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

Development of computational methods for the analysis of proteomics and next generation sequencing data

Pavel Sinitcyn

aus
Jaroslawl, Russland
2020

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Jürgen Cox betreut und von Herrn Prof. Dr. Thomas Carell von der Fakultät für Chemie und Pharmazie vertreten.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Planegg, 17.12.2020

....

Ort, Datum Pavel Sinitcyn

Dissertation eingereicht am:	17.12.2020
1. Gutachter: Professor Dr. Thomas Carell	.....
2. Gutachter: Dr. Jürgen Cox	.....
Mündliche Prüfung am:	27.01.2021

Summary

If DNA is the cookbook of life, and RNA molecules are copies of single recipes, then proteins are the ready-to-serve dishes. Studying protein's expression, degradation, localization, interaction, and modifications is therefore crucial to grasp any biological process. Proteomics attempts to seize all protein aspects through a systematic approach. Proteomics produces very complex data that demands the creation of novel software to extract biological meaning. And with every new proteomics-related method, new algorithms need to be simultaneously developed. The MaxQuant and Perseus are open-access softwares that were developed to make the processing and statistical analysis, respectively, of shotgun proteomics data accessible for all.

This thesis presents the development of several new functionalities of MaxQuant and Perseus. For instance, we developed a visualization tool for mass spectrometry data in MaxQuant which is important for quality control of chromatography and mass spectrometer performance. Furthermore, we made MaxQuant available on Linux. It allows the software to run on high-performance servers, including cloud computing, to drastically reduce the running time for large scale projects. The biggest recent advance in MaxQuant development is arguably the newly added option to analyze DIA mass spectrometry data that has a high impact on clinical proteomics. We extended the well-established algorithms developed for DDA to the analysis of DIA.

Perseus is user-friendly software to do statistical analysis of omics data, including genomics and proteomics. In the past few years, we have developed Perseus's capability to compare gene expression at RNA and protein level cooperatively to answer a variety of fundamental biological questions. For instance, Perseus was successfully used to study how membrane proteins are folded on the endoplasmic reticulum, how diverse the oral microbiome is, and how abundant is alternative splicing on the proteomic level.

We also developed an algorithm to extract non-synonymous mutations from genomics data and incorporate them into the proteomics search. This tool was proved to be particularly useful to identify immuno-peptides that could potentially help the immune system to detect cancerous cells. Further development of this technique could potentially help to create personalized vaccines against the patient's cancer.

Contents

Chapter 1

Introduction

The last two decades were marked by impressive progress in the development of new methods in gene expression analysis, especially in different omics technologies such as mass spectrometry (MS)-based proteomics and deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequencing. Today, thousands of genomes and transcriptomes are sequenced routinely, on a scale of a few days, and at a reasonable price. It is a significant leap compared to the effort behind the Human Genome Project from two decades ago[1]. Similarly, the proteomics technology evolved from a qualitative sequencing of one isolated protein towards the robust quantitative "shotgun proteomics" of complex protein samples with over 10000 proteins[2]. All this progress would not be possible without the corresponding development of computational and statistical workflows[3]. Combining genomics, transcriptomics, and proteomics approaches in one study can provide extensive insight into molecular pathways. In recent years, many studies proved that the concept is right. The first part of the introduction will include a review of improvements in the genomics research field. Next, we will describe the recent progress in proteomics and conclude with some examples of multi-omics studies.

1.1 Advances in Genomics

DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides has revolutionized molecular biology[4]. Although this technology has been successfully applied to numerous milestone projects, it has limited throughput and scalability. At the same time, several ideas have been proposed to overcome this limitation – starting from hybridization rate measurement to DNA spots (DNA microarray) and pyrosequencing strategy (454 technology), to sequencing method based on the reversible dye-terminators (Solexa)[5, 6] - next-generation sequencing (NGS). The Illumina Company officially announced a new sequencing machine HiSeq X. It is the first in the world to break the thousand dollars per human genome barrier[7]. This considerable achievement could lead to breakthroughs in personalized medicine.

Most of the mutations resulting in diseases are found in protein-coding sequences[8]. Therefore, focusing on the 1% of the coding part of the human genome would cost far less than whole genome sequencing (WGS). This idea leads to the development of different target-enrichment strategies to increase the relative concentration of exon subset (whole-exome sequencing (WES))[9], which makes sequencing applicable to the clinic setting[10, 11].

The hybridization-based approach (DNA array) is often used as the cheapest alternative to NGS method[12]. However, this technique has several inherent limitations, which include its reliance upon our existing knowledge about genome sequence, high background levels due to cross-hybridization and significant sensitivity to hybridization conditions, and limited dynamic range[13].

Besides the finding of single nucleotide variants (SNVs) compared to the reference genome or cancer specific mutations (somatic mutations), WGS and WES technologies help to find chromosomal oberations[14]. Deletions and duplications of chromosomal segments (copy number variants (CNVs)) are involved in many development disorders[15] and cancer[16].

Although the WGS and WES technologies provide a significant amount of information about biological samples, they neither can inform us about which portion of the genome is effectively transcribed nor can they give any details on the expression and regulation of each gene. In 2008, a few groups came to an idea of using NGS to study transcriptome on a massive scale, which was not feasible before[17–19]. The principle of this method is to convert all the mRNA present in a cell to DNA fragments using reverse transcriptase. The generated DNA fragments are then sequenced in a highly parallelized manner. This new method is called RNA sequencing (RNA-Seq)[17]. Compare to other transcriptomics methods (expression and tiling microarray, RT-PCR, and EST sequencing), RNA-Seq has

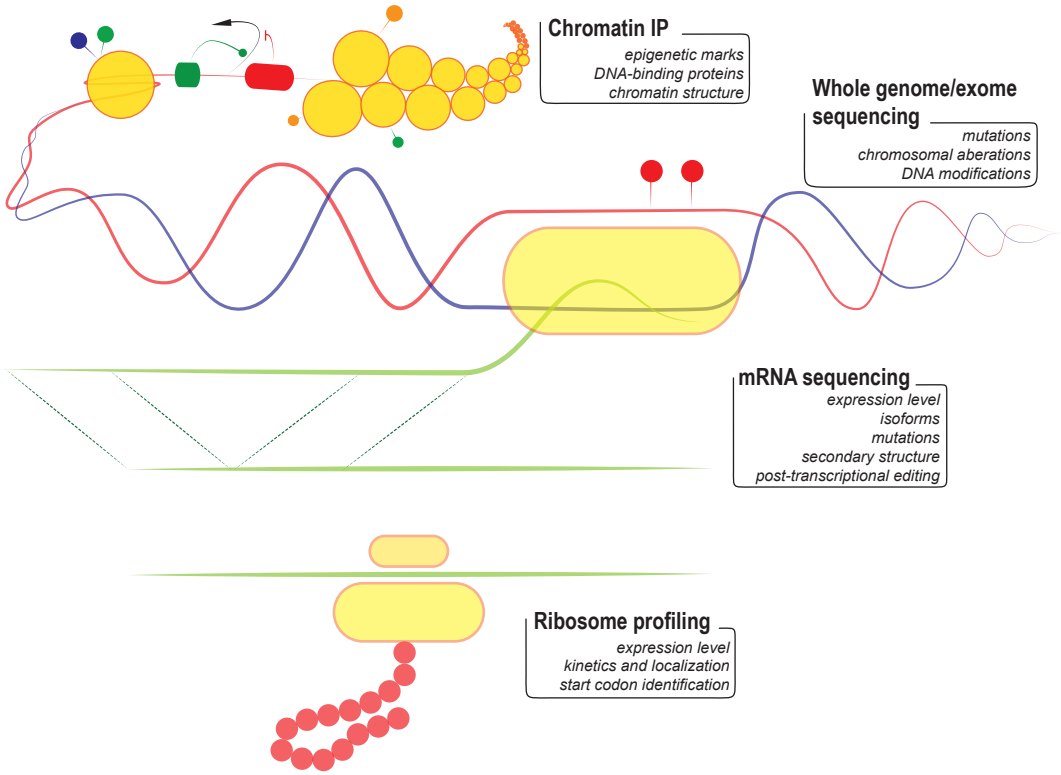


Figure 1.1: **Next Generation Sequencing applications.** WGS and WES allow detecting and localizing mutations, chromosomal aberrations as well as DNA modifications. Combined with Chromatin-Immunoprecipitation (IP), NGS allows to localize epigenetic marks, proteins bound to DNA, and to unravel chromatin structure. NGS can also be used for RNA sequencing, once RNA has been converted to DNA using reverse transcriptase. RNA-Seq is used to determine the gene expression level and which isoforms are produced. It also can serve to define the secondary structure of RNA and to identify post-transcriptional RNA modifications. Moreover, RNA-Seq, as genome sequencing can identify mutations, as long as these mutations reside in a transcribed gene. Finally, NGS is commonly used together with ribosome profiling to determine translation kinetics and localization. This method can also identify start codons.

exceptional advantages: it has a high throughput, low background noise, high sensitivity to expression changes, single-base precision, and more[13]. Therefore, RNA-Seq became the method of choice to understand genes transcriptional control (see Figure 1.1)[20].

Genes tend to express many isoform simultaniously[21]. It has a large impact on many processes in the cell, including cell differentiation[22]. From the very beginning, it became clear that RNA-Seq data would allow making more detailed isoform-specific quantification[17]. But to quantify the relative expression of different isoforms within one gene, one needs to integrate information over many aligned fragments and to add isoform annotations. Thanks to the fact that this inference problem formulation is general, there are many approaches to solve

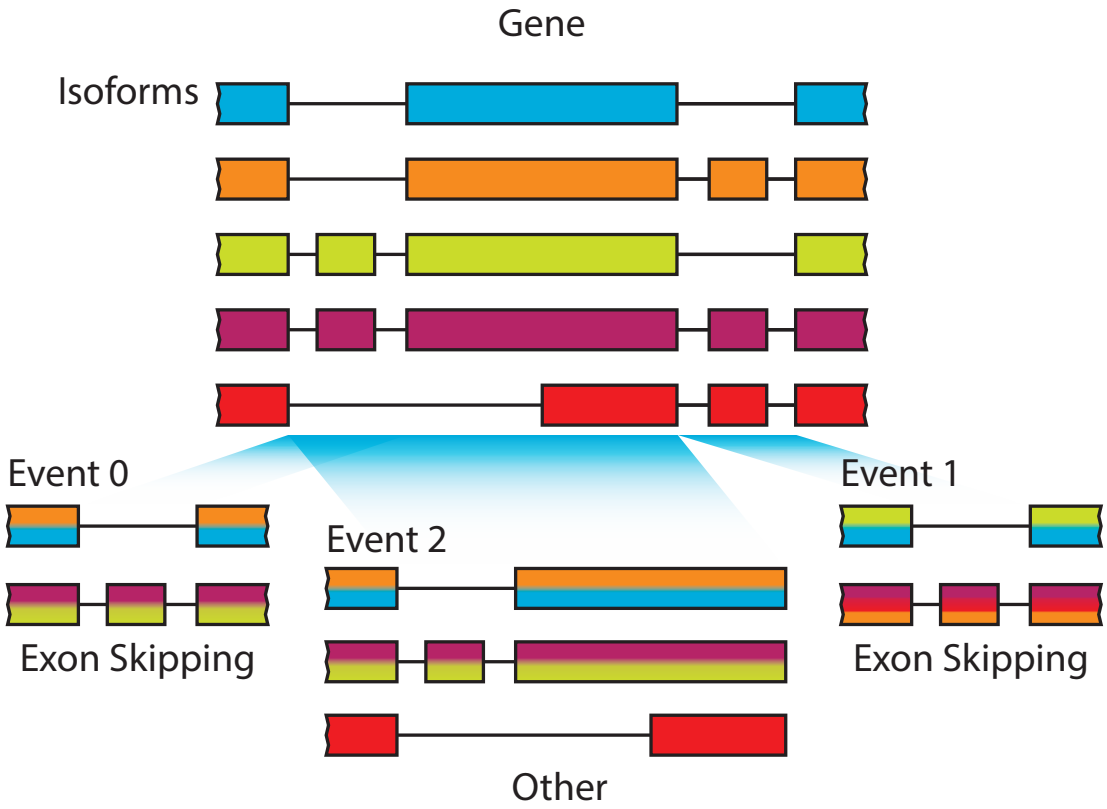


Figure 1.2: **Detection of Splicing Events.** The present gene produced 5 different isoforms. Taking into consideration the short nature of RNA-Seq fragments, it is impossible to distinguish between the first four isoforms. The last isoform however could have specific fragments which bridge between the first two exons. This example illustrates a general complexity of identificatifying and quantifying every expressed isoform. Local event-based strategy allows us to overcome this challenge and it is suitable for the short peptides from proteomics and fragments from genomics.

it[23]. For example, Cufflink software is one of the most popular tools that try to maximize likelihood abundances of isoforms[24].

However, due to the complexity of this problem in the general gene model and considering a situation of constantly growing isoform annotation, a simplified idea to focus on local splice events rather than on the whole isoform quantification has become more prominent (see Figure 1.2)[25]. This analysis is purely based on sequence fragments which are spanning between two exons (splice junctions). From simplistic intron-centric approach[26] to the detection of splice events[27], all these methods show to be robust and informative for a large-scale transcriptomics projects[21, 28]. Efforts of large consortiums helped to qualitatively improve reference isoform annotation and add a lot of newly discovered expressed genes[28, 29], such as long non-coding RNAs[30, 31].

In order to explore and quantify translation regulation, the ribosome profiling (Ribo-Seq) method was developed[32]. Ribo-Seq involves similar

sequencing library preparation to RNA-Seq, but unlike RNA-Seq, Ribo-Seq targets only messenger RNA (mRNA) sequences protected by the ribosome during translation[33]. On top of being a better approximation for the protein expression rate compare to transcriptomics[32], Ribo-Seq allows us to study translation kinetic characteristics of mRNAs[34], organelle specific translation[35–37] and it also allows to enrich gene annotation with the newly detected open reading frames (ORFs)(see Figure 1.1)[38].

All these new methods were found to be ground-breaking in many topics of molecular biology. Microbiology is one of the fields which benefited the most from it[39]. Most of the bacteria cannot be cultivated out of their environment. But the new NGS methods allow studying these bacteria communities in a natural state and on a single cell level[40, 41]. It helps to find out new antibiotics[42], biotechnology methods[43] and to study healthy and pathological developments of the human microbiome[44–47].

1.2 Advances in Proteomics

The term proteomics expresses the ambition to obtain a global view of the proteins world that would be comparable to genomics[48, 49]. The most widely used method for protein MS-based analysis, is the "shotgun proteomics". It strategically bears an analogy to "shotgun genomics" as it reduces long biopolymer to smaller pieces, separates them, sequences each of them independently, and assembles these back to macromolecules. Despite this superficial similarity, there are fundamental differences in the sequencing approaches employed by the two technologies. Proteomics utilize methods based on the measurement of mass to charge (m/z) ratios of individual peptides. Thence, MS analysis of mixed peptide populations results in different peaks corresponding to different mass to charge values. To obtain information on the peptide sequence, the parent ion is subjected to an in-instrument fragmentation and measurement - tandem MS (MS/MS).

From a systems biology perspective, MS-based proteomics offers numerous applications. Expression proteomics determines the relative or absolute amounts of certain proteins in a sample[50, 51]. This is analogous to transcriptomic or ribosome profiling data, except that proteomics automatically takes post-translational regulation of gene expression[49]. Furthermore, with improving instrumental speed and sensitivity it is now becoming possible to obtain sufficient sequence coverage to analyze differential isoform regulation and provides a comprehensive analysis of immune-peptidomes[52, 53].

Another important application of MS-based proteomics lies in the analysis of the interaction of proteins with each other or with other biomolecules[54, 55]. Strong protein-protein interactions are essential for large complexes, but a large number of weak interactions are like "glue" that holds the cellular network together and provides the possibility of highly dynamic regulation[54, 56–58].

It is well known that protein post-translational modifications (PTMs) are diverse, widespread, and responsible for signal transduction in a cell[59–61]. Today MS methods enable unbiased PTM analysis with a single amino acid resolution for the whole proteome and across multiple scales (see Figure 1.3)[62–64]. MS directly measures the presence of a PTM by a defined corresponding shift in the mass of the modified peptide. Thus far, shotgun proteomics was used for the detection of phosphorylation[63], lysine acetylation[60], glycosylation[65], ubiquitylation[66], and methylation[67].

All these advances would not have been possible without the tremendous technological improvement of MS-based proteomics. Starting from standardization and miniaturization of the sample preparation, the invention of soft-ionization methods[68, 69], to continuing hardware[70, 71] and software development[72].

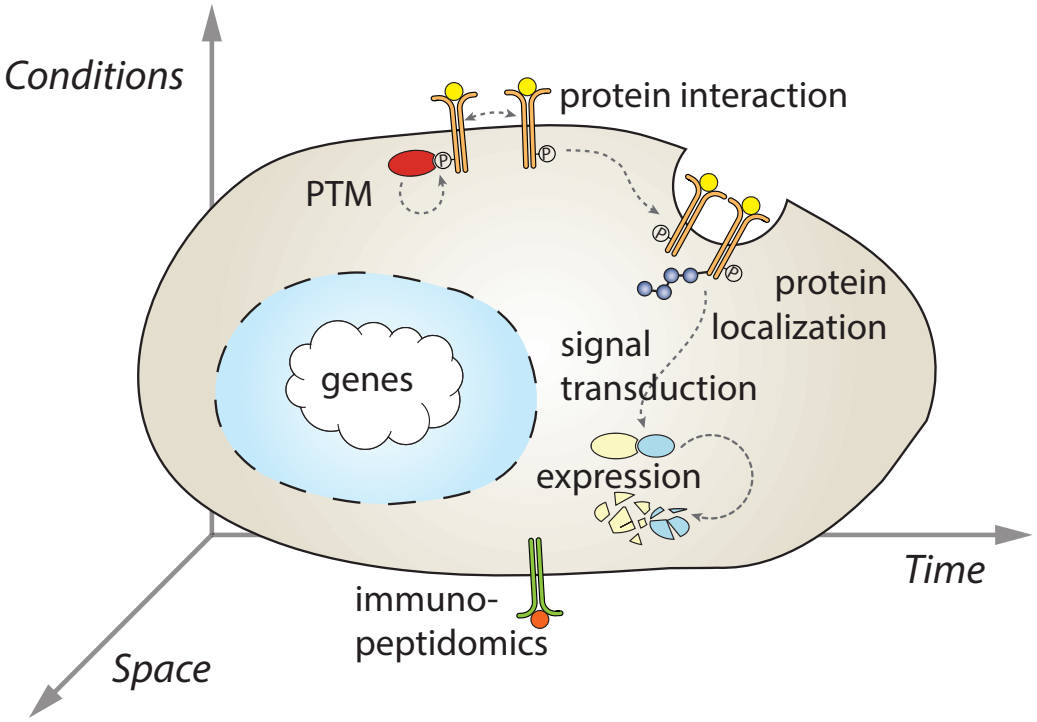


Figure 1.3: **Proteomics applications.** Proteomics is widely used to determine and compare proteins levels in several conditions. It is also one of the most popular methods for the identification of protein-protein applications and is arguably the best system-wide technique for the detection and localization of PTMs such as phosphorylation, lysine acetylation, glycosylation, ubiquitylation or methylation. Proteomics has also been successfully utilized for the characterization of immuno-peptides displayed on the cell surface. Furthermore spatial proteomics has allowed to pinpoint accurately protein localization[73]. Adapted from [74].

1.2.1 Sample preparation

In the workflow of shotgun proteomics, all the proteins present in a sample are first chopped into short peptides using a sequence-specific endoprotease[75]. In most cases, trypsin, which cleaves C-terminal to arginine and lysine, is used as it ensures that every peptide would have at least one positive charge. Also, a positive charge on C-terminus improves the identification of suffix fragments on fragment scan (MS^2). Peptides are then separated by high performance liquid chromatography (HPLC) and evaporated directly into the mass spectrometer. And finally, the peptides are assembled in silico to full proteins, hence the name bottom-up: from peptides to protein. The method is widely used as it provides excellent protein identification and quantification from a complex mixture. This is because peptides are much easier to handle for the liquid chromatography and the mass spectrometer than intact proteins[2].

The top-down approach, however, skips the protein digestion part and attempts to directly analyze intact proteins[76]. While this strategy has the advantages of retaining information about the entire proteins, such as which modifications and exons are found on the same molecule[77], the analysis of the protein mixture is especially challenging. It is due to the complexity of chromatography separation of large molecules and the fundamental limitation of the mass analysers[78]. As a result, top-down proteomics does not reach the level of protein quantification and coverage obtained with bottom-up proteomics. Nonetheless, this method is commonly used for quality control of industrial protein production[79, 80] and shows a lot of promising applications[81, 82].

Besides the classical application of proteomics to measure protein level, shotgun proteomics is the single available option for the large-scale characterization of protein-protein interactions (PPIs) and PTMs.

Immunoprecipitation (IP)-MS used to be a popular technique to characterize stable protein-protein interaction[83]. However, the method requires a very stringent protein purification, usually using a two-step purification procedure, which means that weak, transient interactions were typically lost. The detection of weak interactions demanded to use of other strategies such as the yeast two-hybrid assay, very time-consuming[84]. The increase in MS sensitivity and the rise of reliable label-free quantification allows developing a new strategy for the identification of protein interactors that alleviate the need for stringent protein purification and permit the detection of weak, transient interactors[54]. Moreover, this strategy turns protein contaminants into a key element for reliable quantification across thousands of samples. In this method, called affinity purification mass spectrometry, a single-step affinity enrichment of a tagged protein

and its interactors is performed and followed by single-run, intensity-based label-free quantitative liquid chromatography (LC)-MS/MS analysis[54, 56, 85].

The first step towards understanding the influence of PTMs is their identification and quantifications on a global scale. Shotgun proteomics has proven to be an ideal method for such tasks[2]. MS directly measures the presence of a PTM by a defined corresponding shift in the mass of the modified peptide and a location of the modification within a peptide is found by a corresponding shift in the mass of fragments. So far, shotgun proteomics was used for the detection of phosphorylation, lysine acetylation, glycosylation, ubiquitylation, and methylation. However, the detection and quantification of these PTMs pose several challenges, including the commonly low PTM occupancy. It is therefore often necessary to first enriched peptides for a specific PTM before performing shotgun proteomics to increase PTMs quantity and lower the sample complexity. For the study of phosphorylation events, highly efficient methods to enrich phospho-peptides to up to a 100 fold exist. Metal affinity chromatography, for instance, using titanium dioxide is commonly used.

Today efforts in the sample preparation are focused on miniaturization and increase a high-throughput[86, 87]. Thanks to the current achievements, many proteomics dreams come true, such as single-cell proteomics[88–92] and large-scale clinical proteomics studies[93–97].

1.2.2 Soft ionization

Mass spectrometry is the workhorse of analytical chemistry[98–100]. Its usage is widespread from the analysis of atomic composition with a micro-dalton resolution to the measurement concentration of complex mixtures of molecules. Most biological molecules are polymers and exist in a liquid phase. To analyze such samples with mass spectrometry, one needs to transfer molecules to the gas phase and charge them while preserving the polymer structure.

In 2002, the Nobel Prize in Chemistry was brought to John B. Fenn and Koichi Tanaka "for the development of methods for identification and structural analysis of biological macromolecules". This nomination acknowledges how mass spectrometry revolutionized the study of proteins and their functions.

The term "soft ionization" technique unites two methods for polymer ionization while keeping the molecular structure intact - Matrix-assisted laser desorption/ionization (MALDI), and electrospray ionization (ESI). For the MALDI, the sample is mixed with a suitable matrix material and transferred to a metal plate[69, 101]. Later a pulsed laser irradiates the sample, which helps to evaporate and charge molecules. One of the most famous modern applications of this technique is to analyze the spatial distribution of proteins on histological specimens, so-called proteome imaging[102–104].

ESI ionization is a popular technique to produce ions using a high voltage liquid sprayer[68, 105–107]. It is making a highly charged aerosol that efficiently transfers polymers from a liquid to a gas phase while charging. The key advantage of this technique is the ability to directly connect the HPLC in front of the liquid sprayer. Like that, one can utilize a divide and conquer strategy by simplifying the analysis to just a portion of ion species at a time, and not all at once, as it is in the case of MALDI.

1.2.3 Mass spectometer

Every mass spectrometer instrument consists of several building blocks: an ion source, a mass analyzer, and an ion detector. In the case of shotgun proteomics, where peptides should be charged and transferred to the gas phase with minimum damage, the above described ESI and MALDI techniques well serve the role. There is a wide variety of mass-analyzers and all of them have a special combination of characteristics, which defines the field of usage[108]. However, in shotgun proteomics, time-of-flight (TOF), quadrupole, and Orbitrap are among the most frequently used mass analyzers.

In the TOF analyzer, all ions in a beam receive the same kinetic energy from the accelerator towards the detector[109]. Ions with different masses will have a different speed and, as a result, distinct time of coming to the ion detector. Light ions are going to reach the detector earlier and have a shorter time of flight in comparison to the heavier counterparts.

The quadrupole mass analyzer, as the name implies, consists of four parallel cylindrical rods. By applying a high-frequency voltage, the mass analyzer can filter all ions from a particular m/z window[110]. Even though the quadrupole with a detector is a mass spectrometer by itself, the quadrupole is usually used as a mass filter additionally to the main mass analyzer. In shotgun proteomics, the quadrupole allows the selection of a specific peptide using a small m/z extraction window in data-dependent acquisition (DDA) and a set of peptides from a long m/z range in data-independent acquisition (DIA)[72]. For the development of the quadrupole, Wolfgang Paul received the Nobel prize in physics in 1989.

The Orbitrap is a mass analyzer from the family of Orbital electrostatics traps family[111] that is commonly used for proteomics[112]. The Orbitrap traps ions in a volume between the central rod and an outer electrode[113]. An ion, moving in this volume, will generate an axial frequency, which depends only on the m/z ratio. Thus, multiple ions will generate a linear interference of signals, which can be deconvoluted with Fourier Transformation to one pair of frequencies and amplitudes for each ion[114]. Each frequency can be converted to an actual m/z and each amplitude represents an intensity. Together intensity and m/z make up the mass spectrometry spectra.

The ions injected into the Orbitrap should have a specific set of initial parameters to have a stable orbit. It represented a significant challenge for the practical usage of Orbitrap. However, further developments such as C-trap, to elegantly overcome this issue (see Figure 1.4)[115].

The two separated outer electrodes of the Orbitrap serve as ion detectors. The fact of having both a mass analyzer and a detector as one compact device

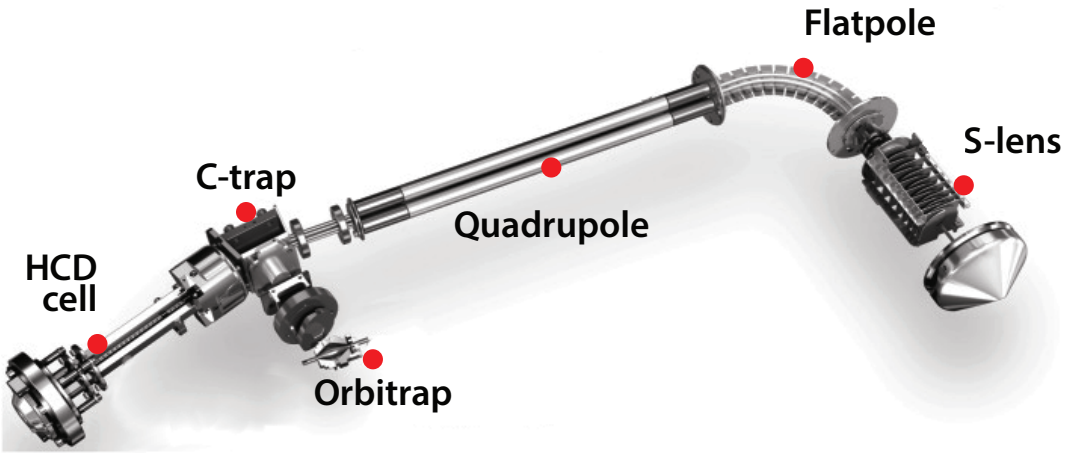


Figure 1.4: **Scheme of the Q Exactive HF instrument with Orbitrap.** After entering the instrument, the beam of molecules is captured and focused by the S-lens. In the next segment, a flatpole bends the beam so that all uncharged species are filtered out. In the peptide scan (MS^1) mode, the ions pass through an inactive quadrupole and accumulate in enough quantity in the C-trap. After reaching the necessary amount of ions, C-trap injects them into the Orbitrap. In the MS^2 mode, the ions exit the flatpole and are filtered for a specific m/z range in the quadrupole. The resulting subset of ions passes to the fragmentation cell (for example, higher-energy collision dissociation cell) where they are subject to the fragmentation. And finally, the produced fragments come to the C-trap and later to the Orbitrap to generate MS^2 spectra. Adapted from [117].

constitutes a distinct feature of the Orbitrap. It allows to miniaturize the benchtop format of the instrument even further and get better performance[116–118]. Today, commercially available mass-spectrometers with Orbitrap represent a state-of-the-art in high-resolution mass spectrometry[119].

The recent decade is marked by the intensive development of additional ion analyzers on an interface between the ion source and mass analyzer - Parallel Accumulation Serial Fragmentation (PASEF)[120] and Field Asymmetric Ion Mobility Spectrometry (FAIMS)[121]. The main purpose of these devices is to reduce the sample complexity per measurement cycle by adding one more dimension - ion mobility and compensatory voltage, respectively. These devices proved to be useful in many proteomics applications[122–126].

1.2.4 Quantitative proteomics

One of the main time bottlenecks in proteomics experiments is the measuring time for each sample. This problem can be mitigated by reducing the length of the HPLC gradient or by acquiring more mass-spectrometry instruments. An alternative would be to mix several samples and analyze them at once. This approach is called multiplexing. Furthermore, due to the high complexity of sample preparation for proteomics, multiplexing serves not just as a way to reduce measurement time, but even more importantly, as a way to alleviate the systematic bias introduced by sample preparation and measurement[2, 72, 127].

Historically, stable isotope labeling by amino acids in cell culture (SILAC), as a representation of peptide scan (MS^1) labeling, was the first multiplexing technique. This method was hugely popular since it was for a long time the best and single available option for large-scale quantitative proteomic studies[127].

Two biological objects are fed either with a growth medium containing normal amino acids (light) or with amino acids labeled with non-radioactive heavy isotopes (heavy). Arginine and lysine amino acids are commonly used since they are essential for most species and nicely combined with trypsin specificity. This insured that all peptides have at least one differential amino acid. All light peptides from one sample will be coeluted with their heavy counterparts from the other samples and located on fixed mass differences relative to each other. This allows for a direct and accurate comparison of peptide levels for each identified protein[128].

Even though SILAC can be easily used for cell culture and single-cell organisms, its usage in higher organisms, including humans, poses ethical problems. Even though few publications used SILAC on whole multicellular organisms[129, 130], the usage in patients is out of the question. A variation of SILAC, called super-SILAC, was therefore introduced to bypass these limitations[131]. The super-SILAC technique utilizes cell lines labeled with heavy amino acids as a spike-in standard for the accurate quantification of the unlabeled samples[131], thereby enabling the quantification of tissues where heavy labels cannot be integrated. The spike-in cells correspond to the heavy channel, while the tissue samples correspond to the light one. Therefore, the choice of the cell line used as spike-in is dictated by the origin of the studied tissue. Super-SILAC approach was successfully used to quantify human tumor proteomes[132]. The concept of using a heavy spike-in as a reference channel can greatly help to characterize changes in the proteomes of model species[133, 134]. The same concept was also elegantly used to map protein localization within a cell[73, 135].

Although SILAC is one of the most popular MS^1 labeling techniques, several other MS^1 labelling techniques exist. In NeuCode, masses of incorporated amino

acids differ by as little as several mDalton (Da)[136, 137] thanks to the different combinatorial set of isotopes - isotopologues[138]. This multiplexing method can be seen as a SILAC alternative with higher multiplexing capability but requires high-resolution proteomics. Another technique is the dimethyl labeling where peptides are chemically modified before mixing of two samples[139].

All MS1 labeling techniques suffer from the same disadvantages: they tend to crowd the MS1 spectra that limit the sequencing depth and limit the number of samples that can be analyzed at once. Hence, another type of multiplexing technique aims to solve this problem by making peptides from different samples indistinguishable on the MS^1, but on the MS^2 level. These methods are called MS^2 or isobaric labeling. tandem mass tag (TMT) is one of the most prominent and scalable technologies available[140]. In TMT, peptides from up to 16 different samples are chemically modified by 16 different tags that have the same mass but a different distribution of isotopes within their atomic structure[141]. As a result, the same peptide from different samples appears as a single peak in MS^1. As the peptide is fragmented before MS^2 measurement, the labels are cleaved into pieces of different masses (reporter ions). The relative abundances of the different reporter ions in MS^2 allows the accurate relative quantification for all samples at once[142]. This fact is decisive in using TMT in the chemical proteomics[143, 144] and in large proteomics projects[145, 146].

Reliable protein quantification without MS^1- or MS^2-labeling has been a long-standing interest in the proteomics field[147]. But this objective was unrealistic for a long time due to imperfections in sample preparation, limitations in the mass spectrometers, and available software solutions. However, improvements in all of these three aspects have opened new opportunities for label free quantification (LFQ). Indeed, the standardization of sample preparation protocols increased the reproducibility within and across different laboratories[86, 148, 149] while the production of a new generation of mass spectrometers created feature-rich spectra with a high-resolution[118, 126]. These two former aspects have permitted the development of new computational approaches that take advantage of the large-scale nature of the produced data[124, 147].

1.2.5 Computational proteomics

The synergic advancement in proteomics sample preparation, mass spectrometry hardware, and software, greatly improved protein identification and quantification, which was mostly manual and inaccurate in the past. Today modern sophisticated and mature algorithms dealing with millions of spectra from complex proteomes[150–154]. Our in-house developed search engine Andromeda accounts for the probability of observed matches between expected and measured fragment masses by chance[155]. This search engine enables analysis of complex proteome datasets in combination with MaxQuant, which provides a user-friendly interface for pre-and post-processing of MS data[150].

This review covers two main areas of computational methods: the identification and quantification of peptides, proteins, and PTMs, as well as the downstream analysis for the biological interpretation of proteomics data[72].

I contributed to conceptualize and write the text. I also have designed and created the figures and acquired the necessary data for them.

Pavel Sinitcyn[*], Jan Daniel Rudolph[*], Jürgen Cox
Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data
(2018) *Annual Review of Biomedical Data Science*
DOI: 10.1146/annurev-biodatasci-080917-013516

[*]these authors contributed equally to this work

ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data

Pavel Sinitcyn,* Jan Daniel Rudolph,* and Jürgen Cox

Computational Systems Biochemistry Research Group, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany; email: cox@biochem.mpg.de

Annu. Rev. Biomed. Data Sci. 2018. 1:207–34

First published as a Review in Advance on May 4, 2018

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-080917-013516

*These authors contributed equally to this article

Keywords

computational proteomics, mass spectrometry, posttranslational modifications, multiomics data analysis, multivariate analysis, network analysis

Abstract

Computational proteomics is the data science concerned with the identification and quantification of proteins from high-throughput data and the biological interpretation of their concentration changes, posttranslational modifications, interactions, and subcellular localizations. Today, these data most often originate from mass spectrometry–based shotgun proteomics experiments. In this review, we survey computational methods for the analysis of such proteomics data, focusing on the explanation of the key concepts. Starting with mass spectrometric feature detection, we then cover methods for the identification of peptides. Subsequently, protein inference and the control of false discovery rates are highly important topics covered. We then discuss methods for the quantification of peptides and proteins. A section on downstream data analysis covers exploratory statistics, network analysis, machine learning, and multiomics data integration. Finally, we discuss current developments and provide an outlook on what the near future of computational proteomics might bear.

INTRODUCTION

Proteins perform nearly all the work in a cell and are the key players in the structure, function, and regulation of cells, tissues, and organs. Collectively they form the proteome (1), a highly dynamic and diverse molecular omics space comprising interactions among proteins and other types of biomolecules. The proteome can be studied comprehensively with mass spectrometry (MS)-based technologies (2–4). Thousands of proteins and posttranslational modifications (PTMs) can be studied quantitatively over a multitude of samples in complex experimental designs. Describing all applications of proteomics is beyond the scope of this review, but among its applications are diverse topics such as cancer immunotherapy (5) and the evolution of extinct species (6).

Computational MS-based proteomics can be roughly subdivided into two main areas: (*a*) the identification and quantification of peptides, proteins, and PTMs and (*b*) downstream analysis, aiming at the biological interpretation of the quantitative results obtained in area *a*. This review follows this subdivision. Computational proteomics is a highly multidisciplinary endeavor attracting scientists from many fields and incorporates other disciplines like statistics, machine learning, efficient scientific programming, and network and time series analysis. Furthermore, the integration of proteomics data with other biological high-throughput data is increasingly gaining importance.

Peptide-based shotgun proteomics, also called bottom-up proteomics (7), needs to be distinguished from top-down proteomics (8–10), in which whole proteins are studied in the mass spectrometer. Data analysis tools and approaches exist for top-down methods (11–13) in which feature deconvolution plays an important part. In targeted proteomics (14–17) (**Figure 1**), a set of key peptides from a target list, which is informative for a set of proteins or PTMs of interest, is quantitatively monitored over many samples using dedicated software (18). Data-independent acquisition (19), as exemplified by the SWATH-MS method, comes with its own computational challenges for which solutions are provided in the literature (20–23). Imaging MS (24) is also a

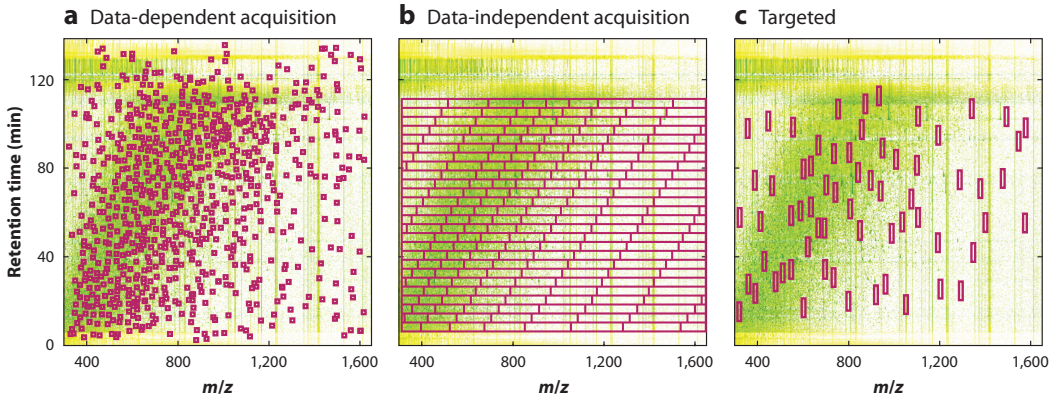


Figure 1

Main formats of mass spectrometry (MS)-based proteomics. Peptide-based bottom-up proteomics is most often done in the data-dependent acquisition mode (*a*). MS2 (second-stage MS) scans are triggered depending on the MS1 (first-stage MS) data features seen in real time. Typically, at a given retention time, the *n* most intense peptide features are selected for fragmentation, dynamically excluding masses that have just been previously selected. In data-independent acquisition (*b*), a set of constant mass ranges, which do not depend on the peptides being analyzed, is isolated for fragmentation. In targeted proteomics (*c*), a list of peptides is targeted based on a list of mass and retention time ranges corresponding to peptides of interest, which are particularly informative of a set of proteins or posttranslational modifications that are the focus of the investigation.

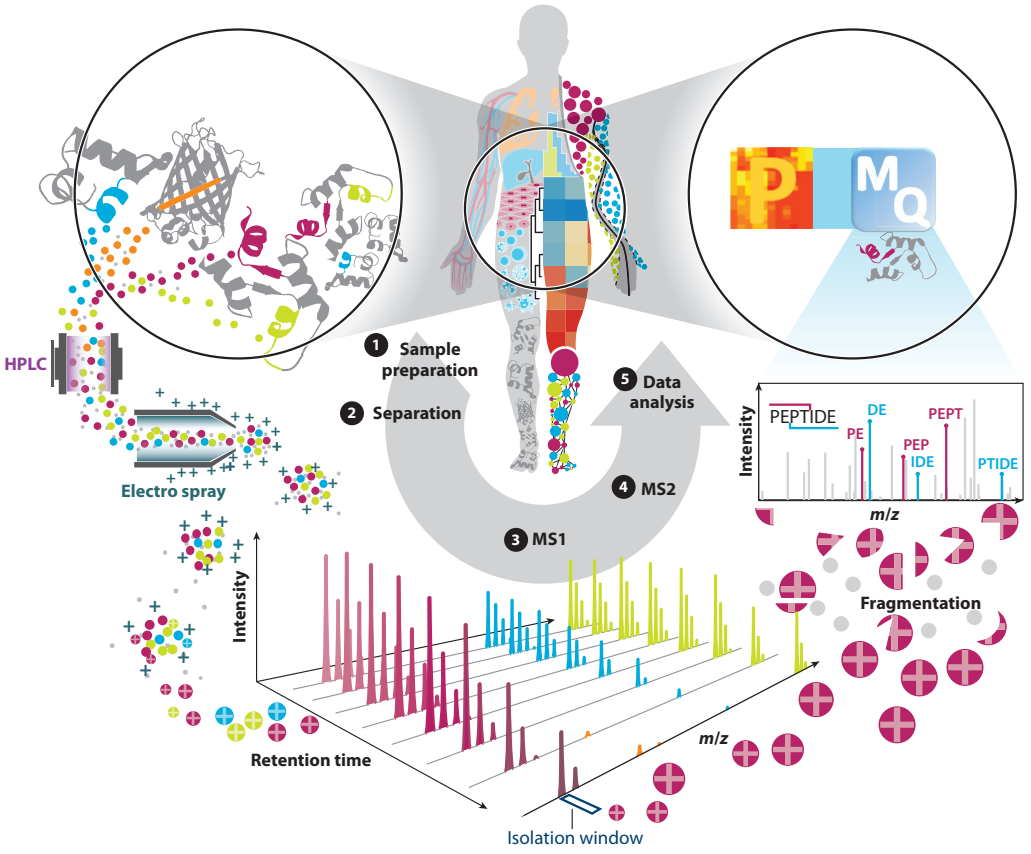


Figure 2

Bottom-up shotgun proteomics workflow. (❶) Proteins are extracted from a sample of interest. Enrichment of organelles or affinity purification may be performed. Proteins are digested to peptides that are optionally enriched for modifications. (❷) After HPLC separation, peptides are ionized (181, 182) and (❸) injected into a high-resolution mass spectrometer (e.g., 183, 184). MS1 spectra containing peptide isotope patterns are recorded in a cycle with a timescale of about one second. (❹) Peptide precursors are selected for fragmentation and fragment (MS2) spectra are recorded. (❺) Both MS1 and MS2 spectra are written to disk, typically resulting in several gigabytes of data per LC-MS run, and then analyzed by computational proteomics software. Abbreviations: HPLC, high-performance liquid chromatography; LC, liquid chromatography; MS, mass spectrometry; MS1, first-stage MS; MS2, second-stage MS.

fruitful area of research that will not be covered here. This review focuses on data-dependent bottom-up or shotgun proteomics (**Figure 2**), which currently is the format most frequently used in proteomics.

It is not the aim of this review to present an exhaustive list of all available software tools. Instead, we focus on explaining concepts and key applications. In several places, we use the MaxQuant (25–27) and Perseus (28) software as concrete examples for the implementation of certain concepts. Alternative software platforms developed in academia (29–31) or offered by mass spectrometer vendors can provide similar functionality. We propose that robustness, ease of use, parallelizability,

and automation of all computational aspects are the key factors to consider in the selection of software tools.

Proteomics research is supported by community tools such as repositories, databases, and annotation sources (32). There are public repositories for the storage and dissemination of MS-based proteomics data (33–39), and submission of raw data is highly recommended for every proteomics publication (34). Protein and peptide sequences are essential for the interpretation of proteomics data. For this purpose, UniProt (universal protein resource) (40) is a comprehensive, high quality, and freely accessible resource of protein sequences and functional information. Since most amino acid sequence identifications can be put into the context of coding nucleic acid sequences—exceptions prove the rule (41)—genome-centric sequence repositories like Ensembl (42) are of high importance as well. Data sharing and dissemination of publicly available proteomics data are facilitated by dedicated software tools for the reanalysis of community data (43, 44).

This review consists of two main parts, the first dealing with the data analysis steps performed on the spectral data itself, going up to the identification and quantification of peptides, proteins, and PTMs. This part is organized in a problem-centric way, where in each subsection, a particular challenge in the MS workflow is described. The second part is about the downstream data analysis. Here, the sections are organized by methodologies rather than application areas, which is a more approachable organization scheme, since the number of different applications is enormous, while the methodologies overlap. The downstream analysis of proteomics data is still an art, and there is not always only one correct way to arrive at biologically meaningful conclusions. Hence, we give a comprehensive overview of the available methods that can be used along the way.

IDENTIFICATION AND QUANTIFICATION OF PEPTIDES, PROTEINS, AND POSTTRANSLATIONAL MODIFICATIONS

Liquid Chromatography-Mass Spectrometry Features

Since the early days of MS, the detection of peaks in a mass spectrum, corresponding to molecular features, played a central role (45). Nowadays, the mass resolution is sufficiently high in general that the isotope pattern of peptides is resolvable (**Figure 3*a***). On the molecular level, a single peak corresponds to an isotopic species with fixed elemental composition and several nucleons. In case of ultrahigh mass resolution, the isotopic fine structure of peptides in the low-mass range can be resolved (46) (**Figure 3*a***), resulting in increased information about the atomic constituents of the peptide. While obtaining isotopic resolution is standard nowadays for peptides, the same is still technically challenging for whole proteins in top-down proteomics. For instance, for each charge state of an antibody, usually only an envelope is detected, while the isotopic peaks remain unresolved.

In proteomics, the mass spectrometer is typically coupled on-line to additional continuous separation dimensions like liquid chromatography (LC) (47) or ion mobility separation (48). MS features can therefore be viewed as higher-dimensional objects. In case of LC-MS, peaks become three-dimensional (3D) objects in the m/z–retention time–intensity space (**Figure 3*b***). Using ion mobility adds another dimension, turning features into 4D objects. Technically, due to its dimensionality, the problem of MS feature detection is equivalent to general-purpose 2D image feature detection or voxel assembly to 3D volume elements (49), respectively. However, since MS data often have additional regularities that can be exploited, the problem is often simpler than generic object recognition. Simplifying assumptions specific to mass spectrometer types should be exploited to apply faster algorithms to the multidimensional feature detection problem. (Readers are referred to the supplement of Reference 25.)

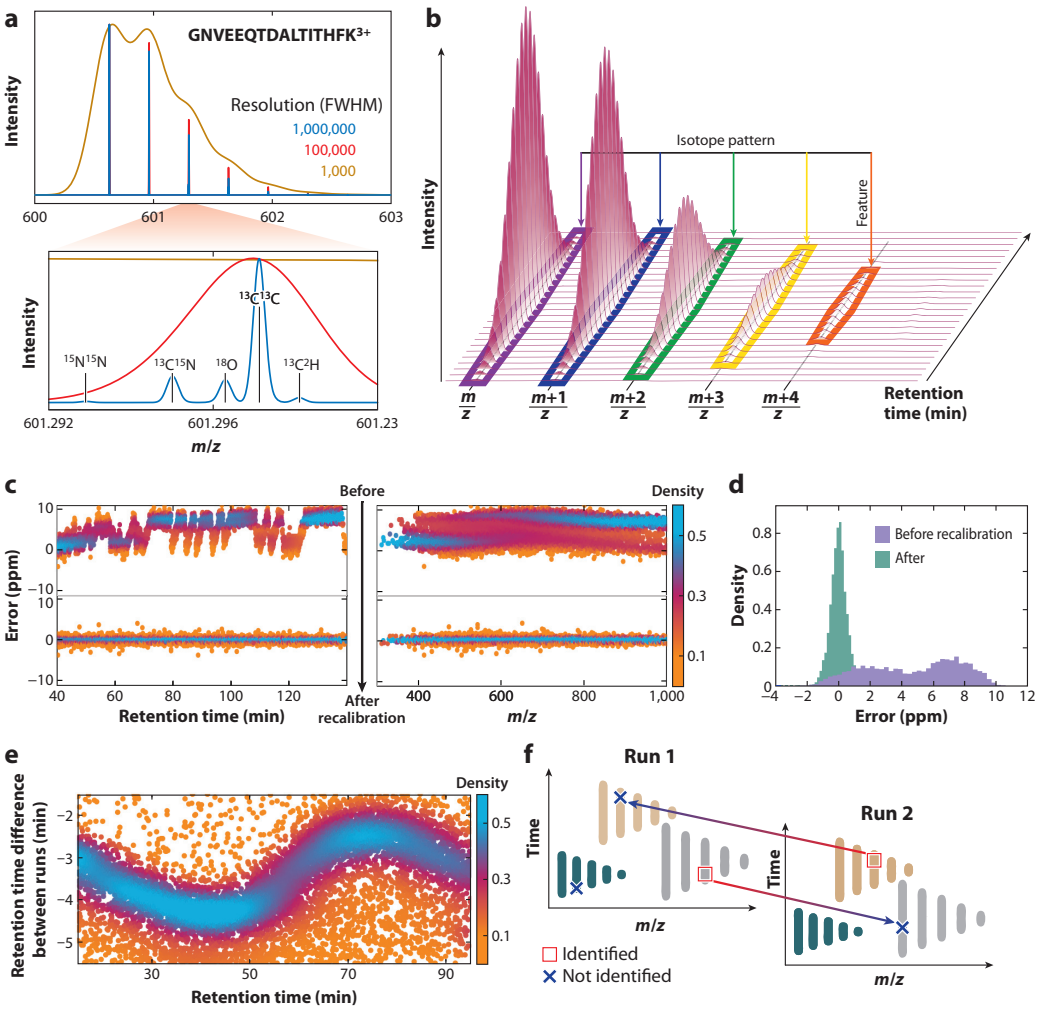


Figure 3

MS1 feature–based computational tasks in a proteomics workflow. (*a*) Theoretical spectrum of an MS1 feature measured in three different resolutions. The lowest resolution (1,000 FWHM) does not resolve the isotope pattern. The ultrahigh resolution (1,000,000) reveals the natural isotopic fine structure. (*b*) A three-dimensional isotope pattern in *m/z*–retention time–intensity space. (*c*) Peptide mass errors as a function of retention time and peptide *m/z* before and after nonlinear recalibration. Clearly, nonlinear systematic errors were present and were then removed by recalibration. (*d*) Mass error distribution before and after recalibration. A large increase in mass accuracy was achieved through nonlinear recalibration. (*e*) Retention time alignment curve between two LC-MS runs. (*f*) Matching between runs. Peptide identities are transferred between LC-MS runs from MS2-identified MS1 features to nonidentified MS1 features in other similar LC-MS runs based on accurate mass and retention time. Abbreviations: FWHM, full width at half maximum; LC, liquid chromatography; MS, mass spectrometry; MS1, first-stage MS; MS2, second-stage MS; ppm, parts per million.

Once features corresponding to isotopic peaks are detected, they are assembled to isotope patterns, effectively deisotoping the spectrum. Different models exist (50–52), one of them being the Averagine model (50), which can be used to explore spectral properties, since nearly all peptides with a given approximate molecular mass have a similar elemental composition. In the model, it is assumed that a peptide is made up of the average number of the 20 amino acids according to their natural occurrence. The model then predicts the mass differences between isotopic peaks in an isotope pattern, as well as their relative heights. This approach is usually sufficient when dealing with data with unresolved isotopic fine structure. When the isotopic fine structure is resolved, one will have to employ the true atomic compositions of the peptide candidates to utilize this information. In the approaches using higher-dimensional features, the exact coelution of isotopic peaks can also be utilized to increase the specificity of assignment of isotope patterns. While in most cases, the spectral information is not sufficient to determine the elemental composition, one will obtain the charge state and a highly precise estimate of the monoisotopic mass from the information contained in the higher-dimensional features.

One can find labeling *n*-plexes of isotope patterns in the MS1 (first-stage MS) data prior to peptide identification, similar to how features are assembled to isotope patterns. This applies to nonradioactive differential isotopic sample labeling techniques (53, 54) like SILAC (stable isotope labeling by amino acids in cell culture) (55) or dimethyl labeling (56, 57). Analogous to the deisotoping step, specific mass differences between the isotope patterns participating in a labeling *n*-plex are expected. This is not the case for ^{15}N labeling (58, 59) in which all nitrogen atoms are completely exchanged with the stable heavy isotope. Isotope patterns belonging to an *n*-plex are usually coeluting, depending on the type of labeling, which can be exploited in the assembly of *n*-plexes.

While mass measurements from modern high-resolution mass spectrometers, in combination with the aforementioned higher-dimensional feature detection, can achieve very-high-mass precision, this does not automatically translate into high-mass accuracies, due to the presence of systematic measurement errors. In **Figure 3*c***, the peptide mass error prior to mass recalibration is displayed as functions of *m*/*z* and of retention time. Systematic errors are typically nonlinear and depend on multiple variables. In addition to *m*/*z* and retention time, the mass error can depend on signal intensity and ion mobility index, if applicable. Nonlinear recalibration on multidimensional parameters is difficult when it must rely on only a few calibration points, as is usually the case if dedicated spike-in molecules are used. Hence, it is typically better in complex samples to use the peptides from the sample itself as calibration points for multivariate recalibration, which is achieved in MaxQuant by a two-level peptide identification strategy (25, 60, 61). The mass accuracy increases by large factors resulting from the applications of these nonlinear recalibration curves obtained in this way (**Figure 3*d***).

Similar to the mass accuracy, the consistency of the retention times of peptide features can also be increased by recalibration. Due to often unavoidable irreproducibility in chromatography, retention times are usually not comparable between LC-MS runs, thereby limiting identification-transfer and quantification between runs. Nonlinear shifts by several minutes are common. Hence, algorithmic approaches were developed to align retention times between multiple runs (**Figure 3*e***). Typically, these retention time corrections need to be nonlinear (62). In MaxQuant, this is achieved with a sample similarity–derived guide tree, which avoids the need for singling out one LC-MS run as the master run (63) that all the other runs are aligned to. Ion mobilities can be aligned between LC-MS runs with similar methods as retention times.

Once masses, retention times, and ion mobilities are recalibrated, one can transfer identifications between related LC-MS runs from peptide features identified by fragmentation to unidentified peptide features by having same mass, charge, retention time, and ion mobility (64)

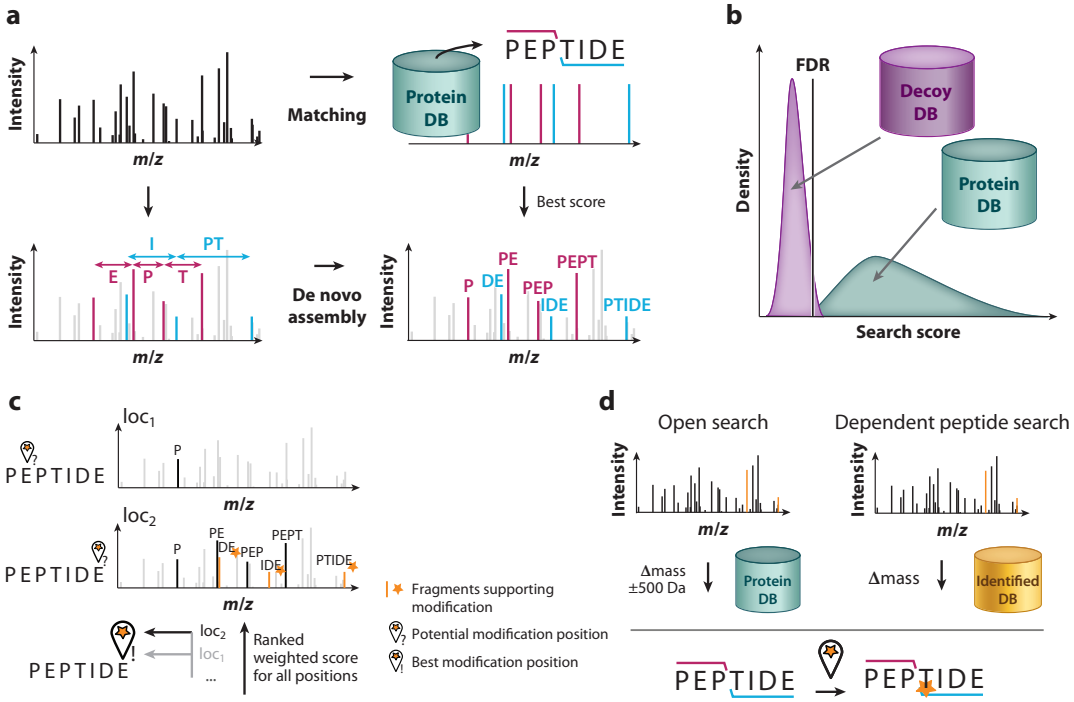
(**Figure 3*f***). Following this strategy, the quantification profiles across many samples become more complete, which partially removes the stochastic behavior of the data-dependent acquisition in bottom-up proteomics. Determining and controlling false discovery rates (FDRs) for these kind of matching approaches is challenging and the subject of current research. However, if samples are similar, error rates caused by matching are in acceptably low ranges.

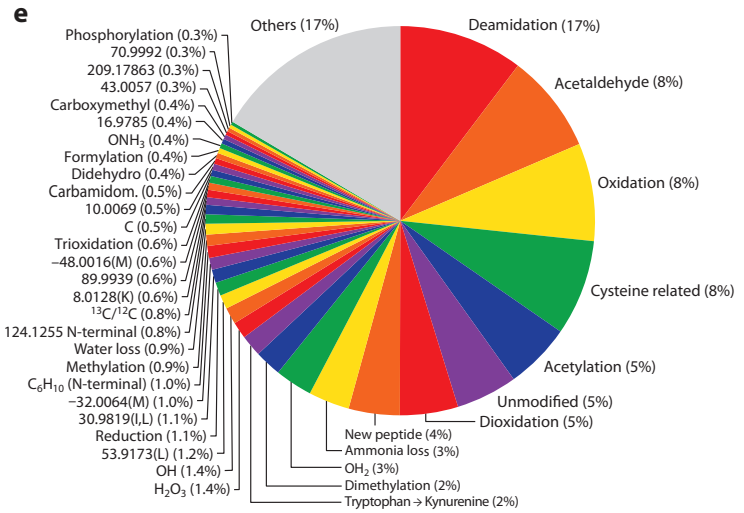
Peptide Identification

Peptide identification tools analyze the fragmentation spectra obtained by the mass spectrometer with the aim of determining the sequence of the peptide. In the most popular approach, database search engines (65–69) utilize a target database of theoretical fragmentation for identification (**Figure 4*a***). The database is generated from all protein sequences that are known or thought to be produced according to the instructions in the genome of an organism. The protein sequences are digested in silico into peptides according to a cleavage rule mirroring the protease used in the experiment (e.g., trypsin, which cleaves after the occurrence of lysine or arginine in the protein sequence). For each of these in silico peptides, the list of expected fragment masses is calculated based on the backbone bond breakages expected for the fragmentation technique used in the experiment. For a given measured fragmentation spectrum, the search engine calculates a match score against all theoretical fragmentation spectra within a specified peptide mass tolerance. The highest-scoring peptide spectrum match (PSM) is taken as a candidate for the identity of the peptide. Since the highest-scoring PSM might still be a false positive, most workflows control the FDR using a target–decoy approach (70) (**Figure 4*b***). In this approach, fragmentation spectra are searched not only against the target database, but also against a decoy database, which is designed to produce false-positive PSMs. Comparing the score distributions of target and decoy PSMs, posterior error probabilities can be calculated and FDRs can be controlled. One procedure to generate decoy sequences is to reverse the target sequences, providing peptides that do not occur in nature.

Additional peptide features besides the search engine score, such as the length of the peptide and the number of missed cleavages, help distinguish true identifications from false positives, leading to more high-confidence identifications. In MaxQuant, the posterior error probability, which is the probability of a PSM being wrongly identified, is conditional on the score and additional peptide properties (25). Other tools such as PeptideProphet (71, 72) and Percolator (73) use linear discriminant analysis or support vector machines (SVMs) with the same aim. Machine learning was used to predict intensity patterns in fragmentation spectra in order to support database scoring and further improve identification (74), but it failed to improve upon the state of the art. In contrast, the application of deep learning to de novo peptide identification did yield improvements (75).

De novo peptide sequencing (**Figure 4*a***) is another technique for identifying peptides from fragmentation spectra. The peptide is identified using only information from the input spectrum and the characteristics of the fragmentation method. Mass differences between certain peak pairs correspond to amino acid masses, which are interpreted as consecutive ions in one of the expected fragment series, for example, y or b ions for collision-induced dissociation. If these mass differences can be continued to a whole series from N- to C-termini, the peptide is identified without reference to a sequence database. An incomplete de novo amino acid series is called a sequence tag and might be completed on either of the termini with a sum of amino acid masses and PTMs. The many existing tools for de novo peptide identification explore different algorithmic approaches, some allowing for de novo sequencing errors and homology searches (76–79). An interesting approach is a hybrid between database search and de novo sequencing (80); it requires only a little de novo information and hence inherits high sensitivity from the database search approach.





(Caption appears on following page)

Figure 4 (*Figure appears on preceding page*)

Overview of peptide identification methods. (*a*) In the peptide database (DB) search engine approach, measured second-stage mass spectrometry (MS2) spectra are scored against a list of theoretical spectra from an in silico digest of protein sequences. De novo peptide identification allows reading the peptide sequence partially or completely out of the MS2 spectrum. (*b*) In the target–decoy approach, true and decoy protein sequences are offered to estimate the false discovery rate (FDR). (*c*) Determining the localization probability for a posttranslational modification on a peptide. (*d*) Open search and dependent peptide search are methods for detecting modifications in an unbiased way. Modifications still must be localized after open search. (*e*) Modifications found in a typical dependent peptide search. Data from Reference 185 were used.

For a peptide that has been identified as having a certain sequence and carrying one or more modifications, the positions of these modifications on the sequence might not be localizable with complete certainty. Hence, a score needs to be calculated that quantifies for each potentially modifiable amino acid in the peptide sequence the certainty of localization at a given locus (**Figure 4*c***). For instance, a peptide might contain several potentially phosphorylated serine, threonine, and tyrosine residues, but from the peptide mass it is known that it is phosphorylated only once. Then one needs to determine which of the sites are phosphorylated and use the spectral evidence to derive each site's probability that it is the one bearing the modification (81–85). The most important spectral features for the calculation of localization probabilities are the site-determining ions, which are fragments that are matched with one hypothetical localization but not with the other. The exact way the localization score is calculated varies between different methods. In MaxQuant, the localization probability is calculated as a weighted average of exponential Andromeda scores over all combinations of phosphorylation configurations (86).

The identification of modified amino acids, either as PTMs such as phosphorylation or as modifications introduced during sample preparation, is usually done by adding these as variable modifications into the database search. While this strategy is highly sensitive, all modifications have to be specified beforehand. The number of modifications that can be specified is limited due to the combinatorial explosion of modified peptides species, leading to a large increase in database size. There are two approaches overcoming these limitations: open search (87) and dependent peptide search (88) (**Figure 4*d***). The open search approach does not extend the sequence database but instead widens the precursor mass tolerance window for the MS1 precursor peptide molecule to, for example, ±500 Da, while keeping the fragment mass tolerance low (87). Therefore, a modified peptide with a mass within the tolerance window can still be matched to the correct unmodified database sequence despite ∼50% of fragment ions being shifted by the modification. The high number of candidate matches makes the open search computationally demanding, but recent approaches make use of fragment ion indexing to speed up the search significantly (89). The dependent peptide search, also implemented in MaxQuant, is a generic approach to retrospectively identify unassigned MS2 (second-stage MS) scans; it relies on the assumption that the sample contains not only the modified dependent peptide, but also its unmodified base peptide counterpart (88). Using any search algorithm will yield identifications, as well as unassigned MS2 spectra. The search now queries all unassigned spectra against all identified spectra, while simultaneously localizing the modification. The mass difference between the peptides is the putative mass of the modification, which is used to generate a shifted ion series for each position in the peptide. The highest-scoring match will therefore determine the sequence of the peptide, as well as the mass and locus of the modification. **Figure 4*e*** shows the most frequent modifications found by dependent peptide search in a typical data set.

There are a number of special topics in peptide identification, starting with dipeptides resulting from cross-linked proteins (90, 91), which have the challenge of a vastly increased search space due to pairing of peptides, for which several popular software packages are available (92–97). In proteogenomics searches (98), peptides are identified based on customized protein sequence

databases generated from genomic or transcriptomic information. Search spaces for proteogenomics searches are typically larger than in conventional searches since they often involve three- or six-frame translations of genomic sequences. Furthermore, these search spaces are heterogeneous, since the sequence content ranges from clearly existing, manually validated protein sequences to in silico–translated genomic regions without any prior evidence for their expression. Hence, extra measures need to be taken in the identification process to account for this heterogeneity. Proteomics of species without sequenced genome requires tools to integrate incomplete sequencing data with homologous sequence data from closely related species (99).

Protein Inference and False Discovery Rate

Protein inference, that is, the assembly of peptides into a list of proteins, is a crucial step in a computational proteomics workflow, since usually the peptides are only technical aids to study proteins. (Readers are referred to Reference 100 for a review.) The relationship between peptides and proteins is many-to-many, since upon digestion a protein gives rise to many peptides, but a peptide can also originate from more than one protein. Furthermore, based on the identified peptides, proteins that share common sequences might not be distinguishable from each other. Hence, a redundancy grouping of protein sequences is necessary.

Peptides that are unique to a protein are more desirable than nonunique ones. On average, longer peptides are more likely to be unique, and hence, more informative. As an order of magnitude estimate, we calculate how often a random peptide of a given length would occur in the human proteome, assuming it is randomly composed out of the 20 amino acids and has the same size as the latest human UniProt release 2017_09, which contains 93,588 protein sequences comprising 37,118,756 amino acids in total. Peptides of length 5 should occur on average 12 times in the proteome, meaning that their information content is nearly worthless. Peptides of length 6 should occur on average 0.6 times, making them only just potentially useful, but many of them can still be expected to be nonunique. In this model, only peptides of length 7 or longer are on average expected to be informative and useful. Although other factors like tryptic peptides and paralog relationships between genes realistically should be considered, the conclusions hold true of real data.

Many tools and algorithms for the protein assembly have been described in the literature. The most frequently applied ones can be roughly subdivided into parsimonious and statistical models. Parsimonious models (25, 101–104) apply Occam's razor principle (105) to the protein inference problem by finding a set of proteins that is as small as possible to explain the observed peptides. Usually, fast greedy heuristics are used to find such a protein set. Statistical models (106, 107) can assemble large amounts of weak peptide identifications to infer the existence of a protein. However, for both types of models, it is worth considering a threshold on peptide identification quality, for example, 1% FDR for PSMs. High-quality peptide identifications allow for solid conclusions about the properties of the identified proteins, while weakly identified peptides can compromise protein quantification accuracy. Ideally, the output of the protein inference step is a list of protein groups. Each protein group contains a set of proteins that cannot be distinguished from each other based on the observed peptides. Either the proteins in a protein group have equal sets of identified peptides or the peptide set of one protein is a proper subset of that of another protein, in which case, based on the peptide identifications, there is no evidence for the existence of the latter protein, assuming that the former protein is in the sample.

The phenomenon of error expansion from peptide to protein identification in large data sets is well known in the field (106, 108). Even if the FDR is thoroughly controlled at the PSM level, if no additional measures are taken, the FDR on protein level can become arbitrarily large. Hence, it

is highly important to use workflows that control FDR on the protein level (25, 106, 108, 109) to limit the number of proteins falsely claimed to be present in the sample, particularly if the number of identified proteins is a relevant outcome of the study.

Quantification

Proteomics becomes more powerful when done quantitatively, as compared to only browsing through lists of identified proteins. Many responses to stimuli on the level of proteins are not switching the expression of a protein on and off completely, but manifest themselves as changes in cellular concentrations that might be small, yet important. Quantitative proteomics approaches can be subdivided into absolute and relative quantification methods. In absolute quantification, one wants to determine copy numbers or concentrations of proteins within a sample, while in relative quantification, a quantitative ratio or relative change of protein concentrations between samples is desired. Both absolute and relative quantification can be done either with the aid of labels or label-free.

Figure 5 shows an overview of relative quantification methods. In label-free quantification, the samples being compared are biochemically processed separately. The distinction between metabolic and chemical labeling is not important from a computational perspective. Instead, the main distinction is between MS1-level labeling, in which the peptide signals corresponding to the multiple samples are compared and form multiplexed isotope patterns in the MS1 spectra, and MS2-level or isobaric labeling, in which the multiplexed signals appear in the fragmentation spectra. Hence, computational methods for relative quantification should be distinguished between label-free, MS1-level labeling, and MS2-level labeling.

In label-free quantification, one faces particular challenges with normalization intensities between LC-MS runs and the compatibility of quantification with prefractionation. In MaxQuant, the MaxLFQ algorithm (110) is implemented for relative label-free quantification. It uses signal intensities of MS1 peptide features as input, optionally including the ones identified by matching between runs, and produces as output relative protein abundance profiles over multiple samples. MaxLFQ accounts for any peptide or protein prefractionation of the samples by applying a sophisticated intensity normalization procedure to the feature intensities of each LC-MS run. A protein intensity profile is constructed that best fits protein ratios determined in all pairwise comparisons between samples. In each of these pairwise comparisons, only peptides that occur in both samples are used, which makes the relative comparison very precise. Hence, MaxLFQ is more accurate than merely summing up all peptide intensities belonging to a protein. By using a sample-similarity network for the intensity normalization step, the algorithm scales well to large data sets and can quantify hundreds of samples against each other.

Stable isotope labeling with sample multiplexing appearing on the level of MS1 spectra (55–57, 111, 112) promises to be more accurate than label-free quantification since the coelution of features in the same LC-MS run can be exploited. The ratio calculation can be performed along the elution profile separately in each MS1 scan and separately for each isotopic peak. This results in many estimates of the ratio, which can be summarized by taking the median. This robust ratio estimate is less sensitive to contamination by other coeluting peptides. In this way, the ratios between MS1-label channels are calculated in a more precise way, as compared to the label-free approach, where feature intensities are calculated separately before their ratio is taken. During MS1-label n-plex assembly, the isotope patterns of parts of the n-plex might be missing, leading to an incomplete quantitative profile. Proper MS1 isotope patterns might be missing for peptides arising from low-abundant proteins. In MaxQuant, the requantification algorithm tries to find traces of these isotope patterns close to the noise level.

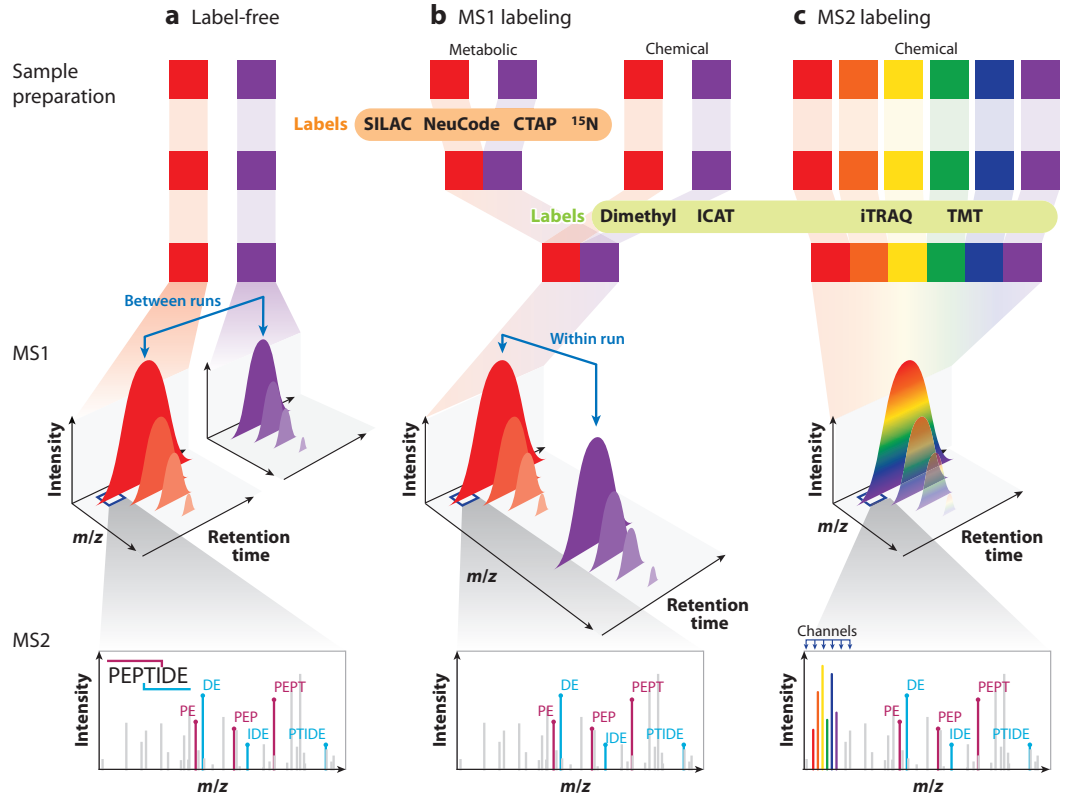


Figure 5

Overview of relative quantification methods. Relative quantification of samples (*colored squares*) can be done in a label-free, metabolic, or chemical labeling approach. For computational approaches, the distinction between MS1 labeling (*b*) and MS2 (isobaric) labeling (*c*) is more crucial. In the label-free approach (*a*), the quantification is done for each peptide feature between extracted ion chromatograms in different LC-MS runs. In MS1 label–based quantification (e.g., SILAC, dimethyl, NeuCode), multiple samples will appear as differentially labeled isotope patterns in the MS1 spectra. For isobaric labeling (e.g., iTRAQ, TMT), the quantification signals appear as reporter ions in the low-mass range of the MS2 spectra. Abbreviations: CTAP, cell type–specific labeling using amino acid precursors; ICAT, isotope-coded affinity tags; iTRAQ, isobaric tags for relative and absolute quantification; LC, liquid chromatography; MS, mass spectrometry, MS1, first-stage MS; MS2, second-stage MS; SILAC, stable isotope labeling with amino acids in cell culture; TMT, tandem mass tags.

One can use one labeling channel as a common standard, as is done in Super-SILAC (113), which allows quantifying unlabeled samples with the added accuracy of labeling by using ratios of ratios to compare samples with each other. Computationally, these hybrid samples are analyzed like MS1-labeled samples in the feature detection, but the downstream analysis proceeds nearly as if they were label-free samples.

In isobaric labeling (114–116), peptides in different samples are labeled with different molecules per sample that have the same mass but that eject different reporter ions upon fragmentation. The biggest advantage of isobaric labeling is its multiplexing capacity. Up to 11 samples can be measured simultaneously with the currently available tandem mass tag reagents. The downside is

that the presence of coeluting peptides in the isolation window for fragmentation leads to ratio compression (117). To be precise, cofragmentation makes ratios wrong in arbitrary and individual ways. However, since it is often a valid assumption that most of the proteins are not changing between samples, the cofragmented peptides are likely to have 1:1 ratios, thus compressing the ratios of changing proteins. There are several experimental strategies to reduce or remove the cofragmentation problem, such as gas-phase purification (118), MultiNotch MS3 (119), and use of complementary ions (120). There are several computational methods that reduce ratio compression. Reporter ions of low intensity are prone to carry more noise and be more affected by cofragmentation signals. Hence, peptides with higher reporter ion intensities should be given higher weights when calculating protein intensities. Another approach is to calculate the fraction of precursor signal divided by the total MS1 signal observed in the isolation window (121, 122), which can be used for filtering peptides used for quantification. To some extent, this quantity can also be used to correct for ratio compression (123).

Approximate measures of absolute protein abundances can be obtained with simple computational prescriptions like the iBAQ or Top3 methods (124, 125). The problem that peptides of a protein have vastly different flyability (a term used to cover the relative efficiencies of ionization, transfer, and detection), making them not directly comparable for quantification, is solved by averaging over many peptides or selecting the most intense ones, which enriches for high flyability. For eukaryotic cells, one can add an absolute scale to these readouts with the proteomic ruler approach (126), which uses the signal of histones, assuming that it is proportional to the amount of DNA in the sample.

The quantification of peptides and PTMs differs from protein quantification in that only a single or few features can be used for quantification, while on the protein level, accuracy is achieved by accumulating quantitative information over many peptides. Hence, the variability of PTM quantification data and the number of missing values is usually higher than it is for proteins. For combined PTM-enriched and proteome data, computational methods exist for calculating occupancies (86, 127), which are the percentages of proteins modified at a given PTM site.

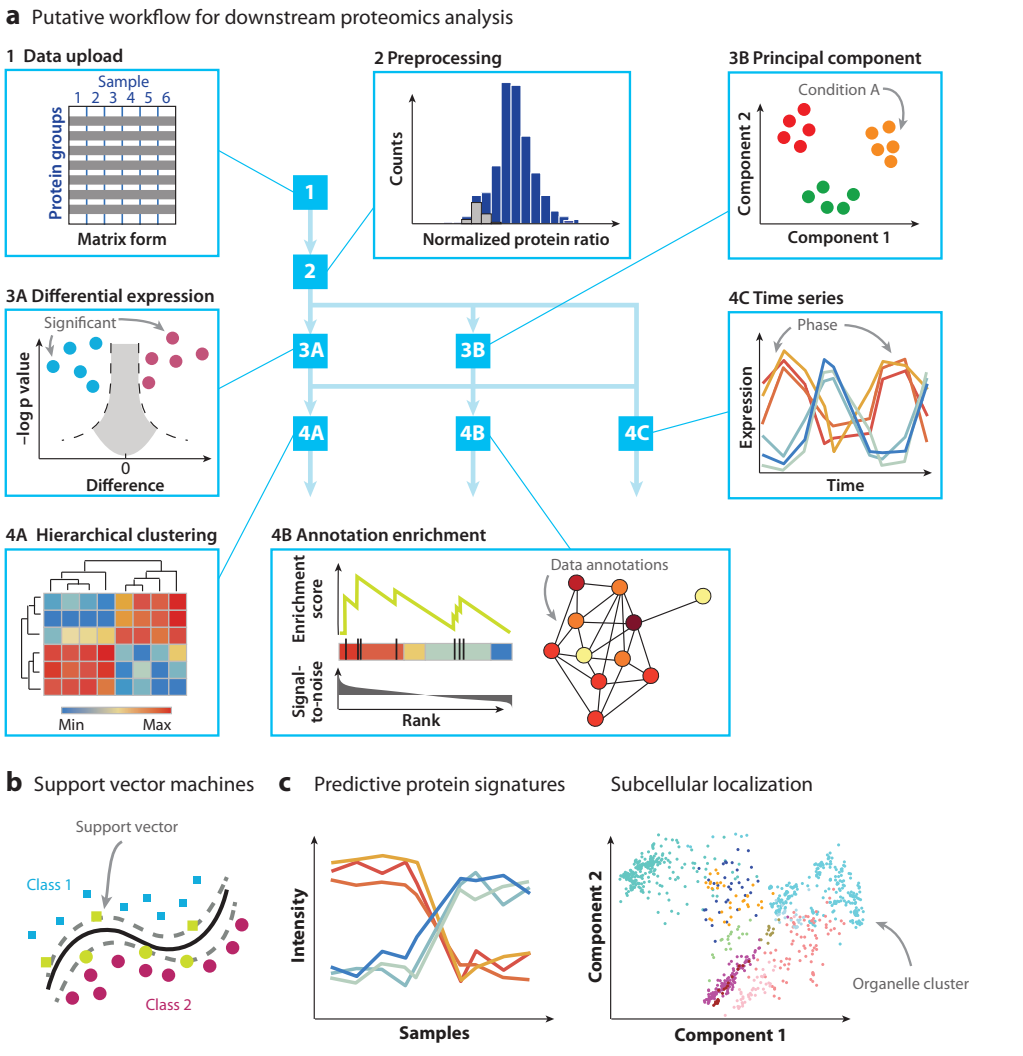
DOWNSTREAM DATA ANALYSIS

Exploratory Statistics

Once proteins have been identified and quantified over many samples, one obtains a matrix with proteins (or protein groups) as rows, samples as columns, and protein abundances or abundance ratios in the matrix cells. Usually, the interpretation of this quantitative protein or PTM data and the translation into significant biological or biomedical findings are the most important and labor-intensive parts of a study. The Perseus platform (28) was developed to support the domain expert in this data exploration. It is workflow based, modular, and extensible through a plugin infrastructure.

There are some preparatory steps preceding most analyses, such as normalization of intensities or ratios, data filtering, and potentially missing-value imputation (**Figure 6*a***). A common task in discovery proteomics is to identify proteins of biological interest and distinguish them from the rest of the proteome. Statistical models are popular tools for identifying differentially expressed proteins. Clustering methods, such as hierarchical clustering, are often used for finding expression patterns of groups of proteins and for their visualization in a heat map. Principal component analysis (PCA) is an alternative method of visualizing the main effects in the data and the relatedness between samples. It also provides information on proteins responsible for a separation of sample groups through the so-called loadings.

The statistical tests *t*-test and ANOVA (analysis of variance, which is the generalization of the *t*-test to more than two groups) are the basic versions of a series of statistical models that test for significant changes between sample groups (128, 129). In more complex experimental designs, one might want to test for the effects of two factors simultaneously (e.g., gender and treatment), in which case two-way ANOVA can be used. ANOVA can be generalized to any number *n* of factors, resulting in *n*-way ANOVA. After retrieving a list of significant proteins from ANOVA, a post hoc test can be applied to pinpoint the sample groups within the experimental design that were changing. If samples are related and independency assumptions are violated, so-called



(*Caption appears on following page*)

Figure 6 (*Figure appears on preceding page*)

Downstream analysis overview. (*a*) Putative workflow for downstream proteomics analysis. After data upload (*Step 1*) and preprocessing (*Step 2*), common analyses include differential expression (*Step 3A*), principal component analysis (*Step 3B*), hierarchical clustering (*Step 4A*), annotation enrichment (*Step 4B*) and time series analysis (*Step 4C*). Data preprocessing (*Step 2*) may involve several steps including data normalization and visual inspection of distributions of protein quantification values in histograms. Differential expression analysis (*Step 3A*) reveals those proteins that are significantly changing their concentrations between two or more conditions. Principal component analysis (*Step 3B*) highlights main trends in the data such as a separation between cellular conditions, as shown in the example. Hierarchical clustering (*Step 4A*) is often done in conjunction with heat map visualization of expression changes and reveals characteristic patterns relating groups of samples to clusters of proteins. Results are often validated using annotation enrichment analysis (*Step 4B*). Time series analysis (*Step 4C*) can distinguish between characteristic temporal patterns such as phases of peaking protein concentrations in a periodic process, as shown in the example. Adapted from Reference 28. (*b*) Support vector machines are a powerful machine learning tool for classification. From training data they learn decision rules that can distinguish between classes of samples based on their protein expression profiles. The decision rule is indicated here by a separating line between the two classes. Support vectors are those samples that contribute most to defining the separating line. Adapted from Reference 28. (*c*) Applications for machine learning in proteomics include finding predictive protein signatures and predicting the subcellular localization of proteins. The colored clusters represent proteins that are localized in same organelles. Data from Reference 147 were used.

repeated measures ANOVA is a valid method of data analysis. For all of the methods above, it is crucial to control false positives due to multiple hypothesis testing, since many tests are done simultaneously. If only a moderate *p*-value cutoff is applied to define significant proteins, the number of false positives will be inflated (130). Benjamini-Hochberg FDR control (131) or permutation-based FDR estimates (132) are efficient methods to deal with this problem.

When an interesting group of proteins has been identified, for instance, by statistical testing, clustering, or PCA, enrichment analysis can be performed to find biological processes, complexes, or pathways common to these proteins. Fisher's exact test checks for contingency between group membership and the property of interest. It clarifies what is common to the cluster-member proteins and might indicate the functional role of the cluster. For this purpose, annotation sources like gene ontology (133), pathway memberships (134), or curated protein complexes (135) are needed.

Biological processes under study often exhibit temporal changes, with proteins following an expected pattern, for instance, as periodic changes in the cell cycle or circadian rhythm. Other studies involve measuring a response to dose changes of stimuli. In these situations, methods can be applied that detect concentration changes following a given model, such as periodic changes with a given periodicity. For this case of periodic temporal changes, the analysis will assign an amplitude of change and a peaking time to each protein (136).

Posttranslational Modifications

Quantitative PTM data can be represented as a matrix resembling proteome-expression data, but with modified peptides or modification sites on the identified proteins as rows. Therefore, PTM studies can be analyzed with methods similar to those used for protein expression. For instance, after suitable normalization and filtering, hierarchical clustering or PCA can be applied to determine dominant patterns of phosphorylation changes (86). As previously discussed, one needs to be aware of the higher variance of PTM-level data compared to protein-level data. This requires a higher number of replicates compared to protein-level data to achieve the same statistical power.

There are several public resources for obtaining PTM specific annotations. UniProt (40) provides comprehensive information on local protein properties at the PTM site or in its vicinity. Specialized databases, such as PhosphoSitePlus (137), Signor (138), and Phospho.ELM (139), cover mostly phosphorylation events. They include functional annotations, as well as kinase–substrate interactions. This information can be used for enrichment analysis to gain information about the processes involved in writing, reading, and erasing the studied PTMs. One can also analyze PTMs in the context of signaling networks, as discussed below.

Machine Learning

Machine learning has several applications in the downstream analysis of proteomics data (**Figure 6*b*,*c***). A very prominent one is the classification of patient-derived samples based on their protein expression patterns (140–142). For artificial intelligence–based diagnosis, a supervised learning algorithm is first trained on samples derived from patient cohorts for which a certain property is known, for instance, the cancer subtype. The trained algorithm is then used to diagnose novel samples, that is, to predict the same property for samples where the property is not known. The same supervised learning approach can be combined with feature selection algorithms to derive predictive protein signatures. Each signature contains proteins that show a distinct expression pattern and can be used for sample classification. Multivariate feature selection methods can take the interdependence of proteins acting within networks into account and can find patterns for which the discriminatory power is not apparent in the expression profiles of single proteins. This makes machine learning–based feature selection a powerful alternative to ANOVA-like methods to determine protein signatures, where a *p*-value is calculated for only one protein at a time, independently from all the other proteins.

Machine learning approaches are most easily validated using cross-validation (143), which provides a measure of how well the prediction performance of a classification or regression model will generalize to independent data not used for model training. Cross-validation helps avoid the notorious problem of model overfitting and can be used to monitor prediction errors when extracting optimal protein sets from the output of feature selection algorithms. SVMs (144) often perform particularly well in classification or regression of samples in omics spaces. This is not surprising, since for most technologies, including proteomics, the number of features (biomolecules) is typically much larger than the number of samples. SVMs were created to perform well in spaces with exactly these properties. Deep learning (145, 146) is gaining traction in proteomics (75) and will likely find more applications in the future.

Machine learning has also been successfully applied to the prediction of subcellular localization with the dynamic organellar maps method (147, 148), which allows global mapping of protein translocation events. First, one generates a database of marker proteins with known localization and absolute copy number information and characteristic fractionation profiles. Then, using SVMs, a model is built for the prediction of cellular localization. This method has dynamic capabilities to capture translocation events upon a stimulation. This enables a widely applicable proteome-wide analysis of cellular protein movements without requiring process-specific reagents.

Network Biology

MS-based proteomics provides researchers with diverse tools for the study of biological networks (149). Enrichment protocols interrogate the interaction partners of a bait protein and provide the basis for the assembly of large-scale protein–protein interaction (PPI) networks (**Figure 7*a***). Affinity enrichment/purification coupled to LC-MS is routinely used to quantify hundreds of physical interaction partners. Since relying only on identification of proteins in the pull-down leads to many false positives, it is crucial to distinguish background binders from significantly enriched bona fide interactors. Statistical tests, such as the two-sample *t*-test, can identify true interactors but require a control to compare against. This control sample either can be a dedicated experiment lacking the bait protein or can be assembled from other orthogonal experiments within the same study (150, 151). Due to its quantitative nature, this approach can probe not only steady-state interactions, but also dynamic rewiring upon stimulation by internal or external stimuli. If intensity-based quantification is used, the missing values problem for enriched samples can be overcome by imputation. Alternative methods

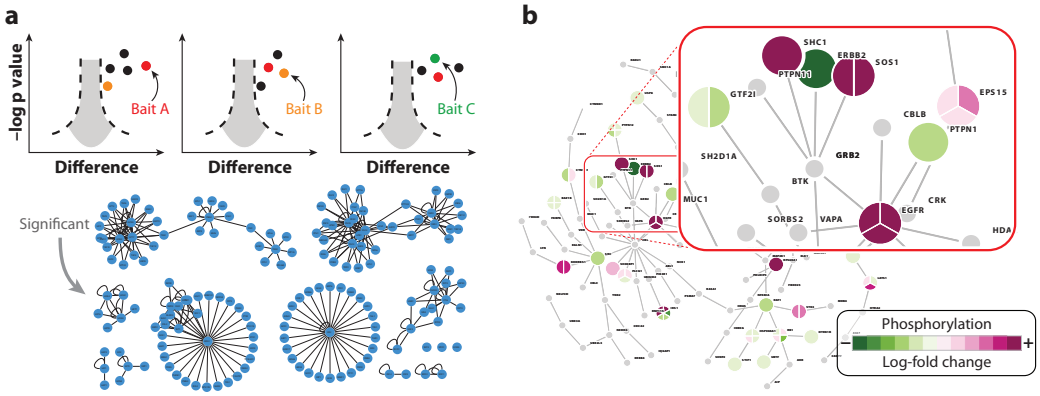


Figure 7

Network analysis. (*a*) Protein–protein interaction networks can be constructed by applying statistical testing to a series of pull-down experiments with different bait proteins. The resulting network of proteins with significant enrichment to any of the bait proteins can be visualized in tools such as Cytoscape. Adapted from Reference 28. (*b*) Signaling pathway reconstructed from phosphoproteomics data derived from MCF7 cells after epidermal growth factor stimulation (160). The pie charts in the network visualize the measured phosphorylation changes on each of the proteins. Proteins with unknown phosphorylation states are colored gray.

relying on spectral counting directly accommodate for the absence or presence of a protein in a sample (152). Both approaches have been used to construct large-scale PPI networks (151, 153).

Cells often achieve signal transduction through PTMs, which are enzymatically written, read, and erased. The interpretation of PTMs in the context of these signaling networks is therefore natural. PTM specific networks, such as kinase–substrate interactions, can be obtained from curated databases, such as PhosphoSitePlus (137). To increase coverage, kinase–substrate relationships can also be predicted by machine learning and PPI network analysis (154). Logic models obtained from, for example, the Signor database (138) can provide a mechanistic interpretation of phosphoproteomic data, indicating active kinases, as well as functional phosphorylation sites. Several computational methods predict kinase activities from kinase–substrate interactions and phosphoproteomics data. For a recent review and benchmark, readers are referred to References 155 and 156. Kinase–substrate enrichment analysis (157) uses parametric tests to compare the changes of the substrates of one kinase to all other substrates. Cluster evaluation (158) clusters phosphorylation sites based on time series data, from which enrichments of kinase–substrate annotations are calculated. Inference of kinase activities from phosphoproteomics (159) uses machine learning to estimate the strength of kinase–substrate interactions, as well as kinase activities. Phosphoproteomic dissection using networks (PHOTON) (160) is a method using general PPI networks for interpreting phosphorylation data within their signaling context. PHOTON identifies proteins that significantly contribute to signaling and uses these proteins to reconstruct the most plausible signaling pathway from the PPI network (**Figure 7*b***).

For general-purpose network analysis, Cytoscape (161) has emerged as the de facto standard. Through its plugin infrastructure, it provides a wealth of analyses and visualizations, often integrating expression-omics technologies with interaction networks. Cytoscape reads networks from various standard formats and can extend them with interactions and pathways from various databases. Tools such as BiNGO (162) can identify significantly enriched gene ontology

terms in these networks. Large-scale networks can be clustered into modules, either by topology (MCODE; see Reference 163) or by differential expression (jActiveModules; see Reference 164). Alternatively, network reconstruction tools, such as ANAT (165), identify a subset of interactions connecting, for example, differentially expressed proteins to their signaling stimulus.

Multiomics Data Analysis

Analyzing data from two omics technologies applied to the same samples becomes straightforward if there is a near one-to-one match between the biomolecules measured in each of the two omics spaces. For instance, when comparing the proteome and the transcriptome, the one-to-one correspondence between transcript and protein sequences holds true with only little deviations due to, for example, translation errors and postprocessing of the protein sequence. Thus, the molecular correspondence is sufficiently valid to conceptually work with matching rows between the two omics matrices. The problem reduces to mapping transcript to protein identifiers and to dealing with the different depth in distinguishable splice variants, for which algorithmic solutions exist (28). A similar molecular correspondence can be applied to the genome–proteome spaces for correlating local genomic properties such as DNA copy number (166) or loss of heterozygosity to protein expression if proteins matching to the same gene model are grouped together. Also, ribosomal profiling data (167) can be brought into molecular correspondence with proteomics data.

Once a correspondence between omics spaces has been established, one can perform pointwise comparisons, as is done in the scatterplots in **Figure 8*a***, in which protein abundances, messenger RNA levels, and ribosomal profiling data are compared. Individual outliers in each of these plots may hint at interesting biology. However, it is difficult to assign significance to individual data points. Hence, researchers developed 2D annotation enrichment (168; **Figure 8*b***) to answer the question, Which classes of gene products show concordant and which show discordant behavior between the different levels of gene expression? While transcriptional regulation is a dominant factor in expression control, there are many known examples of posttranscriptional regulation like microRNA-controlled inhibition of transcripts (169) and directed protein degradation (170), which are detectable by this method.

Further examples of simultaneous multivariate analysis in two matched omics spaces are joint time series analysis, which is exemplified in **Figure 8*c*** for circadian transcriptomics, and proteomics data (136). Here, it was possible to derive time lags between peaks in transcript and protein abundances as a proxy for the time lag between transcription and translation for individual cycling transcripts and their associated proteins. Additionally, joint transcriptomics–proteomics PCA performed on the same data (**Figure 8*d***) indicates global similarities in transcript and protein concentrations, but with a time delay.

When the input is time-resolved data for transcriptome and proteome, protein expression control analysis (PECA) (171, 172) computes the probability of regulation changes between adjacent time intervals. PECA quantitatively dissects protein expression variation into the contributions of mRNA and protein synthesis–degradation rate ratios.

Unlike in the previous examples, when combining proteomics with metabolomics, there is not a one-to-one correspondence between molecules. In biochemical pathways, proteins are associated with reactions between metabolites as catalysts. The required mapping of biomolecules is facilitated by the consensus human metabolic reconstruction Recon 2.2 (173), which has a high potential for integrating and analyzing diverse data types. Recon 2.2 facilitates the integration of proteomics data with an updated curation of relationships between genes, proteins, and reactions.

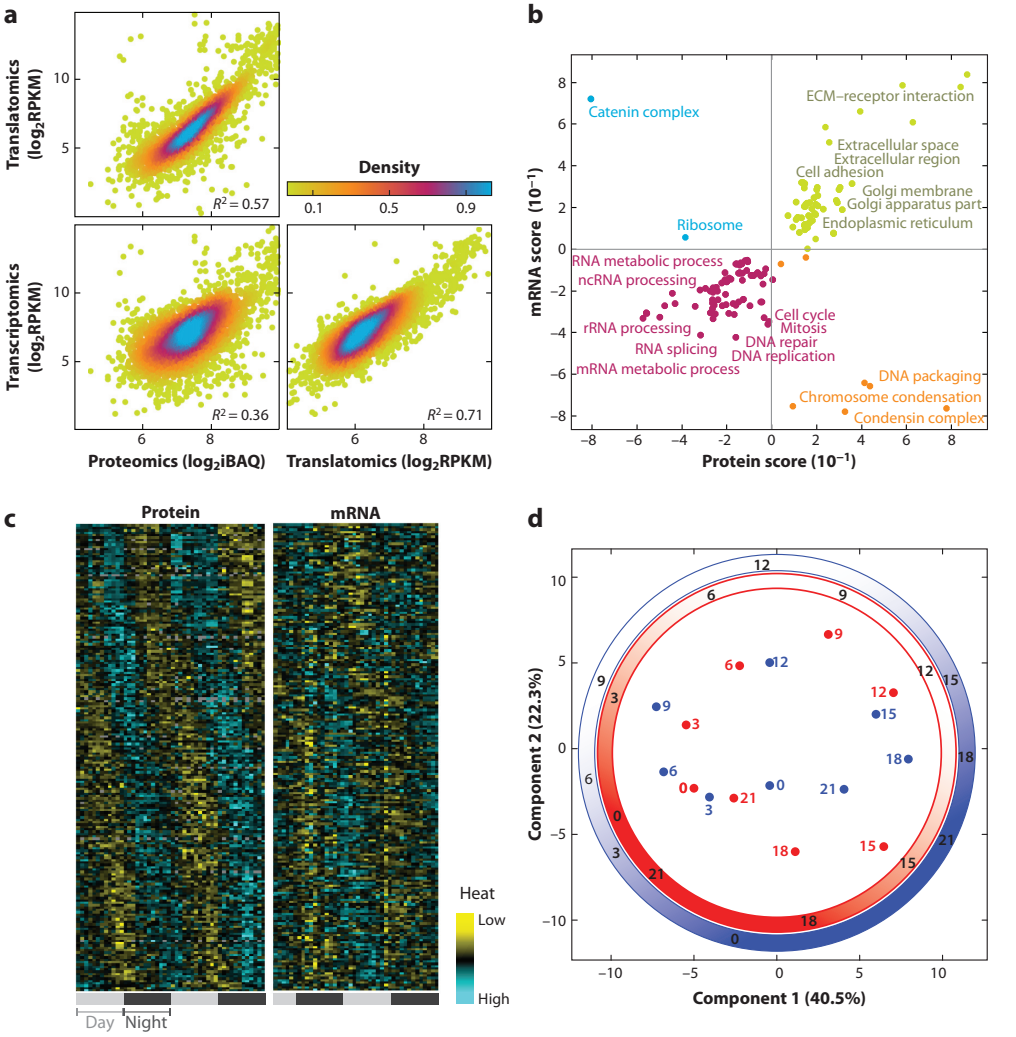


Figure 8

Cross-omics data analysis. (*a*) Comparison of protein abundances, ribosomal profiling data, and mRNA expression. Proteins are quantified with the iBAQ method (124), while RPKM (186) was used for the other two data types. Adapted from Reference 28. (*b*) Output of the two-dimensional enrichment analysis applied to protein and mRNA abundances. Adapted from Reference 28. (*c*) Side-by-side heat maps for daily rhythmic proteins and transcripts showing a cycling pattern. In the rows, samples are ordered by time of extraction, and in the columns, proteins are ordered by time of their peak concentration. Adapted from Reference 136. (*d*) Principal component analysis performed jointly on transcriptomics data (*red*) and proteomics data (*blue*) of two phases of circadian mouse liver data. Labels next to data points denote time in hours. Both transcriptomics and proteomics data points arrange in a periodic time series pattern in the first two principal components. Adapted from Reference 136. Abbreviations: ECM, extracellular matrix; iBAQ, intensity-based absolute quantification; mRNA, messenger RNA; ncRNA, noncoding RNA; RPKM, reads per kilobase per million mapped reads; rRNA, ribosomal RNA.

DISCUSSION AND OUTLOOK

Computational proteomics has matured substantially and is keeping up well with the massive amounts of data produced by modern mass spectrometers. Platforms for identification and quantification of proteins can analyze the data in a reliable and automated way. Therefore, attention is increasingly being shifted to the downstream part of the data analysis, in which the quantification results are interpreted, hypotheses are tested, and novel biological and biomedical knowledge is gained. We anticipate that future developments of computational proteomics tools will be particularly active in these areas, including network biology and cross-omics data analysis. In previous work (28), we made the case for enabling the end users—the researchers from fundamental biology, drug discovery, and medical sciences—to perform large parts of the data analysis themselves, and this is increasingly happening.

Single-cell DNA and RNA sequencing (174) have shed new light onto the heterogeneity and diversity of biological processes behind the cellular averages that are typically monitored in many omics technologies. According to reports in the literature (175), single-cell proteomics is just around the corner and will likely bear many new discoveries. Once it is scalable and sufficiently deep in terms of proteome coverage, it might help define a highly resolved atlas of all cell types and cell states in the human body (176). Certainly, novel computational tools will have to be developed for the particular challenges of single-cell proteomics data, which will likely have unique challenges in terms of normalization and handling of missing data.

There is still a large gap between the generation of large-scale proteomics data and the modeling of signaling pathways and biochemical reactions. The curated knowledge of PTMs currently available in public resources (134, 177) is still limited and needs to be expanded to support more comprehensive analyses. New tools are emerging to reconstruct signaling pathways and translate them into logic models (178). Hopefully, the path from large-scale time series data to kinetic modeling (179, 180) will become more accessible for many interdisciplinary researchers, leading to an improved mechanistic understanding of the biological processes under investigation based on large-scale data.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement 686547 and from the FP7 grant agreement GA ERC-2012-SyG_318987–ToPAG.

LITERATURE CITED

1. James P. 1997. Protein identification in the post-genome era: the rapid rise of proteomics. *Q. Rev. Biophys.* 30(4):279–331
2. Cox J, Mann M. 2011. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* 80:273–99
3. Altelaar AF, Munoz J, Heck AJ. 2013. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 14(1):35–48

4. Aebersold R, Mann M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature* 537(7620):347–55
5. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, et al. 2016. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* 7:13404
6. Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, et al. 2015. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* 522(7554):81–84
7. Wolters DA, Washburn MP, Yates JR. 2001. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 73(23):5683–90
8. Fornelli L, Durbin KR, Fellers RT, Early BP, Greer JB, et al. 2017. Advancing top-down analysis of the human proteome using a benchtop quadrupole-orbitrap mass spectrometer. *J. Proteome Res.* 16(2):609–18
9. Toby TK, Fornelli L, Kelleher NL. 2016. Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* 9:499–519
10. Chait BT. 2006. Mass spectrometry: bottom-up or top-down? *Science* 314(5196):65–66
11. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, et al. 2007. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* 35:W701–6
12. Kou Q, Xun L, Liu X. 2016. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 32(22):3495–97
13. Park J, Piehowski PD, Wilkins C, Zhou M, Mendoza J, et al. 2017. Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* 14(9):909–14
14. Gillette MA, Carr SA. 2013. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nat. Methods* 10(1):28–34
15. Liebler DC, Zimmerman LJ. 2013. Targeted quantitation of proteins by mass spectrometry. *Biochemistry* 52(22):3797–3806
16. Picotti P, Aebersold R. 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* 9(6):555–66
17. Ebhardt HA, Root A, Sander C, Aebersold R. 2015. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* 15(18):9193–208
18. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, et al. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26(7):966–68
19. Doerr A. 2014. DIA mass spectrometry. *Nat. Methods* 12(1):35–35
20. Rosenberger G, Bludau I, Schmitt U, Heusel M, Hunter CL, et al. 2017. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* 14(9):921–27
21. Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, et al. 2017. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteom.* 16(12):2296–309
22. Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, Hopfgartner G, et al. 2015. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* 15(5–6):964–80
23. Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, et al. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* 12(3):258–64
24. McDonnell LA, Heeren RMA. 2007. Imaging mass spectrometry. *Mass Spectrom. Rev.* 262007:606–43
25. Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26(12):1367–72
26. Tyanova S, Temu T, Cox J. 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11(12):2301–19
27. Tyanova S, Temu T, Carlson A, Sinitcyn P, Mann M, Cox J. 2015. Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* 15(8):1453–56
28. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, et al. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* 13(9):731–40
29. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, et al. 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Meth.* 13(9):741–48

30. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, et al. 2010. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10(6):1150–59
31. McIlwain S, Tamura K, Kertesz-Farkas A, Grant CE, Diament B, et al. 2014. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* 13(10):4488–91
32. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA. 2015. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 15(5–6):930–50
33. Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, et al. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 44(D1):D447–56
34. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, et al. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32(3):223–26
35. Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, et al. 2014. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteom.* 13(10):2765–75
36. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509(7502):582–87
37. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. 2014. A draft map of the human proteome. *Nature* 509(7502):575–81
38. Schaab C, Geiger T, Stoehr G, Cox J, Mann M. 2012. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell. Proteom.* 11(3):M111.014068
39. Desiere F. 2006. The PeptideAtlas project. *Nucleic Acids Res.* 34(90001):D655–58
40. UniProt Consort. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45(D1):D158–69
41. Mohimani H, Yang YL, Liu WT, Hsieh PW, Dorrestein PC, Pevzner PA. 2011. Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* 11(18):3642–50
42. Yates A, Akanni W, Amode MR, Barrell D, Billis K, et al. 2016. Ensembl 2016. *Nucleic Acids Res.* 44(D1):D710–16
43. Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. 2011. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 11(5):996–99
44. Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, et al. 2015. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* 33(1):22–24
45. Zhang J, Gonzalez E, Hestilow T, Haskins W, Huang Y. 2009. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genom.* 10(6):388–401
46. Miladinović SM, Kozhinov AN, Gorshkov MV, Tsybin YO. 2012. On the utility of isotopic fine structure mass spectrometry in protein identification. *Anal. Chem.* 84(9):4042–51
47. Snyder LR, Kirkland JJ, Dolan JW. 2010. *Introduction to Modern Liquid Chromatography*. Hoboken, NJ: Wiley
48. Kanu AB, Dwivedi P, Tam M, Matz L, Hill HH. 2008. Ion mobility–mass spectrometry. *J. Mass Spectrom.* 43(1):1–22
49. Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. 2006. Cluster-based analysis of FMRI data. *Neuroimage* 33(2):599–608
50. Senko MW, Beu SC, McLafferty FW. 1995. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 6(4):229–33
51. Rockwood AL, Van Orden SL, Smith RD. 1996. Ultrahigh resolution isotope distribution calculations. *Rapid Commun. Mass Spectrom.* 10(1):54–59
52. Horn DM, Zubarev RA, McLafferty FW. 2000. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 11(4):320–32
53. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *PNAS* 96(12):6591–96
54. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 389(4):1017–31
55. Ong SE, Mann M. 2007. Stable isotope labeling by amino acids in cell culture for quantitative proteomics. *Methods Mol. Biol.* 359:37–52

56. Hsu JL, Huang SY, Chow NH, Chen SH. 2003. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* 75(24):6843–52
57. Boersema PJ, Aye TT, van Veen TA, Heck AJ, Mohammed S. 2008. Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. *Proteomics* 8(22):4624–32
58. Engelsberger WR, Erban A, Kopka J, Schulze WX. 2006. Metabolic labeling of plant cell cultures with $K^{15}NO_3$ as a tool for quantitative analysis of proteins and metabolites. *Plant Methods* 2(3):14
59. Ippel JH, Pouvreau L, Kroef T, Gruppen H, Versteeg G, et al. 2004. In vivo uniform ^{15}N-isotope labelling of plants: using the greenhouse for structural proteomics. *Proteomics* 4(1):226–34
60. Cox J, Michalski A, Mann M. 2011. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* 22(8):1373–80
61. Cox J, Mann M. 2009. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* 20(8):1477–85
62. Podwojski K, Fritsch A, Chamrad DC, Paul W, Sitek B, et al. 2009. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* 25(6):758–64
63. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, et al. 2007. *SuperHirn*—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 7(19):3470–80
64. Pasa-Tolic L, Masselon C, Barry RC, Shen Y, Smith RD. 2004. Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* 37(4):621–36
65. Eng JK, McCormack AL, Yates JR. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5(11):976–89
66. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18):3551–67
67. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. 2004. Open mass spectrometry search algorithm. *J. Proteome Res.* 3(5):958–64
68. Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20(9):1466–67
69. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 10(4):1794–1805
70. Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4(3):207–14
71. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74(20):5383–92
72. Choi H, Nesvizhskii AI. 2008. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* 7(1):254–65
73. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* 4(11):923–25
74. Degroeve S, Martens L, Jurisica I. 2013. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* 29(24):3199–203
75. Tran NH, Zhang X, Xin L, Shan B, Li M. 2017. De novo peptide sequencing by deep learning. *PNAS* 114(31):8247–52
76. Taylor JA, Johnson RS. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 11(9):1067–75
77. Ma B, Zhang K, Hendrie C, Liang C, Li M, et al. 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17(20):2337–42
78. Ma B, Johnson R. 2012. *De novo* sequencing and homology searching. *Mol. Cell. Proteom.* 11(2):O111.014902
79. Han Y, Ma B, Zhang K. 2004. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *Proc. Comput. Syst. Bioinform. Conf., Stanford, Calif., 16–19 Aug.*, pp. 206–15. New York: IEEE
80. Bern M, Cai Y, Goldberg D. 2007. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* 79(4):1393–1400

81. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* 24(10):1285–92
82. Bailey CM, Sweet SMM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. 2009. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *J. Proteome Res.* 8(4):1965–71
83. Lemeer S, Kunold E, Klaeger S, Raabe M, Towers MW, et al. 2012. Phosphorylation site localization in peptides by MALDI MS/MS and the Mascot Delta Score. *Anal. Bioanal. Chem.* 402(1):249–60
84. Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, et al. 2011. Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell. Proteom.* 10(2):M110.003830
85. Taus T, Köcher T, Pichler P, Paschke C, Schmidt A, et al. 2011. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* 10(12):5354–62
86. Sharma K, D'Souza RC, Tyanova S, Schaab C, Wisniewski JR, et al. 2014. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* 8(5):1583–94
87. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, et al. 2015. A mass-tolerant database search—supplementary. *Nat. Biotechnol.* 33(7):743–49
88. Savitski MM. 2006. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteom.* 5(5):935–48
89. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. 2017. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* 14(5):513–20
90. Sinz A. 2006. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass Spectrom Rev.* 25(4):663–82
91. Singh P, Panchaud A, Goodlett DR. 2010. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Anal. Chem.* 82(7):2636–42
92. Hoopmann MR, Zelter A, Johnson RS, Riffle M, MacCoss MJ, et al. 2015. Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* 14(5):2190–98
93. Götze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, et al. 2012. StavroX—a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* 23(1):76–87
94. Liu F, Lössl P, Scheltema R, Viner R, Heck AJR. 2017. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* 8:15473
95. Yang B, Wu YJ, Zhu M, Fan SB, Lin J, et al. 2012. Identification of cross-linked peptides from complex samples. *Nat. Methods* 9(9):904–6
96. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, et al. 2010. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell Proteom.* 9(8):1634–49
97. Chen ZA, Fischer L, Cox J, Rappsilber J. 2016. Quantitative cross-linking/mass spectrometry using isotope-labeled cross-linkers and MaxQuant. *Mol. Cell Proteom.* 15:2769–78
98. Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11(11):1114–25
99. Temu T, Mann M, Räschle M, Cox J. 2016. Homology-driven assembly of NOn-redundant protEin sequence sets (NOmESS) for mass spectrometry. *Bioinformatics* 32(9):1417–19
100. Huang T, Wang J, Yu W, He Z. 2012. Protein inference: a review. *Brief. Bioinform.* 13(5):586–614
101. Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, et al. 2004. DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* 3(5):1002–8
102. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, et al. 2009. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* 8(8):3872–81
103. Slotta DJ, McFarland MA, Markey SP. 2010. MassSieve: panning MS/MS peptide data for proteins. *Proteomics* 10(16):3035–39
104. Alves P, Arnold RJ, Novotny MV, Radivojac P, Reilly JP, Tang H. 2007. Advancement in protein inference from shotgun proteomics using peptide detectability. *Proc. Pac. Symp. Biocomput., Maui, Hawaii, 3–7 Jan.*, pp. 409–20. **http://psb.stanford.edu/psb-online/proceedings/psb07/alves.pdf**
105. Sober E. 2017. *Ockham's Razors: A User's Manual*. Cambridge, UK: Cambridge Univ. Press
106. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75(17):4646–58

107. Serang O, MacCoss MJ, Noble WS. 2010. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* 9(10):5346–57
108. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, et al. 2009. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell Proteom.* 8(11):2405–17
109. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. 2015. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell Proteom.* 14:2394–404
110. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteom.* 13(9):2513–26
111. Gauthier NP, Soufi B, Walkowicz WE, Pedicord VA, Mavrakis KJ, et al. 2013. Cell-selective labeling using amino acid precursors for proteomic studies of multicellular environments. *Nat. Methods* 10(8):768–73
112. Merrill AE, Hebert AS, MacGilvray ME, Rose CM, Bailey DJ, et al. 2014. NeuCode labels for relative protein quantification. *Mol. Cell Proteom.* 13(9):2503–12
113. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. 2010. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* 7(5):383–85
114. Thompson A, Schäfer JJ, Kuhn K, Kienle S, Schwarz J, et al. 2003. Tandem mass tags: a novel quantificaiton strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75(8):1895–1904
115. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. 2004. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteom.* 3(12):1154–69
116. Rauniyar N, Yates JR. 2014. Isobaric labeling-based relative quantification in shotgun proteomics. *J. Proteome Res.* 13(12):5293–303
117. Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. 2009. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly." *J. Proteome Res.* 8(11):5347–55
118. Wenger CD, Lee MV, Hebert AS, McAlister GC, Phanstiel DH, et al. 2011. Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* 8(11):933–35
119. McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, et al. 2014. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* 86(14):7150–58
120. Wühr M, Haas W, McAlister GC, Peshkin L, Rad R, et al. 2012. Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* 84(21):9214–21
121. Savitski MM, Fischer F, Mathieson T, Sweetman G, Lang M, Bantscheff M. 2010. Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *J. Am. Soc. Mass Spectrom.* 21(10):1668–79
122. Michalski A, Cox J, Mann M. 2011. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* 10(4):1785–93
123. Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, et al. 2013. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* 12(8):3586–98
124. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. 2011. Global quantification of mammalian gene expression control. *Nature* 473(7347):337–42
125. Silva JC. 2005. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteom.* 5(1):144–56
126. Wisniewski JR, Hein MY, Cox J, Mann M. 2014. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell Proteom.* 13(12):3497–506
127. Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, et al. 2010. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* 3(104):ra3
128. Krzywinski M, Altman N. 2013. Points of significance: significance, P values and t-tests. *Nat. Methods* 10:1041–42

129. Krzywinski M, Altman N. 2014. Points of significance: Analysis of variance and blocking. *Nat. Methods* 11(7):699–700
130. Noble WS. 2009. How does multiple testing correction work? *Nat. Biotechnol.* 27(12):1135–37
131. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300
132. Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98(9):5116–21
133. Gene Ontol. Consort. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43:D1049–56
134. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, et al. 2016. The reactome pathway knowledgebase. *Nucleic Acids Res.* 44(D1):D481–87
135. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, et al. 2009. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* 38(Suppl.1):D646–50
136. Robles MS, Cox J, Mann M. 2014. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLOS Genet.* 10(1):e1004047
137. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43:D512–20
138. Perfetto L, Briganti L, Calderone A, Perpetuini AC, Iannuccelli M, et al. 2016. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* 44(D1):D548–54
139. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, et al. 2011. Phospho.ELM: A database of phosphorylation sites—update 2011. *Nucleic Acids Res.* 39(Suppl. 1):D261–67
140. Deeb SJ, Tyanova S, Hummel M, Schmidt-Supprian M, Cox J, Mann M. 2015. Machine learning based classification of diffuse large B-cell lymphoma patients by their protein expression profiles. *Mol. Cell Proteom.* 14(11):2947–60
141. Iglesias-Gato D, Wikstrom P, Tyanova S, Lavallee C, Thysell E, et al. 2015. The proteome of primary prostate cancer. *Eur. Urol.* 69(5):942–52
142. Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T. 2016. Proteomic maps of breast cancer subtypes. *Nat. Commun.* 7:10259
143. Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Int. Jt. Conf. Artif. Intell., 14th, Montr., Can., 20–25 Aug.*, pp. 1137–43. San Francisco: Morgan Kaufmann
144. Vapnik VN. 1995. *The Nature of Statistical Learning Theory*. New York: Springer
145. Schmidhuber J. 2015. Deep learning in neural networks: an overview. *Neural Netw.* 61:85–117
146. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
147. Itzhak DN, Tyanova S, Cox J, Borner GHH. 2016. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* 5:e16950
148. Itzhak DN, Davies C, Tyanova S, Mishra A, Williamson J, et al. 2017. A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell Rep.* 20(11):2706–18
149. Bensimon A, Heck AJR, Aebersold R. 2012. Mass spectrometry–based proteomics and network biology. *Annu. Rev. Biochem.* 81:379–405
150. Keilhauer EC, Hein MY, Mann M. 2015. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteom.* 14(1):120–35
151. Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, et al. 2015. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163(3):712–23
152. Sowa ME, Bennett EJ, Gygi SP, Harper JW. 2009. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 138(2):389–403
153. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, et al. 2015. The BioPlex network: a systematic exploration of the human interactome. *Cell* 162(2):425–40
154. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, et al. 2008. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* 36(Suppl. 1):D695–99
155. Dermit M, Dokal A, Cutillas PR. 2017. Approaches to identify kinase dependencies in cancer signalling networks. *FEBS Lett.* 591(17):2577–92

156. Hernandez-Armenta C, Ochoa D, Gonçalves E, Saez-Rodriguez J, Beltrao P. 2017. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* 33(12):1845–51
157. Casado P, Rodriguez-Prados J-C, Cosulich SC, Guichard S, Vanhaesebroeck B, et al. 2013. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* 6(268):rs6
158. Yang P, Zheng X, Jayaswal V, Hu G, Yang JYH, Jothi R. 2015. Knowledge-based analysis for detecting key signaling events from time-series phosphoproteomics data. *PLOS Comput. Biol.* 11(8):e1004403
159. Mischnik M, Sacco F, Cox J, Schneider HC, Schäfer M, et al. 2015. IKAP: A heuristic framework for inference of kinase activities from phosphoproteomics data. *Bioinformatics* 32(3):424–31
160. Rudolph JD, de Graauw M, van de Water B, Geiger T, Sharan R. 2016. Elucidation of signaling pathways from large-scale phosphoproteomic data using protein interaction networks. *Cell Syst.* 3(6):585–93
161. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13(11):2498–2504
162. Maere S, Heymans K, Kuiper M. 2005. *BiNGO*: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21(16):3448–49
163. Bader GD, Hogue CW. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* 4:2
164. Ideker T, Ozier O, Schwikowski B, Siegel AF. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1):S233–40
165. Yosef N, Zalckvar E, Rubinstein AD, Homilius M, Atias N, et al. 2011. ANAT: a tool for constructing and analyzing functional protein networks. *Sci. Signal.* 4(196):pl1
166. Geiger T, Cox J, Mann M. 2010. Proteomic changes resulting from gene copy number variations in cancer cells. *PLOS Genet.* 6(9):e1001090
167. Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15(3):205–13
168. Cox J, Mann M. 2012. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinform.* 13(Suppl. 1):S12
169. He L, Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* 5(7):522–31
170. Hochstrasser M. 1996. Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* 30:405–39
171. Teo G, Vogel C, Ghosh D, Kim S, Choi H. 2014. PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation. *J. Proteome Res.* 13(1):29–37
172. Cheng Z, Teo G, Krueger S, Rock TM, Koh HW, et al. 2016. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol. Syst. Biol.* 12(1):855–855
173. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, et al. 2016. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* 12(7):109
174. Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, et al. 2017. Challenges and emerging directions in single-cell analysis. *Genome Biol.* 18(1):84
175. Budnik B, Levy E, Slavov N. 2017. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. bioRxiv 102681. **https://doi.org/10.1101/102681**
176. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Yosef N. 2017. The Human Cell Atlas. bioRxiv 121202. **http://dx.doi.org/10.1101/121202**
177. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44(D1):D457–62
178. Terfve CDA, Wilkes EH, Casado P, Cutillas PR, Saez-Rodriguez J. 2015. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* 6:8033
179. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, et al. 2006. COPASI–a COmplex PAthway SImulator. *Bioinformatics* 22(24):3067–74
180. Angermann BR, Klauschen F, Garcia AD, Prustel T, Zhang F, et al. 2012. Computational modeling of cellular signaling processes embedded into dynamic spatial contexts. *Nat. Methods* 9(3):283–89
181. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246(4926):64–71

182. Hillenkamp F, Karas M, Beavis RC, Chait BT. 1991. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* 63(24):A1193–1203
183. Eliuk S, Makarov A. 2015. Evolution of orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem.* 8:61–80
184. Meier F, Beck S, Grassl N, Lubeck M, Park MA, et al. 2015. Parallel accumulation-serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* 14(12):5378–87
185. Graumann J, Hubner NC, Kim JB, Ko K, Moser M, et al. 2008. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell Proteom.* 7(4):672–83
186. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5(7):621–28

Chapter 2

List of publications

2.1 Proteomics software development

2.1.1 MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics

In DDA, a subset of peptide precursors is selected from MS^1 for a subsequent fragmentation and identification in MS^2. Due to the selection of the topN most abundant precursors, DDA technique is usually biased towards highly expressed proteins. Therefore low abundant peptides are identified in a stochastic manner. This fact complicates the analysis of thousands of samples because of the numerous missing values. Alternatively, DIA is considered as a state-of-art proteomics technique for large-scale clinical and fundamental studies, due to its robustness and accuracy of protein quantification[153]. The classical DIA experiment requires to first generate a spectral library of peptides using DDA method. The drawback of this technique is that peptides, which are not represented in the library, will not be identified.

MaxQuant is a well-known software for the processing of DDA shotgun proteomics[3, 150]. This manuscript presents new features of MaxQuant, which allows the analysis of DIA data. Those features include an iterative approach for identifying library-to-DIA matches (bootstrap-DIA algorithm) and an accurate false discovery rate estimator enhanced by machine-learning. Furthermore, the MaxLFQ algorithm, that was originally applied to DDA[147], was extended to DIA to normalize intensities of fragment matches. This improvement allows very accurate protein quantification as prooved by multi-species mix experiments.

I contributed to this study by developing and testing MaxQuant-DIA. Also, I implemented and tested the discovery mode DIA pipeline, which does not require

to have a measured DDA library. Recent advances in the prediction of MS^2 spectrum based on peptide sequence[156] allows constructing a library from in silico prediction instead of DDA measurement. Thus, DIA experiment does not depend on the depth of DDA library. From the user perspective, the discovery mode reduces the design complexity of DIA.

Pavel Sinitcyn[*], Hamid Hamzeiy[*], Favio Salinas Soto[*], Daniel Itzhak, Frank McCarthy, Christoph Wichmann, Martin Steger, Uli Ohmayer, Ute Distler, Stephanie Kaspar-Schoenefeld, Nikita Prianichnikov, Şule Yılmaz, Jan Daniel Rudolph, Stefan Tenzer, Yasset Perez-Riverol, Nagarjuna Nagaraj, Sean J. Humphrey and Jürgen Cox
MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics
(2020) *The manuscript is in under revision process at Nature Biotechnology journal*

[*]these authors contributed equally to this work

MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics

Pavel Sinitcyn[1,#], Hamid Hamzeiy[1,#], Favio Salinas Soto[1,#], Daniel Itzhak[2], Frank McCarthy[2], Christoph Wichmann[1], Martin Steger[3], Uli Ohmayer[3], Ute Distler[4], Stephanie Kaspar-Schoenefeld[5], Nikita Prianichnikov[1], Şule Yılmaz[1], Jan Daniel Rudolph[1,6], Stefan Tenzer[4], Yasset Perez-Riverol[7], Nagarjuna Nagaraj[5], Sean J. Humphrey[8] and Jürgen Cox[1,9,*]

[1]Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.

[2]Chan Zuckerberg Biohub, 499 Illinois St., San Francisco, CA 94158, USA.

[3]Evotec München GmbH, Am Klopferspitz 19a, 82152 Martinsried, Germany.

[4]Institute for Immunology, Johannes Gutenberg University, Langenbeckstraße 1, 55131 Mainz, Germany.

[5]Bruker Daltonik GmbH, Farenheitstr. 4, 28359 Bremen, Germany.

[6]Bosch Center for Artificial Intelligence, Robert-Bosch-Campus 1, 71272 Renningen, Germany

[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD UK

[8]School of Life and Environmental Sciences, Charles Perkins Centre, University of Sydney, John Hopkins Drive, Camperdown, NSW 2006, Australia.

[9]Department of Biological and Medical Psychology, University of Bergen, Jonas Liesvei 91, 5009 Bergen, Norway.

[#]These authors contributed equally to the publication.

[*]Correspondence: cox@biochem.mpg.de

Abstract

MaxDIA is a universal platform for analyzing data-independent acquisition proteomics data within the MaxQuant software environment. Using spectral libraries, MaxDIA achieves cutting-edge proteome coverage with significantly better coefficients of variation in protein quantification than other software. MaxDIA is equipped with accurate false discovery rate estimates on both library-to-DIA match and protein levels, also when using whole-proteome predicted spectral libraries. This is the foundation of discovery DIA – a framework for the hypothesis-free analysis of DIA samples without library and with reliable FDR control. MaxDIA performs three- or four-dimensional feature detection of fragment data and scoring of matches is augmented by machine learning on the features of an identification. MaxDIA's novel bootstrap-DIA workflow performs multiple rounds of matching with increasing quality of recalibration and stringency of matching to the library. Combining MaxDIA with two new technologies, BoxCar acquisition and trapped ion mobility spectrometry, both lead to deep and accurate proteome quantification.

Data-independent acquisition (DIA) proteomics[1] promises robust and accurate quantification of proteins over large-scale study designs and across heterogeneous laboratory conditions[2]. In all omics sciences, robust data analysis pipelines are as important as the data acquisition technology itself, and proteomics is no exception. MaxQuant[3–6] is the most widely-used software for analyzing data-dependent acquisition (DDA) proteomics data, providing a vendor-neutral complete end-to-end solution for all common experimental designs. With version 2.0 described here, MaxQuant offers an equally complete DIA software infrastructure, termed MaxDIA. Such a unified framework over all mass spectrometry-based proteomics based on peptide quantification comes with several advantages over existing software[7–10]. DDA libraries and DIA samples can be processed in integrated, consistent ways. Algorithmic parts of the workflow that do not depend on the type of acquisition, like protein quantification algorithms, such as MaxLFQ[11], protein redundancy grouping, or protein-level false discovery rate (FDR) can be applied to all data in exactly the same way, making DDA and DIA studies much more comparable.

The classical approach to DIA data analysis utilizes a spectral library of peptides which are queried in the DIA samples and quantified in case of their presence. In this spectral library-based approach, the rate of false matches can in principle be controlled with techniques similar to those developed in DDA proteomics[12]. For instance, the target-decoy method[13] has been adapted to DIA[9]. Additionally, several library-free approaches exist[14] and spectral predictions have been successfully used for DIA data analysis[15–20]. However, effective control of false discovery rates, in particular on the level of identified proteins with these methods is still a critical aspect. Once this is achieved, DIA can additionally be employed in a discovery mode, without biases imposed by a library and with the certainty that the identified set of proteins contains at most a predefined percentage of false positives, e.g. 1%, as is standardly applied in DDA-based proteomics. Here we demonstrate that MaxDIA fulfills these criteria and can indeed be used in such a discovery DIA mode.

Machine learning is an integral part of MaxDIA. We use the bidirectional recurrent neural network[21] (BRNN) approach termed DeepMass:Prism[15] to create *in silico* very precise libraries of MS/MS spectra for peptides digested from complete proteome sequence databases. BRNNs are also used for the dataset-specific prediction of liquid chromatography retention times. Furthermore, to score library DIA sample matches based on multivariate information derived from properties of the matches, we apply the gradient boosting method XGBoost[22], which is highly superior to only using the matching score itself, and also compared to applying other machine learning approaches.

High-quality three-dimensional (3D) or, in the presence of ion mobility data, 4D feature detection[3,23] of the precursor data is one of the most important ingredients of MaxQuant for DDA data, leading to efficient noise

suppression. In MaxDIA, fragment ions are additionally detected as 3D/4D features. Besides noise removal, this ensures that data is not over-interpreted: The feature detection on fragment data allows to require that all signals belonging to a 3D/4D peak contribute as evidence only to one peptide identification, ensuring that signals at slightly different retention times or ion mobility values, but really belonging to the same feature, are not used as independent evidence for two similar peptides, e.g. differing by a modification or resulting from an amino acid polymorphism.

In MaxDIA we support two new and promising technologies, both of which enable deep quantification of DIA samples. One is to combine DIA with high dynamic-range precursor data obtained by the BoxCar acquisition method[24]. The second is to utilize ion mobility as an extra data dimension on a timsTOF Pro instrument[25–27] for DIA. Both increase the quantified proteome in DIA samples substantially providing highly precise and linear quantification over the whole dynamic range. Furthermore, since the MaxLFQ algorithm has been designed to perform label-free quantification on pre-fractionated samples[11], also MaxDIA has the capability to perform label-free quantification of pre-fractionated samples analyzed by DIA, which opens up applications of DIA requiring ultra-deep proteome quantification. Complete submissions to the PRoteomics IDEntifications[28] (PRIDE) database using an adapted mzTab[29] scheme can also be performed automatically using MaxDIA.

RESULTS

MaxDIA data analysis workflow

MaxDIA is embedded into the MaxQuant software environment (Fig. 1) and shares with it the graphical user interface, computational infrastructure, and many algorithmic workflow components applicable to both. It is vendor-neutral, with direct support for the most common native vendor file formats for reading mass spectra, as well as the open mzML file format[30]. MaxDIA can be operated in a classical library-based approach or in discovery DIA mode. In the former, DIA datasets are interrogated within MaxQunat by spectral libraries generated with MaxQuant, while the latter does not require acquisition of a spectral library. In discovery DIA mode, spectral libraries are generated by DeepMass:Prism[15], a bidirectional recurrent neural network that enables precise prediction of spectral intensities from peptide sequences. Decoy spectra are generated by reverting library sequences under the constraint of preserving the cleavage characteristics of the protease that was used in the experiment and ensuring that the decoy peptide masses, retention times and ion mobility values follow the same multivariate distribution as the target peptides. DIA samples and libraries are then analyzed in an end-to-end workflow for peptide and protein identification and quantification. MaxQuant's three-dimensional (3D) or four-dimensional (4D) feature detection[3,23] (Fig. 2) and de-isotoping is performed on the

precursor data and on all LC-MS/MS or LC-IMS-MS/MS fragment data domains corresponding to precursor selection windows. Defining MS/MS features in a multi-dimensional way is particularly important for fragment data, since it avoids over-interpretation of identification results. This enables the requirement that every MS/MS feature is used at most once in peptide identification. Problems may arise if such precautions are not taken, since features will be double-counted for the identification of peptides that are similar to each other due to sequence homology or due to the presence or absence of a modification, but for which there is insufficient evidence for the existence of both peptide forms.

Bootstrap-DIA

Central to the workflow is bootstrap-DIA, which consists of multiple steps of matching the library spectra to DIA samples (Supplementary Fig. 1). These steps aim to bootstrap the DIA identification process based on the least possible prior knowledge. Bootstrap-DIA replaces and substantially extends the concept of the 'first search-main search' strategy[31] as well as the 'retention time alignment' and 'match between runs' used in DDA MaxQuant. Increasingly more information is gained in each round, with this information utilized in subsequent rounds. For instance, in the first round of matching, no retention time constraint is used. Based on these matches, a linear model is fit between the library and sample retention times, which is used to align runs to one another, even when gradient lengths substantially differ. This linear correction can be applied to the data and in the second round of matching, retention times can be filtered based on a time window that is automatically adapted to the distribution of all retention time differences after linear alignment. This filtering removes sufficiently many false positive matches, so that from the third round of matching a nonlinear retention time recalibration function can be determined. Application of the nonlinear recalibration function allows to subsequently apply more stringent filtering. Similar multi-step recalibration and filtering steps are applied to precursor and fragment masses, as well as to collision cross sections, if applicable. Supplementary Fig. 2 shows how target decoy distributions are affected after each matching step with increasingly more stringent filers. The resulting nonlinear precursor and fragment m/z recalibrations depending on m/z and retention time are shown in Supplementary Figs. 3,4.

A consequence of the bootstrap-DIA process is that precursor and fragment masses, retention times and ion mobility values are nonlinearly aligned between each DIA sample and library without the need for spike-in standards. A prerequisite for this is that the DDA runs in the datasets used for the library are well aligned to each other, since the precision of alignment between library and DIA samples is otherwise limited by the variability of retention times and collision cross sections within the library. Therefore, when processing libraries in MaxQuant, retention time and ion mobility alignments should be activated. A challenging attribute that can be learned from the data are nonlinear retention time mappings between library and samples. This

means that gradients between library and DIA runs do not need to be the same, and label-free quantification is possible even between DIA measurements with different gradients lengths. To evaluate the matching of different DIA gradient durations to a library we generated a DDA library consisting of 16 high pH-reversed phase fractions of a HeLa cell lysate measured with 25-minute gradients, and measured the same sample unfractionated with DIA using 30, 60, 90 and 120-minute gradients. Supplementary Fig. 5 shows retention time alignments between the library and DIA samples, and precise quantification between samples with different gradient lengths are shown in Supplementary Fig. 6. These capabilities greatly enhance the flexibility of MaxDIA, making the software applicable to analyzing a broader range of samples.

Scoring of library-to-sample matches by machine learning

To quantify the quality of match between a library spectrum and a DIA sample at a given retention time (and CCS value) we first find a precursor feature and all fragment features that match to the library spectrum with tolerances for m/z, retention time and CCS, dependent on the matching step in the bootstrap-DIA workflow. To measure the match quality, we then calculate a score which is the sum over all matching features of numbers between zero and one, each quantifying how far away from the apex the respective peak was hit (Supplementary Fig. 7). For a given library spectrum this score is maximized over retention time and ion mobility. It is then ensured, through a second round of scoring, that every feature in a DIA sample is used at most for one library spectrum match.

This score then is enhanced through machine learning. To this end, we construct a feature space that in addition to the score contains various properties of the match (Supplementary Fig. 8), such as mass errors (in p.p.m.) for precursor and fragments compared to masses calculated from elemental compositions, retention time and ion mobility errors, and apex fractions. We employ a classification algorithm to separate target from decoy hits based on this feature space. We define the machine learning based match score as the assignment probability to the target class of the machine learning algorithm. This is not just the binary decision of the classifier, but a number expressing the affinity to the target spectra as opposed to the decoy spectra. To eliminate the risk of overfitting, we determine these machine learning scores in 5-fold cross validation, such that a match for which the machine learning score is calculated has not been used for training the model that is used for its prediction.

We used several different classification algorithms and monitored their effect on the identification performance of MaxDIA. We compared the performances of XGBoost[22], fully connected multi-hidden layer neural networks, random forests[32] and AdaBoost (Supplementary Fig. 9) scanning for each algorithm suitable ranges of meta-parameters. We found that XGBoost performs best among the tested algorithms, in contrast to

Demichev et al.[10] who found neural networks to perform favorably. This choice is also different from DDA where for similar purposes support vector machine based methods are used[33]. XGBoost provides information on the importance of features for classification (Supplementary Fig. 8). We found that in the library-based approach, the feature defining whether the precursor has an isotope pattern assigned or was only seen as a single peak is of greater importance than the raw score itself. Furthermore, retention time, precursor mass errors, number of modifications and missed cleavages were among the top 10 highest ranked features. Also among the top 10 is the 'sample fragment overlap' which quantifies if and to what extent the N- and C-terminal ion series are overlapping in the DIA sample, thereby placing restrictions on the precursor mass.

Identification performance and quantification precision

To evaluate the performance of MaxDIA we ran it, and Spectronaut 13 on a dataset comprising 27 technical replicate injections of peptides derived from the human HepG2 cell line measured in DIA as well as a DDA library created from 12 high pH-reversed phase fractions (see online Methods). Using default parameters in both software, including a 1% FDR on precursor and protein levels, we obtain 6,238 protein groups mapped to Entrez gene identifiers with MaxDIA, compared to 6,015 with Spectronaut with an overlap of 5,549 (Fig. 3a). MaxDIA finds 20% more peptides than Spectronaut at 1% library-to-DIA matches FDR. The length distribution of identified peptides is very similar between the two analysis software (Fig. 3b).

While DIA is believed to be better in terms of data completeness[34,35] compared to DDA, we observe that this depends on the algorithmic details and that there is a tradeoff between data completeness and confidence of protein identification within a specific sample, as opposed to the whole dataset. After identifying peptides and proteins for the whole dataset, we apply a 'transfer q-value' cutoff to the identifications of matches in each sample. Setting it to 1, implies that no sample-specific restrictions are applied and that the peptide is quantified, whenever any evidence is found for its existence. A transfer q-value of 0.01 (equal to the global q-value of library-to-sample matches) results in stringent identification in every sample and hence, certainty about the actual sample-specific presence of peptides and proteins. We scanned through 7 values of the transfer q-value between 0.01 and 1 and monitored the number of proteins which have a certain number or less valid values in terms of LFQ intensities (Fig. 3c). As expected, for larger transfer q-values, the curves are flatter and higher in terms of total protein numbers. When using 1 for the 'minimum ratio count' parameter of the LFQ algorithm, most parts of all curves are above the line for the Spectronaut software. For 'minimum ratio count' = 2, which ensures higher accuracy of quantification, the array of curves is intersecting with the Spectronaut curve. After evaluating the accuracy of benchmark quantification results on several mass spectrometry platforms we decided to select 0.3 as the default value for the transfer q-value. Study-specific objectives

(completeness of quantification vs. certainty of identification in individual samples) may suggest deviations from this default value.

The distribution of coefficients of variation (CVs) (Fig. 3d) indicates substantially higher quantification precision obtained with MaxLFQ (described below) in MaxDIA compared with Spectronaut, with median CVs of 0.072 and 0.109, respectively. Fig. 3e.f show typical log-log scatter plots of protein intensities between replicates displaying less outliers and higher Pearson correlation for MaxDIA. All pair-wise replicate Pearson correlations of logarithmic intensities are represented as a heat map in Fig. 3g for both programs, showing consistently higher correlations for MaxDIA (median 0.993) compared to Spectronaut (median 0.977). We find a good overall agreement between averaged Spectronaut intensities and MaxDIA iBAQ values (Fig. 3h) with a Pearson correlation of 0.87. We performed mRNA vs. protein copy number comparisons based on RPKM[36] and iBAQ[37] values, respectively, using MaxDIA and Spectronaut (Fig. 3i,j). Both comparisons show similar correlations between mRNA and protein levels, which are also compatible with correlations typically found in such studies[38].

Accuracy of FDR estimates and discovery DIA

In order to evaluate the reliability of FDR estimates using MaxDIA's target-decoy strategy, we used a pooled DDA library generated from mixed human and maize samples, with corresponding DIA runs comprising only human samples[34]. Hence, every match identified as being derived from the maize proteome is a known false positive identification (having discarded peptides that are shared between proteins of the two species). This enables calculation of an 'external' FDR which is calculated independently of the 'internal' FDR estimated by the decoy approach in MaxDIA. Fig. 4a compares internal and external FDRs on match, peptide and protein group levels. The curves for internal and external FDR are in very good agreement on all three levels. When comparing the numbers of identified matches, peptides and protein groups at 1% FDR, which is often taken as a default value in shotgun proteomics, the numbers differed only by 3.0%, 3.4% and 5.0%, respectively, between internally and externally controlled FDR. Hence our decoy-based FDR estimates are in good agreement with external FDR calculations.

Given these results, we investigated how accurate the FDR estimates are for cases in which the library is dissimilar to the DIA sample. Hence, we assembled a library of *in-silico* predicted spectra based on DeepMass:Prism[15] consisting of all tryptic peptides digested from all human UniProt[39] sequences (Release 2019_05 containing 20959 proteins) without missed cleavages. We additionally generated predicted retention times for each *in-silico* spectrum based on a bidirectional recurrent neural network used previously for the same purpose[15]. Using this library with the same DIA dataset as in Fig 4a, we generated the same curves for

internal and external FDRs as before (Fig. 4b). Also here we observed good agreement between internal and external FDRs. In particular, at an FDR of 1% the number of identified protein groups differed only by 1.5%. We do however identify 39% more protein groups with the *in-silico* library compared to with the measured library. This highlights that MaxDIA does not require that spectral libraries are generated from matching samples in a project-specific manner, and yet FDRs are still reliably controlled. This enables the use of MaxDIA in a 'discovery' mode (discovery DIA), which is not biased by a library and completely hypothesis-free in terms of which proteins can be found, by using *in-silico* predicted libraries for all protein sequences.

We additionally repeated these analyses using the raw matching score instead of the machine learning-improved score (Fig 4c,d). This revealed that the agreement of internal and external FDR does not depend on whether the XGBoost-based machine learning was used to adjust the scoring. However, the use of machine learning does substantially increase peptide (83% and 58% for library DIA and discovery DIA, respectively) and protein group identifications (28% and 18%, respectively).

MaxLFQ adaptation for DIA

A prime example of the re-use and continued development of algorithms from DDA MaxQuant to MaxDIA is the label-free quantification algorithm, MaxLFQ[11]. Here, quantification is based on first calculating all pair-wise peptide ratios between samples, which are then summarized by the intensity profile that best fits all the pair-wise ratios. This procedure can be generalized to DIA by replacing a single ratio per peptide with multiple ratios derived from precursor intensities and from the most intense fragment peaks (Supplementary Fig. 10). This approach naturally implements hybrid quantification of precursor and fragment intensities.

To benchmark quantification accuracy, we downloaded a four-species dataset with well-defined small ratios between replicate groups[34]. Ratios are expected to be 0%, 10%, 20% or 30%, depending on the species comprising: *H. sapiens*, *C. elegans*, *S. cerevisiae* and *E. coli*. We tested several combinations of precursor, fragment or mixed quantification and fragment intensities summed up or kept separately. We measured the variability as the inter-quartile range of ratios within each species, and summed these over the four species (Fig. 5a). We found that hybrid quantification between precursors and fragments with fragment intensities kept separate for individual ion types in LFQ resulted in the smallest quantification errors measured as the sum of the inter-quartile ranges of ratio distributions over the four species. The accuracy observed exceeded both MS1- and MS2-level quantification reported by Bruderer et al.[34]. A further question is how the filtering of fragments by their intensity improves quantification accuracy. To this end, we used only the top-N intense peaks for quantification while varying N (Supplementary Fig. 11a). We found that accuracy increases with the number of fragments used, indicating that no filtering of fragments by intensity is required. Similarly, we

investigated, if filtering to top-N most intense peptides per protein is beneficial (Supplementary Fig. 11b), finding that it is best to use all available peptides.

Next, we analyzed a quantitative benchmark dataset obtained on SCIEX TripleTOF 6600 instrument, mixing proteomes from three species in defined ratios between replicate groups[2] (Fig. 5b). Using the original library analyzed with MaxQuant and using default values for all parameters, we identify 4,627 protein groups and achieve linear quantification for all three species over the whole dynamic range. In discovery mode with a predicted library allowing for one missed tryptic cleavage, the number of identified protein groups raises by 48% to 6,858 (Fig 5c) with on average improved quantification accuracy for the species with ratios as measured by inter-quartile ranges of species-specific ratio distributions. Importantly, *H. sapiens* which expresses a much larger number of proteins received the largest increase, identifying almost 2-fold more protein groups (4,012 vs. 2,127), while *C. elegans* and *E. coli* received proportionally fewer additional proteins.

We next acquired a quantitative three-species benchmark dataset utilizing ion mobility on a Bruker timsTOF Pro instrument. Using the DDA library acquired on the same instrument type, we identify 10,352 protein groups. We again used MaxLFQ for DIA with hybrid quantification with separate intensities for each fragment ion (Fig. 5d), seeing excellent quantification over the whole dynamic range without nonlinearities. In discovery mode (Fig. 5e), the number of identified protein groups increases to 10,466 with higher quantification accuracy, again judged by the inter-quartile ranges of ratio distributions. Scanning through the transfer q-value, we found that quantification accuracy was best with a value near 0.3 (Supplementary Fig. 12).

BoxCar and fractionated DIA

We recently implemented analysis of data acquired using the BoxCar acquisition method in MaxQuant in the DDA context[24], whose primary goal is to achieve higher dynamic range for the precursor intensities. Since this should be beneficial for DIA as well, we implemented its generalization to combining high-dynamic range precursor measurements with DIA acquisition for the fragments. Furthermore, it is possible with MaxDIA to analyze and quantify DIA samples that have been pre-fractionated on peptide or protein levels. To showcase these features, we acquired both DDA libraries and DIA measurements from HEK cell lysate as single shots and as high-pH reversed phase peptide fractionated samples, which were pooled into eight fractions for MS analysis (see Online Methods). We analyzed all combinations of libraries and samples, and in addition we analyzed the DIA samples in discovery-DIA mode allowing for one missed trypsin cleavage (Fig. 6a). For the fractionated DIA samples we observe an increase in the number of identified protein groups concomitant with

the size of the library, with the most identifications in discovery mode. With single shot samples, the number of identified proteins saturates with library size, having slightly more identifications with the fractionated library. However, comparing identifications for the single shot DIA samples between fractionated library and discovery mode, we find that the results are very similar with 89% overlap of Entrez gene identifier mapped protein groups (Supplementary Fig. 13). This indicates that for both types of DIA samples it is not compulsory to produce a deep, fractionated library, but that comparable or even better results can be achieved in discovery DIA mode. Quantification with MaxLFQ between three replicates of fractionated DIA samples shows very good correlation with a median Pearson correlation of 0.993 (Fig. 6b).

We then compared the results obtained with the three different library-creation approaches to RNA-seq data of HEK cells (see Online Methods). Fig. 6c compares the four sets of identifications based on gene identifiers. Out of the 9,503 genes covered by proteomics methods, 65% were found with all three library methods. Additional 25% were found with both, discovery mode and fractionated library, but not with the single shot library. 608 proteins were uniquely found with the discovery approach, compared to 251 with the deep fractionated library, suggesting preference for the discovery mode from the perspective of results, in addition to its economic advantages. In Fig. 6d, the results from Fig. 6c are displayed according to RPKM intervals of the RNA-seq data. The RNA-seq data shows a bimodal left shoulder that is typical of expression noise[40], genes for which there is only limited proteomic evidence of translation. As expected, highly abundant proteins are recovered with all methods, while at low abundance, both the deep-fractionated library and discovery DIA approach add identifications.

DISCUSSION

Here we introduce MaxDIA, a complete end-to-end DIA workflow embedded into the MaxQuant environment with major new features and broad applicability to established and novel mass spectrometry technologies. We demonstrate the widespread and general utility of the software, including its use in analyzing BoxCar-DIA and ion mobility DIA data, demonstrating very high proteome quantification coverage.

This framework lends itself to several extensions which are currently under development. In particular, while the analysis of posttranslational modifications (PTMs) is possible in principle by providing suitable libraries with spectra from modified peptides, proper localization of the modification on the peptide has to be carefully implemented as an additional process following peptide identification[41]. For these purposes, a PTM score guiding localization needs to be calculated directly from the DIA data and not from extracted spectra. Similarly, extensions to the identification of cross-linked peptides are straightforward[42] and are planned for future releases of MaxDIA.

ONLINE METHODS

HepG2 technical replicate data

Cell culture and MS sample preparation. HepG2 were from ATCC and cultured in MEM and 10% FCS. Cells were washed twice with ice-cold PBS and harvested using freshly prepared SDC buffer (1% SDC, 10 mM TCEP, 40 mM CAA, 75 mM Tris-HCl at pH= 8.5). The SDC lysates were heated to 95°C for 10 min while shaking at 750 rpm in a Thermomixer (Eppendorf) and then sonicated for 10 min (10 x 30 sec on/off cycles) using a Bioruptor® Pico sonication device (Diagenode). Protein concentrations were determined using the 660 nm assay (Thermo Fisher Scientific) and the proteins were digested with trypsin/Lys-C mix (Promega, V5071) overnight at 37°C with a 1:50 enzyme to protein ratio. The digestion was stopped by adding two volumes of 99% ethylacetate/1% TFA, followed by sonication for 1 min using an ultrasonic probe device (energy output ~40%). The samples were then desalted using in-house prepared, 200 µl two plug SDB-RPS StageTips[43] (3M EMPORETM, 2241). SDB-RPS StageTips were conditioned with 60 µl isopropanol, 60 µl 80% ACN/5% NH4OH and 100 µl 0.2% TFA. The SDC/ethylacetate mixture was directly loaded onto the tips followed by two washing steps of 200 µl 0.2% TFA each. Peptides were eluted with 80% ACN/5% NH4OH, speedvac dried and then resupended in 0.1% FA. After estimation of the concentration using a nanodropTM device (Thermo Fisher Scientific), the samples were adjusted to 0.4 µg/µl with 0.1% FA, of which 2 µl (800 ng) were injected into the mass spectrometer.

LC-MS/MS measurements. Peptides were loaded on 40 cm reversed phase columns (75 µm inner diameter, packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin [ReproSil-Pur®, Dr. Maisch GmbH]). The column temperature was maintained at 60°C using a column oven. An EASY-nLC 1200 system (ThermoFisher) was directly coupled online with the mass spectrometer (Q Exactive HF-X, ThermoFisher) via a nano-electrospray source, and peptides were separated with a binary buffer system of buffer A (0.1% formic acid (FA) plus 5% DMSO) and buffer B (80% acetonitrile plus 0.1% FA plus 5% DMSO), at a flow rate of 250 nl/min. The mass spectrometer was operated in positive polarity mode with a capillary temperature of 275°C. The samples were acquired with a DIA method established by Bruderer et al.[34]. Briefly, the method consisted of a MS1 scan (m/z= 300-1,650) with an AGC target of 3x10^6 and a maximum injection time of 60 ms (R= 120,000). DIA scans were acquired at R= 30,000, with an AGC target of 3x10^6, 'auto' for injection time and a default charge state of 4. The spectra were recorded in profile mode and the stepped collision energy was 10% at 25%.

High pH reversed-phase fractionation. HepG2 cells were lysed as described in '*Cell culture and MS sample preparation'*. 150 µg of total protein was digested with a trypsin/Lys-C mix (Promega, V5071) overnight at 37°C with a 1:50 enzyme to protein ratio. The digestion was stopped by adding two volumes of 99%

ethylacetate/1% TFA, followed by sonication for 1 min using an ultrasonic probe device (energy output ~40%). The peptides were desalted using 30 mg (8B-S029-TAK) Strata-X-C cartridges (Phenomenex) as follows: a) conditioning with 1 ml of isopropanol; b) conditioning with 1 ml of 80% ACN/5% NH_4OH; c) equilibration with 1 ml of 99% ethylacetate/1% TFA; d) loading of the sample; e) washing with 2 x 1 ml of 99% ethylacetate/1% TFA; f) washing with 1 ml of 0.2% TFA; g) elution with 2 x 1 ml of 80% ACN/5% NH_4OH. The eluates were snap-frozen in liquid nitrogen and lyophilized overnight. The lyophilized peptides were resuspended in 400 µl 0.1 % FA and fractionated using a 3x250 mm xBridge column (Waters) on an ÄKTA HPLC system (GE Healthcare). Fractionation was performed with a flow rate of 0.5 ml/min and with a constant flow of 10% 25 mM ammonium bicarbonate, pH 10. Peptides were separated using a linear gradient of ACN from 7% to 30% over 15 min, followed by a 5-min increase to 55% ACN and a subsequent ramping to 100% ACN. Fractions were collected at 50-sec intervals in 15 ml Falcon tubes to a total of 36 fractions and then pooled to obtain 12 fractions (A1-B1-C1, A2-B2-C2 etc.). All fractions were acidified by addition of FA to a final amount of 0.1% and then lyophilized. Peptides were subsequently resuspended in 100 µl 0.1% TFA and desalted using in-house prepared C18 STAGE tips[43] as follows: a) equilibration with 100 µl isopropanol, b) Equilibration with 100 µl 0.1% TFA, c) loading of the sample, d) washing with 100 µl 0.1% formic acid (FA), e) elution with 30 µl of 80% Acetonitrile/0.1% FA. Peptides were speed-vac dried, resupended in 20 µl 0.1% FA and the concentration estimated on a nanodropTM device (Thermo Fisher Scientific). The samples were then adjusted to 0.4 µg/µl with 0.1% FA, of which 2 µl (800 ng) were injected into the mass spectrometer.

HeLa data with varying gradients

High-pH reversed phase peptide fractionation. 6 µg of HeLa peptides were loaded onto a Waters BEH130 C18 2.1 × 250 mm column in 90 µL of MS loading buffer at a flow rate of 0.5 mL/min using a Dionex Ultimate 3000 HPLC, and column temperature was maintained at 50°C. After loading, a binary gradient of 10% buffer A (2% acetonitrile, 10 mM ammonium formate pH 9) to 40% buffer B (80% acetonitrile, 10 mM ammonium formate pH 9) was formed over 4.4 minutes, followed by a wash-out from 40–100% buffer B over 1 minute, after which the column was held at 100% buffer B for 10 minutes prior to re-equilibration. Fractions were collected over a period of 6.4 minutes from the first peptide elution, with fraction collection each 8 seconds and automatic concatenation into 16 fractions (200 µL fraction volume). Fractions were dried down in a vacuum concentrator (Eppendorf) and resuspended in MS loading buffer (0.3% TFA, 2% acetonitrile).

MS analysis. Peptides were loaded onto a 40 cm column with a 75 µM inner diameter, packed in-house with 1.9 µM C18 ReproSil particles (Dr. Maisch GmbH). Column temperature was maintained at 60°C with a column oven (Sonation GmbH). A Dionex U3000 RSLC nano HPLC system (Thermo Fisher Scientific) was interfaced with a Q Exactive HF X benchtop Orbitrap mass spectrometer (Thermo Fisher Scientific) using a

NanoSpray Flex ion source (Thermo Fisher Scientific). For all samples, peptides were separated with a binary buffer system of 0.1% (v/v) formic acid (buffer A) and 80% (v/v) acetonitrile/0.1% (v/v) formic acid (buffer B) and peptides eluted at a flow rate of 400 nl/min. Gradient ranges and durations were as follows: 5–40% buffer B over 30 minutes (DDA library); 3–19% buffer B over 10 minutes and 19–41% over 5 minutes (15 min DIA gradient); 3–19% buffer B over 20 minutes and 19–41% over 10 minutes (30 min DIA gradient); 3–19% buffer B over 40 minutes and 19–41% over 20 minutes (1 h DIA gradient); 3–19% buffer B over 60 minutes and 19–41% over 30 minutes (1.5 h DIA gradient); 3–19% buffer B over 80 minutes and 19–41% over 40 minutes (2 h DIA gradient). For the DDA library, peptides were analysed with one full scan (350-1,400 m/z, R=60,000 at 200 m/z) with a target of 3e6 ions, followed by up to 20 data-dependent MS/MS scans with HCD (target 1e5 ions, maximum IT 28 ms, isolation width 1.4 m/z, NCE 27%, intensity threshold 3.7e5), detected in the Orbitrap (R=15,000 at 200 m/z). Dynamic exclusion was enabled (15 s). For DIA measurements, peptides were analysed with one full scan (350-1,400 m/z, R=120,000 at 200 m/z) at a target of 3e6 ions, followed by 48 data-independent MS/MS scans spanning 350–975 m/z with HCD (target 3e6 ions, maximum IT 22 ms, isolation width 14 m/z, NCE 25%), detected in the Orbitrap (R=15,000 at 200 m/z).

Three species timsTOF Pro benchmark data

Sample preparation. Human cervix carcinoma cell line HeLa was purchased from the German Resource Centre for Biological Material (Braunschweig, Germany). Cells were cultured in Iscove's Modified Dulbecco Medium (PAN Biotech) supplemented with 10% (v/v) fetal calf serum (FCS; Thermo Fisher Scientific), 1% (v/v) glutamine (Carl Roth) and 1% (v/v) sodium pyruvate (Serva) at 37 °C in a 5% CO_2 environment. A pure culture of the *Saccharomyces cerevisiae bayanus*, strain Lalvin EC-1118 was obtained from the Institut Oenologique de Champagne (Epernay, France). Yeast cells were grown in YPD media as described by Fonslow *et al*.[44]. *Escherichia coli* (TOP10) cells were purchased from Thermo Fisher Scientific and grown in LB liquid medium. After harvesting, cells were lysed adding a urea-based lysis buffer (7 M urea, 2 M thiourea, 5 mM DTT, 2% (w/v) CHAPS). Lysis was promoted by sonication at 4°C for 15 min using a Bioruptor (Diagenode, Liège, Belgium). After cell lysis, protein amounts were determined using the Pierce 660 nm Protein Assay (Thermo Fisher Scientific) according to manufacturer's protocol. Tryptic digestion applying a modified filter-aided sample preparation[45] protocol was performed as described in detail before[46]. To generate the two hybrid proteome samples, tryptic peptides were combined in the following ratios as detailed previously[2,46]. Sample A was composed of 65% w/w human, 30% w/w yeast, and 5% w/w *E. coli* proteins. Sample B was composed of 65% w/w human, 15% w/w yeast, and 20% w/w *E. coli* proteins.

LC MS analysis. Samples were analyzed by LC-MS on a trapped ion mobility spectrometry – quadrupole time of flight mass spectrometer (timsTOF Pro, Bruker Daltonics), which was coupled online to a nanoElute

nanoflow liquid chromatography system (Bruker Daltonics) via a CaptiveSpray nano-electrospray ion source. Peptides (corresponding to 200 ng) were separated on a reversed-phase C18 column (25 cm x 75 µm i.d., 1.6 µm, IonOpticks, Australia). Mobile phase A was water containing 0.1% (v/v) formic acid, and mobile phase B acetonitrile containing 0.1% (v/v) formic acid. Peptides were separated running a gradient of 2–37% mobile phase B over 100 min at a constant flow rate of 400 nL/min. Column temperature was controlled at 50°C. MS analysis of eluting peptides was performed in diaPASEF mode. For diaPASEF, we adapted the instrument firmware to perform data-independent isolation of multiple precursor windows within a single TIMS separation (100 ms). We used a method with two windows in each 100 ms diaPASEF scan. Sixteen of these scans covered the diagonal scan line for doubly charged and triply charged peptides in the *m/z* – ion mobility plane with narrow 25 *m/z* precursor windows resulting in a total cycle time of 1.6 s.

BoxCar DIA HEK data

Cell Culture and MS Sample preparation. HEK293 cells were grown in DMEM supplemented with penicillin, streptomycin and 10% FCS. Cells were washed twice with ice-cold PBS, before scraping in PBS and centrifuged at 300 x g for 6 mins at 4°C. Supernatant was aspirated and the pellet lysed in 2.5 % SDS buffered with 50 mM Tris pH 8.1, and heated to 95C for 5 minutes, prior to probe sonication. The BCA assay was used to quantify the protein content of centrifuge-clarified lysates prior to precipitation with 5 volumes of acetone. Pellets were resuspended in 50 mM Tris pH 8.1 containing 8 M urea, reduced with 1 mM DTT and alkylated with 5 mM IAA prior to initiation of digestion overnight with LysC at an enzyme to protein ratio of 1:100. The digest mixture was diluted 4-fold, and trypsin was added at an enzyme to protein ratio of 1:100 for 6 hours, followed by an additional aliquot of trypsin overnight. Digestion was stopped by acidification to 1% TFA, placed on ice for 5 minutes and centrifuged to remove insoluble material. Peptides were desalted with a mixed-mode SPE cartridges (Strata-XC, Phenomenex), activated with 100% methanol, conditioned with 80% Acetonitrile, 0.1% TFA and equilibrated with 0.2% TFA, which was followed by sample loading, washing with 99.9% isopropanol 0.1% TFA, washing twice with 0.2% TFA, and washing once with 0.1% formic acid, before elution with 60% acetonitrile 0.5% ammonium hydroxide. Eluate was flash frozen and dried by centrifugal evaporation.

Offline peptide fractionation. Peptides were resuspended in buffer A (10 mM ammonium bicarbonate) and injected onto a 4.6 x 250 mm 3.5µm Zorbax 300 Extend-C18 column. Peptides were separated on a non-linear gradient exactly as described (Mertins et al., 2018, Nature protocols), using the following composition of buffer B (10 mM ammonium bicarbonate, 90 % acetonitrile). Peptide fractions were frozen at -80 °C before centrifugal evaporation. Peptides were resuspended in 1% TFA, and concatenated at by combining every 24th fraction for the library, or every 8th fraction for the fractionated BoxCar DIA runs, using fractions 13 – 90.

Concatenated or non-fractionated samples were desalted with SEP-PAK tC18 SPE cartridges (Waters), activated with 100 % methanol, conditioned with 80 % acetonitrile, 0.1% TFA, and equilibrated with 0.2 % TFA. Following sample loading, cartridges were wash with 0.5, 1, and 3 cartridge volumes of 0.2 % TFA, and eluted with 1 volume of 80% acetonitrile, 0.1 % TFA, then frozen before drying in a centrifugal evaporator.

1ug of peptide were loaded onto an Aurora 25cm x 75µm ID, 1.6µm C18 column (Ionopticks) maintained at 40°C. Peptides were separated with an EASY-nLC 1200 system at a flow rate of 300 nl/min using a binary buffer system of 0.1% formic acid (Buffer A) and 80% acetonitrile with 0.1% formic acid (Buffer B), in a two-step gradient from 3-27% B in 105 min and from 27-40 % B in 15 min. All scans were recorded in the Orbitrap of a Fusion Lumos instrument running Tune version 3.3, equipped with a nanoflex ESI source, operated at 1.6 kV, and the RF lens set to 30%. The scan sequence was initiated with MS1 scans from 350-1650 m/z recorded at 120,000 resolution, with an AGC target of 250%, and maximum injection time of 246 ms. The mass range was divided into 24 segments of variable width, with 3 BoxCar scans (multiplexed targeted SIM scan) isolating 8 segments per scan, comprising every third segment. The segments used were identical to those in the MS2 scans, retaining a 1 m/z overlap between boxes in adjacent scans. The normalized AGC target was 200% per segment, with a maximum injection time of 246 ms. BoxCar scans were also recorded at a resolution of 120,000. This was followed by 24 MS2 scans from 200 – 2000 m/z with windows as previously described (Bruderer et al., 2017 MCP). Fragmentation was induced with HCD using stepped collision energy of 22, 27, and 32% for the window center. Each MS2 scan was recorded at a resolution of 30,000, and an AGC target of 1000 % with a maximum injection time of 60 ms.

Data downloads

In addition to the data measured for this publication, we downloaded the following publicly available datasets. The four-species mixture dataset[34] containing *H. sapiens*, *C. elegans*. *S. cerevisiae* and *E. coli* with ratios of 0%, 10%, 20% and 30%, respectively, between replicate groups was downloaded from ProteomeXchange (PXD005573). SCIEX TripleTOF 6600 three species benchmark data[2] was obtained from ProteomeXchange (PXD002952). The HepG2 RNA-seq data is part of the ENCODE dataset[47] and was downloaded from SRA (SRP014320). The HEK RNA-seq data is part of the Cell Atlas dataset[48] and was downloaded from SRA (SRP017465).

Data analysis

In all MaxQuant analyses for generating libraries and for analyzing DIA samples (MaxDIA) version 2.0.0 was used and for all parameters the default values were used unless stated otherwise. Searches were performed

with the following FASTA files from UniProt: UP000005640_9606 (*H. sapiens*), UP000007305_4577 (*Z. mays*), UP000002311_559292 (*S. cerevisiae*), UP000000625_83333 (*E. coli*), UP000001940 (*C. elegans*). Methionine oxidation and protein N-terminal acetylation were used as variable modifications in all searches, as is default in MaxQuant.

Comparing number of proteins between datasets. Proteins are assembled into protein groups for identification to account for the redundancy of protein sequences with regards to the peptide evidence distinguishing them. This works in MaxDIA in exactly the same way as in the standard DDA usage of MaxQuant. These protein groups are dataset dependent and hence comparisons between two protein groups tables, for instance in Venn diagrams, or between a protein groups table and RNA-seq data are nontrivial. Here, we follow the route of mapping all protein identifiers in a protein group to Entrez gene identifiers[49]. In the vast majority of cases, protein groups map to single gene identifiers. For cases, in which they map to more than one, both gene identifiers are taken into the set. For counting protein group identifications, we always remove protein groups that are flagged as 'reverse' or 'only identified by site'. For human datasets, we removed protein groups denoted as 'potential contaminant' only if they are of non-human origin and kept human proteins, which consist mostly of human keratins. For the dataset containing bovine plasma the proteins in the standard MaxQuant contaminant list of bovine origin were not removed.

FDR curves. For estimating external FDR, we used a combination of human and maize libraries from reference[34] or of human and maize predicted libraries in discovery mode on the human HepG2 DIA samples. For analyzing library-to-DIA-sample matches and peptide identifications in Fig. 4, we do not apply a protein level FDR and scan through the library-to-DIA-sample FDR. It is crucial to take this approach, in particular when comparing numbers of identifications with other software, since when applying protein-level FDR in MaxQuant, peptides which are not mapping to a protein identified at the specified protein FDR are discarded, unlike in most other software packages. For obtaining the protein-level FDR curves in Fig.4 we applied a library-to-DIA-sample match FDR of 1%. Peptides that are shared between human and maize proteins were discarded.

RNA-seq data analysis. Raw reads were filtered using trimmomatic[50] (version 0.36) using default parameters for paired-end data. Filtered reads were mapped to the human reference genome GRCh38 (Ensemble release 100) using STAR[51] aligner (version 2.5.3a). Further processing – sorting, converting from SAM to BAM format and indexing – was done using SAMtools[52] (version 1.6). Gene expression quantification (RPKM) for protein-coding genes was performed in Perseus[53] (version 1.6.14.0).

Spectronaut analysis. Raw MS data were processed using Spectronaut version 13.10.191212 using default settings, using a spectral library generated by searching using MaxQuant version 1.6.10.43.

Software development, requirements, availability and usage

MaxDIA has been developed in conjunction with MaxQuant in C#, runs on Windows and Linux operating systems and requires .NET Core 2.1. In addition, .NET Framework 4.7.2 has to be installed on Windows. The graphical user interface version is currently restricted to Windows. A platform-neutral command line version is available. MaxQuant is efficiently running in parallel on arbitrarily many CPUs on single-node platforms. Having 4Gb of memory per parallel running thread is recommended. Disk space should be at least twice the space that is used by the raw data. MaxQuant is freeware and the code is partially open and available at https://github.com/JurgenCox/compbio-base. MaxQuant including MaxDIA can be downloaded from https://www.maxquant.org/. MaxDIA is included in the standard MaxQuant release from version 2.0.0 onward. (MaxQuant 2.0.0 is included in the PRIDE submission for the reviewers.) How to use MaxDIA in library or discovery mode is described in the accompanying Supplementary Notes document. It also contains a list of all user-definable parameters with a description of their meaning.

PRIDE support

We support complete submissions to the PRoteomics IDEntifications (PRIDE) database[28] for the DIA identification results. We extended the mzTab format[29] to cover DIA data sets. For this purpose, new controlled vocabulary terms were introduced along with additional external reference files. These external reference files contain DIA library matches with mass, intensity and annotation information in a spectral library format (msp-format). MaxQuant will generate a new output folder called ‘combined\msp’ into which these results are written. A user must provide this folder in addition to raw and mzTab files during submission to PRIDE. More details on a complete PRIDE submission are provided in the Supplementary Notes. This is the first instance of complete PRIDE submissions for DIA data sets.

Data availability

The MS proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifiers PXD022582 (DDA data, login/password for review: reviewer_pxd022582@ebi.ac.uk / oKBHzhLq) and PXD022589 (DIA data, containing also MaxQuant version 2.0.0, login/password for review: reviewer_pxd022589@ebi.ac.uk / yui5MuP8).

ACKNOWLEDGEMENTS

We thank Roland Bruderer for providing data, Georgina D. Barnabas for testing, and all members of the Computational Systems Biochemistry Research Group for helpful discussions. This project was partially funded by the German Ministry for Science and Education (BMBF) funding action MSCoreSys, reference number FKZ 031L0214D and 031L0217A and the Deutsche Forschungsgemeinschaft (SFB1292 Z02, to S.T.). S.Y. is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 792536. Y.P.R. is supported by BBSRC, grant number BB/P024599/1. D.I. is a Chan Zuckerberg Biohub Fellow.

AUTHOR CONTRIBUTIONS

P.S., H.H., F.S.S., N.P., C.W., Ş.Y., J.D.R. and J.C designed and developed the code. D.I., S.T., N.N. S.J.H. and J.C. conceptualized the wet-laboratory experiments and mass spectrometric measurements. D.I., F.M., M.S., U.O., U.D., S.K.S., S.T. and S.J.H. carried out the wet-laboratory experiments and mass spectrometric measurements. H.H., Ş.Y and Y.P.R. designed and developed the PRIDE support, P.S., H.H., F.S.S. N.P. C.W. S.J.H. and J.C. analyzed the data, M.S., D.I., U.D., N.N., S.J.H. and J.C. wrote online Methods sections, J.C. wrote the manuscript and directed the project.

COMPETING FINANCIAL INTERESTS

The authors state that they have potential conflicts of interest regarding this work: M.S. and U.O. are employees of Evotec, N.N. and S.K.S. are employees of Bruker and J.D.R. is employee of Bosch.

REFERENCES

1. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35–35 (2014).
2. Navarro, P. *et al*. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3685
3. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
4. Azvolinsky, A., DeFrancesco, L., Waltz, E. & Webb, S. 20 years of Nature Biotechnology research tools. *Nat. Biotechnol.* **34**, 256–261 (2016).
5. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).
6. Sinitcyn, P. *et al.* MaxQuant goes Linux. *Nat. Methods* **15**, 401 (2018).
7. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* (2014). doi:10.1038/nbt.2841
8. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

9. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* (2015). doi:10.1074/mcp.M114.044305
10. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* (2020). doi:10.1038/s41592-019-0638-x
11. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513–2526 (2014).
12. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Meth* **14**, 921–927 (2017).
13. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
14. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
15. Tiwary, S. *et al.* High quality MS/MS spectrum prediction for data-dependent and -independent acquisition data analysis. *Nat Methods* (2019).
16. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* (2019). doi:10.1038/s41592-019-0426-7
17. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* (2020). doi:10.1038/s41467-019-13866-z
18. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* (2020). doi:10.1038/s41467-020-15346-1
19. Lou, R. *et al.* Hybrid Spectral Library Combining DIA-MS Data and a Targeted Virtual Library Substantially Deepens the Proteome Coverage. *iScience* (2020). doi:10.1016/j.isci.2020.100903
20. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* (2019). doi:10.1038/s41592-018-0260-3
21. Graves, A. *et al.* A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 855–868 (2009).
22. Chen, T. & Guestrin, C. XGBoost : Reliable Large-scale Tree Boosting System. *arXiv* (2016). doi:10.1145/2939672.2939785
23. Prianichnikov, N. *et al.* MaxQuant software for ion mobility enhanced shotgun proteomics. *bioRxiv* (2019). doi:10.1101/651760
24. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* (2018). doi:10.1038/s41592-018-0003-5
25. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* (2011). doi:10.1007/s12127-011-0067-8
26. Silveira, J. A., Ridgeway, M. E. & Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* (2014). doi:10.1021/ac501261h
27. Meier, F. *et al.* Online parallel accumulation – serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer.
28. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1106
29. Griss, J. *et al.* The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Mol. Cell. Proteomics* **13**, 2765–2775 (2014).
30. Martens, L. *et al.* mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **10**, R110 000133 (2011).
31. Cox, J., Michalski, A. & Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom* **22**, 1373–1380 (2011).
32. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

33. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
34. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics* mcp.RA117.000314 (2017). doi:10.1074/mcp.RA117.000314
35. Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial . *Mol. Syst. Biol.* (2018). doi:10.15252/msb.20178126
36. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628 (2008).
37. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
38. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* (2020). doi:10.1038/s41576-020-0258-4
39. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
40. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* (2011). doi:10.1038/msb.2011.28
41. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* (2020). doi:10.1038/s41467-020-14609-1
42. Müller, F., Kolbowski, L., Bernhardt, O. M., Reiter, L. & Rappsilber, J. Data-independent acquisition improves quantitative cross-linking mass spectrometry. *Mol. Cell. Proteomics* (2019). doi:10.1074/mcp.TIR118.001276
43. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663–670 (2003).
44. Fonslow, B. R. *et al.* Digestion and depletion of abundant proteins improves proteomic coverage. *Nat. Methods* (2013). doi:10.1038/nmeth.2250
45. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* (2009). doi:10.1038/nmeth.1322
46. Distler, U., Kuharev, J., Navarro, P. & Tenzer, S. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. *Nat. Protoc.* (2016). doi:10.1038/nprot.2016.042
47. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
48. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* (2017). doi:10.1126/science.aal3321
49. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkq1237
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu170
51. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp352
53. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–40 (2016).

FIGURE LEGENDS

Fig. 1: Overview of the MaxDIA workflow. MaxDIA can be operated in library and discovery mode. Many concepts and algorithms, for instance for protein quantification, are re-used from the conventional MaxQuant workflow for DDA data and have been further developed for DIA. This results in an end-to-end DIA software that contains many established MaxQuant concepts, like label-free quantification with MaxLFQ or iBAQ quantification.

Fig. 2: 3D/4D feature detection of precursors and fragments. a, Visualization of precursor and fragments of a peptide measured on an Orbitrap. The raw data can be visualized together with the peak detection results as heat maps and 3D models for precursor and fragment data in the graphical user interface of MaxQuant. **b,** Two peptides with nearly equal mass, both with charge 2 and having very similar retention times are resolved by ion mobility on a timsTOF Pro mass spectrometer. A heat map visualizes intensities as a function of retention time and collision cross section for the precursor isotope patterns. The two respective MS/MS spectra of fragments assigned to the precursors are shown.

Fig. 3: Performance evaluation. 27 technical replicates of HepG2 cell lysate were analyzed on an Orbitrap mass spectrometer (see online Methods). **a,** Number of identified protein groups with 1% FDR on protein and peptide level, and number of peptides at 1% library-to-DIA-sample FDR obtained with MaxDIA and Spectronaut. **b,** Histograms of peptide lengths identified with MaxDIA (blue) and in Spectronaut (red). **c,** Number of proteins with at most x out of 27 valid values for Spectronaut (red), MaxDIA with MaxLFQ minimum ratio count = 1 (blue, dashed) and = 2 (blue, solid). Multiple curves for the two MaxQuant seies of curves correspond to seven different choices for the transfer q-value (0.01, 0.02, 0.05, 0.1, 0.2 and 0.5). **d,** Histograms of coefficients of variation for analyses with default settings in MaxDIA (blue) and in Spectronaut (red). **e,** Log-log scatter plot of LFQ intensities between two representative replicates obtained with MaxQuant. The two replicates were chosen to have the median Pearson correlation of all pair-wise replicate comparisons. **f,** Same as in panel f for Spectronaut intensities. Similarly, the two replicates were chosen to represent the median Pearson correlation coefficient of all pair-wise comparisons. **g,** Heat map with all pair-wise Pearson correlations between the 27 replicates for MaxDIA (upper triangle) and Spectronaut (lower traingle). The two values corresponding to the comparisons in panels e and f are marked with red squares. **h,** Log-log scatterplot of iBAQ protein intensities from MaxDIA against Spectronaut protein intsnsities. **i,** Log-log scatterplot of MaxDIA iBAQ values averaged over the replicates against RPKM values from RNA-seq data. **j,** Same as panel i with protein intensities from Spectronaut.

Fig. 4: Internal and external FDR. a, Number of identifications (blue: matches, green: peptides, red: protein groups) as a function of estimated FDR. The FDR is once estimated with the 'internal' target-decoy method implemented in MaxQuant (solid lines) and once with the 'external' method using mixing maize and human samples for generating the library and using only human sample in the DIA runs (dashed lines). **b,** Same as in panel a but using *in-silico* predicted libraries generated using DeepMass:Prism[15] **c,** Same as panel a but using the raw score instead of the machine learning-derived score. **d,** Same as panel b but using the raw score instead of the machine learning-derived score.

Fig. 5: MaxLFQ for DIA. a, Stacked inter-quartile rages of protein ratio distributions in the small-ratio four-species dataset from Bruderer et al.[34] using different versions of MaxLFQ for DIA and compared to the results from this publication. **b,** Quantification of a three-species benchmark mixture measured on a SCIEX TripleTOF 6600 instrument mixing proteomes from three species in defined ratio[2] with MaxLFQ for DIA. The accompanying DDA library was used. **c,** Same as b, but analyzed with MaxDIA in discovery mode. **d,** Quantification of a three-species benchmark mixture measured on a Bruker timsTOF Pro instrument mixing proteomes from three species in defined ratio using a DDA library. **e,** Same as d, but analyzed in discovery mode.

Fig. 6: BoxCar and fractionated DIA. a, Schedule of libraries and DIA samples. Three different library approaches, single-shot, deep fractionated and discovery mode library were compared to single-shot deep fractionated DIA samples. **b,** MaxLFQ quantification between three replicates of fractionated BoxCar DIA samples analyzed in discovery DIA mode. All pair-wise Pearson correlations are above 0.99. **c,** Venn diagram-like comparison represented as bar plot between RNA-seq data of HEK cells and three different library methods applied to the fractionated DIA samples. All data has been mapped to gene identifiers **d,** Histogram of protein identifications mapped to gene identifiers sorted into bins according to log2 RPKM values of the RNA-seq data.

Figure 1

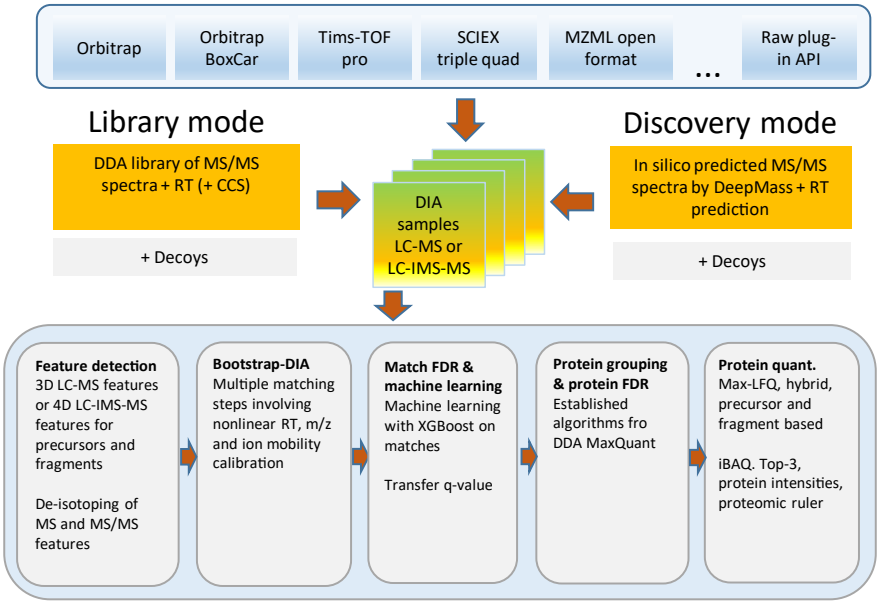

Orbitrap
Orbitrap BoxCar
Tims-TOF pro
SCIEX triple quad
MZML open format
...
Raw plug-in API
Library mode
DDA library of MS/MS spectra + RT (+ CCS)
+ Decoys
DIA samples LC-MS or LC-IMS-MS
Discovery mode
In silico predicted MS/MS spectra by DeepMass + RT prediction
+ Decoys
Feature detection
3D LC-MS features or 4D LC-IMS-MS features for precursors and fragments
De-isotoping of MS and MS/MS features
Bootstrap-DIA
Multiple matching steps involving nonlinear RT, m/z and ion mobility calibration
Match FDR & machine learning
Machine learning with XGBoost on matches
Transfer q-value
Protein grouping & protein FDR
Established algorithms fro DDA MaxQuant
Protein quant.
Max-LFQ, hybrid, precursor and fragment based
iBAQ. Top-3, protein intensities, proteomic ruler

Figure 2

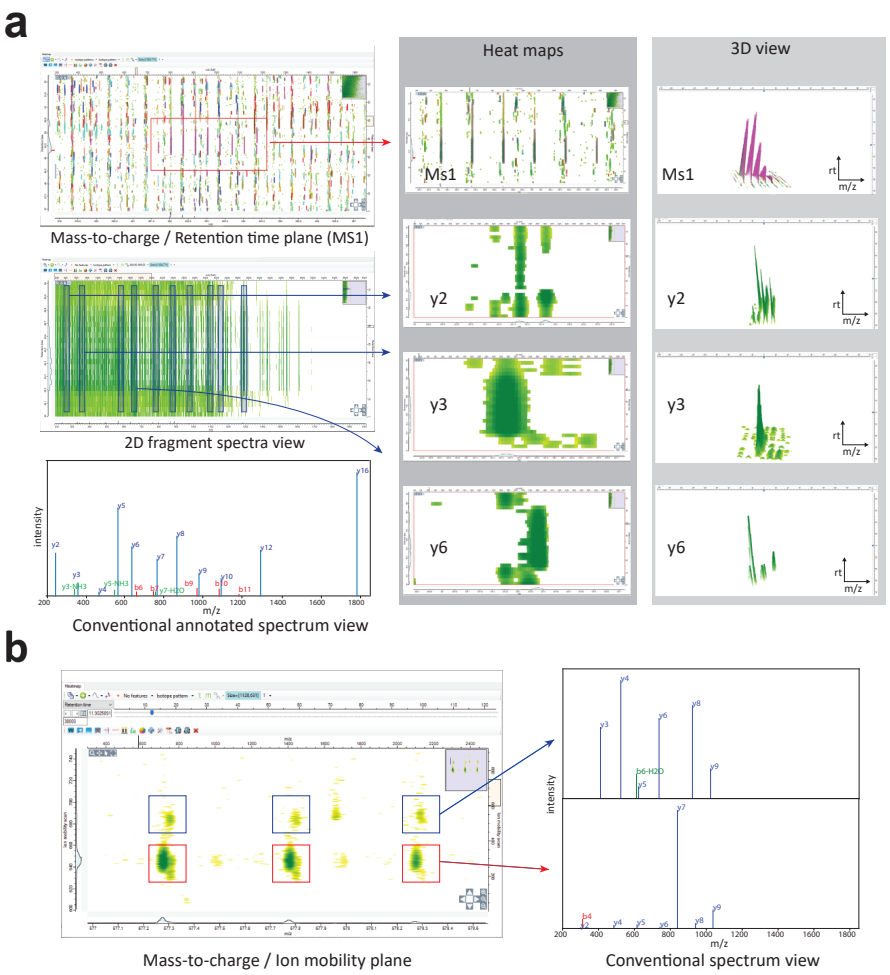

a
Heat maps
3D view
Mass-to-charge / Retention time plane (MS1)
2D fragment spectra view
Conventional annotated spectrum view
Ms1
y2
y3
y6
intensity
m/z
b
Mass-to-charge / Ion mobility plane
Conventional spectrum view

Figure 3

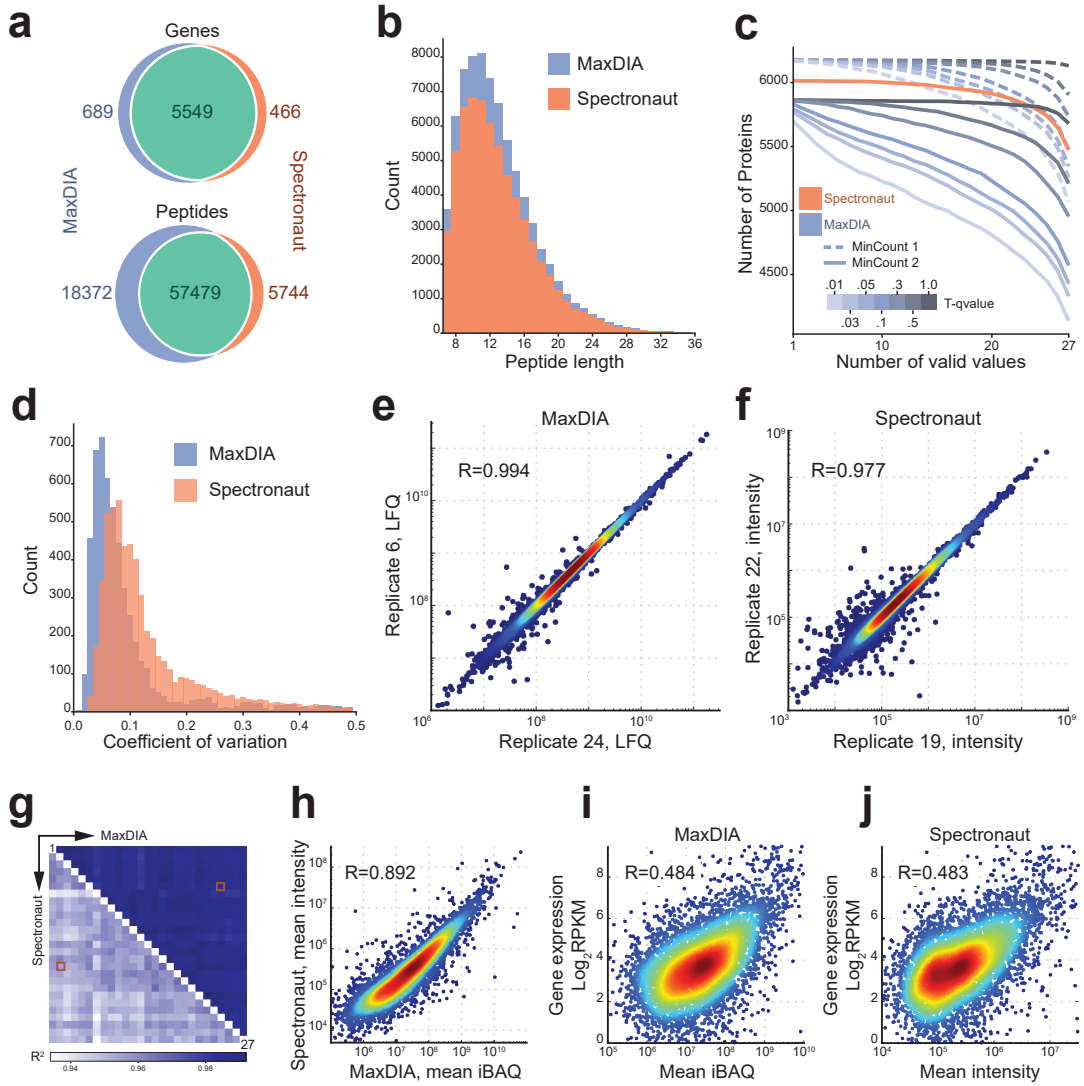

a
Genes
689
5549
466
MaxDIA
Spectronaut
Peptides
18372
57479
5744
b
MaxDIA
Spectronaut
Count
Peptide length
c
Number of Proteins
Spectronaut
MaxDIA
MinCount 1
MinCount 2
.01 .05 .3 1.0
.03 .1 .5
T-qvalue
Number of valid values
d
MaxDIA
Spectronaut
Count
Coefficient of variation
e
MaxDIA
R=0.994
Replicate 6, LFQ
Replicate 24, LFQ
f
Spectronaut
R=0.977
Replicate 22, intensity
Replicate 19, intensity
g
MaxDIA
Spectronaut
R²
h
R=0.892
Spectronaut, mean intensity
MaxDIA, mean iBAQ
i
MaxDIA
R=0.484
Gene expression Log2 RPKM
Mean iBAQ
j
Spectronaut
R=0.483
Gene expression Log2 RPKM
Mean intensity

Figure 4

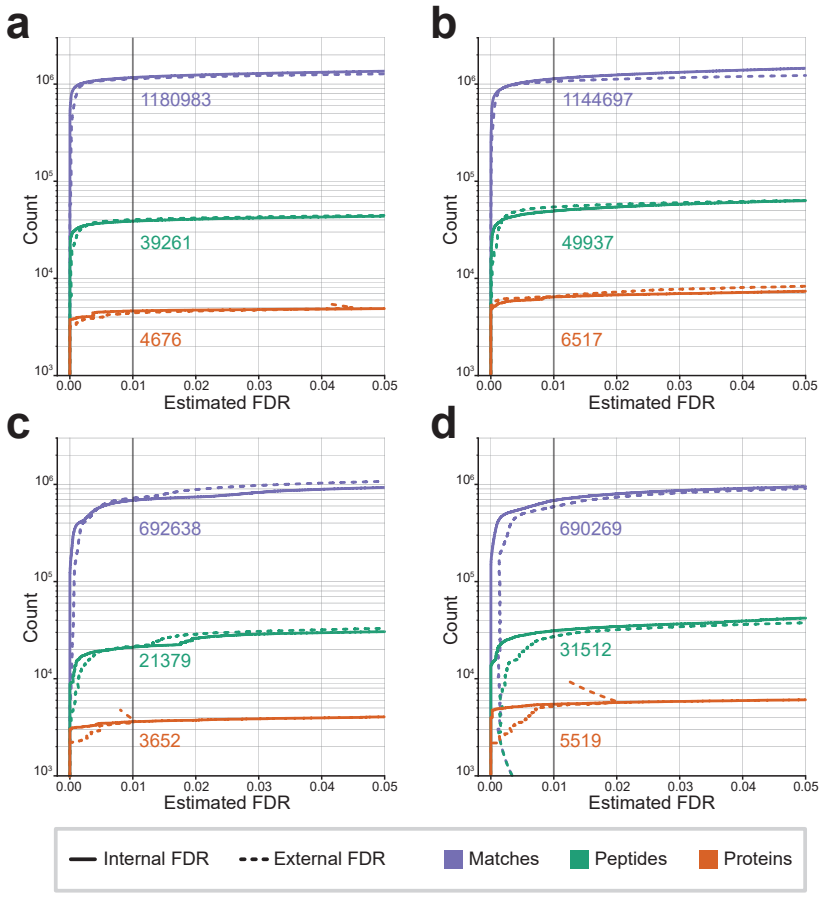

a
b
c
d
Count
Estimated FDR
1180983
39261
4676
1144697
49937
6517
692638
21379
3652
690269
31512
5519
Internal FDR
External FDR
Matches
Peptides
Proteins

Figure 5

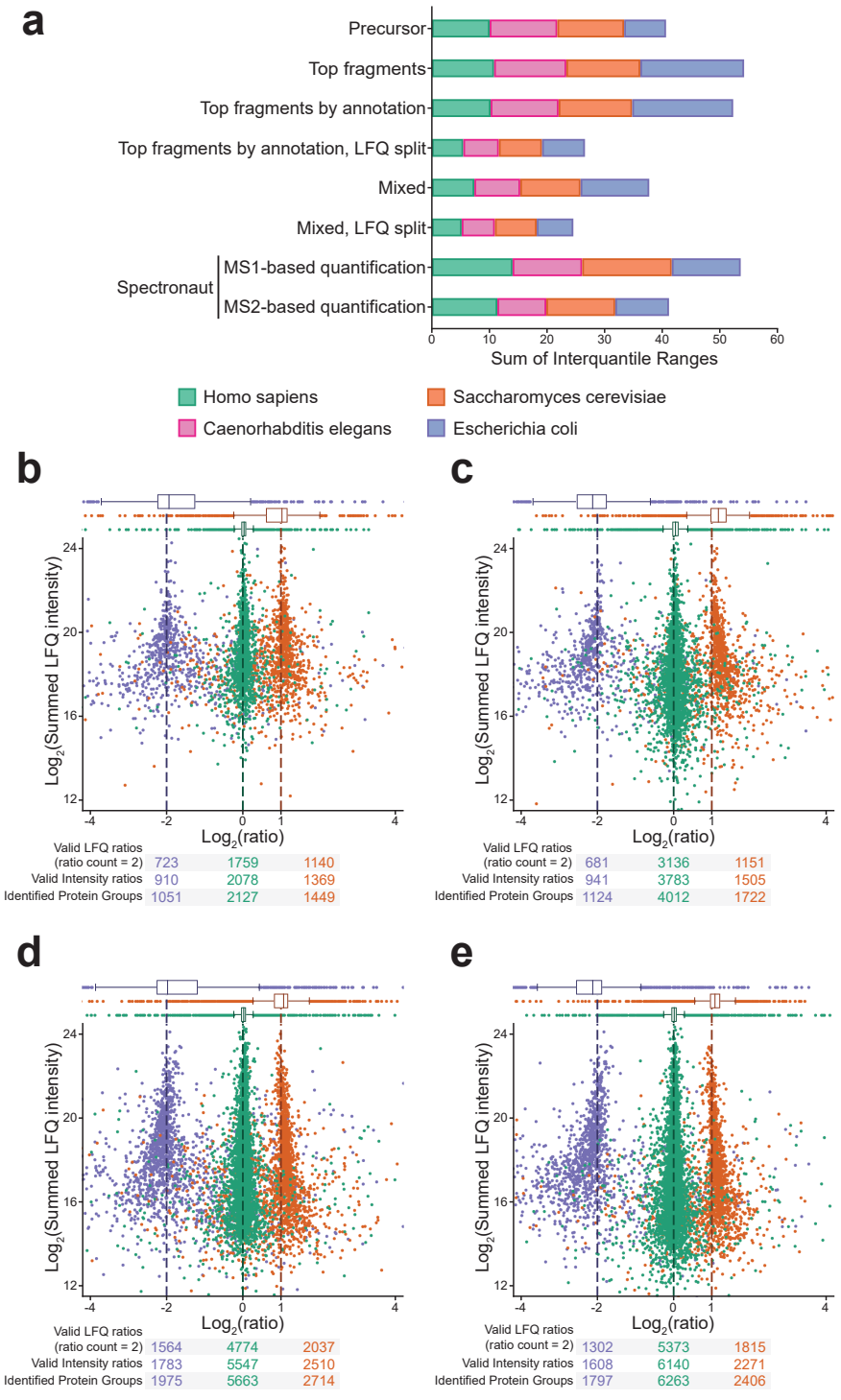
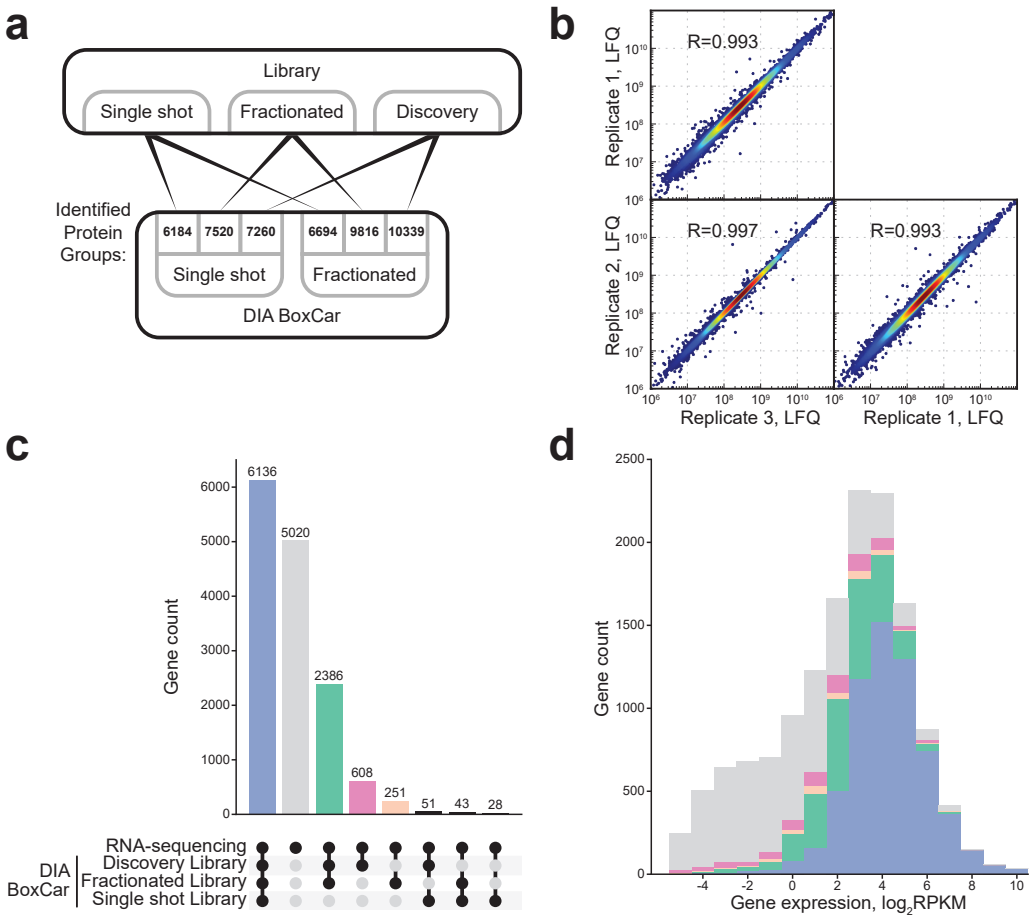

a
Precursor
Top fragments
Top fragments by annotation
Top fragments by annotation, LFQ split
Mixed
Mixed, LFQ split
Spectronaut
MS1-based quantification
MS2-based quantification
0 10 20 30 40 50 60
Sum of Interquantile Ranges
Homo sapiens
Caenorhabditis elegans
Saccharomyces cerevisiae
Escherichia coli
b
c
d
e
Log2(Summed LFQ intensity)
Log2(ratio)
Valid LFQ ratios (ratio count = 2) 723 1759 1140
Valid Intensity ratios 910 2078 1369
Identified Protein Groups 1051 2127 1449
Valid LFQ ratios (ratio count = 2) 681 3136 1151
Valid Intensity ratios 941 3783 1505
Identified Protein Groups 1124 4012 1722
Valid LFQ ratios (ratio count = 2) 1564 4774 2037
Valid Intensity ratios 1783 5547 2510
Identified Protein Groups 1975 5663 2714
Valid LFQ ratios (ratio count = 2) 1302 5373 1815
Valid Intensity ratios 1608 6140 2271
Identified Protein Groups 1797 6263 2406

Figure 6


a
Library
Single shot
Fractionated
Discovery
Identified Protein Groups:
6184
7520
7260
6694
9816
10339
Single shot
Fractionated
DIA BoxCar
b
R=0.993
R=0.997
R=0.993
Replicate 1, LFQ
Replicate 2, LFQ
Replicate 3, LFQ
Replicate 1, LFQ
c
Gene count
6136
5020
2386
608
251
51
43
28
RNA-sequencing
DIA BoxCar
Discovery Library
Fractionated Library
Single shot Library
d
Gene count
Gene expression, log2RPKM

2.1.2 MaxQuant goes Linux

MaxQuant has been successfully used to analyze proteomic data with a broad distribution of experimental designs for more than a decade[3, 150]. The original version of MaxQuant was restricted to Microsoft Windows Operation System, which was also dictated by vendor-provided raw data access libraries. Taking into account that Linux is the most common choice as a high-performance computing environment, this restriction becomes problematic for large scale proteomics projects. This manuscript represents a joint effort of our laboratory to restructure the MaxQuant codebase and make it truly cross-platform[157].

I contributed to this manuscript by developing the codebase and conducting benchmark runs to check the reproducibility of results and performance.

Pavel Sinitcyn, Shivani Tiwary, Jan Rudolph, Petra Gutenbrunner, Christoph Wichmann, Şule Yılmaz, Hamid Hamzeiy, Favio Salinas, Jürgen Cox
MaxQuant goes Linux
(2018) *Nature methods*
DOI: 10.1038/s41592-018-0018-y

correspondence

MaxQuant goes Linux

To the Editor: We report a Linux version of MaxQuant[1] (http://www.biochem.mpg.de/5111795/maxquant), our popular software platform for the analysis of shotgun proteomics data.

One of our main intentions in developing MaxQuant was to 'take the pain out of' quantifying large collections of protein profiles[2]. However, unlike, for instance, the Trans-Proteomic Pipeline[3], the original version of MaxQuant could be run only on Microsoft Windows, and thus its use was restricted in high-performance computing environments, which very rarely use Windows as an operating system. When we began developing MaxQuant, Windows was the only operating system supported by vendor-provided raw data access libraries. Therefore, we wrote MaxQuant in the C# programming language on top of the Windows-only .NET framework. Windows support for cloud platforms is more expensive, and the operating system is harder to use and less scalable compared with Linux.

We recently carried out a major restructuring of the MaxQuant codebase, and we made it compatible with Mono (https://www.mono-project.com/), an alternative cross-platform implementation of the .NET framework. Furthermore, we now provide an entry point to MaxQuant from the command line without the need to start its graphical user interface, which allows execution from scripts or other processing tools. Meanwhile, Thermo Fisher Scientific has released its platform-independent and Mono-compatible implementation of its raw data access library (http://planetorbitrap.com/rawfilereader), and hopefully more vendors will follow soon. Together, this leads to a situation in which large-scale computing of proteomics data with MaxQuant becomes feasible on all common platforms.

When we parallelized the MaxQuant workflow over only a few central processing unit (CPU) cores, we hardly noticed a difference in performance between Linux and Windows (Fig. 1). However, in benchmarking of a highly parallelized MaxQuant run on 120 logical cores, we observed that the Linux version showed highly superior parallelization performance, with speed 64% faster than that observed under a Windows server operating system using identical hardware. MaxQuant uses operating system processes, rather than the intrinsic multi-threading mechanism of C#, to realize parallel execution, and it manages the load-balancing of an arbitrarily large set of raw data files over a specified number of processors by itself. We hypothesize that this allows Linux to optimize parallel execution to the high extent that we observed. A larger benchmark study is under way, in which we will investigate the dependence of the increased speed on hardware such as, for instance, the type of CPU and storage systems.

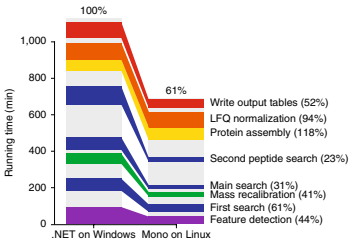


Fig. 1 | Benchmarking MaxQuant on Linux and Windows. We analyzed 300 LC-MS runs with MaxQuant using 120 logical cores in parallel, once with Ubuntu Linux (version 16.04.3) and once with Windows server 2012 R2 as the operating system. We used identical hardware in both cases: four Intel Xeon E7-4870 CPUs and 256 GB of DDR3 RAM. The total running times are shown, and several long-running sub-workflows are highlighted. Percentages indicate the amount of time needed to complete the relevant process in Linux as a percentage of the total time required for the same process in Windows.

MaxQuant has already been adapted in several forms for cloud and high-performance computing applications, as described, for instance, by Judson et al.[4] and on the Chorus platform (https://chorusproject.org). We expect that the number of applications will increase with our Linux-compatible MaxQuant version. We envision that proteomics core facilities, for instance, will benefit from the combination of command-line access and Linux compatibility, which enables standardized high-throughput data analysis. The MaxQuant code base is identical for Windows and for Linux; thus there is only a single distributable running on both operating systems, which can be downloaded from http://www.maxquant.org (version 1.6.1.0). MaxQuant is freeware, and contributions to new functionality are collaboration-based. The code of open source parts is available at https://github.com/JurgenCox/compbio-base. ❐

Pavel Sinitcyn, Shivani Tiwary, Jan Rudolph, Petra Gutenbrunner, Christoph Wichmann, Şule Yılmaz, Hamid Hamzeiy, Favio Salinas and Jürgen Cox*

Computational Systems Biochemistry, Max Planck Institute for Biochemistry, Martinsried, Germany.
**e-mail: cox@biochem.mpg.de*

Published online: 31 May 2018
https://doi.org/10.1038/s41592-018-0018-y

References
1. Cox, J. & Mann, M. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
2. Azvolinsky, A., DeFrancesco, L., Waltz, E. & Webb, S. *Nat. Biotechnol.* **34**, 256–261 (2016).
3. Deutsch, E. W. et al. *Proteomics Clin. Appl.* **9**, 745–754 (2015).
4. Judson, B., McGrath, G., Peuchen, E. H., Champion, M. M. & Brenner, P. In *Proc. 8th Workshop on Scientific Cloud Computing* (eds. Chard, K. et al.) 17–24 (ACM, New York, 2017).

Acknowledgements
This project has received funding from the European Union's Horizon 2020 research and innovation program (grant agreement no. 686547 to J.C., J.R. and S.Y.) and from the FP7 (grant GA ERC-2012-SyG_318987–ToPAG to S.T. and F.S.).

Author contributions
P.S., S.T., J.R., P.G., C.W., S.Y., H.H., F.S. and J.C. developed the software. P.S. conducted the performance analysis. J.C. wrote the manuscript.

Competing interests
The authors declare no competing interests.

2.1.3 Visualization of LC-MS/MS proteomics data in MaxQuant

Modern shotgun proteomics data are highly multidimensional and complex, compare to NGS data. A typical mass-spectrometry experiment consists of a minimum of three dimensions - retention time, mass-to-charge, and intensity. Additionally, there are spectra of peptide fragments on MS^2 that were acquired from specific mass and time windows. Because of such complexity, it is often crucial to verify the technical quality of chromatography, a mass analyzer, and software performance by directly visualizing data.

Additionally, to the analysis of the raw data, MaxQuant allows visualizing raw files as well as identification and quantification results from MaxQuant[158].

My contribution to this paper was to develop and test the visualization features and to edit the manuscript.

Stefka Tyanova, Tikira Temu, Arthur Carlson, **Pavel Sinitcyn**, Matthias Mann, Juergen Cox
Visualization of LC-MS/MS proteomics data in MaxQuant
(2015) *Proteomics*
DOI: 10.1002/pmic.201400449

Proteomics 2015, *15*, 1453–1456 DOI 10.1002/pmic.201400449 1453

TECHNICAL BRIEF

Visualization of LC-MS/MS proteomics data in MaxQuant

Stefka Tyanova[1], Tikira Temu[1], Arthur Carlson[1], Pavel Sinitcyn[1], Matthias Mann[2] and Juergen Cox[1]

[1] Max-Planck-Institute of Biochemistry, Computational Systems Biochemistry, Martinsried, Germany
[2] Max-Planck-Institute of Biochemistry, Proteomics and Signal Transduction, Martinsried, Germany

Modern software platforms enable the analysis of shotgun proteomics data in an automated fashion resulting in high quality identification and quantification results. Additional understanding of the underlying data can be gained with the help of advanced visualization tools that allow for easy navigation through large LC-MS/MS datasets potentially consisting of terabytes of raw data. The updated MaxQuant version has a map navigation component that steers the users through mass and retention time-dependent mass spectrometric signals. It can be used to monitor a peptide feature used in label-free quantification over many LC-MS runs and visualize it with advanced 3D graphic models. An expert annotation system aids the interpretation of the MS/MS spectra used for the identification of these peptide features.

Received: September 19, 2014
Revised: December 12, 2014
Accepted: January 28, 2015

Keywords:
Bioinformatics / Mass spectrometry / LC-MS/MS / Visualization

MaxQuant is a widely used software platform for the analysis of shotgun proteomics data [1, 2]. From version 1.5 on, MaxQuant has been distributed with a unified data exploration component, termed the Viewer, which enables integrated visual inspection of the mass spectrometric raw data together with results from the identification and quantification pipeline. It provides navigation through arbitrarily large datasets by efficient indexing of the underlying data structures and on-demand loading of the required raw data. MaxQuant, including the integrated Viewer described here, is programmed in C# using the .NET framework 4.5. MaxQuant uses the Windows Ribbon Framework to achieve easy navigation and quick access to the graphical user interface. It is freely available and can be downloaded from www.maxquant.org. MaxQuant can read raw data in native vendor formats from Thermo Fisher Scientific, Bruker Daltonics, and AB/Sciex as well as the open mzML format. Concise information on "Getting started," including software requirements, trouble shooting, and a test dataset, is available online (http://141.61.102.17/maxquant_doku/doku.php?id=maxquant:viewer).

Correspondence: Dr. Juergen Cox, Max-Planck-Institute of Biochemistry, Computational Systems Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
E-mail: cox@biochem.mpg.de
Fax: +49-89-8578-2219

Figure 1A displays the fully redesigned graphical user interface of MaxQuant. Organization of main components in tabs allows for easy navigation in the main level control of MaxQuant. Here, we are interested in the "Viewer" tab. Below the tab selector is the command ribbon, which hosts multiple buttons and other control elements acting on the main visual components beneath it. The main display is subdivided into four parts. The upper left component is the map view, which displays the mass spectrometric color-coded intensities, typically at the MS1 level, in the m/z-retention time plane. Peak boundaries are displayed in different coloring schemes, which indicate the grouping of peaks into isotope patterns or of isotope patterns into labeling pairs or triplets. For instance, the peak shapes in Fig. 1A are colored according to the isotope pattern to which they belong. The blue indicator rectangle encloses a peptide whose 3D peak boundaries are shown in blue. There are eight isotopic peaks found for this peptide in this label-free sample. The flat red rectangle indicates the region in the m/z-retention time plane in which ions were collected for the fragmentation spectrum that was recorded in order to identify the peptide colored in blue. By visual inspection one can conclude that no major cofragmentation of other peptide species is expected in this case. The proportion of cofragmented ions is also

Colour Online: See the article online to view Figs. 1 and 2 in colour.

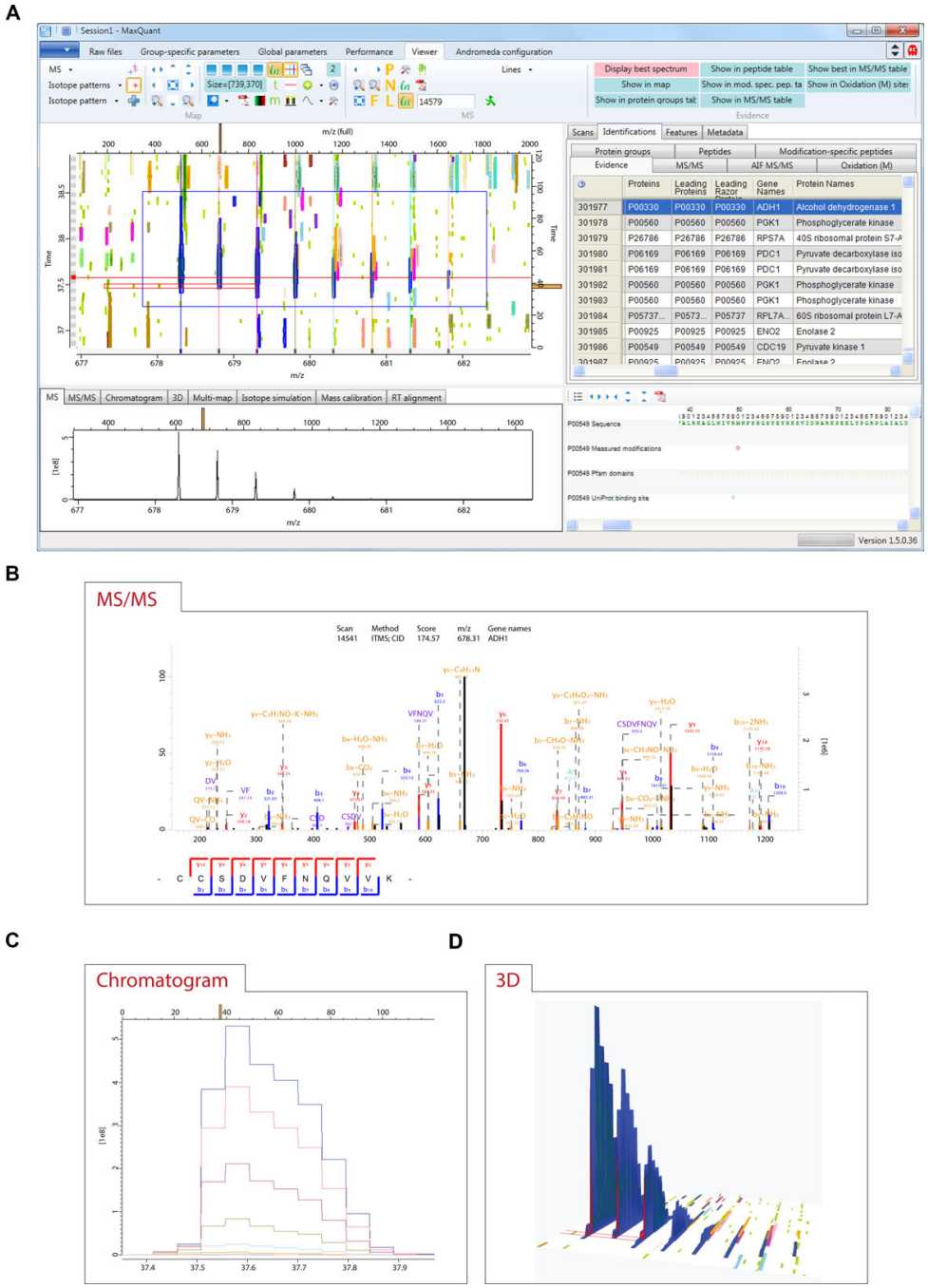


Figure 1. Overview of the MaxQuant Viewer tab. (A) Upon selection of a feature of interest from the evidence table, the MS intensities of the feature (color coded by isotope pattern in blue) are displayed in the *m/z*-retention time map as indicated by the blue rectangle. Underneath the map the MS mode of the feature view is shown. Various modes of the feature view are represented: (B) MS/MS of the selected feature with advanced annotation enabled and display of the peptide sequence; (C) chromatogram with mass traces enabled; (D) 3D view of the isotope peaks.

Proteomics 2015, *15*, 1453–1456

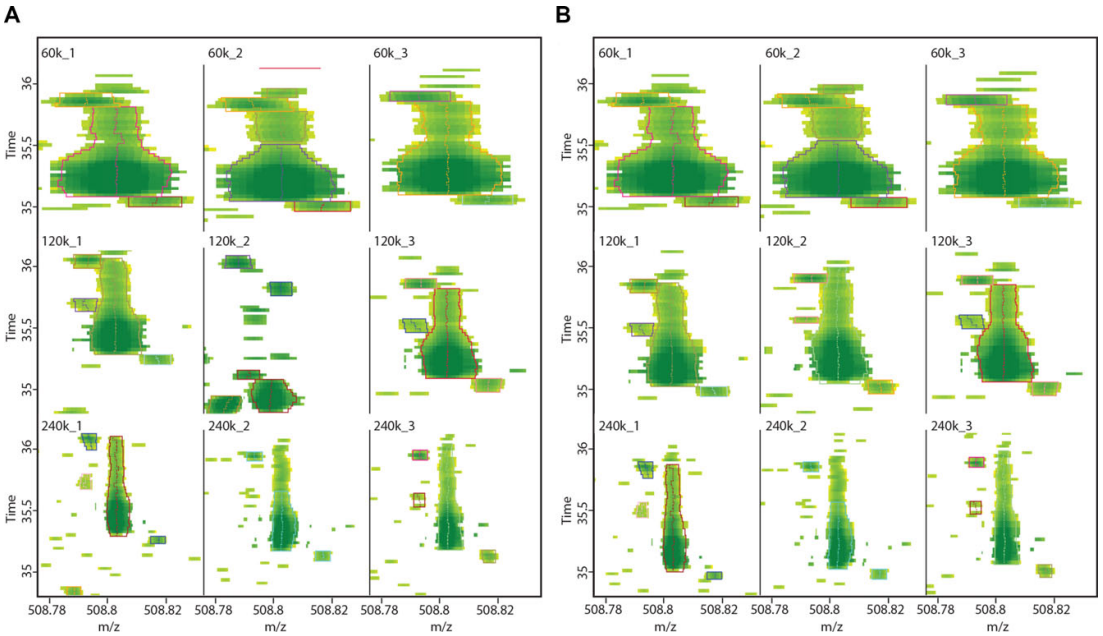


Figure 2. Multi-map view. A view of the same feature in different experiments: three mass resolutions with three replicates for each resolution are shown: (A) without retention time alignment and (B) with alignment.

automatically quantified in MaxQuant for every fragmentation event [3]. The map view can be applied as well to MS/MS data that is continuous in time, as is produced, for instance, during data independent acquisition. For "all ion fragmentation" data [4] MaxQuant can determine 3D peaks as well, which will be displayed in the Viewer in the same way as MS1 level peaks. The map view is fast and responsive because it reads the minimal amount of data from the raw file that is needed for determining the colors of the pixels in the specific zoom level.

In the lower left corner is a subordinate tab document that hosts several detailed displays relating to the content in the current zoom window of the map display. In Fig. 1A the MS spectrum display is visible, which shows the mass spectrum at the horizontal cross-section as indicated by the horizontal red line in the map view. The display can rapidly be moved through consecutive scans with the help of the up-down arrow keys or positioned on a specific retention time by clicking on the left border of the map display. This is useful, for instance, for inspecting the isotopic intensity distribution at the maximum of the elution profile of a peptide.

Another display item is the annotated MS/MS spectrum shown in Fig. 1B that identifies the peptide whose MS1 features are displayed in Fig. 1A. Annotations can include only the main series fragments plus water and ammonia losses, or alternatively an extensive set of peak annotations generated by an expert system [5], which includes internal fragments and a more extensive set of potential neutral losses. The annotated peptide sequence is displayed as well, indicating which main series peaks are identified along the sequence.

The tab "Chromatogram" (Fig. 1C) shows either the time dependence of the total ion current in the given retention time zoom window or similarly multiple intensity profiles in selected mass channels. For that purpose the m/z values are selected in the lower border of the map view. They can also be set automatically to all m/z values occurring in an isotope pattern or in a label pair or triplet, which is particularly useful for inspection of retention time shifts between different labeled versions of a peptide, which can be an issue of concern for particular labeling techniques [6].

Figure 1D shows the 3D visualization of the m/z-retention time area that is indicated by the blue rectangle in Fig. 1A. The 3D rendering is done with the 3D graphics capabilities of Windows Presentation Foundation, which is part of the .NET Framework 4.5 and therefore readily available on every windows computer. For comparison, other displays in this part of the software include an isotope peak simulation display where for a peptide molecule with known elemental composition the intensity profile is calculated with a desired resolution for the purpose of comparison to measured isotopic envelopes. Also the retention time calibration curves resulting from the nonlinear retention time alignment algorithm in MaxQuant can be displayed here at the current zoom levels.

In the upper right area all MaxQuant output tables can be browsed. The information content is the same as in the tab-separated output tables that can be found in the folder "combined\txt" in the processed MaxQuant project. These tables contain all results regarding identification and quantification of peptides and proteins where each table corresponds to a different organization level of the data. The most important of these tables are the protein groups, peptides, modification-specific peptides, evidence, and MS/MS tables. All these tables are interconnected among each other as well as with the displays on the left side. For instance all MS/MS spectra identifying a particular protein of interest can be selected in the corresponding table. The annotated MS/MS spectrum as well as the surroundings of the precursor in the MS1 plane are continuously updated when selecting a row in the MS/MS table. Publication-ready annotated MS/MS spectra can be exported in vector graphics format for any subset of spectra of interest by a single button click.

The lower right area of the main display contains the protein sequence view. For each protein group, sequences of all members are displayed in which the identified peptides are indicated and color-coded according to their uniqueness regarding their occurrence in the protein database. Additional annotation tracks can be displayed along the sequence. As a particularly interesting example, one can display here the PTMs identified in the MaxQuant project in one track and show in another the already known PTMs from a central repository, such as PhosphoSitePlus [7] or MaxQB [8]. A multitude of different sequence-specific annotations as derived from UniProt [9] as well as the Pfam domain structure [10] can be displayed as tracks.

Finally, a very useful feature is the multi-map view (Fig. 2). Here one can monitor a selection region in the *m/z*-retention time plane across many LC-MS runs. In Fig. 2A this can be seen for a particular peak that is viewed over nine different runs. These runs are triplicate groups of measurements of the same biological sample at three different mass resolutions. It is apparent that the retention time of the peak center varies appreciably between the runs, and in the central panel the peak of interest is not even present in the view area. The main map as well as the small maps in the multi-map view can also be displayed with the recalibrated retention time as the vertical axis that is shown in Fig. 2B. Here, the peak occurs at the same retention time in all nine samples. The multi-map view can verify that a particular feature used by the MaxLFQ algorithm for label-free quantification [11] is well aligned across the LC-MS runs involved and that it fulfils the criteria for the feature matching algorithm.

In summary, the Viewer component of MaxQuant has been thoroughly updated and now fulfils the demands of rich content visualization of high resolution proteomics data.

The authors have declared no conflict of interest.

References

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

[2] Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A. et al., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 2011, *10*, 1794–1805.

[3] Michalski, A., Cox, J., Mann, M., More than 100000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* 2011, *10*, 1785–1793.

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

[5] Neuhauser, N., Michalski, A., Cox, J., Mann, M., Expert system for computer-assisted annotation of MS/MS spectra. *Mol. Cell. Proteomics* 2012, *11*, 1500–1509.

[6] Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S., Heck, A. J., Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protocols* 2009, *4*, 484–494.

[7] Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B. et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 2012, *40*, D261–D270.

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

[9] UniProt Consortium, Activities at the universal protein resource (UniProt). *Nucleic Acids Res.* 2014, *42*, D191–D198.

[10] Finn, R. D., Bateman, A., Clements, J., Coggill, P. et al., Pfam: the protein families database. *Nucleic Acids Res.* 2014, *42*, D222–D230.

[11] Cox, J., Hein, M. Y., Luber, C. A., Paron, I. et al., Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 2014, *13*, 2513–2526.

2.1.4 The Perseus computational platform for comprehensive analysis of (prote)-omics data

One of the biggest challenges in proteomics is to extract biological meaning out of peptide/protein/PTMs quantification. The Perseus platform provides a variety of tools for data processing, statistical analysis, and visualization for omics data, including proteomics and genomics data[159].

My contribution to this paper was to develop the ability to analyze genomics data and to integrate it with other omics data.

Stefka Tyanova, Tikira Temu, **Pavel Sinitcyn**, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, Jürgen Cox
The Perseus computational platform for comprehensive analysis of (prote)-omics data
(2016) *Nature methods*
DOI: 10.1038/nmeth.3901

PERSPECTIVE

The Perseus computational platform for comprehensive analysis of (prote)omics data

Stefka Tyanova[1], Tikira Temu[1], Pavel Sinitcyn[1], Arthur Carlson[1], Marco Y Hein[2], Tamar Geiger[3], Matthias Mann[4] & Jürgen Cox[1]

A main bottleneck in proteomics is the downstream biological analysis of highly multivariate quantitative protein abundance data generated using mass-spectrometry-based analysis. We developed the Perseus software platform (http://www.perseus-framework.org) to support biological and biomedical researchers in interpreting protein quantification, interaction and post-translational modification data. Perseus contains a comprehensive portfolio of statistical tools for high-dimensional omics data analysis covering normalization, pattern recognition, time-series analysis, cross-omics comparisons and multiple-hypothesis testing. A machine learning module supports the classification and validation of patient groups for diagnosis and prognosis, and it also detects predictive protein signatures. Central to Perseus is a user-friendly, interactive workflow environment that provides complete documentation of computational methods used in a publication. All activities in Perseus are realized as plugins, and users can extend the software by programming their own, which can be shared through a plugin store. We anticipate that Perseus's arsenal of algorithms and its intuitive usability will empower interdisciplinary analysis of complex large data sets.

A decade ago, proteomics projects were labor intensive and cumbersome, and high-quality results required semimanual analysis of spectra for identification and quantification. Today, mass-spectrometry(MS)-based shotgun proteomics is reaching a level of maturity that makes it a powerful and broadly applicable technology for researchers in biology and biomedical sciences[1,2]. Consistent automatic processing of spectra and the identification of peptides, proteins and post-translational modifications (PTMs) with the help of search engines[3–7] and reliable workflows have become standard computational tasks for which satisfactory solutions exist for single studies as well as community-wide data reanalysis[8–10]. Sophisticated computational proteomics platforms enable the quantification of proteins and PTMs over many samples in a large variety of labeling or label-free formats[11]. Public repositories for the storage and dissemination of MS-based proteomics data exist in practical forms[12,13]. Complete proteome quantification is possible in yeast[14] under many different conditions or stimuli with modest measurement effort[15]. Starting with a cohort of human samples, protein expression matrices with sample-wise ratios or relative abundances can be readily obtained for more than 10,000 proteins[16–19].

These technological advances have shifted the bottleneck to the biological interpretation of quantitative abundance and PTM data and to the translation of high-dimensional molecular data into relevant findings within the domain of a particular biological or medical investigator. Many potentially important findings are not currently extracted from proteomics data simply because the computational methods and algorithms that would highlight them are not in the hands of the researcher with the necessary domain knowledge to appreciate the meaning of the findings. There are often barriers between informatics and biological researchers, which need to be bridged in order to translate omics technologies and data to valuable biological or medical discoveries.

Here, we address this void by creating a computational platform that fulfils two potentially conflicting objectives: (1) all methods should be statistically sound, powerful and comprehensive; (2) the platform should be intuitive and easy to use for the domain expert in a biomedical discipline who is not a computational expert. To reach these goals we developed

[1]Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, California, USA. [3]Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [4]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. Correspondence should be addressed to J.C. (cox@biochem.mpg.de).

RECEIVED 28 JANUARY; ACCEPTED 10 MAY; PUBLISHED ONLINE 27 JUNE 2016; DOI:10.1038/NMETH.3901

the Perseus platform in close collaboration with biologists, with whom we analyzed projects involving multiple, diverse and distinct data types and experimental approaches. Experienced Perseus users can perform essentially all the computational tasks alone, even with little or no formal bioinformatic training. Users can also involve programmers and bioinformatics specialists to extend the functionality of Perseus with plugins that add to the Perseus workflow as custom activities. Here we describe the functionalities available in version 1.5.5.0 of Perseus.

A comprehensive workflow-based data analysis platform

Downstream analysis of proteomic data is a multifaceted and demanding endeavor that integrates many aspects of bioinformatics, statistics and machine learning. It is common practice for biological researchers to involve bioinformaticians to help with various analytical problems. Often these efforts result in multiple small scripts that are tedious to maintain and scale and that require the help of the original developer to be reused or stitched together. This approach is bound to turn downstream data analysis into a major bottleneck for scientific projects and discoveries. Furthermore, the results may be of questionable validity when there is no clear documentation and transparency about the methods and scripts employed. We thus set out to develop the Perseus platform as a holistic software platform that allows continuous expansion of scalable analytical tools, their smooth integration and reusability while providing the user with explicit documentation of the analysis steps and parameters. Details on the implementation and download of Perseus are provided in **Box 1**.

Perseus offers a wide variety of algorithmic activities that cover topics ranging from data normalization through exploratory multivariate data analysis to integration with other omics levels (**Fig. 1**). In what follows, we highlight several computational and statistical tools in Perseus. Many activities in Perseus produce interactive graphical output for the visualization of data analysis results, which scale easily to very large sets of input data and therefore allow for thorough inspection by the user even for large-scale experiments with complex experimental designs and many measured variables. Any plot can be exported in a number of graphical formats and edited in standard vector graphics editors upon release of all clipping masks.

The central data type in Perseus is the 'augmented data matrix', which typically represents expression or abundance values of genes or proteins (rows) and biological samples or technical replicates (columns). It is supplemented by additional data containers for annotation of the rows, columns and cells of the matrix (see **Box 2**). These annotation containers are automatically filled by Perseus with gene or protein information derived from the publicly available ontologies, pathways and annotation databases. Sample annotations are used in many activities to define the study design, such as designating which samples are replicates or which belong to different treatments or time points in a time-series analysis.

The main navigation tool is the workflow panel, which is composed of matrices and activities and which controls the information flow in a Perseus session (**Supplementary Fig. 1**). The interactive workflow allows the user to keep track of all steps in the analysis and to navigate through data matrices and visualization components. It facilitates revisiting intermediate steps in a complex computational workflow, branching off with alternative parameter settings or a different combination of activities and comparing results of alternative branches to each other. The matrix objects move through the workflow and are transformed and modified by activities. The workflow itself is a bipartite graph in which every matrix is connected via an activity to the next matrix. A matrix can have interactive local visualizations attached (e.g., plots, histograms and heat maps). Activities can be of a simple single-input structure or they can receive inputs from several

BOX 1 SOFTWARE IMPLEMENTATION, DOWNLOAD AND MAINTENANCE

Perseus is implemented in the C# programming language from the .NET Framework 4.5 and runs natively on Windows operating systems. Perseus can be downloaded for free from http://www.perseus-framework.org under acceptance of our freeware license agreement and user account registration. No installation is required, and the software can immediately be used upon download and decompression of the zipped folder. Detailed descriptions of the functions and their parameters are available in the online documentation of Perseus, which is linked to the download page and can also be directly accessed from within the software. Other sources of user support include the active Perseus Google group (https://groups.google.com/forum/#!forum/perseus-list) with more than 1,400 users (May 2016), and the YouTube videos demonstrating the use of the software (https://www.youtube.com/c/MaxQuantChannel). Several complete analysis workflows are available on our DokuWiki pages (http://www.coxdocs.org/doku.php?id=perseus:user:use_cases:start) that contain step-by-step descriptions of three standard proteomics project types. Substantial changes are usually reflected in major yearly releases; however, we recommend updating the annotation files at shorter time intervals. Reproducible bugs in the latest available Perseus version can be reported via the YouTrack bug-tracking system (http://maxquant.myjetbrains.com/youtrack/).

Perseus has been codeveloped with MaxQuant[11], which has become a comprehensive and widely accepted environment for the analysis of MS-based proteomics data and which contains further proteomics-specific data visualization tools[70]. As a result, integration between Perseus and MaxQuant is excellent, but these environments are independent and can be used together with any upstream data analysis tool. Most of the data structures and algorithms are programmed from scratch, and only a few external libraries are used. An advantage of this design choice is that it gives us full control over all implementation details and helps improve performance, which can be significantly faster than the performance achieved in other statistical programming environments[71]. Like MaxQuant, Perseus will be continuously maintained and developed with the support of long-term funding by the Max Planck Society for the Advancement of Science.

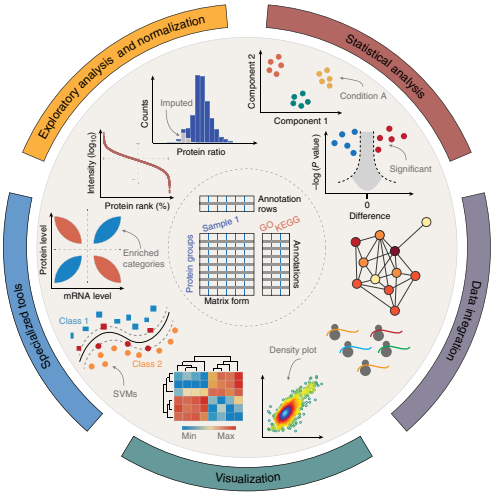


Figure 1 | The Perseus data analysis platform. The core data structure of Perseus is the data matrix, containing samples in columns and expression values (e.g., protein, mRNA) in its cells. Additional information such as GO terms, KEGG pathways and other database sources can be added for each row entry in the form of annotation columns. Perseus incorporates data cleansing and normalization and multiple methods for exploratory analysis such as histogram charts, intensity curves and scatterplots. Classical expression omics data analysis is supported by robust statistical tools including *t*-tests, PCA, correlation analysis and enrichment analysis. Beyond the standard methods, Perseus supports more complex tasks, such as supervised learning, for example Support Vector Machines (SVMs) and PTM data analysis.Furthermore, other types of omics data such as RNA seq data, as well as metadata from external databases such as KEGG and UniProt can easily be uploaded and analysed using the data integration modules. Min, minimum. Max, maximum.

matrices for the purpose of data integration when merging data from two or more different omics levels (see **Box 3**).

A session contains a workflow together with all intermediate results and parameter settings for all activities. Session files can be saved and reloaded and can also be shared with other researchers who can load them into their Perseus instance for collaborative data analysis. Furthermore, the workflow and the session serve as a complete account of the computational methods used in a project, representing an accurate and reproducible description of the data analysis for documentation or publication.

Plugin architecture

Perseus is not a static and monolithic software tool; rather, it is based on a plugin architecture that can be extended by the users (**Supplementary Fig. 2**). Perseus and its plugins are written in the C# programming language and adhere to a standardized application programming interface (API) that consists of a set of interfaces defining the minimum functionality that a plugin must implement. The five main interfaces in Perseus (data upload, export, processing, analysis and multimatrix handling) form the foundation of the extensible plugin architecture (**Supplementary Fig. 2**). Plugins implementing these interfaces are visually distributed along the ribbon control menu of Perseus. The source code of many of the plugins is available from our GitHub repository (http://www.github.com/JurgenCox/perseus-plugins). Using this source code as examples and the plugin architecture of Perseus, developers can easily expand the current functionalities by programming novel independent modules. The compiled Dynamic Link Library (DLL) then has to be placed into the main folder of the Perseus installation, which will completely integrate the DLLs, making them ready for use. A tutorial video on how to program plugins is available at http://www.youtube.com/watch?v=MhS4UM1CMwU.

The API allows any user to program activity plugins in their local development environment independent of the central Perseus code repository. We provide a core set of plugins containing more than 100 activities that are bundled with the standard Perseus download and that can also be reused in newly developed activities (**Supplementary Table 1**). For the majority of these plugins, the source code is provided via GitHub. Once users have programmed a new plugin, they can make it available through the Perseus plugin store (http://www.perseus-framework.org/plugins). As an example, the 'Proteomic ruler' package combines convenient functionality for the absolute quantification of protein copy numbers per cell from generic label-free shotgun proteomics data[20].

Application highlights

Expression proteomics. Many proteomics projects consist of measuring cells or tissues in two or more conditions, in a certain number of biological replicates per condition, for instance, using relative label-free quantification[21] or a common labeled reference standard[22] for enhancing quantification accuracy. These kinds of proteomics data have similarities to transcriptomics microarray data, and their analysis can benefit from the wealth of experience obtained in more than two decades of transcriptome data analysis by a large community. Perseus includes adaptations of some of these algorithms to proteomics workflows.

Before proteomics data can be used for the actual data analysis, they need to be normalized, filtered and potentially subjected to missing-value imputation, for which we provide a multitude of options in the standard set of Perseus activities (**Supplementary Fig. 3**). One common task is to determine which proteins are significantly changing between conditions. Perseus adapts a particularly robust method in the 'two-sample tests' and 'multisample tests' activities, which originate from microarray data analysis that includes a permutation-based false-discovery rate (FDR) and *q*-value estimation[23]. This enables reliable estimation of the percentage of proteins that are mistakenly indicated as changing.

Another frequent task is to find the main clusters of expression patterns in the data and the sets of proteins responsible for the formation of these patterns. We provide a hybrid hierarchical *k*-means clustering algorithm that creates interactive heatmaps and scales to matrices with a large number of rows and/or columns in a short computing time. As an alternative to clustering, Perseus includes principal-component analysis (PCA) based on singular value decomposition[24], a form that computationally performs well on high-dimensional data. PCA detects the main effects in the data and the proteins driving the separation of the proteomic states.

Once an interesting cluster of proteins has been identified, enrichment analysis[25] of biological processes, complexes or pathways is done in a variety of ways, for instance, with the Fisher's exact test checking for contingency between cluster membership

BOX 2 AUGMENTED DATA MATRIX

The central data format of Perseus is the data matrix, in which biological samples are represented as columns, and proteins or other molecular species are represented as rows. Perseus distinguishes several different types of columns. Upon reading new input data, the type of each column must be specified. In case the data comes from the MaxQuant environment[11], the suitable type of most columns of the output tables is automatically detected via the column name. The main data are stored in the 'main columns', which typically contain the protein expression values that are to be subjected to downstream normalization, transformation, etc., and Perseus automatically selects them for statistical tests and data visualization. Other numerical values that serve as annotations, such as sequence length, number of identified peptides or posterior error probabilities, are stored in 'numerical columns'. This data can also be explored by standard summary statistics and visualization tools, but no statistical tests (e.g., for differential expression) are applied to them. Non-numerical information can be stored as 'text' or 'categorical' columns. 'Text' is suitable for storing protein, RNA and gene names and identifiers, and these columns are available as data labels in plots. In data integration, this information is interpreted as identifiers to match rows of different matrices to each other or to an external data source. Categorical columns contain data of an enumerable type about each protein, which often signify membership in biological processes or ontologies. This column type is used in enrichment analysis. The column type 'multi-numerical' can contain multiple numerical values per entry. Most activities make a preselection of columns based on the designated type for a specific context, so it is most convenient that the column types are assigned correctly from the beginning. However, the data type can be changed retroactively if necessary.

Several functions in Perseus rely on additional supplementary data matrices that contain metainformation about the main data matrix (**Supplementary Fig. 5**). Missing values are a common problem of large-scale data in general, as some statistical methods cannot handle missing information and therefore require 'imputation' before the analysis[72,73]. Perseus offers several imputation techniques, including a method drawing random values from a distribution meant to simulate expression below the detection limit (**Supplementary Fig. 3**). Upon imputation, a Boolean background matrix is created (**Supplementary Fig. 5a**), which keeps track of which value was measured and which was imputed. This allows visualization and filtering of imputed values during downstream analysis. Similarly, the user can generate a 'quality matrix', which will be stored in the background as well. The 'quality matrix' contains one corresponding value to each entry in the main data matrix and can be used to filter the main matrix (**Supplementary Fig. 5b**). For example, a 'quality matrix' can be generated from the number of peptides used in the quantification of each protein in each sample. This can be useful to mask all cases where a given protein was quantified with less than two peptides in a given sample. The phosphorylation-site table is another example in which such filtering is desirable, as sites with occupancy errors larger than a fixed threshold can be filtered out using a 'quality matrix' containing the site-specific errors.

Data that characterizes the samples (i.e., information regarding the experimental design) is added to Perseus via row annotations. The groupings used in analysis methods such as *t*-test statistics and machine learning approaches are set as categorical row annotations (or numerical ones in case of continuous data, such as the time point for time-series data) and are automatically recognized by the software in all suitable procedures.

and the property of interest. The FDR is controlled with the Benjamini–Hochberg method[26]. This method elucidates what the cluster-member proteins have in common, and it provides clues about the functional role of the cluster. Similar enrichment functionalities have been developed in the context of genomic technologies[27], and Perseus adapts these enrichment analyses so that they are specifically tailored to the purpose of proteomics. In particular, the reference space for enrichments is always

BOX 3 DATA INTEGRATION

One of the most laborious and error-prone steps in data analysis is the matching and integration of different data types. Through its multiprocessing interface, Perseus offers an easy way to combine matrices and to import information from external databases. Two matrices can be matched based on any identifier that is provided as a column in each of them, and the information to be transferred from one matrix to the other can be selected as well. Cases in which multiple entries from one matrix map to a single entry in the other are handled by the software in user-selectable ways; for instance, multiple numeric values from multiple rows in one matrix can be summarized to a single entry in the other matrix. Furthermore, different omics data sets can easily be mapped through the prebuilt genome lists that can be loaded with a single click.

Interpretation of genome-scale data often incorporates functional information such as pathways, cellular function and localization as well as structural information. In Perseus the user can upload a preprocessed set of annotations from UniProt[74] and use these in filtering and enrichment analysis of the data. Furthermore, PTM-specific annotations such as those obtainable from PhosphoSitePlus[33] and common kinase motifs can be automatically assigned by the software. Integration of user-defined curated annotations is supported in Perseus if certain simple file format requirements are met. The software can read customized annotations from tab-delimited text files, in which the first column contains the identifiers that will be used for matching the annotations to the main matrix, and the header row contains the names of all annotations to be added. All further columns contain the customized annotations.

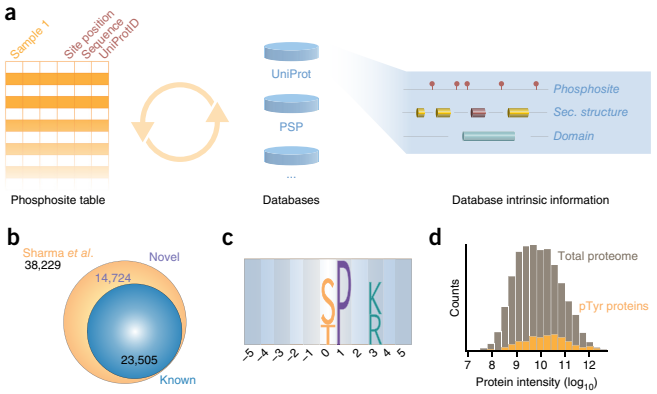


Figure 2 | Post-translational modifications. (**a**) Annotations from various resources including UniProt and PhosphoSitePlus (PSP) can be mapped onto each phosphorylation site via the protein identifier, the modified amino acid and its position. Multiple site-specific annotations from UniProt, including protein secondary (sec.) structure, information as to whether a site is known to be biologically important and domain information, can easily be imported into Perseus. Orange circle illustrates the integration of various databases into the PhosphoSite table. (**b**) Estimation of the number of novel phosphorylation sites detected in an experiment by Sharma *et al.*[31] as compared to known sites stored in public repositories, using Perseus-based analysis. (**c**) A set of short sequences surrounding a modification site (location "0" in the diagram) can be used to generate a sequence logo and scale it by entropy in order to identify possible recognition motifs. (**d**) Comparison of the protein intensity distributions of matched total and phosphotyrosine (pTyr) proteomes showing that phosphotyrosines preferentially appear on more abundant proteins[31].

appropriately chosen to be the subset of measured proteins. Furthermore, proteins that are indistinguishable based on the measured peptides are not double counted in enrichment tests because the occurrence of multiple alternative identifiers in enrichment tests is appropriately handled.

Post-translational modifications. Proteomics software typically generates a table for each PTM type of interest, indicating all positions on the identified proteins that are likely to be modified in at least one of the conditions of a study. In addition to scores reflecting the reliability of identification and the confidence in the localization of each site in the protein sequence[28,29], quantitative information is crucial for understanding the functional role of the modification sites. Relative quantification in the form of site-specific ratios or intensity-based quantification is usually required for the comparison of phosphorylation in different conditions or upon different stimuli. Furthermore, analysis of the proportion of modified to total peptides, i.e., site occupancies, is important for the elucidation of major regulatory phosphorylation events during key cellular processes[30,31].

Reformatting tools are provided in Perseus that transform the site quantification into a matrix that resembles proteome-expression data, which retains information about multiple modification states of peptides. This matrix can then be analyzed with similar methods as introduced in the previous section for expression proteomics, but with some special adaptations. For example, to place phosphorylation events in the context of cellular pathways and signaling events, enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO)[32] terms can be employed. Importantly, as proteins are often characterized by multiple phosphorylation sites, care should be taken to avoid overcounting of protein-derived annotation in PTM-site-based analysis ('protein-relative enrichment').

Integration of external resources is currently a tedious task that requires building access to the databases, parsing the data in the correct format and finally matching identifiers to the in-house data. In Perseus, site-specific annotation, for instance, from PhosphoSitePlus[33], or sequence-position-specific annotation from UniProt are integrated by using an easily operated activity designed for that purpose (**Fig. 2**). This information can be used to generate statistics on which sites in the study are already known from other publications or which are novel and to import experimentally known kinase–substrate relationships into the matrix. Alternatively, kinase motifs are matched to the sequence window surrounding the phosphorylation site which, when combined with clustering and enrichment analysis, often leads to noteworthy conclusions about kinase activity patterns[34]. Reversible phosphorylation is regulated by multiple factors, including increased or decreased concentration of kinases and phosphatases, and the level of phosphorylation may appear to vary on account of changes in the abundance of the modified protein itself. Therefore, Perseus enables straightforward overlaying of modification site and protein abundance to determine the actual quantitative changes in phosphorylation on a certain site and their likely origin.

Interaction proteomics. Affinity-enrichment experiments followed by MS for determining interaction partners can now be performed on a large scale involving more than a thousand bait proteins[35,36]. This works well with intensity-based relative label-free quantification[21], but SILAC- or TMT-based quantification can also be used. Typically, analysis of such data requires comparison of the quantities of individual proteins in specific samples with those in a control group (**Fig. 3a**). The control may derive from cells not expressing a tagged bait protein[37] or cells in which the bait was knocked down[38]. Alternatively, all samples in which unrelated proteins served as bait can be used as negative controls, which we have shown to be the superior control in medium-scale[39] and large-scale data sets[35]. Perseus allows the streamlined calculation of large numbers of tests necessary to derive a list of statistically significant outliers specific to each bait, with permutation-based FDR control for each pair of sample and controls. The resulting network of interactions can automatically be formatted to be uploaded to external tools like Cytoscape[40] for visualization (**Fig. 3b**).

For some experimental setups it is necessary to control the FDR globally instead of on the level of individual samples, for instance, when interactions are measured under different conditions

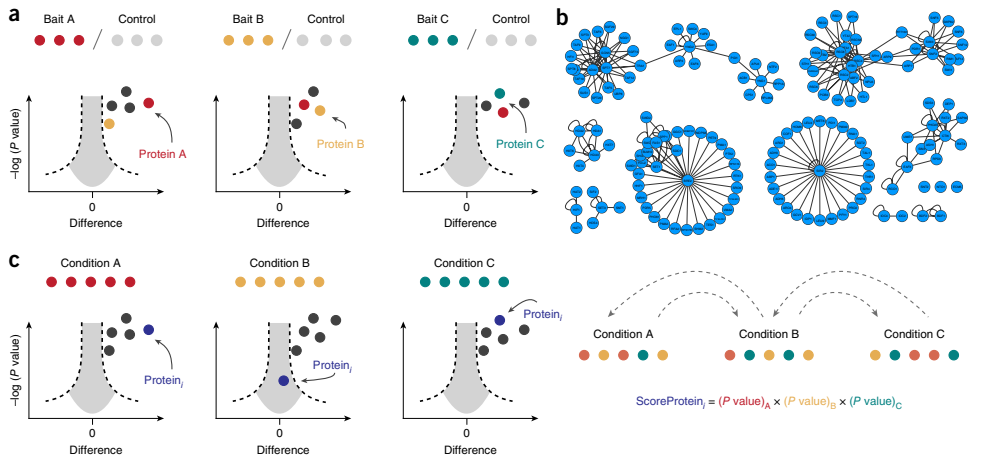


Figure 3 | Interaction proteomics. The interaction partners of the baits in a large set of pulldowns can be determined in a multivolcano analysis using Perseus. Control groups can be defined in multiple ways: (**a**) common control group for all pulldowns (as shown), specific controls for each pulldown or the complement group of each pulldown set (i.e., the union of all other pulldowns). (**b**) Cytoscape visualization of the interaction network generated by Perseus using affinity enrichment data from ref. 75. (**c**) The total set of interactors and interactions from all pulldowns is determined by a global permutation-based FDR approach implemented in Perseus. For each condition a two-sample test is performed with all other conditions serving as control. The global set of interactors at a given value for the FDR is obtained by a permutation involving all conditions. val., value. *i* indicates that the same protein was compared in conditions A, B, and C. ScoreProtein$_i$, global interaction score computed by combining the probabilities from all conditions.

or over a time course[41]. To this end, Perseus offers a method to combine FDR-based cutoffs for multiple samples (**Fig. 3c**). This method is an advantage over methods such as ANOVA because it retains information about the enrichment of each protein in each condition (which is lost in ANOVA) while additionally offering global-level statistics.

Time-series analysis. Many biological processes are controlled by characteristic temporal changes in the concentrations of specific biomolecules. For instance, the cell cycle is accompanied by periodic changes in mRNA and protein expression[42–44]. Likewise, the circadian cycle[45] involves concerted changes in abundances of proteins, protein modifications, mRNAs and metabolites[46]. Perseus contains an FDR-controlled method for detecting expression behavior that is statistically significantly following a given temporal model such as expression with a given periodicity (**Fig. 4**). To derive the length of the cycle from the data, a Fourier-based periodicity analysis can be performed that determines the base frequency of periodic expression changes and also allows screening for other possible cycle lengths (e.g., harmonics of the base frequency). The analysis will assign an amplitude of change and a peaking time to each protein. A specialized annotation enrichment analysis designed for periodic expression changes can then determine which biological processes or pathways are switched on at which points along the time axis, detecting clusters of activity in the time dimension. Side-by-side analysis of transcriptome and proteome reveals the time lag between transcription and translation[46].

Cross-omics data analysis. Perseus has activities for comparing proteomics data to other omics dimensions, such as mRNA levels as measured by RNA-seq[47]. An importer activity loads next-generation sequencing (NGS) short-read information, such as that obtained by the Illumina platform, into a Perseus session. Reads can be aligned by standard spliced-alignment workflows as, for example, those provided by the TopHat[48] or STAR[49] suites, and read-count-based quantification is generated upon upload to Perseus. Multiple reference-genome-aligned read files corresponding to data from multiple samples can be used simultaneously, and a Perseus matrix will be filled with read-count information for each gene. The reads can represent RNA-seq or ribosome-profiling data[50], which are then converted to quantitative expression profiles by, for instance, calculating reads per kilobase of transcript per million mapped reads (RPKM) values[51]. To investigate the relationship between transcription and translation, this matrix can then be matched to another matrix containing protein expression values such as iBAQ values, which are estimates of absolute protein abundances[52,53]. This enables correlation analysis between the two quantitative omics dimensions (**Supplementary Fig. 4**), and for this purpose we routinely use the vast amounts of freely available NGS data ready for download—for example, from the Ensembl[54] (http://www.ensembl.org/info/data/ftp), European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) or Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) databases—most of which are already aligned to the reference genome. Hence, the 'NGS data upload' plugin enables comprehensive analysis of multiple genomics experiments and comparison with proteomics data in a very short time.

To compare functional differences between any two omics types, we implemented the so called '2D annotation enrichment' activity[55] (**Fig. 5**), which identifies annotation terms whose members show statistically significant outlier behavior in the two dimensions chosen. Genome-wide annotation for this purpose

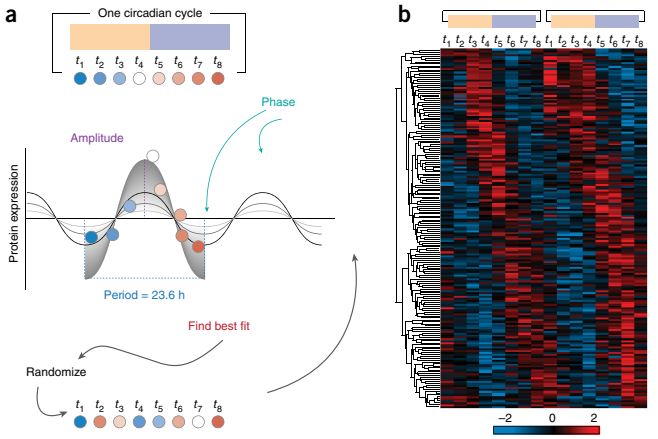


Figure 4 | Time-series analysis. The time-series set of Perseus plugins contains a periodicity-analysis component that allows detection of periodic oscillations in protein expression over time. (**a**) The amplitude (expression level) and phase (upregulation or downregulation) are determined by the software by optimizing a cosine function fit to the data. A permutation-based approach in which the time points are randomly reshuffled multiple times identifies the statistically significantly oscillating proteins, exemplified by global circadian oscillations of the proteome in mouse liver[46]. *t*, time point. (**b**) A total of 180 proteins were found to follow circadian rhythm over two cycles, and characteristic phases of upregulation and downregulation were clearly characterized as illustrated by the red and blue clusters, respectively. 0 in scale bar corresponds to the logarithm of ratio 1.

can be membership of proteins in biochemical pathways, gene ontology terms, subcellular localization, protein domain content or membership in protein complexes. Processes can be simultaneously upregulated or downregulated in both dimensions, or they can lack correlation, such as regulation in one dimension without any corresponding change in the other. We have found 2D enrichment analysis to be a powerful tool to probe regulation of the respective pathways or biological processes including, but not limited to, information about the processes that are predominantly transcriptionally, post-transcriptionally or post-translationally regulated.

Machine learning for detecting subtle biological associations and biomarker discovery

Patients can greatly benefit from a more accurate diagnosis and a subsequently more efficient personalized treatment. Perseus combines powerful machine learning and statistical methods for the classification of proteomics samples. For example, we have applied Perseus to study clinical classification of disease subtypes from proteomic data in lymphoma[56], prostate cancer[57] and breast cancer[58] studies. In Perseus we provide an extensible classification and regression framework that does not rely on a single 'favorite' machine learning technique (**Fig. 6**). Instead, at every stage one algorithm can be exchanged for another and rated, making it possible for the nonspecialist to determine the machine learning method that is best suited for a particular type of data. In addition to the many algorithms for classification, regression and feature selection that are provided together with the standard Perseus release, including a support vector machine[59] implementation, the machine learning framework is extensible, allowing the users to program their own implementations of algorithms. We provide stable APIs for classification and regression models as well as for feature-selection algorithms in the context of classification and regression. For example, we adapted the popular LIBSVM[60] implementation of a support vector machine as an open-source classification plugin.

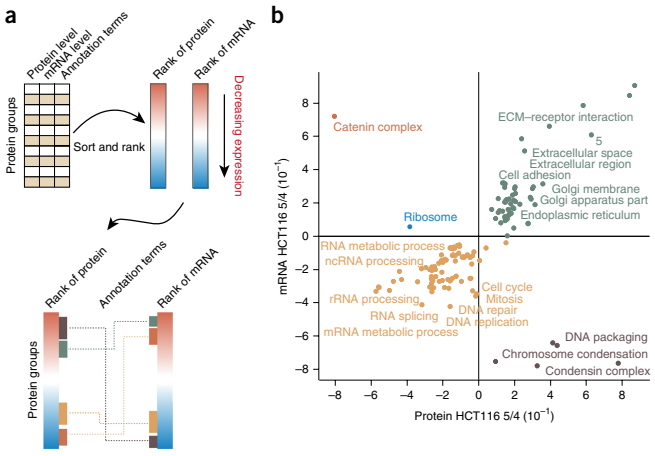


Figure 5 | Cross-omics data comparison by 2D annotation enrichment analysis. (**a**) Proteome and transcriptome expression data are joined into one Perseus matrix. Both omics columns are sorted and transformed into ranks. A bivariate test is performed on each annotation term, checking if the protein–mRNA pairs belonging to a certain process show a common trend, for instance, if they are upregulated in both dimensions. (**b**) The processes and locations represented by green dots show common upregulation at both mRNA and protein levels, whereas the orange dots indicate simultaneous downregulation (data from ref. 76). The processes represented by brown dots exhibit upregulation at the protein level, while the corresponding mRNA levels are collectively downregulated. The blue dot indicates downregulation at the protein level and no significant changes at the mRNA level, whereas the red dot indicates upregulation at the mRNA level and downregulation at the protein level. ncRNA, noncoding RNA; 5, chromosome 5.

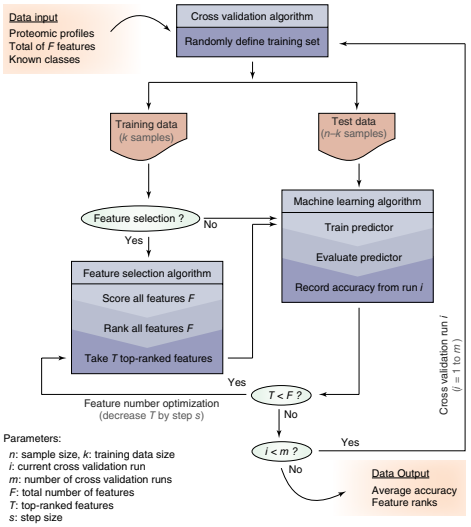


Figure 6 | Machine learning for clinical proteomics and biomarker discovery. The Learning plugin in Perseus provides implementation of classification and regression analyses and implements various feature-selection methods. Estimation of the accuracy of a trained predictor, including the feature-selection step, is performed in a crossvalidation procedure in which the data set is first split into training and test subsets and the classifier is trained on the training set, and its performance is then estimated on the test set. After training, the classification–regression model then assigns a predicted class to the samples of unknown class. The feature-selection procedure outputs the ranks for all proteins with best ranks corresponding to the most discriminative proteins in the data. The learning module is complemented by an algorithm for screening for the optimal parameters of the different classification algorithms to maximize the classifier's performance.

The machine learning framework of Perseus has a crossvalidation structure for the purpose of measuring how the prediction performance of classification or regression will generalize to independent data that have not been used for model building, thereby avoiding notorious problems such as overfitting[61]. The crossvalidation tools allow robust determination of optimal parameter values in linear or nonlinear models used for prediction. Furthermore, they help in extracting optimal protein sets from the output of a feature-selection algorithm that strike a balance between good prediction performance and simplicity. This machine-learning-based feature selection, combined with accurate monitoring of the prediction errors by crossvalidation, offers a complement to *t*-test-like approaches for determining discriminating protein subsets. It detects multivariate patterns in protein expression profiles, for which the discriminatory power might not be apparent in the expression profiles of single proteins. In this way we can retrieve the members of protein response networks that are invisible to univariate feature-selection methods.

Vision and future developments

Perseus integrates a large amount of bioinformatic expertise based on experience in the analysis of diverse types of large-scale proteomics data. The software offers an intuitive interface that enables researchers without formal computational skills to analyze their own data by guiding them through statistical procedures in a rigorous manner and equipping them with various tools for the extraction of maximum information and biological insights from the data. We have strived to lower the 'activation barrier' to the adoption of Perseus through the absence of installation procedures, by making it completely freely available, by ensuring users' ability to visualize every step with intuitive and interactive plots, and through the automatic generation of a complete record of each analysis step and the parameters used. We believe the latter feature is crucial for the scientific community, as it fosters transparency and reproducibility of the reported results. Moreover, the use of a common platform for analysis allows for unbiased comparison of results generated in different groups and enhances collaboration between scientists by simplifying the process of documentation and sharing of protocols.

Through continuous development and maintenance, our goal is to establish Perseus as a comprehensive analysis and visualization tool for systems biology research, similarly to what we have done previously with the MaxQuant software for the analysis of mass spectrometric data[11]. As the experimental designs become more and more complex, the functionality of Perseus will be enriched accordingly, building upon its extensible architecture to offer more tools and to support future data types. In particular, a comprehensive toolset for the analysis of biological networks[62–64] resulting from coexpression or interaction studies will soon be included.

For most of the development of activities in Perseus, we started with proteomics data in mind, as well as their comparison to other omics dimensions. However, we have found that many of the techniques implemented in Perseus are applicable to other data types without major modifications, and Perseus has already become popular in our group for the analysis of nonproteomics data as well. In the future, metabolomics data with relative quantification profiles for a global set of metabolites over several samples, which are similar to label-free-quantification proteomics data, will be accommodated by Perseus with only slight adaptations, such as customization of the annotation of molecular species.

One major challenge and opportunity that will drive the future development of Perseus is the bridging of the current gap between large-scale proteomics data generation and modeling of signaling pathways and biochemical reactions. Modeling studies are still generally performed only on low-throughput data, such as western blots or fluorescence-activated cell sorting (FACS) data. Our goal will be to provide a more automated way to extract quantitative information from large-scale data that can directly be used as input for available modeling platforms[65–67]. Providing automatically meaningful and reliable connections to signaling pathways will also require more extensive knowledge of the behavior of PTM sites in biochemical and signaling pathways than what is currently available in public resources[68,69].

Perseus has already been 'battle tested' in cutting-edge proteomics research. We anticipate that it will allow researchers from many areas of life science, including fundamental biology, drug discovery and medical sciences, to increasingly participate

directly in sophisticated data analyses. Our hope is that this novel platform will contribute to better communication between disciplines and more effective application of computational tools.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 686547 (J.C.) and from the FP7 grant agreement GA ERC-2012-SyG_318987–ToPAG (J.C.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Altelaar, A.F., Munoz, J. & Heck, A.J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).
2. Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80**, 273–299 (2011).
3. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
This publication describes the earliest approach to correlating tandem mass spectra of peptides to theoretical fragment-ion series calculated from *in silico* digests of known protein sequences with the aim of identifying peptides and proteins.
4. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
5. Geer, L.Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
6. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
7. Bern, M., Cai, Y. & Goldberg, D. Lookup peaks: a hybrid of *de novo* sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **79**, 1393–1400 (2007).
8. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3**, 1234–1242 (2004).
9. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).
10. Deutsch, E.W. *et al.* Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.* **9**, 745–754 (2015).
11. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
Perseus has been developed in conjunction with MaxQuant, which comprises a complete quantitative workflow for the analysis of shotgun proteomics data, including support for a large variety of experimental techniques.
12. Vizcaino, J.A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).
13. Vizcaíno, J.A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
14. de Godoy, L.M. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
15. Hebert, A.S. *et al.* The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).
In this paper the authors demonstrate that the yeast proteome can be analyzed within a 1-h measurement time, recovering nearly all expressed cellular proteins.
16. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011).
17. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549 (2011).
18. Munoz, J. *et al.* The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol. Syst. Biol.* **7**, 550 (2011).
19. Mann, M., Kulak, N.A., Nagaraj, N. & Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **49**, 583–590 (2013).
20. Wisniewski, J.R., Hein, M.Y., Cox, J. & Mann, M.A. 'Proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* **13**, 3497–3506 (2014).
21. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, yermed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
Here the MaxLFQ algorithm for relative label-free protein quantification is described. It enabled many researchers to conduct large proteomics studies with complex experimental designs without the need for labeling their samples.
22. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7**, 383–385 (2010).
23. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
A pioneering method is described for the robust detection of significantly changing biomolecules in large omics data sets. It uses repeated permutations of the data to determine FDRs.
24. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
25. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
GSEA is the forerunner of many methods for analyzing molecular profiling data to determine which sets of genes or proteins are correlated with a phenotypic class distinction.
26. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
In this seminal paper a simple yet powerful procedure is shown to control the FDR for multiple testing of many independent hypotheses.
27. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
28. Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
29. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
30. Olsen, J.V. *et al.* Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* **3**, ra3 (2010).
31. Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8**, 1583–1594 (2014).
32. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
33. Hornbeck, P.V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
34. Tyanova, S., Cox, J., Olsen, J., Mann, M. & Frishman, D. Phosphorylation variation during the cell cycle scales with structural propensities of proteins. *PLoS Comput. Biol.* **9**, e1002842 (2013).
35. Hein, M.Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
36. Huttlin, E.L. *et al.* The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
37. Hubner, N.C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals *in vivo* protein interactions. *J. Cell Biol.* **189**, 739–754 (2010).
38. Selbach, M. & Mann, M. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat. Methods* **3**, 981–983 (2006).
39. Keilhauer, E.C., Hein, M.Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* **14**, 120–135 (2015).
40. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

41. Räschle, M. *et al.* DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links. *Science* **348**, 1253671 (2015).
42. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
43. Gauthier, N.P. *et al.* Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res.* **36**, D854–D859 (2008).
44. Eser, P. *et al.* Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Mol. Syst. Biol.* **10**, 717 (2014).
45. Partch, C.L., Green, C.B. & Takahashi, J.S. Molecular architecture of the mammalian circadian clock. *Trends Cell Biol.* **24**, 90–99 (2014).
46. Robles, M.S., Cox, J. & Mann, M. *In vivo* quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genet.* **10**, e1004047 (2014).
47. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
48. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
49. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
51. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
52. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
In this publication a large-scale quantitative analysis of transcription and translation rates is performed, introducing the iBAQ technique for estimating protein abundances from mass-spectrometry data.
53. Aviner, R., Shenoy, A., Elroy-Stein, O. & Geiger, T. Uncovering hidden layers of cell cycle regulation through integrative multi-omic analysis. *PLoS Genet.* **11**, e1005554 (2015).
54. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).
55. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13**, S12 (2012).
56. Deeb, S.J. *et al.* Machine learning-based classification of diffuse large B-cell lymphoma patients by their protein expression profiles. *Mol. Cell. Proteomics* **14**, 2497–2460 (2015).
57. Iglesias-Gato, D. *et al.* The proteome of primary prostate cancer. *Eur. Urol.* **69**, 942–952 (2016).
58. Tyanova, S. *et al.* Proteomic maps of breast cancer subtypes. *Nat Commun.* **7**, 10259 (2016).
59. Vapnik, V.N. *The Nature of Statistical Learning Theory* (Springer, 1995).
60. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).
61. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2001).
62. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
63. Ideker, T. & Krogan, N.J. Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).
64. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
65. Hoops, S. *et al.* COPASI–a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067–3074 (2006).
66. Angermann, B.R. *et al.* Computational modeling of cellular signaling processes embedded into dynamic spatial contexts. *Nat. Methods* **9**, 283–289 (2012).
67. Cowan, A.E., Moraru, II., Schaff, J.C., Slepchenko, B.M. & Loew, L.M. Spatial modeling of cell signaling networks. *Methods Cell Biol.* **110**, 195–221 (2012).
68. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477 (2014).
69. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
70. Tyanova, S. *et al.* Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* **15**, 1453–1456 (2015).
71. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
72. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
73. Liew, A.W., Law, N.F. & Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief. Bioinform.* **12**, 498–513 (2011).
74. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
75. Hosp, F. *et al.* A double-barrel liquid chromatography-tandem mass spectrometry (LC-MS/MS) system to quantify 96 interactomes per day. *Mol. Cell. Proteomics* **14**, 2030–2041 (2015).
76. Stingele, S. *et al.* Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol. Syst. Biol.* **8**, 608 (2012).

2.2 Multi-omics applications

2.2.1 Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry

The Nobel Prize in Physiology or Medicine 2018 was awarded for "the discovery of cancer therapy by inhibition of negative immune regulation."[160, 161] This discovery opens a new area in cancer therapy, where the patient's immune system is used to fight its tumor.

Certain tumors accumulate a large amount of mutations[162]. Some of them happen to be non-synonymous and potentially could be present on the surface of cancerous cells as immunopeptides. As a result, some of these mutations can be recognized by the immune system. Our group[53], as well as another group[163], has tried to explore a possibility to directly identify such peptides from patient's tumors using advanced mass spectrometry analysis.

My contribution to this study was to develop a pipeline that identifies somatic mutations from paired - tumor and peripheral blood mononuclear cell (PBMC) - NGS data and then include them into peptide search database for their potential proteomics identification.

Using this approach, we were able to find one to three peptides with somatic mutations per patient[53]. Further immunological assessment of identified peptides shows that this approach is promising for personalized immunotherapeutics.

Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, **Pavel Sinitcyn**, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, Marc E Martignoni, Angelika Werner, Rüdiger Hein, Dirk H Busch, Christian Peschel, Roland Rad, Jürgen Cox, Matthias Mann, Angela M Krackhardt
Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry
(2016) *Nature communications*
DOI: 10.1038/ncomms13404

nature COMMUNICATIONS

ARTICLE

Received 9 May 2016 | Accepted 30 Sep 2016 | Published 21 Nov 2016

DOI: 10.1038/ncomms13404 OPEN

Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry

Michal Bassani-Sternberg[1,†,*], Eva Bräunlein[2,*], Richard Klar[2], Thomas Engleitner[3,4], Pavel Sinitcyn[1], Stefan Audehm[2], Melanie Straub[5], Julia Weber[3,4], Julia Slotta-Huspenina[5,6], Katja Specht[5], Marc E. Martignoni[7], Angelika Werner[7], Rüdiger Hein[8], Dirk H. Busch[9], Christian Peschel[2,4], Roland Rad[3,4], Jürgen Cox[1], Matthias Mann[1,**] & Angela M. Krackhardt[2,4,**]

Although mutations may represent attractive targets for immunotherapy, direct identification of mutated peptide ligands isolated from human leucocyte antigens (HLA) on the surface of native tumour tissue has so far not been successful. Using advanced mass spectrometry (MS) analysis, we survey the melanoma-associated immunopeptidome to a depth of 95,500 patient-presented peptides. We thereby discover a large spectrum of attractive target antigen candidates including cancer testis antigens and phosphopeptides. Most importantly, we identify peptide ligands presented on native tumour tissue samples harbouring somatic mutations. Four of eleven mutated ligands prove to be immunogenic by neoantigen-specific T-cell responses. Moreover, tumour-reactive T cells with specificity for selected neoantigens identified by MS are detected in the patient's tumour and peripheral blood. We conclude that direct identification of mutated peptide ligands from primary tumour material by MS is possible and yields true neoepitopes with high relevance for immunotherapeutic strategies in cancer.

[1] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, Martinsried 82152, Germany. [2] IIIrd Medical Department, Klinikum rechts der Isar, Technische Universität München, Ismaningerstr. 22, Munich 81675, Germany. [3] IInd Medical Department, Klinikum rechts der Isar, Technische Universität München, Ismaningerstr. 22, Munich 81675, Germany. [4] German Cancer Consortium of Translational Cancer Research (DKTK) and German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [5] Institute of Pathology, Technische Universität München, Ismaningerstr. 22, Munich 81675, Germany. [6] MRI-TUM-Biobank at the Institute of Pathology, Technische Universität München, Ismaningerstr. 22, Munich 81675, Germany. [7] Surgery Department, Klinikum rechts der Isar, Technische Universität München, Ismaningerstr. 22, Munich, 81675, Germany. [8] Dermatology Department, Klinikum rechts der Isar, Technische Universität München, Biedersteiner Str 29, Munich 80802, Germany. [9] Institute for Medical Microbiology, Immunology and Hygiene, Technische Universität München, Trogerstr. 30, Munich 81675, Germany. † Present address: Department of Oncology, UNIL/CHUV, Ludwig Cancer Research Center, Epalinges 1066, Switzerland. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to M.M. (email: mmann@biochem.mpg.de) or to A.M.K. (email: angela.krackhardt@tum.de).

Cancer immunotherapy has demonstrated remarkable efficacy in a large variety of neoplasms and is currently revolutionizing the treatment of malignant diseases. Immune checkpoint modulation, in particular, is emerging as a highly effective therapeutic strategy in an increasing number of cancer entities[1,2]. To further improve current immunotherapeutic approaches, understanding the nature of immunological tumour recognition is of utmost importance. This may be important also for the identification of suitable biomarkers influencing decisions regarding therapeutic sequences and combinations. A number of tumour-associated antigens (TAA) have been evaluated as target antigens in clinical investigations especially in patients with melanoma. These include antigens derived from differentiation antigens and cancer testis antigens[3,4]. However, so far these approaches showed only limited efficacy. Adoptive transfer of T cells transgenic for T-cell receptors (TCR) specific for selected TAA seem to be a reasonable and effective therapeutic development especially using affinity maturated TCR or ones selected from a non-self environment[5,6]. However, severe or even fatal side effects have been observed[5–8]. Response rates observed following treatment with immune checkpoint inhibitors have demonstrated that effective immune responses can be induced in an autologous environment in a significant proportion of melanoma patients[9–11]. Response rates correlate to the mutational load of patients' tumours as shown for melanoma and lung cancer, demonstrating that neoantigens comprising such mutations play a crucial role in anti-cancer immunoreactivity[12–14]. Cancer genomics allows us to precisely determine the landscape of tumour-specific mutations from which such neoantigens may derive[15]. However, our knowledge about defined and clinically relevant tumour-specific antigens (TSA) presented by human leucocyte antigens (HLA) and recognizable by T cells is still very limited. Most efforts to define such antigens in humans and mice currently employ exome and transcriptome analyses followed by in silico epitope prediction and large-scale immunogenicity assays[16–19]. This approach results in many predicted peptide ligands, only few of which have proven to be immunogenic. Peptide ligands selected for therapeutic targeting by prediction may therefore not be clinically effective. Direct identification of neoantigens by tumour-infiltrating T cells is highly laborious and time-consuming[20], and therefore less suitable for clinical translation. There are few reports about the direct identification of neoantigens by the analysis of the tumour ligandome using mass spectrometry (MS) and subsequent matching with exome and transcriptome data[21–23]. Importantly, this approach resulted in the direct identification of therapeutically relevant TSA in two murine models[21,22]. However, so far mutated peptide ligands identified by MS were derived from analyses of monoclonal cell lines only[21–23], not representing the complex heterogeneity of native tumours. Thereby, especially those clonal mutations representing particular promising target antigens for prolonged tumour rejection[24] may be missed. Direct identification of neoantigens from native tumour tissue samples was so far impeded by limitations in sensitivity and bioinformatics. However, translated to human patients, this would represent a major advance for clinical translation of neoantigen-directed immunotherapies.

We hereby report on the application of our recently developed high sensitivity mass spectrometry workflow[25] to the analysis of 25 human native tumour specimens. We provide an unprecedented depth of the tumour-derived ligandome harbouring a broad spectrum of attractive tumour-associated antigens. Most importantly, we discover tumour-specific neoantigens in selected patients validated by the proof of potent patients' derived neoantigen-specific anti-tumour immune responses. Thus, these data demonstrate that high sensitivity MS is a powerful tool to identify neoantigens highly relevant for the development and optimization of personalized immunotherapies in patients with cancer.

Results

In-depth immunopeptidomics on native melanoma tissue samples. Tumour tissue samples from 25 melanoma patients (Supplementary Tables 1 and 2) were used for analysis of biochemically purified HLA class I and II binding peptides. In total, we performed 140 MS measurements of purified peptides by LC–MS/MS analysis (Supplementary Data 1) using a state-of-the-art mass spectrometer, followed by stringent bio(informatics) analyses in the MaxQuant environment[26]. We identify 95,662 unique peptide sequences with a false discovery rate (FDR) of 1% (Fig. 1a, Supplementary Data 2) and report in total 99,355 peptide forms. We discover 78,605 peptides in the HLA class I peptidome from 12,663 proteins and 15,009 in the HLA class II peptidome from 2,832 proteins. In addition, 2,048 peptides from 746 proteins are detected in both classes I and II peptidome samples. The large variability in the number of eluted peptides per patient is in agreement with the amount of eluted HLA complexes. We demonstrate this by showing significant positive correlation between the number of identified peptides in HLA class I peptidome and the amount of recovered beta-2 microglobulin (B2M) in each tissue (Supplementary Fig. 1a and b). Eluted peptides show the characteristic length distribution and the MS-data itself assigns to proper anchor residues of defined HLA allotypes as exemplarily shown for two patients in Fig. 1b,c using the Gibbs clustering approach[27]. Many of the longer peptides (up to 15 amino acids) identified among the eluted HLA class I peptides still show the typical anchor motifs and are therefore likely binders and not contaminants (Fig. 1b,c). Another common approach used to assess purity and overall performance of elution of HLA peptides is the estimation of the affinity of the eluted peptides to the respective HLA molecules by predicting binding affinities[25,28]. This analysis, however, depends very much on the performance of the prediction programs. We predicted the binding affinity of eluted HLA-I peptides from patients Mel15 and Mel16. Due to the difficulty in assignment of peptides with multiple potential restrictions, we filtered the list of peptides to include only 9-mer peptides that bind to only one defined HLA allotype according to the minimum predicted affinity. Instead of using the 500 nM threshold commonly used for peptide binding prediction, we set the threshold for a binding as rank $<2\%$ (standard settings in NetMHC4.0). Using our dataset, we observed that a considerable amount of peptides that was assigned as HLA-B3503 binders fit the binding motif (Pro and Ala in the second position and Leu in the last position). In contrast, the predicted affinities of these binders (rank $<2\%$) were extremely low, with median predicted affinity of 2,806 nM ($n=581$) (Fig. 1d,e). These results differ substantially from analysis of HLA-I peptides assigned to HLA-B0702, an allotype with a rather similar motif, for which we observed a median predicted affinity of 17.7 nM of associated peptides ($n=1,191$) eluted from the tumour of patient Mel16.

Peptide ligands derived from tumour-associated antigens. The depth of the ligandome analysis is demonstrated by the identification of a large number of both known and novel peptide ligands derived from described melanocyte-associated differentiation and cancer testis antigens (Fig. 2; Supplementary Data 3). We detected the highest number of TAA-derived peptide ligands for the well-known melanoma-associated antigen PMEL (gp100), (64 HLA I and 46 HLA II ligands) a few of which were

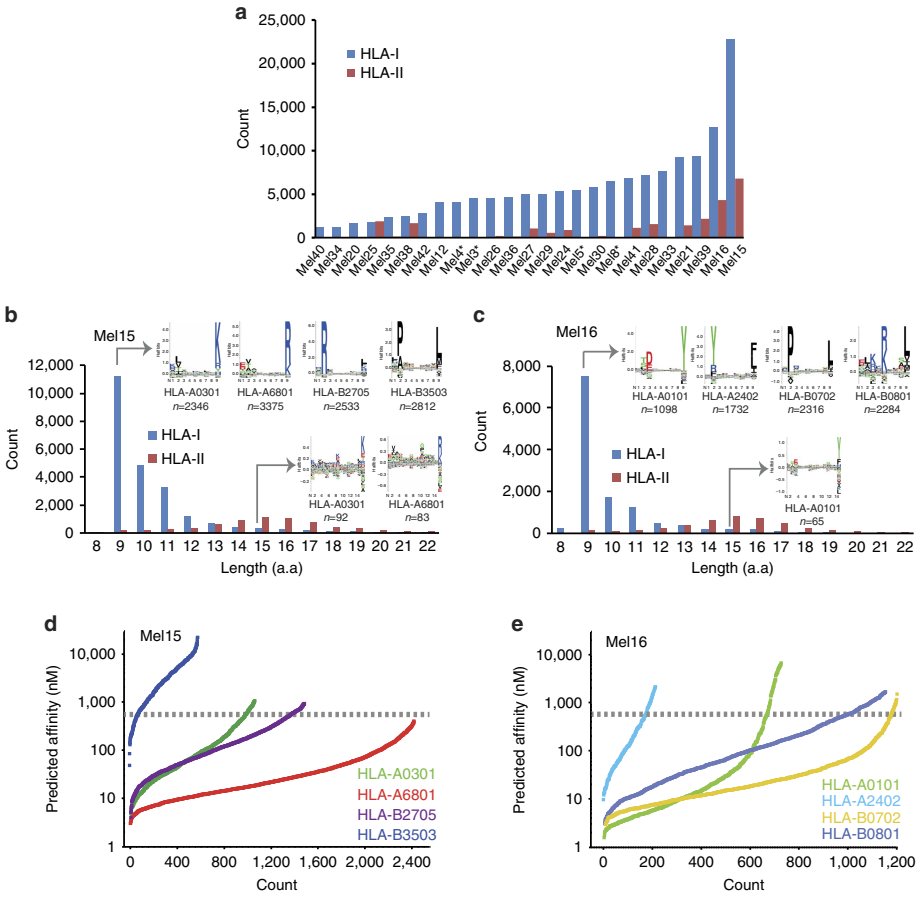


Figure 1 | In depth analysis of the melanoma-associated ligandome. Number of epitopes identified per patient. Asterisk marks samples for which no HLA-II peptidomics have been performed (**a**). Typical length distribution of eluted HLA-I and HLA-II peptides in Mel15 and Mel16. HLA-I peptides clustered to reveal the main binding motifs that fit the patients' HLA type (Supplementary Table 2) (**b**,**c**). Predicted affinity of eluted 9-mer peptides from Mel15 and Mel16 using NetMHC (**d**,**e**) using the threshold of top 2% ranked predicted sequences. The grey line represents the 500 nM threshold of binding affinity.

detected in both classes I and II peptidome samples (Fig. 2a,b). PMEL-derived peptide ligands were distributed over the entire protein but showed hot spot sequences presented by a large number of patients (Fig. 2b). We normalized for each patient the number of peptides derived from three selected TAA (PMEL, tyrosinase and PRAME) to the total amount of eluted peptides from the respective patient and correlated peptide presentation to RNA and protein expression of the defined TAA (Fig. 2c). These analyses resulted in a statistically significant correlation in case of PMEL (Fig. 2d–g). We observed again that, with respect to peptide prediction, many eluted peptide ligands have predicted binding scores of >500 nM according to NetMHC. Yet, they are still considered among the top 2% (ref. 29) (Supplementary Data 3).

Phosphorylated peptides are detected without enrichment. We performed a database search by enabling phosphorylation as a variable modification and although we did not specifically enrich for them, we detected a substantial fraction of phospho-HLA peptides within the eluted immunopeptidome. We filtered the list of identified phospho-HLA peptides by restricting the delta score to >15 and the localization probabilities to >0.75. After applying such stringent filters, we identified 365 phospho-HLA-I and 25 phospho-HLA-II peptides (Supplementary Data 4). About a quarter of the phospho-HLA binding peptides are shared among at least two patients and 6% of the phospho-HLA peptides have been identified independently in four or more patients' tumour samples (Fig. 3a). One third of the sites have not been previously described in the PhosphoSitePlus database[30] (Fig. 3b). Additional relevant information with respect to these sites is provided in Supplementary Data 4. We observed phosphorylation in 78% of the peptides on Serine and in 19% of the peptides on Threonine. The remaining 3% are on Tyrosine (Fig. 3c). To independently check the accuracy of identification of the phospho-HLA peptides, we synthesized 10 of them and all produced identical MS-fragmentation patterns as compared with the patient derived peptides (Supplementary Figs 2–11). To determine the position of these phosphorylations, HLA-I peptides were grouped according to their length, as presented in Fig. 3d. Interestingly, independent of the broad spectrum of HLA allotypes, the modification is most prominent on the fourth position of 9–11 mer HLA-I peptides and on the fourth and sixth

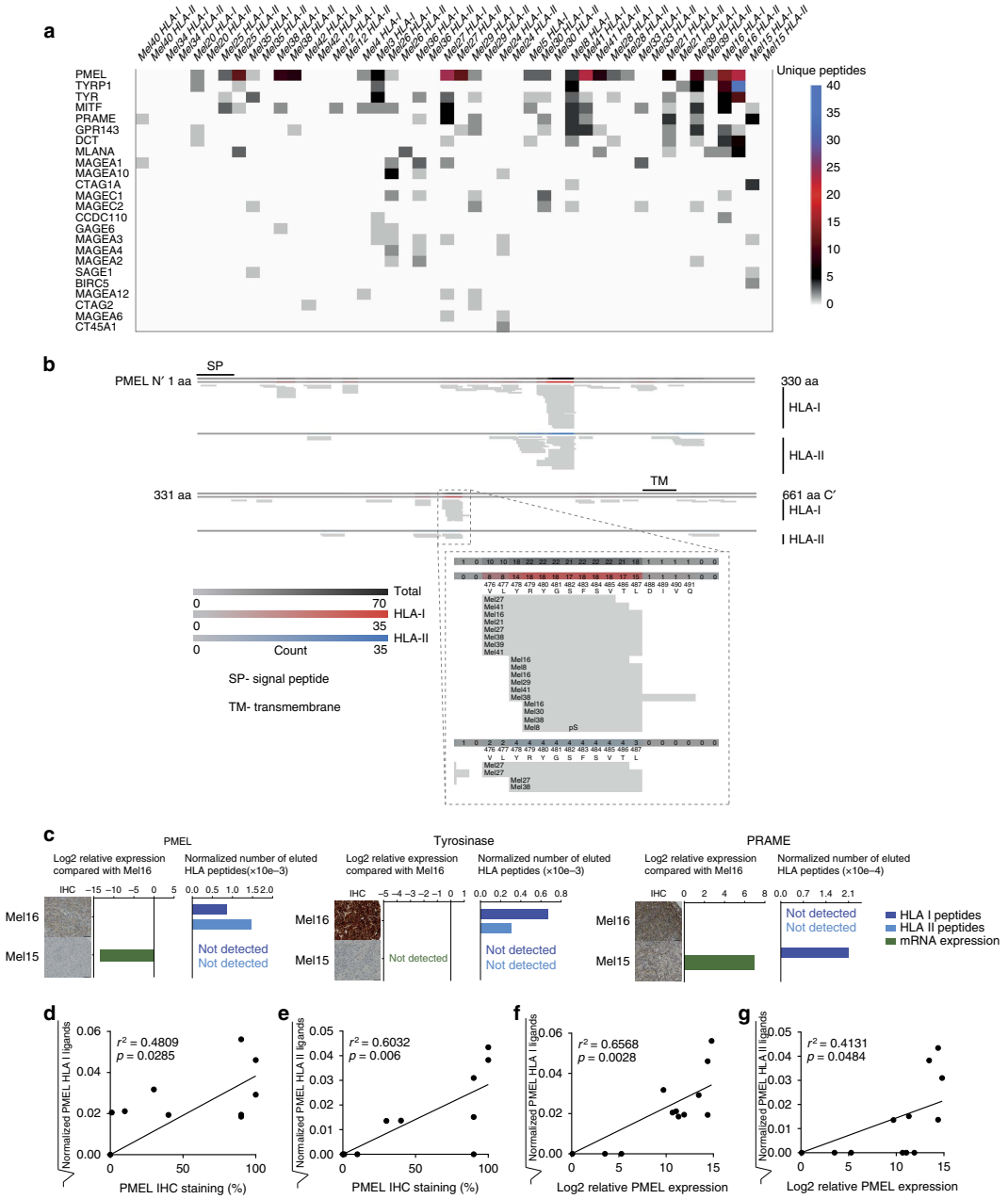


Figure 2 | Presentation of common tumour-associated antigens. Heat map presentation of the number of epitopes per patients that derived from a panel of 24 melanoma antigens (**a**). Alignment of the 99 epitopes from PMEL reveals hot spots presented by several patients in common (**b**). Expression of PMEL, tyrosinase and PRAME on mRNA and protein level is exemplarily compared with the number of HLA ligands for patients Mel15 and Mel16 normalized to the total number of identified HLA ligands. mRNA expression is depicted as log2 relative expression as compared with Mel16. Scale bars of Mel16-PMEL, Mel15-Tyrosinase, Mel15- and Mel16-PRAME: 50 μm; scale bars of Mel15-PMEL, Mel16-Tyrosinase: 100 μm (**c**). The number of PMEL-derived HLA class I or II ligands identified by immunopeptidomics was normalized to the total number of identified HLA ligands in the respective patient sample. Square root transformation was applied to deal with deviations from a normal distribution (**c**). Normalized HLA ligand numbers of PMEL of 12 patients with > 2000 HLA I ligands are plotted against per cent positive cells per tissue section as determined by IHC (**d**,**e**) or against log2 fold mRNA expression relative to a tissue panel consisting of 20 human tissues (**f**,**g**). Pearson correlation was calculated and the respective *p* value was corrected for multiple testing. For visual guidance a regression line is depicted on each panel.

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms13404 ARTICLE

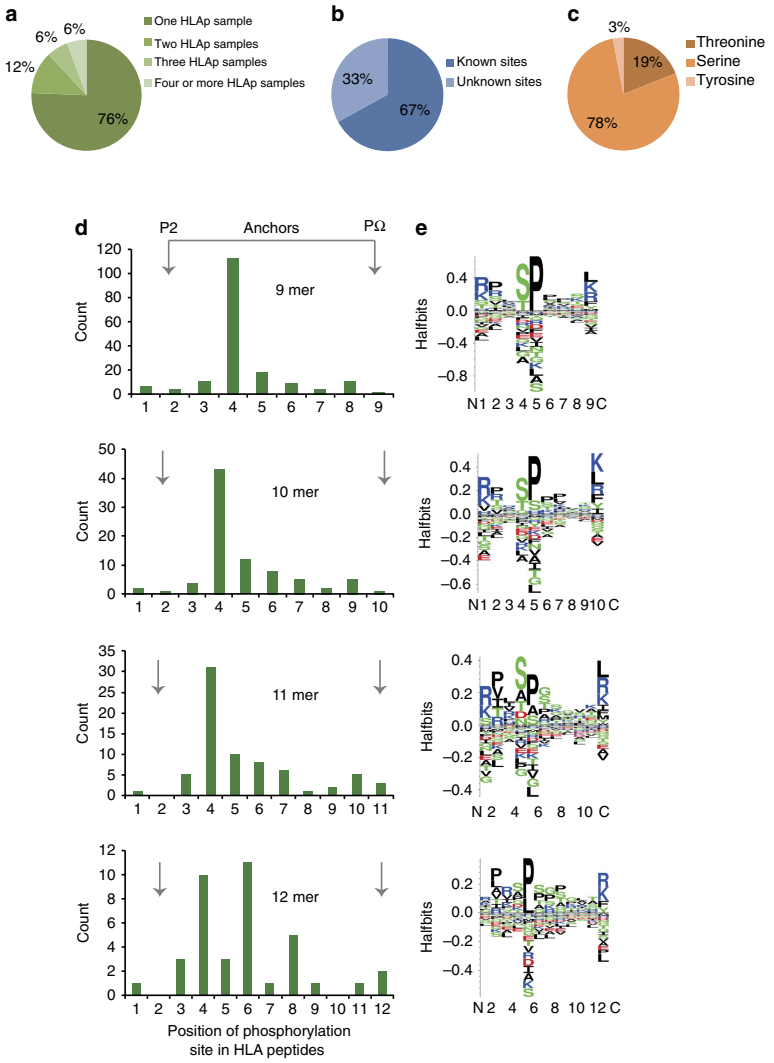


Figure 3 | Characterization of phosphorylation on eluted HLA peptides. Percentage of phospho-HLA peptides commonly detected in two or more patients (**a**) as well as percentage of known phosphorylation sites deposited on the PhosphoSitePlus database (**b**). Percentage of defined amino acids affected by phosphorylation within the phosphopeptide ligandome (**c**). Position of phosphorylation within the eluted phospho-HLA peptides according to the peptide length, from 9 mer to 12 mer peptides (**d**). Logo plots of residue frequency at each position of phospho-HLA peptides according to their length (**e**).

positions of 12 mer peptides (Fig. 3d). Moreover, we observed a preferential usage of Arginine and Lysine on position 1 (Fig. 3e) as previously reported for phospho-HLA peptides[31,32]. Of note, a clear signature of proline-directed phosphorylation is apparent in the sequence logos of the phospho-HLA binding peptides (Fig. 3e) as was reported before[32]. These features therefore seem to be rather HLA allotype-independent and make them attractive to be tested as common target antigen candidates in a broader patient population. Taken altogether, our direct approach provides data about a large melanoma-associated phosphopeptide ligandome potentially attractive for targeted immunotherapies.

Direct identification of mutated peptide ligands by MS. To test if our method provides the depth to identify peptide ligands possibly comprising mutations, as well as validating them as neoantigens (Fig. 4a), we first performed exome sequencing of the DNA extracted from five patients' tumours exemplarily selected due to variable responses to immune checkpoint modulation. Detailed information about patients and the course of disease is provided in Supplementary Table 2 and Supplementary Fig. 12. Stringent somatic single nucleotide variant (SNV) calling was conducted to define each patient's mutational load and to mimic the state-of-the-art approach for neoepitope prediction (Fig. 4b and Supplementary Data 5). Mutations previously known in

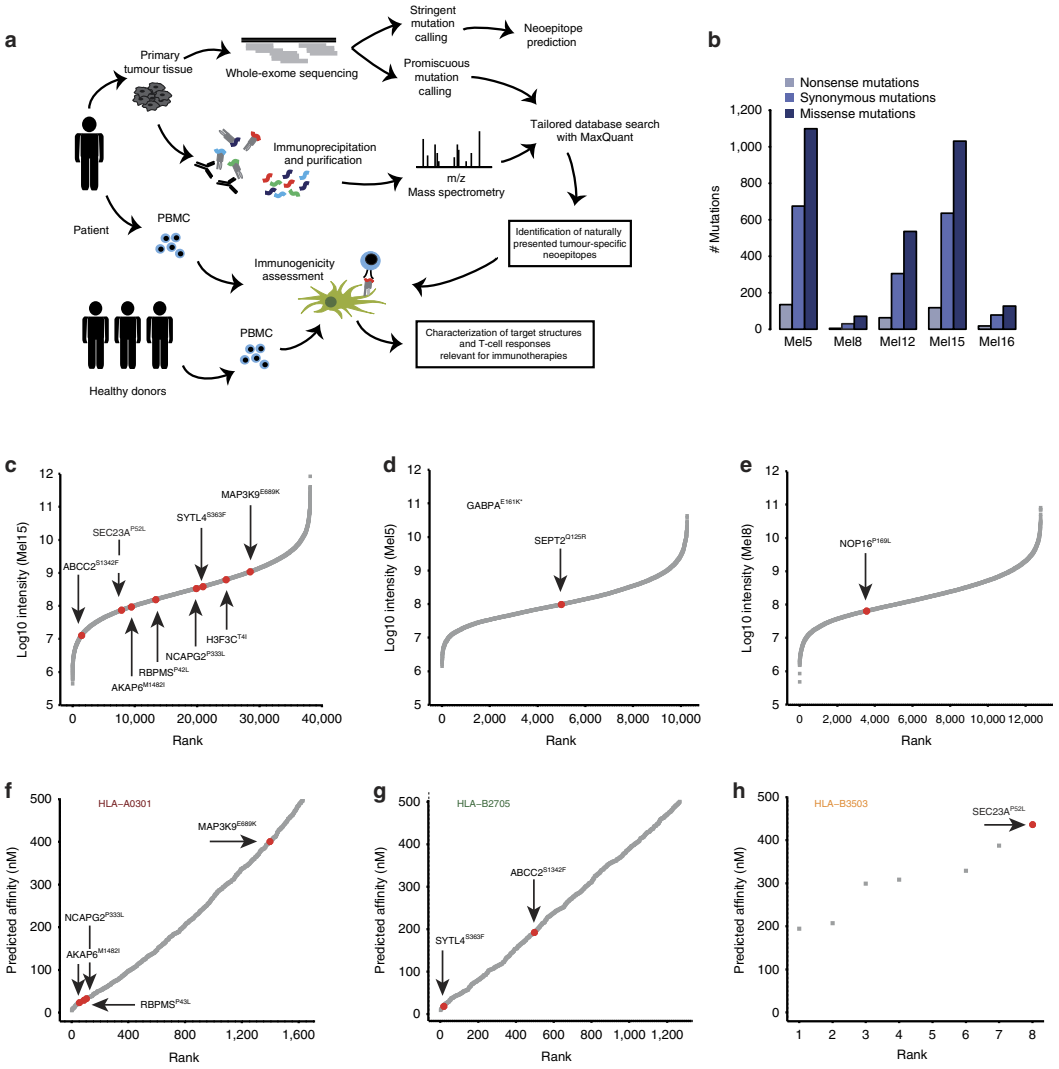


Figure 4 | Identification of mutated peptide ligands by matching exome sequencing and mass spectrometry immunopeptidomics. Overview of the experimental approach. Patient tumour tissue was used for MS analysis and exome sequencing. Mutations were called and matched with MS data. Mutated peptide ligands were then further evaluated for recognition by patient's autologous and matched allogeneic T cells (**a**). Overview of the number of non-synonymous and synonymous mutations per patient (**b**). Ranked intensity values of MS data derived from the immunopeptidome of the three patients with identified mutated peptide ligands (Mel15, Mel5 and Mel8). Positions of the mutated peptide ligands are projected on the curve (**c**–**e**). GABPA^{E161K} was detected at the MSMS level only, therefore no intensity is reported (**d**). Predicted affinity of neoantigen candidates using the 500 nM threshold for binders using NetMHC and ranking of neoepitope candidates for Mel15 with respect to HLA-A0301 ($n=1,632$), HLA-B2705 ($n=1,265$) and HLA-B3503 ($n=8$). The mutated peptide ligands detected by MS are not among the top 10 for HLA-A0301 or - B2705 (**f**-**h**).

selected tumour probes such as cKIT (P10721.1, L576P; Mel8) and BRAF (P15056.1, V600E; Mel16) were detected by this analysis (Supplementary Data 5). Of note, the mutational load among the five patients differed substantially and neither correlated with the number of identified mutated peptide ligands ($r=0.65$, 95% CI: −053 to 0.98) nor with the response to immune checkpoint modulation ($r=-0.03$, 95% CI: −0.89 to 0.88) (Fig. 4b, Supplementary Fig. 12 and Supplementary Data 5). In parallel, we developed a new module in the MaxQuant software that performs mutations calling from NGS data and generates a customized personalized reference database containing all protein isoforms where a detected SNV alters the amino acid sequence. We then performed a non-stringent mutation calling to avoid loss of SNV during database search. This resulted in a high number of non-synonymous mutations in all patients (>15,000 per tumour sample). We searched the raw MS data from the 5 selected patients against this database and, for the first time, directly identified 11 peptide ligands harbouring

Table 1 | List of 11 mutated peptide ligands identified by MS-based immunopeptidomics from human melanoma tissues.

Gene name	Sequence (Position)	a.a Alt	HLA allele predicted affinity nM;%rank;bindLevel	Chr position ENSMBEL transcrip ID	Patient	FDR	Reads tumour Ref:Alt	Reads PBMCRef:Alt	Comments
SYTL4	GRIAF**F**LKY (358-366)	S363F	HLA-*B* 27:05* 18.43; 0.6; SB	ChrX:100687163 ENST00000263033	Mel15	1%	29:9	51:1	WT HLAp GRIAFSLKY detected in Mel15
RBPMS	R**L**FKGYEGSLIK (45-56)	P46L	HLA-*A* 03:01* 29.2; 0.15; SB	Chr8:30474849 ENST00000517860	Mel15	1%	63:18	122:0	WT HLAp RPFKGYEGSL; RPFKGYEGSLI; RPFKGYEGSLIKL detected in Mel8
SEC23A	**L**PIQYEPVL (52-60)	P52L	HLA-*B* 35:03* 436.2; 0.01; SB	Chr14:39095964 ENST00000307712	Mel15	1%	36:9	34:0	
H3F3C	R**I**KQTARK (3-10)	T4I	HLA-*A* 03:01* 1614; 3.0; --	Chr12:31792156 ENST00000340398	Mel15	5%	48:6	63:0	
NCAPG2	K**L**ILWRGLK (332-340)	P333L	HLA-*A* 03:01* 32.6; 0.15; SB	Chr7:158680743 ENST00000409339	Mel15	1%	130:23	107:1	
AKAP6	KLKLP**I**IMK (1477-1485)	M1482I	HLA-*A* 03:01* 23.3; 0.1; SB	Chr14:32822259 ENST00000280979	Mel15	1%	56:20	108:0	
MAP3K9	ASWVVPIDI**K** (680-689)	E689K	HLA-*A* 03:01* 400.9; 1.2; WB	Chr14:70733760 ENST00000555993	Mel15	5%	24:6	41:0	
ABCC2	GRTGAGKS**F**L (1334-1343)	S1342F	HLA-*B* 27:05* 192.9; 0.7; WB	Chr10:99845661 ENST00000370449	Mel15	5%	27:10	50:0	
NOP16	SPGPVKLE**L** (161-169)	P169L	HLA-*B* 07:02* 26.3; 0.12; SB	Chr5:176384171 ENST00000621444	Mel8	5%	80:11	90:0	
GABPA	ETS**K**QVTRW (158-166)	E161K	HLA-*A* 25:01* 3231.1; 0.40; SB	Chr21:25752162 ENST00000354828	Mel5	5%	17:22	87:0	WT HLAp ETSEQVTRW detected in Mel5 and Mel40
SEPT2	YIDE**R**FERY (121-129)	Q125R	HLA-*A* 01:01* 6.0; 0.01; SB	Chr2:241337414 ENST00000391973	Mel5	5%	107:77	148:0	WT HLAp YIDEQFERY detected in Mel3, Mel5, Mel8, Mel12, Mel16, Mel25, Mel26, Mel38, Mel39 and Mel40

Mutated amino acids within the peptides are indicated with bold letters.

mutations from primary human cancer tissues (Table 1, Supplementary Data 6 and Supplementary Table 3). The mutated peptide ligands have different intensity ranks in the patients' specific tumour ligandomes, and most are within the second and third quartiles of the intensity distribution (Fig. 4c-e). Eight mutated peptide ligands have been identified in the tumour of one single patient (Mel15), a tumour sample for which a large peptidome has been discovered. For this patient, four tumour-derived tissue sections have been processed in parallel for the elution of peptides that afterwards were measured sequentially. Most mutated peptide ligands from patient Mel15 were independently identified in several MS measurements, supporting that they are well detected and well presented in the tumour of this patient. Specifically, SYTL4^{S363F}, RBPMS^{P46L}, SEC23A^{P52L}, MAPK3K9^{E689K} and H3F3C^{T4I} were identified in all four tissue probes, while NCAPG2^{P333L} and AKAP6^{M1482I} were detected in three and two probes, respectively. We synthesized peptides for all HLA ligands representing mutations and found their MS/MS spectra and elution times to be identical to the endogenous ones (Supplementary Fig. 13–24 and Supplementary Table 3). Of note, all of the somatic mutations of the 11 neoepitopes were also detected by the stringent SNV calling. In some cases, we detected the wildtype (wt) peptides, either in the same sample, like wt SYTL4 in Mel15 and wt GABPA and wt SEPT2 in Mel5, or in several other patients' samples, for example GABPA and SEPT2 (Table 1). This might indicate that they are located within hot spots for HLA peptide biogenesis, and since the peptides have been purified from a tissue that contains also healthy cells to a variable degree, presentation of the wt peptides is expected.

Comparison of identified mutated to predicted peptides. Prediction of neoepitopes currently represents the standard method to identify mutated peptide ligands potentially representing suitable targets for immunotherapy. To investigate the ranking of our peptides according to standard prediction algorithms, we applied NetMHC (ref. 29) for identification of potential HLA class I-predicted nonamer neoepitopes on all non-synonymous mutations identified by exome sequencing in the tumour probe of patient Mel15. A standard threshold of <500 nM as predicted affinity was set. We then ranked the predicted affinity and projected the ligandome-based identified mutated peptides on the curve (Fig. 4f-h). Notably, none of the identified mutated peptide ligands were within the top 10 candidates for the HLA allotypes HLA-A0301 (best candidate AKAP6^{M1482I}, rank 55) and HLA-B2705 (best candidate SYTL4^{S363F}, rank 18) (Fig. 4f-h) for which thorough database information is available. In the case of HLA-B3503 (Fig. 4h), prediction was again highly limited with only eight neoantigens predicted to bind to this HLA allotype.

Characterization of autologous neoantigen-specific T cells. We next asked if the MS-detected mutated peptide ligands represent neoepitopes that can be recognized by the patient's own T cells. We selected patient Mel15 as for this patient diverse mutated

peptide ligands were identified and miscellaneous biomaterial could be collected. The detailed clinical course including biomaterial collection of the patient is shown in Fig. 5a. For the investigation of recall responses, we stimulated unfractionated PBMC derived from diverse venipunctures in the course of the disease following application of Ipilimumab (Fig. 5a,b). Without any further enrichment, we identified defined T-cell responses by ELISpot as early as two days after stimulation of PBMC (Fig. 5c). Notably, specific responses were repeatedly observed against SYTL4^{S363F} at that early time point (Fig. 5c). Prolonged peptide stimulation of PBMC derived from diverse blood venipunctures resulted in expansion of T cells with specificity for SYTL4^{S363F}, as well as NCAPG2^{P333L} but not for other peptides (Fig. 5d). Dynamic courses of specific responses observed against these two peptides indicate a decline of specific T-cell responses over time. The quality of T-cell responses against SYTL4^{S363F} was superior compared with NCAPG2^{P333L} as shown by higher frequencies of T cells with dual cytokine secretion (Fig. 5e,f). Of note, wt peptides were not recognized (Fig. 5e,f). Specificity of defined T-cell lines was further confirmed by multimer staining of NCAPG2^{P333L}–specific T cells (Fig. 5g). In case of T-cell line PBMC-SYTL4-740, we were able to isolate a specific clone, PBMC-SYTL4clone1, which recognized endogenously processed mutated but not wt peptide after minigene transfer of respective gene sequences of SYTL4 (Fig. 5h).

Two years after application of Ipilimumab, a remaining single lung metastasis progressed and was removed at day 796 (Fig. 6a–d). Interestingly, this metastasis showed areas with vital tumour cells with intensive PD-L1 expression while high T-cell infiltration was apparent only in adjacent tumour areas (Fig. 6c,d). PD-L1 may be predominantly responsible for T-cell exclusion and tumour evasion in this case. The defined SYTL4^{S363F} mutation was detected on genomic DNA, as well as reverse transcribed coding DNA (cDNA) level in this second biopsy (Fig. 6e). Importantly, peripheral blood-derived T-cell lines with specificity for SYTL4^{S363F} from day 740 recognized freshly removed tumour material (Fig. 6f). Moreover, *ex vivo* expanded tumour-infiltrating T cells (TIL) exclusively recognized SYTL4^{S363F} but not other mutated peptide ligands (Fig. 6g). SYTL4^{S363F}-specific TIL-derived T-cell responses were functionally sorted and cloned resulting in expansion and further characterization of T-cell clone TIL-SYTL4clone1. Peptide titration of SYTL4^{S363F} revealed a functional avidity in the nanomolar range but no reactivity against the wt peptide (Fig. 6h). Specificity of TIL-SYTL4clone1 was further confirmed by recognition of endogenously processed mutated but not wt peptide as investigated by cytokine release and cytotoxicity (Fig. 6i).

Validation of neoantigens in the matched allogenic setting. To investigate if mutated peptide ligands may be immunogenic in matched healthy donors, we stimulated naïve T cells isolated from different donors with mutated peptide ligands. We identified additional reactivity against two peptides, AKAP6^{M1482I} derived from Mel15 and NOP16^{P169L} derived from Mel8 (Fig. 7a). An expanded T-cell line, HD1-AKAP6, with specificity for AKAP6^{M1482I} was further characterized. We observed specific binding of respective multimer but not wt multimer (Fig. 7b). In contrast, peptide titration experiments showed recognition of the mutant but also wt peptide, the latter with reduced functional avidity (Fig. 7c). Functional quality of T-cell responses against wt and mutated peptides were additionally investigated in detail with respect to heterogeneous cytokine release and cytotoxicity (Fig. 7d,e). Therefore, target cells either pulsed with defined peptides or transduced with minigenes were used. Of note, cytokine responses against wt peptide were inferior when compared with the mutated counterpart whereas the cytotoxic responses were comparable (Fig. 7d,e).

Discussion

We hereby present for the first time integrative classes I and II immunopeptidomes of native melanoma tissue samples resulting in the identification of almost 100,000 peptide ligands naturally presented on the tumour. With our methodology >95% of the peptides fit the binding motifs of the different HLA-I allotypes[25], supporting the high yield and purity of the eluted peptides. Also, among the long HLA-I peptides, many seem to fit well to the distinct binding motifs as shown for Mel15 and Mel16. This is in concord to other reports about long HLA-I binders[33–35]. We hypothesize that identical peptides that have been detected in both the class I and II peptidome may be related to common cellular processing which need to be tested in future studies[36,37]. The depth of the ligandome is highlighted by the large number of both, known and novel peptide ligands derived from previously described tumour and melanoma-associated antigens like PMEL, tyrosinase, MELAN-A, NY-ESO-1 and several proteins of the MAGE superfamily of cancer testis antigens. In case of PMEL, from which we detected almost 100 different peptide sequences, the magnitude of presentation, estimated by the number of unique peptide ligands per peptidome sample correlated with messenger RNA (mRNA) and protein expression. Alignment of the PMEL derived HLA peptides on the PMEL protein sequence revealed that several domains along the protein sequence are sources of multiple class I and II peptides in several tumour samples derived from diverse patients. We collectively name such domains as 'hot spots'. Other domains may not be as efficiently accessible for the antigen processing and presentation machinery and those were either not presented at all by any of the 25 studied melanoma tumours, or their resulting peptides were below our detection limit. Targeting of PMEL by peptide-based vaccination showed only limited clinical success when compared with results of checkpoint modulation[4] and combination of anti-CTLA-4 treatment with PMEL vaccination did not enhance anti-tumour activity[9]. We hypothesized that our large dataset might shed light on the extent these peptides are presented *in vivo*. Interestingly, the two nonameric peptides, P209 and P280, used previously for vaccination were eluted only from tumour probes of Mel27. In the case of P209, the nonamer peptide sequence is indeed located in the most dominant hot spot for presentation, although the sequence was, with exception of Mel27, included in peptide ligands with a length >14 aa. In the case of P280, the peptide sequence could be detected only in Mel27. These data suggest other PMEL-derived peptides to be potentially more promising for defined targeting in a larger patient cohort.

Even without further enrichment, peptide ligands harbouring posttranslational modifications as phosphorylation were detected. This implies for the high recovery and sensitivity of our method and importantly it avoids the requirement of reservation of dedicated samples for enrichments of phospho-peptides and additional laborious sample processing[38]. Nevertheless, such peptides may contain cancer-specific phosphorylation patterns and therefore potentially represent attractive targets for cancer immunotherapy[31,32]. One third of identified phosphorylation sites have not been reported in the PhosphoSitePlus database[30]. We envision that direct immunopeptidomic analyses have the potential to identify novel sites on protein sequences that may not be compatible with trypsin digestion and therefore may be undetected by shotgun phospho-proteomics[39]. 24% of the phospho-HLA peptides have been identified in tumour samples

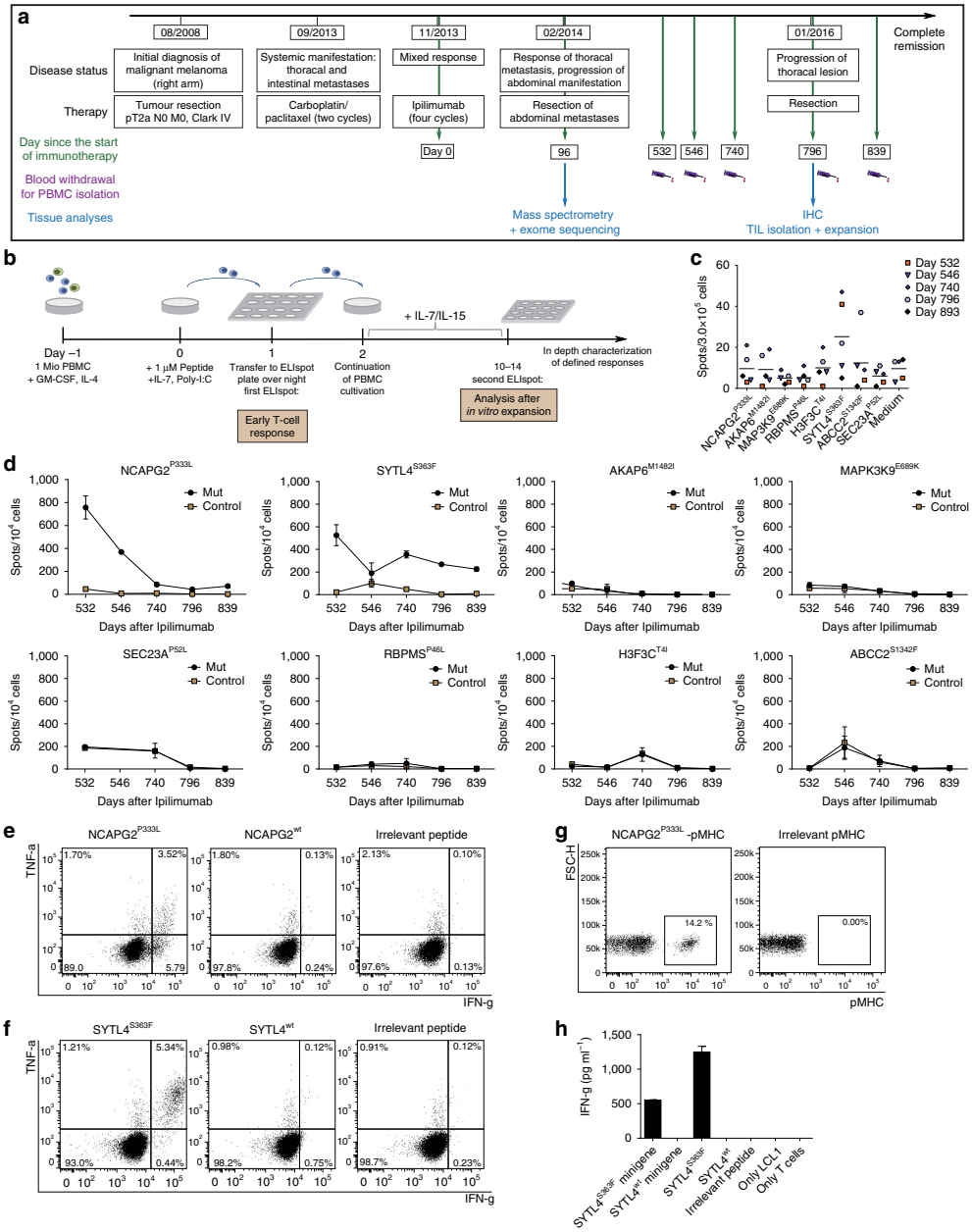


Figure 5 | Immune responses against mutated ligands in PBMC of patient Mel15. Clinical course and retrieval of patient material (**a**). Schematic overview of the experimental design of recall immune responses among blood-derived T cells from patient Mel15 (**b**). Early immune responses detected in PBMC derived from different blood withdrawals two days after *in-vitro* peptide stimulation (**c**). Time course of specific reactivities of blood-derived PBMC obtained at different time points against the eight identified mutated epitopes from patient Mel15. All analyses were performed in duplicates and spot counts were adjusted to 10^4 cells (**d**). Intracellular cytokine staining (ICS) of an expanded NCAPG2^{P333L} specific T-cell line from day 546 (PBMC-NCAPG2-546) after co-incubation with peptide pulsed T2-A3 target cells for 5 h (**e**). ICS of T-cell line PBMC-SYTL4-740 stimulated with SYTL4^{S363F} from day 740 after co-culture with peptide pulsed T2-B27 target cells (**f**). Staining of line PBMC-NCAPG2-546 with the specific multimer in comparison to irrelevant multimer staining (**g**) IFN-g secretion after coincubation of T-cell clone PBMC-SYTL4clone1 derived from line PBMC-SYTL4-740 with peptide pulsed and minigene-transduced LCL1 (results of triplicates) (**h**). Data from experiments with triplicates are shown as mean ± s.d., data resulting from duplicates are shown as mean.

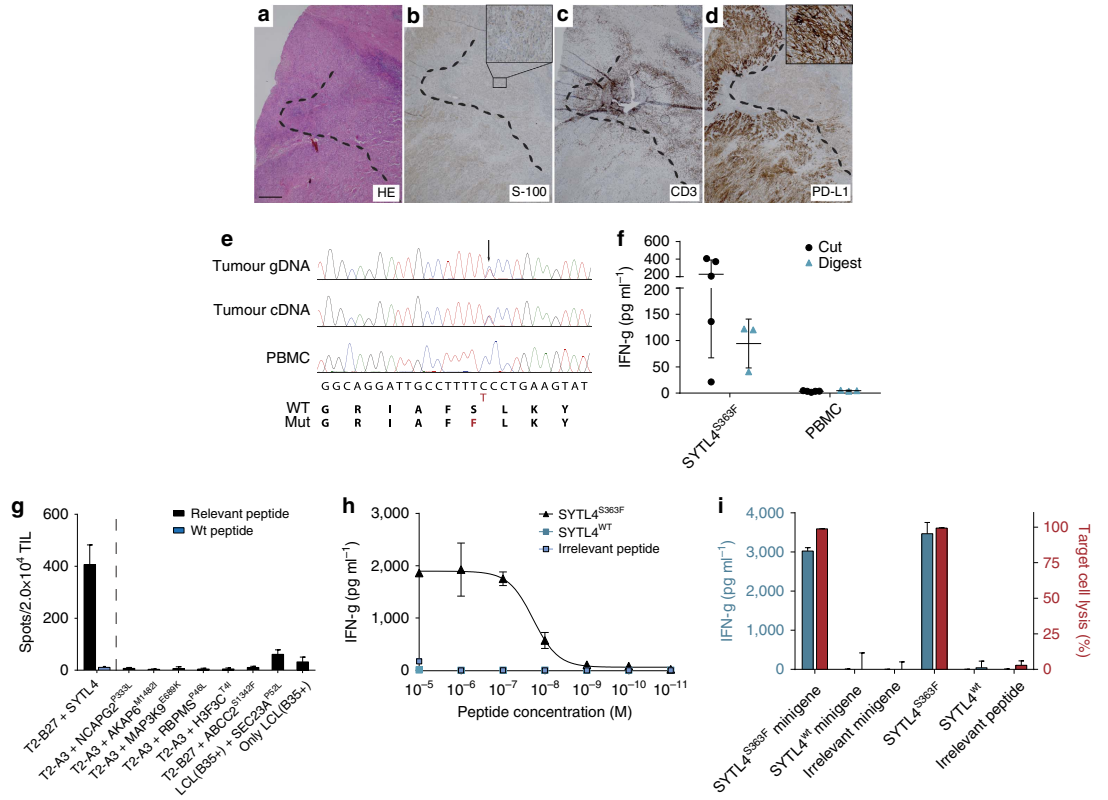


Figure 6 | In-depth characterization of tumour and peptide-reactivity of SYTL4-specific T cells derived from PBMC as well as TILs. HE (**a**) staining of a lung metastasis after metastasectomy (01/2016, day 796) as well as immunohistochemistry stainings with anti-S100 (**b**), anti-CD3 (**c**) and anti-PD-L1 (**d**); (**b**,**d**) Inset: × 20 magnification. Scale bar, 500 µm. Sanger sequencing of the mutated region from SYTL4 in processed tumour material from day 796 using either isolated genomic DNA (gDNA) or coding DNA (cDNA) as template (**e**). IFN-g secretion of T-cell line PBMC-SYTL4-740 on co-culture with cut (5 wells) or digested (3 wells) fresh tumour material for 36 h (**f**). Non-stimulated PBMC from Mel15 served as controls. Horizontal lines and error bars show mean and s.d., respectively. Co-incubation of *in-vitro* expanded TIL with target cells pulsed with mutated peptide ligands (**g**). SYTL4^{wt} served as negative control, analysis was performed using triplicates and depicted as mean ± s.d. Reactivity of the TIL-derived T-cell clone TIL-SYTL4clone1 against T2-B27 target cells pulsed with titrated concentrations of mutated, wt or irrelevant peptide (**h**). Co-culture of TIL-SYTL4clone1 with LCL1 either peptide pulsed or transduced with mutated or wt minigenes (**i**) with results shown as mean of duplicates. Amount of IFN-g secretion was assessed in supernatants (left *Y*-axis) and amount of target cell lysis was analysed by FACS-based acquisition of total number of vital target cells in relation to untreated LCL (right *Y*-axis); coincubation was performed in triplicates and results are shown as mean and s.d.

derived from multiple patients. A clear signature of proline-directed phosphorylation of the detected HLA peptides could be observed and this is likely to be assigned to a defined kinase motif associated to cell proliferation and tumorigenicity[32,40]. Thus, our data point to a common oncogenic phospho-peptide signature potentially attractive for multimodal targeting. Notably, our data revealed a very conserved motif within detected phospho-HLA peptides with preferred Arginine and Lysine in P1 and the phosphorylation site in P4. This canonical motif has been previously described for defined HLA allotypes and structural data suggest that the conserved amino acid usage in P1 may increase peptide binding of low affinity peptides whereas phosphorylation in P4 may improve immunogenicity by direct presentation of the phosphorylation site to the TCR or conformational peptide changes[31]. Our data indicate that such peptides are presented over a broad HLA repertoire making these peptides attractive to be tested as more general target antigens. Reactivity of patients' derived autologous T cells with specificity for these peptides might be limited due to negative thymic depletion of reactive T cells. However, TCR derived from the mismatched or xenogeneic repertoire may still represent attractive therapeutic tools to target self-antigens with cancer-specific expression in adults[41].

In contrast to shared TAA, mutated peptide ligands can be regarded as foreign antigens which have been previously described to be well detectable by autologous T cells in diverse disease settings[16,17,19,20,42–44]. Especially clonogenic ones have been shown to be associated to a durable clinical benefit by immune checkpoint inhibitors[24]. We hereby describe for the first time the identification of mutated peptide ligands derived from patients' non-modified and non-cultivated native tumour tissue samples using our discovery MS approach. Within this proof of concept, with testing of five patients, we detected 11 mutated peptide ligands, 8 of them in one patient. This

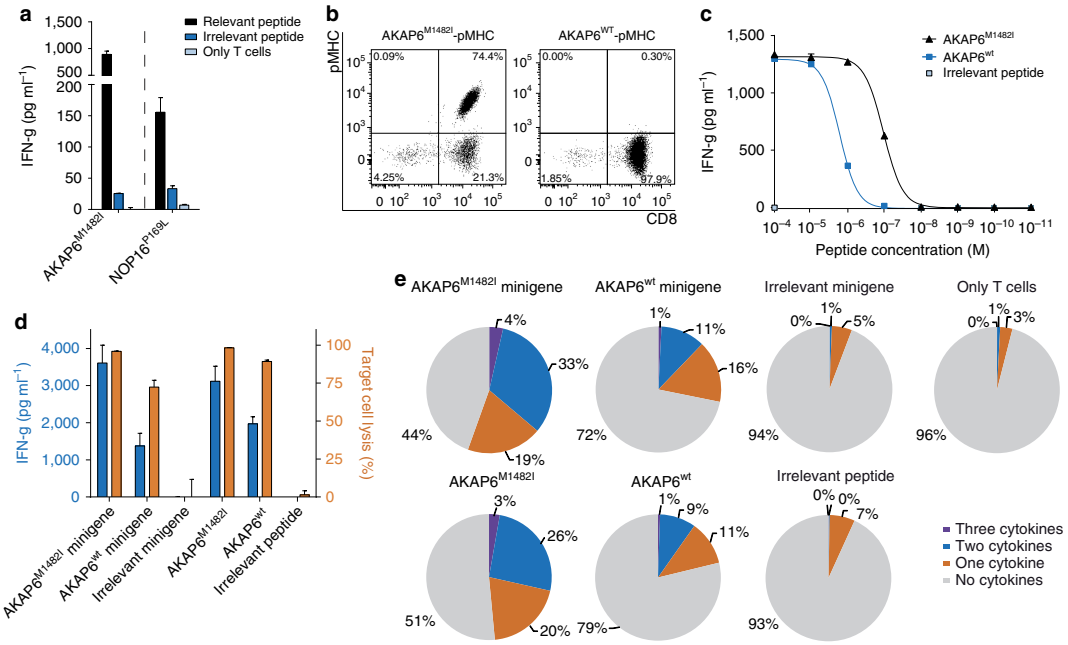


Figure 7 | Characterization of mutant-specific T-cell responses in HLA-matched healthy donors. T-cell responses of two different matched healthy donors against neoepitopes AKAP6^{M1482I} and NOP16^{P169L}. Effector cells were coincubated in duplicates with T2-A3 or T2-B7 pulsed either with the relevant peptide or control peptides with the same HLA restriction as the mutated ligands, results are shown as mean (**a**). Staining of T-cell line HD1-AKAP6 with the mutated or wt multimer (**b**). IFN-g release of the T-cell line on peptide titration of AKAP6^{M1482I} and its non-mutated counterpart using T2-A3 as targets (duplicates are depicted as mean) (**c**). IFN-g secretion (left Y-axis) and target-cell lysis (right *Y*-axis) after coincubation of the T-cell line HD1-AKAP6 with peptide-pulsed and minigene-transduced LCL1 cells performed in triplicates, data shown as mean ± s.d. (**d**). Intracellular cytokine staining (IFN-g, TNF-a and IL-2) on co-culture of the T-cell line HD1-AKAP6 with LCL1 cells, either peptide-pulsed or minigene-transduced, determined by flow cytometry. Cells were gated on ethidium monoazide bromide-negative and CD8-positive events (**e**).

patient experienced prolonged clinical benefit following the application of Ipilimumab, a checkpoint inhibitor involved in the enhancement of primary and memory T-cell responses[45]. Notably, response to treatment in melanoma has been associated to the mutational load, suggesting that mutated peptide ligands are the major source of target antigens[13]. Few mutated peptide ligands have been previously identified by MS using murine or human cell lines as raw material[21–23]. However, the direct identification of neoantigens from non-modified native human tumour tissue represents an important breakthrough for several reasons. Tissue, unlike cell lines, is heterogeneous. Although more challenging, mutated peptide ligands identified from this material are likely among the most well presented peptides and hence the best targets for immune interventions. In fact, most mutated peptide ligands identified from Mel15 were independently identified in several MS measurements of different tissue probes, and their measured MS intensity indicates adequate presentation similar to non-mutated self-antigens. MS-based detection is by nature biased to detect the more abundant peptides and hence will favour identification of clonogenic mutated peptides. In contrast, mutations present only in malignant subclones are likely to be under-represented in the total peptidome and therefore below the detection limit of the current MS-based discovery approach.

Most of the neoantigens we detected are predicted to bind with high affinity to their respective HLA molecules, although they are mostly not within the top 10 predicted ones. In case that amino acid alterations are located in anchor positions, novel ligands may be generated. However, in case that the alterations are not in anchor positions there is a fair chance that also the wt peptides will be detected by MS, especially if the peptides are expected to bind with high affinity. Indeed, we detected the corresponding wt peptides of mutated SYTL4 and GABPA in the peptidomes of Mel15 and Mel5, respectively. Interestingly, we also detected corresponding wt peptides, and also sequences that are shorter or longer versions of the core neoantigen sequences in HLA peptidome samples of other patients with alternative HLA allotypes as e.g. for RPBMS. Multiple peptide sequences homing to a certain location on the protein suggest that this might be again a hot spot for presentation. Thus, mutations that are included in hot spots may have preferred presentation *in vivo*, although larger studies are needed to confirm this hypothesis. However, if this hypothesis is correct, future large-scale immunopeptidomics studies are expected to reveal such hot spots in the human proteome, and consequently in-silico algorithms should prioritize neoantigen candidates that are included within them in order to shorten the target list.

The potential and promise of MS detection of neoantigens is in fact highlighted by the hit rate of mutated peptide ligands with obvious clinical relevance. In fact, two out of eight mutated peptide ligands were detected by blood and TIL-derived autologous T cells of patient Mel15. This indicates a clear advantage compared with the usage of prediction software tools to identify neoantigens[16,19,24,44,46,47]. Moreover, mutation-

specific T-cell responses could be detected as early as two days after peptide stimulation without prior enrichment as previously published[44,46]. Mutation-specific T-cell lines recognized freshly isolated tumour material further emphasizing clinical relevance of these neoantigens. Of note, neoantigen-specific responses were declining over time whereas a single remaining lung metastasis started to progress. These data are highly intriguing and suggest that mutation-specific T-cell responses might be investigated as personalized surrogate biomarkers in future studies. Two other mutated peptide ligands were recognized by matched allogeneic T cells and we suggest that such matched allogeneic T cells may be an attractive source to be used for adoptive T-cell or TCR transfer as recently published and suggested[43]. Interestingly, in-depth characterization of one defined T-cell population demonstrated reactivity also against the wt peptide. The distinct T-cell population was not detected by the wt multimer and wt-peptide-specific cytokine release was clearly inferior if compared with the mutated peptide. However, lysis of minigene-transduced or peptide-pulsed target cells presenting respective wt epitope reached high levels comparable to lysis of targets presenting the mutated ligand. These data suggest that differences in TCR avidity in response to mutated versus wt peptide may follow similar rules compared with observed differences in response to viral antigens[48]. Thus, multi-functional characterization of neoantigen-specific T cells is important to estimate risks of autoimmunity and neoantigen-specific matched allogeneic T-cell populations or TCR need to be carefully selected for adoptive transfer.

The time necessary for direct identification of mutated peptide ligands by MS using native tissue probes can be as short as three weeks, including whole exome sequencing analysis. Therefore, this approach is highly suitable for the development of personalized treatment approaches. In contrast, the usage of cell lines as raw material for MS analysis requires prolonged time for generation of a sufficient number of cells and is not always successful. Similarly, the prediction approach currently predominantly used for identification of presumable neoantigens is highly time consuming and expensive as subsequent large-scale immunogenicity testings are necessary for neoantigen validation. Moreover, the prediction approach harbours the risk for biased or limited results especially in case of binders to rare HLA allotypes. Our data indicate that the commonly used threshold for predicted affinity of 500 nM is by far too low for some HLA allotypes. Moreover, our data may be highly useful as to be a training dataset to further improve the performance of such predictors and consequently to enable more reliable in-silico identification of neoepitopes in the future.

The issue of the sensitivity of this discovery MS-based approach is currently the major limitation as indicated by the limited number of identified neoantigens. More than that, we could not detect neoantigens among the class II peptidome in this study. The latter result was expected as class II molecules are typically expressed on professional antigen presenting cells in the tumour microenvironment, and often not directly on the melanoma cells. Currently, unlike T-cell based assays, the MS approach is not sensitive enough to detect the few copies of neoantigens presented only on professional antigen presenting cells that are in the tumour microenvironment. A more intensive fractionation of the HLA peptides sample prior to the MS analysis may increase the depth but will also significantly increase the investment in MS measurement time. We envision that the new generations of MS instruments, new computational algorithms and more efficient procedures for sample preparations will further improve the sensitivity and therefore this direct antigen discovery approach.

The direct identification of clinically relevant antigens, both shared and private, will foster our understanding of essential characteristics of targets and their respective specific T cells relevant for effective tumour rejection and protection. Defined neoepitopes can be targeted by vaccination and respective T-cell responses can be tracked and used as biomarkers. Neoepitopes and defined common TAA not recognized by the patient's T cells may be attractive for alternative immunotherapeutic strategies such as transfer of effector cells with defined specificity.

Methods

Primary human material and cell lines. Informed consent of all healthy and diseased participants was obtained following requirements of the institutional review board (Ethics Commission, Faculty of Medicine, TU München). All patients included in the analysis were diagnosed for metastatic malignant melanoma and treated at the Klinikum rechts der Isar, TU München. An overview about all patients is given in Supplementary Table 1. More detailed information is provided for patients Mel5, Mel8, Mel12, Mel15 and Mel16 who additionally donated blood for isolation of PBMC and identification of mutated peptide ligands by matching the immunopeptidome with exome sequencing data (Supplementary Fig. 12, Supplementary Table 2 and Supplementary Data 5). Tumour tissue samples were collected from patients, who underwent tumour resection at the Department of Surgery, Klinikum rechts der Isar of the TU München. Immediately after resection (within 30 min), tumour tissue was macroscopically dissected by an experienced pathologist, snap frozen and stored in liquid nitrogen (− 196 °C) at the MRI-TUM-Biobank (MTBIO) until usage. Additional tumour tissue was formalin-fixed and paraffin-embedded (FFPE). Before molecular analysis, tumour diagnosis was confirmed by a pathologist and tumour content was determined by an HE stain taken from the sample going to be used. TIL derived from the tumour tissue of patient Mel15 removed at day 796 after treatment with Ipilimumab were expanded for 2–3 weeks by cultivation of minced tumour tissue pieces with irradiated feeder PBMC, 1,000 U ml^{-1} IL-2 (PeproTech, London, UK) and 30 ng ml^{-1} OKT3 (kindly provided by Elisabeth Kremmer). Change of medium supplemented with 300 U ml^{-1} IL-2 was performed twice a week. PBMC from patients and healthy donors were isolated from whole blood by density-gradient centrifugation (Ficoll/Hypaque, Biochrom, Berlin, Germany) immediately on receipt. T cells were cultivated in T-cell medium, RPMI 1640 (Invitrogen, Carlsbad, CA, USA) supplemented with Penicillin/Streptomycin (Pen/Strep) (PAA, Pasching, Austria), 5% FCS, 5% human serum, 10 mM Hepes (Invitrogen) and Gentamycin (Biochrom, Berlin, Germany), or serum-free AIM-V (Invitrogen) as indicated. Cell lines used in this study: T2 (American Type Culture Collection (ATCC), Manassas, VA), lymphoblastoid cell lines (LCL) LCL1 (IHW09005, HLA-A0301, B2705, C0102) and LCL2 (IHW09216, HLA-A0201, A0301, B3502, B3801 (both kindly provided by Steve Marsh). Morphology and constant growth behaviour of all cell lines were controlled periodically and the absence of mycoplasma infection was routinely confirmed by PCR (Venor GeM mycoplasma detection kit, Minerva Biolabs). T2 were retrovirally transduced with the HLA restriction elements HLA-A0301 (T2-A3), B0702 (T2-B7) or B2705 (T2-B27) as described before[49]. All target cell lines were maintained in RPMI 1640 supplemented with Pen/Strep and 10% FCS.

Purification of HLA peptides. For the preparation of the affinity columns, panHLA-I and panHLA-II antibodies were purified from HB95 cells and HB145 cells (ATCC, Manassas, VA), respectively. We cross-linked the antibodies to Protein-A Sepharose beads (Invitrogen, CA) with 20 mM dimethyl pimelimidate in 0.2 M sodium borate buffer pH9. Tumour amount that has been available for this research varied significantly, from about 0.1 g to 4 × 1 g in Mel15. For the purification of HLA complexes, snap-frozen melanoma tissue samples were homogenized for 10 s on ice using ULTRA-TURRAX (IKA, Staufen, Germany) in a tube containing 5–10 ml of lysis buffer and incubated at 4 °C for 1 h. The lysis buffer contained 0.25% sodium deoxycholate, 0.2 mM iodoacetamide, 1 mM EDTA, 1:200 Protease Inhibitors Cocktail (Sigma-Aldrich, MO), 1 mM PMSF, 1% octyl-β-D glucopyranoside (Sigma-Aldrich, MO) in PBS. The lysates were cleared by 20 min centrifugation at 40,000g. Lysates were passed through a column containing Protein-A Sepharose beads (Invitrogen, CA) to deplete the endogenous antibodies. Subsequently, HLA-I molecules were immunoaffinity purified from cleared lysate with the W6/32 antibody covalently bound to Protein-A Sepharose beads (Invitrogen, Camarillo, CA). HLA-II molecules were then purified by transferring the flow through onto similar affinity columns containing the HB-145 antibody. Affinity columns were washed first with 10 column volumes of 150 mM NaCl, 20 mM Tris–HCl (buffer A), 10 column volumes of 400 mM NaCl, 20 mM Tris–HCl, 10 volumes of buffer A again, and finally with seven column volumes of 20 mM Tris–HCl, pH 8.0. HLA molecules were eluted at room temperature by adding 500 µl of 0.1 N acetic acid, in total seven elutions for each sample[25].

Eluted HLA peptides and the subunits of the HLA complexes were loaded on Sep-Pak tC18 (Waters, MA) cartridges that were prewashed with 80% acetonitrile (ACN) in 0.1% trifluoracetic acid (TFA) and with 0.1% TFA. The peptides were separated from the much more hydrophobic HLA heavy chains and B2M on the

C18 cartridges by eluting them with 30% CAN in 0.1% TFA. They were further purified using a Silica C-18 column tips (Harvard Apparatus, Holliston MA) and eluted again with 30% ACN in 0.1% TFA. The peptides were concentrated and the volume was reduced to 15 µl using vacuum centrifugation. Remaining immunoaffinity purified HLA heavy chains and the B2M molecules were eluted from the Sep-Pak tC18 cartridges with 80% ACN in 0.1%TFA. For western-blot detection, 1% of each of those protein containing samples were used. Anti human B2M antibody EP2978Y (1:5,000, Abcam, Cambridge, United Kingdom) was used and was detected with donkey anti-rabbit IgG HRP conjugate secondary antibody (1:5,000, Thermo Fisher Scientific) in a peroxidase assay using SuperSignal West Pico Chemiluminescent substrate (Thermo Fisher Scientific).

LC–MS/MS analysis of HLA peptides. HLA peptides were separated by a nanoflow HPLC (Proxeon Biosystems, Thermo Fisher Scientific, Odense) and coupled on-line to a Q Exactive or the Q Exactive HF mass spectrometers (Thermo Fisher Scientific, Bremen) with a nanoelectrospray ion source (Proxeon Biosystems). We packed a 50 cm long, 75 µm inner diameter column with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) in 100% methanol. Peptides were eluted with a linear gradient of 2–30% buffer B (80% ACN and 0.5% acetic acid) at a flow rate of 250 nl min^{-1} over 90 min. Data were acquired using a data-dependent 'top 10' method, which isolated them and fragment them by higher energy collisional dissociation. We acquired full scan MS spectra at a resolution of 70,000 at 200 m/z with a target value of 3e6 ions. The most intense ions were sequentially isolated and accumulated to an AGC target value of 1e5 with a maximum injection time of 120 ms. For measurement of HLA-I peptides, in case of unassigned precursor ion charge states, or charge states of four and above, no fragmentation was performed. In addition, we excluded the fragmentation charge state of one from measurement of HLA-II peptides. The peptide match option was disabled. MS/MS resolution was 17,500 at 200 *m/z*. Fragmented *m/z* values were dynamically excluded from further selection for 20 s.

Synthetic peptides. Synthetic peptides for spectra validation were synthesized with the Fmoc solid phase method using the ResPepMicroScale instrument (Intavis AG Bioanalytical instruments, Cologne, Germany).

Mass spectrometry data analysis of HLA peptides. We employed the MaxQuant computational proteomics platform[25,26] version 1.5.3.2. Andromeda, a probabilistic search engine incorporated in the MaxQuant framework[50], was used to search the peak lists against the UniProt databases (Human 85,919 entries, Sep 2014), and a file containing 247 frequently observed contaminants such as human keratins, bovine serum proteins, and proteases. For identification of mutated peptide ligands, customized references databases were used (see below). N-terminal acetylation (42.010565 Da), methionine oxidation (15.994915 Da) and phosphorylation (79.9663304 Da on serine, threonine and tyrosine) were set as variable modifications. The second peptide identification option in Andromeda was enabled. The enzyme specificity was set as unspecific. Andromeda reports the posterior error probability and FDR, which were used for statistical evaluation. A false discovery rate of 0.01 was required for peptides for the global ligandome analysis and for the phospho-HLA peptides identification. We applied in addition a less stringent threshold of 5% for the identification of mutated peptide ligands. No protein false discovery rate nor permutation rules were set in MaxQuant in creating the decoy database. The initial allowed mass deviation of the precursor ion was set to 6 p.p.m. and the maximum fragment mass deviation was set to 20 p.p.m. We enabled the 'match between runs' option, which allows matching of identifications across different replicates that belongs the same patient, in a time window of 0.5 min and an initial alignment time window of 20 min. From the 'peptide.txt' output file produced by MaxQuant, hits to the reverse database and contaminants were eliminated. The resulting list of peptides is provided in Supplementary Data 2.

Variant search using MaxQuant. For the analysis of peptides we used a newly designed module of the MaxQuant software that enables the search for peptides based on genomic variations. MaxQuant takes as input aligned reads from exome data and calls variants as described below. The base search space results from unspecific digestion of all protein sequences utilizing all peptides from length 8 to 25. Variants increase the peptide search space by either including or excluding them on each peptide. In case several variants can be present on the same peptide all combinations of the absence/presence patterns are taken into account. In extreme cases of very many combinatorial possibilities for a peptide, these are cut off at 100 contributing peptides. To account for different a priori probabilities of different peptide classes the posterior error probability is calculated depending on the type of the peptide. For instance, different classes of peptides that are treated separately in the posterior error probability calculation are unmodified peptides without variants, unmodified peptides resulting from a variant, phosphorylated peptides without variants and phosphorylated peptides resulting from a variant. A common PSM-FDR threshold is applied based on this peptide class dependent posterior error probability.

DNA isolation from FFPE tissue. For isolation of genomic DNA (gDNA) from FFPE tissue, paraffin sections (five 10 µm sections per tumour sample) were de-paraffinized using xylene (2 × 5 min) and cleared in absolute ethanol. Tumour tissue was macro-dissected and DNA isolation was performed with DNeasy Blood & Tissue Kit (Qiagen/Hilden, Germany) according to manufacturer's instructions with following modifications: (i) tissue lysis was performed for an extended period of 60 h and (ii) Qiagen MinElute spin columns were used for a reduced elution volume of 50 µL.

Whole-exome sequencing and bioinformatics analysis. DNA fragmentation was performed with Covaris S2/E220 ultrasonicator to yield a fragment size of ~200 bp. The SureSelectXT Human All Exon V5 (Agilent Technologies, Santa Clara, USA) kit was used for library preparation. Sequencing (100 bp paired-end) was performed on the Illumina HiSeq 2000 system. Mutation calling was performed according to a promiscuous or stringent protocol. For promiscuous mutation calling, we excluded positions with quality <13 (equivalent to 0.05 error probability) and used the following thresholds: total read depth of the position should be >10 reads; number of reads which support a variant should be greater than 5 reads and at the same time the minimum variants frequency was set to 5 per cent.

Stringent variant calling was done with Mutect v1.1.7 (ref. 51) using default settings. Mutations were considered as relevant if the frequency was greater or equal 5% and the read depth was greater or equal 10. Raw read sequences were filtered with Prinseq v0.20.4 (ref. 52). Nucleotides with a Phred Score below 20 at 3′ or 5′ end were clipped. Reads were then mapped to the GRCh38.p3 (http://www.ensembl.org/index.html) reference genome with BWA v0.7.12 (ref. 53) using default settings. Duplicates were marked with Picardtools v1.129 (http://picard.sourceforge.net.) and kept for downstream analysis. Realignment and base recalibration was done with GATK v3.3 (ref. 54). Annotation was done with SNPeff Version 4.1g (ref. 55) based on the ENSEMBL GRCh38.78 genome. Only transcripts with CCDS sequences were used for further analysis.

Semi-quantitative realtime PCR. Tumour tissue from melanoma patients was micro-dissected and RNA extraction was performed according to the manufacturer's instructions using High Pure RNA Paraffin Kit (Roche Diagnostics/Mannheim, Germany). As control, RNA from the Human Total RNA Master Panel II (Clontech, Mountain View, USA), from human testis (Clontech) or human adult skin (amsbio, Abingdon, U.K.) was used. cDNA was synthesized by the Superscript II reverse transcriptase (Invitrogen) using random hexamer primers (Roche). qPCR was conducted on a StepOnePlus system (Applied Biosystems) using the KAPA Probe Fast Universal qPCR MasterMix (peqlab, Erlangen, Germany). Relative quantification was calculated by the delta-delta Ct method[56] using the geometric mean of control genes (GAPDH, HMBS and HPRT1) for normalization. The following primers and probes were used: PMEL: 5′-ACCTATCCCTGAGCCTGAAG-3′ (forward primer (fwd)), 5′-GCCCAGGGAACCTGTAATACT-3′ (reverse primer (rev)), 5′-[6FAM]TGCCAGCTCAATCATGTCTACGGA[TAM]-3′ (probe); Tyrosinase: 5′-TGCACAGATGAGTACATGGGA-3′ (fwd), 5′-GGCTACAGACAATCTGCCAAG-3′ (rev), 5′-[6FAM]CTCAGCCCAGCATCATTCTTCTCCT[TAM]-3′ (probe); PRAME: 5′-TATCGCCCAGTTCACCTCTC-3′ (fwd), 5′-ATCACGTGCCTGAGCAACT-3′ (rev), 5′-[6FAM]CAGTCTGCAGTGCCTGCAGGC[TAM]-3′ (probe); GAPDH: 5′-TTCCAATATGATTCCACCCA-3′ (fwd), 5′-GATCTCGCTCCTGGAAGATG-3′ (rev), 5′-[6FAM]TTCCATGGCACCGTCAAGGC[TAM]-3′ (probe); HMBS: 5′-ACGATCCCGAGACTCTGCTTC-3′ (fwd), 5′-GCACGGCTACTGGCACACT-3′ (rev), 5′-[6FAM]CCTGAGGCACCTGGAAGGAGGCTG[TAM]-3′ (probe); HPRT1: 5′-CTGGCGTCGTGATTAGTGAT-3′ (fwd), 5′-CTCGAGCAAGACGTTCAGTC-3′ (rev), 5′-[6FAM]CATTATGCTGAGGATTTGGAAAGGGTG[TAM]-3′ (probe).

Immunohistochemistry. FFPE tumour samples were selected to construct a tissue microarray using a Tissue Microarrayer (Beecher Instruments/Sun Praierie, USA) with a core size of 0.6 mm. At least three tumour cores from tumour center and tumour periphery were taken from areas previously marked by a pathologist.

Immunohistochemistry was performed on 2 µm sections using the following antibodies: S-100 (polyclonal, dilution 1:600, DAKO, Hamburg, Germany), HMB45 (clone HMB-45, dilution 1:200, Cell Marque, Rocklin, USA), MelanA (clone A103, dilution 1:200, Cell Marque, Rocklin, USA), PRAME (polyclonal, dilution 1:150, Sigma-Aldrich), Tyrosinase (clone T311, dilution 1:200, Santa Cruz, Dallas, Texas). Immunohistochemistry on one representative slide of the pulmonary metastasis was performed using the following antibodies: S-100 (Polyclonal, dilution 1:600, DAKO, Hamburg, Germany), CD3 (Clone MRQ 39, dilution 1:500, Cell Marque, Rocklin, USA) and PD-L1 (Clone 28-8, dilution 1:500, Abcam, Cambridge, United Kingdom). Stainings were run on an automated immunostainer with an iVIEW DAB detection kit (Ventana Medical Systems, Roche Diagnostics, Mannheim, Germany). Appropriate positive controls for each antibody were run in parallel. In cases of marked staining heterogeneity, 2 µm sections from FFPE tumour blocks were stained in addition to exclude scoring inaccuracy due to tumour heterogeneity. Immunoreactivity was evaluated regarding the percentage of positive tumour cells. Nuclear and cytoplasmic staining was taken into account. A 4-tiered system was used for scoring: (0) absent, (1) >0–25%, (2) >25–50%, (3) >50–75%, (4) >75–100%.

Statistics. For the analysis of correlation of ligand identification and antigen expression, the square root of the normalized number of PMEL HLA ligands was calculated to deal with deviations from a normal distribution. Pearson correlation was calculated and the respective p value was corrected for multiple testing. A regression line is depicted for visual guidance on each panel.

HLA typing. HLA typing was done for selected patients on gDNA isolated from PBMC by next generation sequencing (Zentrum für Humangenetik und Laboratoriumsdiagnostik, Martinsried, Germany) or using the HLA miner tool[57] for exome sequencing data when limited patient material was available.

Sanger sequencing of DNA and RNA from tumour samples of Mel15. Snap-frozen tumour tissue obtained from the resection at day 792 was homogenized by mechanical disruption. Genomic DNA was obtained using DNA mini kit (Qiagen). RNA was extracted by passing sheared tissue additionally through a QIAshredder Homogenizer (Qiagen) followed by isolation with RNeasy mini kit (Qiagen). Reverse transcription was performed with Affinity Script (Agilent) and oligo(dT) Primers. PCR was conducted with KOD Polymerase (Merck Millipore) and Primers as described for minigene cloning (see below). Products were purified after gel electrophoresis with Nucleospin Gel and PCR Cleanup kit (Macherey-Nagel) and sequenced at MWG Eurofins (Ebersberg, Germany).

Algorithms used for prediction of peptide ligands. Affinity to the corresponding allotypes expressed in Mel15 was predicted for all eluted peptides identified in Mel15 samples using NetMHC4.0 (ref. 29). To be more conservative regarding assignment of peptides with multiple specificities, the list of peptides was filtered to include only 9 mer peptides that bind to only one HLA allotype. The threshold for binding was set to rank $<2\%$ to include weak binders (standard setting according to ref. 58). This resulted in 1,065, 2,518, 1,499 and 581 peptides that fit HLA-A0301, HLA-A6801, HLA-B2705 and HLA-B3505, respectively. Predicted affinities to the HLA supertype representative allotypes were calculated for the TAA-derived peptides using NetMHCcons[59], and are provided in Supplementary Data 3. Clustering of peptides into groups based on sequence similarities was performed using the GibbsCluster-1.1 tool[27] using default settings.

For prediction of affinity scores of mutated peptide ligands, protein transcript sequences associated with non-synonymous mutations were downloaded from ENSEMBL GRCh38.78. A 23-mer small peptide sequence was generated by adding 11 amino acids up and downstream of the altered position. If the mutation is located less than 11 amino acids away from the 3′ or 5′ end, the peptide is shorter, respectively. The peptide was then used as input for NetMHC4.0 (ref. 29) and fragments comprising the mutation were used for further analysis. Ligands with a predicted affinity of <500 nM were included in the graphical analysis.

Cloning and expression of minigenes. Oligonucleotide primers were designed to amplify fragments of gene products ranging between 200 and 400 bp encompassing the mutated base. Generally, forward primer additionally encoded a methionine and the reverse primer contained a stop codon to allow expression of a mutated and non-mutated version of the respective minigene for immunological assays. gDNA isolated out of FFPE melanoma tissue was used as template for PCR amplification of the minigene containing the defined mutation whereas PBMC served for cloning of the wt minigene, except for the NCAPG2-derived mutated and wt minigenes, which were cloned out of customized synthesized vector constructs (GenScript, Piscataway, USA). Following primer sequences were used for cloning of all constructs: AKAP6_fwd 5′- TAGCGGCCGCCACCATGGATGAG GGGGAAAGCAT-3′, AKAP6_rev 5′- TAGTCGACTTCTATATTGCCACTTTT AT-3′; NOP16_fwd 5′- TAGCGGCCGCCACCATGTCCTCGATTCTTTGCA G-3′, NOP16_rev 5′- TAGTCGACAGAGCGGGGAGTGTGCACGT-3′ (control minigenes); SYTL4_fwd 5′- TAGCGGCCGCCACCATGAGTACGATCGGCAGC AT-3′, SYTL4_rev 5′- TAGTCGACCTTGGCTTCATCAGCATAGG-3′; NCAPG2_fwd 5′- TAGCGGCCGCCACCATGTCTCCAGTGCATTCCAA-3′, NCAPG2_rev 5′- TAGTCGACCATGAAGGTTTGGATCC-3′ (control minigenes). Cells were transduced with retroviral vectors coding for the respective minigenes and the fluorescent dye dsRed Express II to allow sorting of transgenic cells. Retroviral particles were generated as described previously[60]. Briefly, retroviral vector plasmids coding for respective minigenes were co-transfected with plasmids carrying retroviral genes for gag/pol derived from Moloney murine leukaemia virus (pcDNA3.1-Mo-MLV) and env (pALF-10A1) into 293T cells using TransIT (Mirus, Göttingen, Germany). After 48 h incubation, supernatants were filtered (45 µm) and used for transduction of LCL1.

***In-vitro* stimulation of effector T cells.** Recall antigen-experienced T-cell responses were investigated by stimulation of PBMC from patients or healthy donors as previously described[61] with slight modifications. Briefly, 0.3–0.5 Mio PBMC per well were cultivated in AIM-V for 24 h in the presence of IL-4 and GM-GSF (PeproTech, London, UK). 1 µM Peptide (GenScript, Piscataway, USA), 0.5 ng ml^{-1} IL-7 (Peprotech) and 20 µg ml^{-1} Poly-I:C (Invitrogen) were added after 24 h. Cells were then transferred to a previously coated IFN-g ELISpot plate and cultured over night at 37 °C. Afterwards, cells were gently re-suspended and re-cultivated in T-cell medium.

For stimulation of naive T cells with defined mutated peptide ligands, monocytes of healthy donors were differentiated into dendritic cells (DC) by plate adherence and incubation with IL-4 (20 ng ml^{-1}) and GM-CSF (100 ng ml^{-1}) (Peprotech) for 48 h. Cells were further matured using TNF-a (10 ng ml^{-1}), IL-1b (10 ng ml^{-1}), IFN-g (5,000 IU ml^{-1}), PGE_2 (250 ng ml^{-1}) (Peprotech) and CL075 (1 µg ml^{-1}) (InvivoGen, San Diego, USA) for 24 h. Naïve T cells from the DC donor were isolated as described previously[49]. After pulsing of DC with 1 µM peptide for 2 h in AIM-V medium (Invitrogen), priming was started at an effector to target ratio of 10:1 in the presence of IL-21 (30 ng ml^{-1}) (Peprotech). In each stimulation procedure, IL-7 and IL-15 (5 ng ml^{-1}) (Peprotech) were added every two to three days.

Multimer staining and further enrichment of specific T cells. HLA multimers were manufactured as previously described[62]. Multimer staining was performed according to current recommendations and protocols of the CIMT Immunoguiding Program (http://www.cimt.eu/workgroups/cip).

For a detailed investigation of T-cell responses against SYTL4 on the clonal level (Figs 5h and 6h,i), expanded T-cell lines were functionally sorted by enrichment of CD137 positive cells after overnight stimulation with irradiated peptide pulsed T2 cells, cloned by limiting dilution and screened for peptide-specific recognition. Relevant clones (PBMC-SYTL4clone1, TIL-SYTL4clone1) were further expanded using irradiated feeder, IL-2 and Okt-3 for assessment of minigene recognition and peptide titration assays.

Functional T-cell analysis. Expanded T cells were co-incubated after 10–14 days with peptide-pulsed (1 µM) target cells or cell lines transduced with different minigene constructs. Respective target cells were pulsed with the mutated peptides, the wt counterpart or irrelevant peptides with the same HLA restriction as ligands of interest (designated as control peptides). Coincubation assays for detection of cytokine secretion were performed in duplicates. ELISpot analysis was performed with IFN-g-coating monoclonal antibody (1-D1K), IFN-g-capture-mAb (7-B6-1-biotin) and Streptavidin-HRP (all Mabtech, Sweden) as recommended by the manufacturer using 20,000 target cells and 20,000–40,000 effector cells per well as indicated. Phorbol 12-myistate 13-acetate (PMA) (Sigma-Aldrich) and Ionomycin (Merck, Germany) were used for a positive control. ELISpot plates were read out on an ImmunoSpot S6 Ultra-V Analyzer using Immunospot software 5.4.0.1 (CTL-Europe, Bonn, Germany).

Co-culture experiments for assessment of IFN-g release were performed with an effecter-to-target ratio of 1:1 using each 10,000 target and effector cells per well. Peptide titration assays were performed at least twice showing comparable results for each reactivity pattern. IFN-g release in cell culture supernatants of coincubation assays was determined using the BD OptEIA Human IFN-g ELISA Kit II (BD Biosciences, Franklin Lakes, USA). Intracellular cytokine staining was performed with IC staining kit (eBioscience). 100,000 effector cells were coincubated with 100,000 target cells. After one hour, 10 µg ml^{-1} Brefeldin A (Sigma-Aldrich) was added and cells were incubated for 4 more hours at 37 °C. Cells were then stained with Ethidium-monoazide bromide (Invitrogen) for life-dead discrimination and anti-CD8-APC (clone RPA-T8). After fixation and permeabilization, intracellular cytokines were stained with anti-IFN-g-AF700 (clone B27), anti-TNF-a-V450 (clone Mab11) and anti-IL-2-BV510 (clone 5344,111) antibodies (all BD Biosciences).

Cytotoxic activity of specific T cells was analysed by coincubation of 50,000 effector cells with 50,000 target cells followed by using FACS-based quantification of remaining target cells after 20 h. Therefore, cocultures of target and effector cells were stained with 7-Aminoactinomycin D (7AAD) (Sigma-Aldrich) for dead cell exclusion, anti-CD8-FITC (clone HIT8a) and anti-CD3-AF700 (clone UCHT1) (all BD Biosciences). Target cells were identified according to their morphology in the FSC/SSC and gated on 7-AAD$^-$/CD8$^-$/CD3$^-$ events. Absolute numbers of cells per well were calculated with AccuCheck COUNTING BEADS (Invitrogen) according to the manufacturer's instructions. Lysis of minigene transduced or peptide-pulsed LCL was then set in relation to untreated LCL cocultered with respective effector cells using the following formula:

$$\text{percentage of lysed LCL} = \left(1 - \frac{\text{absolute number of remaining LCL}}{\text{mean of untreated LCL}}\right) * 100 \quad (1)$$

Cytotoxic experiments were performed in triplicates and depicted results are representative for two independent experiments each. Measurements of all FACS-based assays were performed on a LSR II flow cytometer (BD Biosciences) and samples were analysed using FlowJo Software.

T-cell responses against freshly removed tumour material from patient Mel15 were analysed by using either small non-treated tumour pieces or digested tumour tissue. Non-treated fresh material was prepared by mincing tumour tissue into small pieces of 1 mm length and two tumour fragments were added to each well of a 96-well plate. Tumour digestion was performed as described previously[63] with slight modifications. Briefly, teased tissue (<3 mm^3) was incubated with tumour digestion medium consisting of RPMI supplemented with DNase type I, Hyaluronidase, Collagenase type IV (all Sigma-Aldrich), Pen/Strep and

Gentamycin. Obtained tumour suspension was cocultured with 50,000 cells of expanded T-cell lines or 100,000 cells of freshly isolated PBMC per well.

Data availability. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository[64] with the dataset identifier PXD004894. Whole exome sequencing data has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002050. The authors declare that all the other data supporting the finding of this study are available within the article and its supplementary information files and from the corresponding author on reasonable request.

References

1. Page, D. B., Postow, M. A., Callahan, M. K., Allison, J. P. & Wolchok, J. D. Immune modulation in cancer with antibodies. *Ann. Rev. Med.* **65,** 185–202 (2014).
2. Weber, J. S. *et al.* Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. *Lancet Oncol.* **16,** 375–384 (2015).
3. van Baren, N. *et al.* Tumoral and immunologic response after vaccination of melanoma patients with an ALVAC virus encoding MAGE antigens recognized by T cells. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **23,** 9008–9021 (2005).
4. Schwartzentruber, D. J. *et al.* gp100 peptide vaccine and interleukin-2 in patients with advanced melanoma. *N. Engl. J. Med.* **364,** 2119–2127 (2011).
5. Robbins, P. F. *et al.* Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **29,** 917–924 (2011).
6. Morgan, R. A. *et al.* Cancer regression in patients after transfer of genetically engineered lymphocytes. *Science* **314,** 126–129 (2006).
7. Johnson, L. A. *et al.* Gene transfer of tumor-reactive TCR confers both high avidity and tumor reactivity to nonreactive peripheral blood mononuclear cells and tumor-infiltrating lymphocytes. *J. Immunol.* **177,** 6548–6559 (2006).
8. Linette, G. P. *et al.* Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122,** 863–871 (2013).
9. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
10. Larkin, J. *et al.* Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *N. Engl. J. Med.* **373,** 23–34 (2015).
11. Robert, C. *et al.* Pembrolizumab versus Ipilimumab in Advanced Melanoma. *N. Engl. J. Med.* **372,** 2521–2532 (2015).
12. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348,** 124–128 (2015).
13. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371,** 2189–2199 (2014).
14. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
15. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
16. Linnemann, C. *et al.* High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nat. Med.* **21,** 81–85 (2015).
17. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-Cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol.* **31,** e439–e442 (2013).
18. Kreiter, S. *et al.* Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* **520,** 692–696 (2015).
19. Tran, E. *et al.* Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* **350,** 1387–1390 (2015).
20. Lennerz, V. *et al.* The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc. Natl Acad. Sci. USA* **102,** 16013–16018 (2005).
21. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515,** 572–576 (2014).
22. Gubin, M. M. *et al.* Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515,** 577–581 (2014).
23. Kalaora, S. *et al.* Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* **7,** 5110–5117 (2016).
24. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351,** 1463–1469 (2016).
25. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteom.: MCP* **14,** 658–673 (2015).
26. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).
27. Andreatta, M., Lund, O. & Nielsen, M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* **29,** 8–14 (2013).
28. Caron, E. *et al.* Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol. Cell. Proteom.: MCP* **14,** 3105–3117 (2015).
29. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61,** 1–13 (2009).
30. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** D512–D520 (2015).
31. Mohammed, F. *et al.* Phosphorylation-dependent interaction between antigenic peptides and MHC class I: a molecular basis for the presentation of transformed self. *Nat. Immunol.* **9,** 1236–1243 (2008).
32. Cobbold, M. *et al.* MHC class I-associated phosphopeptides are the targets of memory-like immunity in leukemia. *Sci. Transl. Med.* **5,** 203ra125 (2013).
33. Schittenhelm, R. B., Dudek, N. L., Croft, N. P., Ramarathinam, S. H. & Purcell, A. W. A comprehensive analysis of constitutive naturally processed and presented HLA-C*04:01 (Cw4)-specific peptides. *Tissue Antigens* **83,** 174–179 (2014).
34. Trolle, T. *et al.* The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and mhc allele-specific binding preference. *J. Immunol.* **196,** 1480–1487 (2016).
35. McMurtrey, C. *et al.* Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. *Elife* **5,** e12556 (2016).
36. Apcher, S., Prado Martins, R. & Fahraeus, R. The source of MHC class I presented peptides and its implications. *Curr. Opin. Immunol.* **40,** 117–122 (2016).
37. Rodriguez, A., Regnault, A., Kleijmeer, M., Ricciardi-Castagnoli, P. & Amigorena, S. Selective transport of internalized antigens to the cytosol for MHC class I presentation in dendritic cells. *Nat. Cell Biol.* **1,** 362–368 (1999).
38. Abelin, J. G. *et al.* Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry. *Nat. Protoc.* **10,** 1308–1318 (2015).
39. Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8,** 1583–1594 (2014).
40. Lu, K. P., Liou, Y. C. & Zhou, X. Z. Pinning down proline-directed phosphorylation signaling. *Trends Cell Biol.* **12,** 164–172 (2002).
41. Zarling, A. L. *et al.* MHC-restricted phosphopeptides from insulin receptor substrate-2 and CDC25b offer broad-based immunotherapeutic agents for cancer. *Cancer Res.* **74,** 6784–6795 (2014).
42. Verdegaal, E. M. *et al.* Neoantigen landscape dynamics during human melanoma-T cell interactions. *Nature* **536,** 91–95 (2016).
43. Stronen, E. *et al.* Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science* **352,** 1337–1341 (2016).
44. Gros, A. *et al.* Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat. Med.* **22,** 433–438 (2016).
45. Chambers, C. A., Kuhns, M. S. & Allison, J. P. Cytotoxic T lymphocyte antigen-4 (CTLA-4) regulates primary and secondary peptide-specific CD4(+) T cell responses. *Proc. Natl Acad. Sci. USA* **96,** 8603–8608 (1999).
46. Cohen, C. J. *et al.* Isolation of neoantigen-specific T cells from tumour and peripheral lymphocytes. *J. Clin. Investig.* **125,** 3981–3991 (2015).
47. Lu, Y. C. *et al.* Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumour regressions. *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res.* **20,** 3401–3410 (2014).
48. Jenkins, M. R. *et al.* Visualizing CTL activity for different CD8+ effector T cells supports the idea that lower TCR/epitope avidity may be advantageous for target cell killing. *Cell Death Differ.* **16,** 537–542 (2009).
49. Klar, R. *et al.* Therapeutic targeting of naturally presented myeloperoxidase-derived HLA peptide ligands on myeloid leukemia cells by TCR-transgenic T cells. *Leukemia* **28,** 2355–2366 (2014).
50. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).
51. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).
52. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
54. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
55. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

56. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–408 (2001).
57. Warren, R. L. *et al.* Derivation of HLA types from shotgun sequence datasets. *Genome Med.* **4,** 95 (2012).
58. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32,** 511–517 (2015).
59. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64,** 177–186 (2012).
60. Weigand, L. U. *et al.* Isolation of human MHC class II-restricted T cell receptors from the autologous T-cell repertoire with potent anti-leukaemic reactivity. *Immunology* **137,** 226–238 (2012).
61. Martinuzzi, E. *et al.* acDCs enhance human antigen-specific T-cell responses. *Blood* **118,** 2128–2137 (2011).
62. Knabel, M. *et al.* Reversible MHC multimer staining for functional isolation of T-cell populations and effective adoptive transfer. *Nat. Med.* **8,** 631–637 (2002).
63. Wang, R. F. Molecular cloning and characterization of MHC class I- and II-restricted tumor antigens recognized by T cells. *Curr. Protoc* **84,** 20.10.1–20.10.29 (2009).
64. Vizcaino, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226 (2014).

Acknowledgements

This work was supported by grants from the Wilhelm Sander-Stiftung (2015.030.1), the DFG/SFB824 (C10) and DFG TR/SFB36 (A13). M.B.-S. was supported by the Alexander von Humboldt-Foundation. We are highly thankful to our patients for their cooperation. We also thank Stephanie Rämisch for excellent technical support and Chloe Chong for experimental support during revision. There are no competing interests for patenting.

Author contributions

M.B.-S., R.K., M.M., A.M.K. designed the study, M.B.-S., E.B., R.K., S.A., J.W., M.S., J.S.-H., K.S. and A.M.K. performed and/or analysed experiments, D.H.B. generated MHC multimer reagents, T.E., P.S., R.R. and J.C. performed statistics and bioinformatics, R.H., A.W., M.E.M., C.P. and A.M.K. provided patient material, M.B.-S., E.B., R.K., T.E., R.R., M.M. and A.M.K. wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7,** 13404 doi: 10.1038/ncomms13404 (2016).

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

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

The era of high-throughput genomics brought a considerable advance in the field of microbiome research. Large scale projects as human microbiome project (HMP) studied the collection of thousands of samples from a dozen body cavities and thousands of patients[39]. Prior studies were focused on a limited set of bacteria that could be cultivated outside of the body site. The new studies demonstrate high variability of microbiome between people and conditions as well as broad dynamics within a day.

The oral microbiome is considered the most dynamic biome in the human body[164]. MALDI-TOF proteomics is widely used in clinical studies as a fast detection method for highly abundant bacterial species[165]. However, this approach has a limited dynamic range. In this article we compared in-lab developed shotgun proteomics approach with state-of-art techniques to study the oral microbiome, such as MALDI-TOF mass spectrometry and next-generation sequencing data from HMP[166]. Our data shows a strong agreement between all techniques. However, using the shotgun proteomics we were able to detect bacteria within a large expression range (4 orders of magnitude of LFQ intensity). An additional advantage of our approach is the ability to detect secreted human proteins as well, that are not accessible by genomics methods.

I contributed to this study by developing a quantitative method for bacteria identification. This method is generally applied to proteomics and genomics data. In short, we placed a quantity of identified bacterial peptides as well as NGS DNA fragments, on a taxonomic tree such that these peptides do not allow discrimination of the branches below. Using this method, we were able to make a PCA analysis of our proteomics results together with results from HMP and observe a significant agreement.

Niklas Grassl, Nils Alexander Kulak, Garwin Pichler, Philipp Emanuel Geyer, Jette Jung, Sören Schubert, **Pavel Sinitcyn**, Jürgen Cox, Matthias Mann
Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome
(2016) *Genome medicine*
DOI: 10.1186/s13073-016-0293-0

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

Genome Medicine

RESEARCH Open Access

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

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

Abstract

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

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

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

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

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

Background

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

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

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

Methods

Experimental design

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

Protein digestion and peptide purification

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

Single run and prefractionated liquid chromatography-MS measurement

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

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

Raw data processing of human proteins

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

Data analysis of human proteins

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

Raw data processing of human and bacterial proteins

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

Data analysis of the oral microbiome

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

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

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

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

Microbiological processing of the samples

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

Identification by MALDI-TOF MS

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

Results and discussion

In-depth quantification of the saliva proteome

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

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

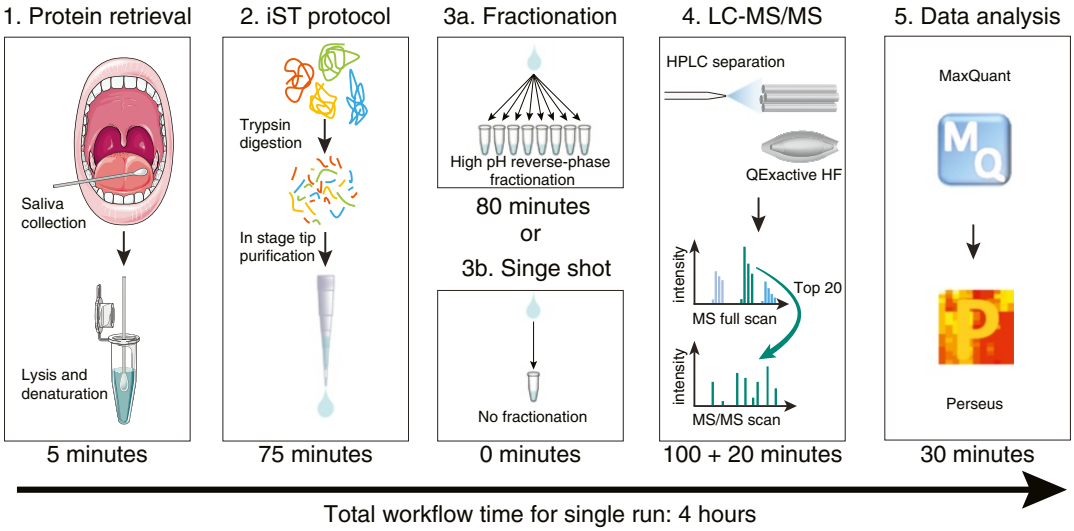


Fig. 1 Workflow for ultra-deep and quantitative saliva proteomics. (1, 2) Saliva is collected with a sterile cotton swab and its proteins are denatured, digested, and purified according to the iST protocol [12]. (3) Depending on the desired proteome depth, samples are either separated into eight fractions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

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

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

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

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

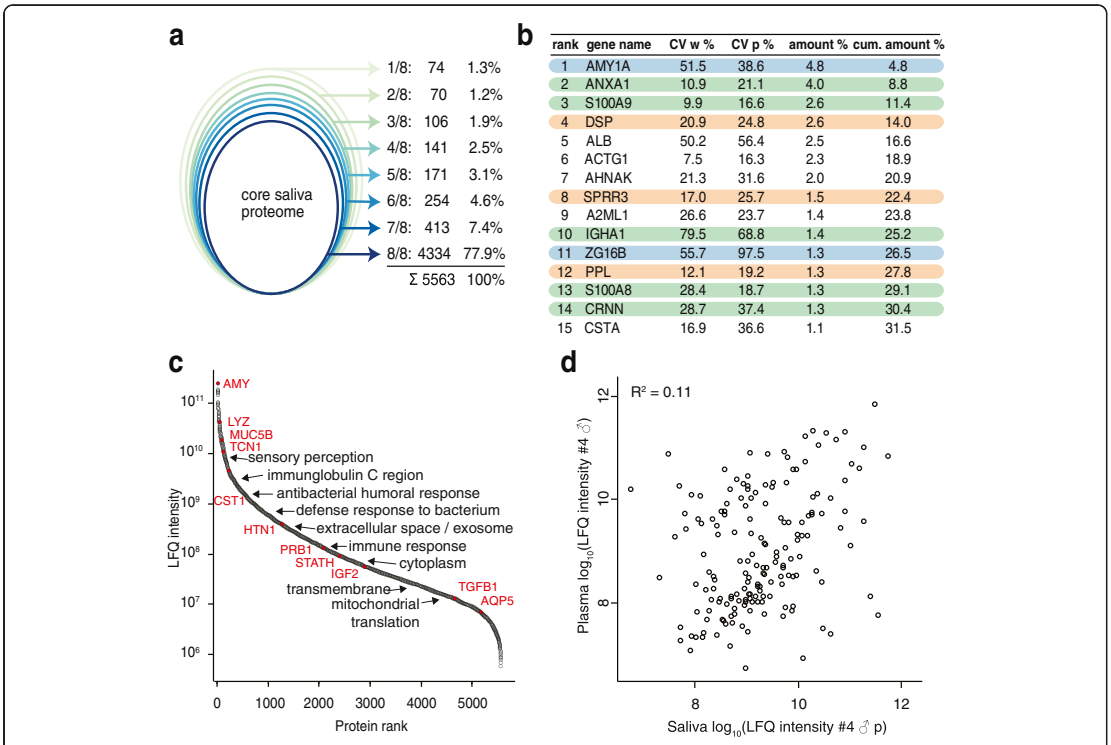


Fig. 2 Deep human saliva proteomes of eight healthy donors. **a** *Ovals* represent the number of saliva proteins shared by the respective number of donors. The *outer oval* contains all proteins that were detected in at least one donor, whereas the inner oval contains all proteins found in each sample—the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

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

A deep single-run workflow

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

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

Dynamics of the saliva proteome in a cohort

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

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

Identification of bacterial proteomes in human saliva

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

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

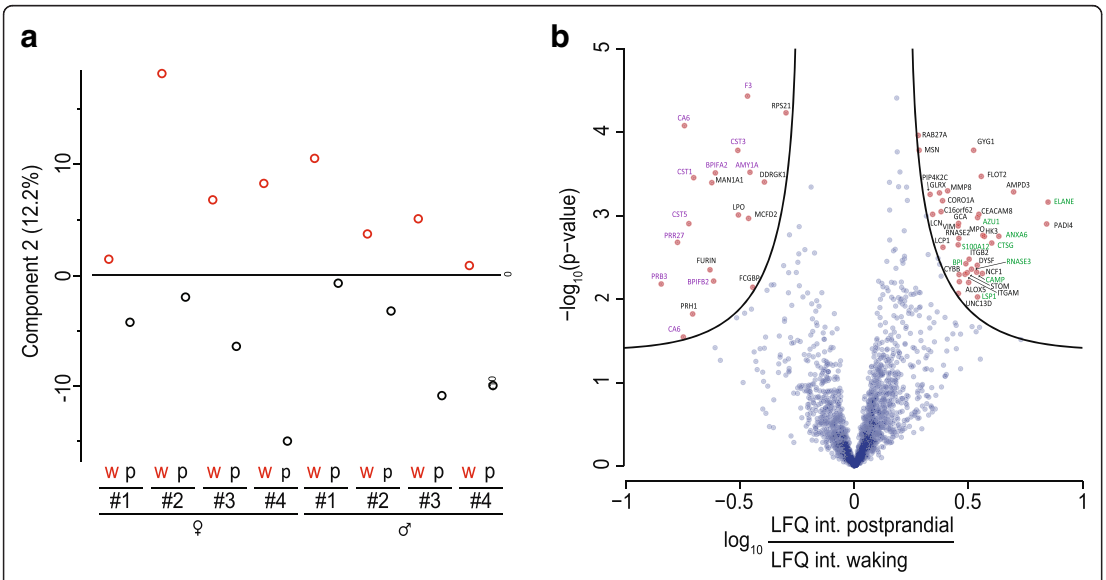


Fig. 3 Intraday dynamics of the human saliva proteome. **a** PCA of the 16 saliva samples showing that component 2 separates samples based on the collection time (*w* = waking and *p* = postprandial). **b** Differentially regulated proteins between w and p as determined by plotting the *t*-test significance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

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

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

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

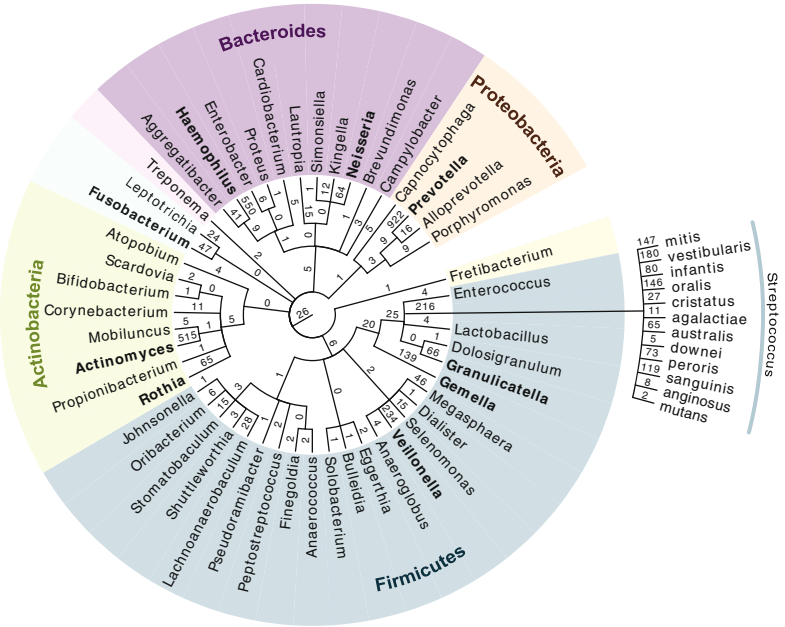


Fig. 4 Taxonomy tree of 50 bacterial genera with evidence at the peptide level. The number of peptides that were specifically attributable to this position on the taxonomic tree are given above the *edges* of the graph. Genera in *bold* were also detected after bacterial culture of the saliva samples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

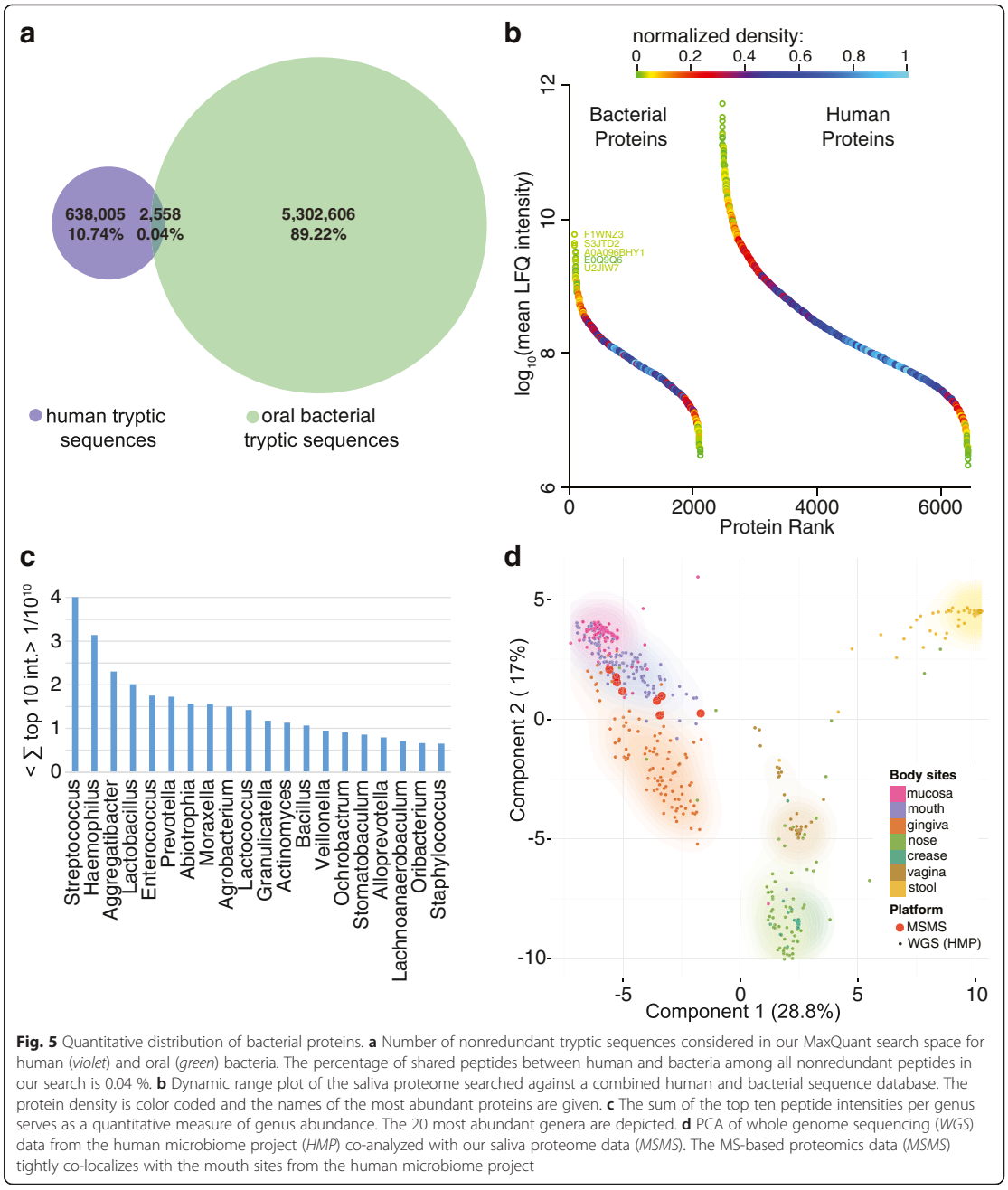


Fig. 5 Quantitative distribution of bacterial proteins. **a** Number of nonredundant tryptic sequences considered in our MaxQuant search space for human (*violet*) and oral (*green*) bacteria. The percentage of shared peptides between human and bacteria among all nonredundant peptides in our search is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

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

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

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

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

The quantitative oral metaproteome

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

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

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

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

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

Variation and dynamics of the metaproteome

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

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

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

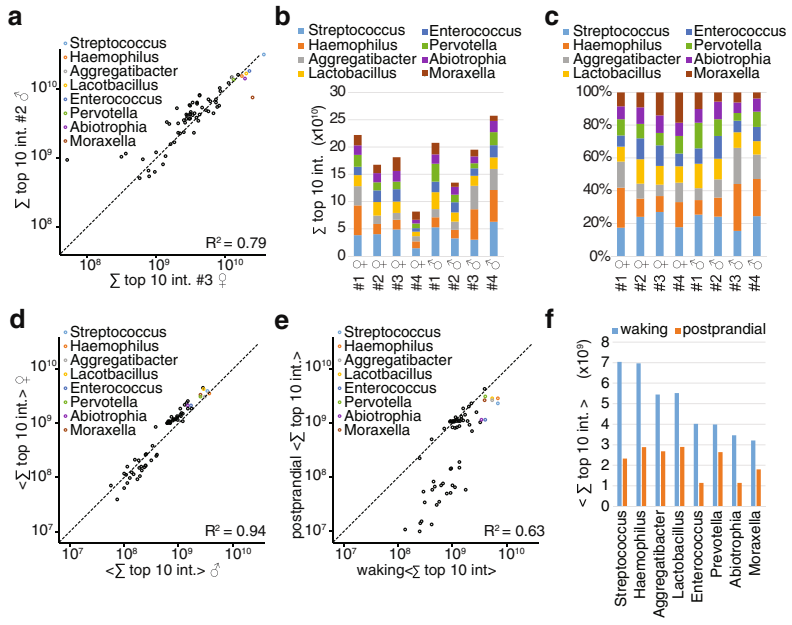


Fig. 6 Bacterial composition across donors and time points. **a** Typical scatter plot for two donors of the bacterial genera quantified by the sum of their top ten peptide intensities from fractionated proteome measurements. The eight most abundant genera are color coded. **b** Absolute quantification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

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

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

Conclusions

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

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

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

Availability of supporting data

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

Additional file

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

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

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

Acknowledgments

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

Author details

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

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

References

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

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

2.2.3 The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins

Many ribosomes are engulfed on the surface of the endoplasmic reticulum (ER) where they synthesize secreted and membrane proteins. The membrane proteins need special processing to be properly oriented and inserted in the ER membrane. Most ion channels cross the membrane multiple times and frequently have functionally important charged amino acids in the middle of the transmembrane domain. Due to their complexity, such proteins need additional chaperon help[167].

In this article, we demonstrated the importance of multisubunit ER membrane complex (EMC) for biogenesis of a range of multipass transmembrane proteins, with a particular enrichment for transporters[37, 168]. For this purpose, two subunits of the EMC were depleted independently. In both cases, the same subset of membrane proteins was degraded while their RNA level and translational rate remain unchanged. This suggests that in the absence of the EMC chaperon, this subset of proteins were not properly folded which resulted in their degradation.

I contributed to this article by analyzing transcriptomics and proteomics data for differential expression. I was involved in the effort to find a common sequence motif among the EMC targets.

Matthew J Shurtleff*, Daniel N Itzhak*, Jeffrey A Hussmann, Nicole T Schirle Oakdale, Elizabeth A Costa, Martin Jonikas, Jimena Weibezahn, Katerina D Popova, Calvin H Jan, **Pavel Sinitcyn**, Shruthi S Vembar, Hilda Hernandez, Jürgen Cox, Alma L Burlingame, Jeffrey L Brodsky, Adam Frost, Georg HH Borner, Jonathan S Weissman
The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins
(2018) *eLife*
DOI: 10.7554/eLife.37018

*these authors contributed equally to this work

eLIFE elifesciences.org

RESEARCH ARTICLE

The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins

Matthew J Shurtleff[1†], Daniel N Itzhak[2†], Jeffrey A Hussmann[1], Nicole T Schirle Oakdale[1,3], Elizabeth A Costa[1], Martin Jonikas[1‡], Jimena Weibezahn[1], Katerina D Popova[1], Calvin H Jan[1§], Pavel Sinitcyn[2], Shruthi S Vembar[4], Hilda Hernandez[5], Jürgen Cox[2], Alma L Burlingame[5], Jeffrey L Brodsky[4], Adam Frost[3,6], Georg HH Borner[2]*, Jonathan S Weissman[1,7]*

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, United States; [2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany; [3]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, United States; [4]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, United States; [5]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, United States; [6]Chan Zuckerberg Biohub, San Francisco, United States; [7]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, United States

***For correspondence:** borner@biochem.mpg.de (GHHB); Jonathan.Weissman@ucsf.edu (JSW)

[†]These authors contributed equally to this work

Present address: [‡]Department of Molecular Biology, Princeton University, Princeton, United States; [§]Calico Life Sciences LLC, San Francisco, United States

Competing interests: The authors declare that no competing interests exist.

Funding: See page 19

Received: 27 March 2018
Accepted: 26 May 2018
Published: 29 May 2018

Reviewing editor: David Ron, University of Cambridge, United Kingdom

Abstract The endoplasmic reticulum (ER) supports biosynthesis of proteins with diverse transmembrane domain (TMD) lengths and hydrophobicity. Features in transmembrane domains such as charged residues in ion channels are often functionally important, but could pose a challenge during cotranslational membrane insertion and folding. Our systematic proteomic approaches in both yeast and human cells revealed that the ER membrane protein complex (EMC) binds to and promotes the biogenesis of a range of multipass transmembrane proteins, with a particular enrichment for transporters. Proximity-specific ribosome profiling demonstrates that the EMC engages clients cotranslationally and immediately following clusters of TMDs enriched for charged residues. The EMC can remain associated after completion of translation, which both protects clients from premature degradation and allows recruitment of substrate-specific and general chaperones. Thus, the EMC broadly enables the biogenesis of multipass transmembrane proteins containing destabilizing features, thereby mitigating the trade-off between function and stability.
DOI: https://doi.org/10.7554/eLife.37018.001

Introduction

As the primary site of transmembrane protein synthesis, insertion, and folding, the endoplasmic reticulum (ER) must accommodate a diverse range of transmembrane proteins destined for locations throughout the cell. Individual transmembrane domains (TMDs) of multipass proteins are cotranslationally inserted into the lipid bilayer, and this step can often be energetically costly (***Cymer et al., 2015***). Some features of diverse transmembrane proteins present a particular challenge for their insertion into and stabilization within the ER membrane. First, the length of the TMD may not match

the thickness of the lipid bilayer in the ER due to differences in membrane composition between the ER and the protein's final destination (**Sharpe et al., 2010**). Additionally, many TMDs contain features that are destabilizing during membrane protein insertion and biosynthesis but are necessary for function. Transporters, transmembrane ATPases, and solute carriers, for example, contain polar and/or charged residues within membrane spanning domains that form aqueous channels within the plane of the membrane (**Tector and Hartl, 1999**). Yet, during biogenesis, these charged helices enter the lipid bilayer unshielded by the remainder of the protein.

Faced with these challenges, membrane protein folding is subject to failure both in normal and disease settings. Misfolded membrane proteins underlie a host of diseases, including cystic fibrosis (due to mutations in the cystic fibrosis conductance regulator – CFTR), Charcot-Marie-Tooth disease (PMP22 and Connexin 32), diabetes insipidis (Aquaporin), retinitis pigmentosa (Rhodopsin), Niemann-Pick disease (NPC1) (**Gelsthorpe et al., 2008**) and others. Disease-associated mutations often cluster within transmembrane helices (**Sanders and Myers, 2004**) and are biased towards missense mutations, resulting in the addition of polar or charged residues within transmembrane helices (**Partridge et al., 2002**). The disease-associated CFTR mutations illustrate the challenges and opportunities associated with membrane protein biogenesis. Even in healthy cells under normal conditions, less than 50% of newly synthesized CFTR folds properly and traffics out of the ER (**Kopito, 1999**). Small molecule enhancers of folding have emerged as a promising therapeutic strategy for disease-associated misfolded proteins. Indeed, the treatment for the most common form of cystic fibrosis (CFTRΔF508) employs a drug combination including a 'folding corrector' to allow increased ER exit and transport to the cell surface and a 'potentiator' to increase chloride transport activity (**Boyle et al., 2014**). Understanding the full complement of machinery used by the ER to ensure membrane protein synthesis, folding and transport is a fundamental problem with clear biomedical implications.

Although the ER has machinery for stabilizing, sensing and degrading misfolded proteins, how proteins with destabilizing features within transmembrane helices are stabilized during and immediately following synthesis remains poorly understood. The ER membrane protein complex (EMC) has emerged as an intriguing player in membrane protein biogenesis in the ER. The EMC was first described in yeast as a complex of 6 co-purifying and conserved proteins (Emc1-6) with strongly correlated phenotypes in a double mutant genetic modifier map of the unfolded protein response (**Jonikas et al., 2009**). The pattern of EMC genetic interactions strongly resembles the pattern of other factors whose loss leads to the accumulation of misfolded membrane proteins, including the overexpression of a misfolded membrane protein (Sec61-2), and these insights first suggested that the EMC may be a TMD protein chaperone (**Jonikas et al., 2009**). The mammalian EMC orthologues were subsequently identified in a mammalian physical interaction map of ER-associated degradation (ERAD) components (**Christianson et al., 2011**). In vivo experiments have shown that loss of the EMC compromises synthesis, stabilization and/or trafficking of specific multipass membrane proteins in *S. cerevisiae* (a Yor1 mutant which mimicked a common disease allele of CFTR, and Mrh1) (**Louie et al., 2012**; **Bircham et al., 2011**), *D. melanogoster* (rhodopsin) (**Satoh et al., 2015**), *C. elegans* (acetylcholine receptors) (**Richard et al., 2013**) and mice (ABCA3) (**Tang et al., 2017**), and that knockdown of an EMC component compromised biogenesis of mutant CFTR expressed in HeLa cells (**Louie et al., 2012**). In addition, the EMC has been implicated in autophagy (**Shen et al., 2016**; **Li et al., 2013**), lipid transfer and tethering between the ER and mitochondria (**Lahiri et al., 2014**), and flavivirus replication (**Zhang et al., 2016**; **Savidis et al., 2016**; **Marceau et al., 2016**; **Ma et al., 2015**; **Krishnan et al., 2008**). Finally, the EMC was recently shown to act as a posttranslational insertase into the ER membrane for the sterol biosynthesis enzyme, squalene synthase (SQS/FDFT1) and a subset of other tail-anchored (TA) proteins. These TA substrates have moderately hydrophobic TMDs rendering them unable to interact with TRC40/Get3, the cytosolic receptor that delivers TA proteins to the dominant ER insertase, GET1/2 (*24*).

Despite its clear importance, many questions regarding EMC function remain: Is the effect on multipass transmembrane protein biogenesis direct or indirect (e.g. due to changes in lipid composition)? If direct, what is the EMC substrate range and does the EMC physically interact with clients? Lastly, at which stage(s) does the EMC act: insertion into the membrane (**Guna et al., 2018**), co- or post-translationally, during folding, or, finally, during ER export?

Here, we used systematic and unbiased in vivo approaches to identify client proteins and to explore the principles of EMC action with minimal perturbations in both yeast and human cells. Our

studies reveal three conserved principles of EMC function: (1) The EMC interacts with and stabilizes a range of client proteins consisting, with a few exceptions, of multipass transmembrane proteins biased towards transporters. (2) The EMC can initiate client interactions cotranslationally and stabilizes newly synthesized, client proteins after initial ER targeting to prevent premature degradation. (3) The EMC can engage client proteins following clusters of TMDs, and client TMDs are enriched in uncommon transmembrane amino acids (especially charged and bulky residues). Thus, the EMC enables the biogenesis and folding of a subset of multipass membrane proteins which present challenges for the canonical membrane protein synthesis and insertion machineries and thereby expands the functional repertoire of the membrane protein proteome.

Results

The EMC physically interacts with multipass transmembrane proteins transiting the ER and substrate-specific chaperones

We initially sought to define the range of proteins that interact with the EMC in the budding yeast *Saccharomyces cerevisiae*. To evaluate EMC interaction partners, we endogenously tagged EMC3 with 3xFLAG epitope at its C-terminus. To maximize retention of interacting partners, we recovered Emc3 using a one-step affinity purification in the presence of digitonin. Following SDS-PAGE analysis, prominent bands were excised and analyzed by mass spectrometry to identify Emc3-interacting proteins (***Figure 1A***). As expected, we identified roughly stoichiometric quantities of all core EMC components (Emc1, Emc2, Emc4, Emc5 and Emc6) and the accessory proteins Sop4 and Emc10 (***Jonikas et al., 2009***). Although Sop4 is a near-stoichiometric interactor with the EMC, it shares only a subset of the characteristic genetic interactions of other core components (***Jonikas et al., 2009***), suggesting that, in contrast to the six core EMC components (Emc1-6), the full function of the EMC complex does not depend on Sop4, and it may play a role distinct from the core complex. In addition to EMC components, Emc3 interacted with several large, multi-pass membrane proteins (Pma1, Fks1, Spf1) and Ilm1, a poorly characterized ER-localized protein (***Lockshon et al., 2007***) (***Figure 1A***).

To confirm that these interactions were specific to the EMC, we performed stable isotope labeling with amino acids in cell culture (SILAC) for Emc3-3xFLAG or N-terminally 3xFLAG tagged Orm1 yeast strains. The Orm complex is an unrelated ER resident complex that serves as a control for specific interactions with the EMC. This quantitative analysis verified interactions with all core (Emc1-6) and accessory EMC components (Sop4 and Emc10), and demonstrated specific interactions between the EMC, multipass membrane proteins (Spf1, Fks1, and Pma1) and Ilm1 (***Figure 1B***). We also detected specific interaction with Erg9, the yeast homolog of a TA protein recently shown to be inserted into the ER in an EMC-dependent manner (***Guna et al., 2018***). Therefore, this approach detects both stable and transient interactions between the EMC and binding partners.

The EMC interacts with specialized membrane protein chaperones

In addition to interacting with putative client transmembrane proteins, our pulldown results implicate the EMC as interacting with membrane protein substrate-specific chaperones, including Sop4 and Gsf2 (***Figure 1B***). Sop4 was previously shown to be a specialized chaperone/transport factor required for the biogenesis of the yeast plasma membrane ATPase (Pma1), (***Luo et al., 2002***), and Gsf2 plays a role in the biogenesis and export of hexose transporter 1 (Hxt1) from the ER (***Sherwood and Carlson, 1999***).

Several considerations suggested to us that the resident ER protein Ilm1 could also act as substrate-specific chaperone for the cell wall synthesis enzyme, Fks1. ILM1 deletion results in enhanced oleic acid sensitivity, likely due to a cell wall defect (***Lockshon et al., 2007***) and showed increased sensitivity to caspofungin, similar to the deletion phenotype for FKS1 (***Markovich et al., 2004***). To explore if Ilm1 functioned as an Fks1 chaperone, we generated yeast expressing Ilm1-3xFLAG and performed pulldowns to identify Ilm1 interacting proteins. We observed a near stoichiometric interaction between Ilm1 and Fks1 (***Figure 1C***). Since mature Fks1 localizes to the plasma membrane and Ilm1 is exclusively in the ER, this likely represents an interaction between Ilm1 and transiting, newly synthesized Fks1. Other prominent bands were identified as components of the EMC (Emc1 and Emc2), ER oxidoreductase 1 (Ero1) and general chaperone proteins (Ssa1, Ssb1 and Kar2). The near

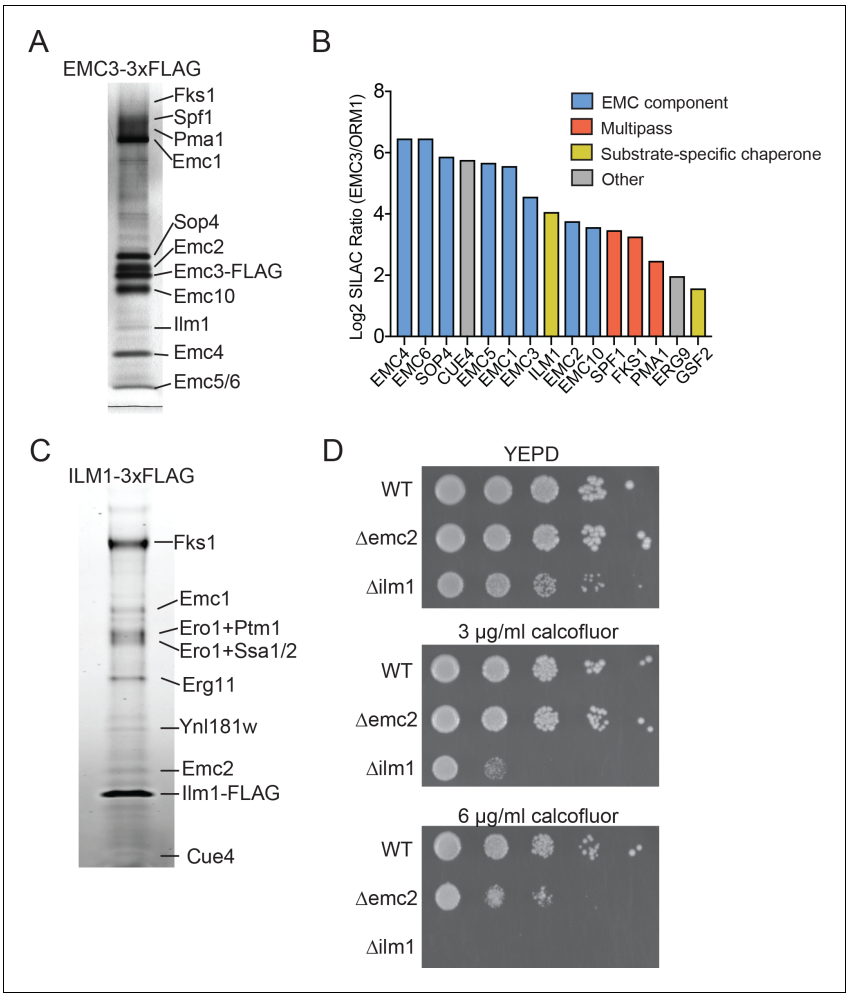


Figure 1. Identifying EMC interaction partners in *S. cerevisiae*. (**A**) SDS-PAGE gel of Emc3-3xFLAG co-immunoprecipitated proteins. Proteins identified from excised bands are indicated. (**B**) SILAC ratios for proteins identified by mass spectrometry for Emc3-3xFLAG (heavy) and ORM1-3xFLAG (light) immunoprecipitations. The most Emc3-3xFLAG enriched proteins are shown. (**C**) SDS-PAGE gel of Ilm1-3xFLAG co-immunoprecipitated proteins. Proteins identified from excised bands are indicated. (**D**) 10-fold serial dilutions of log phase cultures of the indicated strains were plated on YEPD, YEPD containing calcofluor white at the indicated concentrations and incubated at 30°C for 24 hr.
DOI: https://doi.org/10.7554/eLife.37018.002

stoichiometric interaction between Ilm1 and Fks1 and interactions with chaperone proteins further suggested a role for Ilm1 as a co-chaperone for Fks1.

As the catalytic subunit of the 1,3-beta-D-glucan synthase cell wall biosynthesis enzyme, FKS1 deletion results in hypersensitivity to compounds, such as calcofluor white, that affect cell wall assembly (***Ram et al., 1995***). If Ilm1 acts as a co-chaperone for Fks1, we hypothesized that a *Δilm1* strain would show a calcofluor white hypersensitive phenotype. Indeed, we found profound sensitivity with growth decreased at 3 µg/ml and completely inhibited at 6 µg/ml calcofluor white in the

Δ*ilm1* background (***Figure 1D***). We also noted increased sensitivity compared to wild type for Δ*emc2* at 6 µg/ml, further supporting a role for the EMC in maintaining cell wall integrity, possibly by promoting the biosynthesis of Fks1.

EMC interacts with transmembrane protein clients independent of substrate specific chaperones

We first hypothesized that the EMC might directly interact with client-specific membrane protein chaperones, which act to bridge the EMC and multipass membrane protein clients. We tested this hypothesis by performing quantitative mass spectrometry of Emc3-3xFlag interacting proteins in wild type, Δ*sop4* or Δ*ilm1* strains (***Figure 2A***). Rather than a decrease in the interaction between the EMC and multipass proteins, as predicted by the bridging model, we observed a prominent increase in EMC association with Fks1 and the functionally redundant paralog, Gsc2, in the Δ*ilm1* background as well as Elo2, Pma1 and Mrh1 in the Δ*sop4* background (***Figure 2B***). Interestingly, Mrh1 was previously shown to be dependent on the EMC for its biosynthesis and cell surface localization (***Bircham et al., 2011***). Consistent with a role for Ilm1 and Sop4 as membrane protein-specific chaperones, we observed a decrease in general chaperone proteins (Ssa1, Ssb1, and Kar2) and an increase in ribosomal proteins associated with EMC3-3xFLAG in strains missing these factors (***Figure 2B***). Together, these pulldown experiments showed that the EMC interacts with multipass membrane proteins transiting through the ER, membrane protein specific co-chaperones, general chaperones, and the ribosome. The presence of client-specific co-chaperones (Ilm1 and Sop4) enhances the interaction between the EMC and general chaperones and decreases the association with both multipass transmembrane clients and the ribosome (***Figure 2C***). These observations

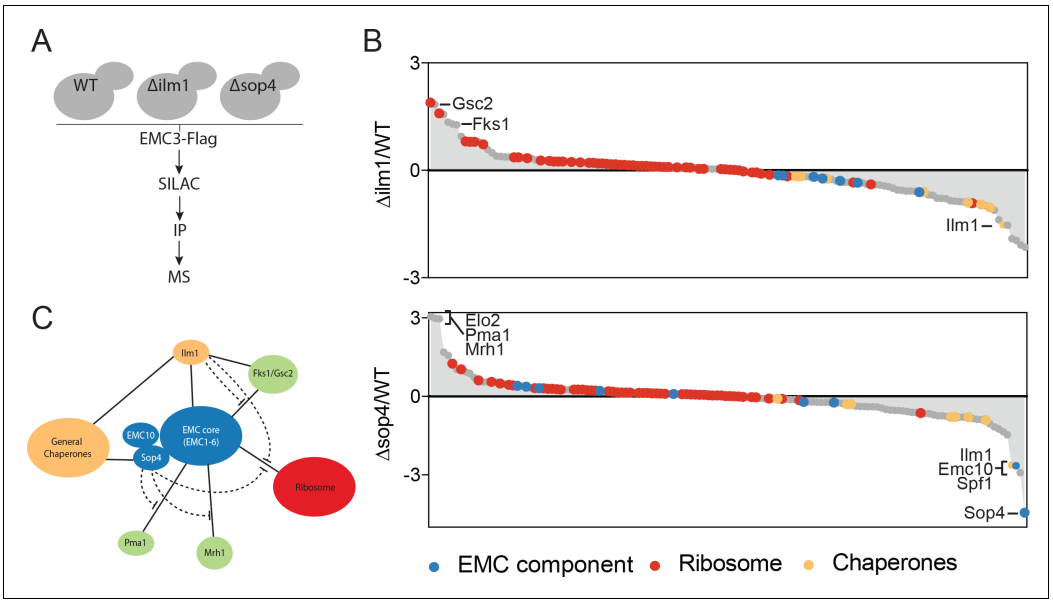


Figure 2. The EMC interacts with multipass client proteins independent of co-chaperones. (**A**) Schematic showing SILAC strategy for comparative analysis of EMC3-3xFLAG interactions in wildtype (WT - light) and Δ*ilm1 (heavy)* and Δ*sop4* (heavy) cells. IP – immunoprecipitation, MS – mass spectrometry. (**B**) Log2 SILAC ratios for all proteins identified in EMC3-FLAG expressing strains (top - Δ*sop4* and WT, bottom - Δ*ilm1* and WT). Enriched multipass proteins and strongly depleted proteins are indicated. (**C**) Schematic showing a summary of physical interactions based on pull downs presented in ***Figures 1*** and ***2***.

DOI: https://doi.org/10.7554/eLife.37018.003

suggest that the EMC interacts directly with multipass transmembrane client proteins early during their synthesis, insertion or folding independent of and possibly prior to chaperone engagement.

The EMC cotranslationally interacts with folding-challenged multipass client proteins

The observed physical interaction between the EMC and Fks1 as well as the increased association of the EMC with ribosomal proteins following chaperone deletion suggested that the EMC may interact with client proteins cotranslationally. To explore this possibility, we performed proximity-specific ribosome profiling (***Jan et al., 2014***) in yeast expressing a fusion protein of an EMC component with biotin ligase (Emc5-BirA) and ribosomes incorporating an AviTag (the substrate of the biotin ligase). Avi-Tagged ribosomes that contact the EMC are biotinylated upon pulse-labelling with biotin. Following streptavidin-based isolation of the biotinylated ribosomes, deep sequencing of the mRNA fragments protected by affinity-purified ribosomes and comparison to the total pool of ribosomes allows identification of messages translated in the proximity of the EMC (***Figure 3A***). In particular, the ratio of pulldown-to-total footprint reads across a message provides a codon-resolution measurement of when translating ribosomes are most accessible to Emc5-BirA.

The profile of this ratio across FKS1 and GSC2 revealed prominent increases in EMC-ribosome proximity immediately following the translation of clusters of TMDs (***Figure 3B***). This enrichment pattern was not observed in strains with two other ER-localized BirA fusions (BirA-Ubc6-TA and BirA-Ssh1), suggesting that the EMC specifically interacts with these nascent chains shortly after synthesis of TMD clusters by the ribosome. The positional enrichment pattern observed for FKS1 and GSC2 motivated a systematic search for other nascent chains cotranslationally engaged by the EMC. To do this, we computed the ratio of total Emc5-BirA enrichment to total BirA-Ubc6-TA enrichment in a sliding window of 101 codons across all genes We then ranked genes according to the highest enrichment ratio attained anywhere in the gene. In addition to the most enriched genes (FKS1 and GSC2), we identified 51 genes that reproducibly showed localized peaks of Emc5-specific enrichment (defined as being in the top 10% of eligible genes in each of two biological replicates; see Materials and methods) (***Figure 3—figure supplement 1***). This list of putative EMC client proteins includes two genes for which the EMC was previously implicated in their biogenesis (PMA1 and YOR1) (***Louie et al., 2012***; ***Luo et al., 2002***). Emc5-BirA positional enrichment for client proteins typically was triggered following synthesis of a cluster of TMDs (***Figure 3—source data 1***). We also performed the same analysis on data from a BirA fusion to Sec63, a component of the Sec translocon, produced for an earlier study (***Aviram et al., 2016***). Intriguingly, patterns of enrichment for Sec63-BirA mirrored some, but not all, of the localized Emc5-BirA peaks (***Figure 3C***).

To gain further insight into the timing of cotranslational engagement of the EMC with client nascent chains, we compared the average enrichments around the first TMD for the various ER localized BirAs (i.e. Emc5, Sec63, Ssh1, and Ubc6-TA). For this analysis, we focused on the set of EMC clients defined above and compared their enrichment to the full set of TMD containing proteins (see TMD annotation in Materials and methods) (***Figure 3D***). Consistent with previous observations (***Jan et al., 2014***), we observed that all of the ER-localized BirAs showed an initial enrichment ~60 codons after the first TMD which likely represents the recruitment of the translating message to the ER by the signal recognition particle (SRP). With the exception of the EMC clients when monitored by Emc5-BirA, this enrichment then levels off or decreases modestly (***Figure 3D***). By contrast, for Emc5-BirA, the EMC clients, but not the full set of TMDs, showed a continued increase in enrichment such that the maximum was achieved following synthesis of an additional ~100 amino acids. These results further indicate that cotranslational EMC association with clients continues after initial targeting, and generally following synthesis of clusters of TMDs (***Figure 3D*** and ***Figure 3—source data 1***).

Since EMC-specific enrichment peaks followed TMD clusters, we analyzed the number and amino acid composition of TMDs in the 51 putative client proteins. Strikingly, the full list of EMC clients was strongly enriched for multipass proteins (***Figure 4A***) and EMC-interacting TMDs were enriched for charged amino acids and depleted for hydrophobic amino acids, including aliphatic residues common in TMDs (***Figure 4B***). Gene ontology term enrichment on the set of putative EMC clients showed enrichment for terms related to transporters; indeed, a majority of clients are classified as transporters (***Figure 4C***). Notably, integral membrane glycosyltransferases (which must overcome the challenge of transferring hydrophilic sugars onto membrane-associated substrates), including beta-glucan synthase genes (FSK1 and GSC2) defined a second class of putative EMC clients

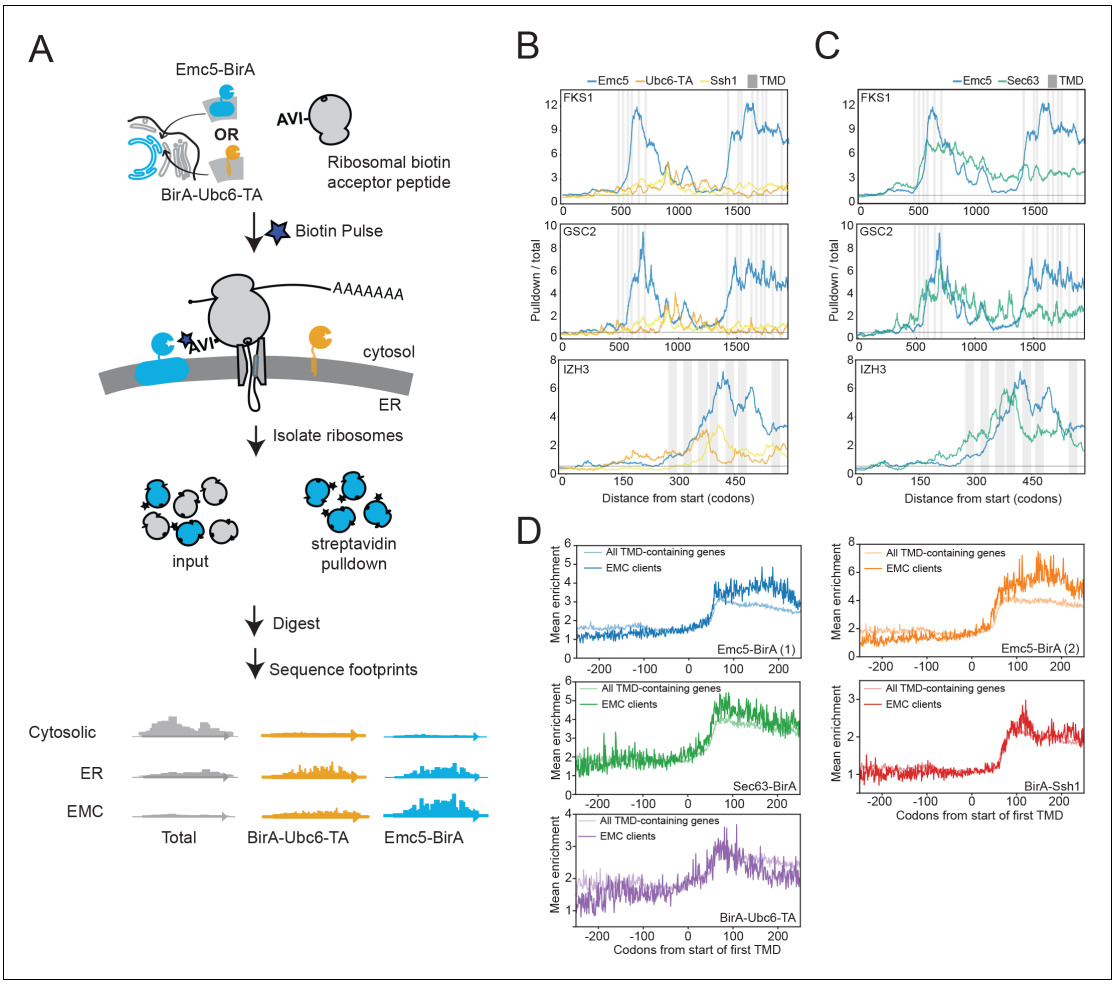


Figure 3. The EMC engages client proteins cotranslationally. (**A**) Schematic for strategy to examine the role of the EMC (Emc5-BirA) in cotranslational interaction with clients using proximity-specific ribosome profiling. (**B**) Positional enrichment plots showing footprint reads across the full-length mRNAs of the genes indicated for Emc5-BirA, BirA-Ubc6-TA and Ssh1-BirA expressing strains. Transmembrane domains (TMDs) are indicated in gray. (**C**) As in (**B**), except comparing Emc5-BirA and Sec63-BirA expressing strains. (**D**) Mean enrichments of all TMD-containing genes and EMC clients (N = 51) following start of first TMD for two independent replicates of Emc5-BirA, Sec63-BirA, BirA-Ssh1 and BirA-Ubc6-TA.

DOI: https://doi.org/10.7554/eLife.37018.004

The following source data and figure supplement are available for figure 3:

Source data 1. Positional enrichment plots across genes in the >90th percentile for Emc5-BirA/Ubc6-BirA 101 codon window enrichments.

DOI: https://doi.org/10.7554/eLife.37018.006

Figure supplement 1. Maximum 101-codon EMC/Ubc6 enrichment ratio windows from two independent Emc5-BirA replicates.

DOI: https://doi.org/10.7554/eLife.37018.005

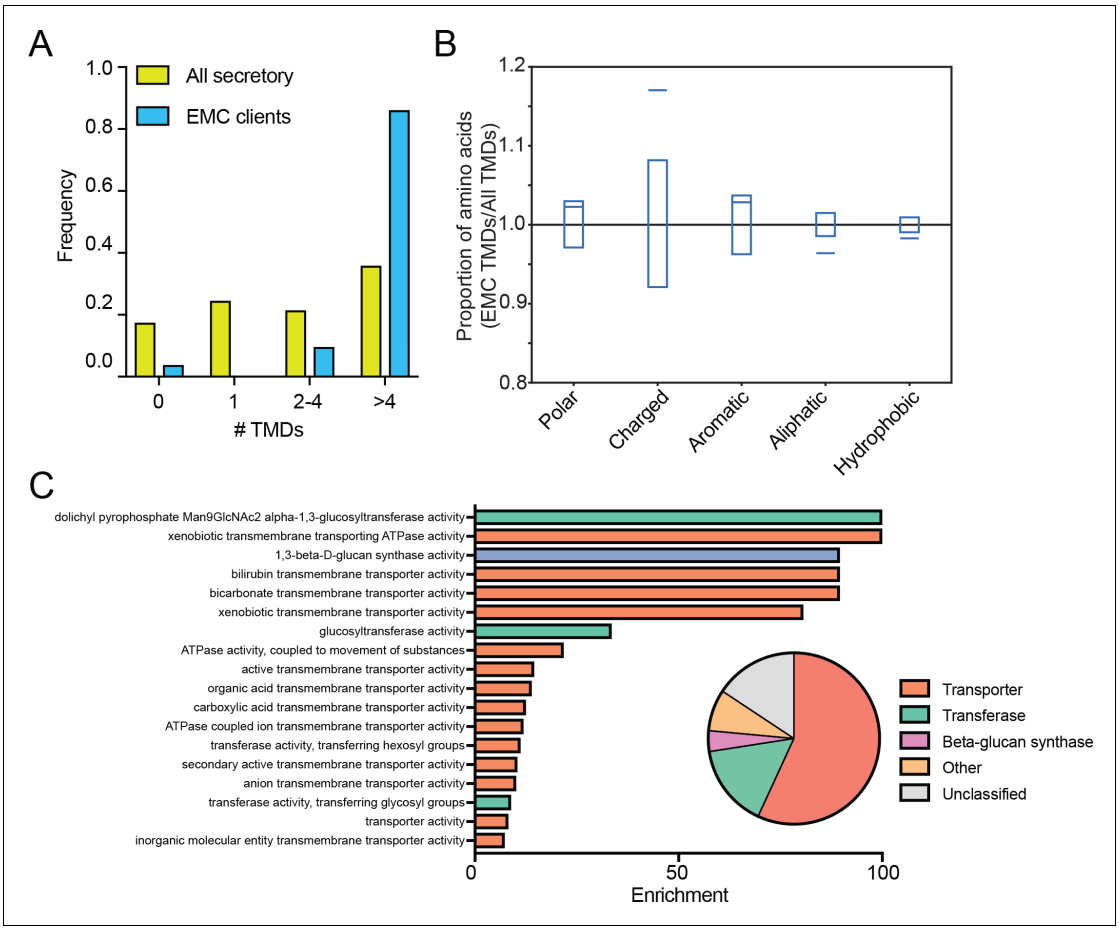


Figure 4. EMC clients are enriched for multipass membrane proteins with sub-optimal transmembrane helices. (**A**) Histogram showing the proportion of proteins containing the given number of TMDs for all proteins that enter the secretory pathway (Uniprot annotations) and EMC clients. (**B**) Fraction of amino acids with the given properties in TMDs from EMC client proteins compared to all secretory proteins. Proportion of EMC TMD amino acids/all TMDs for each property is shown by a blue line. Blue boxes indicate 95% confidence ranges defined by 10,000 random sub-samplings of total TMDs with a pool size equal to EMC TMDs (N = 51). (**C**) Top non-redundant over-represented GOMF terms calculated from PANTHER (FDR < 0.05; redundant terms removed by REViGO). Inset: PANTHER protein classifications pie chart for all clients (N = 51).

DOI: https://doi.org/10.7554/eLife.37018.007

(*Figure 4C*). Together, these results suggest that the EMC cotranslationally engages certain multipass membrane proteins and that these clients are enriched for biochemical features and functions that pose a challenge to the general membrane protein biogenesis machinery of the ER.

Select multipass membrane proteins are EMC clients in mammalian cells

Having identified EMC clients and characterized their properties in yeast, we used an orthogonal approach to identify potential EMC clients in mammalian cells. We reasoned that the EMC may play a role in stabilizing select multipass membrane proteins during synthesis until substrate-specific and general chaperones along with select binding partners, are recruited. This ensemble might then be

required to achieve a structure competent for export or function in the ER. We therefore hypothesized that depletion of the EMC in mammalian cells would result in destabilization and degradation of clients, which could be quantified by global mass spectrometry. Accordingly, we used unbiased quantitative SILAC proteomics to identify putative EMC clients on a proteome-wide scale. Using the CRISPRi system (**Qi et al., 2013**; **Gilbert et al., 2013**), we generated two independent EMC-depleted cell lines in which the expression of either EMC2 or EMC4 was strongly reduced. Whereas EMC2 depletion affects the abundance of all detected members of the EMC complex, depletion of EMC4 has no effect on other EMC members (**Figure 5A**). We thus infer that EMC2 but not EMC4 is structurally integral to the formation of the EMC complex. However, based on the strong similarity in the spectrum of genetic interactions in yeast, *EMC2* and *EMC4* deleted yeast strains, it is likely that the function of the EMC depends on both EMC2 and EMC4 (*9*).

Proteomic changes in these two EMC-depleted cell lines were measured against a control cell line expressing a non-targeting guide RNA (GAL4). Except for EMC components, almost all observed changes were strongly correlated between the EMC2 and EMC4 knockdown cell lines (**Figure 5B**). This supports the notion that depletion of EMC2 and concomitant depletion of the rest of the complex has few additional effects relative to depletion of EMC4 alone, beyond affecting the complex. The identified depleted proteins represent potential EMC client proteins that are destabilized and degraded in the absence of the EMC. To confirm that these clients do not arise from a decreased rate of synthesis, we performed ribosome profiling on EMC2-depleted and control cells, and compared changes in the rate of protein synthesis to changes in steady state proteome abundance (**Figure 5C**). This analysis revealed that the translation rate of the vast majority of potential client proteins is unaffected (with the notable exception of ATP6V0A1, which is decreased at both the translational and protein level), consistent with post-translational degradation. By contrast, proteins that are upregulated at the protein level are also upregulated at the translational level, indicating transcriptional and/or translational induction. As expected from EMC genetic interactions, 2D annotation enrichment (**Cox and Mann, 2012**) shows that proteins upregulated at the translatome and proteome-level are enriched for gene ontology terms related to the unfolded protein response (**Figure 5—figure supplement 1**) (**Jonikas et al., 2009**). By contrast, analysis of the proteins degraded upon EMC-depletion reveals an enrichment for proteins that enter the secretory pathway (**Supplementary File 1**). Moreover, of the 11 proteins decreased by 2-fold or more in both EMC2- and EMC4-depleted cells, 10 have at least one transmembrane domain. Additionally, only 8 of the 37 total hits are soluble proteins with the majority of soluble proteins (5) being localized to the lysosome, and thus likely secondary to depletion of the multipass subunit of the V-ATPase (ATP6V0A1) (**Supplementary file 1**). Confirming a role for the EMC in TA protein insertion (**Guna et al., 2012018**), the levels of mammalian squalene synthase (FDFT1) and TREX1 (*35*) decreased upon EMC depletion. However, similar to that seen in yeast, the cohort of putative EMC client proteins is strongly enriched for multipass transmembrane proteins (**Figure 5D**) and, like the yeast clients, are more likely to contain charged and aromatic residues distributed throughout the TMD (**Figure 5E** and **Figure 5—figure supplement 2**). Further mirroring the yeast EMC client profile, GO terms associated with the identified human clients were enriched for transporter related terms which constituted the largest class of proteins whose abundance decreased upon EMC depletion (**Figure 5F**).

Together, these data support a conserved role for the EMC in facilitating the biogenesis of multipass membrane proteins with destabilizing membrane spanning sequences. Depletion of the EMC renders some client proteins unable to attain normal expression levels.

Discussion

Our work establishes that the predominant function of the EMC is to ensure the biogenesis of a subset of multipass membrane proteins. These studies are consistent with a model in which the EMC cotranslationally interacts with nascent polypeptides. Following synthesis, the EMC stabilizes clients and enables recruitment of substrate-specific and general chaperones to achieve a conformation that is competent for function in the ER or for transport to the Golgi (**Figure 6**). In the absence of the EMC, newly synthesized client proteins are likely extracted from the membrane for degradation by the UPS (**Figure 6**).

Our studies reveal three principles of the EMC's action. First, the EMC directly interacts with and stabilizes a range of client proteins consisting primarily of multipass transmembrane proteins. Several

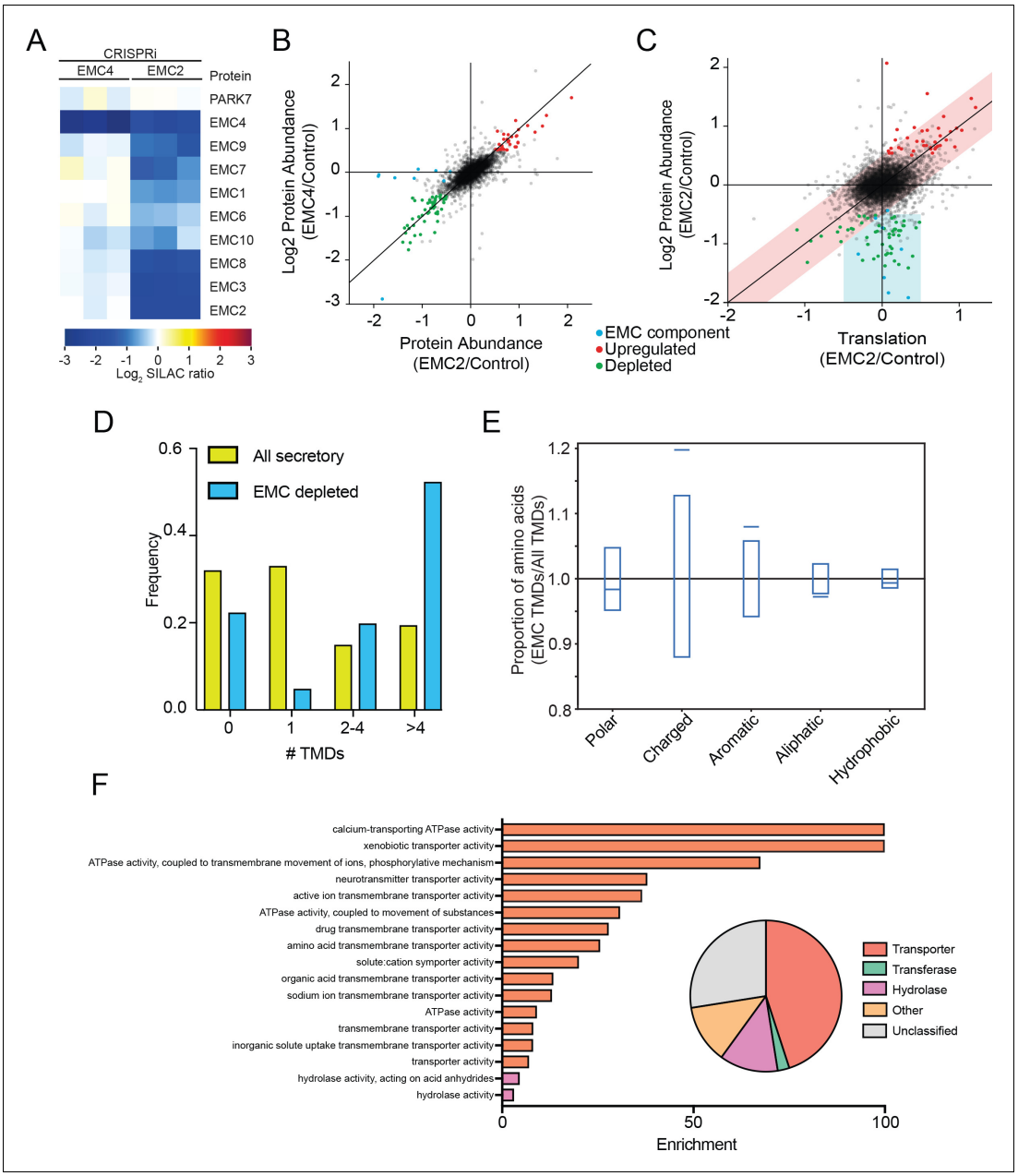


Figure 5. The mammalian EMC stabilizes multipass transmembrane proteins. (**A**) SILAC quantification of EMC components, comparing their expression level in cells with guide RNAs targeting EMC2 or EMC4 with those expressing non-targeting GAL4 guide RNA. PARK7 is shown as a control (***Wiśniewski and Mann, 2016***). (**B**) Full proteome comparison scatter plot of protein abundance change in cells expressing EMC2 guide RNA against abundance change in cells expressing EMC4 guide. Expression is relative to non-targeting GAL4 guide RNA. Proteins colored red are significantly

Figure 5 continued on next page

Figure 5 continued

upregulated in both EMC2 and EMC4 cells. Proteins colored green are significantly downregulated in both EMC2 and EMC4 cells (Log2 >0.5). EMC components are colored blue. (**C**) Comparison of translation change by ribosome profiling with proteome change. (**D**) Histogram showing the proportion of proteins containing the given number of TMDs for all proteins that enter the secretory pathway (as defined in Uniprot) and EMC clients. (**E**) Fraction of amino acids with the given properties in TMDs from EMC client proteins compared to all secretory proteins. Proportion of EMC TMD amino acids/all TMDs for each property is shown by a blue line. Blue boxes indicate 95% confidence ranges defined by 10,000 random sub-samplings of total TMDs with a pool size equal to EMC TMDs (N = 37). (**F**) Top non-redundant over-represented GOMF terms calculated from PANTHER (FDR < 0.05; redundant terms removed by REViGO). Inset: Protein classifications pie chart for EMC client proteins (N = 37).

DOI: https://doi.org/10.7554/eLife.37018.008

The following figure supplements are available for figure 5:

Figure supplement 1. 2D annotation enrichment based on the protein ratios and ribosome profiling ratios of EMC knockdown versus control knockdown cells.

DOI: https://doi.org/10.7554/eLife.37018.009

Figure supplement 2. Amino acid composition of transmembrane domains of EMC clients and background.

DOI: https://doi.org/10.7554/eLife.37018.010

observations support this conclusion. The EMC physically associates with multipass proteins in yeast (***Figure 1 A and B***). Deletion of an EMC component and dedicated membrane protein chaperone (SOP4) results in increased interaction with several multipass membrane proteins that transit the ER, including a previously identified Sop4 substrate (Pma1) (***Luo et al., 2002***) and a membrane protein

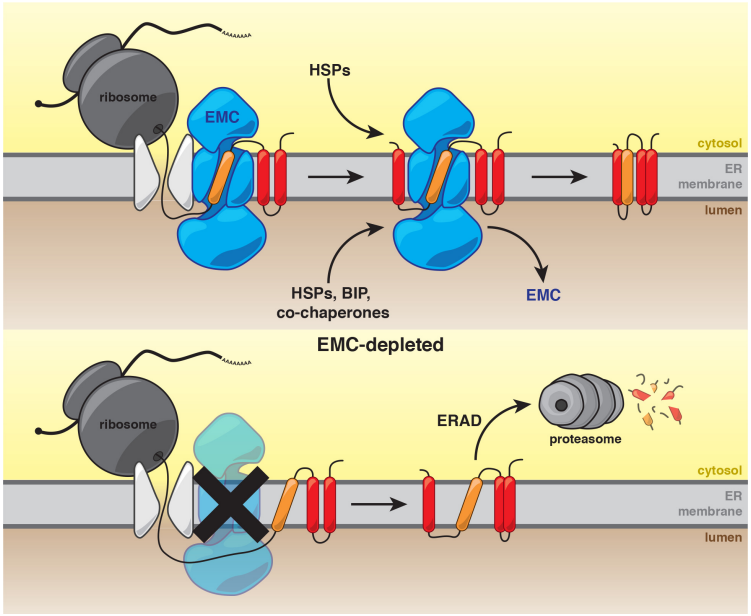


Figure 6. Model for the role of the EMC in multipass transmembrane protein biogenesis. See text for details. Unstable transmembrane domains are shown in orange. Note, while the EMC is depicted here as cooperating with the translocon following insertion, our data do not exclude the possibility that the EMC acts as an insertase for some substrates.

DOI: https://doi.org/10.7554/eLife.37018.011

previously shown to require the EMC to localize to the cell surface (Mrh1) (**Bircham et al., 2011**) (**Figure 2B**). In addition, we show that the poorly characterized ER resident protein Ilm1 acts as a substrate-specific chaperone for Fks1 (**Figure 1 C and D**), and ILM1 deletion results in increased interactions between the EMC and Fks1 (**Figure 2B**). These yeast studies show that the EMC directly interacts with multipass protein clients early in biosynthesis independent of, and likely prior to, engagement by their dedicated chaperones. Multipass proteins are also selectively destabilized in human cells depleted of the EMC by CRISPRi (**Figure 5**), further demonstrating that the biogenesis/chaperoning of multipass membrane proteins is a conserved feature of EMC function.

Second, the EMC can begin to interact cotranslationally and subsequently stabilize newly synthesized client proteins, likely preventing premature degradation by ERAD. Proximity-specific ribosome profiling in yeast revealed that a common feature of the EMC is cotranslational engagement of multipass client proteins (**Figure 3**). The presence of full-length multipass membrane proteins in our pulldown analyses in yeast (**Figures 1** and **2**) indicate that the EMC can engage clients cotranslationally and remain bound after completion of protein synthesis for at least a subset of client proteins. Although the EMC was recently shown to act as a TA protein insertase (**Guna et al., 2018**), our results suggest that the EMC interacts with multipass proteins during synthesis and remains associated post-ER targeting (**Figure 3D**). In addition, all of the EMC clients identified in yeast are dependent on SRP, a factor that cotranslationally delivers substrates to the ER surface, and are cotranslationally targeted to the ER (**Costa et al., 2018**). Similarly, all of the mammalian multipass protein EMC substrates were previously found to be cotranslationally targeted to the ER by proximity-specific ribosome profiling studies (**Jan et al., 2014**). These results, however, do not exclude the possibility that the EMC acts to insert or facilitate flipping of individual helices of multipass membanes proteins following initial targeting to the ER, or that the EMC acts during the initial insertion for a subset of the newly identified client proteins. Overall, our results support a model in which the EMC can engage client proteins cotranslationally, perhaps following insertion via the translocon into the ER membrane, likely acting to stabilize folding intermediates and to forestall degradation.

Third, the EMC engages transporters and other client proteins enriched for sub-optimal residues in transmembrane helices. Our proximity-specific ribosome profiling results indicate that the EMC typically engages transporters and other membrane proteins enriched for charged residues in TMDs (**Figure 4 B and C**) following synthesis of TMD clusters (**Figure 3B**). In addition, membrane proteins that were destabilized in human cells were also enriched for transporters (**Figure 5F**), and human EMC clients were enriched for charged and bulky residues in TMDs (**Figure 5E**). These observed features of EMC clients are inter-related, multipass membrane proteins are more likely to have sub-optimal helices and are enriched for transporter related functions. Indeed, many transporters contain polar and charged residues within TMDs, which are necessary for solute delivery across the lipid bilayer. While our studies in HeLa cells primarily identified solute carriers and transmembrane ATPases as clients, the EMC was previously implicated in the biogenesis of cell-type specific multimeric channel proteins that also contain TMDs with charged residues (**Richard et al., 2013**). Thus, we propose that the EMC promotes the biogenesis of a wide range of human multipass client proteins by stabilizing transmembrane regions during biosynthesis and prior to completion of folding.

How does the EMC act both as a cotranslational TMD chaperone, following ER targeting, and as a post-translational insertase for a subset of TA proteins (**Guna et al., 2018**)? We suggest that these phenomena reflect a common biochemical property of the EMC: the ability to interact with transmembrane helices with low hydrophobicity. This property is consistent with the previously reported role for the TMD of EMC1 in binding to and stabilizing a destabilized, ER membrane embedded SV40 virion (**Bagchi et al., 2016**). Therefore, we propose that a unifying function of the EMC is to accommodate and stabilize the wide diversity of membrane spanning sequences by directly interacting with select membrane proteins with destabilizing features in TMDs. Interestingly, EMC3 may share a common ancestry with the universally conserved YidC/Oxa1/Alb3 protein family in bacteria and mitochondria (**Anghel et al., 2017**). YidC also plays a dual role in both the insertion of membrane proteins (**Samuelson et al., 2000**) and the stabilization of membrane proteins inserted via the translocon by direct interaction with SecYEG (**Nagamori et al., 2004**).

Beyond its role in engaging atypical TMDs, our studies point to a broader role for the EMC as a nexus of TMD folding, which requires chaperone recruitment and protection from the ERAD machinery. The EMC is a large protein complex with significant predicted mass integral to the membrane, as well as soluble mass in the cytosol and lumen. It is counterintuitive that such mass is necessary to

act solely as a transmembrane chaperone and insertase. For example, members of the YidC/Oxa1/Alb3 membrane protein family in bacteria and mitochondria act as insertases and transmembrane chaperones, but they function as smaller monomers without appended soluble domains (**Nagamori et al., 2004**). Indeed, we find that in yeast the EMC also recruits substrate-specific chaperones including Ilm1, Sop4 and Gsf2, as well as general cytoplasmic and lumenal chaperones and oxidoreductases. We propose that, in analogy to the recently described function of the Slp1-Emp65 complex for soluble lumenal ER proteins (**Zhang et al., 2017**), the EMC may also hold folding intermediates in a privileged, ERAD-protected state until the recruitment of substrate-specific chaperones and/or general chaperones to complete folding and allow ER export. Thus, the EMC may compartmentalize multiple functions necessary for membrane protein biogenesis: transmembrane stabilization, recruitment of folding factors and protection of folding intermediates from recognition and degradation by the ER quality control machinery.

Materials and methods

Yeast strains and plasmids

Strains BY4741 (*MATa his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0) and W303 (*ade2-1 leu2-3 his3-11,15 trp1-1 ura3-1 can1-100*) were used as wild-type parental strains. Genomic knockouts and knockins were generated by one-step gene replacement as described (**Rothstein, 1991**). Generation of EMC3-3X-FLAG (BY4741: EMC3-FLAG-Nat^R) was described previously (**Jonikas et al., 2009**). To generate FLAG-tagged Ilm1p, the *ILM1* coding sequence including the open reading frame and ~350 base pairs of the 3'-UTR was amplified from genomic DNA extracted from BY4742 wild-type yeast. The resulting PCR products were inserted into a plasmid immediately upstream of an in-frame fusion with 3X-FLAG and the *NATMX6* coding sequence. Next, a 3XFLAG epitope was introduced at the 3' end of the *ILM1* open reading frame in pILM1-UTR-NAT using a modified site-directed mutagenesis protocol (**Wang and Malcolm, 1999**). For homologous recombination, the resulting *ILM1-3XFLAG-UTR-NAT* sequence was amplified and transformed into BY4741. Positive clones were selected on yeast extract-peptone-dextrose (YEPD) medium supplemented with nourseothricin and analyzed for FLAG-tagged Ilm1p expression by immunoblotting with anti-FLAG antibodies (Santa Cruz Biotechnology).

Immunoprecipitation from microsomes

Immunoprecipitations of Emc3-3xFLAG and Ilm1-3xFLAG were performed as described previously (**Denic and Weissman, 2007**). Yeast were grown in 3L YEPD, harvested and resuspendend in 2 ml lysis buffer (50 mM HEPES-KOH pH 6.8). Resuspended yeast pellets were frozen dropwise in liquid nitrogen and subsequently disrupted by bead beating. 15 ml lysis buffer was added to the frozen yeast powder lysis was performed by 10 strokes in a Dounce homogenizer. The homogenate was centrifuged at 1000Xg for 10 min at 4˚C. The supernatant was transferred to a new tube and the centrifugation was repeated. The supernatant was then transferred to a 50.2 Ti ultracentrifuge tube (Beckman Coulter) and centrifuged at 22,000 RPM for 16 min at 4˚C. The microsome pellet was resuspended in 0.5 ml lysis buffer, flash frozen in liquid nitrogen and stored at −80˚C until use.

Microsomes were solubilized in 15 ml immunoprecipitation buffer (100 mM HEPES-KOH pH 6.8, 300 mM KOAc, 4 mM MgOAc, 2 mM $CaCl_2$, 30% glycerol) for 1 hr at 4˚C. Detergent extracted microsomes were centrifuged for at 22,000 RPM for 16 min at 4˚C. 150 µl of anti-FLAG bead slurry was added the supernatant and incubated for 2 hr at 4˚C. Beads were pelleted at 1,000Xg for 1 min and washed 4 times in 10 ml wash buffer (immunoprecipitation buffer with 0.1% digitonin). Beads were eluted with 150 µl of 2 mg/ml FLAG peptide in was buffer. Eluates were stored at −80˚C until use.

In-gel tryptic digestion and mass spectrometry

Immunoprecipitation eluates were separated by SDS-PAGE and single bands were excised, or alternatively, gels were cut along the lanes in 11 to 13 pieces. Proteins were subjected to in-gel digestion (University of California San Francisco Mass Spectrometry Facility protocol) with trypsin (porcine, side-chain protected, Promega). The extracted digests were vacuum evaporated and resuspended

in 0.1% formic acid in water. The digests were analyzed by capillary HPLC-tandem mass spectrometry. The separation was performed with a C_{18} PepMap 75 µm × 150 mm column (LC Packings, Sunnyvale, CA) used on either an Ultimate HPLC system linked with a FAMOS autosampler (LC Packings, San Francisco, CA) or an Agilent 1100 series HPLC system equipped with an autosampler (Agilent Technologies, Palo Alto, CA). The column effluent was directed to either a QSTAR-Pulsar or QSTAR-Elite tandem mass spectrometer (Applied Biosystems/MDS Sciex, Toronto, Canada). Throughout the chromatographic separation, a 1 s MS acquisition was followed by a 3 s CID acquisition for computer-selected precursor ions in information-dependent acquisition mode. The collision energy was set according to the *m/z* value and charge state of the precursor ion.

Data was analyzed with Analyst QS 1.1 software (Applied Biosystems/MDS Sciex) and peak lists were generated using the mascot.dll script (Mascot.dll 1.6b18, Applied Biosystems). Precursor mass tolerance for grouping was set to 0.2 amu. MS centroiding parameters were 50% peak height and 0.02 amu merge distance. MS/MS centroiding parameters were 50% peak height and 0.05 amu merge distance.

The peak lists were searched in in-house Protein Prospector version 5.3.0 (a public version is available on line). Peptides containing one miscleavage were allowed. The number of modifications was limited to two per peptide. Carbamidomethylation modification of cysteine; acetylation of the N terminus of the protein; oxidation of methionine; and formation of pyro-Glu from N-term Gln were allowed as variable modifications. Mass tolerance for was 150 ppm for precursor and 0.2 Da for fragment ions.

Proximity-Specific ribosome profiling

Proximity-based ribosome profiling was performed essentially as previously described (***McGlincy and Ingolia, 2017***).

Strain construction

The endogenous copies of RPL10a were C-terminally tagged with an engineered HA-TEV-AviTag sequence to allow for detection by western blot, biotinylation, and specific elution after streptavidin pulldown via TEV protease cleavage. BirA fusion proteins EMC5 and SSH1 were endogenously tagged at the C-terminus (EMC5) or N-terminus (SSH1), respectively with a BirA (biotin ligase), allowing the specific biotinylation and streptavidin pull-down of ribosomes in close proximity to the EMC specifically or to the ER. Generation of Sec63-Bir was previously described (***Jan et al., 2014***).

Media and growth conditions

Yeast were grown in biotin-free, synthetic defined media (1.7 g/L YNB-Biotin [Sunrise Science Products], 5 g/L Ammonium sulfate, 20 g/L dextrose, complete amino acids) supplemented with d-biotin (Sigma) to a final concentration of 0.125 ng/mL, at 30°C with vigorous shaking. Twenty milliliters of an overnight culture was used to inoculate a 300 ml culture at an OD600 of 0.05–0.1, and biotin induction was performed at mid-log phase with an OD600 of 0.5–0.6 as in (***Jan et al., 2014***).

Biotin induction and harvesting

Cyclohexamide (CHX) was added to media 2 min prior to the addition of biotin, at a final concentration of 100 µg/mL. To induce biotinylation, D-biotin was added to the media to a final concentration of 10 nM and biotinylation was allowed to proceed for 2 min at 30°C while shaking. After 2 min, cells were harvested by filtration onto 0.45 µm pore size nitrocellulose filters (Whatman), scraped from the membrane, and immediately submerged in liquid nitrogen.

For western-blot quantification, 1 mL aliquots were taken from uninduced cultures and placed into pre-chilled, 1.5 mL siliconized microcentrifuge tubes. Samples were then spun at 20,000 x g at 4°C for 30 s, the supernatant removed, and the pellet-containing tubes immediately placed in liquid nitrogen. For levels of biotin induction quantification, a small patch of induced, filtered cells were scraped from the nitrocellulose filters.

Western blotting and biotinylation quantification

Lysates were prepared from pelleted induced and uninduced yeast by resuspending frozen pellets in 30–50 µL Laemmli buffer, followed by denaturation at 95˚C for 5 min, and clarification at room temperature by spinning at 20,000 x g for 10 min.

Lysates were run on 4–12% Bis-tris gels in MOPS buffer, transferred to nitrocellulose membrane using the BioRad Transfer system (BioRad) according to the manufacturer's instructions, blocked with Odyssey blocking buffer, and subsequently probed. The HA epitope tag was detected using a mouse a high-affinity rat anti-HA antibody at a 1:1000 dilution (Roche 3F10). IR800 anti-rat (Rockland) secondary antibody was then used at 1:5000 dilution. Biotin was detected directly using Streptavidin AlexaFluor 680 (Molecular Probes) at a 1:5000 dilution in TBS-T and a 10 min incubation period after incubation in secondary antibody. All blots were visualized using the Licor (Odyssey) system.

Percent biotinylation was quantified by probing for HA in a streptavidin shift assay, in which clarified lysates were mixed with excess unlabeled streptavidin (Rockland) prior to electrophoresis and immunoblotting. Biotinylated AviTags shift to a higher molecular weight than the corresponding, non-biotinylated AviTags, and percent biotinylation was quantified from the fraction of total signal that was shifted (***Algire et al., 2002***).

Yeast lysis, lysate desalting and monosome isolation

650 µL of polysome lysis buffer (20 mM Tris pH 8.0, 140 mM KCl, 1.5 mM MgCl2, 100 µg/mL CHX, 1% Triton X-100) was dripped into a 50 mL conical tube filled with and immersed in liquid nitrogen, containing the harvested yeast strip/pellet from a mid-log phase 300 mL biotin-induced culture. The frozen cell-buffer mixture was cryogenically pulverized for a minute in a freezer mill. Pulverized cells were thawed and centrifuged for 2 min at 4˚C and 20,000 x g. The supernatant was immediately loaded onto pre-chilled, 2 mL Zeba de-salt spin column previously equilibrated with polysome gradient buffer (20 mM Tris pH 8.0, 140 mM KCl, 5 mM MgCl2, 100 µg/mL CHX, 0.5 mM DTT) according to the manufacturer's instructions. Aliquots of this extract were flash-frozen in liquid nitrogen and stored at −80˚C, typically yielding 0.5–1 mL of extract with A260 of 100–300. A 200–500 µL aliquot of the above lysate was treated with 750 U RNaseI (Ambion) per 50 A260 units of lysate (or 15 U RNaseI per 40 µg RNA, where 1 A260 unit corresponds to 40 µg RNA), and incubated for 1 hr at room temperature on an overhead roller. Reactions were then quenched with 10 µL SUPERase-In RNase inhibitor (Ambion) and stored on ice until loaded onto sucrose density gradients (10–50% w/v) prepared with the polysome gradient buffer described above. Gradients were made in Sw-41 ultracentrifuge tubes (Seton Scientific) using a BioComp Gradient Master (BioComp Instruments) according to the manufacturer's instructions. Samples were spun for 3 hr at 4˚C and 35,000 rpm in an Sw-41 rotor (Beckmann Coulter). Fractionation was performed on the Gradient Master using a BioRad EM-1 Econo UV monitor to continually monitored A260 values. Monosome peaks were collected, flash-frozen in liquid nitrogen, and stored at −80˚C. Typical yields were 2–3 mL of monosomes with A260 of 2–5.

Streptavidin pulldown of biotinylated ribosomes

Biotinylated ribosomes were isolated from the total monosome fraction using MyOne streptavidin C1 magnetic DynaBeads (Invitrogen). The volume of beads used per pulldown was scaled based on 187 µL (1.87 mg) beads per 15 pmol of biotinylated ribosomes, as estimated from the manufacturer's instructions. The pmol of biotinylated ribosomes in a given volume was calculated from (i) the fraction of biotinylated ribosomes as estimated from a streptavidin shift assay and (ii) the total concentration of 80S ribosomes in the fraction (***Jan et al., 2014***). Prior to binding, beads were washed twice with one volume (equal to the initial bead volume) of Buffer A (100 mM NaOH, 50 mM NaCl), once with one volume of Buffer B (100 mM NaCl), and once with one volume of low-salt binding Buffer C (20 mM Tris pH 8.0, 140 mM KCl, 5 mM MgCl2, 100 µg/mL CHX, 0.5 mM DTT, 0.1% Triton X-100). Triton X-100 was added to monosome fractions containing 15 pmol of biotinylated ribosomes, to a final concentration of 0.01%. This solution was added to washed beads, mixed well, and the pulldown was allowed to proceed on an overhead roller for 1 hr at 4˚C. The supernatant was removed and the beads were washed three times with 1 mL high-salt wash Buffer D (20 mM Tris pH 8.0, 500 mM KCl, 5 mM MgCl2, 100 µg/mL CHX, 0.5 mM DTT, 0.1% Triton X-100), each for 20 min

at 4˚C. After the third wash, beads were re-equilibrated in low-salt Buffer C by resuspension in 1 mL, then transferred to a new tube and resuspended in a smaller volume (200 µL) of Buffer C in preparation for elution by TEV protease cleavage. Cleavage was performed by incubation on a nutator with in-house TEV protease for 1 hr at room temperature. Three volumes of Trizol LS (Ambion) were added to both the TEV eluate and a separate, matched input sample consisting of 10–20 pmol of total monosomes.

Library generation

Ten-twenty pmol of monosomes in Trizol LS were extracted using 200 µL chloroform per 750 µL Trizol LS. RNA was precipitated for at least 1 hr at −30˚C (or 30 mins at −80) using GlycoBlue (Invitrogen) and an equal volume of isopropanol, pelleted, resuspended in 11 µL (input) or 5 ul (pulldown) 10 mM Tris pH 7.0, and resolved on a 15% TBE-urea gel. Samples were denatured in 2X TBE-Urea loading buffer at 80˚C for 2 min. Gel was run at 200V for 60 min and visualized after 5 min incubation with SYBR Gold (Invitrogen). Oligoribonucleotide size standard in neighboring lanes was used to excise roughly 28 nt ribosome footprints. Footprints were passively eluted on a tube nutator overnight at 4˚C in 420 µL 0.3 M NaCl after crushing gel slices. After overnight RNA elution from gel, ribosome footprints were then precipitated with GlycoBlue and 2.5 volumes ethanol, resuspended directly in a dephosphorylation master mix containing 8 µL 1.25x T4 polynucleotide kinase (PNK) buffer (New England Biolabs, NEB), and dephosphorylated with 2 µL PNK for 1 hr at 37˚C. This solution was used directly for ligation to 0.5 µg 3' miRNA cloning linker 1 (Integrated DNA Technologies) upon addition of 8 µL 50% PEG (NEB), 1 µL 10x truncated T4 RNA ligase 2 K227Q (rnl2) buffer (NEB), and in-house rnl2 enzyme as in (***Kopito, 1999***) (or 3.5 ul 50% PEG, 0.5 ul µL 10x T4 RNA ligase buffer, 0.5ul T4 RNA ligase rnl2, and 0.05 ul 1M DTT, and a sample-specific barcode linker as in [***McGlincy and Ingolia, 2017***]). Ligation proceeded for 3 hr at 25˚C at which point RNA was precipitated for at least 1 hr at −30˚C, purified on a 10% TBE-urea gel, eluted, and precipitated as above.

rRNA contaminants were removed from ligation products using antisense biotinylated oligos as described (***Brar et al., 2012***). rRNA-depleted ligation products were then reverse-transcribed in a 16.7 µL reaction using SuperScript III (Invitrogen) for 30 min at 48˚C. RNA template was hydrolyzed for 20 min at 98˚C after addition of 1/10 vol 1 M NaOH. Equi-molar HCl was added to quench the reaction and cDNAs were precipitated at −30˚C for at least 1 hr and subsequently purified on a 10% TBE-urea gel, eluted overnight, precipitated, and resuspended in 15 µL nuclease-free water.

cDNAs were circularized using CircLigase (Epicentre) in a 20 µL reaction for 1.5 hr at 60˚C according to the manufacturer's instructions. Circularized products were amplified by 8–16 cycles of PCR using oNTI231 and any of several Illumina indexing primers (IDT) using Phusion polymerase (Finnzymes) in a 17 µL reaction. PCR amplicons were gel purified on 8% non-denaturing TBE gels, eluted, precipitated, resuspended in 10 µL EB, and quantified using the Bioanalyzer High Sensitivity DNA assay (Agilent Technologies). 2 nM dilutions were multiplexed as needed and sequenced via a single-end run on an Illumina HiSeq sequencer.

Identification of co-translationally engaged regions

Sequencing reads were trimmed of adaptor sequences and aligned to yeast coding sequences as previously described (***Hussmann et al., 2015***). To compute the efficiency with which ribosomes translating a particular region of a coding sequence were labeled by a BirA fusion, for each codon position in each coding sequence, for both pulldown and input samples, we calculated the sum of ribosome profiling reads in a 50 codon window on either side (i.e. 101 total codons) of the position normalized to the total number of mapped reads for the sample, then computed the enrichment ratio of pulldown reads to input reads in the window. To quantify the extent to which a gene has any region for which ribosomes translating that region are more efficiently labeled by Emc5-fused BirA than a BirA fused with the tail-anchor from Ubc6 (BirA-Ubc6-TA), we computed the maximum value of the ratio of (pulldown/total enrichment for Emc5-BirA) to (pulldown/total enrichment for BirA-Ubc6-TA) for each position in the gene.

The following filters were applied to restrict the set of relevant genes:

- Uniquely mappable: to exclude artifacts from genes that are too similar in sequence, genes were required to have >80% of positions in the coding sequence uniquely mappable when tested with synthetic 25 nt reads.
- Expression cutoff: to exclude noise from very lowly expressed genes, genes were required to have higher than 0.02 reads per codon per million reads in the input (non-pulldown) ribosome profiling for every BirA, and to have no 101 codon window in which there were 0 total reads in any ribosome profiling sample.
- Localization: to exclude noise from genes that are not translated in the proximity of any BirA at all, genes were required to have RPKM(pulldown)/RPKM(input)>1 for at least one BirA ribosome profiling sample pair.

Amongst genes passing these filters, we identified potential EMC clients as those genes whose maximum positional enrichment ratio was in the top 10% of all genes in both biological replicates of the Emc5-BirA ribosome profiling.

Transmembrane domain annotations

Annotated transmembrane domains were collected from two sources: domains predicted by TMHMM in yeast were downloaded from SGD (date stamp on file: 8/23/2017), and domains annotated in UniProt for yeast and human were extracted from Uniprot (Reviewed Swiss-Prot) databases in xml format (file names uniprot-reviewed%3Ayes + taxonomy%3A4932.xml for yeast and uniprot-reviewed%3Ayes + taxonomy%3A9606.xml for human). For yeast, the sets of domains from both sources were merged, and when a TMHMM prediction overlapped with a domain in Uniprot, the Uniprot domain was chosen. Domains in mitochondrially encoded genes, dubious ORFs, and pseudogenes were excluded. Final sets of transmembrane domains considered are shown in ***Supplementary file 2***.

Amino acid composition of TMDs

To evaluate biochemical properties of TMDs in potential EMC clients, the set of all amino acids in all TMDs of EMC clients was collected, and the fraction of such amino acids that were aliphatic, aromatic, charged, hydrophobic, or polar was calculated. The same calculations were carried out for all annotated TMDs, and the ratio of fraction in TMDs in EMC clients to fraction in all TMDs was computed for each property. To assess statistical significance, random subsets of the same number of TMDs as in the total EMC client set were drawn from the set of all TMDs and the same ratio of fractions was computed. This process was repeated for 10,000 random subsets and the fifth and 95th percentile of the 10,000 ratios produced were recorded.

Targeted downregulation of gene expression by CRISPRi

CRISPRi HeLa cell lines were generated by transducing with pHR-SFFV-dCas9-BFP-KRAB (Addgene ID: 46911) and sorting for BFP positive cells (***Gilbert et al., 2013***). EMC2 (GAGTACGCG TCCGGGCCAA), EMC4 (GTCATTTCCGCCCTGGAAAT) and negative control Gal4-4 (GAACGAC TAGTTAGGCGTGTA) protospacers were cloned into a lentiviral expression plasmid expressing guides from a mouse-derived U6 promoter, BFP and puromycin (pU6-sgRNA EF1-Alpha-puro-T2A-BFP; Addgene ID: 60955). HeLa CRISPRi cells were transduced with guide RNA expression plasmids, selected in puromycin for 72 hr and either directly used for experiments or expanded for SILAC labeling. HeLa cell lines were confirmed by STR analysis and tested as free from mycoplasma contamination.

Mammalian ribosome profiling

Ribosome profiling was performed for HeLa-dCas9-KRAB cells, and HeLa-dCas9-KRAB cells expressing EMC2 or Gal4-4 control guide RNAs. Cells were cultured in 15 cm plates with Dulbecco modified eagle medium (DMEM) with 10% fetal bovine serum (Gibco) until ~80% confluency. Cells were treated with 100 ug/ml CHX for 2 min and then lysed using 500 µl per polysome lysis buffer (20 mM Tris pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 1% Triton x-100, 1 mM DTT, 8% glycerol, 100 µg/ml CHX, 24 U/ml Turbo DNase) per plate using a rubber cell scraper to facilitate lysis. Lysate was centrifuged for at 20,000 x g for 2 min at 4°C and the remaining cleared polysome-containing lysate was flash frozen by immersion in liquid nitrogen and stored at -80°C until digestion. $CaCl_2$ was added to

polysome-containing lysate to a final concentration of 5 mM and 30 µg was digested to monosomes using micrococcal nuclease (8 U/µg) for 1 hr at room temperature and the reaction was terminated by the addition of EGTA (6.25 mM). Digested lysates were equilibrated to 500 µl with polysome gradient buffer (20 mM Tris pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 1% Triton x-100, 1 mM DTT, 100 µg/ml CHX) and loaded on top of a sucrose cushion (polysome buffer containing 1.65 M sucrose) and ultracentrifuged in a TLA-110 rotor (Beckman Coulter) for 4 hours at 4°C. The monosome-containing pellet was resuspended in 700 µl Trizol (Life Technologies). Total RNA was extracted and libraries were prepared as described for proximity-based ribosome profiling.

Quantitative SILAC mass spectrometry

Cell culture and harvesting

All cell culture reagents were obtained from Gibco unless otherwise stated. Cells were cultured for at least seven doublings in SILAC DMEM supplemented with 10% Dialyzed FBS, 20 Units/mL Penicillin, 20 µg/ml Streptomycin, 1 mM Sodium Pyruvate, 10%, and 2 mM L-alanyl-L-glutamine dipeptide and either; 42 mg/L $^{13}C_6$,$^{15}N_4$-L-Arginine HCl (Silantes) together with 73 mg/L $^{13}C_6$,$^{15}N_2$-L-Lysine HCl (Silantes), or 42 mg/L Arginine HCl and 73 mg/L Lysine HCl with standard isotopic constituents (Sigma). Cells were harvested by rinsing twice in ice-cold PBS excluding Calcium Chloride and Magnesium Chloride. Cells from 1 × 10 cm dish were scraped in 1 ml of ice-cold PBS and transferred to a 1.5 mL Eppendorf tube, for centrifugation at 300 x *g* at 4°C. The PBS was aspirated and cells were resuspended in lysis buffer (4% SDS, 100 mM DTT in 100 mM Tris-HCl, pH 7.6 at room temperature) and heated to 95°C for 5 min. Samples were sonicated for 15 × 30 s cycles using a Bioruptor to reduce viscosity of the lysate. Protein concentrations were determined via tryptophan assay.

Filter-aided sample preparation (FASP)

Peptides were generated essentially as described (***Zielinska et al., 2010***). Protein lysate from Gal4 SILAC heavy-labelled cells and EMC2 or EMC4 SILAC light-labelled cells were mixed 1:1, and 100 µg of sample plus 250 µL Urea Buffer (8M urea, 100 mM Tris pH 8.5) was loaded onto Microcon 30 kDa MWCO centrifugal filters. Loaded filters were centrifuged at 10,000 x *g* at 18°C for 20 min. Filters were centrifuged a further two times with 250 µL Urea Buffer. Samples were alkylated at room temperature for 15 min by incubation with 50 mM Iodoacetamide in Urea Buffer. Samples were centrifuged a further three times for 15 min each with 150 µL Urea Buffer, before two times centrifugation with 150 µL digestion buffer (40 mM NH_4HCO_3). Finally 2 µg trypsin (Sigma) or 6.25 µg GluC (NEB) in 40 µL digestion buffer was added to the filters and incubated overnight at 37°C. Peptides were collected by centrifugation followed by a further two washes with elution buffer (1 mM CaCl2, 1 mM MnCl2, 500 mM NaCl in 20 mM TrisHCl, pH 7.3).

Peptide purification

Peptides were acidified with 1% (v/v) TFA, and assuming 50% recovery, 20 µg peptides were loaded directly onto SDB-RPS stage tips. Stage tips were washed twice with 0.1% (v/v) TFA, and sequentially eluted with 20 µL SDB-RPS1 (100 mM Ammonium formate, 40% (v/v) Acetonitrile, 0.5% (v/v) Formic acid), followed by 20 µL SDB-RPS2 (150 mM Ammonium formate, 60% (v/v) Acetonitrile 0.5% (v/v) Formic acid), then 30 µL SDB-RPS3 (1% (v/v) TFA, 80% Acetonitrile). Tryptic peptides were dried to completion in a centrifugal vacuum concentrator (Concentrator 5301, Eppendorf), and volumes were restored to 10 µL with buffer A* (0.1% (v/v) TFA, 2% (v/v) Acetonitrile).

Liquid chromatography coupled to tandem mass spectrometry

LC-MS/MS was performed exactly as described previously (***Itzhak et al., 2016***), with the exception that the LC was coupled to a Q Exactive HF-X Hybrid Quadropole-Orbitrap mass spectrometer, which boasts improved ion transfer (***Kelstrup et al., 2018***).

Processing of mass spectrometry .RAW files

Mass spectrometry .RAW files were processed in MaxQuant (***Tyanova et al., 2016***; ***Cox and Mann, 2008***), version 1.5.5.2. RAW files were organized into two parameter groups to separate trypsin and GluC digested peptides. For both groups, multiplicity for was set to two, with Lys8 and Arg10 selected as heavy labels, re-quantify was turned on, with matching enabled between adjacent

peptide fractions with the same enzyme. The Fasta file Homo_sapiens.GRCh38.pep.all.fa was downloaded from Ensembl.

Acknowledgements

We gratefully acknowledge funding and support from the following institutions and foundations: MJS is a Howard Hughes Medical Institute Fellow of the Helen Hay Whitney Foundation, JSW is an Investigator of the Howard Hughes Medical Institute and, AF is a Searle Scholar and Chan-Zuckerberg Biohub investigator. This work was further supported by the Jane Coffin Childs Memorial Foundation (NTSO), the Howard Hughes Medical Institute (ALB), the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation (ALB,) a Faculty Scholar grant from the Howard Hughes Medical Institute (AF), the Sandler Family Foundation through the UCSF Program for Breakthrough Biomedical Research (AF), the American Asthma Foundation (AF), the Max Planck Society for the Advancement of Science (GHHB, DNI), the German Research Foundation (Gottfried Wilhelm Leibniz Prize MA 1764/2-1 (GHHB), the louis-Jeantet Foundation (DNI), the European Research Council Synergy Grant 'ToPAG – Toxic Protein Aggregation in neurodegeneration' (ERC2012-SyG_318987-ToPAG (DNI), and the National Institutes of Health (NIH) (GM075061 - JLB, AG041826 - JSW, 1DP2GM110772-01 - AF, 8P41GM103481 and 1S10OD16229 –ALB). We thank Kendra Swain, Christopher Williams and Maya Schuldiner for their experimental and intellectual contributions. We also thank Lakshmi Miller-Vedam, Marco Jost and Marco Hein for critical reading of the manuscript, Jeffrey Quinn for producing model figures, and the rest of the Weissman lab for helpful discussions and various contributions.

Additional information

Funding

Funder	Grant reference number	Author
Howard Hughes Medical Institute	Investigator Program	Jonathan S Weissman
National Institutes of Health	GM075061	Jeffrey Brodsky
Helen Hay Whitney Foundation	Postdoctoral Fellowship	Matthew J Shurtleff
Jane Coffin Childs Memorial Fund for Medical Research	Postdoctoral Fellowship	Nicole T Schirle Oakdale
Sandler Foundation	Program for Breakthrough Biomedical Research	Adam Frost
American Asthma Foundation		Adam Frost
Louis-Jeantet Foundation		Daniel N Itzhak
Dr. Miriam and Sheldon G. Adelson Medical Research Foundation		Alma L Burlingame
Max-Planck-Gesellschaft		Daniel N Itzhak
Deutsche Forschungsgemeinschaft	Gottfried Wilhelm Leibniz Prize MA 1764/2-1	Georg HH Borner
European Research Council	ERC2012-SyG_318987-ToPAG	Daniel N Itzhak
Howard Hughes Medical Institute	Faculty Scholar Grant	Adam Frost
National Institutes of Health	AG041826	Jonathan S Weissman
National Institutes of Health	1DP2GM110772-01	Adam Frost
National Institutes of Health	8P41GM103481	Alma L Burlingame
National Institutes of Health	1S10OD16229	Alma L Burlingame

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Matthew J Shurtleff, Conceptualization, Formal analysis, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Daniel N Itzhak, Conceptualization, Data curation, Formal analysis, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Jeffrey A Hussmann, Conceptualization, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Nicole T Schirle Oakdale, Conceptualization, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—review and editing; Elizabeth A Costa, Conceptualization, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—review and editing; Martin Jonikas, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—review and editing; Jimena Weibezahn, Shruthi S Vembar, Investigation, Methodology; Katerina D Popova, Formal analysis, Validation, Investigation, Methodology, Writing—review and editing; Calvin H Jan, Data curation, Formal analysis, Investigation, Visualization, Methodology; Pavel Sinitcyn, Formal analysis, Investigation, Visualization, Methodology; Hilda Hernandez, Resources, Data curation, Formal analysis, Methodology; Jürgen Cox, Software, Supervision, Methodology; Alma L Burlingame, Resources, Funding acquisition, Methodology; Jeffrey L Brodsky, Conceptualization, Supervision, Funding acquisition, Writing—review and editing; Adam Frost, Supervision, Funding acquisition, Writing—review and editing; Georg HH Borner, Conceptualization, Supervision, Funding acquisition, Investigation, Methodology, Writing—review and editing; Jonathan S Weissman, Conceptualization, Supervision, Funding acquisition, Visualization, Writing—original draft, Writing—review and editing

Author ORCIDs

Matthew J Shurtleff http://orcid.org/0000-0001-9846-3051
Elizabeth A Costa http://orcid.org/0000-0001-8365-0436
Pavel Sinitcyn https://orcid.org/0000-0002-2653-1702
Adam Frost https://orcid.org/0000-0003-2231-2577
Georg HH Borner https://orcid.org/0000-0002-3166-3435
Jonathan S Weissman http://orcid.org/0000-0003-2445-670X

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.37018.022
Author response https://doi.org/10.7554/eLife.37018.023

Additional files

Supplementary files

- Supplementary file 1. EMC client proteins identified by mass spectrometry. The number of transmembrane domains (TMDs), whether the protein enters the secretory pathway based on Uniprot annotation and the Log2 fold changes in protein abundance in EMC2 or EMC4 depleted cells compared to control cells expressing a non-targeting sgRNA (GAL4).
  DOI: https://doi.org/10.7554/eLife.37018.012
- Supplementary file 2. Transmembrane domain annotations in yeast. A list of all transmembrane domains annotated in the yeast proteome for analyses performed in ***Figure 3***. See Materials and methods for details on TMD annotation.
  DOI: https://doi.org/10.7554/eLife.37018.013
- Transparent reporting form
  DOI: https://doi.org/10.7554/eLife.37018.014

Data availability

Sequencing data have been deposited in GEO under accession code GSE112891.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Costa EA, Popova KD, Schirle Oakdale NT, Jan CH, Weissman JS	2018	The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112891	Publicly available at the NCBI Gene Expression Omnibus (accession no: GSE112891)

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Aviram N, Ast T, Costa EA, Arakel EC, Chuartzman SG, Jan CH, Haßdenteufel S, Dudek J, Jung M, Schorr S, Zimmermann R, Schwappach B, Weissman JS, Schuldiner M	2016	The SND proteins constitute an alternative targeting route to the endoplasmic reticulum	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85686	Publicly available at the NCBI Gene Expression Omnibus (accession no: GSE85686)
Jan CH, Williams CC, Weissman JS	2014	Principles of ER Co-Translational Translocation Revealed by Proximity-Specific Ribosome Profiling	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61012	Publicly available at the NCBI Gene Expression Omnibus (accession no: GSE61012)

References

Algire MA, Maag D, Savio P, Acker MG, Tarun SZ, Sachs AB, Asano K, Nielsen KH, Olsen DS, Phan L, Hinnebusch AG, Lorsch JR. 2002. Development and characterization of a reconstituted yeast translation initiation system. *RNA* **8**:382–397. DOI: https://doi.org/10.1017/S1355838202029527, PMID: 12008673

Anghel SA, McGilvray PT, Hegde RS, Keenan RJ. 2017. Identification of Oxa1 homologs operating in the eukaryotic endoplasmic reticulum. *Cell Reports* **21**:3708–3716. DOI: https://doi.org/10.1016/j.celrep.2017.12.006, PMID: 29281821

Aviram N, Ast T, Costa EA, Arakel EC, Chuartzman SG, Jan CH, Haßdenteufel S, Dudek J, Jung M, Schorr S, Zimmermann R, Schwappach B, Weissman JS, Schuldiner M. 2016. The SND proteins constitute an alternative targeting route to the endoplasmic reticulum. *Nature* **540**:134–138. DOI: https://doi.org/10.1038/nature20169, PMID: 27905431

Bagchi P, Inoue T, Tsai B. 2016. EMC1-dependent stabilization drives membrane penetration of a partially destabilized non-enveloped virus. *eLife* **5**:e21470. DOI: https://doi.org/10.7554/eLife.21470, PMID: 28012275

Bircham PW, Maass DR, Roberts CA, Kiew PY, Low YS, Yegambaram M, Matthews J, Jack CA, Atkinson PH. 2011. Secretory pathway genes assessed by high-throughput microscopy and synthetic genetic array analysis. *Molecular BioSystems* **7**:2589–2598. DOI: https://doi.org/10.1039/c1mb05175j, PMID: 21731954

Boyle MP, Bell SC, Konstan MW, McColley SA, Rowe SM, Rietschel E, Huang X, Waltz D, Patel NR, Rodman D, VX09-809-102 study group. 2014. A CFTR corrector (lumacaftor) and a CFTR potentiator (ivacaftor) for treatment of patients with cystic fibrosis who have a phe508del CFTR mutation: a phase 2 randomised controlled trial. *The Lancet Respiratory Medicine* **2**:527–538. DOI: https://doi.org/10.1016/S2213-2600(14)70132-8, PMID: 24973281

Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. 2012. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**:552–557. DOI: https://doi.org/10.1126/science.1215110, PMID: 22194413

Christianson JC, Olzmann JA, Shaler TA, Sowa ME, Bennett EJ, Richter CM, Tyler RE, Greenblatt EJ, Harper JW, Kopito RR. 2011. Defining human ERAD networks through an integrative mapping strategy. *Nature Cell Biology* **14**:93–105. DOI: https://doi.org/10.1038/ncb2383, PMID: 22119785

Costa EA, Subramanian K, Nunnari J, Weissman JS. 2018. Defining the physiological role of SRP in protein-targeting efficiency and specificity. *Science* **359**:689–692. DOI: https://doi.org/10.1126/science.aar3607, PMID: 29348368

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**:1367–1372. DOI: https://doi.org/10.1038/nbt.1511, PMID: 19029910

Cox J, Mann M. 2012. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13**:S12. DOI: https://doi.org/10.1186/1471-2105-13-S16-S12, PMID: 23176165

Cymer F, von Heijne G, White SH. 2015. Mechanisms of integral membrane protein insertion and folding. *Journal of Molecular Biology* **427**:999–1022. DOI: https://doi.org/10.1016/j.jmb.2014.09.014, PMID: 25277655

Denic V, Weissman JS. 2007. A molecular caliper mechanism for determining very long-chain fatty acid length. *Cell* **130**:663–677. DOI: https://doi.org/10.1016/j.cell.2007.06.031, PMID: 17719544

Gelsthorpe ME, Baumann N, Millard E, Gale SE, Langmade SJ, Schaffer JE, Ory DS. 2008. Niemann-Pick type C1 I1061T mutant encodes a functional protein that is selected for endoplasmic reticulum-associated degradation due to protein misfolding. *Journal of Biological Chemistry* **283**:8229–8236. DOI: https://doi.org/10.1074/jbc.M708735200, PMID: 18216017

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**:442–451. DOI: https://doi.org/10.1016/j.cell.2013.06.044, PMID: 23849981

Guna A, Volkmar N, Christianson JC, Hegde RS. 2018. The ER membrane protein complex is a transmembrane domain insertase. *Science* **359**:470–473. DOI: https://doi.org/10.1126/science.aao3099

Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH. 2015. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLOS Genetics* **11**:e1005732. DOI: https://doi.org/10.1371/journal.pgen.1005732, PMID: 26656907

Itzhak DN, Tyanova S, Cox J, Borner GH. 2016. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**:e16950. DOI: https://doi.org/10.7554/eLife.16950, PMID: 27278775

Jan CH, Williams CC, Weissman JS. 2014. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346**:1257521. DOI: https://doi.org/10.1126/science.1257521, PMID: 25378630

Jonikas MC, Collins SR, Denic V, Oh E, Quan EM, Schmid V, Weibezahn J, Schwappach B, Walter P, Weissman JS, Schuldiner M. 2009. Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science* **323**:1693–1697. DOI: https://doi.org/10.1126/science.1167983, PMID: 19325107

Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen JV. 2018. Performance evaluation of the Q exactive HF-X for shotgun proteomics. *Journal of Proteome Research* **17**:727–738. DOI: https://doi.org/10.1021/acs.jproteome.7b00602, PMID: 29183128

Kopito RR. 1999. Biosynthesis and degradation of CFTR. *Physiological Reviews* **79**:S167–S173. DOI: https://doi.org/10.1152/physrev.1999.79.1.S167, PMID: 9922380

Krishnan MN, Ng A, Sukumaran B, Gilfoy FD, Uchil PD, Sultana H, Brass AL, Adametz R, Tsui M, Qian F, Montgomery RR, Lev S, Mason PW, Koski RA, Elledge SJ, Xavier RJ, Agaisse H, Fikrig E. 2008. RNA interference screen for human genes associated with west nile virus infection. *Nature* **455**:242–245. DOI: https://doi.org/10.1038/nature07207, PMID: 18690214

Lahiri S, Chao JT, Tavassoli S, Wong AK, Choudhary V, Young BP, Loewen CJ, Prinz WA. 2014. A conserved endoplasmic reticulum membrane protein complex (EMC) facilitates phospholipid transfer from the ER to mitochondria. *PLoS Biology* **12**:e1001969. DOI: https://doi.org/10.1371/journal.pbio.1001969, PMID: 25313861

Li Y, Zhao Y, Hu J, Xiao J, Qu L, Wang Z, Ma D, Chen Y. 2013. A novel ER-localized transmembrane protein, EMC6, interacts with RAB5A and regulates cell autophagy. *Autophagy* **9**:150–163. DOI: https://doi.org/10.4161/auto.22742, PMID: 23182941

Lockshon D, Surface LE, Kerr EO, Kaeberlein M, Kennedy BK. 2007. The sensitivity of yeast mutants to oleic acid implicates the peroxisome and other processes in membrane function. *Genetics* **175**:77–91. DOI: https://doi.org/10.1534/genetics.106.064428, PMID: 17151231

Louie RJ, Guo J, Rodgers JW, White R, Shah N, Pagant S, Kim P, Livstone M, Dolinski K, McKinney BA, Hong J, Sorscher EJ, Bryan J, Miller EA, Hartman JL. 2012. A yeast phenomic model for the gene interaction network modulating CFTR-ΔF508 protein biogenesis. *Genome Medicine* **4**:103. DOI: https://doi.org/10.1186/gm404, PMID: 23270647

Luo WJ, Gong XH, Chang A. 2002. An ER membrane protein, Sop4, facilitates ER export of the yeast plasma membrane [H+]ATPase, Pma1. *Traffic* **3**:730–739. DOI: https://doi.org/10.1034/j.1600-0854.2002.31005.x, PMID: 12230471

Ma H, Dang Y, Wu Y, Jia G, Anaya E, Zhang J, Abraham S, Choi JG, Shi G, Qi L, Manjunath N, Wu H. 2015. A CRISPR-Based screen identifies genes essential for West-Nile-Virus-Induced cell death. *Cell Reports* **12**:673–683. DOI: https://doi.org/10.1016/j.celrep.2015.06.049, PMID: 26190106

Marceau CD, Puschnik AS, Majzoub K, Ooi YS, Brewer SM, Fuchs G, Swaminathan K, Mata MA, Elias JE, Sarnow P, Carette JE. 2016. Genetic dissection of Flaviviridae host factors through genome-scale CRISPR screens. *Nature* **535**:159–163. DOI: https://doi.org/10.1038/nature18631, PMID: 27383987

Markovich S, Yekutiel A, Shalit I, Shadkchan Y, Osherov N. 2004. Genomic approach to identification of mutations affecting caspofungin susceptibility in Saccharomyces cerevisiae. *Antimicrobial Agents and Chemotherapy* **48**:3871–3876. DOI: https://doi.org/10.1128/AAC.48.10.3871-3876.2004

McGlincy NJ, Ingolia NT. 2017. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**:112–129. DOI: https://doi.org/10.1016/j.ymeth.2017.05.028, PMID: 28579404

Nagamori S, Smirnova IN, Kaback HR. 2004. Role of YidC in folding of polytopic membrane proteins. *The Journal of Cell Biology* **165**:53–62. DOI: https://doi.org/10.1083/jcb.200402067, PMID: 15067017

Partridge AW, Therien AG, Deber CM. 2002. Polar mutations in membrane proteins as a biophysical basis for disease. *Biopolymers* **66**:350–358. DOI: https://doi.org/10.1002/bip.10313, PMID: 12539263

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**:1173–1183. DOI: https://doi.org/10.1016/j.cell.2013.02.022, PMID: 23452860

Ram AF, Brekelmans SS, Oehlen LJ, Klis FM. 1995. Identification of two cell cycle regulated genes affecting the beta 1,3-glucan content of cell walls in Saccharomyces cerevisiae. *FEBS Letters* **358**:165–170. DOI: https://doi.org/10.1016/0014-5793(94)01418-Z, PMID: 7828729

Richard M, Boulin T, Robert VJ, Richmond JE, Bessereau JL. 2013. Biosynthesis of ionotropic acetylcholine receptors requires the evolutionarily conserved ER membrane complex. *PNAS* **110**:E1055–E1063. DOI: https://doi.org/10.1073/pnas.1216154110, PMID: 23431131

Rothstein R. 1991. Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. *Methods in enzymology* **194**:281–301. PMID: 2005793

Samuelson JC, Chen M, Jiang F, Möller I, Wiedmann M, Kuhn A, Phillips GJ, Dalbey RE. 2000. YidC mediates membrane protein insertion in bacteria. *Nature* **406**:637–641. DOI: https://doi.org/10.1038/35020586, PMID: 10949305

Sanders CR, Myers JK. 2004. Disease-related misassembly of membrane proteins. *Annual Review of Biophysics and Biomolecular Structure* **33**:25–51. DOI: https://doi.org/10.1146/annurev.biophys.33.110502.140348, PMID: 15139803

Satoh T, Ohba A, Liu Z, Inagaki T, Satoh AK. 2015. dPob/EMC is essential for biosynthesis of rhodopsin and other multi-pass membrane proteins in Drosophila photoreceptors. *eLife* **4**:e06306. DOI: https://doi.org/10.7554/eLife.06306

Savidis G, McDougall WM, Meraner P, Perreira JM, Portmann JM, Trincucci G, John SP, Aker AM, Renzette N, Robbins DR, Guo Z, Green S, Kowalik TF, Brass AL. 2016. Identification of zika virus and dengue virus dependency factors using functional genomics. *Cell Reports* **16**:232–246. DOI: https://doi.org/10.1016/j.celrep.2016.06.028, PMID: 27342126

Sharpe HJ, Stevens TJ, Munro S. 2010. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* **142**:158–169. DOI: https://doi.org/10.1016/j.cell.2010.05.037, PMID: 20603021

Shen X, Kan S, Hu J, Li M, Lu G, Zhang M, Zhang S, Hou Y, Chen Y, Bai Y. 2016. EMC6/TMEM93 suppresses glioblastoma proliferation by modulating autophagy. *Cell Death and Disease* **7**:e2043. DOI: https://doi.org/10.1038/cddis.2015.408, PMID: 26775697

Sherwood PW, Carlson M. 1999. Efficient export of the glucose transporter Hxt1p from the endoplasmic reticulum requires Gsf2p. *PNAS* **96**:7415–7420. DOI: https://doi.org/10.1073/pnas.96.13.7415, PMID: 10377429

Tang X, Snowball JM, Xu Y, Na CL, Weaver TE, Clair G, Kyle JE, Zink EM, Ansong C, Wei W, Huang M, Lin X, Whitsett JA. 2017. EMC3 coordinates surfactant protein and lipid homeostasis required for respiration. *Journal of Clinical Investigation* **127**:4314–4325. DOI: https://doi.org/10.1172/JCI94152, PMID: 29083321

Tector M, Hartl FU. 1999. An unstable transmembrane segment in the cystic fibrosis transmembrane conductance regulator. *The EMBO Journal* **18**:6290–6298. DOI: https://doi.org/10.1093/emboj/18.22.6290, PMID: 10562541

Tyanova S, Temu T, Cox J. 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11**:2301–2319. DOI: https://doi.org/10.1038/nprot.2016.136, PMID: 27809316

Wang W, Malcolm BA. 1999. Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed mutagenesis. *BioTechniques* **26**:680–682. PMID: 10343905

Wiśniewski JR, Mann M. 2016. A Proteomics Approach to the Protein Normalization Problem: Selection of Unvarying Proteins for MS-Based Proteomics and Western Blotting. *Journal of Proteome Research* **15**:2321–2326. DOI: https://doi.org/10.1021/acs.jproteome.6b00403, PMID: 27297043

Yang YG, Lindahl T, Barnes DE. 2007. Trex1 exonuclease degrades ssDNA to prevent chronic checkpoint activation and autoimmune disease. *Cell* **131**:873–886. DOI: https://doi.org/10.1016/j.cell.2007.10.017, PMID: 18045533

Zhang R, Miner JJ, Gorman MJ, Rausch K, Ramage H, White JP, Zuiani A, Zhang P, Fernandez E, Zhang Q, Dowd KA, Pierson TC, Cherry S, Diamond MS. 2016. A CRISPR screen defines a signal peptide processing pathway required by flaviviruses. *Nature* **535**:164–168. DOI: https://doi.org/10.1038/nature18625, PMID: 27383988

Zhang S, Xu C, Larrimore KE, Ng DTW. 2017. Slp1-Emp65: a guardian factor that protects folding polypeptides from promiscuous degradation. *Cell* **171**:346–357. DOI: https://doi.org/10.1016/j.cell.2017.08.036, PMID: 28919078

Zielinska DF, Gnad F, Wiśniewski JR, Mann M. 2010. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* **141**:897–907. DOI: https://doi.org/10.1016/j.cell.2010.04.012, PMID: 20510933

2.2.4 Deep proteomic annotation of mutations and alternative splicing

There is a general scientific debate on the prevalence and impact of alternative splicing on protein complexity[169, 170]. The detection and quantification of alternative isoforms using proteomics would be key to unravel this mystery. One of the ways would be to directly analyze intact proteins using top-down mass spectrometry. However, it is technically not feasible to do on a system biology level[77]. The RNA-Seq was successfully applied to numerous studies focused on alternative splicing regulation at the mRNA level. Compared to RNA-Seq, the standard shotgun proteomics experiment does not even get close to full sequence coverage, thus limiting the detection and quantification of isoforms. One issue for the detection of protein isoforms using bottom-up proteomics is the bias introduced by trypsin cleavage[171].

In this manuscript, we explore the potential of bottom-up proteomics for isoforms and mutations detection by cleaving proteins with various proteases, deep fractionation, and various fragmentation methods. We applied this workflow to six cell lines, which were initially used in the Encode project. Like that, we could reuse the already generated deep NGS data. Also, we explore the feasibility of detecting translated non-synonymous mutations and whether it would be possible to reconstruct proteins de novo from detected peptides.

My contribution to this manuscript was the development of an algorithm for the detection of alternative splicing and single amino acids polymorphisms. The developed solution allowed us to detect around 500 splicing events where both alternatives were present. We also found evidence for more than a thousand expressed mutations per cell line, which were in an agreement with genomic mutations.

Pavel Sinitcyn[*], Alicia L. Richards[*], Dain R. Brademan, Jesse Meyer, Michael S. Westphall, Evgenia Shishkova, Juergen Cox, and Joshua J. Coon
Deep Human Proteome Sequencing Enables Genome Annotation of Mutations and Alternative Splicing
(2020) *The manuscript is in the submission process to Nature Biotechnology journal*

[*]these authors contributed equally to this work

Deep Human Proteome Sequencing Enables Genome Annotation of Mutations and Alternative Splicing

Pavel Sinitcyn,[1,*] Alicia L. Richards, [2,3*] Dain R. Brademan, [4,5] Jesse Meyer, [2,4] Michael S. Westphall, [2,4] Evgenia Shishkova,[2,4] Juergen Cox, [1,^] and Joshua J. Coon[2,3,4,5^]

[1]Computational Systems Biochemistry Research Group, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

[2]National Center for Quantitative Biology of Complex Systems, University of Wisconsin-Madison, Madison, WI 53706

Departments of Chemistry[3] and Biomolecular Chemistry[4], University of Wisconsin-Madison, Madison, WI 53706

[5]Morgridge Institute for Research, Madison, WI 53715

[*]these authors contributed equally
[^]to whom correspondence should be addressed: cox@biochem.mpg.de or coon@wisc.edu

ABSTRACT (150/150 words)

Mature proteomics methods now routinely enable detection of around 10,000 human proteins from a single sample. However, proteins are typically identified by peptide sequences representing a small fraction of all amino acids predicted from the genome. Deeper sequencing – detection of all amino acids – is necessary for discovery and quantitative comparison of all unique proteoforms. Here, we utilized six cell lines, six proteases, and three tandem mass spectrometry (MS/MS) fragmentation methods to collect 2,491 raw MS data files. From these data we identified 1,119,510 unique peptides from 17,717 protein groups with a median sequence coverage of 79.2%, confirming over eight million unique human amino acid residues. We compare our proteomics data with RNA-seq results and demonstrate how such deep proteome coverage can enable detection of over 2,000 protein mutations and over 5,000 alternative splicing event junctions. Our dataset represents a valuable resource as the largest human proteome sequence coverage ever reported.

INTRODUCTION

Persistent developments in mass spectrometry (MS), especially the shotgun strategy, have propelled our ability to analyze proteins. Today near-complete proteomes of simple organisms can be detected following only one hour of analysis.[1-3] For more complex organisms, patient analysts can monitor over 10,000 proteins within a day.[3-6] Draft maps of the human proteome, assembled using copious data from various tissues and cell types, provide evidence for the translation of ~ 90% of known protein-coding genes.[7-9] Although the human genome contains around 20,000 protein-coding genes,[10, 11] it is estimated that alternative splicing events, where mRNA fragments are combined in different arrangements, can yield four to five-fold more unique transcripts.[12, 13] Other alterations, including single nucleotide polymorphisms (SNPs), alternative transcription of start and stop sites, and post-translational modifications (PTM) further increase proteomic complexity.[14, 15] While evidence of some of these events can be obtained from genomic and transcriptomic data, it remains an open question of just how muchvariation exists at the protein level.

A well-known limitation of the shotgun method is that the presence of an entire protein is determined using peptide proxies – sometimes only two or three. Thus, sequence coverage in a proteomics experiment is generally insufficient to fully characterize what Kelleher and Smith have termed proteoforms present within a sample.[14,

16] They persuasively advocate that the ability to precisely monitor these protein molecular forms will be essential to understanding biological systems. Even the current deepest proteomic datasets do not contain enough sequence data to globally identify the alterations named above.[3, 7-9, 17, 18] One approach to achieve proteoform-level detection is top-down MS, a strategy that measures intact protein mass prior to dissociation for sequence determination using tandem mass spectrometry (MS/MS). Ensuring no loss in resolution, the top-down strategy is intellectually appealing. Practical issues with high-mass proteins, sequence coverage, and detection of low abundance species, however, limit its impact.[19] Given the technical hurdles with top-down proteomics, we revisited the shotgun strategy.

Shotgun proteomics generally relies on trypsin as the preferred enzyme used to catalyze hydrolysis of proteomes. Trypsin cleaves C-terminal to lysine and arginine and produces peptides of length and charge distributions most amenable to tandem MS (MS/MS). However, even with the assistance of extensive chromatographic separations, not all portions of the proteome are accessible from tryptic peptides; many of the peptides produced are either too short or too long to be detected using current LC-MS technology. As proteoforms can differ by a small number of amino acids, near-complete sequence coverage is crucial for distinguishing near-identical variants. The use of alternative enzymes in addition to trypsin during digestion can increase the amino acid coverage of individual proteins, phosphorylation sites, and whole proteomes[6, 20-22]. However, given the considerable increased effort it involves, this strategy has not become routine.

We wondered whether digestion of human proteomes with several different proteases coupled with extensive LC fractionation and state-of-the-art MS could produce sufficient sequence depth to allow for global assessment of genomic variation and RNA editing at a proteome level. To test this idea, we cultured six unique human cell lines, harvested proteins from each, and digested them (separately) with six specific proteases. The resulting peptides were extensively fractionated offline and then analyzed on a tribrid Orbitrap mass spectrometer. Peptides derived from all enzymes were dissociated using a variety of fragmentation methods including beam-type collisional activation (HCD),[23] ion trap collisional activation (CAD),[24] and electron transfer dissociation (ETD).[25] Altogether we collected approximately ~20 million high resolution mass spectra and ~164 million MS/MS spectra within ~2,500 nLC-MS/MS experiments. The combination of data from all cell lines allowed for identification of 17,717 proteins with an overall median sequence coverage of 79.2%. Using these data, and comparing them to transcriptomic analysis, we offer a global view of mutations and alternative splicing at the protein level. Further, we have compiled these data into an online resource to enable the intuitive exploration of the on the deepest protein variant datasets to date. This resource is accessible at the web address https://deep-sequencing.app. Using the deep and overlapping peptide sequence information generated, we demonstrate feasibility for *de novo* protein assembly.

RESULTS

Deep human proteome sequencing

In silico tryptic digestion of the ~20,200 reviewed, canonical proteins of the human proteome (Uniprot) predicts ~2.4 million tryptic peptides of suitable size for MS detection (7–35 amino acids, up to one missed cleavage). These peptides comprise ~7.5 million amino acid residues – *i.e.,* only ~65% of the proteome's ~11.3 million total amino acid residues. Next, if we consider digestion of the same 20,000 proteins using the six enzymes in our study (LysC, LysN, AspN, chymotrypsin, GluC, and trypsin), two million peptides suitable for shotgun proteomics, representing 99% of the amino acids contained by the proteins, are generated. Note we have previously demonstrated the increase in sequence coverage of a yeast proteome following digestion with proteases in addition to trypsin.[21]

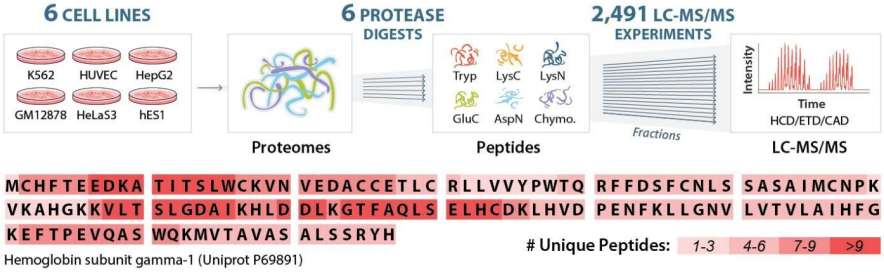


Figure 1. Overview of Deep Proteomics Workflow. Six human cell lines were grown in parallel, their proteomes isolated, and then separate aliquots of each proteome were digested by one of six proteases in parallel. Peptides resulting from each digestion were fractionated by high-pH reversed phase chromatography, and then analyzed separately with nanoLC-MS/MS using HCD, ETD, and CAD. The resulting data was searched with MaxQuant against the human proteome database, and over 17,000 proteins were identified by peptides that produce a median coverage of over 80%. The high coverage achieved is illustrated on the sequence of hemoglobin subunit gamma-1 with color coding to illustrate the number of unique peptides that cover each position.

To test the hypothesis that we can similarly increase coverage of the human proteome, we selected six diverse human cell lines: hES1, an embryonic stem cell line, HeLa S3, from cervical carcinoma, HepG2, from liver carcinoma, GM12878, a blood lymphoblastoid line, K562, from chronic myeloid leukemia, and HUVEC, from umbilical vein epithelial cells (**Figure 1A**). As part of the ENCODE project, these cell lines have a large amount of publicly available genomic and transcriptomic data. Aliquots of each cell line were separately digested with the six proteases named above. To maximize depth, peptides were heavily fractionated (24–80 fractions) and analyzed on either an Orbitrap Fusion (Thermo) or an Orbitrap Lumos (Thermo). All fractions were analyzed using nano flow LC-MS/MS. Dissociation for MS/MS was achieved using HCD, CAD, and ETD. The resulting 2,491 raw files were simultaneously analyzed by database search to identify proteins and peptides using the Andromeda search engine inside MaxQuant[14,15], and results were filtered to 1% protein-level FDR.

Figure 2 summarizes these data by cell line, depth of coverage, and gains provided by multiple-protease digestion. For each cell line, an average of 539,325 unique peptides, corresponding to ~16,000 proteins were identified (**Figure 2A**). The highest number of protein identifications was from the hES1 cell line (17,121), followed by HeLa S3 (16,399), GM12878 (16,344), HepG2 (16,328), HUVEC (16,158) and K562 (16,054). The trypsin dataset contributed (as expected) the largest number of unique peptides (396,782), followed by LysN (194,506), LysC (193,956), GluC (162,784), AspN (152,259) and chymotrypsin (114,152). Notably, within each cell line, data from each enzyme digestion alone identified over 10,000 protein groups. Across all cell lines, data from tryptic peptides contributed the largest number of identifications and unique sequences, totaling 17,631

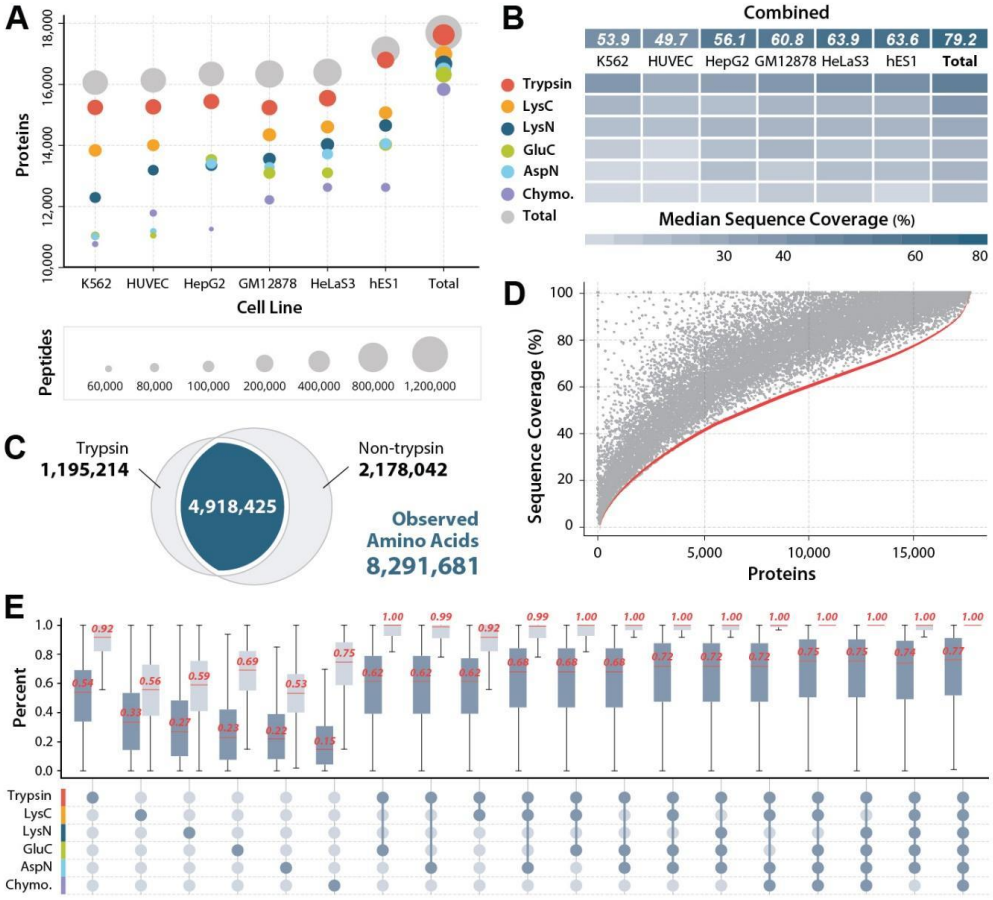


Figure 2. Overview of results from deep proteomics analysis. A. Number of proteins detected for each of the six cells lines and cumulative as a function of peptides from the various proteases. **B.** Median sequence coverage for the various cellular proteomes for combined and individual protease results. **C.** In total over eight million individual amino acids were sequenced in this experiment. **D.** Sequence coverage for each of the detected proteins for the tryptic peptide data (red) and combined (gray). **E.** Theoretical sequence coverage achievable for various combinations of proteases (light gray boxes) and observed sequence coverage (dark gray).

proteins with 56.5% median sequence coverage. Data from LysC-generated peptides mapped to 17,006 protein groups with 37.9% median sequence coverage, followed by chymotrypsin-generated peptides that contributed data supporting 15,833 protein group identifications with 19.6% median sequence coverage. Data from LysN digestion contributed data supporting 16,686 protein group identifications with 32.6% sequence coverage, AspN contributed data supporting 16,479 protein group identifications with 27.2% median sequence coverage, and GluC contributed data supporting 16,318 protein group identifications with 29.0% median sequence coverage.

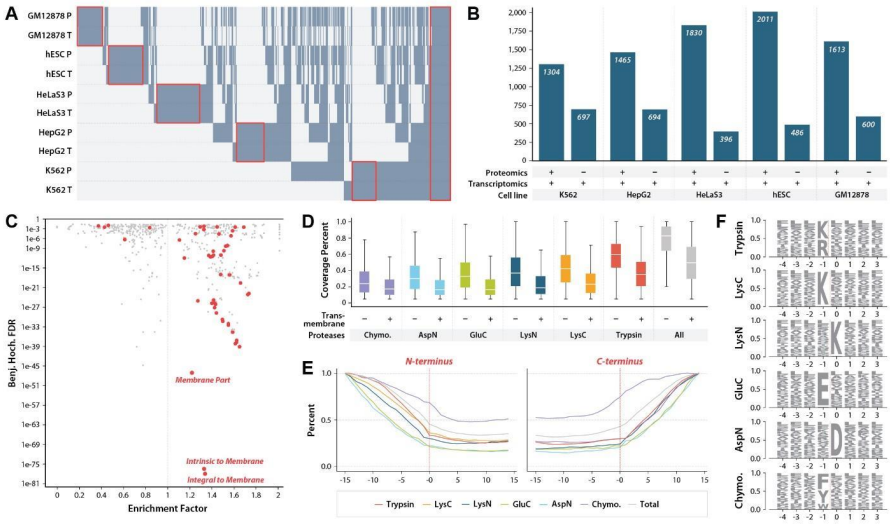
Analysis of all data from all cell lines and protease digestions together identified 12,151,708 peptide spectrum matches (PSMs), 1,119,510 unique peptides, and 17,717 protein groups at a peptide and protein FDR of 1%. These proteins correspond to over 12,000 genes, comprising ~65% of predicted protein-coding genes. Of those, 790 proteins were identified with 100% sequence coverage. The average number of unique peptides per protein was 97 (median = 65). Fifty-four proteins were identified from only one unique peptide; only 1,122 proteins, or 6.3% of the total proteins, were identified by ten or fewer unique peptides. These 17,717 protein groups have a median sequence coverage of 79.2%. Median sequence coverage for the combined dataset and the contributions from subsets is shown in **Figure 2B** (see also **Supplementary Table 3**), and ranges from 49.7% (HUVEC; 16,158 proteins) to 63.9% (HeLa S3; 16,399 proteins). Remarkably, nearly half of all identified proteins were observed with 80–100% sequence coverage (**Supplementary Figure 1**). Only 936 proteins, or 5.3% of the total data, have sequence coverage below 25%.

The addition of enzymes other than trypsin provided a slight increase in the total number of proteins identified but a large increase in the non-redundant amino acids detected. The 17,717 human proteins present in our dataset comprise 12,006,700 amino acid residues. In total, the unique peptides identified in the combined tryptic datasets from all cell lines detected approximately half of these amino acids (6,113,639; see **Supplementary Table 4**). The overlap between the amino acids detected from tryptic peptides and amino acids identified in peptides generated by all other enzymes is plotted in **Figure 2C. Figure 2D** illustrates the median sequence coverage achieved with various combinations of enzymes. In total, the addition of alternate enzymes to trypsin added 2,179,015 amino acids for a total of 8.2 million, or 69.1%, of the residues that make up the proteins we identified. **Figure 2E** illustrates how the various combinations of enzymes contributes to protein coverage distribution for all proteins. Noteworthy, is that all top performing enzyme combinations include trypsin. Our total human proteome coverage is the largest to date, with 2.12 million more residues (a 34.4% increase) over the 6.17 million identified using exclusively tryptic peptides from the entire MassIVE data repository[16].

The impact of these additional amino acids on protein sequence coverage can be seen in **Figure 2D**. The sequence coverage of human proteins identified using trypsin is plotted in red, with sequence coverage obtained for each protein from the combination of all enzymatic datasets plotted in gray. On average, sequence coverage increased by 19% from digestion with six enzymes compared to digestion with trypsin alone. Among all protein identifications, 86 could not be identified using tryptic peptides alone. Four of these proteins – CUGBP Elav-like family member 1 (F5H7M7), Protein transport protein Sec31A (U3KQR3), and predicted genes ENSP00000421703.1, ENSP00000468392.1 – saw increases from no coverage with data from trypsin to 100%

coverage with the addition of other data. Each of these proteins are relatively short, containing either zero or few suitable tryptic peptides and were thus not detected in the trypsin dataset (**Supplementary Figure X?**).

Alternative proteases have previously been utilized to uncover novel portions of the proteome, including membrane proteins. These proteins – essential to many biological processes and representing important drug discovery targets (Wu and Yates 2003) – remain under-represented in proteomics datasets due to their hydrophobic nature. This is also true of our dataset. Gene ontology (GO) cellular component (CC) pathway enrichment analysis of the proteins with sequence coverage below 25% revealed that these low coverage proteins were primarily membrane proteins (**Supplementary Figure 3B**). Peptide sequences of these proteins may be lost due to limited solubility or a lack of appropriate cleavage sites for our proteases, which, with the exception of chymotrypsin, primarily target charged amino acids. To assess each enzyme's ability to produce membrane-spanning peptides suitable for identification by shotgun proteomics, we mapped these peptide sequences to downloaded structural domain information from Uniprot. Structural domains include helix, turn, transmembrane, intramembrane, and strand. No significant enrichment was observed for helix, turn, and strand domains, with each enzyme covering on average >80% of amino acids comprising a specific domain. Significant differences were observed among each enzyme's ability to access transmembrane-spanning sequences. Digestion with AspN, GluC, LysN, and LysC identified 86, 79, 186, and 293 transmembrane-spanning regions, respectively, with each permitting access to an average of 33% of the amino acids in these regions. Following tryptic digestion, 599 transmembrane-spanning regions were identified, with ~60% coverage. Digestion with chymotrypsin allowed the identification of 1,034 transmembrane-spanning regions, albeit at lower coverage than other enzymes due to chymotrypsin's multiple cleavage sites. To further explore the behavior of peptides generated from transmembrane-spanning sequences, we calculated the enzyme-specific coverage of aligned membrane-spanning regions to either the N- or C-terminus (**Figure 2X**). These data demonstrate that because transmembrane regions are depleted for typical protease cleavage sites, peptides suitable for detection by shotgun proteomics are not being observed. This conclusion is further supported by the strong relative performance of chymotrypsin, which is atypical in cleaving at hydrophobic residues, as compared to the other proteases. The overlap of transmembrane regions is presented in **Supplementary Figure X**, highlighting the numerous transmembrane regions identified with use of chymotrypsin that were not present following digestion with the other enzymes in our study.



Majority of hypothetical SAPs are confirmed in the proteome

Single amino acid polymorphisms (SAPs) are variations in the protein sequence resulting from single nucleotide polymorphisms (SNPs) in genomic sequences that in turn lead to a non-synonymous codon change. The HeLa S3 cell line used in this study contains ~4.5 million SNPs. Of these, ~30,000 occur in coding regions, 4,740 of which are non-synonymous (Landry et al. 2013). We wondered whether our deep sequencing data would afford us the ability to determine the extent that these SNPs are translated into SAPs. To that end, we searched for SAPs with a novel MaxQuant module which specializes in the identification of peptide evidence for the translation of genomic variations (see Methods). We found protein-level evidence for individual cell lines up to 2,304, or a total of ~3,000 SAPs. (**Figure 3A, Supplementary Table 5**). For all cell lines except HUVEC, there was high overlap between the mutations detected by transcriptomics and proteomics. Given HUVEC is the only primary cell line (*i.e.* obtained directly from host tissue) in the study, this low overlap is expected; transcriptomic and proteomic data were collected on cells from different donors [citations]. Therefore, we omitted HUVEC from further analysis. **Figure 3A (new)** shows that the majority of non-synonymous SNPs that appear in the transcript also appear at the protein level (median 73% over all studied cell lines). Further, the multi-enzyme data led on average to a doubling of identified SAPs compared to when only trypsin was applied as a protease. To our knowledge, these data represent the first global view of how SNPs are propagated into the proteome.

These data contain 939 mutations mapping to at least one entry in the Online Mendelian Inheritance in Man Database (OMIM) (Hamosh 2002), a catalog of SNPs associated with disease. 607 of these mutations were identified, at least partially, using tryptic peptides, while over a third of the mutations (354) were identified exclusively using enzymes other than trypsin. Two examples of SNPs linked to cancer were identified on gene MSH3. rs26279, resulting from a G > A polymorphism and resulting in an A > T amino acid substitution, has

been linked to an increased cancer risk, particularly for colorectal and breast cancer (Miao et al. 2015). Evidence of this SAP was present in the K562, GM12878, HepG2 and hES1 cell lines, with at least two unique ____ spanning the AA substitution. rs184967, arising from an A > Q polymorphism and resulting in a G > R amino acid substitution, has similarly been linked to increased cancer risk, particularly endometrial (Nat Genetics, 1996) and triple negative breast cancer (Lee et al. 2014). Peptides spanning this SAP were present in the K562, GM12878, HepG2, HeLaS3 and hES1 cell lines, with nine representative PSMs linked to one of the GM12878 peptides. Crystal structure: highlight region where this is. Peptide/overlap of two peptides color-coded, point with arrow to mutated amino acid. Show mutation point, show where peptide is in structure; if space, show tandem mass spectra.

Enrichment analysis. On average, X additional SNPs were identified in transcript measurements; we wondered whether the genes bearing the SNPs that were not observed at the protein level had any functional enrichment. To answer this question, we performed enrichment analysis of the mutations that we found only on the transcriptomic level, which revealed many GO terms associated with membrane protein families (**Supplemental Figure X**). As discussed above, membrane proteins are notoriously challenging to detect using MS-based proteomics; while our multiple-protease data expands our ability to detect this protein class, coverage of them is lower (add number, **Supplemental Figure Y**).

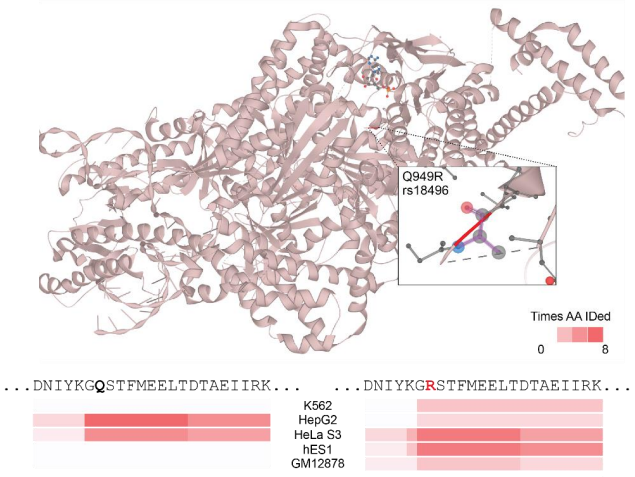


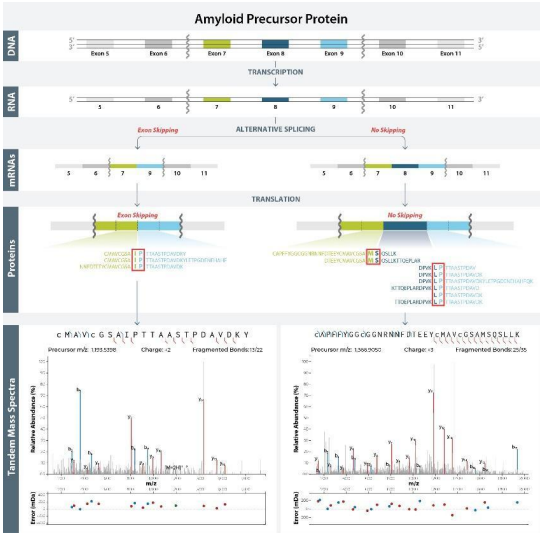
Figure 3C shows the mutations as a function of cell line and whether they are detected at the protein level. Enrichment in each cell line vs. enrichment in islands. Each cell line contains a distinct island of mutations that were confirmed by both RNA-seq and proteomics **(Figure 3C)**. Both proteomics and transcriptomics found 191 proteins mutated in all cell lines. A functional term enrichment analysis of these shared proteins revealed a few groups of terms and proteins with no obvious relation (**Supplemental Figure 5**). No trend was found in protein domain enrichment analysis. From this result, we conclude that the majority of SNPs are translated to SAPs.

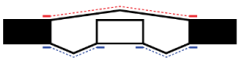
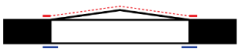
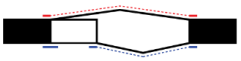
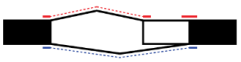
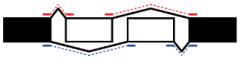
Protein-level Evidence for Alternative Splicing

Discuss current Figure 2E here. Another potential utility of high proteome sequence coverage is the ability to detect the presence of proteoforms that arise from alternative splicing of transcripts. Note we define a splicing event as one that alters the ultimate protein coding region of the transcript. Genome annotation predicts

extensive alternative splicing (~52,000 events with two alternatives using GRCh38 ensemble annotation 94). Further, RNA-seq studies of transcripts have demonstrated that many of these putative events do occur at the transcript level (~17,119) (cite GTX paper). Although many genes undergo alternative splicing, global evidence for the translation of these alternative sequences is lacking (Tress et al. 2017) + more). Specifically, conventional shotgun proteomics relies on incomplete sequence coverage arising from tryptic digestion alone, which is insufficient to identify instances of alternative mRNA splicing on a global scale (Wang et al. 2018; Sheynkman et al. 2013).

Given the deep sequence coverage afforded in this study, we sought to provide the first global assessment of proteome-wide alternative splicing. **Figure 4A** explains the rationale of our strategy and illustrates an example of heterogenous alternative splicing. Here we consider the gene for amyloid precursor protein and the known alternative splicing event that occurs at exon 8. In one form of the transcript, this exon is included, and in another, it is skipped. To distinguish the protein products of these processes, one must observe peptides spanning the junction between exons 7 and 9 (green and light blue region of **Figure 4A**) and the junctions between exons 7 and 8 or between exons 8 and 9 (green and light blue or dark blue and light blue **Figure 4A**). In total, we detected 11 unique peptides spanning these junctions, confirming the presence of both proteoforms.

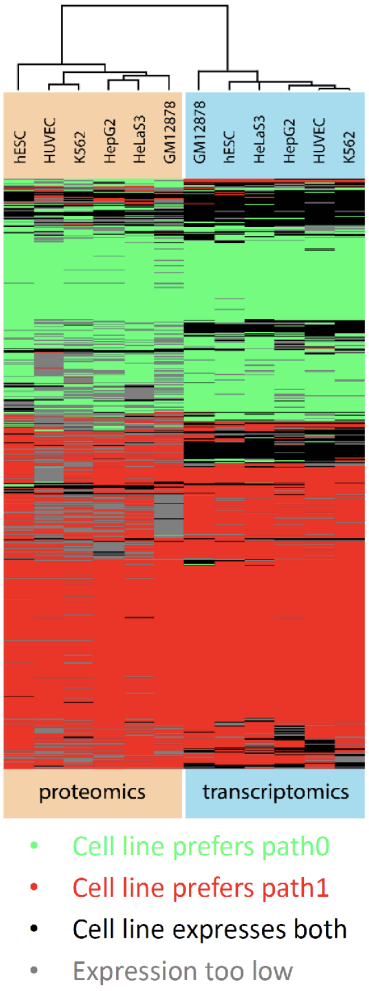


Name	Scheme	Total events Ensemble v94	Transcriptome: Number detected	Transcriptome: Both alternatives detected	Proteome: Number detected	Proteome: Both alternatives detected
Skipped exon		5436	3881	728	2929	270
Retained intron		230	78	20	50	3
Alternative donor		839	482	145	334	59
Alternative acceptor		1613	1045	352	704	141
Mutually exclusive exons		124	77	21	60	12
	Total	8242	5563	1266	4077	485

In addition to the exon skipping example described above, **Table 1** depicts several known types of alternative splicing. Note the five examples shown in **Table 1** represent a small portion of possible configurations but depict the most frequent and most known versions of splicing. First we analyzed the RNAseq data of all cell lines together and revealed approximately 1,266 splicing events with evidence for both alternatives. Next, we merged the mass spectrometry results from all cell lines and searched that comprehensive data to see how many splicing events could be detected. From this analysis we detected protein evidence for 485 of the 1,266 alternative splicing events where both proteoforms were present (**Table 1, Supplementary Table 6**). In other words, approximately 40% of the alternative splicing events detected from the transcript level are translated into proteins. Notably, proteins resulting from intron retention were the most rarely detected (*i.e.*, only 3 of 20). This finding agrees with previous results that intron retention is accompanied by a decrease in protein abundance[19]. The other types of alternative splicing were found at a similar proportion as described by the coding sequence assembly (**Table 1**) both in total and across the individual cell lines.

Among the 485 proteoform pairs where both splicing alternatives were detected there was a significant enrichment in proteins with GO terms related to transcriptional and cell cycle control, including physical separation of cells during mitosis (**Supplementary Figure 7**). A recent paper exploring differential splicing in transcripts along the cell cycle from synchronized cells concluded that alternative splicing primarily occurred in cell cycle-related genes and was under control of a specific kinase[20]. Where we detect both paths of splice of events in these genes at transcript and protein level, our results confirm the relevance of this transcriptomic observation for producing different functional proteins (a relationship that subject to debate[21–23]). Our observation of both paths is most likely due to the heterogeneous cell cycles that occur in cell culture.

We also detected unique peptides spanning at least one path for 4,077 splice events. To investigate the nature of the bias towards one alternative manifests across the transcriptome, the proteome, and cell lines. **Figure 5** plots the path preference for each splicing event as a function of cell line and ome. Here we observe an overall agreement agreement in path preference in across transcript and proteome. That said, there are cases where both transcripts are present an single protein is generated. This observation is intriguing and warrants further investigation as we cannot exclude the possibility that the one protein form is below our detection limit. A second conclusion from these data is that, by in large, cell lines tend to select the same splicing paths.



De novo proteoform assembly

Protein inference is conceptually akin to reference transcriptome assembly in short-read next-generation sequencing (NGS), where a previously assembled proteome or genome database is required to map peptide sequences or nucleic acid reads, respectively. In proteomics, however, genome assemblies for proteome database generation are either not available or not high-quality for many organisms. Several tools are available to assemble NGS reads without a reference genome, such as SOAPdenovo-Trans[24]. But *de novo* assembly of nucleic acid sequences relies on the presence of randomly overlapping sequences, which is not true of proteomic datasets that use only a single enzyme for digestion (*e.g.* trypsin) that should not produce overlapping sequences.

Figure 4A. Example of heterogenous alternative splicing detected in amyloid precursor protein. I), Scheme depicting the layout of amyloid precursor protein's exons showing the inclusion or exclusion of exon 8 with the peptide sequences detected that

However, with the data from six different protease digestions described in this paper, we produce many sequences with partial overlap, which we hypothesized may enable *de novo* protein assembly. To test this hypothesis, we reverse translated all peptide identifications with a non-degenerate codon table and used SOAPdenovo-Trans to assemble proteins. An excellent example for the *de novo* assembly of proteasome subunit alpha type-6 (UniProtKB P60900) with full sequence coverage is shown in **Supplementary Figure 8A**. Overall, the *de novo* assembly produced 35,480 scaffolds, of which 16,496 (~47%) correctly match to 9,695 protein groups. Median sequence coverage from the de *novo* assembly was 18% compared to 79.2% for the reference assembly (**Supplementary Figure 8B, 8C**). Assembled scaffolds have a range of 33 to 358 amino acids with a median length of 45 (**Supplementary Figure 8D**), and an average of two scaffolds were mapped to each protein (**Supplementary Figure 8E**). Our results demonstrate the feasibility of *de novo* proteome assembly using overlapping peptides from multiple protease digestions of the proteome; application of more sophisticated proteomics-specific assembly methods may improve this result in the future[25].

DISCUSSION

Here we used six human cell lines, six parallel protease digestions, and three MS/MS fragmentation methods to generate over 164 million tandem mass spectra from nearly 2,500 nLC-MS/MS analysis (**Figure 1**). Our analysis of the combined data identified over one million unique peptides from 17,717 protein sequences (**Figure 2**). The median protein sequence coverage was 79.2%, which represents 8.29 million unique amino acids. Use of proteases that produce sequences complementary to trypsin were especially helpful in detecting 2.18 million unique amino acids, increasing the average protein's sequence coverage by 19%. Interestingly, the quantitative character of protein identifications shared between peptides unique to each digest group gave similar protein abundance levels as measured by iBAQ, even though different peptide sequences should have fundamentally different ionization efficiencies (**Figure 3**). The protein abundance proxy iBAQ was roughly correlated with the observed sequence coverage of each protein, and low sequence coverage proteins were enriched in GO terms related to membrane proteins.

We demonstrate the value of this deep human proteome sequencing in several ways. We compare the protein sequences we found with transcriptomics results, where high transcript coverage is routine. In doing so we first find that relative protein quantities and transcript quantities can be used to segregate data from each cell type, and that proteomic data clusters with transcriptomic data by PCA. We next find that protein mutations detected by MS are very well correlated with those detected by RNA-seq, providing evidence that events where the wrong tRNA is incorporated are exceedingly rare. Further, we used our peptide identifications to detect protein-level evidence of alternative splicing events that were predicted from RNA-seq and find that the proportions of various splice events that translate into protein mirror the proportions detected from RNA-seq (with the exception of intron retention, which is never translated). Expression of these protein and transcript variants have been compiled into an online resource for exploration by the proteomics community at https://deep-sequencing.app.

Finally, we show proof-of-concept results that such high protein sequence coverage generated from multiple overlapping peptide sequences can be used to assemble proteins *de novo*, which could be useful when genome sequence is unavailable or for detection of exotic novel transcript translation.

Our massive dataset provides a unique resource for the analysis of six standard human cell lines and provides benchmarks for future studies of multi-protease, multi-dissociation studies of the human proteome. We expect this data and analysis to guide future studies seeking to understand the relationship between splicing and translation as well as tolerable mutations in the proteome.

ONLINE METHODS

Cell culture and lysis. HeLa S3 cells (ATCC CCL-22; ATCC, Manassas, VA) were grown at 37°C with 5% CO_2 in F-12K medium (ATCC) supplemented with 10% fetal bovine serum (FBS) and antibiotics. HUVEC cells (Lonza CC-2517; Lonza, Walkersville, MD) were grown at at 37°C with 5% CO_2 in Endothelial Growth Media (EGM) supplemented with EGM Complete Media (Lonza) and antibiotics. HepG2 cells (ATCC HB-8065; ATCC) were grown at 37°C with 5% CO_2 in Eagle's Minimum Essential Medium (EMEM, ATCC) supplemented with 10% FBS and antibiotics. K562 cells (ATCC CCL-243; ATCC) were grown at 37°C with 5% CO_2 in Iscove's Modified Dulbecco's Medium (IMDM, ATCC) supplemented with 10% FBS and antibiotics. GM12878 cells (GM12878 K Order 104598; Coriell Institute for Medical Research, Camden, NJ) supplemented with 15% FBS and RPMI-1640 medium (Sigma Aldrich). Cells were harvested at >70% confluency through centrifugation at 300xg for 5 minutes at 4°C. The supernatant was removed, and cells were washed with phosphate-buffered saline (PBS) and centrifuged at 300xg for 5 minutes at 4°C. The resulting pellet was stored at -80°C. Cell pellets were resuspended in lysis buffer containing 8 M urea, 50 mM tris (pH 8), 5 mM $CaCl_2$, 30 mM NaCl, and protease (Roche) and phosphatase (Roche) inhibitor tablets. The pellet was lysed by four rounds of sonication at 4°C, alternating between 20 seconds on and 20 seconds off. Lysate protein concentration was measured by BCA (Thermo Pierce).

Digestion. Protein was reduced by addition of 5 mM dithiothreitol and incubated for 45 min at 55 °C. The mixture was cooled to room temperature, followed by alkylation of free thiols by addition of 15 mM iodoacetamide in the dark for 30 min. The alkylation reaction was quenched with 5 mM dithiothreitol. For tryptic digestion, a 1 mg protein aliquot was digested overnight with 20 µg trypsin (Promega, Madison, WI) at room temperature in 1 M urea. For LysC digestion, a 1 mg protein aliquot was digested overnight with 20 µg LysC (Wako, Richmond, VA) at room temperature in 4 M urea. For LysN digestion, a 1 mg protein aliquot was digested for four hours with 20 µg LysN (Thermo Pierce) at 37°C in 4 M urea. For GluC digestion, a 1 mg protein aliquot was digested overnight with 25 µg GluC (Roche Diagnostics, Indianapolis, IN) at room temperature in 0.5 M urea. For chymotrypsin digestion, a 1 mg protein aliquot was digested overnight with 12.5 µg of chymotrypsin resuspended in 0.2% FA (Promega, Madison, WI) in 1 M urea. For digestion with AspN, a 1 mg protein aliquot was incubated with 6 µg AspN (Roche Diagnostics, Indianapolis, IN) at room temperature overnight. Each digest was quenched by the addition of TFA and desalted on a 100 mg C_{18} Sep-Pak cartridge (Waters, Milford, MA).

Fractionation. High-pH RP fractionation was performed either using a Surveyor LC quarternary pump or a Dionex UltiMate 3000. Fractionation was performed at a flow rate of 1.0 mL/min using a 5 µm column packed with C18 particles (250-mm by 4.6-mm, Phenomenex) on a Surveyor LC quarternary pump. Samples were resuspended in buffer A and separated using the following gradient: 0-2 min, 100% buffer A and separated by increasing buffer B over a 60-minute gradient at a flow rate of 0.8 mL/minute (buffer A: 20 mM ammonium formate, pH 10; buffer B: 20 mM ammonium formate, pH 10, in 80% ACN). Flow rate was increased to 1.5 mL/minute during equilibration. Fractionation was performed at a flow rate of 0.45 mL/min using a 1.7 µm

column packed with BEH particles (50-mm by 1-mm, Waters) on a Dionex Ultimate 3000 pump (Thermo). Samples were resuspended in buffer A and separated by increasing buffer B over a 45-minute gradient at a flow rate of 0.45 mL/minute (buffer A: 20 mM ammonium bicarbonate; buffer B: 20 mM ammonium bicarbonate in 80% ACN). Trypsin digested H1-hESC cells were first fractionated via strong cation exchange fractionation. Peptides were dissolved in 400 µl of strong cation exchange buffer A (5 mM KH_2PO_4 and 30% acetonitrile; pH 2.65) and injected onto a polysulfoethylaspartamide column (9.4 mm × 200 mm; PolyLC) attached to a Surveyor LC quarternary pump (Thermo Electron, West Chester, PA) operating at 3 ml/min. Fractions were collected every 2 min starting at 10 min into the following gradient: 0–2 min at 100% buffer A, 2–5 min at 0%–15% buffer B (5 mM KH_2PO_4, 30% acetonitrile, and 350 mM KCl (pH 2.65)), and 5–35 min at 15%–100% buffer B. Buffer B was held at 100% for 10 min. Fractions were collected from 8-12 minutes, 12-14 minutes, 14-16 minutes and 16-25 minutes. Each of these four SCX fractions was further fractionated by high-pH RP fractionation on a Surveyor LC quarternary pump, as described above.

LC-MS/MS. Samples were resuspended in 0.2% formic acid (FA) and separated via reversed phase (RP) chromatography. Peptides were injected on to a RP column prepared in-house. Approximately 35 cm of 75 µm-360 µm inner-outer diameter bare-fused silica capillary, each with a laser pulled electrospray tip, were packed with 1.7 µm diameter, 130 Å pore size, Bridged Ethylene Hybrid C18 particles (Waters). Columns were fitted on to either a nanoAcquity (Waters) or Dionex (Thermo) and heated to 60 °C using a home-built column heater. Mobile phase buffer A was composed of water and 0.2% formic acid. Mobile phase B was composed of 70% ACN, 0.2% formic acid, and 5% DMSO. Each sample was separated over a 100-min gradient, including time for column re-equilibration. Flow rates were set at 300-350 µl/min.
Peptide cations were converted to gas-phase ions by electrospray ionization and analyzed on a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo) or a Thermo Orbitrap Lumos (Q-OT-qIT, Thermo). All fractions were analyzed using HCD. Precursor scans were performed from 300 to 1,500 *m/z* at either 60K or 120K resolution (at 400 *m/z*). A 5 x 10^5 ion count target was used on the Orbitrap Fusion, a 1 x 10^6 ion count target was used on the Orbitrap Lumos. Precursors selected for tandem MS were isolated at 0.7 Th with the quadrupole, fragmented by HCD with a normalized collision energy of 30, and analyzed using turbo scan in the ion trap. For some analyses, precursors above 500 *m/z* were fragmented by HCD using the described conditions, while precursors below 500 *m/z* were fragmented by CAD with a normalized collision energy of 30. The maximum injection time for MS^2 analysis was normally set at either 25 or 35 ms, but was set higher for some analyses, with an ion count target of 10^4. Precursors with a charge state of 2-8 were sampled for MS^2. Dynamic exclusion time was set at 15 seconds, with a 10-ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. Analyses were performed in top speed mode with either 3 or 5 second cycles.

LysC, LysN, AspN, GluC and chymotrypsin fractions were analyzed using ETD. To maximize identifications, precursor scans were performed from 200 to 800 *m/z* at either 60K or 120K resolution (at 400 *m/z*). A 5 x 10^5 ion count target was used on the Orbitrap Fusion, a 1 x 10^6 ion count target was used on the Orbitrap Lumos.

Precursors selected for tandem MS were isolated at 0.7 Th with the quadrupole. Precursors were fragmented by ETD using custom reaction times; +3: 40 ms, +4: 22 ms, +5: 14 ms, +6: 10 ms, +2: 70 ms. EThcD was performed on +2 precursors, at 25% supplemental activation collision energy. Precursor ions were selected for fragmentation based on charge state in the following order: +3, +4, +5, +6, +2. Fragment ions were analyzed in the ion trap. Dynamic exclusion time was set at 15 seconds, with a 10-ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. Analyses were performed in top speed mode with either 3 or 5 second cycles.

Fractionated peptides from chymotrypsin-catalyzed proteolysis were analyzed using CAD. Precursor scans were performed from 300 to 1,500 *m/z* at either 60K or 120K resolution (at 400 *m/z*). A 5 x 10^5 ion count target was used on the Orbitrap Fusion, a 1 x 10^6 ion count target was used on the Orbitrap Lumos. Precursors selected for tandem MS were isolated at 0.7 Th with the quadrupole, fragmented by CAD with a normalized collision energy of 30, and analyzed using turbo scan in the ion trap. The maximum injection time for MS^2 analysis was normally set at either 25 or 35 ms, but was set higher for some analyses, with an ion count target of 10^4. Precursors with a charge state of 2-8 were sampled for MS^2. Dynamic exclusion time was set at 15 seconds, with a 10-ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. Analyses were performed in top speed mode with either 3 or 5 second cycles.

Protein Identification.

The raw MS data was searched with the MaxQuant software (version 1.5.7.5). Searches were performed against the following protein sequence databases: UniProt canonical (UP000005640_9606), Uniprot isoform (UP000005640_9606_additional), Ensembl canonical (GRCh38.pep.all), Ensembl isoform (GRCh38.pep.abinitio), and a three-frame translation of Ensembl ncRNA (GRCh38.ncrna). Searches use the default precursor mass tolerances (20 ppm first search and 4.5 ppm main search) and a product mass tolerance of 0.35 Da. The *in silico* digest was set to specific cleavage and a maximum of two missed cleavages. Parameters for each protease (LysC, LysN, chymotrypsin, AspN, GluC, and trypsin) were set in groups. The fixed modification specified were carbamidomethylation of cysteine residues and variable modifications were oxidation of methionine and acetylation of protein N-terminus. Peptides and proteins groups were both filtered to a 1% FDR. Protein groups were filtered for "Only identified by site", "Reverse", and "Contaminant". Gene locus information was mapped to majority protein IDs with HGNC IDs from UniProt and Ensembl BioMart.

Protein coverage calculation.

Sequence coverage for various subsets of runs was calculated with a custom C# application. For each row in the MaxQuant proteinGroups.txt output, all associated peptides were retrieved from peptides.txt. For each peptide, it was first determined if it was found in this subset of runs, using the experiment-based PSM count columns in peptides.txt. If so, the sequence was searched for all occurrences in the sequence of the first major protein of the protein group, ignoring enzyme specificity. A list of unique amino acid residues observed was maintained across all peptides, and at the end the number of residues in the list was divided by the total number

of residues in the major protein sequence. Whenever possible, sequence coverages obtained in this manner were compared with those computed by MaxQuant and included in proteinGroups.txt, and the agreement was excellent. The console C# code is located at https://github.com/cwenger/cwenger.github.io/tree/master/MaxQuantAnalyzer

RNA-seq data and analysis.
The paired RNA-seq data for HeLa S3/HUVEC/HepG2/K562/GM12878/hESC is a part of the ENCODE dataset and was downloaded from SRA (SRP014320). Raw reads were filtered using trimmomatic (version 0.36) using default parameters for paired-end data. Filtered reads were mapped to the human reference genome GRCh38 (Ensemble release 91) using STAR aligner (version 2.5.3a). Further processing – sorting, converting from SAM to BAM format and indexing – was done using SAMtools (version 1.6).

To compare proteomics and transcriptomics data (**Figure 3B**), raw reads per gene were counted in Perseus (version 1.6.14.0), and rows were logorithmised with pseudocount one and normalized by z-scoring for each experiment independently. iBAQ values from the standard proteomics search were summed up for each cell line (through fractions, fragmentation methods, and proteases), logorithmised, z-scored for each cell line independently, and imputed by replacing missing values from the normal distribution (width = 0.3, down shift =1.8) separated for each cell line. After joining the two tables, genes with both protemics and transcriptomics data were used for the PCA plot. Component 1 (accounting for 27.8% of the variance) was not used because it explains the difference between proteomics and transcriptomics data.

Mutation analysis – Transcriptomics.
Non-synonymous mutations were extracted from RNA-seq data of all studied cell lines using “Variation extraction” tool in MaxQuant (Tools/Variation extraction). This tool reports in a fasta file all non-synonymous mutations which pass a list of filters: total reads depth should more or equal than 10, a number of reads with mutations should be more or equal than 5, the frequency of reads with mutations to overall depth should be more or equal than 15%, the base quality, as well as the mapping quality, should be more or equal than 13, which automatically filters out multi-mapped reads. The “Variation extraction” tool generates amongst many output files, *protein.fa* file with all annotated “protein_coding” sequences as well as information about non-synonymous mutations in a header of each sequence.

Mutation analysis – Proteomics.
To enable MaxQuant to use the specified mutations, one has to add the fasta file into the “Fasta files” tab (Global Parameters/Sequences/Fasta files) and change “Variation mode” parameter to “Read from fasta file”. In MaxQuant output `peptides.txt` file an additional columns, such as “Mutated” and “Mutation names” columns, will be created. “Mutated” column reports “No” if one peptide comes from the reference proteome

(without mutations), “Yes” if peptide results from mutation inclusion and “Mixed” if one can find peptides in reference as well as mutated proteome. The “Mutation names” stands for a list of involved mutations.

Splicing Analysis – Transcriptomics and Proteomics.
The analysis of alternative splicing is based on the gene graph structure, where nodes represent the beginning and the end of exons, and edges correspond to exon-exon junctions as well as connections within an exon. Each splicing event in this graph is a local subgraph with multiple paths, however, all paths start from the same node and finish on the same downstream node. It is important to point out that one path can consist of several isoforms. The algorithm is adapted from article. In order to use the same approach for proteomics, protein coordinates of peptides were converted to genome locations, taking into account the intron-exon structure of genes. The modified version of the algorithm is available as a plugin for Perseus software.

***De novo* proteome assembly.**
The peptide spectrum matches (PSM) were extracted from the evidence.txt file and filtered by “Potential contaminant” and “Reverse”. Each PSM was reverse translated into nucleotide sequence with a non-degenerate codon table and written into a FASTA file as input to SOAPdenovo. The SOAPdenovo config file parameters were set to default except for maximal read length to 150. SOAPdenovo-Trans-31mer was run with K-mer length 23 (at least 8 amino acids) and minimum contig length 100 (at least 34 amino acids). Scaffolds from the assembly were matched back to the proteome sequences using brute force string matching.

Data Availability.
All raw mass spectrometry data files and MaxQuant output from the standard search are available at the PRIDE repository. Profiled protein and transcript variants are compiled in the following location.
https://deep-sequencing.app

Description of Supplementary Tables

Supplementary Table 1 gives a summary of the peptide and protein group identifications contributed by various subsets of the data.

Supplementary Table 2 gives a summary of the peptide and PSM counts contributed by various subsets of the data.

Supplementary Table 3 gives the sequence coverage for all identified protein groups and various subsets of the data.

Supplementary Table 4 gives a summary of the non-redundant amino acids detected by trypsin, other enzymes, or all proteases combined.

Supplementary Table 5 gives a summary of all mutations detected in the proteomics and transcriptomics data.

Supplementary Table 6 gives a summary of all splicing events detected in the proteomics and transcriptomics data.

REFERENCES

1. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
2. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S. & Coon, J. J. The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).
3. Richards, A. L., Hebert, A. S., Ulbrich, A., Bailey, D. J., Coughlin, E. E., Westphall, M. S. & Coon, J. J. One-hour proteome analysis in yeast. *Nat Protoc.* **10**, 701–714 (2015).
4. Gholami, A. M., Hahne, H., Wu, Z., Auer, F. J., Meng, C., Wilhelm, M. & Kuster, B. Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Rep.* **4**, 609–620 (2013).
5. Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hogrebe, A., Harder, A. & Olsen, J. V. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
6. Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H. & Pandey, A. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
7. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
8. Tsiatsiani, L. & Heck, A. J. R. Proteomics beyond trypsin. *FEBS J.* **282**, 2612–2626 (2015).
9. Meyer, J. G. *In Silico* Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. *ISRN Comput. Biol.* **2014**, 1–7 (2014).

10. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).
11. Meyer, J. G., Kim, S., Maltby, D., Ghassemian, M., Bandeira, N. & Komives, E. A. Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases. *Mol. Cell. Proteomics* (2014). doi:10.1074/mcp.M113.034710
12. Trevisiol, S., Ayoub, D., Lesur, A., Ancheva, L., Gallien, S. & Domon, B. The use of proteases complementary to trypsin to probe isoforms and modifications. *PROTEOMICS* **16**, 715–728 (2016).
13. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9528–9533 (2004).
14. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
15. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
16. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W. & Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.* **7**, 412-421.e5 (2018).
17. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
18. Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E. G., Van Drogen, A., Borel, C., Frank, M., Germain, P.-L., Bludau, I., Mehnert, M., Seifert, M., Emmenlauer, M., Sorg, I., Bezrukov, F., Bena, F. S., Zhou, H., Dehio, C., Testa, G., Saez-Rodriguez, J., Antonarakis, S. E., Hardt, W.-D. & Aebersold, R. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **37**, 314–322 (2019).
19. Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J. C., Aebersold, R., Venkitaraman, A. R. & Wickramasinghe, V. O. Impact of Alternative Splicing on the Human Proteome. *Cell Rep.* **20**, 1229–1241 (2017).
20. Dominguez, D., Tsai, Y.-H., Weatheritt, R., Wang, Y., Blencowe, B. J. & Wang, Z. An extensive program of periodic alternative splicing linked to cell cycle progression. *eLife* **5**, (2016).

21. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem. Sci.* **42**, 98–110 (2017).
22. Tress, M. L., Abascal, F. & Valencia, A. Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem. Sci.* **42**, 408–410 (2017).
23. Blencowe, B. J. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem. Sci.* **42**, 407–408 (2017).
24. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G. K.-S. & Wang, J. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
25. Guthals, A., Clauser, K. R. & Bandeira, N. Shotgun Protein Sequencing with Meta-contig Assembly. *Mol. Cell. Proteomics* **11**, 1084–1096 (2012).

Chapter 3

Discussion and Outlook

The synergic advancement in proteomics sample preparation, mass spectrometry hardware, and software, greatly improved protein identification and quantification, which was mostly manual and inaccurate in the past[2, 172]. Today, sophisticated algorithms are capable of dealing with millions of spectra from complex proteomes[72]. Computational proteomics has matured substantially to keep up with the massive amounts of data produced by modern mass spectrometers. Platforms for the identification and quantification of proteins can analyze data in a reliable and automated way. For example, our in-house search engine, Andromeda, calculates the probability to observe matches between expected and measured fragment masses by chance[155]. This search engine enables the analysis of complex proteome datasets in combination with MaxQuant, which provides a user-friendly interface for pre-and post-processing of MS data[150]. As a result of these progresses, attention is increasingly being shifted to the downstream part of the data analysis, in which the quantification results are interpreted, hypotheses are tested, and novel biological and biomedical knowledge is gained[159, 173]. Despite all these great advances, there are still key areas in computational proteomics that need automation and the development of new software solutions to enable advances in proteomics workflows and downstream biological applications.

In this thesis, we present key advances in developing novel software solutions that open up new avenues in biological investigations ranging from immunopeptidomics to analyzing microbiomes. For example -

- **Adapting MaxQuant to new data acquisition type adapted to clinical applications - MaxDIA**

  We adopted well-proven practices from DDA to DIA analysis, such as sequential searches with gradually constricting parameters (Bootstrap-DIA), hypersensitive feature detection approach that applied for finding features in the new MS^2 space, and MaxLFQ adaptation for DIA which takes the

pair-wise ratios of fragment peaks too[147]. We also introduce major new features and concepts, such as the usage of machine learning to enhance the score of library-to-DIA matches and to create whole-proteome predicted spectral libraries. The complete end-to-end DIA workflow embedded into the MaxQuant environment allows for analysis of BoxCar-DIA and ion mobility DIA data, demonstrating very high proteome quantification coverage. Future extension of the MaxDIA algorithm could lead to a reliable workflow for detecting PTMs with localization information from DIA data. Since DIA is well known as a state-of-the-art proteomics technique in large scale clinical studies, MaxDIA is going to be battle-tested on many applications.

- **MaxQuant adjusted for scalable and automated processing on Linux systems**

  We have generalized a code structure to meet the expectations of modern proteomics, which is frequently dealing with thousands of samples with complicated experimental structure[157]. Enabling MaxQuant to run on general high computing infrastructures, such as large distributed Linux servers or cloud computers, allow to easily scale any proteomics workflow.

- **Update MaxQuant for easy data visualization - Viewer**

  We developed and thoroughly updated the Viewer component of MaxQuant which now fulfills the demands for rich content visualization of high-resolution proteomics data[158]. The updated MaxQuant version has a map navigation component that steers the users through mass and retention time-dependent mass spectrometric signals. It can be used to monitor a peptide feature used in label-free quantification over many LC-MS runs and visualize it with 3D graphic models. An expert annotation system aids the interpretation of the MS/MS spectra used for the identification of these peptide features[174]. The vector of new instrument developments is pointing toward an increasing dimensionality of the generated data while decreasing sample complexity per volume unit[123]. The future challenge in the visualization of these data would be to find a way to represent it in a user-friendly way.

- **MaxQuant can now extract somatic cancer mutations thereby enabling advanced neoantigen identification workflow**

  We have developed a workflow to extract genomic variants that are translated to the protein sequencing and introduce them into the proteomic search space[53]. Thus enabling a single amino acid variation to the reference proteome. We successfully applied this algorithm to find cell line-specific

mutations and cancer-specific antigens from actual patients. This opens a wide range of future perspectives in personalized cancer treatment and significant challenges[175]. Immunopeptides are famously complicated targets for shotgun proteomics due to the length distribution which is biased toward short values and due to the deficiency of positively charged amino acids. The future improvements in the identification rate of such peptides will likely require innovations in sample preparation and computational proteomics[176]. Hopefully, it will help to bring proteomics closer to the actual clinical applications.

- **Perseus allows the analysis of microbiome composition, genes, and isoform expression**

  We have developed user-friendly tools to analyse expression from NGS data along with proteomics within a Perseus environment[159]. We applied the newly developed algorithm to quantitatively analyze splicing, to detect species composition of saliva microbiome[166] and to find out targets of chaperons[37]. All these discoveries have been done on proteomics and transcriptomics level in parallel. The current development of single-cell technologies[177], including proteomics[88], already creates a demand for novel computational tools that will have to be developed to deal with unique challenges in terms of normalization and handling of missing data within and across different omics[89].

In conclusion, this thesis work demonstrates the successful development of novel algorithms in MaxQuant, Perseus, and multi-omics analysis as listed above. We anticipate that future developments of computational proteomics tools will be particularly active in machine learning[156, 178] and network biology fields[179]. We have continued on our philosophy to enable the end-users - the researchers from fundamental biology, drug discovery[144], and medical sciences - to perform large parts of the data analysis themselves, and this is further demonstrated in our work with neoantigen workflows in Perseus and in developing MaxDIA[53].

Certainly, there is still a large gap between the generation of large-scale proteomics data and the modeling of signaling pathways and biochemical reactions[180]. New tools are emerging to reconstruct signaling pathways and translate them into logic models. With the future development of these tools, large-scale time-series data to kinetic modeling will become more democratically accessible to interdisciplinary researchers, leading to an improved mechanistic understanding of the biological processes. This hopefully will bring us closer to the dream of a reliable in the silico model of a cell.

Acronyms

CNV copy number variant

Da Dalton

DDA data-dependent acquisition

DIA data-independent acquisition

DNA deoxyribonucleic acid

EMC ER membrane complex

ER endoplasmic reticulum

ESI electrospray ionization

FAIMS Field Asymmetric Ion Mobility Spectrometry

HMP human microbiome project

HPLC high performance liquid chromatography

IP Immunoprecipitation

LC liquid chromatography

LFQ label free quantification

m/z mass to charge

MALDI Matrix-assisted laser desorption/ionization

mRNA messenger RNA

MS mass spectrometry

MS/MS tandem MS

MS^1 peptide scan

MS^2 fragment scan

NGS next-generation sequencing

ORF open reading frame

PASEF Parallel Accumulation Serial Fragmentation

PBMC peripheral blood mononuclear cell

PPI protein-protein interaction

PTM post-translational modification

Ribo-Seq ribosome profiling

RNA ribonucleic acid

RNA-Seq RNA sequencing

SILAC stable isotope labeling by amino acids in cell culture

SNV single nucleotide variant

TMT tandem mass tag

TOF time-of-flight

WES whole-exome sequencing

WGS whole genome sequencing

Bibliography

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

[2] Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function (2016).

[3] Azvolinsky, A., DeFrancesco, L., Waltz, E. & Webb, S. 20 years of Nature Biotechnology research tools. *Nature Biotechnology* **34** (2016).

[4] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74** (1977).

[5] Metzker, M. L. Sequencing technologies the next generation (2010).

[6] Mardis, E. R. DNA sequencing technologies: 2006-2016 (2017).

[7] Check Hayden, E. Technology: The $1,000 genome. *Nature* **507** (2014).

[8] Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research* **47** (2019).

[9] Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106** (2009).

[10] Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project (2013).

[11] Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173** (2018).

[12] Heller, M. J. DNA microarray technology: Devices, systems, and applications (2002).

[13] Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63 (2009).

[14] Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number (2009).

[15] Duchon, A. & Herault, Y. DYRK1A, a dosage-sensitive gene involved in neurodevelopmental disorders, Is a target for drug development in down syndrome. *Frontiers in Behavioral Neuroscience* **10** (2016).

[16] Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics* **45**, 1127–1133 (2013).

[17] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5** (2008).

[18] Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320** (2008).

[19] Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5** (2008).

[20] Shendure, J. The beginning of the end for microarrays? *Nature Methods* **5** (2008).

[21] Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

[22] Su, C. H., Dhananjaya, D. & Tarn, W. Y. Alternative splicing in neurogenesis and brain development (2018).

[23] Pachter, L. Models for transcript quantification from RNA-Seq (2011). 1104.3889.

[24] Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28** (2010).

[25] Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS Computational Biology* **4** (2008).

[26] Pervouchine, D. D., Knowles, D. G. & Guigó, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29** (2013).

[27] Sammeth, M. Complete alternative splicing events are bubbles in splicing graphs. In *Journal of Computational Biology*, vol. 16 (2009).

[28] Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** (2015).

[29] ENCODE. An Integrated Encyclopedia of DNA Elements in the Human Genome The ENCODE Project Consortium. *Nature* **489** (2012).

[30] Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458** (2009).

[31] Engreitz, J. M., Ollikainen, N. & Guttman, M. Long non-coding RNAs: Spatial amplifiers that control nuclear structure and gene expression (2016).

[32] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324** (2009).

[33] Ingolia, N. T. Ribosome profiling: New views of translation, from single codons to genome scale (2014).

[34] Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147** (2011).

[35] Jan, C. H., Williams, C. C. & Weissman, J. S. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346** (2014).

[36] Williams, C. C., Jan, C. H. & Weissman, J. S. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science* **346** (2014).

[37] Shurtleff, M. J. *et al.* The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *eLife* **7** (2018).

[38] Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* **109** (2012).

[39] Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486** (2012).

[40] Lasken, R. S. Genomic sequencing of uncultured microorganisms from single cells (2012).

[41] Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348** (2015).

[42] Terekhov, S. S. *et al.* Ultrahigh-throughput functional profiling of microbiota communities. *Proceedings of the National Academy of Sciences of the United States of America* **115** (2018).

[43] Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology* **13** (2015).

[44] Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457** (2009).

[45] Johnson, A. J. *et al.* Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host and Microbe* **25** (2019).

[46] Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574** (2019).

[47] Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569** (2019).

[48] Wilkins, M. R. *et al.* Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and Genetic Engineering Reviews* **13** (1996).

[49] Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual Review of Biochemistry* **80** (2011).

[50] Schwanhüusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473** (2011).

[51] Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Molecular and Cellular Proteomics* **13** (2014).

[52] Blakeley, P., Siepen, J. A., Lawless, C. & Hubbard, S. J. Investigating protein isoforms via proteomics: A feasibility study. *Proteomics* **10** (2010).

[53] Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature Communications* **7**, 13404 (2016).

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

[55] Mateus, A. *et al.* Thermal proteome profiling for interrogating protein interactions. *Molecular Systems Biology* **16** (2020).

[56] Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015). arXiv:1408.1149.

[57] Scaturro, P. *et al.* An orthogonal proteomic survey uncovers novel Zika virus host factors. *Nature* **561** (2018).

[58] Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583** (2020).

[59] Olsen, J. V. & Mann, M. Status of large-scale analysis of posttranslational modifications by mass spectrometry (2013).

[60] Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling (2014).

[61] Millar, A. H. *et al.* The Scope, Functions, and Dynamics of Posttranslational Protein Modifications (2019).

[62] Sharma, K. *et al.* Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Reports* **8**, 1583–1594 (2014).

[63] Humphrey, S. J., Azimifar, S. B. & Mann, M. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nature Biotechnology* **33**, 990–995 (2015).

[64] Kubiniok, P. *et al.* Dynamic Phosphoproteomics Uncovers Signaling Pathways Modulated by Anti-oncogenic Sphingolipid Analogs. *Molecular and Cellular Proteomics* **18** (2019).

[65] Riley, N. M., Hebert, A. S., Westphall, M. S. & Coon, J. J. Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nature Communications* **10** (2019).

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

[67] Ong, S. E., Mittler, G. & Mann, M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nature Methods* **1** (2004).

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

[69] Tanaka, K. *et al.* Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **2** (1988).

[70] Hu, Q. *et al.* The Orbitrap: A new mass spectrometer (2005).

[71] Beck, S. *et al.* The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Molecular and Cellular Proteomics* **14** (2015).

[72] Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science* **1**, 207–234 (2018).

[73] Lundberg, E. & Borner, G. H. Spatial proteomics: a powerful discovery tool for cell biology (2019).

[74] Hein, M. Y., Sharma, K., Cox, J. & Mann, M. Proteomic Analysis of Cellular Systems. *Handbook of Systems Biology* 3–25 (2013).

[75] Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nature Protocols* **11** (2016).

[76] Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms (2016).

[77] Aebersold, R. *et al.* How many human proteoforms are there? (2018).

[78] Chait, B. T. Mass spectrometry: Bottom-up or top-down? (2006).

[79] Rosati, S. *et al.* Exploring an orbitrap analyzer for the characterization of intact antibodies by native mass spectrometry. *Angewandte Chemie - International Edition* **51** (2012).

[80] Srzentić, K. *et al.* Interlaboratory Study for Characterizing Monoclonal Antibodies by Top-Down and Middle-Down Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **31** (2020).

[81] Snijder, J. *et al.* Defining the stoichiometry and cargo load of viral and bacterial nanoparticles by orbitrap mass spectrometry. *Journal of the American Chemical Society* **136** (2014).

[82] Wörner, T. P. *et al.* Resolving heterogeneous macromolecular assemblies by Orbitrap-based single-particle charge detection mass spectrometry. *Nature Methods* **17** (2020).

[83] Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression (2013).

[84] Fields, S. & Song, O. K. A novel genetic system to detect protein-protein interactions. *Nature* **340** (1989).

[85] Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Molecular and Cellular Proteomics* **14**, 120–135 (2015).

[86] Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **6**, 359–362 (2009).

[87] Humphrey, S. J., Karayel, O., James, D. E. & Mann, M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nature Protocols* **13**, 1897–1916 (2018).

[88] Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology* **19** (2018).

[89] Marx, V. A dream of single-cell proteomics. *Nature Methods* **16** (2019).

[90] Kelly, R. T. Single-cell Proteomics: Progress and Prospects (2020).

[91] Cheung, T. K. *et al.* Defining the carrier proteome limit for single-cell proteomics. *Nature Methods* (2020).

[92] Piehowski, P. D. *et al.* Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-μm spatial resolution. *Nature Communications* **11** (2020).

[93] Coscia, F. *et al.* A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *Journal of Pathology* **251** (2020).

[94] Bruderer, R. *et al.* Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Molecular and Cellular Proteomics* **18** (2019).

[95] Geyer, P. E. *et al.* Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Systems* **2** (2016).

[96] Niu, L. *et al.* Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Molecular Systems Biology* **15** (2019).

[97] Zhang, B. *et al.* Clinical potential of mass spectrometry-based proteogenomics (2019).

[98] Aiken, A. C., DeCarlo, P. F. & Jimenez, J. L. Elemental analysis of organic species with electron ionization high-resolution mass spectrometry. *Analytical Chemistry* **79** (2007).

[99] Loos, G., Van Schepdael, A. & Cabooter, D. Quantitative mass spectrometry methods for pharmaceutical analysis (2016).

[100] Niyonsaba, E., Manheim, J. M., Yerabolu, R. & Kenttämaa, H. I. Recent Advances in Petroleum Analysis by Mass Spectrometry (2019).

[101] Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Analytical Chemistry* **60**, 2299–2301 (1988).

[102] Caprioli, R. M., Farmer, T. B. & Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Analytical Chemistry* **69** (1997).

[103] Neagu, A. N. Proteome imaging: From classic to modern mass spectrometry-based molecular histology. *Advances in Experimental Medicine and Biology* **1140** (2019).

[104] Kriegsmann, J., Kriegsmann, M. & Casadonte, R. MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (review). *International Journal of Oncology* **46** (2015).

[105] Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. *Journal of Physical Chemistry* **88** (1984).

[106] Alexandrov, M. L. *et al.* Extraction of ions from solutions under atmospheric pressure as a method for mass spectrometric analysis of

bioorganic compounds. *Rapid Communications in Mass Spectrometry* **22** (2008).

[107] Mann, M. The ever expanding scope of electrospray mass spectrometry—a 30 year journey (2019).

[108] Brunnée, C. The ideal mass analyzer: Fact or fiction? (1987).

[109] Boesl, U. Time-of-flight mass spectrometry: Introduction to the basics (2017).

[110] Paul, W. & Steinwedel, H. Ein neues Massenspektrometer ohne Magnetfeld (1953).

[111] Kingdon, K. H. A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Physical Review* **21** (1923).

[112] Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Analytical Chemistry* **85**, 5288–5296 (2013).

[113] Makarov, A. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry* **72** (2000).

[114] Lange, O., Damoc, E., Wieghaus, A. & Makarov, A. Enhanced Fourier transform for Orbitrap mass spectrometry. *International Journal of Mass Spectrometry* **369** (2014).

[115] Makarov, A. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry* **78** (2006).

[116] Michalski, A. *et al.* Ultra high resolution linear ion trap orbitrap mass spectrometer (orbitrap elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular and Cellular Proteomics* **11** (2012).

[117] Scheltema, R. A. *et al.* The Q exactive HF, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field orbitrap analyzer. *Molecular and Cellular Proteomics* **13**, 3698–3708 (2014).

[118] Bekker-Jensen, D. B. *et al.* A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Molecular and Cellular Proteomics* **19** (2020).

[119] Makarov, A. Orbitrap journey: taming the ion rings (2019).

[120] May, J. C. & McLean, J. A. Ion mobility-mass spectrometry: Time-dispersive instrumentation (2015).

[121] Cooper, H. J. To What Extent is FAIMS Beneficial in the Analysis of Proteins? *Journal of the American Society for Mass Spectrometry* **27** (2016).

[122] Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Analytical Chemistry* **90** (2018).

[123] Prianichnikov, N. *et al.* Maxquant software for ion mobility enhanced shotgun proteomics. *Molecular and Cellular Proteomics* **19** (2020).

[124] Yu, F. *et al.* Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Molecular & cellular proteomics : MCP* **19** (2020).

[125] Meier, F. *et al.* Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *Journal of Proteome Research* **14** (2015).

[126] Meier, F. *et al.* Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Molecular and Cellular Proteomics* **17** (2018).

[127] Mann, M. Functional and quantitative proteomics using SILAC (2006).

[128] Cox, J. & Mann, M. Computational Principles of Determining and Improving Mass Precision and Accuracy for Proteome Measurements in an Orbitrap. *Journal of the American Society for Mass Spectrometry* **20** (2009).

[129] Krüger, M. *et al.* SILAC Mouse for Quantitative Proteomics Uncovers Kindlin-3 as an Essential Factor for Red Blood Cell Function. *Cell* **134** (2008).

[130] Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Molecular and Cellular Proteomics* **12** (2013).

[131] Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods* **7** (2010).

[132] Tyanova, S. *et al.* Proteomic maps of breast cancer subtypes. *Nature Communications* **7** (2016).

[133] Shenoy, A. & Geiger, T. Super-SILAC: Current trends and future perspectives. *Expert Review of Proteomics* **12** (2014).

[134] Walther, D. M. *et al.* Widespread proteome remodeling and aggregation in aging C. elegans. *Cell* **161** (2015).

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

[136] Hebert, A. S. *et al.* Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature Methods* **10** (2013).

[137] Overmyer, K. A. *et al.* Multiplexed proteome analysis with neutron-encoded stable isotope labeling in cells and mice. *Nature Protocols* **13** (2018).

[138] Merrill, A. E. *et al.* NeuCode labels for relative protein quantification. *Molecular and Cellular Proteomics* **13** (2014).

[139] Lau, H. T., Suh, H. W., Golkowski, M. & Ong, S. E. Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics. *Journal of Proteome Research* **13** (2014).

[140] Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry* **75**, 1895–1904 (2003).

[141] Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nature Methods* (2020).

[142] Cox, J., Yu, S. H. & Kyriakidou, P. Isobaric matching between runs and novel PSM-level normalization in maxquant strongly improve reporter ion-based quantification. *Journal of Proteome Research* **19** (2020).

[143] Franken, H. *et al.* Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature Protocols* **10** (2015).

[144] Ball, K. A. *et al.* An isothermal shift assay for proteome scale drug-target identification. *Communications Biology* **3** (2020).

[145] Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180** (2020).

[146] Mertins, P. *et al.* Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. *Nature Protocols* **13** (2018).

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

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

[149] Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications* **8** (2017).

[150] Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008). nbt.1511.

[151] Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews* **113**, 2343–2394 (2013).

[152] Röst, H. L. *et al.* OpenMS: A flexible open-source software platform for mass spectrometry data analysis (2016).

[153] Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial. *Molecular Systems Biology* **14** (2018).

[154] Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics (2020).

[155] Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10** (2011).

[156] Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods* **16** (2019).

[157] Sinitcyn, P. *et al.* MaxQuant goes Linux. *Nature Methods* **15**, 401 (2018).

[158] Tyanova, S. *et al.* Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* **15**, 1453–1456 (2015).

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

[160] Ishida, Y., Agata, Y., Shibahara, K. & Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO Journal* **11** (1992).

[161] Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **271** (1996).

[162] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499** (2013).

[163] Kalaora, S. *et al.* Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* **7** (2016).

[164] Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* (2014).

[165] Angeletti, S. Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology (2017).

[166] Grassl, N. *et al.* Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome Medicine* **8** (2016).

[167] Cymer, F., Von Heijne, G. & White, S. H. Mechanisms of integral membrane protein insertion and folding (2015).

[168] Jonikas, M. C. *et al.* Comprehensive Characterization of Genes Required for Protein Folding. *Science* **323** (2009).

[169] Liu, Y. *et al.* Impact of Alternative Splicing on the Human Proteome. *Cell Reports* **20** (2017).

[170] Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity (2017).

[171] Wang, X. *et al.* Detection of proteome diversity resulted from alternative splicing is limited by Trypsin cleavage specificity. *Molecular and Cellular Proteomics* **17** (2018).

[172] Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **4**, 787–797 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17901868.

[173] Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12** (2015).

[174] Neuhauser, N., Michalski, A., Cox, J. & Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Molecular and Cellular Proteomics* **11** (2012).

[175] Nesvizhskii, A. I. Proteogenomics: Concepts, applications and computational strategies (2014).

[176] Faridi, P., Purcell, A. W. & Croft, N. P. In Immunopeptidomics We Need a Sniper Instead of a Shotgun. *Proteomics* **18** (2018).

[177] Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the science (2016).

[178] Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16** (2019).

[179] Rudolph, J. D. & Cox, J. A Network Module for the Perseus Software for Computational Proteomics Facilitates Proteome Interaction Graph Analysis. *Journal of Proteome Research* **18**, 2052–2064 (2019).

[180] Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nature Biotechnology* **38** (2020).

List of Figures

Acknowledgements

Above all, I would like to thank my supervisor, Dr. Jürgen Cox, for his outstanding training and everyday support. Through all ups and downs of my Ph.D., you help me find solid scientific ground and believe in myself.

I cannot find enough words to express how lucky I am to be a member of the Cox group. You are awesome, my folks!

I thank all current and past members of my thesis committee, Prof. Dr. Thomas Carell, Prof. Dr. Dmitrij Frishman, Dr. Franz Herzog, Prof. Dr. Matthias Mann, Prof. Dr. Klaus Förstemann, and Prof. Dr. Julian Stingele.

Kirti, Tikira, Michal, Jacob, and Alexey - thank you for all scientific discussions without end, collaborations, and your support.

Georg, thank you for all scientific lunch discussions and not taking my jokes too seriously.

I would like to express a very special thanks to my wife, Julie Rojas, for her critical view of my research and thesis writing. Every day you make me read and feel that there is no place like home.

Also many thanks to Julie's family for constantly pushing me to finish my Ph.D. thesis and learn French - I am going to take serious care of the second part soon.

I am grateful to my family for all your unconditional support and the freedom to pursue what I truly like. Спасибо мама и папа!