Aus dem Max von Pettenkofer-Institut für Hygiene und
Medizinische Mikrobiologie
Institut der Ludwig-Maximilians-Universität München
Lehrstuhl: Medizinische Mikrobiologie und
Krankenhaushygiene
Leitung: Prof. Dr. med. Sebastian Suerbaum

Mass spectrometry based proteomics in medical diagnostics

Dissertation
zum Erwerb des Doktorgrades der Medzin
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von
Niklas Graßl aus
München

Jahr
2019

Mit Genehmigung der Medizinischen Fakultät
der Universität zu München

Berichterstatter:	Prof. Dr. Sören Schubert
	Prof. Dr. Andreas G. Ladurner
Mitberichterstatter:	Prof. Dr. Daniel Teupser
	Prof. Dr. Bianca Schaub
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tag der mündlichen Prüfung:	27.06.2019

Zusammenfassung

Die Messung von Proteinen aus Patientenproben ist von überragender Bedeutung für die medizinische Diagnostik. Neben den klassischen Enzyme-linked Immunosorbent Assays in der Labormedizin findet Proteinanalytik in Form der MALDI-TOF-Massenspektrometrie neuerdings maßgeblichen Einsatz zur molekularen Speziesdifferenzierung in der Mikrobiologie. Die Massenspektrometrie eröffnet dabei vollkommen neue Perspektiven für die medizinische Diagnostik, indem sie die Quantifizierung aller in einer Zelle oder einem Organismus vorkommender Proteine, das sogenannte Proteom, anstrebt. Der technologische Fortschritt der vergangenen Jahre ermöglicht es seit kurzem, ein solches Proteom innerhalb weniger Stunden zu messen.

In dieser Arbeit präsentiere ich die Bestimmung des humanen Speichelproteoms von acht gesunden Individuen zu zwei Tageszeiten mittels Massenspektrometrie basierter Proteomik. Die erworbene Proteindatenbank gewährt umfassende Einblicke in die Proteinzusammensetzung und die Dynamik des menschlichen Speichels und stellt mit mehr als 5000 Proteinen das tiefste bislang gemessene Proteom einer Körperflüssigkeit dar. Große Teile dieses Proteoms lassen sich bereits in einem beschleunigten Verfahren mit nur vier Stunden gesamter Analysezeit erheben und erscheinen deshalb attraktiv zur Nutzung in der klinischen Diagnostik. Dabei wird auch das humane Speichelproteom mit dem humanen Plasmaproteom verglichen und daraus ergibt sich der Schluss, dass die Proteinkonzentrationen zwischen Speichel und Plasma kaum korreliert sind. Damit erscheint es für zahlreiche Proteine nicht möglich, ihre Plasmakonzentrationen anhand der Konzentration im Speichel abzuschätzen.

In einem zweiten Schritt wird die direkte Messung des oralen Mikrobioms mit Massenspektrometrie basierter Proteomik vorgestellt. Im Gegensatz zur derzeit in der Mikrobiologie verwendeten MALDI-TOF-Massenspektrometrie benötigt die hier gezeigte Methode keinerlei kulturelle Anzucht der gemessenen Bakterien. Das könnte von großem Interesse für eine schnelle Erregerbestimmung und Resistenztestung sein. Die Ergebnisse zeigen Proteinsignaturen von 50 bakteriellen Genera und erlauben eine quantitative Abschätzung ihrer Veränderungen im Tagesverlauf. Eine Validierung dieser Bestimmung mit MALDI-TOF-Massenspektrometrie und mit Daten des Human Microbiome Project ergibt eine sehr gute Übereinstimmung. Damit stellt die Massenspektrometrie basierte Proteomik eine komplementäre Methode zur Bestimmung des Mikrobioms dar. Die kurze, kulturunabhängige Analyse begründet die Hoffnung, Mikrobakterien in Zukunft noch schneller und präziser charakterisieren zu können.

Abstract

The measurement of proteins from patient samples is of outstanding significance for medical diagnostics. Apart form classical Enzyme-linked Immunosorbent Assays in laboratory medicine protein analysis has recently been increasingly used in the form of MALDI-TOF mass spectrometry for the biotypization in microbiology. Mass spectrometry opens new perspectives for medical diagnostics allowing the quantification of all proteins in a cell or an organism, the so called proteome. The technological progress in the last couple of years enables the measurement of such a proteome within only a few hours.

In this thesis I present the human saliva proteome of eight healthy individuals for two timepoints as determined by mass spectrometry based proteomics. The acquired protein repository provides a comprehensive insight into the protein composition and dynamic of human saliva and represents the deepest body fluid proteome measured to date comprising more than 5000 proteins. A large proportion of this proteome can be obtained by a fast-track method of just four hours of total measurement time rendering it appealing for clinical diagnostics. The comparison of the human saliva proteome and the human plasma proteome reveal that the protein concentrations between saliva and plasma are hardly correlated. Consequently, it does not seem to be possible to determine the concentrations of many proteins in plasma from their concentration in saliva by proxy.

In a second step the direct measurement of the oral microbiome by mass spectrometry based proteomics is presented. By contrast to the currently employed MALDI-TOF mass spectrometry, the method introduced here does not require any cultivation of the measured bacteria. This might be of great interest for the fast species differentiation and for resistance testing. The results show protein signatures of 50 bacterial genera and allow a quantitative estimate of their alterations during the day. A validation of these measurements with MALDI-TOF mass spectrometry as well as with data from the Human Microbiome Project reveal good agreement. Hence, mass spectrometry based proteomics represents a complementary approach to determine the microbiome. The short, culture independent analysis nourish hope that it will be possible to characterize microbacteria even faster and more precisely in the future.

Contents

Abbreviations

ADH Alcohol dehydrogenase

APCS Serum amyloid P-component

APO Apolipoprotein

ATRN Attractin

BMI Body mass index

BSA Bovine serum albumin

CF Complement factor

CID Collision induced dissociation

CRP C-reactive protein

CV Coefficient of variation

EF Enrichment factor

ELISA Enzyme-linked Immunosorbent Assay

ESI Electrospray ionization

FDR False discovery rate

FWHM Full width half maximum

HDL high density lipoprotein

HMP Human Microbiome Project

HPLC High performance liquid chromatography

IMS Ion mobility spectrometry

ITIH3 Inter-alpha-trypsin inhibitor heavy chain H3

iTRAQ Isobaric tags for relative and absolute quantitation

LBP Lipopolysaccharide-binding protein

LC Liquid Chromatography

LDL low density lipoprotein

LFQ Label free quantification

MALDI Matrix-Assisted Laser-Desorption/Ionization

MS Mass spectrometry

NRP1 Neuropilin-1

ORM Alpha-1-acid glycoprotein

PASEF Parallel Accumulation-Serial Fragmentation

PCA Principle component analysis

PCR Polymerase chain reaction

PON1 Serum paraoxonase 1

PRG4 Proteoglycan 4

QTOF Quadrupole time of flight

SA Serum amyloid protein

SDB-RPS Styrenedivinylbenzene-Reversed Phase Sulfonate

SERPINA1 Alpha-1-antitrypsin

SERPINA3 Alpha-1-antichymotrypsin

SERPINA6 Corticosteroid-binding globulin

SERPIND1 Heparin cofactor 2

SERPINF1 Pigment epithelium-derived factor

SHBG Sex hormone-binding globulin

SILAC Stable isotope labelling with amino acids in cell culture

SWATH Sequential window acquisition of all theoretical mass spectra

TIMS Trapped ion mobility spectrometry

TMT Tandem mass tag

TOF Time of flight

1 Introduction

1.1 The nature of the proteome

A comprehensive understanding of the human body requires a detailed knowledge of its molecular constituents and their interaction in health and disease. Over the course of the last century scientists made huge progress in the analysis of biomolecules leading to sophisticated diagnostic tools and specific pharmacological therapies for the diagnosis and treatment of diseases. Yet, the pathogenesis of many diseases is only partially understood and for only few diseases specific diagnostic markers are in place.

At present, the breakthroughs in genome sequencing stir hope that the comparison of genomes of different individuals and of pathogens will revolutionize medical diagnostics and treatment. While genome sequencing will almost certainly have a big impact on personalized medicine it should be noted that in most of the cases genes only predispose for certain diseases. Ultimately, diseases manifest themselves on the level of proteins. Not surprisingly, most laboratory tests determine the level of a particular protein in a probe. While the DNA is merely the construction plan, proteins are the executives of almost all cellular functions. This is best illustrated by a caterpillar that turns into a butterfly - both share the same genome, but they have a vastly different morphology and protein composition [1]. Similarly, different cells in the human body vary greatly in their size, shape and function despite the fact that most of them carry identical copies of the genome.

The proteome is therefore defined in analogy to the genome as all the expressed proteins under a given biological condition. This does not only include the protein amount, but also their modification, location in the cell, interaction partners and their turnover. Hence, the proteome is highly complex and dynamic. Clearly, the proteome is the most relevant entity for the understanding of cellular function and regulation. Proteomics - the study of the proteome - is a systems biology approach to capture the properties of all the proteins of a system as opposed to the investigations of individual gene products in classical molecular biology.

Proteomics has not found widespread application in medical diagnostics yet - with the exception of Matrix-assisted Laser-Desorption/Ionization time of flight (MALDI-TOF) mass spectrometry (MS) that revolutionized biotypization of bacteria and fungi. This is mainly due to the fact that the methodological repertoire to study the entire cellular proteomes in the form of mass spectrometry based proteomics has only been developed recently.

Measuring the protein composition of a tissue is challenging for two main reasons. First, the dynamic range in protein abundance covers several orders of magnitude, which makes it very difficult not to lose signals from proteins with low copy numbers. For example, Nagaraj et al. showed that the proteome of the widely studied cell line HeLa encompasses 12000 proteins with a dynamic range of six orders of magnitude [2]. Secondly, proteins cannot easily be replicated by means of molecular biology in contrast to DNA which can be amplified by polymerase chain reaction (PCR). Consequently, the first characterization of the proteome of a free living organism was only achieved in 2008 for yeast [3]. In comparison, the first whole genome of *Haemophilus influenzae* was already sequenced in 1995 [4].

While the genome is less complex than the proteome, genome sequencing methods are much more mature than proteomics analysis methods. The transcriptome represents an intermediate layer of complexity and can be well studied by means of molecular biology. However, despite the common determination of the transcriptome to map regulatory cellular changes it should be noted that the proteome and the transcriptome are only moderately correlated with a correlation coefficient of $r = 0.7$ or less [5]. Despite the challenges to determine proteomes, huge progress has been made in mass spectrometry based proteomics in the last couple of years. Entire cellular proteomes can now be measured in a matter of hours using shotgun mass spectrometry [6, 7]. In 2014 two groups independently published nearly complete drafts of the human proteome using electrospary ionisation (ESI)-MS based shotgun proteomics [8, 9]. Although their bioinformatic analysis was criticized by the proteomics community [10], the results show that ESI-MS based proteomics might soon be applicable for clinical diagnostics.

1.2 Mass spectrometry-based proteomics

The choice of the analysis method in MS-based proteomics depends on the scientific question to be addressed. Several alternatives exist for key steps of the measurement process such as sample preparation, soft ionization and the choice of mass spectrometer. Nonetheless, it is fair to say that the analysis of complex protein mixtures mainly follows a generic ‚shotgun' workflow that is discussed in the following section and depicted schematically in figure 1. The workflow of the MALDI-TOF mass spectrometer that is used in clinical microbiology nowadays differs substantially from the generic ‚shotgun' workflow, because the requirements in terms of costs, efficiency, handling and depth of proteomic coverage are different. Its characteristics will therefore be briefly described separately in the next subsection.

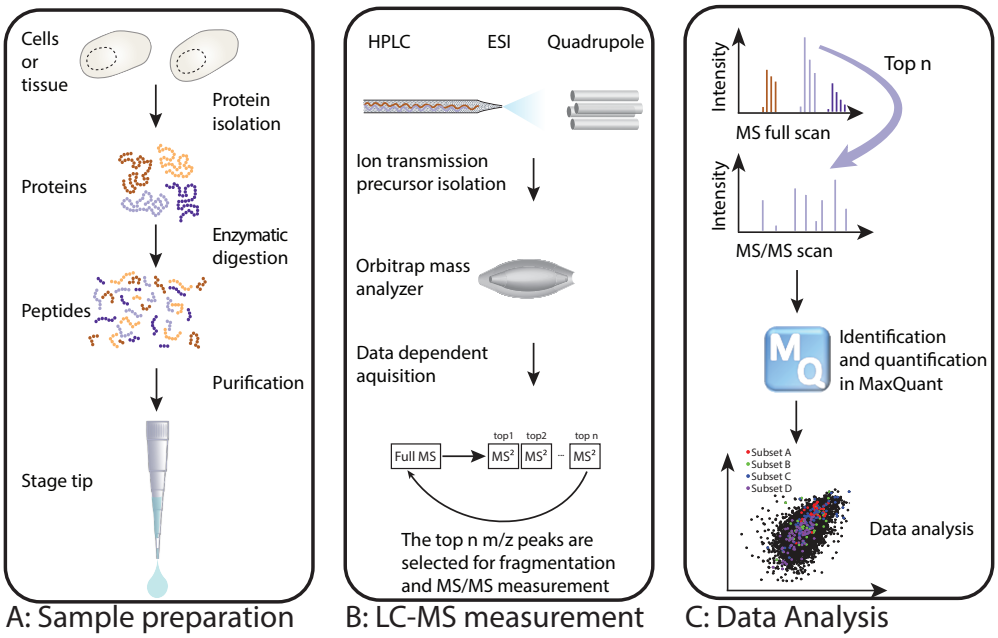


Figure 1: **Generic shotgun proteomics workflow.** A: Sample preparation: The isolated proteins are digested with site specific endoproteases and purified in stage tips. B: HPLC tandem MS: Peptides are separated by HPLC and transferred to the vacuum of the mass spectrometer via ESI. The top n m/z peaks in the full mass scan are isolated for fragmentation with a quadrupole and the resulting MS/MS spectra are recorded. C: Protein identification and quantification: The MS/MS and MS spectra together with sequence databases are used to determine the identity and quantity of the proteins in the original sample. This is done with sophisticated software environments like MaxQuant [11] that take multiple hypothesis testing into account and limit the false discovery rate to a predefined threshold.

A major hurdle for the analysis of large biomolecules by means of MS was their ionization and transfer into the vacuum of a mass analyzer. Franz Hillenkamp and John Fenn developed two different soft ionization methods in the 1980ies that solved this problem, MALDI [12] and ESI [13].

In MALDI the analytes are placed on a solid organic matrix and get desorbed by a laser beam pulse. As they detach from the matrix, the analytes or their fragments accept protons that originate from the organic matrix. The resulting ions are up to 200 kDa in weight and mainly singly charged.

In ESI ions are created by electrostatic dispersion and evaporation as the analytes enter the vacuum of the mass analyzer in small charged droplets of solvent. This idea is realized by applying a voltage of several kilo Volts between the tip of a glass capillary and the entrance funnel to the vacuum of the mass analyzer. The principles of MALDI and ESI are depicted in figure 2. Since a high pressure liquid chromatography (HPLC) can be directly coupled via ESI to a mass spec-

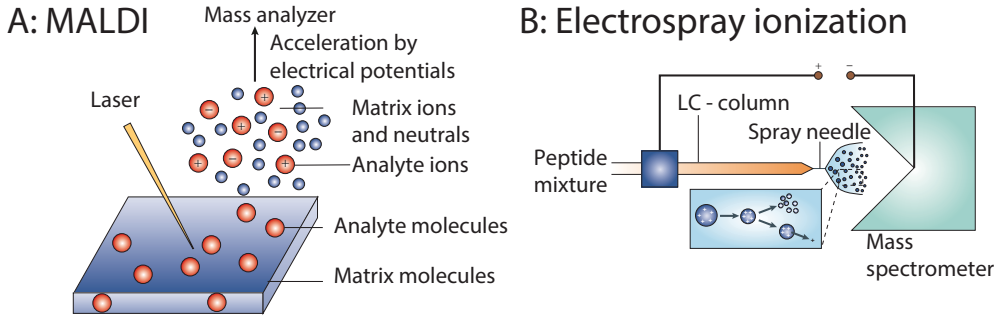


Figure 2: **Soft ionization methods for MS based proteomics.** A: MALDI: The pulsed laser beam desorbs the analytes that are embedded in the organic matrix. The energy deposition of the laser pulse enables analytes and matrix components to separate from the matrix, often in the form of singly charged ions. B: ESI: The application of a strong electric potenial between the needle of an HPLC and the vacuum funnel of a mass spectrometer leads to the formation of a cone beam with subsequent disintegration of the solute in small charged particles. Due to the evaporation of solvent the droplets split into ever smaller droplets in a process called Coulomb explosion resulting in charged analyte molecules in the gas phase. Modified from [14].

trometer, ESI is today the most common soft ionization method. In ESI-MS, the complexity of proteomics samples gets reduced by separating the probes according to their hydrophobicity prior to the analysis in the mass spectrometer. HPLC is a very powerful separation technique for peptides and MS-based proteomics leverages improvements in chromatography for ever deeper proteomics analyses. Consequently, the generic shotgun workflow relies on the coupling of a HPLC via ESI to the mass spectrometer (fig. 1 B).

Although soft ionization methods allow the analysis of proteins as a whole, it is easier and more effective to digest complex protein mixtures with sequence-specific endoproteases like Trypsin or LysC and measure the resulting so called tryptic peptides (fig. 1 A). This approach is called ‚bottom-up‘ principle and relies on subsequent bioinformatic reconstruction of the proteins in the sample (fig. 1 C). The disadvantage of the bottom-up approach is that it includes additional sample preparation steps prior to the measurement resulting in additional costs, sample loss and potential errors such as misscleavages of the endoproteases used. In general, cell lysis, protein purification and protein digestion should be performed carefully, because a good yield of tryptic peptides from the sample is imperative for accurate proteomics [15].

Today two main types of mass analysers are common, time of flight instruments (TOFs) and Orbitrap instruments - leaving the less common Fourier-transform ion cyclotorn resonance analyzers and linear ion traps aside [16, 17]. Both Orbitraps and TOFs offer high mass accuracy, m/z resolution, acquisition speed

and dynamic range of protein abundance [18–21]. Mass accuracy denotes the difference between measured mass and actual mass, m/z resolution describes the capacity to separate two signals with only slightly different m/z values. In daily practice, the robustness of the instrument's performance is also worthwhile consideration and arguably Orbitrap analyzers owe part of their success to their stable operating performance.

The Orbitrap analyzer consists of a spindle shaped electrode and an outer barrel shaped electrode as depicted in figure 3 A. The ions to be measured are collected in a C-trap prior to their eccentric injection into the space between the two electrodes of the Orbitrap. As the ions enter the Orbitrap they oscillate at different frequencies along the axis depending on their m/z ratio and induce currents in the outer electrode. A Fourier transformation of the detected signals allows to calculate the m/z spectrum. The limited space in the C-trap and the Orbitrap leads to space charging effects if too many ions are accumulated - a major disadvantage of this type of mass analyzer. Increasing the size of the Orbitrap and the C-trap is not an option since longer Orbitraps decrease the acquisition speed due to slower oscillation frequencies. Furthermore, the precise injection of more ions from a larger C-trap is challenging. For these and other reasons the potential for improvements of Orbitrap analyzers seem limited.

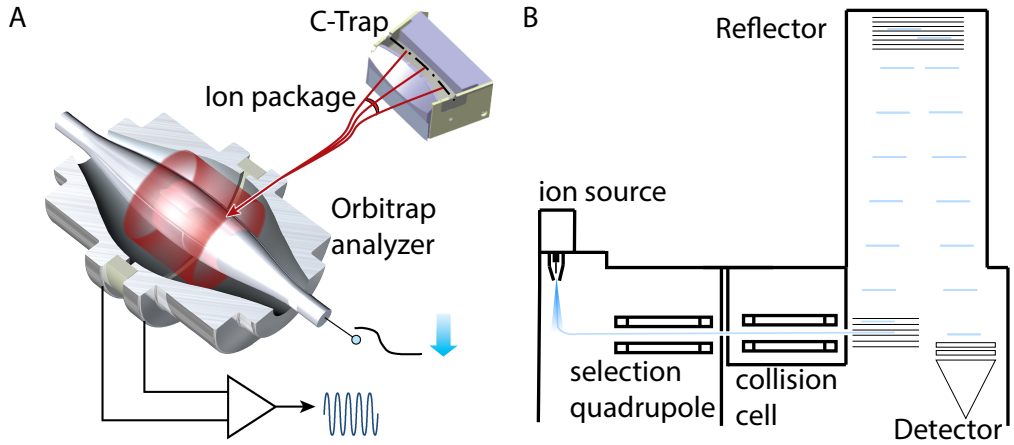


Figure 3: **Two common mass analyzers:** A: In Orbitrap analyzers an ion package is injected from a C-trap into the orbi cell. The frequency of the horizontal oscillation is used to calculate the m/z ratio via a Fourier transformation. Adapted from [22]. B: In TOF instruments, the ions are accelerated vertically into a field free drift tube. In many cases the ions get reflected to double the flight path before they get detected on a microchannel detector plate.

TOF mass analyzers accelerate packages of ions orthogonally in an electric field of known strength. This leads to a velocity dispersion of ions according to their m/z ratio. The difference in velocity between different ion species is subsequently detected by measuring the time of flight of the ions for a defined distance (fig.

3 B). TOF analysers operate at high scan speed and have the potential to reach high sensitivity, accuracy and good signal to noise ratios. Improvements in the guiding of ions along their trajectories and the sensitivity of the microchannel detector plate would make TOF instruments even more attractive. Already now TOF instruments represent a competitive alternative to Orbitrap analyzers for proteomics [20, 23, 24]. Additionally, the coupling of an ion mobility trap with a TOF instrument could render additional performance improvements which might be relevant for medical diagnostics as discussed in the Appendix section C of this thesis.

The mere recording of the m/z spectrum of the tryptic peptides is not sufficient to accurately determine the peptide sequence in most cases. Therefore, most mass spectrometers additionally isolate specific ions with a quadrupole and fragment the selected ions in a collision cell. The m/z spectrum of the fragment ions - the MS/MS spectrum - is in turn determined (fig. 1 B). Commonly, the ten to twenty most abundant peptide species of a given MS spectrum are isolated for fragmentation before the next MS spectrum is recorded - an approach called data dependent acquisition.

The reconstruction of the proteins that were originally in the sample based on the recorded MS and MS/MS spectra is a challenging endeavor. In the early days of mass spectrometry based proteomics human experts had to examine many MS/MS spectra to validate computationally calculated results. Nowadays, sophisticated software offers fast and highly accurate interpretation of mass spectra with stringent statistical cut-offs. The most popular platform is MaxQuant [11] and the following description outlines its processing algorithm, although the same steps are carried out by alternative platforms in more or less the same way [25, 26].

The first step in the processing of the raw mass spectra is the feature detection. To this end, three dimensional hills above the m/z-retention time plane are gathered. Fitting Gaussians to each peak in the MS spectrum and calculating the weighted average of the centroid masses of each feature increases mass precision. The charge state of an ionized peptide can be calculated based on its isotope pattern. Two chemically identical peptides can differ in mass due to a different composition of isotopes and hence form so called isotopologues. These isotopologues can be easily grouped since they differ exactly in multiples of one Dalton divided by the charge state of the peptide. The grouped isotope pattern constitutes a feature of known charge state and m/z ratio.

In a second step the detected features are mapped to their peptides of origin. Ideally, the MS/MS spectrum is so rich in fragments of different length, that the peptides sequence can be reconstructed entirely based on the fragment spectrum. In practice this is rarely the case. Hence, the features are compared to a

list of candidate peptides that is created by in-silico digestion of proteins based on the DNA sequence of the organism studied. In order to avoid false discoveries, a target-decoy approach is used to statistically control the false discovery rate (FDR) [27, 28]. This means that an equal number of fake candidate peptides is created artificially and added to the list of actual candidate peptides. Subsequently a matching score for each peptide in the joint list of candidate peptides is calculated and the peptides are ranked according to their matching score. The peptides with the best matching scores are accepted as hits as long as the total percentage of fake peptides among the accepted peptides does not exceed a predefined threshold - usually 1 %.

Assigning the identified peptides to proteins is not straightforward and has been coined the ‚protein interference problem'. Some peptides can be the products of the tryptic digestion of different proteins. In such a case, it is impossible to reconstruct from which protein they actually originated. Instead, these peptides are assigned to the entire group of proteins from which they could potentially originate. Another difficulty arises by the size of the proteomics dataset that requires to impose a FDR on the protein level in addition to the FDR on the peptide level. This is done by ranking the identified proteins or protein groups based on a score calculated from the posterior error probabilities of the peptides and applying the target decoy approach once again. The aforementioned drafts of the human proteome did not apply a FDR on the protein level which is why the reported results should be treated with caution [8, 9].

The mere identification of proteins is not sufficient to address most biological questions, but quantitative readouts are desired. Hence, sophisticated quantification methods have been developed for mass spectrometry based proteomics. The problem is that different peptides do not result in MS signals that are directly proportional to their original amount in the sample, since they differ in chemical behavior. For example, the ionization efficiency varies largely along the liquid chromatography (LC) gradient and for peptides with different physicochemical properties [29].

There are several ways to circumvent this bias. One of the most accurate quantification strategies is stable isotope labeling by amino acids in cell culture (SILAC) [30]. In SILAC two cell culture populations - i.e. treatment and control - are grown on different media. The medium of one cell culture only contains ^{12}C labeled arginine and the medium of the other cell culture only contains arginine labeled with ^{13}C. Hence, the resulting peptides of the two samples after digestion with trypsin are isotopologues with a mass difference of 1 Da. Consequently, their chemical behavior is identical, in particular their ionization efficiencies. As a consequence, the resulting ratio of the mass intensities of these isotopologues

reflect the ratio of their original protein ratios between the two samples. However, metabolic isotope labelling can be expensive and is impractical for in vivo experiments not to mention experiments in humans. As an alternative the labelling can be performed during sample preparation with chemical labels such as di-methyl labeling [31], isobaric Tags for Relative and Absolute Quantitation (iTRAQ) [32] or tandem mass tags (TMT) [33].

Chemical labeling reagents also allow multiplexing of samples by attaching different labels to different samples [34]. Using TMT reagents 10 samples have already been analyzed in parallel [35] and TMT 18-plex is under development [36]. This could be of great value for the application of proteomics in medical diagnostics, because it greatly reduces measurement time per sample. For the applicability of TMT reagents in medical diagnostics challenges in the reagent's stability and the labeling efficiency still need to be overcome.

Another very popular quantification strategy is label free quantification (LFQ). This approach assumes that the peptides of a protein that are most readily detected have nearly the same ionization efficiency. Using these peptide intensities for quantification of the respective proteins and normalization of the signals yields accurate results for protein quantification [37]. Not surprisingly, quantification is more accurate for proteins with high abundance, but fold changes can still be reasonably detected across several orders of magnitude [37]. The beauty of LFQ is that it can be applied to any proteomics sample without any additional preparation steps. This renders the combination of multiplexing samples with TMT reagent and quantifying them with LFQ algorithms particularly attractive for applications of proteomics in medical diagnostics.

1.3 MALDI-TOF in clinical microbiology

The MALDI-TOF mass spectrometers that are now used for routine clinical diagnostics differ in several aspects from the generic ‚shotgun' workflow explained in the last subsection. The food and drug agency approved the first mass spectrometry system for automated identification of pathogenic bacteria and yeasts in 2013. This was the result of a long development process that started in 1975, when Anhalt and Fenselau first proposed to identify bacteria by mass spectrometry [38]. Twenty years later, several groups managed to differentiate selected microorganisms by MALDI-TOF MS from intact bacterial cells [39, 40]. However, it was only a couple of years ago that scientists managed to reproducibly identify a wide range of clinically relevant bacteria with reasonable accuracy [41, 42].

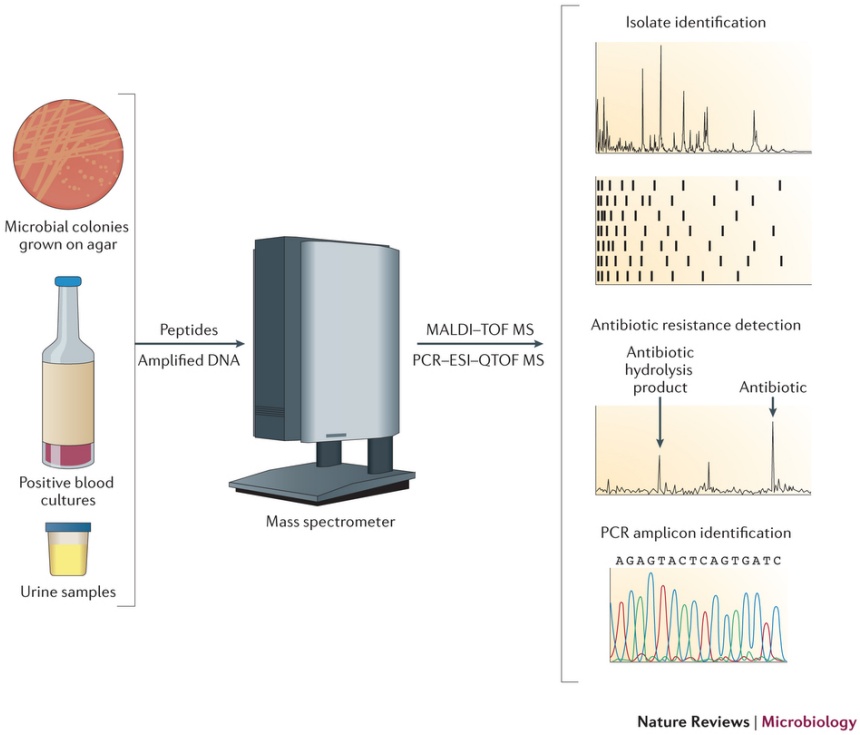


Figure 4: **Application of MS in clinical microbiology:** Bacterial colonies from agar plates, blood cultures or urine samples are transferred to MALDI-TOF mass spectrometers. The resulting mass spectrum is compared to datasets of previously characterized microorganisms to identify the bacterium or yeast under study. Since mass spectrometers can also measure metabolites or oligonucleotides, they allow to determine certain bacterial resistances via the detection of hydrolysis products or to identify bacteria with ESI-quadrupole-TOF MS following PCR amplification. Adapted from [43].

The basic workflow of MALDI-TOF MS is graphically summarized in figure 4. The first step in the discrimination of bacteria or yeasts with MALDI-TOF MS is to grow the microorganism on a culture plate or a broth culture [44]. For urine samples, it has been shown, that bacteria can be directly identified from samples with bacterial concentrations as low as 10^3 colony forming units per ml [45, 46]. The bacterial colonies are either directly transferred to the MALDI target plate together with strong organic acid and the matrix solution or incubated with ethanol, formic acid and acetonitrile for protein extraction prior to the transfer. These additional sample preparation steps offer enhanced biomarker detection and result in higher spectral scores [47]. Subsequently, the mass spectrum of the microorganism is recorded by means of MALDI-TOF.

The used TOF mass spectrometers commonly only record MS spectra, because they lack a quadrupole and a collision cell to select ions for fragmentation. Many

of the mass peaks originate from proteins of ribosomal origin that are particularly useful to discriminate bacteria phylogenetically - a principle also used by 16S-RNA sequencing. The resulting mass spectra are then used to identify the microorganism by comparing the spectrum to a library of mass spectra from microorganism. A score is calculated that expresses the certainty of identification. Note that proteins and their quantities are not determined in this way, but the direct comparison of mass spectra serves to identify bacteria or yeasts. This also implies that non proteinaceous molecules such as metabolites or oligonucleotides may play a part in the identification process. This can be used to detect antibiotic hydrolysis products and hence verify antibiotic resistance [48].

Furthermore, in an alternative method called PCR-ESI-quadrupole-TOF MS, the bacterial DNA can also be amplified using PCR followed by ESI and measurement of the PCR products in a quadrupole TOF (QTOF) mass spectrometer [49]. However, this alternative requires a different type of mass spectrometer and has not managed to establish itself in routine clinical microbiology.

1.4 Scope of the project

MALDI-TOF has fundamentally changed the way we identify bacteria in clinical microbiology and challenged well established diagnostic methods. Given that more complex mass spectrometers enable scientists to accurately quantify the protein inventory of a tissue or bodyfluid within hours [50] the question arises whether this technology might be of significance for other areas of medical diagnostics. The research lab of Professor Mann has a small subgroup working on this question. I had the great privilege to join this subgroup on clinical proteomics for a period of one and a half years. In this period I worked on three projects, one focusing on the determination of the human saliva proteome and the oral microbiome, one devoted to the fast quantification of the human plasma proteome and one method development project that implemented a new operation mode on a trapped ion mobility mass spectrometer. The first of these three projects is the subject of this thesis, because it was mainly conceived and carried out by me under the close supervision of Professor Matthias Mann, Professor Sören Schubert, Dr. Nils Kulak and Dr. Garwin Pichler.

Saliva was chosen, since it can be collected non-invasively and enables a particularly deep coverage of the proteome compared to other bodyfluids. The presented dataset represents the deepest body fluid proteome recorded to date and provides an unprecedented insight into saliva homoeostasis. Furthermore, the oral cavity harbors the microbiome rendering saliva as an ideal medium to explore the capabilities of mass spectrometry based proteomics for the direct characterization of complex bacterial environments without prior culturing. Special emphasis

was put on the simultaneous characterization of the oral microbiome by means of proteomics, because it opens up new possibilities to study host pathogen interactions and poses an alternative to the established sequencing approaches to characterize the microbiome. The main results of this project were published in *Genome Medicine* [51].

My contribution to this project included the conception together with Professor Matthias Mann, the design and the performance of the majority of experiments with the exception of the MALDI-TOF measurements and the plasma proteome measurements that were designed and carried out by Dr. Jette Jung, Professor Sören Schubert and Philipp Geyer respectively. Furthermore, I performed most of the data analysis, receiving support from Dr. Pavel Sinitcyn and Professor Jürgen Cox for the processing and interpretation of next generation sequencing data. The MaxQuant software [11] that I used for the data analysis was written by Professor Jürgen Cox.

The aforementioned plasma project was mainly conducted by Philipp Geyer. Besides many discussions about the clinical focus of the project with Philipp, I carried out several MS-based proteomics measurements of patient plasma. The project demonstrates how plasma proteomics allows to study pathophysiological changes systematically in an unbiased way and key results are presented in the Appendix. The plasma proteomes of 43 obese individuals undergoing an acute weight reduction therapy were determined across seven measurement time points spread over more than one year. The quantification of more than 400 proteins per sample gives a detailed, time resolved account of how lipoproteins and inflammatory factors respond to sustained weight loss. This study was published in *molecular systems biology* [52] and serves as a blueprint for how plasma proteomics can be exploited to discover new biomarkers or to examine the systematic effects of diseases.

The final project was technical in nature and mainly carried out by Florian Meier and Scarlett Beck. It dealt with the implementation of trapped ion mobility spectrometry (TIMS) on a high end quadrupole TOF mass spectrometer. While TIMS has been coupled to less complex mass analyzers in the past, this instrument type is in the middle of development and could lead to substantial improvements in instrument performance. Specifically, the so called Parallel Accumulation-Serial Fragmentation (PASEF) operation mode has the potential to drastically increase sequencing speed and sensitivity thus enhancing throughput.

I was actively involved in the design of the PASEF operation mode and conducted preliminary experiments on the prototype together with Florian Meier and Scarlett Beck. Part of the results of this project were published in the *Journal of proteome research* [53]. Since TIMS can also be realized in a MALDI-TOF

setting the development could eventually also impact the microbiological use of MS that is already in place. Either way, PASEF brings technological advances about that have the potential to make MS based proteomics way more competitive for a wide range of applications in clinical diagnostics. Key findings of this project will be briefly discussed in the Appendix.

1.5 Body fluid proteomics

Fast and comprehensive laboratory diagnostics has become one of the cornerstones of modern high performance medicine. The measurement of hallmark biomarkers from body fluids is a valuable piece of information to make a diagnosis, monitor drug therapy or to determine organ function. The vast majority of these biomarkers are proteins floating in the blood. Medical doctors commonly focus on a small subset of proteins that are known to have altered concentrations in the blood in certain diseases. Out of these, they select those with highest sensitivity and specificity for the hypothetical disease and measure them mostly with Enzyme-linked Immunosorbent Assays (ELISA) - the workhorse of protein analysis in clinical chemistry.

Proteomics would allow to assess the pathophysiological changes in a disease on the systems level. Instead of delivering the changes of a handful of known biomarkers, it can uncover the alterations of all the major proteins in body fluids such as blood, saliva or urine. This much more detailed picture of the health state of an individual would enable doctors to confirm a diagnosis by assessing a panel of several hundreds or thousands of proteins. While many of the measured proteins are likely to have a low sensitivity and specificity for diagnosing the disease in question, a score calculated on the basis of multiple, potentially hundreds of protein biomarkers is likely to yield high sensitivity and specificity. Furthermore, the determination of a patient's body fluid proteome could reveal unexpected pathological alterations that were not suspected from the patient's symptoms. Certainly, correction measures preventing the accumulation of false positives in such a diagnostic test would need to be put in place. The calculation of disease scores could again turn out to be helpful for the prevention of false positives.

Another advantage of body fluid proteomics is that it would allow to efficiently determine individual reference ranges for thousands of protein concentrations in the body of an individual. As discussed in the project on the plasma proteome in the Appendix B, the interindividual variation of plasma protein levels are notably higher than the intraindividual variation of plasma protein levels in a longitudinal study. Individual reference ranges would allow doctors to notice

deviations from healthy homoeostasis earlier. Therefore, proteomics could be of great interest for personalized medicine in the future.

Despite the appeal of this approach, its advantages have been precluded by the difficulty to quantify thousands of proteins in a body fluid accurately in a reasonable measurement time. This is mainly due to the dynamic range problem of proteomics, namely the simultaneous presence of proteins with low abundance and high abundance in the same sample. This makes it exceedingly difficult to measure low abundant proteins and to characterize samples with huge dynamic range down to the least abundant protein species. Over the last years, Professor Mann's laboratory has spent considerable effort on streamlining and simplifying its proteomics platform in order to enable its clinical application. The project on the dynamics of the saliva proteome [51] and the project on the changes of the plasma proteome [54] during weight loss intend to demonstrate the utility of proteomics for medical diagnostics.

1.6 The human saliva proteome

The oral cavity has a highly complex homoeostasis that is maintained by saliva. Saliva's functions include pre digestion, the lubrication of ingested food and bacterial defense. Disruptions in saliva secretion can therefore result in pathological conditions such as tooth decay, gingivitis or pharyngitis. Saliva is mainly secreted by the three pairs of salivary glands, but also contains shed epithelial cells and a multitude of microbes.

The Human Microbiome Project (HMP) has revealed that the oral microbiome and the gut microbiome are the most diverse in the human body and that the two are highly correlated [65]. Increasing evidence suggests a link between the human microbiome and diseases like obesity, allergies or multiple sclerosis [66–69]. This created enormous interest in the human microbiome and its interaction with the human body. With respect to the oral microbiome, it is well established that tooth decay and gingivitis are not caused by a single bacterial species, but by several species acting together [70]. The interest in saliva and its protein composition has been driven by these findings recently and promising steps towards the characterization of the human saliva proteome have already been undertaken [71, 72]. Yet, the authors either used extensive fractionation of their samples hence sacrificing quantitative accuracy and throughput or relied on ELISA based analyses that implied only a coverage of a few hundred proteins.

The analysis of the oral microbiome has so far relied on sequencing approaches or classical microbiology. Culture based analyses of complex microbiomes come at the disadvantage of having to grow the respective organisms outside their

natural environment in single colonies. While some bacteria might not even grow on agar plates, many of them will adapt to the culture conditions hence concealing their in vivo characteristics. On the contrary, sequencing approaches allow to analyze the collected sample with minimal delay at great sensitivity and therefore represent the method of choice for the analysis of highly complex microbiome studies. Yet, the bacterial genomes only help to identify bacteria, but reveal little about their state of being. MS based proteomics opens up the possibility to investigate the microbiome on the basis of its protein constituents and yields information about the host's interaction with his flora at the same time. This approach could also be of interest for bacterial biotypization in the clinic, because it could be faster than culture based assays. Given the success of MALDI-TOF and the fact that the speed of analysis is mainly limited by the growth of the colony in this technology, assays that do not require cultivation seem to be the next logical step. Since there have been hardly any attempts to characterize the oral microbiome by MS based proteomics, the saliva project also aimed to investigate whether in depth MS based proteomics can capture the oral microbiome.

Here we present the quantitative saliva proteomes of eight healthy individuals with unprecedented depth of more than 5500 proteins. In addition, we developed a single-run workflow that requires minimal amount of human saliva and delivers the quantities of thousands of proteins within four hours. This allowed us to investigate dynamic changes and inter-individual differences of the human saliva proteome. We also addressed the long-standing question as to which degree the quantities of proteins in saliva and plasma are correlated. Our measurements also quantified more than 2000 microbial proteins from 50 different bacterial genera. We co-analyzed these proteomics results with next-generation sequencing data from the HMP and compared them to MALDI-TOF mass spectrometry from microbial cultures and found strong agreement in both cases. Finally, we could show that the oral microbiome undergoes substantial changes upon eating and tooth brushing and differs between individuals.

2 Material and Methods

The application of mass spectrometry based proteomics to identify bacterial species has largely been limited to MALDI-TOF technology. Consequently, community standards and manufacturer's recommendations as to how to identify bacterial samples by mass spectrometry only exist for the routine clinical analysis of bacterial cultures with MALDI-TOF mass spectrometry. The use of ESI-MS based shotgun proteomics for the identification and quantification of bacterial communities therefore represented a challenge and the methodology presented here could not rely on established protocols. Instead, the choice of sample preparation, protein database and bacterial quantification represent the Mann lab's experience of how the new questions that arise from the analysis of bacterial samples, can be best addressed. The rational for choosing the following methodology is mainly outlined in the results section below. However, it should be noted, that the chosen methods are likely to be subject to refinement as the use of ESI-MS based shotgun proteomics becomes more widely adopted for the study of microbial communities.

2.1 Study cohort

We collected saliva from four female and four male, healthy, non-smoking donors aged 24 to 40 years with Caucasian backgrounds. All individuals did not take any drugs or antiseptics, regularly visited the dentist, showed no permanent medical condition and no signs of active inflammation, infection or bleeding. All donors provided their written informed consent to participate in this study and to publish the acquired results. The study was approved by the ethics committee of the Max Planck Society. All subjects donated two saliva samples at different timepoints, once at around 7 a.m. immediately after waking, before drinking, eating or tooth brushing and once at around 10 a.m. after the donors had eaten breakfast and brushed their teeth.

We also collected three samples immediately after one another form one donor at 10 a.m., prepared them in parallel and compared the results from this workflow replicate. We found a high reproducibility of $R^2 = 0.92$, consistent with previous bodyfluid analyses on our measurement platform [54]. The high reproducibility of our workflow encouraged us to refrain from replicates and use the available measurement time for the analysis of several donors and timepoints instead.

2.2 Sample collection and preparation

All donors abstained from eating and drinking at least 30 minutes prior to the collections. Saliva was collected with sterile cotton swabs that were subsequently placed in Eppendorf tubes 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). The swabs were thoroughly squeezed at the inner wall of the Eppendorf tube prior to removal. More than 100 μg of protein as estimated by the Bradford protein assay were reproducibly obtained in this way for the ensuing sample preparation with our in-StageTip protocol [15]. To this end, 20 μg of protein were digested by 0.4 μg trypsin and LysC in our lysis buffer at 37 °C shaking for 60 minutes. Subsequently, we added trifluoroacetic acid until a 1 % concentration was reached. 20 μg of acidified peptides were placed on a styrenedivinylbenzene-reversed phase sulfonate (SDB-RPS) StageTip [55]. The peptides bound to the SDB-RPS matrix and were purified by washing them with 200 μl of water. They were finally eluted with 60 μl 80 % acetonitrile (v/v) and 1 %ammonium (v/v). The resulting mixture was dried in a SpeedVac concentrator at 45 °C, and resuspended in A* buffer (2 % acetonitrile (v/v), 0.1 % trifluoroacetic acid (v/v), pH 2) to a concentration of 1 g/l.

For our deep human saliva proteomes, 15 μg of each sample were fractionated in an 80-min basic reversed phase gradient on a 20-cm, 75 μm thick column that was in-house packed with ReproSil-Pur C_{18} beads (Dr. Maisch GmbH, Germany). A fraction lasted 3 minutes of the gradient and the resulting fractions were concatenated to form eight final fractions per sample. These concatenated samples were again dried in a SpeedVac concentrator at 45 °C, and resuspended in A* buffer (2 % acetonitrile (v/v), 0.1 % trifluoroacetic acid (v/v), pH 2) to a concentration of 1 g/l.

The fractionated samples as well as the single run samples were separated with a 100-min chromatography gradient using an ESY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific) and a 40-cm column as above. The column was heated to 50 ° C by an in house designed conductive oven to enable high flow without pressures exceeding 1000 bar. By applying a 2.2 kV potential difference, the column was on line coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific). The column tip faced the entrance funnel of the mass spectrometer at an angle of 30 ° with the horizontal plane.

The MS scan range was from 300 to 1650 m/z with a resolution of 120,000 at 200 m/z and a maximum injection time of 55 ms. We applied a top 15 MS/MS method using higher-energy collisional dissociation with a quadruple isolation width of 1.5 m/s and a dynamic exclusion time of 30 s, meaning, that a given m/z

peak was excluded from isolation for the 30 s after being targeted for isolation. These settings of mass range, injection times, resolution, isolation width and dynamic exclusion have been shown to yield optimal results in this type of mass spectrometer for complex proteomics samples [19].

2.3 Protein database

We used the human reference proteome from Uniprot for our human only analysis (downloaded on 24 June 2015 from http://www.uniprot.org/). For the bacterial analysis, the Uniprot fasta files of all named species of the Human Oral Microbiome Database [56] with more than five protein sequences were downloaded (on 24 June 2015 from http://www.uniprot.org/) and combined it with the human reference proteome for the joint analysis.

In both databases, we only accepted Swiss-Prot reviewed entries, meaning, that the respective protein entries were manually annotated and the records were based on information from literature and curator-evaluated computational analysis rather than computationally analyzed data. Unreviewed databases tend to be much bigger and result in less reliable protein identification or - if a FDR is not imposed - in an accumulation of false positives. The human database comprised 90.5 K sequences and the joint database consisted of 1209.4 K entries.

2.4 Protein identification and quantification

We analyzed our proteomics raw data in MaxQuant (version (1.5.3.15)) [11]. This platform encompasses several algorithms for the statistically robust identification and quantification of peptides and proteins from high-resolution MS data. It has established community standards for statistically sound protein quantification and is now the most widely used computational platform in shotgun proteomics.

The integration of serial mass measurements and corrections of mass offsets into the calculation of peptide mass yield accuracies in the parts per billion range. Internally, MaxQuant uses the Andromeda search engine [57] to match the detected features against the reference proteome. A target decoy approach allows to impose a FDR on both the peptide and the protein level to identify peptides and proteins stringently. Furthermore, the MaxLFQ algorithm provides a highly accurate quantification of label-free proteomics samples [37] - ideally suited for our purpose.

Only tryptic peptides that were at least seven amino acids long and had two missed cleavages at the maximum were considered by MaxQuant. 247 potential

contaminants as listed by MaxQuant [11] were excluded from the analysis, resulting in a database of 0.64 million candidate peptides for the human database and 5.9 million peptides for the joint database, that fulfilled the criteria mentioned above. Search parameters were set to the default setting of MaxQuant version 1.5.3.15. The mass tolerance was thus set to 4.5 parts per million at the MS level and 0.5 Da at the MS/MS level, N-acetylation of proteins' N-termini and oxidation of methionine were accepted as variable modifications and carbamidomethylation of cysteine as fixed modification. We used the MaxLFQ algorithm for relative, label-free quantification with the minimum ratio count set to 1.

We analyzed the fractionated and single run measurements together in order to exploit the match between runs function in MaxQuant [58]. This algorithm optimizes peptide identification by aligning the elution times of different runs of the same tissue sample. Subsequently, precursors that are not selected for fragmentation in one run can be identified from other runs, in which they are selected for fragmentation. This way, even more peptides can be identified and quantified in highly complex peptide mixtures. The search parameters for the joint database were the same as for the human database except for the use of the so called split by taxonomy id algorithm on the phylum level and the exclusive use of unique peptides for protein quantification.

The split by taxonomy id algorithm is more thoroughly discussed in the results section. Briefly, the analysis of samples consisting of several species poses a challenge, because some peptide sequences are potentially shared by organisms. As a consequence, it is not possible to assign the peptide to one organism. Note, that a similar problem arises, as a peptide sequence is shared by several proteins from the same organism. This scenario however is resolved by MaxQuant by creating a protein group that contains all proteins with the given peptide sequence.

For the purposes of this work, the creation of protein groups containing human and bacterial proteins does not pose a viable solution. First, shared peptides could lead to the identification of proteins from other species, simply because peptides from other species shared peptide sequences with the respective protein. This is particularly relevant, if the protein abundances between human proteins and bacterial proteins is great. Figure 9 B demonstrates, that the abundance of bacterial proteins is substantially smaller than the abundance of most human proteins. Second, this abundance difference could lead to substantial quantification distortions, if peptides were not carefully attributed to their organism of origin.

Our analysis of the sequence identity among bacterial and human peptide sequences considered by MaxQuant in figure 9 A demonstrates, that the overlap is rather small, amounting to just 0.04 %. This provides evidence, that the potential distortions from shared peptide sequences is limited. Yet, we cannot assume

that the probability of MS peptide identification is equally distributed across the peptide sequence space. Given that the shared sequences are likely to belong to conserved genomic regions encoding essential proteins, these sequences are potentially overrepresented among the identified peptides.

Therefore, we decided to allow the formation of protein groups on the phylum level only, hence preventing a false assignment from human to bacteria or bacteria to human. Shared peptides were simply excluded from analysis. Along these lines, we only used unique peptides for the quantification of proteins, meaning that only peptides that were unique to a protein were used to estimate the respective protein abundance. This way, an overestimation of a low abundant protein through a shared peptide with a high abundant protein is prevented.

Together, the split by taxonomy id and the quantification with unique proteins provide a robust methodology to analyze multi-organism protein data also for future investigations.

2.5 Data analysis

MaxQuant output tables were further analyzed in Perseus (version 1.5.2.12) [59] - program that is tailored for the biological downstream analysis of highly multivariate quantitative protein abundance data. All proteins from the decoy database as well as proteins only identified by site and all contaminants were removed - including all keratin type I and II proteins. On the one hand, keratin type I and II are certainly expected in saliva due to shedding of epithelial cells, but on the other hand it is highly likely, that the samples are contaminated with keratins from the skin during sample preparation. The quantity of keratins in saliva would therefore be highly overestimated and we decided to exclude them from the analysis altogether.

All quantified proteins were ranked according to the mean LFQ intensities of the fractionated waking and postprandial samples of all donors. The means of low abundant proteins, that were not quantified in all donors were calculated by taking the LFQ mean among the donors in which the protein was quantified. Proteins were annotated with Gene Ontology terms and Uniprot keywords and a one-dimensional annotation enrichment analysis of these terms was calculated [60]. To correct for multiple hypothesis testing a Benjamini-Hochberg FDR of 2 % was imposed.

The comparison of the human plasma and saliva proteome was achieved by triplicate plasma proteome measurements with a 45-min HPLC gradient of two of our saliva donors. The sample preparation and MS measurement of the respective plasma samples was identical to that in Geyer et al. [54]. The resulting six

raw files were analyzed together with the corresponding saliva raw files with the MaxQuant settings described above.

Principle component analysis (PCA) was calculated on the logarithmized LFQ intensities of all 16 single shot saliva measurements. The comparison of the waking and the postprandial proteomes was done by filtering for 100 % valid values in all 16 samples and using a two sided t-test with a Benjamini-Hochberg FDR of 5 %and the s_0 parameter set to 0.1. The significance analysis for the upregulated proteins at the two timepoints was done by calculating a Fisher exact test with 2 % permutation based FDR.

The taxonomic tree in figure 8 was created by assigning peptides to bacteria. The number of peptides that we were able to assign to the respective level of this taxonomic tree was then written above the edges of this tree. Peptides that were shared by certain genera were added to the number of the lowest taxonomy edge shared by these genera (operating taxonomy unit). Genera that did not have at least one unique peptide were excluded because we could not rule out the possibility, that all the identified shared peptides attributed to this genus originated other microorganism.

Since we found 1069 peptides from streptococci, we extended the tree for this genus down to the species level. This was also partly motivated by Dr. Jung's remark, that the species differentiation of Streptococcus species can sometimes be challenging with MALDI-TOF. Bacterial genus abundance was calculated by summing the top ten peptide intensities for each genus in analogy to the top three protein quantification method [61, 62] and neglecting genera with less than ten peptides.

2.6 Comparison to sequencing data

For the comparison of our proteomics measurement of the oral microbiome to the HMP genome sequencing data, we downloaded the protein multifasta file from the HMP and analyzed fractionated and single run files with the MaxQuant settings described above against the human reference proteome and the multifasta file. The genomic quantification was performed by trimming the 764 fastq files from the HMP in Trimmatic [63] and aligning the sequences using Burrows-Wheeler alignment with default parameters [64]. The adapter, the leading and trailing sequences with a Phred quality score below 10 were removed. Additionally, reads with less than 36 nucleotides were not considered in the analysis.

Following Z-score scaling within each sample, a PCA of the reads per genus of the whole genome sequencing dataset together with the top ten peptide intensities per genus from the MaxQuant data was calculated. The body sites

„saliva“, „tongue dorsum“, „attached keratinized gingiva“, „palatine tonsils“ and „throat“ from the HMP were taken together in figure 9 D since they clustered tightly in the PCA and are all connected to the oral cavity.

2.7 Comparison to MALDI-TOF mass spectrometry

For the MALDI-TOF MS analysis all donors collected saliva by passive drooling into a sterile tube immediately after the 10 a.m. collection time point. 50 μl of each sample were instantly plated on one columbia, one chocolate blood agar plate for aerobic bacteria and two Schaedler agar plates for anaerobic bacteria. We incubated the anaerobic cultures under anaerobic conditions at 37 °C for at least five days and the aerobic cultures at 5.8 % CO_2 for 3 days. Different colonies were subcultured for MALDI-TOF MS identification as judged by visual and morphological evaluation by trained microbiologists.

The colonies were analyzed in duplicates following the protocol recommended by Bruker Daltonik. Briefly, a single colony was transferred onto the target plate and covered by 1 μl of matrix solution containing 10 mg/ml of α-cyano-4-hydroxy-cinnamic acid in 50 % acetonitrile/2.5 % trifluoroacetic acid (α-HCCA portioned matrix, Bruker Daltonik GmbH, Bremen, Germany). We used a Microflex LT benchtop instrument operated by felxControl 3.3 software (Bruker Daltonik GmbH, Germany) in linear positive ion mode at 60 Hz laser frequency with a mass range of 2 to 20 kDa for our measurements. The acceleration voltage was set to 20 kV, the IS2 voltage to 18.6 kV and the extraction delay time was 0.2 ms. The spectra were matched with the Bruker Taxonomy database version 4.0.0.1.

3 Results

3.1 In depth quantification of the human saliva proteome

Four female and four male healthy individuals with European genetic background volunteered to donate saliva. They were asked not to eat or drink for at least 30 minutes prior to the saliva collection to minimize disruption and contamination of the oral cavity. For the collection, each donor wiped the vestibule of the oral cavity, the teeth and the sublingual compartment with a sterile cotton swab (fig. 5). The recently developed In-StageTip sample preparation protocol [15] enabled us to perform repeated measurements with the approximately 200 μg of total protein from each swab.

A key step to any proteomics measurement is a high quality sample preparation. Ideally, the vast majority is retained during the purification, modification and digestion steps and any systematic bias is kept to a minimum. Even the best LC-MS measurement technology cannot compensate for any losses or contamination during sample preparation.

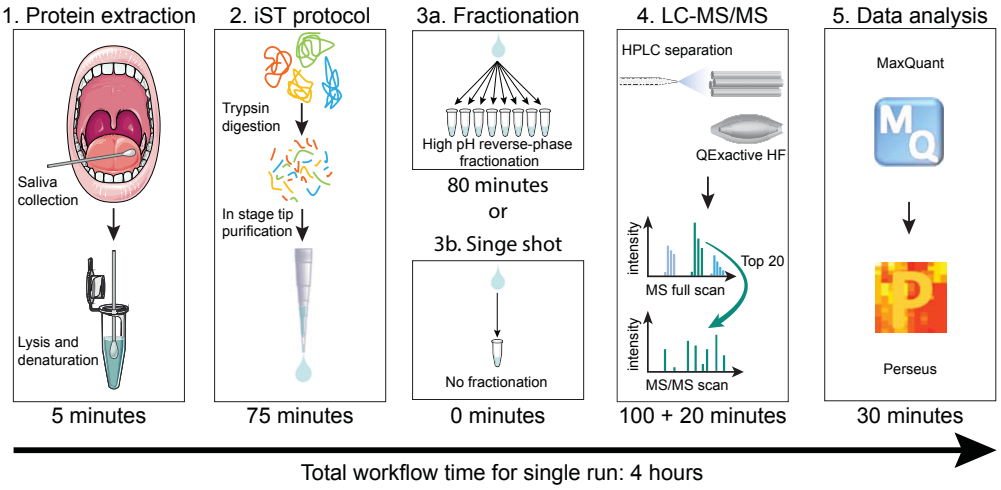


Figure 5: **Workflow for in-depth quantification of the saliva proteome:** (1) Saliva is collected by wiping the oral cavity with a sterile cotton swab thoroughly. Next, the cotton swab is placed in an eppendorf tube filled with lysis and denaturation buffer. (2) Proteins are digested and purified following the In-StageTip protocoll [15]. (3) Purified peptides are either measured directly in single shot measurments or fractionated with high pH reverse-phase chromatography into 8 fractions for in depth measurements. (4) Peptides are separated on a 100 min HPLC gradient and sprayed into a Q Exactive mass spectrometer operating with a top 20 tandem mass spectrometry method. (5) The resulting spectra are analyzed in MaxQuant and the Perseus software environment. The duration of each step is indicated below the respective panel. Adapted from [51].

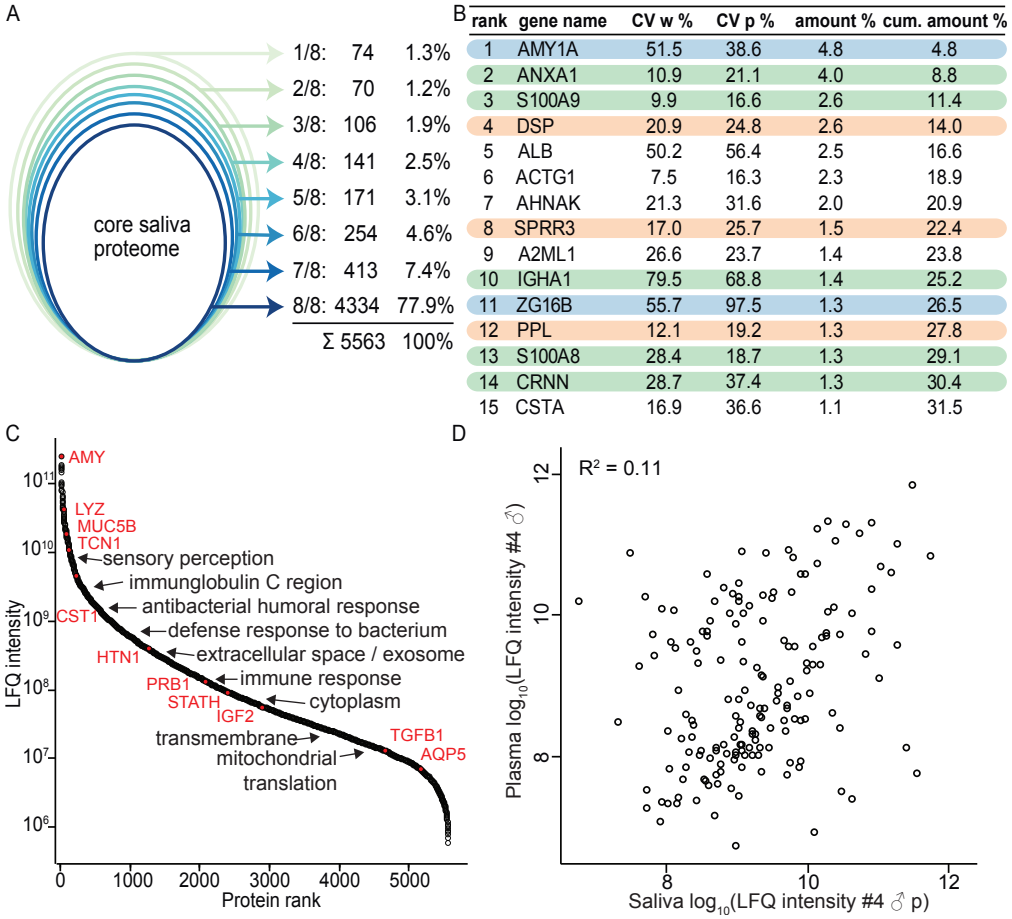
The In-StageTip sample preparation keeps the number of processing steps to a minimum by performing cell lysis, protein solubilization, protein denaturation, reduction, alkylation of cystein residues, enzymatic digestion and peptide purification in the single enclosed volume of a stage tip. This way, the peptide yield and the quantitative accuracy is very high. After a quick protein digestion of one hour with Trypsin and LysC, the peptides were purified and fractionated into eight fractions by basic reversed-phase chromatography. We subsequently measured all fractions with a 100 min LC gradient on a Q Exactive HF mass spectrometer, followed by data analysis in MaxQuant (fig. 5).

We found more than 54,000 sequence-unique peptides across our eight donors and more than 5500 different protein species in our search against the human reference proteome from Uniprot (fig. 6 A).

For both cases, we applied a FDR of 1 % to limit false positives due to multiple hypothesis testing. Note, that if peptides are not attributable to a specific protein, but occur in several proteins, a protein group is formed and such a protein group is only counted once in the 5500 proteins form above. A majority of 78 % of the detected proteins were identified in each of the eight saliva donors, 90 % in at least six donors. Only 1,3 % of the detected proteins were unique to one donor and we expect most of these proteins to be false positives given the FDR of 1 %. From this we conclude that our saliva retrieval and processing is robust allowing to compare thousands of proteins across individuals. We identified 5213 human proteins for one individual donor in the eight fractions. To our knowledge this constitutes the deepest body fluid proteome measured for an individual to date.

A comparison to plasma and urine proteomics revealed the reason for this remarkable protein coverage in saliva compared to other body fluids. The 15 most abundant protein species in saliva only account for 32 % of the total protein mass (fig. 6 B). By comparison the top 15 proteins in plasma and urine make up more than 90 % or 58 % respectively [54, 73]. Digestive proteins, proteins involved in immune defense and proteins belonging to the oral epithelium are among the most abundant saliva proteins (fig. 6 B).

To further investigate the abundance distribution of different saliva proteins, we plotted the LFQ intensities against the protein abundance rank for all quantified proteins and performed a 1D annotation enrichment analysis in the Perseus software environment using GO terms and Uniprot keywords (fig. 6 C) [60]. This analysis reveals the location of proteins with certain functions in the abundance scale. For example, proteins that scored in the upper quartile of this range were disproportionately often annotated with the terms „antibacterial humoral response“ and „defense response to bacterium“. The terms „extracellular space“ and



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

Figure 6: **Deep human saliva proteomes of eight healthy donors:** A: The ovals indicate the number of saliva proteins that are shared by the respective number of donors. On the right the numbers and percentage of proteins are given. B: Gene names of the 15 most abundant saliva proteins, their coefficients of variation (CVs) at waking (w) and postprandial (p), abundances in percentage of the total proteome and cumulative protein abundance (cum. amount). Proteins in blue are digestive proteins, proteins in green are involved in immune defense and proteins in red are of epithelial origin. C: Dynamic range plot of protein abundance in saliva with key salivary proteins in red. 1D annotation enrichment revealed that certain GO terms and Uniprot keywords are significantly enriched in certain abundance regions. D: Scatter plot of the LFQ intensities of proteins shared in the saliva and the plasma proteome. Adapted from [51].

„extracellular exosome“ turned out to be enriched near the median of the distribution. The low abundant proteins had annotations typical for intracellular proteins such as „cytoplasm“ and „mitochondrial translation“. The annotation enrichment therefore helps to grasp the quantitative distribution of functional proteins from a global perspective. Altogether, the protein abundance of saliva

spans a dynamic range of roughly six orders of magnitude as judged by our present measurement depth.

The question, whether saliva is suitable to measure plasma biomarkers by proxy, has been subject of ongoing debates [74]. It would be very convenient, if some of the information usually obtained from blood tests could also be acquired from saliva. Saliva collection is non-invasive, economical and well tolerated by all sorts of patients including children. There have been attempts to use saliva more extensively for medical diagnostics especially as surrogates for plasma markers. To address this issue, we determined the plasma proteomes of two of our saliva donors in triplicate measurements [54]. Since more than 50 % of the quantified plasma proteins were also present in the saliva proteomes of the respective donors, we were able to perform the most comprehensive comparison of protein quantities between saliva and plasma to date. The correlation between the LFQ intensities of the shared proteins is low (fig. 6 D). Across all replicates it never exceeded $R^2 = 0.20$. In addition, we analyzed specific saliva samples separately from the opening of the duct of the parotid gland, the duct of the sublingual and submandibular gland and from gingiva, but each of these samples showed equally low correlation to the plasma proteome (additional fig. C.2). Hence, we conclude that saliva cannot directly be used to infer plasma biomarker levels, because the saliva proteome and the plasma proteome show little overall correlation despite an extensive number of shared proteins. This does not exclude the possibility that individual protein levels show a high correlation between saliva and plasma, nor does it undermine the diagnostic potential of saliva in general.

All our saliva results are publicly available via the proteomeXchange repository (http://www.proteomexchange.org, accession number PXD003028) and via the user friendly MaxQB database [75]. This platform allows to browse our database for certain proteins and reveals the LFQ intensity, the abundance rank and further information for a protein of interest. For example, additional figure C.1 B shows the protein transcobalamin-1 in a protein abundance plot. It is a known component of saliva stabilizing cobalamin and protecting it against acidity of the stomach. It also serves in the transportation of cobalamin in the bloodstream. Cobalamin deficiency is common in the Western world with an estimated prevalence of 20 % in individuals aged 60 or above [76]. Given its clinical significance, its standard values in the blood have been determined in dedicated studies [77, 78]. Our saliva measurements revealed transcobalamin levels in saliva together with thousands of other proteins, hence demonstrating the utility that proteomics could have for clinical chemistry.

3.2 Dynamics of the saliva proteome

In the past, studies have examined the intraday changes of hormones such as cortisol in saliva [79]. However, changes in the saliva proteome during the day have not been studied systematically yet. In the following, we will present the intraday changes of the saliva proteome by comparing two timepoints, once immediately after waking before tooth brushing and one at 10 a.m. after breakfast and tooth brushing. Since this implied multiple measurements, we came up with a more rapid, but still sufficiently deep way to determine the levels of thousands of saliva proteins.

The peptide fractionation we used to obtain our deep saliva proteome comes at the cost of prolonging measurement time by the number of fractions. The throughput in modern clinical chemistry motivated us to speed up our measurement and to recover as many proteins as possible from a single-run experiment. To this end, we skipped the fractionation step and used the same 100 minute HPLC gradient to determine the saliva proteome of our eight donors at two different time points, once immediately upon waking before tooth brushing and once after breakfast and tooth brushing. Despite the shorter total workflow time of only 4 hours from collection to quantitative results we were able to identify 3835 proteins on average. The vast majority of 94 % of these proteins were also quantifiable using LFQ (additional fig. C.3 A). Repeated, simultaneous saliva collection from the same individual followed by independent processing yielded highly reproducible results with a mean coefficient of determination R^2 of 0.92 (additional fig. C.3 B). The interindividual differences were higher with a mean R^2 of 0.89. The CVs of protein LFQ intensities for the biological replicates did not dependent on protein abundance (additional fig. C.3 C). Consequently, our single-run workflow is suitable to reveal biological differences across a wide abundance range.

We started our analysis of dynamic changes of the saliva proteome by performing a PCA on our 16 single-run measurements. The PCA revealed that component 1 weakly separates the samples by sex, whereas component 2 separates the two collection timepoints - waking versus postprandial after tooth brushing (fig. 7 A and figure C.4 A in the appendix). Hence sex and timepoint account for the majority of variance in the data. To identify the proteins that drive this separation, we filtered for 100 % valid LFQ values and plotted Benjamini-Hochberg-corrected p-values versus fold change (fig. 7 B). That means, we only considered proteins that were quantified in all 16 measurements and used the appropriate correction for multiple hypothesis testing for this kind of analysis. Interestingly, the proteins that were more abundant at waking were significantly enriched in the keywords „antimicrobial“ ($p = 6.6 \cdot 10^{-8}$, enrichment factor (EF) = 24) and

„antibiotic“ ($p = 6.6 \cdot 10^{-8}$, $EF = 24$). By contrast, the terms „thiol protease inhibitor“ and „secreted“ were significantly more abundant in the postprandial saliva collection ($p = 3.3 \cdot 10^{-5}$, $EF = 42$ and $p = 8.7 \cdot 10^{-9}$, $EF = 6$, respectively). This suggests that the protein composition of saliva at night is specialized on bacterial defense while secretory digestive proteins dominate at 10 a.m.. As expected the highly abundant digestive enzyme α amylase had constantly higher quantities after breakfast. Hence, body fluid proteomics is now able to accurately capture shifts in protein abundance during the day.

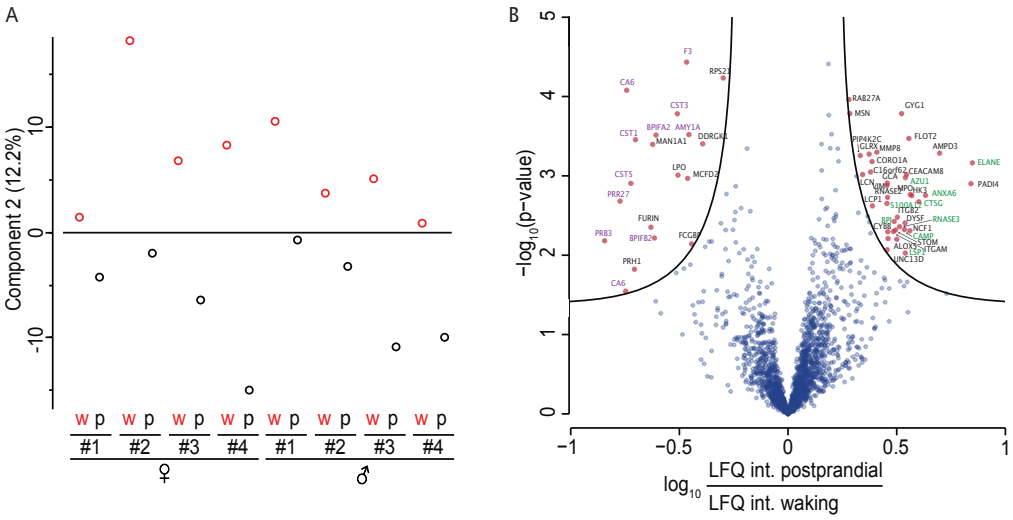


Figure 7: **Intraday variation of the human saliva proteome:** A: Component 2 of the PCA of our 16 saliva samples separates samples based on their collection timepoint. The red circles indicate the samples collected immediately after waking (w), the black circles are from the post prandial samples. B: Volcano plot showing the differentially regulated proteins between the two measurement time points by plotting the t-test significance (5 % permutation based FDR) against the logarithmized fold change of the LFQ intensity. The gene names in green belong to proteins labeled with the Uniprot keywords „antimicrobial“ and „antibiotic“ , whereas the gene names in purple belong to proteins labeled with the Uniprot keyword „secreted“. Adapted from [51].

3.3 Identification of the oral microbiome by proteomic means

The large number of identified saliva proteins encouraged us to look for bacterial protein signatures in our samples. We decided to include the Uniprot protein sequences of all named bacterial species from the Human Oral Microbiome Database that determined the oral microbiome by means of 16S rRNA sequencing [56]. The addition of these bacterial protein sequences to the human reference proteome resulted in a search-space that was eleven times larger than the human

database alone. By limiting the FDR to 1 % the increased search space does not lead to more falsely identified proteins, which is the case if no protein FDR is applied.

The combined search space brings about the challenge to correctly assign peptides to bacterial phyla since certain peptide sequences occur in proteins from different phyla. Closely related bacteria indisputably share many peptide sequences in common, but even not so closely related bacteria share many sequences [80, 81]. There is no community standard how to deal with this issue, since our work is the first that aims to quantify the proteins of a human microbiome accurately and stringently with MS-based proteomics. We therefore had to think of a way to reasonably assign shared peptides of closely related organisms to proteins. We used the so called split by taxonomy id algorithm on the phylum level in MaxQuant to deal with this issue. This algorithm was originally developed by Professor Jürgen Cox to analyze xenografts of human tumors transplanted to mice. We argue that it can also be used to accurately quantify bacterial proteins in complex microbiological samples in the following way.

Usually, a peptide shared between two proteins leads to the formation of a protein group. Split by taxonomy id prevents this from happening for proteins from different phyla and hence neglects these peptides for protein identification. In combination with the exclusive use of unique peptides for protein quantification peptides shared by different phyla do not contribute to the identification or quantification of proteins (compare materials and methods). While we decided to apply the split by taxonomy id algorithm at the phylum level of the taxonomy tree this is by no means canonical. If the cut-off for the formation of protein groups across different species was applied on a lower level of the taxonomy tree a large proportion of peptides would get neglected. However, if no cut-off was applied at all, peptides from one phylum could contribute to proteins from another phylum distorting its identification and quantification.

The issue and our rational for applying the split by taxonomy id algorithm on the phylum level is visualized in figure 8. We placed the number of shared bacterial peptides in our deep saliva proteomes on the respective branches of the taxonomic tree. The peptides on each branch originate from protein sequences that are shared by all bacteria down this branch and cannot be ascribed to one microorganism in particular. Genera that did not have at least one peptide that was unique to them were excluded. We hold that this is the best quantification strategy for this kind of proteomics analysis at the present measurement depth.

In addition, we considered the possibility that the simultaneous presence of human and bacterial proteins in the oral cavity impairs our described human protein quantification. To address this issue, we digested all bacterial protein sequences

and all human protein sequences in our database in silico and identified the nonredundant tryptic peptides that were long enough to be considered by MaxQuant. Among these, only 0.043 % originated from both bacterial and human protein sequences (fig. 9 A). Consequently, the presence of bacterial proteins in our saliva samples does not substantially influence the quantification of human proteins.

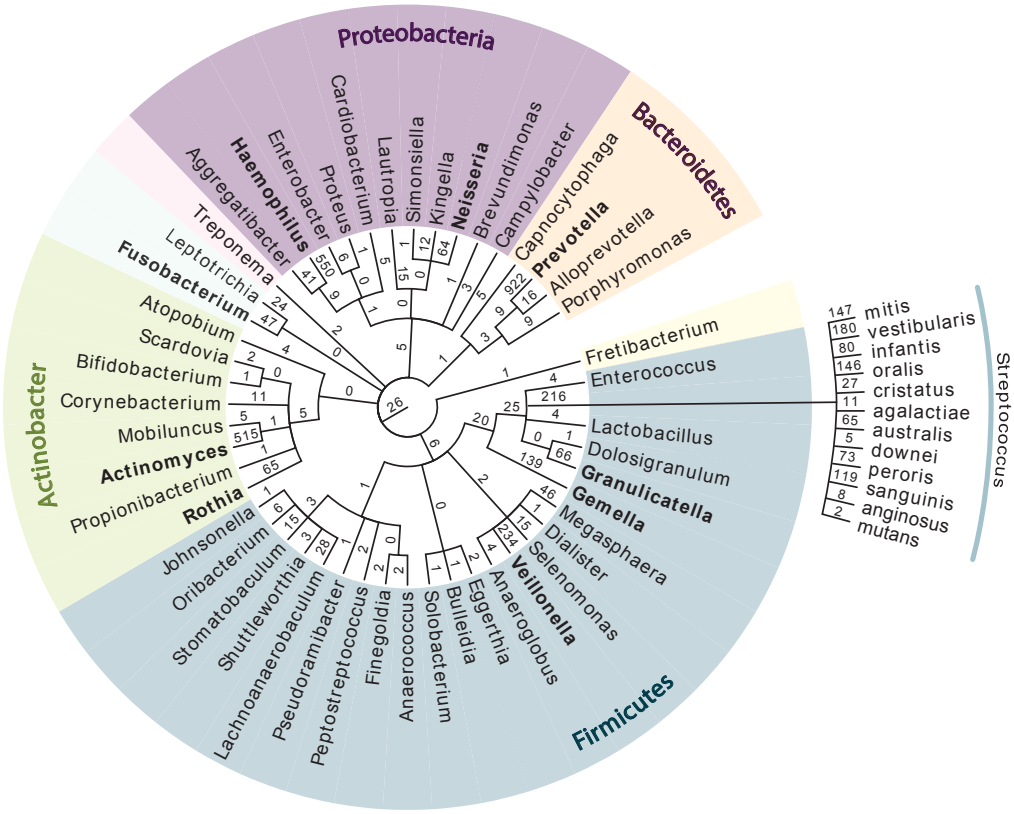


Figure 8: **Distribution of bacterial peptides along the taxonomy tree of 50 bacterial genera:** The numbers above the edges indicate the number of peptides that were attributable to this position in the taxonomic tree. Genera in bold were also found by MALDI-TOF MS of bacterial cultures of our saliva samples. We extended the tree down to the species level for the highly prevalent genus Streptococcus. Modified from [51].

Along these lines we also wondered whether ingested proteins from food might lead to a misquantification of human or bacterial proteins. We therefore performed an analogous analysis for the protein sequences of wheat and bovine as representative parts of a Western breakfast diet. The resulting overlap with bacterial or human peptides with at least seven amino acids in length was in each case far below 1 %, except for the overlap between human and bovine peptide sequences which amounted to 20.7 % (fig. 10). Hence, our in silico analysis does not

exclude the possibility of a quantification bias due to ingested bovine proteins. However, we did not find a substantial difference in the quantity of human milk or human muscle proteins between the waking and the postprandial samples, as we would expect if a bovine diet had an influence on the quantification of human proteins (compare fig.7).

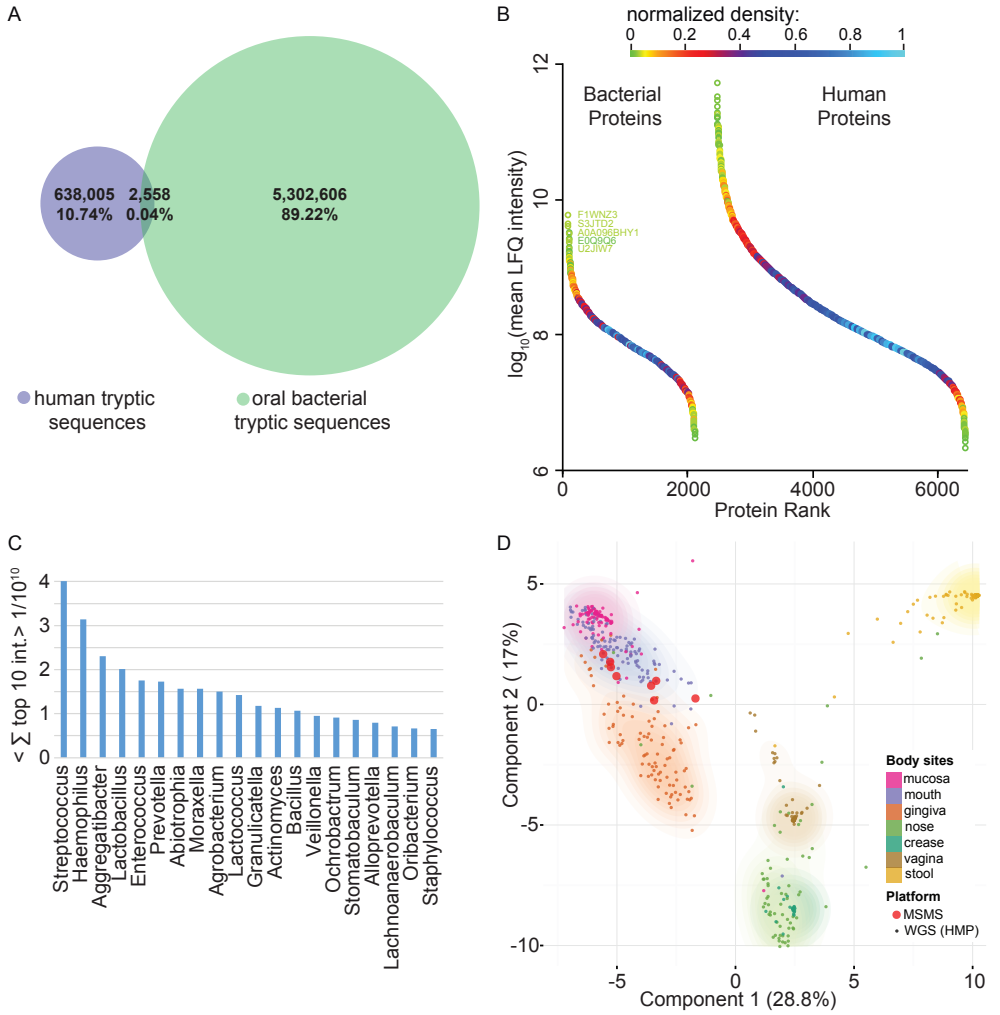


Figure 9: **Quantitative bacterial proteome:** A: Number of non-redundant tryptic peptide sequences in the human (violet) and oral bacterial (green) search space. B: Protein abundance plotted against protein rank. Bacterial proteins are less abundant, but still span several orders of magnitude. C: Quantitative relations of the 20 most abundant bacterial genera as approximated by summing the top ten peptide intensities of each genus. D: PCA of the whole genome sequencing data from the HMP co-analyzed with our saliva proteome data. The respective quantities of the involved bacteria were calculated from the reads per genus for the genomic data and by the top ten peptide intensities for the MS data. The MS-based bacteria quantification tightly co-localizes with the mouth sites from the HMP. Adapted from [51].

We concluded from these considerations that our approach of combining human protein sequences with the protein sequences of the oral microbiome in a joint database is suitable for the simultaneous analysis of human and bacterial proteins. We were surprised that we identified 2234 different bacterial proteins in this way using our standard 1 % FDR at the peptide and protein level. These proteins originated from 50 different bacterial genera from nine different phyla. This corresponds to almost 50 % of the named genera in the Human Oral Microbiome Database with annotated UniProt proteomes. The number of identified peptides differed substantially between different genera, ranging from only one to 1069 for streptococci. This allowed us to extend our analysis and to differentiate twelve different streptococcus species including *Streptococcus mutans* (fig. 8).

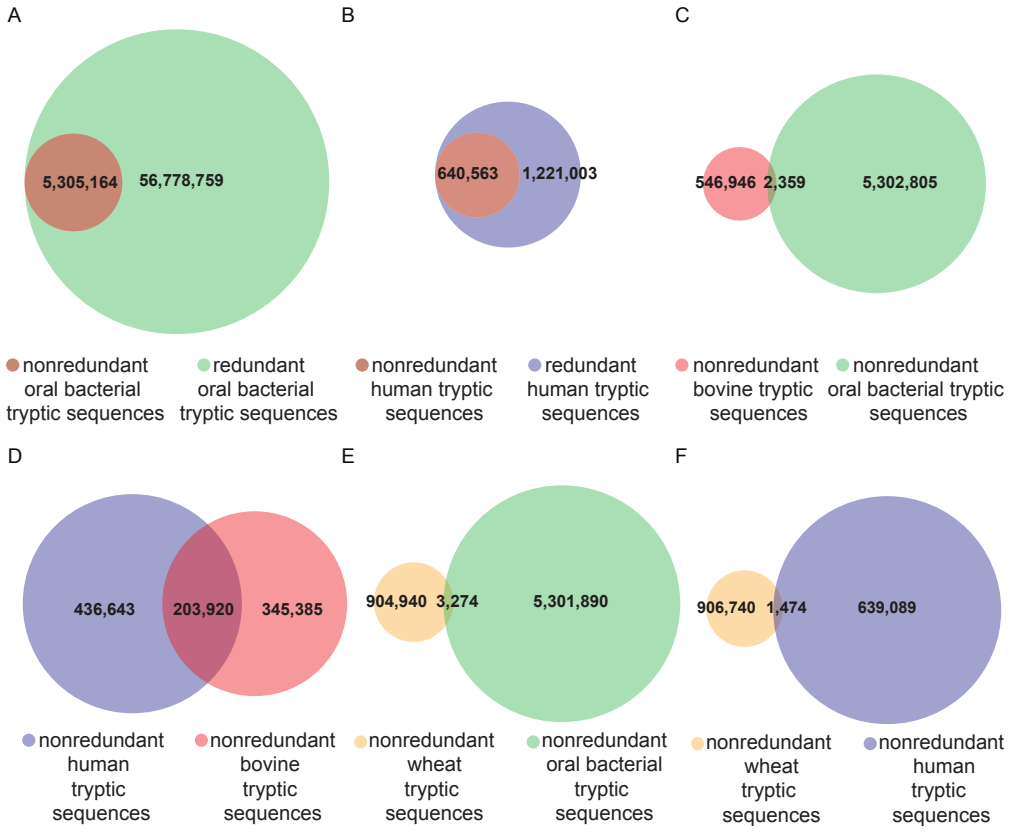


Figure 10: **Intersection of tryptic sequences considered by MaxQuant for different organisms:** The Venn diagrams depict the number of tryptic peptides of at least seven amino acids in length for different organisms. A, B: The oral microbiome has many redundant tryptic sequences compared to humans since many bacteria share sequences. C, D: Bovine share few tryptic peptides with bacteria, but almost 20 % with humans. E, F: In contrast wheat sequences have very little identity with bacterial or human sequences. Adapted from [51].

We also employed MALDI-TOF MS of aerobic and anaerobic cultures of our saliva samples according to standard protocols in clinical microbiology (compare materials and methods). This complementary method found 14 different genera in total, all of which were also identified by shotgun proteomics. The number of identified peptides of these genera suggests that they were comparatively abundant in saliva (fig. 8). Certainly, the aim in clinical microbiology is not to identify all bacterial species present in a sample, but rather to determine a dominant pathogen efficiently and at low costs. Nonetheless, it is striking that unbiased and relatively straightforward shotgun proteomics allows to identify such a high number of bacterial genera without the need for cultivation.

3.4 Quantification of the oral metaproteome

Although bacterial proteins accounted for almost 35 % of the total set of 6000 proteins in the analysis of the joint database, their total protein mass is rather low. This is best illustrated by plotting the cumulative percentage of bacterial proteins as a function of protein abundance rank (fig. 11). For example, the percentage of bacterial proteins among the most abundant 1000 proteins was only 5 %. The proportion of bacterial proteins per 100 identified proteins consequently rises steadily and reaches around 50 % towards the limit of detection. This suggest that as the depth of proteomic coverage improves, further bacterial proteins will be revealed. This seems attractive, since a better coverage of the bacterial proteome would allow to analyze how bacterial abundance and metabolic pathways change across different conditions.

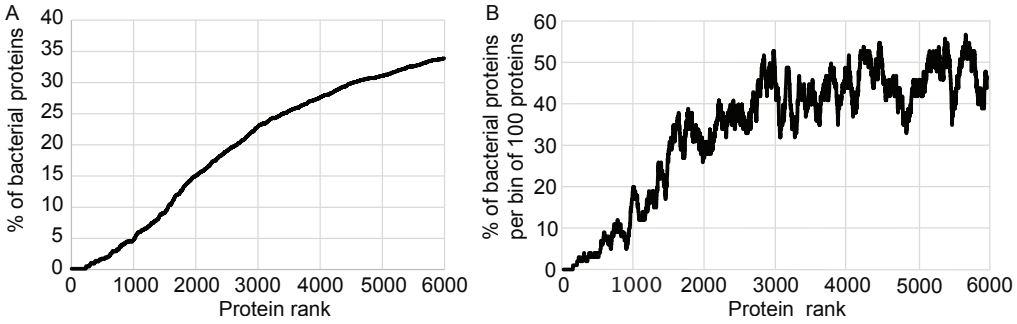


Figure 11: **Distribution of bacterial proteins across the protein abundance range:** A: Percentage of bacterial proteins as a function of protein abundance rank. B: Percentage of bacterial proteins per bin of 100 proteins. Adapted from [51].

Another abundance comparison between bacterial and human salivary proteins is illustrated in the abundance rank plot in figure 9 B. Surprisingly, the most abundant bacterial protein, F1WNZ3, the *Moraxella catarrhalis* homolog of chaperone

protein HscA, was around 100 fold less abundant than α-amylase 1, the most prevalent human protein. Other highly abundant bacterial proteins carry out household functions like A0A096BHY1, a glyceraldehyde-3-phosphate dehydrogenase, or E0Q9Q6, a subunit of DNA polymerase III. Their high abundance is probably in part explained by the high degree of sequence identity across different bacterial species. Hence, their dominance could be explained by the joint contribution of several different genera.

Next, we wanted to come up with a method to estimate bacterial abundance by our proteomics data. The majority of bacterial proteins belonged to one of the four phyla, firmicutes, actinobacteria, bacteroides, proteobacteria, with 300 to 800 uniquely assigned proteins each (C.4 in the appendix). Yet, the number of identified proteins is merely a very rough indicator of bacterial abundance. Instead, we summed the MS intensity of the top ten most abundant peptides of a microorganism across all samples in analogy to the top-three-peptide that is well established for the label-free quantification of protein abundance [61, 62]. Microorganisms with less than ten peptides were excluded from our analysis. Summing the intensities of all peptides of a microorganism would most likely lead to an overestimation of abundance differences, just like the summation of all peptide intensities of a protein causes an overestimation of protein abundance differences.

We decided to consider all peptides of that microorganism for the determination of the ten peptides with highest intensity regardless of whether they were shared with other microorganisms. This puts oral microbiota that share peptide sequences with other oral microbiota at a slight advantage, but the alternative of using only unique peptides would have put them at a great disadvantage. It is important to note that this problem is inherent to any quantification method of a diverse microbial community by means of MS-based proteomics. At the present coverage of the proteome a sophisticated calibration of our quantification method using a complementary quantification method seems impractical. Still the described quantification method of bacterial genera by their top ten peptide intensities should to be considered as rough approximations rather than accurate quantifications.

The abundances of the top 20 bacterial genera in saliva as judged by the described quantification are displayed in figure 9 C. Our results are in accordance with 16S RNA studies of the human oral microbiome [82, 83] that found Streptcoccus and Lactococcus as most abundant bacterial genera.

Another indication that our bacterial quantification measure roughly represents actual bacterial abundances is our comparison to the quantitative results of the HMP. We used the reads per genus of the HMP whole genome sequencing database and processed them together with the genus abundance estimations from

our MS data. The resulting PCA is displayed in figure 9 D and shows a close co-localization of the two datasets. This is remarkable given the quantification of bacteria used two completely different methods and the samples originated from different donors. The similarity of the oral microbiome across individuals is however well established [84].

The HMP analyzed the microbiome from different parts of the body including several separate sites in the oropharynx including *saliva* , *tongue dorsum*, *attached keratinized gingiva* , *palatine tonsils* and *throat*. The sites in the oropharynx all clustered together in the PCA thereby demonstrating that determining the oral microbiome from human saliva alone is a valid strategy.

3.5 Interindividual variation and dynamics of the oral microbiome

Our saliva collection from eight individuals at two timepoints allowed us to compare the oral microbiome across individuals and changes between our two collection timepoints. The interindividual differences in bacterial diversity and abundance were small (fig. 12 A). The mean R^2 was 0.82 between our donors. Nonetheless, certain genera differed up to tenfold between different individuals. We estimated relative differences in the total bacterial mass of the top eight genera by summing their respective abundance estimates (fig. 12 B). It turns out that this bacterial mass differs up to threefold between different donors. The relative contributions of these top eight genera are remarkably similar across individuals (fig. 12 C).

Given that the human saliva proteome clusters by sex in the first component of a PCA we were interested whether there were substantial sex differences in the oral microbiome. To this end, we aggregated male and female saliva metaproteomes and compared the bacterial abundances between these groups. Bacterial abundance was highly correlated with a coefficient of determination of $R^2 = 0.94$ (fig. 12 D), suggesting that there are only minor sex differences in the oral microbiome. Yet, the abundances of oral bacteria changed drastically upon eating breakfast and tooth brushing (fig. 12 E). The abundances of the most prevalent bacteria reduced around 2.5-fold, while the reduction of low abundant bacteria was even greater. The highly prevalent streptococci were reduced almost threefold (fig. 12 F). In the future it would be interesting to study the composition and changes of the oral microbiome under different conditions at greater depth.

A major advantage of our proteomics characterization of the microbiome over genome sequencing approaches is the simultaneous determination of the human saliva proteome. This allows to uncover interactions of the body and its microbiome on the protein level. For instance, our measurements revealed that the

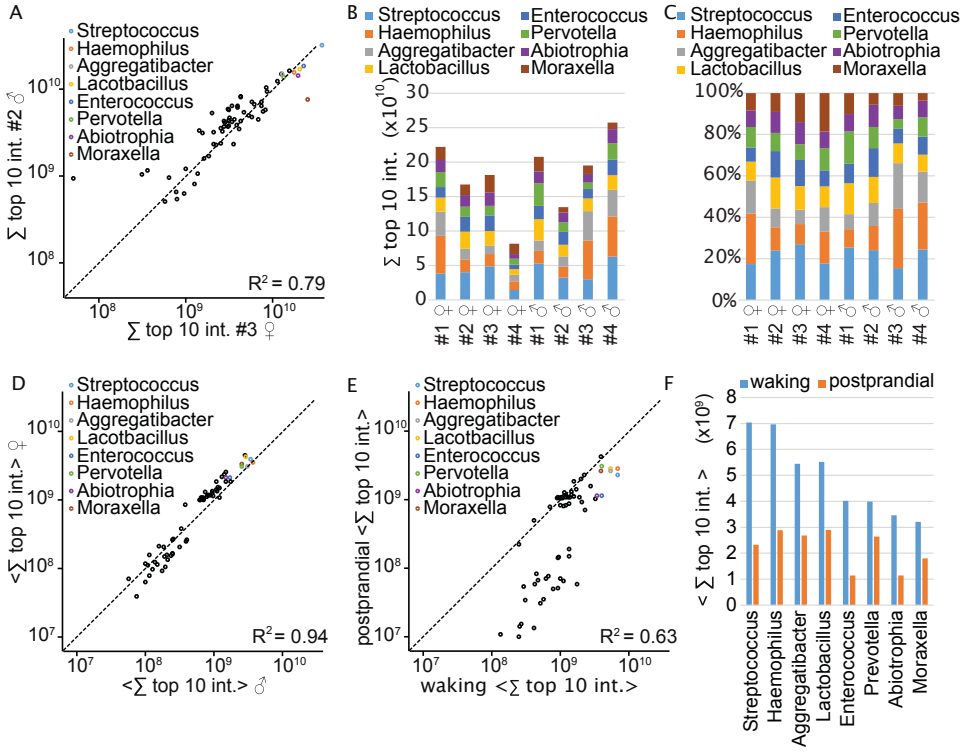


Figure 12: **Interindividual and timepoint differences of the oral microbiome:** A: Typical scatter plot of the bacterial genera abundance between two donors. The eight most abundant genera are highlighted in color. B: Relative quantification of the total bacterial mass of the highlighted eight most abundant genera showing up to threefold interindividual differences. C: The respective percentages of the involved genera are relatively constant across individuals. D: Differences in genera abundance across the averaged quantities in males and females were little. E: Mean bacterial abundance was markedly reduced after breakfast and toothbrushing compared to waking. F: Bacterial abundance of the eight most abundant genera in saliva compared between waking and postprandial. Adapted from [51].

human saliva proteome was enriched with proteins involved in bacterial defense at waking, when bacterial abundance was high. This reflects the body's attempts to limit bacterial growth during the night. Thus metaproteomics approaches like ours open up new ways to study host bacteria communication.

4 Discussion

4.1 Significance of MS based saliva proteomics and determination of the oral microbiome

This work demonstrates that shotgun proteomics is now in the position to characterize body fluids such as saliva at a depth of several thousand proteins. In fact, our deep human saliva proteome represents the deepest body fluid proteome recorded to date despite the fact that it was acquired in limited measurement time. The quantification of so many proteins in saliva represents a valuable resource for the research community. As a matter of fact, several scientists have contacted me since the publication of the saliva proteome asking whether I could check the relative quantities of certain proteins in saliva for them.

The rapid workflow, that allows to determine a human saliva proteome from collection to results in only 4 hours, seems particularly attractive. The measurement time has reached time scales that seem compatible with clinical application of MS-based protomics. This certainly renders saliva proteomics attractive for studies on patient cohorts. Such studies could identify proteins that are of major importance for maintaining oral cavity homeostasis. Potentially this could even enable physicians to develop a risk score to determine the probability of a patient to come down with cavity paving the way to personalized prevention.

In a second step, we demonstrated that proteomics can identify 50 bacterial genera in saliva without the need to cultivate them first. We cross validated our results by comparing them to next-generation sequencing data from the HMP and MALDI-TOF bacteria typing that is used in routine clinical practice. In both cases our proteomics data showed good agreement, indicating that shotgun proteomics is a complementary method for microbiome analyses.

MS-based proteomics seems especially attractive for the study of pathogen host interaction, since proteomics can provide host protein levels of the bacterial environment at the same time. It might help us to understand the interplay between microbial communities and our body on the systems level. In the case of the oral cavity, we showed that the levels of antimicrobial proteins are elevated at waking when the total bacterial mass is high. Yet, a better annotation of bacterial sequences and complete bacterial databases are needed for this approach to reach its full potential. The characterization of human microbiomes by means of MS-based proteomics is still in very early stages, but it opens new frontiers for the culture independent study of microorganisms.

The pace of development in MS-based proteomics is very high and does not appear to slow down. This stirs hope that it could be of huge benefit for the

medical sciences. While proteomics is already extensively used in research, it has not found clinical application yet, except for its use in MALDI-TOF in clinical microbiology. This thesis explored the potential of MS-based proteomics for the characterization of bodyfluids and the oral microbiome. The following two subsections will discuss some of the technological challenges that MS based proteomics is still facing and highlight its potential clinical applications.

4.2 Technological challenges for clinical proteomics

At the moment there are three main technological obstacles for the use of MS-based proteomics in clinical diagnostics beyond its use in biotypization of bacteria. These are the moderate sample throughput, the instability of the instrument performance and the complexity of data interpretation.

Sample throughput in clinical chemistry is enormous. This is partly due to the fact that most clinical parameters, proteins included, are determined in separate, dedicated assays. However, even if we neglect that most patient samples are tested for multiple readouts, the mere number of patients requires efficient and cheap measurement processes.

MS-based proteomics made stunning progress in sample throughput in the last decade. Back in 2008 the milestone of the first complete characterization of the yeast proteome took the authors three months to complete [3]. Nowadays, the same results can be achieved in one hour with considerably better quantitative accuracy [7, 15]. Nonetheless, such measurement times are still not suitable for the analysis of patient samples other than for research. The limited sample throughput also implies high costs, because the investment for instrumentation amounts to several hundred thousand euro. However, the costs of the supplies per sample are low, suggesting, that advances in sample throughput could render MS-based proteomics cost efficient.

The project on PASEF discussed in the appendix aims to overcome some of the described limitations in MS technology. PASEF will substantially increase proteomics coverage or - more importantly - allow to increase sample throughput without losing sensitivity. Furthermore, it is fully compatible with multiplexing strategies that use chemical labeling to parallel MS analysis. The combination of these two approaches is likely to boost sample throughput at least twenty-fold, potentially even more. In conclusion, the innovations that are already under way as well as future developments are likely to overcome the present shortcomings in sample throughput.

The instability of instrument performance is one of the main reasons why MS based shotgun proteomics has not been as widely adopted as might seem appropriate. From Professor Mann's lab's experience, the combination of HPLC and Orbitrap mass spectrometers deviate from the desired performance around once per week. A certain expertise and experience is required to resolve these issues [85] and presently most problems occur on the HPLC side [52]. For clinical applications it would be favorable, if all MS-instrumentation operated robustly guaranteeing reliable and reproducible results without the permanent checking of an MS expert. As proteomics becomes more attractive for clinical application, MS manufacturers will probably increase their efforts to develop robust instrumentation. MALDI-TOF MS serves as a positive example, but it should not be forgotten that it took decades to come up with an instrument that fits clinical needs so well. Therefore, it seems justified to be carefully optimistic that MS instrumentation suitable for non-expert clinical use will be developed in the next decade.

Finally, the complexity of the MS readout seems to prevent its present use for medical diagnostics. The simultaneous quantification of several hundred or thousand proteins is certainly more difficult to interpret than targeted analyses of known disease markers. However, the amount of information of a proteomics readout is likely to enable clinicians to diagnose diseases with greater certainty and assess a patients condition with considerably higher accuracy. In this respect, proteomics and genomics are much alike, though proteomics has the advantage that it reflects both nature and nurture. It will therefore be necessary to develop meaningful scores to reduce the information load and to draw conclusions from these data. This will require some bioinformatic effort and a lot of medical research to define how the insights of proteomics are best used to improve human health.

4.3 Perspectives for clinical proteomics

Given that MS based shotgun proteomics seems well on its way for clinical application, its potential benefits shall be discussed briefly. Deep MS-based analyses of the proteome could be very attractive for clinical microbiology. The characterization of the human oral microbiome by means of protoemics is among the first projects of this kind and there is still much room for improvement. For example, our approximate quantification of bacterial genera is by no means optimal. To come up with more accurate and reliable results a direct comparison of quantitative metaproteomic data to quantitative metagenomic data would be needed. This would allow to gauge different metaproteomics quantification strategies and compare them directly to well established methods.

From a clinical standpoint, it would be highly desirable to extend the scope of the mircobiome analysis down to the species or subspecies level. This is challenging given the huge dynamic range of protein abundance in such samples. However, the concentration of a pathogen in an infection is in most cases considerably higher than the concentration of bacteria from normal flora in healthy individuals. This suggests that the characterization of a pathogen from a infected tissue might be a lot easier. It is therefore well possible that shogun mass spectrometry will soon be able to characterize bacterial species down to the species level without the need to culture them first. Patients suffering from sepsis or severe pneumonia could substantially benefit from faster culture independent typization, because an early specific antibiotic therapy is crucial for patient outcome in these cases.

Likewise, the proteomics approach developed in this thesis is amenable to pathogens that cannot be cultured or grow slowly, i.e. mycobacteria. It would therefore be highly desirable that clinical samples from such infections are analyzed with MS based shotgun proteomics. This would allow to optimize the method for clinical needs and to evaluate its clinical utility.

In the future it might even be possible to directly quantify the proteins that convey antibiotic resistance. Given the rising resistance rates in many health care units fast and reliable antibiotic resistance testing is becoming increasingly important. The quantitative accuracy of proteins responsible for antibiotic resistance could be raised by targeting the respective peptide precursors specifically. Along these lines, it would even be possible to design dedicated MS methods for resistance testing. Since shotgun proteomics omits the cultivation of bacteria, this form of resistance testing is likely to reflect the in vivo situation more accurately than testing of bacterial cultures. In addition, the impact of antibiotic treatment on a pathogen and on the remaining human microbiome could be studied in detail with this technology. For example, the pathogenesis of pseudomembranous colitis from *Clostridium difficile* during antibiotic therapy could be investigated on the systems level. Understanding this process in detail might enable physicians to develop new strategies to prevent this undesirable complication of antibiotic therapy. It seems furthermore ideally suited to investigate the interactions of the human body with its microbiome.

The clinical potential of MS based shotgun proteomics is not limited to microbiology. The longitudinal profile of the weight reduction cohort from the project in the appendix revealed that a surprisingly large proportion of plasma proteins shows greater variation between individuals than in one individual over time. This suggests, that personalized medicine should seek to define reference values for individuals separately. Pathological deviations form homeostatic baseline

levels could thus be detected with considerably higher accuracy compared to deviations form a much broader window of reference values of a standard cohort. Ultimately, patient tailored diagnostic tests could make medicine more precise and increase the informative value of known biomarkers. Such patient specific baseline levels would have to be determined for a plethora of functional proteins, a task for which MS-based proteomics seems ideally suited.

Furthermore, the simultaneous changes of functionally related proteins are likely to increase diagnostic certainty and will help to resolve problems due to multiple hypothesis testing. The difficult interpretation of the proteomics readout could be simplified by calculating scores for diseases based on the levels of several proteins. Body fluid proteomics therefore seems ideally placed for routine clinical health checkups, because it could asses an individual's well being by analyzing thousands of functionally related and unrelated proteins and provide a comprehensive picture of the homeostasis of diverse organ systems and metabolic functions. The saliva proteome measurements revealed that such measurements should ideally take place at the same time of the day since protein levels change during the day. The minimal amounts of sample that are needed for proteomics measurements are another major advantage of this sort of measurement.

Yet, for all these potential benefits of clinical proteomics to be realized, more research will be necessary. We will learn a lot about the human body in the course of this quest and MS based proteomics is likely to play a major role in it.

A Multiplying sequencing speed using Trapped Ion Mobility Mass Spectrometry

A.1 The problem of inaccessible peptide species in data-dependent LC-MS/MS

An important prerequisite for the clinical application of MS-based proteomics is efficient measurements of proteomes at great depth. Despite huge progress in this area in the last years, even better coverage of the proteome would be desirable. Take the bacterial species in saliva as an example. If more low abundant proteins could be measured, some of the less abundant microorganisms would not escape detection and bacterial genera could potentially be differentiated down to the species level. Similarly, the human plasma proteome in the weight loss study described below covered around 437 of the most abundant plasma proteins. However, many important low abundant plasma proteins and peptides are missed, such as prostate specific antigen, troponins or thyroid-stimulating hormone. Consequently, technological progress would facilitate the clinical application of MS-based proteomics. The project described in the following establishes a new operation mode on a prototype mass spectrometer that could substantially increase sequencing speed and boost the depth of analysis in this way.

The determination and quantification of several thousand protein species from one sample cannot be achieved by measuring all the tryptic peptides simulatneously. Instead it has become a standard to separate peptides by liquid chromatography first and measure them successively. Hence, different peptide species are measured independently depending on their elution time. This way, peptides are separated in two dimensions with their HPLC elution time as one axis and their m/z ratio as the other axis. This is frequently illustrated by plotting the detected peptides in a m/z-retention time plane. Ideally, this plane is homogeneously populated by peptides so that the overlap of proteins is kept minimal along both axes. Many peptide species however elute simultaneously and cannot all be sequenced individually by MS/MS. Data suggests that only 16 % out of 100,000 peptide features eluting in the course of a 90 min HPLC gradient are targeted for MS/MS scans, leaving the majority of peptides undetected [86].

Despite advances in sequencing speed and resolving power [87] this situation has not changed much - partially, because the faster acquisition of MS/MS spectra decreases the number of ions detected per MS/MS scan. The instrument method sequential window acquisition of all theoretical mass spectra (SWATH) aims to overcome the problem of inaccessible precursors by cycling through consecutively broader mass windows and fragmenting all peptides per window in parallel [21]. A major disadvantage of this strategy is that it renders the previously discussed

multiplexing with iTRAQ or TMT impossible, because the reporter ion ratios of individual precursors are lost. Even if SWATH is not used and the mass selection window for MS/MS scans is kept small, the percentage of the precursor ion among all ions in the selection window can be small. This increases the complexity of peptide identification and limits quantification accuracy in multiplexing. It is therefore desirable to have a big precursor ion fraction among all ions in the selection window.

Trapped ion mobility spectrometry (TIMS) [88] is a new type of ion mobility mass spectrometer only described in 2015. In a nutshell, it is a mass spectrometer that measures an additional physical property of the analytes, their ion mobility. The ion motility is a measure for how fast an analyte travels in a carrier gas and is consequently a function of its collisional cross section. TIMS determines the ion mobility using an ion trap. This offers several advantages over conventional ion mobility spectrometry. In particular, it allows to fragment several peptides in parallel while still correctly assigning all fragments to their peptides of origin. This is possible, because the ion mobility of a peptide is determined in addition to its m/z ratio and its elution time. The described mode of operation was coined Parallel Accumulation-Serial Fragmentation (PASEF) and enables very accurate measurement of the reporter ion ratios since the peptides that are fragmented in parallel differ in their ion mobilities. In the following, PASEF will be explained and discussed in detail as well as its implementation on a TIMS prototype. Parts of this work were published in the *Journal of proteome research* [53].

A.2 Instrument characteristics of a TIMS-QTOF mass spectrometer

This subsection introduces the instrument prototype on which the PASEF method can be implemented. It differs substantially from conventional ion mobility spectrometers and offers additional set screws to optimize MS measurements.

The PASEF method was realized on a prototype TIMS-QTOF mass spectrometer, which is schematically depicted in figure A.13. TIMS is an advancement of ion mobility spectrometry (IMS) [88]. Compared to normal MS, IMS separates molecules based on their ion mobility in addition to the separation by hydrophobicity on the HPLC and the separation by mass. Traditionally, this comes at the cost of losing ions along the long ion mobility tunnel.

By reversing the flow of the drift gas and trapping the ions along an increasing electric field the ion mobility can be determined in a compact setup with better ion control. To this end, peptides are separated via HPLC and ionized via ESI before entering the instrument through a glass capillary (fig. A.13 A). They are subsequently deflected by 90° to get rid of uncharged molecules and enter

the TIMS tunnel. While the ions are confined radially by a radio-frequency quadrupolar field, they are dragged forward by the atmospheric drift gas (fig. A.13 B). An increasing electric field strength directed in the opposite direction prevents the ions from immediately passing the tunnel and causes ions with high mobilities to accumulate at the entrance of the tunnel and ions with low mobility at the exit of the tunnel (fig. A.13 B). Once enough ions are accumulated the electric potential along the gradient is reduced over time and the ion packages elute from the TIMS tunnel with low mobility ions eluting first. The eluting ions are kept in focus and directed to an analytical quadrupole that can be used as collision cell and to select ion species for MS/MS scans. Subsequently, the ions are accelerated orthogonally into the field free drift region of the flight tube of the TOF mass analyzer. The mass analyzer is essentially the same as in the Bruker impact II instrument [20].

The potential of the TIMS tunnel therefore changes periodically. During a full period length of this cycle several TOF spectra are recorded at a frequency of 8.7 kHz. All the TOF scans that are measured in such a period length together define a ‚TIMS scan‘. The signal to noise ratio of an individual TIMS scan can be insufficient, hence requiring to add several TIMS scans to come up with a reasonable TIMS-MS spectrum. Such a spectrum displays the signal intensity of an ion species in the ion mobility-m/z-plane. At present, the precursor selection is based on the intensity of the ions in the m/z spectrum regardless of their ion mobility. Hence, for the precursor selection all TOF scans from one TIMS scan are summed and the top N ions are selected for fragmentation.

The selection of ions in the TIMS device has two advantages compared to normal MS/MS spectra. First, although there are several different ion species in the m/z selection window, these ion species are separated in ion mobility. Consequently, the precursor ion fraction can be very high when only the ions with the ion mobility of the precursor peptide are considered. This advantage is not exclusive to TIMS, but occurs in normal IMS as well and it is well established that it renders benefits for peptide identification [89]. If one only uses the reporter ions in the ion mobility window of the precursors, quantification in multiplexing experiments is hence very accurate. This feature of TIMS is called ratio compression.

The second advantage arises from the accumulation of ions in the TIMS tunnel. Since ions are accumulated, their signal intensity is very high when they elute. Instead of recording the ion signal continuously, the TIMS tunnel compresses the signal of an ion into the short time interval when it elutes, resulting in a better signal to noise ratio.

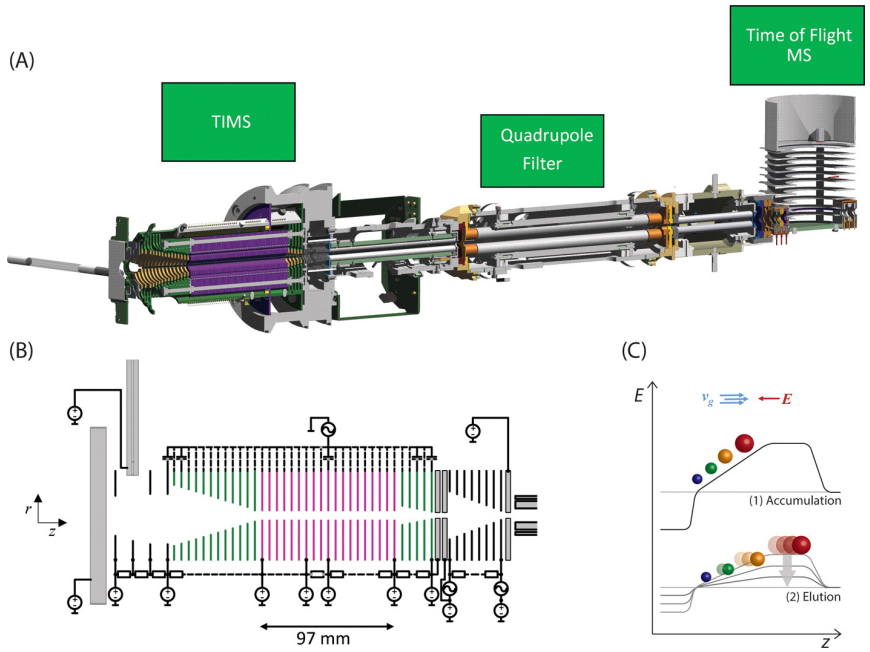


Figure A.13: **TIMS-QTOF mass spectrometer:** A: The TIMS-QTOF consists of three main components, the TIMS tunnel, the quadrupole filter and the TOF mass analyzer. The ions enter the instrument through the entrance funnel on the left, exit the TIMS tunnel in a controlled manner depending on their ion mobility, pass through the quadrupole filter and end up in the flight tower. B: The TIMS cell itself consists of the green entrance funnel on the left, the purple TIMS tunnel, where the ions are accumulated, and the green exit funnel on the right. The tiny bars in the profile represent ring electrodes in a quadrupolar arrangement that are coupled to one another by resistors. This setup allows to realize the linear electric potential in (C) along the z axis while confining the ions in the x-y-plane via an oscillating quadrupolar field. C: First, ions are trapped in the TIMS tunnel by a forward directed gas flow and a linearly rising potential that exerts an opposing electric force (1). Subsequently, the ions serially elute from the TIMS tunnel depending on their ion mobility as the electric filed decreases (2). Adapted from [53].

A.3 Principle of Parallel Accumulation-Serial Fragmentation

The separation of ions in the MS/MS selection window by ion mobility is already beneficial. However, the precursor of interest only elutes in a very small ion mobility window and most of the other TOF scans in the respective TIMS scan can be of little value, especially if they contain only low intensity signals. For the described case, it is more efficient to rapidly change the selection window so that different precursors of different ion mobility can be selected within one TIMS scan. We call this operation mode ‚Parallel accumulation - serial fragmentation‘ (PASEF). A scheme of this operation mode is depicted in figure A.14.

In PASEF several precursors are measured in one TIMS scan. Since the m/z values and the ion mobilities of the precursors are known from the TIMS-MS spectrum, a serial precursor selection of ion mobility and m/z windows takes place with the limitation that the ion mobility windows must not overlap. Also, PASEF requires very short switching times of the selection quadrupole. In this way, the full signal intensities of several precursors are measured in a single TIMS scan, since the precursor intensities are compressed into tiny non overlapping ion mobility windows by the accumulation in the TIMS tunnel. This efficient use of available ions can be used in two ways. Either the speed of MS/MS acquisition is increased by the number of targeted precursors per TIMS scan or the sensitivity is increased by the same factor by targeting the same precursors several times. One can also use a combination of the gains in sensitivity and speed.

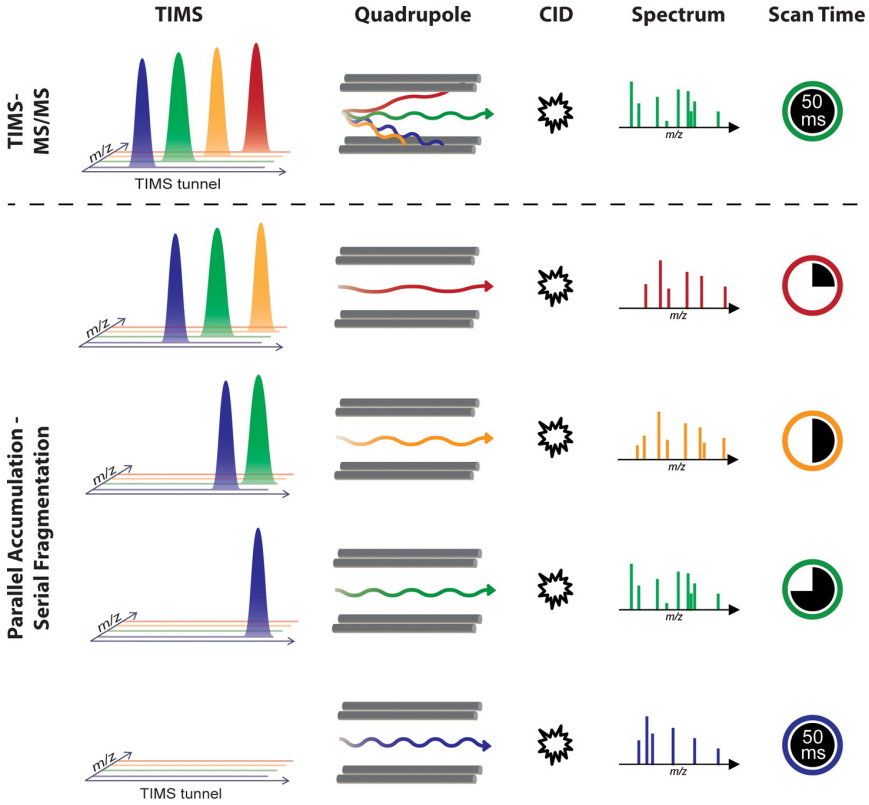


Figure A.14: **Comparison of normal TIMS-MS/MS with PASEF:** The normal TIMS-MS/MS method is illustrated in the top panel. One precursor is selected for fragmentation with collision induced dissociation (CID) in the entire TIMS scan. The lower panel shows the serial selection of different isolation windows for precursors of different ion mobility in a single TIMS scan. This way all four available ions can be used for fragmentation in one TIMS scan. Adapted from [53].

A.4 Realization of Parallel Accumulation-Serial Fragmentation

The PASEF method requires a rapid determination of the precursors that are to be selected for MS/MS scans and a fast switching of the quadrupole. The chromatographic peak width for 90-min gradients in shotgun proteomics is around 7 s in Professor Mann's laboratory as determined by the full width at half-maximum (FWHM) definition [20]. The acquisition of the MS spectrum takes around 0.2 s, the acquisition of one MS/MS scan takes 60 ms for a data-dependent top17 method on our QTOF instruments. Thus, the entire acquisition cycle for a top 17 method amounts to 1.3 s. Since an individual TOF scan only lasts 110 μs, each MS or MS/MS spectrum consists of the sum of several hundred TOF scans.

These figures show, that the selection quadrupole in the QTOF instrument switches its isolation window approximately every 60 ms. Given that ion mobility peaks in our TIMS tunnel have half widths (FWHM) of around one millisecond, the PASEF method requires the selection quadrupole to operate on a much shorter timescale. The build-up time of the power supplies enables to change the isolation window in the desired m/z range within less than 1.5 ms. To capture the entire ion mobility peak at the present resolution of the TIMS tunnel an isolation time of 2.5 to 4 ms is adequate. Consequently, up to 12 - 20 times more MS/MS spectra should be measurable using PASEF without any loss in sensitivity. It seems unlikely that enough different precursors exist to use all these additional MS/MS for the identification of additional precursors. Instead they could be employed to improve identification and quantification by measuring the same precursor several times to obtain better spectral quality and more accurate intensities.

Presently, the instrument controller is unable to calculate and apply the respective isolation windows at the desired speed of less than 1.5 ms. We still wanted to demonstrate the PASEF method in a proof of principle experiment on our present set-up. For this reason we first determined the desired isolation window and elution time from a TIMS precursor scan and then applied the switching times and the isolation window separately via the instrument controller.

In the future, the real time field-programmable gate array will calculate and set the respective TIMS and isolation windows in real time on the fly. This requires improvements in the accurate peak interpretation and 4D precursor determination in real time (fig. A.15).

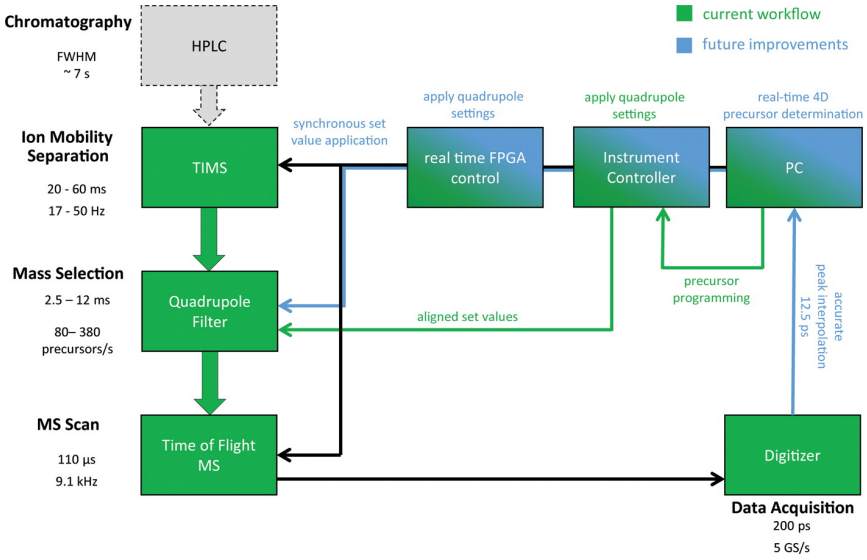


Figure A.15: **Hardware requirements for PASEF method:** The major hardware components of a protoeomics LC-TIMS-MS/MS setup and their connections are depicted schematically. The green and blue flow lines represent exisiting and future implementations that are required to realize the indicated time scales. GS/s, Giga-samples per second. Adapted from [53].

A.5 Parallel Accumulation-Serial Fragmentation in the analysis of a complex peptide mixture

Though the hardware requirements for the real time operation of PASEF remain to be realized, we aimed to demonstrate the benefits of this method for the analysis for complex peptide mixtures in our proof of principle experiment. We therefore injected a mixture of digested alcohol dehydrogenase (ADH), bovine serum albumin (BSA), enolase and phosphorylase b by direct infusion to mimic the simultaneous elution of several peptides from the HPLC. Without the separation of these peptides by their ion mobility, the resulting mass spectrum would be very complex. The resulting two dimensional TIMS-MS spectrum is depicted in figure A.16 A. Note that the m/z spectrum without the separation in ion mobility is simply the projection of the horizontal ion mobility axis on the vertical m/z axis. Due to their higher collision cross section, heavier particles elute earlier and show lower mobility values. Consequently, m/z and ion mobility are inversely correlated to some extent. Furthermore, differently charged ions populate different regimes of the TIMS-MS spectrum in figure A.16 A. This correlation implies that a reasonably high resolution in ion mobility is necessary for the separation in ion mobility to yield additional information.

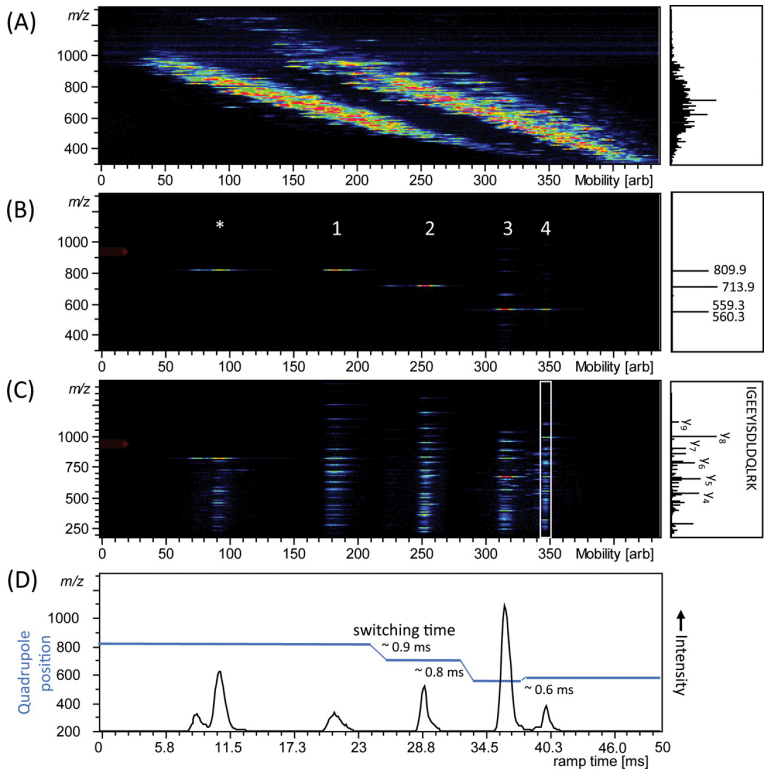


Figure A.16: **TIMS-QTOF measurement of a directly infused complex peptide mix from digested ADH, BSA, phosphorylase b and enolase:** A: TIMS-MS spectrum from a full MS scan. The projection on the right represents the mass spectrum obtained without mobility separation. B: Serial isolation of four precursors with non-overlapping ion mobility peaks. C: PASEF of the four precursors in B. The resulting fragemnts remain separated in ion mobility and the detected intensities correspond to the intensities registered in the normal TIMS operation mode. D: The arrival time distribution of the summed fragments from C reveals that more non overlapping ion mobility windows could have been chosen. Adapted from [53].

In our experiment we set the TIMS tunnel time to ramp down the TIMS tunnel gradient to 50 ms. This was sufficient to realize a resolution (FWHM) of above 40 that separated many ions of equal m/z. Next, we selected four precursors with m/z 810.3 (1), 714.3 (2), 559.3 (3) and 560.6 (4) and selected the corresponding switching times for the precursor selection by hand. The resulting PASEF scan in figure A.16 B shows the isolation of the entire mobility peak of each of the four precursors following their parallel accumulation. Note that precursors (3) and (4) only differ in one Thomson, hence they would probably end up in the precursor isolation window of one another. Their separation in ion mobility allows to differentiate these two precursors. The feature marked with an asterisk represents two singly charged ion species that belong to the singly charged ion

population and were only isolated, because the precursor isolation window was set to the isolation of feature (1) from the beginning.

Subsequently, we determined the MS/MS spectra of the four selected precursors using PASEF. As figure A.16 C demonstrates, the ion mobilites are preserved despite the fragmentation of the precursors. The fragment spectra allowed us to identify all our precursors, namely VLGIDGGEGKEELFR (1) from enolase, HLQIIYEINQR (2), VAAAFPGDVDR (3) and IGEEISDLDQLRK (4) from phosphorylase B. Figure A.16 D shows that the respective fragment ion peaks were only around 3 ms wide allowing to select more than ten precursors during the 50 ms ramp time.

We also wondered whether these results would hold true, if we selected more different precursors across the entire mass range from 418 to 956 Th and across the entire mobility range with elution times ranging from 17 to 44 ms with an average half width of 0.8±0.2 ms. To this end we chose ten sets consisting of four precursors each and performed the experiment from above on all sets. The results are summarized in Table A.1 in the appendix. The full signal intensity was recovered compared to the normal TIMS method even though the acquisition of MS/MS spectra was four fold faster.

A.6 Relevance of Parallel Accumulation-Serial Fragmentation for clinical proteomics

Ion mobility spectrometry combined with a mass spectrometer provides an additional separation dimension. Consequently, it could be of great value for the analysis of complex protein mixtures such as human tissue or body fluids. We introduced the PASEF method that is compatible with any ion mobility mass spectrometer with a scan speed in the sub-millisecond range. In our compact TIMS-QTOF analyzer, we managed to select several precursors for fragmentation in a single 50 ms ion mobility scan, hence demonstrating PASEF for the first time.

Since the signal intensity was not decreased due to the serial fragmentation, PASEF could increase the sequencing speed around 10-fold without loss in quantitative accuracy. These gains in sequencing speed are probably best used by a combination of targeting more precursors and targeting low intensity precursors several times. By contrast to other methods that target several precursors simultaneously, such as SWATH, PASEF is fully data dependent. This implies that the mass isolation window targets a single precursor. PASEF is therefore suitable for chemical labeling strategies such as iTRAQ or TMT. This would allow to multiplex the analysis of patient samples. For PASEF to be used for

entire proteomics samples, several improvements in the hardware communication of our TIMS-QTOF analyzer remain to be implemented. Once these challenges are met, PASEF is likely to render faster and more in depth analysis of patient proteomics samples possible.

A.7 Materials and Methods of Parallel Accumulation-Serial Fragmentation

A profound description of the TIMS analyzer used has been published elsewhere [88, 90]. We used nitrogen as a bath gas at room temperature and the pressure difference between the entrance funnel and the exit funnel was 1 mbar creating a gas flow velocity of around $1.5 \cdot 10^2$ m/s. The ion accumulation time was set to 50 ms in the in instrument control software OtofControl (Bruker Daltonik, Bremen, Germany). The electric field gradient was realized by setting the entrance of the TIMS tunnel to - 180 V and the exit to -40 V. The time for decreasing the slope of the ramp was set to 435 TOF scans of 115 μs each, resulting in around 50 ms.

Purified, digested standard solutions of ADH, BSA, enolase and phosphorylase b from Waters GmbH (Eschborn, Germany) were solved separately in 0.1 % formic acid and diluted to a concentration of 10 pmol/μL each. These solutions were mixed to form an equimolar solution that was subsequently diluted in 50 % water/50 % acetonitrile/0.1 % formic acid (v/v/v) to a final concentration of 100 fmol/μL. This solution was directly infused into our TIMS-QTOF prototype at a flow rate of 3 μL/min to simulate the simultaneous elution of several peptides from an HPLC.

The precursor masses for PASEF were determined from a full TIMS-MS scan. Subsequently, the appropriate isolation windows, switching times and collision energy were manually set at the instrument controller. We chose 3 Th for the isolation window with and collision energies between 30 to 60 eV depending on the m/z of the respective precursor. The resulting ion mobility-mass spectra were analyzed in a prototype software of Bruker's DataAnalysis.

B Effects of sustained weight loss on the human plasma proteome

B.1 Evaluating weight loss effects on the body from a systems perspective

The prevalence of severe obesity in the Western world is increasing at a dramatic pace. In US children, numbers have climbed from 4 % during 1999-2004 to 6 % during 2011-2012 [91, 92]. This represents a major public health burden, giving rise to serious secondary diseases such as type 2 diabetes, hypertension or dyslipidemia among many others [93, 94]. The best treatment for obese patients suffering from the metabolic syndrome is sustained weight reduction [95]. Yet, the underlying mechanisms of the benefit of weight loss in obese patients are insufficiently understood and so is the extent to which they vary between individuals [96]. The metabolic health status of an individual is commonly evaluated by plasma metabolites and plasma proteins such as cholesterol, blood glucose, diverse lipoproteins, C-reactive protein or HbA1c [97]. Given that cholesterol and triglyceride levels correlate well with plasma proteins like apolipoprotein A1 and B, plasma proteomics seems uniquely positioned to study the processes during weight loss from a systems perspective. Although it is well established that weight loss leads to substantial changes of plasma proteins like the sex hormone-binding globulin [98], the impact of weight loss on the plasma proteome has not been assessed globally, but only protein by protein with antibody based methods.

The unbiased quantitative account of a multitude of plasma proteins by MS-based proteomics has fascinated researchers and clinicians since the early days of this technology [99]. However, the huge dynamic range of plasma makes it particularly challenging to analyze the plasma proteome at sufficient coverage for an entire study population (compare subsection 2.2 and [100]). The substantial progress in MS technology in recent years [101, 102] motivated Professor Mann's laboratory to develop an automated, relatively high-throughput workflow that allows the reproducible, quantitative analysis of hundreds of plasma proteomes [54]. In this project, we wanted to evaluate the utility of our workflow by measuring the proteomes of 43 individuals at seven time points over 14 months from a longitudinal prospective cohort study on sustained weight loss [103]. This represents the first global analysis of the impact of weight loss on the plasma proteome.

B.2 Study design and plasma proteomics analysis

The study aimed to assess the effect of weight loss on the human plasma protein. An existing randomized controlled trial on 52 obese individuals seemed ideally suited for that purpose. It originally investigated the effect of GLP-1 receptor agonist treatment on maintaining body weight loss and free leptin levels [103]. Briefly, 52 healthy individuals with body mass index (BMI) 30 - 40 kg/m^2 aged between 18 and 65 years were recruited for the study. Subjects completed an 8-week weight loss program (800 kcal per day; Cambridge Weight Plan, Corby, UK [104]), loosing 12 % body weight on average. Following the weight loss phase, participants were randomized into two groups, 27 received the GLP-1 receptor agonist liraglutide 1.2 mg per day and the remaining 25 persons served as control group. Both groups were monitored for one year with a total of 13 visits with the study dietitian. Each of the 43 participants that adhered to the study protocol for the entire study period donated fasting blood samples at week -8, week 0, week 4, week 13, week 26, week 39 and week 52 (fig. B.17 A). Since neither the original study nor our analysis revealed significant differences between the treatment and control group we did not differentiate between the two study arms in our analysis.

This resulted in 319 plasma samples each measured in quadruplicates to obtain highly accurate label-free protein quantification. In order to increase the depth of our proteomic coverage, we created a MS plasma proteome library that enables the identification of precursors not targeted for fragmentation. This is achieved by comparing MS spectra recorded at equal elution times between runs and checking whether the precursors in questions were targeted for MS/MS in other runs [58]. We therefore took blood from three healthy male and female volunteers and depleted these samples twice with antibodies for the top 20 most abundant plasma proteins. Subsequent sample processing was identical to the study samples. Note that this depletion leads to substantial distortions of relative plasma protein levels due to unspecific binding. It is therefore not suitable to obtain an unbiased quantification of the plasma proteome. It is however well suited for the creation of a deep library that improves precursor identification.

Altogether, the study measurements and the creation of the plasma proteome library amounted to 1294 MS measurements - to our knowledge the largest plasma proteomics study in a clinical framework. It took us 10 weeks to complete the data acquisition with a measurement time of 30 minutes per sample if no complications occurred. The reproducibility of our measurements was high with a mean correlation coefficient R^2 of 0.97 for the quadruplicate measurements. We identified 737 plasma proteins across all study participants excluding potential

contaminants such as keratins. On average 437 proteins were detected per individual with a standard deviation of 23.

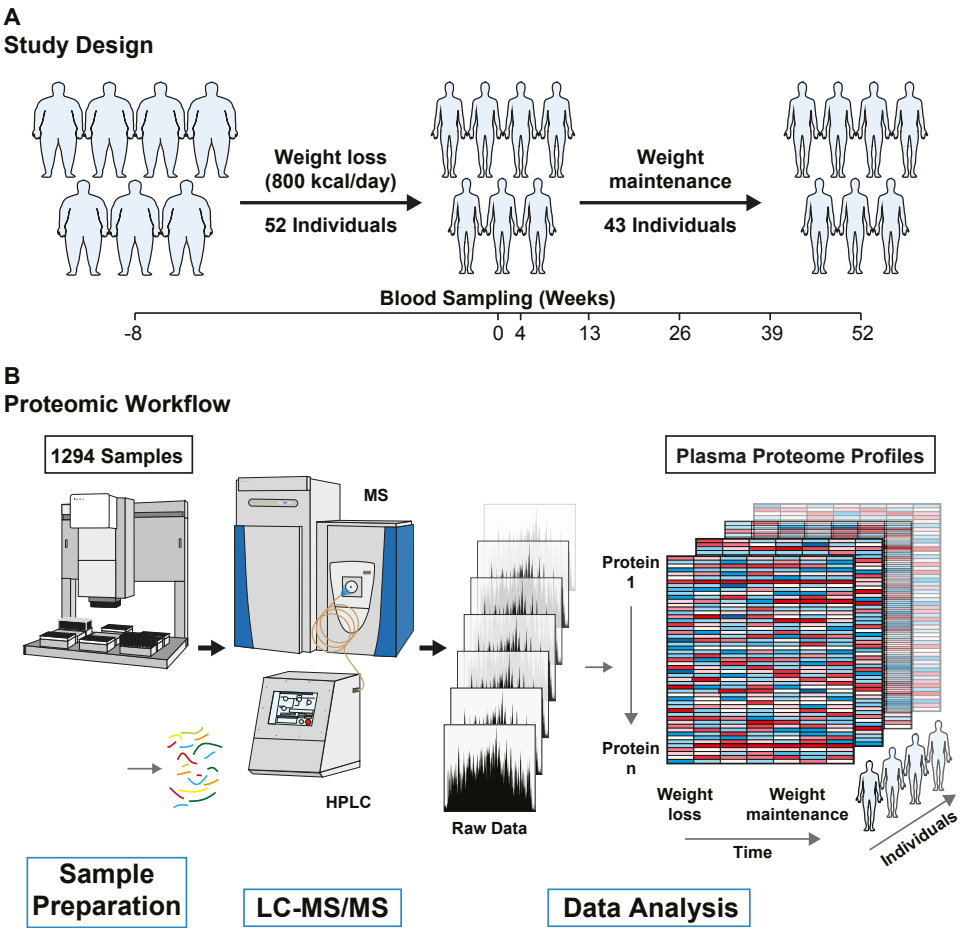


Figure B.17: **Study design and plasma proteomics workflow:** A: A total of 52 healthy obese study participants lost 12 % of their body weight on average during an 8 week period of caloric restriction. This acute weight loss was followed by a weight maintenance period of 52 weeks by 43 individuals. Blood samples from each participant were taken at the indicated timepoints. B: An automated liquid handling platform prepared quadruplicates of all samples for LC-MS/MS measurement on a Q-Exactive HF mass spectrometer. We optimized the data analysis in MaxQuant with a matching library and created plasma proteome profiles for all 52 individuals. Adapted from [52].

B.3 Interindividual variation of the plasma proteome

The five follow up plasma samples per donor after the acute weight loss lead us to investigate how constant plasma protein levels are in a one-year period and

how they vary between individuals. We compared the plasma protein levels of all 448 proteins that were quantified at all five time points for at least one individual across all participants. Remarkably, 25 % of these proteins differed more than five-fold from the group average and 69 % proteins differed more than two-fold (fig. B.18 A).

A

	Percentage of all proteins			
Fold difference	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

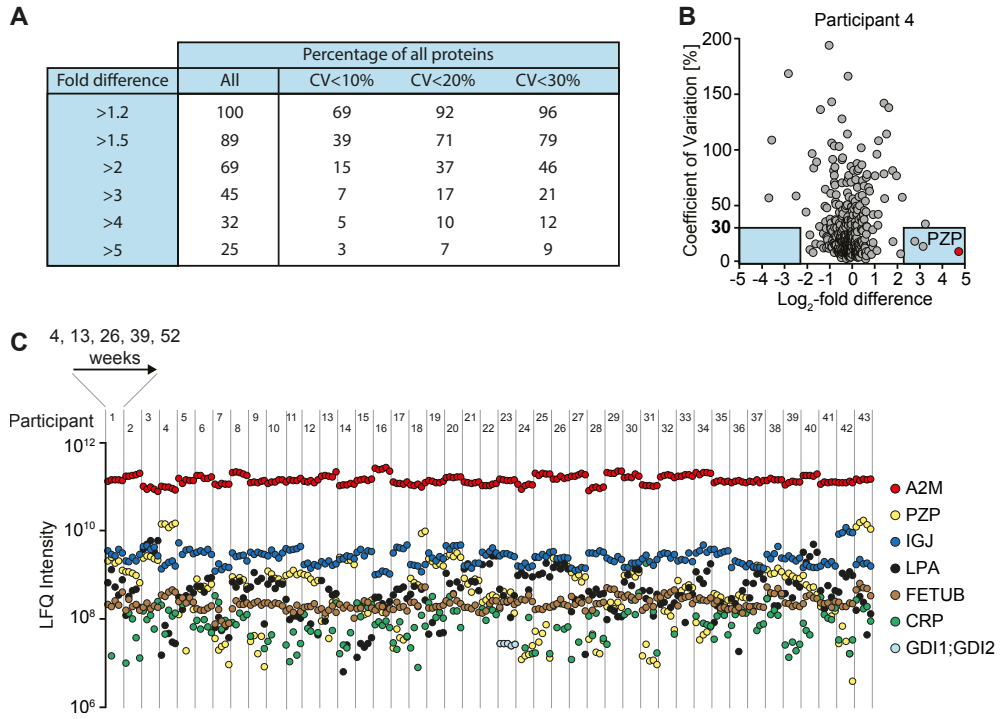


Figure B.18: **Personalized plasma protein levels across the five longitudinal plasma samples:** A: We calculated the coefficients of variation (CVs) for all proteins across all five post weight loss samples in all participants. The percentage of proteins below the certain CV thresholds are given. B: CVs of quantified proteins across the five post weight loss samples from participant four plotted against the fold change difference compared to the average label free quantification (LFQ) intensity of the study cohort. The blue boxes contain proteins with a CV below 30 % and a fold change above 5, hence satisfying our criteria for individual specific proteins. C: LFQ Intensities of seven highly individual specific plasma proteins of all 43 participants across all five longitudinal samples. 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. Adapted from [52].

While some of these interindividual differences are explained by inaccurate quantification of low abundant proteins, numerous proteins showed substantial interindividual differences despite low CVs across the five measurement time points. It turns out that 46 % of all proteins across all participants deviate by more

than two fold from the group average while having CVs below 30 % across the five samples. Proteins fulfilling these criteria for a certain participant will be referred to as individual specific for this individual in the remainder of this thesis. The concept of individual specific proteins can be illustrated by pregnancy zone protein (PZP) that was 26 times more abundant in donor 4 than in the group average but remained constant over one year with a CV below 10 % (fig. B.18 B). Other typical individual specific proteins detected in all participants included α-2-macroblobuline, lipoprotein (a), immunoglobulin J chain, fetuin-B and C-reactive protein (fig. B.18 C). In the case of lipoprotein(a), the variation is genetically explained by the difference in the number of kringle domains that reduce secretion into the bloodstream [105]. Other proteins hardly varied across individuals or time indicating that a strict control of their levels is crucial for body homeostasis. These proteins were serum albumin, complement factor C3, vitamin D-binding protein, kininogen-1, hemopexin, complement factor H and clusterin among others.

B.4 Impact of weight loss on the plasma proteome

The follow up blood donations in the study allowed to investigate both the acute and the long term effects of weight reduction on the plasma proteome. To evaluate the immediate effect of weight loss on the plasma proteome, we calculated a one-sample t-test with 5 % FDR after Benjamini-Hochberg correction of all detected proteins comparing the samples of week -8 to the samples of week 0.

It revealed that 63 proteins were significantly decreased and 30 were significantly increased upon weight loss (fig. B.19 A). The fold changes were moderate compared to stimulation experiments in cell culture as expected given the homeostatic control of blood constituents. For example, levels of Albumin, the main plasma protein, were relatively increased by 8 %. Apolipoprotein F (APOF) and Inter-α-trypsin inhibitor heavy chain H3 (ITIH3) changed more drastically with an increase of 37 % and 34 % respectively. The most significant reduction in relative protein abundance upon weight loss was recorded for Pigment-epithelium derived factor (SERPINF1) that changed by -16 % and is known to be secreted by adipocytes [106]. By contrast, the greatest significant increase in relative abundance was observed for sex hormone binding globulin (SHBG) with 117 %. This is in accordance with previous analyses targeting SHBG specifically [98]. Importantly, there were always exceptions of these regulatory changes for a hand full of individuals, suggesting that complex metabolic transformations are best studied on the systems level rather than studying just one single biomarker (fig. B.19 B).

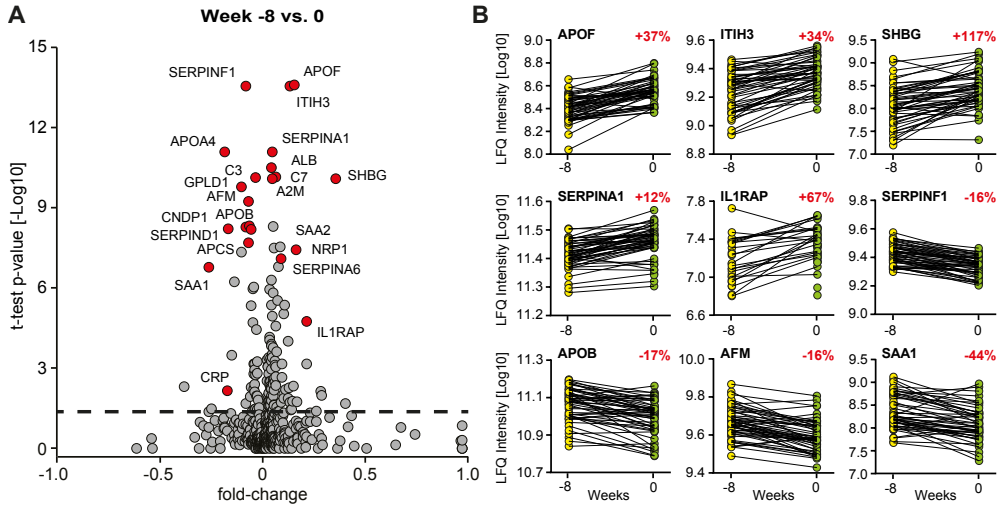


Figure B.19: **Personalized plasma protein levels:** A: t-test p-values plotted against fold change before (week -8) and directly after weight loss (week 0) of plasma proteins (5 % FDR after Benjamini-Hochberg correction). B: Label free quantification (LFQ) changes on an individual basis are displayed for certain proteins with the median change indicated in the top right corner in red. APOF: Apolipoprotein F, ITIH3: Inter-α-trypsin-inhibitor heavy chain H3, SHBG: Sex hormone binding globulin, SERPINA1: α-1-antitrypsin, IL1RAP: Interleukin-1 receptor accessory protein, SERPINF1: Pigment-epithelium derived factor, APOB: Apolipoprotein B-100, AFM: Afamin, SAA1: Serum amyloid A-1 protein. Adapted from [52].

Previous investigations discovered that the levels of freely circulating cortisol decrease upon weight loss [98]. Although we did not determine steroid hormones with our method, we still observed indirect signs of this change in the form of a 12 % increase of the Corticosteroid-binding globulin (SERPINA6), which binds 80 % of circulating cortisol. In addition, the increase in Albumin is likely to contribute to the reduction in freely circulating cortisol.

Next, we investigated the chronic effects of sustained weight loss on the plasma proteome. We found 84 proteins that were changed with high significance ($p < 5 \cdot 10^{-4}$) between the baseline level and at least one timepoint after weight loss in the one year weight maintenance period. They clustered into seven groups depending on their adaptation to the acute weight loss (fig. B.20).

Proteins from the first group markedly decreased in response to the weight loss, but steadily turned back to normal in the weight maintenance period. Hence, they seem to remain in a steady state irrespective of the body weight, but decrease upon a sudden reduction in energy supply.

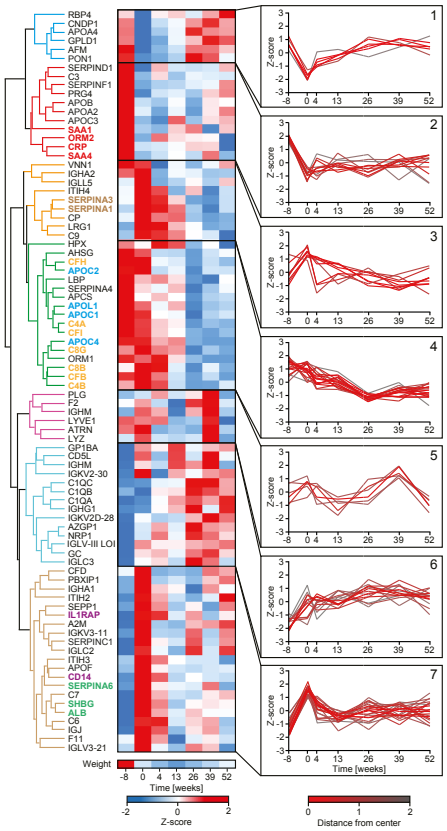


Figure B.20: **Long-term impact of sustained weight loss on the plasma proteome:** Hierarchical clustering of Z-scored label free quantification intensities for highly singnificant proteins ($p < 0.0005$). The time course of protein abundance of the seven resulting clusters is dispalyed in the inlets with color coding for the distance from the center of the respective cluster. The color coding of the gene names on the left mark inflammatory proteins in red, serine protease inhibitors in brown, members of the complement system in orange, apolipoproteins in blue, anti inflammatory proteins in purple and steroid transporting proteins in green. Adapted from [52].

They included APOA4, Afamin, a vitamin E transporting protein and Serum paraoxonase (PON1). PON1 is known to be less active in obese children and adults [107].

The second group of proteins decreased upon weight loss and maintained at a low level throughout the one year follow up period. It consisted mainly of inflammatory proteins (CRP, SAA1, SAA4 and ORM2) and apolipoproteins (APOA2, APOB, APOC3), both of which will be discussed in detail in the next subsection.

Similar to the second group, the proteins in group four decreased upon weight reduction, but it took them several months to arrive at a new state of equilibrium, now on a lower abundance level. The clustering is biologically plausible, because related proteins ended up in the same or adjacent groups indicating that they were collectively regulated. Interestingly, the levels of neurophilin-1 (NRP1) increased almost twofold in the first year after acute weight loss and so did the levels of Vitamin D-binding protein. By contrast, the proteogylcan 4 (PRG4), that occurs in joints decreased by -19 % on average and heparin cofactor 2 (SERPIND1), a thrombin inhibitor, by -9 %.

B.5 Effects of weight loss on the apolipoprotein profile

Apolipoproteins facilitate the transport of insoluble fats in the body. Consequently, their levels are of major importance to diagnose and monitor lipo-

metabolic disorders that are highly prevalent worldwide and a major contributor to cardiovascular morbidity [108]. Our plasma measurements revealed longitudinal profiles of 18 different lipoproteins (fig. B.21). LPA levels showed the largest increase of 95 % upon the acute weight loss and maintained elevated throughout the observational period. Members of the Apolipoprotein C family (APOC1, APOC2, APOC4) collectively decreased permanently to about 70 % of their base level. Other lipoproteins such as APOF responded with an increase to the acute weight loss, but reverted to base levels in the course of the one year maintenance period. Similarly, APOA4 acutely decreased by 36 %, but also returned to baseline within a year.

Given this differential regulation of different lipoproteins, we subsequently correlated the observed changes with the changes of cholesterol, triglyceride, glucose, high density lipoprotein (HDL), low density lipoprotein (LDL) levels and BMI. APOB strongly correlated with LDL ($R^2 = 0.72$) as expected since a LDL forms around one APOB molecule. The question whether APOB might therefore be a more accurate measure for cardiovascular risk assessment and prevention than LDL is the subject of ongoing debates [109]. Likewise, the levels of APOA1 were highly correlated with HDL levels ($R^2 = 0.64$). Furthermore, APOF was negatively correlated with triglyceride levels ($R^2 = -0.50$) and APOB, APOC2, APOC3, APOC4 and APOE were positively correlated with triglycerides ($R^2 = 0.33$, 0.45, 0.52, 0.38 and 0.45 respectively).

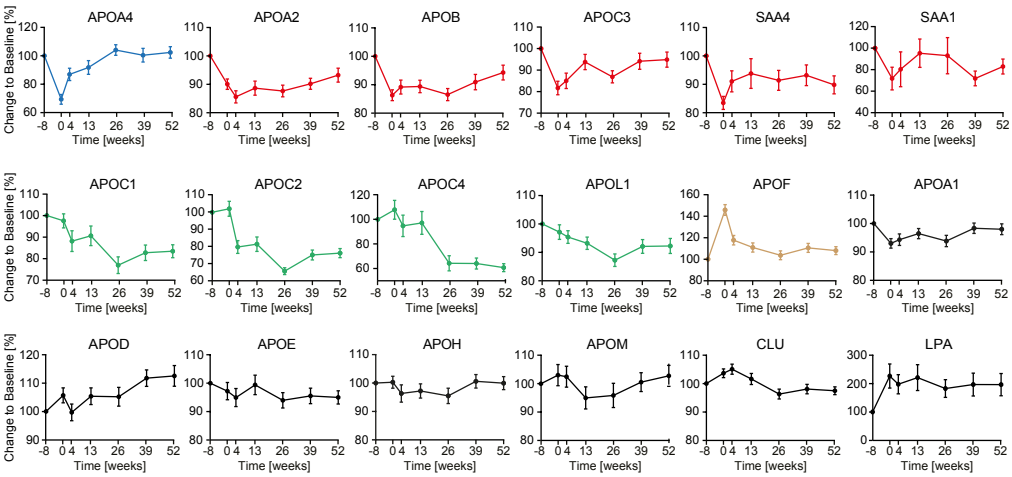


Figure B.21: **Longitudinal apolipoprotein profiles of acute weight loss:** Mean normalized label free quantification (LFQ) intensities for 18 apoliporproteins are plotted across the entire observation period. LFQ values were normalized to the initial levels before weight loss (week -8) and the bars represent the standard error of the mean. The coloring refers to the clusters in figure B.18. Adapted from [52].

B.6 Effects of weight loss on inflammatory proteins

Weight loss resulted in a marked reduction of inflammatory proteins, in particular acute phase proteins. CRP, SAA1, SAA4 and ORM2 all clustered in group 2 of our cluster analysis, meaning that they were immediately reduced following acute weight loss (compare fig. B.20). The reduction of the two cardiovascular risk markers SAA1 and CRP was particularly pronounced with -43 % and -35 % respectively, compared to the slower reduction of ORM1, Serum amyloid P-component (APCS) and Lipopolysaccharide-binding protein (LBP) (-16 %, -10 % and -16 % respectively at week 52). Other acute phase proteins such as Alpha-1-antitrypsin (SERPINA1) and Alpha-1-antichymotrypsin (SERPINA3) were upregulated immediately after weight loss, but declined in the weight maintenance period.

Subsequently, we determined the correlation of plasma protein levels to clinical parameters upon weight loss (fig. B.22 A-F). Remarkably, the five factors that showed the most significant correlation with BMI were Complement Factor H (CFH), C3, APCS, ORM2 and CFI, all inflammatory proteins. Likewise, other inflammatory proteins including CRP, SAA4, ORM1, Attractin (ATRN) and CFB also correlated with BMI. Though the link between inflammation and obesity is well established [110] this analysis of the plasma proteome provides one of the most comprehensive analyses of the impact of weight loss on inflammation. Alongside known inflammatory factors in obesity like the dipeptidase ATRN [111], the results show that the described multitude of inflammatory factors responds to weight reduction, suggesting a systematic reduction in inflammation.

Consequently, we wanted to come up with a systemic inflammation profile for all study participant across the observation period. To this end, we only considered highly significantly changed proteins (compare fig. B.20) for the Uniprot keywords „acute phase“, „immunity“ and „inflammatory response“. We also added the two known acute phase proteins APCS and LBP to our profile list and excluded the anti-inflammatory protein CD14. Consequently, we arrived at 24 inflammation-related proteins, 20 of which decreased in intensity in the observation period. Among these, we chose to include all ten proteins that were significantly correlated with BMI into the protein panel of our inflammation profile. We therefore calculated the standard score for each protein in the panel over the individual time series and performed a hierarchical clustering on the level of study participants. The resulting cluster and the heat map of the longitudinal inflammation profiles for all 42 individuals is displayed in figure (fig. B.22 G).

Before weight reduction the levels of inflammatory proteins in our panel were high for almost all participants. While some profiles and proteins in the panel

were immediately reduced following the weight loss, others only decreased gradually. The trend however is clear across almost all individuals and all proteins in the panel: Weight reduction leads to a marked decrease in the levels of the inflammatory proteins in the panel. The study participants clustered in three main groups. The first group consisting of seven individuals (top in fig. B.22 G) showed a strong increase in inflammatory proteins in weeks 4 or 13 after weight loss, probably due to infections following the metabolic changes. The members of this group did not regain weight in this time period. For example, participant 31 from this group had a 24-fold increase in CRP and a 54-fold increase in SAA1 in week 13 compared to her normal levels of these inflammation markers, pointing towards an infection as the cause. The main cluster comprised 72 % of the study participants and is therefore likely to represent the common inflammatory response to weight loss. Figure B.22 H depicts the changes of the panel proteins in this group with the median across all panel proteins and individuals shown in black for each time point. Judged by figure B.22 H it seems that inflammation mostly decreases during weight loss and in the first months of weight maintenance followed by constantly reduced inflammation levels.

We also wanted to assess for each individual whether inflammation reduced following the body weight reduction. We did so by calculating the mean of all Z-scored ten panel proteins for each individual for each time point and evaluating the resulting trend. 39 of the 42 individuals had a negative slope over the observation period corresponding to beneficial effect on inflammation. Two of the three individuals, that had a positive slope, actually gained weight in the observation period and the third individual had elevated inflammation markers in week 13 and 26, probably due to an infection. Taken together the data support the theory, that weight reduction can reduce the low grade inflammation observed in obese individuals. The simultaneous, unbiased measurement of a multitude of inflammatory molecules in our study provides us with a comprehensive and systematic evaluation of inflammatory alterations accompanying and following weight loss.

B.7 Clinical significance of plasma proteomics changes upon weight loss

The study investigated the effects of eight weeks of weight reduction followed by one year of weight maintenance in 43 obese individuals by state of the art MS-based proteomics. The automated sample preparation workflow allows to quantitatively measure more than thousand samples in a matter of weeks paving the way for the large scale analysis of clinical studies. This high throughput did not come at the cost of a drastic reduction of proteome coverage, because the 400

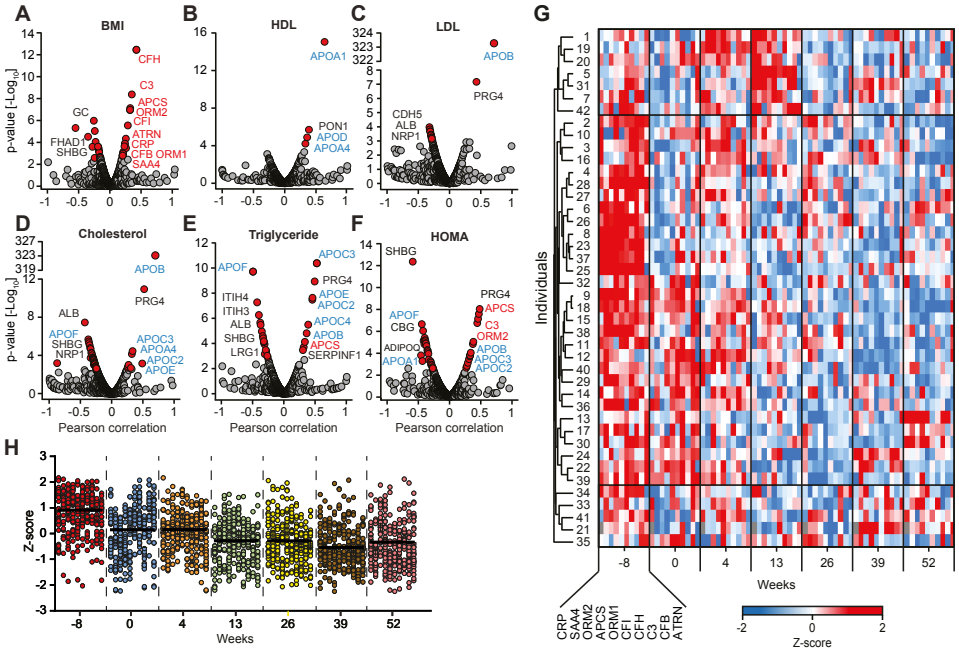


Figure B.22: **Longitudinal profile of inflammatory proteins following acute weight loss:** A-F: Correlation of the indicated clinical parameters with all quantified plasma proteins plotted against p-values. G: Longitudinal inflammation profiles for all 42 participants across the entire observation period as judged by a panel of ten inflammatory proteins. The MS-intensities of the ten panel proteins were Z-transformed for each individual and a hierarchical clustering was calculated for the individual longitudinal inflammation profiles. H: Dot plot of the panel proteins in the same order as in G for the central cluster. The median Z-score of the panel per time point is indicated by the black line. Adapted from [52].

routinely quantified proteins comprised all clinically relevant lipoproteins, many markers of low grade inflammation and many other functional plasma proteins.

A previous plasma proteomics study with a clinical focus relied on the depletion of highly abundant plasma proteins and multiplexing to obtain a reasonable proteomic coverage of 190 proteins per sample [112]. While multiplexing appears to be a promising strategy to further enhance throughput, the unspecific binding of depletion kits compromises quantitative accuracy. From our quadruplicate measurements of each sample, we were able to conclude that our protein quantification was highly reliable. Among other things the longitudinal analysis of such a large cohort revealed that a surprisingly large proportion of plasma proteins shows greater variation between individuals than in one individual over time. The acute weight loss had a profound impact on the levels of multiple plasma proteins and while some protein levels changed permanently others reverted to their initial baseline following a short adaptation period. In several cases the

studied proteins had previously been associated with obesity like CRP, S100-A9 [112] or SERPINF1 [106, 113], but their response to weight loss had not been studied systematically. Our results on the reduction of SERPINF1 suggest that it might be a useful indicator for fat mass reduction as opposed to lean mass reduction.

Most inflammatory proteins and acute phase proteins in particular showed a marked reduction upon weight loss. Prominent examples were CRP (-35 % reduction) and SAA1 (-44 % reduction) both of which are associated with increased risk for cardiovascular diseases. Their longitudinal behavior clustered together with functionally related proteins, but also some proteins with no previously known functional connections, suggesting that some of these proteins might be useful for the assessment of cardiovascular morbidity. Another prominent group of proteins that was substantially altered by weight loss were apolipoproteins. We recorded the expected correlations between LDL, HDL and triglycerides with their respective apolipoproteins. Combinations of these protein levels might render fruitful clinical parameters to evaluate the risk of cardiovascular diseases.

Both the inflammatory markers and the apolipoprotein profiles of individuals suggest, that proteomics might be of clinical use to stratify patients based on their cardiovascular risks and their potential benefit from weight loss. This could contribute to a better, more personalized treatment of patients suffering from the metabolic syndrome.

B.8 Materials and Methods of the plasma proteome project

The study design is described in detail in [103]. Briefly, 58 individuals with BMI between 30 and 40 kg/m^2 were enrolled for the study. The exclusion criteria comprised chronic illness other than obesity, medical treatment affecting glucose or lipid metabolism, appetite or food intake, pregnancy or breast feeding and fasting glucose levels above 7 mmol/l. All participants kept a very low calorie diet (800 kcal per day; Cambridge Weight Plan, Corby, UK [104]) for eight weeks resulting in a weight loss of at least 7.5 %. This period of acute body weight reduction was followed by a weight maintenance period of one year with a daily energy uptake of 600 kcal below the estimated need. Since the study originally aimed to test the incretin mimetic drug liraglutide, half the study cohort received 1.2 mg liraglutide after weight loss. Given that both groups were equally successful in maintaining the weight loss and we could not find significant differences between their plasma proteomes, we did not distinguish between the two groups in our analyses. Blood samples were collected before weight loss (week -8), immediately after (week 0) and at weeks 4, 13, 26, 39 and 52 after weight loss.

The ethical committee in Copenhagen approved the study (reference number: H-4-210-134) and it was conducted in accordance with the Helsinki Declaration II and ICH-GCP practice. All participants were volunteers and had the possibility to retract their consent at any time. ClinicTrials.gov identifier: NCT02094183.

Our sample preparation protocol and instrument settings were previously described in [54]. An EAYS-nLC 1000 from Thermo Fisher Scientific was used to separate the purified peptides on a 18-min HPLC gradient prior to the transfer to a Q Exactive HF Orbitrap (Thermo Fisher Scientific) with identical settings to [54]. For the construction of our matching library the 20 most abundant plasma proteins were depleted from the samples of three healthy men and three healthy women using two immunodepletion kits successively according to the manufacturer's protocol (Agilent Multiple Affinity Removal Spin Cartridge and ProteoPrep20 Plasma Immunodepletion Kit). Upon depletion the samples were digested and measured in triplicates with identical settings as the non-depleted samples of the weight loss study, except for the use of 45-min gradient instead of the 18-min gradient.

MS spectra were analyzed in MaxQuant version 1.5.3.23 [11] using the homo sapiens Uniprot FASTA database from June 2015. We filtered for contaminants with the predefined contaminants database of the Andromeda search engine and imposed a FDR of 0.01 on all identified peptides and proteins as described above. The match between runs feature was used to optimize MS identification of peptides in the short runs from our deep plasma proteomics matching library. Label-free protein quantification by the MaxLFQ algorithm [37] used a minimum ratio count of 1.

Downstream analysis was carried out in Perseus (version 1.5.2.12) [59]. All hypothesis tests were corrected for multiple hypothesis testing via Benjamini-Hochberg using a FDR of 5 %. A detailed description of the 1 D annotation enrichment analysis can be found in [60].

C Supplementary material

C.1 Additional figures of the saliva proteome project

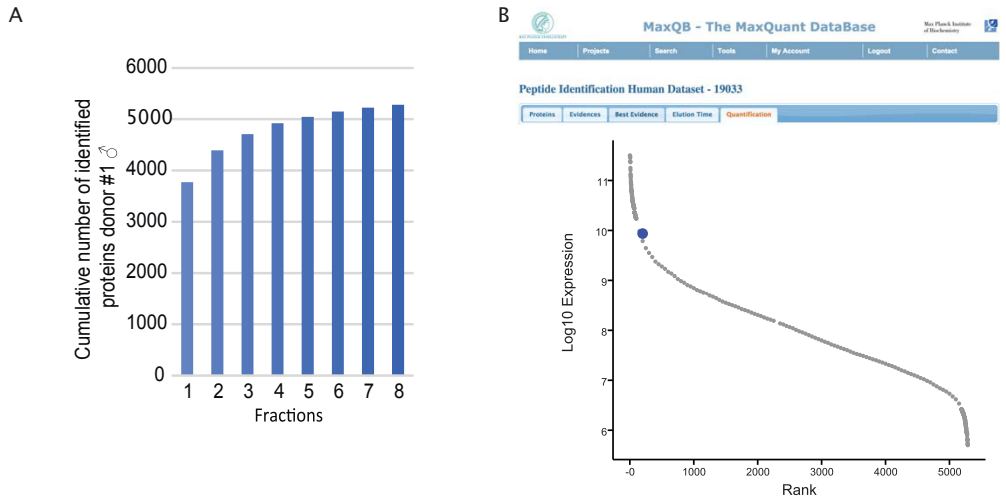


Figure C.1: **Cumulative benefit of fractionation and MaxQB:** A: Cumulative number of identified human saliva proteins from male donor 1 from eight basic reverse phase fractions. B: MaxQB protein rank plot for Transcobalamin 1, an important transporter of Vitamin B12. Adapted from [51].

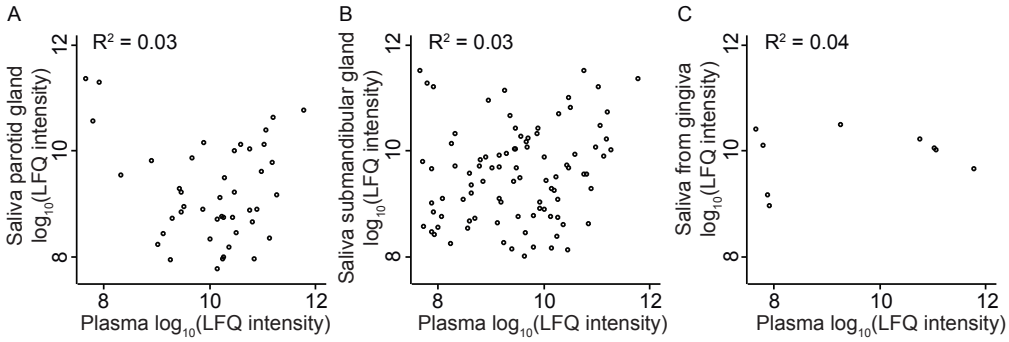


Figure C.2: **Correlation of plasma proteome and saliva proteome for specific sites:** Scatter plots of LFQ values of shared plasma proteins and slaiva proteins form the parotid gland (A), the opening of the submandibular and sublingual duct (B) and from gingiva (C). Adapted from [51].

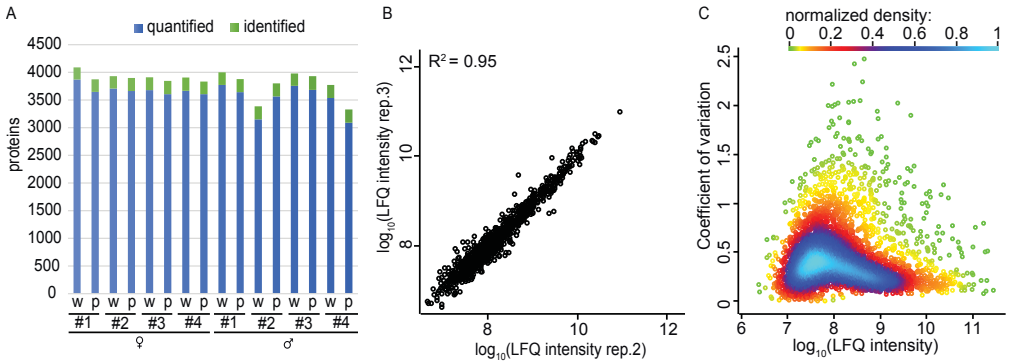


Figure C.3: **Single shot measurements of eight donors at two timepoints:** A: Number of quantified or only identified human proteins from single shot measurements of the waking (w) and the postprandial (p) saliva samples of our eight donors. B: Reproducibility of LFQ intensities for two workflow replicates. C: CVs of LFQ protein intensities of our 16 single shot measurements plotted against protein abundance. Adapted from [51].

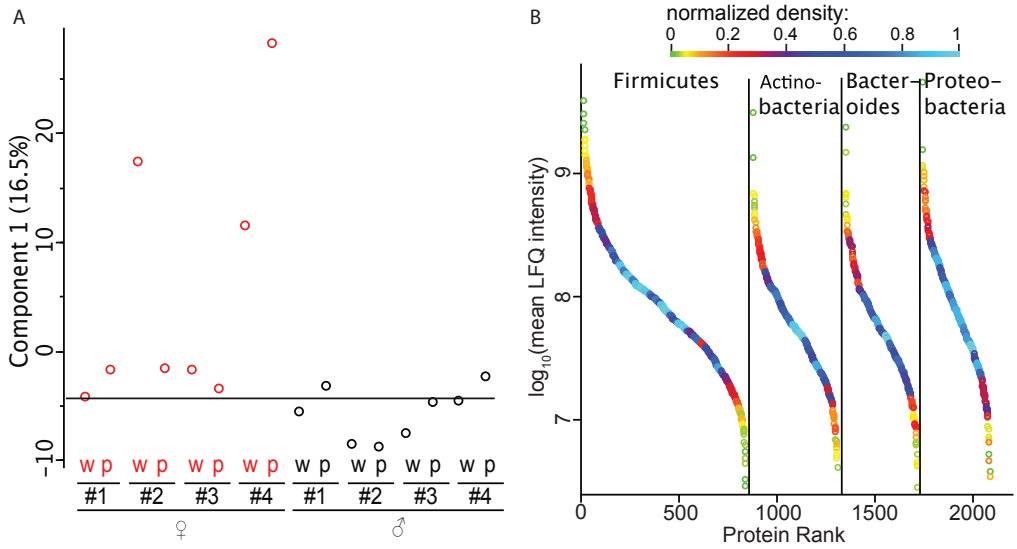


Figure C.4: **PCA of the human saliva proteome and dynamic range plot of bacterial phyla:** A: Component 1 of the PCA of our eight waking (w) and eight postprandial (p) saliva samples separates samples weakly based on sex. B: Dynamic range plot of the bacterial proteins for the four most abundant phyla. The protein density is color coded. Adapted from [51].

C.2 PASEF analysis of 40 precursors from a complex peptide mixture

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

Table A.1: **PASEF Analysis of 40 peptides from an equimolar mixture of ADH, BSA, Enolase and Phosphorylase B:** Ten sets of four precursors each were selected for PASEF Scans in analogy to the experiment in figure A.16 . [a] Median summed fragment ion intensities were extracted for each precursor from PASEF (N = 331) and TIMS-MS/MS (N = 9) scans with identical quadrupole isolation settings. [b] As an artifact resulting from the asynchronous operation of TIMS and quadrupole, these two precursors were not isolated in each PASEF scan. Adapted from [53].

References

[1] N.C. Mishra and G. Blobel. Introduction to Proteomics: Principles and Applications. *Methods of Biochemical Analysis. Wiley*, 2011. (↑ p. 1)

[2] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Paabo, and M. Mann. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, 7:548, 2011. (↑ p. 2)

[3] L. M. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther, and M. Mann. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254, Oct 2008. (↑ pp. 2 and 38)

[4] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, Jul 1995. (↑ p. 2)

[5] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, 4(9):117, 2003. (↑ p. 2)

[6] N. Nagaraj, N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm, and M. Mann. 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(3):M111.013722, Mar 2012. (↑ p. 2)

[7] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon. The one hour yeast proteome. *Mol. Cell Proteomics*, 13(1):339–347, Jan 2014. (↑ pp. 2 and 38)

[8] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014. (↑ pp. 2 and 7)

[9] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, May 2014. (↑ pp. 2 and 7)

[10] Adam Bonislawski. As field closes in on human proteome, issues of methodology, validation take stage at hupo 2014. *genomeweb*, Okt 2014. (↑ p. 2)

[11] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26(12):1367–1372, Dec 2008. (↑ pp. 3, 6, 11, 17, 18, and 65)

[12] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.*, 60(20):2299–2301, Oct 1988. (↑ p. 3)

[13] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, Oct 1989. (↑ p. 3)

[14] H. Steen and M. Mann. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, 5(9):699–711, Sep 2004. (↑ p. 4)

[15] N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, and M. Mann. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Meth-*

ods, 11(3):319–324, Mar 2014. (↑ pp. 4, 16, 23, and 38)

[16] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev*, 17(1):1–35, 1998. (↑ p. 4)

[17] J. C. Schwartz, M. W. Senko, and J. E. Syka. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.*, 13(6):659–669, Jun 2002. (↑ p. 4)

[18] A. Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.*, 72(6):1156–1162, Mar 2000. (↑ p. 5)

[19] R. A. Scheltema, J. P. Hauschild, O. Lange, D. Hornburg, E. Denisov, E. Damoc, A. Kuehn, A. Makarov, and M. Mann. 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(12):3698–3708, Dec 2014. (↑ p. 17)

[20] S. Beck, A. Michalski, O. Raether, M. Lubeck, S. Kaspar, N. Goedecke, C. Baessmann, D. Hornburg, F. Meier, I. Paron, et al. The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics. *Mol. Cell Proteomics*, 14(7):2014–2029, Jul 2015. (↑ pp. 6, 45, and 48)

[21] L. C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics*, 11(6):O111.016717, Jun 2012. (↑ pp. 5 and 43)

[22] S. Eliuk and A. Makarov. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu Rev Anal Chem (Palo Alto Calif)*, 8:61–80, 2015. (↑ p. 5)

[23] D. Helm, J. P. Vissers, C. J. Hughes, H. Hahne, B. Ruprecht, F. Pachl, A. Grzyb, K. Richardson, J. Wildgoose, S. K. Maier, H. Marx, M. Wilhelm, I. Becher, S. Lemeer, M. Bantscheff, J. I. Langridge, and B. Kuster. Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell Proteomics*, 13(12):3709–3715, Dec 2014. (↑ p. 6)

[24] U. Distler, J. Kuharev, P. Navarro, Y. Levin, H. Schild, and S. Tenzer. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods*, 11(2):167–170, Feb 2014. (↑ p. 6)

[25] A. Keller and D. Shteynberg. Software pipeline and data analysis for MS/MS proteomics: the trans-proteomic pipeline. *Methods Mol. Biol.*, 694:169–189, 2011. (↑ p. 6)

[26] H. Weisser, S. Nahnsen, J. Grossmann, L. Nilse, A. Quandt, H. Brauer, M. Sturm, E. Kenar, O. Kohlbacher, R. Aebersold, et al. An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res.*, 12(4):1628–1644, Apr 2013. (↑ p. 6)

[27] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, Mar 2007. (↑ p. 7)

[28] A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods*, 4(10):787–797, Oct 2007. (↑ p. 7)

[29] J. B. Fenn. Ion formation from charged droplets: Roles of geometry, energy, and time. *J. Am. Soc. Mass Spectrom.*, 4(7):524–535, Jul 1993. (↑ p. 7)

[30] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*,

1(5):376–386, May 2002. (↑ p. 7)

[31] J. L. Hsu, S. Y. Huang, N. H. Chow, and S. H. Chen. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.*, 75(24):6843–6852, Dec 2003. (↑ p. 8)

[32] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics*, 3(12):1154–1169, Dec 2004. (↑ p. 8)

[33] A. Thompson, J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed, and C. Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, 75(8):1895–1904, Apr 2003. (↑ p. 8)

[34] P. J. Boersema, R. Raijmakers, S. Lemeer, S. Mohammed, and A. J. Heck. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc*, 4(4):484–494, 2009. (↑ p. 8)

[35] J. A. Paulo, J. D. O'Connell, R. A. Everley, J. O'Brien, M. A. Gygi, and S. P. Gygi. Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 S. cerevisiae proteins across 10 carbon sources. *J Proteomics*, 148:85–93, Jul 2016. (↑ p. 8)

[36] G. C. McAlister, E. L. Huttlin, W. Haas, L. Ting, M. P. Jedrychowski, J. C. Rogers, K. Kuhn, I. Pike, R. A. Grothe, J. D. Blethrow, et al. Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.*, 84(17):7469–7478, Sep 2012. (↑ p. 8)

[37] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteomics*, 13(9):2513–2526, Sep 2014. (↑ pp. 8, 17, and 65)

[38] J. P. Anhalt and C. Fenselau. Identification of bacteria using mass spectrometry. *Analytical Chemistry*, 47:219–225, Feb 1975. (↑ p. 8)

[39] R. D. Holland, J. G. Wilkes, F. Rafii, J. B. Sutherland, C. C. Persons, K. J. Voorhees, and J. O. Lay. Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, 10(10):1227–1232, 1996. (↑ p. 8)

[40] M. A. Claydon, S. N. Davey, V. Edwards-Jones, and D. B. Gordon. The rapid identification of intact microorganisms using mass spectrometry. *Nat. Biotechnol.*, 14(11):1584–1586, Nov 1996. (↑ p. 8)

[41] A. Mellmann, J. Cloud, T. Maier, U. Keckevoet, I. Ramminger, P. Iwen, J. Dunn, G. Hall, D. Wilson, P. Lasala, et al. Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry in comparison to 16S rRNA gene sequencing for species identification of nonfermenting bacteria. *J. Clin. Microbiol.*, 46(6):1946–1954, Jun 2008. (↑ p. 8)

[42] G. Marklein, M. Josten, U. Klanke, E. Muller, R. Horre, T. Maier, T. Wenzel, M. Kostrzewa, G. Bierbaum, A. Hoerauf, et al. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. *J. Clin. Microbiol.*, 47(9):2912–2917, Sep 2009. (↑ p. 8)

[43] P. E. Fournier, M. Drancourt, P. Colson, J. M. Rolain, B. La Scola, and D. Raoult. Modern clinical microbiology: new challenges and solutions. *Nat. Rev. Microbiol.*, 11(8):574–585, Aug 2013. (↑ p. 9)

[44] A. Freiwald and S. Sauer. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nat Protoc*,

4(5):732–742, 2009. (↑ p. 9)

[45] L. Ferreira, F. Sanchez-Juanes, M. Gonzalez-Avila, D. Cembrero-Fucinos, A. Herrero-Hernandez, J. M. Gonzalez-Buitrago, and J. L. Munoz-Bellido. Direct identification of urinary tract pathogens from urine samples by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.*, 48(6):2110–2115, Jun 2010. (↑ p. 9)

[46] M. Inigo, A. Coello, G. Fernandez-Rivas, B. Rivaya, J. Hidalgo, M. D. Quesada, and V. Ausina. Direct Identification of Urinary Tract Pathogens from Urine Samples, Combining Urine Screening Methods and Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. *J. Clin. Microbiol.*, 54(4):988–993, Apr 2016. (↑ p. 9)

[47] S. Q. van Veen, E. C. Claas, and E. J. Kuijper. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. *J. Clin. Microbiol.*, 48(3):900–907, Mar 2010. (↑ p. 9)

[48] M. Kostrzewa, K. Sparbier, T. Maier, and S. Schubert. MALDI-TOF MS: an upcoming tool for rapid detection of antibiotic resistance in microorganisms. *Proteomics Clin Appl*, 7(11-12):767–778, Dec 2013. (↑ p. 10)

[49] D. J. Ecker, R. Sampath, C. Massire, L. B. Blyn, T. A. Hall, M. W. Eshoo, and S. A. Hofstadler. Ibis T5000: a universal biosensor approach for microbiology. *Nat. Rev. Microbiol.*, 6(7):553–558, Jul 2008. (↑ p. 10)

[50] M. Mann, N. A. Kulak, N. Nagaraj, and J. Cox. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell*, 49(4):583–590, Feb 2013. (↑ p. 10)

[51] N. Grassl, N. A. Kulak, G. Pichler, P. E. Geyer, J. Jung, S. Schubert, P. Sinitcyn, J. Cox, and M. Mann. Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome Med*, 8(1):44, 2016. (↑ pp. 11, 13, 23, 25, 28, 30, 31, 32, 33, 36, 67, and 68)

[52] P. E. Geyer, N. J. Wewer Albrechtsen, S. Tyanova, N. Grassl, E. W. Iepsen, J. Lundgren, S. Madsbad, J. J. Holst, S. S. Torekov, and M. Mann. Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.*, 12(12):901, Dec 2016. (↑ pp. 11, 39, 55, 56, 58, 59, 60, and 63)

[53] F. Meier, S. Beck, N. Grassl, M. Lubeck, M. A. Park, O. Raether, and M. Mann. Parallel Accumulation-Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.*, 14(12):5378–5387, Dec 2015. (↑ pp. 11, 44, 46, 47, 49, 50, and 69)

[54] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, and M. Mann. Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst*, 2(3):185–195, Mar 2016. (↑ pp. 13, 15, 19, 24, 26, 53, and 65)

[55] J. Rappsilber, M. Mann, and Y. Ishihama. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc*, 2(8):1896–1906, 2007. (↑ p. 16)

[56] T. Chen, W. H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, and F. E. Dewhirst. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)*, 2010:baq013, Jul 2010. (↑ pp. 17 and 28)

[57] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, 10(4):1794–1805, Apr 2011. (↑ p. 17)

[58] N. Nagaraj, N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning,

O. Vorm, and M. Mann. 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(3):M111.013722, Mar 2012. (↑ pp. 18 and 54)

[59] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods*, 13(9):731–740, Sep 2016. (↑ pp. 19 and 65)

[60] J. Cox and M. Mann. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics*, 13 Suppl 16:S12, 2012. (↑ pp. 19, 24, and 65)

[61] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011. (↑ pp. 20 and 34)

[62] J. C. Silva, M. V. Gorenstein, G. Z. Li, J. P. Vissers, and S. J. Geromanos. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell Proteomics*, 5(1):144–156, Jan 2006. (↑ pp. 20 and 34)

[63] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, Aug 2014. (↑ p. 20)

[64] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010. (↑ p. 20)

[65] T. Ding and P. D. Schloss. Dynamics and associations of microbial community types across the human body. *Nature*, 509(7500):357–360, May 2014. (↑ p. 13)

[66] V. Tremaroli and F. Backhed. Functional interactions between the gut microbiota and host metabolism. *Nature*, 489(7415):242–249, Sep 2012. (↑ p. 13)

[67] J. L. Round, S. M. Lee, J. Li, G. Tran, B. Jabri, T. A. Chatila, and S. K. Mazmanian. The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science*, 332(6032):974–977, May 2011.

[68] K. Atarashi, T. Tanoue, T. Shima, A. Imaoka, T. Kuwahara, Y. Momose, G. Cheng, S. Yamasaki, T. Saito, Y. Ohba, et al. Induction of colonic regulatory T cells by indigenous Clostridium species. *Science*, 331(6015):337–341, Jan 2011.

[69] K. Berer, M. Mues, M. Koutrolos, Z. A. Rasbi, M. Boziki, C. Johner, H. Wekerle, and G. Krishnamoorthy. Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. *Nature*, 479(7374):538–541, Nov 2011. (↑ p. 13)

[70] P. Belda-Ferre, L. D. Alcaraz, R. Cabrera-Rubio, H. Romero, A. Simon-Soro, M. Pignatelli, and A. Mira. The oral metagenome in health and disease. *ISME J*, 6(1):46–56, Jan 2012. (↑ p. 13)

[71] N. Delaleu, P. Mydel, I. Kwee, J. G. Brun, M. V. Jonsson, and R. Jonsson. High fidelity between saliva proteomics and the biologic state of salivary glands defines biomarker signatures for primary Sjögren's syndrome. *Arthritis Rheumatol*, 67(4):1084–1095, Apr 2015. (↑ p. 13)

[72] J. M. Yoshizawa, C. A. Schafer, J. J. Schafer, J. J. Farrell, B. J. Paster, and D. T. Wong. Salivary biomarkers: toward future clinical and diagnostic utilities. *Clin. Microbiol. Rev.*, 26(4):781–791, Oct 2013. (↑ p. 13)

[73] N. Nagaraj and M. Mann. Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. *J. Proteome Res.*, 10(2):637–645, Feb 2011. (↑ p. 24)

[74] B. Cuevas-Cordoba and J. Santiago-Garcia. Saliva: a fluid of study for OMICS. *OMICS*, 18(2):87–97, Feb 2014. (↑ p. 26)

[75] C. Schaab, T. Geiger, G. Stoehr, J. Cox, and M. Mann. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell Proteomics*, 11(3):M111.014068, Mar 2012. (↑ p. 26)

[76] A. Hunt, D. Harrington, and S. Robinson. Vitamin B12 deficiency. *BMJ*, 349:g5226, Sep 2014. (↑ p. 26)

[77] R. Carmel, R. Green, D. W. Jacobsen, K. Rasmussen, M. Florea, and C. Azen. 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.*, 70(5):904–910, Nov 1999. (↑ p. 26)

[78] R. Carmel, S. Brar, and Z. Frouhar. Plasma total transcobalamin I. Ethnic/racial patterns and comparison with lactoferrin. *Am. J. Clin. Pathol.*, 116(4):576–580, Oct 2001. (↑ p. 26)

[79] C. Kirschbaum and D. H. Hellhammer. Salivary cortisol in psychoneuroendocrine research: recent developments and applications. *Psychoneuroendocrinology*, 19(4):313–333, 1994. (↑ p. 27)

[80] T. Opperman and J. P. Richardson. Phylogenetic analysis of sequences from diverse bacteria with homology to the Escherichia coli rho gene. *J. Bacteriol.*, 176(16):5033–5043, Aug 1994. (↑ p. 29)

[81] S. Schubert, J. L. Sorsa, S. Cuenca, D. Fischer, C. A. Jacobi, and J. Heesemann. HPI of high-virulent Yersinia is found in E. coli strains causing urinary tract infection. Structural, functional aspects, and distribution. *Adv. Exp. Med. Biol.*, 485:69–73, 2000. (↑ p. 29)

[82] J. A. Aas, B. J. Paster, L. N. Stokes, I. Olsen, and F. E. Dewhirst. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.*, 43(11):5721–5732, Nov 2005. (↑ p. 34)

[83] E. M. Bik, C. D. Long, G. C. Armitage, P. Loomer, J. Emerson, E. F. Mongodin, K. E. Nelson, S. R. Gill, C. M. Fraser-Liggett, and D. A. Relman. Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J*, 4(8):962–974, Aug 2010. (↑ p. 34)

[84] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, Jun 2012. (↑ p. 35)

[85] A. G. Paulovich, D. Billheimer, A. J. Ham, L. Vega-Montoto, P. A. Rudnick, D. L. Tabb, P. Wang, R. K. Blackman, D. M. Bunk, H. L. Cardasis, et al. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell Proteomics*, 9(2):242–254, Feb 2010. (↑ p. 39)

[86] A. Michalski, J. Cox, and M. Mann. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.*, 10(4):1785–1793, Apr 2011. (↑ p. 43)

[87] C. D. Kelstrup, R. R. Jersie-Christensen, T. S. Batth, T. N. Arrey, A. Kuehn, M. Kellmann, and J. V. Olsen. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J. Proteome Res.*, 13(12):6187–6195, Dec 2014. (↑ p. 43)

[88] K. Michelmann, J. A. Silveira, M. E. Ridgeway, and M. A. Park. Fundamentals of trapped ion mobility spectrometry. *J. Am. Soc. Mass Spectrom.*, 26(1):14–24, Jan 2015. (↑ pp. 44 and 52)

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

4160, Aug 2010. (↑ p. 45)

[90] F. A. Fernandez-Lima, D. A. Kaplan, and M. A. Park. Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev Sci Instrum*, 82(12):126106, Dec 2011. (↑ p. 52)

[91] J. A. Skelton, S. R. Cook, P. Auinger, J. D. Klein, and S. E. Barlow. Prevalence and trends of severe obesity among US children and adolescents. *Acad Pediatr*, 9(5):322–329, 2009. (↑ p. 53)

[92] A. C. Skinner and J. A. Skelton. Prevalence and trends in obesity and severe obesity among children in the United States, 1999-2012. *JAMA Pediatr*, 168(6):561–566, Jun 2014. (↑ p. 53)

[93] Evaluation National Cholesterol Education Program (NCEP) Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation*, 106(25):3143–3421, Dec 2002. (↑ p. 53)

[94] N. Stefan, F. Schick, and H. U. Haring. Ectopic fat in insulin resistance, dyslipidemia, and cardiometabolic disease. *N. Engl. J. Med.*, 371(23):2236–2237, Dec 2014. (↑ p. 53)

[95] P. R. Schauer, D. L. Bhatt, J. P. Kirwan, K. Wolski, S. A. Brethauer, S. D. Navaneethan, A. Aminian, C. E. Pothier, E. S. Kim, S. E. Nissen, et al. Bariatric surgery versus intensive medical therapy for diabetes–3-year outcomes. *N. Engl. J. Med.*, 370(21):2002–2013, May 2014. (↑ p. 53)

[96] The Look AHEAD Research Group. Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. *N. Engl. J. Med.*, 369(2):145–154, Jul 2013. (↑ p. 53)

[97] N. Esser, S. Legrand-Poels, J. Piette, A. J. Scheen, and N. Paquot. Inflammation as a link between obesity, metabolic syndrome and type 2 diabetes. *Diabetes Res. Clin. Pract.*, 105(2):141–150, Aug 2014. (↑ p. 53)

[98] M. Azrad, B. A. Gower, G. R. Hunter, and T. R. Nagy. Intra-abdominal adipose tissue is independently associated with sex-hormone binding globulin in premenopausal women. *Obesity (Silver Spring)*, 20(5):1012–1015, May 2012. (↑ pp. 53, 57, and 58)

[99] L. Anderson. Six decades searching for meaning in the proteome. *J Proteomics*, 107:24–30, Jul 2014. (↑ p. 53)

[100] N. L. Anderson. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin. Chem.*, 56(2):177–185, Feb 2010. (↑ p. 53)

[101] R. A. Zubarev and A. Makarov. Orbitrap mass spectrometry. *Anal. Chem.*, 85(11):5288–5296, Jun 2013. (↑ p. 53)

[102] J. Munoz and A. J. Heck. From the human genome to the human proteome. *Angew. Chem. Int. Ed. Engl.*, 53(41):10864–10866, Oct 2014. (↑ p. 53)

[103] E. W. Iepsen, J. Lundgren, C. Dirksen, J. E. Jensen, O. Pedersen, T. Hansen, S. Madsbad, J. J. Holst, and S. S. Torekov. Treatment with a GLP-1 receptor agonist diminishes the decrease in free plasma leptin during maintenance of weight loss. *Int J Obes (Lond)*, 39(5):834–841, May 2015. (↑ pp. 53, 54, and 64)

[104] B. F. Riecke, R. Christensen, P. Christensen, A. R. Leeds, M. Boesen, L. S. Lohmander, A. Astrup, and H. Bliddal. Comparing two low-energy diets for the treatment of knee osteoarthritis symptoms in obese patients: a pragmatic randomized clinical trial. *Osteoarthr. Cartil.*, 18(6):746–754, Jun 2010. (↑ pp. 54 and 64)

[105] G. Utermann. The mysteries of lipoprotein(a). *Science*, 246(4932):904–910, Nov 1989. (↑ p. 57)

[106] P. Wang, E. Mariman, J. Keijer, F. Bouwman, J. P. Noben, J. Robben, and J. Renes. Profiling of the secreted proteins during 3T3-L1 adipocyte differentiation leads to the identification of novel adipokines. *Cell. Mol. Life Sci.*, 61(18):2405–2417, Sep 2004. (↑ pp. 57 and 64)

[107] I. Seres, L. Bajnok, M. Harangi, F. Sztanek, P. Koncsos, and G. Paragh. Alteration of PON1 activity in adult and childhood obesity and its relation to adipokine levels. *Adv. Exp. Med. Biol.*, 660:129–142, 2010. (↑ p. 59)

[108] T. A. Jacobson. Opening a new lipid "apo-thecary": incorporating apolipoproteins as potential risk factors and treatment targets to reduce cardiovascular risk. *Mayo Clin. Proc.*, 86(8):762–780, Aug 2011. (↑ p. 60)

[109] J. H. Contois, G. R. Warnick, and A. D. Sniderman. Reliability of low-density lipoprotein cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B measurement. *J Clin Lipidol*, 5(4):264–272, 2011. (↑ p. 60)

[110] M. F. Gregor and G. S. Hotamisligil. Inflammatory mechanisms in obesity. *Annu. Rev. Immunol.*, 29:415–445, 2011. (↑ p. 61)

[111] J. S. Duke-Cohan, J. Gu, D. F. McLaughlin, Y. Xu, G. J. Freeman, and S. F. Schlossman. 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. U.S.A.*, 95(19):11336–11341, Sep 1998. (↑ p. 61)

[112] O. Cominetti, A. Nunez Galindo, J. Corthesy, S. Oller Moreno, I. Irincheeva, A. Valsesia, A. Astrup, W. H. Saris, J. Hager, M. Kussmann, and L. Dayon. Proteomic Biomarker Discovery in 1000 Human Plasma Samples with Mass Spectrometry. *J. Proteome Res.*, 15(2):389–399, Feb 2016. (↑ pp. 63 and 64)

[113] P. Wang, E. Smit, M. C. Brouwers, G. H. Goossens, C. J. van der Kallen, M. M. van Greevenbroek, and E. C. Mariman. Plasma pigment epithelium-derived factor is positively associated with obesity in Caucasian subjects, in particular with the visceral fat depot. *Eur. J. Endocrinol.*, 159(6):713–718, Dec 2008. (↑ p. 64)

Acknowledgements

I want to express my deep gratitude to everyone who contributed to this work:

First and foremost, I would like to thank Professor Sören Schubert for his great support and his supervision of my thesis. My discussions with him about my project were very helpful to focus my project and taught me to think about the potential of proteomics from the perspective of a clinician and microbiologist. He inspired me in many ways and I finally succeeded in publishing my first first-author paper together with him.

I am extremely grateful for the amazing support from Professor Matthias Mann from the MPI of biochemistry. I was so fortunate to get to know the nuts and bolts of mass spectrometry based proteomics at first hand by working in the lab of one of the founding fathers of the entire field. Our numerous discussions were extremely supporting and delivered me a most precious insight into how to cutting edge research is done.

Dr. Jette Jung helped and supported me throughout my project with great commitment. It was a pleasure to get to work with her and I am particularly grateful for her kind way and her motivating words.

My work would not have been possible without the permanent support of Philipp Geyer, Nils Kulak and Garwin Pichler. They taught me so much about mass spectrometry, came up with brilliant ideas and created a great working atmosphere.

I had inspiring discussions with Florian Mayer and Scarlett Beck about all aspects of the prototype. They as well as the developers from Bruker Daltonics were of great help for acquiring and interpreting the data.

Alina Bartzick, Korbinian Mayr, Igor Paron and Gaby Sowa provided excellent support on all issues concerning the mass spectrometers and sample preparation.

Finally, the entire department welcomed and supported me throughout the entire project and I am very grateful for that.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, 23.07.2019 Niklas Severin Graßl