DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER FAKULTÄT FÜR CHEMIE UND PHARMAZIE
DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

PLASMA PROTEOME PROFILING TO ASSESS HUMAN HEALTH AND DISEASE

PHILIPP EMANUEL GEYER

AUS
Penzberg, DEUTSCHLAND

2017

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 08.06.2017

_____

Philipp Emanuel Geyer

Dissertation eingereicht am 09.06.2017

1. Gutachter: Prof. Dr. Matthias Mann _____

2. Gutachter: Prof. Dr. Dr. Lesca Miriam Holdt _____

Mündliche Prüfung am 12.07.2017

Abstract

The majority of diagnostic decisions are made on results from blood-based tests, and protein measurements are prominent among them. However, current assays are restricted to individual proteins, whereas it would be much more desirable to measure all of them in an unbiased, hypothesis-free manner. Therefore, characterization of the plasma proteome by mass spectrometry holds great promise for clinical application.

Due to great technological challenges and study design issues, plasma proteomics has not yet lived up to its promises: no new biomarkers have been discovered, plasma proteomics has not entered clinical diagnostics and few biologically meaningful insights have been gained. As a consequence, relatively few groups still continue to pursue plasma proteomics, despite the undiminished clinical need.

The overall aim of my PhD thesis was to pave the way for biomarker discovery and clinical applications of proteomics by precision characterization of the human blood plasma proteome. First, we streamlined the standard, time consuming and labor-intensive proteomic workflow, and replaced it by a rapid, robust and highly reproducible robotic platform. After optimization of digestion conditions, peptide clean-up procedures and LC-MS/MS procedures, we can now prepare 96 samples in a fully-automated way within 3h and we routinely measure hundreds of plasma proteomes. Our workflow decreases hands-on time and opens the field for a new concept in biomarker discovery, which we termed ‘Plasma Proteome Profiling’.

It enables the highly reproducibility (CV<20% for most proteins), and quantitative analysis of several hundred proteins from 1 µl of plasma, reflecting an individual’s physiology. The quantified proteins include inflammatory markers, proteins belonging to the lipid homeostasis system, gender-related proteins, sample quality markers and more than 50 FDA-approved biomarkers. One of my major goals was to demonstrate that MS-based proteomics can be applied to large cohorts and that it is possible to gain biologically and medically relevant information from this. We achieved this aim with our first large scale plasma proteomic study in which we analyzed by far the largest plasma proteomics study with almost 1,300 proteomes, which allowed us to define inflammatory and insulin resistance panels in a weight loss cohort.

In summary, this PhD thesis has developed the concept and practice of Plasma Proteome Profiling as a fundamentally new approach in biomarker research and medical diagnostics – the system-wide phenotyping of humans in health and disease.

Zusammenfassung

Der Großteil aller diagnostischen Entscheidungen basiert auf Bluttests, wobei Proteine den größten Anteil der untersuchten Analyten einnehmen. Die klinisch eingesetzten Assays sind jedoch auf einzelne Proteine beschränkt und es wäre erstrebenswert möglichst alle Proteine in einer einzelnen Messung hypothesenfrei und objektiv zu erfassen. Deshalb wäre die massenspektrometrische Charakterisierung des Plasma Proteomes ein sehr vielversprechender Ansatz.

Große technologische Herausforderungen und schlecht konzipierte Studien führten jedoch dazu, dass die Massenspektrometrie (MS)-basierende Proteomics die hohen Erwartungen bis heute nicht erfüllen konnte: kein einziger neuer Biomarker wurde mittels Proteomics entdeckt, die Massenspektrometrie hat nicht den Sprung in die klinische Anwendung geschafft und es war nicht möglich bedeutende biologische Erkenntnisse aus dem Plasma zu gewinnen. Dies führte dazu, dass heute nur noch relativ wenige Forschungsgruppen das Plasma Proteom untersuchen, obwohl der medizinische Bedarf noch immer genauso groß ist.

Das Ziel meiner Doktorarbeit war es den Weg für die Entdeckung neuer Biomarker und der klinischen Anwendung von Proteomics zu ebnen. Der typische Arbeitsablauf in der Proteomics ist sehr zeitintensiv und aufwendig. Deshalb haben wir ihn zuerst grundlegend vereinfacht und auf Schnelligkeit, Robustheit und Reproduzierbarkeit optimiert. Nach der Verbesserung von Verdaubedingungen, der Peptid-Aufreinigung und der Instrumentenparameter sind wir nun in der Lage 96 Proben vollautomatisiert innerhalb von 3 Stunden vorzubereiten und hunderte von Plasma Proteomen am Stück zu messen. Dieser Arbeitsablauf vereinfacht nicht nur die proteomische Anwendung im Allgemeinen, sondern eröffnet auch die Möglichkeit eines neuen Konzepts in der Biomarkerforschung.

Dieses neue Konzept bezeichnen wir als „Plasma Proteome Profiling“. Es erlaubt die hochreproduzierbare (CV<20%) Quantifizierung von hunderten von Proteinen aus einem Mikroliter Plasma und liefert damit eine Reflektion des Gesamtzustandes eines Menschen. Unter andern messen wir Entzündungsmarker, Proteine des Fettstoffwechselsystems, geschlechts-spezifische Proteine, Qualitätsmarker und zudem über 50 verschiedene bereits klinisch angewendete Biomarker. Eines meiner zentralen Ziele war es die Messung großer Kohorten zu ermöglichen. So haben wir die bis heute größte Plasma Proteomics Studie mit annähernd 1300 Plasma Proteomen analysiert und dabei klinisch bedeutende Informationen über den Entzündungsstatus und die Insulin-Resistenz-Neigung von Studienteilnehmern entdeckt.

Das in dieser Doktorarbeit entwickelte Konzept von Plasma Proteom Profiling ist ein grundsätzlich neuer Ansatz in der Biomarkerforschung und auch für die medizinische Diagnostik, die zum Phänotypisieren von Menschen mittels minimaler Blutmengen eingesetzt werden kann. Bereits heute setzen wir Plasma Proteomics Profiling auf täglicher Basis zur Erforschung neuer krankheitsrelevanter Biomarker in verschiedenen Studien ein. Auch weiterhin investieren wir viel Energie in die Erforschung neuer Technologien um unsere Idee der proteomischen Phänotypisierung mittels Plasma Proteom Profiling noch weiter auszubauen und sie schließlich in die klinische Anwendung zu übertragen.

Table of contents

1. Introduction

1.1. Mass spectrometry-based exploration of the proteome

Proteins control and execute the vast majority of biological processes. Changes in their expression levels, activity, localization or interaction characterize different states of biological systems. The proteome is defined as the entirety of all proteins in a biological system and proteomics is the technology and approach for its large-scale investigation. The proteomics field has benefited from continuous development over the last 20 years. Starting from gel electrophoresis-based to high technology mass spectrometry (MS)-based methods, proteomics now allows the holistic investigation of diverse biological conditions and processes (Aebersold and Mann, 2016; Larance and Lamond, 2015).

The initial breakthrough for MS-based proteomics was the development of soft ionization techniques for large molecules in the late 1980s. Especially two technologies allowed the ionization and vaporization of proteins and peptides: Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ES) (Fenn et al., 1989; Karas et al., 1985; Tanaka, 1988). ES became especially popular in research because it can be easily coupled with a liquid chromatography system (LC). In ES, the analytes are volatized and ionized directly out of a solution by dispersion of the liquid into very small, charged droplets that rapidly evaporated, transferring charges to desolvated, labile analytes (Kebarle and Tang, 1993). John Fenn was awarded a share of the chemistry Nobel Prize for his invention of electrospray ionization for large molecules in 2002. Further technological breakthroughs were the combination of peptide sequence tag algorithms for the identification of peptides in DNA databases and the development of highly sensitive nanoelectrospray (Mann and Wilm, 1994; Wilm and Mann, 1996).

In principle, purified intact proteins can directly be analyzed by MS-based proteomics, a technology called ‘top-down’ proteomics (Catherman et al., 2014). However, ions with lower mass are more sensitive in MS-based analysis and intact protein measurements are not as informative, which has meant that top-down proteomics is confined to special niches such as protein drug characterization. In contrast ‘bottom-up’ proteomics has been broadly successful and is the mainstay of MS-based proteomics today. For a long time and still today proteins have been analyzed by polyacrylamide gels but it was not possible to continue from gels to MS analysis. The development of ‘in-gel digestion’ protocols enabled the efficient isolation of peptides from polyacrylamide gels and high sensitivity analysis of biological systems (*Saccharomyces cerevisiae*) (Shevchenko et al., 2006; Shevchenko et al., 1996). Combined with nanoelectrospray and peptide

sequence tags, this for the first time made mass spectrometry applicable to low level analysis of important proteins and MS has been the method of choice for their analysis ever since (Mann, 2016; Wilm et al., 1996).

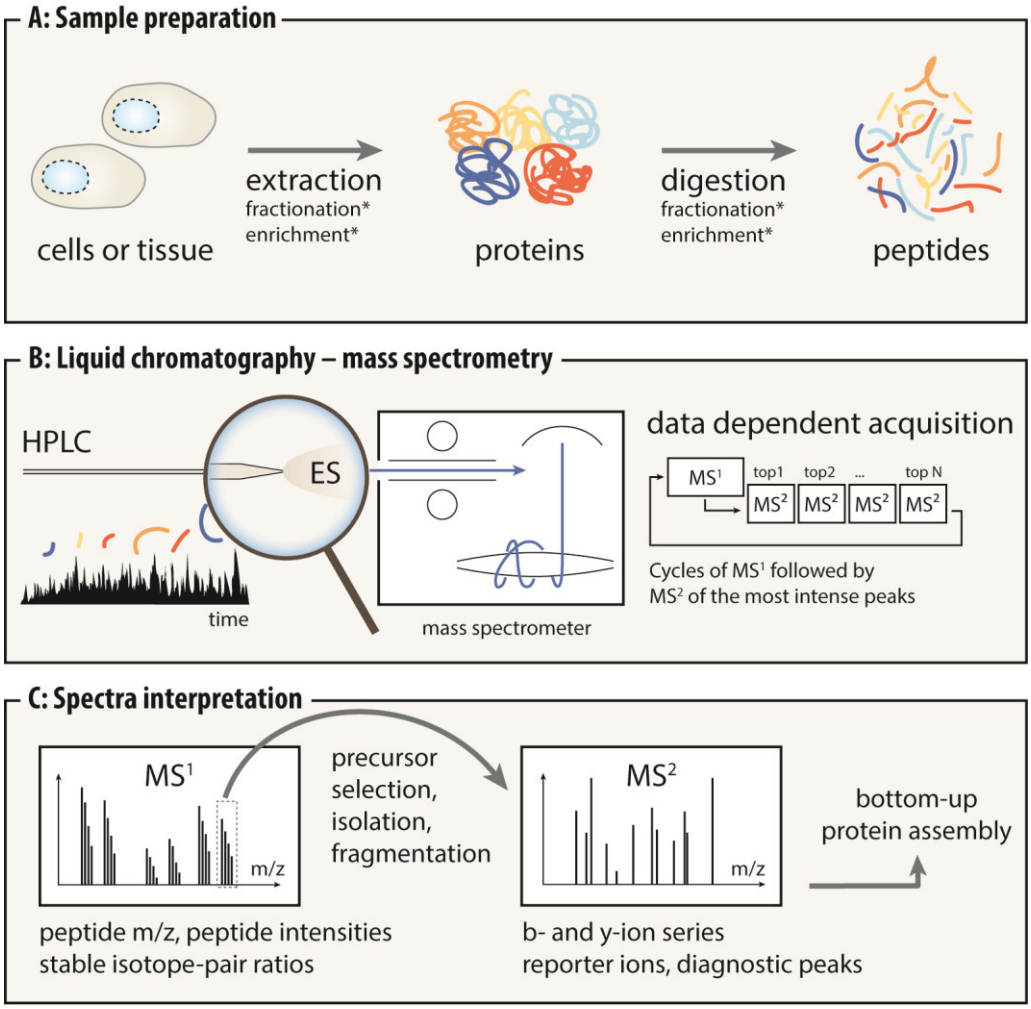


Figure 1: Shotgun proteomics workflow. (A) In the sample preparation step, proteins are extracted from cells or tissue and enzymatically digested to peptides. Fractionation can be applied at the protein or peptide levels to increase proteome coverage. **(B)** Peptides are separated by a high-performance liquid chromatography (HPLC) system and ionized by electrospray for subsequent mass spectrometry (MS) analysis. The typical schema of a data dependent top N method for data acquisition is depicted (one MS^1 scan followed by n MS^2 scans). **(C)** Bioinformatic spectra interpretation uses the information from the full MS (MS^1) and MS^2 spectra for data searching. From (Hein et al., 2013).

Over the years, sample preparation has remained a key component of proteomics (Figure 1 A). The aim of a typical sample preparation workflow is to harvest proteolytic peptides suitable for bottom-up MS. The process usually starts with the lysis of cells or tissue, followed by the reduction of intra- and inter-protein disulfide bonds. Alkylation is necessary in order to prevent the reactive thiol groups of cysteine residues from forming disulfide bridges again. The next step in generating peptides is digestion by sequence

specific enzymes. These generate predictable terminal amino acids, supplying further constraints for bioinformatics peptide identification. Trypsin and/or LysC are almost always used because they generate particularly favorable peptides for MS fragmentation and identification.

Past protocols often used strong detergents like sodium-dodecyl-sulfate (SDS) for cell lysis, which also results in the denaturation of the digestion enzymes and is in any case incompatible with ES. Protocols have been developed to remove the detergents e.g. by protein precipitation or 'Filter-Aided Sample Preparation' (FASP) (Wisniewski et al., 2009). However, apart from recurrent issues regarding reproducibility, these multi-step protocols frequently suffered from remaining detergent contamination, negatively affecting digestion efficiency. In 2014, Kulak et al. published the 'in-StageTip' digestion protocol, an all-in-one reaction buffer system for cell lysis, reduction and alkylation of cysteine residues and highly efficient digestion on the solid-phase extraction matrix (Kulak et al., 2014). Solid phase extraction then delivers clean peptides, ready for MS analysis. This protocol not only radically increased digestion efficiency but also virtually eliminated hands on time.

In the next step of the proteomic workflow, peptides are separated according to hydrophobic interactions with initially a mobile and later a stationary phase in a high-pressure liquid chromatography (HPLC) system (Figure 1 B). Peptides elute from the column in a sequential manner and are immediately ionized in the electrospray source. HPLC systems operating in the nano-flow range have proven to be especially efficient for peptide separation and the following ionization, resulting in high sensitivity.

Today's mass spectrometers are highly complex systems. They consist of a large number of components that focus the ion beam with lenses in the vacuum, effect the ions flight path by dynamic electric fields, allow to filter for ions with distinct mass to charge (m/z) ratios and break them into smaller fragments at selectable energies. The fact that an ion behavior in the vacuum is strictly dependent on its m/z can be used for identification and quantification. Many types of mass analyzers and detectors are used for this purpose.

A typical MS-measurement uses two steps (MS^1 and MS^2) to acquire the necessary information for peptide identification. In the MS^1 step (full scan), a broad-range mass spectrum (e.g. m/z=300-1,650 Th) is acquired, delivering m/z values for all intact peptide masses at a distinct time point during the LC run. In the MS^2 scan a single ion species is selected and fragmented according to an intensity-based priority list (top N method) and the fragments masses are determined. The MS^1 and MS^2 measurements deliver

the data that is used in the identification of the peptides (Figure 1 C). As peptides are combinations of amino acids with distinct masses it is in principle possible to determine their sequences. The MS^1 spectrum provides the intact peptide mass and thus constrains the peptide's amino acid composition. For all peptide sequences in the database with a compatible mass that satisfies the enzyme specificity, the MS^2 spectra are calculated. The number of matches to the measured fragments is converted to a score that reflects the likelihood that these matches occurred by chance. False Discovery Rates (FDRs) are then rigorously determined by comparison to the number of peptide and protein matches in a sequence reversed database. In this thesis, peptide and protein identification and quantification were all performed in the MaxQuant suite of computational tools (Cox and Mann, 2008; Tyanova, 2016).

The last years have seen dramatic improvements in all areas of the MS-based proteomics workflow, ranging from sample preparation to measurement and subsequent bioinformatic analysis (Aebersold and Mann, 2016; Bantscheff et al., 2012; Cox and Mann, 2011; Munoz and Heck, 2014). Together, these advances have enabled the broad application of quantitative proteomics in biological research, resulting in thousands of publications each year. Today, specialized proteomic laboratories identify quasi-complete proteomes of more than 10,000 proteins in mammalian model organisms (Bekker-Jensen et al., 2017; Geiger et al., 2012; Kulak et al., 2017; Richards et al., 2015) and apply their workflows to a diverse array of cell biological, biochemical and medical processes (Figure 2). This has also answered basic questions relating to the regulation of the proteome, including mRNA translation efficiency – in this case showing a highly protein specific regulation (Lahtvee et al., 2017; Nagaraj et al., 2011). Moreover, temporal regulation of protein expression and modification during the cell cycle (Ly et al., 2014; Olsen et al., 2010) and spatial distribution of proteins with subcellular organelle maps haven been used to investigate protein dynamics on a global scale (Andersen et al., 2005; Itzhak et al., 2016). Interaction partners of a protein of interest can be revealed by 'pull-down' experiments. The global application of this technology resulted in drafts of the human interactome, an extensive network analysis of the connections of thousands of proteins (Hein et al., 2015; Huttlin et al., 2017). Even symbiotic association can be disentangled at the proteome level, such as the one between legumes and nitrogen-fixing bacteria (Marx et al., 2016). Finally, the analysis of signaling pathways by enrichment of phosphorylated peptides allows researchers to uncover complex signal transduction pathways in vitro and in vivo (Humphrey et al., 2015).

The integration of several types of human tissue proteomes combined with data generated by the community resulted in two first 'drafts of the human proteome' (Kim et al., 2014; Wilhelm et al., 2014). Although these maps were by no means complete and their analysis methods are controversial, they illustrate the desire to determine the complete proteome as are a first step towards an understanding of the complex protein composition in the human body. More focused projects revealed cell type and region specific maps of whole mammalian organs, providing insights into biological processes in model animals and humans (Aye et al., 2010; Azimifar et al., 2014; Sharma et al., 2015).

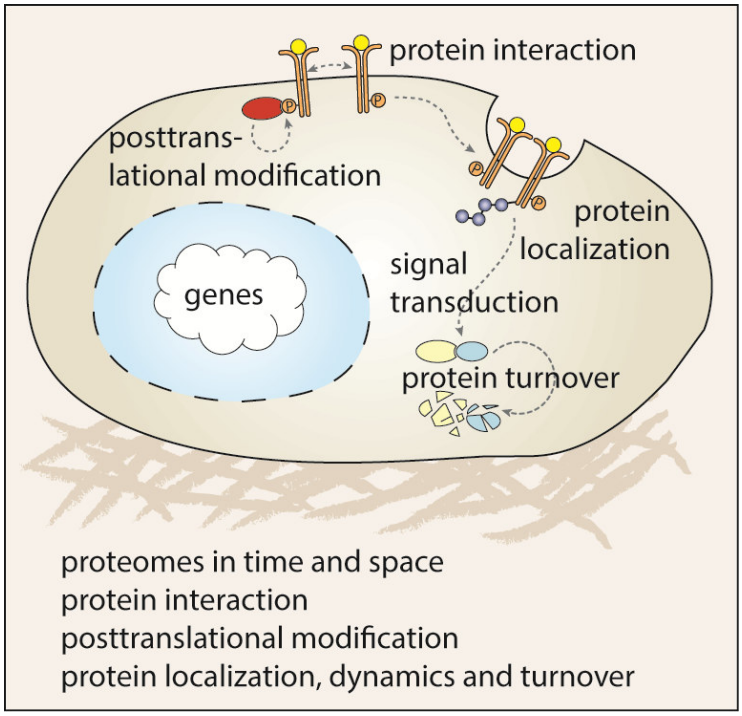


Figure 2: Proteomics exploration of a cell. Proteins in their various forms and modifications can be investigated by different proteomics techniques to explain diverse cellular processes on the molecular level. From (Hein et al., 2013).

1.2. Clinical proteomics

Precision medicine is a key aim of modern medical science, directly connected to individualized treatment and disease prevention (Collins and Varmus, 2015). It is driven by the idea that inter-individual biological variances determines the differences in disease presentation and subsequent response to treatment. Large scale technologies that have the power to differentiate between individuals such as genomics, transcriptomics and proteomics all have the promise to personalize medicine. The proteome is the most direct molecular representation of the phenotype, levels of

individual proteins are already widely applied as indicators of diseases in clinical practice and most drugs influence the activities or concentrations of proteins. Therefore, proteomics should in principle be an ideal technology to investigate disease mechanisms. Clinical proteomics could be used in a multi-faceted manner to deliver on the promises of personalized medicine: diagnosing diseases in early stages, correlating protein patterns for disease sub-classification, predicting disease progression and finding causal molecular targets for new treatment strategies.

Research groups around the globe strive to address the diverse medical needs of society. Some clinical questions can be readily answered using cell line systems. In this context, proteomics has been successfully applied to disentangle the respective mechanisms of actions of drug treatments, see for example (Sacco et al., 2016), and it has also been used to find off-target effects of therapeutics (Bantscheff et al., 2007; Klaeger et al., 2016). Only a minority of diseases have obvious causal mechanisms such as monogenetic disorders; rather pathogenicity usually depends on the accumulation of multiple diverse epigenetic and environmental factors. Even in cancer, where the underlying defect is gene mutation, there is complexity in the cumulative tumor heterogeneity. Cells within a single tumor can exhibit diverse mutations due to clonal evolution. Further heterogeneity comes from the immediate microenvironment level (blood vessels, interaction with stroma, nutrients), while there is a nearly infinite variability at the level of the host that may influence tumor progression (immune response, microbial response, age of host, environment exposure). Thus model organisms can only help in providing a generalized, simplified overview. Investigation of diseased tissues or body fluids of individuals can shed light on protein-based molecular mechanisms and proteomics has already been successfully applied to the investigation of tumor samples. Disease specific patterns have been identified and further sub-stratification of individuals within one disease has been achieved, see for example (Deeb et al., 2015; Mertins et al., 2016; Tyanova et al., 2016; Zhang et al., 2014).

Tissues are generally only available for diseases where surgery is a necessary treatment step. In other cases, they have to be obtained post mortem. In contrast, body fluids like saliva, urine, stool and tears are sampled non-invasively or in the case of blood by minimally invasive procedures. Evidently, they represent a unique opportunity in terms of potential clinical utility and research potential. As a consequence, there are clinically established tests for all of them. Blood and blood derived matrices like plasma and serum (collectively referred to a ‘plasma in this thesis) are by far the most important ones for diagnostic purposes and will be discussed separately below.

In the past decades, the proteomics field has collectively endeavored to search for new biomarkers in body fluid proteomes. For instance, in stool samples chemical and immunological tests are used to detect blood in the context of colorectal cancer screening (Rex et al., 2009). Stool mainly consists of bacteria and this has become the focus of much current research. The microbiome has been investigated in a wide range of conditions and diseases by next-generation sequencing (NGS) methods and these analyses clearly reveal the profound influence of the microbiome in many diseases (Lynch and Pedersen, 2016). A recent proteomic study in human and mice achieved high proteome coverage of more than 30,000 microbial and host proteins in mice and 19,000 in humans (Zhang et al., 2016). In our group we investigated the microbial community of human saliva by proteomics, quantifying 5,500 human and 2,000 microbial proteins. We found drastic remodeling of the microbiome in response to food intake and tooth brushing (Grassl et al., 2016). In the clinic, saliva samples are routinely tested for a broad range of diseases like HIV or helicobacter pylori (Malamud, 2011). A broad range of analytes, especially small molecules, are determined in urine, but total protein levels are also tested to detect increased glomerulus permeability e.g. in infectious diseases, diabetes, hypertension and general kidney malfunction. Human chorionic gonadotropin (hCG) is the commonly detected substance in pregnancy tests, which is also performed in urine. The proteomic community has investigated the urine proteome extensively, achieving a depth of nearly 3,500 proteins (Santucci et al., 2015). Tear fluid is also interesting for diagnosis and proteomic studies report more than 1,500 identified proteins (Aass et al., 2015).

In the past and present, plasma forms the basis of standard clinical diagnosis and this will in all likelihood continue in the future. In the realm of proteomics, plasma has also emerged as a center of attention. This fact is clearly reflected in a comparison of the collective number of publications regarding urine, stool, saliva and tears vs. those that investigate either blood, plasma and serum. As of May 2017 this ratio stands at 1,500 to 7,700. In light of this, it is unfortunate that, due to a variety of technical and conceptual shortcomings, the exploration of the human plasma proteome has proven to be somewhat of a disappointment, with essentially no proteomics-derived biomarker having been integrated into clinical practice.

1.3. The blood plasma proteome

Blood is considered the foremost bodily fluid and around 5 L are circulating in the human body. It serves as the medium through which a vast array of functions is executed:

oxygen and nutrients are provided, metabolites are carried and removed, signaling molecules are transported for inter-organ communication, body temperature is regulated and pathogens are fought by the immune system.

Blood is a suspension, consisting of a cellular (~40%) and a liquid component (~60%) (Fischbach, 2009). Its cellular portion can be classified into erythrocytes, thrombocytes and leucocytes. Erythrocytes are the most abundant ones (~5*10^6 cells/µL), responsible for the transport of oxygen and for pH buffering. Thrombocytes (~1-4*10^5 cells/µL) are the protagonists of haemostasis, initiating repair upon injuries. Leucocytes (5-10*10^3 cells/µL) constitute a broad class of immune cells of which granulocytes and monocytes are responsible for the unspecific and lymphocytes (B cells, T cells, NK cells) for the specific immune response.

The straw-coloured liquid portion of blood is called plasma, in which all components are retained whereas serum remains after activation of the coagulation cascade. In our experience, serum and plasma are equally suited to proteomic analysis.

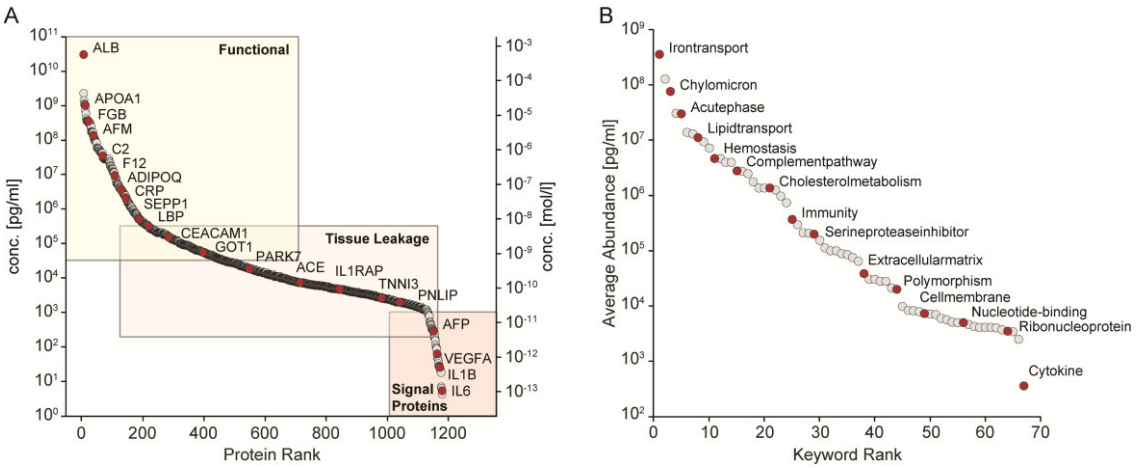


Figure 3: Functional annotation of the plasma proteome. (A) Plasma proteins are spread over a dynamic range of at least ten orders of magnitude. Typical serum proteins are annotated in the abundance plot. The three boxes reflect classification into functional, tissue leakage and signal proteins according to the proteins and their keyword annotation from Anderson (Anderson and Anderson, 2002). **(B)** Keyword annotation and one-dimensional enrichment analysis provide a functional reflection of the plasma proteome based on bioinformatic analysis (Cox and Mann, 2012). Protein concentrations were derived from the Plasma Proteome Database (Nanjappa et al., 2014). Adapted from (Geyer et al., 2017).

The plasma proteome – the entirety of all proteins present in plasma – can be categorized into three general classes based on functionality (Anderson and Anderson, 2002; Surinova et al., 2011): highly abundant proteins with specific roles in plasma, medium abundant tissue leakage proteins with no dedicated purpose in plasma, and low abundant signalling proteins (Figure 3 A). The concentration difference between the most abundant protein serum albumin (ALB) at around 50 mg/mL and the lowest

concentrated cytokines e.g. interleukin 1 beta (IL-1β) with less than 5 pg/mL, results in a dynamic range spanning more than ten orders of magnitude. Note that the categorization into abundance and functional classes is only approximate; for instance, there are very low abundance tissue leakage proteins with no functional role.

In the high abundant class, albumin maintains the osmotic pressure, the apolipoprotein family transports insoluble molecules such as lipids, haptoglobin sequesters free haemoglobin that would otherwise harm the kidneys, serotransferrin recycles free iron, acute phase proteins defend the body against pathogens and the proteins of the coagulation cascade initiate wound healing. Tissue leakage proteins may be released by shedding into the circulation such as the apolipoprotein receptors SRB1, LRP1 and LDLR from the liver or by tissue damage like prostate specific antigen (PSA) (Vihko et al., 1978), which is elevated in prostate cancer patients. Likewise, increased levels of the cardiac muscle troponin T (TNNT2) may indicate a myocardial infarction (Hamm et al., 1992). The third class consists of messenger molecules like small protein or peptide hormones (e.g. insulin or ghrelin) and cytokines, which typically have very low abundances at steady state and are upregulated on demand.

The diverse functions of the plasma proteome, distributed over the entire concentration range, are displayed in figure 3 B. Keyword annotations of a list of 1,176 proteins and a subsequent ‘one-dimensional enrichment analysis’ (Cox and Mann, 2012) identified 67 significantly enriched terms. These cover only keywords that are connected to multiple proteins and thus even underestimate the functional complexity of the plasma proteome. For example, copper transport is executed by a single protein (ceruloplasmin), whose function is not enriched in such an analysis.

1.4. Biomarkers and the clinical plasma proteome

According to the US National Institutes of Health (NIH) Biomarkers Definitions Working Group, a biomarker is a defined characteristic that can be quantified as an indicator of a normal biological process, pathogenic process, or a response to an exposure or intervention (Biomarkers Definitions Working, 2001). The BEST resource (FDA-NIH:Biomarker-Working-Group, 2016) of the American Food and Drug Administration (FDA) classifies biomarkers into seven categories (Figure 4 A).

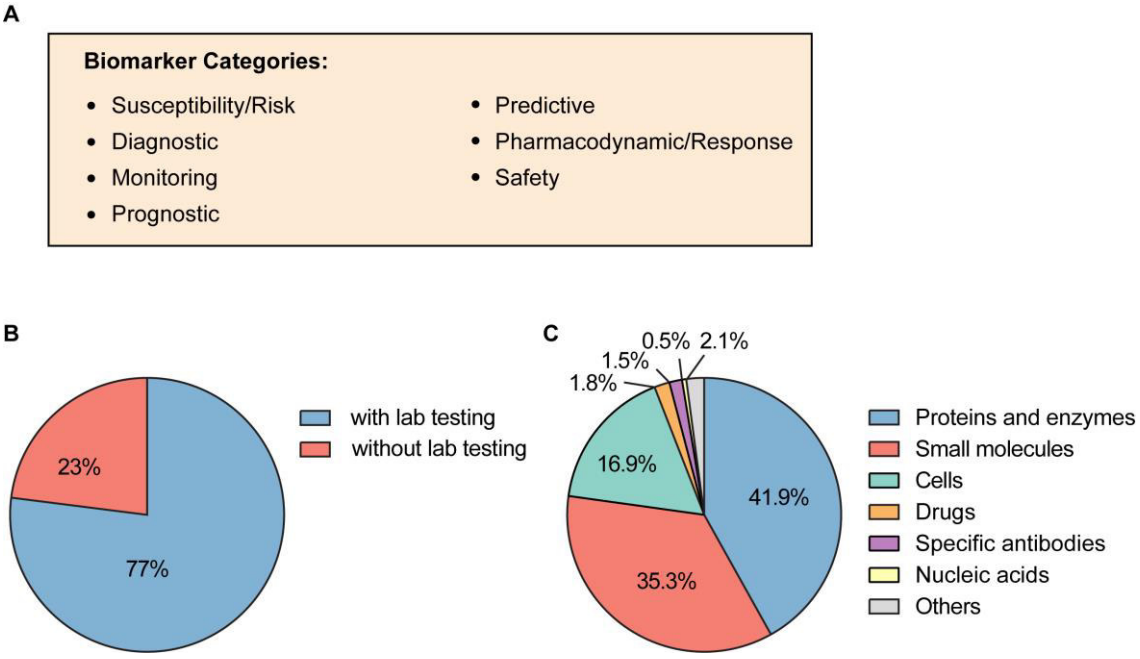


Figure 4: Biomarkers and their clinical application. (A) Biomarkers are divided by the FDA into the seven indicated categories according to the BEST resource (2016). **(B)** Proportion of clinical decisions that are made based on laboratory testing. **(C)** Proportion of clinical tests that are based on different molecule types. Adapted from (Geyer et al., 2017).

Biomarkers have a profound role in clinical decision making. According to a survey of our collaborators at the Institute of Laboratory Medicine at the Klinikum Großhadern – one of the largest University Hospitals in Germany – 77% of all clinical decisions are based on laboratory tests (Figure 4 B) (Geyer et al., 2017). The largest group of these (42%) measures the concentrations or enzymatic activities of proteins (Figure 4 C). In daily clinical practice, the quantitative analysis of individual plasma proteins is overwhelmingly performed with immuno- or enzymatic-assays that target single proteins. This is because these tests have inherent limitations regarding multiplexing and antigen-antibody recognition. Such limitations include cross-reactivity, non-linear responses (Hook effect) and interference by background molecules such as triglycerides (Hoofnagle and Wener, 2009; Wild, 2013). Furthermore, there are a plethora of clinically important protein variants that are difficult to detect by antibody-based assays. One example is apolipoprotein(a), a marker for the assessment of cardiovascular disease risk (Danesh et al., 2000). Apolipoprotein(a) contains a number of kringle IV domains that is genetically determined. These affect the structure of the protein and the affinity of the antibody towards it (McConnell et al., 2014). Another example is vitamin D binding protein of which there are three common isoforms in humans, each reacting differently in clinical immunoassays. A frequent polymorphism in African Americans has resulted in the underestimation of vitamin D binding protein levels and in the mistaken notion that

African Americans have lower concentrations of this proteins in general (Powe et al., 2013).

To date, the concept of protein biomarker discovery and measurement is generally synonymous with single protein tests, with the unstated implication that there should be a biomarker for each disease. However, this notion suffers from an inherent conceptual limitation: there are only about 20,000 human genes and the number of different human diseases is nearly as large – 14,400 according to the International Classification of Diseases (ICD). Biologically, it appears unlikely that a distinct protein-based biomarker exists for each and every disease (and it would already be arithmetically impossible). Even adding non-protein compounds such as metabolites, would not change the situation appreciably. A more promising concept would be to combine proteins into 'multi-biomarker panels'. This generates a very large number of degrees of freedom – many more than the number of different disease. Even a small panel consisting of five arbitrary proteins with binary states would result in about $20{,}000^5 = 3*10^{21}$ possible combinations. Apart from potentially enabling many more potential patterns than single protein assays, multi-biomarker panels could also more readily account for inter-individual variability. For example, one of the studies described in this thesis defines a multi-protein inflammation panel consisting of 10 proteins that reflects low level inflammation in the body (Geyer et al., 2016a). Interestingly, in clinical practice there are some examples of multi parameter diagnostic scores like the sFlt-1/PIGF ratio for the diagnosis of preeclampsia (Levine et al, 2004) or the integration of albumin, bilirubin, quick test, ascites and encephalopathy into the Child-Pugh-score for liver cirrhosis (Pugh et al., 1973).

Currently, there are only about 100 FDA cleared or approved clinical plasma or serum tests available. Furthermore, more than 80% of these have been implemented more than 20 years ago. In the past two decades, the rate of discovery of new biomarkers has remained constant or even declined, with less than two new biomarkers incorporated into clinical practice per year (Anderson, 2010; Geyer et al., 2017).

Given the fundamental limitations of individual protein assays, MS-based methods are in principle an attractive alternative for clinical applications as well as biomarker research as they are inherently capable of discovering multi-protein panels. However, the technological challenges are daunting and call for drastic improvements in robustness, sensitivity and throughput compared to what is available at the moment.

1.4.1. Challenges of plasma proteomics

1.4.1.1. Pitfalls of the past

Given the attractiveness of plasma proteomics, many research groups around the globe have attempted to mine the human plasma proteome in search of new biomarkers over the last decades. Unfortunately, despite these individual efforts and those of the Human Proteome Organization's Plasma Proteome Project (Omenn et al., 2005), this major goal of our community has not fulfilled its initial promises. In retrospect, this is clearly due to several intractable technological challenges, which were not sufficiently addressed at the time.

On the biomarker side, the only case in which plasma proteomics was partially successful was the OVA1 test (Rai et al., 2002; Zhang et al., 2004). OVA1 consists of a five protein panel used to distinguish between benign and malignant ovarian tumors in very specific indications. Four of the proteins are the highly abundant plasma proteins beta-2 macroglobulin, apolipoprotein A1, serotransferrin and pre-albumin. They were identified by proteomics but are very unlikely to be specific to ovarian cancer status. Their levels are combined with the already known biomarker CA125 and the patient´s menopausal status into a risk score.

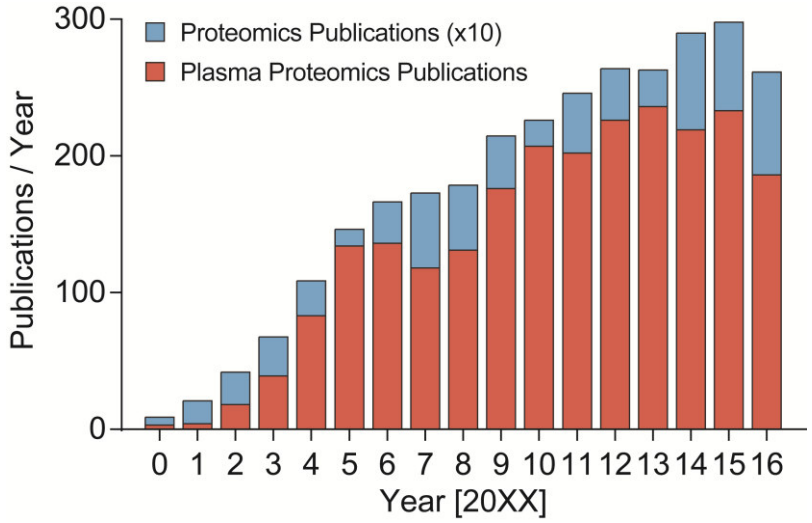


Figure 5: Literature review of plasma proteomics. The total number of publications using MS-based proteomics are more than 10-times higher than the plasma proteomics literature (search terms: [proteomics AND mass spectrometry]; [plasma AND proteomics AND mass spectrometry]). The low level and fluctuations in the number of publications per year for plasma proteomic are in stark contrast to the steadily increase in MS-based proteomic publications in general.

In reviewing the history of plasma proteomics, we found that there have been two periods of particular activity towards the discovery of new biomarkers. This is reflected in spikes of publication numbers in plasma proteomics compared to the total number of publications in the field of proteomics (Figure 5). Disregarding an initial phase in which two dimensional gel electrophoresis was employed, the first period started already 2000 with a publication peak in 2006. This included claims of early cancer detection on the basis of very low resolution MALDI spectra using 'serum patterns', rather than actual protein identifications (Petricoin et al., 2002). It was later shown that biases in the experimental procedures were responsible for the claimed classification success and this resulted in a severe setback for plasma proteomics (Baggerly et al., 2004). The next period with increasing numbers of publications extends from 2009 to 2013, followed by stagnation, presumably due to the fact that no new biomarkers had been discovered. The number of publications remains low and even dropped to a minimum in 2016. This becomes even more remarkable against the backdrop of an ever expanding community of researchers using proteomics and their steadily increasing output of publications. Today, relatively few groups continue to pursue plasma proteomics, despite the undiminished medical need for new biomarkers and the success of MS-based proteomics in other areas. This raises the question of what holds back the field of MS-based plasma proteomics.

1.4.1.2. Technological limitations

Finding new biomarker requires high samples throughput to obtain statistically robust results. However, in current MS-based proteomics, the preparation of peptides from a biological sample typically requires more than 24h and long, 2-4h gradients are usually employed. Furthermore, sample preparation workflows are not standardized, much less over a period of years. Plasma contains lipids and other small molecules that act as impurities or contaminants in proteomic workflows, if not removed. This can result in clogging of the HPLC columns that are coupled online to the MS as well as in frequent cleaning of the instruments. Together, this has made plasma proteomics very time consuming and expensive. Clearly, proteomics based biomarker research requires a robust, highly reproducible and ideally automatable workflow. Such a workflow should allow the preparation of large numbers of samples in a short time and their highly reproducible measurement, without down time of the instrumentation.

Plasma is generally considered to be the most complex of all body tissues for proteomic analysis, due to its high dynamic range of at least ten orders of magnitude combined

with the need for very high sensitivity. This restricts analysis of the plasma proteome to about six to seven orders of magnitudes with current state of the art instruments. As LC-MS/MS is based on peptides that are separated by a gradient and that elute in a time-ordered manner from the column, co-elution and electrospray ionization of very highly abundant and low abundant peptides decreases the probability to detect the low abundant ones. In plasma, the 41 tryptic peptides of serum albumin or the 312 peptides of apolipoprotein B (fully tryptic peptides with 7-30 amino acids) present particular challenges because of their extreme abundances and high numbers. On Orbitrap analyzers, in particular, the space charge limit of the ion trap can be almost completely taken up by such abundant peptides in a very short time (< 1 ms), 'crowding out' the low abundant ones.

1.4.1.3. Cohort intrinsic problems

Another problem for biomarker discovery is the fact that the levels of plasma proteins can be individual-specific. This can be genetically determined, for instance the concentration of the above-mentioned apolipoprotein(a) decreases with increasing numbers of kringle IV domains and the levels of pregnancy zone protein (PZP) are gender specific (Christensen et al., 1989; Utermann, 1989). Despite its potential impact, this issue has rarely been recognized by the community before being recently addressed by MS-based proteomics (Geyer et al., 2016a; Geyer et al., 2016b; Liu et al., 2015). Another, often neglected issue in clinical studies is the sample quality (Hassis et al., 2015; Kaisar et al., 2016). Samples may be collected by medical doctors that are under constant time pressure and whose primary aim is to take care of patients. Quality cannot always be guaranteed under such circumstances, calling for markers to identify problematic samples.

The very design of proteomics based biomarker studies can also be an issue. Our literature search revealed that only 47% of the studies had any kind of validation of the results in the discovery phase (Figure 6 A). In half of the cases the follow up experiments were simple western blots or immunoassays of candidate proteins performed with the same sample rather than an independent cohort. Moreover, in 30% of all studies, cases and controls were pooled (Figure 6 B). This is usually a consequence of the low throughput of the workflow employed but is justified by the argument that it will reduce individual specific differences (Weinkauf et al., 2006). However, proteins such as pregnancy zone protein or the clinically important C-reactive protein (CRP) can be 10,000 fold increased in single individuals, skewing the levels in the entire pool. I found

similar concentration differences for quality marker like carbonic anhydrase 1 (CA1), indicating erythrocyte lysis (Geyer et al., 2016a). Even more problematic, several studies in our literature search reported these quality markers as potential biomarkers. Furthermore, pooling can remove subgroup specific effects in a cohort, which are by definition the basis of personalized medicine. For example, the discovery that HER2/neu was expressed in just 30% of women with breast cancer enabled the therapeutic antibody Herceptin to pass all clinical phases, whereas a pooling strategy might have denied patients this lifesaving therapy (Ullrich et al., 1984).

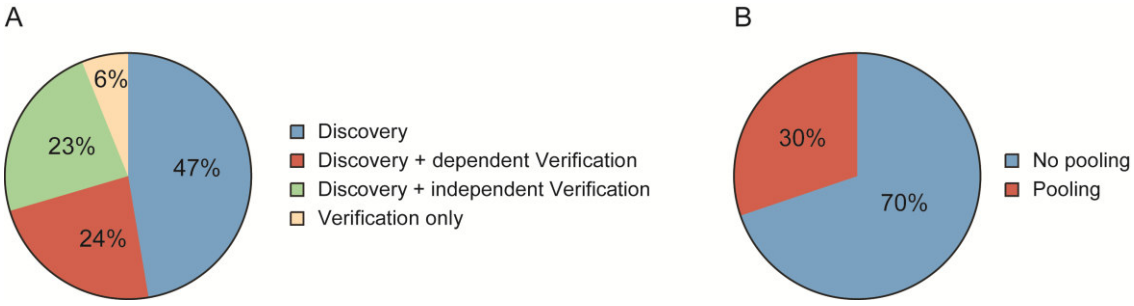


Figure 6: Literature review. (A) Pie chart of the proportion of studies, using discovery and validation phases. **(B)** Percentage of studies investigating pooled samples.

1.4.1.4. Traditional plasma proteomic workflows

Proteomic researchers are accustomed to large numbers: in a standard preparation of cancer cell lines like HeLa, it is readily possible to identify more than 40,000 peptides, corresponding to more than 4,000 proteins using 2h gradients on quadrupole-Orbitrap instruments (Q Exactive HF). This can even be increased to more than 10,000 proteins and 100,000 peptides with more elaborate workflows (Bekker-Jensen et al., 2017; Geiger et al., 2012; Kulak et al., 2017). In stark contrast, we could only detect around 2,000 peptides and 200 proteins with very similar workflows and more elaborate versions did not drastically improve those numbers. This was mainly due to the extreme dynamic range in conjunction with sensitivity challenges as mentioned above.

To partially overcome this challenge, researchers have applied very extensive fractionation and depletion of the most abundant plasma proteins. Sample pre-fractionation can easily be implemented, but decreases throughput and reproducibility, drawbacks that are especially problematic for biomarker studies. The aim of depletion is to remove high abundance proteins from plasma and thereby to enrich the lower abundant and potential more interesting ones. There are two common strategies: The

first is based on 'ProteoMiner' hexa-peptides that are immobilized on beads (Thulasiraman et al., 2005). Proteins bind to the hexa-peptides with different probabilities, partially 'randomizing' the plasma proteome. The second strategy uses bead-immobilized antibodies against the most abundant plasma proteins. Target proteins bind to the antibodies and the unbound portion can be collected and analyzed. Different vendors sell depletion kits for the highest 1, 2, 6, 12, 14 or even 20 proteins (called top X depletion). Using a combination of immunodepletion and extensive fractionation has led to the identification of more than 1,000 (Addona et al., 2011; Cao et al., 2012; Paczesny et al., 2010) or even more than 5,000 proteins in plasma (Keshishian et al., 2015). The latter number was achieved with so-called 'supermix depletion', which pushes this technique to its extreme (Qian et al., 2008). The polyclonal antibody mixtures used in chromatographic supermix depletion are generated by immunizing hens with top 14 depleted human plasma and subsequently purifying IgY antibodies from eggs.

Although attractive in principle, depletion suffers from unspecific removal of proteins cross-reacting with the antibody targets or sticking to the chromatographic material. This problem can be illustrated by comparing our deepest dataset from undepleted plasma samples of a cohort of more than 40 individuals to a much used plasma dataset with supermix depletion of four individuals (Keshishian et al., 2015). Notably, this resulted in poor correlation over the entire abundance range with an R^2 value of only 0.23. Proteins were separated into two clouds, the lower of which is presumably caused by unintended 'off target' depletion. There are even proteomic researchers who endeavor to identify the hundreds of proteins bound to albumin – the 'albuminome' – and who also discuss the effect of albumin depletion on the plasma proteome (Gundry et al., 2007; Gundry et al., 2009; Holewinski et al., 2013; Lowenthal et al., 2005). Moreover, depletion columns can be very expensive (2,000-31,000 €) and are only intended to be used for up to 200 depletions. Their efficiency will also change uncontrollable over time, making the reproducible analysis of large cohorts very difficult. That said, applying depletion strategies very carefully to a strictly controlled set of samples can be a suitable means to reach a sufficient proteome coverage for biomarker discovery (Keshishian et al., 2015; Li et al., 2013), however, results should be verified and validated in independent cohorts (Rifai et al., 2006). In this thesis, depletion is used only for the generation for 'plasma peptide libraries'.

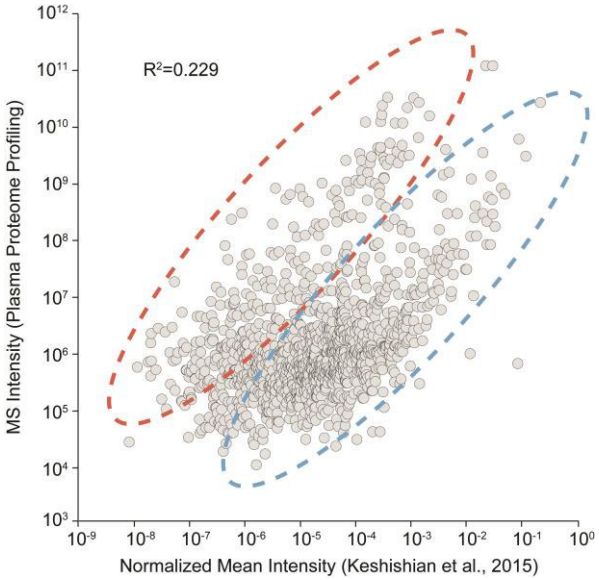


Figure 7: Correlation of an undepleted to a depleted plasma proteome. Correlation of a dataset resulting from supermix depletion of four individuals (Keshishian et al., 2015), to our deepest, quantitative Plasma Proteome Profiling dataset of 47 study participants. This resulted in two populations of proteins, of which the red population was decreased in the supermix depletion dataset, presumably due to unintended de-enrichment.

A survey of the literature revealed that biomarker research has so far focused on areas that reflect diagnostic interests of the medical community rather than current technological possibilities of plasma proteomics: About one third of all publications deal with cancer, followed by cardiovascular diseases, topics in human physiology, inflammation, diabetes and Alzheimer´s disease. Even with a combination of a supermix depletion strategy with extensive fractionation, it is still questionable if the very high proteome coverage and sensitivity that would be required for some of these diseases could be reached.

The decrease in throughput inherent in fractionation can partially be recovered by multiplexing. After digestion, peptides can be chemically labeled with isobaric tags such as iTRAQ or TMT (Bantscheff et al., 2008). The tags are constructed such that they add to the same total mass but give rise to different low mass reporter ions. Generally, between four and ten samples can be combined with such a strategy. Quantification is achieved by fragmenting the peptides and quantifying the relative ratios of low mass reporter ions. To date, a major disadvantage of these techniques is the 'ratio compression', the distortion of the peptide ratios caused by co-isolated peptide species that contribute to the same reporter ion. In principle this can be addressed by more elaborate scan modes such as MS3 (Ting et al., 2011), but currently at the cost of speed and sensitivity.

Partly as a consequence of the demands on instrument time, rarely more than 30 plasma samples have been analyzed at a time and only 5 studies exceeded more than 500. Thus the number of proteins that are measured by proteomics results in a severe challenge of multiple hypothesis testing, which becomes the more problematic that the protein numbers exceed the sample numbers. As a consequence, most studies only report 'potential biomarkers'. Rigorous follow up experiments would be required to confirm these potential biomarkers in independent cohorts. However, usually the only verification has been the re-measurement of the same cohort by another technological platform like MRMs or immuno-assays.

In summary, technological limitations, unawareness of potential pitfalls and issues in study designs have all contributed to prevent the identification of true biomarkers so far.

1.4.2. The 'triangular strategy' for biomarker research

By its nature, MS-based discovery proteomics is a hypothesis free approach with no assumptions regarding the origin or identity of possible biomarker candidates. This is in contrast to the analysis of single proteins by immunoassays or targeted proteomics, which are always hypothesis driven. Therefore, in principle MS-based proteomics should be an ideal tool for the discovery of novel biomarkers. In reality, however, the above-mentioned challenges have so far prevented the identification and validation of biomarkers by proteomics.

As mentioned, the low sample throughput in relation to the number of quantified proteins has resulted in a division of the biomarker research process into several steps. The resulting 'triangular strategy' is generally accepted as the gold standard for biomarker discovery (Rifai et al., 2006). In this strategy, the number of individuals increases over three study phases from just a few to several hundreds, whereas the number of investigated proteins decreases from up to several thousands to one or a few proteins. This results in a triangular shape for the numbers of study participants and an inverted triangle for the number of proteins (Figure 8).

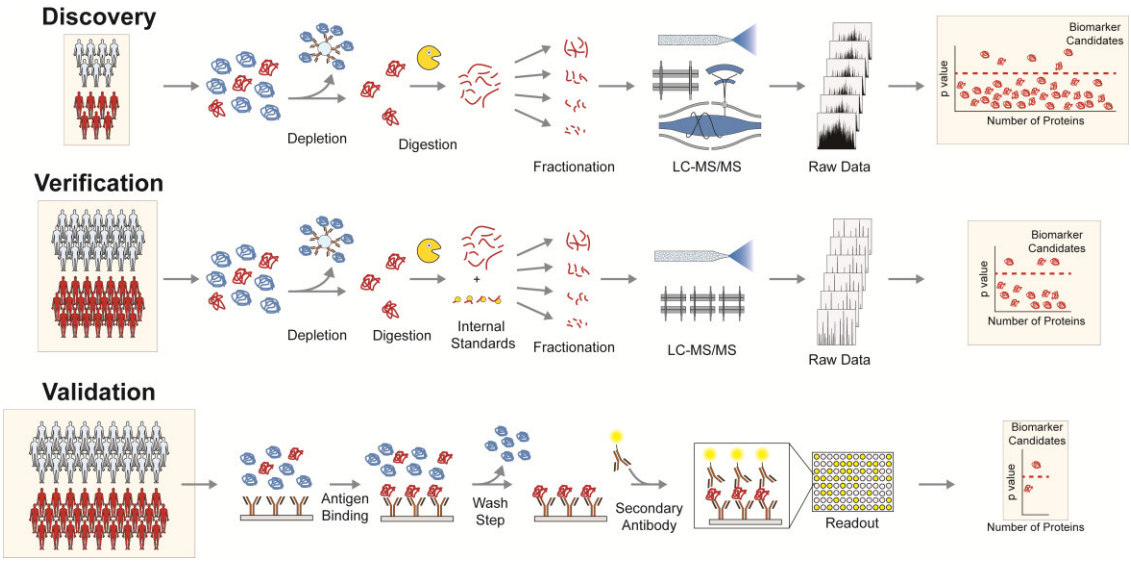


Figure 8: Triangular workflow for biomarker discovery. The triangular strategy is based on three phases with increasing numbers of samples and decreasing numbers of investigated proteins over the different stages. In the discovery phase, plasma of a small cohort is harvested and typically depleted of the highest abundant proteins (blue). The remaining proteins (red) are digested to peptides, which are optionally labeled with isobaric tags for multiplexing. Peptides of different individuals labeled with unique heavy isobaric tags are combined and fractionated. Each fraction is separately measured by LC-MS/MS. Next, the raw data are processed and analyzed to find new biomarker candidates. In the verification phase, targeted proteomics is applied. Ideally, heavy labeled peptides of the targeted proteins are added as internal standards for absolute quantification. Fractionation is applied to quantify low abundant proteins but multiplies the number of required measurements. Triple quadrupole MS are the typical MS instruments for targeted analysis. In the validation phase, one or a small number of biomarker candidates are screened by immuno-assays against individual proteins in a large cohort. In enzyme-linked immunosorbent assays (ELISA), antibodies bind the candidates (red) and non-bound proteins (blue) are removed. A secondary antibody linked to a reporter-fluorophore is used in a sandwich configuration for specific quantification. Proteins that significantly discriminate between cases and controls would be considered as true biomarkers.

In the first – discovery – phase, shotgun proteomics is applied using the above mentioned strategies with low sample throughput and with a proteomic coverage that is as high as possible. Typically, this stage results in a smaller number of proteins (~10s) that are termed as ‘potential biomarker’ or ‘biomarker candidates’, which refers to the need of further evaluation in the second – verification – phase, preferably in a larger and independent cohort compared to the discovery phase. This stage can also be done by MS, however, employing medium-throughput and targeted techniques such as multiple reaction monitoring (MRM), (Carr et al., 2014; Ebhardt et al., 2015). In MRM, one or more unique peptides for each biomarker candidate are selected and their characteristics like retention time and optimal fragmentation energy are experimentally determined to establish the MRM assay. In principle, as the MS is only focusing on a small number of peptides, high sensitivity and accurate quantification can be achieved with less extensive sample preparation steps, resulting in higher throughput. Even

though inter-laboratory studies have achieved good reproducibility in proof-of-principle studies (Abbatiello et al., 2015; Addona et al., 2009), the reported sensitivities do not reach the low ng/mL concentration range and demonstrated multiplexing capabilities are typically less than 50 peptides (Oberbach et al., 2014; Percy et al., 2013; Shi et al., 2013; Wu et al., 2015). Absolute quantification of individual proteins is preferable to relative quantification and this can be achieved in a highly accurate manner with internal standards. For this purpose heavy isotopically labeled, synthesized peptides are typically used. Even more accurate would be the addition of recombinant expressed proteins (SILAC-PrESTs) to the sample before digestion to control for variations during the complete workflow from adding the first buffers to the sample until the MS measurement (Edfors et al., 2014; Geyer et al., 2016a; Zeiler et al., 2012). The last step in the triangular strategy is the validation phase and its purpose is the further evaluation of the biomarker candidates that have passed the previous stages. The great advantage of immunoassay in this phase is their high throughput combined with high sensitivity, which enables testing candidates in hundreds or even thousands of samples. However, establishing specific immunoassays is time-consuming and far from trivial. Note that this 'gold standard triangular strategy', is quite demanding and that there are few if any examples, where it has been applied in its entirety.

The lack of success in finding new biomarkers resulted in many recommendations for proper study design, quality standards, workflows and evaluation of results (Hoofnagle et al., 2016; Luque-Garcia and Neubert, 2007; Mischak et al., 2010; Parker and Borchers, 2014; Paulovich et al., 2008; Skates et al., 2013; Surinova et al., 2011). However, this just serves to underline the fact that there are still no validated plasma biomarkers that had been discovered by proteomics.

1.4.3. Plasma Proteome Profiling

1.4.3.1. The Concept

The technological developments in our departments before and during this PhD thesis enabled the development of a new concept, which we termed 'Plasma Proteome Profiling', a novel way to attack the plasma proteome with a systems-wide view. Our primarily aim is not necessarily to find new biomarker per se, but to establish a powerful way for deep phenotyping of humans. With Plasma Proteome Profiling, we wish to obtain a better understanding of human biology, starting from basic questions like how the plasma proteome responds to different environmental influences such as simple life

style changes and continuing on to more complex processes, for example disease progression or the response to a treatment. Our principal strategy is to gather as much information on as many proteins over as many conditions as possible. Apart from reviving biomarker research, a knowledgebase that integrates all this information would have very broad applications – ranging from the selection of optimal lifestyle changes to monitoring the effectiveness of medical interventions. Storing information about protein changes in response to widely different circumstances would also help to evaluate biomarker candidates of any particular clinical study. Below we describe how some of our proteomics quality panels could have helped to discard biomarker candidates already in the discovery phase of other studies. This would have eliminated time consuming and costly follow up. In my own work, I was able to combine the results from two studies to differentiate between the effects of caloric restriction and bariatric surgery induced weight loss and to address one of the major questions in the field of human metabolism from a novel angle (Albrechtsen et al., 2017).

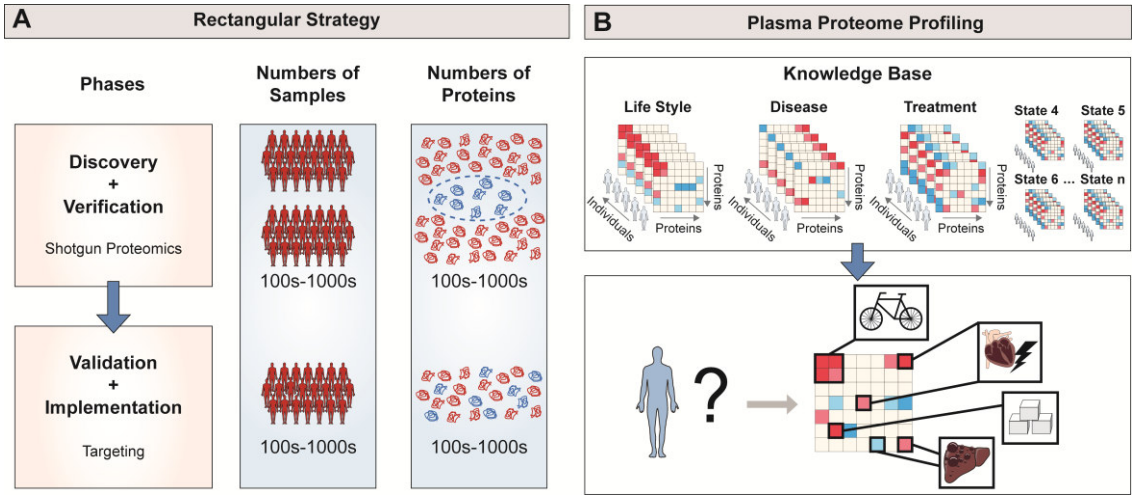


Figure 9: A 'rectangular strategy' evolved from the Plasma Proteome Profiling concept. (A) In the rectangular strategy, many individuals will be screened by Plasma Proteome Profiling, resulting in quantitative information about a large number of proteins. Two independent cohorts are measured and potential biomarkers must be significant in both of them. In the validation and implementation phase. The biomarker candidates can then be validated in yet another cohort, and clinically implemented either in the same form as in the discovery phase or with spike-in internal standards (SILAC-PrESTs). **(B)** Plasma Proteome Profiling aims at the high throughput screening of as many proteins and in as many conditions as possible in large studies. This would result in a large 'knowledge-base' with quantitative information – ideally about all proteins and conditions. Data mining can be used to interpret an individual's Plasma Proteome Profile and disentangle the possible influences that add up to his or her health and disease state (a human phenotype). Adapted from (Geyer et al., 2017).

This concept of deep phenotyping humans is in stark contrast to the current gold standard for biomarker discovery or plasma proteomics described above. One of the

main requirements to asses many conditions is to create a high throughput plasma proteomics pipeline. Such a pipeline must be very robust and reproducible to deliver highly accurate and valuable information. The combination of very robust and accurate measurements with a deep proteome coverage would make such a strategy very interesting for biomarker discovery and could even result in a change of the current paradigm for finding biomarkers.

There is a similarity to other technologies that evolved from low to high throughput workflows such as genome-wide association studies (GWAS). Because of the high cost of genotyping, which necessitates determination of thousands of potential genetic markers in thousands of subjects, researchers traditinally followed a workflow similar to the triangular one in proteomic biomarker studies (Satagopan et al., 2002; Thomas et al., 2004). This involved genotyping a few samples on many markers at first, followed by validation of a smaller number of candidate markers in a larger cohort in a second phase to reach statitistically significant results. It was then demonstrated that jointly analyzing data from both stages would nearly always increase statistical power compared to the two step approach (Skol et al., 2006), and this strategy has been adopted in subsequent GWAS studies. A high throughput plasma proteomics pipeline would allow us to implement a similar strategy, where discovery and verification are handed in parallel and are followed by a verification and implementation phase.

Even in the proposed rectangular strategy, a multi-phase approach will still be indispensable to verify results in independent cohorts to control for study-specific effects and biases. However, our high throughput plasma proteomics pipeline would allow us to shift from the classical triangular to a rectangular workflow. In this new strategy a large number of proteins would be quantified across a large number of samples, which would result in much stronger candidates for the following validation phase(s) (Figure 9 A). Instead of a discovery study that is followed by a verification phase, Plasma Proteomic Profiling is applied to two large and independent cohorts, neither of which is privileged over the other. The set of overlapping, significant proteins then constitutes the verified biomarkers. In addition to delivering more robust ‘first phase candidates’, this approach offers the opportunity to verify several biomarkers at once. Repeating this process in many studies for a large diversity of conditions will by itself build up the ‘knowledge base’ described above, that connects the plasma proteome to actionable human phenotype information (Figure 9 B).

1.4.3.2. Sample preparation

MS-based proteomic workflows consist of multiple steps, namely sample preparation, on-line liquid chromatography, MS measurements, followed by computational data analysis and bioinformatics interpretation. The extensive sample preparation procedure begins with the extraction and solubilization of proteins, followed by denaturation, reduction and alkylation of cysteine residues and enzymatic digestion. The peptides are cleaned up and separated by long gradients (2-4h) where the HPLC is on-line coupled to electrospray ionization and data acquisition by the mass spectrometer. In contrast to such workflows, which usually aim to maximize protein identifications, we here focused on quantitative accuracy and throughput to develop a rapid, robust and highly reproducible workflow from sample preparation to data analysis that could be used for clinical applications.

By optimizing the digestion buffer system, Kulak et al. simplified the sample preparation protocol, removed bias prone precipitation steps and increased digestion rates (Kulak et al., 2014). To obtain an even more rapid workflow, I further removed unnecessary steps like repeated sample boiling, ultrasound-treatment and overnight digestion. Further minimization of starting material allowed decreasing the amount of expensive digestion enzymes and the combination these (trypsin and LysC) in the same digestion mixture increased throughput and proteome coverage. Using this protocol we observed that suitable digestion occurred already after 1h with low 'missed cleavage rates', similar to the standard overnight digestion, and very low coefficients of variations (CVs) for the majority of all proteins.

Contaminants that can result in clogged HPLC columns and frequent cleaning cycles of the MS cause increased instrument down times and low sample throughput. Therefore, one of our main objectives was to establish optimal washing conditions for peptide cleanup. This was achieved by extensive testing of a large variety of solvents and mixture conditions as well as different solid phase extraction matrices. The washing procedure need to remove interfering buffer components, lipids and other contaminants, yet retain the peptides. The final protocol combined one solvent condition with extensive mixing of the digest with the washing buffer (100 pipetting cycles) and a particular solid phase extraction matrix (polystyrene-divinylbenzene – reverse phase sulfonate, SDB-RPS). Mixing dissolves all contaminants and makes it possible to separate them from the peptides by a cleanup over the solid phase extraction matrix which retains peptides.

Next we wished to improve reproducibility and high throughput by automation. For this purpose, we installed an Agilent Bravo liquid handling system with disposable pipet tips

and transferred our protocol to this platform. The robotic protocol was optimized and evaluated over several month, resulting in a reproducible, reliable and error-free system. I developed several liquid handling schemes, starting with finding optimal lab-ware that does not introduce polymers or absorbs peptides during processing. Pipetting strategies like pre and post air aspiration and tip touch procedures for highly accurate handling of small volumes were incorporated. Further adjustments of sample volumes as well as the washing buffers were necessary to transfer the protocol to the robotic platform.

The following statistics illustrate the initial challenges that we were facing and the progress that we have made: Prior to optimization, a single cleanup workflow needed 24-48h and we were only able to analyze 20-30 samples before contaminants led to deterioration of our HPLC system. Today, we prepare 96 samples in a fully-automated manner within 3h and we regularly measure hundreds of samples without any problems. This decreases hands-on time, which can be spend on data analysis instead, improves reproducibility, and makes optimal upkeep of expensive mass spectrometers much less stressful.

We termed our concept of analyzing whole, undepleted plasma in a rapid manner with a very robust and reproducible workflow, ‘Plasma Proteome Profiling’ and described it in a manuscript that became the featured article in the journal Cell Systems *(Geyer et al., 2016a)*.

1.4.3.3. LC-MS/MS optimization

The optimization of sample preparation described above was the first step necessary for the analysis of large cohorts. It enables high throughput on the sample preparation side by preparing purified peptides in a short time that are ready for MS-analysis. However, typically MS-based proteomics is maximized for the number of proteins that can be identified and this usually entails long HPLC gradients, which would impede throughput. In highly complex samples with hundred thousands of peptides like cell lines or tissues, long gradients provide the peptide fragmentation time necessary for high proteome coverage. In plasma, in contrast, there are just thousands of peptides in a sensitivity range that makes them accessible to MS^1 and MS^2 analysis. As a consequence, we found that shorter gradients result in nearly the same number of identified proteins. In particular, a 20 min gradient lost only 5% of protein identifications compared to the standard 100 min gradient (Geyer et al., 2016a). We then investigated increasing the flow rate, decreasing loading volumes, optimizing gradients and shortening the HPLC column length to assure optimal usage of the short gradients. In

this way we are now able to analyze almost 50 samples per day. Later we also implemented DMSO as an additive to increase peptide ionization efficiency (Hahne et al., 2013).

One disadvantage of short gradients is that they use the MS instruments inefficiently. Even after optimization, loading and equilibration requires more than 10 minutes on high-end HPLC systems, which for the 20 min gradients, would mean that the mass spectrometer is unused a 1/3rd of the time. This could potentially be avoided by more sophisticated LC set ups, but here we decided to increase utilization of the MS instrument time with a somewhat longer gradient (45 min; about 24 samples/day) and combine this with the 'library matching' approach, that will be explained below.

1.4.3.4. Library matching strategy

Data-dependent acquisition strategies use a combination of MS^1 and MS^2 scans. Following their detection in the MS^1 spectra, peptide precursor ions are ranked by intensity and the analytical quadrupole selects them in this order with a small isolation window (typically 1.4 Th) centered on the measured m/z. The peptides are fragmented in the collision cell and the masses of these fragments are recorded with high accuracy in the Orbitrap analyzer. Precursor mass and fragment masses are then used in a database search to determine the sequence and therefore the identity of the peptide. Because relative peptide elution varies somewhat between runs, this strategy results in a semi-stochastic selection of precursors depending on the intensity-based ranking (top N). Furthermore, the fragmentation spectra may be of sufficient quality for identification in one run but not another.

Our laboratory has developed the software package MaxQuant for peptide and protein identification and quantitative analysis of MS data (Cox and Mann, 2008). MaxQuant incorporates an optional feature termed 'match between runs' to transfer peptide identification based on information about retention time and accurate m/z ratios from one LC-MS run where the peptide was identified by an MS^2 scan to another HPLC run where the required MS^2 data is not present (Geiger et al., 2012; Nagaraj et al., 2012). In plasma we observed that the matching strategy was also advantageous. However, because the numbers of identified peptides in the libraries was low, there were still few matched peptides in the single runs. I therefore acquired successively deeper plasma peptide libraries. The first consisted of single runs of plasma that was depleted for the top 6 most abundant proteins, but we then realized that the combination with a top 14 depletion column produced superior results. Importantly, we use the libraries only for

matching and therefore it is not required that depletion is quantitatively accurate. In this way we were able to boost protein identifications by almost 40% (Geyer et al., 2016a). Later we combined the double depletion with high pH reversed phase fractionation (Kulak et al., 2017) to acquire very deep libraries of more than 1,500 proteins (see below). Without any matching applied, identification was limited to around 200 proteins in 45 min gradients. The 'depleted library' approach allowed us to cover more than 500 proteins within the same time.

To further boost the depth of the measured plasma proteome, we made use of a recent development in our laboratory that dramatically increases the dynamic range of detection in MS^1 scans, which are limited to about one million ions due to space charge effects (Meier et al., 2017). In this 'BoxCar' acquisition method, the m/z range is broken up into multiple narrow m/z windows, which are filled with much longer injection times than the MS^1 scan. This results in 'normalizing' the full scan and in effect boosts the intensity of low abundance ions ten-fold or more. BoxCar works particularly well in proteomes with a high dynamic range such as plasma where a few very high abundant peptides otherwise mask co-eluting, lower abundant peptides in the MS^1 scan (Meier et al., 2017). Remarkably, combination of the above-mentioned improvements with BoxCar scans allowed us to identify over 800 proteins in single and more than 1,000 proteins in triplicate measurements.

1.4.3.5. Deep quantitative plasma proteomes

To achieve deep proteome coverage in complex biological samples, an additional step of peptide fractionation is widely used. Splitting a tryptic peptide mixture into several fractions, while loading the analytical column to capacity, will increase the detectability of low abundance proteins, because the peptides are separated from each other and more material can be injected into the MS in total. High pH reversed-phase fractionation in combination with concatenation as a first dimensional separation step has proven to be highly efficient (Delmotte et al., 2007; Gilar et al., 2005b; Manadas et al., 2009).

Because of the larger diameter columns used, such fractionation approaches typically require rather large sample amounts (in the mg-range) and the concatenation procedure can be error-prone and time consuming. To tackle this problem, we developed a high pH reversed-phase fractionation and concatenation device – called 'Spider Fractionator' – for automated off-line chromatography separation of small peptide amounts (Kulak et al., 2017). In cell lines this approach allowed us to identify more than 150,000 peptides and almost 12,000 proteins with 24 fractions. In plasma, I used several depleted plasma

samples and constructed a library of 2,000 proteins, reflected by 14,000 sequence unique peptides.

For the physiological interpretation of protein levels in plasma, we also wished to construct a deep quantitative plasma proteome. With eight 'spider-fractionations' measured as singlets in 45 gradients, I quantified nearly 1,500 proteins. As expected, 'functional plasma proteins' were generally of comparatively high abundance, 'tissue leakage proteins' were scattered among the middle and low abundance range, whereas most of the cytokines have exceedingly low levels in normal plasma and were therefore not detected. To my knowledge, this is the first deep and quantitative plasma proteome and it should be a useful resource to the community.

1.4.3.6. Throughput vs. deep measurements

In addition to the throughput required to realize the concept of Plasma Proteome Profiling, it clearly also requires a certain depth of coverage, to combine highly accurate MS acquired data with clinical data on a global scale. This mandates an optimal combination of throughput and proteome coverage.

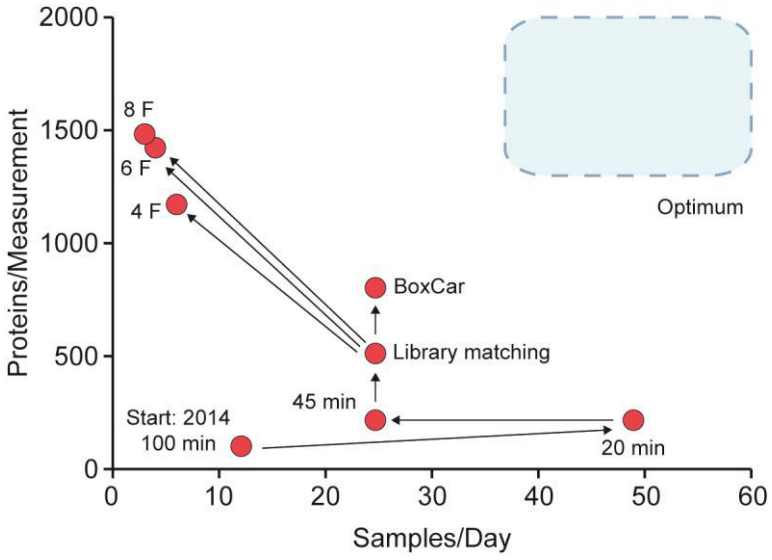


Figure 10: Technological developments and strategic assessments of Plasma Proteome Profiling. This PhD thesis started with a standard proteomic workflow in 2014. Optimizations on various stages allowed us to develop the Plasma Proteome Profiling concept. Implementation of further technological developments increased the number of identified proteins and the throughput as shown. Unless indicated otherwise, 45 min gradients were used. The numbers of fractions are displayed for the fractionation strategy.

To visualize progress towards the goals of Plasma Proteome Profiling, one can plot the total number of identified proteins against the number of samples that can be measured

per day (Figure 10). In this plot the far right and top corner is the most desired region and the figure illustrates successive improvements in throughput and proteomics depths. With the technology developed in this thesis, we can now identify on average 800 proteins in plasma in 45 min gradients at a throughput of 24 samples per day and per instrument (8 per day in case of triplicate measurements). Alternatively, collecting 4, 6 or 8 fractions, resulted in 1160, 1429 and 1487 identified proteins, respectively. An optimum between measurement time and number of proteins seemed to be reached with six fractions. Compared to existing literature, such numbers are exceptional for plasma proteomics, especially taking into account that we obtain quantitative proteomes from undepleted plasma. However, the throughput of only four samples per day with six fractions is not yet compatible with our goal of analyzing large cohorts. This will require either an even deeper coverage in single runs or a multiplexing strategy to increase sample throughput after fractionation.

1.4.3.7. Quality marker panels

Samples for thousands of clinical studies are being collected at any given time worldwide, adding to the countless clinical studies that are already stored in biobanks with the aim to find biomarkers. There is a large variety of protocols for sample collection and storage, potentially confounding subsequent analysis. The community is acutely aware of these issues, but could not address them in a systematic manner so far (Lombardi et al., 2012; Pickup et al., 2017; Zhao et al., 2012). Given the fact that it is a basic requirement that the samples are of high quality and the potential for systematic errors to affect study outcomes, we asked whether Plasma Proteome Profiling could solve this issue.

Samples of poor quality contribute to the variation between individuals, one of the major focus areas in biomarker research (Mischak et al., 2010; Surinova et al., 2011). The small discovery cohorts in the triangular strategy of biomarker research are especially prone to suffer from samples with poor quality. As pointed out above, it is a wide-spread practice to pool clinical samples within case and control groups to decrease measurement time and “equalize” individual-specific differences. The danger in such practices lies in single contaminated samples that can result in a systematic bias. Our literature search illustrated the dimension of the problem, and suggests that a sizable proportion of the literature has reported potentially incorrect results because of this.

According to the same literature search, shotgun proteomic results are often ‘verified’ in the same cohort just with a different technology. Importantly, proteins enriched in case

or control solely due to quality issues, can still pass such experimental designs and may be considered for further validation in larger, time and cost intensive cohorts. Furthermore, genuine biomarkers may be obscured by protein variation introduced through quality issues. In view of these issues, it is unfortunate that to date, there are no protein-based markers available to monitor the quality of blood-based samples. This is in contrast to the situation in metabolomics, which faces the same challenges. Here, it is possible to assess some quality criteria with a metabolomics-based marker panel, which take blood coagulation and storage time into account (MxP Quality Control Plasma (Kamlage et al., 2014)). However, this test does not consider the blood sampling procedure, erythrocyte contamination and is restricted to EDTA-plasma.

We reasoned that the unbiased coverage of Plasma Proteome Profiling and the robust nature of the workflow would lend themselves to address the issue of sample quality. We performed a wide variety of experiments to investigate this question and uncovered three different kinds of sample quality marker classes: those for blood sampling procedure, erythrocyte contamination and coagulation.

The first class includes all markers that originate from blood sampling. Blood taking with different equipment, including needles with different diameters, collection tubes and the use of products from different vendors can all influence results. In daily clinical practice, blood is sampled by trained nurses, medical doctors, but also by rather untrained medical personal like students, contributing to this class of quality issues. We reasoned that smooth muscle and endothelial cell specific proteins would be candidates for this class of markers. Each blood vessel consists of different layers. The outer ones consist of smooth muscle cells and the inner one of endothelial cells. The above mentioned issues can lead to the collection of different proportions of these cell types together with the blood. To investigate this at the protein level, we looked for smooth muscle and endothelial cell specific proteins (SMECs) in our datasets and confirmed them by analyzing the proteomes of blood vessels from humans (Figure 11 A, B).

Erythrocytes are the dominant cell type in our blood stream. Blood is centrifuged to yield plasma or serum, which are the preferred matrices for clinical tests. Delay in the time until start of centrifugation, inappropriate centrifugation speed and time, handling, transportation, harvesting of the separated plasma and recontamination after centrifugation may all result in variable presence of highly abundant erythrocyte-specific proteins (HAEP) in plasma. This is especially striking when plasma is repeatedly harvested from the same person (Figure 11 C). We used spike-in experiments to identify HAEP panels and were able to quantify HAEPs down to an erythrocyte to plasma ratio of 1:10.000 (Figure 11 D) (Geyer et al., 2016a).

Activation of the coagulation cascade is necessary to produce serum from blood. In contrast, to obtain plasma, blood must be instantly mixed with an anti-coagulant, otherwise partial coagulation will result. To find markers for unintended blood coagulation, we compared serum to plasma from the same individuals. The levels of fibrinogen alpha (FGA), beta (FGB) and gamma (FGG) chains were decreased and the platelet-specific proteins platelet factor 4 variant (PF4V1) and platelet basic protein (PPBP) were increased in serum compared to plasma (Figure 11 E, F) (Geyer et al., 2016a). Using the coagulation influenced proteins (CIP) as a quality panel, we have so far detected systematic bias in one out of 13 studies.

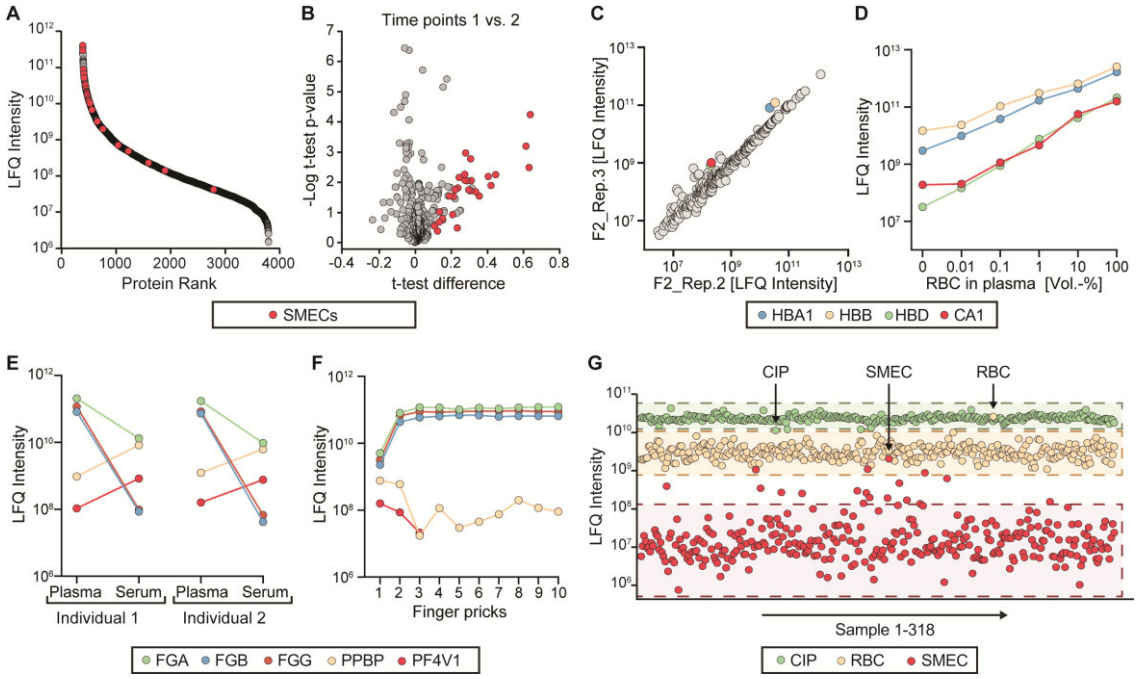


Figure 11: Quality marker panels. (A) Smooth muscle and endothelial cell (SMECs) protein markers (red) are among the highest abundant proteins in the blood vessel proteome. Proteins are ranked according to their abundances. **(B)** Volcano plot comparing two time points (1 vs. 2) of a clinical study that had a bias, indicated by increased levels of SMECs (red). **(C)** Scatterplot of repeated finger pricks of one individual (replicate 2 vs. replicate 3) showing that erythrocyte-specific proteins were elevated as a group of four proteins. HBA1, hemoglobin subunit alpha; HBB, hemoglobin subunit beta; HBD, hemoglobin subunit; CA1, carbonic anhydrase 1. **(D)** Spike-in of erythrocytes into plasma resulted in an increase of these proteins as a group. **(E)** In a comparison of plasma and serum of two individuals, the levels of FGA, FGB, and FGG were decreased, and PPBP as well as PF4V1 were elevated in serum. FGA, FGB, FGG, fibrinogen chains alpha, beta, gamma; PPBP, platelet basic protein; PF4V1, platelet factor 4 variant. **(F)** Blood was processed from ten different fingers of one individual after finger pricking, and mass spectrometric (LFQ) intensities of FGA, FGB, FGG, PPBP, and PF4V1 are plotted. In samples 1 and 2, fibrinogens were decreased, whereas platelet-specific proteins are increased. **(G)** Sample quality assessment of a study consisting of 318 samples. One protein of each quality marker panel was chosen for this illustration. The reference range of each marker is highlighted in the color of the quality marker class (green, yellow, red). Outliers of these categories, like the three indicated samples, are of poor quality. CIP: Coagulation influenced proteins; RBC: Red blood cell specific proteins. **(C-F)** were adapted from (Geyer et al., 2016a).

Quality marker panels would be especially valuable in hospitals for assessment of clinical samples and in the selection of cohorts to be used in biomarker studies. On the basis of such tests, clinical samples of poor quality can be discarded to avoid reporting incorrect clinical test results and in general as an internal quality check of the clinical laboratory (Figure 11 G). Assessing the quality of existing studies should drastically decrease the fruitless follow up of spurious biomarker candidates and conversely, it would increase the probability of finding real biomarkers by certifying high quality of the investigated clinical samples and cohorts.

1.4.3.8. SILAC-PrESTs as internal standards for absolute protein quantification

In clinical diagnostics, the quantification of the analyte of interest by immunoassays is almost always uses standards of known concentration for calibration (external reference). In contrast to immunoassays, MS-based methods can accommodate internal standards, promising accurate and absolute quantification. Heavy isotopically labeled analytes are added to the sample at the earliest possible time point of sample processing. As the internal standard is exposed to the same influences as the analyte, it automatically corrects for variations during the workflow, resulting in highly accurate quantification. Such MS-based assays with internal standards are already applied in clinical practice for small molecules, including several metabolites such as phenylalanine in the phenylketonuria screen for newborns.

In a collaboration with the Uhlen group in Stockholm, our laboratory had developed ‘Stabile Isotope-Labeled Protein Epitope Signature Tags’ (SILAC-PrESTs) for absolute quantification of multiple proteins (Edfors et al., 2014; Zeiler et al., 2012). They are constructed as fusion proteins consisting of a histidine-tag, the albumin binding protein (ABP) and a unique sequence stretch of the protein of interest. First, the construct is recombinantly expressed in *E. coli* in heavy labeled form, growing in media with stable isotope labeled amino acids and purified by the histidine-tag (Figure 12). The ABP enhances solubility of the recombinant proteins and is used for determination of the concentration of the expressed SILAC-PrESTs. In this second step, the absolute concentration of a ‘light version’ (not isotopically labeled) of the ABP can be determined very accurately. This light ABP is digested and measured together with the heavy labeled SILAC-PrEST and their ratio is used to calculate the concentration of the SILAC-PrEST. The third step is the crucial one for absolute quantification. The SILAC-PrEST is spiked into the sample and the SILAC ratios of the PrEST peptides to the peptides of the target protein can be determined. Importantly, this approach is suitable for

multiplexing, which has already been demonstrated for 40 SILAC-PrESTs quantifying HeLa proteins. Multiplexed SILAC-PrEST assays could allow the absolute quantification of many clinical interesting proteins in single measurement instead of multiple, separate protein based immunoassays.

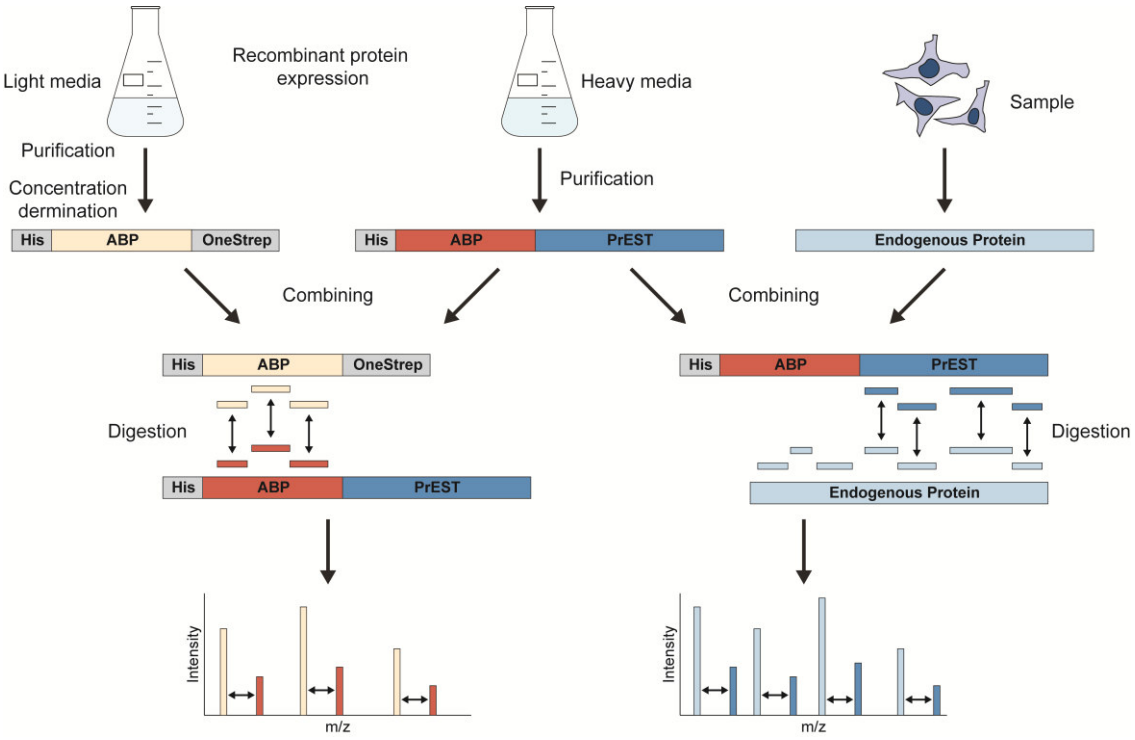


Figure 12: Absolute protein quantification based on SILAC-PrESTs. (A) The albumin binding protein (ABP) is recombinantly expressed in *E. coli* growing in light amino acid containing media. It is purified with a His-tag and an OneStrep-tag and the concentration is determined. The SILAC-PrEST construct is expressed in media with heavy amino acids. After purification the SILAC-PrEST and the light ABP are combined and digested together. The peptides are measured in the MS and the ratios are used to calculate the absolute concentration of the SILAC-PrEST. **(B)** The SILAC-PrEST is combined with the sample of interest (e.g. cells or plasma) and they are digested together. The ratio of heavy to light peptides of the endogenous protein and its SILAC-PrEST homologous sequence are used to calculate the absolute concentration.

2. Aims of the thesis

The overall aim of my PhD thesis was to pave the way for biomarker discovery and clinical applications of proteomics by precision characterization of the human blood plasma proteome, a major goal of mass spectrometry (MS)-based proteomics for decades. Due to great technological challenges, misguided concepts and study designs, the analysis of the plasma proteome by MS has not yet lived up to its promises: no new biomarkers have been discovered, plasma proteomics has not entered clinical diagnostics and new biologically meaningful insights have been gained. As a consequence of these unsuccessful efforts, relatively few groups continue to pursue plasma proteomics, despite its undiminished potential for research and medicine.

One requirement for a revival of plasma proteomics would be a rapid, robust and reproducible workflow. Towards this aim had to tackle several challenges. First of all, we had to find strategies that allowed us to efficiently measure many plasma samples. We further had to develop sample preparation procedures and optimize them for reproducibility. As proteomic workflows are generally very time consuming and labor-intensive, we streamlined the process by shortening several steps and discarding of others. MS-based workflows are usually also not automated, a requirement for a truly robust and a high throughput sample preparation procedure. Furthermore, proteomic LC-MS/MS systems are currently not optimized for high throughput and several developments were necessary to streamline this part of the workflow.

Another major challenge of plasma is the high dynamic range of protein concentrations. We addressed this obstacle by developing and implementing several strategies and technologies like library matching, the ‘Spider Fractionator’ and ‘BoxCar’ scans, dramatically increasing the number of quantified proteins.

A central aim of the thesis was to demonstrate that MS-based proteomics can be applied to large cohorts and that it is possible to gain biologically and medically relevant information. We achieved this aim with our first large scale plasma proteomic study in which we analyzed more than 1,000 proteomes and defined inflammatory and insulin resistance panels.

The plasma proteomics community implicitly subscribes to a particular strategy in biomarker research, which we find to be problematic. In my PhD thesis we developed new strategies and concepts that set proteomics-based biomarker discovery on a solid foundation.

3. Publications

3.1. Article 1: Plasma Proteome Profiling to Assess Human Health and Disease

Authors: Philipp E. Geyer[1,2], Nils A. Kulak[1], Garwin Pichler[1], Lesca M. Holdt[3], Daniel Teupser[3], and Matthias Mann[1,2]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[3]Institute of Laboratory Medicine, Ludwig-Maximilians-University Munich, 80539 Munich, Germany

In my manuscript 'Plasma Proteome Profiling to Assess Human Health and Disease', I tackle directly many issues in the analysis of plasma proteomes like the labor-intensive workflow, problematic contaminations, sample quality assessment and the analytical variability. I streamlined the proteomic workflow, resulting in a rapid, robust and highly reproducible pipeline, which was further automated by implementing it on a liquid handling platform. Optimization of digestion conditions, peptide clean-up procedures and LC-MS/MS settings enables us to prepare 96 samples in a fully-automated manner within 3h. We now regularly measure hundreds of samples without any problems. The high throughput of this technology opens up for new concepts in biomarker discovery, which we describe more fully in our review article 'Revisiting Biomarker Discovery by Plasma Proteomics' later in this thesis.

We call our technology for the analysis of blood and its derivatives 'Plasma Proteome Profiling'. This term is meant to imply that the information of the plasma proteome can mirror human physiology. Our pipeline offers highly reproducible and quantitative information for several hundred proteins (CV<20% for most proteins), including more than 40 clinical applied biomarkers from a single fingerprick with a 30 minute measurement. The quantified proteins include inflammatory markers, proteins belonging to the lipid homeostasis system, gender-specific proteins, disease relevant allele variations and quality markers.

Furthermore, we provide proof-of-principle to transfer Plasma Proteome Profiling into clinical practice by introducing a SILAC-PrESTs panel of five proteins to plasma samples. These were used as internal standards, controlling variations during the entire

sample preparation workflow and allowing accurate relative as well as absolute quantification.

This manuscript was the featured article in the journal Cell Systems and it was described as one of the journals highlights in the 2016 end-of-the-year review and it is also one of three listened article on the homepage of the Human Plasma Proteome Project. In our laboratory we are building on the developments described in this article to assess human health and disease states with the aim to fundamentally alter biomarker research and clinical diagnostics.

Cell Systems
Article

Plasma Proteome Profiling to Assess Human Health and Disease

Philipp E. Geyer,[1,2] Nils A. Kulak,[1] Garwin Pichler,[1] Lesca M. Holdt,[3] Daniel Teupser,[3] and Matthias Mann[1,2,*]
[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[3]Institute of Laboratory Medicine, Ludwig-Maximilians University Munich, 80539 Munich, Germany
*Correspondence: mmann@biochem.mpg.de
http://dx.doi.org/10.1016/j.cels.2016.02.015

SUMMARY

Proteins in the circulatory system mirror an individual's physiology. In daily clinical practice, protein levels are generally determined using single-protein immunoassays. High-throughput, quantitative analysis using mass-spectrometry-based proteomics of blood, plasma, and serum would be advantageous but is challenging because of the high dynamic range of protein abundances. Here, we introduce a rapid and robust "plasma proteome profiling" pipeline. This single-run shotgun proteomic workflow does not require protein depletion and enables quantitative analysis of hundreds of plasma proteomes from 1 μl single finger pricks with 20 min gradients. The apolipoprotein family, inflammatory markers such as C-reactive protein, gender-related proteins, and >40 FDA-approved biomarkers are reproducibly quantified (CV <20% with label-free quantification). Furthermore, we functionally interpret a 1,000-protein, quantitative plasma proteome obtained by simple peptide pre-fractionation. Plasma proteome profiling delivers an informative portrait of a person's health state, and we envision its large-scale use in biomedicine.

INTRODUCTION

Blood, plasma, and serum are the predominant samples used for diagnostic analyses in clinical practice and are available in biobanks from thousands of clinical studies (Végvári et al., 2011). The quantitative analysis of individual plasma proteins by immunoassays is used in daily clinical diagnostics. However, immunoassays have inherent limitations with regard to multiplexing, their specificity for protein isoforms, and their incompatibility with hypothesis-free investigations. Mass spectrometry (MS)-based proteomics is a technology that could address all of these limitations and that should be capable of discovering biomarkers in this easily accessible body fluid (Anderson, 2014). However, MS-based plasma proteomics is extremely challenging for a number of reasons, most prominently the extremely large dynamic range of protein abundances (Anderson and Anderson, 2002; Omenn, 2005). There is also a lack of very reproducible, robust, and high-throughput proteomic workflows to identify and verify potential biomarker in large cohorts. As a result, only few novel biomarkers have been established—fewer than 1.5 per year in the 15 years before 2010 (Anderson, 2010)—and this has generally been done by immunoassay-based technologies, such as prostate-specific antigen, one of the best known biomarkers in medicine (Vihko et al., 1978).

Dramatic improvements in the technology of MS-based proteomics over the last few years (Cox and Mann, 2011; Geiger et al., 2010; Muñoz and Heck, 2014) have rekindled an interest in plasma proteomics. Using such technology and combining it with immunodepletion of high- and medium-abundance proteins as well as very extensive peptide fractionation methods, it has now become possible to identify more than 1,000 (Addona et al., 2011; Cao et al., 2012; Paczesny et al., 2010) or even more than 5,000 proteins (Keshishian et al., 2015) in plasma. However, immunodepletion may lead to biases because of cross-reactions of the antibodies used or by proteins bound to carrier proteins such as albumin (Bellei et al., 2011; Tu et al., 2010). Furthermore, extensive pre-fractionation decreases throughput, which is undesirable in clinical practice. Accordingly, the paradigm in biomarker discovery by MS has been to analyze a small number of samples in as much depth as possible, whereas the verification phase was to be done on larger cohorts but with targeted methods and a small number of candidate markers. The final clinical test for a biomarker identified by MS was to be performed with classical immunoassays (Anderson et al., 2009; Surinova et al., 2011). Although this scheme is practical with current technology, it is very laborious and loses much of the promise of systemwide and unbiased investigation of the plasma proteome. Using another approach, Liu et al. (2015) constructed a list of plasma peptide transitions, which they used to interpret the signals in sequential window acquisition of all theoretical MS (SWATH) runs of plasma samples of twins. In this way, the contribution of heritable and environmental changes to the plasma proteome could be distinguished.

In contrast to previous approaches, we here focused on developing a robust and highly streamlined shotgun plasma proteomics workflow. For the MS readout, we used very short liquid chromatography (LC)-MS/MS gradients and recent advances in label-free quantification (Cox et al., 2014). We hypothesized that the resulting "plasma proteome profile" would have a high yield of information about the health state of an individual and that it can be obtained for a large number of clinical samples.

RESULTS

Rapid, Robust, and Highly Reproducible Plasma Proteomic Workflow

Past efforts in shotgun plasma proteomics endeavored to maximize protein identifications, whereas generally less emphasis was placed on quantitative accuracy or throughput. Here we wished to develop a convenient workflow, from sample preparation to data analysis, that can potentially be used in a clinical context. We reasoned that such a workflow should be rapid, optimal for high-throughput, robust, and highly reproducible. Therefore, it should minimize all preparation and analysis steps, while still quantifying clinically interesting proteins accurately. With this in mind, we decided to omit any depletion steps of high-abundance plasma proteins.

Building on the recently described in-StageTip (iST) method (Kulak et al., 2014), we further streamlined the procedure for plasma (Experimental Procedures; Figure 1A). Starting with 1 μl of plasma from a single finger prick, all preparation steps were performed in a single reaction vial. Using ordinary amounts of digestion enzymes, we found that adequate protein digestion had already occurred after 1 hr (protein coefficients of variation [CVs] and tryptic missed cleavage rates were similar to overnight digestion; Table S1). Peptides were then eluted and ready for LC-MS/MS analysis. The entire up-front procedure took less than 2 hr and can readily be performed in a 96-well format and automated in a liquid handling platform, if desired.

Starting with single-run gradient times typical of proteomics experiments, we successively reduced them to determine the maximum information content per unit time. We found that the number of identified proteins decreased very slowly with decreasing time, down to 20 min (only 12 additional identified protein groups in 100 min versus 20 min gradients; Table S1). Below this time, loading and equilibration times become dominant, and therefore we chose 20 min gradients as our standard (33 min between injections, about 50 samples/day). The combination of optimized sample preparation and LC setup allowed for hundreds of plasma proteome analyses, whereas previously clogging of columns was a common occurrence with plasma samples.

We used MaxQuant for quantitative label-free analysis of the LC-MS/MS data (Cox et al., 2014; Cox and Mann, 2008) and for transferring peptide identifications from one LC run to other LC runs in which the peptide was not sequenced ("match between runs"). In combination with a matching library consisting of undepleted plasma of ten different individuals as well as plasma depleted of the 20 highest abundant proteins, this boosted protein identification by 39% (Experimental Procedures; Figures S1A and S1B). Of the 347 protein groups identified in total in the 20 min gradients, 285 were detected in all ten individuals (Figures S1C and S1D). The entire workflow, including the finger-prick procedure and the data analysis, takes less than 3 hr (Figure 1A).

Accuracy of Label-free Quantification of the Plasma Proteome

To investigate the quantitative reproducibility of our workflow (intra-assay variability), we sampled blood by venous puncture from one individual and harvested plasma after centrifugation. We performed the entire workflow 15 separate times on 1 μl aliquots of this stock and correlated protein abundances across the whole measuring range of each of the replicates. The mean R^2 correlation value of the quantified protein signals between individual replicates was excellent at 0.980, with a range of 0.966–0.994 (Figure 1B, excluding keratins; Table S2). We performed 96 blood plasma analyses using multiplexed preparation on a liquid handling robot and short measurement times (5 hr and 51 hr in total, respectively) and achieved a mean R^2 value of 0.97 (Figure S2).

On average, 284 ± 5 different proteins were quantified (total 313): the large majority in all 15 samples and only 3% uniquely in single LC runs (Figure 1C). We picked six well-characterized plasma proteins across a million-fold abundance range and found that quantification was highly reproducible (Figure 1D). We compared different conditions by the proportion of proteins with CVs less than 20%, because this is a commonly used cutoff for in vitro diagnostic assays (U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Veterinary Medicine, 2001). Notably, 67% of quantified proteins were within the 20% cutoff range, and 30% had CVs below 10% (Figure 1E).

To determine the variability caused by LC-MS/MS analysis alone (analytical variability), peptides from one sample preparation were injected and measured 15 times. This resulted in only slightly better reproducibility (71% with a CV less than 20% and 37% with a CV less than 10%), indicating that up-front sample preparation contributed little to overall quantitative variability (Figure 1F). A notable exception to this trend were certain keratin proteins, which had very small analytical variability but sometimes had a large quantitative difference between repeated analysis of the same sample. This is readily explained by contamination with exogenous keratins during sample preparation. Nevertheless, it is clinically relevant, because we found that plasma proteomes of the same person clustered together much better after excluding keratins and other proteins introduced by sample processing such as hemoglobins (see below).

Intra- and Inter-individual Variability of the Plasma Proteome

The high-throughput of our workflow allowed us to extensively characterize the quantitative variation within and between individuals. To determine inter-individual variability, we performed finger pricks on one person four times a day over 8 consecutive days and analyzed all 32 blood proteomes with less than 24 hr of measuring time. This revealed stability of the plasma proteome over time (55% of proteins below 20% CV; Figure 2A). The proteins with large CVs were the aforementioned keratins, as well as high-abundance erythrocyte-specific proteins. The latter are caused by a slightly different extent of erythrocyte lysis during plasma preparation or by contamination of plasma with erythrocytes during plasma harvesting.

To determine inter-individual variability, we harvested plasma from five female and five male donors in triplicate by finger pricks. The average R^2 value within the technical workflow triplicates was 0.976, excluding keratins and erythrocyte-specific protein groups. For CVs of the technical replicates of all individuals, see Table S3. Of 345 proteins quantified, only a minority was under the CV cutoff. This indicates that overall, the plasma

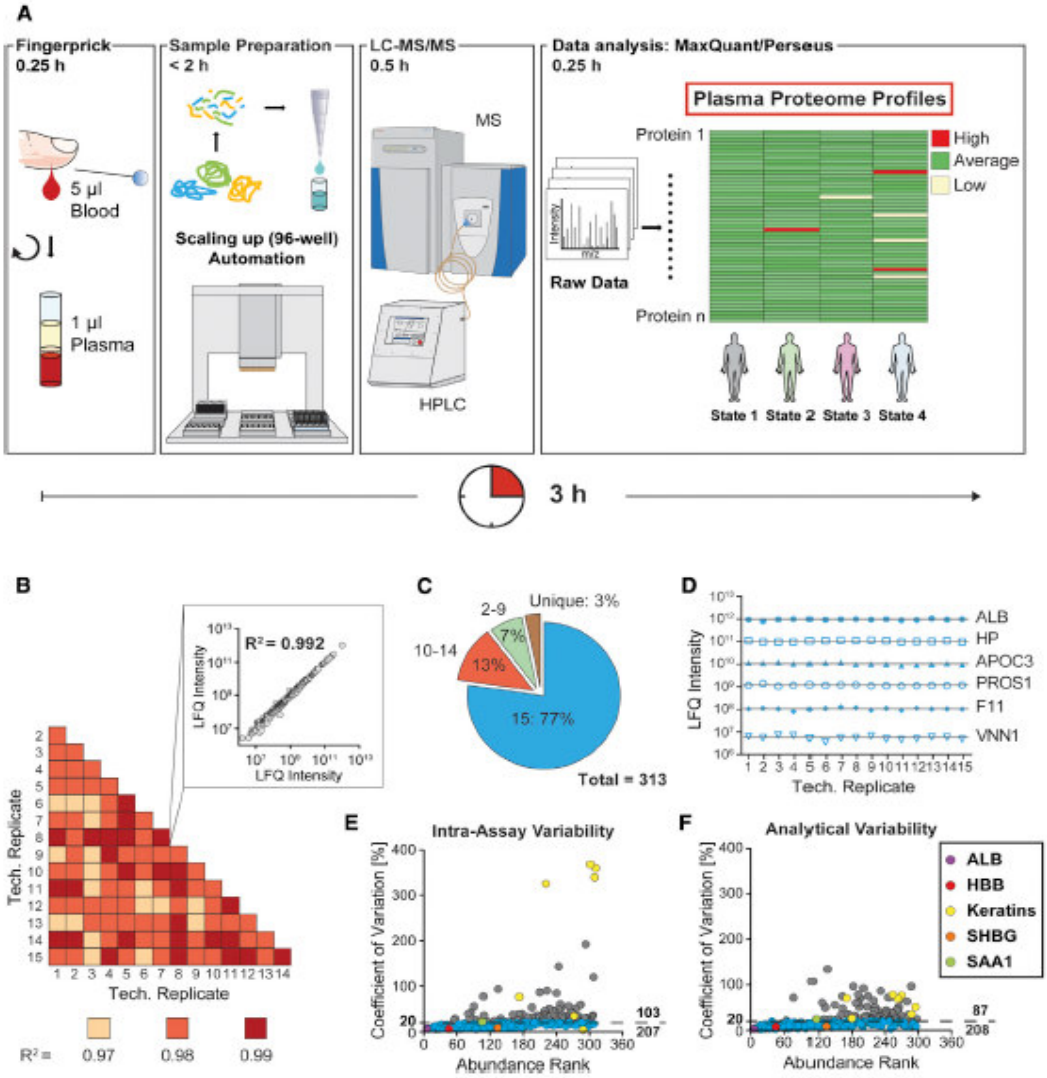


Figure 1. Technological Aspects of Plasma Protein Profiling

(A) Schematic depiction of the workflow. Blood volumes of 5 µl are routinely used to harvest 1 µl of plasma. The workflow is based on the iST protocol and consists of denaturation, reduction, alkylation of cysteines, short 1 hr enzymatic digestion, and purification of peptides. Automation for liquid handling platforms is also indicated. Peptides are separated with optimized short 20 min HPLC gradients and measured online by LC-MS/MS. Data analysis is performed by MaxQuant and Perseus, which deliver information about hundreds of plasma proteins that could reflect an individual's state as symbolized by the plasma proteome profiles.

(B) Color-coded R^2 values for the binary comparison of 15 technical workflow replicates. R^2 values up to 0.994 demonstrate high reproducibility.

(C) Frequency of protein quantification, which was present in all 15 workflow replicates, in 10–14, in 2–9, or only in 1.

(D) Reproducibility of the LFQ intensities of six proteins covering nearly six orders of magnitude for 15 workflow replicates. The line represents the mean values for ALB (serum albumin), HP (haptoglobin), APOC3 (apolipoprotein C-III), PROS1 (vitamin K-dependent protein S), F11 (coagulation factor XI), and VNN1 (pantetheinase).

(E) To determine the intra-assay variability, CVs of all quantified proteins were calculated for the 15 workflow replicates and are plotted according to their abundance. Proteins with CVs < 20% are colored in blue and those with CVs > 20% in gray. HBB, hemoglobin subunit beta. SHBG, sex hormone-binding globulin; SAA1, serum amyloid A-1 protein.

(F) Fifteen repeated injections were used to determine the analytical variability, which includes variability of the LC-MS/MS analysis.

proteome has much higher inter- than intra-individual variability (19% and 55% of proteins within a CV of 20%, respectively; Figure 2B). These general trends have been observed previously (for a recent example, see Liu et al., 2015). Here they suggest that our label-free workflow is well suited to capture the natural or pathological variation of protein levels between individuals.

To directly test this notion, we asked if we could discern systematic differences between the plasma proteomes of women

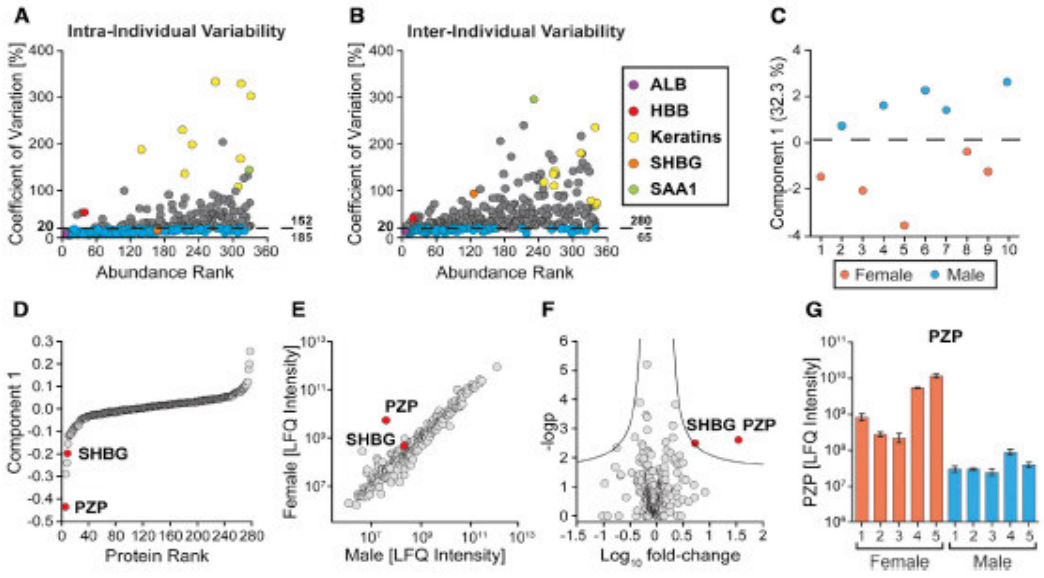


Figure 2. Intra- and Inter-individual Variability of the Plasma Proteome

(A) Intra-individual variability was assessed by finger pricks four times per day for 8 days, and CVs for all quantified proteins were calculated. Proteins with CVs < 20% are colored in blue and those with CVs > 20% in gray.

(B) Inter-individual variation of five women and five men.

(C) Female and male proteomes in one-dimensional PCA.

(D) Proteins and their contribution to the gender separation.

(E) Direct comparison of female subject 4 (F4) and male subject 5 (M5) depicts the extreme difference of PZP between women and men against the background of all other quantified proteins.

(F) Volcano plot of female against male proteomes (x axis, fold change of females to males serving as t test difference; y axis, p value). The black curves show the threshold for statistical significance, where we used a false discovery rate of 5% and an S0 of 0.8.

(G) LFQ intensities for PZP in all ten individuals.

and men, a question that to our knowledge has not been addressed by shotgun proteomics before. Indeed, one-dimensional principal-component analysis (PCA) was already sufficient for complete separation (Figure 2C). Inspection of the drivers of the PCA separation revealed that several of them are known to be regulated by estrogen (Figure 2D) (Christensen et al., 1989; Ottosson et al., 1981; Sand et al., 1985). Direct comparison of the plasma proteome profiles of a woman and a man shows that pregnancy zone protein (PZP) and sex hormone-binding globulin (SHBG) are of high absolute abundance in the plasma proteome of women and can be as high as 1% of human serum albumin (Figure 2E). This suggests a functional role in plasma, and indeed, SHBG binds estrogen, whereas PZP traps proteases (Figure 2F). On average, PZP levels were 33-fold higher in women compared with men. Furthermore, two women had 10- to 100-fold higher levels than the other three, likely because of highly elevated levels of estrogen (Figure 2G).

Rapid Assessment of Sample Quality by Plasma Proteomes

A frequently discussed issue in plasma proteomics as well as in clinical laboratory medicine is the potentially deleterious effects of inconsistent sample handling, such as variable time between blood taking and workup. We reasoned that our rapid and highly reproducible workflow might also allow the determination of protein markers of sample quality.

In clinical practice, a certain degree of hemolysis is not uncommon. Starting from our observation that high-abundance erythrocyte-specific proteins often showed high variability (Figures 2A, 2B, and 3A), we deliberately spiked in increasing amounts of erythrocyte lysates to pure plasma. We obtained a proportionate increase of erythrocyte-specific proteins, specifically, hemoglobin subunits alpha, beta, delta, and carbonic anhydrase 1. Notably, these proteins increased linearly (R^2 = 0.99), and even an admixture of 1 in 10,000 could easily be spotted (Figure 3B). This demonstrates that plasma proteome profiling readily indicates even small amounts of cellular contamination, in which case the values of pertinent proteins could be disregarded or corrected. The importance of this analysis step is illustrated by triplicate plasma proteome analysis, in which the samples from individual donors clustered together much more tightly in PCA when keratins and prominent red blood cell proteins were removed (Figure S3).

The blood coagulation system is primed for clotting in case of injury and wound repair. Although serum is harvested by inducing coagulation, harvesting of plasma requires addition of appropriate amounts of anticoagulants. Our plasma proteome profile contained many proteins with a function in the coagulation cascade, and we next evaluated the coordinate behavior

of these proteins as a quality control for appropriate plasma preparation. Plasma from each of the fingers of one individual was processed (Experimental Procedures). The levels of fibrinogen alpha (FGA), fibrinogen beta (FBA), and fibrinogen gamma (FGG) were lower in two of the samples. In addition, platelet basic protein (PPBP) and platelet factor 4 variant (PF4V1), which are released from activated blood platelets, were increased only in these same samples (Figure 3C), suggesting that partial coagulation had occurred. To test this hypothesis, we collected plasma and serum from two individuals and carried out sample preparation in triplicates. Indeed, levels of FGA, FGB, and FGG were much lower and levels of PPBP and PF4V1 much higher in serum compared with plasma (Figure 3D).

These observations prompted us to investigate coagulation and erythrocyte status in optimally prepared plasma. For this purpose, we obtained reference samples from a blood bank, which had gone through an extremely rigorous sample collection procedure (Experimental Procedures). They had very low and constant levels of red blood cell-specific proteins, and none had evidence of partial clotting. Although our plasma samples were also virtually coagulation free, this is in our experience not always the case with samples obtained from clinical studies (Figure S4).

Quantification of Clinically Interesting Markers in Short Gradients

Apolipoproteins are functional blood proteins involved in lipid homeostasis. They therefore reflect an individual's metabolic status, and some of them are classical markers of cardiovascular risk and metabolic disorders such as diabetes (Jenkins et al., 2014; Jensen et al., 2014). We quantified 15 apolipoproteins at each of 32 different time points in one individual. Apolipoprotein-a (LPA) had the strongest variation (CV = 20%), whereas APOB had the lowest (CV = 6%). The distribution of LPA levels in the population is skewed toward zero, with most individuals having low LPA levels but some (~20%) having higher levels. The successful quantification of the apolipoproteins in 32 plasma proteomes demonstrates the feasibility of a longitudinal measurement of risk factors known to be associated with an individual's propensity for certain diseases (Figure 3E).

Some of the apolipoproteins have allelic variants occurring with high frequency in populations that can easily be detected by MS (Krastins et al., 2013; Martinez-Morillo et al., 2014). The apolipoprotein allele APOE4 in the homozygous form is the largest known risk factor for late-onset Alzheimer's disease with a 10-fold higher risk compared with the homozygous APOE3 form (Tanzi, 2012). APOE4 has an arginine at position 112 instead of a cysteine residue in APOE2 and APOE3. In the 20 min LC-MS/MS data, we were able to clearly distinguish between the peptides LGADMEDVR (APOE4) and LGADMEDVCGR (APOE2, APOE3). In our group of ten individuals, two had one APOE4 allele (Figures 3F and 3G). The second allele was either an APOE3 or the APOE2 allele.

Serum amyloid A-1 protein (SAA1) and C-reactive protein (CRP) are acute phase proteins that are routinely measured in the clinic. Both are correlated with inflammatory states, and chronic elevation is strongly associated with increased risk for future cardiovascular events (Hua et al., 2009; Wilson et al., 2008). We found that their expression levels varied up to 1,000-fold among the ten individuals, and in a correlated manner (R^2 = 0.6; Figures 3H and 3I). In the plasma proteome with the highest levels of SAA1 and CRP, these are by far the largest differences to the plasma proteomes of the other healthy individuals, and this is presumably caused by recovery from a common cold (Figure 3J).

Next, we asked if our rapid proteome profiles contained information on any further known biomarkers. We scanned the raw data of the 15 technical workflow replicates to calculate the CVs for Food and Drug Administration (FDA)-cleared or FDA-approved biomarkers, as listed Anderson (2010). In total, 49 FDA-approved biomarkers were present in this data set (46 quantified in all 15 workflow replicates); 41 of them had CVs of less than 20%, and 28 had CVs even less than 10% (Figure 3K). When dividing these FDA-approved biomarkers into different classes (Anderson, 2010), 45 fell into "act in plasma," 2 into "tissue leakage," and 1 into "receptor ligand," and 1 was lysozyme, which had not been assigned to any category. The 20 min gradients already covered 45 of a total of 54 proteins among the "act in plasma" biomarkers. Interestingly, 42 of them were among the 180 highest abundance proteins, whereas the next 133 proteins contained only 7 known biomarkers (Figure 4A; Table S2).

Plasma Protein Epitope Signature Tags as Internal Standards for Protein Quantification

In clinical applications, quantification is almost always performed with internal standards. To add this capability to our fast workflow, we investigated the use of stable isotope labeling of amino acids in cell culture (SILAC)-protein epitope signature tags (PrESTs), which are recombinant expressed stable isotope-labeled protein fragments. This approach has the advantage that it controls for digestion efficiency, alkylation rate, and other workflow aspects and that a "master mix" of dozens of proteins of interest can be readily prepared and quantified (Edfors et al., 2014; Zeiler et al., 2012). We used APOA1, APOA4, APOB, APOE, and SHBG to construct a master mix for quantification of multiple plasma proteins in short gradients. Samples from ten individuals were prepared in triplicate and measured (Figure S5, Table S4). This resulted in low CVs for these proteins (APOA1 = 2.3%, APOA4 = 3.8%, APOB = 5.3%, APOE = 3.8%, and SHBG = 14.7%). Optimized targeted methods applied to peptides resulting from the PrESTs could improve these CVs even further.

A Quantitative Proteome of 1,000 Plasma Proteins

The above experiments highlight the value of quantifying hundreds of proteins in a very short analysis time. To obtain estimates of abundances for a deeper plasma proteome, we used a combination of peptide pre-fractionation, a matching library consisting of depleted plasma, and 100 min high-performance LC (HPLC) gradients. With 16 hr of measurement time, we identified 1,040 proteins in non-depleted plasma, of which 965 had label-free protein quantification (LFQ) values. Although MS signals for these proteins span more than six orders of magnitude, the majority of them were confined to a 100-fold abundance range (Figure 4B). The deep proteome data can be assessed in Table S5 and in the MaxQB database (Schaab et al., 2012), which also displays the mass spectrometric evidence and MS/MS transitions for all identified peptides.

Unexpectedly, the deep plasma proteome contained only 14 additional FDA-approved biomarkers compared with the 49 already found in the 20 min gradients. Nine of them were

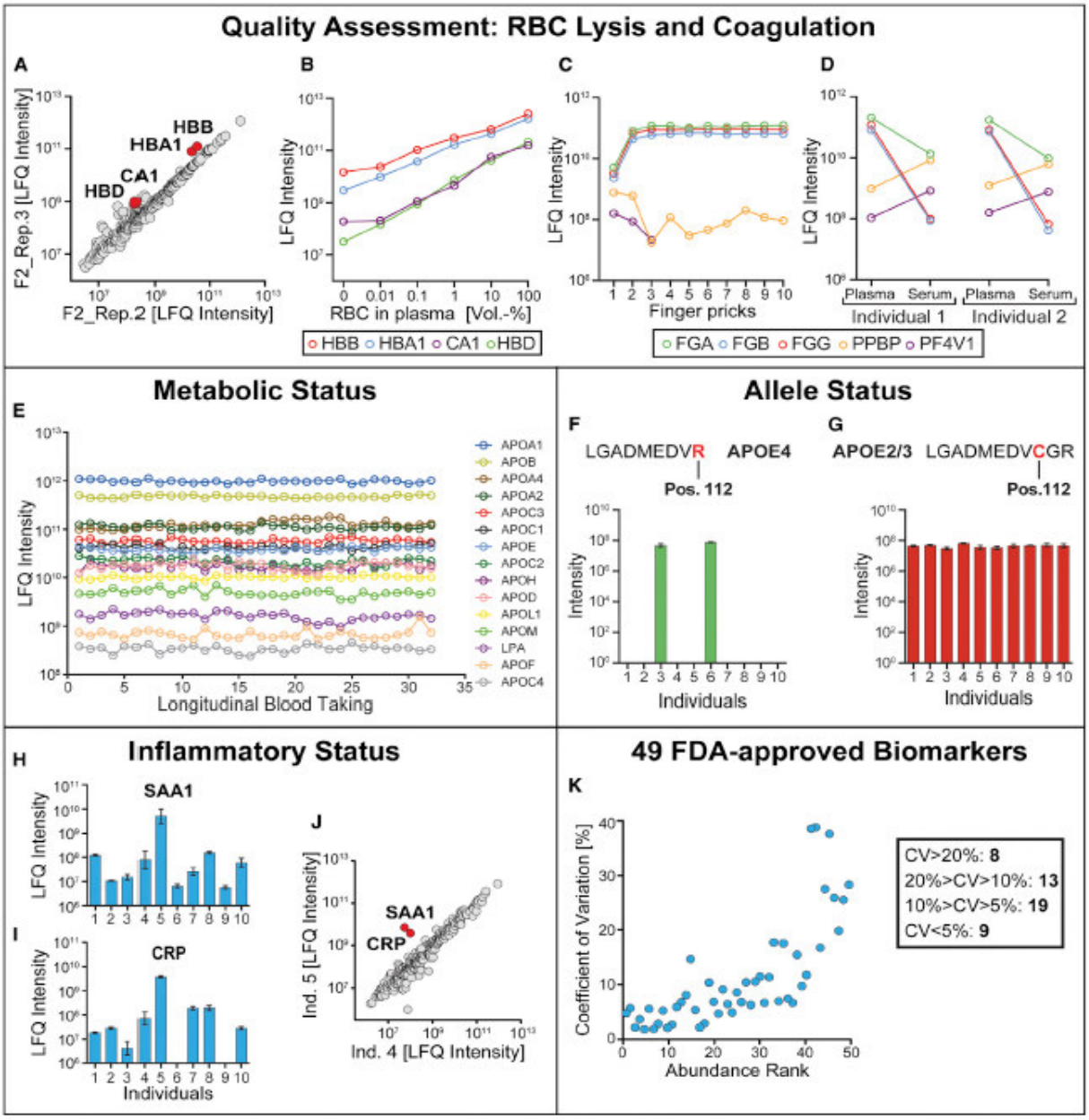


Figure 3. Quantification of Clinically Interesting Proteins

(A) Scatterplot of repeated finger pricks of one individual (replicate 2 versus replicate 3) showing that erythrocyte-specific proteins were elevated as a group. HBA1, hemoglobin subunit alpha; HBB, hemoglobin subunit beta; HBD, hemoglobin subunit; CA1, carbonic anhydrase 1.

(B) Spike-in of erythrocytes into plasma resulting in an increase of these proteins.

(C) Blood was processed from ten different fingers of one individual after finger pricking, and LFQ intensities of FGA, FGB, FGG, PPBP, and PF4V1 are plotted. In samples 1 and 2, fibrinogens are decreased, whereas platelet-specific proteins are increased.

(D) FGA, FGB, and FGG levels are decreased, and PPBP as well as PF4V1 levels are elevated in serum compared with plasma in two individuals.

(E) Fifteen apolipoproteins were quantified without any missing value after longitudinal collection of 32 plasma samples of one individual (four finger pricks per day over 8 days).

(F) The peptide LGADMEDVR is specific for the APOE4 allele and was present and quantified in two of ten individuals.

(G) Presence of at least one APOE2 or APOE3 allele in all ten individuals.

(legend continued on next page)

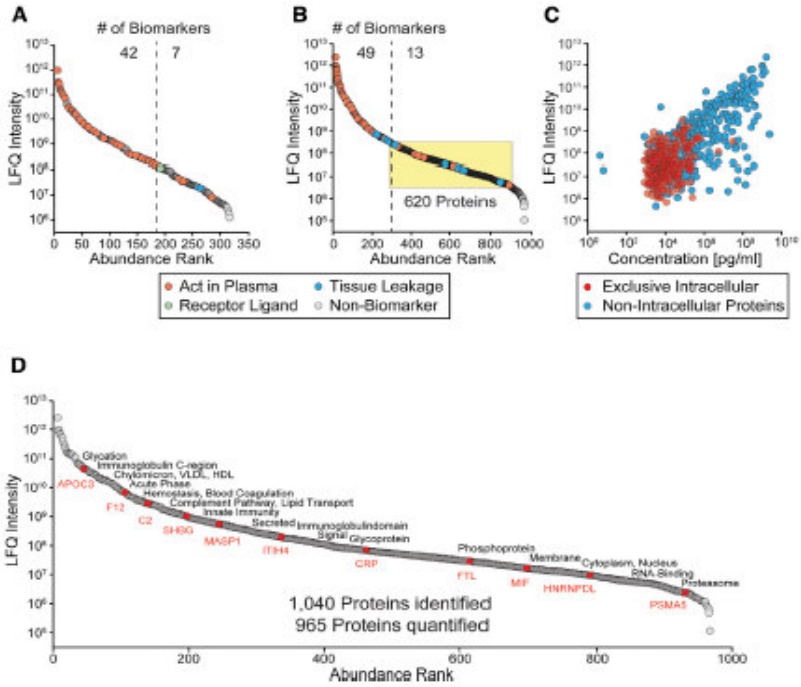


Figure 4. FDA-Approved Biomarker and Deep Plasma Proteomics

(A) Distribution of proteins quantified in the 20 min gradients of the 15 workflow replicates. FDA-approved biomarkers are color coded. The dashed line separate the regions densely populated and sparsely populated by biomarkers.

(B) LFQ intensities of 965 proteins quantified after separating peptides into eight fractions. The yellow rectangle encloses a 100-fold range, which contains the majority of the measured plasma proteome.

(C) Correlation of LFQ intensities of the deep plasma proteome data set and absolute concentrations from the Plasma Proteome Database.

(D) UniProtKB keyword annotations and their enrichment along the whole abundance range as determined by 1D enrichment (see main text). Exemplary proteins contributing to keywords are highlighted in red. APOC3, apolipoprotein C-III; F12, coagulation factor XII; C2, complement C2; MASP1, Mannan-binding lectin serine protease 1; ITIH4, inter-alpha-trypsin inhibitor heavy chain H4; HNRNPDL, heterogeneous nuclear ribonucleoprotein D-like; PSMA5, proteasome subunit alpha type 5.

classified as "tissue leakage," only 4 as "act in plasma," and 1 was not assigned to any category. A depth of 450 plasma proteins would be sufficient to cover 87% of the FDA-approved biomarkers present in our deep data set according to Anderson (2010).

The fact that we did not use protein depletion allowed us to investigate the quantitative nature of the plasma proteome. Bioinformatics analysis revealed that 457 of all quantified proteins had an extracellular annotation and 651 an intracellular one, with 221 overlapping proteins. Interestingly, the 430 proteins with exclusive intracellular annotation, which also have an independent abundance estimate in the Plasma Proteome Database (Nanjappa et al., 2014), are almost completely excluded from the top three order of magnitudes of protein abundance (Figure 4C). These proteins are likely of tissue origin and have been released by normal tissue damage, without necessarily having a function in blood. In contrast to the deep proteome, these intracellular proteins were largely absent in the 20 min measurements (25 of 313 proteins; Table S2). Reassuringly, 91% of proteins identified in plasma by us had also been identified in at least one of the studies collected in the PeptideAtlas repository (Farrah et al., 2014). In the absence of MS-derived quantitation of a deep, non-depleted proteome, we turned to the Plasma Proteome Database, which lists absolute concentrations of 597 of the proteins that are also quantified in our data set. Although these concentrations derive from the literature from a wide variety of individuals, health states, and quantification methods, we found a reasonable correlation, with an R^2 value of 0.53. This analysis also confirmed that we had quantified many proteins of clinical interest in the lower abundance range, such as the plasma protein ferritin (FTL) (12 ng/ml), which is widely used to diagnose dysregulation of iron homeostasis, or the cytokine macrophage migration inhibitory factor (MIF) (10 ng/ml). A total of 183 proteins in our data set have reported concentrations below 10 ng/ml.

To bioinformatically analyze the functional nature of the plasma proteome, we used the "1D annotation" algorithm in the MaxQuant software (Cox and Mann, 2008, 2012), to assign UniProtKB keywords to distinct abundance ranges. This resulted in 58 statistically significant features (Table S6). Classical characteristics of the plasma proteome were typically located in the high-abundance range. These include "glycation," "immunoglobulin," and "chylomicrons," which are expected because functional plasma proteins are typically glycated, a large proportion of functional plasma proteins are antibodies, and apolipoproteins are the structural components of chylomicrons. In the low-abundance tail of the distribution, we found highly abundant intracellular complexes such as the proteasome as well as RNA-binding and processing proteins. "Phosphoprotein" was situated close to the middle of the distribution, and above "membrane," "cytoplasm," and "nucleus," presumably because most intracellular proteins have by now been shown to be phosphorylatable, in addition to some of the extracellular ones. The

(H) Variation of the acute phase protein SAA1 in ten individuals.
(I) Variation of the inflammatory marker CRP in the same ten individuals.
(J) Direct comparison of two individuals to visualize the magnitude of SAA1 and CRP in the background of the other quantified plasma proteins.
(K) The CVs of 49 FDA-approved biomarkers from 15 workflow replicates as a function of protein abundance rank.

mean of functionally important Gene Ontology annotated biological processes, such as "protein-lipid complex assembly," "sterol and cholesterol transport," "acute phase response," and "regulation of coagulation" processes all scored in the upper third of the distribution, highlighting that these functions are overwhelmingly carried out by high-abundance plasma proteins. The above analysis can also be used to infer the likely function or lack thereof of a protein found at a certain concentration in normal plasma. For instance, the hormone-binding protein SHBG is in the upper range of the plasma proteome, which correlates well with its carrier function for an abundant circulating hormone (Ottosson et al., 1981) (Figure 4D).

DISCUSSION

Using state-of-the-art shotgun proteomics technology, in particular the recently described iST preparation (Kulak et al., 2014), Orbitrap instrumentation with very high sequencing speed (Kelstrup et al., 2014; Scheltema et al., 2014), and advances in label-free quantification (Cox et al., 2014), we here developed a streamlined and robust workflow for shotgun plasma proteomics. Sample preparation steps are minimized without loss of performance, and the procedure can be performed in 96-well format by a liquid handling platform. In this way, hundreds of plasma proteomes can be processed and sample preparation is not a limiting step for plasma proteomics in our workflow. We found that even extremely short measurements of 20 min still allowed the identification of more than 300 proteins, which was aided by a reference data set and the "match between runs" functionality. Accuracy and precision of the label-free workflow were excellent with intra-assay correlation of about $R^2 = 0.98$ and CVs smaller than 20% for the majority of quantified proteins.

Starting from only a finger prick of blood, the entire workflow, including database search and label-free quantification, can be performed in less than 3 hr. Previous plasma proteome studies typically started from milliliter amounts of blood, used depletion, and extensive pre-fractionation and therefore required days for completion (Cao et al., 2012; Keshishian et al., 2015; Liu et al., 2015; Such-Sanmartín et al., 2014). The ability to use small sample amounts makes blood testing much less invasive, improves cost-efficiency, and is clinically attractive in many situations, including the testing of infants as well as elderly patients (Bai et al., 2013). Likewise, fast response time is frequently important such as in the case of myocardial infarction. Our procedure uses a very short digestion time (60 min), which could be reduced by further optimization, so that the entire procedure could conceivably be performed in less than 1 hr.

Our very short LC-MS/MS runs contain nearly 50 proteins that are already subject to FDA-approved diagnostic tests, whereas the deep proteome only added few additional ones. Furthermore, the proportion of functional plasma proteins was very high, in contrast to the lower abundance range, which was dominated by tissue-derived "leakage proteins." Nevertheless, the deeper proteome still contained many proteins of known clinical significance, and it is interesting to speculate whether the relative paucity of approved biomarkers in this range is due to the greater difficulties associated with studying these proteins. Even in the short analysis runs, the lower half of the distribution has not yet been associated with specific patient states. We suggest that mixtures of recombinant isotope-labeled protein fragments, so-called SILAC-PrESTs (Edfors et al., 2014; Zeiler et al., 2012), could routinely be added to plasma samples. This would enable very high accuracy in absolute quantification for the discovery and validation of such biomarkers at high-throughput.

Further throughput improvements can be achieved with chemical labeling strategies, for instance with isobaric chemical tags such as TMT (Thompson et al., 2003). For 10-plex encoding, this could increase throughput to hundreds of patients per day. This compares favorably with metabolomics studies, which are already performed at large scale in plasma cohorts (Suhre et al., 2011), while providing equally useful and complementary information. Alternatively, TMT could be used to label patient samples before peptide pre-fractionation. This should result in deep proteome coverage, while keeping effective MS measurement time reasonably short at 1–2 hr per patient and compatible with large-scale studies.

The proteins characterized in our short workflow already contain a plethora of useful information. For example, it was easy to distinguish the gender of the donor and to obtain some risk-associated genotype information. The spectrum of apolipoproteins, as well as inflammatory markers, was excellently quantified, reflecting the cardiovascular and metabolic health state. Unexpectedly, the global nature of shotgun proteomics supplies us with valuable information about sample quality, which is not tested in routine clinical practice but can influence test results and medical decisions.

As mentioned above, the current strategy in plasma biomarker discovery by MS-based proteomics involves a narrowing down and widening strategy: a small number of patients and controls are analyzed in great depth with unbiased and relatively low-throughput methods. Resulting potential biomarkers are then envisioned to be validated with targeted MS-based methods or classical immunoassays in much larger cohorts (Anderson, 2014; Keshishian et al., 2015; Surinova et al., 2011).

Here we suggest an additional strategy, which we term "plasma proteome profiling." It consists of the measurement of large numbers of plasma proteomes at the greatest possible depth with streamlined and high-throughput technologies as described in this paper. This allows us to retain one of the basic attractions of unbiased, systemwide methodologies, namely, that associations do not have to be predefined but emerge naturally from "big data mining." Although our current work is only a first step in this direction, we believe that rapid development in the underlying technology will make this strategy more and more attractive. Given the low resource requirements, large cohorts could be investigated in the future, and one can even envision individuals routinely and repeatedly have their plasma proteome profile recorded. These high-dimensional profiles could indicate current disease risk as well as efficacy of lifestyle changes or pharmacological interventions and thereby contribute to individual and public health.

EXPERIMENTAL PROCEDURES

Tryptophan Fluorescence Emission Assay for Protein Quantification

Protein concentrations were determined after solubilizing of samples 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). From this, we estimated that 0.1 µg/µl tryptophan is equivalent to the emission of 7 µg/µl of human protein extract, assuming that tryptophan on average accounts for 1.3% of the human protein amino acid composition.

Blood Collection from Finger Pricks and Venous Blood Sampling

Blood was taken by lancets (Vitrex Sterilance Lite II) to obtain small quantities of capillary blood, and 5 µl of blood was transferred by a pipette into a pipette-tip-based centrifugal devices containing 0.56 µl 106 mM trisodium citrate (end concentration 10.6 mM trisodium citrate, as commonly used in blood collection tubes). The pipette-tip-based centrifugal device was made by melting the end of a pipette tip to seal it. When larger amounts of plasma were needed, blood was taken by venipuncture using a commercially available winged infusion set into collection tubes containing sodium citrate. The blood was centrifuged for 15 min at 2,000 × *g*, and plasma was harvested. Blood was sampled from healthy donors, who provided written informed consent, with prior approval of the ethics committee of the Max Planck Society.

Plasma taken by venipuncture was used to determine analytical and intra-assay variability, because in this case, larger amounts of plasma (15 µl) were needed.

Plasma for intra-individual variability was taken from one person by four finger pricks (at 6 a.m., 9 a.m., 12 p.m., and 3 p.m.) per day for 8 days. To determine the inter-individual variability, blood was taken by finger pricking of ten different individuals in triplicate (five women and five men), and samples were randomized within the gender groups. Furthermore, blood was taken from all ten fingers of one individual to one individual plasma and by venipuncture from two individuals for the comparison of plasma and serum.

Highly reliable plasma samples (Plasma^{Ref} Panels) were obtained from the blood bank Blutspendedienst des Bayerischen Roten Kreuzes.

High-Abundance Protein Depletion for Building a Matching Library

A combination of two immunodepletion kits was used for optimal removal of the 20 highest abundance plasma proteins with the purpose of establishing a peptide library for matching between runs (Nagaraj et al., 2012). First we used the Agilent Multiple Affinity Removal Spin Cartridge for removal of the top six high-abundance proteins (albumin, IgG, IgA, antitrypsin, transferrin, and haptoglobin), followed by ProteoPrep20 Plasma Immunodepletion Kit for the 20 highest abundance proteins from human plasma (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). Both depletion steps were carried out according to the manufacturer's instructions. The depleted plasma was digested and measured as described below. Raw data of the depleted plasma of one individual and undepleted plasma of ten different individuals served as a "library" for matching between runs for the 20 min gradients.

Sample Preparation: Protein Digestion and iST Purification

Sample preparation was performed as described previously (Kulak et al., 2014) with optimization for blood plasma as follows: 24 µl of SDC reduction and alkylation buffer (Kulak et al., 2014) were added to 1 µl of blood plasma. The mixture was boiled for 10 min to denature proteins. After cooling down to room temperature, the proteolytic enzymes LysC and trypsin were added in a 1:100 ratio (micrograms of enzyme to micrograms of protein). Digestion was performed at 37°C for 1 hr. Peptides were acidified to a final concentration of 0.1% trifluoro-acetic acid (TFA) for SDB-RPS binding, and 20 µg was loaded on two 14-gauge StageTip plugs. Ethylacetate/1% TFA (125 µl) was added, and the StageTips were centrifuged using an in-house-made StageTip centrifuge (a centrifuge with identical specifications is available from Sonation) for up to 2,000 × *g*. After washing the StageTips using two wash steps of 100 µl ethylacetate/1% TFA and one of 100 µl ddH_2O/0.2% TFA consecutively, purified peptides were eluted by 60 µl of elution buffer (80% acetonitrile, 19% ddH_2O, 1% ammonia) into auto sampler vials. The collected material was completely dried using a SpeedVac centrifuge at 45°C (Eppendorf, Concentrator plus). Peptides were suspended in buffer A* (2% acetonitrile, 0.1% TFA) and afterward sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510).

For the deep plasma data set, 20 µg purified and digested plasma peptides were fractionated using basic reversed-phase pre-fractionation. The peptides were loaded onto a reversed-phase C18 column (1.9 µm Reprosil-Pur C18 beads; Dr. Maisch) and were eluted using an EASY-nLC 1000 system (Thermo Fisher Scientific). A gradient was generated by using a dual-buffer system with buffer A (ddH_2O) and buffer B (ddH_2O, 80% ACN) adjusted to pH 10 with ammonium hydroxide. Peptides were separated and eluted from 5% B to 40% B in 50 min followed by a linear increase to 60% B in 10 min. The gradient was followed by a 12 min washout with 60%–95% B. We concatenated the 46 collected fractions into 8 fractions (concatenation scheme: 1 + 9 + 17 + 25, 2 + 10 + 18 + 26, etc.). A total of 1 µg of each concatenated fraction was loaded and measured by LC-MS/MS as described below.

Plasma samples from two individuals were dispensed into a 96-well plate (48 samples for each individual), and the complete sample preparation, with the exception of the centrifugation steps, was performed on an Agilent Bravo liquid handling platform.

Design, recombinant expression, purification and quantification of plasma PrESTs was as described in (Zeiler et al., 2012). Plasma PrESTs of the proteins APOA1, APOA4, APOB, APOE, and SHBG were combined in a master mix, which was added together with the SDC reduction and alkylation buffer to the blood plasma. The subsequent steps for sample preparation workflow are described above.

Ultra-High-Pressure LC and MS

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1000 ultra-high-pressure system (Thermo Fisher Scientific) coupled via a nano-electrospray ion source (Thermo Fisher Scientific) to a Q Exactive HF Orbitrap (Thermo Fisher Scientific) (Scheltema et al., 2014). Purified peptides were separated on 40 cm HPLC-columns (internal diameter 75 µm; in-house packed into the tip with ReproSil-Pur C18-AQ 1.9 µm resin; Dr. Maisch). For each LC-MS/MS analysis, about 1 µg peptides were used for 20 min runs and for each fraction of the deep plasma data set.

Peptides were loaded in buffer A (0.1% v/v formic acid) and eluted with a linear 15 min gradient of 10%–50% of buffer B (0.1% v/v formic acid, 60% v/v acetonitrile), followed by a 5 min 98% wash at a flow rate of 450 nl/min. Column temperature was kept at 60°C by a Peltier element-containing, in-house-developed oven, and parameters were monitored in real time by the SprayQC software (Scheltema and Mann, 2012). MS data were acquired with a Top5 data-dependent MS/MS scan method (topN method). 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 25 ms and a resolution of 60,000 at m/z 400. A 1.5 m/z isolation window and a fixed first mass of 100 m/z was used for MS/MS scans. Fragmentation of precursor ions was performed by higher energy C-trap dissociation with a normalized collision energy of 27 eV. MS/MS ssans were performed at a resolution of 15,000 at m/z 200 with an ion target value of 5×10^4 and a maximum injection time of 25 ms. Dynamic exclusion was set to 15 s to avoid repeated sequencing of identical peptides.

Data Analysis

MS raw files were analyzed by MaxQuant software version 1.5.2.10 (Cox and Mann, 2008), and peptide lists were searched against the human Uniprot FASTA database (version June 2014) and a common contaminants database by the Andromeda search engine (Cox et al., 2011) with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidations as variable modifications. The false discovery rate was set to 0.01 for both proteins and peptides with a minimum length of seven amino acids and was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin as protease, and a maximum of two missed cleavages were allowed in the database search. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm. Matching between runs was performed with depleted plasma and undepleted plasma of ten different individuals serving as a library. Proteins matching to the reversed database were filtered out. LFQ was performed with a minimum ratio count of 1 (Cox et al., 2014).

Bioinformatics Analysis

All bioinformatics analyses were performed with the Perseus software of the MaxQuant computational platform (Cox and Mann, 2008). Absolute quantification of protein abundances was computed using peptide label-free

quantification values, sequence length, and molecular weight (Cox et al., 2014). For enrichment analysis, a false discovery rate of <0.02 after Benjamini-Hochberg correction was used.

Statistical Analysis

Reproducibility was analyzed by calculating R^2 values for direct comparison of the LFQ intensities of any two LC-MS/MS runs. CV values were calculated on the basis of LFQ intensities. To determine the analytical and intra-assay variability, we used 15 raw data files, for intra-individual variation 32 files, and for inter-individual variation 30 files, and triplicates for each individual were combined before determining the CV.

ACCESSION NUMBERS

The accession number for the raw and processed data reported in this paper is PRIDE proteomeXchange: PXD002854.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.02.015.

AUTHOR CONTRIBUTIONS

Conceptualization, M.M., P.E.G., and N.A.K.; Methodology, P.E.G., N.A.K., G.P., L.M.H., and D.T.; Validation, P.E.G., N.A.K., and G.P.; Formal Analysis, P.E.G.; Investigation, P.E.G.; Writing – Original Draft: M.M., P.E.G., N.A.K., and G.P.; Supervision, M.M., L.M.H., and D.T.; Project Administration, M.M.; Funding Acquisition, M.M.

ACKNOWLEDGMENTS

We thank all members of the Proteomics and Signal Transduction Group for help and discussions and in particular Igor Paron, Korbinian Mayr, Gaby Sowa for MS technical assistance, Jürgen Cox for bioinformatic tools, and Niklas Grassl and Sean Humphrey for fruitful discussions. Nils Kulak and Garwin Pichler received an m^4 award from the Bio^M Munich Biotech Cluster funded by the Bavarian government. The work carried out in this study was partially supported by the Max Planck Society for the Advancement of Science and by the Novo Nordisk Foundation (grant NNF15CC0001).

Received: October 12, 2015
Revised: January 19, 2016
Accepted: February 24, 2016
Published: March 23, 2016

REFERENCES

Addona, T.A., Shi, X., Keshishian, H., Mani, D.R., Burgess, M., Gillette, M.A., Clauser, K.R., Shen, D., Lewis, G.D., Farrell, L.A., et al. (2011). A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. Nat. Biotechnol. *29*, 635–643.

Anderson, N.L. (2010). The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clin. Chem. *56*, 177–185.

Anderson, L. (2014). Six decades searching for meaning in the proteome. J. Proteomics *107*, 24–30.

Anderson, N.L., and Anderson, N.G. (2002). The human plasma proteome: history, character, and diagnostic prospects. Mol. Cell. Proteomics *1*, 845–867.

Anderson, N.L., Anderson, N.G., Pearson, T.W., Borchers, C.H., Paulovich, A.G., Patterson, S.D., Gillette, M., Aebersold, R., and Carr, S.A. (2009). A human proteome detection and quantitation project. Mol. Cell. Proteomics *8*, 883–886.

Bai, J.P., Barrett, J.S., Burckart, G.J., Meibohm, B., Sachs, H.C., and Yao, L. (2013). Strategic biomarkers for drug development in treating rare diseases and diseases in neonates and infants. AAPS J. *15*, 447–454.

Bellei, E., Bergamini, S., Monari, E., Fantoni, L.I., Cuoghi, A., Ozben, T., and Tomasi, A. (2011). High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. Amino Acids *40*, 145–156.

Cao, Z., Tang, H.Y., Wang, H., Liu, Q., and Speicher, D.W. (2012). Systematic comparison of fractionation methods for in-depth analysis of plasma proteomes. J. Proteome Res. *11*, 3090–3100.

Christensen, U., Simonsen, M., Harrit, N., and Sottrup-Jensen, L. (1989). Pregnancy zone protein, a proteinase-binding macroglobulin. Interactions with proteinases and methylamine. Biochemistry *28*, 9324–9331.

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

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

Cox, J., and Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics *13* (*Suppl 16*), S12.

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.

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.

Edfors, F., Boström, T., Forsström, B., Zeiler, M., Johansson, H., Lundberg, E., Hober, S., Lehtiö, J., Mann, M., and Uhlen, M. (2014). Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins. Mol. Cell. Proteomics *13*, 1611–1624.

Farrah, T., Deutsch, E.W., Omenn, G.S., Sun, Z., Watts, J.D., Yamamoto, T., Shteynberg, D., Harris, M.M., and Moritz, R.L. (2014). State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. J. Proteome Res. *13*, 60–75.

Geiger, T., Cox, J., and Mann, M. (2010). Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol. Cell. Proteomics *9*, 2252–2261.

Hua, S., Song, C., Geczy, C.L., Freedman, S.B., and Witting, P.K. (2009). A role for acute-phase serum amyloid A and high-density lipoprotein in oxidative stress, endothelial dysfunction and atherosclerosis. Redox Rep. *14*, 187–196.

Jenkins, A.J., Toth, P.P., and Lyons, T.J., eds. (2014). Lipoproteins in Diabetes Mellitus (Humana Press).

Jensen, M.K., Bertoia, M.L., Cahill, L.E., Agarwal, I., Rimm, E.B., and Mukamal, K.J. (2014). Novel metabolic biomarkers of cardiovascular disease. Nat. Rev. Endocrinol. *10*, 659–672.

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.

Keshishian, H., Burgess, M.W., Gillette, M.A., Mertins, P., Clauser, K.R., Mani, D.R., Kuhn, E.W., Farrell, L.A., Gerszten, R.E., and Carr, S.A. (2015). Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. Mol. Cell. Proteomics *14*, 2375–2393.

Krastins, B., Prakash, A., Sarracino, D.A., Nedelkov, D., Niederkofler, E.E., Kiernan, U.A., Nelson, R., Vogelsang, M.S., Vadali, G., Garces, A., et al. (2013). Rapid development of sensitive, high-throughput, quantitative and highly selective mass spectrometric targeted immunoassays for clinically important proteins in human plasma and serum. Clin. Biochem. *46*, 399–410.

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.

Liu, Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., Cheng, L.Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T.D., et al. (2015). Quantitative variability of 342 plasma proteins in a human twin population. Mol. Syst. Biol. *11*, 786.

Martínez-Morillo, E., Nielsen, H.M., Batruch, I., Drabovich, A.P., Begcevic, I., Lopez, M.F., Minthon, L., Bu, G., Mattsson, N., Portelius, E., et al. (2014). Assessment of peptide chemical modifications on the development of an accurate and precise multiplex selected reaction monitoring assay for apolipoprotein e isoforms. J. Proteome Res. *13*, 1077–1087.

Muñoz, J., and Heck, A.J. (2014). From the human genome to the human proteome. Angew. Chem. Int. Ed. Engl. *53*, 10864–10866.

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

Nanjappa, V., Thomas, J.K., Marimuthu, A., Muthusamy, B., Radhakrishnan, A., Sharma, R., Ahmad Khan, A., Balakrishnan, L., Sahasrabuddhe, N.A., Kumar, S., et al. (2014). Plasma Proteome Database as a resource for proteomics research: 2014 update. Nucleic Acids Res. *42*, D959–D965.

Omenn, G.S. (2005). Exploring the Human Plasma Proteome, *Volume 5* (John Wiley).

Ottosson, U.B., Damber, J.E., Damber, M.G., Selstam, G., Solheim, F., Stigbrand, T., Södergård, R., and von Schoultz, B. (1981). Effects of sex hormone binding globulin capacity and pregnancy zone protein of treatment with combinations of ethinyl-oestradiol and norethisterone. Maturitas *3*, 295–300.

Paczesny, S., Braun, T.M., Levine, J.E., Hogan, J., Crawford, J., Coffing, B., Olsen, S., Choi, S.W., Wang, H., Faca, V., et al. (2010). Elafin is a biomarker of graft-versus-host disease of the skin. Sci. Transl. Med. *2*, 13ra2.

Sand, O., Folkersen, J., Westergaard, J.G., and Sottrup-Jensen, L. (1985). Characterization of human pregnancy zone protein. Comparison with human alpha 2-macroglobulin. J. Biol. Chem. *260*, 15723–15735.

Schaab, C., Geiger, T., Stoehr, G., Cox, J., and Mann, M. (2012). Analysis of high accuracy, quantitative proteomics data in the MaxQB database. Mol. Cell. Proteomics *11*, M111.014068.

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

Scheltema, R.A., Hauschild, J.P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014). The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. Mol. Cell. Proteomics *13*, 3698–3708.

Such-Sanmartín, G., Ventura-Espejo, E., and Jensen, O.N. (2014). Depletion of abundant plasma proteins by poly(N-isopropylacrylamide-acrylic acid) hydrogel particles. Anal. Chem. *86*, 1543–1550.

Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., et al.; CARDIoGRAM (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature *477*, 54–60.

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

Tanzi, R.E. (2012). The genetics of Alzheimer disease. Cold Spring Harb. Perspect. Med. *2*, 2.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. *75*, 1895–1904.

Tu, C., Rudnick, P.A., Martinez, M.Y., Cheek, K.L., Stein, S.E., Slebos, R.J., and Liebler, D.C. (2010). Depletion of abundant plasma proteins and limitations of plasma proteomics. J. Proteome Res. *9*, 4982–4991.

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Veterinary Medicine (2001). Guidance for Industry, Bioanalytical Method Validation (U.S. Department of Health and Human Services).

Végvári, A., Welinder, C., Lindberg, H., Fehniger, T.E., and Marko-Varga, G. (2011). Biobank resources for future patient care: developments, principles and concepts. J. Clin. Bioinforma. *1*, 24.

Vihko, P., Sajanti, E., Jänne, O., Peltonen, L., and Vihko, R. (1978). Serum prostate-specific acid phosphatase: development and validation of a specific radioimmunoassay. Clin. Chem. *24*, 1915–1919.

Wilson, P.W., Pencina, M., Jacques, P., Selhub, J., D'Agostino, R., Sr., and O'Donnell, C.J. (2008). C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. Circ Cardiovasc Qual Outcomes *1*, 92–97.

Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M., and Mann, M. (2012). A protein epitope signature tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. Mol. Cell. Proteomics MCP *11*, O111.009613.

3.2. Article 2: Proteomics Reveals the Effects of Sustained Weight Loss on the Human Plasma Proteome

Authors: Philipp E. Geyer[1,2,†], Nicolai J. Wewer Albrechtsen[2,3,4,†], Stefka Tyanova[1], Niklas Grassl[1], Eva W. Iepsen[3,4], Julie Lundgren[3,4], Sten Madsbad[4,5], Jens J. Holst[3,4], Signe S. Torekov[3,4], and Matthias Mann[1,2]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[3]Institute of Laboratory Medicine, Ludwig-Maximilians University Munich, 80539 Munich, Germany
[†]These authors contributed equally to this work.

The 'weight loss study' is our first clinical study, and it demonstrates that Plasma Proteome Profiling can live up to its promises. Before it was neither possible to measure large cohorts nor to find biological meaningful information in the plasma proteome. In this study we analyzed the largest cohort in the field of plasma proteomics with almost 1,300 separately prepared and measured plasma samples.

Weight loss and sustained weight maintenance are of central concern in modern society, research and medicine. Obesity and the metabolic syndrome are major public health burdens, predisposing to several diseases, including type 2 diabetes and cardiovascular diseases and increasing the overall likelihood of early death. However, not everyone agrees on how universal the positive effects of weight loss on cardiovascular and metabolic risk factors are. We investigated a longitudinal cohort of 52 obese study participants by measuring their plasma proteomes over an initial weight loss period of 8 weeks, followed by one year of weight maintenance.

Applying a matching library strategy by using double depleted plasma, we were able to quantify 437 proteins per individual. The obtained Plasma Proteome Profiles revealed the comprehensive systemic effects of weight loss on individual plasma proteins: Of a total of 737 investigated proteins, 63 were decreased and 30 were increased directly in response to weight loss. The longitudinal study design allowed us to monitor long-term regulation of proteins. We were able to follow the reduction of fat mass by the adipocyte secreted protein SERPINF1 which – together with the apolipoprotein F (APOF) – was the most significantly regulated protein in weight loss. Comprehensive quantification of 18 members of the apolipoprotein family – the main lipid homeostasis mediators – delivered information on metabolic and cardiovascular risk factors that were strongly influenced by weight loss.

The meta-data available in the study supplied us with physiological and clinical laboratory parameters, which we correlated with the plasma proteomes, establishing novel dependencies. Remarkably, a panel of eight plasma proteins showed a higher correlation with insulin resistance than the known biomarker adiponectin. Moreover, we defined an inflammation panel consisting of proteins that was assessed for each study participant. By combining these data on an individual-resolved level, we connected low-grade inflammation and insulin resistance. Most of the individuals with high levels of inflammation also had an unfavorable insulin resistance profile, but importantly, individuals in all groups benefited from weight loss.

With this study we demonstrated that it is possible to measure large cohorts and to extract biologically and medically meaningful information in the human plasma proteome. Another aim of this study was to identify bottlenecks for further optimization. In this regard, we found that most of the down time was caused by HPLC issues that were not connected to the sample quality itself. Higher robustness in this area would directly result in higher throughput and clinical applicability.

Published online: December 22, 2016

Article

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Proteomics reveals the effects of sustained weight loss on the human plasma proteome

Philipp E Geyer[1,2,†], Nicolai J Wewer Albrechtsen[2,3,4,†], Stefka Tyanova[1], Niklas Grassl[1], Eva W Iepsen[3,4], Julie Lundgren[3,4], Sten Madsbad[4,5], Jens J Holst[3,4], Signe S Torekov[3,4] & Matthias Mann[1,2,*]

Abstract

Sustained weight loss is a preferred intervention in a wide range of metabolic conditions, but the effects on an individual's health state remain ill-defined. Here, we investigate the plasma proteomes of a cohort of 43 obese individuals that had undergone 8 weeks of 12% body weight loss followed by a year of weight maintenance. Using mass spectrometry-based plasma proteome profiling, we measured 1,294 plasma proteomes. Longitudinal monitoring of the cohort revealed individual-specific protein levels with wide-ranging effects of losing weight on the plasma proteome reflected in 93 significantly affected proteins. The adipocyte-secreted SERPINF1 and apolipoprotein APOF1 were most significantly regulated with fold changes of −16% and +37%, respectively ($P < 10^{-13}$), and the entire apolipoprotein family showed characteristic differential regulation. Clinical laboratory parameters are reflected in the plasma proteome, and eight plasma proteins correlated better with insulin resistance than the known marker adiponectin. Nearly all study participants benefited from weight loss regarding a ten-protein inflammation panel defined from the proteomics data. We conclude that plasma proteome profiling broadly evaluates and monitors intervention in metabolic diseases.

Keywords diabetes; mass spectrometry; metabolic syndrome; obesity; plasma proteome profiling

Subject Categories Metabolism; Post-translational Modifications, Proteolysis & Proteomics; Systems Medicine

DOI 10.15252/msb.20167357 | Received 29 September 2016 | Revised 6 December 2016 | Accepted 7 December 2016

Mol Syst Biol. (2016) 12: 901

Introduction

Obesity and the metabolic syndrome represent a major public health burden, predisposing to several diseases including type 2 diabetes and cardiovascular syndromes and increasing the overall likelihood of early death (Eckel *et al*, 2005; Grundy, 2015). The chances of developing the metabolic syndrome can be reduced considerably by sustained weight loss in obese individuals, through its positive effects on a broad range of metabolic risk factors (Hansen & Bray, 2008). However, it is not entirely clear how weight loss exerts these beneficial effects and to what extent they may differ between individuals (Look *et al*, 2013). The metabolic state is reflected in the levels of lipid transport proteins in the blood, most prominently the apolipoprotein family that is involved in lipid turnover. Several apolipoproteins, for instance A1 and B, correlate with cholesterol and triglycerides (Dominiczak & Caslake, 2011). Obesity is also associated with increased systemic low-grade inflammation, as indicated by plasma levels of specific markers such as C-reactive protein (Esser *et al*, 2014). These proteins are normally quantified individually by antibody-based assays, providing only a partial picture of changes in the entirety of proteins in this body fluid, the plasma proteome. In the case of weight loss, particular proteins like sex hormone-binding globulin are known to change (Azrad *et al*, 2012), but a global view of the dynamic changes in the plasma proteome is currently lacking.

Human blood plasma and serum are the predominant matrices for clinical analysis as they are easily accessible and clearly reflect an individual's metabolism. Mass spectrometry (MS)-based proteomics should be an optimal technology to investigate changes in the human plasma proteome, because this holistic approach can in principle yield specific and quantitative information on all proteins in an unbiased way. Due to several technological challenges, including the large "dynamic range" (the difference between most abundant and least abundant proteins), the proteomic analysis of plasma has remained a very specialized endeavor, precluding the analysis of large numbers of individual plasma proteomes (Anderson, 2010, 2014). The technology of MS-based proteomics has drastically improved over the last years (Mann *et al*, 2013; Zubarev & Makarov, 2013; Munoz & Heck, 2014; Aebersold & Mann, 2016), and several groups have reinvestigated the plasma proteome recently (Liu *et al*, 2015; Cominetti *et al*, 2016;

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 Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
4 NNF Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
5 Department of Endocrinology, Hvidovre University Hospital, Hvidovre, Denmark
*Corresponding author. Tel: +49 89 8578 2557; E-mail: mmann@biochem.mpg.de
†These authors contributed equally to this work

Published online: December 22, 2016
Molecular Systems Biology

Malmstrom *et al*, 2016). Our laboratory has developed an automated, rapid, and robust shotgun proteomics workflow that allows the streamlined analysis of hundreds of plasma proteins from a single drop of blood, a technology that we call "plasma proteome profiling" (Geyer *et al*, 2016). These profiles provide quantitative information on the majority of the classical, functional plasma proteins (Surinova *et al*, 2011), and we speculated that the metabolic status of individuals during weight loss and maintenance would be reflected by their plasma proteomes. We selected a longitudinal prospective cohort (Iepsen *et al*, 2015) from which we measured the plasma proteomes of 43 individuals at seven time points over 14 months. This allowed us to analyze the global changes related to lipid metabolism and inflammatory processes resulting from a fundamental life style change in the plasma proteome for the first time.

Results

Measurement of 1,294 plasma proteomes in a weight loss study

We recently described a highly sensitive proteomics sample preparation method that can be performed with a minimum number of steps in a single reaction vial (Kulak *et al*, 2014). On this basis, we subsequently developed an automatable workflow for plasma, which allows the robust measurement of this challenging body fluid in < 1 h (Geyer *et al*, 2016). We reasoned that this technology might enable analysis of relatively large studies, involving longitudinal monitoring of a substantial cohort.

To investigate the biological impact of losing weight on the human plasma proteome, we made use of a study in which 52 obese individuals were enrolled for an 8-week-long, diet-induced weight loss intervention of 800 kcal/day during which they lost on average 12% body mass. A total of 43 of these individuals were followed for an additional year of successful weight maintenance (Iepsen *et al*, 2015). We obtained plasma from subjects that were fasted overnight, sampled before and after weight loss as well as at five time points over the subsequent year (Fig 1A).

For the purpose of constructing an MS proteome library, we doubly depleted reference plasma samples from three healthy women and three healthy men for the 20 most abundant plasma proteins. The resulting data files were used to increase the depths of analysis in the subsequent cohort measurements by transferring peptide identifications between liquid chromatography tandem mass spectrometry (LC-MS/MS) runs (Geyer *et al*, 2016). For accurate label-free quantification, we measured quadruplicates of 319 plasma samples of the cohort. The resulting 1,294 plasma proteome measurements (including 18 samples for library construction) represent to our knowledge by far the largest plasma proteomics study in a clinical context (Fig 1B). Data acquisition could be accomplished in a reasonable time (10 weeks), and interestingly, we found that remaining challenging points in the LC-MS/MS measurements were concentrated on the chromatographic rather than the MS side (Fig EV1A).

Across the individuals, we identified 737 plasma proteins (subtracting contaminants such as keratins) and an average of 437 (± 23) per individual. Quantitative accuracy was high as reflected by a mean Pearson correlation coefficient of 0.97 for quadruplicate measurements (Fig EV1B).

As shotgun proteomics is not limited to the analysis of a predefined set of proteins, our measurements contained additional information, for instance, on sample quality (Geyer *et al*, 2016). Erythrocyte lysis, which is indicated by increased levels of highly abundant erythrocyte-specific proteins (HBA1, HBB, HBD, CA1), occurred only in one sample, and minor coagulation events during blood taking were present in five of the 318 samples (Fig EV1C and D). This suggests excellent sample handling procedures throughout the study.

Plasma protein levels are individual-specific

Study participants were followed longitudinally for 1 year after weight loss, which enabled us to identify individual-specific protein levels as well as intra-individual variability of the plasma proteome profiles. For this analysis, we removed the first two time points, which covered the weight loss intervention (weeks −8 and 0) to minimize weight loss-induced effects. For each of 448 proteins that were quantified in all five time points of at least one individual, we calculated the average level per person and for the entire cohort. Strikingly, 69% of proteins differed from the group average more than twofold and 25% more than fivefold (Fig 2A and Table EV1). To investigate the individual variations across time points, we additionally determined longitudinal protein-specific coefficients of variation (CVs). Volcano plots visualize the proteins with a minimum difference from the group average and a maximum variation in the individual. For instance, in participant 4, pregnancy zone protein (PZP) was 26-fold higher than the group average, but varied less than 10% over time, making it a highly individual-specific protein by these criteria (Fig 2B). For the whole dataset, at a twofold difference and 30% CV cutoff, 46% of all proteins would be individual-specific. This represents a lower limit because any measurement errors would tend to decrease the apparent number of individual-specific proteins. We also note that "individual-specific" contains "sub-group"-specific proteins, because each individual could be a representative of a sub-group.

Overall, protein levels tended to vary considerably between participants, but to remain quite constant over time within each individual, as exemplified by seven individual-specific proteins in Fig 2C. Alpha-2-macroglobulin (A2M) is 10 times higher in some individuals compared to others, but has a mean CV of 6% over the 48 weeks for all of the individuals. The high inter-individual variation in lipoprotein(a) (LPA) over three orders of magnitude has a genetic reason: Individuals vary in the number of LPA kringle domains and secretion into the circulation depends on the size of the protein (Utermann, 1989). Some proteins were detected in only a minority of individuals, for instance the intracellular protein Rab GDP dissociation inhibitor (GDI1/2). It was robustly quantified but only in a single person and may therefore be present in this individual's blood due to tissue leakage. A few proteins including complement factor C3, serum albumin (ALB), vitamin D-binding protein (GC), kininogen-1 (KNG1), hemopexin (HPX), complement factor H (CFH), and clusterin (CLU) show very low variation between individuals and over time (fold difference < 1.3; CV < 20%), indicating tight biological control of their levels (Table EV2).

Published online: December 22, 2016
Philipp E Geyer et al Proteome profiling during sustained weight loss Molecular Systems Biology

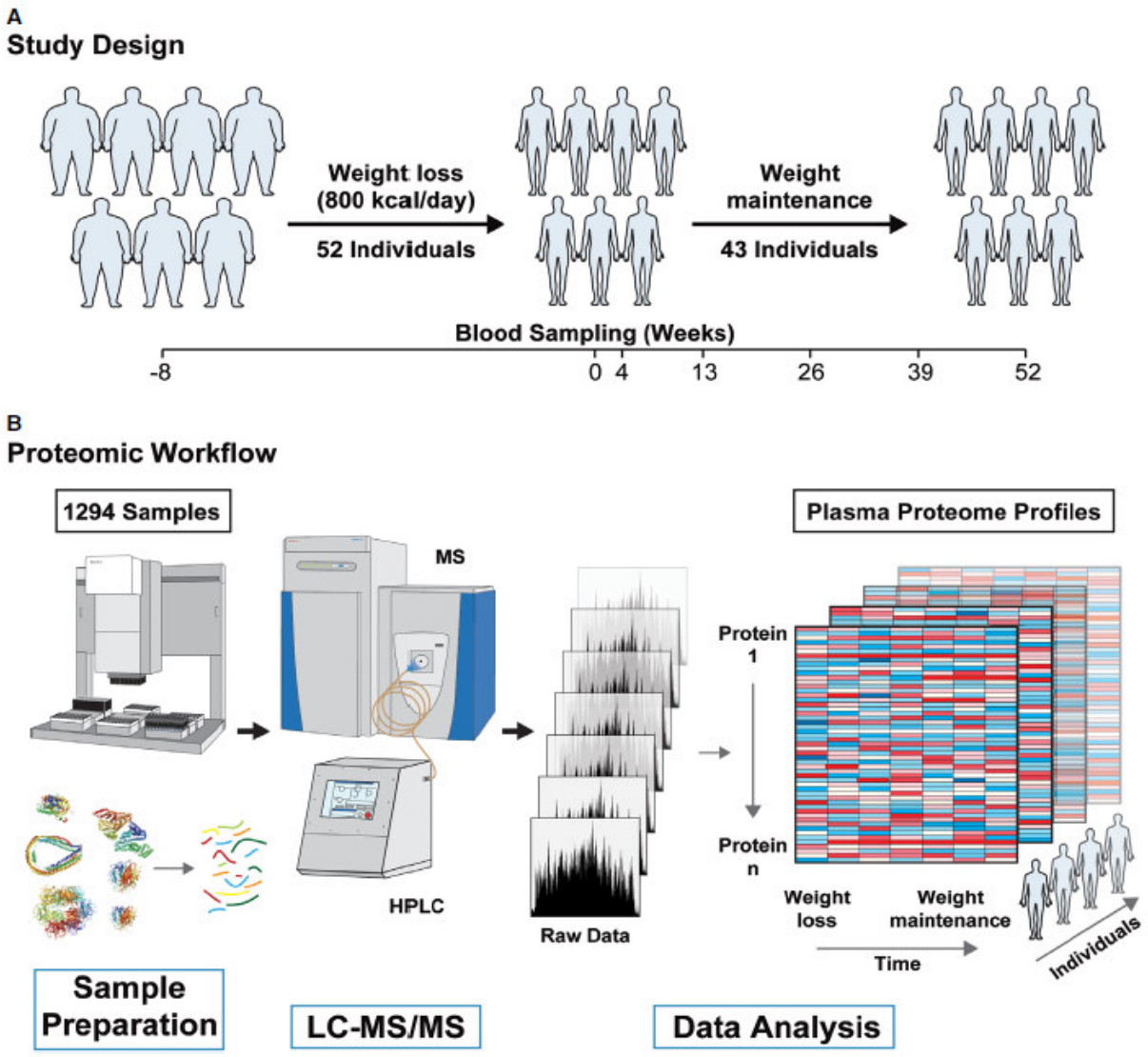


Figure 1. Study design and the plasma proteome profile pipeline.

A The study cohort consisted of 52 obese individuals, who lost on average 12% of their body mass during 8 weeks of calorie restriction (800 kcal/day). The acute weight loss was followed by a 52-week weight maintenance period by 43 of the study participants with longitudinal blood sampling at the indicated time points.

B Quadruplicates of the samples and the establishment of a matching library resulted in 1,294 plasma proteomes, which were separately prepared by an automated liquid handling platform. The LC-MS/MS data, which we analyzed by MaxQuant and Perseus, resulted in 319 individual plasma proteome profiles for 52 participants.

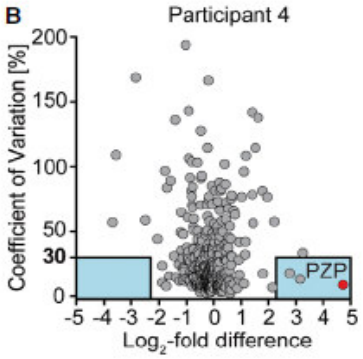
Weight loss changes the plasma proteome profile

Focusing on the effect of weight loss, we analyzed the plasma proteome changes of the study participants from before weight loss to after weight loss (baseline to the 8-week time point). We used a one-sample *t*-test to take individual-specific protein levels into account. Weight loss had a comprehensive systemic effect on the blood plasma proteome profile with 63 decreased and 30 increased protein levels; however, the magnitude of the changes was not large (Fig 3A and Table EV3). For instance, apolipoprotein F (APOF) and inter-alpha-trypsin-inhibitor heavy chain H3 (ITIH3) displayed extremely significantly changing protein levels (both $P < 10^{-13}$) and they increased by 37 and 34%, respectively (Fig 3B). Pigment epithelium-derived factor (SERPINF1) changed with a similar

Published online: December 22, 2016
Molecular Systems Biology

A

Fold difference	Percentage of all proteins			
	All	CV<10%	CV<20%	CV<30%
>1.2	100	69	92	96
>1.5	89	39	71	79
>2	69	15	37	46
>3	45	7	17	21
>4	32	5	10	12
>5	25	3	7	9



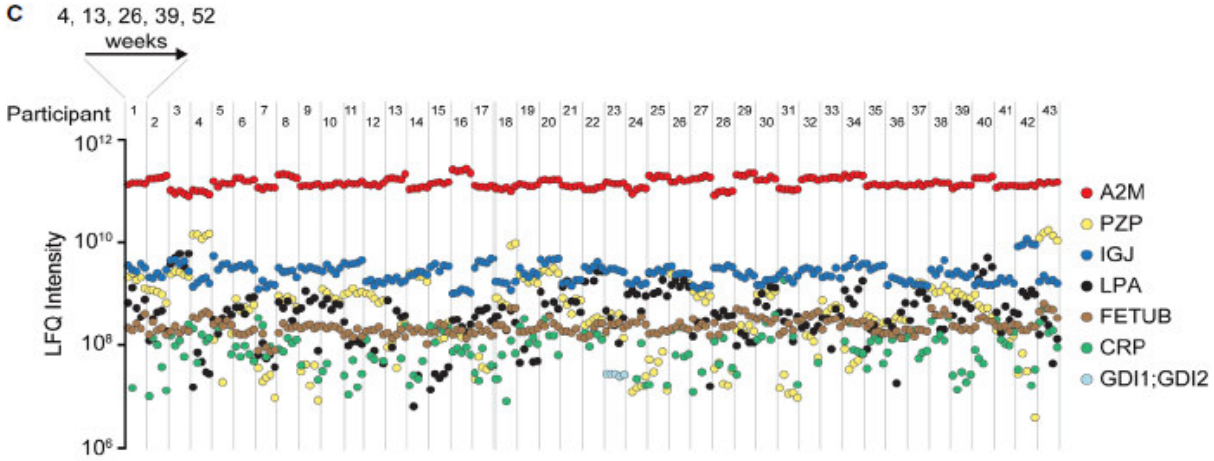


Figure 2. Individual-specific plasma protein levels.

A Coefficients of variation (CVs) for all proteins were calculated in all participants for the five longitudinal samples. Combination of these longitudinal CVs and fold differences allows estimation of how many proteins are individual-specific.

B CVs plotted against the $\log_2$-fold difference of one participant compared to the average label-free quantitation (LFQ) intensity of the study cohort. Considering a fivefold difference and a CV below 30% yields the proteins in the blue boxes as specific for this participant.

C LFQ intensities of seven proteins plotted during weight maintenance (weeks 4–52) for all 43 individuals. These proteins are stable over time within individuals, but strongly vary between individuals. A2M: alpha-2-macroglobulin; PZP: pregnancy zone protein; IGJ: immunoglobulin J chain; LPA: apolipoprotein(a); FETUB: fetuin-B; CRP: C-reactive protein; GDI1/GDI2: Rab GDP dissociation inhibitor alpha/beta.

significance, and here, the average fold difference was only −16%. SERPINF1 is known to be secreted by adipocytes (Wang *et al*, 2004), highlighting the ability of plasma proteomics to pinpoint biologically meaningful but very small changes. Albumin itself, which constitutes about half of plasma proteome mass, increased highly significantly ($P < 10^{-10}$), but only by 8%. Sex hormone-binding globulin (SHBG) changed most strongly due to weight loss with an increase of 117%. A similar effect of weight loss on SHBG levels has been observed before by non-proteomic analysis, serving as a further positive control of our results (Azrad *et al*, 2012). Nevertheless, in all these cases, there are some individuals that deviated from the rest of the cohort. For SHBG, this might be due to its dependence on estrogen levels and thereby age and gender, illustrating the richness of information potentially encoded in the plasma proteome (Geyer *et al*, 2016). Corticosteroid-binding globulin (SERPINA6), which binds 80% of circulating cortisol, is increased by 12% upon weight loss and together with greater albumin levels may contribute to the decrease in freely circulating cortisol levels upon weight loss (Lewis *et al*, 2005).

Long-term effect of weight loss due to weight maintenance

Having determined individual-specific and acute weight loss-induced proteins, we next investigated the dynamics of the plasma proteome profile over the 1-year weight maintenance period. We considered proteins that changed highly significantly ($P < 5 \times 10^{-4}$) between baseline and at least one time point after weight loss. Proteins that could have been introduced during the blood sampling procedure, such as keratins, were excluded. In total, 84 proteins fulfilled these criteria. According to their behavior, they clustered into

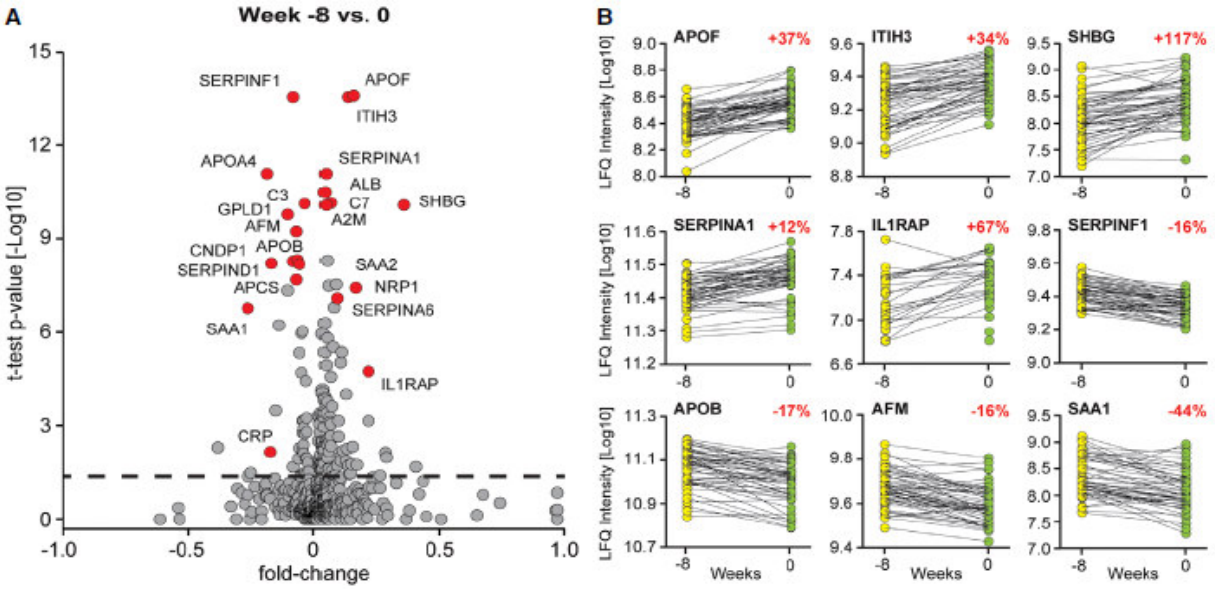


Figure 3. The effect of weight loss.

A Volcano plot of proteomes before (week −8) and directly after weight loss (week 0), with the *x*-axis depicting the fold change in protein levels and the *y*-axis the $-\log_{10}$ *t*-test *P*-value of the quadruplicates.

B Changes in protein abundance are shown for each individual. LFQ intensities from before (yellow) and after weight loss (green) are connected by a line to highlight individual-specific protein levels and their changes. The median increase or decrease for each protein over the population is indicated in red.

seven groups, reflecting different aspects of functional adaption to sustained weight loss (Fig 4).

Proteins in the first group decreased rapidly in response to acute weight loss and steadily reverted toward their initial levels over the weight maintenance period. These proteins change transiently in response to the acute energy-deficient state and appear to be regulated toward steady-state levels independent of body weight and fat mass. They included APOA4, afamin (AFM), an alternative transporter for vitamin E, and serum paraoxonase (PON1). PON1, the levels of which are normally determined by enzymatic activity assays, is associated with HDL and is known to exhibit lower activity in obese children and adults (Seres *et al*, 2010; Ceron *et al*, 2014).

Group 2 comprised proteins of special clinical interest as they decreased concomitantly with weight loss and remained low during weight maintenance. They are mainly apolipoproteins (APOA2, APOB, APOC3) and markers of inflammation (CRP, SAA1, SAA4, and ORM2) and will be discussed below. Further members of these two groups were present in a related cluster (group 4) and also exhibited a long-term decrease, but their levels take many months to reach equilibrium. Seven increased and 13 decreased proteins had very significant long-term effects during weight loss (Fig EV2A and B). Apart from the known upregulation of SHBG and downregulation of CRP, novel findings include a near doubling over the entire year of follow-up of the levels of neurophilin-1 (NRP1), whose soluble isoform can bind vascular endothelial growth factor A (VEGFA) (Soker *et al*, 1998). Vitamin D-binding protein (GC) is another protein that was significantly and persistently upregulated. Proteoglycan 4 (PRG4), normally associated with lubrication of articulating joints, showed a significant long-term decrease upon weight loss (median: −19%), as did heparin cofactor 2 (SERPIND1) (median: −9%), a thrombin inhibitor.

Overall, groups 1, 2, 3 and 7 showed rapid changes in response to lower body weight. In contrast, the adaption of the proteins in groups 4 and 6 was much slower. Zinc-alpha-2-glycoprotein (AZGP1) is an example of a protein that increased slowly over time. It stimulates lipid degradation in adipocytes (Hirai *et al*, 1998), and its sustained elevation may reflect an improved lipid turnover. Likewise, the long-term decrease in the adipocyte-secreted protein SERPINF1 very closely parallels the rapid loss of fat mass followed by long-term maintenance (Fig EV2C).

Related proteins tended to be in the same or related clusters, arguing for the technical reliability of the results and suggesting common and biologically meaningful regulation. For instance, subunits of the same proteins clustered closely together such as C1QA, B, and C. Examples of functionally related proteins include the two serine protease inhibitors SERPINA1 and SERPINA3 (alpha-1-antitrypsin and alpha-1-antichymotrypsin) in group 3 and functionally related members of the complement cascade in group 4 (C4A, C4B, CFI, C8G, C8B, CFB).

The participants of the study had been randomized to treatment with the incretin GLP-1, which showed a significant but small difference in weight loss (Iepsen *et al*, 2015). However, this difference did not correlate with statistically significant changes in our plasma proteome measurements.

Published online: December 22, 2016
Molecular Systems Biology

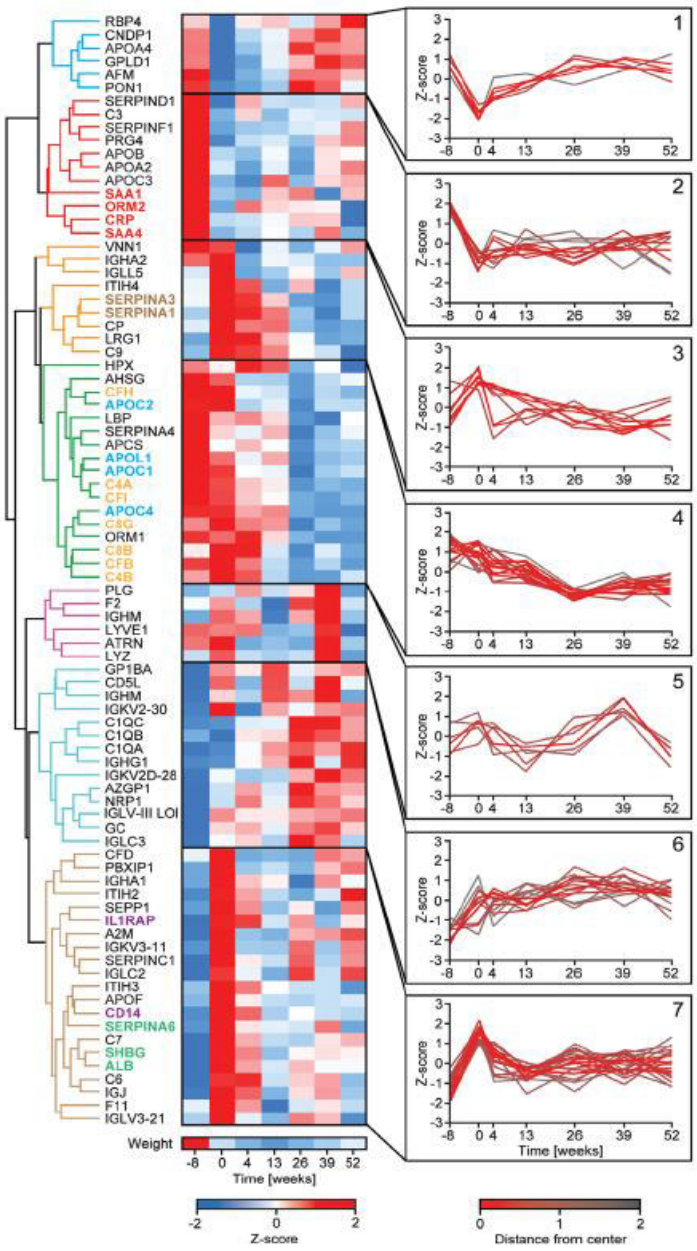


Figure 4. Long-term effects of weight loss on the plasma proteome profile.
Hierarchical clustering of *Z*-scored median LFQ intensities for highly significant proteins ($P < 0.0005$) resulted in seven longitudinal weight regulated protein clusters. Scale bar for *Z*-scored weight is displayed below the protein clusters. Insets show *Z*-scores for proteins in different clusters as a function of time, color-coded for the distance from the center. The highlighted protein names with the same color indicate functionally connected proteins (red: inflammatory markers; brown: serine protease inhibitors; orange: complement system; blue: apolipoproteins; purple: anti-inflammatory-acting proteins; green: steroid transport proteins).

Published online: December 22, 2016

The effect of weight loss on systemic inflammation factors

Our results show that weight loss changes multiple components of the plasma proteome. Several acute phase proteins were downregulated. CRP, SAA1, SAA4, and ORM2 cluster closely together in group 2, indicating a fast response upon weight loss (Fig 4). Directly after weight loss, SAA1 and CRP, two prominent risk markers for cardiovascular disease, showed median decreases of 43% and 35%, respectively. ORM1, APCS and LBP decreased more gradually over time (16%, 10% and 16%, respectively, at week 52) and are thus part of group 4. SERPINA1 and SERPINA3 (alpha-1-antitrypsin, alpha-1-antichymotrypsin) are also categorized as acute phase proteins and they are upregulated initially after weight loss (week 0) and start to decline afterward. Altogether, the cluster of highly significant proteins contained a group of 15 complement factors of the classical and alternate complement pathways.

Consistent with decreasing systemic inflammation upon weight loss, the soluble form of the anti-inflammatory protein, interleukin-1 receptor accessory protein (IL1RAP), a known antagonist of the major pro-inflammatory cytokine interleukin-1 (IL-1) (Smith *et al*, 2003), rose upon acute weight loss by an average of 67%. However, this decreased to 18% at the end of the weight maintenance period. Soluble CD14 has been reported to dampen inflammation (Thompson *et al*, 2003), and its levels also increased due to weight loss, but reverted nearly to baseline after 1 year of weight maintenance.

Next, we correlated the quantified plasma proteins with classical laboratory parameters including BMI, HDL, LDL, cholesterol, triglyceride levels, and insulin resistance (HOMA-IR) to investigate whether they were mirrored in the plasma proteome (Fig 5A–F and Table EV4 and see below). Remarkably, of all proteins in our dataset, the five proteins most significantly correlating with BMI were inflammation factors (CFH, C3, APCS, ORM2, and CFI; Fig 5A). For each, the *P*-value was lower than 10^{-8} and Pearson correlation coefficients ranged from 0.3 to 0.4. Five other inflammation-related proteins (CRP, SAA4, ORM1, ATRN, and CFB) also correlated significantly with BMI (Table EV5). ATRN (attractin) is a dipeptidase involved in inflammatory responses, but has also been linked to obesity (Duke-Cohan *et al*, 1998; Laudes *et al*, 2010).

Serum amyloid P component (APCS) correlated with HDL, as reported previously (Li *et al*, 1998), but otherwise we found no significant dependency of the above-mentioned inflammatory proteins to other clinical parameters, perhaps because these were only available at three time points (Fig 5B–F).

To calculate a longitudinal systemic inflammation profile for all individuals, we filtered the highly significantly changing proteins (Fig 4) for the keywords "acute phase", "inflammatory response", and "immunity", which resulted in 23 proteins (C1QA/B/C counted as one). From this list, we removed the anti-inflammatory protein CD14 and added two further non-annotated, but known acute phase proteins (APCS and LBP), which were not keyword annotated for the filtered terms. The median MS intensity of the resulting 24 inflammation-related proteins was *Z*-scored, and we calculated the slope over time for each protein. Of these, 20 inflammation proteins had a negative slope (decreased levels due to weight loss) and 10 further significantly correlated with BMI (Fig 5A and Table EV5). We *Z*-scored each protein of the ten-protein panel over the individual time series to make them comparable, followed by hierarchical clustering on the level of the study participants. The resulting heat map is a longitudinal inflammation profile for each of the 42 individuals (Fig 5G).

High levels (red color) are clearly predominant before or directly after weight loss (left side of the heat map), whereas low values are mainly found at the later time points. A group of seven participants is clustered at the top of the heat map and is distinguished by several red patches indicating raised inflammation levels at several time points during weight maintenance. This was not connected to regain of weight, suggesting infection as the cause. For instance, in participant 31, levels of CRP were 28-fold and SAA1 54-fold increased at week 13 compared to her average levels. Focusing on the central 72% of the inflammatory profiles, we calculated the median level of the ten-protein panel and plotted it over time. This revealed that the inflammatory state decreased substantially from before weight loss at week −8 until week 13 and stayed constant at the lower level from then on (Fig 5H).

In addition to these global trends, our proteomic dataset resolves the trajectories of both the individual participants and the individual proteins. For instance, some proteins such as CRP and SAA4 react much faster than others to weight change (Figs 4 and 5G). Moreover, panel values tended to be uniformly high at the beginning and uniformly low at the end, whereas they were more mixed at intermediate time points. To answer the question of how many study participants profited from weight loss regarding their inflammation profile, we averaged *Z*-scores of the ten-protein panel for each time point and calculated the slope over time. In total, 39 of the 42 individuals had a negative slope, indicating a positive effect on the inflammatory profile of the overwhelming majority of individuals (Table EV6). Investigation of the three individuals that had a positive slope revealed that two gained some weight and the third had high inflammation profile levels at weeks 13 and 26. Moreover, weight regain was present in two further individuals out of five that showed very small positive effects in response to sustained weight loss (Table EV6).

Proteomic inflammation markers and insulin resistance

Nearly 40 plasma proteins correlated significantly with HOMA-IR (homeostasis model assessment—insulin resistance). This included adiponectin (ADIPOQ), the protein with the highest known correlation (Weyer *et al*, 2001), but remarkably nine plasma proteins were even more significant (Fig 5F and Table EV4). Excluding an IgG chain, this allowed us to define a positively correlating panel of four proteins (pro-IR) and a negatively correlating panel of five proteins (including adiponectin) (anti-IR).

To compare levels of insulin resistance-related proteins within the study population, we *Z*-scored each of these proteins along all individuals for each time point. As expected, proteins in the two panels were co-regulated, whereas the panels themselves were anti-regulated (Fig 6A). Nearly all individuals that had high values in the pro-IR panel showed low values in the anti-IR panel. The 24 participants that were in at least one of these groups were considered as individuals with high insulin resistance for the following analyses.

To investigate the known connection between low-grade inflammation and IR (van Greevenbroek *et al*, 2013) at the proteome level, we used the inflammation panel and analyzed it

Published online: December 22, 2016
Molecular Systems Biology

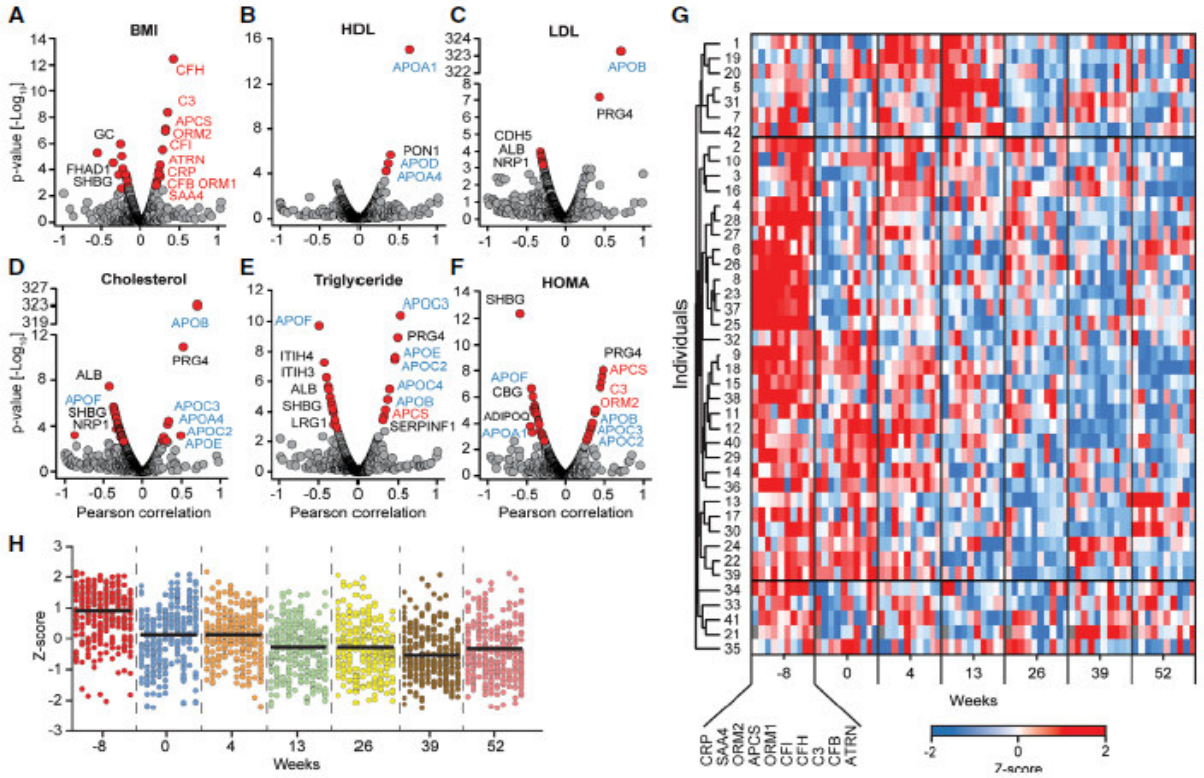


Figure 5. Plasma proteins and the longitudinal inflammation profile.

A Correlation of body mass index (BMI) with all quantified proteins in our dataset.

B–F Correlation analysis of the indicated clinical parameter with plasma protein levels over time.

G For each individual, *Z*-scores were calculated for each of the proteins of the ten-protein panel over the seven time points. The proteins were arranged in the indicated order and hierarchical clustering was performed on the level of the different individuals, resulting in a longitudinal inflammation profile.

H Dot plots for the ten-protein panel in the same order as in (G) for the central cluster. The black line indicates the median *Z*-score of the inflammation panel for each time point.

Data information: Significant proteins are displayed by red dots, and non-significant ones with gray dots. Red letters indicate inflammation factors that correlate with the BMI and that were used to generate panel (G). A Benjamini–Hochberg FDR of 0.05 was used for significance in all correlation analyses (A–D).

together with the IR panel. We *Z*-scored the proteins along all individuals and time points, which clearly separated the cohort into higher and lower inflammatory sub-groups. Remarkably, there was an overlap of 14 individuals that were both in the 16-member group with high inflammation and in the 24-member group with the high IR panel values. Thus, a sub-group with high metabolic burden (those with increased plasma levels of markers previously linked to cardiovascular and metabolic diseases) can be determined entirely from the plasma proteome profiling data.

To compare the effects of weight loss and weight maintenance of these 14 individuals to the other 28 study participants, we *Z*-scored the proteins of the anti-IR/pro-IR and inflammation panel for each protein over the whole study period and all individuals. Both the high- and the comparatively low-risk groups were able to lower their insulin resistance as well as their systemic inflammation levels, as reflected by the panels. The benefit regarding HOMA-IR was even higher for the high metabolic burden group (HOMA-IR: −19% vs. −39%, Fig 6B). Nevertheless, the high metabolic burden group was only able to adjust the IR and inflammation panels to about the start levels of the low-risk group.

Levels of 80 plasma proteins correlated with leptin levels as determined by ELISA. Inflammation factors like CRP, SAA1, SAA4, C3, CFH, and APCS had highly positive correlations (Table EV4). Six leptin-correlating proteins are part of the ten-protein inflammation panel, which confirms the connection between insulin resistance and inflammation. Leptin positively correlated with all four proteins from the pro-HOMA-IR panel and two of the five proteins of the anti-HOMA-IR panel (NRP1 and APOF) were also anti-correlated with leptin levels.

Changes in the apolipoprotein family during weight loss

Levels of apolipoprotein family members are of central importance in determining the risk of cardiovascular and metabolic diseases,

Published online: December 22, 2016
Philipp E Geyer et al Proteome profiling during sustained weight loss Molecular Systems Biology

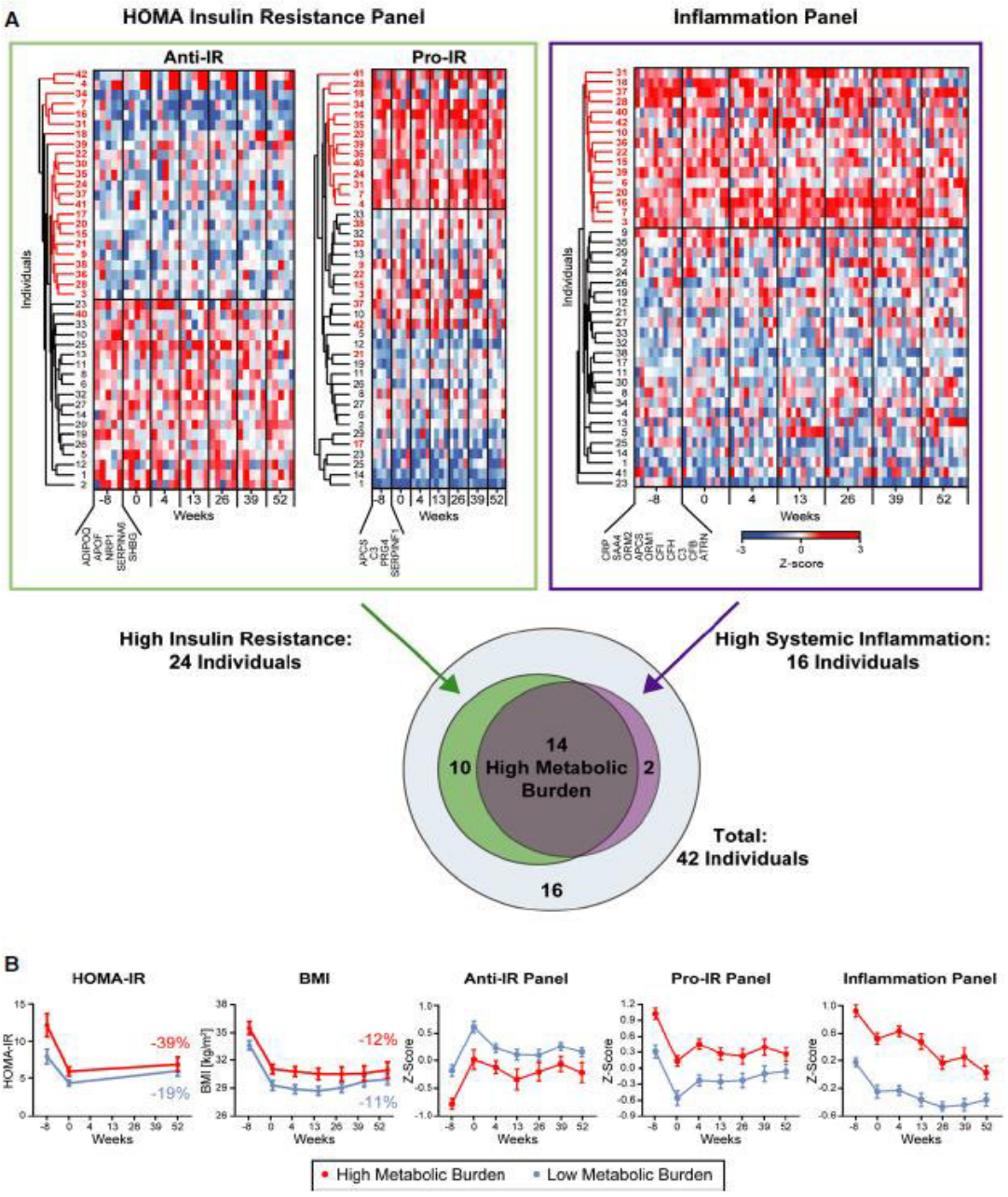


Figure 6. Insulin resistance and systemic inflammation.

A The five proteins with the highest positive and the four proteins with the highest negative correlation with IR were used to define a pro- and an anti-IR panel. These panels separated the study cohort in a high and a low IR group and the 24 individuals that were present in at least one of the IR panels are highlighted (numbered in red at the *y*-axes). Likewise, the ten-protein inflammation panel separates the cohort in individuals with high and low systemic inflammation levels and individuals with high levels were highlighted (numbered in red). Of 16 study participants with high inflammation levels, 14 were also present in the high IR group as illustrated by a Venn diagram, indicating a high metabolic burden group as defined by plasma proteome profiling.

B HOMA-IR levels and the BMI are compared between the high and the low metabolic burden group with indicated changes in percent at the study endpoint. These changes are linked to longitudinal changes in the anti-IR, the pro-IR, and the inflammation profile. The means are plotted with SEM as error bars over time.

Published online: December 22, 2016
Molecular Systems Biology

but current immuno-based assays only measure one or a few of them at a time. In contrast, shotgun proteomics should be able to comprehensively profile the entire family, and indeed, we successfully recorded longitudinal profiles for 18 different apolipoproteins (Fig 7A). Twelve of these changed significantly at least at one point during weight loss or maintenance and six (APOA2, APOB, APOC2, APOL1, SAA1, and SAA4) showed significant long-term effects (Fig EV2A and B). Of the rapid responding apolipoproteins, APOF increased by 37% and APOA4 decreased by 36% upon weight loss, but the levels of both reverted to baseline over the course of a year. In contrast, levels of APOC1, APOC2, and APOC4 consistently decreased and stayed at about 70% of their initial levels. Apolipoprotein(a) (LPA) had the largest absolute change on average as a response to weight loss (increase of 95%).

Next, we correlated the dynamics of the apolipoproteins with BMI, cholesterol, triglyceride, glucose, HDL, and LDL levels. Of these, APOF had a high negative correlation with triglyceride levels (-0.50; $P < 6 \times 10^{-8}$) and APOB, APOC2, APOC3, APOC4 as well as APOE a strong positive correlation (0.33, 0.45, 0.52, 0.38, and 0.45, respectively; Fig 5E). APOB, APOC2, APOC3, APOE, and APOF also correlated with total cholesterol (Fig 5D). APOB further strongly correlated with LDL (Fig EV3) (0.72; $P < 3 \times 10^{-323}$), which is expected as each LDL particle contains one APOB molecule (Dominiczak & Caslake, 2011). Similarly, APOA1 is a constituent of HDL and accordingly, it was highly correlated with HDL levels (0.64; $P < 9 \times 10^{-16}$). As mentioned above, APOA4 and PON1 are in the rapid response cluster (group 1 of Fig 4) and both proteins and APOD correlate with HDL measurements (Fig 5B). Several non-apolipoproteins also showed a good correlation with LDL, for instance the above-mentioned PRG4, which furthermore correlated significantly with triglycerides (0.52, 0.48; $P < 3 \times 10^{-11}$, 1×10^{-9}; Fig 5C).

Interestingly, the ratio of APOB to APOA1, which is used to assess cardiovascular disease risk, decreased due to weight loss by 8% and remained lower over time (week 52: 7%) for 25 of the 42 study participants.

To investigate the general response of lipoprotein particles and metabolic process during weight loss on the basis of the plasma proteome, we used gene ontology (GO). This assigned apolipoproteins to five main lipoprotein particles: chylomicrons, high-density lipoprotein (HDL), intermediate-density lipoprotein (LDL), low-density lipoprotein (LDL), and very low-density lipoprotein particle (vLDL) (Fig 7A and B). Of the 12 apolipoproteins that occur in high-density lipoprotein (HDL) or low-density lipoprotein (LDL) particle, 11 changed significantly. Moreover, we observed a fast response for seven significantly changed apolipoproteins (belonging to clusters 1, 2, and 7 of Fig 4). Globally, the level of the different lipoprotein particles changed most rapidly during weight loss and tended to remain at a lower level during weight maintenance. Performing the same analysis at the level of gene ontology defined "biology processes" likewise showed that most of these that were related to lipoproteins, lipids, cholesterol, and fat decreased with body weight (Figs 7C and EV4). Thus, plasma proteome profiling revealed the dynamics of metabolic changes during weight loss both at the level of individual proteins and at the global levels of lipoprotein particles and processes.

Discussion

Losing weight and maintaining the weight loss are central topics in modern society, research, and medicine. Although generally viewed as desirable, their effects on cardiovascular and general metabolic risk at the individual level are far from universally agreed (Goodpaster *et al*, 2010; Casazza *et al*, 2013; Look *et al*, 2013; Kushner & Ryan, 2014). Here, we wished to contribute to this debate by deciphering the plasma proteome at a global level, using state-of-the-art MS-based proteomics technologies. We used an automated and robust plasma proteome profiling workflow and successfully measured 1,294 plasma proteomes from 52 obese individuals, revealing dynamic changes in response to 8 weeks of diet-induced weight loss followed by a year of weight maintenance. The depth of proteome coverage obtained here—more than 400 proteins per individual—was sufficient for covering all clinically relevant lipoproteins and markers of low-grade inflammation as well as many other functional blood proteins. Quadruplicate measurements as well as the measurement of time profiles of 43 participants allowed us to pinpoint relatively small changes (< 20%) with very high statistical significance, which compares favorably with standard, antibody-based laboratory tests. Further advantages of MS-based proteomics are that large numbers of proteins can be analyzed simultaneously and with very high specificity, as there is no "cross-reactivity" in MS measurements. Furthermore, measurements are unbiased in the sense that the identity of analytes does not have to be known beforehand. We found that the global nature of plasma proteomics also allows us to quickly assess the quality of individual samples and entire studies on the basis of erythrocyte lysis markers and proteins involved in coagulation (Fig EV1).

Omics technologies have already been brought to bear on the study of obesity. GWASs have linked specific loci with genetic propensity for this trait, whereas transcriptome studies have investigated tissues such as fat, muscle, or white blood cells. These studies are by their nature not directly connected to changes in protein levels in the plasma. In contrast, metabolomics has been performed on plasma in the context of weight loss, which demonstrated that several metabolic markers change after weight loss and similar to what we have reported here that individual-specific levels seem to predominate (Piccolo *et al*, 2015; Wahl *et al*, 2015; Newgard, 2016). At the protein level, individual regulators of lipid transport have been studied in depth, but our study provides a first proteomic view of changes in the plasma. Although not subject of this study, it would be interesting in the future to connect the different omics datasets to obtain a more comprehensive understanding of physiological changes during weight loss and maintenance.

Our study design allowed us to separate the influence of weight loss and weight maintenance on the plasma proteome and also provided a systematic view of the variations in the levels of hundreds of plasma proteins in a human cohort. We defined "individual-specific protein levels" as those whose variation over time in each individual was small compared to the difference between the individuals. By these criteria, a surprisingly large part of the plasma proteome was individual- or sub-group-specific as nearly half of the proteins varied more than twofold, while their longitudinal CV was less than 30%. Thus, levels of many proteins remain essentially similar over long time periods, but vary between different individuals. Such observations have already been made

Published online: December 22, 2016
Philipp E Geyer et al Proteome profiling during sustained weight loss Molecular Systems Biology

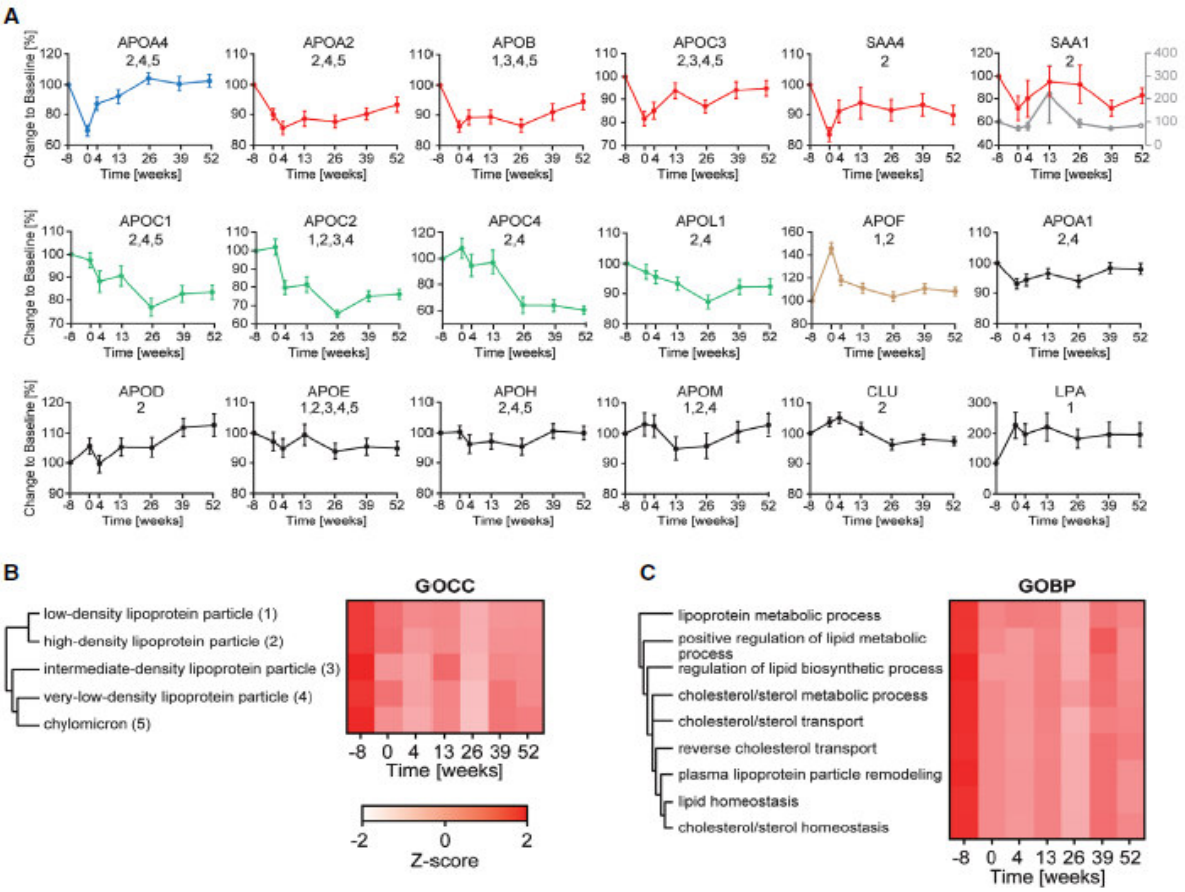


Figure 7. Effect of weight loss on the apolipoprotein family.

A The initial LFQ intensity before weight loss (week −8) was set to 100% to normalize protein abundance within each participant to account for individual-specific protein levels. Colors derive from clusters in Fig 4. The means are plotted with SEM as error bars over time. For the acute phase protein SAA1, the peak at week 13 was caused by very high levels in one individual (gray curve and right *y*-axis in the 6th panel). Excluding this individual results in the red curve. Numbers below the protein name refer to their presence in the lipoprotein particle in panel (C).

B Annotation for gene ontology cellular component (GOCC) was *Z*-scored and filtered for main lipoprotein particles.

C Gene ontology biological process (GOBP) annotations were filtered for the keywords lipid, lipoprotein, fat, and cholesterol, leading to the displayed group of GOBPs, which decreased due to weight loss.

before, but only with selected proteins and generally in smaller longitudinal studies (Crawford & Elisens, 2006; Kamstrup *et al*, 2008; Carlsson *et al*, 2010; Anderson, 2014). Our results suggest that it would be beneficial to determine baseline levels and variations in proteins by MS-based proteomics in even larger populations and to determine the underlying causes. Furthermore, these baseline levels could be determined in each patient in the context of precision medicine. This would enable patient-adjusted diagnostic tests and cutoff levels that take an individual's protein expression values as a reference rather than population-based reference intervals, which are used in clinical diagnostic tests today (Anderson, 2010). In the context of plasma proteome profiling, this could be crucial for a proper interpretation of the plasma proteome in health and disease. Our results suggest that longitudinal plasma proteome profiles can circumvent problems associated with the natural variability of protein levels within and between individuals to a large degree.

Our study establishes that weight loss has a wide effect on the plasma proteome profile with a large proportion of quantified proteins changing significantly (93 proteins). Overall, we observed a strong difference between before and after weight loss followed by an adaption of the protein levels during the yearlong weight maintenance period. Many of the changes in the plasma proteome profile are readily explainable from the underlying biology and physiology.

Published online: December 22, 2016
Molecular Systems Biology

For instance, levels of SERPINF1, which is secreted by adipocytes (Famulla *et al*, 2011), decrease with very high statistical significance, mirroring the loss of fat mass. SERPINF1 has already been associated with obesity before (Wang *et al*, 2008) but our study quantitatively establishes its fast downregulation in response to weight loss. This behavior and its consistency across the study population (in contrast to the known weight loss marker SHBG) could make SERPINF1 of possible interest in a clinical context.

Weight maintenance is a key challenge of any weight loss intervention. We therefore compared the initial plasma proteomes of poor (19 individuals), intermediate (13 individuals), and good (nine individuals) low-weight maintainers using body weight data from 2-year follow-up. This generated some interesting trends in plasma protein expression; however, these were not statistically significant (Table EV7). Nevertheless, future plasma proteomic studies may consider including such perspectives, potentially allowing identification of markers of individuals with a high probability of weight regain.

Monitoring the adaption of protein levels after weight loss yields new insights into the regulation of the plasma proteome. We identified several groups of highly significantly changed and functionally connected proteins with the same longitudinal behavior. Other proteins also clustered closely, but had no known functional connection. The tight cluster of four acute phase proteins, including CRP and SAA1, appears to represent systemic low-grade inflammation status in response to weight loss. The connection between high levels of CRP and obesity is well known (Yudkin *et al*, 1999; Selvin *et al*, 2007). Additionally, both proteins are associated with increased risk for cardiovascular diseases (CVD), where an increase of 10% of CRP levels leads to a 5.5% increase in CVD risk and a twofold increase of SAA1 to a 17% increase (Ridker *et al*, 2002; McEneny *et al*, 2015). In our study, weight loss induced a lowering of the individual's median levels of CRP by 35% and SAA1 by 44%, commonly accepted markers for CVD risk. The APOB/APOA1 ratio, another CVD risk marker, likewise decreased due to weight loss. In this way, plasma proteome profiling links previously established risk markers to weight loss. Apart from the specific aims of this study, plasma proteome profiling now provides the clinician with a new toolbox to investigate potentially important risk markers of CVD or other metabolic-related disease.

For the first time, MS-based plasma proteomics delivered a comprehensive picture of the response of proteins involved in lipid transport, including 18 apolipoproteins and other proteins that play a role in lipid metabolism. We found the expected correlations between LDL, HDL, total cholesterol, and triglycerides with constituent apolipoproteins, and these correlations may be even higher if the classical laboratory values would be reported more precisely. It would be interesting to investigate whether combinations of some of the top correlating proteins could be useful and robust risk markers. NRP1 and PRG4, which we identified by their longitudinal profiles and correlation with clinical parameters, are examples of promising candidates for further investigation. NRP1 was significantly increased at all time points after weight loss and the fact that NRP1 binds VEGF (Pellet-Many *et al*, 2008) makes it interesting to investigate a possible mechanism involving this interaction during weight loss. PRG4 was downregulated in response to weight loss. This proteoglycan lubricates articulating joints, and its presence in plasma may indicate tissue leakage. However, the strikingly strong correlation with LDL, triglyceride, and cholesterol levels may implicate PRG4 in lipid metabolism and in any case make it a potential biomarker related to LDL levels.

The anti-inflammatory proteins IL1RAP and CD14 increased after weight loss. This raises the possibility that the IL1 binding activity of IL1RAP results in antagonism of IL1 action and could play a role in lowering systemic inflammation (Smith *et al*, 2003). Levels of soluble IL1RAP and CD14 are known to be lower in obese individuals compared to controls (Bozaoglu *et al*, 2014; Laugerette *et al*, 2014). Our finding that IL1RAP and CD14 levels increased after weight loss therefore provides a possible mechanism that contributes to reduced low-grade inflammation.

We defined a panel of ten inflammation proteins and analyzed their correlation with BMI, to evaluate which individuals would benefit the most from weight loss based on a longitudinal inflammation profile. In our dataset, 39 of 42 individuals showed a clear positive effect in response to weight loss, and the three "non-profiting" individuals had increased inflammation levels apparently in part because of regain of weight. This indicates that the vast majority of obese individuals would profit from weight loss by improving their inflammatory profiles including known CVD and metabolic risk factors.

For further risk stratification of the cohort, we combined the inflammation panel with an insulin resistance panel, which was also defined by plasma proteome profiling. There was a high but not complete overlap of individuals in the two panels, pinpointing individuals with a high metabolic burden. Clearly, both these "high-risk" individuals and the other study participants greatly benefitted from weight loss and these effects persisted or even continued to improve over the 1-year weight maintenance period. Interestingly, the initial average values of the high-risk group in the inflammation and IR panel decreased to those of the other participants over the observational period.

From a clinical perspective, one could speculate that MS-based plasma proteomic may be used for patient stratification of obese subjects in a low and high metabolic burden profile, thereby providing a new diagnostic tool to intensify and optimize both pharmacological and non-pharmacological treatment of obese subjects with an elevated risk of cardiovascular disease. On the other hand, a global plasma proteomic analysis, as the one reported here, may target potential unknown metabolic regulators, thereby fostering future experiment setups by using a knockout approach of proteins of interest in rodents or in cell lines.

Incorporating an additional step of peptide fractionation would allow quantification of more than 1,000 proteins (Geyer *et al*, 2016), and adding multiplexing would further increase throughput and perhaps measurement precision. We envision such a capability to be available soon, which would enable routine measurements of studies such as this one in even greater depth and still in a reasonable amount of time. In the future, it would be interesting to use a workflow with even deeper coverage and higher throughput on a wide variety of clinical studies related to weight loss and other life style or pharmacological interventions. This would help to define novel risk markers and to disentangle correlations between them and existing clinical parameters. The resulting knowledge extracted from the plasma proteome could predict the individual gains expected from different interventions on the health or disease state.

Published online: December 22, 2016

Materials and Methods

Study design

Details of the weight loss study design are published elsewhere (Iepsen *et al*, 2015). In total, 58 obese study participants were recruited with the following inclusion criteria: healthy individuals with a BMI between 30 and 40 kg/m^2 and an age between 18 and 65 years. Excluded were participants with any acute or chronic illness other than obesity, any medical treatment with known effects on glucose and lipid metabolism, appetite or food intake, pregnancy or breast feeding and fasting plasma glucose levels of ≥7 mmol/l.

Study participants followed a weekly supervised very low-calorie powder diet (800 kcal per day; Cambridge Weight Plan, Corby, UK) for 8 weeks to achieve a weight loss of at least 7.5% of the initial body weight after 8 weeks (Riecke *et al*, 2010).

During the weight maintenance phase, the calorie intake of the study participants was restricted to the estimated daily energy needs subtracted by 600 kcal. In the case of weight gain, up to two meals a day during the weight maintenance period were allowed to be replaced by Cambridge Weight Plan products to ensure weight maintenance. Half of the participants also received 1.2 mg of liraglutide (daily) after weight loss. Both groups equally successfully maintained the weight loss with no significant change in weight from after weight loss to 1 year of weight maintenance (Iepsen *et al*, 2015). The difference in plasma proteomes between liraglutide-treated and non-treated individuals was subtle at our depth and accuracy of proteome measurement and was not further pursued in our analysis.

Blood samples were taken before weight loss (week −8), directly after (week 0) and at five time points during weight loss (weeks 4, 13, 26, 39, 52). Weight was measured at each visit, and further data for each participant were acquired for time points −8, 0, and 52.

The study was approved by the ethical committee in Copenhagen (reference number: H-4-210-134) and was performed in accordance with the Helsinki Declaration II and with ICH-GCP practice. Participation in the investigation was voluntary and the individuals could at any time retract their consent to participate. ClinicITrials.gov identifier: NCT02094183.

Highly abundant protein depletion for building a matching library

We built up a matching library and used a depletion of the top 20 most abundant plasma proteins by a combination of two immunodepletion kits (Nagaraj *et al*, 2012; Geyer *et al*, 2016). Plasma samples of three women and three men were obtained from a highly reliable reference blood bank (Plasma^{Ref} Panels) from the Blutspendedienst des Bayerischen Roten Kreuzes. The Agilent Multiple Affinity Removal Spin Cartridge was used for the depletion of the top six highly abundant proteins (albumin, IgG, IgA, antitrypsin, transferrin, and haptoglobin), followed by ProteoPrep20 Plasma Immunodepletion Kit for the 20 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). Samples were depleted, digested, and measured in triplicate in the same way as the non-depleted sample set of the weight loss study.

Sample preparation: protein digestion and in-StageTip purification

Sample preparation was carried out as described in Geyer *et al* (2016) and Kulak *et al* (2014) with the automated setup on an Agilent Bravo liquid handling platform. Plasma samples were diluted 1:10 with $_{dd}H_2O$ and 10 µl of the sample was mixed with 10 µl twofold concentrated SDC buffer. Reduction and alkylation were carried out at 95°C for 10 min. Trypsin and LysC (1:100 µg of enzyme to micrograms of protein ratio) were added to the mixture after a 5-min cooling step at room temperature. Digestion was performed at 37°C for 1 h. The digest was acidified by adding 40 µl of 1% trifluoroacetic acid (TFA) in isopropanol. An amount of 20 µg of peptides was loaded on two 14-gauge StageTip plugs, followed by the addition of 100 µl 1% trifluoroacetic acid (TFA) in isopropanol and strong mixing. The StageTips were centrifuged using a 3D-printed in-house-made StageTip centrifugal device at 1,500 *g*. After washing the StageTips two times using 100 µl 1% trifluoroacetic acid (TFA) in isopropanol and one time using 100 µl 0.2% TFA in $_{dd}H_2O$, purified peptides were eluted by 60 µl of elution buffer into autosampler vials. The collected material was completely dried using a SpeedVac centrifuge at 60°C (Eppendorf, Concentrator plus). Peptides were suspended in buffer A* (Kulak *et al*, 2014) and afterward sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). It was not possible to obtain any peptides from one of the samples (# 44_1).

Ultra-high-pressure liquid chromatography and mass spectrometry

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1000 ultra-high-pressure system (Thermo Fisher Scientific), which was combined with a Q Exactive HF Orbitrap (Thermo Fisher Scientific) and a nano-electrospray ion source (Thermo Fisher Scientific) (Scheltema *et al*, 2014). Purified peptides were separated on 40-cm HPLC columns [ID: 75 µm; in-house packed into the tip with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH)]. For each LC-MS/MS analysis, around 1 µg peptides was used for 45-min runs and for each fraction of the deep plasma dataset.

Peptides were loaded in buffer A (0.1% (v/v) formic acid) and eluted with a linear 18-min gradient of 5–20% of buffer B (0.1% (v/v) formic acid, 60% (v/v) acetonitrile), followed stepwise by a 12-min increase to 35% of buffer B, a 6 min to 50% of buffer B, 5.5-min increase to 98% of buffer B, followed by a 3.5-min wash of 98% buffer B at a flow rate of 350 nl/min. Column temperature was kept at 60°C by a Peltier element containing in-house-developed oven, 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 (topN method). 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

Published online: December 22, 2016
Molecular Systems Biology

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

Mass spectrometry raw files were analyzed by MaxQuant software version 1.5.3.23 (Cox & Mann, 2008), and peptide lists were searched against the human Uniprot FASTA database (version June 2015). A contaminants database by the Andromeda search engine (Cox *et al*, 2011) with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidations as variable modifications was used. 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, 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, and 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 performed after constructing a matching library consistent of depleted and all the undepleted plasma samples from the weight loss study. 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 1 (Cox *et al*, 2014).

Bioinformatics analysis

All bioinformatics analyses were done with the Perseus software of the MaxQuant computational platform (Cox & Mann, 2008; Tyanova *et al*, 2016). For statistical analysis of significantly changed proteins before (week −8) and directly after weight loss (week 0), a one-sample *t*-test was used with a false discovery rate of < 0.05 after Benjamini–Hochberg correction. We only considered highly significant proteins with a *P*-value of $P < 0.0005$ for the hierarchical clustering in Fig 4. For all correlation analyses, a false discovery rate of < 0.05 after Benjamini–Hochberg correction was applied.

All data needed for correlation analysis of classical clinical parameters like BMI, weight, levels of cholesterol, leptin, HDL, LDL, and triglycerides as well as HOMA-IR to MS-based proteomic acquired LFQ intensities are available for all study participants and time points (Table EV8).

Data and materials 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 PXD004242.

Expanded View for this article is available online.

Acknowledgements
We thank all members of the Proteomics and Signal Transduction Group for help and discussions and in particular Korbinian Mayr, Igor Paron, and Gaby Sowa for technical assistance 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 and by the Novo Nordisk Foundation (Grant NNF15CC0001).

Author contributions
PEG designed, performed, and interpreted the MS-based proteomic analysis of patient plasma and wrote the paper and generated the figures. NJWA designed and interpreted the MS-based proteomic analysis of patient plasma and generated article text. ST provided statistical assistance, interpretation of the proteomic analysis, and revised the manuscript. NG performed MS-based proteomic analysis of patient plasma and revised the manuscript. EWI, JL, SM, JJH, and SST provided patient material and clinical data and revised the manuscript. MM designed and interpreted the MS-based proteomic analysis of patient plasma, supervised and guided the project, and wrote the paper.

Conflict of interest
The authors declare that they have no conflict of interest.

References

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

Anderson NL (2010) The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem* 56: 177 – 185

Anderson NL (2014) Six decades searching for meaning in the proteome. *J Proteomics* 107: 24 – 30

Azrad M, Gower BA, Hunter GR, Nagy TR (2012) Intra-abdominal adipose tissue is independently associated with sex-hormone binding globulin in premenopausal women. *Obesity (Silver Spring)* 20: 1012 – 1015

Bozaoglu K, Attard C, Kulkarni H, Cummings N, Diego VP, Carless MA, Shields KA, Johnson MP, Kowlessur S, Dyer TD, Comuzzie AG, Almasy L, Zimmet P, Moses EK, Goring HH, Curran JE, Blangero J, Jowett JB (2014) Plasma levels of soluble interleukin 1 receptor accessory protein are reduced in obesity. *J Clin Endocrinol Metab* 99: 3435 – 3443

Carlsson L, Lind L, Larsson A (2010) Reference values for 27 clinical chemistry tests in 70-year-old males and females. *Gerontology* 56: 259 – 265

Casazza K, Pate R, Allison DB (2013) Myths, presumptions, and facts about obesity. *N Engl J Med* 368: 2236 – 2237

Ceron JJ, Tecles F, Tvarijonaviciute A (2014) Serum paraoxonase 1 (PON1) measurement: an update. *BMC Vet Res* 10: 74

Cominetti O, Nunez Galindo A, Corthesy J, Oller Moreno S, Irincheeva I, Valsesia A, Astrup A, Saris WH, Hager J, Kussmann M, Dayon L (2016) Proteomic biomarker discovery in 1000 human plasma samples with mass spectrometry. *J Proteome Res* 15: 389 – 399

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

Published online: December 22, 2016
Philipp E Geyer et al Proteome profiling during sustained weight loss Molecular Systems Biology

Crawford PT, Elisens WJ (2006) Genetic variation and reproductive system among North American species of Nuttallanthus (Plantaginaceae). *Am J Bot* 93: 582 – 591

Dominiczak MH, Caslake MJ (2011) Apolipoproteins: metabolic role and clinical biochemistry applications. *Ann Clin Biochem* 48: 498 – 515

Duke-Cohan JS, Gu J, McLaughlin DF, Xu Y, Freeman GJ, Schlossman SF (1998) Attractin (DPPT-L), a member of the CUB family of cell adhesion and guidance proteins, is secreted by activated human T lymphocytes and modulates immune cell interactions. *Proc Natl Acad Sci USA* 95: 11336 – 11341

Eckel RH, Grundy SM, Zimmet PZ (2005) The metabolic syndrome. *Lancet* 365: 1415 – 1428

Esser N, Legrand-Poels S, Piette J, Scheen AJ, Paquot N (2014) Inflammation as a link between obesity, metabolic syndrome and type 2 diabetes. *Diabetes Res Clin Pract* 105: 141 – 150

Famulla S, Lamers D, Hartwig S, Passlack W, Horrighs A, Cramer A, Lehr S, Sell H, Eckel J (2011) Pigment epithelium-derived factor (PEDF) is one of the most abundant proteins secreted by human adipocytes and induces insulin resistance and inflammatory signaling in muscle and fat cells. *Int J Obes (Lond)* 35: 762 – 772

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

Goodpaster BH, Delany JP, Otto AD, Kuller L, Vockley J, South-Paul JE, Thomas SB, Brown J, McTigue K, Hames KC, Lang W, Jakicic JM (2010) Effects of diet and physical activity interventions on weight loss and cardiometabolic risk factors in severely obese adults: a randomized trial. *JAMA* 304: 1795 – 1802

van Greevenbroek MM, Schalkwijk CG, Stehouwer CD (2013) Obesity-associated low-grade inflammation in type 2 diabetes mellitus: causes and consequences. *Neth J Med* 71: 174 – 187

Grundy SM (2015) Adipose tissue and metabolic syndrome: too much, too little or neither. *Eur J Clin Invest* 45: 1209 – 1217

Hansen BA, Bray GA (2008) *The metabolic syndrome: epidemiology, clinical treatment, and underlying mechanisms*. Totowa, NJ: Humana Press

Hirai K, Hussey HJ, Barber MD, Price SA, Tisdale MJ (1998) Biological evaluation of a lipid-mobilizing factor isolated from the urine of cancer patients. *Cancer Res* 58: 2359 – 2365

Iepsen EW, Lundgren J, Dirksen C, Jensen JE, Pedersen O, Hansen T, Madsbad S, Holst JJ, Torekov SS (2015) Treatment with a GLP-1 receptor agonist diminishes the decrease in free plasma leptin during maintenance of weight loss. *Int J Obes (Lond)* 39: 834 – 841

Kamstrup PR, Benn M, Tybjaerg-Hansen A, Nordestgaard BG (2008) Extreme lipoprotein(a) levels and risk of myocardial infarction in the general population: the Copenhagen City Heart Study. *Circulation* 117: 176 – 184

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

Kushner RF, Ryan DH (2014) Assessment and lifestyle management of patients with obesity: clinical recommendations from systematic reviews. *JAMA* 312: 943 – 952

Laudes M, Oberhauser F, Schulte DM, Schilbach K, Freude S, Bilkovski R, Schulz O, Faust M, Krone W (2010) Dipeptidyl-peptidase 4 and attractin expression is increased in circulating blood monocytes of obese human subjects. *Exp Clin Endocrinol Diabetes* 118: 473 – 477

Laugerette F, Alligier M, Bastard JP, Drai J, Chanseaume E, Lambert-Porcheron S, Laville M, Morio B, Vidal H, Michalski MC (2014) Overfeeding increases postprandial endotoxemia in men: inflammatory outcome may depend on LPS transporters LBP and sCD14. *Mol Nutr Food Res* 58: 1513 – 1518

Lewis JG, Bagley CJ, Elder PA, Bachmann AW, Torpy DJ (2005) Plasma free cortisol fraction reflects levels of functioning corticosteroid-binding globulin. *Clin Chim Acta* 359: 189 – 194

Li XA, Yutani C, Shimokado K (1998) Serum amyloid P component associates with high density lipoprotein as well as very low density lipoprotein but not with low density lipoprotein. *Biochem Biophys Res Commun* 244: 249 – 252

Liu Y, Buil A, Collins BC, Gillet LC, Blum LC, Cheng LY, Vitek O, Mouritsen J, Lachance G, Spector TD, Dermitzakis ET, Aebersold R (2015) Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* 11: 786

Look ARG, Wing RR, Bolin P, Brancati FL, Bray GA, Clark JM, Coday M, Crow RS, Curtis JM, Egan CM, Espeland MA, Evans M, Foreyt JP, Ghazarian S, Gregg EW, Harrison B, Hazuda HP, Hill JO, Horton ES, Hubbard VS *et al* (2013) Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. *N Engl J Med* 369: 145 – 154

Malmstrom E, Kilsgard O, Hauri S, Smeds E, Herwald H, Malmstrom L, Malmstrom J (2016) Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nat Commun* 7: 10261

Mann M, Kulak NA, Nagaraj N, Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* 49: 583 – 590

McEneny J, Daniels JA, McGowan A, Gunness A, Moore K, Stevenson M, Young IS, Gibney J (2015) A cross-sectional study demonstrating increased serum amyloid a related inflammation in high-density lipoproteins from subjects with type 1 diabetes mellitus and how this association was augmented by poor Glycaemic control. *J Diabetes Res* 2015: 351601

Munoz J, Heck AJ (2014) From the human genome to the human proteome. *Angew Chem* 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

Newgard CB (2016) Metabolomics and metabolic diseases: where do we stand? *Cell Metab* doi: 10.1016/j.cmet.2016.09.018

Pellet-Many C, Frankel P, Jia H, Zachary I (2008) Neuropilins: structure, function and role in disease. *Biochem J* 411: 211 – 226

Piccolo BD, Keim NL, Fiehn O, Adams SH, Van Loan MD, Newman JW (2015) Habitual physical activity and plasma metabolomic patterns distinguish individuals with low vs. high weight loss during controlled energy restriction. *J Nutr* 145: 681 – 690

Ridker PM, Rifai N, Rose L, Buring JE, Cook NR (2002) Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *N Engl J Med* 347: 1557 – 1565

Riecke BF, Christensen R, Christensen P, Leeds AR, Boesen M, Lohmander LS, Astrup A, Bliddal H (2010) Comparing two low-energy diets for the treatment of knee osteoarthritis symptoms in obese patients: a pragmatic randomized clinical trial. *Osteoarthritis Cartilage* 18: 746 – 754

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

Scheltema RA, Hauschild JP, Lange O, Hornburg D, Denisov E, Damoc E, Kuehn A, Makarov A, Mann M (2014) The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol Cell Proteomics* 13: 3698 – 3708

Selvin E, Paynter NP, Erlinger TP (2007) The effect of weight loss on C-reactive protein: a systematic review. *Arch Intern Med* 167: 31 – 39

Published online: December 22, 2016
Molecular Systems Biology

Seres I, Bajnok L, Harangi M, Sztanek F, Koncsos P, Paragh G (2010) Alteration of PON1 activity in adult and childhood obesity and its relation to adipokine levels. *Adv Exp Med Biol* 660: 129 – 142

Smith DE, Hanna R, Della F, Moore H, Chen H, Farese AM, MacVittie TJ, Virca GD, Sims JE (2003) The soluble form of IL-1 receptor accessory protein enhances the ability of soluble type II IL-1 receptor to inhibit IL-1 action. *Immunity* 18: 87 – 96

Soker S, Takashima S, Miao HQ, Neufeld G, Klagsbrun M (1998) Neuropilin-1 is expressed by endothelial and tumor cells as an isoform-specific receptor for vascular endothelial growth factor. *Cell* 92: 735 – 745

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

Thompson PA, Tobias PS, Viriyakosol S, Kirkland TN, Kitchens RL (2003) Lipopolysaccharide (LPS)-binding protein inhibits responses to cell-bound LPS. *J Biol Chem* 278: 28367 – 28371

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

Utermann G (1989) The mysteries of lipoprotein(a). *Science* 246: 904 – 910

Wahl S, Vogt S, Stuckler F, Krumsiek J, Bartel J, Kacprowski T, Schramm K, Carstensen M, Rathmann W, Roden M, Jourdan C, Kangas AJ, Soininen P, Ala-Korpela M, Nothlings U, Boeing H, Theis FJ, Meisinger C, Waldenberger M, Suhre K *et al* (2015) Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Med* 13: 48

Wang P, Mariman E, Keijer J, Bouwman F, Noben JP, Robben J, Renes J (2004) Profiling of the secreted proteins during 3T3-L1 adipocyte differentiation leads to the identification of novel adipokines. *Cell Mol Life Sci* 61: 2405 – 2417

Wang P, Smit E, Brouwers MC, Goossens GH, van der Kallen CJ, van Greevenbroek MM, Mariman EC (2008) Plasma pigment epithelium-derived factor is positively associated with obesity in Caucasian subjects, in particular with the visceral fat depot. *Eur J Endocrinol* 159: 713 – 718

Weyer C, Funahashi T, Tanaka S, Hotta K, Matsuzawa Y, Pratley RE, Tataranni PA (2001) Hypoadiponectinemia in obesity and type 2 diabetes: close association with insulin resistance and hyperinsulinemia. *J Clin Endocrinol Metab* 86: 1930 – 1935

Yudkin JS, Stehouwer CD, Emeis JJ, Coppack SW (1999) C-reactive protein in healthy subjects: associations with obesity, insulin resistance, and endothelial dysfunction: a potential role for cytokines originating from adipose tissue? *Arterioscler Thromb Vasc Biol* 19: 972 – 978

Zubarev RA, Makarov A (2013) Orbitrap mass spectrometry. *Anal Chem* 85: 5288 – 5296

3.3. Article 3: Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics

Authors: Nils A. Kulak[1,†], Philipp E. Geyer[1,2,†], and Matthias Mann[1,2]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[†]These authors contributed equally to this work.

Pre-fractionation of peptides is a key enabler for the very deep and large-scale proteome characterization. In this manuscript we build upon the popular combination of high pH pre-separation in a first dimension and online low pH separation in the second dimension. We developed and characterized an automated rotor-valve based nano-flow fractionation and concatenation device, which we called 'Spider Fractionator'.

In existing approaches, samples are separately collected after high flow fractionation and afterwards combined. In contrast, the rotor valve of the Spider Fractionator automatically splits the flow of separated peptides after the column into time dependent packages, directed to a number of tubes corresponding to the number of desired fraction to be analyzed. The system allows the researcher many degrees of freedom for the experiments, e.g. the amount of fractionated material can range from just 1 µg up to more than 100 µg, collection of 2-96 fractions is possible and any desired time interval of eluting peptides can be collected.

Instead of the typical setup for high pH fractionation, we used columns with smaller inner diameter, much lower flow-rates and no intermediate collection points. This makes our system much less prone to sample loss, which we proved by a comparison to two commercially available high pH fractionation systems. This analysis showed that our device has little if any detectable sample loss, whereas the commercial systems lost substantial amounts of sample during fractionation. We demonstrated that the Spider Fractionator enables extraordinary sensitivity: As little as one µg of peptides allowed the identification of more than 10,000 protein in HeLa cells after fractionation. We further used different fractionation strategies to obtain a deep proteome in as little time as possible. We applied our optimized conditions to quantify the proteomes of twelve human cell lines to a median depth of more than 11,000 different proteins while fractionating only 20 µg of starting material – by far the deepest proteome measurements yet achieved with such low sample amounts. This experiment also revealed molecular differences that recapitulated the developmental origin of the cell lines.

The ability to efficiently fractionate low sample amounts is also beneficial because a decrease in the starting amounts will also result in a reduction in reagent costs, which is especially important for samples that have been derivatized with expensive mass tags like iTRAQ or TMT. In our laboratory the device is now routinely used for any project involving pre-fractionation and it has already proven to be robust in dozens of projects. Together, the many advantages of using small sample amounts should make the Spider Fractionator attractive to the proteomics community as indicated by the fact that the manuscript has been downloaded more than 600 times within the first month.

For Plasma Proteomic Profiling, the Spider Fractionator is one of the key elements as it enables very deep proteomes from depleted plasma samples for the library matching strategy and it allows us to obtain very deep quantitative proteomes by fractionation of undepleted samples.

Technological Innovation and Resources

⌘ Author's Choice

Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics*[S]

Nils A. Kulak‡§¶, Philipp E. Geyer‡¶‖, and Matthias Mann‡‖**

Recent advances in mass spectrometry (MS)-based proteomics now allow very deep coverage of cellular proteomes. To achieve near-comprehensive identification and quantification, the combination of a first HPLC-based peptide fractionation orthogonal to the on-line LC-MS/MS step has proven to be particularly powerful. This first dimension is typically performed with milliliter/min flow and relatively large column inner diameters, which allow efficient pre-fractionation but typically require peptide amounts in the milligram range. Here, we describe a novel approach termed "spider fractionator" in which the post-column flow of a nanobore chromatography system enters an eight-port flow-selector rotor valve. The valve switches the flow into different flow channels at constant time intervals, such as every 90 s. Each flow channel collects the fractions into autosampler vials of the LC-MS/MS system. Employing a freely configurable collection mechanism, samples are concatenated in a loss-less manner into 2–96 fractions, with efficient peak separation. The combination of eight fractions with 100 min gradients yields very deep coverage at reasonable measurement time, and other parameters can be chosen for even more rapid or for extremely deep measurements. We demonstrate excellent sensitivity by decreasing sample amounts from 100 μg into the sub-microgram range, without losses attributable to the spider fractionator and while quantifying close to 10,000 proteins. Finally, we apply the system to the rapid automated and in-depth characterization of 12 different human cell lines to a median depth of 11,472 different proteins, which revealed differences recapitulating their developmental origin and differentiation status. The fractionation technology described here is flexible, easy to use, and facilitates comprehensive proteome characterization with minimal sample requirements. ***Molecular & Cellular Proteomics 16: 10.1074/mcp.O116.065136, 694–705, 2017.***

From the ‡Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany; §PreOmics GmbH, Martinsried, Germany; and ‖Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Received October 30, 2016, and in revised form, January 25, 2017
Published, MCP Papers in Press, January 26, 2017, DOI 10.1074/mcp.O116.065136

Author contributions: N.A.K., P.E.G., and M.M. designed research; N.A.K. and P.E.G. performed research; N.A.K. and P.E.G. contributed new reagents or analytic tools; N.A.K., P.E.G., and M.M. analyzed data; N.A.K., P.E.G., and M.M. wrote the paper.

Mass spectrometry (MS)-based bottom-up proteomic workflows consist of multiple steps, namely sample preparation, on-line liquid chromatography (LC) coupled with MS measurements, followed by computational data analysis, and interpretation. LC-MS/MS technologies have improved drastically from the initial identification of one or a few proteins using manual, complex, and time-consuming protocols to essentially complete proteomic coverage of microorganisms in a rapid and streamlined manner (1–4). These advances are based on multiple breakthroughs in the analytical and computational sides of the proteomic workflow over the last decade and now make MS-based proteomics a powerful player in systems biology (5). However, for complex proteomes, such as human cell lines, organs, and body fluids, very deep characterization still involves great effort, sample amounts, and costs. Therefore, there is a continuing need for more powerful workflows, and here we contribute to these efforts in the important area of peptide pre-separation before the LC-MS/MS analysis.

To yield in-depth proteomes of complex biological samples, two-dimensional separation approaches at the peptide level are attractive because they are more universally applicable than protein level fractionation. First dimension separation techniques range from isoelectric focusing (6–9) and pipette-based approaches such as StageTip fractionation (10–12) to off-line HPLC systems (13–17). High pH reversed-phase fractionation, alternatively termed basic reversed-phase, as a first off-line chromatography separation in combination with the low pH reversed-phase fractionation in the second on-line dimension was first demonstrated more than 10 years ago. In comparison with other methodologies, it benefits from the uniform first dimensional peptide elution profiles achievable with high pH reversed-phase separation and the high peptide separation efficiency in both dimensions (18, 19). Because the two separation dimensions are not completely orthogonal (meaning that peptide retention times are still correlated), direct application of high pH fractionation would lead to non-uniform filling of the gradient in the second dimension. The key advance that solved this problem was the combination of fractions that elute at substantially different times in the

first dimension (13). This "concatenation" tends to uniformly fill the gradients, leading to deeper proteomes independent of the nature of the sample, while maintaining throughput (20–22). This two-dimensional separation technique, combining high pH fractionation with concatenation, compares favorably with other approaches and is increasingly being applied by the proteomics community (20, 23–27).

Despite the success of high pH reversed-phase fractionation in the deep characterization of complex proteomic samples, a current limitation is the requirement for rather large amounts of starting material. This is due to the large column diameters, flow rates, and number of fractions that are collected before concatenation to preserve peak separation from the first dimension and to maintain collection volumes that can easily be handled. Therefore, instead of the nanoflow systems typical of on-line separation, much larger columns and flow rates are almost always employed. This in turn requires large sample sizes, and milligram amounts of starting material are typical for high pH reversed-phase fractionation. Unfortunately, this implies high reagent costs, for instance for proteolytic enzymes or for the chemical labeling reagents used in multiplexing. Furthermore, it restricts deep proteomes preliminary to cases where comparatively large protein amounts are available and excludes the investigation of rare cellular subpopulation or laser micro-dissected cells in tumor tissues, for instance. High pH fractionation with higher flow rates and larger sample amounts is also used to investigate post-translational modifications in great depth by the combination of isobaric mass tag labeling of peptides after digestion, followed by high pH fractionation and consecutive enrichments (28). In such cases, large sample amounts and high volumes are necessary because of the subsequent enrichment. However, post-translational modification analysis could benefit from a drastic scale-down in high pH fractionation, if already enriched peptides are fractionated. This would necessitate high sensitivity of the fractionation step and be economically attractive in terms of labeling reagents.

Here, we describe a novel approach that allows efficient sample concatenation without using large volumes. Instead, nanoflow systems are employed, and the intermediate sample collection step is eliminated. We demonstrate the operating principle of our spider fractionator, show that fractionation efficiency remains very high, and establish that the flexibility of the system allows choosing an optimum balance between measurement time and desired depth of proteome coverage. Very low sample amounts can be separated without apparent fractionation-induced sample losses. We demonstrate the sensitivity of the system by the analysis of 12 human cell lines to a depth of about 10,000 proteins with only 1 μg of sample.

MATERIALS AND METHODS

Cell Culture—HeLa cells were cultured in high glucose DMEM with 10% fetal bovine serum and 1% penicillin/streptomycin (all from Life Technologies, Inc.). Cell lines were cultured in Dulbecco's modified Eagle's medium (Invitrogen) containing 10% dialyzed fetal bovine serum and penicillin/streptomycin. Cells were counted using a countess cell counter (Invitrogen), and aliquots of 10^6 cells were snap-frozen and stored at −80 °C.

Tryptophan Fluorescence Emission Assay for Protein Quantification—Protein concentrations were determined after solubilizing the samples 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). From this, we estimated that 0.1 μg/μl tryptophan are equivalent to the emission of 7 μg/μl of human protein extract, assuming that tryptophan on average accounts for 1.3% of human protein amino acid composition.

Sample Preparation, Protein Digestion, and in-StageTip Purification—Sample preparation was performed as described previously (3) with the following adaptations. 300 μg of cells were suspended in 50 μl of SDC reduction and alkylation buffer (3). We used 2-chloro-N,N-diethylacetamide as alkylating reagent for the comparison of the 13 cell lines and 2-chloro-acetamide for all other experiments. The cells were kept at 95 °C for 5 min to denature proteins and afterward sonicated to shear DNA and enhance cell disruption with a water bath sonicator (Bioruptor, model UCD-200, Diagenode) for 15 min at the maximum level. The proteolytic enzymes LysC and trypsin were added in a 1:100 ratio (micrograms of enzyme to micrograms of protein), and the solution was incubated for 4 h at 37 °C.

Peptides were acidified by adding 100 μl of ethyl acetate, 1% TFA and extensive mixing for 2 min, and 20 μg were transferred into StageTips containing two 14-gauge SDB-RPS (poly(styrene-divinylbenzene) reversed phase sulfonate) plugs. Afterward, the StageTips were washed with 100 μl of ethyl acetate, 1% TFA to strip SDC and lipids from the digested cells. This was followed by a wash step with 100 μl of ddH_2O[1], 0.2% TFA. The purified peptides were eluted with 60 μl of 80% acetonitrile, 19% ddH_2O, 1% ammonia in autosampler vials. For all steps, the StageTips were centrifuged at 2,000 × *g* until the solutions were rinsed through completely. The collected material was dried using a SpeedVac centrifuge at 60 °C (Eppendorf, Concentrator Plus). Peptides were suspended in 2% acetonitrile, 0.1% TFA in ddH_2O and sonicated for 15 min in a water bath sonicator (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). Moreover, 6,600 HeLa cells, the equivalent to 1 μg of starting material (29), were separately digested using the in-StageTip protocol (3) with the above mentioned adaptations.

Pre-fractionation—We constructed a software-controlled, fully automated, rotor-valve-based fraction collector system coupled on line to a nanoflow HPLC (EASY-nLC 1000 system, Thermo Fisher Scientific), and we used this for all high pH reversed-phase pre-fractionations. The fraction collector system was named Spider Fractionator and is under commercial development by PreOmics GmbH, Martinsried, Germany. We provide a list of components used in constructing the fractionator (supplemental Table 1). For the work reported here, we standardized on a first dimension column of 250 μm inner diameter and a length of 30 cm, which was packed with 1.9 μm C18 particles (ReproSil-Pur C18-AQ 1.9 μm resin by Dr. Maisch GmbH) and has an estimated loading capacity of at least 100 μg. The column is available from PreOmics GmbH (Article No. P.O. 00007). All columns (first dimension and on-line dimension) were passivated by a single run of BSA to saturate irreversible binding sites. For separation into eight pooled fractions, we loaded 20 μg (or other amounts where indicated) of purified and digested peptides onto a reversed-phase C_{18} column. A gradient was generated by using a dual buffer system (buffers A and B) also from PreOmics GmbH (Article No. P.O. 00009).

[1] The abbreviations used are: ddH_2O, double distilled H_2O; SMC, smooth muscle cell; EC, embryonic carcinoma.

TABLE I
Concatenation scheme

No. pooled fractions	4	8	16	24
Peptide amount (μg)	20	20	40	60
No. of non-pooled fractions	54	54	54	54
Pooling scheme	1;5;9;13;17;21;25;29;33;37;41;45;49;53 2;6;10;14;18;22;26;30;34;38;42;46;50;54; etc.	1;9;17;25;33;41;49 2;10;18;26;34;42;50; etc.	1;17;33;49 2;18;34;50; etc.	1;25;49 2;26;50; etc.

Peptides were eluted from 3% B to 30% B in 45 min followed by a linear increase to 60% B in 17 min. This gradient was followed by a further linear increase to 95% B in 5 min and a 3 min wash at 95% B, followed by a 10 min decrease to 3% B. The last segments ensure that the output lines (volume about 800 nl) are emptied, and none of the remaining peptides are lost. The flow rate was kept at a constant 2 μl/min. The 96-well plate was moved by a stepper motor-driven linear actuator. Software was implemented on a Raspberry microcontroller.

We separated peptides into 4, 8, 16, and 24 fractions using rotor valve shifts of 90 s. Fractions were collected into 0.2-ml thin-walled 8-tube strips (Thermo Fisher Scientific). We loaded 20 μg of starting material for 4 and 8 fractions, 40 μg for 16 fractions, and 60 μg for 24 fractions.

The concatenation scheme of table I was used for pooling. For a more detailed version of the fractionation schemes see supplemental Fig. 1.

The pooled fractions were dried using a SpeedVac centrifuge at 60 °C (Eppendorf, Concentrator Plus). Peptides were suspended in 2% acetonitrile, 0.1% TFA in ddH_2O and sonicated for 15 min in a water bath sonicator (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). A total of 2 μg of each concatenated fraction was loaded and measured by LC-MS/MS as described below.

Comparison of the Spider Fractionator to Other High pH Fractionation Systems—We used the same buffers, gradients, and pooling scheme as for the spider fractionator system in comparison with a higher flow system and to a recently introduced spin column system. For all three systems, the same HeLa digest was used to fractionate 1 or 20 μg of peptides. The higher flow system consisted of an XBridge peptide BEH C18 column (2.5 μm particle size, 2.1 × 250 mm, Waters) with a Shimadzu HPLC system at a 60 °C run at a flow rate of 150 μl/min. The fractions were manually pooled. For the 1 μg sample, all fractions were re-pooled into a single vial to determine sample loss. For the 20 μg sample, we manually concatenated samples according to the same scheme as automatically done by the spider fractionator. On the spin system (high pH reversed-phase peptide fractionation kit, Pierce catalog number 84868), separation was done according to the manufacturer's instructions resulting in eight fractions but no concatenation.

Ultra-high Pressure Liquid Chromatography and Mass Spectrometry—Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific) coupled via a nanoelectrospray ion source (Thermo Fisher Scientific) to a hybrid quadrupole Orbitrap mass spectrometer (Q Exactive HF Orbitrap from Thermo Fisher Scientific) (30, 31). Purified peptides were separated on 40 cm HPLC columns (75 μm inner diameter; in-house packed into the tip) at 60 °C with ReproSil-Pur C18-AQ 1.9 μm resin by Dr. Maisch GmbH).

For all measurements, peptides were loaded in buffer A (0.1% formic acid, 5% DMSO (32)) and eluted with a linear 55 min gradient of 2–20% of buffer B (0.1% formic acid, 5% DMSO, 80% acetonitrile), followed by an increase to 40% buffer B within 40 min and afterward within 2 min to 98% buffer B and a 2 min wash at 98% buffer B. The flow rate was kept at 350 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 (33).

MS data was acquired with the Thermo Xcalibur software version 3.0.63, a topN method where *N* could be up to 100. This method in principle allows a very large number of precursor peaks to be picked for fragmentation but is in practice limited by the number of precursors with sufficient ion intensity. In the entire data set, *N* was 15 on average. 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 15 ms. Transient times corresponding to a resolution of 60,000 at *m/z* 200 were chosen. A 1.5 *m/z* isolation window and a fixed first mass of 100 *m/z* were used for MS/MS scans. Fragmentation of precursor ions was performed by higher energy C-trap dissociation (34) with a normalized collision energy of 27 eV. MS/MS scans were performed at a resolution of 15,000 at *m/z* 200 with an ion target value of 5×10^4 and a maximum injection time of 25 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides.

Data Analysis—MS raw data files were analyzed by MaxQuant software version 1.5.3.31 (35), and peptide lists were searched by the Andromeda search engine (36) against the human Uniprot FASTA database to which common contaminant proteins had been added (86,746 entries) with cysteine diethylcarbamidomethylation as a fixed modification for the comparison of the 13 cell lines and cysteine carbamidomethylation as a fixed modification for all other experiments. N-terminal acetylation and methionine oxidations were used as variable modifications in all experiments. The false discovery rate was set to 0.01 for both proteins and peptides with a minimum length of 7 amino acids and was determined by searching a reverse database. Enzyme specificity was set to trypsin and a maximum of two missed cleavages were allowed in the database search. Peptide identification were performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm. The MaxQuant feature "match between runs" was enabled within the dataset of the pooled eight fractions and the single shot samples for all cell line samples. Proteins matching the reversed database were filtered out. Label-free protein quantification was done with a minimum ratio count of 1 (37). All bioinformatics analyses were performed within the Perseus software of the MaxQuant computational platform (35, 37).

RESULTS

Spider Fractionator—The principle of the fractionator is depicted in Fig. 1. The post-column flow from the first dimension separation enters the input port of an eight-port flow-selector rotor valve. At pre-determined time intervals, the valve switches to a new output port. Each of the outputs is connected via a narrow bore capillary to different output lines in a sample collection device, distributing the sample flow into consecutive tubes for the pooled fractions. Once one cycle has been completed, the valves switches back to the first output port and the next fluid volume is added to the already

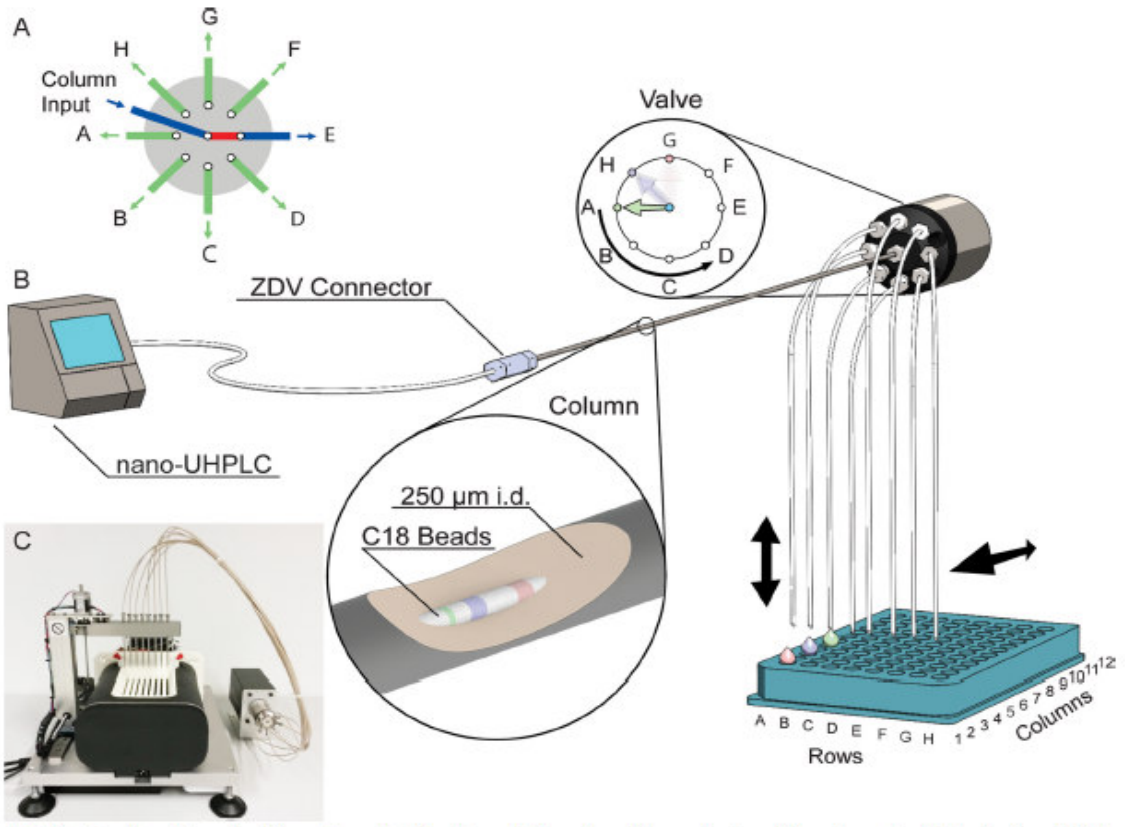


FIG. 1. **Spider fractionation principle and practical implementation.** *A*, switch mechanism of the rotor valve, illustrating how the flow from the first dimension separation is divided to eight output lines. *B*, schematic of the implementation of the spider fractionator. The first dimension separation is realized as a 250 μm inner diameter column, connected upstream through a zero dead volume connector to a nano-HPLC pump (an ultra high pressure unit is depicted but not required). The zoom-in is a cut-away symbolizing different peptide bands being separated in the column by different colors. Downstream, the column is connected to the rotor valve from *A*. The output lines feed into tubes that are filled in turn, according to the concatenation scheme. The spider-like appearance of the output lines give the name to the device. The *arrows* indicate that the output lines can be moved to a new set of tubes for a new separation process. After separation, the tubes are inserted into the autosampler of an UHPLC for LC-MS/MS analysis of the fractions. *C*, photo of the prototype spider fractionator used in this work.

collected first fraction. In this way, the device automatically concatenates and pools the samples, without requiring different collection tubes or the combination of separately collected effluent volumes. Therefore, the volumes are not constrained to a minimum size, which would otherwise be necessary to handle them in separate tubes. We routinely employ a 250 μm inner diameter column in the first dimension at 2 μl/min and switch the valve every 90 s, thus concatenation volumes are only 3 μl. The system is fully programmable, allowing collection not only into multiples of the eight output channels (A–H) but also into as few as two or as many fractions as there are collection tubes in the device (96 in our setup). Furthermore, an arbitrary number of samples can be fractionated, and the rotor valve shifts can be defined by the user. For example, 12 samples could be scheduled for concatenation into eight fractions each in a total of 24 h using 80 min gradients.

When operating in a high pH reversed-phase mode, we use the first column with a buffer at pH 10, which is devoid of non-volatile constituents. The column inner diameter and consequently the flow rate are chosen such that the desired peptide amounts are present in the sample tubes after concatenation. For instance, using second dimension columns with a loading capacity of 2 μg, which is typical of the 75 μm inner diameter columns used in many proteomics laboratories, would call for a sample amount of at least 16 μg to be loaded on the first dimension column. Therefore, the first column should have a capacity of at least the second column multiplied by the number of desired fractions. In principle, the entire system can be scaled up or down as required. Within the constraints mentioned above, different size columns and separation principles can be combined as long as they are at least partially orthogonal. For the work reported here, we standardized on a first dimension column of 250 μm inner

diameter and a length of 30 cm, which is packed with 1.9 μm C_{18} particles and has an estimated loading capacity of at least 100 μg (see "Materials and Methods").

For the subsequent on-line LC separation, no alterations compared with standard laboratory procedures are necessary. In the work reported here, the columns were 40 cm long, 75 μm inner diameter, and packed with 1.9 μm C_{18} particles. The 0.1% formic acid in our buffers ensured low pH compared with the first dimension.

The overall fractionation systems was realized by coupling an EasyLC nanobore HPLC to the first dimension column (see "Materials and Methods"). Note that back-pressure was only 250 bar. Because no high pressure capability is required, a wide range of nanoflow pumps used in proteomics would therefore be suitable. The fractionator principle itself is embodied in an apparatus containing a column oven to maintain 60 °C, the flow-selector valve for fractionation, the required column, two-dimensional axes for automated multi-collection plate position selection, a cooling unit to retain fractions at about 6 °C, a microprocessor control unit for automated contact closure and HPLC interaction, and a driver software to control, log, and monitor all the parameters (Fig. 1*C*). The control unit maintains communications to the upfront HPLC system, to the rotor valve, and to the downstream fraction collection system. The collection system is designed to be fully flexible. Peptides eluting from the column are separated into packages defined by a time interval by rotor valve shifts. The shift in valve position directs each package into one of the eight output lines. Each of these are placed into one of the eight "rows" (A–H) of a 96-well layout. Output line A elutes into row A, line B into row B, and so forth. Eight shifts of the rotor valve will deposit peptides from the column into each tube of the first column, and the next shift will enter the next output line and therefore again fill the first row A. In this way, for eight or less fractions, the output line holder stays at column 1 of the 12 possible positions of a 96-well plate. In case separation into more than eight fractions is desired, the output line holder will move from column 1 to column 2 after eight packages (H1 is followed by A2, supplemental Fig. 1). For 16 fractions, the output holder will move back to column 1 after eight rotations (H2 is followed by A1). Likewise, 24, 32, 40, 48, 56, and 96 fractions can be realized. Fractionation into less than eight fractions or any other desired number between 1 and 96 is also possible as the rotation valve can be programmed to direct packages to arbitrary output lines. For instance, in the case of four fractions, the rotation valve switches directly from D1 to A1. The collection tubes are maintained cooled and can be placed in a SpeedVac and subsequently into the auto sampler of the on-line LC-MS/MS system.

Separation Performance—With the column connected to the Spider fractionator, we first collected each of 54 fractions (90 s duration) in their own tubes. Starting from fraction three, we chose every 6th fraction and analyzed these fractions separately in 100 min gradients on the 40 cm analytical column. The 90 s elution windows from the first dimension eluted roughly in the same region as expected if they had been separated on a low pH analytical column except that their elution range was expanded considerably due to the different pH values (Fig. 2*A*). However, generally the bulk of the peptides was still concentrated within only about 20–50% of the total gradient.

Next, we specified an eight fraction concatenation scheme, meaning that the rotor valve combined the 54 fractions into eight equally filled gradients. Fractions 3, 11, 19, 27, 35, 43, and 51, which were measured separately above, were automatically combined by the rotor valve. Analyzing this concatenated fraction on the 100 min gradient of the analytical column resulted in a peptide elution profile that was filled over the entire range, resembling the super position of the separately measured fractions (Fig. 2*B*). Repeating this experiment in triplicate yielded essentially identical elution profiles, demonstrating reproducibility of the spider fractionator (supplemental Fig. 2).

A desirable feature of a pre-fractionation apparatus is that it concentrates each individual peptide into as few fractions as possible. Note that in a two-dimensional separation scheme there will always be peptides that will be collected into different tubes because the fractionation will occur during peak elution for a certain percentage of them ("peak cutting"). Furthermore, peptides may or may not be sequenced and identified in different fractions, depending on the complexity of the sample and the sequencing speed and sensitivity of the mass spectrometer. Therefore, it is necessary to quantify peptide abundance across all fractions. To evaluate the separation efficiency of the fractionator and pooling scheme, we therefore employed the "match between runs" option in the MaxQuant software (11, 38), which allowed the transfer of identifications between the fractions. Note that the label-free algorithms in MaxQuant in any case normalize the contributions of the fractions and add the contributions for peptides found in more than one (37).

For the majority of peptides (68%), their total intensities were concentrated in one fraction to more than 75% (Fig. 2*C* and supplemental Fig. 3). This is roughly in line with a model in which the peptide distribution in the analytical dimension is largely a function of the cutting of peaks in the first dimension. (In our case, a peak width of 15 s in the first dimension and a 90 s collection window would result in about 15/90 = 16.6% of "cut peaks").

Evaluating the Optimal Number of Fractions—For any sample, the spider fractionator allows choosing the desired number of fractions. A large number of fractions will increase proteome coverage in two ways. First, at any chosen gradient length, the time available for sequencing peptides will increase with the number of fractions. In complex samples, this will lead to more identified peptides and proteins. Second, as there is a maximal loading capacity of the analytic column, a larger number of total fractions increases the total material

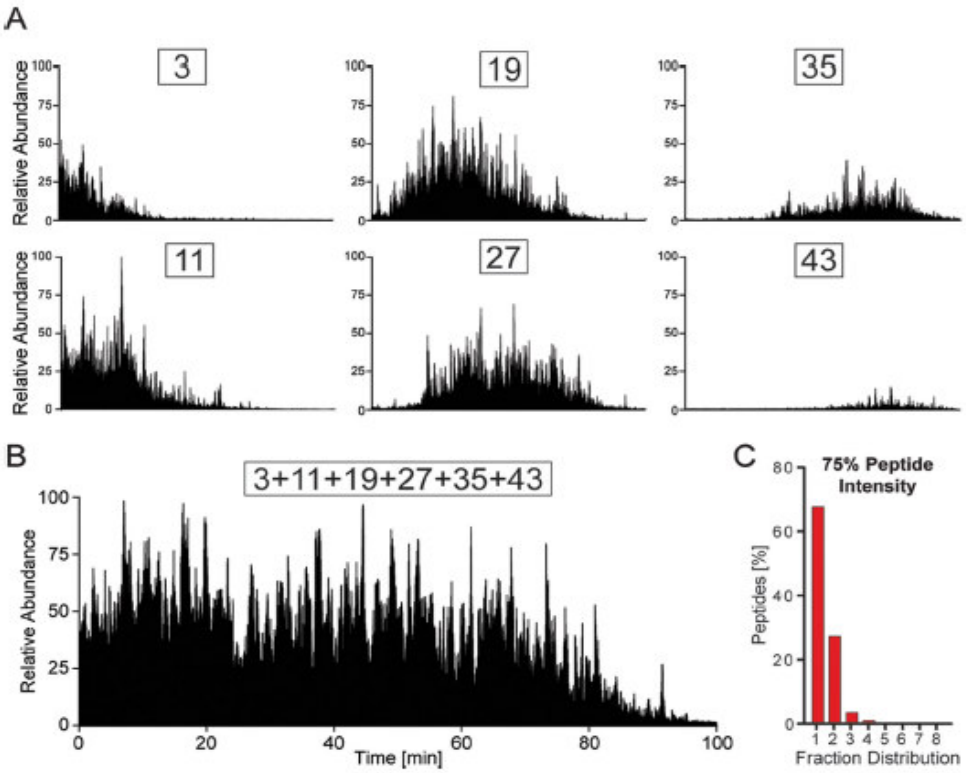


FIG. 2. **Comparison of pooled and non-pooled peptide mixtures and separation efficiency.** *A*, total ion current of separately collected, 90-s elution cuts from the 1st dimension column. *B*, total ion current of automatically pooled fractions corresponding to the ones in *A*. *C*, *histogram* of peptides containing at least 75% of their total mass over all fractions in the indicated number of fractions.

that can be used in a proteomic analysis and therefore its sensitivity. Conversely, many fractions imply long measurement times per sample and may be less beneficial if sample size is limited. In practice, a good compromise, maximizing the effort/gain balance, needs to be found according to the parameters and the goals of the experiment at hand.

To investigate this, we employed HeLa digest as a typical complex proteome and determined the number of identified peptides and proteins as a function of fraction number. We separated peptides into 4, 8, 16, and 24 fractions. We loaded 20 μg of starting material for 4 and 8 fractions, 40 for 16 fractions, and 60 for 24 fractions, so as not to be sample limited for the individual LC MS/MS runs, in which an estimated 2 μg were injected in each case. As expected, separation into 24 fractions, followed by 48 h of total MS measurement time, yielded the largest number of different peptides and proteins groups. In total, 128,966 sequence unique peptides and 10,769 different protein groups were identified by MaxQuant in the HeLa cells with 1% false discovery rate at the protein and peptide levels. Match between runs to all files acquired in this project increased these numbers to 159,024 sequence unique peptides and 11,897 protein groups (supplemental Table 2).

Strikingly, using 16 fractions (32 h measuring time) and 8 fractions (16 h) still resulted in 98 and 95% of those protein identifications, respectively. Even the four fraction experiments identified 90% of the proteins in 8 h, corresponding to only 1/6th of the measuring time of the 24 fractions. However, although the loss of protein identifications was very moderate with decreasing fraction number, this was not as pronounced at the peptide level, where only 91, 78, and 62% of peptides were still found (supplemental Table 2). This observation is explained by the fact that increasing depth of measurement will result in a saturating number of identified proteins, whereas the number of peptides and the sequence coverage of the proteins still increase. Accordingly, Fig. 3*A* shows a rapid rise of identified peptides when accumulating the results of subsequent fractions within one experiment. The first fractions of each experiment add newly identified peptides at an almost linear rate. Here, the four-fraction experiment has a clear advantage as it identifies 28,000 peptides (47,000 with matching) in the first fraction, whereas the first fraction of the 24-fraction experiment only results in 19,000 peptides (32,000 with matching). This reflects the fact that in the four-fraction experiment each of the fractions contains a quarter of total peptides, whereas the 24-fraction experi-

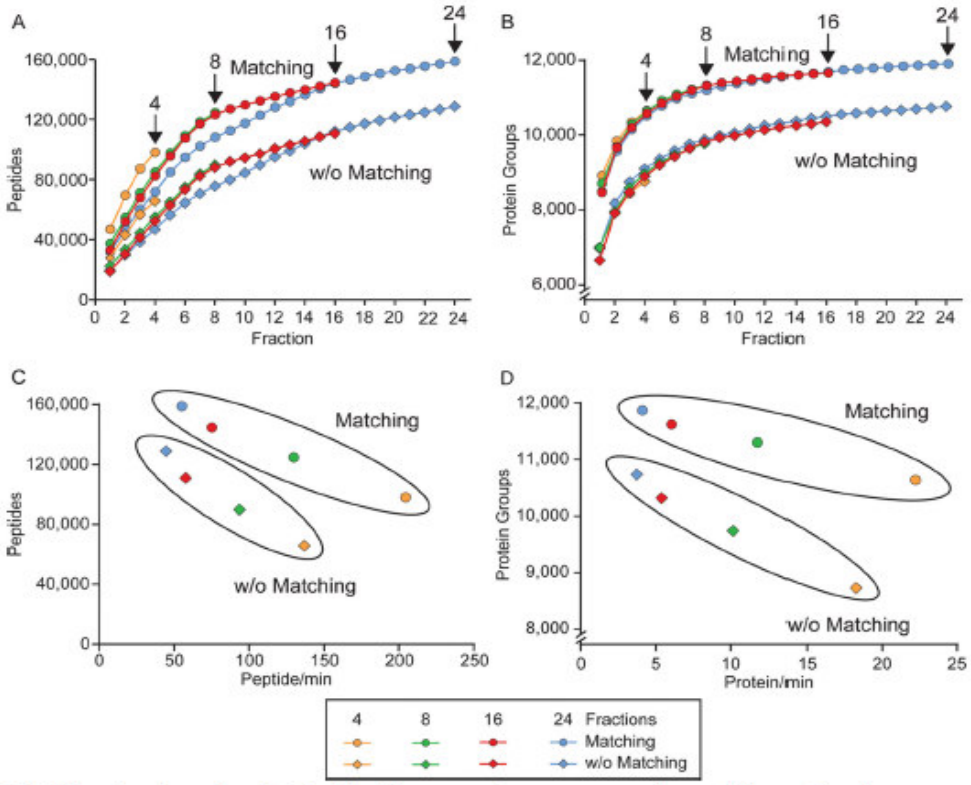


FIG. 3. **Effect of different numbers of pooled fractionations on proteome coverage.** *A*, cumulative number of sequence unique peptides as a function of fraction number for a 4, 8, 16, and 24 fractionation scheme. The *upper curves* (*circles*) are obtained with match between runs enabled in MaxQuant and the *lower curves* (*diamonds*) without match between runs. The last fraction of the experiment is labeled in each case. *B*, same as *A* but for protein numbers. *C*, number of peptides identified per min in 100 min gradient runs as a function of total number of peptides identified. Values enclosed in the *upper ellipse* are those employing match between runs and in the *lower ellipse* without match between runs. High values on the *x* and on the *y* axis are desirable (large number of identifications per min as well as high number of identified peptides). *D*, same as *C* but for protein numbers.

ment leads a smaller number of indefinable peptides despite the higher amount per peptide. The total number of peptides identified in the four-fraction experiment is already matched between 6 and 7 fractions for the 24-fraction experiment, which goes on to yield almost twice the total number of peptides. At the protein level, the identification numbers are essentially only a function of the number of fractions, and that is to say the cumulative number of proteins per fraction are almost identical. The saturation of the curve has largely occurred by fraction 8, especially when using match between runs (Fig. 3*B*).

For the decision of how many fractions the experimenter should choose, the total number of proteins or peptides as well as the effort/gain balance need to be considered, as already mentioned above. For this purpose, we plotted the total number of proteins and peptides against the peptide- or protein-to-time ratio (Fig. 3, *C* and *D*). Again, it appears that eight fractions result in an optimum regarding both factors.

Comparison of the Spider Fractionator to Other High pH Fractionation Systems—The spider fractionator setup was compared with a high flow system (150 μl/min) coupled to a 2.1 mm $\times$ 250 mm C18 column and to a recently released spin column-based high pH reversed-phase peptide fractionation kit (see under "Materials and Methods").

To analyze potential sample losses, we fractionated 1 μg of peptides from the same HeLa digest on all three systems, combined the total eluted volume, and compared the median peptide intensity to a measurement of 1 μg of the same unfractionated peptides. For such low sample amounts, the high flow and the spin column system resulted in much less recovered peptides than the spider fractionator (recoveries were 15, 24, and 81% of the unfractionated sample) (supplemental Fig. 4*A*). Moreover, we fractionated 20 μg of the same HeLa digest with all three systems and compared the median peptide intensities, numbers of identified peptides, and protein groups. The spin column setup allowed only fractionation into eight samples without any concatenation, and for the high

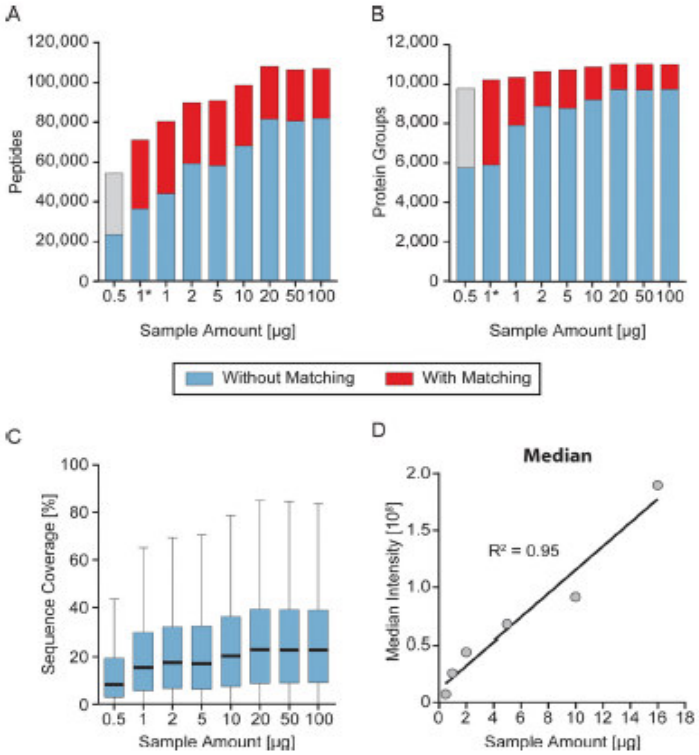


FIG. 4. **Dependence of proteome coverage on sample amount.** *A*, fractionation of a total of 0. 5, 1, 2, 5, 10, 20, 50, and 100 µg of HeLa peptides resulted in the indicated number of identified peptides. For the sample amount 1*, we started with 6,600 HeLa cells, which is equal to 1 µg of starting material, for an in-StageTip digestion with subsequent peptide cleanup and fractionation. *Blue* represents peptides identified by MS/MS and *red* those identified by match between runs in MaxQuant. The *gray bar* indicates that the false discovery rate for match between runs was not validated at this very low sample amount. In the case of 20, 50, and 100 µg of starting material, the volume corresponding to 2 µg of peptide material was injected to avoid overloading the analytical column. *B*, same as *A* for the number of identified proteins. *C*, sequence coverage as a function of starting peptide material displayed as Tukey plots. The *bold black lines* represent the median of all proteins. The *blue box* marks the upper and lower quartile of the sequence coverage and the whiskers the 1.5-fold interquartile range. *D*, median intensity determined as label-free intensity values by MaxQuant for all proteins that were quantified in the dilution series are plotted as a function of initial peptide sample amount. Each value is the median of all protein quantifications.

flow system the samples were concatenated manually. The spider setup resulted in the highest median peptide intensity, identified peptides, and protein groups, followed by the high flow and the spin column system (supplemental Fig. 4, *B–F*).

The experiments for 1 and 20 µg fractionation amounts point in the direction that the spider fractionator had by far the lowest sample loss. Major sample losses could have occurred due to the interaction surfaces in the high flow and the spin column systems.

Moreover, the fully automated concatenation of the spider fractionator saved a lot of hands-on time compared with the two other systems. The major bottlenecks of the high flow system were losses by handling the high volumes and several pipetting and concatenation steps as well as the very long SpeedVac times of up to 6 h for the 12 ml of fractionated volume.

Spider Fractionator Allows Loss Less Fractionation— Because of sample losses associated with high flow rate HPLC systems, fractionation is generally only employed when large sample amounts are available. However, due to its operating principle, the spider fractionator should not have these limitations. To investigate this potential advantage in detail, we fractionated different amounts of digested HeLa peptides (0.5, 1, 2, 5, 10, 20, 50, and 100 µg) into eight pooled fractions each. To minimize potential issues associated with carry-over, we measured the lowest amounts first and on a new column.

First, we analyzed the behavior at the higher sample amounts. For the three highest sample loadings and assuming an equal distribution of peptides in all fractions, more than 2 µg were available per LC MS/MS run, but only this amount was injected. This yielded the same number of identified peptides and proteins (around 11,000 proteins and nearly 110,000 peptides), demonstrating that the spider fractionator equipped with the 250 µm inner diameter column can handle these amounts of sample or more (Fig. 4, *A* and *B*). The average sequence coverage of the proteome was 26% for fractionation of more 10 µg, decreasing gradually to 20% for 1 µg (Fig. 4*C*).

As expected from the smaller amount of peptide material injected into the analytical column, the total number of peptides identified decreased from 10 to 1 µg of starting material (maximum of 1.25 to 0.125 µg per injection). However, the loss of identification was much less than linear, decreasing to about half with 10-fold lower peptide amount. Remarkably, there was very little loss of protein identifications in the same range. In particular, when using matching, the 1 µg total loading still resulted in more than 10,000 different protein groups (7,800 without matching). Loading less than 1 µg did result in a considerable reduction of proteins and peptides.

However, 5,724 proteins were still identified by MS/MS from 23,765 peptides even in this case. Note that this may still not reflect a limitation of the fractionator but instead simply be due to the nanogram amounts of peptide loaded onto the analytical column.

To further investigate the apparent absence of sample losses of the spider fractionator, we next plotted the median intensities of all individual proteins against the amount of injected material up to the 20 μg value (Fig. 4*D*). This resulted in a linear relationship down to the lowest amounts tested, demonstrating that any potential sample losses incurred by the spider fractionator, if they occur at all, are so small that they are not detectable in the setup used here.

To show the applicability of our workflow for a limited amount of starting material, we prepared peptides directly from 6,600 HeLa cells (about 1 μg (29)) by using the in-StageTip protocol (3). The complete material of digested and purified peptides was fractionated using the spider fractionator resulting in 5,869 protein groups and 37,000 peptides without and 10,165 protein groups and 72,110 peptides with matching (Fig. 4, *A* and *B*).

In-depth Measurement of Human Cell Lines—Having established optimal fractionation parameters and sample requirements, we next employed the spider fractionator for the in-depth measurement of 12 different human cell line proteomes. Specifically, we used the 8-fraction, 100 min gradient scheme, resulting in a total measuring time of 16 h, including column loading and equilibration and the 20 μg loading, which was the minimum amount that saturates the number of identifiable peptides. Thus, the entire experiment consumed only 8 days of measurement time and much less than a small cell culture dish for each cell line (corresponding to about 100,000 HeLa cells).

Combined with the one replicate HEK293 cell line (see below), the experiment yielded a total of 199,882 sequence unique tryptic peptides corresponding to 12,444 different protein groups (supplemental Table 3). The median sequence coverage of these protein groups for this cell line dataset was a remarkable 41.3%. In the 13 cell line experiment the median number of identified peptides was 87,769, and this number ranged from 72,471 in the HeLa sample to 105,487 in the GAMG cell line. Applying the match between run algorithm boosted median peptide identifications by a further 47% (Fig. 5*A*).

The median number of proteins identified in each of the singlet experiments was 11,472, and this was very consistent between cell lines (minimum 11,340 in HeLa and maximum 11,634 in GAMG). These are among the deepest proteome results reported for cell lines so far, which is particularly remarkable given the small sample consumption and measurement time. The proteome of the 13 cell lines (12,444 protein groups; 199,882 peptides) mapped to 11,442 protein-coding genes, which made up more than 57% of the 20,154 entries listened in SwissProt at the time of writing.

The large majority of total identified proteins (78%) was also identified in each singlet experiment and almost 90% of them in at least 10 of the 13 experiments (Fig. 5*C*). This implies that the proteomes of these different cell lines are quite similar in terms of expressed and identifiable proteins. It further implies that our data set, acquired with a data-dependent acquisition strategy, is substantially complete and can only have a very small percentage of missing values, despite the use of data driven shotgun proteomics.

Cell lines, including the ones used here, have often been in culture for years or decades and even those that are nominally the same can develop differences over time. Here we had obtained the human embryonic kidney cell line HEK293 from two institutes and treated them as separate entities for the purpose of comparison with different cell lines. Nearly the same number of proteins as well as peptides was identified after fractionation in both cases, and the overlap of each of the fractionated, matched datasets to all identified proteins was 97.1%, with the unique proteins at the lower levels of expression in proteome. The Pearson coefficient for the abundance rank order of the proteomes calculated from the scatter plot in Fig. 5*D* was 0.96, implying that they were very similar at the quantitative level as well. Thus, a simple fractionation experiment establishes that in this case the same cell lines with different origins are very similar at the proteome level.

We calculated the Pearson correlation between all combinations of the 13 cell lines and plotted the result as a heat map (Fig. 5*E*). As expected, the two instances of the HEK293 cells had the highest correlation, whereas the median correlation was 0.83. Against this very high median correlation, a few cell types show a considerably divergent behavior. One of these is EC, an embryonic carcinoma cell line, whose proteome had a correlation to other cell lines down to 0.77. This observation can be readily explained by the fact that EC is the sole undifferentiated cell line in our set. Interestingly, the only cell line to which EC has a high correlation is SMC, another outlier cell line. The proteome of SMC likewise showed a lower overall correlation to the other cell lines (down to 0.73), and in this case the biological explanation is that muscle is developmentally derived from the mesenchyme, whereas the other cell lines are primarily of epithelial origin. Finally, HepG2 likewise correlates less well than an average cell type, presumably reflecting the specialized organismal role of this model of liver function.

To illustrate how readily acquired deep proteomes can shed light on cellular function, we quantitatively compared SMC against a cell line whose proteome had typical correlation values to the other cell lines. For this, we chose LNCaP, a widely used cell model of prostate cancer. The correaltion between SMC and LNCaP was comparatively poor (R=0.76), and the scatter plot reveals a large number of proteins that were expressed at drastically different levels (Fig. 5*F*). Among these, we found the epithelial cell adhesion molecule, which is the classical positive marker used in immunohistochemistry to stain cells of an epithelial origin, to be strongly increased in

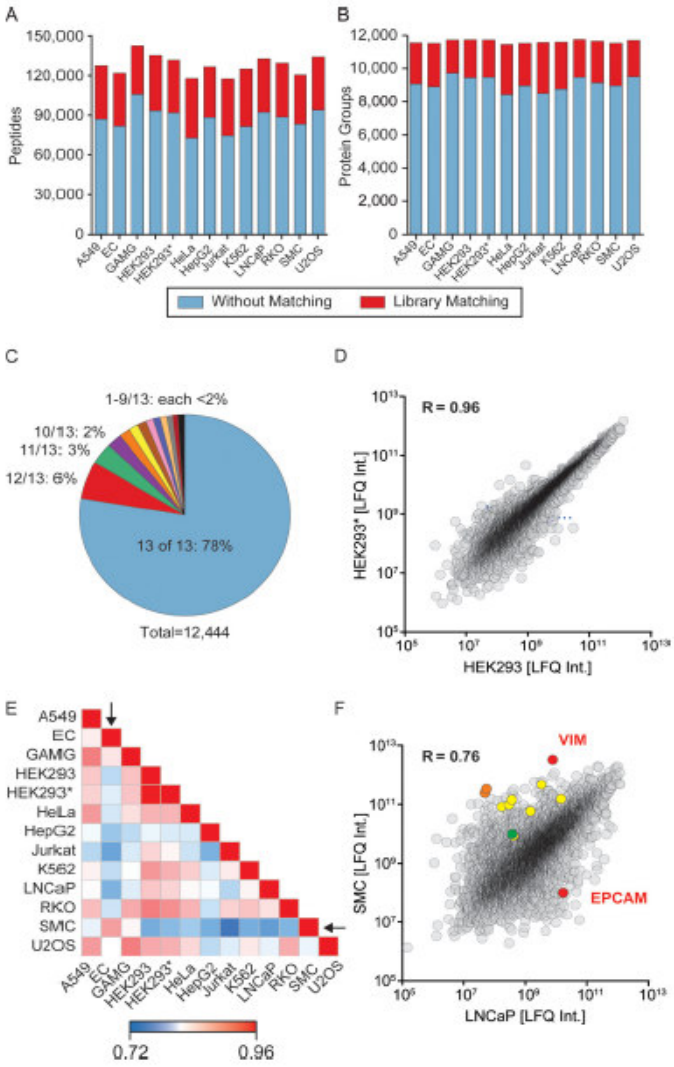


FIG. 5. **Rapid and sensitive sequencing of 13 human cell line proteomes.** *A*, number of sequences of unique peptides identified for the different cell lines indicated on the *x* axis (see supplemental Table 2 for cell line abbreviations). *Blue* indicates the proportion identified by MS/MS and *red* the additional peptides identified by match between runs in MaxQuant. *B*, same as *A* for identified protein numbers. *C*, *pie chart* of the proportion of proteins identified in the indicated number of cell lines. A total of 89% of the proteins identified are also identified in at least 10 of the 13 cell lines. *D*, scatter plot of the label-free intensity (*LFQ*) assigned by MaxQuant to the same protein in two different instances of the same HEK293 cell line (termed HEK293 on the *x* axis and HEK293* on the *y* axis, respectively). *E*, heat map of the rank order correlation of the 13 different proteomes. The SMC and EC cell lines are outliers with respect to their correlations to the others and are indicated by *arrows*. *F*, scatter plot of the proteins quantified in both the LNCaP (epithelial origin) and the SMC cell line (mesenchymal origin). The known epithelial marker epithelial cell adhesion molecule is much more highly expressed in LNCaP, whereas the known mesenchymal marker vimentin is extremely highly expressed in SMC. Vimentin together with LARP6 (colored in *green*) stabilizes type I collagen mRNA for CO1A1 and CO1A2 (colored in *orange*). Several other collagens (COL12A1, -3A1, -5A1, -6A1, -6A2, -6A3, and -7A1 colored in *yellow*) follow the same pattern.

LNCaP. Conversely, the classical mesenchymal marker vimentin was strongly expressed in SMC. It is known that vimentin, together with LARP6, stabilizes type I collagen mRNAs, which in turn leads to up-regulation of the collagens CO1A1 and CO1A2 (39). Our data show that several other collagen isoforms are also strongly expressed in this mesenchymal cell line, suggesting that they may be up-regulated by similar mechanisms (Fig. 5*F*).

DISCUSSION

In the quest for very deep and large scale proteome characterization, pre-fractionation of peptides occupies a pivotal role. We build upon the success of high pH pre-separation as a first dimension coupled to concatenated fractionation sample pooling for the second dimension of analysis. Samples have been separately collected and then combined in these approaches, whereas in the spider fractionator introduced here, concatenation is implemented by a rotating valve. This valve automatically directs sections of the eluent of the first column into a number of tubes corresponding to the number of desired fractions to be analyzed. In this way, any number of pooled fractions with any concatenation volume can in principle be realized. First dimension column diameters and flow rates are much smaller than those typically used in high pH-based proteomics workflows, and the absence of intermediate collection points means that there are no obvious points of sample loss. We implemented the spider fractionator as an assembly of the first dimension column and its acces-

sories, an automated valve, temperature controls, and an automatic fraction collection system for unattended multi-sample fractionation. The device is now routinely used in our laboratory for any project involving pre-fractionation and has proven robust in dozens of projects already.

Here, we characterized the spider fractionator in different dimensions of performance. Comparison of individually combined and pooled samples gave very similar results at high sample amounts, demonstrating the automated pooling scheme correctly implements the concatenated high pH strategy. We obtained quantitative intensity profiles over the pooled fractions for tens of thousands of peptides, which showed that the bulk of each individual peptide mass is localized to a single fraction. The fractionator can be operated in a parameter space defined by the number of fractions and the width of the volume that is concatenated. Likewise, the diameter, flow rate, and stationary phase of its column can be chosen to fit the desired objectives, within at least the range of up to 100 μg, above which a standard high pH setup may be just as effective. Using the same C_{18} material as in our standard LC MS/MS setup, we investigated the influence of the number of fractions on the depth of proteome coverage. Four fractions already led to a very good proteome coverage, and adding additional fractions up to 24 fractions resulted in asymptotic gain at the protein level, while peptide coverage still improved. Considering the tradeoffs in measuring time and available sample quantity in terms of proteins identified per min, we conclude that an eight-fraction scheme is a good compromise in many situations.

Using these parameters, we then demonstrated that the spider fractionator enables extraordinary profiling sensitivity and depth in high pH fractionation experiments. As little as 1 μg of peptide sample, when fractionated, enabled the identification of more than 10,000 proteins. Analysis of protein signal as a function of increased loading of the first dimension column demonstrated that the device has little if any detectable sample loss. We then applied the spider fractionator to the rapid analysis of small amounts of cell line material, a typical challenge for proteomics. In only 16 h we reached a proteome coverage of a median of 11,472 different protein groups (a total of 12,444 different protein goups for all cell lines). In the past, our group employed much longer measurement times and larger sample amounts and still only reached smaller total numbers in cell line systems (11, 40). To our knowledge, these results are also larger than those currently described in any given cell line system in the literature, in any case when considering the amount of protein used and the total measuring time. Furthermore, coverage was extremely consistent between singlet measurements of different cell lines, due to the fact that cell lines tend to have qualitatively similar proteomes (11, 41) and because the depth of proteome coverage reached by our workflow makes our results very robust against 'missing values' that can occur in shotgun proteomics. Although we used a "match between run" strategy in the experiments described here, which resulted in substantial gains, the identification numbers without matching are also very high. Indeed, because of the near absence of sample loss, the maximum amount of peptide material is available for fragmentation and identification. The increased measuring time due to fractionation implies more sequencing events, and thus the nano-fractionator is arguably less reliant on the transfer of peptide identifications.

Various developments can be envisioned to further improve on the results shown here. For instance, the depth of the matching library could be increased, which could be used to reduce the number of fractions without compromising coverage. Although not shown here, the spider fractionator would work equally well with peptide samples that have been derivatized with isotopically labeled mass tags such as iTRAQ or TMT. In this case, a 10-fold decrease in initial sample amount, for instance, would directly translate into a 10-fold reduction in reagent costs. Furthermore, the first dimension column could be further scaled down to enable even smaller sample amounts to be efficiently fractionated and ultra-narrow bore columns and/or ultralow flow rates could also be used in the on-line dimension. Apart from total proteome measurements, the spider fractionator could also be applied to the analysis of post-translational modifications, an area where sensitivity is especially desired. Finally, the scheme presented here is agnostic in regards to acquisition strategies (data-dependent acquisition, data-independent acquisition, or targeted acquisition).

DATA AVAILABILITY

The RAW and processed data associated with this project is deposited in PRIDE proteomeXchange (project accession number PXD005141).

Acknowledgments—We thank all members of the Department of Proteomics and Signal Transduction for help and discussions and in particular Igor Paron, Korbinian Mayr, Gaby Sowa for mass spectrometry assistance; Jürgen Cox for bioinformatic tools; Marco Y. Hein, Garwin Pichler, and Nagarjuna Nagaraj for data interpretation and discussion; and Alina Bartzick, Florian Meier, and Sophia Doll for technical assistance. We gratefully acknowledge an m[4] award by the Bio[M] Munich Biotech Cluster funded by the Bavarian Government, which enabled development of the prototype spider fractionator described here.

* This work was supported in part by the Max Planck Society for the Advancement of Science and by the Novo Nordisk Foundation Grant NNF15CC0001. N. A. K. is a founder and CEO of PreOmics, a start-up company that wishes to commercialize the spider fractionator. The other authors have no conflicts of interest.

S This article contains supplemental material.

** To whom correspondence should be addressed. E-mail: mmann@biochem.mpg.de.

¶ These authors contributed equally to this work.

REFERENCES

1. Beck, M., Claassen, M., and Aebersold, R. (2011) Comprehensive proteomics. *Curr. Opin. Biotechnol.* **22,** 3–8

2. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., and Coon, J. J. (2014) The one hour yeast proteome. *Mol. Cell. Proteomics* **13,** 339–347
3. 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
4. Muñoz, J., and Heck, A. J. (2014) From the human genome to the human proteome. *Angewandte Chemie* **53,** 10864–10866
5. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355
6. Hörth, P., Miller, C. A., Preckel, T., and Wenz, C. (2006) Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Mol. Cell. Proteomics* **5,** 1968–1974
7. Essader, A. S., Cargile, B. J., Bundy, J. L., and Stephenson, J. L., Jr. (2005) A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics* **5,** 24–34
8. Hubner, N. C., Ren, S., and Mann, M. (2008) Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* **8,** 4862–4872
9. Krijgsveld, J., Gauci, S., Dormeyer, W., and Heck, A. J. (2006) In-gel isoelectric focusing of peptides as a tool for improved protein identification. *J. Proteome Res.* **5,** 1721–1730
10. 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
11. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11,** M111.014050
12. Wiśniewski, J. R., Zougman, A., and Mann, M. (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* **8,** 5674–5678
13. Gilar, M., Olivova, P., Daly, A. E., and Gebler, J. C. (2005) Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J. Sep. Sci.* **28,** 1694–1703
14. Toll, H., Oberacher, H., Swart, R., and Huber, C. G. (2005) Separation detection, and identification of peptides by ion-pair reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry at high and low pH. *J. Chromatogr. A* **1079,** 274–286
15. Gilar, M., Olivova, P., Daly, A. E., and Gebler, J. C (2005) Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.* **77,** 6426–6434
16. Gauci, S., Helbig, A. O., Slijper, M., Krijgsveld, J., Heck, A. J., and Mohammed, S. (2009) Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. *Anal. Chem.* **81,** 4493–4501
17. Ducret, A., Van Oostveen, I., Eng, J. K., Yates, J. R., 3rd, and Aebersold, R. (1998) High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci.* **7,** 706–719
18. Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., and Huber, C. G. (2007) Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis. *J. Proteome Res.* **6,** 4363–4373
19. Manadas, B., English, J. A., Wynne, K. J., Cotter, D. R., and Dunn, M. J. (2009) Comparative analysis of OFFGel strong cation exchange with pH gradient, and RP at high pH for first-dimensional separation of peptides from a membrane-enriched protein fraction. *Proteomics* **9,** 5194–5198
20. Wang, Y., Yang, F., Gritsenko, M. A., Wang, Y., Clauss, T., Liu, T., Shen, Y., Monroe, M. E., Lopez-Ferrer, D., Reno, T., Moore, R. J., Klemke, R. L., Camp, D. G., 2nd, and Smith, R.D. (2011) Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11,** 2019–2026
21. Dwivedi, R. C., Spicer, V., Harder, M., Antonovici, M., Ens, W., Standing, K. G., Wilkins, J. A., and Krokhin, O. V. (2008) Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal. Chem.* **80,** 7036–7042
22. Song, C., Ye, M., Han, G., Jiang, X., Wang, F., Yu, Z., Chen, R., and Zou, H. (2010) Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides. *Anal. Chem.* **82,** 53–56
23. Stephanowitz, H., Lange, S., Lang, D., Freund, C., and Krause, E. (2012) Improved two-dimensional reversed-phase-reversed-phase LC-MS/MS approach for identification of peptide-protein interactions. *J. Proteome Res.* **11,** 1175–1183
24. Wang, H., Sun, S., Zhang, Y., Chen, S., Liu, P., and Liu, B. (2015) An off-line high pH reversed-phase fractionation and nano-liquid chromatography-mass spectrometry method for global proteomic profiling of cell lines. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **974,** 90–95
25. Cao, Z., Tang, H. Y., Wang, H., Liu, Q., and Speicher, D. W. (2012) Systematic comparison of fractionation methods for in-depth analysis of plasma proteomes. *J. Proteome Res.* **11,** 3090–3100
26. Keshishian, H., Burgess, M. W., Gillette, M. A., Mertins, P., Clauser, K. R., Mani, D. R., Kuhn, E. W., Farrell, L. A., Gerszten, R. E., and Carr, S. A. (2015) Multiplexed quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. *Mol. Cell. Proteomics* **14,** 2375–2393
27. Batth, T. S., and Olsen, J. V. (2016) Offline high pH reversed-phase peptide fractionation for deep phosphoproteome coverage. *Methods Mol. Biol.* **1355,** 179–192
28. Mertins, P., Yang, F., Liu, T., Mani, D. R., Petyuk, V. A., Gillette, M. A., Clauser, K. R., Qiao, J. W., Gritsenko, M. A., Moore, R. J., Levine, D. A., Townsend, R., Erdmann-Gilmore, P., Snider, J. E., Davies, S. R., *et al.* (2014) Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. Proteomics* **13,** 1690–1704
29. Finka, A., and Goloubinoff, P. (2013) Proteomic data from human cell cultures refine mechanisms of chaperone-mediated protein homeostasis. *Cell Stress Chaperones* **18,** 591–605
30. Scheltema, R. A., Hauschild, J. P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014) The QExactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell. Proteomics* **13,** 3698–3708
31. 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
32. Hahne, H., Pachl, F., Ruprecht, B., Maier, S. K., Klaeger, S., Helm, D., Médard, G., Wilm, M., Lemeer, S., and Kuster, B. (2013) DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods* **10,** 989–991
33. Scheltema, R. A., and Mann, M., (2012) SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11,** 3458–3466
34. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712
35. 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
36. 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
37. 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
38. Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111.013722
39. Challa, A. A., and Stefanovic, B. (2011) A novel role of vimentin filaments: binding and stabilization of collagen mRNAs. *Mol. Cell. Biol.* **31,** 3773–3789
40. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548
41. Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6,** 450

3.4. Article 4: Revisiting Biomarker Discovery by Plasma Proteomics

Authors: Philipp E. Geyer[1,2], Lesca M. Holdt[3], Daniel Teupser[3], and Matthias Mann[1,2]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[3]Institute of Laboratory Medicine, Ludwig-Maximilians-University Munich, 80539 Munich, Germany

Despite its longstanding allure, plasma proteomics has not lived up to its promise of revolutionizing biomarker research and clinical diagnostics. In this review we investigate the reasons that have held plasma proteomics back over the years. We perform a systematic literature research of 381 plasma proteomics publications that aimed to discover new biomarkers. We classify the publications by the approaches that they applied, such as depletion of high abundance proteins, extensive fractionation and chemical labeling. We evaluate problems in the design of the investigated studies, for example small numbers of cases and controls or sample pooling.

This review also discusses current paradigms of biomarker research and develops alternative concepts. Briefly, technological progress and our high throughput plasma proteomics workflow enable us to pursue biomarker research with a 'rectangular' shaped process. Here, many individuals are investigated in each of the phases of the study by shotgun proteomics. Exploring a large cohort already at the discovery stage will result in much more likely biomarker candidates. The further testing of these candidates in a verification and a validation cohort with shotgun proteomics then proceeds with the testing of biomarker panels instead of individual proteins. Investigating as many conditions as possible for as many people and proteins as possible will over time generate a 'big data' matrix. This knowledge base itself will be an incomparable resource that can be mined for connections between different diseases or condition by advanced machine learning algorithms. It can also form the background for the deep phenotyping of humans.

Moreover, we give an overview about the modern clinical laboratory, in which single biomarker are tested. This serves as a basis to discuss how proteomics and multi-protein panels could be translated into clinical practice and we describe first steps in this direction.

Published online: September 26, 2017

Review

OPEN ACCESS

molecular systems biology

Revisiting biomarker discovery by plasma proteomics

Philipp E Geyer[1,2], Lesca M Holdt[3], Daniel Teupser[3] & Matthias Mann[1,2,*]

Abstract

Clinical analysis of blood is the most widespread diagnostic procedure in medicine, and blood biomarkers are used to categorize patients and to support treatment decisions. However, existing biomarkers are far from comprehensive and often lack specificity and new ones are being developed at a very slow rate. As described in this review, mass spectrometry (MS)-based proteomics has become a powerful technology in biological research and it is now poised to allow the characterization of the plasma proteome in great depth. Previous "triangular strategies" aimed at discovering single biomarker candidates in small cohorts, followed by classical immunoassays in much larger validation cohorts. We propose a "rectangular" plasma proteome profiling strategy, in which the proteome patterns of large cohorts are correlated with their phenotypes in health and disease. Translating such concepts into clinical practice will require restructuring several aspects of diagnostic decision-making, and we discuss some first steps in this direction.

Keywords biomarkers; diagnostic; mass spectrometry; plasma proteomics; systems medicine
DOI 10.15252/msb.20156297 | Received 7 June 2017 | Revised 4 August 2017 | Accepted 15 August 2017
Mol Syst Biol. (2017) 13: 942

Introduction

The central and integrating role of blood in human physiology implies that it should be a universal reflection of an individual's state or phenotype. Its cellular components are erythrocytes, thrombocytes, and lymphocytes. The liquid portion is called plasma, when all components are retained, and serum, when the coagulation cascade has been activated (blood clotting). For simplicity, we will use the term "plasma" rather than "serum", since most conclusions apply to both.

Concentrations of various plasma components are routinely determined in clinical practice. These include electrolytes, small molecules, drugs, and proteins. The proteins constituting the plasma proteome can be categorized into three different classes (Fig 1A and B). The first contains abundant proteins with a functional role in blood. These include human serum albumin (HSA, roughly half of total protein mass); apolipoproteins, which have crucial roles in lipid transport and homeostasis; acute phase proteins of the innate immune response; and proteins of the coagulation cascade. The second class are tissue leakage proteins without a dedicated function in the circulation. Examples are enzymes such as aspartate aminotransferase (ASAT) and alanine aminotransferase (ALAT), which are used for the diagnosis of liver diseases, as well as low-level, tissue-specific isoforms of proteins such as cardiac troponins. The third class are signaling molecules like small protein hormones (for instance, insulin) and cytokines, which typically have very low abundances at steady state and are upregulated when needed. Baseline levels of the cytokine interleukin-6 (IL-6) are 5 pg/ml, establishing a minimum 10^{10}-fold dynamic range of the plasma proteome when compared to the concentration of the most abundant protein, HSA, with about 50 mg/ml.

In accepted use, "a biomarker is a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or a response to an exposure or intervention" (FDA-NIH: Biomarker-Working-Group, 2016). For the purpose of this review, we focus specifically on protein or protein modification-based biomarkers. In this sense, there are more than 100 FDA-cleared or FDA-approved clinical plasma or serum tests, mainly in the abundant, functional class (50%), followed by tissue leakage markers (25%), and the rest include receptor ligands, immunoglobulins, and aberrant secretions (Anderson, 2010). Most of these are decades old, and the current introduction rate of novel markers is less than two per year (Anderson *et al*, 2013). A typical test consists of an enzymatic assay or immunoassay against a single target. Clinicians interpret the results in conjunction with other patient information, based on their expert knowledge. Ratios of abundances are only employed in specific cases. Examples are the 60-year-old De Ritis ratio of ASAT/ALAT to differentiate between causes of liver disease (De-Ritis *et al*, 1957) or the more recent sFlt-1/PlGF ratio for diagnosis of preeclampsia (Levine *et al*, 2004).

In contrast to enzymatic and antibody-based methods, mass spectrometry (MS)-based proteomics measures the highly accurate mass and fragmentation spectra of peptides derived from sequence-specific digestion of proteins. Because the masses and sequences of

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
2 Faculty of Health Sciences, NNF Center for Protein Research, 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

Published online: September 26, 2017
Molecular Systems Biology

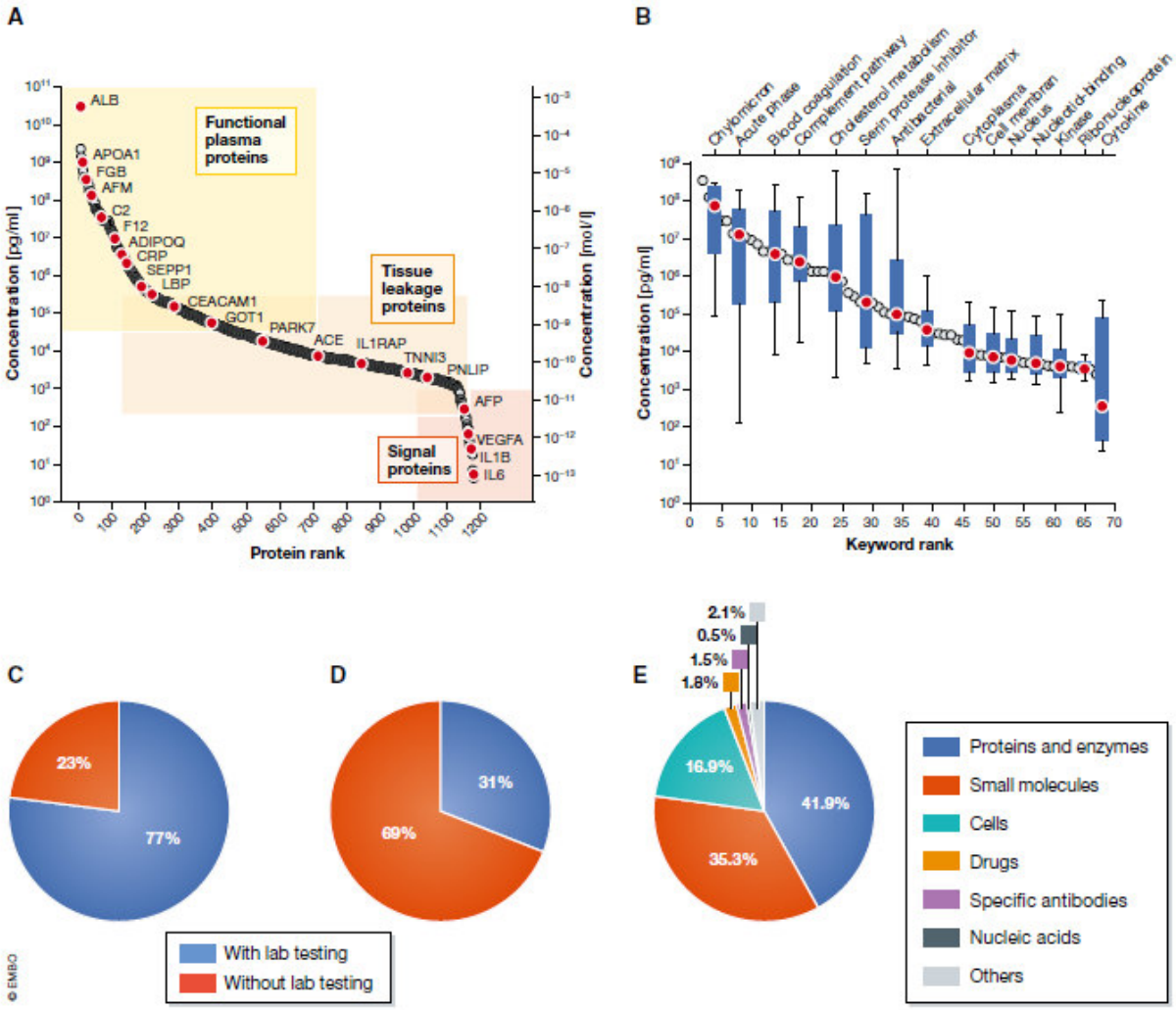


Figure 1. Blood-based laboratory testing in a clinical setting.

(A) Concentration range of plasma proteins with the gene names of several illustrative blood proteins (red dots). Concentrations are in serum or plasma and measured with diverse methods as retrieved from the plasma proteome database in May 2017 (http://www.plasmaproteomedatabase.org/) (Nanjappa *et al*, 2014). (B) Bioinformatic keyword annotation of the plasma proteome database. The blue boxplots with the 10–90% whiskers visualize the range of diverse proteins contributing to distinct functions. (C) Percentage of inpatient admissions receiving blood-based laboratory testing. Numbers are based on 9 million tests performed in the year 2016 at the Institute of Laboratory Medicine, University Hospital Munich. (D) Percentage of outpatient admissions receiving blood-based laboratory testing. (E) Distribution of laboratory tests based on frequency of request. Examples of test for different classes of analytes are as follows: Proteins and enzymes—*liver enzymes, inflammatory proteins, tumor markers*; Small molecules—*electrolytes, substrates, vitamins*; Cells—*red, white blood cells, and platelets*; Drugs—*immunosuppressants, antibiotics, and drugs of abuse*; Specific antibodies—*autoantibodies and antibodies against infectious agents*; and Nucleic acids—*viruses and genetic variants*.

these peptides are unique, proteomics is inherently specific, a constant problem with colorimetric enzyme tests and immunoassays (Wild, 2013). In principle, MS-based proteomics can analyze all the proteins in a system—its proteome—and is in this sense unbiased and hypothesis-free (Aebersold & Mann, 2016). Furthermore, MS methods are ideally suited to discover and quantify post-translational modifications (PTMs) on proteins. These PTMs can also be the basis of diagnostic tests, such as HbA1c levels that serve as a readout of long-term glucose exposure in the context of diabetes. Nevertheless, none of the routinely performed laboratory tests in plasma is based on proteins that were identified by mass-spectrometric approaches, and in routine analysis, MS is so far only employed for measuring small molecules such as drugs and metabolites (Vogeser & Seger, 2016).

Over the past years, the technology of MS-based proteomics has dramatically improved, and it is now a mainstay of all biological research that involves proteins (Cox & Mann, 2011; Altelaar & Heck, 2012; Richards *et al*, 2015; Zhang *et al*, 2016). In

Published online: September 26, 2017

particular, its performance has robustly matured into a sensitivity and dynamic range that makes it interesting for biomarker studies. This review will focus on the prospects of determining proteins in blood by mass spectrometry. We start by empirically assessing the role of proteins in clinical diagnostic today and exhaustively review the literature on previous attempts at finding biomarkers in plasma by MS-based proteomics. So far, proteomics strategies have involved extensive investigations of few samples, to be followed up by targeted approaches in larger cohorts. We discuss how recent advances in technology now enable a new strategy in which deep proteomes are measured for many time points and participants with the prospect to find new biomarkers and biomarker panels. We believe that proteomics will become part of the instrumental routine in the clinical laboratory within the next decade and may even eliminate current technologies in the far future.

The current extent of clinical protein-based diagnostics

Laboratory tests of blood and body fluids aim at disease diagnosis or confirmation, risk prediction, prognosis monitoring, and evaluating treatment effectiveness. It is commonly assumed that 70% of diagnoses are informed by blood testing, even though this number has not been well substantiated. At the Institute of Laboratory Medicine of the University Hospital Munich, laboratory testing is ordered for the vast majority of inpatients at some point during hospitalization (77%; Fig 1C). This fraction is much smaller in patients seen in one of the Hospital's outpatient clinics (31%; Fig 1D). These numbers indicate that hospitalized patients, who are usually sicker, are more likely to receive laboratory tests than ambulatory patients. Based on numbers of requested analyses, clinical routine is dominated by proteins (42% of analyses), followed by small molecules (35%) and cells (17%) (Fig 1E). Thus, already today proteins are the most frequently assayed class of laboratory analytes in clinical practice. We also note that methods suitable for determining plasma proteins have the largest share of the worldwide *in vitro* diagnostics.

Laboratory assays for plasma proteins are based either on classical clinical chemistry, utilizing enzymatic activities of certain plasma proteins, or on antibody-based immunoassays. The costs of enzymatic assays are only in the cent-range, and they run on high-throughput automated analyzers, delivering up to 10,000 test results per hour. In contrast, immunoassays are more expensive (usually several euros/dollars per sample) and throughput of the respective automated analyzers is about 1,000 tests/hour. Large clinical chemistry as well as immunoassay-based analyzers may carry reagents for more than 100 different analytical parameters. Main advantages of immunoassays are a greater degree of flexibility due to the accessibility to plasma proteins devoid of enzymatic activity and a significantly higher sensitivity. Another, clinically relevant issue is the time required per laboratory test. Due to the necessity of immediate decision-making, the majority of enzymatic assays and several immunoassays have to be scaled down to analysis times of < 10 min. In general, immunoassays tend to take longer than enzymatic assays; nevertheless, the vast majority of current automated immunoassays require no more than 30 min.

Systematic review of MS-based plasma proteomics in biomarker research

Plasma proteins had already been investigated by two-dimensional gel electrophoresis in the 1990s, sometimes in combination with MS identification of excised spots. However, these generally identified only a few dozen proteins, and as they preceded MS-based proteomics, they are not discussed in this review. Claims of early cancer detection based on very low-resolution MALDI spectra of plasma that produced patterns but no protein identifications (Petricoin *et al*, 2002) have not been substantiated (Baggerly *et al*, 2004), and these technologies have largely been abandoned today.

To obtain a comprehensive collection of publications dealing with plasma biomarker research and employing MS-based proteomics, we performed an unrestricted PubMed search specifying co-occurrence of the terms "biomarker", "plasma OR serum", "proteome", "proteomics", and "mass spectrometry". This yielded an initial list of 947 publications of which 103 were reviews. We further subtracted studies that did not deal with human subjects or did not involve plasma or serum, leaving 381 original publications (Dataset EV1).

Publications started to appear in 2002 and reached a maximum of 33 per year in 2005, when the special issue on the plasma proteome was released by the Human Proteome Organization (HUPO) (Omenn *et al*, 2005). Two further maxima appeared in 2011 and 2014 with 39 and 43 publications per year, followed by drops in 2013 to 24 and in 2016 to only 20 publications per year (Fig 2A). The observed dynamics contrasts with an ever-expanding community of researchers using proteomics, which is reflected in thousands of publications per year, with a clear upward trend. The ratio of plasma proteome publications to total proteome publications is now < 1% and continues to drop. Given the clear medical need for plasma biomarkers and the success of MS-based proteomics in other areas, this raises the question as to what holds back the field of plasma proteomics.

Of the 381 primary publications, about half dealt with the analytical descriptions of the workflow employed in plasma analysis, whereas the remainder investigated a physiological or pathophysiological question (Fig 2B). About a third of the latter focused on cancer, followed by cardiovascular disease (CVD), topics in human biology, inflammation, diabetes, and infectious diseases (Fig 2B). Clearly, this ordering reflects the interest in the diseases rather than the likelihood of finding relevant changes with the available technology. Only 47% of the studies had any kind of validation of the primary findings (Fig 2C). In half of the cases (24%), these were simple Western blots or ELISAs of candidate proteins performed with the same samples rather than an independent cohort as is usual practice in clinical studies. Only 36 papers used MS-based proteomics to validate potential biomarkers that were proposed independently (Dataset EV1).

The extremely high dynamic range of plasma still makes it difficult to identify more than a few hundred of the most abundant proteins by LC-MS/MS. To partially overcome this challenge, highly abundant plasma proteins are often depleted, generally through columns with immobilized antibodies directed against the top 1 to 20 proteins (Fig 2D). However, these antibodies are never entirely specific and bound proteins—such as HSA—themselves have an affinity for several other proteins (Tu *et al*, 2010; Bellei *et al*, 2011).

Published online: September 26, 2017
Molecular Systems Biology

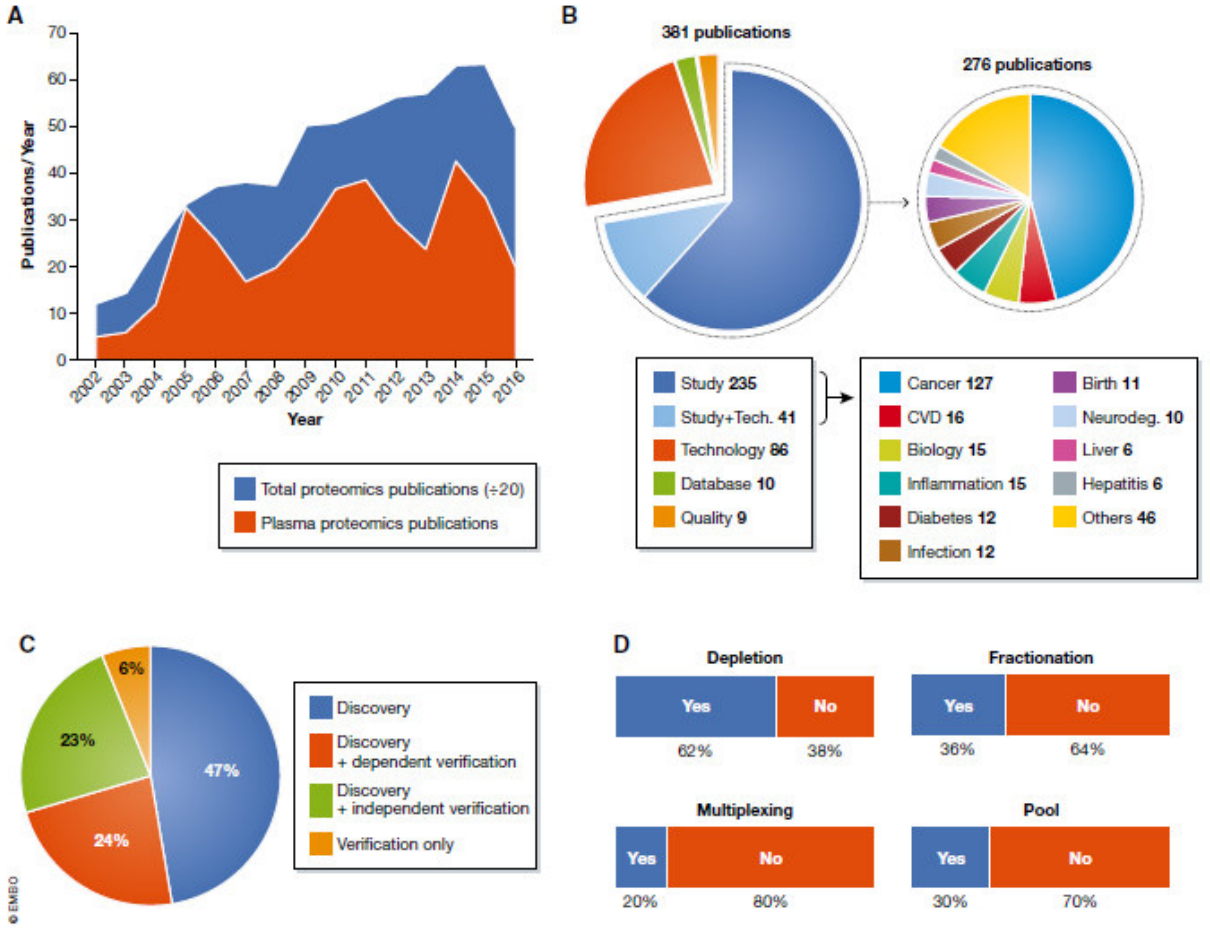


Figure 2. Comprehensive literature review.
(A) Publications using MS-based proteomics in plasma biomarker research (red) compared to the total number of publications in proteomics (blue). (B) Pie charts about the intentions of the investigated studies and proportions of investigated diseases. (C) Overview of the percentage of studies, using discovery and validation phases. (D) Studies using pooled samples, depletion, fractionation, and multiplexing in plasma biomarker research using MS-based proteomics.

Thus, the depleted plasma sample is not a quantitative representation of the original proteome. This is especially true when using "super-depletion" (Qian *et al*, 2008)—a broad mixture of polyclonal antibodies raised against whole plasma—or beads with hexameric peptide mixtures that non-specifically "normalize" the plasma proteome (Thulasiraman *et al*, 2005). Furthermore, these procedures introduce variability and additional expense into the workflow, generally precluding accurate quantification of plasma proteins. Therefore, their use is currently restricted to small discovery projects.

A second strategy to deal with the dynamic range and sensitivity challenge is extensive plasma fractionation, which can be done in various ways at the protein or peptide level. Several studies aiming at in-depth coverage of the plasma proteome by combined depletion and extensive separation (up to hundreds of fractions) identified from several hundred to several thousand proteins (Liu *et al*, 2006; Pan *et al*, 2011; Cao *et al*, 2012; Cole *et al*, 2013; Keshishian *et al*, 2015; Lee *et al*, 2015). Note that many plasma proteome studies continue to use much less stringent statistical identification criteria than the 1% peptide and protein false discovery rates (FDR) that have become standard in MS-based proteomics.

The decrease in throughput implicit in fractionation can partially be recovered by multiplexing. For example, between four and ten samples have been analyzed together using the iTRAQ or TMT strategies, in which samples are labeled with mass neutral tags that give rise to different low mass reporter ions (Kolla *et al*, 2010; Zhou *et al*, 2012; Cominetti *et al*, 2016). Quantification is achieved by fragmenting peptides and quantifying the relative ratios of the reporter ions (Bantscheff *et al*, 2008). Although attractive in principle, these techniques generally suffer from ratio distortion caused by co-isolated peptide species that all contribute to the same reporter ion pattern ("ratio compression"). Regulation of very low-level proteins or those with small but disease-relevant changes may be completely obscured. In shotgun proteomics, eluting

Published online: September 26, 2017

peptides are fragmented in order of intensity (data-dependent acquisition), a semi-stochastic process that may lead to missing values across LC-MS/MS runs. Recently introduced data-independent acquisition strategies more consistently identify peptides across runs (Picotti & Aebersold, 2012; Sajic *et al*, 2015). However, they are incompatible with reporter-ion-based multiplexing because one would quantify the average of groups of peptides.

In about 30% of the studies, plasma samples were pooled to reach a desired plasma proteome coverage within the available measuring time. This approach sacrifices within-group variances and outlier or contaminant proteins in individual samples can skew the whole group, making it all but impossible to assess whether proteins that are different between groups are actually significant on a person-by-person basis.

Partly as a consequence of the demands on instrument time, generally no more than 20–30 samples were analyzed and only few exceeded 500 (Garcia-Bailo *et al*, 2012; Cominetti *et al*, 2016; Lee *et al*, 2017). Considering the large number of measurement points within samples, these are small sample numbers. Accordingly, most studies proposed a few "potential biomarkers", defined as proteins that differ between cases and controls. Furthermore, many of these candidates are unlikely to be specific indicators of the disease in question, because they belong to biological categories that are at best indirectly related to the disease or are likely artifacts of sample preparation (such as keratins and red blood cell proteins). In summary, limitations in proteomics technology and experimental design have prevented the identification of true biomarkers in the published literature to date. To our knowledge, the only possible exception is the OVA1 test, in which the levels of the highly abundant plasma proteins beta-2 macroglobulin, apolipoprotein 1, serum transferrin, and pre-albumin were combined with the previously established ovarian cancer marker CA125 in a narrow, FDA-approved indication (Rai *et al*, 2002; Zhang *et al*, 2004).

Triangular MS-based biomarker discovery and validation strategy

The principal advantage of hypothesis-free MS-based proteomics is that no assumptions need to be made regarding the possible nature and number of potential biomarkers, in stark contrast to single protein measurements in classical biomarker research. Conceptually, MS-based proteomics combines all possible hypothesis-driven biomarker studies for each disease into one and furthermore defines the relation of potential biomarkers to each other. In practice, the challenges of plasma proteomics have so far prevented in-depth and quantitative studies on large cohorts. Instead, a stepwise or "triangular" strategy for biomarker discovery has been advocated, with several phases in which the number of individuals increases from a few to many, whereas the number of proteins decreases from hundreds or thousands to just a few (Rifai *et al*, 2006; Fig 3A).

The typical workflow for hypothesis-free discovery proteomics in plasma is similar to that used in other areas of bottom-up proteomics (Aebersold & Mann, 2016; Altelaar & Heck, 2012; Fig 3B). Briefly, proteins are enzymatically digested into peptides, which are separated by high-pressure liquid chromatography (HPLC) coupled to electrospray ionization. Peptide masses and abundances are measured in the mass spectrometer in full MS scans, whereas a further step of peptide fragmentation produces MS/MS spectra for peptide identification. Well-established proteomics software platforms automatically and statistically rigorously identify peptides in database searches and quantify them (Cox & Mann, 2008; MacLean *et al*, 2010; Rost *et al*, 2014). Furthermore, plasma contains blood components such as lipids that can easily clog HPLC columns, which necessitates dedicated peptide cleanup procedures (Geyer *et al*, 2016a).

Targeted proteomics for candidate verification is a second phase of the triangular strategy (Fig 3C). A relatively small number of proteins (typically < 10) with differential expression in the discovery phase are tested in a larger and ideally independent cohort. Since immunoassays are often not available, targeted MS methods can be employed. The most widespread of these is "multiple reaction monitoring" (MRM—sometimes also called single or selected reaction monitoring—SRM) (Picotti & Aebersold, 2012; Carr *et al*, 2014; Ebhardt *et al*, 2015). For each protein, a set of suitable peptides is selected and their elution and fragmentation behavior is assessed to define an MRM assay. During analysis, the mass spectrometer is programmed to continuously fragment only these peptides as they elute. By monitoring several fragments per peptide, sensitive and specific quantification can be achieved even with low-resolution mass spectrometers. The advantage of MRM over shotgun proteomics for verification is its higher sensitivity and throughput. Inter-laboratory studies have achieved good reproducibility (Addona *et al*, 2009; Abbatiello *et al*, 2015), but reported sensitivities typically do not reach the low ng/ml concentration range and practically achieved multiplexing capabilities are limited to dozens of peptides (Percy *et al*, 2013; Shi *et al*, 2013; Oberbach *et al*, 2014; Wu *et al*, 2015). Nevertheless, two recent studies have reported the targeting of 82 and 192 proteins, respectively (Ozcan *et al*, 2017; Percy *et al*, 2017). The sensitivity of MRM can be improved to the low ng/ml or even high pg/ml ranges by more extensive sample preprocessing with depletion or fractionation (Burgess *et al*, 2014; Kim *et al*, 2015; Nie *et al*, 2017).

Absolute and accurate quantification requires internal standards —generally heavy isotope versions of the monitored peptides. Synthesized heavy peptides are added after digestion, creating a source of quantitative inaccuracy since the variability of protein digestion is not taken into account. This can be addressed by embedding the peptide in its original sequence context, for instance, in the SILAC-PrEST strategy, in which a 150- to 250-amino acid stretch of each protein of interest, fused to a quantification tag, is recombinant expressed in a heavy form (Zeiler *et al*, 2012; Edfors *et al*, 2014; Geyer *et al*, 2016a).

Targeted methods can also be combined with immunoenrichment of proteins or peptides. For instance, in "stable isotope standards and capture by anti-peptide antibodies" (SISCAPA) specific peptides are immunoprecipitated together with their heavy-labeled counterparts, followed by rapid MS-based readout (Anderson *et al*, 2004; Razavi *et al*, 2016). This combines the enrichment capabilities of antibodies with the specificity of MS detection; however, development of assays can be difficult and time-consuming—narrowing the advantage compared to purely antibody-based methods.

The final phase in the triangular strategy is the validation with immunoassays, a field that has matured over decades. For maximum specificity, sandwich assays are typically preferred (Fig 3D). While they are costly and laborious to develop, they can achieve

Published online: September 26, 2017
Molecular Systems Biology

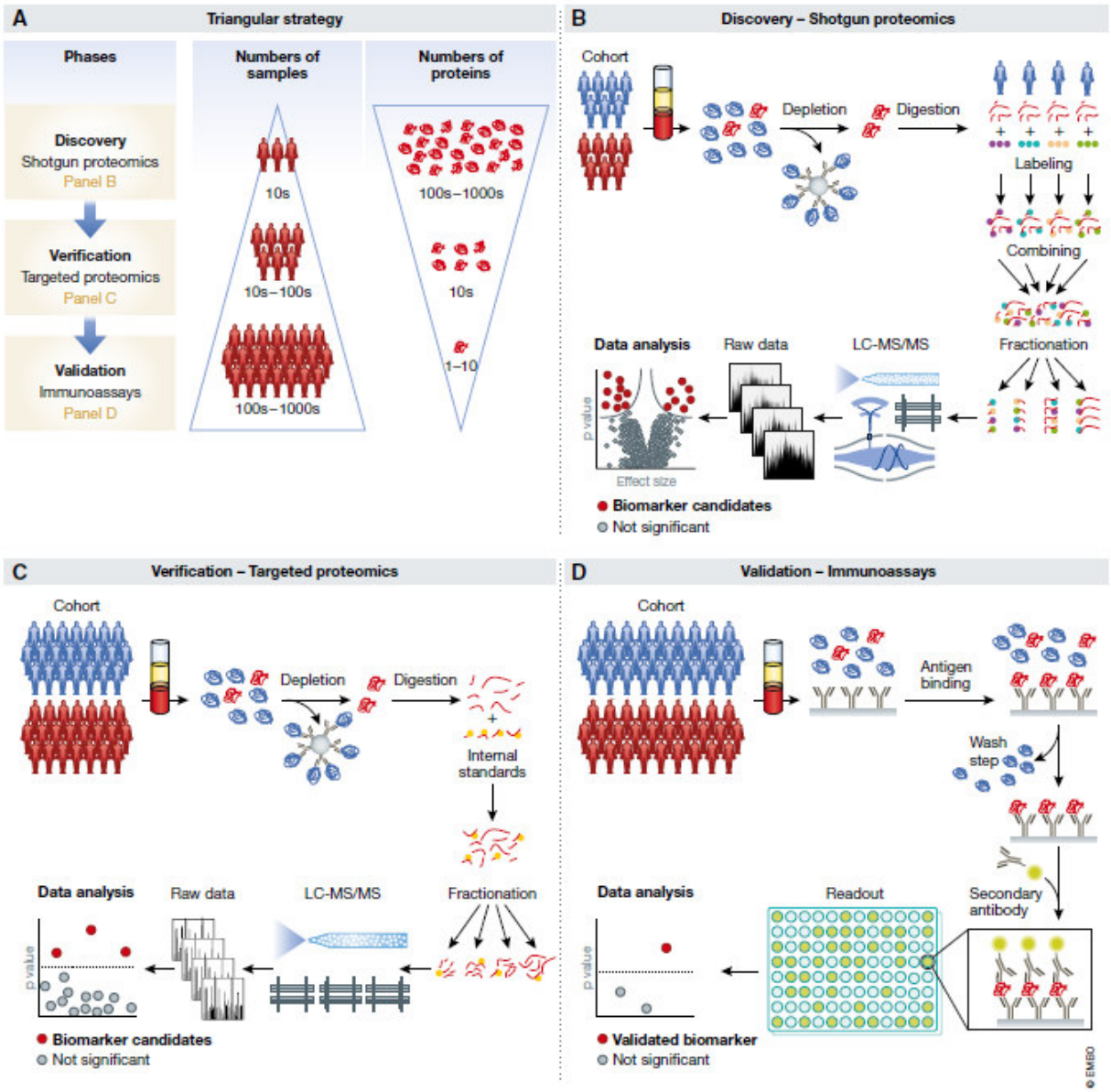


Figure 3. Current paradigms in plasma biomarker research ("triangular approach").
(A) A relatively small number of cases and controls are analyzed by hypothesis-free discovery proteomics in great depth, ideally leading to the quantification of thousands of proteins (top layer in the panel). This may yield tens of candidates with differential expression that are screened by targeted proteomics methods in cohorts of moderate size (middle layer). Finally, for one or a few of the remaining candidates, immunoassays are developed, which are then validates in large cohorts and applied in the clinic (bottom layer). (B) Workflow for hypothesis-free discovery proteomics. (C) Targeted proteomics for candidate verification. (D) Development of immunoassays for clinical validation and application.

high sensitivity and high throughput. Even cohorts with thousands of participants can be tested with this technology, but only for one or a few candidate biomarkers. Such large numbers may be necessary to establish specificity not only against controls but also with respect to other diseases. Standard requirements include insuring adequate statistical power and replication in an independent population. Today, such clinical studies can be expensive multi-year endeavors, partly explaining the paucity of new biomarkers.

Immunoassays have some inherent limitations, mostly related to antigen-antibody recognition. These include cross-reactivity,

Published online: September 26, 2017

interference by background molecules such as triglycerides, and non-linear response ("hook effect") (Hoofnagle & Wener, 2009; Wild, 2013). Furthermore, not all clinically important protein variants are easily recognizable by antibody-based assays. Given these limitations, MS-based methods would be attractive alternatives in at least some large-scale clinical trials, but this requires much more robust, sensitive, and higher throughput technologies than those available today.

Over the last decade, the proteomics community has developed guidelines for proper development of biomarkers that discuss quality standards and emphasize the importance of selecting adequate cohorts that ensure statistical significance of the findings as well as specificity of potential biomarkers and their potential clinical application (Luque-Garcia & Neubert, 2007; Paulovich *et al*, 2008; Mischak *et al*, 2010; Surinova *et al*, 2011; Skates *et al*, 2013; Parker & Borchers, 2014; Hoofnagle *et al*, 2016).

Not surprisingly in view of the rigorous requirements of the triangular strategy, there are few, if any, reports in which it has been applied completely and successfully. This may also partly be due to the fact that three different technologies—shotgun proteomics, targeted proteomics, and immunoassay development—are involved. Many publications just describe the first phase or only combine it with immunoassay verification in the same cohort (Dataset EV1).

Among the studies with more than a few participants and with some verification, the majority selected candidates of interest and performed Western blotting, ELISA, or MRM assays. A representative example is the study by Zhang *et al* (2012) in which depleted plasma of 10 colorectal cancer patients versus controls was labeled with iTRAQ and fractionated, leading to the identification of 72 proteins. Among several up- or downregulated proteins, ORM2 was followed up by ELISAs in 419 individuals. Since this protein is a part of the innate immune system (like the other two upregulated candidates), it is unlikely to be a specific cancer marker. In another study, super-depletion, iTRAQ labeling, and fractionation identified 830 proteins in a discovery cohort of 751 patients with cardiovascular events and controls that had been reduced to 50 pooled samples (Juhasz *et al*, 2011). The known markers CRP and fibronectin were selected from the list of candidates and found to be significantly upregulated in the original cohort by immunoassays against these proteins. In a heart transplantation study, analysis of depleted and iTRAQ-labeled plasma from 26 patients at five time points before and after surgery identified a total of more than 900 proteins (273 per individual; Cohen *et al*, 2013). MRM assays and ELISAs against five medium-abundant proteins in a partially independent follow-up cohort of 43 individuals served to develop a computational pipeline for risk markers for organ rejection. In an approach of potential clinical utility, depleted plasma from a mouse model of breast cancer allowed the identification of more than 1,000 plasma proteins from which 88 were selected for MRM assays in an independent verification cohort of 80 animals (Whiteaker *et al*, 2011).

Rectangular biomarker strategy and plasma proteome profiling

In the last few years, the community has substantially improved all aspects of the workflow of MS-based proteomics. In sample preparation, laborious, multi-stage preparation workflows have been replaced by robust, single-vial processing with a minimum of manipulation steps. This also helps with automation and increases throughput. The sensitivity and sequencing speed of MS instruments have improved severalfold. The entire LC-MS/MS system has become much more robust, although this is still far from what will be needed for routine clinical application. Finally, bioinformatic analysis of the results is now statistically sound and straightforward to use and increasingly enables correlation of MS results with a wide range of other classical clinical and additional "omics" data. Illustrating the power of cutting edge MS-based proteomics, cell lines can now routinely be quantified to a depth of more than 10,000 different proteins in a relatively short time, sometimes even without any fractionation (Mann *et al*, 2013; Richards *et al*, 2015; Sharma *et al*, 2015; Bekker-Jensen *et al*, 2017).

Given this technological progress of proteomics in cell line and tissue samples, we asked whether one could also develop a fast and automated workflow that would quantify the plasma proteome in depth in a large number of samples (Geyer *et al*, 2016a). We reasoned that this would then enable a "rectangular strategy" in which as many proteins as possible are measured for as many individuals and conditions as possible. In contrast to the triangular workflow, the initial discovery cohort would be much larger, ideally encompassing hundreds or thousands of participants, resulting in a greater likelihood to reveal any patterns that might differentiate the investigated groups or conditions. These larger initial numbers of plasma proteomes would allow the discovery of statistically significant, but small differences and changes associated with a group of proteins. In the proposed rectangular strategy, discovery and validation cohorts would both be measured by shotgun proteomics in great depth. This removes the dependency of validation on discovery, meaning that both cohorts can be analyzed together (Fig 4A). Moreover, having separate cohorts allows unmasking study-specific confounders. A further advantage of the rectangular strategy is its ability to discover and validate protein patterns that are characteristic of particular health or disease states, in addition to single biomarker candidates, something that is unattainable with the triangular approach.

Interestingly, an analogous change of concept has already happened a number of years ago for genome-wide association studies (GWAS). Researchers in this field found that joint analysis of as many samples as possible was superior to a sequential pipeline (Skol *et al*, 2006). In proteomics, the obvious challenge is achieving sufficient proteomics depth in a short time, ideally without depletion and in a robust workflow. This goal has not been achieved at the time of writing, but the current rate of technological improvements promises to make it feasible in the near future. Below, we discuss four examples of this emerging approach.

The first of these investigated a cohort of 36 monozygotic and 22 dizygotic twin pairs to determine the influence of genetic background on the levels of plasma proteins (Liu *et al*, 2015). The authors established a spectral library using depleted, fractionated, and pooled samples and measured their samples with data-independent acquisition (DIA). A total of 232 plasma samples were then measured with 35-min gradients in a data-independent mode, leading to the consistent quantification of 1,904 peptides and 342 proteins. Interestingly, protein levels were often relatively stable within individuals as compared to between individuals.

Published online: September 26, 2017
Molecular Systems Biology

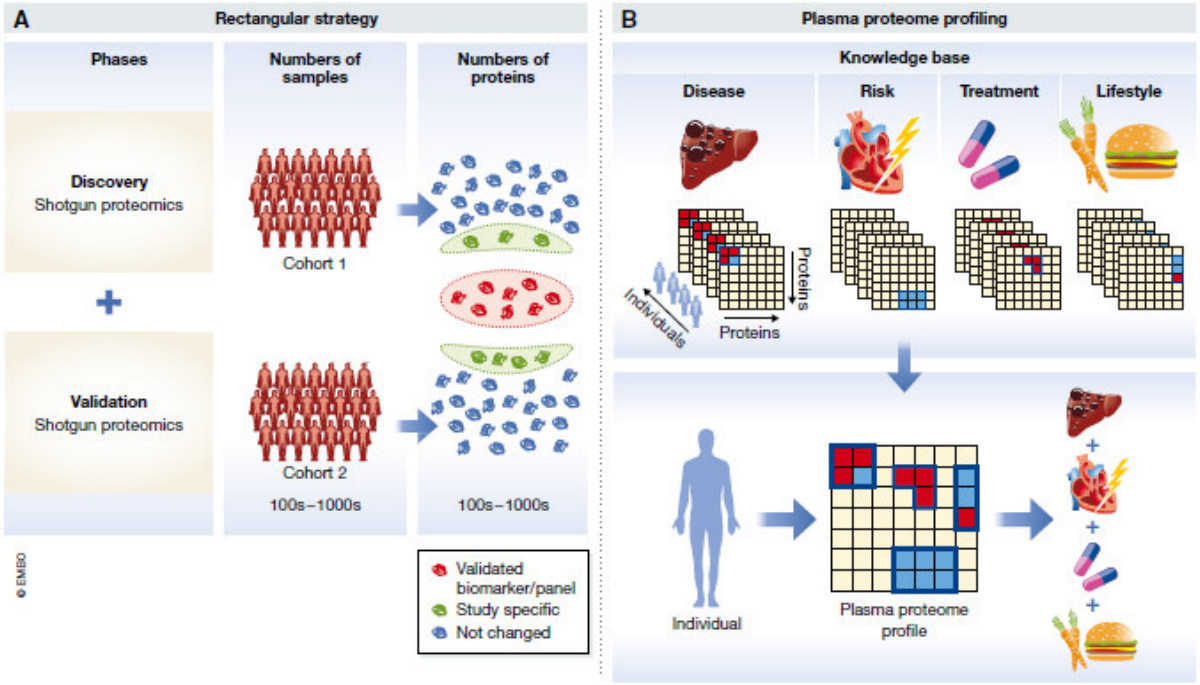


Figure 4. Rectangular workflow.
(A) A large cohort is investigated in the discovery phase with as much proteome coverage as possible. In the validation phase, another cohort is analyzed to confirm the biomarker candidates, but it uses the same technology and similar cohort size. Both cohorts can be analyzed in parallel, but only the proteins that are statistically significantly different in both studies (orange as opposed to green circle in the right-hand part of panel A) are validated biomarkers. (B) Plasma proteome profiling of diverse lifestyle, disease, treatment, or other relevant alterations will over time build up a knowledge base that connects plasma protein changes to perturbations in a general manner (upper panel). The plasma proteome profile of a given individual can then be deconvoluted using the information and algorithms associated with the knowledge base (lower panel).

Furthermore, there were clear indications for the levels of some proteins to be under genetic control. For instance, processes connected to "immune response" and "blood coagulation" tended to be heritable, whereas those associated with "hormone response" did not. Although a pioneering study, the number of plasma proteomes analyzed was relatively small in view of the generality of the research question posed. Generally, genetics studies routinely investigate thousands of participants to tease out subtle heritable effects, illustrating the need for much higher throughput in clinical proteomics.

Malmström *et al* (2016) induced sepsis in mice by injecting *S. pyogenes* and followed their plasma proteomes through three time points on non-depleted, non-fractionated samples. A library of diverse mouse tissues was employed to support data-independent identifications as well as to determine the origin of tissue damage proteins. In this way, 2-h runs quantified an average of 786 mouse proteins, although it should be noted that proper FDR criteria for inferring peptide identities in the complex DIA MS/MS spectra are still being discussed (Nesvizhskii *et al*, 2007; Bruderer *et al*, 2017; Rosenberger *et al*, 2017). Several expected categories of plasma proteins increased during sepsis, as well as some markers associated with damage to the vascular system. Some of the changes were related to mobilization of the immune system against the pathogen, and others appeared to be correlated with necrosis in severely affected animals.

In a workflow termed "plasma proteome profiling", we focused on the rapid and robust analysis of only 1 µl of undepleted plasma from a single fingerpick (Geyer *et al*, 2016a). Total gradient time was only 20 min, enabling extensive investigation of analytical, intra-assay, intra-individual, and inter-individual variation of the plasma proteome. Based on the quantification of 300 plasma proteins, about 50 FDA-approved biomarkers were covered with label-free quantification (CV < 20%). Rapid analysis of a wide range of samples also revealed different sets of quality markers that clearly classified samples with evidence of red blood cell lysis, those with partial activation of the coagulation cascade due to inappropriate sample handling, and those with exogenous contaminations such as keratins. Even though this study provided a useful overview of the information content of the plasma proteome, the depth of coverage was not yet sufficient to address low-level, regulatory plasma proteins. A single step of fractionation yielded a quantitative plasma proteome of about 1,000 proteins, including 183 proteins with a reported concentration of < 10 ng/ml, however at the cost of longer measurement times per sample.

An improved version of the plasma proteome profiling workflow allowed the robotic preparation and measurement of nearly 1,300 plasma proteome samples in a weight loss study (Geyer *et al*, 2016b). Quadruplicate analysis of individuals captured the dynamics of an average of 437 proteins upon losing weight and over a year of

Published online: September 26, 2017

weight maintenance. Weight loss itself had a broad effect on the human plasma proteome with 93 significantly changed proteins. Quantitative differences were often small but physiologically meaningful, such as a 16% reduction of the adipocyte-secreted factor SERPINF1. The longitudinal study design in which the individuals sustained an average 12% weight loss for 1 year allowed capturing the long-term dynamics of the plasma proteome and categorizing it into proteins stable within versus between individuals. Multi-protein patterns reflected the lipid homeostasis system (apolipoprotein family), low-level inflammation, and insulin resistance. These patterns quantified the benefits of weight loss at the level of the individual, potentially opening up for individualized treatment and lifestyle recommendations.

Together, these studies also highlight the advantages of longitudinal over cross-sectional study designs, because the plasma proteome tends to be much more constant within an individual over time than between different individuals. Furthermore, they are similar in that they use less bias-prone undepleted plasma, and identify many proteins in a given analysis time (up to 20 proteins/min).

Regarding the question of how many proteins should be covered, we found that a proteomic depth of more than 1,500 proteins in undepleted plasma allows the coverage of tissue leakage proteins such as liver-based lipoprotein receptors and is within reach of technological capabilities that are currently being developed. Among the first 300 highest abundant proteins, every fourth protein is a biomarker, whereas in the next 1,200 proteins, it is only every 25th protein (Fig 5). As there is no *a priori* reason that biomarkers should have a skewed abundance distribution, this suggests that many biomarkers are still to be found. We believe that the real promise of plasma proteome profiling using the rectangular strategy is that it can discover proteins and protein patterns that have not been considered as biomarkers yet. The exponential increase in the underlying LC-MS/MS technology will stimulate a matching increase in the number of plasma proteome datasets recorded in laboratories around the world. This will create an extensive database of plasma proteomes and their dynamics, involving many clinical studies and individuals. Such data could then be aggregated to build up a knowledge base that connects proteome states to a wide diversity of "perturbations", including diseases, risks, treatments, and lifestyles. At a minimum, this approach will reveal all the different conditions in which a given set of biomarkers is involved, in addition to the specific context where they were discovered. Proteome overlap between disease conditions could reveal commonalities between them (Fig 4B, upper panel). An individual's plasma proteome profile and its dynamics could then be interpreted by comparing it to the global knowledge base. This could be used to deconvolute co-morbidities and to guide treatment and monitor effectiveness (Fig 4B, lower panel).

Standardization of the proteomic biomarker discovery pipeline

It has been suggested that the current lack of biomarkers making their way into the market may be the result of various technical, scientific, and political aspects including undervaluation, resulting from inconsistent regulatory standards, and lack of evidence for analytical validity and clinical utility (Hayes *et al*, 2013). To overcome these challenges, systematic pipelines for biomarker development have been advocated (Pavlou *et al*, 2013; Duffy *et al*, 2015). In the context of moving from a triangular to a rectangular strategy of biomarker discovery, it will be particularly important to consider the following principles.

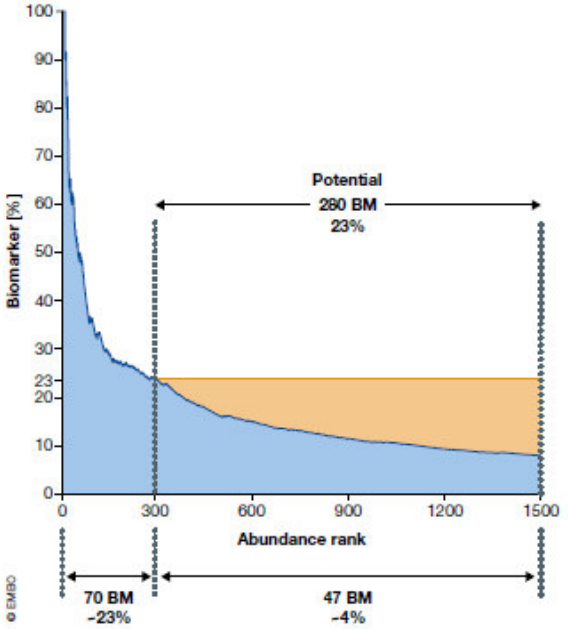


Figure 5. Biomarker distribution across the abundance range.
The blue area illustrates the percentage of biomarker (BM) as a function of increasing depth of the plasma proteome. Within the 300 most abundant proteins, 23% are already known biomarkers. The top of the yellow region extrapolates this proportion to the remainder of the plasma proteome. If the portion of biomarkers remained as high as it is in the 300 most abundant proteins, there are at least 233 potential biomarkers to be discovered (yellow area of the figure).

(1) Analytical performance characteristics: Analytical validity is the capacity of a test to provide an accurate and reliable measurement of a biomarker. Establishment of analytical validity of the plasma proteomics methodology will be key, because the same method will often be carried on from discovery to application. Detailed standards to determine analytical validity have been developed by the Clinical and Laboratory Standards Institute (CLSI) (www.clsi.org). An overview can be found in Grant and Hoofnagle (2014) and Jennings *et al* (2009). Some of these standards have been recognized by the U.S. Food and Drug Administration (FDA) and are accepted for bringing *in vitro* diagnostic test to the market (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfstandards/search.cfm). Even though starting off with a full analytical validation conforming to FDA standards might be prohibitive in biomarker discovery, at least some of the key criteria, such as carryover, accuracy, precision, analytical sensitivity, analytical specificity, and limit of quantification, should be tested early on. This is in line with what we advocate in the context of the rectangular strategy and is also in the interest of saving resources, because the step following biomarker discovery is biomarker validation, where analytical validity will be mandatory.

Published online: September 26, 2017
Molecular Systems Biology

(2) Clinical performance characteristics: Clinical validity relates to the associated diseases and clinical conditions of patients and is different from analytical validity, which focuses on the correct measurement of analytes targeted by the assay. According to International Standard Organization (ISO) 15189 and ISO 17025, validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled". Therefore, establishing clinical performance is the main goal in the validation phase of a biomarker. Clinical performance characteristics include (i) defining normal reference ranges by measuring cohorts of apparently healthy individuals, (ii) determining clinical sensitivity, which is defined as the proportion of individuals who have the disease and are tested positive, and (iii) determining clinical specificity, which is defined as the proportion of disease-free individuals who are tested negative. Derived statistics such as receiver operating characteristic (ROC) plots are particularly helpful in assessing the clinical performance of biomarkers (Zweig & Campbell, 1993; Obuchowski *et al*, 2004).

(3) Study design and pre-analytics: Careful study design and well-controlled pre-analytical conditions are key requirements at any time during a biomarker study. With respect to study design, it is mandatory to clearly define the clinical question and the medical need that should be addressed by the biomarker. A common problem in biomarker studies is that samples from cases and controls have been collected independently and are mismatched for age, ethnicity, sex, and other factors that may or may not lead to unintentional bias (Duffy *et al*, 2015). Methods against bias include proper study design as well as precise and deep clinical phenotyping of participants, using systematic classifications such as the International Statistical Classification of Diseases (http://apps.who.int/classifications/icd10/browse/2016/en) or the human phenome ontology (Kohler *et al*, 2017). In this way, if a person has multiple disease conditions, this can be properly accounted for. Sample collection is important as well, and it is imperative that all samples (including cases and controls) are treated equally from blood drawing to the analytical phase. Another critical step in many biomarker studies is biobanking. When employing ELISAs, we have found that storage of protein-based biomarkers for 3 months requires temperatures of −80°C or below (Zander *et al*, 2014). Sample stability for longer periods is only poorly investigated. However, in our experience, shotgun proteomics has a high tolerance for variation in sample history, because there are no protein epitopes that need to be preserved and even partial protein degradation may be tolerable as long as the majority of subsequently generated proteolytic peptides remain unaltered.

The road to clinical application

The current progress in plasma proteomics opens exciting novel avenues for research and the clinic. How likely is it, given all the aforementioned precautions that the outlined approaches will lead to the discovery of novel protein-based biomarkers? And what will the proteomic biomarker of the future look like? A key theme in this context is the discriminative power of a biomarker to distinguish between the presence and absence of a particular disease state or risk, in other words its clinical performance. Examples of currently used biomarkers with high specificity and high sensitivity are cardiac troponins, which are structural proteins specifically expressed in cardiomyocytes and therefore highly specific for myocardial damage. For this reason, cardiac troponins have even been incorporated into the universal definition of myocardial infarction (Roffi *et al*, 2016).

It is likely that proteomics approaches will succeed in the identification of additional biomarkers with similar performance, at least for certain diseases. In fact, we need to be aware that most biomarkers used today are either highly abundant or originate from a known pathophysiological context. As a thought experiment, we have extrapolated the ratio of the number of biomarkers relative to the number of proteins in the high abundance range to lower abundance protein range, which indicates the potential for several hundred novel biomarkers, which might be accessible with appropriate technology (Fig 5). In analogy to GWAS, where a significant number of hits turned out to be related to previously unknown pathophysiology of the investigated disease (Holdt & Teupser, 2013; Manolio, 2013), it is quite likely that new markers, which have hidden below the radar of previous strategies, will be identified by novel systematic proteomics approaches. These biomarkers may also have the potential to improve our understanding of disease pathophysiology not only in diagnostics but also for therapy. Note, however, that the identified biomarkers might not always be directly involved in the disease pathophysiology but may only be associated with it.

The human genome encodes for about 20,000 protein coding genes, which is opposed to more than 14,500 diseases classified by an ICD code. This makes it even conceptually difficult to imagine that one gene or protein is associated with each disease condition, as is often implied in current efforts to find biomarkers. In contrast, the rectangular strategy, allowing to screen large cohorts for multiple markers, holds great promise to discover and validate protein patterns that are characteristic of particular health or disease states. Indeed, multi-marker combinations may achieve higher specificity and sensitivity compared to single markers and first tools for selecting accurate marker combinations out of omics data have been developed (Mazzara *et al*, 2017). However, a common problem with new biomarkers combined with existing ones is that they frequently only lead to minor classification improvements, in particular when added to well-performing ones (Pencina *et al*, 2010). Contrary to common and intuitive assumptions, it has been shown that correlation (especially negative correlation) between predictors can be beneficial for discrimination (Demler *et al*, 2013). More research in this area is clearly warranted, and new proteomics technologies will provide the data required for the validation of appropriate statistical methods.

Finally, how will these markers be applicable in a clinical setting? We favor in-depth measurement of the entire plasma proteome regardless of the occasion, as this provides the most complete information. Over time, it adds to the longitudinal plasma proteome profile that could usefully be obtained even of healthy subjects. As mentioned above, plasma protein levels tend to generally be stable but person-specific, allowing individual-specific interpretation instead of population-based cutoff values. Furthermore, co-morbidities are the rule rather than the exception in many patient groups. These are much more easily and economically addressed by a generic diagnostic test such as plasma proteomic profiling rather than a succession of individual ELISA tests. Nevertheless, there would clearly be many situations in which a universal test will not be appropriate because it may inadvertently uncover other

Published online: September 26, 2017

conditions. Similar issues arise with other technologies such as genome sequencing or imagining techniques, where individuals may not want to learn about predispositions that they can do little about. In these cases and generally to avoid the risk of overdiagnosis (Hofmann & Welch, 2017), clinicians may prefer plasma proteomics tests of a more directed nature that focuses on a particular disease context. This could be accomplished by the above-mentioned MS techniques targeting a panel of proteins, rather than the entire proteome.

For either whole-proteome diagnostic tests or panel-based tests, the question arises how doctors would deal with the resulting multi-dimensional data. Figure 6A shows the current single/oligo biomarker diagnostics, which is integrated into decision-making largely based on clinical knowledge and intuition. New biomarkers clearly hold the promise of better informed clinical decisions, but also imply the risk of generating patterns exceeding the human cognitive capacity of interpretation (Fig 6B). A solution to this problem might be the algorithmic combination of multiple biomarkers into a quantitative panel, possibly combined with clinical metadata, which might substantially aid clinical decision-making (Fig 6C). Given rapid developments in "deep learning" and "big data", it will be very interesting to see whether this combination can provide powerful and unprecedented associations. We note that there are already multi-parameter scores in clinical practice today. For instance, the Child–Pugh score and the Framingham Risk Score have each combined several blood values with patient data, to aid

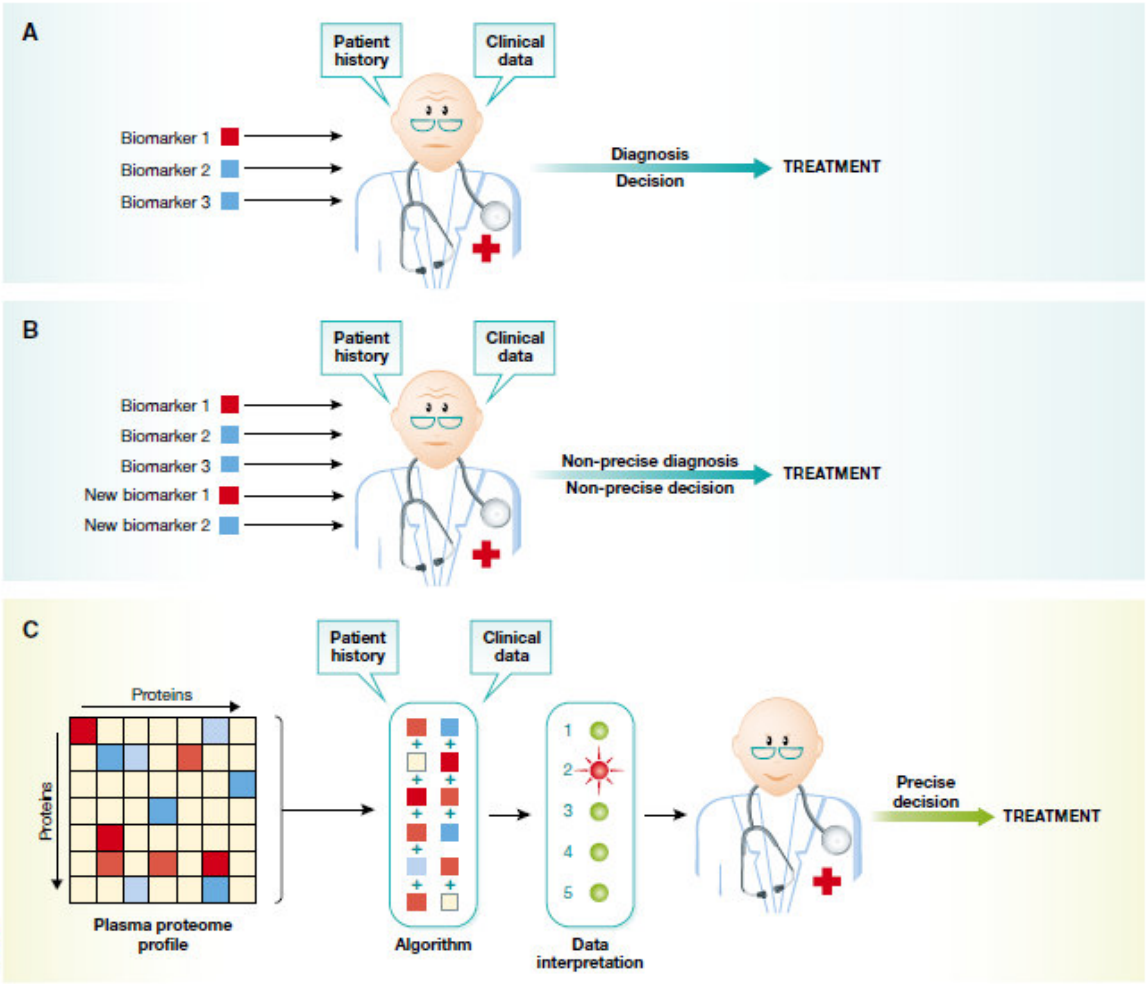


Figure 6. Implementation of proteomic data in clinical decisions.
(A) Currently, physicians make treatment decisions on the basis of a few plasma biomarker tests, combined with patient history and clinical data (upper panel). (B) Adding new biomarkers would quickly overwhelm the current paradigm—leading to suboptimal clinical decisions. (C) Multi-protein panels and the data from past studies (the knowledge base in Fig 4B) are combined algorithmically. This will aid the physician in making more precise recommendations for treatment, while still taking patient history and other clinical data into account.

Published online: September 26, 2017
Molecular Systems Biology

clinician's decision in treating liver disease and cardiovascular treatment, respectively, for decades. This also suggests a way how plasma proteomics could be accepted into evidence-based medical practice, a huge challenge given the many parameters and parameter combinations involved, which clearly cannot all be validated with separate clinical trials. A pragmatic alternative might be to devise trials in which doctors randomly obtain the proteomic information and associated decision support. It would then be straightforward to determine whether there is a significant benefit in patient outcomes.

Conclusions

Staking stock of the current practice in laboratory medicine shows that the majority of treatment decisions are made on the basis of blood tests and that protein measurements are even today the most prominent among them. Despite successfully being carried out by the millions every year, these assays are almost always directed against individual proteins and the pace of introduction of new protein tests has slowed to a trickle.

MS-based proteomics clearly has the potential for multiplexed and highly specific measurements, in which protein patterns rather than single biomarkers could be the relevant readout. Our review of the literature revealed that past efforts were held back by the great analytical challenges of the plasma proteome, something that is only now giving way to exciting technological developments. We argue that the analysis of large numbers of conditions and participants in all stages of the discovery and validation process has the potential to produce biomarker panels that are likely to be of clinical value. When coupled to large knowledge bases of changes in protein patterns in defined conditions, such a plasma proteome profiling strategy could in principle exploit the entire information contents of this body fluid.

To make this vision a reality, further improvements in throughput, depth of proteome coverage, robustness, and accessibility of the underlying workflow are crucial. Furthermore, plasma proteomics can also be extended to the analysis of post-translation modifications. Likewise, plasma metabolomics also uses MS-based workflows and could routinely be integrated with plasma proteomics in the future. We are confident that the required technological developments can and will all be achieved over time. At least as much of a challenge will be conceptual and "political", as the proteomic information deluge needs to be turned into actionable data for the physician and the healthcare system. This will require a dedicated and untiring commitment from all partners involved. We believe that the promise of much more precise and specific diagnostics will amply reward such efforts.

Expanded View for this article is available online.

Acknowledgements

We thank all members of the Proteomics and Signal Transduction and the Clinical Proteomics groups for help and discussions, in particular Peter V. Treit for assistance with the literature search and Sophia Doll, Lili Niu, and Atul Deshmukh for helpful comments. The work carried out in this project was partially supported by the Max Planck Society for the Advancement of Science and by the Novo Nordisk Foundation (grant NNF15CC0001).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abbatiello SE, Schilling B, Mani DR, Zimmerman LJ, Hall SC, MacLean B, Albertolle M, Allen S, Burgess M, Cusack MP, Gosh M, Hedrick V, Held JM, Inerowicz HD, Jackson A, Keshishian H, Kinsinger CR, Lyssand J, Makowski L, Mesri M *et al* (2015) Large-scale interlaboratory study to develop, analytically validate and apply highly multiplexed, quantitative peptide assays to measure cancer-relevant proteins in plasma. *Mol Cell Proteomics* 14: 2357 – 2374

Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, Spiegelman CH, Zimmerman LJ, Ham AJ, Keshishian H, Hall SC, Allen S, Blackman RK, Borchers CH, Buck C, Cardasis HL, Cusack MP, Dodder NG, Gibson BW, Held JM *et al* (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol* 27: 633 – 641

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

Altelaar AF, Heck AJ (2012) Trends in ultrasensitive proteomics. *Curr Opin Chem Biol* 16: 206 – 213

Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW (2004) Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* 3: 235 – 244

Anderson NL (2010) The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem* 56: 177 – 185

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

Baggerly KA, Morris JS, Coombes KR (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20: 777 – 785

Bantscheff M, Boesche M, Eberhard D, Matthieson T, Sweetman G, Kuster B (2008) Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol Cell Proteomics* 7: 1702 – 1713

Bekker-Jensen DB, Kelstrup CD, Batth TS, Larsen SC, Haldrup C, Bramsen JB, Sørensen KD, Høyer S, Ørntoft TF, Andersen CL, Nielsen ML, Olsen JV (2017) An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst* 4: 587 – 599.e4

Bellei E, Bergamini S, Monari E, Fantoni LI, Cuoghi A, Ozben T, Tomasi A (2011) High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. *Amino Acids* 40: 145 – 156

Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, Gomez-Varela D, Reiter L (2017) Heralds of parallel MS: data-independent acquisition surpassing sequential identification of data dependent acquisition in proteomics. *Mol Cell Proteomics* https://doi.org/10.1074/mcp.M116.065730

Burgess MW, Keshishian H, Mani DR, Gillette MA, Carr SA (2014) Simplified and efficient quantification of low-abundance proteins at very high multiplex via targeted mass spectrometry. *Mol Cell Proteomics* 13: 1137 – 1149

Cao Z, Tang HY, Wang H, Liu Q, Speicher DW (2012) Systematic comparison of fractionation methods for in-depth analysis of plasma proteomes. *J Proteome Res* 11: 3090 – 3100

Carr SA, Abbatiello SE, Ackermann BL, Borchers C, Domon B, Deutsch EW, Grant RP, Hoofnagle AN, Huttenhain R, Koomen JM, Liebler DC, Liu T, MacLean B, Mani DR, Mansfield E, Neubert H, Paulovich AG, Reiter L, Vitek

Published online: September 26, 2017

O, Aebersold R *et al* (2014) Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol Cell Proteomics* 13: 907 – 917

Cohen Freue GV, Meredith A, Smith D, Bergman A, Sasaki M, Lam KK, Hollander Z, Opushneva N, Takhar M, Lin D, Wilson-McManus J, Balshaw R, Keown PA, Borchers CH, McManus B, Ng RT, McMaster WR, Biomarkers in T, the NCECPoOFCoET (2013) Computational biomarker pipeline from discovery to clinical implementation: plasma proteomic biomarkers for cardiac transplantation. *PLoS Comput Biol* 9: e1002963

Cole RN, Ruczinski I, Schulze K, Christian P, Herbrich S, Wu L, Devine LR, O'Meally RN, Shrestha S, Boronina TN, Yager JD, Groopman J, West KP Jr (2013) The plasma proteome identifies expected and novel proteins correlated with micronutrient status in undernourished Nepalese children. *J Nutr* 143: 1540 – 1548

Cominetti O, Nunez Galindo A, Corthesy J, Oller Moreno S, Irincheeva I, Valsesia A, Astrup A, Saris WH, Hager J, Kussmann M, Dayon L (2016) Proteomic biomarker discovery in 1000 human plasma samples with mass spectrometry. *J Proteome Res* 15: 389 – 399

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

Demler OV, Pencina MJ, D'Agostino RB Sr (2013) Impact of correlation on predictive ability of biomarkers. *Stat Med* 32: 4196 – 4210

De-Ritis F, Coltorti M, Giusti G (1957) An enzymic test for the diagnosis of viral hepatitis - the transaminase serum activities. *Clin Chim Acta* 2: 70 – 74

Duffy MJ, Sturgeon CM, Soletormos G, Barak V, Molina R, Hayes DF, Diamandis EP, Bossuyt PM (2015) Validation of new cancer biomarkers: a position statement from the European group on tumor markers. *Clin Chem* 61: 809 – 820

Ebhardt HA, Root A, Sander C, Aebersold R (2015) Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* 15: 3193 – 3208

Edfors F, Bostrom T, Forsstrom B, Zeiler M, Johansson H, Lundberg E, Hober S, Lehtio J, Mann M, Uhlen M (2014) Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins. *Mol Cell Proteomics* 13: 1611 – 1624

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)

Garcia-Bailo B, Brenner DR, Nielsen D, Lee HJ, Domanski D, Kuzyk M, Borchers CH, Badawi A, Karmali MA, El-Sohemy A (2012) Dietary patterns and ethnicity are associated with distinct plasma proteomic groups. *Am J Clin Nutr* 95: 352 – 361

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

Grant RP, Hoofnagle AN (2014) From lost in translation to paradise found: enabling protein biomarker method transfer by mass spectrometry. *Clin Chem* 60: 941 – 944

Hayes DF, Allen J, Compton C, Gustavsen G, Leonard DG, McCormack R, Newcomer L, Pothier K, Ransohoff D, Schilsky RL, Sigal E, Taube SE, Tunis SR (2013) Breaking a vicious cycle. *Sci Transl Med* 5: 196cm196

Hofmann B, Welch HG (2017) New diagnostic tests: more harm than good. *BMJ* 358: j3314

Holdt LM, Teupser D (2013) From genotype to phenotype in human atherosclerosis–recent findings. *Curr Opin Lipidol* 24: 410 – 418

Hoofnagle AN, Wener MH (2009) The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. *J Immunol Methods* 347: 3 – 11

Hoofnagle AN, Whiteaker JR, Carr SA, Kuhn E, Liu T, Massoni SA, Thomas SN, Townsend RR, Zimmerman LJ, Boja E, Chen J, Crimmins DL, Davies SR, Gao Y, Hiltke TR, Ketchum KA, Kinsinger CR, Mesri M, Meyer MR, Qian WJ *et al* (2016) Recommendations for the generation, quantification, storage, and handling of peptides used for mass spectrometry-based assays. *Clin Chem* 62: 48 – 69

Jennings L, Van Deerlin VM, Gulley ML, College of American Pathologists Molecular Pathology Resource C (2009) Recommended principles and practices for validating clinical molecular pathology tests. *Arch Pathol Lab Med* 133: 743 – 755

Juhasz P, Lynch M, Sethuraman M, Campbell J, Hines W, Paniagua M, Song L, Kulkarni M, Adourian A, Guo Y, Li X, Martin S, Gordon N (2011) Semi-targeted plasma proteomics discovery workflow utilizing two-stage protein depletion and off-line LC-MALDI MS/MS. *J Proteome Res* 10: 34 – 45

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

Kim YJ, Sertamo K, Pierrard MA, Mesmin C, Kim SY, Schlesser M, Berchem G, Domon B (2015) Verification of the biomarker candidates for non-small-cell lung cancer using a targeted proteomics approach. *J Proteome Res* 14: 1412 – 1419

Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Ayme S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV *et al* (2017) The human phenotype ontology in 2017. *Nucleic Acids Res* 45: D865 – D876

Kolla V, Jeno P, Moes S, Tercanli S, Lapaire O, Choolani M, Hahn S (2010) Quantitative proteomics analysis of maternal plasma in Down syndrome pregnancies using isobaric tagging reagent (iTRAQ). *J Biomed Biotechnol* 2010: 952047

Lee SE, West KP Jr, Cole RN, Schulze KJ, Christian P, Wu LS, Yager JD, Groopman J, Ruczinski I (2015) Plasma proteome biomarkers of inflammation in school aged children in Nepal. *PLoS ONE* 10: e0144279

Lee SE, Stewart CP, Schulze KJ, Cole RN, Wu LS, Yager JD, Groopman JD, Khatry SK, Adhikari RK, Christian P, West KP Jr (2017) The plasma proteome is associated with anthropometric status of undernourished nepalese school-aged children. *J Nutr* 147: 304 – 313

Levine RJ, Maynard SE, Qian C, Lim KH, England LJ, Yu KF, Schisterman EF, Thadhani R, Sachs BP, Epstein FH, Sibai BM, Sukhatme VP, Karumanchi SA (2004) Circulating angiogenic factors and the risk of preeclampsia. *N Engl J Med* 350: 672 – 683

Liu T, Qian WJ, Gritsenko MA, Xiao W, Moldawer LL, Kaushal A, Monroe ME, Varnum SM, Moore RJ, Purvine SO, Maier RV, Davis RW, Tompkins RG, Camp DG II, Smith RD, Inflammation, the Host Response to Injury Large Scale Collaborative Research P (2006) High dynamic range characterization of the trauma patient plasma proteome. *Mol Cell Proteomics* 5: 1899 – 1913

Liu Y, Buil A, Collins BC, Gillet LC, Blum LC, Cheng LY, Vitek O, Mouritsen J, Lachance G, Spector TD, Dermitzakis ET, Aebersold R (2015) Quantitative

Published online: September 26, 2017
Molecular Systems Biology

variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* 11: 786

Luque-Garcia JL, Neubert TA (2007) Sample preparation for serum/plasma profiling and biomarker identification by mass spectrometry. *J Chromatogr A* 1153: 259 – 276

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26: 966 – 968

Malmström E, Kilsgard O, Hauri S, Smeds E, Herwald H, Malmstrom L, Malmstrom J (2016) Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nat Commun* 7: 10261

Mann M, Kulak NA, Nagaraj N, Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* 49: 583 – 590

Manolio TA (2013) Bringing genome-wide association findings into clinical use. *Nat Rev Genet* 14: 549 – 558

Mazzara S, Rossi RL, Grifantini R, Donizetti S, Abrignani S, Bombaci M (2017) CombiROC: an interactive web tool for selecting accurate marker combinations of omics data. *Sci Rep* 7: 45477

Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, Bennett SE, Bischoff R, Bongcam-Rudloff E, Capasso G, Coon JJ, D'Haese P, Dominiczak AF, Dakna M, Dihazi H, Ehrich JH, Fernandez-Llama P, Fliser D, Frokiaer J, Garin J *et al* (2010) Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* 2: 46ps42

Nanjappa V, Thomas JK, Marimuthu A, Muthusamy B, Radhakrishnan A, Sharma R, Ahmad Khan A, Balakrishnan L, Sahasrabuddhe NA, Kumar S, Jhaveri BN, Sheth KV, Kumar Khatana R, Shaw PG, Srikanth SM, Mathur PP, Shankar S, Nagaraja D, Christopher R, Mathivanan S *et al* (2014) Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Res* 42: D959 – D965

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

Nie S, Shi T, Fillmore TL, Schepmoes AA, Brewer H, Gao Y, Song E, Wang H, Rodland KD, Qian WJ, Smith RD, Liu T (2017) Deep-dive targeted quantification for ultrasensitive analysis of proteins in non-depleted human blood plasma/serum and tissues. *Anal Chem* 89: 9139 – 9146

Oberbach A, Schlichting N, Neuhaus J, Kullnick Y, Lehmann S, Heinrich M, Dietrich A, Mohr FW, von Bergen M, Baumann S (2014) Establishing a reliable multiple reaction monitoring-based method for the quantification of obesity-associated comorbidities in serum and adipose tissue requires intensive clinical validation. *J Proteome Res* 13: 5784 – 5800

Obuchowski NA, Lieber ML, Wians FH Jr (2004) ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 50: 1118 – 1125

Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H *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

Ozcan S, Cooper JD, Lago SG, Kenny D, Rustogi N, Stocki P, Bahn S (2017) Towards reproducible MRM based biomarker discovery using dried blood spots. *Sci Rep* 7: 45178

Pan S, Chen R, Crispin DA, May D, Stevens T, McIntosh MW, Bronner MP, Ziogas A, Anton-Culver H, Brentnall TA (2011) Protein alterations associated with pancreatic cancer and chronic pancreatitis found in human plasma using global quantitative proteomics profiling. *J Proteome Res* 10: 2359 – 2376

Parker CE, Borchers CH (2014) Mass spectrometry based biomarker discovery, verification, and validation–quality assurance and control of protein biomarker assays. *Mol Oncol* 8: 840 – 858

Paulovich AG, Whiteaker JR, Hoofnagle AN, Wang P (2008) The interface between biomarker discovery and clinical validation: the tar pit of the protein biomarker pipeline. *Proteomics Clin Appl* 2: 1386 – 1402

Pavlou MP, Diamandis EP, Blasutig IM (2013) The long journey of cancer biomarkers from the bench to the clinic. *Clin Chem* 59: 147 – 157

Pencina MJ, D'Agostino RB, Vasan RS (2010) Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med* 48: 1703 – 1711

Percy AJ, Chambers AG, Yang J, Borchers CH (2013) Multiplexed MRM-based quantitation of candidate cancer biomarker proteins in undepleted and non-enriched human plasma. *Proteomics* 13: 2202 – 2215

Percy AJ, Michaud SA, Jardim A, Sinclair NJ, Zhang S, Mohammed Y, Palmer AL, Hardie DB, Yang J, LeBlanc AM, Borchers CH (2017) Multiplexed MRM-based assays for the quantitation of proteins in mouse plasma and heart tissue. *Proteomics* https://doi.org/10.1002/pmic.201600097

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359: 572 – 577

Picotti P, Aebersold R (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* 9: 555 – 566

Qian WJ, Kaleta DT, Petritis BO, Jiang H, Liu T, Zhang X, Mottaz HM, Varnum SM, Camp DG II, Huang L, Fang X, Zhang WW, Smith RD (2008) Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. *Mol Cell Proteomics* 7: 1963 – 1973

Rai AJ, Zhang Z, Rosenzweig J, Shih I, Pham T, Fung ET, Sokoll LJ, Chan DW (2002) Proteomic approaches to tumor marker discovery - Identification of biomarkers for ovarian cancer. *Arch Pathol Lab Med* 126: 1518 – 1526

Razavi M, Leigh Anderson N, Pope ME, Yip R, Pearson TW (2016) High precision quantification of human plasma proteins using the automated SISCAPA Immuno-MS workflow. *N Biotechnol* 33: 494 – 502

Richards AL, Merrill AE, Coon JJ (2015) Proteome sequencing goes deep. *Curr Opin Chem Biol* 24: 11 – 17

Rifai N, Gillette MA, Carr SA (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 24: 971 – 983

Roffi M, Patrono C, Collet JP, Mueller C, Valgimigli M, Andreotti F, Bax JJ, Borger MA, Brotons C, Chew DP, Gencer B, Hasenfuss G, Kjeldsen K, Lancellotti P, Landmesser U, Mehilli J, Mukherjee D, Storey RF, Windecker S (2016) 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). *G Ital Cardiol* 17: 831 – 872

Rosenberger G, Bludau I, Schmitt U, Heusel M, Hunter CL, Liu Y, MacCoss MJ, MacLean BX, Nesvizhskii AI, Pedrioli PGA, Reiter L, Röst HL, Tate S, Ting YS, Collins BC, Aebersold R (2017) Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Methods* 14: 921 – 927

Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, Wolski W, Collins BC, Malmstrom J, Malmstrom L, Aebersold R (2014)

Published online: September 26, 2017

OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 32: 219 – 223

Sajic T, Liu Y, Aebersold R (2015) Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics Clin Appl* 9: 307 – 321

Sharma K, Schmitt S, Bergner CG, Tyanova S, Kannaiyan N, Manrique-Hoyos N, Kongi K, Cantuti L, Hanisch UK, Philips MA, Rossner MJ, Mann M, Simons M (2015) Cell type- and brain region-resolved mouse brain proteome. *Nat Neurosci* 18: 1819 – 1831

Shi T, Fillmore TL, Gao Y, Zhao R, He J, Schepmoes AA, Nicora CD, Wu C, Chambers JL, Moore RJ, Kagan J, Srivastava S, Liu AY, Rodland KD, Liu T, Camp DG II, Smith RD, Qian WJ (2013) Long-gradient separations coupled with selected reaction monitoring for highly sensitive, large scale targeted protein quantification in a single analysis. *Anal Chem* 85: 9196 – 9203

Skates SJ, Gillette MA, LaBaer J, Carr SA, Anderson L, Liebler DC, Ransohoff D, Rifai N, Kondratovich M, Tezak Z, Mansfield E, Oberg AL, Wright I, Barnes G, Gail M, Mesri M, Kinsinger CR, Rodriguez H, Boja ES (2013) Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. *J Proteome Res* 12: 5383 – 5394

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 38: 209 – 213

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

Thulasiraman V, Lin S, Gheorghiu L, Lathrop J, Lomas L, Hammond D, Boschetti E (2005) Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. *Electrophoresis* 26: 3561 – 3571

Tu C, Rudnick PA, Martinez MY, Cheek KL, Stein SE, Slebos RJ, Liebler DC (2010) Depletion of abundant plasma proteins and limitations of plasma proteomics. *J Proteome Res* 9: 4982 – 4991

Vogeser M, Seger C (2016) Mass spectrometry methods in clinical diagnostics - state of the art and perspectives. *Trac-Trends Analyt Chem* 84: 1 – 4

Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, Yan P, Schoenherr RM, Zhao L, Voytovich UJ, Kelly-Spratt KS, Krasnoselsky A, Gafken PR, Hogan JM, Jones LA, Wang P, Amon L, Chodosh LA, Nelson PS, McIntosh MW *et al* (2011) A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat Biotechnol* 29: 625 – 634

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

Wu HY, Goan YG, Chang YH, Yang YF, Chang HJ, Cheng PN, Wu CC, Zgoda VG, Chen YJ, Liao PC (2015) Qualification and verification of serological biomarker candidates for lung adenocarcinoma by targeted mass spectrometry. *J Proteome Res* 14: 3039 – 3050

Zander J, Bruegel M, Kleinhempel A, Becker S, Petros S, Kortz L, Dorow J, Kratzsch J, Baber R, Ceglarek U, Thiery J, Teupser D (2014) Effect of biobanking conditions on short-term stability of biomarkers in human serum and plasma. *Clin Chem Lab Med* 52: 629 – 639

Zeiler M, Straube WL, Lundberg E, Uhlen M, Mann M (2012) A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol Cell Proteomics* 11: O111.009613

Zhang Z, Bast RC Jr, Yu Y, Li J, Sokoll LJ, Rai AJ, Rosenzweig JM, Cameron B, Wang YY, Meng XY, Berchuck A, Van Haaften-Day C, Hacker NF, de Bruijn HW, van der Zee AG, Jacobs IJ, Fung ET, Chan DW (2004) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res* 64: 5882 – 5890

Zhang X, Xiao Z, Liu X, Du L, Wang L, Wang S, Zheng N, Zheng G, Li W, Zhang X, Dong Z, Zhuang X, Wang C (2012) The potential role of ORM2 in the development of colorectal cancer. *PLoS ONE* 7: e31868

Zhang X, Ning Z, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M, Mack D, Stintzi A, Figeys D (2016) MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* 4: 31

Zhou C, Simpson KL, Lancashire LJ, Walker MJ, Dawson MJ, Unwin RD, Rembielak A, Price P, West C, Dive C, Whetton AD (2012) Statistical considerations of optimal study design for human plasma proteomics and biomarker discovery. *J Proteome Res* 11: 2103 – 2113

Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 39: 561 – 577

3.5. Article 5: Ultra-deep and Quantitative Saliva Proteome Reveals Dynamics of the Oral Microbiome

Authors: Niklas Grassl[1], Nils A. Kulak[1,2], Garwin Pichler[1,2], Philipp E. Geyer[1,3], Jette Jung[4], Sören Schubert[4], Pavel Sinitcyn[5], Juergen Cox[5] and Matthias Mann[1,3]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]PreOmics GmbH, Am Klopferspitz 18, Martinsried, Germany
[3]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[4]Max von Pettenkofer-Institut für Hygiene und Medizinische Mikrobiologie, Marchioninistr. 17, D-81377, München, Germany
[5]Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152, Martinsried, Germany

The concepts and much of the workflow of Plasma Proteomics Profiling can be applied to many other body fluids. In our laboratory we have already done this for cerebrospinal fluid, urine, tears and saliva. In this publication, we investigate saliva, an easily accessible body fluid with potential for clinical diagnostics. The oral cavity contains a rich community of microorganisms, which is of great current interest as the microbiome has a pivotal role for health and disease states.

We describe the application of our workflow, which we optimized and streamlined for saliva, starting with a simple cotton swab for sample collection. In single run analysis, this yielded a remarkable depth of 3,700 human proteins. Moreover, we used high pH reversed phase fractionation for in depth characterization. This resulted in more than 5,500 identified human proteins, which is the largest body fluid proteome so far. We further searched this data against a database of microbial organisms and found more than 2,000 bacterial proteins, originating from more than 50 genera, with a similar distribution between different individuals.

Next, we applied the streamlined workflow for a first ‘clinical study’ in which eight study participants collected saliva in the morning before and after teeth brushing. This revealed drastic quantitative changes of the oral microbiome in the different individuals.

Grassl et al. Genome Medicine (2016) 8:44
DOI 10.1186/s13073-016-0293-0

Genome Medicine

RESEARCH Open Access

Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome

Niklas Grassl[1], Nils Alexander Kulak[1,2], Garwin Pichler[1,2], Philipp Emanuel Geyer[1,3], Jette Jung[4], Sören Schubert[4], Pavel Sinitcyn[5], Juergen Cox[5] and Matthias Mann[1,3*]

Abstract

Background: The oral cavity is home to one of the most diverse microbial communities of the human body and a major entry portal for pathogens. Its homeostasis is maintained by saliva, which fulfills key functions including lubrication of food, pre-digestion, and bacterial defense. Consequently, disruptions in saliva secretion and changes in the oral microbiome contribute to conditions such as tooth decay and respiratory tract infections. Here we set out to quantitatively map the saliva proteome in great depth with a rapid and in-depth mass spectrometry-based proteomics workflow.

Methods: We used recent improvements in mass spectrometry (MS)-based proteomics to develop a rapid workflow for mapping the saliva proteome quantitatively and at great depth. Standard clinical cotton swabs were used to collect saliva form eight healthy individuals at two different time points, allowing us to study inter-individual differences and interday changes of the saliva proteome. To accurately identify microbial proteins, we developed a method called "split by taxonomy id" that prevents peptides shared by humans and bacteria or between different bacterial phyla to contribute to protein identification.

Results: Microgram protein amounts retrieved from cotton swabs resulted in more than 3700 quantified human proteins in 100-min gradients or 5500 proteins after simple fractionation. Remarkably, our measurements also quantified more than 2000 microbial proteins from 50 bacterial genera. Co-analysis of the proteomics results with next-generation sequencing data from the Human Microbiome Project as well as a comparison to MALDI-TOF mass spectrometry on microbial cultures revealed strong agreement. The oral microbiome differs between individuals and changes drastically upon eating and tooth brushing.

Conclusion: Rapid shotgun and robust technology can now simultaneously characterize the human and microbiome contributions to the proteome of a body fluid and is therefore a valuable complement to genomic studies. This opens new frontiers for the study of host–pathogen interactions and clinical saliva diagnostics.

* Correspondence: mmann@biochem.mpg.de
[1]Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany
[3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark
Full list of author information is available at the end of the article

Background

Using saliva for the diagnosis of medical conditions would be particularly attractive because it can be collected non-invasively and economically [1], but the complexity of the oral cavity and the multiple entities contributing to its homeostasis make this challenging. In addition to the secretions of oral grands, saliva contains cells shed from the epithelium of the oral cavity and harbors the oral microbiome. Promising steps towards the establishment of saliva protein biomarkers have already been undertaken [2, 3]. However, these studies either only considered around 100 proteins with antibody-based assays or employed relatively low throughput mass spectrometry (MS)-based proteomics with extensive fractionation, which generally precluded quantification [4].

Further interest in saliva has recently been fueled by the discovery that the oral microbiome and the gut microbiome are the most diverse ones of the human body and that they correlate well with each other [5]. There is now compelling evidence for a link between the human microbiome and conditions such as obesity, allergies, and even autoimmune diseases like multiple sclerosis [6–8]. In addition, tooth decay and other diseases of the oral cavity are known to be caused by bacteria but turn out to be insufficiently explained by one species alone [9, 10]. Therefore, first metagenomics and then metaproteomics studies have already aimed to relate bacterial composition to caries incidence [10, 11]. However, reproducible identification and consistent quantification of bacteria remain challenging. Dynamic, quantitative studies would be of great help to uncover the functional connections between microbial communities and the prevalent pathologies of the oral cavity.

During the past few years, our laboratory has focused on simplifying and streamlining the proteomics workflow, with the aim of bringing the technology closer to clinical applications. Here we set out to characterize the saliva proteome at the greatest depth possible while still minimizing steps that could compromise quantification. We also developed a rapid single-run analysis workflow, starting from standard clinical cotton swabs and delivering results in a few hours, while retaining a quantification depth of thousands of proteins. This allowed us to investigate changes in the saliva proteome upon perturbation in a healthy cohort. We also analyzed inter-individual differences in the saliva proteome and quantitatively addressed the long-standing question of the degree to which the plasma and saliva proteomes are correlated. Finally, we asked if our in-depth workflow can characterize the oral microbiome and its dynamics and confirmed detected species by the established method of culturing followed by Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) as well as data from next-generation sequencing projects.

Methods

Experimental design

We collected saliva at two different time points from four female and four male, healthy, non-smoking individuals aged 24 to 40 years with Caucasian backgrounds. All subjects were asymptomatic, did not take any drugs or antiseptics, visited the dentist regularly, and showed no signs of inflammation, bleeding, or infection as judged by a medical student (N.G.). The study was approved by the ethics committee of the Max Planck Society and all donors provided their written informed consent to participate in this study and to publish the acquired results. The first collection was immediately after waking, before eating, drinking, or tooth brushing. The second collection took place at 10 a.m., at least 30 min after the donors had eaten breakfast and brushed their teeth. In addition, we collected three samples immediately after one another from the same donor, processed them in parallel, and determined the reproducibility of our workflow. Because this showed very high reproducibility (mean $R^2 = 0.92$, Additional file 1: Figure S3b), we did not perform technical replicates in this study but decided to use our measurement time for the analysis of several donors and proteome states.

Protein digestion and peptide purification

Following collection, the swabs were transferred to an Eppendorf tube containing 200 µl of lysis buffer (1 % sodium dodecyl carbonate (v/v), 10 mM tris (2-carboxyethyl) phosphine, 40 mM 2-chloroacetamide, 100 mM Tris buffer pH 8.5), thoroughly squeezed against the inner wall of the Eppendorf tube, and removed. We reproducibly recovered more than 100 µg of protein in this way as estimated by the Bradford protein assay. Sample preparation followed essentially the in-StageTip protocol [12]. Briefly, a total of 20 µg of protein was digested by adding 0.4 µg trypsin and LysC to our lysis buffer and incubating for 60 min at 37 °C while shaking. Following this short digestion, we acidified the peptides to a final concentration of 1 % trifluoroacetic acid (TFA) and loaded them on an SDB-RPS StageTip [13]. The filter was then washed and peptides were finally eluted with 60 µl 80 % acetonitrile (ACN) (v/v) and 1 % ammonium (v/v), dried in a SpeedVac concentrator, and resuspended in A* buffer (2 % ACN (v/v), 0.1 % TFA (v/v), pH 2) to a concentration of 1 g/l.

Single run and prefractionated liquid chromatography-MS measurement

To obtain a deep saliva proteome, we used basic reversed phase chromatography to fractionate our eight waking samples prior to liquid chromatography (LC)-MS measurement. Approximately 15 µg of peptides were separated in an 80-min gradient on a 20-cm, 75-µm inner diameter column that was in-house packed with

ReproSil-Pur C_{18} beads (Dr. Maisch GmbH, Germany). Concatenated fractions [14, 15] were dried in the SpeedVac concentrator and resuspended in A* buffer to a concentration of 1 g/l. Both the fractionated and the single run samples were subjected to a 100-min chromatography gradient using an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific) and an in-house-made 40-cm column of the type described above. The chromatography was on-line coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific) by applying a spray voltage of 2.2 kV. The MS scan resolution was set to 120,000 at m/z 200, the scan range to 300 to 1650 m/z, and the maximum injection time to 55 ms. The 15 most intense ions per MS scan were selected for higher-energy collisional dissociation (HCD) fragmentation with an isolation width of 1.5 m/z and were measured at a resolution of 30,000. Dynamic exclusion was used with an exclusion time of 30 s.

Raw data processing of human proteins

The raw files were analyzed in MaxQuant [16] (version 1.5.3.15). We analyzed the single runs and the fractionated samples together in order to exploit the match between runs algorithm, which enables the identification of peptides that were not selected for fragmentation in one run by checking whether these peptides were sequenced in another run (the maximum time deviation was 30 s of the recalibrated retention times) [17]. We used the Andromeda search engine [18] to search the detected features against the human reference proteome from Uniprot (downloaded on 24 June 2015; 90.5 K sequences, 3.2 million unique peptides of which 0.64 million were seven amino acids or more in length) and a list of 247 potential contaminants [16]. Only tryptic peptides that were at least seven amino acids in length with up to two missed cleavages were considered. The initial allowed mass tolerance was set to 4.5 ppm at the MS level and 0.5 Da at the MS/MS level. We set N-acetylation of proteins' N-termini (42.010565 Da) and oxidation of methionine (15.994915 Da) as variable modifications and carbamidomethylation of cysteine as a fixed modification (57.021464 Da). A false discovery rate (FDR) of 1 % was imposed for peptide-spectrum matches (PSMs) and protein identification using a target–decoy approach. Relative quantification was performed using the default parameters of the MaxLFQ algorithm [19] with the minimum ratio count set to 1.

Data analysis of human proteins

The "proteinGroups.txt" file produced by MaxQuant was further analyzed in Perseus (version 1.5.2.12). Proteins from the reverse database, proteins only identified by site, and contaminants were removed. We decided to consider all keratin type I and II proteins contaminants because we could not exclude the possibility that their presence in our samples was due to skin desquamation. Proteins were ranked according to the mean label-free quantification (LFQ) intensities of the fractionated waking and the postprandial samples of all donors. We performed one-dimensional (1D) annotation enrichment of the resulting logarithmized LFQ distribution for Gene Ontology (GO) terms and Uniprot keywords with a Benjamini–Hochberg FDR cutoff of 2 % as described [20]. For the comparison of plasma and saliva proteomes, we used triplicate plasma proteomes of two of our saliva donors measured with 45-min HPLC gradients [21]. These six raw files were processed together with the single run saliva files from the two donors using the MaxQuant settings from above. Principal component analysis (PCA) was done on the logarithmized LFQ intensities of all 16 single shot runs. The differences between the waking and postprandial proteomes were analyzed by filtering the list of quantified proteins for 100 % valid values in all 16 single run analyses and performing a two sided *t*-test on the logarithmized LFQ intensities with a Benjamini–Hochberg FDR cutoff of 5 % and the s0 parameter set to 0.1. We determined whether the significantly upregulated proteins at waking were enriched for certain Uniprot keywords compared with the entire proteome using a Fisher exact test with 2 % permutation-based FDR. The analogous analysis was performed for the significantly upregulated postprandial proteins.

Raw data processing of human and bacterial proteins

For the analysis of human and bacterial proteins, we downloaded the fasta files of all named species of the human oral microbiome database [22] with more than five protein sequences (downloaded 24 June 2015; 1118.9 K bacterial protein sequences in total). Together with the human sequences the resulting database contained 1209.4 K protein sequences which correspond to 58.6 million unique peptides after in silico digestion and 5.9 million peptides seven amino acids or more in length, which we considered in our MaxQuant settings. Search parameters were essentially identical to the raw file processing of human proteins alone, except that we applied the split by taxonomy feature on the phylum level and only used unique peptides for quantification. Due to the split by taxonomy on the phylum level, peptides that are part of human and bacterial proteins or peptides that occur in proteins from two different phyla are neglected for protein identification. This, as well as using only unique peptides rather than razor peptides for quantification, guarantees that peptides shared by different phyla are not attributed to the wrong organism.

Data analysis of the oral microbiome

For creating the taxonomic tree in Fig. 4, we determined the number of peptides that uniquely belonged to one

species of our database and wrote this number above the respective edge of the genus. Peptides shared by certain genera were added to the number of the lowest taxonomy edge shared by these genera (Operating Taxonomy Unit). For Fig. 4 we excluded all genera that did not have at least one unique peptide. We extended the analysis for streptococci down to the species level. Bacterial genus abundance was estimated by adding the ten peptides of highest intensity per genus in analogy to the protein quantification in [23, 24]. Genera with less than ten peptides were excluded from quantification.

Co-analysis with whole genome sequencing data from the human microbiome project

To compare our data with results obtained from whole genome sequencing (WGS), protein multifasta (PEP) was downloaded from the Human Microbiome Project (HMP) [25]. Fractionated and single run raw files were analyzed with the MaxQuant settings described above against the human reference proteome from Uniprot and the fasta file from HMP (3.8 million protein sequences, 127.3 million unique peptides). From the genomic side we downloaded 764 fastq files from the HMP (release of 2012) and trimmed them using Trimmomatic [26] (we removed adapter as well as leading and trailing sequences with quality lower than 10 Phred quality score; we also did not accept reads for further analysis with lengths less than 36 nucleotides) and aligned using BWA with default parameters [27]. A PCA of the reads per genus of the WGS dataset together with the top ten peptide intensities per genus across the median of all samples from MaxQuant was performed after Z-score scaling within each sample (Fig. 5d). We combined the body sites "saliva", "tongue dorsum", "attached keratinized gingiva", "palatine tonsils" and "throat" from the HMP for our definition of mouth because these sites clustered tightly in a PCA. Furthermore, we performed hierarchical clustering (Euclidean distance coupled with Ward's agglomeration method was used) on the resulting dataset and visualized the genus abundance per sample in a heatmap (using the R package *heatmap.2*) (Additional file 1: Figure S1).

Microbiological processing of the samples

Together with the cotton swab collection after waking, all donors also collected whole saliva by passive drooling into a sterile tube. Samples were processed immediately after collection as follows. One Columbia and one chocolate blood agar plate for the aerobic and two Schaedler agar plates for the anaerobic culture were plated out with 50 µl saliva each. Aerobic cultures were incubated for 3 days at 37 °C and 5.8 % CO_2. Anaerobic cultures were grown under anaerobic conditions at 37 °C for a minimum of 5 days. Plates were evaluated visually and all morphologically different colonies were subcultured for identification by MALDI-TOF MS.

Identification by MALDI-TOF MS

Samples were measured in duplicates according to the standard protocol recommended by the manufacturer. In brief, a thin layer of bacteria taken from a single colony was smeared onto a polished steel target and overlaid with 1 µl of matrix solution containing 10 mg/ml of α-cyano-4-hydroxy-cinnamic acid in 50 % acetonitrile/2.5 % TFA (α-HCCA portioned matrix, Bruker Daltonik GmbH, Bremen, Germany). For measurements, a Microflex LT benchtop instrument operated by flexControl 3.3 software (Bruker Daltonik GmbH, Germany) was used. Spectra were acquired in the linear positive ion mode at a laser frequency of 60 Hz within a mass range of 2 to 20 kDa. The acceleration voltage was 20 kV, the IS2 voltage was maintained at 18.6 kV, and the extraction delay time was 200 ns. For data analysis, spectra were matched with the Bruker Taxonomy database version 4.0.0.1.

Results and discussion

In-depth quantification of the saliva proteome

We obtained saliva from four male and four female healthy individuals using sterile cotton swabs as is done in routine clinical practice (Fig. 1, "Methods"). Donors were required to abstain from eating and drinking for at least 30 min prior to the collection to avoid food-based contamination or dilution effects. They were instructed to wipe the vestibule of the oral cavity, followed by the teeth and the sublingual compartment. Around 200 µg of total protein was recovered from each swab, an ample amount for repeated measurement using our recently developed in-StageTip digestion procedure [12]. Following an immediate digestion for one hour and purification, the resulting peptides were separated into eight fractions with basic reversed-phase chromatography [14, 15]. Each fraction as well as unfractionated sample was measured with a 100-min LC gradient on a Q Exactive HF mass spectrometer [28, 29]. Data were analyzed using the MaxQuant environment [16, 19].

Across our eight donors we identified more than 54,000 sequence-unique peptides and more than 5500 proteins, both at a false discovery rate (FDR) of 1 %. A total of 78 % of these proteins were detected in each donor, 90 % in at least six of eight donors, and only 1.3 % were unique to single donors (Fig. 2a). Thus, our sample collection protocol is robust and allows comparison of thousands of saliva proteins across individuals. For an individual donor, we identified a remarkable 5213 human proteins in the eight fractions—to our knowledge the deepest body fluid proteome recorded from an individual to date (Additional file 1: Figure S2a). To investigate the reasons for this extensive coverage, we inspected

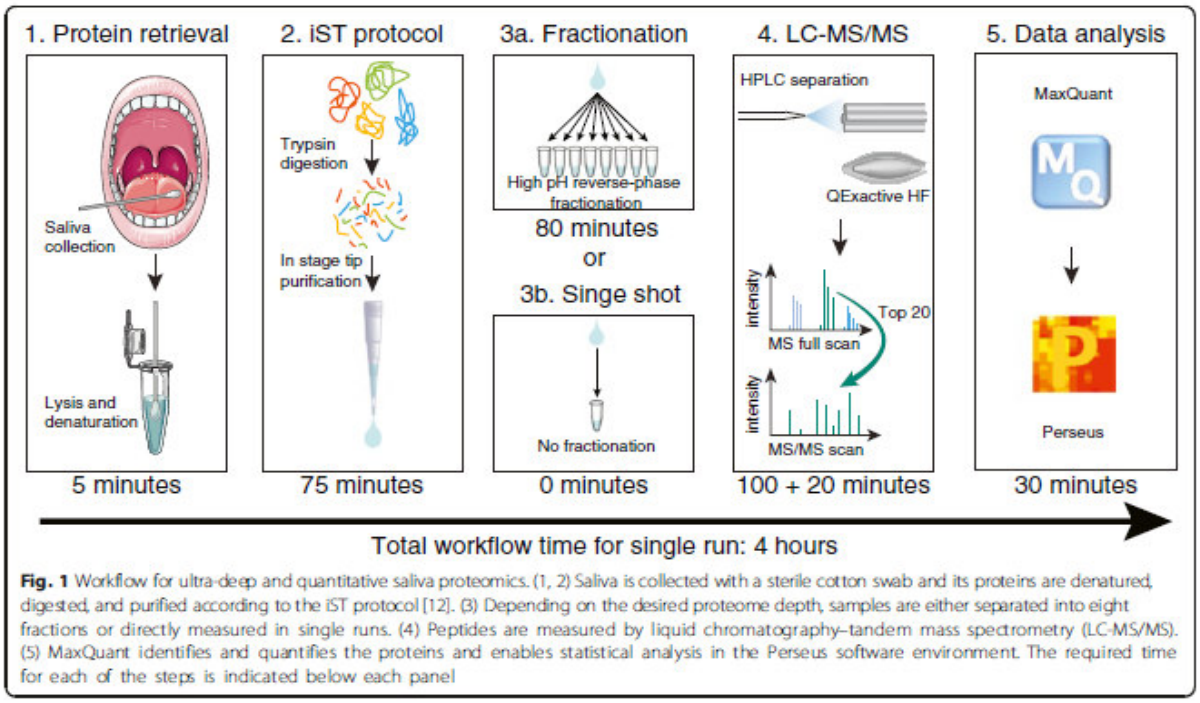


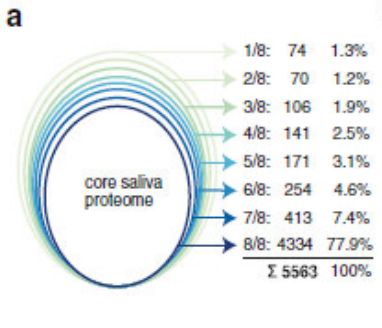
Fig. 1 Workflow for ultra-deep and quantitative saliva proteomics. (1, 2) Saliva is collected with a sterile cotton swab and its proteins are denatured, digested, and purified according to the iST protocol [12]. (3) Depending on the desired proteome depth, samples are either separated into eight fractions or directly measured in single runs. (4) Peptides are measured by liquid chromatography–tandem mass spectrometry (LC-MS/MS). (5) MaxQuant identifies and quantifies the proteins and enables statistical analysis in the Perseus software environment. The required time for each of the steps is indicated below each panel

the MS signal of the most abundant proteins. Unlike other body fluids, the 15 most abundant proteins in saliva make up only 32 % of the total proteome mass (Fig. 2b), whereas in plasma and urine they already account for more than 90 % and 58 % of the total, respectively [30, 31].

The abundance ranked plot of the entire measured saliva proteome spans a dynamic range of six orders of magnitude of estimated absolute abundance (Fig. 2c). To bioinformatically investigate the saliva proteome as a function of abundance, we used 1D annotation enrichment in the Perseus environment for GO terms and Uniprot keywords [20]. "Antibacterial humoral response" and "defense response to bacterium" scored in the upper part of the abundance distribution (Fig. 2c). "Extracellular space" and "Extracellular exosome" were significant near the median, indicating that proteins making up this category are somewhat less abundant than most of the functional saliva proteins. The terms in the lowest abundance range included typical intracellular terms such as "cytoplasm" and "mitochondrial translation".

There is an ongoing debate as to the extend that easily obtainable saliva could be used to measure plasma biomarkers by proxy [32]. We measured the plasma proteomes of two of our saliva donors in singe-run triplicate measurements [21] and compared them with the single-run saliva proteomes of the same donors. Due to the dynamic range challenges, fewer proteins were identified in plasma but more than 50 % of these were also identified in saliva. A scatter plot of the label-free quantification (LFQ) intensities of the proteins [19] that were identified in both body fluids reveals little correlation between these values ($R^2 = 0.11$; Fig. 2d). Over the two individuals and all replicates, it was never higher than $R^2 = 0.20$. We also considered the possibility that particular saliva components might show a higher correlation with the plasma proteome and collected one saliva sample from the opening of the duct of the parotid gland, one from the opening of the sublingual and submandibular gland, and one from gingiva. All these saliva proteomes revealed R^2 values below 0.1 (Additional file 1: Figure S3). Thus, we conclude that the plasma and saliva proteomes show little overall correlation and that saliva cannot directly be used as a substitute for the determination of plasma protein levels.

To make our saliva results available to the community in a user-friendly format, we uploaded them to the MaxQB database [33]. For each protein of interest, a query will reveal whether it is present in our saliva proteome, its abundance rank, estimated absolute abundance, and other protein level information (Additional file 1: Figure S2b). Additionally, peptide evidence leading to protein identification as well as high-resolution precursor–fragment relationships are available for constructing targeted assays. The protein illustrated in Additional file 1: Figure S2b is transcobalamin-1 (TCN1), which is known to be secreted by the salivary



b

rank	gene name	CV w %	CV p %	amount %	cum. amount %
1	AMY1A	51.5	38.6	4.8	4.8
2	ANXA1	10.9	21.1	4.0	8.8
3	S100A9	9.9	16.6	2.6	11.4
4	DSP	20.9	24.8	2.6	14.0
5	ALB	50.2	56.4	2.5	16.6
6	ACTG1	7.5	16.3	2.3	18.9
7	AHNAK	21.3	31.6	2.0	20.9
8	SPRR3	17.0	25.7	1.5	22.4
9	A2ML1	26.6	23.7	1.4	23.8
10	IGHA1	79.5	68.8	1.4	25.2
11	ZG16B	55.7	97.5	1.3	26.5
12	PPL	12.1	19.2	1.3	27.8
13	S100A8	28.4	18.7	1.3	29.1
14	CRNN	28.7	37.4	1.3	30.4
15	CSTA	16.9	36.6	1.1	31.5

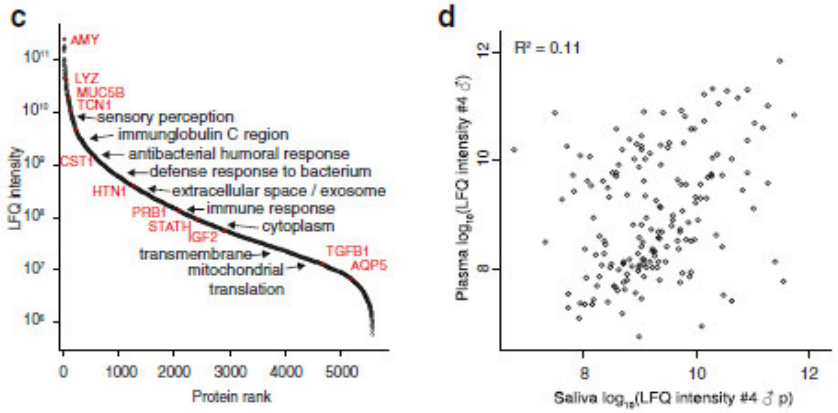


Fig. 2 Deep human saliva proteomes of eight healthy donors. **a** *Ovals* represent the number of saliva proteins shared by the respective number of donors. The *outer oval* contains all proteins that were detected in at least one donor, whereas the inner oval contains all proteins found in each sample—the core proteome. The *numbers* on the *right* indicate the numbers of proteins exactly found in one donor, in two donors, and so on. **b** Gene names of the 15 most abundant saliva proteins, their coefficients of variation (CVs) across eight donors at waking (*w*) and after breakfast and tooth brushing (*p*), as well as their abundances in percentage of the total proteome and the cumulative protein abundances (*cum. amount*). The proteins in *blue* are digestive proteins, the proteins in *green* are part of immune defense, and the proteins in *red* are of epithelial origin. **c** Dynamic range plot of the saliva proteome with some key proteins in saliva highlighted in *red*. Significantly enriched GO terms or Uniprot keywords in specific abundance regions as determined by 1D annotation are listed. **d** Scatter plot of the LFQ intensities of the saliva proteome and the plasma proteome

glands and to protect cobalamin or vitamin B12 against acidity of the stomach. In addition, TCN1 functions as a transport protein in the blood, carrying excess cobalamin to the liver for storage. Cobalamin deficiency occurs in 20 % of individuals over the age of 60 years [34] and causes anemia, demyelinating disease, or both [35]. Due to cobalamin's clinical significance, the physiological levels of TCN1 in blood have been characterized extensively in dedicated studies [36, 37], whereas here its levels are determined in the context of our system-wide investigation of thousands of other saliva proteins.

A deep single-run workflow

The high proteome coverage achieved using fractionation motivated us to determine how much of the saliva proteome could be retrieved in a single-run or "single-shot" experiment [17]. We used the same 100-min gradients as before and measured saliva proteomes from the eight individuals mentioned above, each at two different time points, once immediately after waking before tooth brushing and once post-prandial after tooth brushing. Remarkably, an average of 3835 proteins could be identified and almost all of them (94 %) were also quantifiable (Additional file 1: Figure S4a). The results from three swabs taken at nearly the same time and processed independently but equally were highly similar with a mean coefficient of determination R^2 of 0.92 (Additional file 1: Figure S4b). The difference between individuals was somewhat higher, with an R^2 of 0.89, indicating that biological differences between individuals can also be captured by single-run measurements. Plotting the CVs for saliva proteome variation between the individuals showed that they did not primarily depend on protein abundance (Additional file 1: Figure S4c). This suggests that single-run analysis should be able to determine biological differences across a wide abundance range. As the single-shot proteome still quantifies more than 3700 proteins, which include nearly all the functional

categories described above, very rapid and medium throughput characterization of saliva may be possible in the clinic.

Dynamics of the saliva proteome in a cohort

The oral cavity is subject to a variety of conditions in daily life. Despite several studies investigating, for instance, changing cortisol levels [38], to our knowledge intraday changes in the saliva proteome have not yet been investigated in depth.

To uncover dynamic changes, we first performed a principal component analysis (PCA) on all 16 single-run proteomes. Component 1 of the PCA separated weakly by sex (Additional file 1: Figure S5), whereas component 2 separated the two proteome states (waking versus post-prandial after tooth brushing) and this difference was even more pronounced when inspected on a person-by-person basis (Fig. 3a). To determine the proteins responsible for the PCA clustering, we filtered for 100 % valid LFQ values and plotted significance (5 % FDR) versus fold change (Fig. 3b). The proteins that were significantly upregulated at waking were enriched in the keywords "antibiotic" ($p = 7.7 \times 10^{-9}$, enrichment factor (ef) = 33) and "antimicrobial" ($p = 6.6 \times 10^{-8}$, ef = 24). The proteins with significantly higher abundance in the postprandial state were enriched for the terms "thiol protease inhibitor" and "secreted" ($p = 3.3 \times 10^{-5}$, ef = 42, and $p = 8.7 \times 10^{-9}$, ef = 6, respectively). Serving as a positive control, levels of alpha amylase (AMY1A), a protein that initiates the breakdown of complex oligosaccharides, were consistently upregulated after the meal. Thus, the shifts in protein abundance between our two measurement time points demonstrate that MS-based proteomics can now robustly capture biologically meaningful dynamic changes in body fluid proteomes.

Identification of bacterial proteomes in human saliva

Due to the prominent role of the oral microbiome in health and disease, we investigated whether we could detect bacterial species in the deep saliva proteomes. For this purpose, we downloaded the complete Uniprot protein sequences of all named oral bacterial species that had been identified by 16S rRNA sequencing in a recent study [22]. The resulting database was about 11 times larger than the human one alone.

In metaproteomics it is not straightforward to assign peptides to bacterial phyla because some amino acid sequences are part of proteins from different phyla. We addressed this issue by applying the "split by taxonomy" feature in MaxQuant, which avoids the formation of protein groups between different phyla. Together with the exclusive use of unique peptides for protein quantification, this functionality prevents the same peptide from contributing to the identification and quantification of

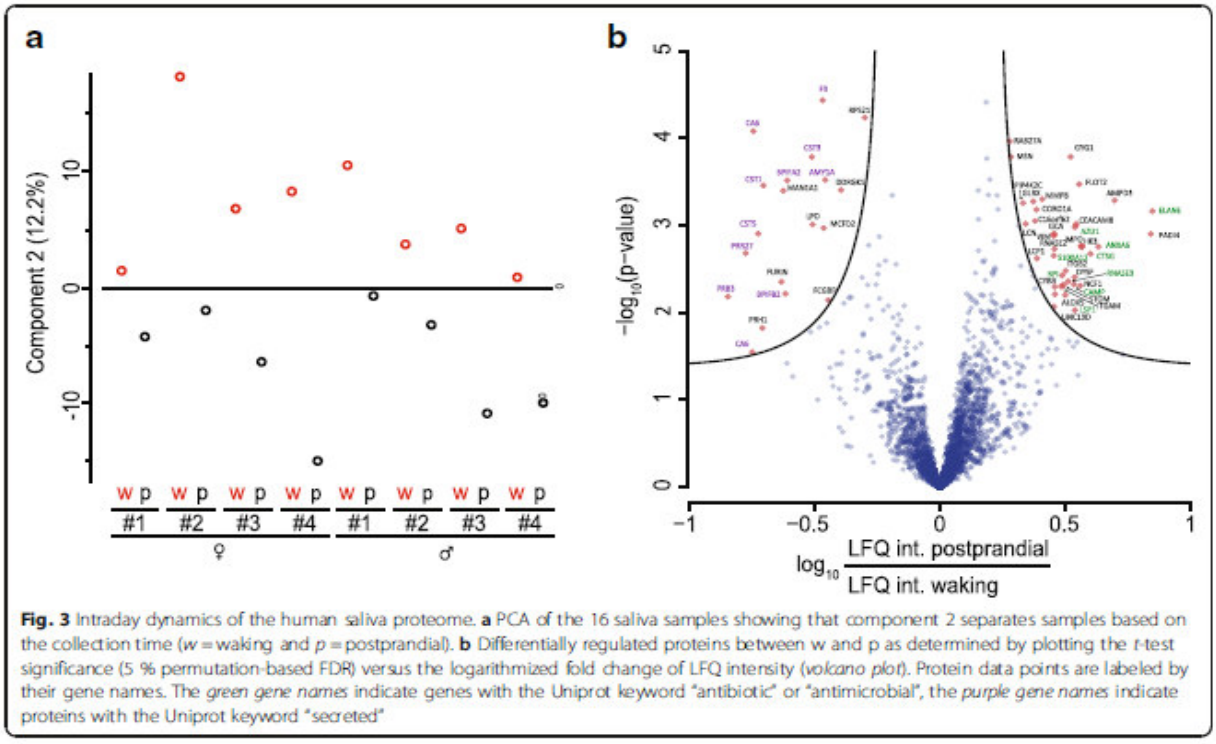


Fig. 3 Intraday dynamics of the human saliva proteome. **a** PCA of the 16 saliva samples showing that component 2 separates samples based on the collection time (*w* = waking and *p* = postprandial). **b** Differentially regulated proteins between w and p as determined by plotting the *t*-test significance (5 % permutation-based FDR) versus the logarithmized fold change of LFQ intensity (*volcano plot*). Protein data points are labeled by their gene names. The *green gene names* indicate genes with the Uniprot keyword "antibiotic" or "antimicrobial", the *purple gene names* indicate proteins with the Uniprot keyword "secreted"

proteins in different phyla ("Methods"). Split by taxonomy id is, therefore, relevant only for protein identification but not for peptide identification or quantification. However, bacteria in the oral cavity can have substantial sequence identity (Additional file 1: Figure S6a, b) [39]. As closely related bacteria share many sequences, one therefore needs to find the most appropriate taxonomy rank for applying the split by taxonomy id. To address this question, we placed identified bacterial peptides on a taxonomic tree such that the number of shared peptides is noted on each branch (Fig. 4). These shared peptides do not allow discrimination of the branches below. Split by taxonomy at a certain taxonomic rank prevents peptides shared at the ranks above from contributing to the identification of proteins. As in the case of human and microbial proteins above, this prevents the misassignment of peptides to phyla from which they do not necessarily originate. Placing the split at the phylum level turned out to be a good compromise between use of peptides for identification and quantification on the one hand and stringency of identification of bacteria on the other hand (Additional file 1: Figure S6) and we used this setting for all following analyses.

The presence of bacteria in the oral cavity also raises the question of whether proteins from them might considerably impair the human protein quantification presented above. To address this question we determined the nonredundant tryptic peptides that were seven or more amino acids long in our human and our oral bacteria database, which is the minimum length considered in our analysis. Among these tryptic peptides, the percentage of peptides with identical sequences between humans and bacteria was only 0.043 % (Fig. 5a). Hence, the quantification bias of human proteins due to bacteria is marginal. This analysis also indicates that bacterial contamination of mammalian proteome samples does not impair protein quantification considerably as long as only peptides of seven amino acids or more in length are considered.

Similarly, ingested proteins from food could, in principle, be erroneously assigned to human or bacterial proteins. To estimate the magnitude of these effects, we performed an analogous analysis on bovine and wheat as representative parts of a Western breakfast diet and determined the number of sequence identical peptides to humans and bacteria (Additional file 1: Figure S7). Except for bovine and human the percentage of overlapping peptide sequences is far below 1 %. Due to an overlap of 20.7 % among the considered human and bovine peptides, our in silico analysis does not exclude the possibility of quantification bias. However, proteins that substantially differ between waking and the postprandial

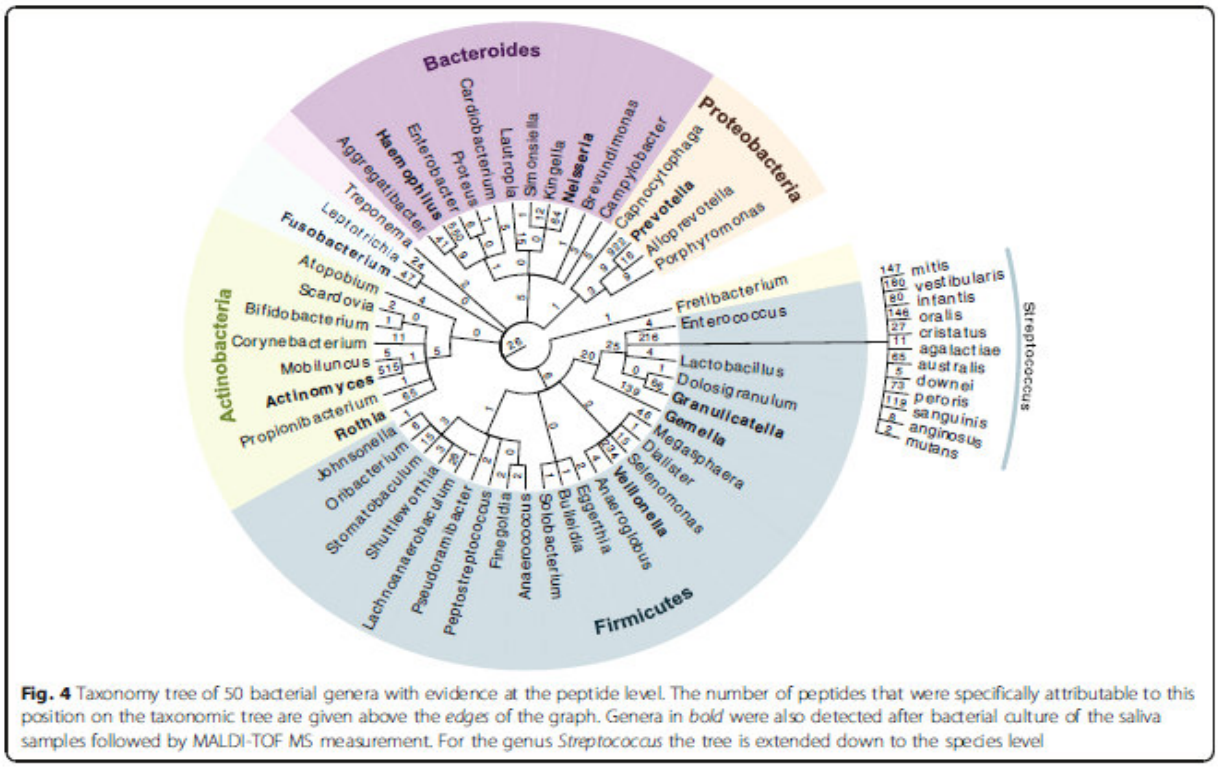


Fig. 4 Taxonomy tree of 50 bacterial genera with evidence at the peptide level. The number of peptides that were specifically attributable to this position on the taxonomic tree are given above the *edges* of the graph. Genera in *bold* were also detected after bacterial culture of the saliva samples followed by MALDI-TOF MS measurement. For the genus *Streptococcus* the tree is extended down to the species level

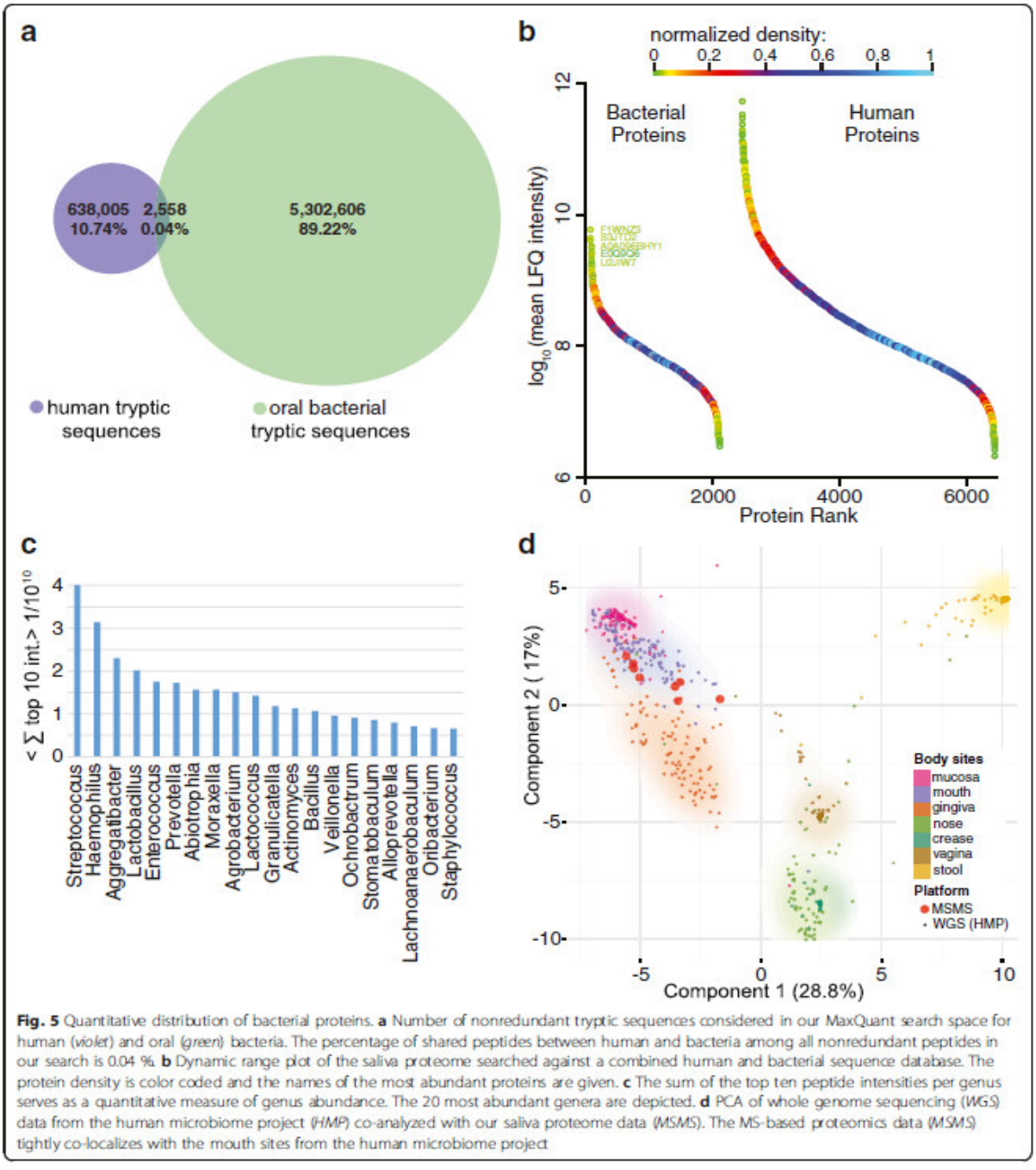


Fig. 5 Quantitative distribution of bacterial proteins. **a** Number of nonredundant tryptic sequences considered in our MaxQuant search space for human (*violet*) and oral (*green*) bacteria. The percentage of shared peptides between human and bacteria among all nonredundant peptides in our search is 0.04 %. **b** Dynamic range plot of the saliva proteome searched against a combined human and bacterial sequence database. The protein density is color coded and the names of the most abundant proteins are given. **c** The sum of the top ten peptide intensities per genus serves as a quantitative measure of genus abundance. The 20 most abundant genera are depicted. **d** PCA of whole genome sequencing (*WGS*) data from the human microbiome project (*HMP*) co-analyzed with our saliva proteome data (*MSMS*). The MS-based proteomics data (*MSMS*) tightly co-localizes with the mouth sites from the human microbiome project

state in Fig. 3 do not include proteins from human milk or human muscle, as would be expected if these differences were due to a bovine diet.

Remarkably, a search of our deep saliva proteome data sets using our standard, stringent search criteria (1 % FDR at the peptide and protein levels) resulted in the identification of 2234 different bacterial proteins. In total, we found evidence for 50 different bacterial genera from nine different phyla. This represents 50 % of the named genera identified by next-generation sequencing with

corresponding, annotated UniProt proteomes and therefore present in our database. The proteomic coverage of bacterial genera is remarkably high given the restricted database and the modest measuring time. The distribution of peptides specific for particular genera was highly unequal, ranging from only 1 to 1069 for the genus *Streptococcus*, for which Fig. 4 shows a detailed taxonomic tree down to the species level. At least 12 different such *Streptococcus* species were present in our deep saliva proteome. The most abundant species was *Streptococcus mitis*, but we also detected peptides unique to *Streptococcus mutans*, a main contributor to dental caries formation.

Standard MALDI-TOF MS as now routinely used in clinical microbiology found evidence of 14 different genera in our saliva samples, with an average of six genera per donor ("Methods"). In each case, shotgun proteomics had also identified the genus in the same sample without the need to cultivate the bacteria prior to processing. A rough comparison with the number of MS-identified peptides for genera identified by MALDI-TOF MS suggests that they were generally the more abundant ones (Fig. 4). While the goal in clinical microbiology is to identify the presence of one or a few pathogens responsible for an infection, rather than a total inventory of the microbiome, it is nevertheless notable that unbiased and relatively straightforward shotgun proteomics of saliva identified these bacteria without intervening cultivation directly from a cotton swab. This identification would presumably have been much easier still in the case of a dominating pathogen.

The quantitative oral metaproteome

To further investigate the unexpectedly large number of bacterial protein identifications, we plotted their cumulative percentage as a function of abundance rank (Additional file 1: Figure S8). Among the first 1000 proteins only 5 % were bacterial proteins. This proportion increased steadily until it reached 35 % for the total set of about 6000 proteins. Expressed as the percentage of bacterial proteins per 100 proteins, the chance to identify bacterial proteins reached more than 50 % towards the limit of detection. This suggests that increasing the depth of proteomic analysis would preferentially uncover further bacterial proteins and that our coverage of the oral metaproteome is far from saturation. As the depth of our bacterial detection increases in the future, it may also be possible to analyze bacterial pathways and how they change across different conditions of the oral cavity.

The simultaneous detection of bacterial and human proteomes in our samples allowed us to directly compare them quantitatively (Fig. 5b). The most abundant bacterial protein was F1WNZ3, the *Moraxella catarrhalis* homolog of chaperone protein HscA, which is involved in maturation of iron-sulfur-containing proteins. Its abundance was only 100-fold lower than the top human protein, alpha-amylase 1. Further highly abundant proteins of the bacterial metaproteome included proteins with household functions, such as A0A096BHY1, which is a glyceraldehyde-3-phosphate dehydrogenase, or E0Q9Q6, a subunit of DNA polymerase III. Sequence alignment in Perseus showed that many of the very abundant bacterial proteins were highly conserved. Therefore, peptides from different species likely contribute to their abundance.

The number of significantly identified human proteins decreased to about 4000 in the combined search space (Fig. 5b). Thus, almost a third of the overall protein count of 6197 is due to the microbiome. The bacterial proteins originated from four main phyla, with 300 to 800 uniquely assigned proteins, each of which spanned the entire abundance range (Additional file 1: Figure S9). In analogy to the top-three-peptide method commonly used in label-free abundance estimation of proteins [23, 24], we defined an approximate quantitative measure of the abundance of a bacterial genus as the summed MS intensity of the top ten most abundant peptides across all samples. These data were available for nearly all genera and, as in the protein case, comparing just the ten highest peptide intensities should be a better measure than summing all peptides, which would tend to overestimate abundance differences. The top ten peptides were determined among all peptides of a genus, not just unique peptides. This comes at the disadvantage that peptides shared by two genera could lead to an overestimation of the taxon's abundance. Considering only unique peptides would have put genera with large sequence identity at a great disadvantage compared with genera with relatively distinct peptide sequences. However, this shows that adequate quantification of bacterial genera by their proteomes is challenging and at the present coverage our quantitative readouts should be considered as approximations rather than exact quantifications.

We applied our bacterial quantification measure to all detected genera and plotted the abundance of the top 20 (Fig. 5c). As expected from quantification performed by 16S RNA sequencing [40, 41], *Streptococcus* was the most abundant genus. The top ten genera did not show drastic differences in abundance (the integrated MS peptide signal of the top ten peptides was 4.0×10^{10} for *Streptococcus* and 1.4×10^{10} for *Lactococcus*). While we believe that the quantitative trends between bacteria are correct, more accurate quantification would require deeper sequence coverage of the bacterial proteomes.

The Human Microbiome Project (HMP) has generated large datasets of human microbiomes using next-generation sequencing [25]. We compared our quantitative bacterial proteomes with the whole genome sequencing data of the HMP in a PCA (Fig. 5d) and a heatmap of genera against samples (Additional file 1: Figure S1). The different body sites clustered separately in the genome data, with our proteomic data strikingly co-localizing with the oral microbiome. We did not expect such close co-localization given that both datasets originate from different samples and individuals. However, these results are in agreement with previous findings showing that the oral microbiome has relatively low diversity among individuals (beta diversity) [25]. The human microbiome study had collected samples from different locations in the mouth, but these data cluster together in the PCA, suggesting that the microbiome is similar throughout the oral cavity.

Variation and dynamics of the metaproteome

Apart from estimating bacterial abundances, our data allow a quantitative comparison of the same genus upon perturbation or across individuals. Overall, individuals varied little in their bacterial diversity in accordance with the HMP [25]. A scatterplot of two typical donors reveals that bacterial abundances are similar for many of them, with a strong mean R^2 of 0.82 (Fig. 6a shows a typical scatter plot). However, there are also genera that varied up to tenfold.

The cumulative abundances of the top eight bacterial genera across all donors indicate differences in total bacterial mass of up to threefold (Fig. 6b). Variation in the relative abundance of genera is much smaller (Fig. 6c) and the same analysis at the level of the five most abundant phyla showed similar variation.

When aggregating males and females separately, the two groups exhibited very comparable bacterial abundances that were highly correlated ($R^2 = 0.94$; Fig. 6d). Thus, proteomics indicates that sex differences in the oral microbiome are minor. In contrast, bacterial abundance changed drastically after eating breakfast and tooth brushing. The high abundance bacterial genera were reduced 2.5-fold on average, while the lower abundant ones generally showed even stronger reduction (Fig. 6e, f). The *Streptococcus* genus, which contains *S. mutans*, was reduced by almost threefold after

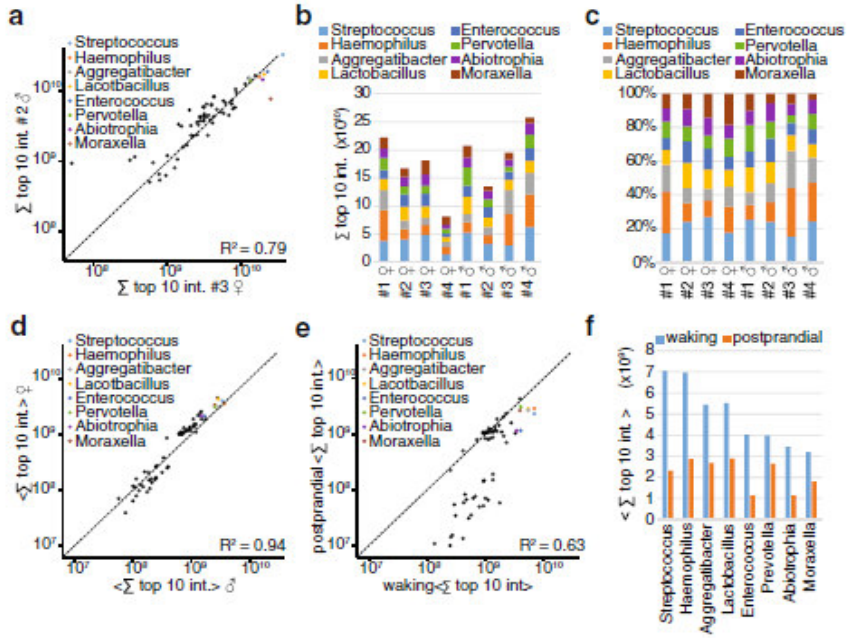


Fig. 6 Bacterial composition across donors and time points. **a** Typical scatter plot for two donors of the bacterial genera quantified by the sum of their top ten peptide intensities from fractionated proteome measurements. The eight most abundant genera are color coded. **b** Absolute quantification of the eight most abundant bacterial species depicted for each individual donor and (**c**) normalized to 100 %. **d** Mean genus quantities between males and females treated as groups from single run measurements. **e** Comparison of mean bacterial abundances at waking to the mean bacterial abundances in the postprandial state. **f** Reduction in bacterial abundance between the waking saliva samples and the postprandial samples

tooth brushing (Fig. 6f). It has been established that the *S. mutans* species is not the only one involved in cavity formation [42] and it would now be interesting to study the effects of different oral hygiene regimes on the oral bacterial community at the proteome level.

Our deep saliva proteomes also allow combined analysis of the human and bacterial proteome changes in response to the same perturbations. For instance, at waking, when bacterial abundance is high, the human saliva proteome was primed towards bacterial defense with substantial enrichment of proteins annotated with the Uniprot keywords "antibiotic" and "anti-microbioal". Given the higher abundance of the microbiome at waking, this likely reflects the body's effort to limit bacterial proliferation during the night when these populations are relatively undisturbed. This example illustrates the utility of the simultaneous detection of the human and bacterial proteomes for the study of the interplay of the host and microbiome.

Conclusions

Here we employed shotgun proteomics with a state of the art workflow and identified more than 5500 proteins, the largest number of human proteins in a body fluid to date. Comparison with the plasma proteome established that the quantitative protein levels do not correlate.

We showed that shotgun proteomics can now readily determine 50 bacterial genera in saliva but the sequence coverage of bacterial proteins and organisms suggests that we have only scratched the surface of the oral bacterial proteome. Quantitative comparison to next-generation sequencing data from the HMP [25] revealed excellent agreement, suggesting that proteomics could provide a valuable complement to sequencing-based measurements of the human microbiome. Furthermore, proteomics appears uniquely positioned to study the interplay of the human immune system with commensurate and pathogenic bacteria on the protein level. With improving technology, our workflow might even become attractive for clinical microbiology since bacteria do not need to be grown and rapid bacterial resistance testing could become possible by directly measuring proteins that confer resistance to antibiotics. An important task for the future is to better characterize and annotate bacterial sequences in order to provide comprehensive, non-redundant databases for bacterial proteomics.

In conclusion, the depth and relatively straightforward nature of our workflow should make it a powerful new tool in the detection of biomarkers of diseases of the oral cavity as well as facilitate complementary studies of the microbiome in different contexts. In particular, proteomics appears uniquely positioned to study the interplay of the human immune system with commensurate and pathogenic bacteria at the systems level. We hope that such approaches will help to open new avenues in clinical application and for microbiology in the future.

Availability of supporting data

The data sets supporting the results of this article are available in the proteomeXchange repository (http://www.proteomexchange.org), accession number PXD 003028.

Additional file

Additional file 1: Supplementary **Figures S1-S9**. (PDF 13519 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NG and MM conceived the project; NG, NK, GP, and SS designed the experiments; NG, PG, and JJ performed the experiments; NG, PS, and JC interpreted the data; and NG and MM wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Marco Hein, Sean Humphrey, Igor Paron, Korbinian Mayr, and Gaby Sowa for help and fruitful discussions and Marco Hein for critical reading of the manuscript. This work was supported by the Max Planck Society for the Advancement of Science

Author details

[1]Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. [2]PreOmics GmbH, Am Klopfespitz 19, D-82152 Martinsried, Germany. [3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark. [4]Max von Pettenkofer-Institut für Hygiene und Medizinische Mikrobiologie, Marchioninistr. 17, D-81377 München, Germany. [5]Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany.

Received: 18 January 2016 Accepted: 24 March 2016
Published online: 21 April 2016

References

1. Shpitzer T, Hamzany Y, Bahar G, Feinmesser R, Savulescu D, Borovoi I, et al. Salivary analysis of oral cancer biomarkers. Br J Cancer. 2009;101:1194–8.
2. Delaleu N, Mydel P, Kwee I, Brun JG, Jonsson MV, Jonsson R. High fidelity between saliva proteomics and the biologic state of salivary glands defines biomarker signatures for primary Sjogren's syndrome. Arthritis Rheumatol. 2015;67:1084–95.
3. Yoshizawa JM, Schafer CA, Schafer JJ, Farrell JJ, Paster BJ, Wong DT. Salivary biomarkers: toward future clinical and diagnostic utilities. Clin Microbiol Rev. 2013;26:781–91.
4. Bandhakavi S, Stone MD, Onsongo G, Van Riper SK, Griffin TJ. A dynamic range compression and three-dimensional peptide fractionation analysis platform expands proteome coverage and the diagnostic potential of whole saliva. J Proteome Res. 2009;8:5590–600.
5. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014;509:357–60.
6. Berer K, Mues M, Koutrolos M, Rasbi ZA, Boziki M, Johner C, et al. Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. Nature. 2011;479:538–41.
7. Tremaroli V, Backhed F. Functional interactions between the gut microbiota and host metabolism. Nature. 2012;489:242–9.
8. Willyard C. Microbiome: Gut reaction. Nature. 2011;479:S5–7.
9. Aas JA, Griffen AL, Dardis SR, Lee AM, Olsen I, Dewhirst FE, et al. Bacteria of dental caries in primary and permanent teeth in children and young adults. J Clin Microbiol. 2008;46:1407–17.

10. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, Pignatelli M, et al. The oral metagenome in health and disease. ISME J. 2012;6:46–56.
11. Belda-Ferre P, Williamson J, Simon-Soro A, Artacho A, Jensen ON, Mira A. The human oral metaproteome reveals potential biomarkers for caries disease. Proteomics. 2015;15:3497–507.
12. Kulak NA, Pichler G, Paron I, Nagaraj N, Mann M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat Methods. 2014;11:319–24.
13. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. 2007;2:1896–906.
14. Wang Y, Yang F, Gritsenko MA, Wang Y, Clauss T, Liu T, et al. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. Proteomics. 2011;11:2019–26.
15. Gilar M, Olivova P, Daly AE, Gebler JC. Orthogonality of separation in two-dimensional liquid chromatography. Anal Chem. 2005;77:6426–34.
16. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72.
17. Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, 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. 2012;11:M111.013722.
18. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10:1794–805.
19. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014;13:2513–26.
20. Cox J, Mann M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics. 2012;13 Suppl 16:S12.
21. Philipp G, Nils AK, Garwin P, Lesca H, Daniel T, Matthias M. Plasma Proteome Profiling to Assess Human Health and Disease. Cell Syst. 2016;2:185–195.
22. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford). 2010;2010:baq013.
23. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011;473:337–42.
24. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics. 2006;5:144–56.
25. HMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
27. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
28. Kelstrup CD, Jersie-Christensen RR, Batth TS, Arrey TN, Kuehn A, Kellmann M, et al. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. J Proteome Res. 2014;13:6187–95.
29. Scheltema RA, Hauschild JP, Lange O, Hornburg D, Denisov E, Damoc E, et al. The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. Mol Cell Proteomics. 2014;13:3698–708.
30. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1:845–67.
31. Nagaraj N, Mann M. Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. J Proteome Res. 2011;10:637–45.
32. Cuevas-Cordoba B, Santiago-Garcia J. Saliva: a fluid of study for OMICS. Omics. 2014;18:87–97.
33. Schaab C, Geiger T, Stoehr G, Cox J, Mann M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. Mol Cell Proteomics. 2012;11:M111.014068.
34. Hunt A, Harrington D, Robinson S. Vitamin B12 deficiency. BMJ. 2014;349:g5226.
35. Stabler SP. Vitamin B12 deficiency. N Engl J Med. 2013;368:2041–2.
36. Carmel R, Brar S, Frouhar Z. Plasma total transcobalamin I. Ethnic/racial patterns and comparison with lactoferrin. Am J Clin Pathol. 2001;116:576–80.
37. Carmel R, Green R, Jacobsen DW, Rasmussen K, Florea M, Azen C. Serum cobalamin, homocysteine, and methylmalonic acid concentrations in a multiethnic elderly population: ethnic and sex differences in cobalamin and metabolite abnormalities. Am J Clin Nutr. 1999;70:904–10.
38. Kirschbaum C, Hellhammer DH. Salivary cortisol in psychobiological research: an overview. Neuropsychobiology. 1989;22:150–69.
39. Opperman T, Richardson JP. Phylogenetic analysis of sequences from diverse bacteria with homology to the Escherichia coli rho gene. J Bacteriol. 1994;176:5033–43.
40. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol. 2005;43:5721–32.
41. Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF, et al. Bacterial diversity in the oral cavity of ten healthy individuals. ISME J. 2010;4:962–74.
42. Takahashi N, Nyvad B. The role of bacteria in the caries process: ecological perspectives. J Dent Res. 2011;90:294–303.

3.6. Article 6: HCD Fragmentation of Glycated Peptides

Authors: Eva C. Keilhauer[1], Philipp E. Geyer[1,2], and Matthias Mann[1,2]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

High glucose level is the main symptom and diagnostic criterion of diabetes, an ever growing disease that affects more than 400 million humans around the globe. Today the gold standard for its diagnosis is to calculate the percentage of hemoglobin that has a distinct glycation, the so called HbA1c value.

Glycation is a non-enzymatic reaction of glucose with amino groups of proteins, and its products reflect the glucose concentration in the blood stream. Because of the known mass that is added by the glycation reaction to lysine or arginine residues, it is possible to detect and quantify these products by MS-based proteomics. In the past, extensive sample preparation with affinity enrichment strategies have been used to identify glycated peptides in plasma samples. Typically, researchers employ a dedicated workflow to obtain the glycation information. This makes the measurement of glycation sites very time consuming and incompatible with high throughput measurements in the context of clinical or biomarker research. Furthermore, in the past specialized fragmentation methods, including electron transfer dissociation (ETD) were applied, which are only available on some MS instruments.

In this publication we show that the standard fragmentation method 'higher-energy collisional dissociation' (HCD) alone can efficiently identify glycated peptides. We establish optimal fragmentation parameters and identify early glycation products on *in vitro* glycated proteins. Furthermore, we apply this workflow to plasma samples and detect more than 100 glycation sites in single run analysis of undepleted and unenriched plasma samples. We have now incorporated the identification of glycated peptides in our Plasma Proteome Profiling pipeline, which gives us additional information for biomarker research. Future research will focus on accurate quantification of glycation, which would allow the diagnosis of uncontrolled glucose levels in all plasma samples.

Journal of proteome research

Technical Note

pubs.acs.org/jpr

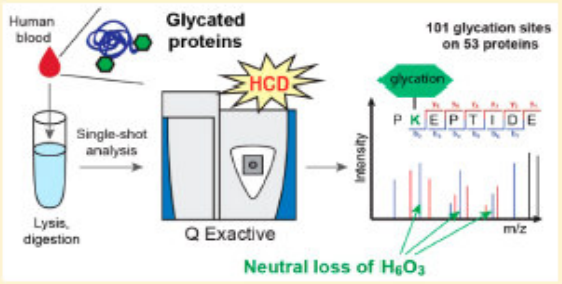
HCD Fragmentation of Glycated Peptides

Eva C. Keilhauer, Philipp E. Geyer, and Matthias Mann*

Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

Supporting Information

ABSTRACT: Protein glycation is a concentration-dependent nonenzymatic reaction of reducing sugars with amine groups of proteins to form early as well as advanced glycation (end-) products (AGEs). Glycation is a highly disease-relevant modification but is typically only studied on a few blood proteins. To complement our blood proteomics studies in diabetics, we here investigate protein glycation by higher energy collisional dissociation (HCD) fragmentation on Orbitrap mass spectrometers. We established parameters to most efficiently fragment and identify early glycation products on in vitro glycated model proteins. Retaining standard collision energies does not degrade performance if the most dominant neutral loss of H_6O_3 is included into the database search strategy. Glycation analysis of the entire HeLa proteome revealed an unexpected intracellular preponderance for arginine over lysine modification in early and advanced glycation (end-) products. Single-run analysis from 1 μL of undepleted and unenriched blood plasma identified 101 early glycation sites as well as numerous AGE sites on diverse plasma proteins. We conclude that HCD fragmentation is well-suited for analyzing glycated peptides and that the diabetic status of patients can be directly diagnosed from single-run plasma proteomics measurements.



KEYWORDS: *protein glycation, higher-energy collisional dissociation, diabetes, blood plasma, AGEs*

INTRODUCTION

Protein glycation, in contrast with enzyme-mediated glycosylation, is produced by the nonenzymatic reaction of glucose molecules or other reducing sugars with amine groups of proteins and is also known as Maillard reaction.[1] Glucose first attaches to form a Schiff base, which then rearranges into the relatively stable Amadori compound,[2] to which we refer here as "early glycation product". Glycated proteins can further react to form advanced glycation endproducts (AGEs), or proteins can directly react with glucose-derived reactive dicarbonyls like methylglyoxal to form AGEs.[3] Glucose is an essential and omnipresent energy source in humans and is tightly regulated in a narrow concentration band in healthy individuals. Dysregulation of glucose levels is the principal feature of diabetes, a growing health epidemic, currently affecting an estimated 415 million individuals worldwide according to the International Diabetes Federation (IDF) Diabetes Atlas (7th edition). The extent of protein glycation and AGEs is increased in proportion to the glucose concentration, and the glycation level of one particular blood protein, hemoglobin, is routinely assessed in the diagnosis of diabetes as well as for long-term monitoring of blood glucose levels of diabetes patients. More specifically, glycation of the N-terminal valine of the hemoglobin beta-chain is assessed, a clinical parameter known as HbA1c.[4,5] Because the lifespan of erythrocytes and hence hemoglobin is around 120 days, the HbA1c value reflects the average blood glucose concentration of the last 6 to 8 weeks.[2,6] Hence the HbA1c-test is often more robust than oral glucose tolerance tests that can be influenced by various factors such as recent food intake, exercise, and blood sampling time. If the HbA1c value can be stabilized close to normal levels, patients have a much better prognosis and less diabetic complications than those with poorly controlled HbA1c values.[7] Glycation and AGEs are central to the development of typical diabetic complications and also play a role in aging and neurodegenerative and cardiovascular diseases.[8−13]

The current and strong focus on glycated hemoglobin and a few more proteins is presumably due to a lack of appropriate methods to robustly detect, characterize, and quantify other glycated proteins. Owing to its extreme complexity and extraordinary dynamic range, blood plasma is the most challenging proteome;[14−16] however, investigation of other glycated proteins could help to better diagnose, monitor, and understand metabolic conditions such as diabetes. For example, measuring several glycated proteins with different lifespans might yield a more detailed picture of blood glucose levels of patients over the last days to weeks.[17−19]

Mass spectrometry (MS) is the method of choice to investigate post-translational protein modifications (PTMs) in an unbiased manner.[20] Analysis of glycation in body fluids has been challenging because of its low stoichiometry and enrichment strategies such as boronate affinity chromatography (BAC) are typically employed.[21,22] Sample complexity is often

Received: May 20, 2016
Published: July 18, 2016

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

additionally reduced by depleting the most abundant plasma proteins or fractionating the plasma on the peptide level. In this way, and by pooling and fractionating a large number of diverse samples, the most comprehensive study to date found evidence of around 1100 glycated proteins from human plasma.[23] Such elaborate protocols are useful for generating glycation site resources; however, they are not practical for clinical tests. We have recently reported a method called "plasma proteome profiling", which allows measuring hundreds of plasma proteins from only 1 μL of plasma in a single-run format without depletion or fractionation.[24] We therefore wondered if we could complement the patient information gained from plasma proteome profiling with the diabetic status by determining glycation of plasma proteins.

Glycated peptides have been studied by MS/MS using various fragmentation techniques.[6] Collision-induced dissociation (CID)[25] in ion traps suffers from dominant neutral losses of the labile Amadori compound, often leading to insufficient fragmentation of the peptide backbone for identification of the peptide sequence and the glycation site.[26,27] Neutral loss-triggered MS^3 scans partly alleviate this problem but at the cost of lower throughput and sensitivity.[28] Electron-transfer dissociation (ETD),[29] a technology generally known to be well-suited for investigating labile modifications, is very effective for glycated peptides. Using ETD, no neutral losses and almost complete series of c- and z-ions were observed;[30] however, ETD is only implemented on specialized mass spectrometers and not on the benchtop Orbitrap instruments that are routinely used in many laboratories. Initial promising results have also been obtained for higher energy collisional dissociation (HCD),[31] however, so far always in combination with other techniques.[28] As the benchtop Orbitrap instruments (Q Exactive) exclusively feature HCD fragmentation, we therefore set out to systematically evaluate how well-glycated peptides can be fragmented and analyzed with HCD-MS2 scans alone.

■ EXPERIMENTAL SECTION

In Vitro Glycation of BSA and HSA

Both bovine serum albumin (BSA) and human serum albumin (HSA) (human fraction 5 powder) were purchased from Sigma-Aldrich. BSA (100 mg/mL) was incubated with 1 M glucose in 50 mM Tris HCl buffer pH 7.5 at room temperature for the indicated times. HSA (10 mg/mL) was incubated with 1 M glucose in the same buffer for 48 h. Both BSA and HSA were digested with trypsin (Promega) with an enzyme to protein ratio of around 1:20 to 1:50 in digestion buffer (2 M urea and 1 mM dithiothreitol (DTT) in 50 mM TrisHCl pH 7.5). After 20 min, 5 mM chloroacetamide (CAA) was added to the samples; then, they were incubated overnight to ensure a complete digest. On the next day, the digestion was stopped by the addition of 1 μL of trifluoroacetic acid (TFA) per sample. The peptides were desalted and purified on StageTips (self-made pipet tips containing two layers of C_{18} material) according to the standard protocol.[32] The StageTips were stored at 4 °C until the sample was measured. BSA and HSA samples were eluted from the C_{18} StageTips with 2 × 20 μL of buffer B (80% acetonitrile (ACN), 0.5% acetic acid). The organic solvent was removed in a SpeedVac concentrator for 20 min; then, the peptide mixture was acidified with buffer A* (2% ACN, 0.1% TFA) to a final sample size of 5 μL.

Preparation of HeLa Digests

HeLa cells were cultured in high glucose DMEM with 10% fetal bovine serum and 1% penicillin−streptomycin (all from Life Technologies). Around 5×10^7 cells were harvested and lysed in 6 M urea/2 M thiourea. Proteins were reduced with 1 mM DTT for 30 min at room temperature, then alkylated with 5 mM iodoacetamide (IAA) for 20 min in the dark. Proteins were digested overnight with LysC and trypsin. The digest was stopped by adding TFA; then, peptides were purified on StageTips as described above.

Preparation of Whole Blood and Blood Plasma Samples: Protein Digestion and in-StageTip Purification

Sample preparation for plasma was done as previously described.[33] In brief, 1 μL of plasma was mixed with 24 μL of SDC reduction and alkylation buffer.[34] After protein denaturation by boiling for 10 min, LysC and trypsin were added in a 1:100 ratio (μg enzyme to μg protein), and digestion was performed for 1 h at 37 °C. Peptides were acidified by adding 125 μL of ethyl acetate/1% TFA, and 20 μg was transferred to StageTips, containing two 14-gauge SDB-RPS plugs. Washing steps included two times 100 μL of ethyl acetate/1% TFA and one time 100 μL of ddH_2O/0.2% TFA. The purified peptides were eluted with 60 μL of elution buffer (80% ACN, 19% ddH_2O, 1% ammonia) into auto sampler vials. The collected material was dried to completion using a SpeedVac centrifuge at 45 °C (Eppendorf, Concentrator plus). Peptides were suspended in 2% ACN, 0.1% TFA and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510) prior to analysis.

The sample preparation procedure for whole blood included an additional sonication step of 15 min by a Diagenode Bioruptor prior to digestion.

LC−MS/MS Measurement of HSA and BSA

Samples were analyzed by nanoflow liquid chromatography (LC−MS/MS) on an EASY-nLC HPLC system (Thermo Fisher Scientific) that was online coupled to either a Q Exactive plus or a Q Exactive HF mass spectrometer (all Thermo Fisher Scientific) through a nanoelectrospray ion source (Thermo Fisher Scientific). A 50 cm column with a 75 μm inner diameter in-house packed with 1.9 μm reversed-phase silica beads (ReproSil-Pur C_{18}-AQ, Dr. Maisch) was used for the chromatography. Peptides were separated using a linear gradient from 5.6 to 25.6% ACN in 0.1% formic acid at a constant flow of 250 nL/min, then directly electrosprayed into the mass spectrometer. Overall gradient length was 1 h. The column oven (Sonation) was heated to 55 °C. The spray voltage was set to 2.4 kV and the heated capillary temperature was set to 250 °C.

BSA/HSA samples were measured using a data-dependent top10 method, and the BSA glycation time course was measured in a top1 method. Instruments were controlled by Tune Plus 2.0 and Xcalibur 2.0. On the Q Exactive plus, full scans (m/z 300−1650) were acquired with a resolution of 70 000 at 200 m/z and an AGC target of 3E06 ions and fragmentation scans with a resolution of 17 500 at 200 m/z and an AGC target of 1E05 ions. Maximum ion accumulation times were 20 ms for the full scans and 120 ms for the fragmentation scans. On the Q Exactive HF, full scans (m/z 300−1650) were acquired with a resolution of 60 000 at 200 m/z and an AGC target of 3E06 ions and fragmentation scans with a resolution of 16 000 at 200 m/z and an AGC target of 1E05 ions. Maximum ion accumulation times were 120 ms for both full scans and

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

fragmentation scans. The most intense ions from the full scans were isolated with an isolation width of 1.4 m/z and fragmented using HCD, with a normalized collision energy (NCE) of 25% (Q Exactive plus) or 27% (Q Exactive HF) unless otherwise specified in the text. Dynamic exclusion was enabled for a duration of 20 s.

LC–MS/MS Measurement of HeLa Digests

Samples were measured on a Q Exactive HF essentially as described for BSA and HSA with the following alterations: Gradient length was 2 h, HeLa samples were measured in top15 mode, and the maximum ion accumulation time for fragmentation scans was 25 ms.

LC–MS/MS Measurement of Blood Plasma/Whole Blood

Samples were measured on a Q Exactive HF essentially as described for BSA and HSA with the following alterations: Column length was 40 cm and the column oven temperature was set to 60 °C. Gradient length was 100 min and samples were measured using a data-dependent top15 method. Full scans (m/z 300–1650) were acquired with a resolution of 120 000 at 200 m/z, an AGC target of 3E06 ions and a maximum injection time of 55 ms. An isolation window of 1.5 m/z and a fixed first mass of 100 m/z were used for MS/MS scans. HCD fragmentation was performed with an NCE of 27. MS/MS scans were acquired with a resolution of 30 000 at 200 m/z with an AGC target of 1E05 ions and a maximum injection time of 55 ms. Dynamic exclusion was enabled for a duration of 30 s.

Data Analysis

All raw data were analyzed using the MaxQuant[35] software environment (version 1.5.3.0). The software searched the derived peak list using the built-in Andromeda search engine[36] against either a bovine reference proteome downloaded from Uniprot (http://www.uniprot.org/) on February 2016 (24 481 sequences) or a human reference proteome downloaded from Uniprot in May 2013 (88 847 sequences). In all cases, a file containing 247 frequently observed contaminants such as human keratins and proteases was included in the search. Trypsin was chosen as the protease with strict specificity for cleavage C-terminal to K or R required. Up to two missed cleavages per peptide were allowed. The minimum peptide length was set to seven amino acids. Because of the sample preparation, carbamidomethylation of cysteine was set as a fixed modification (57.021464 Da). N-acetylation of protein N-termini (42.010565 Da) and oxidation of methionine (15.994915 Da) were set as variable modifications. For glycation/AGE analysis, the corresponding modification with/without different neutral losses was defined in Andromeda configuration and added to the variable modifications, as stated in the text (Glycation: 162.052823 Da, CML: 58.005479 Da, CEL: 72.021129 Da, MG-H1:54.010565 Da, Argpyr: 80.026215 Da, 3DG-H1:144.042259 Da). All other parameters were left at standard settings. Peptide and protein identifications were filtered at a false discovery rate (FDR) of 1%. The "match between runs" option was used where specified in the text with a match time window of 0.7 min and an alignment time window of 20 min.

Further analysis of the MaxQuant output tables was performed using the Perseus software (version 1.5.3.0), which is part of the MaxQuant environment. Plots were produced in R (version 2.15.3).

Data Availability

Raw data and MaxQuant output files are accessible via ProteomeXchange[37] with identifier PXD004182.

■ RESULTS AND DISCUSSION

HCD Fragmentation of Glycated Peptides

Orbitrap mass spectrometers have proven to be powerful instruments for proteomics in general and clinical proteomics in particular and today are standard in many laboratories. The widespread benchtop quadrupole Orbitrap instruments (Q Exactive family) feature only HCD as fragmentation method. Because previous work on glycated peptides had employed ETD or a combination of other fragmentation methods with HCD, we here set out to investigate whether glycated peptides can be identified solely on the basis of HCD–MS/MS scans. As glycation is typically studied in blood plasma, we chose HSA as a model protein. We glycated HSA in vitro, digested it with trypsin, and measured the resulting peptides on a Q Exactive HF without optimizing the instrument in any way. In the MaxQuant data analysis software,[35] we included protein glycation ($C_6H_{10}O_5$; 162.0528 Da) as a variable modification on lysine, which is the major target for glycation by glucose, and on arginine. The "matching between runs" algorithm was enabled between the three technical replicates, which transfers peptide identifications to LC–MS/MS runs where the same peptide was present but was not sequenced. Surprisingly, in view of the complex experimental setup previously employed in the analysis of glycation, already this first experiment identified 45 unique glycation sites on HSA. Most sites (42) were located on lysine, consistent with the fact that this residue is the primary target for this type of glycation, and only three sites were found on arginine. Thus, the large majority of the 59 lysines in mature HSA can be glycated in vitro by incubation with high glucose concentrations. Interestingly, UniProt lists only 20 of the 42 lysine sites as glycated in vitro or in vivo, while 22 of them were incorrectly annotated as "not glycated" in UniProt (See Table 1). Of the eight reported in vivo glycation sites, we found six in our in vitro setting, namely, K36, K257, K341, K375, K549, and K558, and interestingly, we found no evidence of the two other in vivo sites K305 and K463. Instead, we did find good evidence of glycation on K460, which has not been reported to be glycated before. In general, the fact that we identified such a high number of sites on this widely used model protein suggests that standard HCD–MS/MS scans are remarkably well-suited for the characterization of glycated peptides.

Optimizing the Collision Energy for Glycated Peptides

In addition to backbone fragmentation, glycated peptides can also fragment by losing all or part of the Amadori product during CID and HCD fragmentation.[27,39] Therefore, collision energies for HCD might be different for the identification of glycated peptides compared with unmodified peptides, which was suggested by the relatively low identification scores of the glycated HSA peptides described above. Using in vitro glycated BSA as a model protein, we performed LC–MS/MS runs with six different normalized collision energies (NCEs) centered around the standard NCE that we use in our shotgun proteomics experiments. Plotting the number of unmodified BSA peptides identified at each collision energy confirmed that an NCE of 25% on the instrument employed (Q Exactive Plus) was indeed optimal for these peptides (Figure 1A). The same

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881–2890

Journal of Proteome Research Technical Note

Table 1. Detected Glycation Sites on HSA[a]

Amino acid	Position	Status	Identification score	Mean log2 site intensity
K	28	not glycated	60.398	21.5041
K	36	glycated	98.048	32.5308
K	44	not glycated	158.13	28.4838
K	75	*in vitro* glycated	87.652	26.4733
K	88	not glycated	107.3	27.6692
K	97	not glycated	121.13	29.1766
K	130	not glycated	61.253	26.8779
K	160	glycated*	90.475	29.021
K	161	*in vitro* glycated	82.885	28.7883
K	183	not glycated	58.246	30.5033
K	186	*in vitro* glycated	93.371	24.9165
K	198	not glycated	84.653	29.4034
K	205	not glycated	61.444	27.4259
K	214	not glycated	47.774	31.1864
K	223	*in vitro* glycated	73.26	32.3641
K	236	not glycated	62.546	30.431
K	249	*in vitro* glycated	86.014	25.6519
K	257	glycated	84.759	34.4102
K	264	not glycated	86.449	26.1479
K	286	not glycated	153.39	26.7257
K	300	*in vitro* glycated	105.65	29.85
K	337	*in vitro* glycated	72.913	26.7576
K	341	glycated	59.067	29.6695
K	347	*in vitro* glycated	128.74	26.7064
K	375	glycated	76.332	31.5972
K	383	not glycated	124.25	27.0238
K	396	not glycated	69.016	27.2544
K	402	*in vitro* glycated	90.453	32.1664
K	413	glycated*	111.93	28.9697
K	426	not glycated	147.4	29.7922
K	437	*in vitro* glycated	66.022	27.5527
K	438	not glycated	102.64	33.1738
K	460	not glycated	60.255	24.0608
K	490	not glycated	49.559	28.0705
K	499	not glycated	114.86	29.1763
K	548	not glycated	59.227	30.3853
K	549	glycated	102.87	33.5487
K	558	glycated	92.051	24.7948
K	565	not glycated	56.122	25.7802
K	569	*in vitro* glycated	103.13	29.3973
K	597	*in vitro* glycated	57.034	26.0855
K	598	not glycated	87.667	26.6658
R	141	-	49.66	28.4491
R	210	glycated*	104.22	21.751
R	361	-	122.34	26.1596

[a]Glycation sites ordered by position with additional information about the amino acid, the status in UniProt and/or in a recent review[38] if marked by an asterisk, the mean identification score, and the mean log2-transformed intensity.

analysis revealed a broad optimum NCE for the number of identified glycated peptides, centered between 20 (43 sites) and 25% (42 sites) (Figure 1B). An NCE of 40%, in contrast, dramatically reduced identification success. Next we investigated for each identified glycation site in which of the measurements at the different NCEs it was best localized to a particular amino acid (localization probability) and where it obtained the maximum database identification score. By these measures, an NCE of 20% appeared to be optimal for both localization and identification (Figure 1C,D).

When we examined the fragmentation spectra of the glycated peptides more closely, we found that at higher NCEs there were typically no fragments carrying the full modification of 162.053 Da. Furthermore, b-ions were mostly absent from the spectra, and often a number of intense peaks in the higher mass range were unexplained by standard backbone fragmentation (for an example, see spectrum in Figure 2A). The Amadori compound can lose several water molecules and formaldehyde during CID and HCD fragmentation,[27,39] resulting in residual modification masses of 144.0423, 126.0317, 108.0211, 96.0211, and 78.0106 Da (Figure 2B). Additionally, we also observed loss of the entire glucose moiety from the fragments and the intact peptide. After annotating the spectrum in Figure 2A with these reduced forms of glycation using the expert system for fragment annotation,[40] we were able to explain basically all of the peaks in the spectrum (Figure 2C). Essentially the complete series of backbone fragments was represented in at least one of the possible modification states, with the exception of cleavage between the N-terminal phenylalanine and the glycated lysine. In general, while the loss of only one water molecule leading to the 144.0423 Da modification seemed to occur rarely, other pathways appeared to be more dominant: the loss of three water molecules leading to the 108.0211 Da modification and the loss of three water and one formaldehyde molecules leading to the 78.0106 Da modification.

We reasoned that by taking the neutral losses into account we might be able to use our standard collision energy of 25% (or 27% on the Q Exactive HF) to both obtain efficient backbone fragmentation as well as confidently identify glycation sites. Because the MaxQuant software supports only one neutral loss per modification to avoid combinatorial explosion, we next determined the most common neutral loss in a systematic way. We defined seven different versions of glycation for the search engine: without any neutral loss and with a neutral loss of H_2O, H_4O_2, H_6O_3, CH_6O_3, CH_8O_4, and finally $C_6H_{10}O_5$ corresponding to the entire modification. Inter-

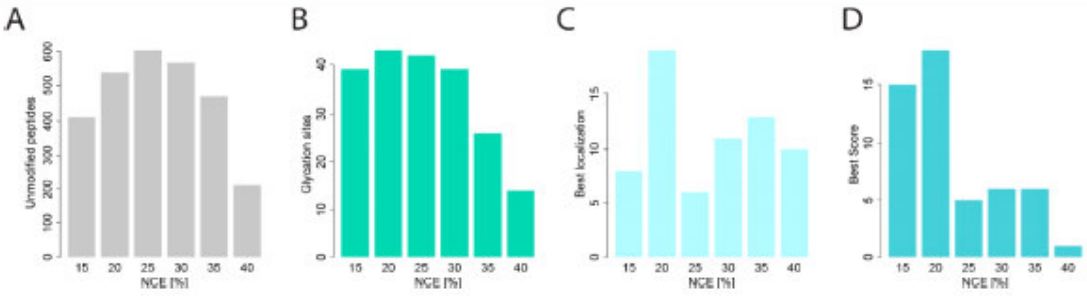


Figure 1. Evaluation of different collision energies. (A) Number of unmodified BSA peptides identified with six different normalized collision energies (NCEs) from 15 to 40%. (B) Glycation sites identified when searching for glycation on K and R. (C) Localization score as a function of the NCE. (D) Andromeda database identification score[36] as a function of the NCE.

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

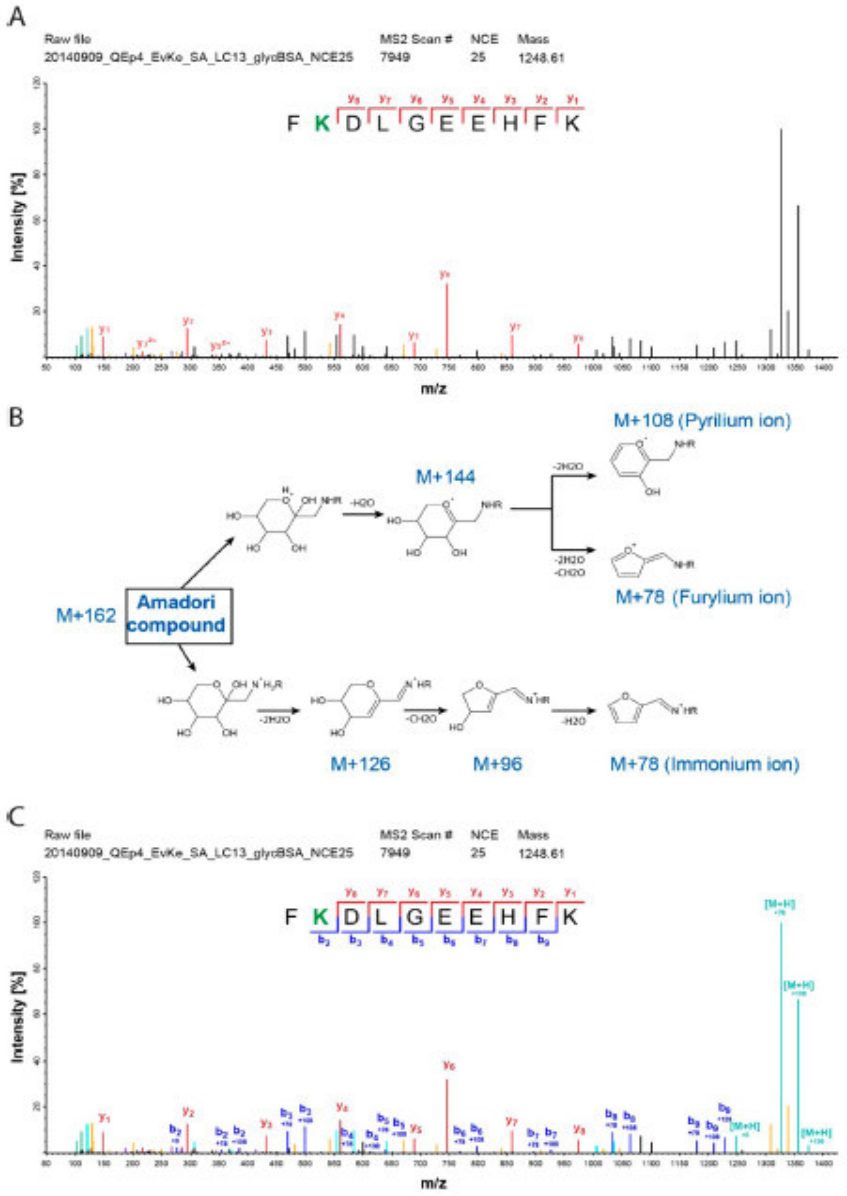


Figure 2. HCD fragmentation behavior of glycated peptides. (A) Spectrum of the glycated BSA peptide FK*DLGEEHFK with an NCE of 25% (the asterisk or green color denotes the position of glycation). An almost complete y-ion series is apparent, however, not a single b-ion was found and many peaks in the spectrum are unexplained. (B) Scheme of proposed pathways generating different neutral losses during CID/HCD fragmentation (adapted from ref 39). (C) The same spectrum as in panel A now manually annotated with the different neutral losses, which explains essentially all fragments.

rogating the data file obtained at the NCE of 25% with the seven different versions of glycation on lysine, we found that a neutral loss of three water molecules (H_6O_3) leading to a residual mass of 108.0211 Da yielded most glycation sites in total (47 sites, see Figure 3A). This search mode also produced most high confidence sites, for example, 44 sites with an identification score over 75. CH_8O_4, with a residual modification mass of 78.0106 Da, was the next most common neutral loss, followed by loss of the entire glucose moiety. With these optimized collision and search settings, we now found an additional 17 high confidence glycation sites on BSA compared with the search without neutral loss (Figure 3A). Figure 3B illustrates that the neutral loss of three water molecules explains the majority of peaks in the MS/MS spectra. Having established the dominant neutral loss in HCD fragmentation at the standard (and optimal) collision energy of 25% to be the loss of three water molecules, we subsequently routinely included this neutral loss in the search for glycated peptides.

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

Journal of Proteome Research Technical Note

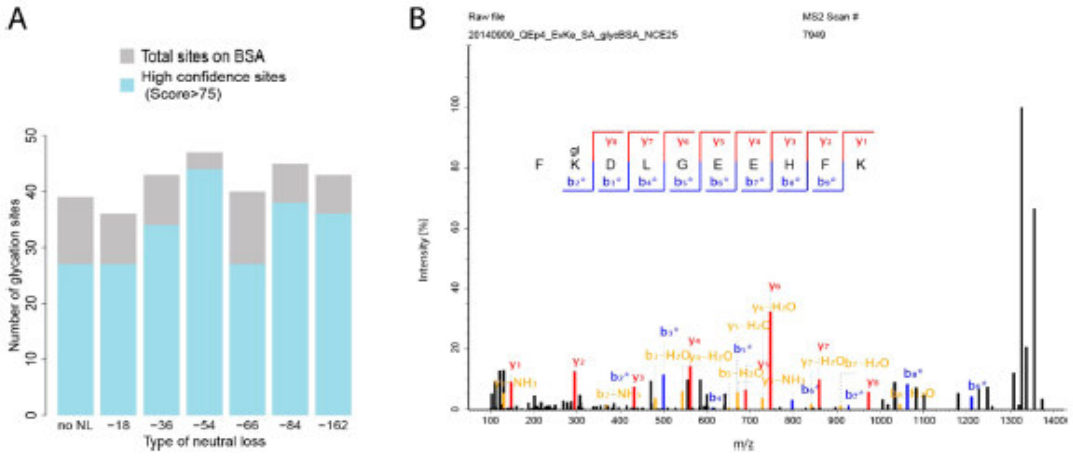


Figure 3. Evaluation of the different neutral losses. (A) Number of glycation sites identified in seven different MaxQuant runs of the same data file with no neutral loss (no NL), neutral loss of H_2O (−18 Da), two H_2O (−36 Da), three H_2O (−54 Da), CH_6O_3 (−66 Da), CH_8O_4 (−84 Da), and of the entire Amadori compound (−162 Da). (B) Same spectrum as in Figure 2A,C now annotated with an almost complete b-ion series due to integrating the neutral loss of three water molecules in the database search. Asterisks on the b-ions indicate that they carry the residual modification after neutral loss of H_6O_3 (standard annotation feature in the MaxQuant viewer).

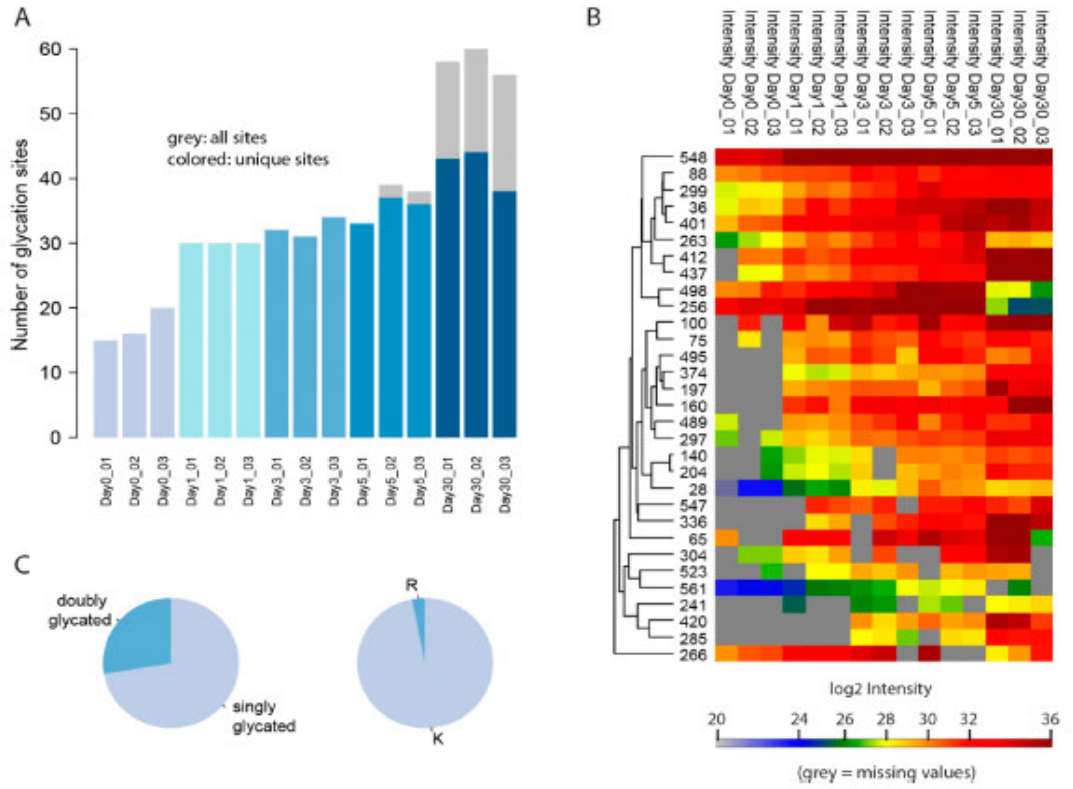


Figure 4. Time dependency of the in vitro glycation reaction. (A) Number of identified glycation sites in triplicate analysis of BSA in vitro glycated with 1 M glucose for 1−30 days. (B) Heatmap of intensities of those glycated lysine sites with >50% valid values over the course of the experiment. (C) Analysis of multiplicity and residue of all glycation sites identified on day 30.

Applying the three water neutral loss analysis to our previous analysis of in vitro glycated HSA increased the number of unique glycation sites from 45 to 54. Among those are 50 lysine residues, meaning that a remarkable 85% of all lysine residues in the mature HSA sequence can be glycated in vitro. This can be explained by the fact that lysine as a charged amino acid is

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

Journal of Proteome Research Technical Note

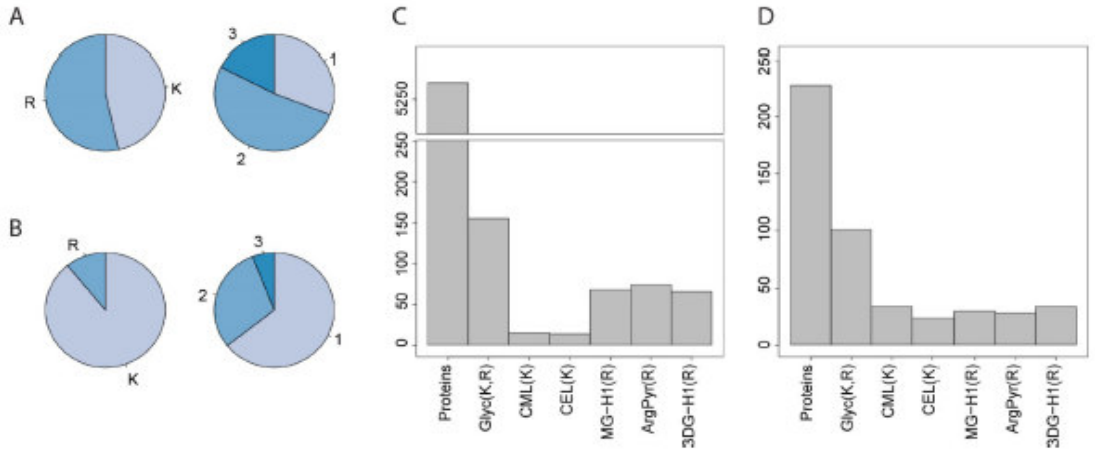


Figure 5. Properties of glycated peptides and AGE analysis. (A) Analysis of glycated peptides identified in HeLa lysate, showing the preferred site of glycation and their multiplicity i.e. whether identified peptides were glycated one, two or three times. (B) Same analysis as in (A) for glycated peptides identified in blood plasma. (C) Barplot depicting the number of proteins, glycation sites and some major AGE sites identified in the HeLa samples. (D) Barplot depicting the number of proteins, glycation sites and some major AGE sites identified in the blood plasma samples.

typically surface exposed. The number of glycated arginines went up from three to four, and the additional site at R184 has been reported before (see Supplementary Table S-1). Regarding the previously reported in vivo glycation sites, we now additionally identify K305; however, we still find no evidence of glycated K463.

Assessing the effect of including the dominant neutral loss on the collision energy evaluation, we found that an NCE of 25% now resulted in the most BSA glycation sites, and the total number of sites increased from 43 to 60 (Supplementary Figure S-1A). The best localization was now obtained with an NCE of 30%, while the highest score was clearly obtained with an NCE of 25% (Supplementary Figure S-1B,C). Thus, the overall optimal collision energy should be between 25 and 30%. Considering that 25% is the optimal setting for unmodified peptides and hence peptide backbone fragmentation and that localization of the glycation site is generally not problematic, we recommend an NCE of 25%, as also optimal for fragmenting glycated peptides, provided the neutral loss of H_6O_3 is taken into account. (Optimal NCEs depend slightly on the specific model, and we find an NCE of 27% to be optimal for glycated and nonglycated peptides on the Q Exactive HF ref 41.)

Time Dependency of Protein Glycation

To investigate the increase in protein glycation over time in vitro, we incubated BSA with 1 M glucose for 0−30 days because after 30 days the equilibrium of the reaction forming the Amadori product should have been reached.[42] Samples were analyzed in triplicate for glycation on K and R, allowing for a neutral loss of H_6O_3 and without matching between runs. Interestingly, our results revealed some glycation events already on the purchased BSA before in vitro incubation with glucose. These are presumably in vivo glycations that have remained stably associated with the protein after purification from bovine blood, processing, and storage. We identified 11 such sites in all three replicates: K28, K36, K88, K256, K263, K266, K299, K401, K498, K548, and K561. (Note that if comparing BSA to HSA sites there is a plus one difference in amino acid position starting from position 140.) Two of these (K36 and K256) correspond to known HSA in vivo sites. With longer incubation the number of detected glycation sites increased substantially (Figure 4A). The Figure shows a near-doubling of detected sites already after 1 day. This means we are initially detecting the Schiff base adduct because several days are needed to convert the Schiff base to the more stable Amadori product.[43] A clear quantitative increase in glycation over time becomes apparent (Figure 4B). The time course analysis reveals an important challenge for quantification of protein glycation: Several sites suddenly drop in intensity on day 30 after gradually increasing before, due to the appearance of doubly glycated species. For example, the sites K256 and K263 are detected as singly glycated peptides until day 5, but at day 30 the doubly glycated peptide AEFVEVTK(gl)LVTDLTK(gl)-VHK that contains both sites appears. Because the intensity of the doubly modified peptide is reported separately, the intensity of the two individual sites goes down (see Figure 4B). Consistent with this, we detected a substantial number of doubly glycated peptides on day 30 (Figure 4C), while none had been detected on day 0. The increase in doubly glycated peptides stems from the fact that glycation inhibits tryptic cleavage. On day 30, almost all sites were still found on lysine, so longer incubation times do not influence the amino acid preferences (Figure 4C).

Analyzing Protein Glycation in Cell Lysate and Blood Plasma

To evaluate the feasibility of detecting glycated peptides in a complex matrix without applying any enrichment step, we chose HeLa lysate as a first test matrix. Because glucose concentrations under standard cell culture conditions are already around five times higher than the physiological concentrations in the body (4.5 mg/mL glucose vs 0.75 to 1.15 mg/mL in normal human blood[44]), we chose to not further expose the cells to glucose. HeLa lysates were trypsin-digested in four workflow replicates, measured in single-shot 2 h measurements on a Q Exactive HF and analyzed for glycation as described before with matching between runs. Even in the absence of any enrichment, we identified 155 glycation sites on 94 different proteins, with a mean localization probability of 0.95. Surprisingly, and in stark contrast with our model plasma

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

proteins, the most frequently modified amino acid was arginine (83 sites) and not lysine (72 sites) (Figure 5A). This indicates that in an intracellular system arginine and lysine are about equally reactive as targets for glycation by glucose. Mean identification scores of the two classes of modification were nearly identical, and manual inspection of the spectra likewise did not reveal marked differences.

We next investigated the possible formation of AGEs in the HeLa proteome. Intracellularly, AGEs may not form by reaction with glucose and via the Amadori product but instead by direct reactions with glucose metabolites.[45] Therefore, we additionally included some major in vivo AGEs derived from glyoxal, methylglyoxal, or 3-deoxyglucosone into the analysis: carboxymethyllysine (CML), carboxyethyllysine (CEL), methylglyoxal-derived hydroimidazolone (MG-H, on arginine), argpyrimidine (on arginine), and 3-deoxyglucosone-derived hydroimidazolone (3DG-H, on arginine). We indeed found many sites for all of those AGEs and interestingly detected about 5 times more arginine AGEs than lysine AGEs (see Figure 5B). This is consistent with what we found for early glycation and with the fact that methylglyoxal is more reactive toward arginine than lysine.[46] Unexpectedly, argpyrimidine was the most common AGE, even though its half-life under physiological conditions has been reported to be shorter than that of MG-H1 (2−9 days vs 2−6 weeks).[47] All HeLa glycation and AGE sites are listed in Supplementary Table S-2. For peptide-centric tables on the HeLa and also the plasma and whole blood data set, see Supplementary Table S-5.

We next went on to test our method on human blood plasma. Exploiting the high scan speed of the Q Exactive HF, we set out to detect glycation sites directly from less than a single drop of human plasma, without depletion of high abundance proteins, peptide fractionation, or enrichment of glycated peptides. We performed the plasma analysis in three technical replicates and analyzed the purified peptides in 100 min gradients using a Top15 method. This yielded 101 glycation sites located on 53 proteins. Similar numbers were obtained in a 2008 study using immunodepletion and boronate affinity enrichment, however, with 5000 times the input material and substantially longer sample processing times.[48] The protein carrying the most glycation sites was albumin with 16 sites, 11 of which were identified with very high localization scores (>0.99): K36, K44, K161, K214, K223, K249, K257, K375, K402, K549, and K598. Although identified in a direct and relatively straightforward analysis in normal human blood, three of these sites have not been reported to be glycated before according to UniProt (see Table 1). Many other typical plasma proteins were found to be glycated, among them apolipoprotein A1 (8 sites), alpha-1-antitrypsin (4 sites), serotransferrin (3 sites), fibrinogen alpha and beta chain (1 site each), and interestingly, many antibody chains. Overall, the plasma glycation sites had a mean localization probability of 0.95 and a mean absolute mass error of only 0.12 ppm (Supplementary Table S-3). The vast majority of glycations in plasma was found to be located on lysine (90 vs 11 sites; Figure 5C). This was similar to what we observed on the model proteins before but very different from the glycated HeLa proteins (see Figure 5A). Furthermore, while in the cell lysate the majority of the peptides were glycated twice, in plasma the majority of the peptides carried only one glycation. We also searched the plasma samples for the five AGEs mentioned above and found at least 20 sites for each of them, with CML and 3DG-H1 being the most abundant AGEs at 34 sites each (see Supplementary Table S-3). In contrast with HeLa cells, lysine and arginine AGEs were similarly abundant in plasma (Figure 5D).

In a final experiment, we measured whole human blood with all cellular components. Thus, it includes the hemoglobin beta-chain (HBB) and its glycation site on the N-terminal valine, which is clinically used to determine the HbA1c value from which diabetes can be diagnosed. We digested and measured whole blood as described before for plasma and analyzed the resulting samples for glycation on valine as well as on lysine and arginine (always including the neutral loss of H_6O_3). We indeed clearly identified the modified valine in position two of HBB (N-terminal position when considering the loss of the initiating methionine) on the easily detectable peptide V*HLTPEEK. Additionally, we found four of the five known lysine glycation sites on HBB as well as two additional sites that have not been reported before. We also detected all four known lysine glycation sites on the hemoglobin alpha chain (HBA) plus two additional ones (see Supplementary Table S-4 for all hemoglobin sites).

■ CONCLUSIONS AND OUTLOOK

Blood plasma is one of the most challenging proteomes, spanning more than 10 orders of magnitude in abundance from the highest to the lowest known plasma protein. Furthermore, PTMs on plasma proteins add another layer of complexity to the inherently intricate plasma proteome. Previous investigations of glycated plasma proteins had relied on extensive sample fractionation, enrichment of glycated peptides, and different peptide fragmentation methods.

In the context of our interest in diabetes, we here asked if modern benchtop Orbitrap platforms are capable of the analysis of glycated peptides in plasma. This would be particularly attractive if it could be incorporated into a routine and robust workflow for plasma proteomics.[33]

We evaluated the fragmentation behavior of glycated peptides and found that HCD−MS/MS scans with the standard collision energy also used by us in proteome measurements are very well-suited for identifying and localizing glycation sites. This requires that the prevailing neutral loss of H_6O_3 is taken into account. In this way, we developed a straightforward workflow to detect glycated peptides directly from blood plasma without applying time-consuming depletion, fractionation, or enrichment steps. We additionally screened for several well-known AGEs and found that they can also be efficiently detected from plasma. Our study demonstrates that straightforward plasma proteome analysis can identify early and advanced protein glycation in this challenging body fluid as part of the routine plasma proteome profiling workflow. Together, this successfully establishes HCD fragmentation for the investigation of protein glycation in general and early glycation in particular.

It may be interesting to determine the reasons for the marked differences in the glycation behavior of intracellular proteomes and the plasma proteome, in particular, the overwhelming preference for lysine over arginine glycation in plasma in contrast with equal occurrence in the cellular proteome.

In the future, we plan to implement a quantification strategy for glycated peptides from patient material because this would allow us to directly assess the level of blood sugar control in any individual in a proteomic study. Clearly, this would be very challenging with label-free methods because of the required

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

accuracy: Normal HbA1c values of below 5.7% need to be robustly distinguished from the prediabetic range (5.7 to 6.4%) and diabetic values of >6.5% (values according to the World Health Organization report on the use of HbA1c in the diagnosis of diabetes, 2011). We envision the use of isotopic labels that can be introduced into patient material via chemical-labeling strategies, such as iTRAQ or TMT; however, ratio compression, which can occur with these techniques, would not be clinically acceptable, and additional challenges connected to the fact that trypsin or LysC do not cleave at glycated lysine residues will have to be overcome.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00464.

Table S-1: Additional HSA sites after reanalysis. Figure S-1: NCE evaluation on HSA after reanalysis. (PDF)
Table S-2: HeLa glycation and AGE sites. (XLSX)
Table S-3: Plasma glycation and AGE sites. (XLSX)
Table S-4: Hemoglobin glycation sites. (XLSX)
Table S-5: Glycated peptides from HeLa, plasma, and whole blood. (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mmann@biochem.mpg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Gaby Sowa, Igor Paron, and Korbinian Mayr for technical assistance. We thank Jürgen Cox and Richard Scheltema for advice regarding data analysis. This work was supported by the Max-Planck Society for the Advancement of Science.

■ ABBREVIATIONS

3DG-H, 3-deoxyglucosone-derived hydroimidazolone; ACN, acetonitrile; AGE, advanced glycation end-product; BAC, boronate affinity chromatography; BSA, bovine serum albumin; CAA, chloroacetamide; CEL, carboxyethyllysine; CID, collision-induced dissociation; CML, carboxymethyllysine; DTT, dithiothreitol; ETD, electron-transfer dissociation; FDR, false discovery rate; HBA, hemoglobin alpha chain; HbA1c, clinical parameter, glycation on the N-terminal valine of the hemoglobin beta chain; HBB, hemoglobin beta chain; HCD, higher-energy collisional dissociation; HPLC, high-pressure liquid chromatography; HSA, human serum albumin; IAA, iodoacetamide; IDF, International Diabetes Federation; LC−MS/MS, liquid chromatography tandem mass spectrometry; MG-H, methylglyoxal-derived hydroimidazolone; MS, mass spectrometry; NCE, normalized collision energy; PTM, post-translational modification; SDB-RPS, poly(styrenedivinylbenzene) reversed-phase sulfonate; SDC, sodiumdeoxycholate; TFA, trifluoroacetic acid

■ REFERENCES

(1) Maillard, L. C. Action of amino acids on sugars. Formation of melanoids in a methodical way. *Compte Rendu* **1912**, *154*, 66−68.
(2) Bunn, H. F.; Gabbay, K. H.; Gallop, P. M. The glycosylation of hemoglobin: relevance to diabetes mellitus. *Science* **1978**, *200* (4337), 21−7.
(3) Thornalley, P. J.; Battah, S.; Ahmed, N.; Karachalias, N.; Agalou, S.; Babaei-Jadidi, R.; Dawnay, A. Quantitative screening of advanced glycation endproducts in cellular and extracellular proteins by tandem mass spectrometry. *Biochem. J.* **2003**, *375* (Pt 3), 581−92.
(4) Bunn, H. F.; Haney, D. N.; Kamin, S.; Gabbay, K. H.; Gallop, P. M. The biosynthesis of human hemoglobin A1c. Slow glycosylation of hemoglobin in vivo. *J. Clin. Invest.* **1976**, *57* (6), 1652−9.
(5) Koenig, R. J.; Peterson, C. M.; Jones, R. L.; Saudek, C.; Lehrman, M.; Cerami, A. Correlation of glucose regulation and hemoglobin A1c in diabetes mellitus. *N. Engl. J. Med.* **1976**, *295* (8), 417−20.
(6) Zhang, Q.; Ames, J. M.; Smith, R. D.; Baynes, J. W.; Metz, T. O. A perspective on the Maillard reaction and the analysis of protein glycation by mass spectrometry: probing the pathogenesis of chronic disease. *J. Proteome Res.* **2009**, *8* (2), 754−69.
(7) Wang, J.; Yan, G.; Qiao, Y.; Wang, D.; Ma, G.; Tang, C. Different levels of glycosylated hemoglobin influence severity and long-term prognosis of coronary heart disease patients with stent implantation. *Exp. Ther. Med.* **2014**, *9* (2), 361−366.
(8) Ulrich, P.; Cerami, A. Protein glycation, diabetes, and aging. *Recent Prog. Horm. Res.* **2001**, *56*, 1−22.
(9) Pamplona, R.; Naudi, A.; Gavin, R.; Pastrana, M. A.; Sajnani, G.; Ilieva, E. V.; Del Rio, J. A.; Portero-Otin, M.; Ferrer, I.; Requena, J. R. Increased oxidation, glycoxidation, and lipoxidation of brain proteins in prion disease. *Free Radical Biol. Med.* **2008**, *45* (8), 1159−66.
(10) Stirban, A.; Gawlowski, T.; Roden, M. Vascular effects of advanced glycation endproducts: Clinical effects and molecular mechanisms. *Mol. Metab.* **2014**, *3* (2), 94−108.
(11) Thorpe, S. R.; Baynes, J. W. Role of the Maillard reaction in diabetes mellitus and diseases of aging. *Drugs Aging* **1996**, *9* (2), 69−77.
(12) Ahmed, N.; Thornalley, P. J. Advanced glycation endproducts: what is their relevance to diabetic complications? *Diabetes, Obes. Metab.* **2007**, *9* (3), 233−45.
(13) Goh, S. Y.; Cooper, M. E. Clinical review: The role of advanced glycation end products in progression and complications of diabetes. *J. Clin. Endocrinol. Metab.* **2008**, *93* (4), 1143−52.
(14) Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002**, *1* (11), 845−67.
(15) Baker, E. S.; Liu, T.; Petyuk, V. A.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Anderson, G. A.; Smith, R. D. Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med.* **2012**, *4* (8), 63.
(16) Omenn, G. S. Exploring the human plasma proteome. *Proteomics* **2005**, *5* (13), 3223−3225.
(17) Misciagna, G.; Michele, G.; Trevisan, M. Non enzymatic glycated proteins in the blood and cardiovascular disease. *Curr. Pharm. Des.* **2007**, *13* (36), 3688−95.
(18) Duncan, B. B.; Heiss, G. Nonenzymatic glycosylation of proteins–a new tool for assessment of cumulative hyperglycemia in epidemiologic studies, past and future. *Am. J. Epidemiol* **1984**, *120* (2), 169−89.
(19) Kim, K. J.; Lee, B. W. The roles of glycated albumin as intermediate glycation index and pathogenic protein. *Diabetes Metab J.* **2012**, *36* (2), 98−107.
(20) Doll, S.; Burlingame, A. L. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem. Biol.* **2015**, *10* (1), 63−71.
(21) Gould, B. J.; Hall, P. M. m-Aminophenylboronate affinity ligands distinguish between nonenzymically glycosylated proteins and glycoproteins. *Clin. Chim. Acta* **1987**, *163* (2), 225−30.
(22) Zhang, Q.; Tang, N.; Brock, J. W.; Mottaz, H. M.; Ames, J. M.; Baynes, J. W.; Smith, R. D.; Metz, T. O. Enrichment and analysis of nonenzymatically glycated peptides: boronate affinity chromatography coupled with electron-transfer dissociation mass spectrometry. *J. Proteome Res.* **2007**, *6* (6), 2323−30.

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

(23) Zhang, Q.; Monroe, M. E.; Schepmoes, A. A.; Clauss, T. R.; Gritsenko, M. A.; Meng, D.; Petyuk, V. A.; Smith, R. D.; Metz, T. O. Comprehensive identification of glycated peptides and their glycation motifs in plasma and erythrocytes of control and diabetic subjects. *J. Proteome Res.* **2011**, *10* (7), 3076−88.
(24) Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M. Plasma proteome profiling to assess human health and disease. *Cell Systems* **2016**, *2*, 185.
(25) Wells, J. M.; McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* **2005**, *402*, 148−85.
(26) Lapolla, A.; Fedele, D.; Reitano, R.; Arico, N. C.; Seraglia, R.; Traldi, P.; Marotta, E.; Tonani, R. Enzymatic digestion and mass spectrometry in the study of advanced glycation end products/peptides. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (4), 496−509.
(27) Frolov, A.; Hoffmann, P.; Hoffmann, R. Fragmentation behavior of glycated peptides derived from D-glucose, D-fructose and D-ribose in tandem mass spectrometry. *J. Mass Spectrom.* **2006**, *41* (11), 1459−69.
(28) Priego-Capote, F.; Scherl, A.; Muller, M.; Waridel, P.; Lisacek, F.; Sanchez, J. C. Glycation isotopic labeling with 13C-reducing sugars for quantitative analysis of glycated proteins in human plasma. *Mol. Cell. Proteomics* **2010**, *9* (3), 579−92.
(29) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (26), 9528−33.
(30) Zhang, Q.; Frolov, A.; Tang, N.; Hoffmann, R.; van de Goor, T.; Metz, T. O.; Smith, R. D. Application of electron transfer dissociation mass spectrometry in analyses of non-enzymatically glycated peptides. *Rapid Commun. Mass Spectrom.* **2007**, *21* (5), 661−6.
(31) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709−12.
(32) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896−906.
(33) Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M. Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* **2016**, *2* (3), 185−195.
(34) 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* **2014**, *11* (3), 319−24.
(35) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−72.
(36) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10* (4), 1794−805.
(37) Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P. A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H. J.; Albar, J. P.; Martinez-Bartolome, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−6.
(38) Rondeau, P.; Bourdon, E. The glycation of albumin: structural and functional impacts. *Biochimie* **2011**, *93* (4), 645−58.
(39) Priego-Capote, F.; Ramirez-Boo, M.; Finamore, F.; Gluck, F.; Sanchez, J. C. Quantitative analysis of glycated proteins. *J. Proteome Res.* **2014**, *13* (2), 336−47.
(40) Neuhauser, N.; Michalski, A.; Cox, J.; Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Mol. Cell. Proteomics* **2012**, *11* (11), 1500−9.
(41) Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell. Proteomics* **2014**, *13* (12), 3698−708.
(42) Mortensen, H. B.; Christophersen, C. Glucosylation of human haemoglobin a in red blood cells studied in vitro. Kinetics of the formation and dissociation of haemoglobin A1c. *Clin. Chim. Acta* **1983**, *134* (3), 317−26.
(43) Higgins, P. J.; Bunn, H. F. Kinetic analysis of the nonenzymatic glycosylation of hemoglobin. *J. Biol. Chem.* **1981**, *256* (10), 5204−8.
(44) Kratz, A.; Ferraro, M.; Sluss, P. M.; Lewandrowski, K. B. Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Laboratory reference values. *N. Engl. J. Med.* **2004**, *351* (15), 1548−63.
(45) Johansen, M. B.; Kiemer, L.; Brunak, S. Analysis and prediction of mammalian protein glycation. *Glycobiology* **2006**, *16* (9), 844−53.
(46) Rabbani, N.; Thornalley, P. J. The critical role of methylglyoxal and glyoxalase 1 in diabetic nephropathy. *Diabetes* **2014**, *63* (1), 50−2.
(47) Sousa Silva, M.; Gomes, R. A.; Ferreira, A. E.; Ponces Freire, A.; Cordeiro, C. The glyoxalase pathway: the first hundred years··· and beyond. *Biochem. J.* **2013**, *453* (1), 1−15.
(48) Zhang, Q.; Tang, N.; Schepmoes, A. A.; Phillips, L. S.; Smith, R. D.; Metz, T. O. Proteomic profiling of nonenzymatically glycated proteins in human plasma and erythrocyte membranes. *J. Proteome Res.* **2008**, *7* (5), 2025−32.

DOI: 10.1021/acs.jproteome.6b00464
J. Proteome Res. 2016, 15, 2881−2890

4. Discussion

The majority of diagnostic decisions are made on the basis of blood-based tests, and protein measurements are prominent among them. However, current assays are restricted to individual proteins, whereas it would be much more desirable to measure all of them in an unbiased, hypothesis-free manner. Therefore, characterization of the plasma proteome by mass spectrometry holds great promise for a new era of biomarker research. However, MS-based proteomics has fallen short of the great expectations that were initially placed in it.

This is mainly due to the tremendous technological challenges in the analysis of the plasma proteome, in which abundance differences between different proteins are more than ten orders of magnitude. In the past, researchers have followed a 'triangular workflow' in which a small cohort is measured in the discovery phase in great depth and differentially expressed proteins are followed up by targeted techniques in a verification and a validation phase (Rifai et al., 2006). The depth of proteome coverage aimed for in the first phase necessitated extensive fractionation and depletion, severely compromising throughput and quantitative fidelity and consequently no biomarkers have yet emerged from these approaches (Bellei et al., 2011; Keshishian et al., 2015; Li et al., 2013; Tu et al., 2010).

In my PhD thesis I questioned the dogmas of current plasma proteomics and biomarker research. First, we set out to radically redesign proteomic workflows with a view to make them truly applicable to the analysis of the plasma proteome. We reduced analysis steps to a minimum by eliminating many of them entirely and streamlining others. This resulted in a rapid, reproducible and very robust pipeline, allowing the automated preparation and measurement of hundreds of plasma samples.

On the basis of this new workflow, we break with previous concepts and introduce a 'rectangular strategy' in which large cohorts are already explored in the first phase in as great a proteomics depths as is compatible with uncompromised throughput. An independent cohort should be analyzed at the same time and the set of proteins that are concordant in both sets qualify as 'verified biomarkers', ready for the validation stage. Using large cohorts already in this discovery phase should have a much higher probability to report true biomarker candidates for further investigation. We call our concept 'Plasma Proteome Profiling' and believe that it has the potential to transform the discovery of new disease indicators.

A basis of Plasma Proteome Profiling is that we aim for the lowest possible variation in our workflow, so as to identify even small fold changes between different study conditions. For this purpose, our group developed and combined several concepts and technological breakthroughs for the highly reproducible quantification of as many proteins in as many samples as possible. The major challenge of Plasma Proteomic Profiling lies in reaching an adequate depth of proteome coverage. As detailed below, we started with only a few hundred proteins but have now broken through the 1,000 protein barrier in single-run, triplicate analyses.

In our first publication, we described our concept by phenotyping a small cohort (Geyer et al., 2016a). We demonstrate that undepleted plasma from a single finger prick (5 µL of blood) provides ample material for Plasma Proteome Profiling. I also developed 'quality marker panels' that allow the assessment of any cohort as well as individual samples. Subsequently, I demonstrated the broad applicability of Plasma Proteome Profiling by analyzing the largest number of plasma samples so far. The investigated study was concerned with the effects of weight loss but also provided us a treasure-trove of new knowledge about the plasma proteome in general (Geyer et al., 2016b). Recently, we implemented novel approaches such as deep peptide libraries and BoxCar scans (patterned fill scans with high dynamic range), which resulted in unprecedented proteomic depth of undepleted plasma in a bariatric surgery study (Albrechtsen et al., 2017; Kulak et al., 2017; Meier et al., 2017). In total, the bariatric surgery study quantified 1,438 proteins in a cohort of 47 individuals, with signal intensities ranging across seven orders of magnitude. Our throughput is now reasonable for medium sized studies and results are quantitatively accurate with around 1,000 plasma proteins per sample.

To illustrate the potential of such a proteome depth, I matched this dataset with a list of 169 approved biomarkers with known concentrations. This revealed a highly unequal distribution of biomarkers across the abundance range: 21% of the proteins within the 300 most abundant proteins were biomarkers and only 4% of the next 1,100 proteins (Figure 13 A). As there should be no physiological reason that biomarkers must be of high abundance, this observation raises the hope that there may be many yet undiscovered biomarkers that are accessible to our technology.

Repeated application of Plasma Proteomics Profiling in many projects and cohorts will generate an extensive amount of information that can be used to construct a universal knowledge base of the plasma proteome. In a new departure for biomarker research, such a knowledge could be data-mined to reveal connections between proteins, to evaluate the value of a biomarker candidate and even to phenotype humans. Apart from the required throughput, one bottleneck of this concept is the availability of high quality,

large-scale studies. In general, it is already difficult to obtain access to one cohort for one disease. Finding a second or third cohort with a suitable design is possible, but in our experience requires time and effort and also involves reconciliation of the interests of the different partners and stakeholders.

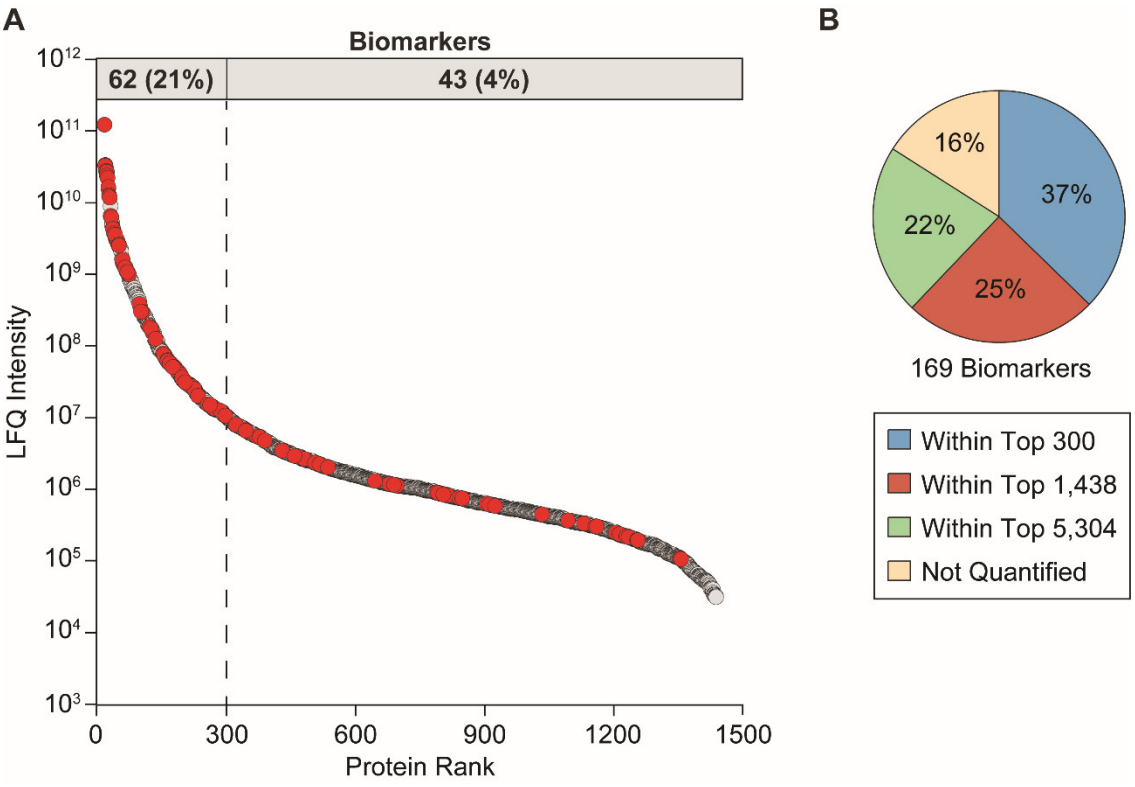


Figure 13: Biomarker coverage and potential for biomarker discovery of MS-based proteomics. (A) The average MS-intensities of nearly 1,500 proteins, which were quantified in one of our Plasma Proteome Profiling projects. Biomarkers (red dots) were annotated using information from our collaboration partners at the Institute of Laboratory Medicine of the Ludwig-Maximilians-University and a list from (Anderson, 2010). The percentage of biomarkers in the indicated abundance rank is provided. **(B)** Pie chart of the percentage of biomarkers in different abundance regions. The data for the Top 300 and Top 1,438 proteins were generated from the above mentioned dataset and the percentage of biomarkers in the Top 5,300 proteins was determined using the data from (Keshishian et al., 2015).

Of all biomarkers, 37% are within the first 300 proteins, 25% in the next 1,100 and only 22% in the following 4,100 proteins (Figure 13 B). A total of 16% of all approved biomarkers were not within the 5,300 plasma proteins reported by proteomics, presumably because they are only increased in specific diseases or because they are exceedingly low abundant.

In the future, we aim to further develop our workflow to reach greater proteomic depth and even higher throughput. For instance, we can now target eluting peptides in real time and at a large scale, an ability that is further helped by applying BoxCar windows.

In a plasma proteome of 2,000 proteins, targeting of three peptides for each protein would result in having to track 6,000 peptides. This appears to be within the technological possibilities of our instruments and software, especially when fractionation is employed. A main advantage of this strategy would be the guaranteed acquisition of MS^2 spectra for multiple peptides per proteins, allowing very high confidence in protein identification. 'Global targeting' could also be useful to individually adjust ion fill times and optimized collision energies on a peptide by peptide basis. Several additional strategies – for instance the combination of targeting from BoxCar windows of depleted and fractionated plasma – could result in even deeper libraries and more targetable peptides.

We are also considering to use isobaric mass tagging. Multiplexing could increase our throughput dramatically in single run measurements or alternatively enable fractionation while maintaining high throughput. Currently, we measure triplicates to ensure high accuracy. In a multiplexing approach, samples are directly compared to each other in the same spectrum, potentially obviating the need for these technical replicates. The throughput would be increased by the product of the multiplexing factor and this factor of three, resulting in up to 18-fold to 30-fold higher capacity (6-plex and 10-plex, respectively). With separation into six fractions, the increase would still be 3-fold to 5-fold, with the advantage of much greater proteome depth.

Moreover, we plan to implement a novel, highly robust and reproducible chromatographic set up. This builds on a rapid elution concept (Falkenby et al., 2014), extended to the formation of pre-formed gradients that already contain the sample. This system has very short overhead times as loading and equilibration are done in parallel and therefore has the potential to increase throughput, especially in short LC-runs. The increased utilization rate of our MS instruments would make it attractive to revert to very short gradients with injection-to-injection times of around 30 min. With fractionation, it may become possible to reach a depth of 1,500 plasma proteins by using the combination of the above mentioned developments: separation into 6 fractions and 6-plex labeling would then result in a throughput of up to 50 plasma proteomes per day and instrument (Figure 14). This combination of throughput and depth would surely provide the basis to elevate plasma proteomics to the next level.

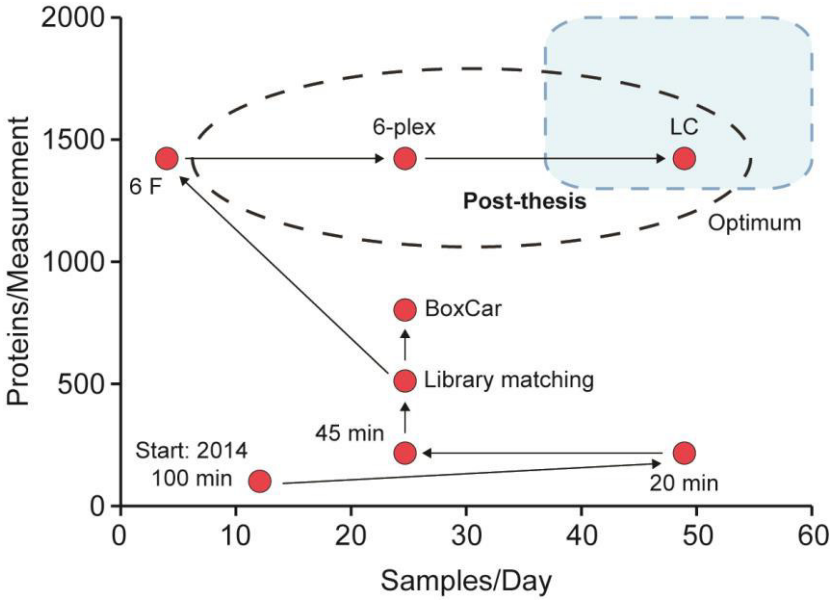


Figure 14: Potential future technological developments in Plasma Proteome Profiling. Throughput and depth of coverage would benefit from fractionation into an optimum of 6 fractions and further 6-plex multiplexing and incorporation of a novel LC concept. This would be one possible path to reach an area where effective biomarker discovery is possible (shaded in blue).

This PhD thesis has paved the way for high throughput screening of clinical cohorts and already delivered the first large-scale proof-of-principle studies. The next step will be to implement the above mentioned technologies, which should make it possible to tackle clinical studies of many disease in high throughput. For this purpose, we have already established contacts to clinicians and researchers to create a pipeline of existing studies that can be analyzed. One of our aims is to measure the first population-based cohorts by plasma proteomics in search of predictive biomarkers and biomarker panels. In the past, such studies could only be assessed for the levels of candidate proteins, but nevertheless have delivered established risk marker like the C-reactive protein (CRP) or low density lipoprotein particle (LDL) (Ridker et al., 2002).

Application of MS-based proteomics directly to patients in the clinic would be the next step for Plasma Proteome Profiling. In this regard, our plan is to initially use SILAC-PrESTs for already established biomarkers for absolute and highly accurate quantification. SILAC-PrESTs can be easily incorporated in our workflow. It is even possible to store them on the StageTip matrix together with the digestion buffer. In a clinical setting, the plasma would be added and the mixture would be automatically processed to peptides.

For some clinical tests time plays a crucial role, which means that samples have to be processed in a sequential manner rather than in batches like in our 96 sample set up.

For this purpose, processing workflows such as those implemented in clinical high throughput platforms can be adapted. In some case, samples have to be analyzed faster than the 3h from sample to result that we have already demonstrated (Geyer et al., 2016a). However, even clinical assays that have to be performed in minutes, such as the troponin tests for myocardial infarction, are not necessarily beyond the reach of proteomics. Immobilized trypsin can in principle digest proteins in a flow-through system and the novel HPLC system alluded to above could start measurements within a few minutes. It is clear that many developments would be necessary to implement proteomics in the clinical laboratory and a complete automatized pipeline would call for partners in industry.

A principal advantage of MS-based proteomics over immunoassays is its ability to multiplex without interference. Immunoassays are constrained by cross-reactions of the antibodies and their interactions with other molecules, which can result in compromised accuracy and severely limits the number of simultaneous assays (Ellington et al., 2010). In contrast, MS-based proteomics allows the simultaneous analysis of as many proteins as desired for biomarker panels. Such panels could be designed to cover markers for differential diagnosis of diseases with similar symptoms as well as quality marker panels to exclude samples of poor quality.

Even proteins that are not included in the SILAC-PrEST mixtures are not entirely lost because they can still be quantified in a label-free manner at lower accuracy. Such data could be blinded and stored together with anonymized patient metadata. Alternatively, if participants have given consent, it would be possible to use a wide range of patient data, including treatment history amongst others, via patient unique IDs. This is already the practice in Denmark where every person has a Civil Personal Registration (CPR) number, which is linked to all their data. This allows ‘big data’ mining for cross-correlations, subclassifications of diseases and identification of potential predictive external factors like drug prescription or life style factors. Such data mining approaches from available clinical data have already been investigated in other contexts (Beck et al., 2016; Ellesoe et al., 2016; Jensen et al., 2014).

A seemingly minor feature of our Plasma Proteome Profiling approach is that it only uses microliter amounts from fingerpricks instead of the large volumes of venous blood that are routinely taken. We already demonstrated that fingerpricks result in highly reproducible quantification and we have identified the few proteins prone to variation due to the finger pricking process (Geyer et al., 2016a). Fingerpicks can be obtained with much higher frequency from adults and are even appropriate for infants. Even dried

blood spots could be analyzed. This would dramatically expand the application of Plasma Proteome Profiling in health and disease.

In summary, this PhD thesis has developed the concept and practice of Plasma Proteome Profiling as a fundamentally new approach in biomarker research and medical diagnostics, leading to a system-wide phenotyping of humans in health and disease.

5. References

Aass, C., Norheim, I., Eriksen, E.F., Thorsby, P.M., and Pepaj, M. (2015). Single unit filter-aided method for fast proteomic analysis of tear fluid. Anal Biochem *480*, 1-5.
Abbatiello, S.E., Schilling, B., Mani, D.R., Zimmerman, L.J., Hall, S.C., MacLean, B., Albertolle, M., Allen, S., Burgess, M., Cusack, M.P.*, et al.* (2015). Large-Scale Interlaboratory Study to Develop, Analytically Validate and Apply Highly Multiplexed, Quantitative Peptide Assays to Measure Cancer-Relevant Proteins in Plasma. Mol Cell Proteomics *14*, 2357-2374.
Addona, T.A., Abbatiello, S.E., Schilling, B., Skates, S.J., Mani, D.R., Bunk, D.M., Spiegelman, C.H., Zimmerman, L.J., Ham, A.J., Keshishian, H.*, et al.* (2009). Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. Nat Biotechnol *27*, 633-641.
Addona, T.A., Shi, X., Keshishian, H., Mani, D.R., Burgess, M., Gillette, M.A., Clauser, K.R., Shen, D.X., Lewis, G.D., Farrell, L.A.*, et al.* (2011). A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. Nature Biotechnology *29*, 635-U119.
Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature *537*, 347-355.
Albrechtsen, N.J.W., Geyer, P.E., Doll, S., Bojsen-Moller, K.N., Martinussen, C., Torekov, S.S., Keilhauer, E., Treit, P.V., Meier, F., Holst, J.J.*, et al.* (2017). Disentangling Effects of Bariatric Suergery on the Human Plasma Proteome. In preparation.
Andersen, J.S., Lam, Y.W., Leung, A.K., Ong, S.E., Lyon, C.E., Lamond, A.I., and Mann, M. (2005). Nucleolar proteome dynamics. Nature *433*, 77-83.
Anderson, N.L. (2010). The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clin Chem *56*, 177-185.
Anderson, N.L., and Anderson, N.G. (2002). The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics *1*, 845-867.
Aye, T.T., Scholten, A., Taouatas, N., Varro, A., Van Veen, T.A., Vos, M.A., and Heck, A.J. (2010). Proteome-wide protein concentrations in the human heart. Mol Biosyst *6*, 1917-1927.
Azimifar, S.B., Nagaraj, N., Cox, J., and Mann, M. (2014). Cell-type-resolved quantitative proteomics of murine liver. Cell Metab *20*, 1076-1087.
Baggerly, K.A., Morris, J.S., and Coombes, K.R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics *20*, 777-785.
Bantscheff, M., Boesche, M., Eberhard, D., Matthieson, T., Sweetman, G., and Kuster, B. (2008). Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. Mol Cell Proteomics *7*, 1702-1713.
Bantscheff, M., Eberhard, D., Abraham, Y., Bastuck, S., Boesche, M., Hobson, S., Mathieson, T., Perrin, J., Raida, M., Rau, C.*, et al.* (2007). Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol *25*, 1035-1044.
Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem *404*, 939-965.
Beck, M.K., Westergaard, D., Jensen, A.B., Groop, L., and Brunak, S. (2016). Temporal Order of Disease Pairs Affects Subsequent Disease Trajectories: The Case of Diabetes and Sleep Apnea. Pac Symp Biocomput *22*, 380-389.
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.*, et al.* (2017). An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Systems.

Bellei, E., Bergamini, S., Monari, E., Fantoni, L.I., Cuoghi, A., Ozben, T., and Tomasi, A. (2011). High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. Amino Acids *40*, 145-156.
Biomarkers Definitions Working, G. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther *69*, 89-95.
Cao, Z., Tang, H.Y., Wang, H., Liu, Q., and Speicher, D.W. (2012). Systematic comparison of fractionation methods for in-depth analysis of plasma proteomes. J Proteome Res *11*, 3090-3100.
Carr, S.A., Abbatiello, S.E., Ackermann, B.L., Borchers, C., Domon, B., Deutsch, E.W., Grant, R.P., Hoofnagle, A.N., Huttenhain, R., Koomen, J.M.*, et al.* (2014). Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. Mol Cell Proteomics *13*, 907-917.
Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top Down proteomics: facts and perspectives. Biochem Biophys Res Commun *445*, 683-693.
Christensen, U., Simonsen, M., Harrit, N., and Sottrupjensen, L. (1989). Pregnancy Zone Protein, a Proteinase-Binding Macroglobulin - Interactions with Proteinases and Methylamine. Biochemistry *28*, 9324-9331.
Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. N Engl J Med *372*, 793-795.
Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol *26*, 1367-1372.
Cox, J., and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem *80*, 273-299.
Cox, J., and Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics *13 Suppl 16*, S12.
Danesh, J., Collins, R., and Peto, R. (2000). Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies. Circulation *102*, 1082-1085.
Deeb, S.J., Tyanova, S., Hummel, M., Schmidt-Supprian, M., Cox, J., and Mann, M. (2015). Machine Learning-based Classification of Diffuse Large B-cell Lymphoma Patients by Their Protein Expression Profiles. Mol Cell Proteomics *14*, 2947-2960.
Ebhardt, H.A., Root, A., Sander, C., and Aebersold, R. (2015). Applications of targeted proteomics in systems biology and translational medicine. Proteomics *15*, 3193-3208.
Edfors, F., Bostrom, T., Forsstrom, B., Zeiler, M., Johansson, H., Lundberg, E., Hober, S., Lehtio, J., Mann, M., and Uhlen, M. (2014). Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins. Mol Cell Proteomics *13*, 1611-1624.
Ellesoe, S.G., Johansen, M.M., Bjerre, J.V., Hjortdal, V.E., Brunak, S., and Larsen, L.A. (2016). Familial Atrial Septal Defect and Sudden Cardiac Death: Identification of a Novel NKX2-5 Mutation and a Review of the Literature. Congenit Heart Dis *11*, 283-290.
Ellington, A.A., Kullo, I.J., Bailey, K.R., and Klee, G.G. (2010). Antibody-based protein multiplex platforms: technical and operational challenges. Clin Chem *56*, 186-193.
Falkenby, L.G., Such-Sanmartin, 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.
FDA-NIH:Biomarker-Working-Group (2016). BEST (Biomarkers, EndpointS, and other Tools) Resource. In (Maryland: Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US)).
Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science *246*, 64-71.

Fischbach, F.D.I., M.B. (2009). A Manual of Laboratory and Diagnostic Tests (Wolters Kluwer, Lippincott Williams and Wilkins).
Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics *11*, M111 014050.
Geyer, P.E., Holdt, L.M., Teupser, D., and Mann, M. (2017). Revisiting Biomarker Discovery by Plasma Proteomics. Molecular Systems Biology *Submitted*.
Geyer, P.E., Kulak, N.A., Pichler, G., Holdt, L.M., Teupser, D., and Mann, M. (2016a). Plasma Proteome Profiling to Assess Human Health and Disease. Cell Syst *2*, 185-195.
Geyer, P.E., Wewer Albrechtsen, N.J., Tyanova, S., Grassl, N., Iepsen, E.W., Lundgren, J., Madsbad, S., Holst, J.J., Torekov, S.S., and Mann, M. (2016b). Proteomics reveals the effects of sustained weight loss on the human plasma proteome. Mol Syst Biol *12*, 901.
Grassl, N., Kulak, N.A., Pichler, G., Geyer, P.E., Jung, J., Schubert, S., Sinitcyn, P., Cox, J., and Mann, M. (2016). Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. Genome Med *8*, 44.
Gundry, R.L., Fu, Q., Jelinek, C.A., Van Eyk, J.E., and Cotter, R.J. (2007). Investigation of an albumin-enriched fraction of human serum and its albuminome. Proteomics Clin Appl *1*, 73-88.
Gundry, R.L., White, M.Y., Nogee, J., Tchernyshyov, I., and Van Eyk, J.E. (2009). Assessment of albumin removal from an immunoaffinity spin column: critical implications for proteomic examination of the albuminome and albumin-depleted samples. Proteomics *9*, 2021-2028.
Hahne, H., Pachl, F., Ruprecht, B., Maier, S.K., Klaeger, S., Helm, D., Medard, G., Wilm, M., Lemeer, S., and Kuster, B. (2013). DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. Nat Methods *10*, 989-991.
Hamm, C.W., Ravkilde, J., Gerhardt, W., Jorgensen, P., Peheim, E., Ljungdahl, L., Goldmann, B., and Katus, H.A. (1992). The prognostic value of serum troponin T in unstable angina. N Engl J Med *327*, 146-150.
Hassis, M.E., Niles, R.K., Braten, M.N., Albertolle, M.E., Ewa Witkowska, H., Hubel, C.A., Fisher, S.J., and Williams, K.E. (2015). Evaluating the effects of preanalytical variables on the stability of the human plasma proteome. Anal Biochem *478*, 14-22.
Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F.*, et al.* (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712-723.
Hein, M.Y., Sharma, K., Cox, J., and Mann, M. (2013). Handbook of Systems Biology: Concepts and Insights: Proteomic Analysis of Cellular Systems. Academic Press, 3-25.
Holewinski, R.J., Jin, Z., Powell, M.J., Maust, M.D., and Van Eyk, J.E. (2013). A fast and reproducible method for albumin isolation and depletion from serum and cerebrospinal fluid. Proteomics *13*, 743-750.
Hoofnagle, A.N., and Wener, M.H. (2009). The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. J Immunol Methods *347*, 3-11.
Hoofnagle, A.N., Whiteaker, J.R., Carr, S.A., Kuhn, E., Liu, T., Massoni, S.A., Thomas, S.N., Townsend, R.R., Zimmerman, L.J., 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.
Humphrey, S.J., Azimifar, S.B., and Mann, M. (2015). High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. Nat Biotechnol *33*, 990-995.
Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H.*, et al.* (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505-509.
Itzhak, D.N., Tyanova, S., Cox, J., and Borner, G.H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. Elife *5*.

Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesoe, S.G., Eriksson, R., Schmock, H., Jensen, P.B., Jensen, L.J., and Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun *5*, 4022.
Kaisar, M., van Dullemen, L.F., Thezenas, M.L., Zeeshan Akhtar, M., Huang, H., Rendel, S., Charles, P.D., Fischer, R., Ploeg, R.J., and Kessler, B.M. (2016). Plasma degradome affected by variable storage of human blood. Clin Proteomics *13*, 26.
Kamlage, B., Maldonado, S.G., Bethan, B., Peter, E., Schmitz, O., Liebenberg, V., and Schatz, P. (2014). Quality markers addressing preanalytical variations of blood and plasma processing identified by broad and targeted metabolite profiling. Clin Chem *60*, 399-412.
Karas, M., Bachmann, D., and Hillenkamp, F. (1985). Influence of the Wavelength in High-Irradiance Ultraviolet-Laser Desorption Mass-Spectrometry of Organic-Molecules. Analytical Chemistry *57*, 2935-2939.
Kebarle, P., and Tang, L. (1993). From Ions in Solution to Ions in the Gas-Phase - the Mechanism of Electrospray Mass-Spectrometry. Analytical Chemistry *65*, A972-A986.
Keshishian, H., Burgess, M.W., Gillette, M.A., Mertins, P., Clauser, K.R., Mani, D.R., Kuhn, E.W., Farrell, L.A., Gerszten, R.E., and Carr, S.A. (2015). Multiplexed, Quantitative Workflow for Sensitive Biomarker Discovery in Plasma Yields Novel Candidates for Early Myocardial Injury. Mol Cell Proteomics.
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.*, et al.* (2014). A draft map of the human proteome. Nature *509*, 575-581.
Klaeger, S., Gohlke, B., Perrin, J., Gupta, V., Heinzlmeir, S., Helm, D., Qiao, H., Bergamini, G., Handa, H., Savitski, M.M.*, et al.* (2016). Chemical Proteomics Reveals Ferrochelatase as a Common Off-target of Kinase Inhibitors. ACS Chem Biol *11*, 1245-1254.
Kulak, N.A., Geyer, P.E., and M., M. (2017). Loss-less nano-fractionator for high sensitivity, high coverage proteomics. Molecular Cellular Proteomics; Manuscript in Press.
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.
Lahtvee, P.J., Sanchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., and Nielsen, J. (2017). Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst *4*, 495-504 e495.
Larance, M., and Lamond, A.I. (2015). Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol *16*, 269-280.
Li, X.J., Hayward, C., Fong, P.Y., Dominguez, M., Hunsucker, S.W., Lee, L.W., McLean, M., Law, S., Butler, H., Schirm, M.*, et al.* (2013). A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. Sci Transl Med *5*, 207ra142.
Liu, Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., Cheng, L.Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T.D.*, et al.* (2015). Quantitative variability of 342 plasma proteins in a human twin population. Mol Syst Biol *11*, 786.
Lombardi, G., Lanteri, P., Colombini, A., and Banfi, G. (2012). Blood biochemical markers of bone turnover: pre-analytical and technical aspects of sample collection and handling. Clin Chem Lab Med *50*, 771-789.
Lowenthal, M.S., Mehta, A.I., Frogale, K., Bandle, R.W., Araujo, R.P., Hood, B.L., Veenstra, T.D., Conrads, T.P., Goldsmith, P., Fishman, D.*, et al.* (2005). Analysis of albumin-associated peptides and proteins from ovarian cancer patients. Clin Chem *51*, 1933-1945.
Luque-Garcia, J.L., and Neubert, T.A. (2007). Sample preparation for serum/plasma profiling and biomarker identification by mass spectrometry. J Chromatogr A *1153*, 259-276.
Ly, T., Ahmad, Y., Shlien, A., Soroka, D., Mills, A., Emanuele, M.J., Stratton, M.R., and Lamond, A.I. (2014). A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. Elife *3*, e01630.

Lynch, S.V., and Pedersen, O. (2016). The Human Intestinal Microbiome in Health and Disease. N Engl J Med *375*, 2369-2379.
Malamud, D. (2011). Saliva as a diagnostic fluid. Dent Clin North Am *55*, 159-178.
Mann, M. (2016). The Rise of Mass Spectrometry and the Fall of Edman Degradation. Clin Chem *62*, 293-294.
Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem *66*, 4390-4399.
Marx, H., Minogue, C.E., Jayaraman, D., Richards, A.L., Kwiecien, N.W., Siahpirani, A.F., Rajasekar, S., Maeda, J., Garcia, K., Del Valle-Echevarria, A.R.*, et al.* (2016). A proteomic atlas of the legume Medicago truncatula and its nitrogen-fixing endosymbiont Sinorhizobium meliloti. Nat Biotechnol *34*, 1198-1205.
McConnell, J.P., Guadagno, P.A., Dayspring, T.D., Hoefner, D.M., Thiselton, D.L., Warnick, G.R., and Harris, W.S. (2014). Lipoprotein(a) mass: a massively misunderstood metric. J Clin Lipidol *8*, 550-553.
Meier, F., Geyer, P.E., Cox, J., and Mann, M. (2017). BoxCar method enables single shot proteomics at a depth of 10,000 proteins in 100 minutes. Under revision.
Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F.*, et al.* (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55-62.
Mischak, H., Allmaier, G., Apweiler, R., Attwood, T., Baumann, M., Benigni, A., Bennett, S.E., Bischoff, R., Bongcam-Rudloff, E., Capasso, G.*, et al.* (2010). Recommendations for biomarker identification and qualification in clinical proteomics. Sci Transl Med *2*, 46ps42.
Munoz, J., and Heck, A.J. (2014). From the human genome to the human proteome. Angew Chem Int Ed Engl *53*, 10864-10866.
Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol Cell Proteomics *11*, M111 013722.
Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol *7*, 548.
Nanjappa, V., Thomas, J.K., Marimuthu, A., Muthusamy, B., Radhakrishnan, A., Sharma, R., Ahmad Khan, A., Balakrishnan, L., Sahasrabuddhe, N.A., Kumar, S.*, et al.* (2014). Plasma Proteome Database as a resource for proteomics research: 2014 update. Nucleic Acids Res *42*, D959-965.
Oberbach, A., Schlichting, N., Neuhaus, J., Kullnick, Y., Lehmann, S., Heinrich, M., Dietrich, A., Mohr, F.W., von Bergen, M., and Baumann, S. (2014). Establishing a reliable multiple reaction monitoring-based method for the quantification of obesity-associated comorbidities in serum and adipose tissue requires intensive clinical validation. J Proteome Res *13*, 5784-5800.
Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A.*, et al.* (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci Signal *3*, ra3.
Omenn, G.S., States, D.J., Adamski, M., Blackwell, T.W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B.B., Simpson, R.J., Eddes, J.S.*, 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.
Paczesny, S., Braun, T.M., Levine, J.E., Hogan, J., Crawford, J., Coffing, B., Olsen, S., Choi, S.W., Wang, H., Faca, V.*, et al.* (2010). Elafin is a biomarker of graft-versus-host disease of the skin. Sci Transl Med *2*, 13ra12.

Parker, C.E., and Borchers, C.H. (2014). Mass spectrometry based biomarker discovery, verification, and validation--quality assurance and control of protein biomarker assays. Mol Oncol *8*, 840-858.
Paulovich, A.G., Whiteaker, J.R., Hoofnagle, A.N., and Wang, P. (2008). The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. Proteomics Clin Appl *2*, 1386-1402.
Percy, A.J., Chambers, A.G., Yang, J., and Borchers, C.H. (2013). Multiplexed MRM-based quantitation of candidate cancer biomarker proteins in undepleted and non-enriched human plasma. Proteomics *13*, 2202-2215.
Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C*., et al.* (2002). Use of proteomic patterns in serum to identify ovarian cancer. Lancet *359*, 572-577.
Pickup, L., Atkinson, S., Hollnagel, E., Bowie, P., Gray, S., Rawlinson, S., and Forrester, K. (2017). Blood sampling - Two sides to the story. Appl Ergon *59*, 234-242.
Powe, C.E., Evans, M.K., Wenger, J., Zonderman, A.B., Berg, A.H., Nalls, M., Tamez, H., Zhang, D., Bhan, I., Karumanchi, S.A*., et al.* (2013). Vitamin D-binding protein and vitamin D status of black Americans and white Americans. N Engl J Med *369*, 1991-2000.
Pugh, R.N., Murray-Lyon, I.M., Dawson, J.L., Pietroni, M.C., and Williams, R. (1973). Transection of the oesophagus for bleeding oesophageal varices. Br J Surg *60*, 646-649.
Qian, W.J., Kaleta, D.T., Petritis, B.O., Jiang, H., Liu, T., Zhang, X., Mottaz, H.M., Varnum, S.M., Camp, D.G., 2nd, Huang, L*., et al.* (2008). Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. Mol Cell Proteomics *7*, 1963-1973.
Rai, A.J., Zhang, Z., Rosenzweig, J., Shih, I., Pham, T., Fung, E.T., Sokoll, L.J., and Chan, D.W. (2002). Proteomic approaches to tumor marker discovery - Identification of biomarkers for ovarian cancer. Archives of Pathology & Laboratory Medicine *126*, 1518-1526.
Rex, D.K., Johnson, D.A., Anderson, J.C., Schoenfeld, P.S., Burke, C.A., Inadomi, J.M., and American College of, G. (2009). American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. Am J Gastroenterol *104*, 739-750.
Richards, A.L., Merrill, A.E., and Coon, J.J. (2015). Proteome sequencing goes deep. Curr Opin Chem Biol *24*, 11-17.
Ridker, P.M., Rifai, N., Rose, L., Buring, J.E., and Cook, N.R. (2002). Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. N Engl J Med *347*, 1557-1565.
Rifai, N., Gillette, M.A., and Carr, S.A. (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol *24*, 971-983.
Sacco, F., Silvestri, A., Posca, D., Pirro, S., Gherardini, P.F., Castagnoli, L., Mann, M., and Cesareni, G. (2016). Deep Proteomics of Breast Cancer Cells Reveals that Metformin Rewires Signaling Networks Away from a Pro-growth State. Cell Syst *2*, 159-171.
Santucci, L., Candiano, G., Petretto, A., Bruschi, M., Lavarello, C., Inglese, E., Righetti, P.G., and Ghiggeri, G.M. (2015). From hundreds to thousands: Widening the normal human Urinome (1). J Proteomics *112*, 53-62.
Satagopan, J.M., Verbel, D.A., Venkatraman, E.S., Offit, K.E., and Begg, C.B. (2002). Two-stage designs for gene-disease association studies. Biometrics *58*, 163-170.
Sharma, K., Schmitt, S., Bergner, C.G., Tyanova, S., Kannaiyan, N., Manrique-Hoyos, N., Kongi, K., Cantuti, L., Hanisch, U.K., Philips, M.A*., et al.* (2015). Cell type- and brain region-resolved mouse brain proteome. Nat Neurosci *18*, 1819-1831.
Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc *1*, 2856-2860.
Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal Chem *68*, 850-858.

Shi, T., Fillmore, T.L., Gao, Y., Zhao, R., He, J., Schepmoes, A.A., Nicora, C.D., Wu, C., Chambers, J.L., Moore, R.J*., et al.* (2013). Long-gradient separations coupled with selected reaction monitoring for highly sensitive, large scale targeted protein quantification in a single analysis. Anal Chem *85*, 9196-9203.
Skates, S.J., Gillette, M.A., LaBaer, J., Carr, S.A., Anderson, L., Liebler, D.C., 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.
Skol, A.D., Scott, L.J., Abecasis, G.R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet *38*, 209-213.
Surinova, S., Schiess, R., Huttenhain, R., Cerciello, F., Wollscheid, B., and Aebersold, R. (2011). On the development of plasma protein biomarkers. J Proteome Res *10*, 5-16.
Tanaka, K.W., H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. (1988). Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of flight Mass Spectrometry. Rapid Commun Mass Spectrom *2 (20)*, 151–153.
Thomas, D., Xie, R., and Gebregziabher, M. (2004). Two-Stage sampling designs for gene association studies. Genet Epidemiol *27*, 401-414.
Thulasiraman, V., Lin, S., Gheorghiu, L., Lathrop, J., Lomas, L., Hammond, D., and Boschetti, E. (2005). Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. Electrophoresis *26*, 3561-3571.
Ting, L., Rad, R., Gygi, S.P., and Haas, W. (2011). MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. Nat Methods *8*, 937-940.
Tu, C., Rudnick, P.A., Martinez, M.Y., Cheek, K.L., Stein, S.E., Slebos, R.J., and Liebler, D.C. (2010). Depletion of abundant plasma proteins and limitations of plasma proteomics. J Proteome Res *9*, 4982-4991.
Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., and Geiger, T. (2016). Proteomic maps of breast cancer subtypes. Nat Commun *7*, 10259.
Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M., Geiger, T., Mann, M., Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods.
Ullrich, A., Coussens, L., Hayflick, J.S., Dull, T.J., Gray, A., Tam, A.W., Lee, J., Yarden, Y., Libermann, T.A., Schlessinger, J*., et al.* (1984). Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. Nature *309*, 418-425.
Utermann, G. (1989). The mysteries of lipoprotein(a). Science *246*, 904-910.
Vihko, P., Sajanti, E., Janne, O., Peltonen, L., and Vihko, R. (1978). Serum prostate-specific acid phosphatase: development and validation of a specific radioimmunoassay. Clin Chem *24*, 1915-1919.
Weinkauf, M., Hiddemann, W., and Dreyling, M. (2006). Sample pooling in 2-D gel electrophoresis: a new approach to reduce nonspecific expression background. Electrophoresis *27*, 4555-4558.
Wild, D. (2013). The immunoassay handbook : theory and applications of ligand binding, ELISA, and related techniques, 4th edn (Oxford ; Waltham, MA: Elsevier).
Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H*., et al.* (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582-587.
Wilm, M., and Mann, M. (1996). Analytical properties of the nanoelectrospray ion source. Anal Chem *68*, 1-8.

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature *379*, 466-469.
Wisniewski, J.R., Zougman, A., and Mann, M. (2009). Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res *8*, 5674-5678.
Wu, H.Y., Goan, Y.G., Chang, Y.H., Yang, Y.F., Chang, H.J., Cheng, P.N., Wu, C.C., Zgoda, V.G., Chen, Y.J., and Liao, P.C. (2015). Qualification and Verification of Serological Biomarker Candidates for Lung Adenocarcinoma by Targeted Mass Spectrometry. J Proteome Res *14*, 3039-3050.
Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M., and Mann, M. (2012). A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. Mol Cell Proteomics *11*, O111 009613.
Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S.*, et al.* (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382-387.
Zhang, X., Ning, Z., Mayne, J., Moore, J.I., Li, J., Butcher, J., Deeke, S.A., Chen, R., Chiang, C.K., Wen, M.*, et al.* (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome *4*, 31.
Zhang, Z., Bast, R.C., Jr., Yu, Y., Li, J., Sokoll, L.J., Rai, A.J., Rosenzweig, J.M., Cameron, B., Wang, Y.Y., Meng, X.Y.*, et al.* (2004). Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res *64*, 5882-5890.
Zhao, X., Qureshi, F., Eastman, P.S., Manning, W.C., Alexander, C., Robinson, W.H., and Hesterberg, L.K. (2012). Pre-analytical effects of blood sampling and handling in quantitative immunoassays for rheumatoid arthritis. J Immunol Methods *378*, 72-80.

6. Acknowledgements

I want to express my deepest gratitude to all the wonderful people that contributed to this work.

First of all, I want to thank my fantastic mentor Prof. Matthias Mann. He is a great inspiration and his enthusiasm guided me on my way to this PhD thesis. Our discussions resulted in the idea of 'Plasma Proteome Profiling' – the proteomic phenotyping of humans from a single finger prick of blood to 'assess health and disease'. Thank you for believing in this idea that has the potential to change medicine to the greater good of mankind.

I also want to thank my fantastic family for their support over the years. They combine the best of the world in a small group of wonderful people. Thank you that you are always there. In particular, I want to thank my Mama being the nicest person on earth, my Papa for being such a cool guy, my brother knowing what really counts, my sister for sharing honey and eggs and of course my grandma for starting with dinner before everyone else had the chance to get a piece.

Ann-Kathrin Jörger just for being the most wonderful human on this planet.

Nils Kulak and Garwin Pichler for being my lab supervisors and for their most valuable advices regarding research, fantastic food, wine and beer. Also thank you guys for so many headaches. All you need is a Spider…

Sean Humphrey and his lovely wife Emily for being such unbelievable kind Australians, for teaching me how not to burn myself with a heated capillary and all other important lessons about mass spectrometers and HPLCs.

When we speak about LC-MS instrumentation, I must say thank you to Korbinian Mayr and Igor Paron. They are amazingly skilled guys that find every piece of dirt in a MS and the most hidden leaks.

Of course, thousand thanks to Gabriele Maria Sowa for being the Column Queen.

Florian Meier for being such a great fellow in sharing beer, black humor, not very smart comments and the most brilliants ideas like cars in boxes, which should be too small for cars.

Sophia Doll for her laughing, her hunger for sweets and to be a real (half real, half French) Bavarian that always supports another Bavarian, whenever he needs help.

Peter V. Treit for being a king amongst kings – a legend. No more comment necessary. Or maybe one: Don´t investigate the microbiome with your eyes.

Lili Niu for the great fun to work with, her great enthusiasm, her self-irony and unbelievable dancing skills.

Lesca Holdt und Daniel Teupser for being more than just collaboration partners. Thank you for your advices, for sharing our dreams and for the wonderful food.

Atul Deshmukh for being a real friend, for all the fun that we have, for his company in sharing so many evenings with beer and wine at so many nice places.

Nicolai Jacob Wewer Albrechtsen for wonderful discussions about obesity and for being a Viking in a Lederhosen who is dreaming of catching a mountain goat once in his life.

Alison Dalfovo, the most organized person in the department that keeps everything running and that takes care of everyone.

Martin Steger for being another very important advisor on questions about sterilization reagents and how to behave in case of intoxication.

Ozge Karayel for bringing all the sweets and tastes of Turkey to Munich.

Katharina Zettl for exhausting hours in front of the centrifuges.

Marco Hein for spreading Burber and Beer around the world and for keeping beers with the original crowd.

Niklas Graßl for his great enthusiasm, awards and real friendship.