA-enriched fraction it is 0.82±0.1, and for the HSA-depleted fraction - 0.75±0.1. Thus, we turned to the proteomic measurements of the same sample set - aiming for the identification of individual proteins responsible for the observed changes in the IMFs.

In line with previous research,[42,59–65] we find a number of proteins that demonstrate p-values below 0.0005 (Table S4). However, the purpose of this study is not the search for specific biomarking candidates; instead, we wish to evaluate whether lung cancer results in a pattern of changes in protein concentrations responsible for its IR signature.

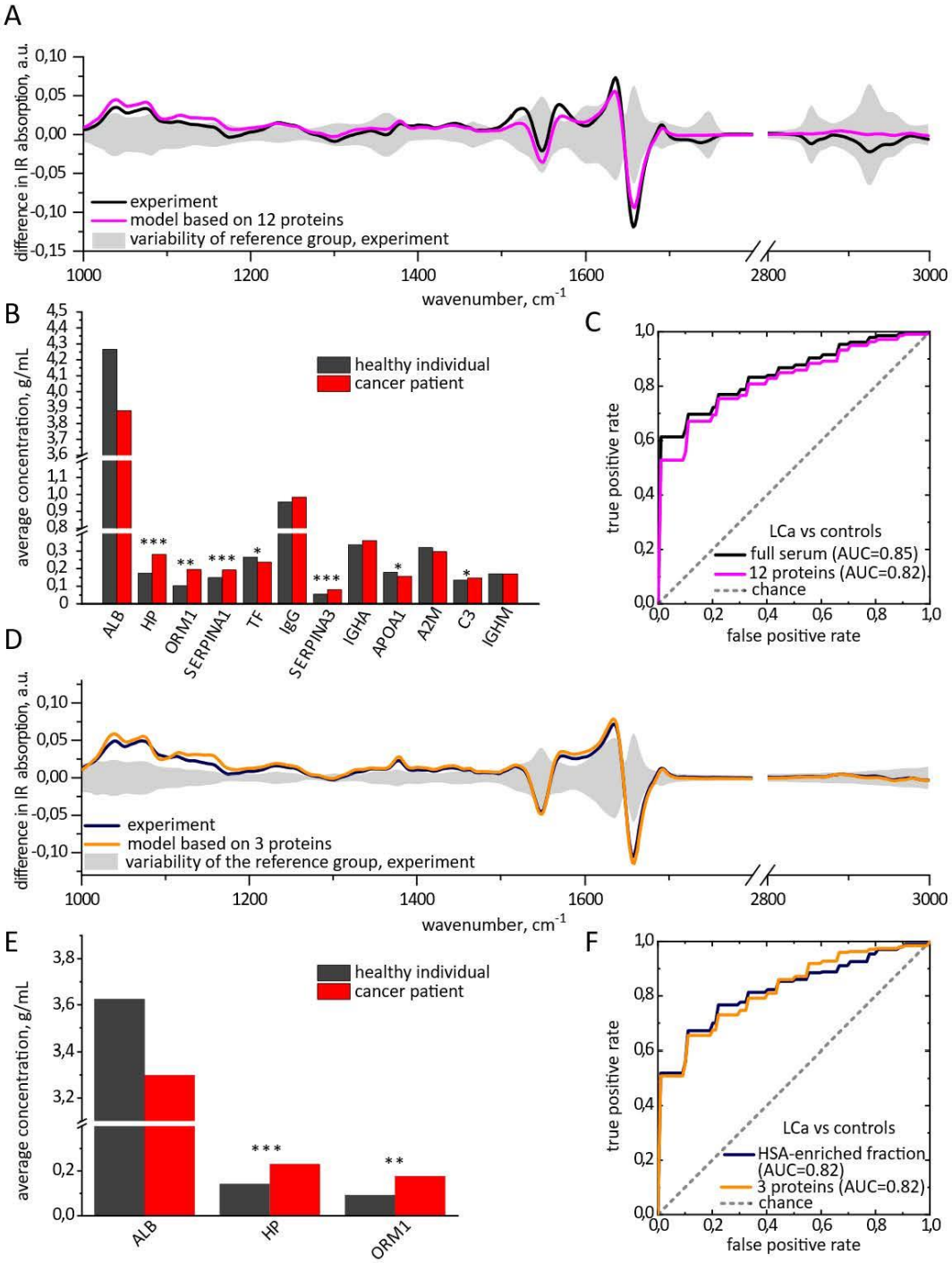

A
difference in IR absorption, a.u.
0,10
0,05
0,00
-0,05
-0,10
-0,15
experiment
model based on 12 proteins
variability of reference group, experiment
1000
1200
1400
1600
2800
3000
wavenumber, cm-1
B
average concentration, g/mL
healthy individual
cancer patient
ALB
HP
ORM1
SERPINA1
TF
IgG
SERPINA3
IGHA
APOA1
A2M
C3
IGHM
C
true positive rate
false positive rate
LCa vs controls
full serum (AUC=0.85)
12 proteins (AUC=0.82)
chance
D
difference in IR absorption, a.u.
experiment
model based on 3 proteins
variability of the reference group, experiment
wavenumber, cm-1
E
average concentration, g/mL
healthy individual
cancer patient
ALB
HP
ORM1
F
true positive rate
false positive rate
LCa vs controls
HSA-enriched fraction (AUC=0.82)
3 proteins (AUC=0.82)
chance

Figure 3. Lung cancer-related molecular changes in blood serum, based on comparison between 55 lung cancer patients and 93 reference individuals. (A) Differential fingerprints of lung cancer in full sera: experimentally measured and modeled based on the levels of 12 proteins. The shaded area shows the standard deviation of the IMFs of the reference group. (B) Change in the concentrations of proteins in blood serum caused by lung cancer, measured by proteomics. The proteins are ordered according the absolute difference in the concentrations in lung cancer and control individuals. *- p-value below 0.05, ** - p-value below 0.0005, ***- p-value below 10^{-6}, no star – p-value above 0.05. (C) ROC curves based on the experimental measurement of IMF of full serum and the set of 12 proteins measured by proteomics. The STDs are 0.1 for AUC in panels (C) and (F). (D) Differential fingerprints of lung cancer in HSA-enriched fraction: experimentally measured and modeled based on the levels of 3 proteins. (E) Change in the concentrations of proteins in HSA-enriched fraction caused by lung cancer, measured by proteomics. (F) Comparison between the ROC curves based on the experimental measurement of IMF of HSA-enriched fraction and the corresponding set of 3 proteins.

The first question we have addressed is: which proteins do we have to consider in order to model the differences in the IMFs between the lung cancer patients and reference individuals. The *differential* fingerprint is affected by the disease-related *absolute change* in the protein concentration due to the linear character of the absorption measurement. Therefore, we ranked all detected proteins according to the absolute difference in average concentration between lung cancer and reference samples, as measured by MS (Table S5).

Out of ten proteins that are most extensively changing, eight are also among the ten most abundant proteins in the blood sera. We further identify other proteins reflecting the differences between the two sample sets, such as alpha-1-acid glycoprotein-1 and alpha-1-antichymotrypsin: although their concentrations in non-symptomatic subjects are below the ten most abundant proteins, they are changing significantly in lung cancer patients and thus have to be taken into account to accurately model the disease differential fingerprint. In total, we considered twelve proteins for the model of lung cancer differential fingerprint, as shown in Figure 3*B*: ten most abundant ones and two additional ones that are changing most significantly.

After we have modelled the IMF of every individual as described above, the differential fingerprint of lung cancer was calculated as the difference between the average fingerprint of lung cancer patients and reference individuals. The resulting curve of this twelve-protein model very closely resembles the measured differential fingerprint, reflecting all the important features (pink line in Figure 3*A*). Moreover, the binary classification of lung cancer cases versus reference individuals based on the concentrations of the twelve identified proteins produces an AUC of 0.82±0.1, which is

close to the value for experimentally measured serum spectra (0.85±0.1). These findings suggest that most of the information in IMFs regarding lung cancer status stems from the molecular changes in these twelve proteins. Moreover, such kind of information can be measured in time- and cost-efficient manner by applying FTIR, without the need to measure the concentrations of each of the protein separately.

Interestingly, the three proteins that change the most between the lung cancer patients and the reference group (namely, HSA, haptoglobin and alpha-1-acid glycoprotein 1, Figure 3*B* and 3*E*) remain predominantly contained in the HSA-enriched fraction during the fractionation procedure. This explains the high AUC obtained for this protein fraction: 0.82±0.1, blue line in Figure 3*F*. It further suggests that most of the molecular information about the presence of lung cancer is encoded in the concentrations of the three proteins named above, out of all twelve proteins analyzed. Indeed, the SVM binary classification based on the concentrations of these three proteins reveales the AUC of 0.82±0.1, the same as based on all 12 proteins considered above.

We modeled the IMFs of the HSA-enriched fraction as detailed above, taking into account the proportion of each protein in HSA-enriched fraction compared to full serum (Table S1 and Figure S1). In line with only a minor contribution of low-abundant proteins and metabolites to the IR spectra of HSA-enriched fraction, we find that the model very well reproduces the experimental curve (Figure 3*D*).

In summary, we observe statistically significant differences between the IMFs of blood serum of lung cancer pateints when compared to the IMFs of reference individuals. Biochemical fractionation and proteomic profiling of the very same sample set facilitated identification of the compounds responsible for these differences and revealed previously unappreciated pattern of changes in the concentrations of well known proteins that we find to be characteristic of lung cancer.

Discussion

Although FTIR has been used over decades and blood-based studies suggested the applicability of this approach to disease diagnostics, the molecular nature of blood-based infrared molecular fingerprints (IMFs) and changes therein has not been well understood. Being cost- and time-efficient, suitable for high-throughput approaches, IMFs could

greatly contribute to clinical diagnostics if their robust correlation with any given condition is reproducibly demonstrated. Molecular understanding of the IMFs along with computational models may open up a path towards informed choice of biofluid (e.g. serum *vs* plasma), improved sample preparation and possibly even initial steps of the biomarker identification. Here we took advantage of a prospective clinical study and examined the samples with two independent techniques - IR spectroscopy and mass spectrometry (MS)-based proteomics - with the goal to elucidate the molecular entities dominating human blood-based IMFs.

As a first step to decompose chemical complexity of IMFs, we established a protocol for highly-reproducible fractionation of crude human blood sera into three fractions: human serum albumin (HSA)-enriched proteins, HSA-depleted proteins, and metabolites. The strongest IR absorption signal in human blood serum arises from proteins. We therefore measured their relative concentrations in the samples using MS-based proteomic profiling and used the concentrations of ten most abundant proteins to reconstruct individual spectra of the human blood serum. This concept is shown in the bottom part of Figure 4 for the general case of any omic technology. Indeed, the model built in this study can be further developed by adding highly abundant metabolites and additional proteins until the model reproduces measured IMFs within their noise limit. In particular, it has been shown previously that in addition to the proteins discussed here, FTIR spectra of blood plasma provide information about the levels of lactate, urea, apolipoproteins B and C, as well as immunoglobulin D.[26] However, the data presented here suggest that our 10-protein-based approach leaves little room for improvement in modelling IMFs measured by FTIR spectroscopy. The ultimate limitation of such modeling lies in the linearity of the model, disregarding any interaction between different blood components.

Infrared molecular fingerprints acquired by field-resolved spectroscopy[66] may drastically increase the precision of infrared molecular fingerprinting by reducing the noise limit. This will render smaller molecular contributions significant, uncovering thereby more molecular information just as the combination of further biochemical fractionation (e.g. by liquid chromatography) with field-resolved spectroscopy will do. Both may allow more

lower-abundance molecules to contribute to the identification of a pathophysiological condition.

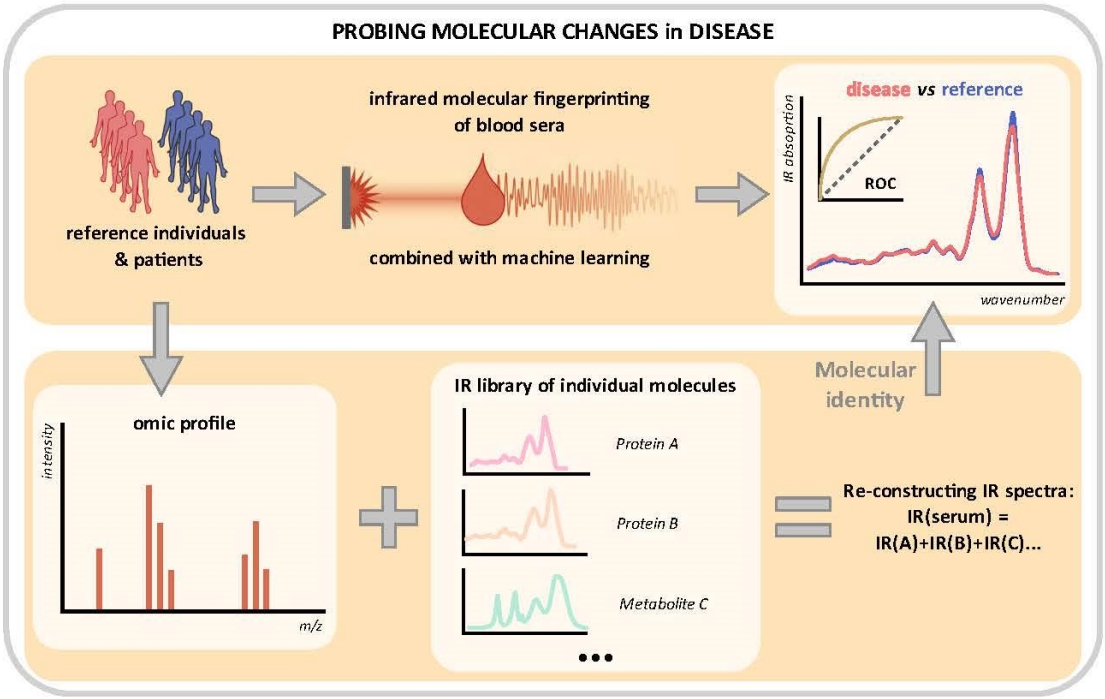


Figure 4. General workflow for probing molecular changes in disease. The infrared absorption spectra of blood sera are reconstructed as a linear combination of the spectra from individual molecular constituents, while the concentrations of the latter are measured using an omics technology. The resulting model is compared to the measured IMFs of blood sera and used to explain disease-related features therein. A similar workflow can potentially be applied to detection of any phenotype in human biofluids.

In this study we use lung cancer as a case scenario of a medical condition, the outcome of which could significantly benefit from early detection. We find that IMFs of sera samples of lung cancer patients differ significantly from that of reference individuals. Using MS-based proteomics, we identify a pattern of known highly-abundant proteins that determine the observed change in the IMFs of blood sera (Figure 4). Some of them have been previously linked to cancer: unexplained hypoalbuminaenia has been assosiated with increased cancer risk,[67] and low pre-treatment albumin level – with poor survival rate. [68] Moreover, in line with our findings, the levels of haptoglobin, complement

component C3, alpha-1 antytrypsin and alpha-1-acid glycoprotein were previously shown to rise in blood of lung cancer patients.[60–62,65]

Importantly, although these proteins are not specifically challenging to detect and measure, they have previously not been used in a combined fashion to help detect or diagnose lung cancer. It is meanwhile widely accepted that using multiple biomarking molecules together, as a pattern, is more effective and robust for detecting a particular health condition.[30,64,69,70] Infrared fingerprinting of human blood serum takes this approach to a new level: here we effectively combine a wide range of molecules into a single IR spectrum, that can be easily measured and interpreted. To illustrate that, we considered the levels of all 114 proteins detected by proteomics in every sample. Importantly, the binary classification efficiency based on all these proteins measured separately is not higher than the efficiency based on a single IMF measurement (Table S6).

Lung cancer induces a number of changes in the levels of blood serum proteins that have been previously linked to acute-phase response, and it is well-known that cancer is often associated with inflammatory states.[41,71] In line with the general discussion in the field,[22] our findings underscore the need for additional clinical studies that would look into the specificity of IMFs. A well-designed reference cohort should include individuals with potentially similar pattern of changes in the blood composition: for example, in the case of lung cancer, with chronic or acute inflammation. Due to cost-efficiency and rapidity of blood-based infrared molecular fingerprinting, it could still find a wide range of applications, even if its specificity proves insufficient for screening applications. Thus, general molecular-level understanding of the disease-related changes in IMFs will help establish better clinical study design, and ultimately lead to improved approaches to medical diagnostics.

Conclusion

As the focus of future healthcare is shifting from treatment to early detection and prevention, such rapid, cost-effective and holistic approaches as infrared molecular fingerprinting of body liquids will become ever more relevant. So far, infrared spectral changes in complex bioliquids were linked to multiple diseases but have remained

uninterpretable with regard to which specific molecule accounts for a spectral change. In this study we looked systematically into the contributions of different constituents of blood serum to the overall IMF. In particular, we showed that the IMFs of blood serum can be to a high extent modelled using the concentrations of the ten most abundant proteins. With non-metastatic lung cancer as an example of a medical condition, we showed that a number of highly abundant acute-phase proteins are up- and down-regulated in cancer patients compared to the reference group, leading to an observable change in the IMFs of blood serum. Accompanied by a meaningful molecular annotation, this change is more likely to find its use in everyday clinical practice.

The paradigm presented here could in principle be used for any pathophysiological condition. After having recorded the IMFs of patients and compared them to matched reference individuals, one could use biochemical fractionation to determine which molecular class is responsible for the disease-related differences and perform in-depth omics profiling of the identified fraction (Figure 4). This would provide insights into the nature of information that infrared molecular fingerprinting is able to provide and into its additional value compared to well-established clinical tests. Moreover, combining biochemical fractionation with field-resolved spectroscopy-based infrared molecular fingerprinting[66] might yield deeper molecular insight along with higher specificity and sensitivity for disease detection. Ultimately, the larger clinical studies with purposefully chosen reference groups, stratified and controlled for comorbidities, may bring IMF – a cheap and time-efficient method – closer to everyday clinical use.

Acknowledgements

This work was funded by the Center for Advanced Laser Applications (CALA) of the Ludwig Maximilians University Munich (LMU), Department of Laser Physics, and the Max Planck Institute of Quantum Optics (MPQ), Laboratory for Attosecond Physics, Germany. We would like to thank Frank Fleischmann, Catherine Vasilopoulou, Jacqueline Hermann, Katja Leitner, Sigrid Auweter, Daniel Meyer, Beate Rank and Incinur Zellhuber for their help with this study. In particular, we wish to acknowledge the efforts of many individuals who participated as volunteers in the clinical study reported here. We also thank A. Barth for his insightful suggestions.

References

[1] P. R. Griffiths, J. A. de Haseth, in *Fourier Transform Infrared Spectrom.*, **2007**, pp. 1–18.

[2] L. Lechowicz, M. Chrapek, J. Gaweda, M. Urbaniak, I. Konieczna, *Mol. Biol. Rep.* **2016**, *43*, 1321–1326.

[3] J. Titus, E. Viennois, D. Merlin, A. G. Unil Perera, *J. Biophotonics* **2017**, *10*, 465–472.

[4] F. Elmi, A. F. Movaghar, M. M. Elmi, H. Alinezhad, N. Nikbakhsh, *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2017**, *187*, 87–91.

[5] X. Yang, T. Fang, Y. Li, L. Guo, F. Li, F. Huang, L. Li, *Optik (Stuttg).* **2019**, *180*, 189–198.

[6] H. J. Byrne, I. Behl, G. Calado, O. Ibrahim, M. Toner, S. Galvin, C. M. Healy, S. Flint, F. M. Lyng, *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2021**, *252*, 119470.

[7] I. Maitra, C. L. M. Morais, K. M. G. Lima, K. M. Ashton, R. S. Date, F. L. Martin, *Analyst* **2019**, *144*, 7447–7456.

[8] A. L. Mitchell, K. B. Gajjar, G. Theophilou, F. L. Martin, P. L. Martin-Hirsch, *J. Biophotonics* **2014**, *7*, 153–165.

[9] P. Carmona, M. Molina, M. Calero, F. Bermejo-Pareja, P. Martínez-Martín, A. Toledano, *J. Alzheimer's Dis.* **2013**, *34*, 911–920.

[10] H. J. Butler, P. M. Brennan, J. M. Cameron, D. Finlayson, M. G. Hegarty, M. D. Jenkinson, D. S. Palmer, B. R. Smith, M. J. Baker, *Nat. Commun.* **2019**, *10*, 1–9.

[11] M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, M. Trubetskov, N. Harbeck, F. Krausz, M. Žigman, *Nat. Commun.* **2021**.

[12] M. Paraskevaidi, C. L. M. Morais, K. M. G. Lima, J. S. Snowden, J. A. Saxon, A. M. T. Richardson, M. Jones, D. M. A. Mann, D. Allsop, P. L. Martin-Hirsch, F. L. Martin, *Proc. Natl. Acad. Sci.* **2017**, *114*, E7929–E7938.

[13] T. G. Mayerhöfer, S. Pahlow, J. Popp, *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2021**, *251*, 119411.

[14] S. Roy, D. Perez-Guaita, D. W. Andrew, J. S. Richards, D. McNaughton, P. Heraud, B. R. Wood, *Anal. Chem.* **2017**, *89*, 5238–5245.

[15] W. Petrich, K. B. Lewandrowski, J. B. Muhlestein, M. E. H. Hammond, J. L. Januzzi, E. L. Lewandrowski, R. R. Pearson, B. Dolenko, J. Früh, M. Haass, M. M. Hirschl, W. Köhler, R. Mischler, J. Möcks, J. Ordóñez-Llanos, O. Quarder, R. Somorjai, A. Staib, C. Sylvén, G. Werner, R. Zerback, *Analyst* **2009**, *134*, 1092–1098.

[16] B. Bird, M. Miljkovi, S. Remiszewski, A. Akalin, M. Kon, M. Diem, *Lab. Investig.* **2012**, *92*, 1358–1373.

[17] A. Sala, D. J. Anderson, P. M. Brennan, H. J. Butler, J. M. Cameron, M. D. Jenkinson, C. Rinaldi, A. G. Theakstone, M. J. Baker, *Cancer Lett.* **2020**, *477*, 122–130.

[18] M. J. Baker, S. R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. a. Lukaszewski, G. Thiéfin, G. D. Sockalingum, *Chem. Soc. Rev.* **2016**, *45*, 1803–1818.

[19] V. Balan, C. Mihai, F. Cojocaru, C. Uritu, G. Dodi, D. Botezat, I. Gardikiotis, *Materials (Basel)*. **2019**, *12*, 2884.

[20] L. Morandini, T. Deprá, M. Alva, S. Martinho, *Vib. Spectrosc.* **2019**, *100*, 195–201.

[21] C. Paluszkiewicz, E. Pięta, M. Woźniak, N. Piergies, A. Koniewska, W. Ścierski, M. Misiołek, W. M. Kwiatek, *J. Mol. Liq.* **2020**, *307*, DOI 10.1016/j.molliq.2020.112961.

[22] M. Diem, *J. Biophotonics* **2018**, *11*, 1–6.

[23] R. A. Shaw, S. Kotowich, M. Leroux, H. H. Mantsch, *Ann. Clin. Biochem. An Int. J. Biochem. Lab. Med.* **1998**, *35*, 624–632.

[24] I. Elsohaby, J. T. McClure, C. B. Riley, J. Bryanton, K. Bigsby, R. A. Shaw, *J. Pharm. Biomed. Anal.* **2018**, *150*, 413–419.

[25] K. Spalding, F. Bonnier, C. Bruno, H. Blasco, R. Board, I. Benz-de Bretagne, H. J. Byrne, H. J. Butler, I. Chourpa, P. Radhakrishnan, M. J. Baker, *Vib. Spectrosc.* **2018**, *99*, 50–58.

[26] C. Petibois, G. Cazorla, A. Cassaigne, G. Déléris, *Clin. Chem.* **2001**, *47*, 730–8.

[27] G. Hoşafçi, O. Klein, G. Oremek, W. Mäntele, *Anal. Bioanal. Chem.* **2007**, *387*, 1815–1822.

[28] H. J. Byrne, F. Bonnier, J. McIntyre, D. R. Parachalil, *Clin. Spectrosc.* **2020**, *2*, 100004.

[29] H. Fabian, P. Lasch, D. Naumann, *J. Biomed. Opt.* **2005**, *10*, 031103.

[30] N. L. Anderson, N. G. Anderson, *Mol. Cell. Proteomics* **2002**, *1*, 845–867.

[31] F. Bonnier, H. Blasco, C. Wasselet, G. Brachet, R. Respaud, L. F. C. S. Carvalho, D. Bertrand, M. J. Baker, H. J. Byrne, I. Chourpa, *Analyst* **2017**, *142*, 1285–1298.

[32] F. Bonnier, G. Brachet, R. Duong, T. Sojinrin, R. Respaud, N. Aubrey, M. J. Baker, H. J. Byrne, I. Chourpa, *J. Biophotonics* **2016**, *9*, 1085–1097.

[33] D. R. Parachalil, C. Bruno, F. Bonnier, H. Blasco, I. Chourpa, J. McIntyre, H. J. Byrne, *Analyst* **2019**, *144*, 4295–4311.

[34] C. Hughes, M. Brown, G. Clemens, A. Henderson, G. Monjardez, N. W. Clarke, P. Gardner, *J. Biophotonics* **2014**, *7*, 180–188.

[35] F. Bonnier, M. J. Baker, H. J. Byrne, *Anal. Methods* **2014**, *6*, 5155–5160.

[36] E. J. Want, G. O'Maille, C. A. Smith, T. R. Brandon, W. Uritboonthai, C. Qin, S. A. Trauger, G. Siuzdak, *Anal. Chem.* **2006**, *78*, 743–752.

[37] D. A. Colantonio, C. Dunkinson, D. E. Bovenkamp, J. E. Van Eyk, *Proteomics* **2005**, *5*, 3831–3835.

[38] S. Tulipani, R. Llorach, M. Urpi-Sarda, C. Andres-Lacueva, *Anal. Chem.* **2013**, *85*, 341–348.

[39] D. Perez-Guaita, S. Garrigues, M. de la Guardia, *TrAC - Trends Anal. Chem.* **2014**, *62*, 93–105.

[40] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, M. Mann Correspondence, M. Mann, *Cell Syst.* **2016**, *2*, 185–195.

[41] S. B. Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, C. Dive, *Open Biol.* **2017**, *7*, 170070.

[42] J. Mo Ahn, J. Yoel Cho, *J. Mol. Biomark. Diagn.* **2015**, *s4*, 1–7.

[43] J. Ollesch, D. Theegarten, M. Altmayer, K. Darwiche, T. Hager, G. Stamatis, K. Gerwert, *Biomed. Spectrosc. Imaging* **2016**, *5*, 129–144.

[44] X. Wang, X. Shen, D. Sheng, X. Chen, X. Liu, *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2014**, *122*, 193–197.

[45] D. K. R. Medipally, D. Cullen, V. Untereiner, G. D. Sockalingum, A. Maguire, T. N. Q. Nguyen, J. Bryant, E. Noone, S. Bradshaw, M. Finn, M. Dunne, A. M. Shannon, J. Armstrong, A. D. Meade, F. M. Lyng, *Ther. Adv. Med. Oncol.* **2020**, *12*, 1–23.

[46] H. Ghimire, C. Garlapati, E. A. M. Janssen, U. Krishnamurti, G. Qin, R. Aneja, A. G. Unil Perera, *Cancers (Basel).* **2020**, *12*, 1–17.

[47] C. Hughes, G. Clemens, B. Bird, T. Dawson, K. M. Ashton, M. D. Jenkinson, A. Brodbelt, M. Weida, E. Fotheringham, M. Barre, J. Rowlette, M. J. Baker, *Sci. Rep.* **2016**, *6*, 20173.

[48] J. Ollesch, M. Heinze, H. M. Heise, T. Behrens, T. Brüning, K. Gerwert, *J. Biophotonics* **2014**, *7*, 210–221.

[49] K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch, F. L. Martin, *Analyst* **2013**, *138*, 3917–3926.

[50] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, *IEEE Access* **2020**, *8*, 42200–42216.

[51] A. Barth, *Biochim. Biophys. Acta* **2007**, *1767*, 1073–1101.

[52] K. Z. Liu, R. A. Shaw, A. Man, T. C. Dembinski, H. H. Mantsh, *Clin. Chem.* **2002**, *48*, 499–506.

[53] D. Vuckovic, *Anal. Bioanal. Chem.* **2012**, *403*, 1523–1548.

[54] R. Aebersold, M. Mann, *Nature* **2003**, *422*, 198–207.

[55] R. Aebersold, M. Mann, *Nature* **2016**, *537*, 347–355.

[56] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, M. Mann, *Mol. Cell. Proteomics* **2014**, *13*, 2513–2526.

[57] T. Fournier, N. Medjoubi-N, D. Porquet, *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **2000**, *1482*, 157–171.

[58] N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T. L. Pedersen, S. R. Smith, F. Bamforth, R. Greiner, B. McManus, J. W. Newman, T. Goodfriend, D. S. Wishart, *PLoS One* **2011**, *6*, DOI 10.1371/journal.pone.0016957.

[59] P. Zhao, J. Wu, F. Lu, X. Peng, C. Liu, N. Zhou, M. Ying, *BMC* **2019**, *19*, 1–11.

[60] P. Dowling, C. Clarke, K. Hennessy, B. Torralbo-Lopez, J. Ballot, J. Crown, I. Kiernan, K. J. O'Byrne, M. J. Kennedy, V. Lynch, M. Clynes, *Int. J. Cancer* **2012**, *131*, 911–923.

[61] W. M. Gao, R. Kuick, R. P. Orchekowski, D. E. Misek, J. Qiu, A. K. Greenberg, W. N. Rom, D. E. Brenner, G. S. Omenn, B. B. Haab, S. M. Hanash, *BMC Cancer* **2005**, *5*, 1–10.

[62] P. A. Ganz, M. Baras, P. Y. Ma, R. M. Elashoff, *Cancer Res.* **1984**, *44*, 5415 LP – 5421.

[63] R. Gasparri, G. Sedda, R. Noberini, T. Bonaldi, L. Spaggiari, *Proteomics - Clin. Appl.* **2020**, *14*, 1–13.

[64] Y. I. Kim, J. M. Ahn, H. J. Sung, S. S. Na, J. Hwang, Y. Kim, J. Y. Cho, *J. Proteomics* **2016**, *148*, 36–43.

[65] T. N. Zamay, G. S. Zamay, O. S. Kolovskaya, R. A. Zukov, M. M. Petrova, A. Gargaun, M. V. Berezovski, A. S. Kichkailo, *Cancers (Basel)*. **2017**, *9*, 1–22.

[66] I. Pupeza, M. Huber, M. Trubetskov, W. Schweinberger, S. A. Hussain, C. Hofer, K. Fritsch, M. Poetzlberger, L. Vamos, E. Fill, T. Amotchkina, K. V. Kepesidis, A. Apolonski, N. Karpowicz, V. Pervak, O. Pronin, F. Fleischmann, A. Azzeer, M. Žigman, F. Krausz, *Nature* **2020**, *577*, 52–59.

[67] F. Hamilton, R. Carroll, W. Hamilton, C. Salisbury, *Br. J. Cancer* **2014**, *111*, 1410–1412.

[68] D. Gupta, C. G. Lis, *Nutr. J.* **2010**, *9*, 69.

[69] S. Ma, W. Wang, B. Xia, S. Zhang, H. Yuan, H. Jiang, W. Meng, X. Zheng, X. Wang, *EBioMedicine* **2016**, *11*, 210–218.

[70] Y. Fan, S. Wang, F. Zhang, *Angew. Chemie Int. Ed.* **2019**, *58*, 13208–13219.

[71] J. Watson, C. Salisbury, J. Banks, P. Whiting, W. Hamilton, *Br. J. Cancer* **2019**, *120*, 1045–1051.

3.7. Article 7: Cohort Profile: MUNICH Preterm and Term Clinical Study (MUNICH-PreTCl)

Authors:
Susanne Pangratz-Fuehrer[1*], Orsolya Genzel-Boroviczény[1], Wolfgang Bodensohn[1], Robin Eisenburger[1], Janne Scharpenack[1], Philipp E. Geyer[2,3], Johannes B. Müller-Reif[2], Nadja van Hagen[1], Alina M. Müller[1], Majken K. Jensen[4,5], Christoph Klein[1], Matthias Mann[2] and Claudia Nussbaum[1]

1 Division of Neonatology, Department of Pediatrics, Dr. von Hauner Children´s Hospital, LMU University Hospital, Munich, Germany,

2 Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany, 3OmicEra Diagnostics GmbH, Planegg Germany,

3 OmicEra Diagnostics GmbH, Planegg, Germany,

4 Departments of Nutrition & Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts,

5 Section of Epidemiology, Department of Public Health, University of Copenhagen, Denmark.

The manuscript ‘Cohort Profile: MUNICH Preterm and Term Clinical Study (MUNICH-PreTCl)’ originates from a collaborative effort between the MPI of Biochemistry and the Neonatology of the LMU. The overall aim of the study was to collect a unique cohort of term and preterm born dried blood spots and plasma for omics analyses and support this with an extensive clinical and epidemiological record. To this end, surveys with the participating parents were conducted and routine laboratory parameters recorded to build up detailed meta data. The cohort profile is a description and examination of the participating parents and newborn infants and holds all the epidemiological information to make the follow up work more comprehensible. With our description of simple correlations between parental age and body weight or infant gender with the gestational age at birth we confirm and extend previous studies’ findings of risk factors for preterm delivery. In the future, with the unbiased investigation of the cohort’s blood proteome, we hope to uncover new biomarkers to stratify infant subgroups with long-term impact from postnatal complications and generally understand the effect of pre- and postnatal factors on long-term health.

Cohort Profile: MUNICH Preterm and Term Clinical Study (MUNICH-PreTCI)

Susanne Pangratz-Fuehrer[1]*, Orsolya Genzel-Boroviczény[1], Wolfgang Bodensohn[1], Robin Eisenburger[1], Janne Scharpenack[1], Philipp E. Geyer[2,3], Johannes B. Müller-Reif[2,3], Nadja van Hagen[1], Alina M. Müller[1], Majken K. Jensen[4,5], Christoph Klein[1], Matthias Mann[2] and Claudia Nussbaum[1]

[1]Division of Neonatology, Department of Pediatrics, Dr. von Hauner Children´s Hospital, LMU University Hospital, Munich, Germany, [2]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany, [3]OmicEra Diagnostics GmbH, Planegg, Germany, [4]Departments of Nutrition & Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, [5]Section of Epidemiology, Department of Public Health, University of Copenhagen, Denmark.

* To whom correspondence should be addressed:
Susanne Pangratz-Fuehrer
Susanne.PangratzFuehrer@med.uni-muenchen.de
Kinderklinik und Kinderpoliklinik
Im Dr. von Haunerschen Kinderspital
Campus Innenstadt, Lindwurmstraße 4, 80337 München

Running Header: Munich Preterm and Term Cohort
Number of figures: 1
Number of tables: 6
Number of pages: 21
Word count: 3779 (+abstract 284)

We thank the staff of the NICU and the newborn ward for their excellent and kind cooperation.

Abbreviations:

ASD: atrial septal defect, BMI: bodymass index, BPD: bronchopulmonary disease, GA: gestational age, GDM: gestational diabetes mellitus, GHT: gestational hypertension, GWG; gestational weight gain, LGA: large for GA, SGA: small for GA, ICH: intracerebral hemorrhage, IVH: intraventricular hemorrhage, PDA: patent ductus arteriosus, PFO: patent foramen ovale, RDS: respiratory distress syndrome, VSD: ventricular septal defect

ABSTRACT

Purpose: The MUNICH Preterm and Term Clinical (MUNICH-PreTCl) prospective birth cohort was established to uncover pathological processes contributing to infant/childhood morbidity and mortality. We collected comprehensive medical information of healthy and sick newborns and their families, together with infant blood samples for proteomic analysis. MUNICH-PreTCl aims to identify mechanism-based biomarkers in infant health and disease to deliver more precise diagnostic and predictive information for disease prevention. We particularly focused on risk factors for pregnancy complications, family history of genetically influenced health conditions such as diabetes, and pediatric long-term health - all to be further monitored and correlated with proteomics data in the future.

Participants: Newborns and their parents were recruited from the Perinatal Center at the LMU University Hospital, Munich, between February 2017 and June 2019. Infants without congenital anomalies, delivered at 23–41 weeks of gestation, were eligible.

Findings: Findings to date concern the clinical data and extensive personal patient information. A total of 662 infants were recruited, 44% were female (36% in preterm, 46% in term). 90% of approached families agreed to participate. Neonates were grouped according to gestational age: extremely preterm (< 28 weeks, *N*=28), very preterm (28 to < 32 weeks, *N*=36), late preterm (32 to < 37 weeks, *N*=97) and term infants (> 37+0 weeks, *N*=501). We collected over 450 data points per child-parent set, (family history, demographics, pregnancy, birth, and daily follow-ups throughout hospitalization) and 841 blood samples longitudinally. The completion rates for medical exams and blood samples were 100%, and 95% for the questionnaire.

Future Plans: The correlation of large clinical datasets with proteomic phenotypes, together with the use of medical registries, will enable future investigations aiming to decipher mechanisms of disorders in a systems biology perspective.

This observational cohort is registered with DRKS (00024189).

STRENGTHS AND LIMITATIONS

- The MUNICH-PreTCI study is a prospective birth cohort consisting of 662 infants (501 full term and 161 preterm) recruited from a single Perinatal Center in Germany.
- We collected phenotypical information including clinical data from maternal and neonatal medical records, family history and demographics survey and blood samples at defined time points for proteomic screening.
- Recruiting at the university hospital and restricting the survey to German and English-speaking parents likely introduced some bias as mostly caucasian families from urban living environment and with a higher educational level participated in the study.
- Combining population-based cohort studies with proteomic screening provides an opportunity to relate the functional protein network status to specific pre- and postnatal factors as well as clinical outcomes recorded at time of birth and during follow-up.
- This hypothesis-free approach may enable the identification of biomarkers for the etiologic understanding of complex multi-factorial short and long-term diseases in infants.

INTRODUCTION

Early-life exposure to environmental impact factors (e.g., air pollution, noise, chemicals and pesticides) and family medical history can alter ontogenic trajectories in fundamental and often unforeseen ways. These often result in clinically important outcomes during pregnancy, at birth and in the long-term. In particular, preterm birth is still associated with an increased incidence of complications, despite advances in neonatal intensive care [1]. The pathogenesis is thought to be of multifactorial origin, involving the exacerbating interaction of genetic components with a multitude of environmental risk factors [2]. Furthermore, it is conceivable, that poor outcomes among preterm infants are not solely associated with being born too early, but that the underlying reasons for prematurity itself could be of even greater importance [3].

This raises the question why some infants develop diseases, while others are resilient despite potentially carrying a heightened risk for morbidity. Resilient individuals may provide important clues for improved disease prevention but can only be identified when patients are compared within a particular group sharing a defining characteristic or common event, such as birth. This highlights the need for population-based-pregnancy groups and birth cohorts to investigate environmental and genetic factors with the purpose of increasing our understanding of the origins of health and disease - starting as early as in pre-pregnancy. Previous neonatal cohort studies have used maternal biomarkers to explore the influence of potentially disease-causing environmental factors on long-term health outcomes of the child [4 5]. Moreover, pre-pregnancy obesity and abnormal weight gain during early gestation have been associated with an adverse cardio-metabolic profile in the offspring, including but not limited to higher childhood body mass index (BMI), body fat and systolic blood pressure [6]. In particular, the third trimester - a critical period for brain and lung development as well as metabolic programming - is now understood to be remarkably sensitive to disruptive factors like alcohol and drug abuse, stress, and malnutrition [7]. The advantage of combining population-based cohort studies with collection of body fluids during hospitalization, is that it provides an opportunity for an unbiased assessment of circulating proteins in blood and plasma by highly informative OMICS technologies, such as plasma proteome profiling, in effect phenotyping the individuals [8]. The results of biospecimen measurements can then be related to specified pre- and postnatal factors as well as clinical outcomes recorded at time of birth and during follow-up. This hypothesis-free approach may enable the identification of biomarkers and disease-modifying effects for the etiologic understanding of complex multi-factorial short and long-term diseases in infants.

COHORT DESCRIPTION

Study aims

MUNICH-PreTCI was designed as a comprehensive cohort study, enrolling both, preterm and term infants. Our overall objectives were to firstly elucidate the role of gestational age, pre- and postnatal environmental exposures, and family demographic and medical history in determining the risk of neonatal morbidity among preterm and term infants. Secondly, we wanted to establish a thorough baseline assessment for future follow-ups regarding disease occurrence. To this end, we systematically collected comprehensive datasets on healthy infants with and without specific susceptibilities to diseases, on sick infants and on critically ill infants. We monitored the medical status of enrolled neonates, recorded their families' backgrounds and obtained blood samples for proteomic analysis from each infant. With these data at hand, we can investigate the impact of specific medical and environmental conditions on neonatal health outcomes in an attempt to preventively improve the lives of preterm infants, critically ill patients, and their families.

Study design

The MUNICH-PreTCI study is a prospective cohort of 662 neonates. Between February 2017 and June 2019, we included 501 full-term neonates recruited from our maternity ward and 161 preterm infants upon admission to our neonatal intensive care unit at the Perinatal Center, Campus Innenstadt, University Hospital, LMU, Munich, Germany. For preterm infants, we defined three subgroups as follows: infants born at less than 28 completed weeks of gestation (*extremely* preterm, *N*=28), infants born between 28+0 to 31+6 weeks (*very* preterm, *N*=36) and infants born between 32+0 to 36+6 weeks (late preterm, *N*=97).

Data collection

Participant recruitment and informed consent

Mothers of eligible infants were approached for enrolment after giving birth at the Perinatal Center, Campus Innenstadt, University Hospital, LMU, Munich, Germany. Full informed consent was given by mothers for the formation of a comprehensive dataset derived from maternal and infant medical records, a parental questionnaire, and for collecting blood samples from the infant during routine blood sampling. MUNICH-PreTCI was approved by the ethics board of the Medical Faculty of the Ludwig-Maximilians-University, Munich, Germany.

Our patient recruitment process consisted of three stages, with stage 1 and 3 being specifically dedicated to preterm infants. During stage 1 (pilot phase), we enrolled neonates with a GA of 23 to 36 completed weeks, born between February 2017 and May 2018, aiming to establish a

workflow for patient recruitment and sample collection for proteomic analyses. We collected data throughout the hospitalization of mothers and infants including their medication and medical procedures, but we did not record their detailed family history and demographics.
Stage 2 commenced in June 2018 and continued through December 2018. During this stage we expanded the recruitment focus by enrolling preterm and term infants, as well as their parents, who received a comprehensive health survey with detailed medical and family history to determine the role of potentially modifiable factors contributing to long-term developmental outcome. Stage 3 was launched in January 2019 essentially constituting an extension of stage 2 but with the objective of collecting a maximum number of extremely preterm and very preterm infants, as well as augmenting specific groups of interest, such as neonates with neonatal infections, diabetic mothers or being "small for gestational age" (< 10^{th} age adjusted weight percentile) *(Figure 1).*

We identified two main reasons for failure to enroll eligible patients in MUNICH-PreTCI:
First, the language barrier - insufficient German or English language proficiency, and second, organizational and logistical challenges due to parents' absence. Notably, some parents enrolled their newborn in our study, but declined to complete the questionnaire. This was rarely observed for families of term infants (5%), but mainly for parents of extremely (45%) and very preterm infants (23%), who appeared to be exhausted having to deal with longterm hospitalization and the illness of their premature child. During stage 1, 94 families out of 294 eligible infants were approached by our study team, and 80 of these subsequently enrolled their child, while parents of 14 infants declined recruitment (85% participation). For stage 2, we identified 1173 eligible patients of preterm and term infants. Parents of 619 of these eligible patients were approached and 549 consented to study participation (89%). For 520 of them a completed survey could be secured (95% participation). In Stage 3 we contacted the families of 38 of 265 eligible patients, of which 33 were enrolled and 5 declined consent (87% participation).

Clinical data and questionnaire

Data were collected from enrolled infants throughout their hospital stay using medical records, results from medical and laboratory examinations and parental questionnaires to characterize current and previous pregnancies and births, education, life-style patterns, as well as chronic health outcomes of their families. Events before or during pregnancy were documented retrospectively. GA was estimated by experienced obstetricians using the mother's last menstrual period, as well as the first trimester ultrasound. GA is expressed in completed weeks (or completed days), such that events occurring 210 to 216 completed days after the onset of the last period

were considered to be at 30 weeks of gestation. Data were then entered on-site into a secure and pseudonymized database by trained doctoral students with password protection for confidentiality. *Table 1* provides an overview of domains and measurements collected in the course of MUNICH-PreTCI.

Blood samples

Infant blood samples were taken at pre-defined time points: At first routine blood sampling after birth, at newborn screening 36 - 48 hours after birth, pre- and post-antibiotic treatment, at adjusted 32 weeks of GA and at discharge from the clinic. Blood spots were collected on Whatman cards, which were stored frozen at -80° for mass spectrometry-based proteomic analysis.

Domains	Child	Parents	Siblings	Ext. Family	Assessment
Infants: birth characteristics, measurements, medical complications, treatments (*Table 2*)	X				Med. records
Pregnancy: prenatal screenings, influences on pregnancy, substance abuse (*Table 3*)		X	X		Med. records; self-report
Delivery : duration, anesthesia, previous deliveries, abortions (*Table 4*)		X	X		Med. records; self-report
Anthropometric / Demographic Data: physical measurements, education, ethnicity (*Table 5*)		X			Med. records; self-report
Family Medical History: allergies, cardiovascular, endocrine, neurological disorders (*Table 6*)		X	X	X	Med. records; self-report

Table 1: Overview of domains and measurements for MUNICH-PreTCI

FINDINGS TO DATE

Study population

Over a period of 27 months, 662 infants (501 full term and 161 preterm) and their parents were enrolled in MUNICH-PreTCI. The questionnaire was handed out to the parents of 582 infants and completed by 550 (95% participation rate). Since the survey did not start until stage 2, a subsequently smaller number of preterm infants´ parents submitted the family questionnaire (78 out of 90 preterm infants and 472 out of 492 term infants). Furthermore, the survey participation rate for families with extremely preterm infants was only 55% and thus much lower compared to the other preterm age groups with 77% for very preterm and 94% for late preterm infants.

Baseline characteristics of neonatal study cohort

The baseline characteristics of all enrolled infants are presented in *Table 2*. Overall, a smaller proportion of participants in this cohort is female (44%), in particular within the group of preterm infants (36%) compared to the term group (46%). These numbers are in accordance with previous studies that report preterm birth to be more common (55%) in male infants[9]. Not only the gestational age, but also the birth weight is pivotal in the classification of an infant´s condition. In our cohort, the mean birth weight was 3422 g for term infants, 725 g for extremely preterm, 1299 g for very preterm infants, and 2240 g for late preterm infants. As per the World Health Organization (WHO), the term "low birth weight" (LBW) is defined as an absolute weight of < 2500 g, regardless of gestational age, and can be further categorized into very low birth weight (VLBW, <1500 g) and extremely low birth weight (ELBW, <1000 g), which generally comprises the youngest preterm infants with highest risk for complications. Among all preterm infants, 33 (21%) can be categorized as ELBW and 62 (39%) as VLBW. Low birth weight can also be an indicator for the infant being too "small for gestational age" (SGA), which refers to infants whose birth weight is below the 10th percentile for GA, due to slow prenatal growth rates caused by maternal health issues, placental complications, or genetics[10]. Neonates born "large for gestational age" (LGA), defined as weight above the 90th percentile, are also associated with significantly higher rates of neonatal morbidity[11]. There was a higher proportion of SGA infants in the group of extremely preterm infants (21%) compared to any other group (6-12%), while the distribution of LGA infants was about equal for each GA (3-5%).

GA *(completed weeks)*	Preterm *Extreme* < 28 weeks	Preterm *Very* 28 to 31 weeks	Preterm *Late* 32 to 36 weeks	Term > 37 weeks	Total 23 to 41 weeks
Infants *N*	*28*	*36*	*97*	*501*	*662*
Birth Assessment: Infant					
GA (weeks) *M* (SD)	25 (1.3)	29.3 (1.2)	34.2 (1.4)	39.4 (1.2)	37.5 (4)
GA (days) *M* (SD)	178 (9)	208 (9)	243 (10)	279 (9)	265 (28)
Sex, female *N* (%)	12 (43)	8 (22)	42 (43)	232 (46)	294 (44)
Birth weight (g) *M* (SD)	725 (174)	1300 (315)	2241 (511)	3423 (474)	3020 (900)
BW < 1000 g (ELBW) *N* (%)	26 (93)	6 (17)	1 (1)	0 (0)	33 (5)
BW < 1500 g (VLBW) *N* (%)	28 (100)	28 (78)	6 (6.2)	0 (0)	62 (9.4)
Pctl. BW (%) *M* (SD)	45 (31.5)	45 (22)	43 (25)	47 (27)	46 (26)
SGA (low BW for GA) *N* (%)	6 (21)	2 (6)	12 (12)	50 (10)	70 (11)
LGA (high BW for GA) *N* (%)	1 (4)	1 (3)	4 (4)	26 (5)	32 (5)
Birth length (cm) *M* (SD)	32 (3)	39 (3)	45 (4)	52 (3)	49 (6)
Pctl. Birth length (%) *M* (SD)	53 (33)	45 (28)	54 (30)	60 (28)	58 (29)
Head Circumf. (cm) *M* (SD)	23 (2)	28 (2)	32 (2)	35 (1)	34 (3)
Pctl. Head Circumf. (%) *M* (SD)	54 (27)	59 (28)	58 (26)	53 (28)	54 (27)
APGAR Score 1, *M* (SD)	5.6 (2.2)	6.7 (2.3)	8 (1.6)	8.8 (1.7)	8.4 (1.9)
APGAR Score 5, *M* (SD)	7.6 (1.9)	8.4 (1.6)	9.1 (1.1)	9.6 (1.3)	9.4 (1.4)
APGAR Score 10, *M* (SD)	8.6 (1.6)	9.2 (1)	9.6 (0.7)	10 (1)	9.7 (1.1)
Multiple Births					
Singles *N* (%)	20 (71)	19 (53)	58 (60)	485 (97)	582 (88)
Multiples *N* (%)	8 (29)	17 (47)	39 (40)	16 (3)	80 (12)
Birth Mode					
Spontaneous vaginal *N* (%)	6 (21)	9 (25)	28 (29)	238 (48)	281 (42)
Induced vaginal birth *N* (%)	0 (0)	0 (0)	7 (7)	63 (13)	70 (11)
Vacuum extract., forceps *N* (%)	0 (0)	0 (0)	11 (11)	77 (15)	88 (13)
C-Section *N* (%)	22 (79)	27 (75)	51 (53)	123 (25)	223 (34)
Significant Diagnosis					
Asphyxia *N* (%)	1 (4)	0 (0)	1 (1)	12 (2)	14 (2)
Cardiovascular *N* (%)	23 (82)	13 (36)	9 (9)	29 (6)	74 (11)
Hypo- /Hypertension *N* (%)	7 (25)	5 (14)	4 (4)	8 (2)	24 (4)
ASD or PFO *N* (%)	9 (32)	10 (28)	4 (4)	6 (1)	29 (4)
VSD *N* (%)	0 (0)	0 (0)	0 (0)	3 (0.6)	3 (0.5)
PDA *N* (%)	16 (57)	4 (11)	0 (0)	4 (0.8)	24 (4)
Hematological *N* (%)	22 (79)	10 (28)	14 (14)	19 (4)	65 (10)
Thrombocytopenia *N* (%)	11 (39)	5 (14)	8 (8)	11 (2)	35 (5)
Anemia *N* (%)	22 (79)	6 (17)	0 (0)	3 (0.6)	31 (5)
Polyglobulia *N* (%)	0 (0)	1 (3)	5 (5)	5 (1)	11 (2)
Coagulation disorder *N* (%)	4 (14)	3 (8)	1 (1)	1 (0.2)	9 (1)
Infections *N* (%)	21 (75)	5 (14)	7 (7)	41 (8)	74 (11)
Suspect. Infections *N* (%)	7 (25)	29 (81)	46 (47)	45 (9)	127 (19)
Neurological *N* (%)	12 (43)	14 (39)	11 (11)	15 (3)	52 (8)
ICH or IVH, any grade *N* (%)	9 (32)	5 (14)	2 (2)	0 (0)	16 (2)
HIE *N* (%)	0 (0)	0 (0)	0 (0)	10 (2)	10 (2)
Abnormal ABR *N* (%)	2 (7)	1 (3)	2 (2)	18 (4)	23 (4)
Respiratory *N* (%)	28 (100)	36 (100)	45 (46)	51 (10)	160 (24)
RDS *N* (%)	27 (96)	32 (89)	25 (26)	11 (2)	95 (14)
Resp. insufficiency *N* (%)	22 (79)	17 (47)	16 (17)	18 (4)	73 (11)
Apnea *N* (%)	18 (64)	21 (58)	7 (7)	1 (0.2)	47 (7)
BPD, any grade *N* (%)	13 (46)	3 (8)	0 (0)	0 (0)	16 (2)
Pneumothorax *N* (%)	7 (25)	1 (3)	1 (1)	5 (1)	14 (2)
ROP, any grade *N* (%)	15 (54)	3 (8)	0 (0)	0 (0)	18 (3)
Treatments					
Antibiotics *N* (%)	28 (100)	34 (94)	53 (55)	86 (17)	201 (30)
Antimycotic Prophylaxis *N* (%)	27 (96)	34 (94)	51 (53)	83 (17)	195 (30)
Blood Transfusion *N* (%)	14 (50)	4 (11)	2 (2)	2 (0.4)	22 (3)
Surfactant *N* (%)	27 (96)	22 (61)	3 (3)	2 (0.4)	54 (8)
Ventilatory Supp. *N* (%)	28 (100)	35 (97)	44 (45)	39 (8)	136 (21)

Table 2: Baseline characteristics of study participants

Infants born at the earliest GA are at the highest risk for severe morbidities and adverse outcome. As expected, postnatal complications were much more frequent in the preterm group and decreased with each advancing week of gestation, which is reflected in the following data: The incidence of *cardiovascular* conditions typically associated with prematurity, including arterial hypo- and hypertension, atrial septal defects (ASD), patent foramen ovale (PFO), and patent ductus arteriosus (PDA), was highest in the youngest preterm infants. The percentage of infants with *hematological* diagnoses (anemia of prematurity, polycythemia, thrombocytopenia, coagulation disorders) was 79% for extremely preterm and 4% for term infants. Furthermore, as expected, the prevalence of *infections* was much higher in the extremely preterm infant group compared to the term infant group (75% for extremely preterm; 8% for term). For *neurological* abnormalities, such as intraventricular and intracerebral hemorrhages (IVH and ICH), hypoxic-ischemic encephalopathy (HIE) and increased latencies of auditory brainstem responses (ABR), the highest percentage was 41% for extremely and very preterm infants, compared to 3% for term infants. *Respiratory* complications, such as respiratory distress syndrome, respiratory insufficiency, apnea, bronchopulmonary dysplasia (BPD) and pneumothorax were most prevalent in patients <32 weeks of GA (89-100%) compared to term infants (3%). Due to our focus on recruiting early preterm infants and infants with infections during stage 3 of our study, the high percentage of neonates at less than 32 weeks´ gestation who had received antibiotic treatment (95%) was predictable. As could be expected, the majority of extremely and very preterm infants required surfactant treatment (79%) in addition to ventilatory support (100% for extremely and 97% for very preterm), only 45% of late preterm infants and 8% of term infants needed ventilation.

Baseline characteristics of prenatal care and pregnancy

Table 3 provides an overview of prenatal care and potential influences on pregnancy. The data shown were obtained from clinical records and additional information was collected through the questionnaire (marked with * in the table). Prenatal care in Germany starts at 10-12 weeks´ gestation and consists of 12 regular check-up appointments, one every four weeks until week 32, and every two weeks thereafter. Among all mothers, more than 80% received their first check-up within the initial 10 weeks of pregnancy.

GA *(completed weeks)*	Preterm *Extreme* < 28 weeks	Preterm *Very* 28 to 31 weeks	Preterm *Late* 32 to 36 weeks	Term > 37 weeks	Total 23 to 41 weeks
Mothers *N*	*24*	*27*	*77*	*493*	*621*
*** Survey data: Parents** *N*	*6*	*9*	*51*	*461*	*527*
1st Prenatal Check-up					
≤ 10 weeks *N* (%)	18 (75)	20 (74)	63 (82)	407 (83)	508 (82)
11 - 20 weeks *N* (%)	4 (17)	3 (11)	5 (7)	70 (14)	82 (13)
Unknown *N* (%)	2 (8)	4 (15)	9 (12)	16 (3)	31 (5)
Number of prenatal visits					
0 - 5 *N* (%)	6 (25)	1 (4)	0 (0)	3 (0.6)	10 (2)
6 - 10 *N* (%)	11 (46)	14 (52)	25 (33)	56 (11)	106 (17)
> 10 *N* (%)	3 (13)	8 (30)	44 (57)	410 (83)	465 (75)
Prenatal BMI *M* (SD)	24.7 (6)	23.6 (3.4)	23.0 (4.6)	22.8 (3.6)	22.9 (3.9)
Adipositas Score					
normal *N* (%)	14 (58)	18 (67)	49 (64)	358 (73)	439 (71)
pre-adipose *N* (%)	4 (17)	5 (19)	10 (13)	80 (16)	99 (16)
adipose *N* (%)	4 (17)	2 (7)	6 (8)	18 (4)	30 (5)
underweight *N* (%)	0 (0)	1 (4)	5 (7)	20 (4)	26 (4)
missing *N* (%)	2 (8)	1 (4)	7 (9)	17 (3)	27 (4)
GWG (Kg) *M* (SD)	6.3 (3.4)	10.4 (3.4)	11.5 (4.9)	14 (5.1)	13.3 (5.3)
Prenatal Diabetes Screening					
Negative *N* (%)	7 (29)	15 (55)	60 (78)	386 (78)	468 (75)
Positive *N* (%)	0 (0)	1 (4)	5 (6)	32 (7)	38 (6)
Unknown/not yet done *N* (%)	17 (71)	11 (41)	12 (16)	75 (15)	115 (19)
Genetic Screening	1 (4)	7 (26)	8 (10)	68 (14)	84 (14)
Risk pregnancy *N* (%)	18 (75)	21 (78)	59 (77)	331 (67)	429 (69)
No *N* (%)	3 (13)	3 (11)	14 (18)	142 (29)	162 (26)
Unknown *N* (%)	3 (13)	3 (11)	4 (5)	20 (4)	30 (5)
Assisted Reprod. Med. *N* (%)	5 (21)	6 (22)	12 (16)	40 (8)	63 (10)
Multiplicity *N* (%)	5 (21)	8 (30)	20 (26)	9 (2)	42 (7)
Multiplicity / Ass. Repr. Med. *N* (%)	2 (40)	3 (50)	8 (67)	4 (10)	17 (27)
Influences on pregnancy					
Bleeding during pregnancy *N* (%)	3 (13)	3 (11)	4 (5)	10 (2)	20 (3)
Diabetes Mellitus (incl. GDM) *N* (%)	0 (0)	3 (11)	7 (9)	43 (9)	53 (9)
Hypertension *N* (%)	4 (17)	8 (30)	12 (16)	14 (3)	38 (6)
Infection during pregnancy *N* (%)	10 (42)	6 (22)	13 (17)	34 (7)	63 (10)
Infection as cause for delivery N (%)	7 (30)	6 (22)	6 (8)	5 (1)	24 (4)
Isthmocervical Insuff. *N* (%)	5 (21)	5 (19)	3 (4)	6 (1)	19 (3)
Placenta dysfunction *N* (%)	3 (13)	2 (7)	5 (7)	16 (3)	26 (4)
Substance abuse during pregnancy					
***Alcohol** *N* (%)	1 (17)	0 (0)	1 (2)	4 (1)	6 (1)
No *N* (%)	5 (83)	9 (100)	49 (98)	457 (99)	520 (99)
***Drug abuse** *N* (%)	0 (0)	0 (0)	0 (0)	3 (1)	3 (1)
No *N* (%)	6 (100)	9 (100)	50 (100)	458 (99)	523 (99)
Smoking	0 (0)	2 (7)	2 (3)	16 (3)	20 (3)
No smoking	18 (75)	21 (78)	67 (87)	466 (95)	572 (92)
Unknown	6 (25)	4 (15)	8 (10)	11 (2)	29 (5)
***Travel during pregnancy**					
Women who travelled *N* (%)	1 (17)	3 (33)	12 (24)	161 (34)	177 (33)
Women who didn`t travel *N* (%)	5 (83)	6 (67)	39 (77)	310 (66)	360 (67)
***Trips during pregnancy** *N*	1	3	12	188	204
Europe *N* (%)	0 (0)	1 (33)	10 (83)	107 (57)	118 (58)
Africa *N* (%)	1 (100)	1 (33)	1 (8)	15 (8)	18 (9)
Asia *N* (%)	0 (0)	0 (0)	0 (0)	28 (15)	28 (14)
North America N (%)	0 (0)	0 (0)	1 (8)	22 (12)	23(11)
South America	0 (0)	1 (33)	0 (0)	3 (2)	4 (2)
Other	0 (0)	0 (0)	0 (0)	13 (6)	13 (6)

Table 3: Baseline characteristics of prenatal care and pregnancy

During routine prenatal screenings, gestational diabetes mellitus was detected in 6%. As the test for GDM is routinely performed between 24-28 weeks of gestation, the majority of mothers of extremely preterm infants (71%) delivered their child before the test was performed. Only a small percentage of mothers (14%) underwent testing for chromosomal abnormalities. Due to specific risk factors, such as advanced maternal age (47% ≥ 35 years), nicotine abuse or individual maternal health problems, nearly 70% of pregnancies were defined as "risk pregnancies". Within this group, 76% were mothers of preterm and 67% of term infants. The majority of mothers (>95%) was tested for "TORCH"-infections (toxoplasmosis, others, rubella, cytomegalovirus, herpes), that could be passed on to their fetuses during pregnancy (data not shown). Another risk factor for pregnancy complications is overweight. Obesity and excessive gestational weight gain (GWG) are associated with increased risk for gestational diabetes (GDM) and hypertension (GHT), preeclampsia, delivery of LGA infants and a higher incidence of congenital defects[12]. Guidelines for pregnant women are recommending a BMI of 18.5-24.9 and GWG of 11.5-16 kg[13]. The women´s prenatal BMI shown in *Table 2* was recorded during their first prenatal check-up (usually between 10-12 weeks of the pregnancy). In comparison with the pre-pregnancy BMI (shown in *Table 5*), there were only minor changes. The GWG calculated for the entire length of the pregnancy was for the term group 14 kg, which is within the limits for women with a healthy BMI. Multiplicity is another strong risk factor for preterm birth and postnatal complications[14]. Among all mothers of this study, there was a clear group difference for multiplicity: For term infants, 2% (9 out of 493) of mothers were pregnant with multiples, compared to 26% (33 out of 128) within the preterm group, and 13 of these 33 mothers (40%) had conceived via assisted reproductive technology. This is in particular interesting, as a growing body of evidence describes an increased risk of cerebral palsy in children conceived by assisted reproduction which is strongly associated with the high proportion of multiplicity and preterm delivery in these pregnancies [15 16]. Furthermore, we screened for well-researched associations of GA and determinants for preterm birth. While infections play a key role in the pathogenesis of prematurity, it is necessary to distinguish between intrauterine infections, such as the Amniotic infection syndrome (AIS), mostly resulting in preterm delivery, and other types of maternal infections, including Influenza, Lyme disease or Herpes virus infection. As expected, there was a higher prevalence for maternal infections in the preterm group, where 21 out of 29 affected women (72%) had AIS (not shown). For 19 (65%) of these mothers, this led to induced preterm delivery. For the term group, a total of 7% of pregnancies were either affected by common infections (e.g., Influenza) or infections manifested as "fever sub partu" and only rarely resulted in induction of delivery. Other factors that determined preterm birth were isthmocervical insufficiency (20% of mothers of preterm infants ≤ 32 weeks GA; 1% of mothers of

term infants), GHT (21% of mothers of preterm infants; 3% of mothers of term infants) and placental dysfunction (13% of mothers of extremely preterm infants; 3% of mothers of term infants). There was no correlation between GA and self-reported drug or alcohol abuse.

GA *(completed weeks)*	Preterm *Extreme* < 28 weeks	Preterm *Very* 28 to 31 weeks	Preterm *Late* 32 to 36 weeks	Term > 37 weeks	Total 23 to 41 weeks
Mothers *N*	*24*	*27*	*77*	*493*	*621*
Mode of delivery (see Table 2)					
Spontaneous *N* (%)	5 (21)	6 (22)	26 (34)	240 (49)	277 (45)
C-Section (prim/sec/emgy) *N* (%)	19 (80)	21 (78)	40 (52)	118 (24)	198 (32)
Induced vaginal birth *N* (%)	0 (0)	0 (0)	5 (7)	60 (12)	65 (10)
Vacuum extraction, forceps *N* (%)	0 (0)	0 (0)	6 (8)	75 (15)	81 (13)
Anesthesia during delivery					
No anesthesia *N* (%)	4 (17)	6 (22)	22 (29)	162 (33)	194 (31)
Epidural *N* (%)	0 (0)	0 (0)	18 (23)	227 (46)	245 (40)
Spinal block *N* (%)	8 (33)	12 (44)	22 (29)	74 (15)	116 (19)
General anesthesia *N* (%)	4 (17)	3 (11)	3 (4)	10 (2)	20 (3)
Missing *N* (%)	8 (33)	6 (22)	12 (16)	20 (4)	46 (7)
Duration of delivery					
< 2 h *N* (%)	14 (58)	14 (52)	29 (38)	69 (14)	126 (20)
2 - 5 h *N* (%)	4 (17)	5 (19)	23 (30)	117 (24)	149 (24)
> 5 h *N* (%)	4 (17)	3 (11)	19 (25)	282 (57)	308 (50)
Missing *N* (%)	2 (8)	5 (19)	6 (8)	25 (5)	38 (6)
Pregnancies (prev. + current)					
Primigravida *N* (%)	11 (46)	16 (59)	42 (55)	244 (50)	313 (50)
Multigravida (2-3) *N* (%)	9 (38)	10 (37)	28 (36)	210 (43)	257 (41)
Multigravida (4-8) *N* (%)	4 (17)	1 (4)	7 (9)	39 (8)	51 (8)
Deliveries (prev. + current)					
Primiparous *N* (%)	13 (54)	20 (74)	54 (70)	293 (60)	380 (61)
Multiparous (2-3) *N* (%)	10 (42)	7 (26)	21 (27)	191 (39)	229 (37)
Multiparous (4-7) *N* (%)	1 (4)	0 (0)	2 (3)	9 (2)	12 (2)
Prev. Preterm Deliveries *N* (%)	5 (21)	7 (26)	10 (13)	14 (3)	36 (6)
No *N* (%)	19 (79)	20 (74)	67 (87)	479 (97)	585 (94)
Prev. Term Deliveries *N* (%)	9 (38)	5 (19)	18 (23)	191 (39)	233 (38)
No *N* (%)	15 (62)	22 (81)	59 (77)	302 (61)	398 (64)
Prev. Miscarriages *N* (%)	5 (21)	8 (30)	18 (23)	95 (19)	126 (20)
No *N* (%)	19 (79)	19 (70)	59 (77)	398 (81)	495 (80)
Prev. Stillborn Deliveries *N* (%)	0 (0)	1 (4)	2 (3)	7 (1)	10 (2)
No *N* (%)	24 (100)	26 (96)	75 (98)	485 (98)	608 (98)
Prev. Abortions *N* (%)	1 (4)	0 (0)	5 (6)	16 (3)	22 (4)
No *N* (%)	23 (96)	27 (100)	72 (94)	474 (96)	596 (96)

Table 4: Baseline characteristics of deliveries

Baseline characteristics of deliveries

An overview of delivery characteristics is provided in *Table 4*. In total, 70% of preterm infants versus 24% of term infants were delivered via Cesarian (C-) section. The largest number of term neonates was born spontaneously (49%) and only a small proportion was delivered by vacuum extraction or forceps. For all infant groups (extremely preterm, very preterm, late preterm and term), over 60% of mothers required some form of anesthesia, mainly delivered via epidural (40%) or spinal (19%) administration.

GA *(completed weeks)*	Preterm *Extreme* < 28 weeks	Preterm *Very* 28 to 31 weeks	Preterm *Late* 32 to 36 weeks	Term > 37 weeks	Total 23 to 41 weeks
Mothers *N*	*24*	*27*	*77*	*493*	*621*
*** Survey data: Parents** *N*	*6*	*9*	*51*	*461*	*527*
Age: Mother (years) *M* (SD)	35 (6.1)	34.9 (5.5)	34.4 (5.3)	34.0 (4.6)	34.2 (4.8)
< 35 years *N* (%)	11 (46)	13 (48)	33 (43)	269 (55)	326 (53)
≥ 35 years *N* (%)	13 (54)	14 (52)	44 (57)	224 (45)	295 (47)
***Father** (years) *M* (SD)	43 (6.1)	37.3 (4.8)	36.6 (5.8)	36 (5.8)	36.2 (5.9)
< 35 years *N* (%)	1 (17)	3 (33)	18 (35)	187 (41)	209 (40)
≥ 35 years *N* (%)	5 (83)	6 (67)	33 (65)	274 (59)	318 (60)
***BMI: Mother** (prior pregnancy) *M* (SD)	24.5 (4.4)	24.2 (4.2)	22.8 (4.6)	22.4 (3.3)	22.5 (3.5)
***Father** *M* (SD)	24.1 (1.2)	25.7 (2.8)	25.6 (3.2)	25.1 (3.2)	25.2 (3.1)
***Adipositas Score: Mother**					
Normal *N* (%)	2 (33)	6 (67)	32 (63)	350 (76)	390 (74)
Pre-adipose *N* (%)	1 (17)	1 (11)	7 (14)	68 (15)	77 (15)
Adipose *N* (%)	0 (0)	1 (11)	5 (10)	14 (3)	20 (4)
Underweight *N* (%)	0 (0)	0 (0)	3 (6)	26 (6)	29 (5)
Missing *N* (%)	3 (50)	1 (11)	4 (8)	3 (1)	11 (2)
***Father**					
Normal *N* (%)	5 (83)	5 (56)	25 (49)	246 (53)	281 (53)
Pre-adipose *N* (%)	1 (17)	4 (44)	20 (39)	171 (37)	196 (37)
Adipose *N* (%)	0 (0)	0 (0)	5 (10)	32 (7)	37 (7)
Underweight *N* (%)	0 (0)	0 (0)	0 (0)	1 (0.2)	1 (0.2)
Missing *N* (%)	0 (0)	0 (0)	1 (2)	10 (3)	10 (3)
***Ethnic Background: Mother**					
Western *N* (%)	3 (50)	7 (70)	45 (92)	433 (94)	488 (93)
Asian *N* (%)	0 (0)	1 (10)	0 (0)	7 (2)	8 (1)
African and middle East *N* (%)	3 (50)	0 (0)	2 (4)	15 (3)	20 (4)
Latin-American *N* (%)	0 (0)	2 (20)	1 (2)	2 (0.4)	5 (0.9)
Indian *N* (%)	0 (0)	0 (0)	1 (2)	4 (0.8)	5 (0.9)
***Father**					
Western *N* (%)	3 (50)	8 (89)	45 (88)	433 (94)	489 (93)
Asian *N* (%)	0 (0)	0 (0)	1 (2)	2 (0.4)	3 (0.6)
African and middle East N (%)	3 (50)	0 (0)	1 (2)	15 (3)	19 (4)
Latin-American *N* (%)	0 (0)	1 (11)	3 (5)	6 (1)	10 (2)
Indian *N* (%)	0 (0)	0 (0)	1 (2)	5 (1)	6 (1)
***Education: Mother**					
Cert. < 10 years school *N* (%)	0 (0)	1 (11)	1 (2)	21 (5)	23 (4)
Cert. ≥ 10 years school *N* (%)	2 (33)	4 (44)	11 (22)	73 (15)	90 (17)
Univ. degree *N* (%)	3 (50)	4 (44)	37 (72)	357 (78)	401 (76)
No / other certificate *N* (%)	1 (17)	0 (0)	2 (4)	9 (2)	12 (2)
***Father**					
Cert. < 10 years school *N* (%)	0 (0)	1 (11)	5 (9)	20 (4)	26 (5)
Cert. ≥ 10 years school N (%)	0 (0)	0 (0)	10 (19)	76 (16)	86 (16)
Univ. degree *N* (%)	4 (80)	8 (89)	34 (64)	344 (75)	390 (74)
No / other certificate *N* (%)	1 (20)	0 (0)	2 (4)	14 (3)	17 (3)
Missing N (%)	1 (20)	0 (0)	0 (0)	7 (2)	8 (2)
***Living Environment**					
Urban *N* (%)	3 (50)	7 (78)	38 (74)	412 (89)	460 (86)
Rural *N* (%)	1 (17)	2 (22)	8 (16)	30 (7)	41 (8)
Mixed *N* (%)	0 (0)	0 (0)	4 (8)	14 (3)	18 (4)
Missing N (%)	2 (33)	0 (0)	1 (2)	5 (1)	8 (2)
***Survey**					
distributed *N*	11	13	66	492	582
completed *N*	6	10	62	472	550
response rate (%)	55	77	94	96	95

Table 5: Baseline characteristics of parental study participants

Previous pregnancies and deliveries

50% of women in the entire parental cohort were primigravida. Mothers of preterm infants had

a higher percentage (13- 21%) of previous preterm deliveries compared to the term group (3%). The percentage of women who had experienced a miscarriage, which refers to pregnancy loss at less than 20 weeks´ gestation, was only slightly higher for the preterm (25%) compared to the term group (19%). The number of mothers who had lost more than one pregnancy was twice as high (13%) for the group of preterm infants < 32 weeks GA compared to term (6%, data not shown). Only an exceedingly small percentage of all mothers had induced abortion (0 - 6%) or stillbirth (0 - 4%).

Baseline characteristics of parental study participants

Anthropometric and demographic characteristics of participating parents are listed in *Table 5*. The mean age of women from the entire cohort is 34.2 ± 4.8 years. Among all mothers, the percentage of women older than 35 years at delivery was higher in all three preterm groups (54% for extremely preterm, 52% for very preterm, 57% for late preterm) compared to the term group with 45%. Accordingly, the mean age of women who gave birth to extremely preterm infants was higher (35 ± 6.1 years) in comparison to those who delivered term infants (34.1 ± 4.5 years). Due to language barriers, obtaining accurate and extensive self-reported data from non-German and non-English speaking parents was difficult. Consequently, the MUNICH PreTCI cohort displays limited ethnic parental diversity, with over 90% of the participating parents reporting a Western European ethnicity. Data on the entire parental cohort portrait a greater proportion of participants that are higher educated than the general population (>75% with University degree) living in urban environments (75-90%).

Overview of family medical data

The family medical history data on parents and their siblings, siblings of enrolled infants, and grandparents are shown in *Table 6* and include information on allergies, asthma, cardiovascular conditions, coagulation disorders, diabetes mellitus, as well as neurological and thyroid disorders. A correlation between neonatal or early life infections and allergy or increased risk for asthma has not yet been established in the literature. Previous research suggested that an increased infectious burden in the first 24 month is associated with a decreased prevalence of IgE-mediated allergy during childhood[17].

GA *(completed weeks)*	Preterm *Extreme* < 28 weeks	Preterm *Very* 28 to 31 weeks	Preterm *Late* 32 to 36 weeks	Term > 37 weeks	Total 23 to 41 weeks
Mothers *N*	*24*	*27*	*77*	*493*	*621*
*** Survey data: Parents** *N*	*6*	*9*	*51*	*461*	*527*
Allergies					
Mothers N (%)	4 (17)	10 (37)	29 (38)	222 (45)	265 (43)
No ALL N (%)	4 (17)	1 (4)	24 (31)	248 (50)	277 (45)
Unknown *N* (%)	16 (67)	16 (60)	24 (31)	22 (5)	78 (13)
*Mat. families N (%)	0 (0)	2 (22)	13 (25)	145 (31)	160 (30)
***Fathers** *N (%)*	2 (33)	6 (67)	19 (37)	176 (38)	203 (39)
*Pat. families N (%)	0 (0)	1 (11)	14 (27)	120 (26)	135 (26)
***Asthma**					
***Mothers** *N* (%)	0 (0)	2 (22)	2 (4)	34 (7)	39 (7)
*Mat. families *N* (%)	0 (0)	2 (22)	9 (18)	53 (11)	64 (12)
***Fathers** *N* (%)	0 (0)	0 (0)	5 (10)	32 (7)	36 (7)
*Pat. families *N* (%)	0 (0)	0 (0)	4 (8)	46 (10)	50 (10)
***Cardiovascular diseases**					
***Mothers** N (%)	1 (17)	0 (0)	3 (6)	16 (4)	20 (4)
*Mat. families *N* (%)	1 (17)	0 (0)	12 (24)	85 (18)	98 (19)
***Fathers** *N* (%)	2 (33)	1 (11)	4 (8)	12 (3)	19 (4)
*Pat. families *N* (%)	0 (0)	1 (11)	1 (2)	18 (4)	20 (4)
***Coagulation disorders**					
***Mothers** *N* (%)	0 (0)	1 (11)	5 (10)	28 (6)	34 (6)
***Fathers** *N* (%)	0 (0)	0 (0)	0 (0)	4 (1)	4 (1)
Diabetes Mellitus (excl. GDM)					
Mothers *N* (%)	0 (0)	2 (7)	2 (3)	9 (2)	13 (2)
*Mat. families *N* (%)	1 (17)	2 (22)	8 (16)	108 (43)	119 (23)
***Fathers** *N* (%)	0 (0)	0 (0)	0 (0)	4 (1)	4 (1)
*Pat. families *N* (%)	0 (0)	1 (11)	12 (24)	80 (17)	93 (18)
***Neurological disorders**					
***Mothers** *N* (%)	0 (0)	0 (0)	2 (4)	12 (3)	14 (3)
*Mat. families *N* (%)	0 (0)	0 (0)	1 (2)	19 (4)	20 (4)
***Fathers** *N* (%)	0 (0)	0 (0)	0 (0)	3 (1)	3 (1)
*Pat. families *N* (%)	0 (0)	1 (11)	1 (2)	19 (4)	21 (4)
Thyroid disorders					
Mothers *N* (%)	4 (17)	4 (15)	18 (23)	128 (26)	154 (25)
Hypo / Hyper	3/1	4/0	15/0	113/6	135/7
Missing *N*	0 (0)	0 (0)	3 (3)	9 (2)	12 (2)
***Fathers** *N* (%)	1 (17)	0 (0)	1 (2)	16 (3)	18 (3)
Hypo / Hyper	1/0	0/0	0/1	3/13	4/14

Table 6: Family Medical History

In the MUNICH-PreTCI cohort, the percentage of mothers reporting allergies is 43% and for fathers 39%. Only about half of these fathers´ and mothers´ parental generation experienced allergic symptoms. The prevalence of asthma in parents and grandparents was low with 6-10%. Additionally, only a small percentage of mothers and fathers reported cardiovascular diseases, coagulation disorders, or neurological abnormalities. There was only a small difference in the prevalence of DM (type 1 and type 2, but not GDM) between women and men. The percentage of thyroid disorders was significantly higher in women than in men (23% in women and 3% in men),

with hypothyroidism affecting predominantly women (19% mothers, 0.6% fathers), probably also due to thyroid screening in pregnancy.

Strength and Limitations of the Study

The MUNICH-PreTCI study is a prospective birth cohort including preterm and term neonates born over a period of 27 months at the Perinatal Center Campus Innenstadt, University Hospital, LMU, Munich. The study contains phenotypical information including clinical data from maternal and neonatal medical records, demographics survey and large medical datasets for all families and their neonates, as well as blood samples at defined time points for proteomic screening. The main strength of this study is the combination of a population-based cohort with state-of-the-art proteomic screening. This enables us to relate the status of the functional protein network to certain pre- and postnatal factors as well as specific clinical outcomes which were recorded at time of birth and during follow-up.

Our cohort has some limitations. Recruiting at the university hospital and restricting the survey to German and English-speaking parents likely introduced some bias, as mostly caucasian families from urban living environment and with a higher educational level participated in the study. It is not possible to characterize the confounding effect of language barriers to study participation and/or answer accuracy compared to a situation under which the questionnaire would have also been distributed in additional languages, thus being more representative of the "typical" community-based population.

The distribution of the questionnaire started with stage 2 of the recruitment process. Consequently, families enrolled in stage 1 (data collection and proteomic analysis method establishment) did not have the opportunity to participate in the survey. Confounding by indication provides another challenge in data analysis. For stage 3, we enrolled infants with the objective of augmenting specific groups of interest, such as extremely and very preterm infants, neonates with infections, with diabetic mothers or being "small for gestational age". Furthermore, in order to obtain as many data points as possible, we did not exclude families who did not want to participate in the survey.

The interdisciplinary collaboration of experts from various disciplines, such as clinicians, proteomic experts and epidemiologists will allow a systematic translational approach to find evidence for novel targets that can be applied in clinical practice to improve identification of neonates at risk and advance patient care for a better outcome of preterm infants.

Conclusion

In conclusion, MUNICH-PreTCI offers a comprehensive assessment of a birth cohort combined with a collection of reusable dried blood samples obtained at birth and at defined time-points throughout hospitalization, providing the opportunity for further phenotyping by using OMICS technologies. These technologies bear great promise to generate extensive and detailed datasets even from very small blood samples and will be an excellent foundation for future systems medicine approaches intended to advance the understanding of complex multi-factorial diseases in neonatal and pediatric health.

Patient and public involvement

We regret, that we were not aware of patient involvement when we designed and conducted this study. Primarily, our plans did include sharing the study´s results with the nurses of the NICU and the newborn ward because they mentioned great interest in our findings. In addition, we will now also contact participating families to disseminate our study results and provide an opportunity to meet in-person to discuss specific questions.

For our future research, we will definitely implement active patient contribution.

References

1. Frey HA, Klebanoff MA. The epidemiology, etiology, and costs of preterm birth. *Semin Fetal Neonatal Med* 2016;21(2):68-73. doi: 10.1016/j.siny.2015.12.011 [published Online First: 2016/01/23]

2. Green ES, Arck PC. Pathogenesis of preterm birth: bidirectional inflammation in mother and fetus. *Seminars in immunopathology* 2020;42(4):413-29. doi: 10.1007/s00281-020-00807-y

3. Wilcox AJ, Weinberg CR, Basso O. On the pitfalls of adjusting for gestational age at birth. *American journal of epidemiology* 2011;174(9):1062-8. doi: 10.1093/aje/kwr230 [published Online First: 2011/09/29]

4. Arbuckle TE, Fraser WD, Fisher M, et al. Cohort profile: the maternal-infant research on environmental chemicals research platform. *Paediatric and perinatal epidemiology* 2013;27(4):415-25. doi: 10.1111/ppe.12061 [published Online First: 2013/06/19]

5. Priliani L, Oktavianthi S, Prado EL, et al. Maternal biomarker patterns for metabolism and inflammation in pregnancy are influenced by multiple micronutrient supplementation and associated with child biomarker patterns and nutritional status at 9-12 years of age. *PLOS ONE* 2020;15(8):e0216848. doi: 10.1371/journal.pone.0216848

6. Gaillard R, Welten M, Oddy WH, et al. Associations of maternal prepregnancy body mass index and gestational weight gain with cardio-metabolic risk factors in adolescent offspring: a prospective cohort study. *BJOG : an international journal of obstetrics and gynaecology* 2016;123(2):207-16. doi: 10.1111/1471-0528.13700 [published Online First: 2015/11/04]

7. Bouyssi-Kobar M, du Plessis AJ, McCarter R, et al. Third Trimester Brain Growth in Preterm Infants Compared With In Utero Healthy Fetuses. *Pediatrics* 2016;138(5):e20161640. doi: 10.1542/peds.2016-1640

8. Geyer PE, Holdt LM, Teupser D, et al. Revisiting biomarker discovery by plasma proteomics. *Molecular systems biology* 2017;13(9):942. doi: 10.15252/msb.20156297 [published Online First: 2017/09/28]

9. Blencowe H, Cousens S, Chou D, et al. Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health* 2013;10 Suppl 1(Suppl 1):S2-S2. doi: 10.1186/1742-4755-10-S1-S2 [published Online First: 2013/11/15]

10. Kramer MS. Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization* 1987;65(5):663-737. [published Online First: 1987/01/01]

11. Mendez-Figueroa H, Truong VTT, Pedroza C, et al. Large for Gestational Age Infants and Adverse Outcomes among Uncomplicated Pregnancies at Term. *American journal of perinatology* 2017;34(7):655-62. doi: 10.1055/s-0036-1597325 [published Online First: 2016/12/08]

12. Poston L, Harthoorn LF, van der Beek EM, et al. Obesity in Pregnancy: Implications for the Mother and Lifelong Health of the Child. A Consensus Statement. *Pediatric Research* 2011;69(2):175-80. doi: 10.1203/PDR.0b013e3182055ede

13. Rasmussen KM, Catalano PM, Yaktine AL. New guidelines for weight gain during pregnancy: what obstetrician/gynecologists should know. *Current opinion in obstetrics & gynecology* 2009;21(6):521-6. doi: 10.1097/GCO.0b013e328332d24e [published Online First: 2009/10/08]

14. Dodd JM, Grivell RM, CM OB, et al. Prenatal administration of progestogens for preventing spontaneous preterm birth in women with a multiple pregnancy. *The Cochrane database of systematic reviews* 2019;2019(11) doi: 10.1002/14651858.CD012024.pub3 [published Online First: 2019/11/21]

15. Hvidtjørn D, Grove J, Schendel D, et al. Multiplicity and early gestational age contribute to an increased risk of cerebral palsy from assisted conception: a population-based cohort study. *Human reproduction (Oxford, England)* 2010;25(8):2115-23. doi: 10.1093/humrep/deq070 [published Online First: 2010/06/18]

16. Chang HY, Hwu WL, Chen CH, et al. Children Conceived by Assisted Reproductive Technology Prone to Low Birth Weight, Preterm Birth, and Birth Defects: A Cohort Review of More Than 50,000 Live Births During 2011-2017 in Taiwan. *Frontiers in pediatrics* 2020;8:87. doi: 10.3389/fped.2020.00087 [published Online First: 2020/04/02]

17. Nilsson C, Linde A, Montgomery SM, et al. Does early EBV infection protect against IgE sensitization? *Journal of Allergy and Clinical Immunology* 2005;116(2):438-44. doi: https://doi.org/10.1016/j.jaci.2005.04.027

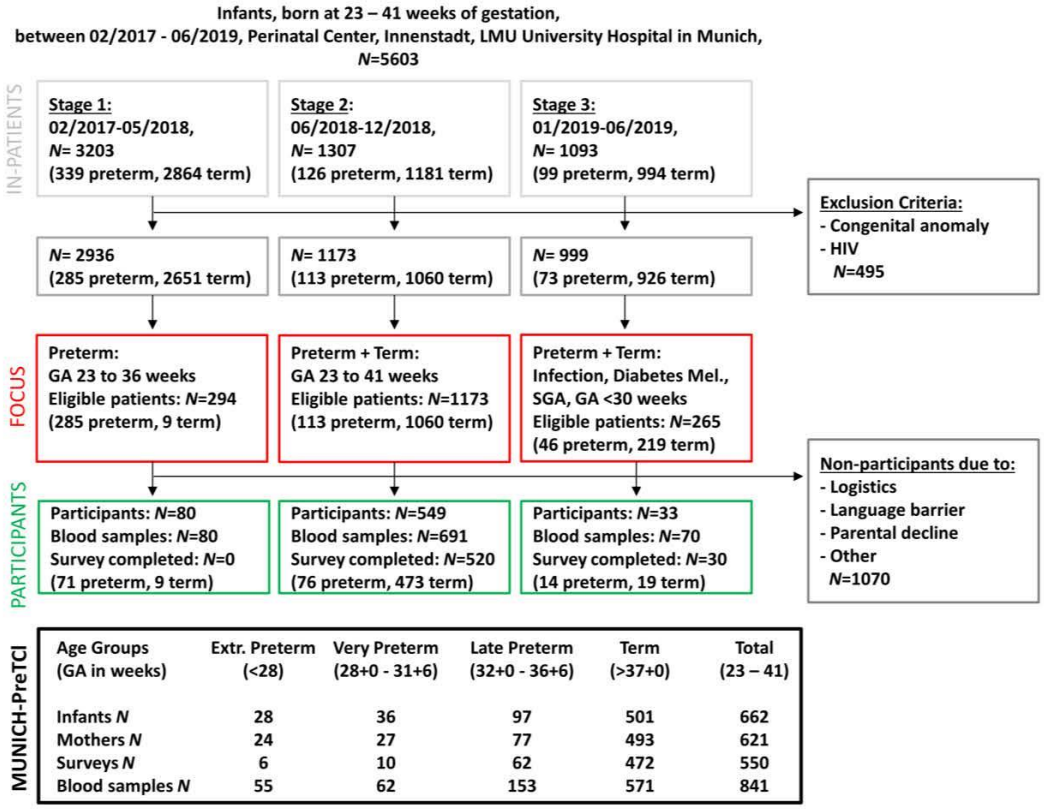


Figure 1: Study flow chart

3.8. Article 8: Cotranslational N-degron masking by acetylation promotes proteome stability in plant

Authors: Eric Linster[1], Francy L. Forero Ruiz[1], Pavlina Miklankova[1], Thomas Ruppert[2], Johannes Mueller[3], Laura Armbruster[1], Giovanna Serino[4], Matthias Mann[3], Rüdiger Hell[1], Markus Wirtz[1,*]

[1]Centre for Organismal Studies Heidelberg, Heidelberg University, Im Neuenheimer Feld 360, Heidelberg, 69120, Germany
[2]Center for Molecular Biology Heidelberg, Heidelberg University, Im Neuenheimer Feld 282, Heidelberg, 69120, Germany
[3]Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, Martinsried, 82152, Germany
[4]Department of Biology and Biotechnology, Sapienza Università di Roma, Rome, 00185, Italy

In the previous mentioned studies, the focus of proteome analysis was on the presence and abundance of proteins and the resulting consequences for the analyzed biological system or specimen. Here, in contrast, we present a study of proteome turnover changes and specifically the control of proteome turnover by N-degron masking in cotranslational manner in the plant *Arabidopsis*. In brief, NatA, a ribosomal associated protein complex recognizes a N-terminal amino acid encoded sequence and stabilizes proteins by N-terminal acetylation. I performed whole proteome measurements of NatA depleted and wilt-type Arabidopsis leaf to identify regulated proteins. The majority of down regulated proteins upon NatA depletion were identified as targets of the NatA complex which points towards the decreased stability of those proteins in the absence of NatA. Together with methods for the identification of proteasome activity and ubiquitination levels, this could be linked to a shift in steady state of NatA substrates, with an increased degradation and translation rate.

Title: Cotranslational N-degron masking by acetylation promotes proteome stability in plants

Authors:

Eric Linster[1], Francy L. Forero Ruiz[1], Pavlina Miklankova[1], Thomas Ruppert[2], Johannes Mueller[3], Laura Armbruster[1], Giovanna Serino[4], Matthias Mann[3], Rüdiger Hell[1], Markus Wirtz[1,*]

Affiliation

1, Centre for Organismal Studies Heidelberg, Heidelberg University, Im Neuenheimer Feld 360, Heidelberg, 69120, Germany

2 Center for Molecular Biology Heidelberg, Heidelberg University, Im Neuenheimer Feld 282, Heidelberg, 69120, Germany

3 Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, Martinsried, 82152, Germany

4 Department of Biology and Biotechnology, Sapienza Università di Roma, Rome, 00185, Italy

***Corresponding author**

Markus Wirtz

e-mail: markus.wirtz@cos.uni-heidelberg.de

Tel.: +49 6221 54 5334

Abstract/Synopsis (200 words)

Abstract (200)

N-terminal protein acetylation (NTA) is a prevalent and highly abundant protein modification that is widely conserved across eukaryotic kingdoms and essential for viability in animals and plants. The principle executor of NTA is the N^{α}-acetyltransferase A (NatA) complex, which is tethered to the ribosome and accounts for cotranslational acetylation of 40% of the proteome. Despite its prevalence, the impact of NTA on protein fate is still enigmatic. Here, we found that depletion of NatA activity led to a 4-fold increase in global protein turnover via the ubiquitin-proteasome system in Arabidopsis. Surprisingly, a concomitant increase in translation, actioned via enhanced Target-of-Rapamycin activity, was also observed, implying that defective NTA triggers feedback mechanisms to maintain steady-state protein abundance. Quantitative analysis of the proteome, the translatome, and the ubiquitome revealed that NatA substrates accounted for the bulk of this enhanced turnover. A targeted analysis of NatA substrate stability revealed that the absence of NTA triggers protein destabilization via a previously undescribed and widely conserved nonAc/N-degron. Hence, the imprinting of the proteome with acetylation marks is essential for coordinating proteome stability. Given the strong conservation of NTA machineries in eukaryotes, we propose that this may represent an evolutionary conserved system for controlling cellular proteostasis.

Introduction

As sessile organisms, plants have to fight environmental challenges on site. The proteome's dynamic plasticity is one of the most critical mechanisms allowing plants to acclimate to changes in their environment rapidly. This notion is supported by the substantially elaborated ubiquitin-proteasome system (UPS) in plants compared to humans [1]. Despite the essential importance of protein degradation, we are only now beginning to understand how plants control proteostasis upon stress and under favorable growth conditions. Protein modifications have been identified as crucial determinants of protein stability in eukaryotes and are highly regulated upon diverse plant stress conditions. One of the most pervasive protein modifications is N-terminal protein acetylation (NTA). NTA occurs on 80-90% of human and Arabidopsis soluble proteins and is executed by up to five ribosome-associated N-terminal acetyltransferases (Nat) complexes, of which NatA, NatB and NatC are conserved in all eukaryotes [2]. Disturbance of NTA in humans causes fatal diseases like Ogden syndrome, whilst enhanced NTA is associated with deregulated cell proliferation in specific cancer types [3, 4].

The NatA complex consists of the catalytically active subunit NAA10 and the ribosome-anchoring subunit NAA15, and targets nascent chains of proteins after the initiator methionine (iMet) is cleaved by methionine aminopeptidase (MetAP). In Arabidopsis and humans, 40 % of proteins are subjected to this N-terminal protein trimming. In plants, Nat complexes are particularly important for the resistance towards diverse abiotic and biotic environmental stresses [5, 6, 7, 8, 9]. The dynamic regulation of the NatA abundance by the phytohormone abscisic acid (ABA) is essential for drought stress resilience. However, the NatA-dependent mechanism for the regulation of drought stress responses remains to be determined. Only in a few cases has NTA has been reported to affect protein functionality [2, 10]. Thus, controlling the activity of individual proteins is unlikely to explain the pervasive NTA of bulk proteins [11].

In yeast and humans, NTA can create N-degrons recognized by the Ac/N-degron pathway and leading to the destruction of proteins by the UPS. On the contrary, another set of proteins was stabilized by NTA [12 13]. Taken together a direct impact of NTA on protein stability has been documented for less than 30 proteins in all eukaryotic model species. Thus the impact of NTA on global proteome stability remains unclear in eukaryotes [12, 13, 14, 15, 16 8, 17].

Here we show that impairment of NTA by NatA results in a global destabilization of the proteome in Arabidopsis and discover a novel degron that marks the majority of non-acetylated cytosolic proteins for degradation via the ubiquitin system.

Results

Since loss of NatA causes embryo-lethality in plants [9], we independently down-regulated both subunits of the NatA complex by an amiRNAi-approach and tested the global protein degradation rates in leaves after feeding of isotope-labeled amino acids. The depletion of the NatA complex substantially enhanced the protein degradation rate, causing up to 4-fold faster protein destruction (Fig. 1a), when the NAA10 abundance was decreased to 25% or 20% of wild type level in *amiNAA10* lines 18 or 24, respectively [9]. The combined activity of metallo-, serine-, acid-, and sulfhydryl-type proteases was not enhanced in any of the NatA-depleted plants (Extended Data Fig. 1). However, the NatA-depletion triggered a specific increase of the proteasome activity (Fig. 1b), which negatively correlated with the previously demonstrated decreased growth of the individual amiRNAi lines with depleted NatA levels [9]. The finding of increased proteasome activity in NatA depleted plants was independently confirmed by the immunological detection of the accumulation of the lid and the core subunit of the 26S proteasome RPN10 and PBA1, respectively, in *amiNAA10* (Extended Data Fig. 2a-d). The endogenous ubiquitination rate also increased in NatA depleted plants and resulted in most significant accumulation of poly-ubiquitinated proteins in the transgenic line with the most substantial depletion of NatA activity. In line with these observations, all NatA depleted plants accumulated higher amounts of ubiquitinated proteins than the wild type after pharmacological inhibition of the proteasome (Fig. 1c, Extended Data Fig. 2e). Furthermore, enhanced neddylation of Cullin 1 demonstrated that Cullin-RING E3 ligases (CRLs, [18]) contributed to the enhanced in vivo ubiquitination activity in NatA depleted plants (Fig. 1d).

Next, we aimed to identify the proteins that were destroyed by the UPS when NatA was depleted. Affinity enrichment of ubiquitinated proteins with the UbiQapture-Q matrix resulted in 1.6-fold more poly-ubiquitinated proteins captured in NatA depleted plants when compared to wild type as detected with a ubiquitin-specific antiserum (Fig. 1e) and a comparable increase of total protein after Ubi-Qapture enrichment from *amiNAA10* plants (1.5-fold increase, $p < 0.05$, Fig. 1f). Out of the 232 identified proteins that were significantly enriched by the Ubi-Qapture matrix in NatA depleted plants 162 (70%) were canonical NatA substrates (Extended Data Table 1 and 2, Fig. 1g), implying a significant enrichment of NatA substrates in the fraction of the poly-ubiquitinated proteins (Fisher´s exact test, p-value < 0.0001). A gene ontology enrichment analysis revealed that poly-ubiquitinated proteins in NatA depleted plants were predominantly cytosolic proteins (141 of 232) and associated with the responses to diverse stresses and protein-folding (Extended Data Table 3).

Despite the induction of the UPS for degradation of NatA substrates, the total protein content and the protein profile of soluble proteins were almost not affected in leaves of NatA depleted plants (Fig 2a, Extended Data Fig. 3a). A quantitative analysis of the steady-state protein levels by shotgun mass

spectrometry uncovered that the abundance of only 92 out of the 1.238 detected proteins was significantly affected in plants depleted for the catalytic subunit of NatA (Fig. 2b, Extended Data Table 4). However, 83 % of the proteins that displayed lowered steady-state levels in NatA depleted plants were canonical NatA substrates (Extended Data Table 5, permutation-based FDR ≤ 1%, Fisher´s exact test, p-value < 0.0001 for the enrichment of NatA substrates in the fraction of low abundant proteins in *amiNAA10*). We selected the decreased protein glutathione reductase 1 (termed ^{ARK}GR1 due to excision of the iMet, AT3G24170, -2.8-fold, $p < 0.05$) and the not-significantly accumulated protein *O*-acetylserine(thiol)lyase A (^{ASR}OAS-TL A, AT4G14880, 1.2-fold) for time-resolved destabilization assays since both cytosolic proteins are canonical NatA substrates. Prior to this analysis, the steady-state protein levels of ^{ASR}OAS-TLA and ^{ARK}GR1 in NatA depleted plants were independently confirmed by immunological detection with specific antisera (Extended Data Figure 3b, c). The cycloheximide chase assays for ^{ARK}GR1 and ^{ASR}OAS-TL A demonstrated significantly enhanced degradation of both NatA substrates in NatA depleted plants (Fig. 2c, Extended data Figure 3d). In contrast, cytosolic proteins that are not recognized as substrates and thus not acetylated by NatA, like the OAS-TL A interacting protein ^{MPP}SAT5 (Serine-Acetyl-Transferase 5, AT1G55920) and ^{MED}COI1 (COronatine-Insensitive protein 1, AT2G39940) were not destabilized in NatA depleted plants (Fig. 2d, Extended Data Fig. 3d). Since the steady-state level of the destabilized ^{ASR}OAS-TL A was unaffected by NatA depletion, we tested the accumulation of ^{ASR}OAS-TL A after proteasome inhibition by MG132. Short-term inhibition of the proteasome resulted in significantly faster accumulation of ^{ASR}OAS-TL A in NatA depleted plants when compared to wild type (Fig. 2e, Extended Data Fig. 4), suggesting that the unaffected steady-state levels of the destabilized ^{ASR}OAS-TL A was a result of enhanced ^{ASR}OAS-TL A translation in NatA depleted plants. Importantly, enhanced translation was not observed for ^{ARK}GR1, and proteins that are not recognized by NatA (^{MPP}SAT5 and ^{MRE}TUBB4, tubulin β4).

Since 40% of the proteome is acetylated by NatA, we hypothesized that translation must be significantly upregulated in NatA depleted plants to maintain the steady-state proteome level. It should be noted in this context that the costs for translation can reach up to 38% of total cellular ATP consumption in wild type Arabidopsis leaves [19], and that protein turnover is known to negatively correlate with the growth rate in the diverse Arabidopsis accessions [20]. Despite the substantial costs of translation in wild type plants, incorporation of isotope-labeled ^{35}S-Met and ^{35}S-Cys into proteins increased up to 4-fold in leaves of NatA depleted plants (Fig 3a). This higher global translation rate was due to the selective enhancement of translation for diverse proteins (Fig 3b). In plants, the sensor kinase Target of Rapamycin (TOR) is a critical regulator of the ribosome amount due to phosphorylation of the kinase S6K (Small-ribosome subunit 6 Kinase)[21] and the translation efficiency of stress-related genes due to phosphorylation of the translation initiation factor eIF3h [22]. NatA depletion triggered the increase of TOR activity (Extended data Fig. 4), resulting in up to 4-fold higher phosphorylation of S6K

at T^{449} and, consequently, a significant accumulation of rRNA (Fig 3c-d). We applied time-resolved biorthogonal non-canonical amino acid tagging (BONCAT) to identify the more efficiently translated proteins in NatA depleted plants after selective enrichment of the translatome [23]. The incorporation of the trackable Met-analogue azidohomoalanine was linear during the period of the analysis and independently confirmed the higher translation rate in NatA depleted plants (Fig 3e, f). Quantitative proteomics of the newly translated proteins (913 proteins detected, Extended Data Table 6) in the wild type and NatA depleted plants uncovered that 45% of identified proteins were more translated upon NatA depletion (Extended Data Table 7). The vast majority of these proteins were NatA substrates (72 %, Fisher´s exact test for enrichment of NatA Substrates, p-value < 0.0001**,** Fig 3g). Comparison of the ubiquitome and the translatome of NatA depleted plants revealed that 65 proteins (30% of the NatA depletion-induced ubiquitome alteration) were more ubiquitinated and more translated. 72% of the proteins with enhanced turnover in NatA depleted plants were canonical NatA substrates (Extended Data Table 8, Fisher´s exact test for enrichment of NatA Substrates, p-value < 0.0001), suggesting that selective destabilization of NatA substrates due to impaired N-terminal acetylation was counteracted by their enhanced translation to maintain their steady-state level in the mutant lines (Fig 2b). In support of this hypothesis, we found the translation of the OAS-TL A protein to be significantly enhanced (1.7-fold, p = 0.02), while GR1 translation was unaffected by NatA depletion (Extended Data Table 6), explaining the difference in the steady-state levels of the two destabilized NatA substrates (Fig. 2c). To provide direct evidence for the enhanced protein turnover by decreased NTA of NatA substrates, we assessed OAS-TL A turnover using the tandem-Fluorescent timer system, which allows for non-invasive quantification of protein half-life time in plants [24]. OAS-TL A's half-life time with native N-terminus (^{ASR}OAS-TL) was significantly lower in NatA depleted plants compared to wild type (Fig 4a). In contrast, the half-life time of proteins that are not targeted by NatA (e.g., tubulin β4 (^{MRE}TUBB4) or SAT5 (^{MPP}SAT5) was unaffected in NatA depleted plants (Fig 4b, c). A proline residue at position 3 is known to inhibit substrate recognition by the NatA complex in diverse metazoa[25] and plants[9]. To ultimately prove that the absence of NTA of alanine in position 2 (Ala2) is causing the destabilization of the NatA substrate ^{ASR}OAS-TL A, we genetically engineered an OAS-TL A protein mutant with impaired NTA by inserting a proline at position 3 (^{APS}OASTL A). The ^{APS}OASTL was significantly destabilized in the wild type and displayed a similar protein half-life time to that of the native ^{ASR}OAS-TL in NatA depleted plants. Remarkably, ^{APS}OASTL was not further destabilized in NatA depleted plants, providing ultimate evidence that the absent NTA of Ala2 is responsible for the destabilization of OAS-TL A in plants (Fig 4a). Based on these findings, we named this novel destabilizing signal nonAc-X^2/N-degron (X^2 = Ala). To judge the generality of nonAc-X^2/N-degron-induced destabilization, we randomly selected ten cytosolic NatA substrate candidates and tested the impact of NTA on their stability by applying the same strategy. The candidates' identity as proper NatA substrates and the inhibitory

impact of proline at position 3 for recognition by the NatA in the three tested NatA substrate groups was verified (X^2 = Ala: group 1, Gly; group 2 or Ser: group 3, Extended Data Fig. 6). Eight out of the ten NatA substrates were significantly destabilized by inhibiting NTA at position 2. This destabilization also occurred when Ser or Gly occupied position 2 (Fig 4d). Seven of these NatA substrates were found to be destabilized in NatA depleted plants, and one NatA substrate, NHO1, could not be detected in the cytosol of NatA depleted plants. This demonstrated that the absence of NTA was the causal effect for decreased half-life time of these proteins (Fig 4d, Extended Data Fig. 7). We could further show that the destabilizing effect of NTA inhibition at position 2 was not restricted to highly stable NatA substrates (Extended Data Fig. 8). In two cases, CYP19 and UGE1, NTA inhibition at position 2 had no impact on protein stability in the wild-type Arabidopsis plants. In agreement, CYP19 and UGE1 were also not destabilized in NatA depleted plants. These findings support the notion that nonAc-X^2/N-degron functionality requires additional parameters encoded in appropriately positioned domains downstream of the N-terminus, e.g., surface-exposed Lys-residues for ubiquitination [26] or accessibility of the N-terminus for recognition [16].

Discussion

Since NatA imprints 40% of the proteome in plants, we suggest that absent masking of the nonAc-X^2/N-degron in many NatA substrates substantially contributes to the observed higher protein turnover in NatA depleted plants. In agreement with this notion, the enhanced proteome turnover was predominantly based on the degradation of NatA substrates, and N-terminally acetylated NatA substrates were significantly overrepresented in the fraction of stable abundant proteins in plants [27]. Since the depletion of the ribosome-anchoring subunit NAA15 resulted in an increase of proteome turnover comparable to that caused by depletion of the catalytically active subunit NAA10, we conclude that cotranslational NTA is required to chemically block the N-terminus to ensure stabilization of the nascent polypeptide and prevent its unwanted turnover. A similar destabilizing effect was previously shown in humans for the non-acetylated ^{MQ}RGS2 protein that was degraded by the Arg/N-degron pathway. Surprisingly, acetylated ^{MQ}RGS2 was recognized by the Ac/N-degron pathway, demonstrating that NTA of the iMet can redirect proteins between different branches of the N-degron pathway system [28]. In contrast to other protein modifications, e.g., ubiquitination or Lys-ε-acetylation, NTA is irreversible [10], implying that the stability of many NatA substrates is intrinsically determined at the moment when these proteins are synthesized. However, co-translationally imprinted N-degrons can also contribute to conditional protein quality control when they are unshielded upon stress-induced protein misfolding or exposed in subunits of multi-protein complexes produced in non-stoichiometric amounts [16, 29]. As a result of its cotranslational mode and its

irreversibility, NTA has been suggested as static in the past [10]. This view on NTA is obsolete in plants for two reasons: first, NTA is rapidly regulated upon environmental stimuli by the phytohormone-system [9] and second, a significant fraction of NatA substrates is only partially N-terminally acetylated [5, 6, 7, 9], implying additional regulatory mechanisms controlling the activity of NatA on the nascent chains extruding from the ribosome exit tunnel. Crystallization of the trimeric metazoan NatA-HYPK complex uncovered a structural basis for NatA regulation by its binding partner HYPK in vitro [30, 31]. In two companion studies, we identify the HYPK orthologue in the monocotyledonous plant *Oryza sativa* (rice) and the dicotyledonous plant Arabidopsis. In both plant species, loss-of-HYPK decreases in vivo NatA activity and enhances global protein turnover [32, 33], demonstrating that global control of protein turnover by NatA is evolutionary conserved in the plant lineage of eukaryotes for more than 150 Mya [34]. The recent identification of E3 ubiquitin ligases (N-recognins), specifically recognizing non-acetylated NatA substrates, provides evidence for the existence of this novel N-degron pathway in humans [14, 35, 36]. If the human nonAc-X^2/N-degron pathway is as ubiquitous as in plants is unclear [37]. However, the concept that the N-terminus and C-terminus of proteins are hotspots for determining protein stability is currently emerging in eukaryotes [38]. Our findings define the nonAc-X^2/N-degron-mediated degradation as a novel hormone-regulated branch of the N-degron pathways in plants targeting a vast number of long-lived cytosolic proteins (Extended Data Fig. 8 and [27]). The previously established Arg/N-degron pathway predominantly targets short-lived regulatory proteins, whose N-degrons are conditionally generated by post-translational processing in plants (e.g., by Cys-oxidation or internal cleavage) [39, 40, 41]. Thus, the nonAc-X^2/N-degron pathway and the Arg/N-degron pathway address different types of protein subsets, causing a potentially different impact on bulk protein turnover. Unlike mutants affected in masking the nonAc-X^2/N-degron pathways, loss-of-Arg/N-degron pathway mutants grow like the wild type plants under non-stressed conditions [38, 39]. We conclude from our results that proteostasis of a large number of cytosolic NatA substrates is substantially affected by a tightly controlled ribosome-associated protein modifier that is essential in Arabidopsis and humans and determines the half-life time of proteins when they are synthesized.

Online content

Any **methods**, additional references, source data, **extended data** (list see below), supplementary information, **acknowledgements**, **details of author contributions** and **competing interests**; and statements of data and code availability are also uploaded.

Figures

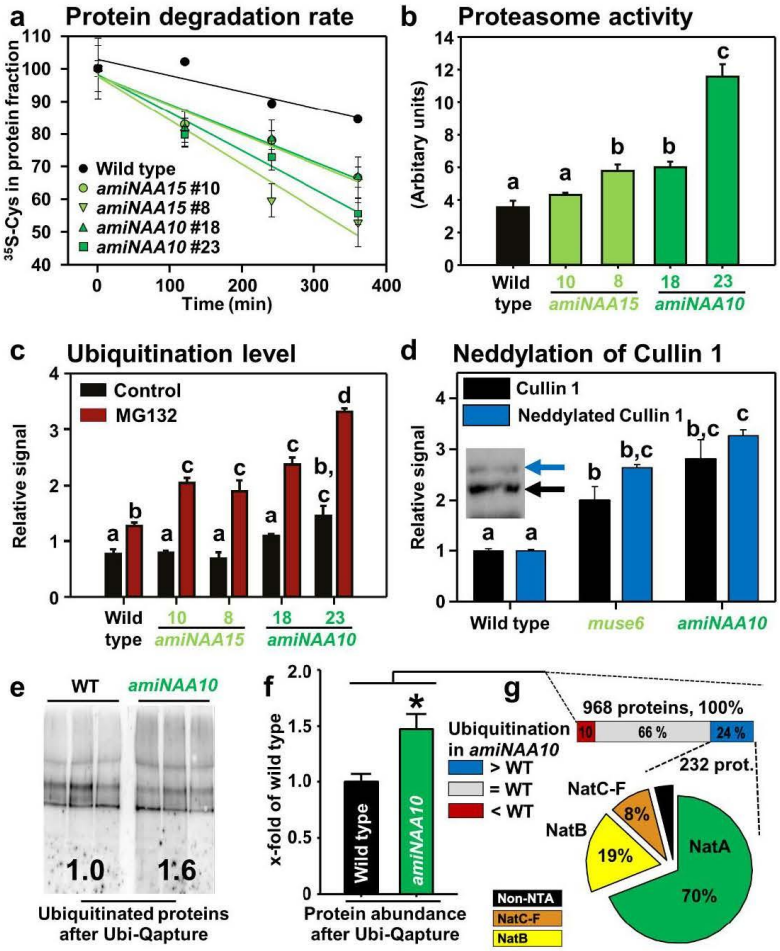


Figure 1. Depletion of NatA activity causes enhanced degradation of proteins via the ubiquitin-proteasome system. **a**, Time-resolved analysis of protein degradation rate in leaves of wild type (black) and four individual lines depleted for the ribosome anchoring (NAA15, light green) or the catalytically active subunit (NAA10, dark green) of the NatA complex ($p < 0.05$, n = X). **b**, Proteasome activity in wild type and NatA activity depleted plants ($p < 0.05$, n = 4). **c**, Relative level of poly-ubiquitinated proteins as determined with the ubiquitin-specific antiserum (α-UBQ11; Agrisera) in leaves of wild type and NatA depleted plants in the presence (red) or absence (black) of the proteasome inhibitor MG132. **d**, Abundance and activation status of Cullin-RING E3 ligase (CRL) complexes as determined by neddylation of Cullin isoform 1 in wild type and transgenic lines depleted of NAA15 (*muse6*) and NAA10 (*amiNAA10*). **e**, Immunological detection of poly-ubiquitinated proteins using an ubiquitin-specific antibody (α-UBQ11; Agrisera) in the wild type and *amiNAA10* line23 (*amiNAA10*) after selective enrichment with the Ubi-Qapture Q^{TM} matrix. **f**, Protein amount of affinity enriched poly-ubiquitinated proteins. (n = 3) **g**, Quantitative proteomics of poly-ubiquitinated proteins revealed that 24 % of quantified proteins (232 proteins) were significantly ($p < 0.05$) more ubiquitinated in *amiNAA10* (> 1.5-fold more than wild type) or either only found in all aminNAA10 replicates (n = 3). The pie chart depicts the classification of Nat substrates in the fraction of significantly more ubiquitinated proteins in *amiNAA10* plants.

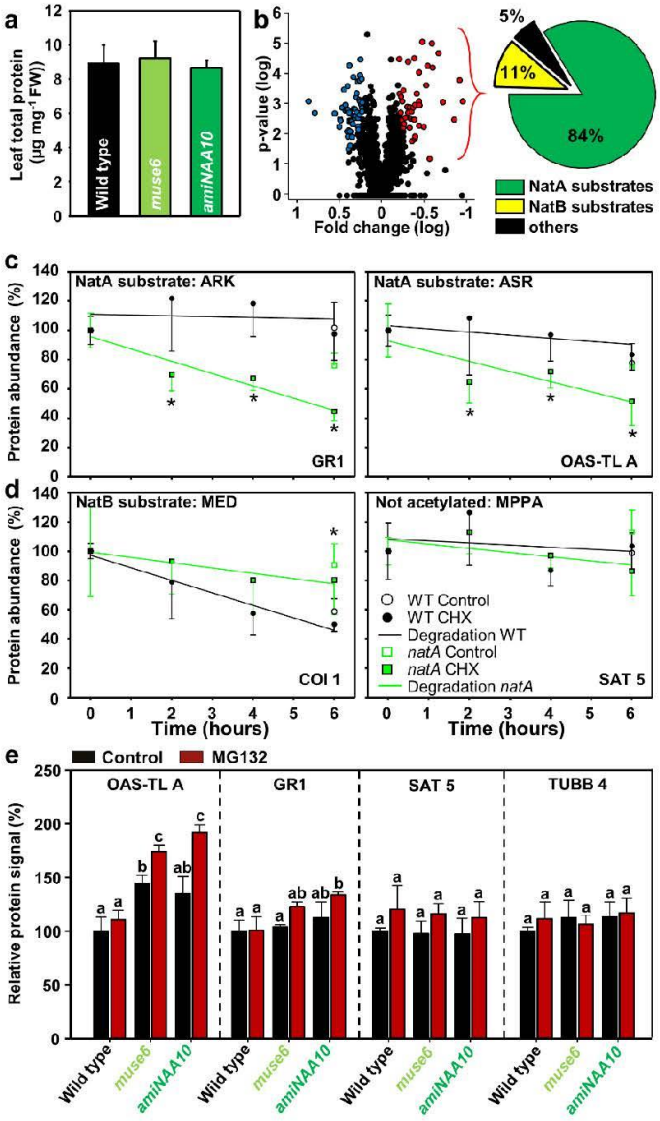


Figure 2. Depletion of NatA activity does not affect total protein steady-state level but significantly destabilize selected NatA substrates. **a**, Concentration of total proteins in leaves of wild type and NatA depleted plants (*muse6* and *amiNAA10*). **b**, Comparison of leaf proteins in wild type and *amiNAA10* as volcano plot to identify changes in the leaf-proteome due to depletion of the catalytic NatA subunit (*amiNAA10* versus wild type). Significantly altered proteins in amiNAA10 are labeled in color (red, decreased, blue, accumulated, FDR < 0.01, n = 4). The pie diagram displays the classification of Nat substrates in the fraction of significantly decreased proteins in *amiNAA10*. **c, d** Time-resolved degradation analysis of selected NatA substrates (**c**, GR1, OAS-TL A) and proteins that are not N-terminally acetylated by NatA (**d**, COI1, SAT5) in the wild type (circle) and *muse6* carrying a point mutation causing lowered NatA activity [8], box) in the presence (filled) or absence (control, empty) of the translation inhibitor cycloheximide (CHX). **e**, Relative level of OAS-TL A, GR1, SAT5 and TUBB4 proteins as determined with the specific antisera in leaves of wild type and NatA depleted plants in the presence (red) or absence (black) of the proteasome inhibitor MG132. Data represent mean ± standard deviation (n = 3, $p < 0.05$).

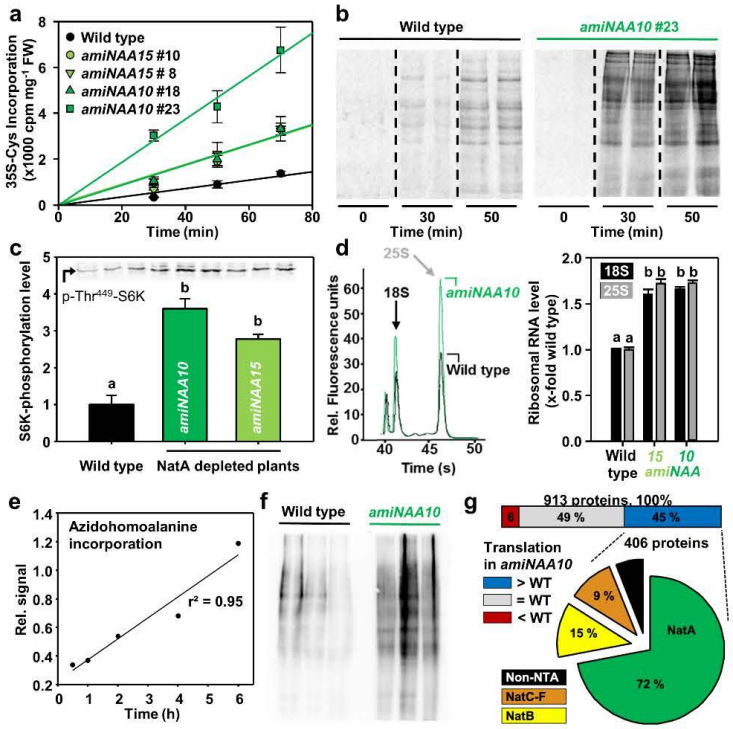


Figure 3. NatA depleted plants display higher translation rates of NatA substrates that are facilitated by TOR-induced production of ribosomes. **a**, Time-resolved incorporation of isotope-labeled sulfur amino acids into foliar proteins of the wild type and four NatA depleted lines (*amiNAA10*, *amiNAA15*). **b**, Auto-radiogram of SDS-PAGE separated foliar proteins from wild type and *amiNAA10* after incorporation of isotope-labeled sulfur-amino acids for indicated time points. **c**, Phosphorylation of T^{449} of S6K by the sensor kinase Target of Rapamycin in leaves of the wild type and NatA depleted plants as determined with a phospo-specific antiserum ($p < 0.05$, $n = 3$). **d**, Quantification of 18S and 25S ribosomal-RNAs in leaves of wild type and NatA depleted plants ($p < 0.0.5$, $n = 4$). **e**, Verification of linear azidohomoalanine incorporation into proteins derived from leaves of wild type for indicated time points (n=2). **f**, Comparison of azidohomoalanine incorporation for three hours into foliar proteins of wild type and NatA depleted plants after azidohomoalanine-mediated biotin labeling (n=3). **g**, Proteomic analysis of newly translated proteins after selective enrichment in leaves of wild type and *amiNAA10* plants. The pie diagram depicts the classification of Nat substrates in the fraction of more efficiently translated proteins in *amiNAA10* plants.

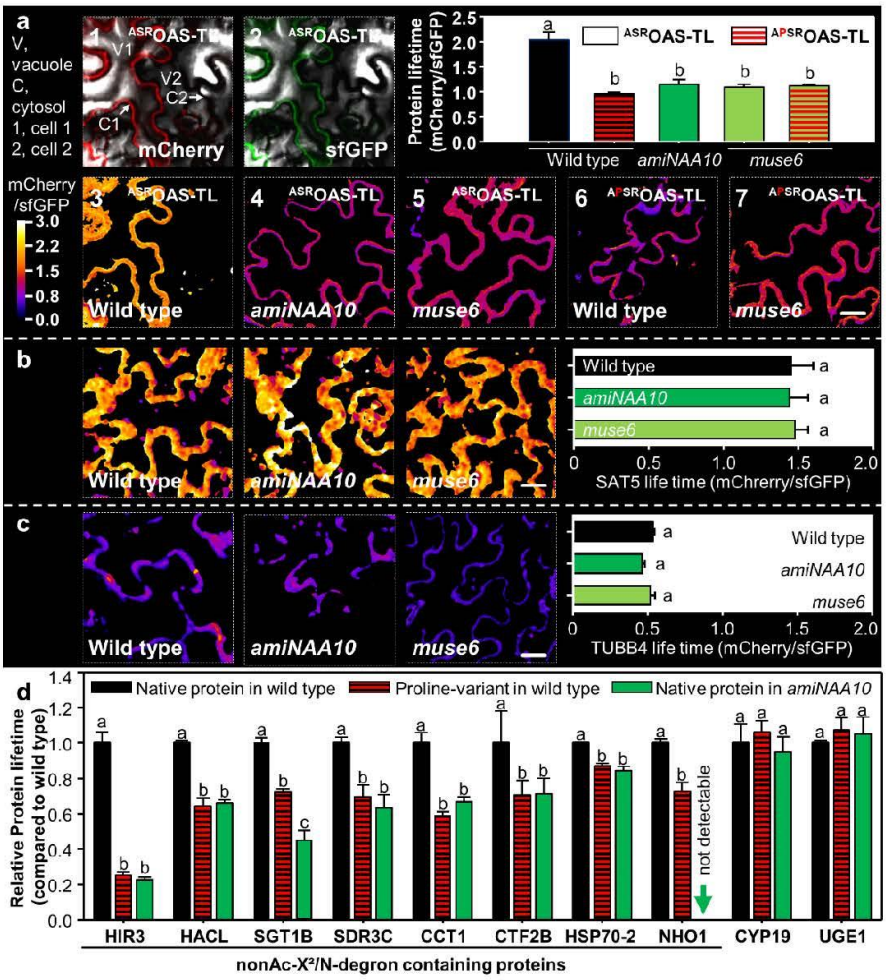


Figure 4. Non-invasive *in vivo* determination of protein half-life times in NatA depleted plants. **a**, Quantification of protein half-life times with the tandem-Fluorescence Timer (tFT) is based on the different maturation times of the fluorescent mCherry (a1) and the super-folding green fluorescent protein (sfGFP, a2) encoded on the same polypeptide chain in fusion with the protein of interest (wild type OAS-TLA, ^{ASR}OAS-TL-tFT). The mCherry/sfGFP signal ratio (a3, black bar) is a direct readout for the age of the polypeptide chain pool in the cytosol of transiently transformed epidermal leaf cells (cell 1) and positively correlates with the stability of the POI-tFT. Expression of the NatA substrate ^{ASR}OAS-TL-tFT in transgenic plants depleted for the catalytic (*amiNAA10*, a4, dark green) or the ribosome anchoring subunit of NatA (*muse6*, a5, light green) resulted in significant lower ^{ASR}OAS-TL-tFT protein half-life time. Inhibition of NTA of OAS-TL by the introduction of a proline at position 3 (^{APSR}OAS-TL-tFT, red shaded) decreased the half-life time in the wild type (a6) but did not further destabilize the protein in the NatA mutant *muse6* (a7). ($p<0.05$, n = 4 - 11). **b**-**c**, The protein half–lifetime of the non-NatA substrates ^{MPP}SAT5-tFT (**b**) and ^{MRI}TUBB4-tFT (**c**) was not affected by NatA depletion (n = 4 - 5). Scale bar, 15 µm **d**, Protein lifetime of ten cytosolic NatA substrates in wild type (black) and *amiNAA10* (green). Definition as canonical nonAc-X²/N-degron containing protein is based on destabilization of the protein by absent NTA due to protein engineering in the wild type (red shaded) or expression of the native protein in NatA depleted plants (green, *amiNAA10*). The NHO1-tFT protein abundance was below the detection limit in the cytosol of *amiNAA10* and *muse6* (Extended Data Fig. 7h).

Extended Data files:

Extended Data Fig. 1. Protease activity in leaves of wild type and NatA depleted plants.

Extended Data Fig. 2. Immunological detection of proteasome subunits, poly-ubiquitinylated proteins and Cullin 1 in leaves of wild type and NatA depleted plants.

Extended Data Fig. 3. Steady-state levels and stability of selected soluble proteins extracted from leaves of wild type and NatA depleted plants.

Extended Data Fig. 4. Accumulation of selected proteins in leaves of wild type and NatA depleted plants after inhibition of the proteasome.

Extended Data Fig. 5. Target of Rapamycin activity in leaves of wild type and NatA depleted plants.

Extended Data Fig. 6. Confirmation of candidate NatA substrates and proline-induced inhibition of N-terminal acetylation of NatA substrate candidates by in vitro NatA activity tests

Extended Data Fig. 7. Quantification of protein half-life times of NatA substrates in leaves of wild type and NatA depleted plants.

Extended Data Fig. 8. Protein half-life times of selected NatA substrates in the wild type.

Extended Data Table 1. Mass-spectrometry based identification of ubiquitinated proteins in leaves of wild type and NatA depleted plants.

Extended Data Table 2. List of proteins that were more ubiquitinated in NatA depleted plants.

Extended Data Table 3. Gene ontology enrichment analysis of proteins displaying higher poly-ubiquitination level in NatA depleted plants.

Extended Data Table 4. Protein steady-state levels in leaves of wild type and NatA depleted plants.

Extended Data Table 5. Classification of significantly decreased proteins in *amiNAA10* plants with respect to their recognition by the ribosome-associated N-terminal modification machinery.

Extended Data Table 6. Mass-spectrometry based identification of actively translated proteins in leaves of wild type and NatA depleted plants.

Extended Data Table 7. List of proteins that were more translated in NatA depleted plants.

Extended Data Table 8. List of NatA substrates displaying enhanced protein turnover in leaves of NatA depleted plants.

Extended Data Table 9. List of applied primers.

Extended Data Table 10. List of applied peptides.

References:

1. Vierstra RD. The ubiquitin-26S proteasome system at the nexus of plant biology. *Nat Rev Mol Cell Biol* **10**, 385-397 (2009).

2. Aksnes H, Drazic A, Marie M, Arnesen T. First Things First: Vital Protein Marks by N-Terminal Acetyltransferases. *Trends Biochem Sci* **41**, 746-760 (2016).

3. Kalvik TV, Arnesen T. Protein N-terminal acetyltransferases in cancer. *Oncogene* **32**, 269-276 (2013).

4. Rope Alan F, *et al.* Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency. *The American Journal of Human Genetics* **89**, 28-43 (2011).

5. Linster E, *et al.* The Arabidopsis Nα-acetyltransferase NAA60 locates to the plasma membrane and is vital for the high salt stress response. *New Phytologist* **228**, 554-569 (2020).

6. Huber M, *et al.* NatB-Mediated N-Terminal Acetylation Affects Growth and Biotic Stress Responses. *Plant Physiology* **182**, 792-806 (2020).

7. Armbruster L, *et al.* NAA50 Is an Enzymatically Active N^{α}-Acetyltransferase That Is Crucial for Development and Regulation of Stress Responses. *Plant Physiology* **183**, 1502-1516 (2020).

8. Xu F, *et al.* Two N-terminal acetyltransferases antagonistically regulate the stability of a nod-like receptor in Arabidopsis. *Plant Cell* **27**, 1547-1562 (2015).

9. Linster E, *et al.* Downregulation of N-terminal acetylation triggers ABA-mediated drought responses in Arabidopsis. *Nat Commun* **6**, 7640 (2015).

10. Aksnes H, Ree R, Arnesen T. Co-translational, Post-translational, and Non-catalytic Roles of N-Terminal Acetyltransferases. *Mol Cell* **73**, 1097-1114 (2019).

11. Mogk A, Bukau B. When the beginning marks the end. *Science* **327**, 966-967 (2010).

12. Kim HK, Kim RR, Oh JH, Cho H, Varshavsky A, Hwang CS. The N-terminal methionine of cellular proteins as a degradation signal. *Cell* **156**, 158-169 (2014).

13. Kats I, *et al.* Mapping Degradation Signals and Pathways in a Eukaryotic N-terminome. *Mol Cell* **70**, 488-501 e485 (2018).

14. Myklebust LM, *et al.* Biochemical and cellular analysis of Ogden syndrome reveals downstream Nt-acetylation defects. *Hum Mol Genet* **24**, 1956-1976 (2015).

15. Hwang CS, Shemorry A, Varshavsky A. Two proteolytic pathways regulate DNA repair by cotargeting the Mgt1 alkylguanine transferase. *Proc Natl Acad Sci U S A* **106**, 2142-2147 (2009).

16. Shemorry A, Hwang CS, Varshavsky A. Control of protein quality and stoichiometries by N-terminal acetylation and the N-end rule pathway. *Mol Cell* **50**, 540-551 (2013).

17. Li Z, *et al.* N-Terminal Acetylation Stabilizes SIGMA FACTOR BINDING PROTEIN1 Involved in Salicylic Acid-Primed Cell Death. *Plant Physiology* **183**, 358-370 (2020).

18. Schwechheimer C. NEDD8-its role in the regulation of Cullin-RING ligases. *Curr Opin Plant Biol* **45**, 112-119 (2018).

19. Li L, Nelson CJ, Trosch J, Castleden I, Huang S, Millar AH. Protein Degradation Rate in Arabidopsis thaliana Leaf Growth and Development. *Plant Cell* **29**, 207-228 (2017).

20. Ishihara H, *et al.* Growth rate correlates negatively with protein turnover in Arabidopsis accessions. *Plant J* **91**, 416-429 (2017).

21. Dong Y, *et al.* Sulfur availability regulates plant growth via glucose-TOR signaling. *Nat Commun* **8**, 1174 (2017).

22. Schepetilnikov M, Dimitrova M, Mancera-Martinez E, Geldreich A, Keller M, Ryabova LA. TOR and S6K1 promote translation reinitiation of uORF-containing mRNAs via phosphorylation of eIF3h. *EMBO J* **32**, 1087-1102 (2013).

23. Glenn WS, *et al.* Bioorthogonal Noncanonical Amino Acid Tagging (BONCAT) Enables Time-Resolved Analysis of Protein Synthesis in Native Plant Tissue. *Plant Physiol* **173**, 1543-1553 (2017).

24. Zhang H, *et al.* Tandem Fluorescent Protein Timers for Noninvasive Relative Protein Lifetime Measurement in Plants. *Plant Physiol* **180**, 718-731 (2019).

25. Goetze S, *et al.* Identification and functional characterization of N-terminally acetylated proteins in *Drosophila melanogaster*. *PLoS Biol* **7**, e1000236 (2009).

26. Gibbs DJ, Bailey M, Tedds HM, Holdsworth MJ. From start to finish: amino-terminal protein modifications as degradation signals in plants. *New Phytol* **211**, 1188-1194 (2016).

27. Martinez A, *et al.* Extent of N-terminal modifications in cytosolic proteins from eukaryotes. *Proteomics* **8**, 2809-2831 (2008).

28. Park SE, *et al.* Control of mammalian G protein signaling by N-terminal acetylation and the N-end rule pathway. *Science* **347**, 1249-1252 (2015).

29. Hwang CS, Shemorry A, Varshavsky A. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science* **327**, 973-977 (2010).

30. Weyer FA, Gumiero A, Lapouge K, Bange G, Kopp J, Sinning I. Structural basis of HypK regulating N-terminal acetylation by the NatA complex. *Nat Commun* **8**, 15726 (2017).

31. Gottlieb L, Marmorstein R. Structure of Human NatA and Its Regulation by the Huntingtin Interacting Protein HYPK. *Structure* **26**, 925-935 e928 (2018).

32. Miklánková P, *et al.* HYPK promotes activity of the essential N^{α}-acetyltransferase A complex to determine proteostasis of nonAc-X^2/N-degron containing proteins. *Nature Plants*, (submitted).

33. Gong X, *et al.* OsHYPK-mediated protein N-terminal acetylation coordinates rice development and stress responses through dynamic shifts in protein turnover. *Nature Plants*, (submitted).

34. Chang C-C, Chen H-L, Li W-H, Chaw S-M. Dating the Monocot-Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. *Journal of molecular evolution* **58**, 424-441 (2004).

35. Timms RT, Zhang Z, Rhee DY, Harper JW, Koren I, Elledge SJ. A glycine-specific N-degron pathway mediates the quality control of protein *N*-myristoylation. *Science* **365**, eaaw4912 (2019).

36. Mueller F, *et al.* Overlap of NatA and IAP substrates implicates N-terminal acetylation in protein stabilization. *Science Advances* **7**, eabc8590 (2021).

37. Yi CH, *et al.* Metabolic regulation of protein N-alpha-acetylation by Bcl-xL promotes cell survival. *Cell* **146**, 607-620 (2011).

38. Varshavsky A. N-degron and C-degron pathways of protein degradation. *Proc Natl Acad Sci U S A* **116**, 358-366 (2019).

39. Graciet E, *et al.* The N-end rule pathway controls multiple functions during Arabidopsis shoot and leaf development. *Proc Natl Acad Sci U S A* **106**, 13618-13623 (2009).

40. Potuschak T, Stary S, Schlogelhofer P, Becker F, Nejinskaia V, Bachmair A. PRT1 of Arabidopsis thaliana encodes a component of the plant N-end rule pathway. *Proc Natl Acad Sci U S A* **95**, 7904-7908 (1998).

41. Gibbs DJ, *et al.* Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants. *Nature* **479**, 415-418 (2011).

4. Discussion

MS-based proteomics has become one of the most practical and universally applicable tools in biological, medical and life sciences over the last decades (Ruedi Aebersold & Mann, 2003, 2016). As with other omics technologies, this does not mean that new developments are not necessary any more. Especially for omics methods in which multiplexed data is acquired, they can always be faster, more sensitive, accurate or cheaper. In the key publication of my thesis 'The proteome landscape of the kingdoms of life' we demonstrate how ubiquitously applicable MS-based proteomics has in fact become by now, by quantitatively measuring the proteomes of 100 organisms across the entire tree of life. From a biological standpoint this means that every organism whose genome is sequenced and from which proteins can be extracted, those proteins can be quantified accurately and routinely. This may have a dramatic impact on the way evolutionary science can be thought and done in the future. In contrast it is apparent from the literature that most protein quantifications are still done by affinity-based methods and even reviewers in prominent journals ask for those methods to validate e.g. MS-based derived data (**Figure 11**).

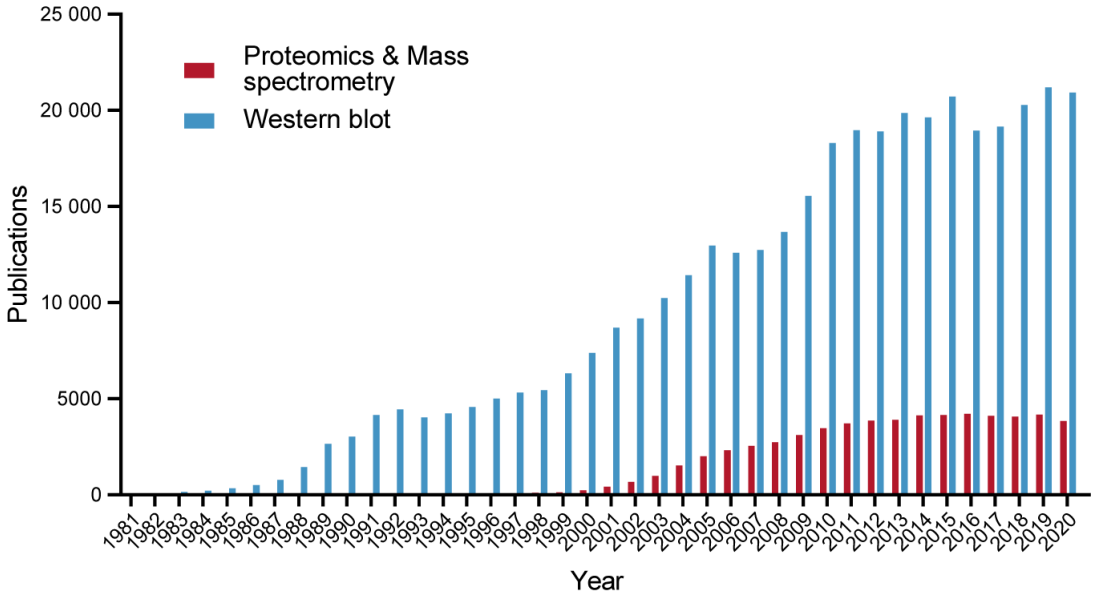


Figure 11: PubMed literature search of terms 'Western blot' respective 'Proteomics' and 'Mass spectrometry'.

With the ability to quantify proteins from every sequenced organism, the advantages of MS-based methods to affinity-based methods have become even clearer, as there rarely are available antibodies for uncommon proteins and usually none in less studied organisms. In the 'Bear proteome' publication that is in preparation, we also demonstrate this with an example where exactly this challenge occurs. The organism *Ursos arctos* commonly known as brown bear has no commercially available antibodies at all. We

screened the plasma and thrombocytes of active and hibernating bears to find biological explanations for the bear's avoidance of venous thrombosis during hibernation. Even if the needed antibodies were available, it would not be reasonable to set up a study, looking at proteins of interest known from humans with e.g. ELISA techniques in the bear samples. This is a good example why this evolutionary medicine approach to explore biology and method to find potential new drug targets or biomarkers is rarely pursued. I believe that it is of major interest to the scientific community to spread the knowledge about the applicability and power of these methods. From the impact of our publication 'The proteome landscape of the kingdoms of life' it is clear that the ideas we promote here are still unfamiliar despite the fact that MS-based proteomics and genome sequencing technique have been ready to enable such experiments since more than a decade.

The impact of 'The proteome landscape of the kingdoms of life' of course also lies in the insights which can be drawn from our dataset. Different proteomes can be compared in abundance distribution and it has become clear in our cross section of the tree of life that the exponential distribution of protein abundances (best described by a beta distribution) within the proteome is universal. Of special interest are the number of poorly characterized or undescribed proteins even among the high abundant ones for a great number of organisms. These highlight proteins that are apparently important biologically and might be worth studying for biotechnological applications and researchers specializing in these organisms. On a functional level we describe the overall most abundant biological processes and protein subdomains and abundances of related proteins are comparable between organisms by homology information. Although these outcomes are not necessarily new in themselves, it has never been possible to study them in such a comprehensive way. Single organisms, proteins or functional activities may have been compared, but our proteomics approach gives a more complete view by subjecting all organisms to the same workflow and analysis method.

Additionally, we provide first solid evidence for the existence of thousands of proteins which had only been predicted from genome sequencing but never actually been observed. This also manifests in the interest of database providers like UniProt to integrate our data into their knowledge pool, which is now planned.

The other major topics in my thesis are developments of LC tools for MS-based proteomics and implementation of clinical proteomics studies.

The Evosep One LC which was developed in the beginning of my PhD time has since been applied especially for every large-scale plasma proteomics project and new

specific applications for longer gradients and high sensitivity workflows are on the way. The reproducibility and throughput of the system represent a milestone in hardware development for clinical proteomics and will provide a valuable tool in the future. The packing station project for high throughput multiplexed capillary column production with high pressure arose from the need to provide material for multiple groups with minimal hands on time. Since the implementation of the new technique, it has proven valuable in reducing the time in providing columns to a minimum. Where previously a single person was literally employed full time with this task, by now this can be managed in a few working days per month in our laboratory. The discussed advantages of commercial packed columns or chip-columns like employed in 'The proteome landscape of the kingdoms of life' may make packed capillary emitter production redundant in the future, but the reasons they are employed at the moment - flexibility and affordability - will still make them a good alternative in many cases.

The clinical projects in my PhD thesis, a biomarker study for Alzheimer's disease in CSF, quality markers in plasma samples by proteomics and the description of a cohort of term and preterm born infants with clinical metadata for a follow up proteomics study of dried blood spots (DBS) represent the wide range of MS-based proteomics. All three projects provided valuable insights into clinical cohorts and we were able to propose a biomarker panel for Alzheimer's disease patients in CSF. The quality marker panel for plasma proteomics will be a valuable dataset and tool for every future study. We found that those proteins are often proposed as biomarkers and our data will prevent this misinterpretation in the future. The newborn study is a long-term cooperation project with a Munich Neonatology clinic and the data provided in the presented manuscript gives insight to the unique cohort we collected for proteomics experiments. The first results from proteomics measurements of dried blood spots already are striking and show that we are able to describe preterm development by blood proteomics.

5. References

Aebersold, R., & Goodlett, D. R. (2001). Mass spectrometry in proteomics. *Chemical Reviews*, *101*(2), 269–295. https://doi.org/10.1021/cr990076h

Aebersold, Ruedi, & Mann, M. (2003). Mass spectrometry-based proteomics. In *Nature* (Vol. 422, Issue 6928, pp. 198–207). Nature Publishing Group. https://doi.org/10.1038/nature01511

Aebersold, Ruedi, & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*(7620), 347–355. https://doi.org/10.1038/nature19949

Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Herman, J. S., Pessaux, P., Baumert, T. F., & Grün, D. (2019). A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, *572*(7768), 199–204. https://doi.org/10.1038/s41586-019-1373-2

Andersen, J. S., Lam, Y. W., Leung, A. K. L., Ong, S. E., Lyon, C. E., Lamond, A. I., & Mann, M. (2005). Nucleolar proteome dynamics. *Nature*, *433*(7021), 77–83. https://doi.org/10.1038/nature03207

Andersson, L., & Porath, J. (1986). Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. *Analytical Biochemistry*, *154*(1), 250–254. https://doi.org/10.1016/0003-2697(86)90523-3

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., Mayr, C. H., Greiffo, F. R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T. M., Nagendran, M., Desai, T., Eickelberg, O., Mann, M., Theis, F. J., & Schiller, H. B. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-08831-9

Bard, F., & Chia, J. (2016). Cracking the Glycome Encoder: Signaling, Trafficking, and Glycosylation. In *Trends in Cell Biology* (Vol. 26, Issue 5, pp. 379–388). Elsevier Ltd. https://doi.org/10.1016/j.tcb.2015.12.004

Batth, T. S., Tollenaere, M. A. X., Rüther, P., Gonzalez-Franquesa, A., Prabhakar, B. S., Bekker-Jensen, S., Deshmukh, A. S., & Olsen, J. V. (2019). Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Molecular and Cellular Proteomics*, *18*(5), 1027–1035. https://doi.org/10.1074/mcp.TIR118.001270

Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen, C. L., Nielsen, M. L., & Olsen, J. V. (2017). An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Systems*, *4*(6), 587-

599.e4. https://doi.org/10.1016/j.cels.2017.05.009

Bernhardt, O., Selevsek, N., Gillet, L., Rinner, O., Picotti, P., Aebersold, R., Reiter, L., Bernhardt, O., Selevsek, N., Gillet, L., Rinner, O., Picotti, P., Aebersold, R., & Reiter, L. (2014). Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data . *F1000Research*, *5*. https://doi.org/10.7490/F1000RESEARCH.1096450.1

Bian, Y., Zheng, R., Bayer, F. P., Wong, C., Chang, Y. C., Meng, C., Zolg, D. P., Reinecke, M., Zecha, J., Wiechmann, S., Heinzlmeir, S., Scherr, J., Hemmer, B., Baynham, M., Gingras, A. C., Boychenko, O., & Kuster, B. (2020). Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-019-13973-x

Blume, J. E., Manning, W. C., Troiano, G., Hornburg, D., Figa, M., Hesterberg, L., Platt, T. L., Zhao, X., Cuaresma, R. A., Everley, P. A., Ko, M., Liou, H., Mahoney, M., Ferdosi, S., Elgierari, E. M., Stolarczyk, C., Tangeysh, B., Xia, H., Benz, R., … Farokhzad, O. C. (2020). Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nature Communications*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-17033-7

Bustin, S. A., Benes, V., Nolan, T., & Pfaffl, M. W. (2005). Quantitative real-time RT-PCR - A perspective. In *Journal of Molecular Endocrinology* (Vol. 34, Issue 3, pp. 597–601). J Mol Endocrinol. https://doi.org/10.1677/jme.1.01755

Chalkley, R. J., MacCoss, M. J., Jaffe, J. D., & Rost, H. L. (2019). Initial guidelines for manuscripts employing data-independent acquisition mass spectrometry for proteomic analysis. In *Molecular and Cellular Proteomics* (Vol. 18, Issue 1, pp. 1–2). American Society for Biochemistry and Molecular Biology Inc. https://doi.org/10.1074/mcp.E118.001286

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Ctortecka, C., & Mechtler, K. (2021). The rise of single-cell proteomics. *Analytical Science Advances*, ansa.202000152. https://doi.org/10.1002/ansa.202000152

Dawson, P. H. (1986). Quadrupole mass analyzers: Performance, design and some recent applications. *Mass Spectrometry Reviews*, *5*(1), 1–37. https://doi.org/10.1002/mas.1280050102

De Beeck, J. O., Pauwels, J., Van Landuyt, N., Jacobs, P., De Malsche, W., Desmet, G., Argentini, A., Staes, A., Martens, L., Impens, F., & Gevaert, K. (2018). A well-

ordered nanoflow LC-MS/MS approach for proteome profiling using 200 cm long micro pillar array columns. In *bioRxiv* (p. 472134). bioRxiv. https://doi.org/10.1101/472134

De Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., & Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, *455*(7217), 1251–1254. https://doi.org/10.1038/nature07341

Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, *17*(1), 41–44. https://doi.org/10.1038/s41592-019-0638-x

Dempster, A. J. (1918). A new method of positive ray analysis. *Physical Review*, *11*(4), 316–325. https://doi.org/10.1103/PhysRev.11.316

Doerr, A. (2014). DIA mass spectrometry. In *Nature Methods* (Vol. 12, Issue 1, p. 35). Nature Publishing Group. https://doi.org/10.1038/nmeth.3234

Doll, S., Dreßen, M., Geyer, P. E., Itzhak, D. N., Braun, C., Doppler, S. A., Meier, F., Deutsch, M. A., Lahm, H., Lange, R., Krane, M., & Mann, M. (2017). Region and cell-type resolved quantitative proteomic map of the human heart. *Nature Communications*, *8*(1), 1–13. https://doi.org/10.1038/s41467-017-01747-2

Doll, S., Kriegmair, M. C., Santos, A., Wierer, M., Coscia, F., Neil, H. M., Porubsky, S., Geyer, P. E., Mund, A., Nuhn, P., & Mann, M. (2018). Rapid proteomic analysis for solid tumors reveals LSD1 as a drug target in an end-stage cancer patient. *Molecular Oncology*, *12*(8), 1296–1307. https://doi.org/10.1002/1878-0261.12326

Douglas, D. J. (2009). Linear quadrupoles in mass spectrometry. *Mass Spectrometry Reviews*, *28*(6), 937–960. https://doi.org/10.1002/mas.20249

Dyring-Andersen, B., Løvendorf, M. B., Coscia, F., Santos, A., Møller, L. B. P., Colaço, A. R., Niu, L., Bzorek, M., Doll, S., Andersen, J. L., Clark, R. A., Skov, L., Teunissen, M. B. M., & Mann, M. (2020). Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin. *Nature Communications*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-19383-8

Edman. (1949). A method for the determination of amino acid sequence in peptides. *Arch Biochem*, *22*(3), :475.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Emmett, M. R., & Caprioli, R. M. (1994). *Micro-Electrospray Mass Spectrometry: Ultra-High-Sensitivity Analysis of Peptides and Proteins*. https://doi.org/10.1021/JASMS.8B00583

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. In *Science*. https://doi.org/10.1126/science.2675315

Fischer, E. (1906). *Untersuchungen über Aminosäuren, Polypeptide und Proteïne*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. https://doi.org/https://doi.org/10.1002/cber.19060390190

Geiger, T., Velic, A., MacEk, B., Lundberg, E., Kampf, C., Nagaraj, N., Uhlen, M., Cox, J., & Mann, M. (2013). Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Molecular and Cellular Proteomics*, *12*(6), 1709–1722. https://doi.org/10.1074/mcp.M112.024919

Geyer, P. E., Holdt, L. M., Teupser, D., & Mann, M. (2017). Revisiting biomarker discovery by plasma proteomics. *Molecular Systems Biology*, *13*(9), 942. https://doi.org/10.15252/msb.20156297

Geyer, P. E., Kulak, N. A., Pichler, G., Holdt, L. M., Teupser, D., & Mann, M. (2016). Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Systems*, *2*(3), 185–195. https://doi.org/10.1016/J.CELS.2016.02.015

Geyer, P. E., Wewer Albrechtsen, N. J., Tyanova, S., Grassl, N., Iepsen, E. W., Lundgren, J., Madsbad, S., Holst, J. J., Torekov, S. S., & Mann, M. (2016). Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Molecular Systems Biology*, *12*(12), 901. https://doi.org/10.15252/msb.20167357

Giegé, R. (2013). A historical perspective on protein crystallization from 1840 to the present day. In *FEBS Journal* (Vol. 280, Issue 24, pp. 6456–6497). FEBS J. https://doi.org/10.1111/febs.12580

Giles, K., Pringle, S. D., Worthington, K. R., Little, D., Wildgoose, J. L., & Bateman, R. H. (2004). Applications of a travelling wave-based radio-frequency-only stacked ring ion guide. *Rapid Communications in Mass Spectrometry*, *18*(20), 2401–2414. https://doi.org/10.1002/rcm.1641

Grassl, N., Kulak, N. A., Pichler, G., Geyer, P. E., Jung, J., Schubert, S., Sinitcyn, P., Cox, J., & Mann, M. (2016). Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome Medicine*, *8*(1), 44. https://doi.org/10.1186/s13073-016-0293-0

Gritti, F., & Gilar, M. (2019). Impact of frit dispersion on gradient performance in high-throughput liquid chromatography. *Journal of Chromatography A*, *1591*, 110–119.

https://doi.org/10.1016/j.chroma.2019.01.021

Gross, J. H. (2017). *Mass Spectrometry*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54398-7

Guan, S., & Marshall, A. G. (1996a). Stacked-ring electrostatic ion guide. *Journal of the American Society for Mass Spectrometry*, *7*(1), 101–106. https://doi.org/10.1016/1044-0305(95)00605-2

Guan, S., & Marshall, A. G. (1996b). Stacked-ring electrostatic ion guide. *Journal of the American Society for Mass Spectrometry*, *7*(1), 101–106. https://doi.org/10.1016/1044-0305(95)00605-2

Hansen, F., Tanzer, M., Brüning, F., Bludau, I., Schulman, B., Robles, M., Karayel, O., & Mann, M. (2020). Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. *BioRxiv*, 2020.07.24.219055. https://doi.org/10.1101/2020.07.24.219055

Hebert, A. S., Prasad, S., Belford, M. W., Bailey, D. J., McAlister, G. C., Abbatiello, S. E., Huguet, R., Wouters, E. R., Dunyach, J. J., Brademan, D. R., Westphall, M. S., & Coon, J. J. (2018). Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Analytical Chemistry*, *90*(15), 9529–9537. https://doi.org/10.1021/acs.analchem.8b02233

Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., & Coon, J. J. (2014). The one hour yeast proteome. *Molecular and Cellular Proteomics*, *13*(1), 339–347. https://doi.org/10.1074/mcp.M113.034769

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I. A., Weisswange, I., Mansfeld, J., Buchholz, F., Hyman, A. A., & Mann, M. (2015). A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell*, *163*(3), 712–723. https://doi.org/10.1016/j.cell.2015.09.053

Hein, M. Y., Sharma, K., Cox, J., & Mann, M. (2013). Proteomic Analysis of Cellular Systems. In *Handbook of Systems Biology* (pp. 3–25). Elsevier Inc. https://doi.org/10.1016/B978-0-12-385944-0.00001-0

Hellgren, E. C. (1998). Physiology of hibernation in bears. *Ursus*, *10*, 467–477. https://doi.org/10.2307/3873159

Hofmeister, F. (1889). Ueber die Darstellung von krystallisirtem Eieralbumin und die Krystallisirbarkeit kolloidaler Stoffe. *Z..Physiol. f. Chemie*, *14*, S. 165-172.

Horváth, C., Melander, W., & Molnár, I. (1976). Solvophobic interactions in liquid chromatography with nonpolar stationary phases. *Journal of Chromatography A*, *125*(1), 129–156. https://doi.org/10.1016/S0021-9673(00)93816-0

Hosp, F., Scheltema, R. A., Eberl, H. C., Kulak, N. A., Keilhauer, E. C., Mayr, K., &

Mann, M. (2015). A double-barrel liquid chromatography- Tandem mass spectrometry (LC-MS/MS) system to quantify 96 interactomes per day. *Molecular and Cellular Proteomics*, *14*(7), 2030–2041. https://doi.org/10.1074/mcp.O115.049460

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., & Cooks, R. G. (2005). The Orbitrap: A new mass spectrometer. In *Journal of Mass Spectrometry* (Vol. 40, Issue 4, pp. 430–443). John Wiley & Sons, Ltd. https://doi.org/10.1002/jms.856

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., & Krijgsveld, J. (2019). Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols*, *14*(1), 68–85. https://doi.org/10.1038/s41596-018-0082-x

Humphrey, S. J., Karayel, O., James, D. E., & Mann, M. (2018). High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nature Protocols*, *13*(9), 1897–1916. https://doi.org/10.1038/s41596-018-0014-9

Ishihama, Y., Rappsilber, J., Andersen, J. S., & Mann, M. (2002). Microcolumns with self-assembled particle frits for proteomics. *Journal of Chromatography A*, *979*(1–2), 233–239. https://doi.org/10.1016/S0021-9673(02)01402-4

Itzhak, D. N., Tyanova, S., Cox, J., & Borner, G. H. H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *ELife*, *5*(JUN2016). https://doi.org/10.7554/eLife.16950

J. B. Beccari. (1731). De frumento. *De Bononensi Scientiarum et Artium Instituto Atque Academia Commentarii*, *Tomus II*(Pars I), p.122-127.

Jiang, L., He, L., & Fountoulakis, M. (2004). Comparison of protein precipitation methods for sample preparation prior to proteomic analysis. *Journal of Chromatography A*, *1023*(2), 317–320. https://doi.org/10.1016/j.chroma.2003.10.029

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Johnson, R. S., & Biemann, K. (1989). Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides.

Biomedical & Environmental Mass Spectrometry, *18*(11), 945–957. https://doi.org/10.1002/bms.1200181102

Karas, M., & Hillenkamp, F. (1988). Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. In *Analytical Chemistry* (Vol. 60, Issue 20, pp. 2299–2301). Anal Chem. https://doi.org/10.1021/ac00171a028

Keilhauer, E. C., Hein, M. Y., & Mann, M. (2015). Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Molecular and Cellular Proteomics*, *14*(1), 120–135. https://doi.org/10.1074/mcp.M114.041012

Kelly, R. T., Tolmachev, A. V., Page, J. S., Tang, K., & Smith, R. D. (2010). The ion funnel: Theory, implementations, and applications. *Mass Spectrometry Reviews*, *29*(2), 294–312. https://doi.org/10.1002/mas.20232

Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hogrebe, A., Harder, A., & Olsen, J. V. (2018). Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *Journal of Proteome Research*, *17*(1), 727–738. https://doi.org/10.1021/acs.jproteome.7b00602

Kennedy, R. T., & Jorgenson, J. W. (1989). Preparation and Evaluation of Packed Capillary Liquid Chromatography Columns with Inner Diameters from 20 to 50 μm. *Analytical Chemistry*, *61*(10), 1128–1135. https://doi.org/10.1021/ac00185a016

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., … Pandey, A. (2014). A draft map of the human proteome. *Nature*, *509*(7502), 575–581. https://doi.org/10.1038/nature13302

Kim, T., Tolmachev, A. V., Harkewicz, R., Prior, D. C., Anderson, G., Udseth, H. R., Smith, R. D., Bailey, T. H., Rakov, S., & Futrell, J. H. (2000). Design and implementation of a new electrodynamic ion funnel. *Analytical Chemistry*, *72*(10), 2247–2255. https://doi.org/10.1021/ac991412x

Kolakowski, B. M., & Mester, Z. (2007). Review of applications of high-field asymmetric waveform ion mobility spectrometry (FAIMS) and differential mobility spectrometry (DMS). In *Analyst* (Vol. 132, Issue 9, pp. 842–864). Royal Society of Chemistry. https://doi.org/10.1039/b706039d

Kondrat, R. W., McClusky, G. A., & Cooks, R. G. (1978). Multiple Reaction Monitoring in Mass Spectrometry/Mass Spectrometry for Direct Analysis of Complex Mixtures. *Analytical Chemistry*, *50*(14), 2017–2021. https://doi.org/10.1021/ac50036a020

Kovalchuk, S. I., Jensen, O. N., & Rogowska-Wrzesinska, A. (2019). FlashPack: Fast and simple preparation of ultrahigh-performance capillary columns for LC-MS. *Molecular and Cellular Proteomics*, *18*(2), 383–390. https://doi.org/10.1074/mcp.TIR118.000953

Krahmer, N., Najafi, B., Schueder, F., Quagliarini, F., Steger, M., Seitz, S., Kasper, R., Salinas, F., Cox, J., Uhlenhaut, N. H., Walther, T. C., Jungmann, R., Zeigerer, A., Borner, G. H. H., & Mann, M. (2018). Organellar Proteomics and Phospho-Proteomics Reveal Subcellular Reorganization in Diet-Induced Hepatic Steatosis. *Developmental Cell*, *47*(2), 205-221.e7. https://doi.org/10.1016/j.devcel.2018.09.017

Kulak, N. A., Geyer, P. E., & Mann, M. (2017). Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Molecular & Cellular Proteomics : MCP*, *16*(4), 694–705. https://doi.org/10.1074/mcp.O116.065136

Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., & Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods*, *11*(3), 319–324. https://doi.org/10.1038/nmeth.2834

Levsen, K., & Schwarz, H. (1976). Stoßaktivierungsmassenspektrometrie - eine neue Sonde zur Strukturbestimmung von Ionen in der Gasphase. *Angewandte Chemie*, *88*(18), 589–599. https://doi.org/10.1002/ange.19760881802

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., & Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, *17*(7), 676–682. https://doi.org/10.1038/10890

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, *72*(6), 1156–1162. https://doi.org/10.1021/ac991131p

Mann, M., & Wilm, M. (1994). Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry*, *66*(24), 4390–4399. https://doi.org/10.1021/ac00096a002

Marx, V. (2019). A dream of single-cell proteomics. *Nature Methods*, *16*(9), 809–812. https://doi.org/10.1038/s41592-019-0540-6

McLafferty, F. W., Bente, P. F., Kornfeld, R., Tsai, S., & Howe, I. (1973). Metastable ion characteristics. XXII. Collisional activation spectra of organic ions. *Journal of the American Chemical Society*, *95*(7), 2120–2129. https://doi.org/10.1021/ja00788a007

Meier, F., Beck, S., Grassl, N., Lubeck, M., Park, M. A., Raether, O., & Mann, M.

(2015). Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *Journal of Proteome Research*, *14*(12), 5378–5387. https://doi.org/10.1021/acs.jproteome.5b00932

Meier, F., Brunner, A. D., Frank, M., Ha, A., Bludau, I., Voytik, E., Kaspar-Schoenefeld, S., Lubeck, M., Raether, O., Bache, N., Aebersold, R., Collins, B. C., Röst, H. L., & Mann, M. (2020). diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods*, *17*(12), 1229–1236. https://doi.org/10.1038/s41592-020-00998-0

Meier, F., Brunner, A. D., Koch, S., Koch, H., Lubeck, M., Krause, M., Goedecke, N., Decker, J., Kosinski, T., Park, M. A., Bache, N., Hoerning, O., Cox, J., Räther, O., & Mann, M. (2018). Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Molecular and Cellular Proteomics*, *17*(12), 2534–2545. https://doi.org/10.1074/mcp.TIR118.000900

Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J., & Mann, M. (2018). BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods*, *15*(6), 440–448. https://doi.org/10.1038/s41592-018-0003-5

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., Samaras, P., Richter, S., Shikata, H., Messerer, M., Lang, D., Altmann, S., Cyprys, P., Zolg, D. P., Mathieson, T., Bantscheff, M., Hazarika, R. R., Schmidt, T., Dawid, C., … Kuster, B. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature*, *579*(7799), 409–414. https://doi.org/10.1038/s41586-020-2094-2

Messner, C., Demichev, V., Bloomfield, N., Ivosev, G., Wasim, F., Zelezniak, A., Lilley, K., Tate, S., & Ralser, M. (2019). ScanningSWATH enables ultra-fast proteomics using high-flow chromatography and minute-scale gradients. In *bioRxiv* (p. 656793). bioRxiv. https://doi.org/10.1101/656793

Meyer, J. G., Niemi, N. M., Pagliarini, D. J., & Coon, J. J. (2020). *Quantitative shotgun proteome analysis by direct infusion*. https://doi.org/10.1038/s41592-020-00999-z

Miller', P. E., & Denton, M. B. (n.d.). *The Quadrupole Mass Filter: Basic Operating Concepts*.

Mulder, G. J. (1839). Ueber die Zusammensetzung einiger thierischen Substanzen. *Journal Für Praktische Chemie*, *16*(1), 129–152. https://doi.org/10.1002/prac.18390160137

Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., & 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. *Molecular & Cellular Proteomics : MCP*, *11*(3), M111.013722. https://doi.org/10.1074/mcp.M111.013722

Nättinen, J., Jylhä, A., Aapola, U., Mäkinen, P., Beuerman, R., Pietilä, J., Vaajanen, A., & Uusitalo, H. (2019). Age-associated changes in human tear proteome. *Clinical Proteomics*, *16*(1), 11. https://doi.org/10.1186/s12014-019-9233-5

Niu, L., Geyer, P. E., Wewer Albrechtsen, N. J., Gluud, L. L., Santos, A., Doll, S., Treit, P. V, Holst, J. J., Knop, F. K., Vilsbøll, T., Junker, A., Sachs, S., Stemmer, K., Müller, T. D., Tschöp, M. H., Hofmann, S. M., & Mann, M. (2019). Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Molecular Systems Biology*, *15*(3). https://doi.org/10.15252/msb.20188793

Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., & Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, *4*(9), 709–712. https://doi.org/10.1038/nmeth1060

Pai, M. Y., Lomenick, B., Hwang, H., Schiestl, R., McBride, W., Loo, J. A., & Huang, J. (2015). Drug affinity responsive target stability (DARTS) for small-molecule target identification. *Methods in Molecular Biology*, *1263*, 287–298. https://doi.org/10.1007/978-1-4939-2269-7_22

Paul, W., & Raether, M. (1955). Das elektrische Massenfilter. *Zeitschrift Für Physik*, *140*(3), 262–273. https://doi.org/10.1007/BF01328923

Paul, Wolfgang. (1990). Electromagnetic traps for charged and neutral particles. *Reviews of Modern Physics*, *62*(3), 531–540. https://doi.org/10.1103/RevModPhys.62.531

Paul, Wolfgang, & Steinwedel, H. (1953). Ein neues Massenspektrometer ohne Magnetfeld. In *Zeitschrift fur Naturforschung - Section A Journal of Physical Sciences* (Vol. 8, Issue 7, pp. 448–450). De Gruyter. https://doi.org/10.1515/zna-1953-0710

Peake, D. (2018). *High-resolution compound identification in metabolomics: a review of current practices*. https://assets.thermofisher.com/TFS-Assets/CMD/Reference-Materials/wp-65356-ms-high-res-compound-id-metabolomics-wp65356-en.pdf

Pepelnjak, M., de Souza, N., & Picotti, P. (2020). Detecting Protein–Small Molecule Interactions Using Limited Proteolysis–Mass Spectrometry (LiP-MS). In *Trends in Biochemical Sciences* (Vol. 45, Issue 10, pp. 919–920). Elsevier Ltd. https://doi.org/10.1016/j.tibs.2020.05.006

Pinkse, M. W. H., Uitto, P. M., Hilhorst, M. J., Ooms, B., & Heck, A. J. R. (2004). Selective isolation at the femtomole level of phosphopeptides from proteolytic

digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Analytical Chemistry*, *76*(14), 3935–3943. https://doi.org/10.1021/ac0498617

Price, P. (1991). *Standard Definitions of Terms Relating to Mass Spectrometry A Report from the Committee on Measurements and Standards of the American Society for Mass Spectrometry*. https://pubs.acs.org/sharingguidelines

Rauniyar, N. (2015). Parallel reaction monitoring: A targeted experiment performed using high resolution and high mass accuracy mass spectrometry. In *International Journal of Molecular Sciences* (Vol. 16, Issue 12, pp. 28566–28581). MDPI AG. https://doi.org/10.3390/ijms161226120

Richards, A. L., Merrill, A. E., & Coon, J. J. (2015). Proteome sequencing goes deep. In *Current Opinion in Chemical Biology* (Vol. 24, pp. 11–17). Elsevier Ltd. https://doi.org/10.1016/j.cbpa.2014.10.017

Riley, N. M., Hebert, A. S., & Coon, J. J. (2016). Proteomics Moves into the Fast Lane. In *Cell Systems* (Vol. 2, Issue 3, pp. 142–143). Cell Press. https://doi.org/10.1016/j.cels.2016.03.002

Robles, M. S., Humphrey, S. J., & Mann, M. (2017). Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. *Cell Metabolism*, *25*(1), 118–127. https://doi.org/10.1016/j.cmet.2016.10.004

Roux, K. J., Kim, D. I., Raida, M., & Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *Journal of Cell Biology*, *196*(6), 801–810. https://doi.org/10.1083/jcb.201112098

Russell, J. D., & Murphy, S. (2016). *Agilent AssayMAP Bravo Technology Enables Reproducible Automated Phosphopeptide Enrichment from Complex Mixtures Using High-Capacity Fe(III)-NTA Cartridges*. https://www.agilent.com/cs/library/applications/5991-6073EN.pdf

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*(4839), 487–491. https://doi.org/10.1126/science.2448875

Schlack, P., & Kumpf, W. (1926). Über eine neue Methode zur Ermittlung der Konstitution von Peptiden. *Hoppe-Seyler's Zeitschrift Fur Physiologische Chemie*, *154*(1–3), 125–172. https://doi.org/10.1515/bchm2.1926.154.1-3.125

Shaffer, S. A., Prior, D. C., Anderson, G. A., Udseth, H. R., & Smith, R. D. (1998). An Ion Funnel Interface for Improved Ion Focusing and Sensitivity Using Electrospray Ionization Mass Spectrometry. *Analytical Chemistry*, *70*(19), 4111–4119. https://doi.org/10.1021/ac9802170

Shishkova, E., Hebert, A. S., & Coon, J. J. (2016). Now, More Than Ever, Proteomics

Needs Better Chromatography. In *Cell Systems* (Vol. 3, Issue 4, pp. 321–324). Cell Press. https://doi.org/10.1016/j.cels.2016.10.007

Shishkova, E., Hebert, A. S., Westphall, M. S., & Coon, J. J. (2018). Ultra-High Pressure (>30,000 psi) Packing of Capillary Columns Enhancing Depth of Shotgun Proteomic Analyses. *Analytical Chemistry*, *90*(19), 11503–11508. https://doi.org/10.1021/acs.analchem.8b02766

Sielaff, M., Kuharev, J., Bohn, T., Hahlbrock, J., Bopp, T., Tenzer, S., & Distler, U. (2017). Evaluation of FASP, SP3, and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. *Journal of Proteome Research*, *16*(11), 4060–4072. https://doi.org/10.1021/acs.jproteome.7b00433

Sikosek, T., & Chan, H. S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. In *Journal of the Royal Society Interface* (Vol. 11, Issue 100). Royal Society of London. https://doi.org/10.1098/rsif.2014.0419

Silveira, J. A., Ridgeway, M. E., & Park, M. A. (2014). High resolution trapped ion mobility spectrometery of peptides. *Analytical Chemistry*, *86*(12), 5624–5627. https://doi.org/10.1021/ac501261h

Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry–based proteomics. *The Biochemist*, *42*(5), 64–69. https://doi.org/10.1042/bio20200057

Stoeckli, M., Chaurand, P., Hallahan, D. E., & Caprioli, R. M. (2001). Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, *7*(4), 493–496. https://doi.org/10.1038/86573

Swearingen, K. E., & Moritz, R. L. (2012). High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. In *Expert Review of Proteomics* (Vol. 9, Issue 5, pp. 505–517). NIH Public Access. https://doi.org/10.1586/epr.12.50

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., & Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9528–9533. https://doi.org/10.1073/pnas.0402700101

Tanzer, M. C., Frauenstein, A., Stafford, C. A., Phulphagar, K., Mann, M., & Meissner, F. (2020). Quantitative and Dynamic Catalogs of Proteins Released during Apoptotic and Necroptotic Cell Death. *Cell Reports*, *30*(4), 1260-1270.e5. https://doi.org/10.1016/j.celrep.2019.12.079

Taylor, G. (1964). Disintegration of water drops in an electric field. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*,

280(1382), 383–397. https://doi.org/10.1098/rspa.1964.0151

The Gene Ontology Resource: 20 years and still GOing strong. (2019). *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Tyndall, A. M., Overton, H., & Grindley, G. C. (1926). The mobility of ions in air. Part II.- Positive ions of short age. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *110*(754), 358–364. https://doi.org/10.1098/rspa.1926.0020

UniProt: a worldwide hub of protein knowledge. (2019). *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., & Yates, J. R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, *1*(1), 39–45. https://doi.org/10.1038/nmeth705

Virreira Winter, S., Karayel, O., Strauss, M. T., Padmanabhan, S., Surface, M., Merchant, K., Alcalay, R. N., & Mann, M. (2020). Urinary proteome profiling for stratifying patients with familial Parkinson's disease. *BioRxiv*, 2020.08.09.243584. https://doi.org/10.1101/2020.08.09.243584

Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F., Uhlen, M., Gagneur, J., Hahne, H., & Kuster, B. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, *15*(2). https://doi.org/10.15252/msb.20188503

Weis, J. H., Tan, S. S., Martin, B. K., & Wittwer, C. T. (1992). Detection of rare mRNAs via quantitative RT-PCR. *Trends in Genetics*, *8*(8), 263–264. https://doi.org/10.1016/0168-9525(92)90242-V

Welinder, K. G., Hansen, R., Overgaard, M. T., Brohus, M., Sønderkær, M., Von Bergen, M., Rolle-Kampczyk, U., Otto, W., Lindahl, T. L., Arinell, K., Evans, A. L., Swenson, J. E., Revsbech, I. G., & Frøbert, O. (2016). Biochemical foundations of health and energy conservation in hibernating free-ranging subadult brown bear ursus arctos. *Journal of Biological Chemistry*, *291*(43), 22509–22523. https://doi.org/10.1074/jbc.M116.742916

Wewer Albrechtsen, N. J., Geyer, P. E., Doll, S., Treit, P. V., Bojsen-Møller, K. N., Martinussen, C., Jørgensen, N. B., Torekov, S. S., Meier, F., Niu, L., Santos, A., Keilhauer, E. C., Holst, J. J., Madsbad, S., & Mann, M. (2018). Plasma Proteome Profiling Reveals Dynamics of Inflammatory and Lipid Homeostasis Markers after Roux-En-Y Gastric Bypass Surgery. *Cell Systems*, *7*(6), 601-612.e3. https://doi.org/10.1016/j.cels.2018.10.012

Whitehouse, C. M., Dreyer, R. N., Yamashita, M., & Fenn, J. B. (1985). Electrospray Interface for Liquid Chromatographs and Mass Spectrometers. *Analytical Chemistry*, *57*(3), 675–679. https://doi.org/10.1021/ac00280a023

Wichmann, C., Meier, F., Winter, S. V., Brunner, A.-D., Cox, J., & Mann, M. (2019). MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Molecular & Cellular Proteomics*, *18*(5), 982–994. https://doi.org/10.1074/MCP.TIR118.001131

Wierer, M., & Mann, M. (2016). Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. In *Human Molecular Genetics* (Vol. 25, Issue R2, pp. R106–R114). Oxford University Press. https://doi.org/10.1093/hmg/ddw208

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., … Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, *509*(7502), 582–587. https://doi.org/10.1038/nature13319

Wishart, D. S. (2011). Advances in metabolite identification. In *Bioanalysis* (Vol. 3, Issue 15, pp. 1769–1782). Future Science Ltd London, UK . https://doi.org/10.4155/bio.11.155

Wiśniewski, J. R. (2017). Filter-Aided Sample Preparation: The Versatile and Efficient Method for Proteomic Analysis. In *Methods in Enzymology* (Vol. 585, pp. 15–27). Academic Press Inc. https://doi.org/10.1016/bs.mie.2016.09.013

Wolff, M. M., & Stephens, W. E. (1953). A pulsed mass spectrometer with time dispersion. *Review of Scientific Instruments*, *24*(8), 616–617. https://doi.org/10.1063/1.1770801

Yang, K., & Han, X. (2016). Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. In *Trends in Biochemical Sciences* (Vol. 41, Issue 11, pp. 954–969). Elsevier Ltd. https://doi.org/10.1016/j.tibs.2016.08.010

Yost, R. A., & Enke, C. G. (1978). Selected Ion Fragmentation with a Tandem Quadrupole Mass Spectrometer. *Journal of the American Chemical Society*, *100*(7), 2274–2275. https://doi.org/10.1021/ja00475a072

Zeleny, J. (1914). The Electrical Discharge from Liquid Points, and a Hydrostatic Method of Measuring the Electric Intensity at Their Surfaces. *Physical Review*, *3*(2), 69–91. https://doi.org/10.1103/PhysRev.3.69

Zhang, X., Deeke, S. A., Ning, Z., Starr, A. E., Butcher, J., Li, J., Mayne, J., Cheng, K., Liao, B., Li, L., Singleton, R., Mack, D., Stintzi, A., & Figeys, D. (2018).

Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications*, *9*(1), 1–14. https://doi.org/10.1038/s41467-018-05357-4

Zubarev, R. A., Kelleher, N. L., & McLafferty, F. W. (1998). Electron capture dissociation of multiply charged protein cations. A nonergodic process. In *Journal of the American Chemical Society* (Vol. 120, Issue 13, pp. 3265–3266). UTC. https://doi.org/10.1021/ja973478k

6. Acknowledgements

After 3 years of PhD-journey it is time to express my gratitude to all the people which travelled along and did lead the way for me.

I would like to thank foremost Matthias Mann. The enthusiasm with which you lead, guide and supervise your groups is unreached. It is inspiring seeing your passion to not only make great science possible but also pushing it all out to the world for people to use and benefit from.

It is valid to say, that Philipp Geyer shaped my path in a major way. He recruited me to join Matthias' group and went on to be the best supervisor/scientific team mate ever. Thank you for always listening to my 'stupid ideas' and discussing every aspect of MS-based proteomics from technology to strategy. At this point, I also want to thank Klaus-Peter Janssen who is 'equally guilty' of shaping my career path by introducing me to Philipp. Thank you for your friendship that has developed since I was your student.

Sebastian Virreira Winter, Sophia Doll were the core part of our clinical office together with Philipp, and I want to thank all of you for the massive support. Exploring MS-based proteomics with three experienced persons around initiated this incredible productive period for me and it is great to see that these synergies will also shape my world in the future.

Peter Treit, having you as a team mate proved that being productive and having fun can go hand in hand. Thank you for all your help and discussions from music to multiple comparison hypotheses.

Lisa Schweizer, my desk neighbor, was a helping hand on every project and problem. Thank you for sharing projects and ideas with me and making every working day more fun!

Fynn Hansen and Jonathan Swietlik, do you remember Barcelona? That was a time to be alive! Thank you for your friendship!

Jakob Bader for discussions and shared projects. We had a fun start together, working on depletion and enrichments, thank you for your help!

Robin Eisenburger, Wolfgang Bodensohn and Janne Scharpenack, thank you for your indispensable help in collecting the Newborn study, all the patient questionnaires and data.

Susanne Pangratz-Fuehrer, for organizing every detail around the Newborn Study and giving us the opportunity to measure such a unique dataset.

Nils Kulak and Lesca Holdt for being part of my TAC committee.

Igor Paron for all the help with the instruments.

Mario Oroshi for all the help with anything IT related, from communication with web developers to the hosting of a web app.
My former office mates Florian Meier, Catherine Vasilopoulou and Andreas-David Brunner, thank you for your support.
Alison Dalfovo, Theresa Schneider, Christian Deiml for taking care of all things that you don't want to block your entire day when you are busy with experiments. Someone who takes organizational or trouble shooting tasks off your shoulders is just invaluable.
Mel Park, Craig Whitehouse, Mark Ridgeway and Jake Meyer for making my work with Bruker so interesting and smooth, be it in Billerica or remotely.
Benjamin ''Benji" Heim, the best best man, true master of protein purification and visiting Metal concerts, without your friendship and support my life would be much more boring. Thank you for everything!
My parents, Fritz and Biggi and my sister Annatina: Thank you for all your support, be it financially, emotionally and practically.
Finally, my deepest gratitude goes to Sabrina Reif, whom I may call my wife since the bittersweet summer of 2020. It sometimes must have looked as if mass specs and columns were more important, but you are the true virtue of my life. Thank you for always having my back and supporting me!