TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Proteomik und Bioanalytik

An in-memory platform for the exploration and analysis of big data in biology

Mathias Wilhelm

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.- Prof. Dr. Dmitrij Frischman
Prüfer der Dissertation:	1. Univ.- Prof. Dr. Bernhard Küster 2. Univ.- Prof. Dr. Hans-Werner Mewes 3. Univ.- Prof. Dr. Oliver Kohlbacher

Die Dissertation wurde am 22.12.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 07.03.2017 angenommen.

“We can only see a short distance ahead,
but we can see plenty there that needs to be done.”
- Alan Turing

Abstract

Mass spectrometry-based proteomics has become the leading technology to identify and quantify thousands of proteins in a single experiment and has plenty of applications in discovery and targeted experiments ranging from characterizing biological samples, over drug-protein interaction analysis to biomarker discovery and patient classification. Providing access to previously conducted experiments is key to make use of the wealth of data in order to correlate or cross-compare studies. In the past, multiple databases and platforms have been developed to address questions arising in both wet and dry lab, but these lack depth, performance and versatility. If data are available in public repositories, their annotation is often superficial and the data generation and processing platforms are of varying capability, performance and maturity. Importantly, there is also a significant challenge in making 'big data' more widely accessible to the scientific community because the development of scalable analysis tools is only in its infancy.

Chapter 2 describes the implementation and design choices made to build a versatile and performant database to store and analyze bottom-up mass spectrometry-based proteomics data, termed ProteomicsDB. Due to the use of the in-memory database technology SAP HANA, this system not only allows the integration of thousands of proteomic experiments on both identification and quantification level but also to perform complex queries. The addition of the experimental design, a versatile data model to portray the heterogeneity of proteomics experiments, allows ProteomicsDB to model and visualize complex experimental setups. This is illustrated on two assays that are used to find protein-drug interactions. The integration of multiple experiments and transcriptomic data allows cross experiment analysis and illustrates the utility beyond protein expression profiles. ProteomicsDB thus enables the navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

In chapter 3 the assembly and analysis of a first draft of the human proteome is described. For this purpose, re-analyzed results of more than 16,000 raw data files were imported into ProteomicsDB providing peptide level evidence for about 18,000 human genes. The information assembled from human tissues, cell lines and body fluids allowed estimating the size of the protein coding genome and identified organ-specific proteins. Analysis of mRNA and protein expression profiles of human tissues revealed conserved control of protein abundance, integration of drug sensitivity data allowed the identification of proteins predicting resistance or sensitivity to targeted cancer drugs and proteome profiles also hold considerable promise for analyzing the composition and stoichiometry of protein complexes. However, this assembly also highlighted major issues in the standard approach of calculating the protein FDR in 'big data sets'.

Chapter 4 addresses the issue of estimating the protein FDR in large scale studies, databases and repositories. Here, a simple and readily available adjustment, termed the 'picked approach', of the commonly used classical protein FDR model is described which allows an unbiased, scalable and precise estimation of the number of false positive identification. The picked protein FDR approach treats target and decoy sequences of the same protein as a pair rather than individual entities and chooses either the target or the decoy sequence depending on which receives the highest score. The results suggest that this method scales to any size, is less susceptible to low quality, noisy data and can be used on both protein and gene level while also increasing the number of identified proteins at low FDR cutoffs.

Zusammenfassung

Massenspektrometrie-basierte Proteomik ist zur Standardmethode in der Proteinanalytik geworden und ermöglicht die gleichzeitige Identifizierung und Quantifizierung tausender Proteine. Ihr Anwendungsbereich reicht von der Charakterisierung unbekannter Proben, der Analyse von Protein-Wirkstoff-Wechselwirkungen bis zur Identifizierung von Biomarkern und zur Klassifizierung von Patienten. Die Bereitstellung erhobener und bereits publizierter Daten erlaubt es, Wissenschaftlern eigene Ergebnisse und veröffentlichte Experimente zu vergleichen und zu validieren. Für diesen Zweck wurden einige Datenbanken entwickelt, jedoch fehlt es vielen an Tiefe, Performance und Vielseitigkeit, um komplexe Experimente miteinander zu vergleichen. Daten in öffentlichen Datenbanken sind oft nur unzureichend annotiert und von variierender Qualität. Zudem befindet sich die Bereitstellung von großen proteomischen Datenmengen noch in der Anfangsphase, da skalierbare Methoden und Applikationen für die interaktive Exploration und Analyse weitgehend fehlen.

Kapitel 2 beschreibt die Entwicklung und Implementierung von ProteomicsDB, einer vielfältig einsetzbaren und performanten Datenbank für die Speicherung und Analyse von Daten aus *bottom-up*-Proteomik-Experimenten. Der Einsatz von HANA, einer *In Memory* Datenbank, entwickelt von SAP als Datenbankmanagementsystem, erlaubt hierbei die Integration von Ergebnissen aus tausenden Experimenten. Die Ablage des experimentellen Designs ermöglicht sowohl die Visualisierung komplexer Zusammenhänge als auch vergleichenden Analysen von Ergebnissen aus verschiedenen Experimenten. Deutlich gemacht wird dies am Beispiel der Wirkstoffforschung, indem zwei Experimenttypen zur Aufklärung von Protein-Wirkstoff-Beziehungen unterstützt werden. Diese Integration und die Erweiterung des Datenbankmodells auf quantitative Transkriptdaten zeigt die Nützlichkeit von ProteomicsDB jenseits von Proteinexpressionsdaten.

In Kapitel 3 wird die Zusammenstellung und Analyse eines ersten Entwurfs des menschlichen Proteoms beschrieben. Hierfür wurden die Ergebnisse von mehr als 16.000 Datensätzen in ProteomicsDB importiert, die zusammengenommen Evidenz für rund 18.000 menschliche Gene beinhalten. Die Daten umfassen Experimente von menschlichen Zelllinien, Geweben und Körperflüssigkeiten und erlauben damit in erster Näherung eine Abschätzung des translatierten menschlichen Genoms und die Identifizierung Gewebe-spezifischer Proteine. Desweiteren wird gezeigt, dass Proteinexpressionsdaten sowohl für die Vorhersage von Sensitivität- und Resistenzmarkern für Wirkstoffe als auch zur Bestimmung der Zusammensetzung und Stöchiometrie von Proteinkomplexen verwendet werden kann. Die integrative Analyse von mRNA- und Proteinexpressionsdaten offenbarte ein konserviertes Verhältnis von Transkript- und Proteinmenge, das zur Vorhersage von Proteinabundanzen verwendet werden kann. Die Zusammenstellung der Daten legt jedoch auch offen, dass die Bestimmung der *false discovery rate* auf Proteinebene in großen Datensätzen mit Hilfe des Standardmodells nur sehr eingeschränkt möglich ist.

Kapitel 4 beschäftigt sich mit der Abschätzung von Protein *false discovery rate* in großen Einzelstudien, Datenbanken und Repositorien. Hier wird eine einfache und leicht einsetzbare Methode, der sogenannte „picked"-Ansatz, vorgestellt, der die Standardmethode erweitert und eine genaue und unvoreingenommene Abschätzung der *false discovery rate* liefert. Im Vergleich zur Standardmethode werden beim „picked"-Ansatz Target- und Decoy-Sequenzen eines Proteins als zusammengehörendes Paar und nicht als einzelne Sequenzen interpretiert. Vor Berechnung der FDR wird pro Paar jeweils nur die Sequenz mit höchster Konfidenz ausgewählt und die Akkumulierung von falsch-Positiven verhindert. Die Ergebnisse zeigen, dass die hier vorgestellte Methode auf beliebige Datensatzgrößen skalierbar ist, weniger anfällig gegenüber schlechter Datenqualität ist und sowohl auf Protein- als auch Gene-Ebene anwendbar ist.

Table of contents

Chapter 1

General Introduction

Contents

"It always takes longer than you expect, even when you take into account Hofstadter's Law"
- Hofstadter's Law; Gödel, Escher, Bach: An Eternal Golden Braid

1 From genomics to proteomics

More than a decade ago, an international research effort changed todays view on all major areas of life. The completion of the Human Genome Project[1] led to many technological[2] and scientific[3] advancements and even affected modern legislation and politics[4]. With the advent of genomics and its technologies, culminating in high-throughput next generation sequencing[5], the routine sequencing of entire genomes and quantification of ribonucleic acid (RNA) even in single cells is possible. Today, this allows the systematic interrogation of the dynamics of transcription[6,7], enabling us to investigate processes such as alternative splicing, mRNA processing and gene expression[8].

While the genome is generally viewed as the blueprint of an organism, the complexity of a living organism is largely determined by the dynamic and versatile nature of its products. Following transcription, messenger RNA (mRNA) is translated into proteins, which carry out almost all chemical reactions in cells. Proteomics, the study of proteomes and their function, provides a complementary approach to study the molecular processes of living organisms by adding yet another level of complexity. Fueled by the ability to decipher the genetic code, proteomics enables the methodical interrogation of processes such as mRNA translation, protein stability, protein-protein interactions, protein localization, and post-translational modifications. However, the ultimate goal of proteomics is to identify and quantify all protein isoforms including their present modifications in any living system simultaneously.

Proteomics has experienced a significant evolution within the past decades. Starting with two-dimensional gel electrophoresis[9,10], especially the advances in mass spectrometry[11,12] enabled the identification and quantification of more than 10,000 proteins in single cell lines[13,14], covering up to 10 orders of magnitude in dynamic range of expression[15]. Mass spectrometers have proven its applicability in a broad range of applications in history and led to major scientific discoveries even on distant planets[16]. Today, mass spectrometry is an irreplaceable tool for the analysis of a wide range of (bio) molecules[17-20], enabling the exploration of almost all biological processes. However, especially the large number of applications in the field of proteomics has led to its vital role in bio sciences[12,21,22], ranging from discovery[23] to targeted experiments[24,25] covering e.g. the characterization of biological samples[13,14], biomarker discovery[26], patient classification[27], signaling pathway analysis[28] and drug discovery[22]. However, in order to reach the ultimate goal of proteomics, many technical but also computational challenges lay ahead.

The amount and complexity of data generated by genomics, transcriptomics and proteomics propelled the development of automatic processing, annotation and storage tools[29]. Covered by the field of bioinformatics[30], particularly today, mass spectrometry-based proteomics requires novel and sophisticated algorithms and tools to address both data processing and data integration. This need gave rise to computational proteomics[31], a field of research dedicated to improve and simplify data acquisition, processing, analysis, integration and interpretation[26,32-35].

Due to the complexity of proteomic experiments, defining a unified facility to store well annotated results is challenging. While many efforts to collect and integrate publicly available proteomics datasets exist[36,37], it is often difficult to retrieve a comprehensive list of identified proteins in a specific biological source or a list of biological sources where a specific protein or post-translational modification is present. Moreover, the lack of integrated meta data[24,38] and quantitative information often only enables the interaction with identification data, rendering this valuable part of the data inaccessible and futile.

2 Mass spectrometry-based proteomics

Mass spectrometry-based proteomics is divided into two prevalent paradigms[39]. The "top-down"[40] approach studies intact proteins and thus enables the identification of proteoforms. However, due to the large diversity of proteins with respect to biochemical and physical properties, both sample preparation and data acquisition are difficult and hinder the systematic and automatic analysis of complex mixtures. The alternative is the more commonly applied "bottom-up" approach[17,23,41,42]. Here, proteins are digested into peptides using site-specific proteases prior to mass-spectrometric analysis (see Figure 1.1). Depending on the complexity of the resulting peptide mixture, subsequent on- or off-line separation, most commonly utilizing liquid chromatography (LC), is necessary. Modern (tandem) mass spectrometers (MS) enable the identification and quantification of tens of thousands of peptides to infer the presence and abundance of proteins in almost any biological sample.

Due to the scope of this thesis, the main focus is on bottom-up mass spectrometry-based proteomics. This section briefly covers all general aspects of sample acquisition starting with the sample preparation, basics in mass spectrometry and its application and acquisition methods in proteomics and quantification approaches.

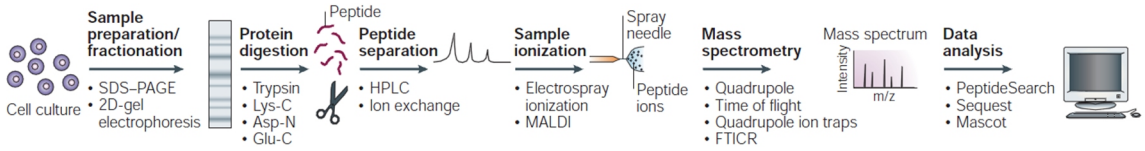


Figure 1.1 | Generic bottom-up proteomics workflow. A typical mass spectrometry-based proteomics workflow includes (1) protein extraction, (2) protein digestion, (3) peptide separation, (4) sample ionization, (5) MS measurement and (6) data analysis. Figure from[42].

2.1 Sample preparation

The general goal of sample preparation is to resolve and identify as many proteins in a complex biological matrix as possible or to enrich sub-proteomes or complexes which are otherwise not accessible[43,44]. Even though a generic bottom-up proteomics workflow exists (Figure 1.1), individual steps can be achieved and modified by several different means. After the extraction of proteins from a biological sample such as cells, tissues or body fluids, the protein mixture can be further separated[45] by immunoassays or other affinity extractions, chromatography and, more commonly, electrophoresis[46]. Subsequently, proteins are digested into peptides using sequence-specific proteases. The most commonly used protease is trypsin, which specifically cleaves proteins on the carboxyl-terminal side of lysine and arginine residues[47]. Trypsin enjoys great popularity since the resulting peptides contain a basic residue at the C-terminus and an average length of 10-14 amino acids, both highly desired properties for subsequent MS analysis. However, alternative proteases such as Lys-C, Asp-N, Glu-C can be used to generate complementary peptides which can significantly increase the sequence coverage in comparison to solely using trypsin[48,49] and also enable access to proteins which do not generate MS-accessible or any tryptic peptide.

The resulting complex mixture of peptides typically exceeds the capacity on any analytical acquisition method used to date[50]. Especially when trying to analyze post-translationally modified (PTM) peptides additional purification steps are necessary to enrich these typically low abundant

peptide species[51-55]. Thus, prior to injection of the analyte into the mass spectrometer, additional offline peptide separation techniques can be employed to further decrease the complexity of the sample and thus increase the number of identified peptides[56]. Sample acquisition benefits most when orthogonal dimensions of separation are used. Common approaches for peptide separation utilizing orthogonal separation techniques are isoelectric focusing (IEF)[57,58], strong cation or anion exchange chromatography (SCX and SAX)[59] or hydrophilic interaction chromatography (HILIC)[60].
Liquid chromatography (LC) can be coupled directly (online) to the mass spectrometer (LC-MS). Due to the high compatibility of the solvent components (water, acetonitrile, organic acids), reversed phase ion-pairing separation is almost exclusively used in mass spectrometry-based proteomics applications. The general principle is that under aqueous acidic conditions the protonated peptides are retained on the C_{18} material (stationary phase) of the chromatographic column. This force is reduced when an increasing percentage of organic solvent is added to the mobile phase, thus increasing the hydrophobicity. A typical setup used for online peptide separation in nanoflow LC-MS uses column lengths of 10 to 50 cm with an inner diameter of 50 to 100 μm, a particle size of 1 to 5 μm and an applied flow rate of 100 to 500 nL/min.

2.2 Mass spectrometry

A mass spectrometer generally is comprised of three parts: 1) an ionizer, 2) a mass analyzer and 3) an ion detector. After ionization, the analyte is transferred into the mass spectrometer via an electrostatic potential. Subsequent separation and detection of ions reaching the detector generate a mass spectrum which records the measured signal at an *m/z* (mass-to-charge) value.

2.2.1 Sample ionization

Ionization describes the process of adding charges to a molecule of interest. Several ionization methods are available, where the principle is the same: Removing or adding protons or electrons causes the molecule to retain one or multiple charges.
Electrospray Ionization (ESI) is the most commonly used ionization technique in mass spectrometry-based proteomics[61]. It allows an automated analysis of peptide mixtures by means of LC, due to the direct infusion of the sample into the detector (Figure 1.2). For this purpose, a small needle is filled with a solution containing the molecules of interest. A high voltage is applied between the needle and the detector entrance, which separates the charges at the surface of the fluid and forces the fluid to emerge from the needle, creating an aerosol. The resulting droplets are attracted to the entrance of the detector and, in the transition, the volatile solvents (mostly acetonitrile) evaporate (desolvation) until they become unstable upon reaching the Rayleigh limit. Due to the ever-decreasing size of the droplets, the electrostatic repulsion becomes more powerful than the surface tension of the droplets, which leads to Coulomb fission, whereby the original droplets explode. The newly created droplets again undergo desolvation and Coulomb fission. The exact mechanisms of how gas-phase ions are produces is still debated, but two main theories, the ion evaporation model and the charge residue model, exist[62,63]. In bottom-up proteomics, ESI generates mostly doubly or higher charged peptides. The efficiency of ionization can be enhanced by the introduction of additives, such as DMSO[64].

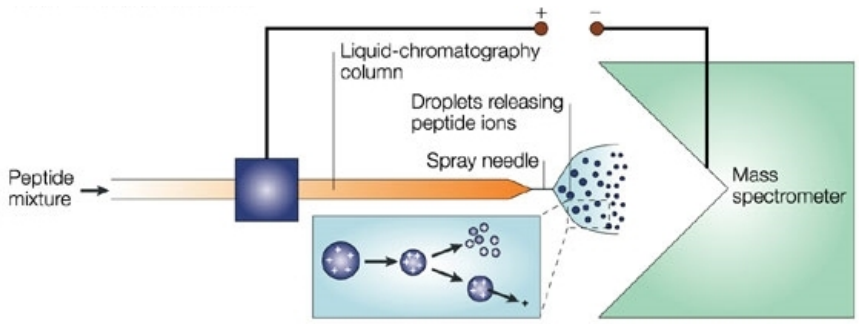


Figure 1.2 | Schematic visualization of electrospray ionization. A complex mixture of peptides is separated by liquid chromatography. Prior to the injection into the mass spectrometer, peptides are ionized by a applying a high voltage between the electrospray needle and the entrance to the mass spectrometer. The emerging droplets release charged peptides which can be manipulated and measured by a mass analyzer and detector. Figure from[42].

Nanoelectrospray Ionization (NanoESI) is a variant of ESI with a very small needle diameter[65]. The ionization efficiency is increased and thus smaller amounts of the sample are needed. A convenient side effect of having a smaller needle and smaller droplets is less evaporation is needed, which means that solvent impurities are less concentrated as compared to ionization with ESI.

2.2.2 Mass analyzers and ion detectors

A mass analyzer measures ions with respect to their mass-to-charge (*m/z*) ratio by separating them in space or time. Their performance is generally described by two different terms. First, the resolution or resolving power R depends on the full width at half maximum of an *m/z* peak and the expected *m/z* of an ion, and describes the ability to differentiate an ion from any other. Contrary, the accuracy describes the ability to calibrate an instrument against a known entity.

Electron multipliers are ion detectors commonly used in combination with mass analyzers which do not have an integrated detector. They consists of a series of dynodes which emit electrons upon the impact of an ion. Each dynode in this series is held at a higher potential, causing more electrons to be expelled in each step. This cascade results in a detectable electrical signal which can be recorded.

Ion trap[66,67] mass analyzers or quadrupolar traps are typically composed of four parallel rod-shaped electrodes. The general mechanisms of confining ions in space is realized by applying direct (DC) and alternating current (AC) to opposing rods. The AC, also referred to as main RF due to its frequency, confines ions radially, whereas DC creates a potential well for axial confinement. Trapped ions are in a cork-screw like (secular) motion, proportional to the main RF amplitude and the mass of the ion. Ion trap scanning employs “resonance ejection”, which is realized by applying an additional AC to the exit rods within the trap. For this purpose, both the additional AC and the main RF are ramped so that ions of different *m/z* enter resonance with the exit rod and are ejected through their slits. During scanning, the number of ejected ions are recorded using electron multiplier. For isolation, all frequencies necessary for the ejection of unwanted ions are superimposed, resulting in a complex AC waveform. The most widely used two dimensional linear ion trap enables ions to spread out axially more than in three dimensional traps, increasing their capacity.

Quadrupole (Q) mass filters, similar to ion traps, consist of four parallel metal rods[67,68]. In contrast to traps ions are, depending on the field, moving through the quadrupole. This is possible because a DC with equal amplitude but opposing sings is applied to pairs of rods. Similar to traps, changing the DC and AC amplitude influences the ions' movement depending on the *m/z*. The secular motion of the ions is altered by the AC and can be used to let ions with small m/z collide with or pass through the rods (high mass pass filter). Similarly, the DC is used to "eject" ions with high *m/z* (low mass pass filter). While the quadrupole is often used as a mass filter, enabling the isolation of ions within a specific *m/z* range, it can also be used for scanning, but also relies on subsequent ion detectors such as the electron multiplier.

Time-of-flight (TOF) mass analyzers utilize an electric field to accelerate ions in a high vacuum to the same kinetic energy[69]. Due to their higher velocity, lighter ions will reach the ion detector earlier than heavier ions. The time an ion needs to arrive at the detector is used to calculate the *m/z* of the ion. The reflectron TOF (re-TOF) uses a constant electrostatic field to reflect ions before they arrive at the detector. It combines the TOF technology with an electrostatic mirror, the reflector, to increases the time ions need to reach the ion detector. The reflector also reduces the variance of the kinetic energy of the ions and, in combination with the increased flight path, results in a higher resolution.

Fourier transform (FT) mass analyzers use the principle of monitoring the motion of ions in a magnetic field[70]. After excitation, the ions orbit at their cyclotron frequency as a coherent cluster. The induced image current at two electrodes is recorded and by performing an Fourier transform the mass to charge ratio of the oscillating ions can be deduced. Frequencies can be measured with very high accuracy and thus the resolution of FT mass analyzers increases when acquiring the FT spectrum for a longer time (increased transient time). Typically, several quadrupole ion guides are used to select and direct the ions into the FT mass analyzer. An implementation of an FT MS is the Fourier transform ion cyclotron resonance mass spectrometer (FT ICR)[71].

Orbitrap mass analyzers are part of the FT MS family[72,73]. An electric field is applied between an outer barrel-like electrode and an inner spindle-like electrode. Ions are injected tangentially to the electric field which causes the ions to move in a stable orbit around the inner electrode, balanced by their centrifugal force. This equilibrium also forces ions with lower *m/z* closer to the inner spindle. In contrast to FT MS, the ions show an axial oscillating movement along the inner electrode. The frequency of the axial movement is inversely proportional to the square root of the *m/z* value, which is used to calculate the *m/z* of the ions. Recent advancements further increased the mass accuracy and resolving power by the introduction of a compact high-field Orbitrap and an enhanced Fourier transform algorithm[11,74,75], rendering Orbitrap mass analyzers one of the most commonly used mass spectrometer in discovery bottom-up mass spectrometry-based proteomics experiments.

2.3 Tandem mass spectrometry

Tandem mass spectrometry enables the identification of the primary sequence of a peptide. For this purpose a first mass spectrum of the intact peptides is recorded, referred to as full or survey scan (Figure 1.3). These spectra are typically recorded with high mass accuracy and resolution to allow the precise calculation of the neutral peptide mass. To derive sequence information, a subsequent tandem fragmentation spectrum (product ion spectrum or MS/MS spectrum) is acquired. Therefore, a peptide ion of interest is selected (precursor selection) and a population of

that ion is collected for fragmentation, which introduces random breaks in the peptide backbone thus generating a population of fragment ions. The position and differences of the resulting fragment ions can be used to determine the sequence of the selected peptide.

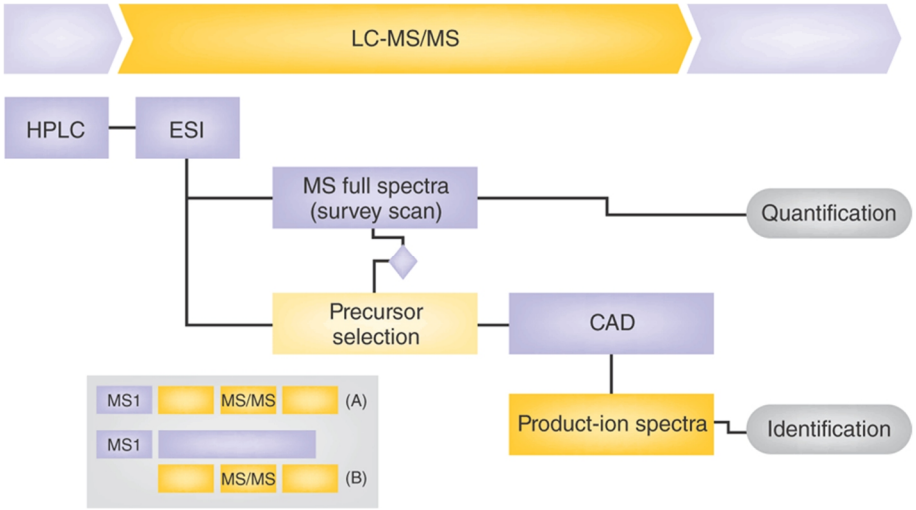


Figure 1.3 |Acquisition schema of a bottom-up shotgun proteomics experiments. The peptide mixture is separated by HPLC and analyzed by MS in full-scan mode. Using simple data-dependent acquisition heuristics based on signal intensity, peptide ions are selected for fragmentation and dissociated by collisional activation. The resulting MS/MS spectra permit determination of the amino acid sequence of the fragmented peptide. The intensity of the precursor ion signal in the survey scan is used for quantification. The insert indicates the different modes of acquisition; either sequential MS and MS/MS analysis as performed using a quadrupole/time-of-flight instrument (A), or parallel analysis as performed on a linear ion trap/Orbitrap mass spectrometer (B). Figure from[26].

A commonly employed method to select precursor ions is data-dependent acquisition (DDA) which chooses intact peptide ions based on their signal intensity in the survey scan. To avoid multiple selection of the same peptide ion, the selected neutral mass is temporarily stored in a dynamic exclusion list which is maintained by the mass spectrometer.

2.3.1 Fragmentation

A variety of fragmentation techniques were developed and implemented[76-78] to derive structural information about a peptide. For this purpose, random breaks in the backbone of the peptide are induced and in an optimal case produce all possible fragment ions along the peptide backbone. In principle, the peptide backbone can break at three positions (Figure 1.4). The nomenclature of the resulting peptide fragments was first described by Roepstorff and Fohlmann[79], followed by Johnson et al.[80]. Fragments containing the N-terminal site of the peptide are termed a_n, b_n, and c_n-ions whereas C-terminal containing fragments are named x_n, y_n and z_n-ions. Here, *n* indicates the position of the break within the peptide backbone.

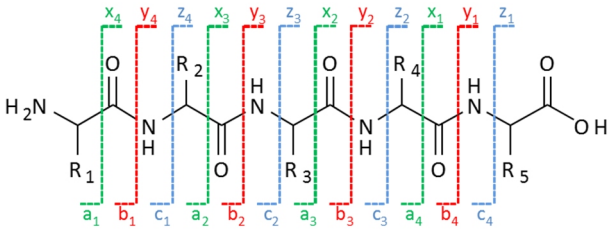


Figure 1.4 | Peptide fragmentation nomenclature according to Roepstorff and Fohlmann[79,80]. N-terminal fragments are named a_n, b_n, c_n ions and C-terminal fragments are x_n, y_n, z_n.

In collision-induced dissociation (CID) ions are vibrationally excited by an electrical potential to a high kinetic energy[81]. This is typically performed in an ion trap using the same principle as for ejection. However, because the ion trap is filled with an inert gas, such as helium or nitrogen, ions stay in resonance without ejection. To avoid the balance between scanning/filtering and fragmentation, the dual linear ion trap consists of a low- (scanning) and high- (fragmentation) pressure cell. In the high-pressure cell, ions will eventually collide with molecules of the inert gas. The collision converts some of the kinetic energy into internal energy, which causes the weak peptide bonds to break and thus lead to fragmentation. Because the resulting fragment ions have lower *m/z*, the applied AC does not further excite them, thus preventing further fragmentation. However, this leads to problems in the analysis of labile modifications, such as phosphorylation. Weak bonds preferentially break and because of no further fragmentation, no structural information about the peptide is generated. CID generates predominantly b, and y-ions.
Higher-energy collisional dissociation (HCD), also termed high-energy CID or beam-type CID uses the same principle as CID but with higher collision energies[82]. This is achieved by accelerating ions by a stronger electric field which is commonly applied between the first mass analyzer and a dedicated collision chamber. Again, the collision chamber is filled with an inert gas but in contrast to CID, peptide bonds fragment almost instantaneously, leading to information-rich spectra with mostly b and y-ions as well as internal and immonium fragments. Because fragment ions can further collide, HCD is preferentially used for the analysis of labile modifications.
Electron-transfer dissociation (ETD) fragments multiple protonated molecules by transferring electrons[83]. It utilizes radical anions, such as fluoranthene, to break the backbone of peptides by generating a charge-reduced species with an unpaired electron (odd-electron molecule). Side chains and peptide modifications are generally left intact. ETD predominantly produces c and z-ions and is, for instance, implemented in the Orbitrap XL. However, ETD requires higher charge states to induce efficient fragmentation and is thus not often used for tryptic peptides which predominately ionize as doubly-charged peptides during ESI.

Figure 1.5 shows the MS/MS spectrum of the doubly-charged peptide LTQLGTFEDHFLSLQR upon HCD fragmentation. Here, the entire y- (red) and almost entire b-ion (blue) series was generated leading to a complete sequence coverage of the peptide. For clarity, the annotation of the singly-charge precursor (*m/z* 1904.98653), neutral losses (predominantly $-H_2O$, $-NH_3$ on b- and y-ions) and immonium ions were suppressed, but are able to explain the majority of the non-annotated peaks (black) in the fragment spectrum.

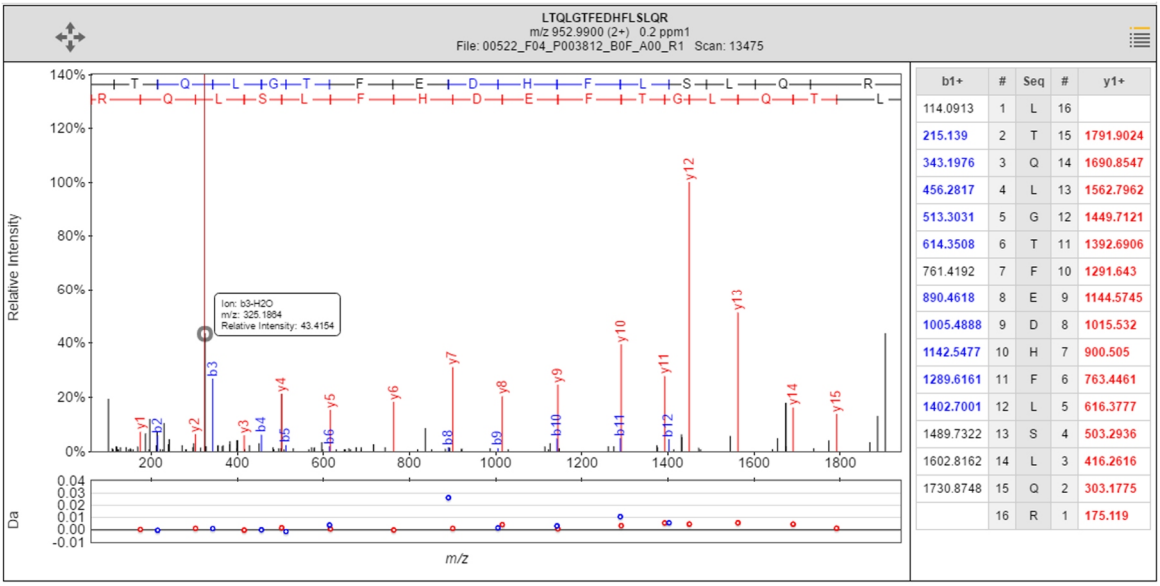


Figure 1.5 | Annotated MS/MS (MS2) spectrum of the peptide LTQLGTFEDHFLSLQR. Visualization of an MS/MS spectrum (left panel) acquired on an Orbitrap Q Exactive Plus of a doubly charged precursor mass of 952.99 including the mass deviation plot of the annotated fragment peaks (bottom panel). The table on the right lists the expected masses of all theoretical fragment ions. Numbers in bold (blue for b- and red for y-ions) indicate that this fragment is annotated in the MS/MS spectrum.

2.3.2 Tandem mass spectrometer

Triple-Quadrupole (QQQ or triple quad) mass analyzers utilize three consecutively placed quadrupoles[84]. The first quadrupole is in scanning mode and selects ions of interest, which are fragmented in the second quadrupole (collision cell). The fragment ions are analyzed in the third quadrupole. By deactivating the selection of ions in the first quadrupole and the collision cell, full MS1 spectra can be acquired.

Quadrupole-TOF (QTOF) and TripleTOF combine the stability of a quadrupole with the advantages of a TOF mass analyzer[85,86]. The quadrupole scans, selects and isolates the precursor ion, which are introduced into the collision cell. The resulting fragments are analyzed by the TOF reflectron mass analyzer.

The LTQ Orbitrap Velos[87] combines a quadrupole, ion trap and Orbitrap to allow rapid low resolution scanning performed in the low pressure cell (ion trap) or high resolution scans in the Orbitrap mass analyzer. Furthermore, both CID and HCD fragmentation can be performed by utilizing the high pressure cell or HCD collision cell (Figure 1.6) offering a wide range of acquisition schemes.

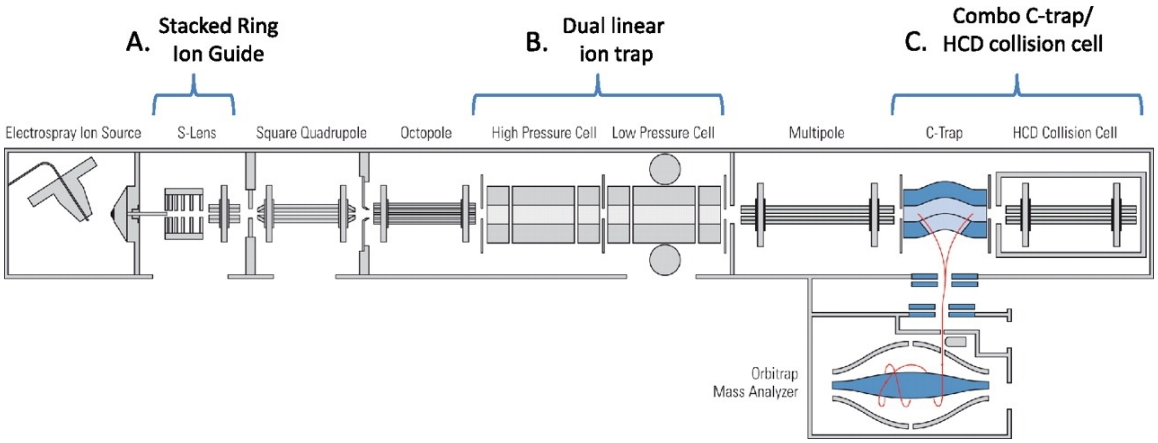


Figure 1.6 | Schematic of the LTQ Orbitrap Velos MS instrument. A, the stacked ring ion guide (S-Lens) increases the ion flux from the electrospray ion source into the instrument by a factor 5–10 in comparison to older machines. B, the dual linear ion trap design enables efficient trapping and activation in the high-pressure cell (left) and fast scanning and detection in the low pressure cell (right). C, the combo C-trap and HCD collision cell with an applied axial field with improved fragment ion extraction and trapping capabilities. Figure and caption from[87].

2.3.3 Alternative acquisition methods

Data-dependent acquisition (DDA) has become a standard method in mass spectrometry-based proteomics and is used in a wide range of applications. While this method is particularly designed to measure samples of unknown composition (discovery proteomics), the stochastic nature of selecting the top N most intense precursor ions for subsequent fragmentation within two MS scans hampers the acquisition of data which require very high reproducibility and accuracy[26,27,88,89]. As a matter of fact, even technical replicates acquired by measuring the same analyte twice do not typically result in the same identification and quantification results[90] as mostly low abundant features are randomly selected and furthermore not always generate interpretable spectra due to for example a low signal intensity.

To circumvent this, the acquisition can be "directed" (Extended Figure C1.1 in the Appendix) by entering an inclusion list[26]. This list contains precursor masses and their expected elution time which will be, despite their intensity, preferentially selected for subsequent fragmentation. If none of the specified precursor masses is present, the classical DDA approach is used to select precursor ions. While this method increases the reproducibility, the large dynamic range and the high complexity of the peptide mixture can still result in missed identifications.

The emerging class of data-independent acquisition (DIA) methods offers an alternative. Targeted peptide measurements[25,27] implemented in single reaction monitoring (SRM), multiple reaction monitoring (MRM) and parallel reaction monitoring (PRM)[91] allows the precise and reproducible quantification of analytes[88]. In contrast to the directed DDA approach, here only user defined transitions are recorded (targeted proteomics). A transition consists of a precursor mass and a fragment mass. Depending on the implementation, either only some (SRM/MRM) or all fragment ions (PRM) are recorded, sometimes with a survey or full scan (Figure 1.7). While this method offers precise quantification due to the increased signal to noise ratio, higher dynamic range and lower limit of detection and quantification, generally SRM And MRM experiments cannot be used to identify peptides[26] and rely on prior knowledge and experiments.

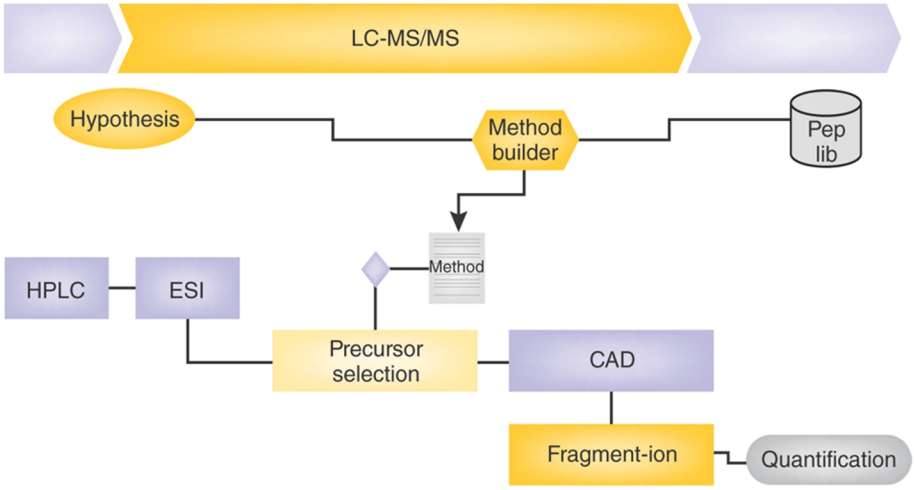


Figure 1.7 | Acquisition schema of a targeted bottom-up shotgun experiments. As the experiment is hypothesis-driven, it targets a very specific subset of peptides uniquely associated with the proteins of interest. An instrument method is built using existing proteomic resources (peptide spectral libraries) required for a target analysis and is typically performed using a triple-quadrupole (QQQ) instrument. For each peptide, a series of transitions (pairs of precursor and fragment ion m/z values) are monitored during a time that specifically corresponds with its predicted elution time. This enables the reproducible analysis of hundreds of peptides in a single experiment. Figure from[26].

The unbiased but still reproducible identification and quantification of peptides is promised by acquisition methods such as SWATH[92,93], AIF[94] or Waters HDMS[e], where all or a major slice of all precursor ions are fragmented simultaneously (multiplexed fragmentation). The increased complexity and size of the raw files renders manual interpretation and validation of results almost impossible. Even though new methods and algorithms were developed to analyze data from these multiplexed fragmentation methods[95,96], severe challenges remain. The large dynamic range of the analytes often results in the identification and quantification of only high abundant proteins. Furthermore, the increased complexity of the fragmentation spectra hinders the ability to identify PTMs and completely prevents multiplexing different samples at MS/MS level. However, the promise of acquiring a digital map of the proteome, which can be reanalyzed at any time, sounds very intriguing.

2.4 Quantification

Mass spectrometry-based proteomics has become the method of choice not only to identify but also to quantify peptides and proteins[23,41,97]. However, in all bottom-up proteomics experiments, the abundance of proteins cannot be measured directly, but instead has to be inferred from the quantification of their peptides[98]. Quantitative proteomics can generally be divided into two groups. First, label-free quantification which compares the mass spectrometric response of two or more conditions from separate acquisitions (Figure 1.8, right most column). Second, label-based quantification[99] which induces a mass shift that can be recognized by the mass spectrometer thus separating multiple conditions and permitting separate quantification and comparison within one acquisition (Figure 1.8, first two columns).

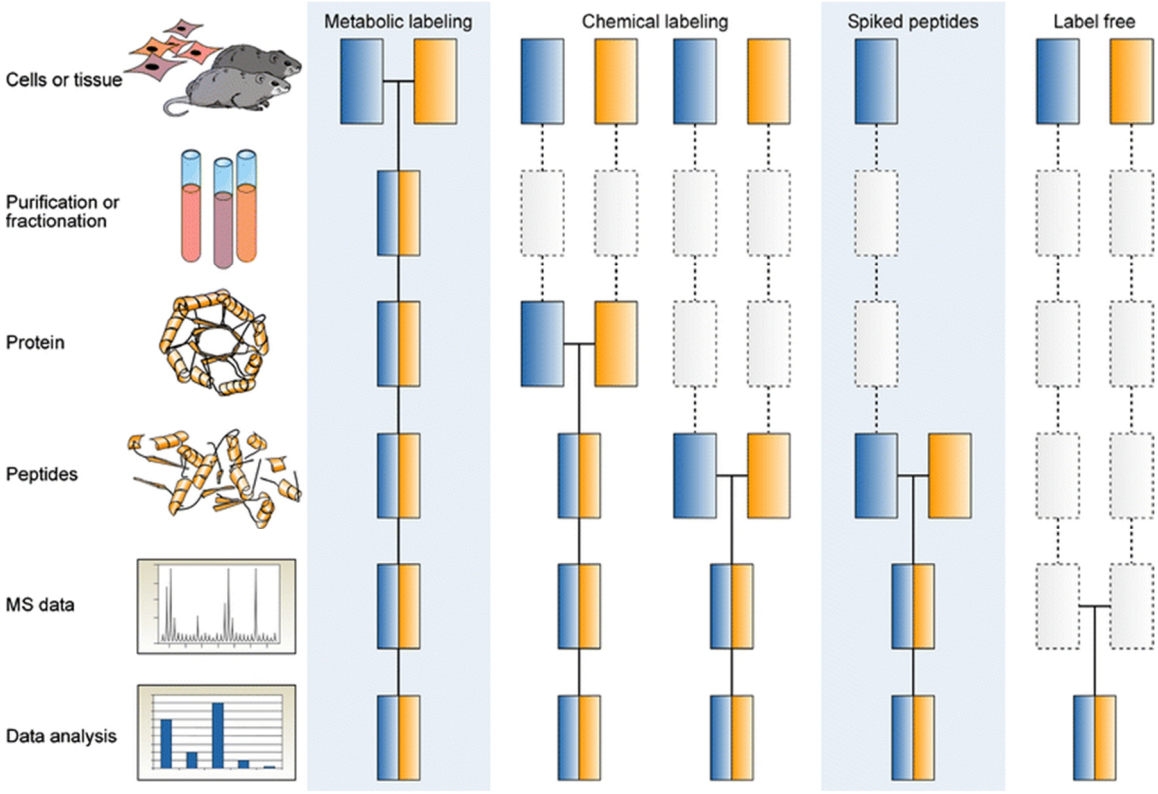


Figure 1.8 | Quantitative mass spectrometry workflows. Boxes in blue and yellow represent two conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur. Figure from[41].

To date, label-free quantification can be performed using spectral counting[100-103] or integration of MS signal intensities[104]. The first approach exploits the fact that in a classical DDA experiment (without or with limited dynamic exclusion) peaks for further fragmentation are selected based on their signal intensity. Due to the correlation of signal intensity to absolute concentration, highly abundant peptides are selected more frequently and thus those peptides (and proteins that match to these peptides) accumulate more MS/MS events. In contrast to spectral counting, intensity-based quantification uses the area of the extracted ion intensity chromatogram (XIC) of the precursor or intensities of specific reporter fragments in the MS/MS spectrum as a direct readout of the peptide abundance. This approach only requires one MS/MS event per precursor and thus more fragmentation events can be used to sample low abundant and previously undetected peptide species (dynamic exclusion). Using the signal intensities of ions thus offers many advantages over spectral counting methods. Both the precursor signal recorded at MS level as well as fragment intensities recorded at MS/MS level can be used for quantification, and thus allow its use independent of the labeling technique.

Label-free quantification allows the comparative analysis of an unlimited number of samples, but at the expense of data acquisition time and require careful parallel sample handling. Label-based quantification enables the simultaneous quantification of multiple samples. This is realized by the incorporation of stable isotopes into peptides either via change of the growth medium or the addition of isotope-coded chemical tags via amine reactive groups. This is based on the assumption that the physiochemical properties of the labeled and native version of a peptide are identical and thus behave identical during sample preparation and mass spectrometric analysis.

However, some studies show minor effects on the chromatographic behavior of peptides labeled with deuterium[105] which in turn requires the labeling of the native peptides with similar light counterparts. Nonetheless, stable isotope labeling has become a standard technique in quantitative proteomics and over the past two decades, multiple strategies were developed to measure the abundance of peptides over multiple conditions[23,106]. Most of them are used for relative quantification since absolute quantification strategies ideally involve the spike-in of stable isotope-labeled peptides as an internal standard (Figure 1.8 third column) to mimic the native peptides, like AQUA[107]. In practice, two major variants of label-based quantification exist, using either MS or MS/MS spectra for the quantitative readout.

2.4.1 MS-based quantification

Peptide quantification methods using the signal intensity of peptides at the precursor level are more stable and exhibit less noise as compared to spectral counting or MS/MS-based quantification due to higher sample statistics and higher signal-to-noise ratio. In order to multiplex samples at MS level, peptides are either metabolically or chemically labeled.

Stable isotope labeling by amino acids in cell culture (SILAC)[108] is the prime example of metabolic labeling (Figure 1.8, first column) and introduces isotope-labeled heavy or medium amino acids. The culture media contains isotope labeled amino acids which are incorporated into proteins during synthesis. An extension to this approach was published allowing higher multiplexing by exploiting the mass defect[109]. Acquiring MS1 spectra at high mass resolution reveals the isotopologue-embedded peptide signals and thus allows quantification.

Metabolic labeling is impractical for clinical samples or higher organisms, although in principle possible[110]. Comparatively cheap and easy alternatives are methods such as ICAT[111] and dimethyl[112] labeling. Here, a chemical modification carrying different isotopes is incorporated after or during protein digestion (Figure 1.8, second column).

However, the introduction of a second or third condition in one MS run using MS1-based quantification doubles and triples the number of features eluting at any time due to the mass shift of the differently labeled peptides. Given the limited number of MS/MS scans possible in order to maintain a reasonable duty cycle between MS1 scans to track the elution of a peptide species, the incorporation of stable isotopes typically results in less peptide and protein identifications since often both the light and heavy counterparts are selected for fragmentation.

2.4.2 MS/MS-based quantification

To circumvent the addition of additional MS1 features by labeling peptides species with different isotopes, MS/MS-based quantification offers the simultaneous quantification of up to 10 samples while maintaining the same number of MS1 features. Perhaps the most popular methods are isobaric tags for relative and absolute quantification (iTRAQ)[113] and tandem mass tags (TMT)[114]. Both target primary amines of the peptide and protein N-terminus and the ε-amino group of lysine using NHS (N-Hydroxysuccinimide) chemistry. Each sample is labeled at the peptide level with an isobaric group, resulting in the same precursor mass shift. However, the isobaric group consists of two components, the reporter group for quantification and a balancer group to generate the same precursor mass shift. Upon fragmentation the tag dissociates whereas only the reporter retains a charge and is thus visible in the lower mass region of the MS2 scan. The ratio between the reporter fragments can be used for absolute and relative quantification.

However, while MS/MS-based quantification offers precise and sensitive multiplexed quantification, isolation windows are typically not free of peptide-interference. This leads to ratio compression as the resulting reporter fragments are identical for all isolated peptides and thus show the sum of their intensities. While there are methods to circumvent[115] or repress[116] ratio compression, they typically come at the expense of peptide identifications due to a more complex data acquisition method leading to less MS/MS spectra and loss in coverage.

2.4.3 Sources of variance

Multiplexing samples offers the reduction of technical variances at the expense of higher sample complexity, thus leading to a lower identification rate. As depicted in Figure 1.8, in label-free experiments, both technical and biological variations are carried to the data analysis. Differences in peptide purification and fractionation, protein digestion and MS performance can impair subsequent analysis. Metabolically labeled samples can be pooled directly after sample collection due to the incorporation of heavy amino acids into newly synthesized proteins. However, this step requires separate cultivation and thus introduces biological variance. Chemical labeling allows the pooling of samples typically at the peptide level after digest, keeping technical variations at a minimum due to the possibility to perform subsequent sample handling steps on the combined pool.

Not only the choice of labeling, but also general sample preparation and acquisition methods affect the overall variance and have to be taken into account when designing an experiment. Each method offers specific advantages (e.g. good cross-experiment comparability of MS-based peptide intensities) and disadvantages (e.g. metabolic labeling not possible for patient derived samples) and choosing the most appropriate for an experiment is a challenging task because cost per acquisition hour, sensitivity and comparability have to balanced.

2.5 Mass spectrometric data

The raw data acquired on (most) mass spectrometers typically consist of a simple but ever increasing list of spectra. A spectrum again is list of tuples containing *m/z* and intensity information of detected ions. Annotated with additional information such as acquisition time, type of mass spectrum (MS vs. MS/MS) and acquisition parameters, these information can be aggregated into different views (e.g. XICs, TICs). The large number of spectra acquired on modern machines renders manual data interpretation impossible, thus requiring automated processing.

3 Computational proteomics

Mass spectrometry-based proteomics has developed into a high-throughput technology generating huge amounts of data per single study rendering manual interpretation of raw data impossible[43,56,117-119]. Automatic data processing tools and pipelines, such as MaxQuant[120] or OpenMS[121-123], are central and critical for the success of any proteomics experiment[124] and perform numerous computational steps to turn raw MS data into interpretable information (Figure 1.9)[23,31,33,41,125]. Depending on the analysis pipeline used[126], raw MS data, often only readable using proprietary libraries supplied with the mass spectrometer, have to be converted into open data formats in order to allow the interpretation, validation and quantification of peptides. The resulting list of peptides enables the identification and quantification of proteins. Although this process is fully automatic and aided by empirical, statistical and machine learning approaches[32], some of these steps require manual data inspection which especially for large studies remains a challenge[24,127].

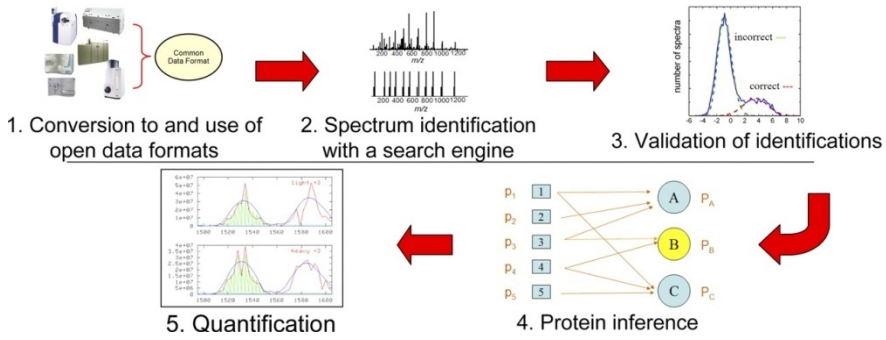


Figure 1.9 | Schematic overview of the typical analysis workflow of MS data. After data acquisition, multiple computationally intense steps are necessary to identify, validate and interpret the raw MS data. Starting with the conversion of raw data into open data formats, the processed spectra are then submitted to a search engine for identification. Subsequent validation and inference enables the quantification of peptides and proteins for successive statistical analysis. Figure from[124].

3.1 Data formats

Mass spectrometry data are acquired on a wide variety of mass analyzer technologies and brands, delivering datasets in various proprietary formats. In order to allow vendor-neutral and independent analysis of the raw data, multiple open-source data formats have been proposed and implemented. The Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification[128]. In 2008 the mzML-format[129] was released and provides an open platform XML-based data format for mass spectrometry data. However, with the increasing amount of data generated by modern mass spectrometers, the original XML-format had to be adjusted to stay on par with the high computational demand[130]. Alternatives providing higher read performance and smaller file sizes, such as an HDF5-based format termed mz5[131] and an approach using standard database principles[132], were developed but lack the traction to be fully employed and supported. A versatile tool which not only allows the conversion of raw MS data into the open standard is ProteoWizard[133,134] and serves as the reference implementation of the HUPO-PSI standards.

In addition to the raw MS data formats, multiple other data formats, such as mzIdentML[135] and mzQuantML[136] for identification and quantification data respectively, intended to replace older formats such as pepXML and protXML[137], were released by the HUPO-PSI in order to harmonize data comparison and exchange of processed data.

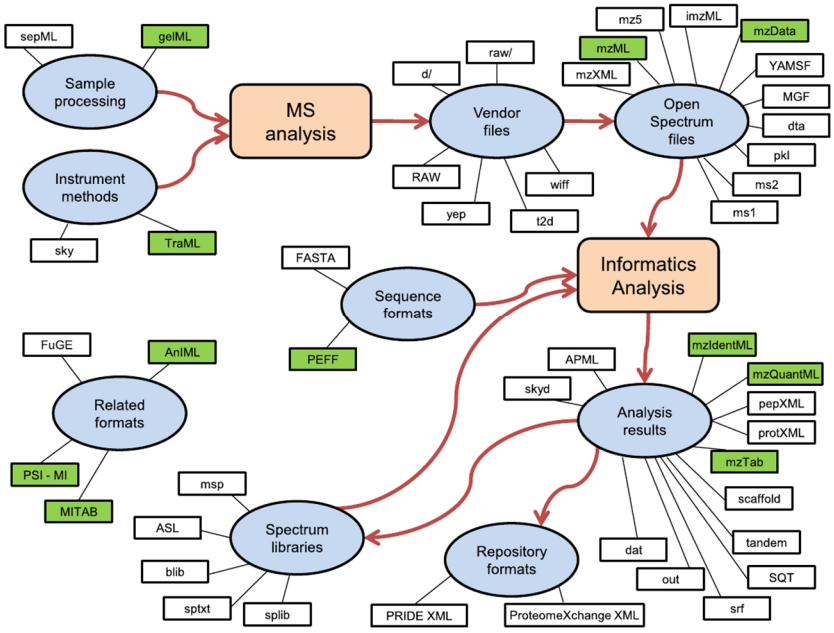


Figure 1.10 | Data format landscape in proteomics. The large number of file formats can be divided into two main groups consisting of mostly unprocessed (MS analysis) or processed (informatics analysis) data. MS files cover data acquisition methods as well as proprietary (vendor) or open data formats for storing the peaks lists acquired by the mass spectrometer. After processing and analysis, a wide variety of open data formats exists to store identification (e.g. mzIdentML) and quantification (e.g. mzQuantML) results of peptides and proteins. Figure from[138].

3.2 Raw data processing

Unprocessed mass spectra are subject to numerous impurities and contain many partially unwanted features[31,139,140]. One of the first steps to reduce the number of features and increase the signal-to-noise ratio is baseline correction and noise reduction. Both methods aim to increase the intensity of the analyte of interest while filtering or reducing the amount of electronic or chemical noise[126]. Commonly applied techniques are simple intensity filters, local maximum filtering, wavelet analysis and intensity normalization[126,141,142]. Additionally, the analyte of interest may be present in different charge states and due to the natural isotope distribution of its atomic constituents is split into the monoisotopic and isotopic peaks. Both effects can be reduced by de-charging and de-isotoping the spectra[31,139], but rely on the accurate determination of the charge state and the sensitivity and accuracy to detect isotopic peaks. Especially for MS/MS spectra, the collation of fragment peaks into one singly-charged monoisotopic peak increases the signal-to-noise ratio and enables a more accurate identification of peptides[143].

3.3 Peptide identification and validation

Central for the analysis of any proteomics dataset is the interpretation of MS/MS spectra, eventually generating a list of confidently observed peptides[125,144,145]. This process is composed of two main steps. First, for each MS/MS spectrum, an ordered list of peptide sequences, which are able to explain the acquired fragmentation spectrum, is generated. The ordering within this list reflects the "likelihood" (score) for this spectrum to be generated by the peptide. However, due to incomplete fragmentation and noise, this process is error prone and generates false matches. In the second step, statistical measures of confidence, such as p- and q-values, are assigned to the peptide identifications to enable subsequent filtering. Starting with the first step of this process, two main approaches exist:

De novo identification methods[146] try to identify the peptide sequence *ab initio* (Figure 1.11, bottom row). Here, typically graph-based algorithms find the peptide sequence whose fragment peaks can explain the peaks in the experimental spectrum best[147,148]. Empirical or probabilistic scoring schemes are used to assign a measure of confidence to the identification.

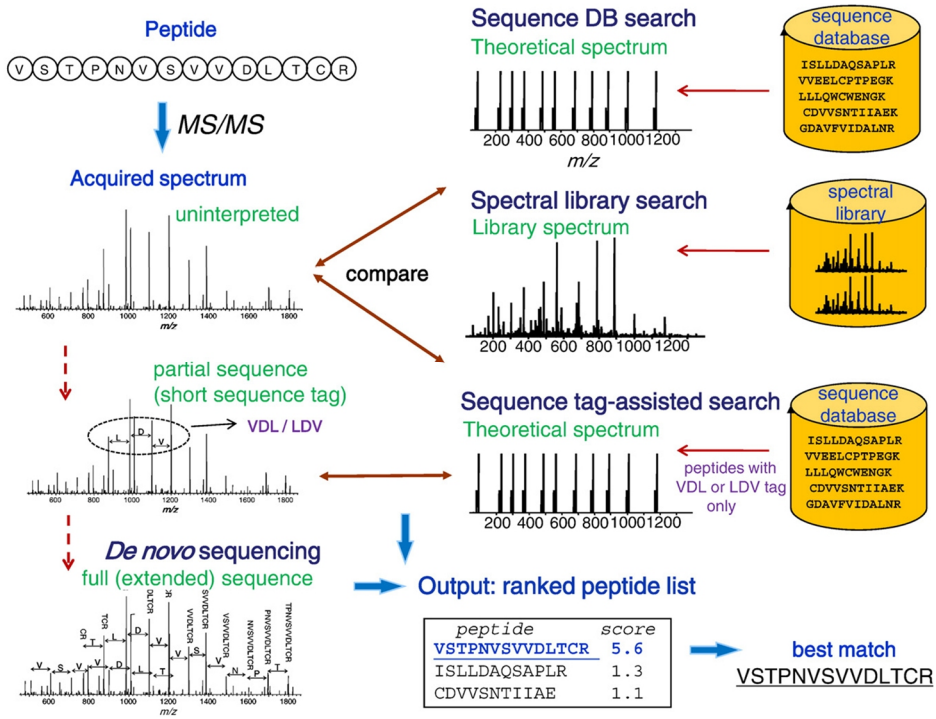


Figure 1.11 | Peptide identification strategies. Peptide identification can be performed by correlating the experimental MS/MS spectrum against a theoretical spectrum predicted for a peptide of interest (first row; sequence DB search), or against previously recorded spectra in a spectral library (second row; spectral library search). Alternatively, *de novo* methods can be used to directly extract sequence information from the MS/MS spectrum (fourth row; *de novo* sequencing). Hybrid approaches submit partial sequences from *de novo* identifications to the database search to further limit the number of peptide candidates for matching (third row; sequence tag-assisted search). Figure from[145].

The database search approach (Figure 1.11, top 2 rows) correlates the experimental spectrum against *in silico* generated spectra or spectra from a reference database. To this end, either a database of peptide sequences is used to generate (*in silico*) theoretical fragmentation spectra (Figure 1.11, top row), or previously recorded and annotated spectra stored in spectral libraries are used for comparison (Figure 1.11, second row).

While database searching is only applicable in cases where the peptide (and protein) sequence in question is known, *de novo* (Figure 1.11, bottom row) methods can be applied to almost all kinds of data, but are almost exclusively used when the peptide sequences in question are unknown[149]. Hybrid approaches, which interpret a high quality segment of the spectrum using *de novo* methods followed by a database search against peptide sequences which contain the partial sequence (Figure 1.11, third row) exist[150-152], but are less frequently used.

3.3.1 Database searching

Database searches require a peptide or protein sequence database to assign amino acid sequences to acquired spectra[144,145]. To this end, the search engine first generates an *in silico* digest of the expected proteins. The resulting list of peptides is filtered by the precursor mass of the experimental MS/MS spectrum where the allowed mass range depends on the resolution and accuracy of the mass analyzer. For each peptide candidate left, an *in silico* spectrum is generated by populating the theoretical spectrum with all possible fragment ions, taking into account the used fragmentation technique, and is then matched against the experimental spectrum. Different matching algorithms are used to score the experimental spectrum against the *in silico* generated one and range from simply counting the number of shared peaks[153], to (cross) correlations[154] and probabilistic models (binomial distributions)[155,156]. The result of this process is a list of peptide spectrum matches (PSMs). Generally, the peptide sequence whose theoretical spectrum matches the most features in the experimental spectrum is at the top of this list (rank 1 match). However, score based systems typically do not provide statistically meaningful significance measures such as a p-values or E-values[157]. However, different methods were developed to associate p- and E-values to PSMs[158-160]. Additionally, it was observed that features such as peptide length, post-translational modifications, precursor charge and mass tolerance can introduce a bias thus require special attention and calibration[161].
Widely used search engines are Mascot[155], SEQUEST[154], X!Tandem[162], OMSSA[163], Andromeda[156], Comet[164], Morpheus[153] and MyriMatch[165]. While each search engine has its strengths, combining results of multiple search engines is tricky and requires a unified statistical framework[159] but it has been shown to increase the number of identified spectra[166].

Database searching can also be performed using libraries of well annotated spectra which scored statistically significant in a previous run. In this case, the experimental spectra are compared against the reference spectra[167-169]. However, due to the immense search space when dealing with multiple PTMs, potentially missed cleavages sites, different collision energies and fragmentation techniques, it seems very unlikely that in discovery-type experiments spectral libraries become the preferred method for identification. However, data generated in DIA experiments, especially in SWATH acquisition methods, requires such prior knowledge.

3.3.2 False discovery rates

The process of assigning peptide sequences to spectra contains deficiencies resulting in either false positive (type I error) or false negative (type II error) identifications[145,170,171]. These errors can arise by using nonrestrictive search parameters, wrong settings with regard to the search space or acquisition method, or simply by chance due to noise. While false negative identifications do not hamper the downstream analysis, false positive identifications can have a detrimental and

misleading effect on the results of an MS-based proteomics experiment. Nonetheless, even under near perfect conditions false positive identifications will randomly occur given the large amount of MS/MS spectra acquired by a mass spectrometer.

A commonly used approach to control the number of type I errors is the false discovery rate (FDR). If the FDR can be calculated, the list of events can be filtered to contain at most a desired number (or percent) of false discoveries. This is often done by using q-values that describe at which FDR cutoff a particular event is present in the result list[171]. The FDR is thus a global measure of significance of a list of events, here PSMs. Similarly, local measures such as the posterior probability or the posterior error probability (PEP) give an estimate of the chances that an individual event is a false discovery[171].

However, *a priori* it is not known which of the events (here identification events such as PSMs) are true and false positive matches, thus calculating the PSM FDR is difficult and requires special methods[145]. Figure 1.12 shows an example of how to estimate the posterior probabilities for a list of PSMs. After performing a protein sequence database search of N MS/MS spectra and retaining only the rank 1 (highest/best) matches, a simple score histogram (bottom right panel) can be computed. If the matching score S is well calibrated, true positive matches should generally exhibit larger scores in comparison to false positive matches and thus a bimodal distribution is visible. Assuming that the low scoring part of the distribution (dashed line in bottom right panel) contains mostly false positive matches and the high scoring part (dotted line bottom right panel) consists of mostly positive matches, a mixture model can be fitted using for example an expectation maximization algorithm[172,173]. The fitted distributions can now be used to calculate both the posterior probability as well as the FDR for any arbitrary score S. The global FDR is calculated by dividing the number of (likely) false positive matches (area under the dashed curve) by the number of (likely) true matches (area under the dotted curve) with a score equal or higher than the selected score. The local FDR is calculated by dividing the absolute (likely) false matches by the (likely) true matches at the selected score.

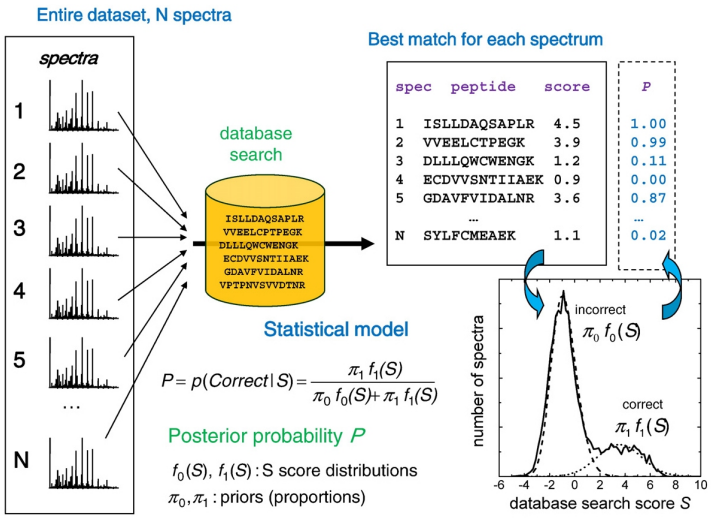


Figure 1.12 | Mixture model approach for computing posterior probabilities. All MS/MS spectra from an experiment are searched against a protein sequence database. The best database match for each spectrum is selected for further analysis. The most likely distributions among correct (dotted line) and incorrect (dashes) PSMs are fitted to the observed data (solid line). A posterior probability is computed for each peptide assignment in the dataset by dividing the number of likely false matches by the number of total matches. The parameters of the distributions, including the mixture proportion π1 are learned from the data using e.g. the EM algorithm. Figure from[145].

This method allows the estimation of the type I error, but requires well calibrated scores with well separated distributions of likely false and true matches. An alternative is the target decoy strategy (TDS)[174,175], a simple yet effective way to estimate the size, location and shape of the distribution of false positive matches. The general concept is to extend the search space by introducing decoy sequences which are by construction false positive matches. It builds on the assumption that spectra giving rise to false positive identifications have an equal chance of being matched into the target or decoy space. The decoy sequences are tagged and thus can be differentiated. When used correctly[176,177], the error prone process of fitting a distribution can be replaced by simply dividing the number of decoy and target matches equal (local) or larger (global) the score S (Figure 1.13)[145,171,178,179]. Once a desired FDR level is reached, this corresponding score can be used as a threshold.

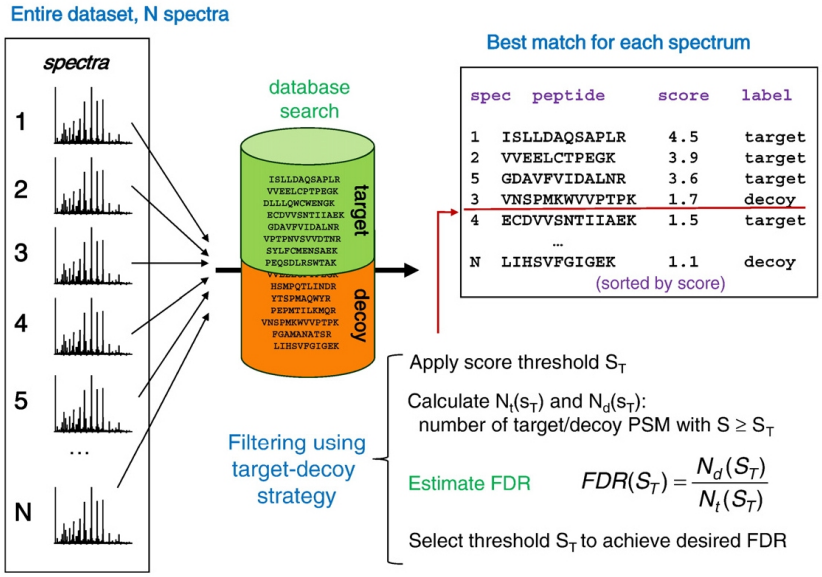


Figure 1.13 | Target decoy strategy for FDR assessment. All MS/MS spectra from an experiment are searched against a composite target plus decoy protein sequence database. The best peptide match for each spectrum is selected for further analysis. The number of matches to decoy peptides are counted and used to estimate the false discovery rate (FDR) resulting from filtering the data using various score thresholds. Figure from[145].

While multiple approaches exist to construct and search data against the decoy database, only minor differences in the result were observed[180,181]. Commonly, the target protein sequence database is reversed (with or without using the protease cleavage sites as fixed amino acids) and concatenated to the target protein sequence database. This ensures that a) the decoy database is of similar size (in terms of number of proteins and peptides); b) the amino acid composition of the decoy peptides is similar to that of the target peptides; and c) MS/MS spectra leading to false positive identifications have an equal chance of being matched against the concatenated target-decoy database.

The target decoy strategy became the standard to estimate global and local type I errors for both PSMs and peptides and is implemented in a wide variety of tools[145,182]. Furthermore, this concept can be extended and is also used in spectral library matching[183] and the analysis of targeted proteomics[184].

3.3.3 Identification of PTMs and unknown modifications

While MS-based proteomics has the capability of identifying thousands of transient and stable PTMs, commonly used scoring models and FDR estimation procedures are not designed to cope with such data. Allowing the presence of a variable PTM such as phosphorylation increases the search space drastically (combinatorial explosion of all cases). This results in the generation of theoretical modified peptide sequences which are sometimes only differentiable by a very small number of fragment peaks. Due to the drastic increase in search space, similar concepts to FDR are necessary to avoid false positive matches. Site localization probabilities and false localization rates (FLR) using the presence of site-determining ions and score differences to the next best PSM can be used to determine score cutoffs[185-190].

Notably, the identification of unknown modifications is also possible by using blind, unrestrictive or dependent searches[120,191-193]. Here, for instance, the precursor mass tolerance window is broadened to include the unmodified peptide sequence, even if a modified species was picked for fragmentation. Depending on the scoring scheme and position of the modification(s), the precursor mass difference between the measured and matched peptide allows to infer which and how many PTMs, unknown modifications or single amino acid polymorphisms are present. Similar to classical PTMs, site determining ions can be used to pin-point the modification within the peptide sequence.

3.4 Protein identification and quantification

The identification of proteins using the bottom-up strategy is (strictly speaking) not possible. Since proteins are digested into peptides prior to the injection into the mass spectrometer, only the identification of peptides is possible. The presence of proteins can only be inferred from a list of identified peptides (protein inference). However, this process is challenging and complicates the analysis and biological interpretation of the data especially in the case of higher eukaryote organisms. The same peptide sequence can be present in multiple different protein isoforms or genes. Such shared peptides therefore can lead to ambiguities in determining the presence and abundance of proteins[194].

3.4.1 Protein inference and grouping

Figure 1.14 illustrates 6 scenarios of the protein inference problem[194]. The simplest case is when proteins are distinct and do not share any peptides (Figure 1.14a). Here, any peptide evidence will lead to the unambiguous identification a single protein. In case some of the peptides are shared (Figure 1.14b), only the identifications of unique peptides (here peptide 1 or 4) can be used to identify the presence of either protein. No decisive conclusion can be drawn if peptide 2 or 3 are identified since the presence of either or both proteins can lead to the occurrence of these peptides. It is generally impossible to undoubtedly identify a protein if all peptides are shared with one (Figure 1.14 c and d) or many (Figure 1.14 e and f) other proteins. These groups of proteins are classified into indistinguishable (no single peptide can distinguish these proteins), subset (a protein contains only peptides which are shared with another differentiable protein) or subsumable (a protein contains only peptides which are shared with multiple but distinguishable proteins) proteins.

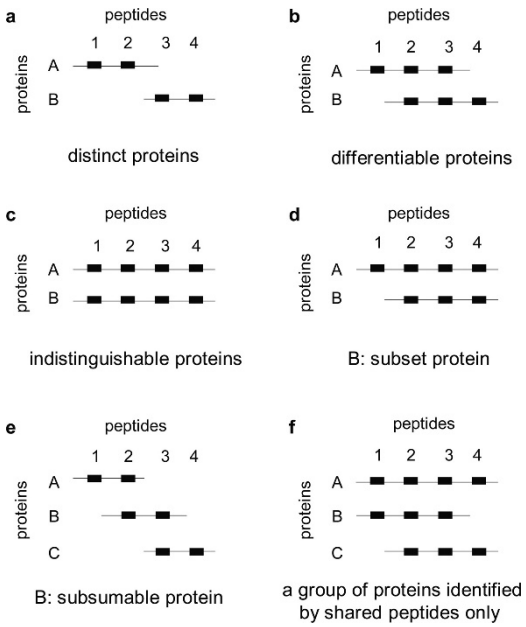


Figure 1.14 | Peptide grouping scenarios. a, distinct protein identifications. b, differentiable protein identifications. c, indistinguishable protein identifications. d, subset protein identification. e, subsumable protein identification. f, an example of a protein group where one protein can explain all observed peptides, but its identification is not conclusive. Figure from[194].

In practice, this often leads to reporting of protein groups instead of single proteins. A protein group consists of at least one protein which, given the peptide evidence, is not distinguishable from one other proteins (Figure 1.15). Here, peptides (rectangles) and proteins (circles) are connected with arrows indicating their possible origin. If an experiment resulted in the identification of the depicted 13 peptides, the distinct protein A and differentiable proteins B, C and E can be uniquely identified (inferred). However, protein D is a subset of protein E and thus no conclusive evidence about its presence in the sample can be drawn. While proteins F and G are indistinguishable and collapsed into a single entry, protein H, I and J are grouped, since each of them can be explained by a different set of peptides.

Common practice is to report the smallest list of protein groups that is sufficient to explain all observed peptides[195]. The process of grouping proteins and assigning a score or probability of its presence is, however, difficult and multiple approaches exist[194-196]. Similar to PSM and peptide FDRs, protein (group) FDRs can be calculated globally and locally using the same or similar methods. However, especially the task of computing protein level FDRs on data sets of increasing complexity and size is a challenge and the FDR estimation process is biased[197]. The separate step of controlling the protein level FDR after applying PSM and peptide FDR score cutoffs is necessary since the error rates are amplified when aggregating information from PSM to peptide to protein levels[145].

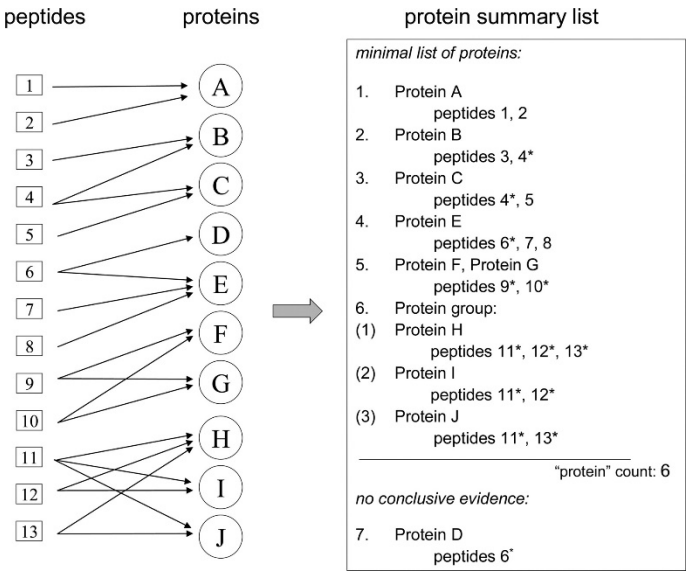


Figure 1.15 | Example of a protein summary list. Peptides are assigned to all their corresponding proteins, and the minimal list of proteins that can explain all observed peptides is derived. Proteins that are impossible to differentiate on the basis of the identified peptides are collapsed into a single entry (F and G) or presented as a group (H, I, and J). Shared peptides are marked with an asterisk. Proteins that cannot be conclusively identified are shown at the end of the list but do not contribute to the protein count. Figure from[194].

3.4.2 Protein abundance estimation

One of the main reasons why MS-based proteomics finds such a wide range of applications is its ability to identify but more important to quantify thousands of proteins. After FDR adjustment and protein inference, two different methods exist on comparing protein abundances across samples. Either peptide intensities are summarized on protein level or peptide-based models are used to compare the expression of proteins across multiple conditions[198]. Two commonly used protein level summarization methods are top3 (sum of the top 3 most intense peptide intensities of a protein group)[199] and iBAQ (length normalized sum of all peptide intensities of a protein group)[200]. These intensity-based approaches allow a straightforward and precise quantification[201] and in contrast to spectral counting, additional retention time alignment enables the reproducible quantification of peptides across samples by matching unidentified features between multiple LC-MS to identified peptide species[120]. Recently, an alternative approach was published allowing the calculation of the number of molecules per cell for any given protein by assuming that the number of histones in a (human) cell is constant[202]. In contrast to DDA experiments, the quantification of proteins acquired in targeted and DIA experiments require specialized software which utilize the fragment intensities[184,203].

3.5 Statistical analysis and data interpretation

Most proteomic experiments aim to find differences between two or more conditions. The goal is to identify proteins (or peptides in case of PTM studies) which show a significant up- or down-regulation within two or more classes, such as treated vs untreated or normal vs disease. These proteins can act as biomarkers for subsequent sample classification or targets for potential therapies. However, due to the diversity of experimental designs and analysis steps, most of the

tasks necessary cannot be performed in a fully automated fashion[33,98] and describing all possibly methods and tools will go beyond the scope of the thesis. Instead, this section focuses on the most basic tasks needed to analyze and interpret the results of a proteomic experiment (Figure 1.16).

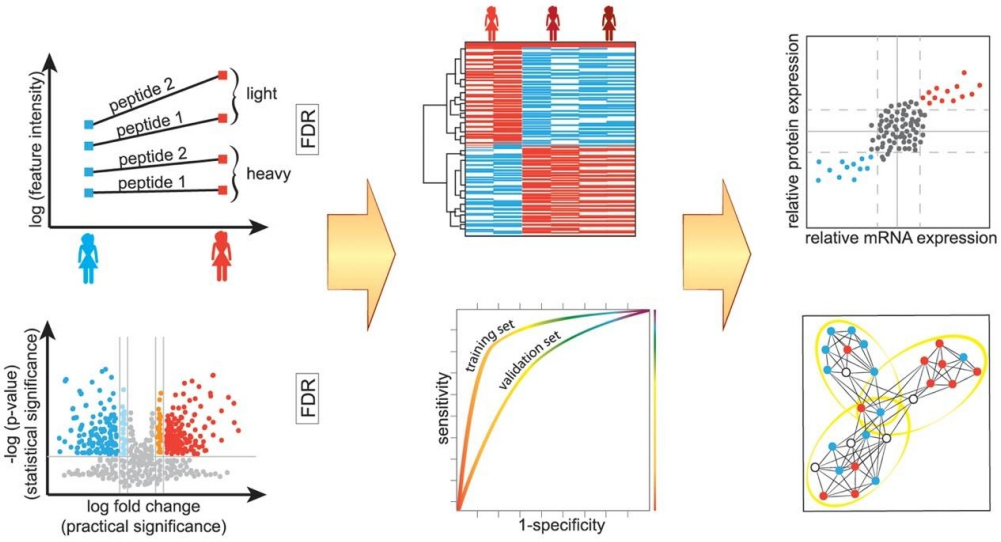


Figure 1.16 | Computation and statistics. After the identification and quantification of peptides and proteins, typical steps for data analysis include: peptide and protein significance analysis (left panel), class discovery and prediction (middle panel), and data integration and for instance pathways and signaling analysis (right panel). Especially the last step is affected by uncertainty in protein identities and the incomplete sampling of the proteome. Figure from[33].

Protein intensities are subject to technical (e.g. different sample loading, MS performance) and biological (e.g. cell culture, amount of sample) variations. To account for that, different normalization and batch removal techniques are available, ranging from simple median centering or quantile and total sum normalizations to more sophisticated probabilistic and regression models[33,204-207]. While some methods are designed for intensities only, some others are also applicable to normalize ratios (e.g. median centering). A general assumption to enable normalization is that most of the peptides and proteins do not exhibit a significant change, thus on average no difference between two samples is expected. While this holds true for most experiments, the applicability and choice of normalization method is experiment and experiment design dependent.

After normalization, a typical next step is the identification of differentially regulated peptides and proteins. The ability to identify differential features is dependent on the reproducible identification and quantification of peptides and protein across samples. However, the lack of information does not imply the absence of the peptide or protein in a sample. Especially low abundant features often lead to poor fragmentation spectra due to ion counting statistics. This is further aggravated by the incomplete sampling of the proteome, resulting from the semi-stochastic nature of the DDA approach in combination with the complexity of the analyte (even in cases where purification steps were performed). While processing pipelines try to circumvent this as much as possible by matching unidentified features across samples and missing value imputation can reduce this issues even further, missing values may introduce additional biases as it is typically unclear which distribution of intensity values represents the features best. This is

because the feature is actually absent, or at least below the limit of detection, or the feature is present but not identified or matched. As for most post-processing steps, the applicability of missing value imputation strongly depends on the experimental design, the underlying hypothesis and acquisition method[208,209].

Even without these issues, the identification of significantly changing peptides and proteins is a challenging task and relies on a carefully designed experiment. After log-transformation to ensure a Gaussian distribution of intensities and ratios, common statistical methods such as t-tests, ANOVAs and f-tests can be applied to assign p-values to proteins. Each statistical test is (largely) independent, thus multiple testing correction is essential to control the type I error[210,211]. After assigning q-values to peptides and proteins, differentially changing features are selected by filtering for both practical (fold change) and statistical (p- and q-values) significance (Figure 1.16 left panel).

After identifying proteins which show a significant regulation across samples, the discovery of classes using supervised or unsupervised methods such as clustering or machine learning refers to the process of identifying features which show a similar trend and reproducibly divide samples into classes (e.g. normal and disease; Figure 1.16 middle column)[32,35,212]. This process aims to identify peptides and proteins which can be used for subsequent class prediction[213].

Furthermore, sets of proteins which show a similar quantitative profile across biological samples can also be used to generate new functional and biological insight. For example, proteins, which show a similar trend upon drug treatment, are likely part of a larger functional network and pathway. Molecular, functional and biological enrichment analysis of significantly differentially regulated proteins[214-216] enables the determination of common features, which can be used to annotate groups of proteins. Briefly, the fraction of proteins assigned to a specific molecular, functional or biological process are compared to the expected fraction. If, e.g. kinases, show a significant enrichment within the differentially regulated proteins in comparison to the entire proteome, kinases might be a key factor for the response.

Integrative analysis with other 'omics'-data[217,218], such as transcriptomics, genomics and metabolomics, and correlation to previously conducted studies[219] can provide additional insights into the underlying biology (Figure 1.16 right panel) and broaden our understanding of the molecular processes on a system-wide level. Even though most underlying mechanisms are still not fully understood, each omic technology in itself offers specific advantages and disadvantages, enabling scientist to retrieve confident information about e.g. mutation status, gene expression and activation status of proteins. However, these steps require thoroughly and extensively described resources and knowledge bases, which are in turn dependent on well-performed, freely available and detailed data.

4 Proteomic and annotation resources

Our ability to analyze and interpret the results of an MS-based proteomic experiment is heavily dependent on prior knowledge. During the analysis, protein sequence databases are used to transform raw MS data into peptide and protein result lists. Subsequently, known functions and annotations of proteins[220,221], their interaction or contribution to complexes and metabolic or signaling pathways[222-227] allow us to draw functional and biological conclusions.

The integration of different experiments enables the continuous increase in knowledge on how biological systems work and act upon different stimuli. Thus, there is great potential in storing and providing access to as many well annotated experiments as possible[219]. For this, both raw and result files, containing identification and quantification data, have to be archived and made available to also non expert researchers. This will also help other disciplines to build and validate their own findings and further advance their hypotheses. Furthermore, the integration of multiple types of data could lead to novel findings, which could not have been uncovered by a single lab or experiment alone.

It has become good practice to share experimental data to support novel findings[38,219]. However, due to the ever-increasing amount of data generated, organizing and storing raw data has become a challenge, especially in the field of proteomics. How to best store these was a long discussion in the scientific community[228] and gave rise to many data repositories and databases[38]. While there are many challenges associated with the storage of data, especially in terms of annotating experimental factors and conditions used to study biological systems, even the integration of comparatively simple studies can broaden our knowledge. For example, the integration of multiple isolated studies measuring the expression of proteins (full proteomes) in model systems (e.g. cell lines) can help researchers to design better experiments by providing an expression map of proteins.

This section aims to introduce some of the aspects discussed here by describing state of the art resources available to analyze and interpret MS-based proteomics data. Last but not least, a brief overview over databases and repositories used in the field of proteomics to share both raw and result files is given.

4.1 Sequence database

Annotated sequence databases are essential in pre-processing mass spectrometry-based proteomics experiments as both identification and quantification of proteins heavily depend on them. A large variety of protein sequence databases exists spanning the entire range of simple sequence databases to manually curated and enriched repositories/knowledge bases covering all species[220,229]. The number of identifiers (IDs) mapping proteins to transcripts, genes and other resources is steadily growing. Further aggravated, discontinued sequence databases such as IPI[230] are not fully integrated into existing databases resulting in a “loss” of up to 20% IDs[231]. This culminated in large number of identifiers requiring specialized tools and services to map them to other databases and external resources[232].

RefSeq, the Reference Sequence database, maintained by the National Center for Biotechnology Information (NCBI), is a collection of integrated, well annotated and non-redundant set of

sequences, including genomic DNA, transcripts and proteins. Sequence entries are generated from selected assembled genomes available in GenBank[233].

The Ensembl project[234] is a collaboration between the European Bioinformatics Institute (EMBI-EBI) and the Wellcome Trust Sanger Institute (WTSI) that provides protein, transcript and gene identifiers from an automatic annotation of genomes, integrated with other available biological data. Ensembl provides the commonly used human reference assembly GRCh37, now in version GRCh38.

UniProt is the most commonly used protein sequence database. It is maintained by the Universal protein resource (UniProt)[220], a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). UniProtKB is a comprehensive resource of protein sequences enriched by additional annotations and consists of multiple databases (Figure 1.17), namely the UniProt Knowledgebase (UniProtKB), the UniProt Reference Cluster (UniRef) and the Uniprot Archive (UniParc). UniProtKB is composed of two sections, UniProtKB/SwissProt and UniProtKB/TrEMBL. SwissProt is a manually annotated high quality and non-redundant protein sequence database which brings experimental results, computed features and scientific conclusions together. TrEMBL is the result of increased dataflow from the genome projects and contains high quality computationally analyzed entries.

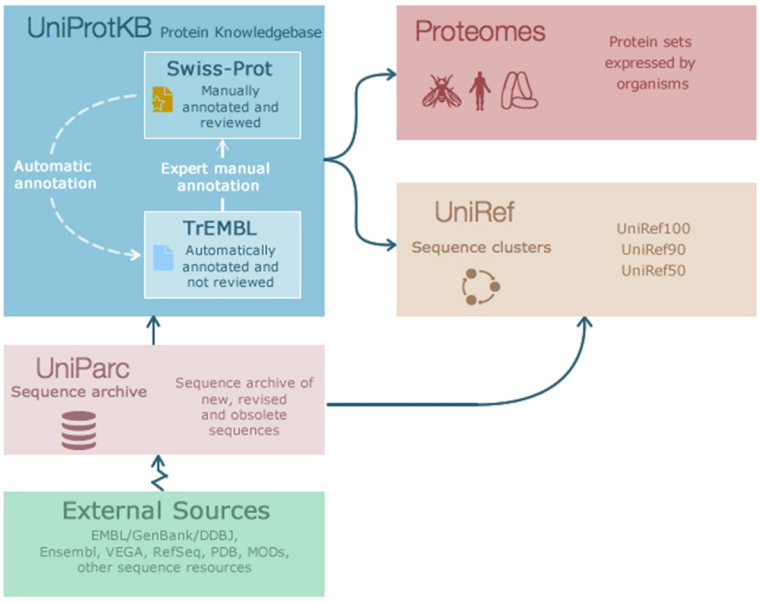


Figure 1.17 | Schematic overview of the Universal protein resource (UniProt). Uniprot consists of three databases, the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). UniProtKB, uses multiple external sources, such as Ensembl and RefSeq, to build the automatically annotated protein sequence database UniProt TrEMBL and the manually annotated and reviewed protein sequence database UniProt Swiss-Prot. Figure from http://www.uniprot.org/help/about.

4.2 Annotation resources

In order to interpret lists of proteins identified in any MS-based proteomics study and assign functional and biological meaning, researchers rely on resources providing annotations of proteins on as many levels as possible[34]. Large efforts exist to integrate and combine as many annotation resources as possible.

GO[221], the Gene Ontology, is a major bioinformatics initiative to develop a representation of how genes encode biological functions at the molecular and cellular level. For this purpose, GO provides a controlled vocabulary of terms to describe gene characteristics and annotations by automatically and manually annotating gene functions based on experiments reported in peer-reviewed scientific papers. It offers these terms in three categories: i) biological process, ii) molecular function and iii) cellular component. While providing a controlled annotation basis for genes, GO terms can be used for enrichment analysis on gene sets[214-216].

KEGG[222], Kyoto Encyclopedia of Genes and Genomes, mostly known for its pathway and molecular network database, is an integrated database resource for biological interpretation of genome sequences and other high-throughput data. For this purpose, the KEGG Orthology (KO) database stores associations between molecular functions of proteins with ortholog groups. Networks of KO nodes represent high-level functions of cells and organisms and can be used for enrichment analysis or simply to check and validate the functions of genes and their products.

Reactome[223,224] is a manually curated resource of human pathways and reactions. It describes these as chemical reactions closely mirroring the actual physical interactions in cells. Accessible via a web interface, 6744 reactions from over 7000 human proteins can be viewed and analyzed. Its data model allows the annotation of cancer and other disease processes to accommodate changes in the amino acid sequence or the formation of fusion proteins. Similar to KEGG, the annotations provided by Reactome can be used as additional information in gene set enrichment analysis to provide supplementary information about de-regulated proteins.

IntAct[225] is an open molecular interaction database developed by the European Bioinformatics Institute (EBI). Its database is populated by either curated data from the scientific literature or direct submissions of interaction data. Recently, the content of the molecular interaction database (MINT)[235] was fully integrated into IntAct to maximize the curation output.

BioGrid[236] is a public database, similar to IntAct, that archives and disseminates genetic and protein interaction data. It holds over 830,000 interactions derived from high-throughput datasets and literature mining. Focusing on areas of biology which help to build insights into networks and pathways relevant to human health, BioGrid holds data for several model organisms and provides links to other resources.

STRING[226] is a database of known and predicted protein interactions. It aims to provide a critical assessment and integration of protein-protein interactions. It contains both direct as well as indirect associations between proteins and covers more than 2000 organisms and 9 million proteins. An interactive viewer allows users to enter a list of genes which is then used to build a network based on the interactions stored in STRING.

PhosphoSitePlus[227] is an open and curated resource for studying experimentally observed PTMs. Besides its comprehensive coverage of protein phosphorylation, this resource also contains information about acetylation, methylation, ubiquitination and O-glycosylation sites. For this purpose, it holds structural and functional information about the topology, biological function and regulatory significance of specific modification sites to allow users to mine and interpret their data with respect to the biological regulation.

HPA[237,238], the Human Protein Atlas, contains gene expression and localization information of the corresponding proteins acquired from both RNA and protein data. For this purpose, high-resolution images of 44 different normal human tissues, 20 different cancer types and 46 different human cell lines were stained using antibodies to show the spatial distribution down to substructures and cell types of tissues. In addition, the transcriptomics data provides quantitative

data on gene expression levels. The recent release contains transcriptomic and proteomics evidence for 99.9% and 86% of the predicted human genes, respectively.
GeneCards[239] is a comprehensive compendium of annotations of human genes developed for biomedical researchers. The content is automatically mined and integrated from over 80 digital sources. Developed for the past 15 years, it is a common entry point for researchers to access the wealth of information stored in its database and provides gene expression data on multiple levels.
NeXtProt[240] is a protein-centric knowledgebase for human proteins. It aims to provide a constantly updated view on human biology capturing a wide range of data and annotations. In order to do this, NeXtProt includes multiple other annotation resources such as IntAct, PhosphoSitePlus and GO via services provided by UniProt and provides cross references to many additional resources, publications and even experimental data.

4.3 Proteomics databases and repositories

Due to the speed and advancements in MS-based proteomics, the amount of raw data but also processed result lists require a large amount of storage space. In order to make this amount of data available to the scientific and public domain, it is becoming a common practice to store data in public repositories[37,38,228].
Over the past years, many different proteomic repositories and compendia were developed, such as MaxQB[241], Human Proteinpedia[242], PaxDB[243] and Tranche[244]. The latter was a distributed repository for redundant storage and dissemination of datasets and offered scientists many features, such as prepublication access control and licensing options. For storage, it utilized an encrypted peer-to-peer system that splits incoming data across multiple servers, making it hard to be controlled. Unfortunately, mostly due to the lack of funding but also because of the distributed design and de-centralized organization, Tranche was discontinued after a couple of years. However, its disappearance triggered the proteomic community to stabilize and advance current solutions.

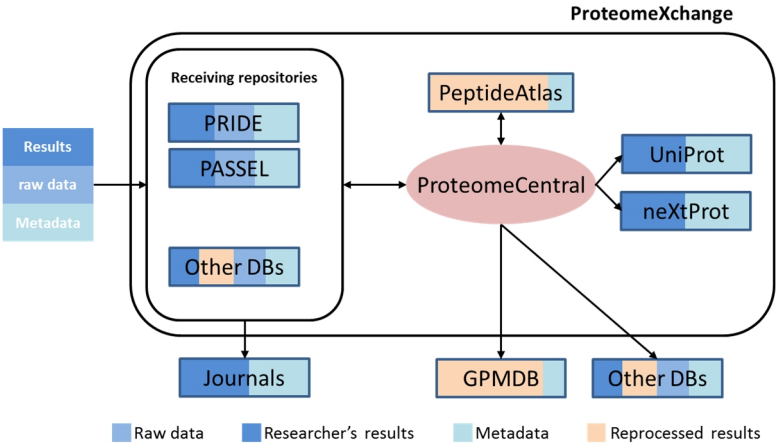


Figure 1.18 | Schematic overview of the ProteomeXchange consortium. The ProteomeXchange consortium coordinates the submission of MS proteomics data to the main existing proteomic repositories and is the central contact point for Journals and annotation resources such as UniProt and NeXtProt for experimental data. For this purpose, ProteomeCentral assigns unique identifiers to submitted datasets and provides a queryable interface for other resources such as GPMDB and PeptideAtlas. Figure modified from https://www.ebi.ac.uk/training/online/course/proteomexchange-submissions-pride/data-resources-proteomexchange

ProteomeXchange[36] fills the gap of de-centralized organization. It is a consortium with the goal to coordinate the submission of MS-based proteomics data (Figure 1.18). It acts as a central hub for multiple proteomics repositories to encourage optimal data dissemination. Resources within ProteomeXchange are classified into two kinds: i) archival resources which store processed data as published by the authors and ii) secondary data resources which store primary data. Each dataset submitted to ProteomeXchange or a partnering resource is identified by a unique identifier generated by ProteomeCentral. In addition, this resource allows other databases to query datasets of interest.

PeptideAtlas[245,246] is developed by the Seattle Proteome Center and part of the ProteomeXchange consortium. As a prime example of a secondary data resource, it allows data submission via its own interface, but mostly retrieves data from ProteomeXchange. The main goal is the full annotation of eukaryotic genomes though a thorough validation of expressed proteins. For this purpose, it summarizes peptide identifications in various ways. PeptideAtlas reprocesses all incoming data using its own pipeline. First, MS/MS spectra are searched using SEQUEST[154,247] and X!Tandem[162] or SpectraST[167]. The initial peptide identifications are rescored and filtered using PeptideProphet[248] and the results are submitted to ProteinProphet[195] for protein identification. In addition, MAYU is used to control the protein FDR[197]. The results are stored and made available in regular precompiled builds. PASSEL[249], the PeptideAtlas SRM Experiment Library, is a component of PeptideAtlas and is designed for the reuse of SRM experimental results.

PRIDE[250,251], short for proteomics identifications, was established as a public data repository by the EBI to support the publication of MS studies. It stores peptide and protein identifications, as well as associated metadata, such as the experimental design of the study. In contrast to PeptideAtlas, PRIDE does not reprocess submitted data "to represent the submitter's view of the data". As the prime archival resource of ProteomeXchange, it became the recommended submission point for several journals.

Chorus (https://chorusproject.org) is a cloud-based repository that provides researchers means to securely store, analyze and share their MS data. Recently released, it aims to create a complete catalogue of the world's MS data and to make it openly and freely accessible to both the scientific and the public domain. Chorus is developed for and on Amazon Web Services such as the Amazon Elastic Compute Cloud, Amazon Simple Storage Service and Amazon Glacier. Its backend uses MapReduce[252] to distribute custom data analysis tools over multiple virtual machines to allow parallel and distributed computing. Currently, it supports viewing chromatographic and spectral data as well as protein sequence database searches. However, due to the use of the Amazon services and the lack of sustained funding, users are charged to store and analyze larger amounts of data.

GPMDB, the Global Proteome Machine database (http://gpmdb.thegpm.org/), and the underlying database GPM[253] servers, was constructed to aid the process of validating MS/MS spectra and protein sequence coverage patterns. It allows users to compare their experimental results with results published previously. It supports different organisms and, similar to PeptideAtlas, is connected to ProteomeCentral. An automatic processing pipeline, using X!Tandem[162] as the main search engine, analyzes recently published data and integrates the results into GPMDB.

5 Objectives

The central task of this thesis was to implement a central database, which can be used by scientists to validate and build new hypotheses, aid researchers in experiment design and to validate new computational tools tackling open and arising issues in computational proteomics. For this purpose, a high performant and simple to use database was proposed, which enables access to large amounts of proteomic data (Chapter 2). One of its first applications was the assembly of a first draft of the human proteome from roughly 16,000 MS raw files on both protein identification and quantification level (Chapter 3). The assembled data also revealed major shortcomings in the standard approach to estimate protein FDR. For this purpose, a novel protein FDR estimation method termed 'picked' protein FDR was developed (Chapter 4).

6 Abbreviations

CETSA	cellular thermal shift assays
CID	collision-induced dissociation
CRM	charge residue model
DDA	data-dependent acquisition
DIA	data independent acquisition
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ESI	electrospray Ionization
ETD	electron-transfer dissociation
FDR	false discovery rate
FLR	false localization rate
FT ICR	fourier transform ion cyclotron resonance (mass spectrometer)
FT	fourier transform
HCD	higher-energy collisional dissociation
HILIC	hydrophilic interaction chromatography
HUPO	Human Proteome Organization
IAM	ion evaporation model
IEF	isoelectric focusing
iTRAQ	isobaric tags for relative and absolute quantification
KO	KEGG orthology
LC	liquid chromatography
LC-MS	liquid chromatography couple to mass spectrometer
LC-MS/MS	liquid chromatography tandem mass spectrometry
LTQ	linear trap quadrupole (mass spectrometer)
m/z	mass (m) to charge ratio (z)
MALDI	matrix-assisted laser desorption/ionization
MRM	multiple reaction monitoring
MS	mass spectrometer and mass spectrum
MS/MS	tandem mass spectrometry and tandem mass spectrum
NanoESI	nano flow electrospray Ionization
PEP	posterior error probability
PIR	protein information resource
PRM	parallel reaction monitoring
PSI	Proteomics Standards Initiative
PSM	peptide (to) spectrum match
PTM	post-translational modification
Q	quadrupole
QQQ	triple-quadrupole (mass spectrometer)
QTOF	quadrupole time of flight (mass spectrometer)
RNASeq	RNA-sequencing
SAX	strong anion exchange chromatography
SCX	strong cation exchange chromatography
SILAC	stable isotope labeling by amino acids in cell culture
SRM	single reaction monitoring
TDS	target decoy strategy
TMT	tandem mass tags
TOF	time-of-flight
UniParc	Uniprot archive
UniProt	universal protein resource
UniProtKB	UniProt knowledgebase
XIC	extracted ion chromatogram
XML	extended markup language

7 References

1 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945, doi:10.1038/nature03001 (2004).

2 van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* 30, 418-426, doi:10.1016/j.tig.2014.07.001 (2014).

3 Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* 470, 187-197, doi:10.1038/nature09792 (2011).

4 Cox, D. B., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nature medicine* 21, 121-131, doi:10.1038/nm.3793 (2015).

5 Mardis, E. R. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics* 9, 387-402, doi:10.1146/annurev.genom.9.081307.164359 (2008).

6 Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660-665, doi:10.1126/science.aaa0355 (2015).

7 Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nature biotechnology* 33, 306-312, doi:10.1038/nbt.3080 (2015).

8 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10, 57-63, doi:10.1038/nrg2484 (2009).

9 O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *The Journal of biological chemistry* 250, 4007-4021 (1975).

10 Klose, J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 26, 231-243 (1975).

11 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 & cellular proteomics : MCP* 13, 3698-3708, doi:10.1074/mcp.M114.043489 (2014).

12 Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering* 11, 49-79, doi:10.1146/annurev-bioeng-061008-124934 (2009).

13 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* 7, 548, doi:10.1038/msb.2011.81 (2011).

14 Beck, M. *et al.* The quantitative proteome of a human cell line. *Molecular systems biology* 7, 549, doi:10.1038/msb.2011.82 (2011).

15 Mitchell, P. Proteomics retrenches. *Nature biotechnology* 28, 665-670, doi:10.1038/nbt0710-665 (2010).

16 Hand, E. Planetary Science. Mars rover finds long-chain organic compounds. *Science* 347, 1402-1403, doi:10.1126/science.347.6229.1402 (2015).

17 Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nature biotechnology* 28, 695-709, doi:10.1038/nbt.1658 (2010).

18 Glish, G. L. & Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nature reviews. Drug discovery* 2, 140-150, doi:10.1038/nrd1011 (2003).

19 Finehout, E. J. & Lee, K. H. An introduction to mass spectrometry applications in biological research. *Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology* 32, 93-100, doi:10.1002/bmb.2004.494032020331 (2004).

20 Guengerich, F. P. Thematic minireview series on biological applications of mass spectrometry. *The Journal of biological chemistry* 286, 25417, doi:10.1074/jbc.R111.266700 (2011).

21 Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature methods* 7, 681-685, doi:10.1038/nmeth0910-681 (2010).

22 Schirle, M., Bantscheff, M. & Kuster, B. Mass spectrometry-based proteomics in preclinical drug discovery. *Chemistry & biology* 19, 72-84, doi:10.1016/j.chembiol.2012.01.002 (2012).

23 Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry* 404, 939-965, doi:10.1007/s00216-012-6203-4 (2012).

24 Oveland, E. *et al.* Viewing the proteome: how to visualize proteomics data? *Proteomics* 15, 1341-1355, doi:10.1002/pmic.201400412 (2015).

25 Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods* 9, 555-566, doi:10.1038/nmeth.2015 (2012).

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

27 Domon, B. & Gallien, S. Recent advances in targeted proteomics for clinical applications. *Proteomics. Clinical applications* 9, 423-431, doi:10.1002/prca.201400136 (2015).

28 Harsha, H. C., Pinto, S. M. & Pandey, A. Proteomic strategies to characterize signaling pathways. *Methods in molecular biology* 1007, 359-377, doi:10.1007/978-1-62703-392-3_16 (2013).

29 Prins, P. *et al.* Toward effective software solutions for big biology. *Nature biotechnology* 33, 686-687, doi:10.1038/nbt.3240 (2015).

30 Hogeweg, P. The roots of bioinformatics in theoretical biology. *PLoS computational biology* 7, e1002021, doi:10.1371/journal.pcbi.1002021 (2011).

31 Colinge, J. & Bennett, K. L. Introduction to computational proteomics. *PLoS computational biology* 3, e114, doi:10.1371/journal.pcbi.0030114 (2007).

32 Kelchtermans, P. *et al.* Machine learning applications in proteomics research: how the past can boost the future. *Proteomics* 14, 353-366, doi:10.1002/pmic.201300289 (2014).

33 Kall, L. & Vitek, O. Computational mass spectrometry-based proteomics. *PLoS computational biology* 7, e1002277, doi:10.1371/journal.pcbi.1002277 (2011).

34 Villavicencio-Diaz, T. N., Rodriguez-Ulloa, A., Guirola-Cruz, O. & Perez-Riverol, Y. Bioinformatics tools for the functional interpretation of quantitative proteomics results. *Current topics in medicinal chemistry* 14, 435-449 (2014).

35 Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. & Bacardit, J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics : a journal of integrative biology* 17, 595-610, doi:10.1089/omi.2013.0017 (2013).

36 Vizcaino, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* 32, 223-226, doi:10.1038/nbt.2839 (2014).

37 Riffle, M. & Eng, J. K. Proteomics data repositories. *Proteomics* 9, 4653-4663, doi:10.1002/pmic.200900216 (2009).

38 Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H. & Vizcaino, J. A. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 15, 930-949, doi:10.1002/pmic.201400302 (2015).

39 Chait, B. T. Chemistry. Mass spectrometry: bottom-up or top-down? *Science* 314, 65-66, doi:10.1126/science.1133987 (2006).

40 Yates, J. R., 3rd & Kelleher, N. L. Top down proteomics. *Analytical chemistry* 85, 6151, doi:10.1021/ac401484r (2013).

41 Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* 389, 1017-1031, doi:10.1007/s00216-007-1486-6 (2007).

42 Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews. Molecular cell biology* 5, 699-711, doi:10.1038/nrm1468 (2004).

43 Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* 162, 425-440, doi:10.1016/j.cell.2015.06.043 (2015).

44 Medard, G. *et al.* Optimized chemical proteomics assay for kinase inhibitor profiling. *Journal of proteome research* 14, 1574-1586, doi:10.1021/pr5012608 (2015).

45 Egas, D. A. & Wirth, M. J. Fundamentals of protein separations: 50 years of nanotechnology, and growing. *Annual review of analytical chemistry* 1, 833-855, doi:10.1146/annurev.anchem.1.031207.112912 (2008).

46 Jafari, M. *et al.* Comparison of in-gel protein separation techniques commonly used for fractionation in mass spectrometry-based proteomic profiling. *Electrophoresis* 33, 2516-2526, doi:10.1002/elps.201200031 (2012).

47 Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & cellular proteomics : MCP* 3, 608-614, doi:10.1074/mcp.T400003-MCP200 (2004).

48 Guo, X., Trudgian, D. C., Lemoff, A., Yadavalli, S. & Mirzaei, H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Molecular & cellular proteomics : MCP* 13, 1573-1584, doi:10.1074/mcp.M113.035170 (2014).

49 Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of proteome research* 9, 1323-1329, doi:10.1021/pr900863u (2010).

50 Eriksson, J. & Fenyo, D. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nature biotechnology* 25, 651-655, doi:10.1038/nbt1315 (2007).

51 Villen, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature protocols* 3, 1630-1638, doi:10.1038/nprot.2008.150 (2008).

52 Ruprecht, B. *et al.* Comprehensive and reproducible phosphopeptide enrichment using iron immobilized metal ion affinity chromatography (Fe-IMAC) columns. *Molecular & cellular proteomics : MCP* 14, 205-215, doi:10.1074/mcp.M114.043109 (2015).

53 Kettenbach, A. N. & Gerber, S. A. Rapid and reproducible single-stage phosphopeptide enrichment of complex peptide mixtures: application to general and phosphotyrosine-specific phosphoproteomics experiments. *Analytical chemistry* 83, 7635-7644, doi:10.1021/ac201894j (2011).

54 Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nature reviews. Molecular cell biology* 15, 536-550, doi:10.1038/nrm3841 (2014).

55 Udeshi, N. D., Mertins, P., Svinkina, T. & Carr, S. A. Large-scale identification of ubiquitination sites by mass spectrometry. *Nature protocols* 8, 1950-1960, doi:10.1038/nprot.2013.120 (2013).

56 Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell reports* 8, 1583-1594, doi:10.1016/j.celrep.2014.07.036 (2014).

57 Essader, A. S., Cargile, B. J., Bundy, J. L. & Stephenson, J. L., Jr. A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics* 5, 24-34, doi:10.1002/pmic.200400888 (2005).

58 Branca, R. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods* 11, 59-62, doi:10.1038/nmeth.2732 (2014).

59 Quan, Q. *et al.* Fully Automated Multidimensional Reversed-Phase Liquid Chromatography with Tandem Anion/Cation Exchange Columns for Simultaneous Global Endogenous Tyrosine Nitration Detection, Integral Membrane Protein Characterization, and Quantitative Proteomics Mapping in Cerebral Infarcts. *Analytical chemistry* 87, 10015-10024, doi:10.1021/acs.analchem.5b02619 (2015).

60 Horie, K. *et al.* Hydrophilic interaction chromatography using a meter-scale monolithic silica capillary column for proteomics LC-MS. *Analytical chemistry* 86, 3817-3824, doi:10.1021/ac4038625 (2014).

61 Wilm, M. Principles of electrospray ionization. *Molecular & cellular proteomics : MCP* 10, M111 009407, doi:10.1074/mcp.M111.009407 (2011).

62 Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes* 136, 167-180, doi:http://dx.doi.org/10.1016/0168-1176(94)04024-9 (1994).

63 Iribarne, J. V. & Thomson, B. A. On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics* 64, 2287-2294, doi:doi:http://dx.doi.org/10.1063/1.432536 (1976).

64 Hahne, H. *et al.* DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nature methods* 10, 989-991, doi:10.1038/nmeth.2610 (2013).

65 Wickremsinhe, E. R., Singh, G., Ackermann, B. L., Gillespie, T. A. & Chaudhary, A. K. A review of nanoelectrospray ionization applications for drug metabolism and pharmacokinetics. *Current drug metabolism* 7, 913-928 (2006).

66 Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. *Mass spectrometry reviews* 24, 1-29, doi:10.1002/mas.20004 (2005).

67 Savaryn, J. P., Toby, T. K. & Kelleher, N. L. A researcher's guide to mass spectrometry-based proteomics. *Proteomics* 16, 2435-2443, doi:10.1002/pmic.201600113 (2016).

68 Schwartz, J. C. Quadrupole ion traps and a new era of evolution. *Online* (2004).

69 Guilhaus, M. Principles and instrumentation in time-of-flight mass spectrometry. *Journal of Mass Spectrometry* 30, 1519-1532 (1995).

70 Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier transform mass spectrometry. *Molecular & cellular proteomics : MCP* 10, M111 009431, doi:10.1074/mcp.M111.009431 (2011).

71 Römpp, A. *et al.* Examples of Fourier transform ion cyclotron resonance mass spectrometry developments: from ion physics to remote access biochemical mass spectrometry. *European journal of mass spectrometry* 11, 443-456, doi:10.1255/ejms.732 (2005).

72 Makarov, A., Denisov, E., Lange, O. & Horning, S. Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *Journal of the American Society for Mass Spectrometry* 17, 977-982, doi:10.1016/j.jasms.2006.03.006 (2006).

73 Makarov, A., Denisov, E. & Lange, O. Performance evaluation of a high-field Orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry* 20, 1391-1396, doi:10.1016/j.jasms.2009.01.005 (2009).

74 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 & cellular proteomics : MCP* 11, O111 013698, doi:10.1074/mcp.O111.013698 (2012).

75 Kelstrup, C. D. *et al.* Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *Journal of proteome research* 13, 6187-6195, doi:10.1021/pr500985w (2014).

76 Nielsen, M. L., Savitski, M. M. & Zubarev, R. A. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Molecular & cellular proteomics : MCP* 4, 835-845, doi:10.1074/mcp.T400022-MCP200 (2005).

77 Shen, Y. *et al.* Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. *Journal of proteome research* 10, 3929-3943, doi:10.1021/pr200052c (2011).

78 Voinov, V. G., Beckman, J. S., Deinzer, M. L. & Barofsky, D. F. Electron-capture dissociation (ECD), collision-induced dissociation (CID) and ECD/CID in a linear radio-frequency-free magnetic cell. *Rapid communications in mass spectrometry : RCM* 23, 3028-3030, doi:10.1002/rcm.4209 (2009).

79 Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry* 11, 601, doi:10.1002/bms.1200111109 (1984).

80 Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T. & Watson, J. T. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Analytical chemistry* 59, 2621-2625 (1987).

81 Wells, J. M. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods in enzymology* 402, 148-185, doi:10.1016/S0076-6879(05)02005-7 (2005).

82 Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* 4, 709-712, doi:10.1038/nmeth1060 (2007).

83 Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 101, 9528-9533, doi:10.1073/pnas.0402700101 (2004).

84 Yost, R. A. & Enke, C. G. Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society* 100, 2274-2275, doi:10.1021/ja00475a072 (1978).

85 Chernushevich, I. V., Loboda, A. V. & Thomson, B. A. An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of mass spectrometry : JMS* 36, 849-865, doi:10.1002/jms.207 (2001).

86 Andrews, G. L., Simons, B. L., Young, J. B., Hawkridge, A. M. & Muddiman, D. C. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Analytical chemistry* 83, 5442-5446, doi:10.1021/ac200812d (2011).

87 Olsen, J. V. *et al.* A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Molecular & cellular proteomics : MCP* 8, 2759-2769, doi:10.1074/mcp.M900375-MCP200 (2009).

88 Savitski, M. M. *et al.* Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *Journal of the American Society for Mass Spectrometry* 21, 1668-1679, doi:10.1016/j.jasms.2010.01.012 (2010).

89 Parker, C. E., Pearson, T. W., Anderson, N. L. & Borchers, C. H. Mass-spectrometry-based clinical proteomics--a review and prospective. *The Analyst* 135, 1830-1838, doi:10.1039/c0an00105h (2010).

90 Paulo, J. A. Practical and Efficient Searching in Proteomics: A Cross Engine Comparison. *WebmedCentral* 4, doi:10.9754/journal.wplus.2013.0052 (2013).

91 Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & cellular proteomics : MCP* 11, 1475-1488, doi:10.1074/mcp.O112.020131 (2012).

92 Huang, Q. *et al.* SWATH enables precise label-free quantification on proteome scale. *Proteomics* 15, 1215-1223, doi:10.1002/pmic.201400270 (2015).

93 Liu, Y., Huttenhain, R., Collins, B. & Aebersold, R. Mass spectrometric protein maps for biomarker discovery and clinical research. *Expert review of molecular diagnostics* 13, 811-825, doi:10.1586/14737159.2013.845089 (2013).

94 Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & cellular proteomics : MCP* 9, 2252-2261, doi:10.1074/mcp.M110.001537 (2010).

95 Tsou, C. C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* 12, 258-264, 257 p following 264, doi:10.1038/nmeth.3255 (2015).

96 Bilbao, A. *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* 15, 964-980, doi:10.1002/pmic.201400323 (2015).

97 Wasinger, V. C., Zeng, M. & Yau, Y. Current status and advances in quantitative proteomic mass spectrometry. *International journal of proteomics* 2013, 180605, doi:10.1155/2013/180605 (2013).

98 Matzke, M. M. *et al.* A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics* 13, 493-503, doi:10.1002/pmic.201200269 (2013).

99 Chahrour, O., Cobice, D. & Malone, J. Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *Journal of pharmaceutical and biomedical analysis* 113, 2-20, doi:10.1016/j.jpba.2015.04.013 (2015).

100 Gilchrist, A. *et al.* Quantitative proteomics analysis of the secretory pathway. *Cell* 127, 1265-1281, doi:10.1016/j.cell.2006.10.036 (2006).

101 Liu, H., Sadygov, R. G. & Yates, J. R., 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry* 76, 4193-4201, doi:10.1021/ac0498563 (2004).

102 Washburn, M. P., Wolters, D. & Yates, J. R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology* 19, 242-247, doi:10.1038/85686 (2001).

103 Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology* 25, 117-124, doi:10.1038/nbt1270 (2007).

104 Higgs, R. E., Knierman, M. D., Gelfanova, V., Butler, J. P. & Hale, J. E. Comprehensive label-free method for the relative quantification of proteins from biological samples. *Journal of proteome research* 4, 1442-1450, doi:10.1021/pr050109b (2005).

105 Boutilier, J. M., Warden, H., Doucette, A. A. & Wentzell, P. D. Chromatographic behaviour of peptides following dimethylation with H2/D2-formaldehyde: implications for comparative proteomics. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* 908, 59-66, doi:10.1016/j.jchromb.2012.09.035 (2012).

106 Minogue, C. E. *et al.* Multiplexed quantification for data-independent acquisition. *Analytical chemistry* 87, 2570-2575, doi:10.1021/ac503593d (2015).

107 Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America* 100, 6940-6945, doi:10.1073/pnas.0832254100 (2003).

108 Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP* 1, 376-386 (2002).

109 Hebert, A. S. *et al.* Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature methods* 10, 332-334, doi:10.1038/nmeth.2378 (2013).

110 Zanivan, S., Krueger, M. & Mann, M. In vivo quantitative proteomics: the SILAC mouse. *Methods in molecular biology* 757, 435-450, doi:10.1007/978-1-61779-166-6_25 (2012).

111 Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology* 17, 994-999, doi:10.1038/13690 (1999).

112 Hsu, J. L., Huang, S. Y., Chow, N. H. & Chen, S. H. Stable-isotope dimethyl labeling for quantitative proteomics. *Analytical chemistry* 75, 6843-6852, doi:10.1021/ac0348625 (2003).

113 Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP* 3, 1154-1169, doi:10.1074/mcp.M400129-MCP200 (2004).

114 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).

115 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods* 8, 937-940, doi:10.1038/nmeth.1714 (2011).

116 Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of proteome research* 12, 3586-3598, doi:10.1021/pr400098r (2013).

117 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61-70, doi:10.1038/nature11412 (2012).

118 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330-337, doi:10.1038/nature11252 (2012).

119 Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell reports* 4, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).

120 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, doi:10.1038/nbt.1511 (2008).

121 Aiche, S. *et al.* Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics* 15, 1443-1447, doi:10.1002/pmic.201400391 (2015).

122 Sturm, M. *et al.* OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* 9, 163, doi:10.1186/1471-2105-9-163 (2008).

123 Deutsch, E. W. *et al.* A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10, 1150-1159, doi:10.1002/pmic.200900375 (2010).

124 Deutsch, E. W., Lam, H. & Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological genomics* 33, 18-25, doi:10.1152/physiolgenomics.00298.2007 (2008).

125 Vitek, O. Getting started in computational mass spectrometry-based proteomics. *PLoS computational biology* 5, e1000366, doi:10.1371/journal.pcbi.1000366 (2009).

126 Perez-Riverol, Y. *et al.* Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochimica et biophysica acta* 1844, 63-76, doi:10.1016/j.bbapap.2013.02.032 (2014).

127 Gehlenborg, N. *et al.* Visualization of omics data for systems biology. *Nature methods* 7, S56-68, doi:10.1038/nmeth.1436 (2010).

128 Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nature biotechnology* 25, 887-893, doi:10.1038/nbt1329 (2007).

129 Deutsch, E. mzML: a single, unifying data format for mass spectrometer output. *Proteomics* 8, 2776-2777, doi:10.1002/pmic.200890049 (2008).

130 Martens, L. *et al.* mzML--a community standard for mass spectrometry data. *Molecular & cellular proteomics : MCP* 10, R110 000133, doi:10.1074/mcp.R110.000133 (2011).

131 Wilhelm, M., Kirchner, M., Steen, J. A. & Steen, H. mz5: space- and time-efficient storage of mass spectrometry data sets. *Molecular & cellular proteomics : MCP* 11, O111 011379, doi:10.1074/mcp.O111.011379 (2012).

132 Shah, A. R. *et al.* An efficient data format for mass spectrometry-based proteomics. *Journal of the American Society for Mass Spectrometry* 21, 1784-1788, doi:10.1016/j.jasms.2010.06.014 (2010).

133 Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 46, 13 24 11-19, doi:10.1002/0471250953.bi1324s46 (2014).

134 Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534-2536, doi:10.1093/bioinformatics/btn323 (2008).

135 Jones, A. R. *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & cellular proteomics : MCP* 11, M111 014381, doi:10.1074/mcp.M111.014381 (2012).

136 Walzer, M. *et al.* The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Molecular & cellular proteomics : MCP* 12, 2332-2340, doi:10.1074/mcp.O113.028506 (2013).

137 Keller, A., Eng, J., Zhang, N., Li, X. J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular systems biology* 1, 2005 0017, doi:10.1038/msb4100024 (2005).

138 Deutsch, E. W. File formats commonly used in mass spectrometry proteomics. *Molecular & cellular proteomics : MCP* 11, 1612-1621, doi:10.1074/mcp.R112.019695 (2012).

139 Listgarten, J. & Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & cellular proteomics : MCP* 4, 419-434, doi:10.1074/mcp.R500005-MCP200 (2005).

140 Matthiesen, R. Extracting monoisotopic single-charge peaks from liquid chromatography-electrospray ionization-mass spectrometry. *Methods in molecular biology* 367, 37-48, doi:10.1385/1-59745-275-0:37 (2007).

141 Na, S. & Paek, E. Quality assessment of tandem mass spectra based on cumulative intensity normalization. *Journal of proteome research* 5, 3241-3248, doi:10.1021/pr0603248 (2006).

142 Du, P., Kibbe, W. A. & Lin, S. M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22, 2059-2065, doi:10.1093/bioinformatics/btl355 (2006).

143 Renard, B. Y. *et al.* When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *Proteomics* 9, 4978-4984, doi:10.1002/pmic.200900326 (2009).

144 Hubbard, S. J. Computational approaches to peptide identification via tandem MS. *Methods in molecular biology* 604, 23-42, doi:10.1007/978-1-60761-444-9_3 (2010).

145 Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* 73, 2092-2123, doi:10.1016/j.jprot.2010.08.009 (2010).

146 Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A. & Pevzner, P. A. De novo peptide sequencing and identification with precision mass spectrometry. *Journal of proteome research* 6, 114-123, doi:10.1021/pr060271u (2007).

147 Ma, B. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* 17, 2337-2342, doi:10.1002/rcm.1196 (2003).

148 Pan, C. *et al.* A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC bioinformatics* 11, 118, doi:10.1186/1471-2105-11-118 (2010).

149 Renard, B. Y. *et al.* Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Molecular & cellular proteomics : MCP* 11, M111 014167, doi:10.1074/mcp.M111.014167 (2012).

150 Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical chemistry* 66, 4390-4399 (1994).

151 Tanner, S. *et al.* InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical chemistry* 77, 4626-4639, doi:10.1021/ac050102d (2005).

152 Zhang, J. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics : MCP* 11, M111 010587, doi:10.1074/mcp.M111.010587 (2012).

153 Wenger, C. D. & Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of proteome research* 12, 1377-1386, doi:10.1021/pr301024c (2013).

154 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. *Journal of the American Society for Mass Spectrometry* 5, 976-989, doi:10.1016/1044-0305(94)80016-2 (1994).

155 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, doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (1999).

156 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* 10, 1794-1805, doi:10.1021/pr101065j (2011).

157 Brosch, M. & Choudhary, J. Scoring and validation of tandem MS peptide identification methods. *Methods in molecular biology* 604, 43-53, doi:10.1007/978-1-60761-444-9_4 (2010).

158 Fenyo, D. & Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry* 75, 768-774 (2003).

159 Alves, G. *et al.* Calibrating E-values for MS2 database search methods. *Biology direct* 2, 26, doi:10.1186/1745-6150-2-26 (2007).

160 Matsuda, F., Tsugawa, H. & Fukusaki, E. Method for assessing the statistical significance of mass spectral similarities using basic local alignment search tool statistics. *Analytical chemistry* 85, 8291-8297, doi:10.1021/ac401564v (2013).

161 Klammer, A. A., Park, C. Y. & Noble, W. S. Statistical calibration of the SEQUEST XCorr function. *Journal of proteome research* 8, 2106-2113, doi:10.1021/pr8011107 (2009).

162 Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467, doi:10.1093/bioinformatics/bth092 (2004).

163 Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *Journal of proteome research* 3, 958-964, doi:10.1021/pr0499491 (2004).

164 Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* 13, 22-24, doi:10.1002/pmic.201200439 (2013).

165 Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research* 6, 654-661, doi:10.1021/pr0604054 (2007).

166 Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. Combining results of multiple search engines in proteomics. *Molecular & cellular proteomics : MCP* 12, 2383-2393, doi:10.1074/mcp.R113.027797 (2013).

167 Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655-667, doi:10.1002/pmic.200600625 (2007).

168 Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry* 78, 5678-5684, doi:10.1021/ac060279n (2006).

169 Craig, R., Cortens, J. C., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research* 5, 1843-1849, doi:10.1021/pr0602085 (2006).

170 Choi, H. & Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of proteome research* 7, 47-50, doi:10.1021/pr700747q (2008).

171 Kall, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of proteome research* 7, 40-44, doi:10.1021/pr700739d (2008).

172 Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* 74, 5383-5392 (2002).

173 Choi, H. & Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of proteome research* 7, 254-265, doi:10.1021/pr070542g (2008).

174 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* 4, 207-214, doi:10.1038/nmeth1019 (2007).

175 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology* 604, 55-71, doi:10.1007/978-1-60761-444-9_5 (2010).

176 Chalkley, R. J. When target-decoy false discovery rate estimations are inaccurate and how to spot instances. *Journal of proteome research* 12, 1062-1064, doi:10.1021/pr301063v (2013).

177 Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry* 22, 1111-1120, doi:10.1007/s13361-011-0139-3 (2011).

178 Shen, C. *et al.* On the estimation of false positives in peptide identifications using decoy search strategy. *Proteomics* 9, 194-204, doi:10.1002/pmic.200800330 (2009).

179 Kall, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research* 7, 29-34, doi:10.1021/pr700600n (2008).

180 Wang, G., Wu, W. W., Zhang, Z., Masilamani, S. & Shen, R. F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical chemistry* 81, 146-159, doi:10.1021/ac801664q (2009).

181 Blanco, L., Mead, J. A. & Bessant, C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *Journal of proteome research* 8, 1782-1791 (2009).

182 Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & cellular proteomics : MCP* 10, M111 007690, doi:10.1074/mcp.M111.007690 (2011).

183 Lam, H., Deutsch, E. W. & Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *Journal of proteome research* 9, 605-610, doi:10.1021/pr900947u (2010).

184 Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature methods* 8, 430-435, doi:10.1038/nmeth.1584 (2011).

185 Chalkley, R. J. & Clauser, K. R. Modification site localization scoring: strategies and performance. *Molecular & cellular proteomics : MCP* 11, 3-14, doi:10.1074/mcp.R111.015305 (2012).

186 Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology* 24, 1285-1292, doi:10.1038/nbt1240 (2006).

187 Bailey, C. M. *et al.* SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *Journal of proteome research* 8, 1965-1971, doi:10.1021/pr800917p (2009).

188 Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research* 10, 5354-5362, doi:10.1021/pr200611n (2011).

189 Lemeer, S. *et al.* Phosphorylation site localization in peptides by MALDI MS/MS and the Mascot Delta Score. *Analytical and bioanalytical chemistry* 402, 249-260, doi:10.1007/s00216-011-5469-2 (2012).

190 Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & cellular proteomics : MCP* 10, M110 003830, doi:10.1074/mcp.M110.003830 (2011).

191 Na, S., Jeong, J., Park, H., Lee, K. J. & Paek, E. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Molecular & cellular proteomics : MCP* 7, 2452-2463, doi:10.1074/mcp.M800101-MCP200 (2008).

192 Liu, C., Yan, B., Song, Y., Xu, Y. & Cai, L. Peptide sequence tag-based blind identification of post-translational modifications with point process model. *Bioinformatics* 22, e307-313, doi:10.1093/bioinformatics/btl226 (2006).

193 Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* 33, 743-749, doi:10.1038/nbt.3267 (2015).

194 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* 4, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).

195 Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry* 75, 4646-4658 (2003).

196 Serang, O., MacCoss, M. J. & Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research* 9, 5346-5357, doi:10.1021/pr100594k (2010).

197 Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* 8, 2405-2417, doi:10.1074/mcp.M900317-MCP200 (2009).

198 Goeminne, L. J., Argentini, A., Martens, L. & Clement, L. Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines. *Journal of proteome research* 14, 2457-2465, doi:10.1021/pr501223t (2015).

199 Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Molecular & cellular proteomics : MCP* 5, 144-156, doi:10.1074/mcp.M500230-MCP200 (2006).

200 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* 473, 337-342, doi:10.1038/nature10098 (2011).

201 Fabre, B. *et al.* Label-free quantitative proteomics reveals the dynamics of proteasome complexes composition and stoichiometry in a wide range of human cell lines. *Journal of proteome research* 13, 3027-3037, doi:10.1021/pr500193k (2014).

202 Wisniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Molecular & cellular proteomics : MCP* 13, 3497-3506, doi:10.1074/mcp.M113.037309 (2014).

203 MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966-968, doi:10.1093/bioinformatics/btq054 (2010).

204 Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1, S96-104 (2002).

205 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* 13, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).

206 Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nature biotechnology* 28, 83-89, doi:10.1038/nbt.1592 (2010).

207 Clough, T. *et al.* Protein quantification in label-free LC-MS experiments. *Journal of proteome research* 8, 5275-5284, doi:10.1021/pr900610q (2009).

208 Webb-Robertson, B. J. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research* 14, 1993-2001, doi:10.1021/pr501138h (2015).

209 Liew, A. W., Law, N. F. & Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics* 12, 498-513, doi:10.1093/bib/bbq080 (2011).

210 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289-300, doi:10.2307/2346101 (1995).

211 Bland, J. M. & Altman, D. G. Multiple significance tests: the Bonferroni method. *Bmj* 310, 170 (1995).

212 Meunier, B. *et al.* Assessment of hierarchical clustering methodologies for proteomic data mining. *Journal of proteome research* 6, 358-366, doi:10.1021/pr060343h (2007).

213 Hathout, Y. Proteomic methods for biomarker discovery and validation. Are we there yet? *Expert review of proteomics* 12, 329-331, doi:10.1586/14789450.2015.1064771 (2015).

214 Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* 10, 48, doi:10.1186/1471-2105-10-48 (2009).

215 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4, 44-57, doi:10.1038/nprot.2008.211 (2009).

216 Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448-3449, doi:10.1093/bioinformatics/bti551 (2005).

217 Huang, S. S. & Fraenkel, E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Science signaling* 2, ra40, doi:10.1126/scisignal.2000350 (2009).

218 Haider, S. & Pal, R. Integrated analysis of transcriptomic and proteomic data. *Current genomics* 14, 91-110, doi:10.2174/1389202911314020003 (2013).

219 Vaudel, M. *et al.* Exploring the potential of public proteomics data. *Proteomics*, doi:10.1002/pmic.201500295 (2015).

220 UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* 43, D204-212, doi:10.1093/nar/gku989 (2015).

221 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 25-29, doi:10.1038/75556 (2000).

222 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, doi:10.1093/nar/gkv1070 (2015).

223 Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* 4, 1180-1211, doi:10.3390/cancers4041180 (2012).

224 Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic acids research* 42, D472-477, doi:10.1093/nar/gkt1102 (2014).

225 Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* 42, D358-363, doi:10.1093/nar/gkt1115 (2014).

226 Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* 43, D447-452, doi:10.1093/nar/gku1003 (2015).

227 Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research* 43, D512-520, doi:10.1093/nar/gku1267 (2015).

228 Martens, L. *et al.* Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* 5, 3501-3505, doi:10.1002/pmic.200401302 (2005).

229 Rebhan, M. Protein sequence databases. *Methods in molecular biology* 609, 45-57, doi:10.1007/978-1-60327-241-4_3 (2010).

230 Kersey, P. J. *et al.* The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 4, 1985-1988, doi:10.1002/pmic.200300721 (2004).

231 Griss, J. *et al.* Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. *Proteomics* 11, 4434-4438, doi:10.1002/pmic.201100363 (2011).

232 Huang, H. *et al.* A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* 27, 1190-1191, doi:10.1093/bioinformatics/btr101 (2011).

233 Benson, D. A. *et al.* GenBank. *Nucleic acids research* 42, D32-37, doi:10.1093/nar/gkt1030 (2014).

234 Cunningham, F. *et al.* Ensembl 2015. *Nucleic acids research* 43, D662-669, doi:10.1093/nar/gku1010 (2015).

235 Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic acids research* 40, D857-861, doi:10.1093/nar/gkr930 (2012).

236 Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic acids research* 43, D470-478, doi:10.1093/nar/gku1204 (2015).

237 Ponten, F., Jirstrom, K. & Uhlen, M. The Human Protein Atlas--a tool for pathology. *The Journal of pathology* 216, 387-393, doi:10.1002/path.2440 (2008).

238 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419, doi:10.1126/science.1260419 (2015).

239 Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database : the journal of biological databases and curation* 2010, baq020, doi:10.1093/database/baq020 (2010).

240 Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: current status. *Nucleic acids research* 43, D764-770, doi:10.1093/nar/gku1178 (2015).

241 Schaab, C., Geiger, T., Stoehr, G., Cox, J. & Mann, M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Molecular & cellular proteomics : MCP* 11, M111 014068, doi:10.1074/mcp.M111.014068 (2012).

242 Mathivanan, S. *et al.* Human Proteinpedia enables sharing of human protein data. *Nature biotechnology* 26, 164-167, doi:10.1038/nbt0208-164 (2008).

243 Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15, 3163-3168, doi:10.1002/pmic.201400441 (2015).

244 Smith, B. E., Hill, J. A., Gjukich, M. A. & Andrews, P. C. Tranche distributed repository and ProteomeCommons.org. *Methods in molecular biology* 696, 123-145, doi:10.1007/978-1-60761-987-1_8 (2011).

245 Desiere, F. *et al.* The PeptideAtlas project. *Nucleic acids research* 34, D655-658, doi:10.1093/nar/gkj040 (2006).

246 Deutsch, E. W. *et al.* State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *Journal of proteome research* 14, 3461-3473, doi:10.1021/acs.jproteome.5b00500 (2015).

247 Washburn, M. P. The H-Index of 'An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database'. *Journal of the American Society for Mass Spectrometry* 26, 1799-1803, doi:10.1007/s13361-015-1181-3 (2015).

248 Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC bioinformatics* 13 Suppl 16, S1, doi:10.1186/1471-2105-13-S16-S1 (2012).

249 Farrah, T. *et al.* PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* 12, 1170-1175, doi:10.1002/pmic.201100515 (2012).

250 Martens, L. *et al.* PRIDE: the proteomics identifications database. *Proteomics* 5, 3537-3545, doi:10.1002/pmic.200401303 (2005).

251 Vizcaino, J. A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research* 41, D1063-1069, doi:10.1093/nar/gks1262 (2013).

252 Dean, J. & Ghemawat, S. MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 107-113, doi:10.1145/1327452.1327492 (2008).

253 Craig, R., Cortens, J. P. & Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *Journal of proteome research* 3, 1234-1242, doi:10.1021/pr049882h (2004).

Chapter 2

A protein centric in-memory database to facilitate the analysis of LC-MS/MS data sets

Contents

"So, what you can do in Microsoft Word is what Bill Gates has decided. What you can do in Oracle Database is what Larry Ellison and his crew have decided."

- Ted Nelson

1 Introduction

Mass spectrometry-based proteomics is rapidly evolving into the technology of choice to identify and quantify thousands of proteins in a single experiment, which allows the large-scale interrogation of biological systems. However, most studies focus their analysis and interpretation on a few central points, such as the difference between a normal and disease state, the expression landscape of proteins in a set of cell lines, tissues or body fluids or the elucidation of specific signaling pathways. While this allows the in-depth investigation of specific phenomena, this approach does not or only to a limited degree make use of previously conducted experiments. An alternative approach is the integration of published data and thus knowledge to acquire new insight. Multiple different databases storing the acquired data and results have been released in the past years[1]. Typically, they include a repository to allow researchers from all fields to gain access to the raw data and result files, which facilitates re-use, re-analyses and cross experiment comparison. However, besides the long-term storage and access to the data, most of these resources provide no or only limited access to analytical features to cross compare or integrate different studies. This renders the interaction with previous results without manual reprocessing and reanalysis difficult, if not impossible, and even the comparison of protein expression patterns across multiple experiments within a single repository is often not possible.
Three main features are prerequisite to allow the real-time analysis of data originating from multiple studies: (i) a generic queryable data model for identification and quantification results, (ii) the integration of the data with the underlying experimental design, and (iii) a performant and scalable data management system. While the last prerequisite is only necessary for real-time analysis, the lack or partial implementation of any of the first two renders the system incapable of providing comprehensive analysis tools. The implementation of a public repository for data dissemination is optional, but further facilitates data sharing and transparency across the scientific community.

The Proteomics identification archive (PRIDE)[2] provides central storage and archiving of mass spectrometry data enabling users to upload raw data and search results including quantification results. Yet, the data is not queryable and researchers with a specific question, for example which cell line to choose to study a protein of interest, cannot make direct use of the large number of studies stored in PRIDE. PeptideAtlas provides queryable access to identification data from mass spectrometry-based experiments for many organisms[3]. The data is stored in (yearly) builds, yet is limited to identification results. Specific identifications can be tracked to single experiments, but PeptideAtlas lacks a comprehensive user interface to easily search for proteins in certain biological sources. The protein abundance database (PAXDB)[4] on the other hand stores quantification data from publicly available data, yet lacks the underlying peptide identification results. Unfortunately, no additional metadata about the experimental design is recorded and thus a cross experiment comparison is not available.

While most repositories require a basic annotation of data, the experimental design is often not stored in a programmatically accessible format. Besides the capabilities of proteomics to map proteomes of biological systems where specific treatment parameters are of inferior importance, many more applications show the power of this technology in deciphering protein dynamics[5-7]. For correct data processing and interpretation, all experimental factors have to be recorded to

enable a context-sensitive analysis and integration[8]. For example, mass spectrometry-based proteomics is becoming a suitable tool in preclinical drug discovery by providing means to analyze protein-drug interactions in both a targeted and unbiased manner. Its applications range from target selection, deconvolution and validation to lead selection/optimization and pre-clinical testing[9]. In these experiments, experimental factors, such as doses, incubation times and temperatures are essential for individual analysis and cross-experiment comparisons.

Due to constant technological advances, manifested in higher multiplexing capabilities as well as increased acquisition speed and thus throughput, enabling the analysis of larger sample sizes, the amount of data generated in single experiments is continuously rising[10]. This poses new challenges on existing databases and repositories. The CHORUS repository enables the storage and analysis of MS data within a cloud environment, reducing the need of local storage and computing resources. The general aim is to create a catalog of MS data which can be openly access by both the general and scientific community. While this approach circumvents the challenges associated with big data for single labs, CHORUS currently does not support cross-experiment comparisons.
On top, regular refinements of gene models require constant changes to the underlying data of a proteomics database as identification properties of peptides and proteins change[11]. In order to keep these consistent with the current gene model, regular adjustments are necessary, increasing the need for performant database management systems. In contrast to classical disk storage, most commonly used for (relational) database management systems, in-memory database systems utilize the main memory as the primary data storage medium. This reduces disk seek when querying data and, ultimately, results in faster data retrieval. Implementing common mass spectrometry algorithms within the database exploits the fast access to the main memory and allows both rapid reprocessing as well as real-time data analysis.

This chapter presents a novel publicly available database, termed ProteomicsDB, utilizing the in-memory database management system SAP HANA[12]. ProteomicsDB allows the real-time interactive exploration of large collections of mass spectrometry-based proteomics data. The protein-centric interface not only enables users to quickly access quantification information across all experiments stored in ProteomicsDB, but also to view individual peptide evidence. If available, the integrated spectrum viewer automatically selects and presents reference data from synthetic peptides to validate peptide identification events. Using modern web-browser technologies, multiple interactive visualizations are available and enable the real-time exploration of multiple proteomes at the same time. Furthermore, the implementation of an experimental design enables ProteomicsDB to utilize meta-data attached to an experiment, exemplified on dose- and temperature-dependent assay data. This allows the analysis of off- and on-target analysis of drugs as well as the theoretical exploration of combination treatments.

1.1 SAP HANA

SAP HANA bundles the calculation, control flow and presentation logic into a single system (Figure 2.19). In contrast to regular databases, SAP HANA stores all data in a hardware optimized in-memory database. The central component of the database is the index server, which processes all incoming queries utilizing additional engines, such as the query plan optimizer and execution engine. In distributed environments, the name server stores the topology of the servers and is

responsible for locating components. The data can be stored in both row- and column-oriented storage. To decrease the memory footprint and increase performance, efficient data compression is applied to the main data. By default, each column is first compressed with a dictionary mapping all distinct values in a column to consecutive numbers. Additional compression algorithm such as prefix encoding, run length encoding or sparse encoding are automatically evaluated and applied.

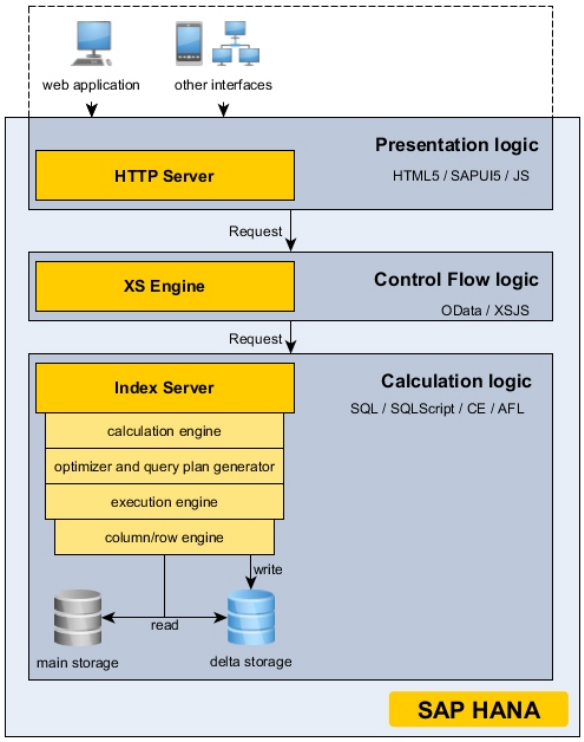


Figure 2.19 | SAP HANA architecture overview. SAP bundles a database management system, the calculation logic, an application server, the control flow logic, and a web server, the presentation logic.

The in memory storage reduces disk seek operation to a minimum and allows to make full use of the speed of the processor and main memory allowing real-time analysis of large amounts of data. Furthermore, due to the speed and compression of data, materializing views and aggregates are in most cases not necessary. For example, multi column joins use self-generated helper indices. However, setting secondary multi-column indices is possible and can additionally improve the performance.

Write operations on this compressed data are computationally demanding and instead performed on a separate data structure, called delta storage, which uses less efficient compression. This enables the database to move changes in collections from the delta storage to the main storage after which the main storage is persisted to the disk while the delta storage itself only exists in the main memory. This implies that changes to the delta storage have to be written to the disk, in form of a delta log. In case the system fails before the delta was merged into the main storage, the database will use the delta log to replicate the last changes.

Due to the strict separation of the calculation logic, control flow and presentation layer (Figure 2.19), SAP HANA can be used as a regular database management system (calculation logic), an application server (including control flow) or an entire web-application server (including presentation layer). HANA-based web applications typically make use of the integrated extended

application services (XS engine) and web/http server (HTTP Server; webdispatcher). This enables the use of data services (XSOData), server-side JavaScript (XSJS) processed and executed in the XS engine and SAPUI5/OPENUI5 to build modern and interactive HTML5 applications without the need of any additional servers. This decreases the number of individual servers necessary and thus overhead. This work is based on HANA 1.X utilizing the XS classic engine.
SAP HANA supports multiple (open) standards for data access and manipulation such as Open Database Connectivity (ODBC), Java Database Connectivity (JDBC), Object Linking and Embedding Database for Online Analytical Processing (ODBO) and Multidimensional Expressions (MDX). Additionally, different strategies are implemented to connect or integrate other external data sources as data providers (e.g. HADOOP), external applications (e.g. R via an R server) or external C++ libraries via the application function library. These services can be implemented directly in HANA using custom procedures or functions.

1.2 Open data protocol

The open data protocol (OData) defines the best practice for building and consuming queryable representational state transfer (RESTful) APIs (application programming interface) using simple hypertext transfer protocol (HTTP) or secure HTTP (HTTPS) requests. Although it is possible to alter data using the OData standard, in this context it is only used to expose database views. It handles the request and response headers, status codes, HTTP methods, URL conventions, payload formats and query options. Each OData method is described in machine-readable format, lists the input parameters and returns objects including their types, and thus allows the creation of generic clients consuming data from ProteomicsDB. Available output formats are extended markup language (XML) and JavaScript object notation (JSON).
OData supports various kind of query options, such as selecting a subset of the provided output properties (keyword *$select*), filtering based on the provided properties (keyword *$filter*), ordering (keyword *$orderby*) or to select only the top entries (keyword *$top*), skip a specified number of entries (keyword *$skip*) or counting the number of entries (keyword *$count*). These options are particularly helpful for querying tables with large amounts of data. Query options can be combined and provide a great variety of filtering and restricting the response.
SAP HANA fully supports the definition and rollout of OData services. These application programming interfaces (APIs) are directly coupled to previously defined views and procedures.

1.3 Kinobeads

Kinases are enzymes that transfer a phosphate group either to serine and threonine or tyrosine residue of their substrate. This influences the activity, cellular localization or the interaction spectrum of the substrate. They are one of the most important classes of drug targets, because they are considered the key regulators of cellular signaling and malfunction is highly associated with several human disease, like inflammation, diabetes and particularly cancer[13-15]. Furthermore, protein kinases are primary regulators of all hallmarks of cancer[16-18].
Small molecule inhibitors against protein kinases are designed to imitate ATP binding in the active site[19]. This binding is in competition with the co-substrate ATP and thus inhibits the activity of the kinase. However, this binding pocket is rather well conserved across most protein kinases[20], thus these small molecules often lack selectivity. Consequently, it is of great importance to deconvolute

the target space of a kinase inhibitor. There are 518 protein kinases encoded in the human genome[20]. Due to their low abundance, standard quantitative proteomics cannot readily be applied. Although many alternative approaches exist[21,22], they often suffer from several shortcomings such as the absence of regulatory domains and interacting proteins or the misrepresentation of the proteins' conformational state and post translational modification (PTMs) status.

One method to overcome these limitations is a targeted chemical proteomic assay, termed Kinobeads[23]. Here, bioactive molecules are coupled to e.g. sepharose beads for the specific enrichment of a sub-proteome. Kinobeads uses multiple immobilized unselective kinase inhibitors as probes to fish up to 350 kinases and other ATP and nucleotide binding proteins out of lysate[24]. The affinity binding constants of these proteins are monitored by increasing the concentration of a free drug, which compete for binding with the immobilized inhibitors (see Figure 2.20). Targets of the free drug lose their ability to bind to the matrix, thus showing a dose dependent depletion. This process is monitored at each dose using quantitative proteomics, resulting in the ability to calculate half maximal inhibitory concentrations (IC_{50}) or half maximal effective concentration (EC_{50}) for each target. In order to apply this method, the small molecule of interest does not need to be chemically modified. However, the method is targeted and only applicable if the drug uses the same binding mode as the immobilized probes.

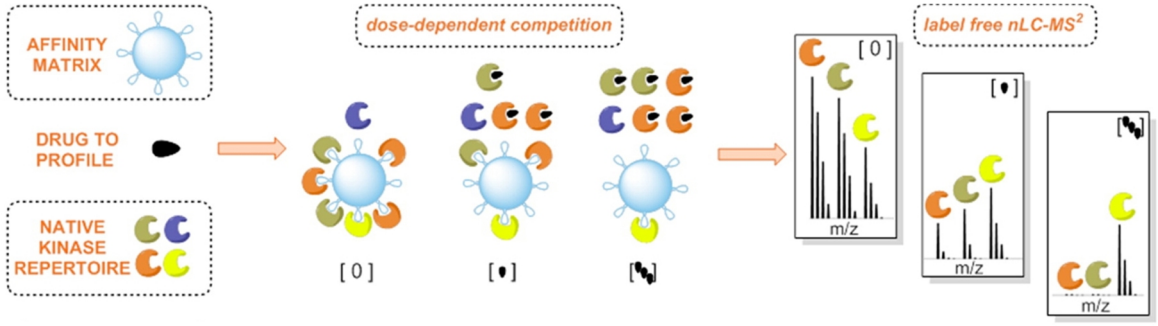


Figure 2.20 |Schematic illustration of a competitive Kinobeads pull-down. In increasing concentration, a free drug is first incubated with cell lysate (native kinase repertoire). Subsequently, Kinobeads fish unoccupied proteins. LC-MS readout using label-free or label-based approaches enables the quantification of multiple dosages. Subsequent normalization and curve fitting allows the estimation of protein-drug potencies for all interaction. Proteins, which do not interact with the free drug, show no dose-dependent effect (yellow protein). Figure from[24].

1.4 Cellular thermal shift assay

Kinobeads provide researchers with an easy to use assay to monitor the target space of kinase inhibitors. However, if the target space of the molecule is not known, a bioactive modified analogue is coupled to beads and used as the probe. The synthesis of such probes is typically difficult and prone to errors since both the bioactivity and binding mode have to match that of the unmodified version.

Thermal shift assays neither require the chemical modification of the small molecule nor do they rely on the same binding mode as the coupled probe[25-27]. Thermal shift assays, such as microscale thermal shift assay[28], are used to study the thermal stabilization of proteins upon ligand binding or folding and have been used in industry and academia to detect all kinds of (protein) interactions. The basic principle behind this assay is a gradual exposure to heat, which denatures

proteins and allows the determination of their melting point. Upon ligand binding, the melting point of a protein may change, as the ligand (typically) increases the energy needed to unfold the protein. This shift in the melting point can be monitored and used as an indication of interaction. This concept has been extended to allow its application in a cellular format, termed cellular thermal shift assay (CETSA)[25-27]. Briefly, cells are treated with a compound of interest, heated and lysed. A subsequent centrifugation separates the cell debris and aggregated proteins from the soluble proteins. Bound proteins are still in solution, even at elevated temperatures, and can be detected in the supernatant. This process is performed at different temperatures to derive the temperature-dependent fraction of non-denatured proteins. While the initial assay uses antibodies for readout, limiting this assay to proteins for which suitable antibodies exist, this method has been further extended to utilize quantitative mass spectrometry (see Figure 2.21a)[29,30]. This significantly increases the applications of CETSA since no prior hypothesis of protein-drug interaction is needed and thousands of proteins can be measured in a single experiment.

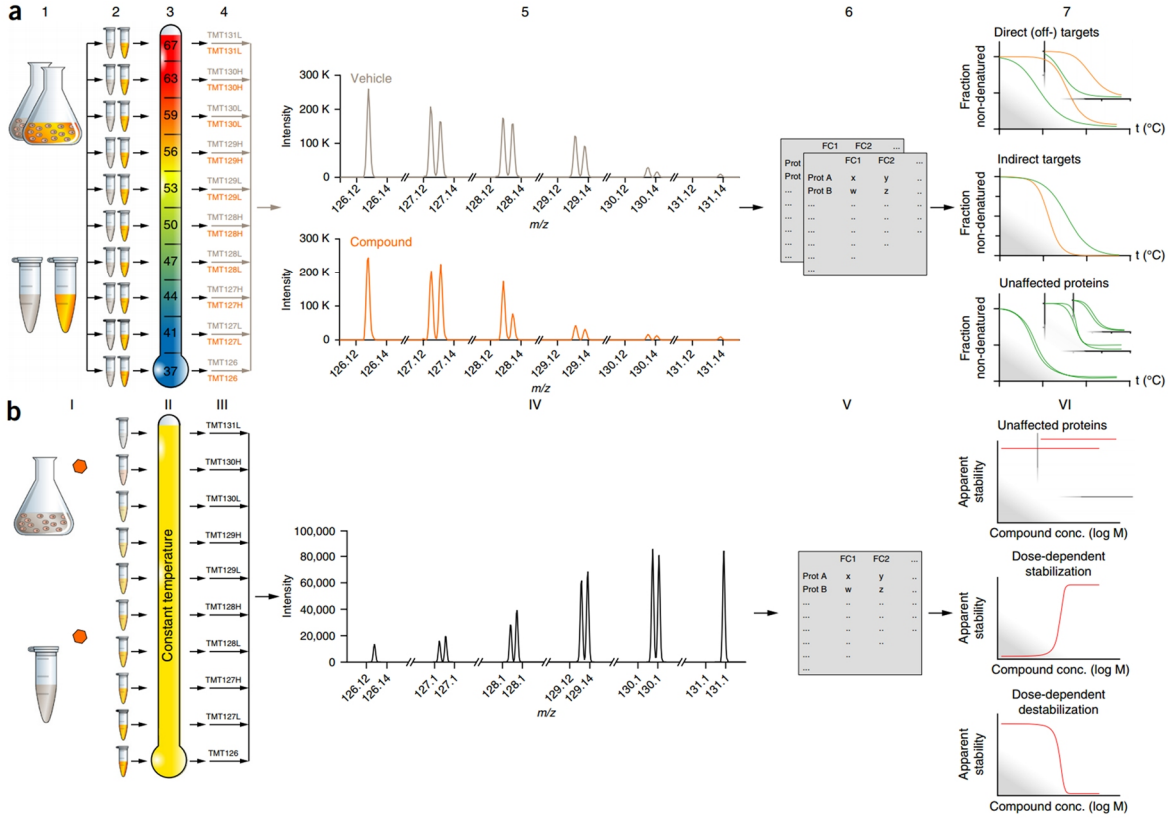


Figure 2.21 | Schematic illustration of an in-cell or in-lysate CETSA and ITDR experiment. a, To define the target space of a drug, the vehicle (gray) and drug (orange) treated samples are aliquoted and subjected to increasing heat. Each sample is then digested and labeled with one TMT isotope tag. Samples from drug or vehicle treated experiments are combined and analyzed using LC-MS allowing the quantification of up to 10 temperatures. After protein identification and quantification, the normalized melting curves of drug and vehicle treated samples of each protein are compared. Significant changes in the melting temperatures suggest a direct or indirect protein-drug interaction. b, To determine EC and IC values of protein-drug interactions, cells are treated with vehicle or drug (here 9 concentrations). Each sample is heated to the same temperature where for example the proteins-drug of interest are expected to show the biggest relative difference in fraction non-denatured (see CETSA). After digestion, labeling and LC-MS measurement, protein identification and quantification results are used to fit dose-response curves which

enable to determination of EC and IC values. Non drug-binding proteins show no effect, while proteins interacting with the drug can be either stabilized or destabilized. Figure from[30].

While CETSA only allows the detection of interactions of proteins and drugs in a cellular format, the results cannot be transformed into binding constants, such as an EC_{50}. For this purpose, the temperature is kept constant and the lysate or cell lines are treated with increasing concentration of the free drug (see Figure 2.21b). For maximal dynamic range, the temperature with the largest difference in the melting behavior is chosen. This experiment enables the determination of the EC_{50} and K_d of the protein-drug interaction, since the effect of binding can be monitored as the response of the change in the melting curve. This assay is termed isothermal dose response (ITDR).

2 Methods and implementation

2.1 Database structure

At the time of writing, ProteomicsDB consists of 80+ tables which can be grouped into four modules: i) static and metadata, which contains general purpose tables, such as annotations, controlled vocabulary ontologies and fragmentation rules; ii) repository, which contains the raw data including their annotation and metadata; iii) identification data, which stores spectra and the associated PSMs and reference data; and iv) quantification data, which holds peptide, protein and transcript expression data. For clarity, each module will be depicted as a simplified entity relationship model only containing the most important attributes.

2.1.1 Static and Metadata

ProteomicsDB uses multiple ontologies and controlled vocabularies (CV) for internal representation and annotation (Figure 2.22). At the time of writing, the following ontologies are used: PSI-MS (terms for proteomics and MS), BTO (BRENDA tissue ontology), UO (unit ontology), GO (gene ontology), sep (terms for chromatographic/separation methods). Terms not defined in the imported ontologies are created manually with the prefix PDB. Furthermore, user defined free-text input (e.g. drugs) is controlled in CVs as well. Terms, their definition and relation are stored in triplestore format enabling the representation of complex relations.

To model the design of an experiment, ProteomicsDB defines treatments as sets of predefined experimental factors which represent the sequential steps performed during sample preparation (Figure 2.22). An experimental factor is either an entity for numeric values, such as time and dose associated with a unit, or for (controlled) free-text fields, such as drugs and baits. At the time of writing, more than 20 treatments (e.g. heat treatment, drug treatment) are defined. Manifestations of factors used in the experimental design are stored as conditions in the repository.

The protein and peptide sequence space is essential for the analysis of proteomic data. For this, the complete protease-specific *in silico* digest of the human proteome (Figure 2.23) is stored to enable access to uniqueness information of all peptides. Additionally, general protein annotations, such as gene names, loci, synonyms and cross references as well as GO and domain information are stored (Figure 2.24).

In order to model the fragmentation of peptides, ProteomicsDB stores an internal representation of the fragmentation behavior of peptides to enable the context-specific generation of fragment ions of any type[31]. For this, masses of amino acids, immonium ions and other central atomic building blocks such as protons and electrons are stored alongside the definition of all major fragment ion types (ion series), such as b-ions or internal fragment ions. Depending on the fragmentation technique and charge-state of the precursor, only specific ion series are generated for annotation. Similarly, the annotation of neutral losses is amino acid and PTM specific and thus depends on the composition of the fragment ion.

2.1.2 Repository

To enable the organized and structured storage of raw data, three hierarchical layers for annotation are implemented (Figure 2.22). A project, belonging to a specific user, groups multiple experiments together. An experiment, apart from a name and description, requires the annotation of a scope. Further, experiments contain one or many samples, which again are represented by one or many raw MS files. However, due to multiplexing (e.g. SILAC or TMT), a single raw file can contain identification and quantification information of multiple samples. The annotation of samples contains information such as the mass spectrometer, biological material and treatment conditions used.

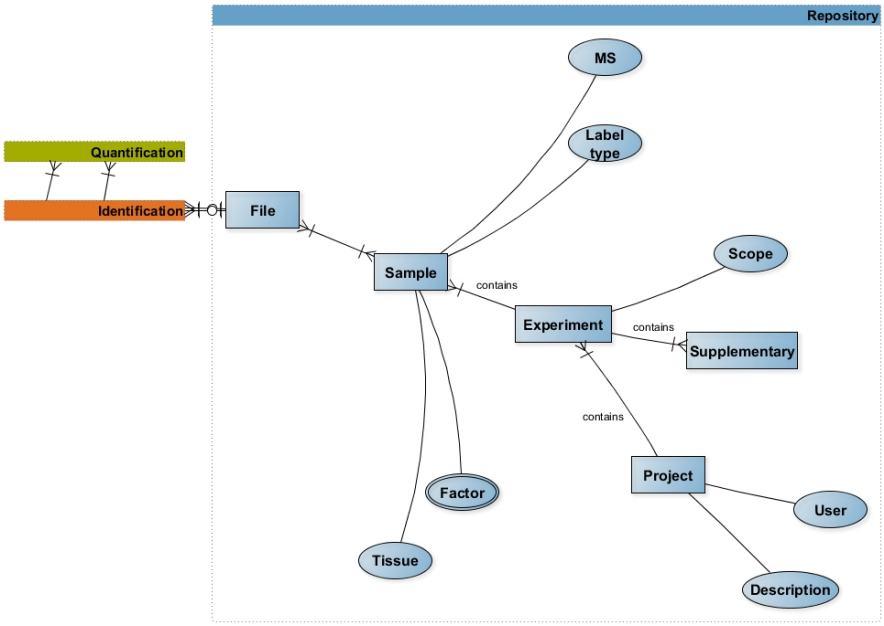


Figure 2.22 | Entity relationship model of the repository.

2.1.3 Identification data

ProteomicsDB is designed to enable the storage of any search result (Figure 2.23). For this purpose, a search is linked to files stored in the repository and is described by search parameters and the search engine used. Peptide spectrum matches (PSMs) associated with a search link a spectrum from an MS-file to a peptide sequence. Besides the search engine score, the FDR and other information about the precursor, a PSM can have multiple modifications attached to it. These map residues and localization probabilities within the peptide sequence to predefined modifications. While a PSM is linked to a peptide sequence from the *in silico* digest of the human proteome, the identified sequence as reported from the search engine is stored as well. This enables a later mapping if a different sequence database was used. In addition to the identified experimental spectra, ProteomicsDB allows the storage of reference spectra acquired from e.g. synthetic peptide standards. These are stored separately from the experimental spectra, but contain similar information on PSM and spectrum level. While experimental spectra are annotated using the fragmentation model stored in ProteomicsDB, the annotation of fragment peaks in reference spectra can be stored directly to enable manual annotation but also to avoid wrong assignment of peaks.

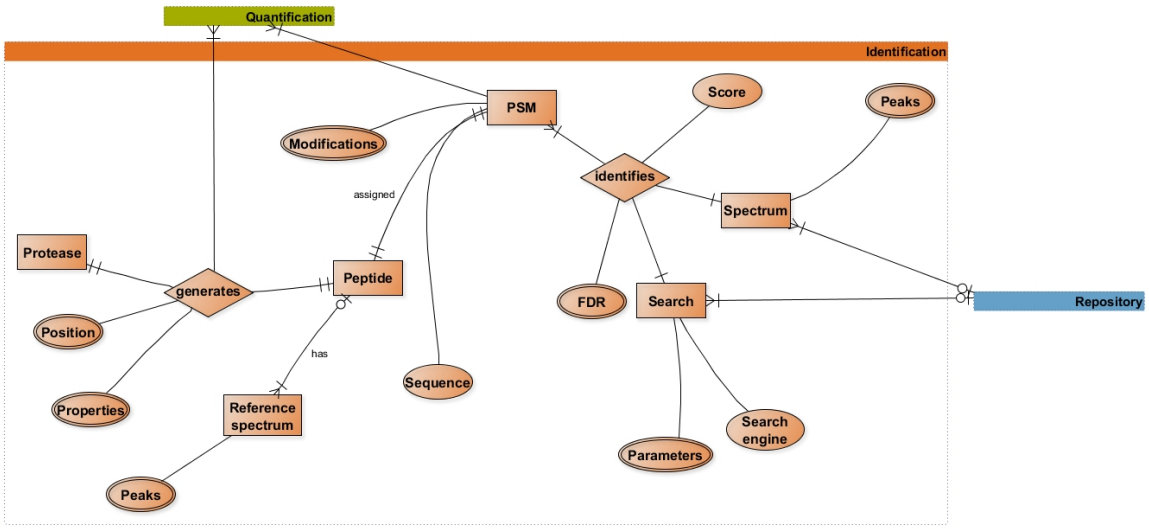


Figure 2.23 | Entity relationship model of identification data.

2.1.4 Quantification data

A core requirement of ProteomicsDB is the storage and evaluation of expression data (Figure 2.24). For this purpose, each PSM is associated with a label-dependent quantification of the corresponding MS-feature (e.g. XIC in label-free or reporter intensity in MS/MS-level quantification). With the mapping of PSMs to peptides, all available quantified features of a protein can be used to estimate a method-dependent expression. Besides expression estimates of proteins, transcript abundance measures can be stored in ProteomicsDB as well. In addition to this, the storage of (arbitrary) models which relate experimental factors (in form of conditions) from the experimental design to the expression of proteins. This enables the analysis of e.g. temperature- or dose-dependent data to identify differential behavior within or across experiments. For this, all fitted parameters and related properties such as the goodness of fit of a model are stored.

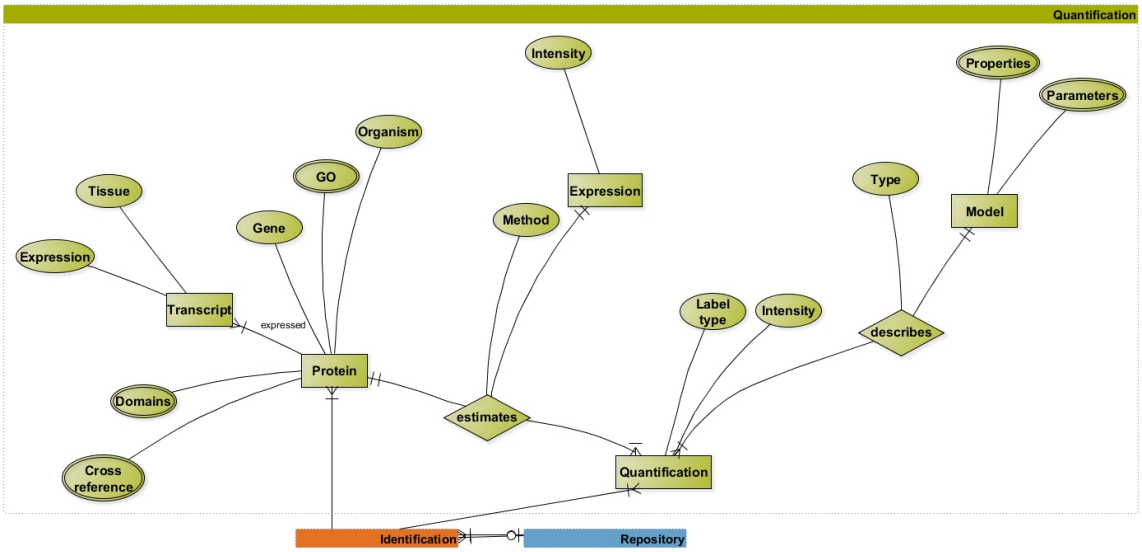


Figure 2.24 | Entity relationship model of quantification data.

2.2 Data import and processing

The entire import process consists of 6 steps (Figure 2.25): i) data selection, ii) manual annotation, iii) peptide identification with Mascot/Percolator[32,33], iii) peptide identification and quantification with Andromeda/Maxquant[34,35], iv) metadata extraction, v) data import and vi) post import data processing.

The primary data sources are the Chair of Proteomics and Bioanalytics and data from other laboratories either from public repositories such as PRIDE, PeptideAtlas and MassIVE/Tranche or direct communication. Datasets selected for import had to fulfil three criteria: (i) high resolution MS1 spectra, (ii) well annotated with respect to sample processing, data acquisition and used material, and (iii) increase the coverage of the human body by providing quantitative data on additional tissues, fluids and cell lines.

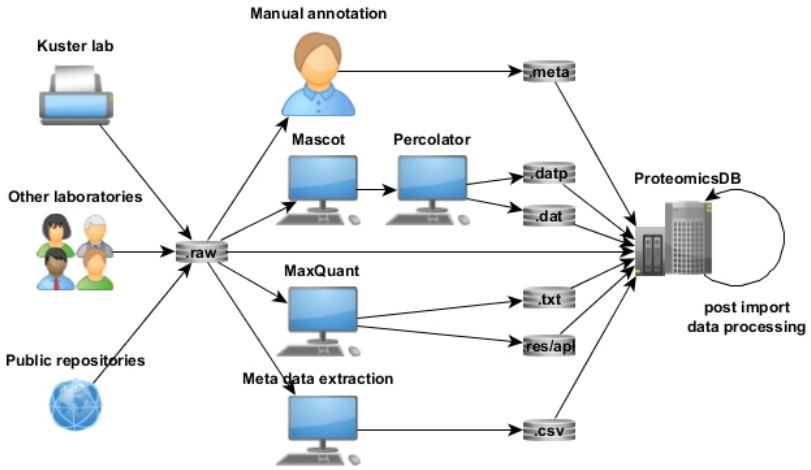


Figure 2.25 | Data processing pipeline. Raw MS-data is collected from three primary sources and subsequently processed via a uniform pipeline. After manual annotation of the data (e.g. used cell line, quantification technique, experimental design), the raw-files are processed via Andromeda/MaxQuant and Mascot/Percolator. Additional metadata can be extracted using an in-house Proteowizard-based application.

2.2.1 Data processing

After selection, each dataset (project) is manually annotated to record basic parameters of data acquisition as well the experimental setup such as utilized quantification technique, biological entities measured and the mapping of raw-files to samples. The resulting annotation is imported into ProteomicsDB, which generates the basic layout of projects, experiments and samples, as described earlier. Afterwards, all raw-files are processed with Mascot/Percolator and Andromeda/Maxquant. Additionally, a custom ProteoWizard-based[36] application was used to extract metadata associated with the mass spectrometer, the acquisition and each spectrum stored in the raw-files, such as utilized mass analyzers, isolation windows and precursor m/z.

2.2.2 Data import

Data import is controlled by a central queue implemented in bash. A simple text file of incoming jobs is monitored by a service on the database server. Depending on the file type (e.g. *dat*-file, *combined*-folder or *raw*-file), the respective import module is triggered. Each module performs an

initial sanity check to validate the incoming data for consistency with the annotation stored in ProteomicsDB.

The import of Mascot search results expects two input files, a *dat*- and *datp*-file. The *dat*-file contains the search results as provided by Mascot. The *datp*-file contains the percolated results of Mascot which provides posterior error probabilities (PEPs) for each peptide spectrum match (PSM). Before the import of the search engine results it is tested whether the processed *raw*-files are present in ProteomicsDB and assigned to a sample with the same quantification properties, e.g. SILAC or TMT. If the incoming data pass this test, the import process is triggered and all target and decoy (including up to rank 10 matches) PSMs are imported into ProteomicsDB.

The Maxquant importer expects a combined folder generated by Maxquant containing the *mqpar.xml*, a *txt*-folder and the respective *res*- and *apl*-files. After the input data passed the same tests as performed by the Mascot importer, the *res*- and *apl*-files containing the processed MS/MS spectra are indexed to enable fast data import. For this purpose, the index contains a mapping of all spectra and their respective peptide sequence and score within all *res*-files. The corresponding *apl*-file contains the deconvoluted spectrum which is imported into ProteomicsDB alongside the identification data, such as score, PEP and modifications.

2.2.3 Post import data processing

After data import, each LC-MS/MS run is separately FDR adjusted. For this purpose, the highest scoring PSMs per spectrum from one search engine are sorted by score in descending order. A first iteration initializes the q-values with the ratio of the number of decoy matches to the number of target matches with scores greater or equal the current selection. To allow consistent filtering, a second iteration in reverse order sets the final q-value to the minimum q-value observed for PSMs with a score less than or equal to the selected one. This reduces sampling artifacts and generates a continuously decreasing q-value curve when plotted against the search engine score. After assigning q-values to each PSM, a subsequent step aggregates spectral information on the protein level (e.g. the number of PSMs, maximum score and minimum q-value) and stores the results in a separate table.

Two of the most popular approaches of estimating protein expression values have been implemented in ProteomicsDB. Protein abundance is estimated using the iBAQ approach[37] as well as the top3 intensity approach[38,39]. For iBAQ, peptide intensities for all peptides in a sample are obtained, summed and divided by the number of theoretically observable peptides. For the top3 approach, peptide intensities of the three most intense peptides in a sample are obtained and summed. In order to be able to compare protein abundances across multiple samples, experiments and projects, iBAQ or top3 intensity values are normalized based on the total sum of the respective protein intensities. Protein abundance is calculated for different label types (e.g. light and heavy SILAC) separately, as these represent different samples. In the case of label-free quantification, SILAC and dimethyl labeling experiments, iBAQ and top3 intensities are derived from MS1 intensity measurements. For isobaric labeling (TMT or iTRAQ), iBAQ and top3 values were calculated based on MS2 reporter intensities. However, iBAQ and top3 abundance estimates derived from MS1 and MS2 intensity-based quantification are not directly comparable.

2.2.4 Spectrum annotation

All experimental spectra are annotated in real-time using the fragmentation rule set described in the CID/HCD expert annotation system[31]. This allows the incorporation of new fragmentation rules without the necessity to recalculate all spectrum annotations. The fragmentation rules stored in ProteomicsDB are used to generate lists of theoretically observable fragment ions and neutral losses. Based on the peptide sequence, precursor charge, present modifications and fragmentation method of the selected PSM, a procedure implemented in L generates all possible fragment ions (ion series). All annotated ions from the experimental spectrum are then used to generate all possible neutral loss ions. Possible neutral losses for all fragment ions are generated based on the amino acid composition and modification status. If multiple annotations are possible for a single peak in an experimental fragment spectrum, the one(s) with the highest priority are chosen.

2.2.5 Curve fitting

All curve models were fitted using an in-house R-script utilizing the R packet *drc*[40]. Dose dependent data were fitted with a log-logistic four-parameter regression model, while for temperature-dependent data a logistic four-parameter regression model was used. Briefly, protein abundance estimates were normalized to the vehicle control and each model was fitted separately. The four model-parameters and the corresponding coefficient of determination[41] (R^2) and the Bayesian information criterion[42] (BIC) were imported into ProteomicsDB.

2.3 Data retrieval and visualization

With minor exceptions, all data shown on the user interface of ProteomicsDB are requested from dedicated OData or XS services using Ajax (asynchronous JavaScript and XML). The response is served in JSON-format. Depending on the complexity of the underlying data, these services access attribute-, analytical- or calculation-views. Attribute views are simple database views without additional aggregation. Analytical views typically provide access to measures which allow additional aggregation. Most services, however, call calculation views which enable complex calculations (e.g. observed and theoretical sequence coverage or the annotation of experimental spectra) not possible without custom SQL-, L- and R-procedures. For data visualization, the JavaScript library *D3* was used.

3 Results

The main goal of ProteomicsDB is to expedite the identification of the human proteome and foster its use across the scientific community. For this, four main features are necessary: (i) data deposition and annotation, (ii) data access, (iii) online data exploration and (iv) online analytical tools supporting the cross-experiment comparison. The following four sections describe these features in more detail. Briefly, data deposition and annotation is possible via the integrated repository. The repository enables the structured storage and description of MS data, which is essential for proper data dissemination and integration. Multiple OData services can be used to retrieve data stored in ProteomicsDB. These services, described in the second section, can be accessed via the browser, programmatically or a custom light-weight Java application. The third section highlights the visualization of identification and quantification data and use-cases covered by the protein-centric visualization. Lastly, ProteomicsDB enables cross-experiment data comparison and combination via browser-based analytical tools. Three specific use-cases are implemented at the time of writing and described in the last section.

3.1 Repository

ProteomicsDB provides a simple repository for data storage and annotation of MS data. However, the repository is not only intended to enable access to the data. The annotation is essential for multi-experiment comparisons or other meta-analyses as it allows to set different experiments into context of each other or to interpret the data within the context of the experimental design.

Raw MS-files can be uploaded to ProteomicsDB once a project and an associated experiment has been created. While a project does not need any further annotations, the scope of the experiment has to be set, allowing different classification options such as full proteome, affinity purification or dose-dependent inhibition. After the initial creation of a project, the project's visibility is "private" and thus not visible to any other visitors. The owner can share a secure link with the public to allow others "read-only" access to the project. Additionally, the owner can modify the visibility by altering the status of the project to frozen, preventing further changes to the underlying data, and subsequently to public. After that, no further changes to either the data or the annotations are possible and the project is publicly accessible.

Two additional annotation layers exist besides the scope of the experiment. First, multiple raw MS-files can be assigned to samples within an experiment to allow both the combined expression estimation of pre-fractionated samples (one sample many *raw*-files) as well as separate expression estimation (one *raw*-file many samples, e.g. SILAC). A sample provides multiple basic annotations, such as digestion conditions, sample collection, biological origin and MS settings. However, especially in more complex experiments such as dose- or temperature-dependent assays, experimental factors such as the specific concentration, the temperature and the used inhibitors are necessary to correctly process, analyze and interpret the results. These information can be provided by adding an experimental design. Figure 2.26 illustrates this on a CETSA experiment (see Extended Data Figure A1 in the Appendix for a Kinobeads experimental design). As described earlier, conditions and treatments are organized into columns and rows in a matrix layout, respectively. An example of a treatment is "inhibition" which is described by a dose, a duration and the inhibitor (see Figure 2.26, first column). Conditions are manifestations of a group

of treatments, such as 5000 nM Dasatinib followed by heat treatment at 52 °C (see Figure 2.26, last row). Adding a sample to a specific condition is done by dragging and dropping an unassigned sample (Figure 2.26; list of samples) into the desired condition and biological replicate column. While biological replicates are manually added by the user, samples assigned to the same condition and biological replicate are considered technical replicates. The example shown in Figure 2.26 depicts two treatments with two biological replicates. With this annotation, ProteomicsDB is able to visualize the temperature-dependent effects of Dasatinib on any identified protein within this experiment and also allows to compare different experiments with each other.

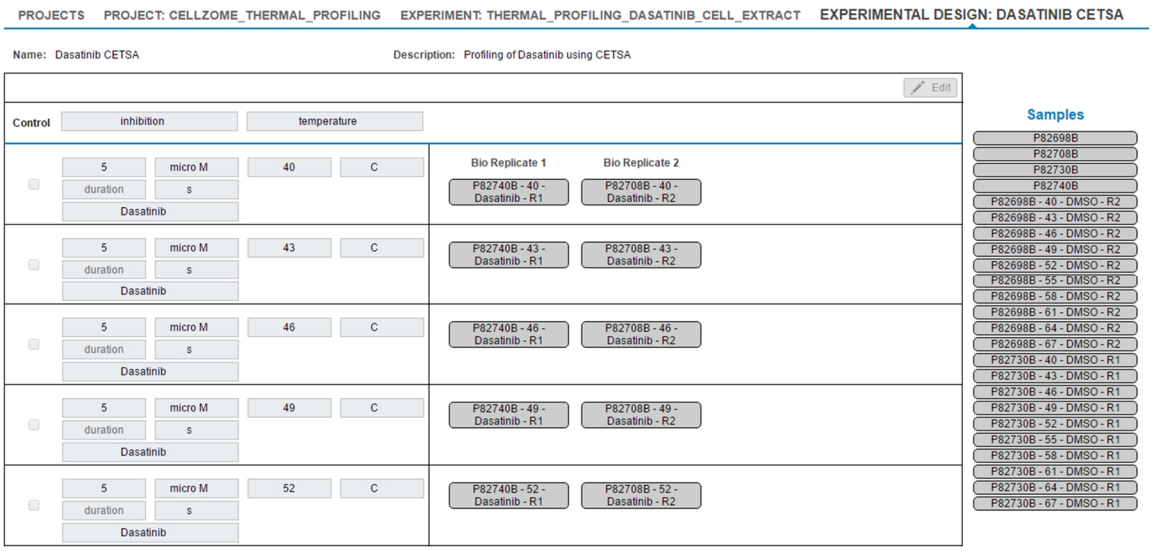


Figure 2.26 | Screenshot of an experimental design of a CETSA experiment from ProteomicsDB. The experiment consists of two treatments (inhibition and temperature). Each condition (e.g. 5 mM Dasatinib with a subsequent heat treatment of 40 °C) was measured with two biological replicates. All samples assigned to the experiment, but not yet used in the experimental design, are listed on the right-hand side and can be moved by drag-and-drop to specific conditions.

3.2 Data access

ProteomicsDB offers a wide variety of APIs using the OData specification for programmatic access to data. Listing 1 shows an example how to query the API `proteinpeptideresult` for all peptides (sequence, score and search engine via `$select`) weighing more than 1000 Da (by using `$filter`) of the protein Q92769 (`InputParams PROTEINFILTER`).

```
https://www.proteomicsdb.org/proteomicsdb/logic/api/proteinpeptideresult.xsodata/InputParams(P
ROTEINFILTER='Q92769')/Results?$select=PEPTIDE_SEQUENCE,SCORE,SEARCH_ENGINE&$filter=PEPTIDE_MA
SS gt 1000
```

Listing 1 | Example HTTPS OData request to **`proteinpeptideresult`**. The columns `sequence`, `score` and `search engine` are requested for protein Q92769 of peptides with a mass of >1000 Da.

In order to use the API, a user has to register with ProteomicsDB and be granted the necessary privileges. After that, the user can use the APIs via the browser or programmatically. Listing 2 briefly shows how to use an OData service in python. After specifying the type of connection and the username and password, a connection to an API via an URL can be opened. The response is either delivered in JSON or XML format.

```
import urllib2, urllib, httplib;
import base64

try:
    import ssl
except ImportError:
    print "error: no ssl support"

class ExampleAccess():
    def __init__(self, username, password):
        self.default_headers = { "Authorization" : "Basic %s" % base64.encodestring( "%s:%s" %
              ( username, password) ).rstrip('\n') }
        self.port = 443
        self.host = 'www.proteomicsdb.org'
        self.url = '''/proteomicsdb/logic/api/proteinpeptideresult.xsodata/...'''

    def connectAndRetrieve(self):
        hconn = httplib.HTTPSConnection( "%s:%d" % (self.host,self.port) )
        hconn.request("GET", self.url, headers = self.default_headers)
        resp = hconn.getresponse()
        print resp.status, resp.reason
        body = resp.read()
        print body
        hconn.close()

if __name__ == "__main__":
    USERNAME = "ProteomicsDBUserName"
    PASSWORD = "ProteomicsDBPassword"

    example = ExampleAccess(USERNAME, PASSWORD)
    example.connectAndRetrieve()
```

Listing 2 | Python example on how to use an OData service from ProteomicsDB.

Ten different services are currently available and described in more detail on the web interface. Briefly, to retrieve:

API#1: All peptide identifications for any protein

This service returns all peptide identifications for a given protein of all public projects. The results contain qualitative information, such as the q-value, PEP and score of the identification, as well as metadata, such as the scope of the experiment in which it was identified.

API#2: All peptide and protein identifications for any experiment

This service requires an experiment ID and returns a full list of identified peptides and proteins. As described, OData allows to add custom filters, which can be used to filter for example the q-value of the peptide identification.

API#3: The expression values for any protein

This API can be used to retrieve protein expression values (iBAQ or top3) for any protein of interest. Additional filters enable the selection of specific tissues and experimental scopes. Some of the filters are designed to prevent the direct comparison of protein expression values between e.g. MS1- and MS2-based quantification or full proteomes and affinity purifications.

API#4: The available quantification types for any protein

This service functions as a helper for API#5. The same procedure is used in the web interface to limit the number of check boxes for the calculation of the experimental proteotypicity.

API#5: The experimental proteotypicity for any protein

Taking into account the specified labeling techniques, this service returns both experimental proteotypicity and the experimental cumulative proteotypicity for a given protein. Both values are calculated in real-time.

API#6: All expression values in an available organ, cell line or body fluid

This service only takes full proteomes into account and returns a full list of proteins found to be expressed in the given biological system.

API#7: A list of all available organs, cell lines and body fluids

This service functions as a helper for API#6 and returns all available biological sources for which full proteomes are available.

API#8: All PSMs for any peptide

Given the peptide sequence of interest and a custom q-value filter, this web service returns a list of all PSMs available for this peptide. This includes, but is not limited to, scores, PTMs, retention time, mass error and the experiment they were observed in.

API#9: A comprehensive list of peptides identified for any protein

In contrast to API#8, this service aggregates peptides based on their sequence, charge and variable modification string. The best PSM is selected and chosen as a representative.

API#10: A list of all peptides identified in any organ, cell line or body fluid

Similarly to API#9, this service compiles a list of best PSMs per sequence, charge and modification string which have been observed in a given biological source, irrespective of the scope of the experiment.

As both the XML and JSON output is not designed to be imported into spreadsheet programs and not all users might have neither the experience nor the tools at hand to use the API to its full effect, the RequestTool, a standalone Java application (Figure 2.27) was implemented to simplify the access to the data. This application allows registered users to login and use the API. At startup, all currently available APIs including their required in- and output parameters are automatically retrieved from ProteomicsDB. The RequestTool enables users to download data in csv output, however, additional writers can be added easily by attaching new modules. Furthermore, an input parameter syntax and a first-in-first-out request queue allows users to specify batch jobs which either generate a single csv per job or append the output to a single file. This enables the automatic retrieval of data from hundreds or thousands of proteins.

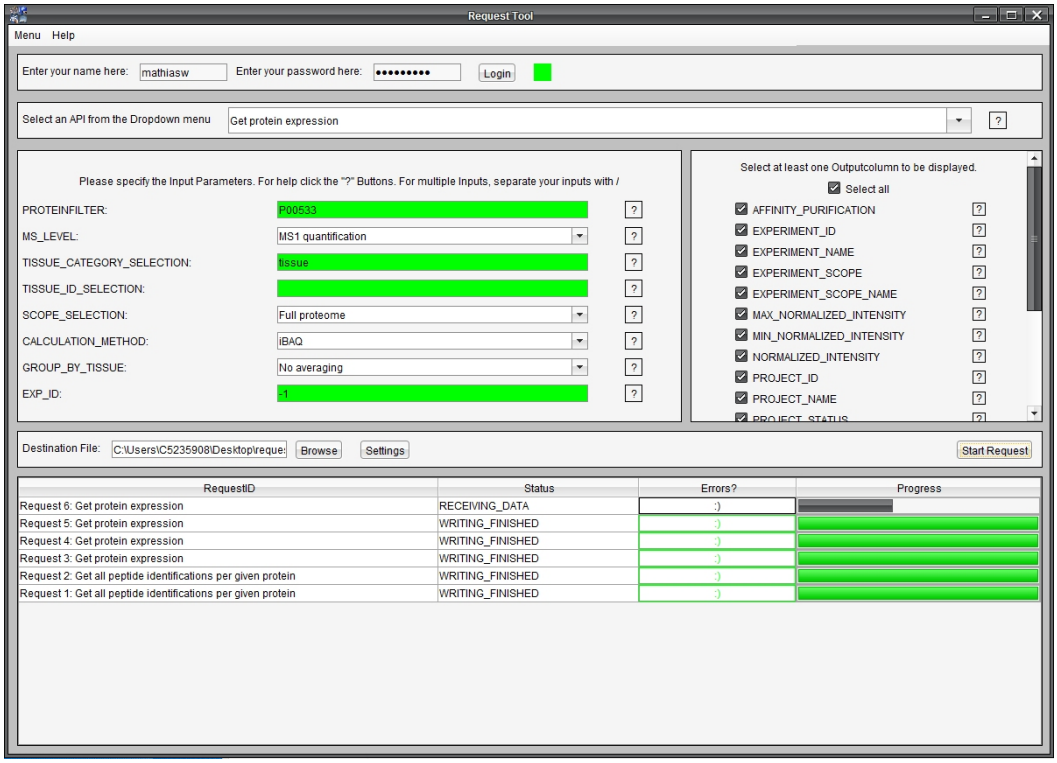


Figure 2.27 | RequestTool to access the ProteomicsDB API. The RequestTool allows any ProteomicsDB user to download data and store data in a user-friendly and simplified process. At startup, this application connects to ProteomicsDB and retrieves all currently available APIs including their in- and output parameters.

3.3 Protein-centric visualization

ProteomicsDB is designed to enable researchers the quick interrogation of available identification and quantification information on single proteins. For this purpose, ProteomicsDB provides views and tools to check and manipulate available data in a protein-centric presentation. Different use-cases are organized into tabs, which are described in more detail in the following subsections.

3.3.1 Protein Summary

The starting page of every protein shows a brief summary (see Figure 2.28). This includes the number of peptides which have been detected (shared and unique on either gene or protein level), the sequence coverage and some basic annotations such as GO terms, chromosomal location, external links and evidence status. If high quality unique (gene or protein level) peptides have been identified in any of the imported projects, the evidence traffic light for this protein is green. Yellow indicates that the best peptide identification for this protein is of poorer quality and the identification of the protein is questionable. If no unique peptide match exists for this protein, or its evidence likely is a false match, this protein is marked red. Additionally, the protein sequence coverage which depicts all identified peptides is shown. Similar to genomic sequence alignments, overlapping peptides are drawn in a stacked manner. Aligned to this, a visualization of the protein domain structure and observed PTMs is shown underneath. This quickly enables users to see which part of the protein is covered by observable peptides (e.g. here the tyrosine kinase catalytic domain; TyrKc). This is especially important for researchers who are interested in specific peptides covering e.g. regions of PTMs or mutations of a protein to build targeted assays.

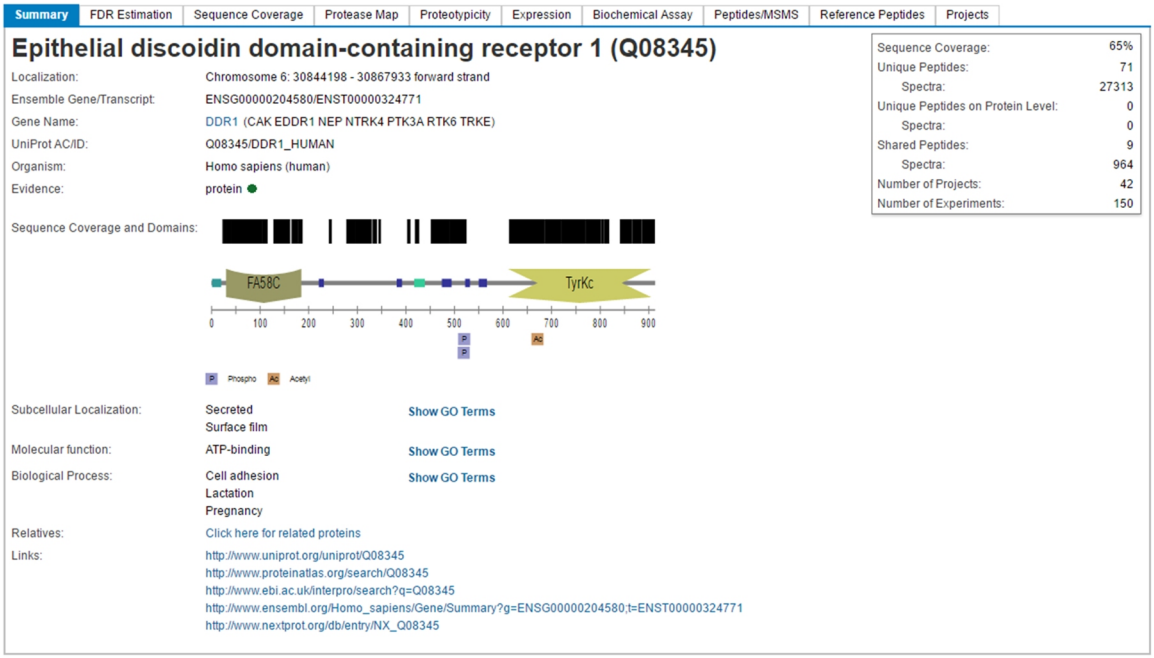


Figure 2.28 | Screenshot of the DDR1 protein summary page from ProteomicsDB. This view depicts general information about the protein, such as gene name, chromosomal location, GO annotation and external links, as well as a brief overview of the MS-evidence, such as number of unique peptides and sequence coverage, stored in ProteomicsDB. The graphical representation of the protein includes the MS-accessible primary structure, the annotation of known and predicted domains and sequence features, as well as observed PTMs and their location.

3.3.2 Sequence coverage

This tab shows the entire sequence of the selected protein including properties such as mass, length and coverage. Amino acid stretches which are supported by the identification of a peptide are marked in red whereas black stretches are not yet covered by confident peptide identifications. This tab is designed to check whether certain peptides are MS-accessible to enable e.g. the identification of SNPs or PTMs.

3.3.3 Protease Map

This tab offers an *in silico* digest of proteins. After selecting the proteases of interest, the user can filter the resulting peptides by number of missed cleavages, length and mass. Each selected protease is visualized with a sequence coverage pattern plot, which shows all possible peptides passing the selected filters. Even though each protease is considered separately, the table underneath the plots also shows the combined maximum expected sequence coverage when using multiple proteases in separate digestion and analysis steps. This view is designed to guide users who are interested in maximizing the sequence coverage of a single protein or who are interested in knowing which protease to choose when a particular area of the protein is of interest.

3.3.4 Proteotypicity

This view is designed to guide researchers which are interested in directed or targeted acquisition methods such as SRM/MRM or PRM assays by ordering peptides according to their potential for reproducible and consistent identification and quantification. For this purpose, peptides are sorted by their experimental proteotypicity. The experimental proteotypicity of a peptide is the ratio of the number of experiments in which this peptide was identified to the number of experiments in which the protein was identified. High values indicate good MS sample preparation and acquisition accessibility. However, some multiplexing techniques strongly influence the detectability of peptides, thus the user can choose between different labeling approaches in which this protein was identified. After selecting one or multiple labeling strategies, the table depicts the experimental proteotypicities. In addition to this, the cumulative proteotypicity predicts how many peptides are necessary in order to reproducibly identify a protein with a certain probability in unknown samples.

This view strongly benefits from using an in-memory database. Pre-calculating all possible combinations of options will result in a large overhead, which has to be stored. ProteomicsDB calculates the experimental proteotypicity in real-time on request. In principle, this allows the incorporation of even more options, such as user defined PSM or peptide FDR cutoffs, charge states or biological sources without losing performance.

3.3.5 Expression

ProteomicsDB is designed to enable the comparison of thousands of samples from a wide range of biological sources such as tissues, body fluids and cell lines. The expression tab allows researchers to explore the expression of a single protein in human samples and cell lines and further allows to trace its expression down to single samples. Consequently, a user is able to infer the conditions under which this protein was identified.

The expression view is built out of two major components for data selection (Figure 2.29a) and visualization (Figure 2.29b-d). In order to enable proper cross-experiment comparison of expression values, only data from similar sources can be selected. As MS1 and MS2 quantifications techniques cannot be compared directly, the filters only support the selection of either type. Likewise, the comparison of full proteome data (unbiased expression analysis) with affinity type experiments (biased expression analysis) is not possible and thus does not allow a direct comparison. In addition, the user can choose between different biological sources and the expression estimation technique.

The data visualization is composed of three interactive connected elements: (i) a heatmap-like body map (Figure 2.29b), (ii) a cell type aggregated bar chart (Figure 2.29c), and (iii) a sample specific bar chart (Figure 2.29d). The body map depicts the median expression of a protein in the human body. A linear color gradient from green to red is used to map low and high abundant proteins to the tissue of origin. The cell type aggregated bar chart shows the median and, if replicate measurements are available, the minimum and maximum expression of a protein in the corresponding tissue, fluid or cell line. The sample-specific bar chart visualizes the expression of a protein without aggregation across multiple selected biological sources.

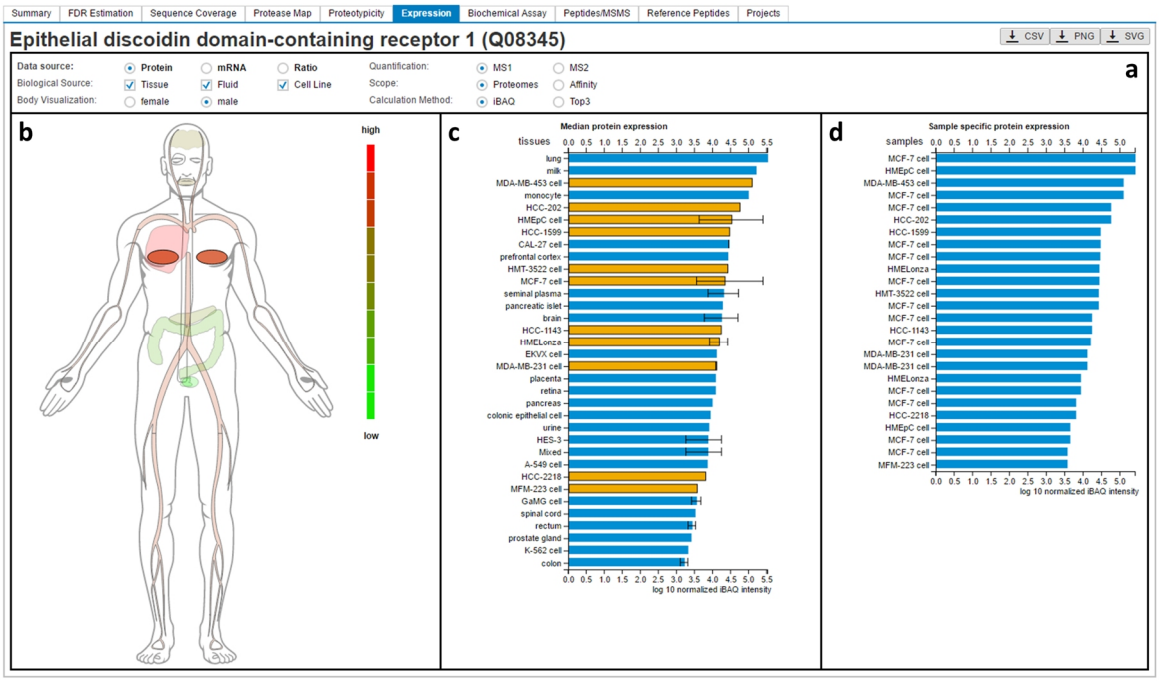


Figure 2.29 | Screenshot of the protein expression of DDR1 from ProteomicsDB. a, The data selection interface enables users to select the data source (protein, mRNA or ratio), biological sources (tissues, fluids and cell lines) and depending on the choice of the data source, different quantification and calculation methods. b, The median expression of a protein is superimposed in a heatmap-like representation (high and low expression depicted in red and green, respectively) onto the human body. The protein expression values of cell lines are mapped to the tissue of origin using the BRENDA tissue ontology. c, The cell type aggregated bar chart depicts the median and, if replicate measurements or multiple studies on the same cell type are available, the minimum and maximum observed expression of a protein. Highlighted (orange bars in c and opaque tissue in b) are all biological sources originating from human breast tissue. d, The sample-specific bar chart visualizes the expression estimates from specific measurements. A click onto a specific bar in the representation opens a popup showing basic sample preparation and acquisition parameters from the experimental design. The body map and the two bar charts are connected, thus selecting a specific tissue in the body map will both highlight all corresponding cell types as well as open the sample-specific bar chart, and vice versa.

These visualization are connected to each other, enabling the interactive exploration of expression patterns. For example, selecting colon in the body map will highlight all bars in the cell type aggregated visualization (including cell lines) which originated from colon. Selecting (by clicking) one or multiple biological sources (bars) in the aggregated visualization will both highlight the tissue of origin in the body map as well as trigger the sample specific visualization to show all available single measurements for the selected cell type. Additionally, a brief summary of the sample preparation details can be opened by clicking a bar in the sample specific bar chart.

The generic implementation of the expression tab and the versatility of ProteomicsDB also enables the storage and visualization of other omics data sources, such as RNASeq data[43]. During data selection, the user can choose mRNA as the primary data source and thus explore the expression of mRNA across the human body using the same mechanism as described for proteins.

3.3.6 Biochemical Assay

To showcase the potential of ProteomicsDB besides visualizing expression values estimated from full proteome data, three visualizations of data originating from biochemical assays used to study protein-drug interactions were implemented. The biochemical assay tab shows a comprehensive overview of all Kinobeads and CETSA experiments stored in ProteomicsDB.

Starting with Kinobeads experiments, the biochemical assay tab shows all available competition binding curves available in ProteomicsDB. These (Figure 2.30) can be filtered by adjusting the sliders for the EC_{50}, R^2 (R2) and BIC. Depending on the protein and the selected filters, the table will highlight small molecules which show a dose dependent effect. The experimental data is plotted using black circles, whereas the blue line shows the fitted dose response curve. The orange error bar indicates the uncertainty (one standard error) associated with the estimation of the EC_{50} resulting from the curve fit. Deactivating all filters will list all small molecules profiled with Kinobeads and also visualizes experiments with no dose-dependent effect. A direct link to DrugBank[44] (Inhibitor) as well as the experimental and experiment design (Repository) are provided in the table.

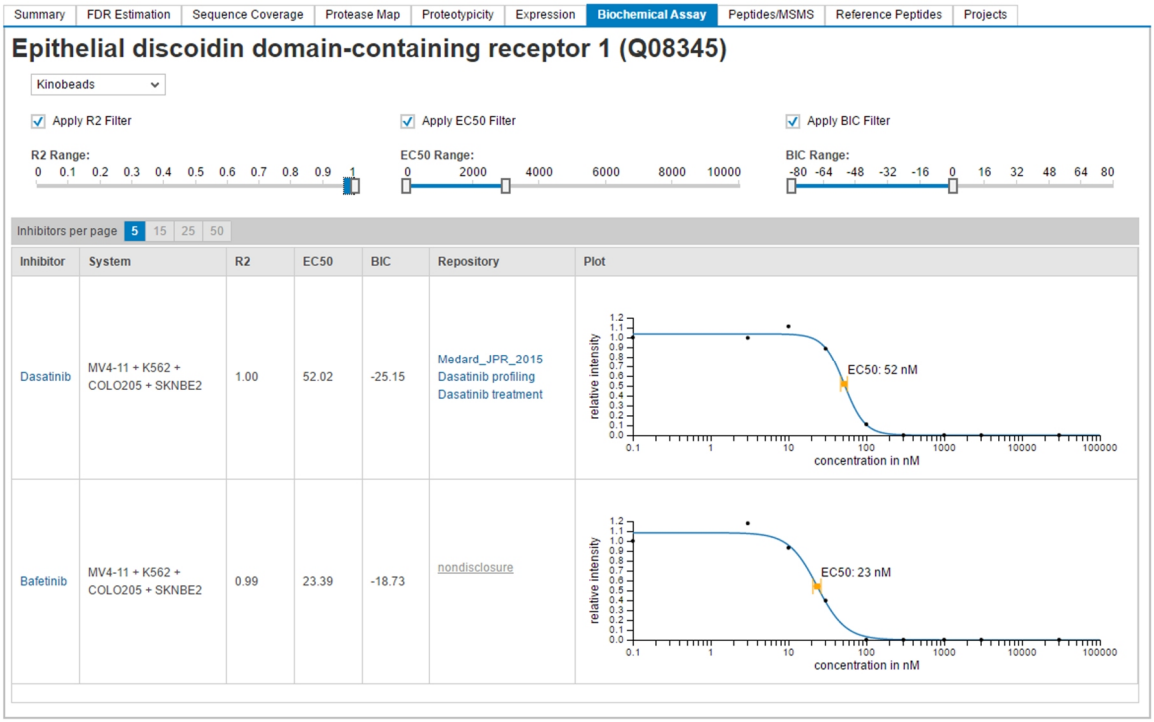


Figure 2.30 | Screenshot of the biochemical assay tab – Kinobeads – from ProteomicsDB. Here, DDR1 is shown which shows a clear inhibition upon treatment with Dasatinib (EC_{50} 52 nM) and Bafetinib (EC_{50} 23 nM). The underlying data as well as the curve fit are stored in ProteomicsDB to allow cross-experiment comparison and analysis.

Two similar views are available for CETSA experiments. First, selecting the option "Melting proteome" shows the melting behavior of the selected protein in CETSA vehicle control experiments (mostly DMSO). This view can be used to investigate whether a protein was ever observed in a CETSA experiment and enables researchers to estimation the CETSA accessible proteome. The second view (see Figure 2.31) shows CETSA vehicle control and drug experiments, similar to Kinobeads. The table lists the difference in the melting point, as well as the mean R^2 of the control and treatment curve. Major shifts in the melting behavior (large Delta T_m) indicate a

direct or indirect protein-drug interaction. Different replicates, annotated in the experimental design, are differentiated by the line type (e.g. solid vs dashed).

Dasatinib was initially design to target BCR-ABL, SRC, Ephrins and GFR[45] for the treatment of chronic myelogenous leukemia. However, both the Kinobeads (Figure 2.30) as well as the CETSA assay (Figure 2.31) independently support the hypothesis that Dasatinib binds to DDR1 suggesting that DDR1 is an off-target of Dasatinib. Additionally, the Kinobeads experiment provides an estimate of the potency of the interaction with an EC_{50} of 52 nM. Kinobeads and its CETSA equivalent, the isothermal dose response (ITDR), are both supported in ProteomicsDB, however, no ITDR data is available yet.

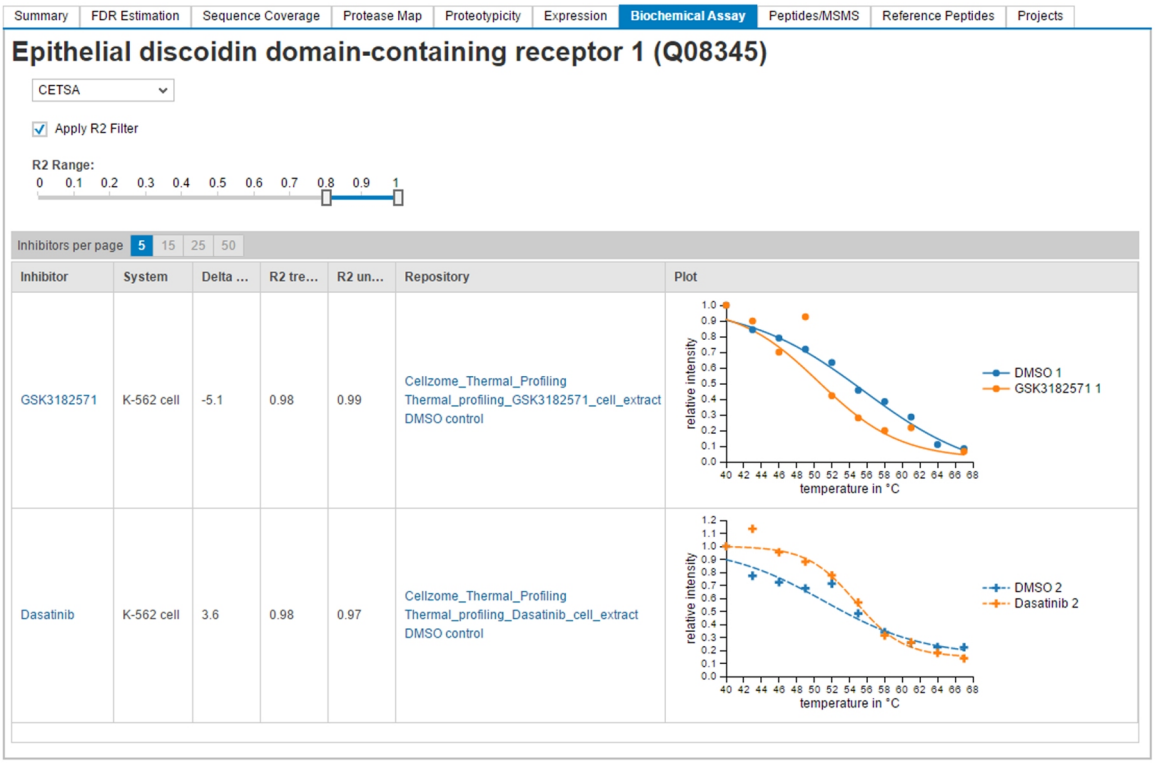


Figure 2.31 | Screenshot of the biochemical assay tab – CETSA – from ProteomicsDB. Here, DDR1 is shown exhibiting a stabilizing effect upon Dasatinib treatment, further validating the observation that Dasatinib binds to DDR1 from Figure 2.30.

3.3.7 Peptides/MSMS

The peptides/MSMS tab is dedicated to show all available peptide evidence for the selected protein. The initial view lists all observed peptides including meta data such as mass, length, uniqueness and the number of observations, as well as its identification score, q-value and PEP. Each spectrum used for inference can be visualized in ProteomicsDB using the built-in spectrum browser (Figure 2.32). Selecting a peptide of interest opens an overlay which lists all available PSMs for this peptide (Figure 2.32a). Similar to the peptide overview, this table shows all available information, such as individual search engine score, q-value, PEP, modifications and source. By default, the PSM with the lowest PEP is selected and its spectrum is displayed below the table (Figure 2.32c). The spectrum viewer offers various different options for configuration and selection (Figure 2.32b). Starting with the annotation of the fragment ions, the spectrum viewer

can be configured to show or hide all major fragment ions as well as specific neutral losses. By default, the expert system annotation is used. The result is visualized in the middle panel (top spectrum). Fragment ions are color coded and their annotations are shown above. A mass deviation plot underneath the spectrum can be used to judge the correctness of the annotation.

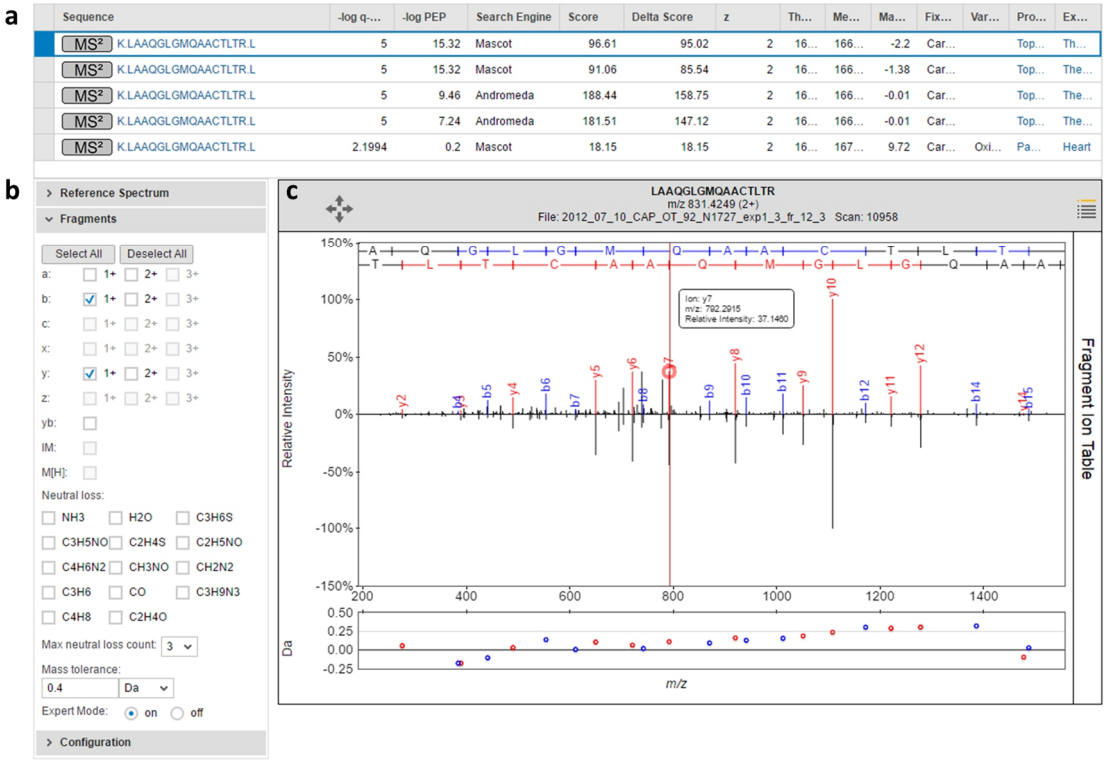


Figure 2.32 | Screenshot of the fragment spectrum of the peptide LAAQGLGMQQCTLTR with the integrated spectrum viewer from ProteomicsDB. a, This table lists all available PSMs stored in ProtoemicsDB and provides links to the respective projects and experiment in the repository. b, The annotation of fragment ions can be controlled by (de)-selection of the respective options in the "Fragments" sub menu. Furthermore, the "Reference Spectrum" submenu can be opened to choose a different reference spectrum for visualization and the "Configuration" submenu enables the configuration of visualization options. c, The spectrum viewer visualizes the observed (top spectrum) and, if available, a reference spectrum (bottom spectrum) of a synthetic peptide in a mirror view.

An integrated feature of the spectrum viewer is the mirror representation of a reference spectrum (bottom spectrum) if available. These spectra originate from e.g. synthetically generated peptides which were measured separately and can be used to validate the identification of the peptide and thus protein. In case a reference spectrum is available, the highest scoring PSM is chosen matching to the precursor and modification status of the selected PSM. As depicted in Figure 2.32, the experimental spectrum identified as the peptide LAAQGLGMQAACTLTR (top spectrum) matches the reference spectrum (bottom spectrum) in terms of fragment masses and their relative intensities. Further reference spectra for comparison can be chosen by opening the "Reference spectrum" tab on the left. While the observed spectrum of LAAQGLGMQAACTLTR is almost completely annotated, leaving none of the most intense peaks unexplained, this is not always the case. Especially for poor quality or chimeric spectra this feature can increase the confidence for a true positive match even though the search engine score would suggest otherwise. In contrast to regular database search approaches, this view allows the comparison of relative fragment intensities.

3.3.8 Reference peptides

As described in the previous section, ProteomicsDB also stores reference spectra for some peptides. All reference spectra are accessible via the "Reference peptides" tab of a protein. This representation is similar to the "Peptide/MSMS" tab and each spectrum can be viewed. Opening a spectrum of a reference peptide again depicts the spectrum with respect to other available reference spectrum, as it uses the same implementation.

3.3.9 Projects

The "Projects" tab lists all projects and experiments in which the selected protein was identified, providing links to the repository (if set to public) and shows the number of peptides and PSMs which led to the identification as well as sequence coverage and evidence status of the identification.

3.4 Browser-based analytical tools

On major application of ProteomicsDB is its integrated solutions for multi experiment comparisons. To highlights these capabilities, three use cases were implemented. First, an expression analysis over multiple experiments utilizing the highly interactive heatmap implemented in ProteomicsDB. Second, an application which can be used to determine the most suitable drug for a given target. Last, the extension of the latter to the exploration of the dose-dependent effects of on- and off-targets using multiple inhibitors.

3.4.1 Analysis of protein sets

The comparison of expression profiles across different tissues, fluids and cell lines can give rise to new hypothesis and biological applications. While the expression tab of a single protein allows the analysis of expression patterns over multiple biological sources, it does not enable to analysis of expression of multiple proteins. This analysis is visually supported by the presentation of data in a heatmap showing proteins and biological sources as rows and columns, respectively.

For this purpose, any list of gene names or Uniprot identifiers can be supplied to ProteomicsDB. Similar to the expression analysis of a single protein, different options exist to filter the results based on biological sources, quantification methods and calculations.

The heatmap shown on ProteomicsDB supports zooming (mouse wheel) and panning (drag). If the resulting number of proteins or biological sources is too large, the names of one or both are hidden and only shown once the width or height of the cells is large enough to print the names. Two dendrograms show the results of the hierarchical clustering in both dimensions. One or multiple sub-trees can be selected to either remove the respective rows or columns, or to perform a GO enrichment analysis using DAVID[46,47]. For large heatmaps, the Ctrl-modifier can be used to directly zoom into a specific sub-tree of the heatmap.

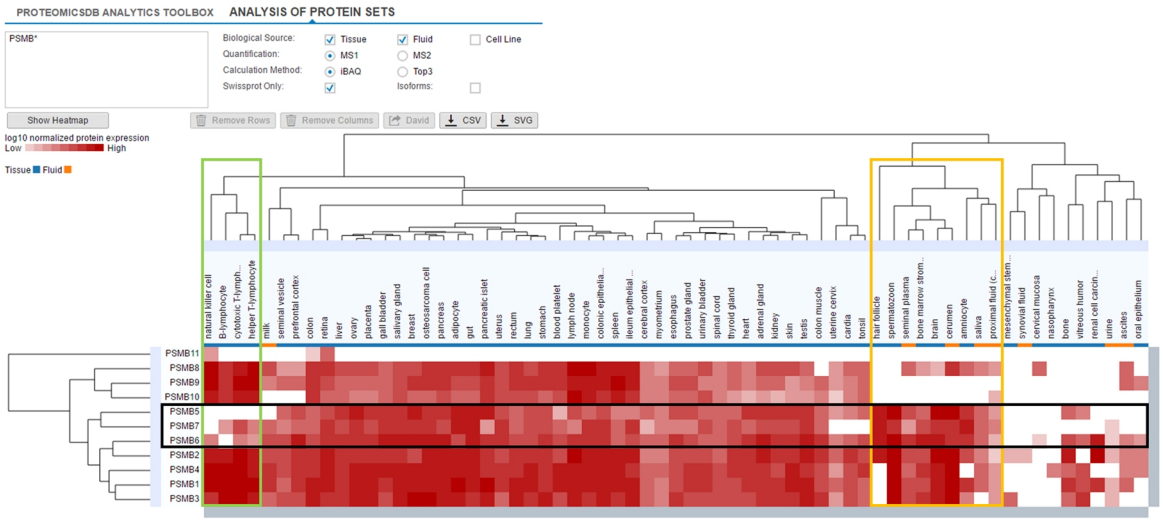


Figure 2.33 | Screenshot of the protein expression heatmap of all major components of the proteasome complex from ProteomicsDB. The online heatmap allows to search for a custom set of proteins (here PSMB*). Different choices of biological sources, calculation methods and protein subsets exists. An online clustering of the results is performed to help visualize differences in expression patterns. The heatmap provides multiple interactive functions, such as zooming, panning, removing of rows and columns and a link to DAVID to perform a GO enrichment analysis. The raw data, as well as the heatmap can be exported as *svg* or *png* files. The subunits PSMB 5-7 (highlighted in black) are found in the constitutive proteasome, while PSMB 8-10 can be found in the immunoproteasome. This is highlighted for immune cells (green) and tissues or fluids expressing mostly the constitutive proteasome (orange).

Figure 2.33 shows the resulting heatmap when searched for "PSMB*" in tissues and body fluids (see Extended Data Figure A2 in the Appendix for a heatmap containing all biological sources). These proteins are the major components of the regular and induced proteasome. Three major observations are directly visible: First, the immunoproteasome (induced) is highly and almost exclusively expressed in immune related tissues (natural killer cells, B-lymphocyte, cytotoxic T-lymphocytes and helper T-lymphocytes; Figure 2.33 green box). This is visible by the lack of the PSMB 5-7 subunits (Figure 2.33 black box), which are replaced by subunits PSMB 8-10. Second, while only some tissues seem to solely express the constitutive (non-induced) type (e.g. hair follicle, spermatozoon; Figure 2.33 orange box), most of the other tissues seem to express both types of the proteasome, with some differences in relative amount (e.g. lymph node vs cerebral cortex). Third, no or only some subunits of the protease are detected in the tissues and fluid on the right. While this observation can be confounded by a large dynamic range of protein expression, their lack or very low abundance is expected (e.g. synovial fluids and urine).

The heatmap is able to show the expression of hundreds of proteins in hundreds of tissues. The example shown in Extended Data Figure A3 in the Appendix displays the expression patterns of all proteins which contain the word "kinase". The interactive zoom and panning enables the identification of interesting features, such as the small cluster in the middle of the heatmap. This cluster contains exclusively neuronal tissues (brain, retina, spinal cord, prefrontal cortex). The 15 proteins, such as the protein kinases CAMK2A, included in this cluster are associated with learning and memory. However, not all of these proteins are currently known to play a major role in these processes. Their co-expression with proteins of known function could be an indication of their function and relevance.

3.4.2 Inhibitor potency/selectivity analysis

On common application in both research and clinics is to find the most selective and potent drug against a target of interest. This use case is implemented in ProteomicsDB and allows the comparison of multiple small molecules which share a common user-defined target.

Starting with the selection of the target protein, the user can additionally filter the dose-dependent models for the EC_{50} range of the targets, the R^2 and BIC (Figure 2.34a). The pEC_{50} ($-\log_{10}$ EC_{50} in nM) distribution of all targets meeting the filter criteria for each drug showing a dose-dependent effect on the selected target are plotted in separate violin charts (Figure 2.34b). Additionally, the red marker indicates the EC_{50} of the selected target protein for each drug. The number of targets with smaller or larger EC_{50} with respect to the target protein is shown below and above of the red marker, respectively. This representation visually aids the user to identify the most selective and/or potent drug currently available in ProteomicsDB.

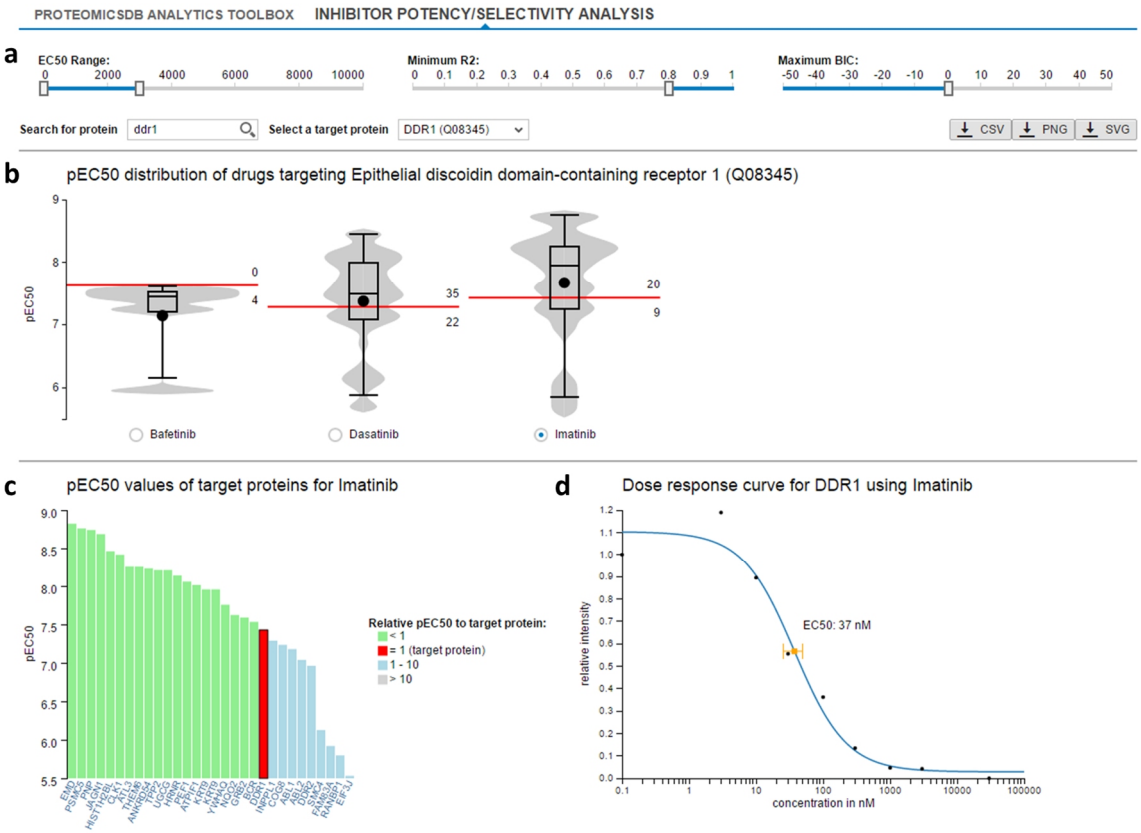


Figure 2.34 | Screenshot of a potency analysis of drugs targeting DDR1 from ProteomicsDB. a, At the time of writing, three kinase inhibitors showed a dose-dependent effect below ~3000 nM with an R2 of >0.8 and BIC < 0 on DDR1. b, The three pEC_{50} distributions of all protein-drug interactions passing the filter criteria is shown using violin charts. The red line highlights the pEC_{50} of DDR1 in the respective drug. The number of target which show a higher (number above red line) and lower (number below red line) pEC_{50} indicates the selectivity of the drug. c, Upon drug selection (radio button in b), the entire target space of that drug is shown (bottom left panel). Targets with a lower pEC_{50} in comparison to the selected protein (here DDR1, red bar) are shown in green. Targets with an EC_{50} up to 10-times of the selected protein are highlighted in blue. All other targets are shown in gray. Individual protein-drug interactions can be further investigated by selecting a specific bar. d, The corresponding dose-response curve is loaded. The orange error bar indicates the uncertainty of the EC_{50} estimation associated with the curve fit.

For additional analysis, each drug can be selected separately (Figure 2.34b radio button). After selection, a third layer (Figure 2.34c and d) visualizes all targets of that drug and, additionally, the dose-dependent residual binding curve of a protein. Within the barplot, proteins with a lower EC_{50} than the selected target protein are highlighted in green, while proteins with a higher EC_{50} are drawn in blue and gray if their EC_{50}s are at most 10-times higher or more than 10-times higher, respectively.

Figure 2.34 shows the results when searching for "DDR1". Three drugs, namely Dasatinib, Bafetinib and Imatinib, show a dose-dependent effect on DDR1. Both, Dasatinib and Imatinib are fairly unselective as 35 and 20 targets have a smaller EC_{50} than DDR1, respectively. This renders both kinase inhibitors not suitable for selective inhibition of DDR1. Bafetinib, however, is a promising candidate for a selective inhibition of DDR1 with an EC_{50} of 23 nM and the smallest number of off-targets (Figure 2.35). Bafetinib is a dual BCR-ABL/LYN inhibitor and was originally designed against Imatinib resistant chronic myeloid leukemia[48,49] and is awaiting clinical approval.

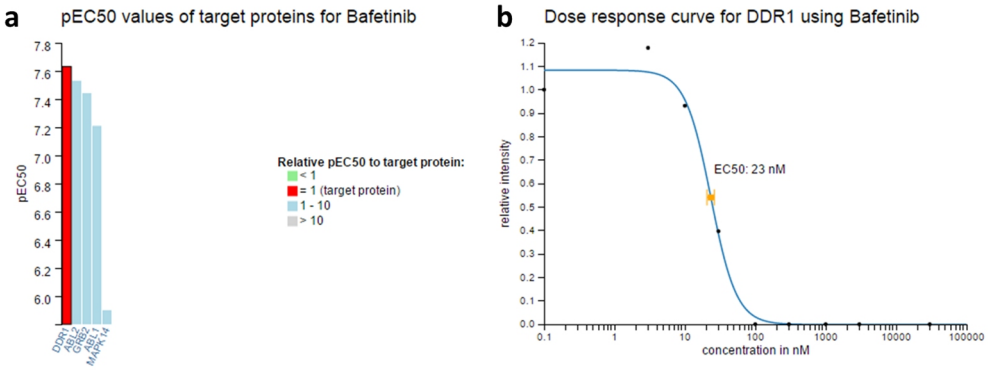


Figure 2.35 | Screenshot of the target space of Bafetinib with respect to DDR1 from ProteomicsDB. a, The drug profiling of Bafetinib shows that DDR1 is the most potent target. b, The binding curve of DDR1 shows a clear dose-dependent inhibition with an EC_{50} of 23nM.

3.4.3 Dose-dependent protein-drug interaction analysis

The potency analysis provides an interface to select an inhibitor for a given target. However, in some applications, e.g. to suppress resistance formation, targeting multiple proteins can lead to a more effective treatment. The dose-dependent protein-drug interaction analysis provides an interface to explore the predicted dose-dependent effects of multiple drugs on multiple proteins to enable the selection of the most promising drug-combination to inhibit a set of proteins with the least amount of off-target effects. Two views are available which show the predicted target profile of the selected drugs at a certain dosage as i) a protein-drug interaction graph and ii) a table showing the predicted inhibition effects. Both views are based on the inhibition/competition curves stored in ProteomicsDB.

The "Target Proteins" search field accepts sets of protein names, accession numbers or keywords in a semicolon, comma, tab or line break separated format (Figure 2.36 top panel). On this basis, all drugs which show at least one inhibitory effect on one of the proteins are taken into consideration. Alternatively, the "Target Drugs" search field can be used to manually select a set of drugs. In case both fields are used, the union of all drugs, either inhibiting at least one of the target proteins or being selected manually, is used.

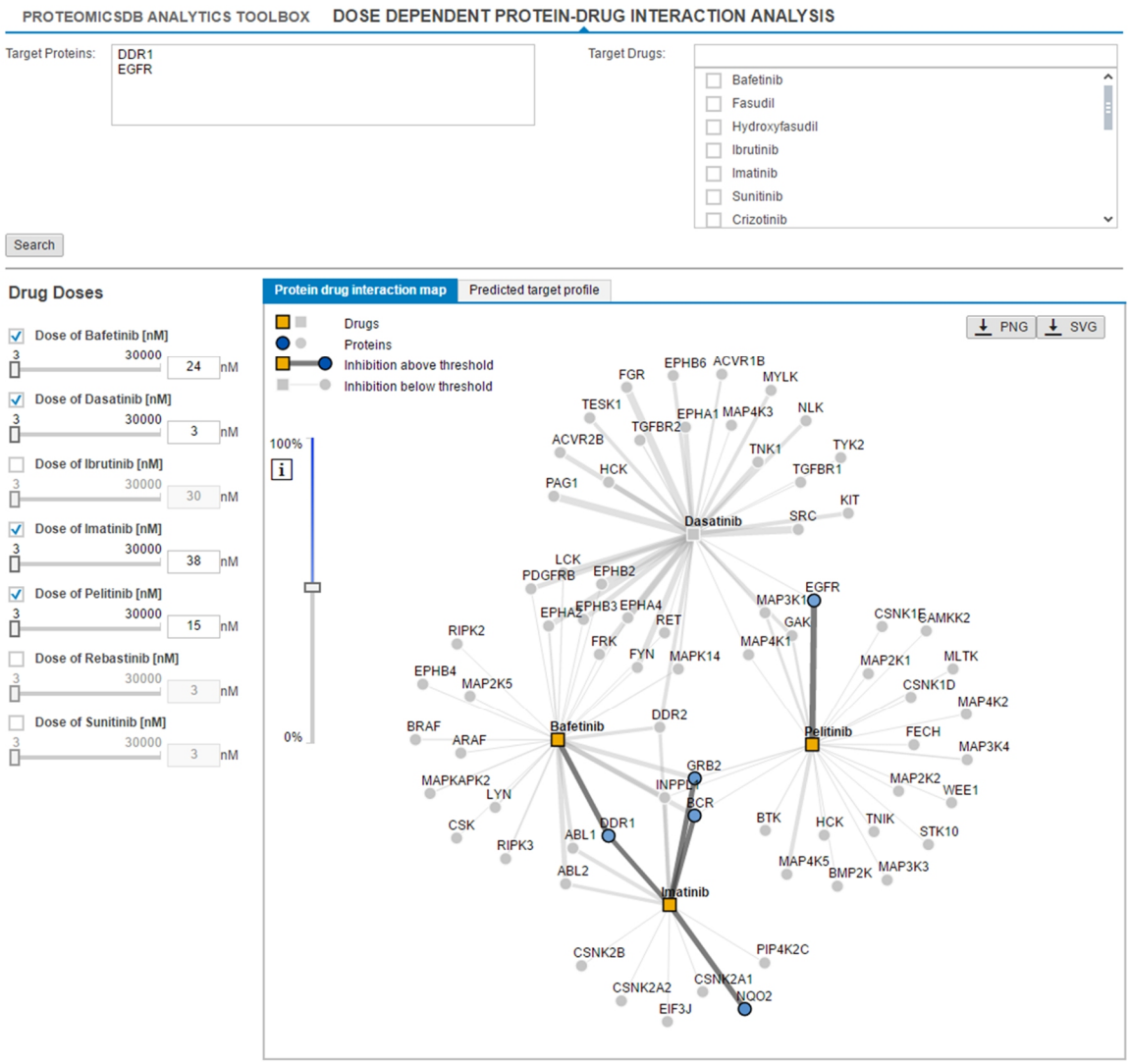


Figure 2.36 | Screenshot of a dose-dependent protein-drug interaction map from ProteomicsDB. The protein-drug interaction landscape of four selected (checkbox) kinase inhibitors (squares) which target either DDR1 or EGFR enables the visual inspection of inhibitors and their combination. Edges, indicating a dose-dependent effect of a drug on a protein, highlighted in dark gray pass the inhibition threshold (vertical slider). The corresponding proteins (circles) are highlighted in blue if inhibited or gray.

Each drug selected for the analysis is displayed on the left hand side of the view (Figure 2.36 bottom left panel). The checkbox can be used to disable (hide) a drug from both views. Additionally, the dosage of each drug can be adjusted by moving the slider or by manually entering a desired drug-concentration. The predicted inhibition of a particular protein in both the graph and the table view are updated in real-time based on the given concentration of a drug. The calculation uses the dose-dependent model fit (inhibition curve) stored in ProteomicsDB.

The graph-view shows the protein-drug interaction landscape of the selected drugs (see Figure 2.36 bottom right panel). Proteins (circles) are connected to drugs (squares) if a binding/inhibition curve is available for this combination. The strength of the edge (width) indicates the EC_{50} of the interaction, with potent interactions being represented with stronger edges. The vertical slider can be used to define a threshold to filter inhibitory effects based on the predicted inhibition. If an inhibition exceeds the selected threshold, the corresponding edge between a drug and proteins is highlighted in dark gray and the protein is colored in light or dark blue corresponding to a low or high inhibition. If all protein-drug interactions of a drug fall below the selected threshold, the

drug (square) is also colored in gray to indicate no major effect. In cases where proteins have multiple drug interactions, the strongest binding/inhibitory effect is chosen as the representative value.
Multiple interactions in the graph-view are possible. Each node of the graph (proteins and drugs) can be moved by drag-and-drop to enable manual disentanglement of the graph. Zoom, using the mouse-wheel, and panning, drag-and-drop on the background, are also enabled by default. Clicking on an edge between a protein and drug sets the dosage/concentration of that drug to the EC_{50} of this interaction.
The predicted target profile can also be explored in the table view (Extended Data Figure A4 in the Appendix). The table lists all proteins and drugs present in the graph and shows the EC_{50} and predicted effective inhibition in percent based on the concentration of the drug. Clicking a column-header allows to sort this table based on these values. For each protein, the column representing a drug with the strongest effect is highlighted with a blue background.

The example provided in Figure 2.36 shows the protein-drug interaction map of the kinase inhibitors Bafetinib, Dasatinib, Imatinib and Pelitinib, which target either DDR1 or EGFR. The inhibitors Ibrutinib, Rebastinib and Sunitib are hidden. Except for the multi kinase inhibitor Dasatinib, the inhibitor concentration of each drug is set to the EC_{50} of DDR1 or EGFR. Based on this, an effective combination of drugs to inhibit DDR1 and EGFR is Pelitinib and Bafetinib at, respectively, 15 nM and 25 nM. At these concentrations, no additional (off) targets show an inhibition greater than 50%. Replacing Bafetinib by 38 nM Imatinib however will result in the addition of GRB2, NQO2 and BCR (BCR-ABL fusion), as indicated by the dark gray edges.

4 Discussion

Here, a novel database, termed ProteomicsDB, combining identification and quantification data from mass spectrometry-based proteomics experiments is described. While multiple databases and repositories exist for either data type, only some provide comprehensive information on both levels[1]. Especially the introduction of a programmatically accessible experimental design in ProteomicsDB enables data interpretation in context of the treatment conditions and thus allows to build complex models based on experimental data. This was shown on dose- and temperature-dependent assay data to elucidate the target space of small molecules. With all this, multi experiment comparisons are possible and allow comprehensive data analysis of e.g. expression patterns over multiple studies or combined treatment effect prediction when using multiple small molecules. The generic database design of ProteomicsDB facilitates the storage of all conceivable fractionation and labeling techniques applied in proteomics experiments. Furthermore, extension to other omics-data is possible and was shown for RNASeq data. The web interface was built using modern JavaScript frameworks for HTML5. On three online real-time analytical use-cases it was shown that ProteomicsDB permits the integrated data analysis and exploration of large-scale data without reprocessing and manual annotation.

ProteomicsDB has been implemented using SAP's HANA in-memory computing technology. The system currently is backed with 50 TB hard-drive storage, 2 TB main memory and 80 processing units. A direct interface to the programming languages L, C++ and R allows calculations not possible with standard SQL and thus further broadens the applicability and possibilities to analyze and integrate multiple studies. This will facilitate the development and implementation of published or novel algorithms on open scientific questions such as false discovery estimation or protein abundance estimation. Especially the development of novel large-scale data analysis tools will benefit from the in-memory storage of all data since computationally intense steps do not rely on slow hard-drive read and write operations and, hence, allows quick adjustments to the underlying algorithms.

However, due to the constant increase of data both in terms of number of studies as well as raw storage, any database system storing results will come to its limits. ProteomicsDB currently stores all data, from the annotations down to single de-convoluted tandem mass spectra. Driven by the complexity of each data type, spectra, identification data and quantification information on peptides and proteins require vastly different amounts of storage and memory capacity. The acquired MS/MS spectra are in most use-cases of inferior importance and yet, especially due to its redundancy caused by the DDA approach, make up to 60% of the data. The development of novel spectral compression algorithms, storage of representative spectra resulting from spectral clustering and spectral libraries and the usage of modern distributed storage and computation systems will decrease the storage and memory footprint and thus provides alternatives for large-data scalability. In contrast, the most valuable information, which is peptide and protein quantification values, only account for 10% and 1% of the data, respectively. While the storage of expression values can be realized in classical database management systems without a significant drop in performance, access to billions of identification and spectral information in the main memory allows real-time calculations on more complex questions, such as the experimental proteotypicity or transition interference.

Proteomic experiments offer a huge variety of options with regard to experimental design, acquisition method, data processing, filter criteria and analysis methods. To enable meaningful data comparison, a uniform analysis pipeline is mandatory as each processing step (search engine, quantification method, false discovery estimation and filter) can have a significant impact on data comparability and even results. In order to circumvent this, data imported into ProteomicsDB is processed and normalized using a uniform pipeline after manual annotation. This enables the direct comparison of results, especially when data was acquired in different labs. The combination of different experimental conditions, platforms and acquisition approaches can be used to avoid the miss-classification and categorization of repeating signals as biologically meaningful while being experimental artifacts. Especially the incorporation of other "omics" data sources can help to decrease the number of experimental artifacts by providing orthogonal measurements. However, while in theory replicate measurements of independent samples enable the separation of real signals from noise, the complete annotation of the treatment conditions and sample origin is essential in order to attribute and account for variations introduced during sample preparation. This is an additional challenge since our understanding of the underlying factors is still limited. Further aggravated by unrecognized impurities and contaminations our ability to fully characterize and understand biological systems is hampered, often leading to irreproducible results.

5 Outlook

ProteomicsDB in its current state does not exploit the full potential of the in-memory capabilities of the underlying database management system. One major goal is to foster the generation, expansion and validation of hypothesis. For this purpose, ProteomicsDB already enables the exploration of expression patterns across multiple experiments, which, for example, enables researchers to select cell lines which express a particular set of proteins. The extension to context-sensitive queries could enable researchers to browse and explore quantitative data more efficiently. Such context-specific queries could be "visualize all proteins which are x-fold differentially expressed in comparison to the median across all tissues" or "proteins which are significantly differentially regulated between two or more biological entities". Furthermore, the analysis of co-expressed proteins could provide new biological insight into the function of uncharacterized proteins. Depending on the allowed complexity of the co-expression pattern, these calculations generate a huge number of cross-correlations. However, ProteomicsDB could enable researchers to investigate of such patterns in real time on a proteome-wide scale. All these applications rely on fast and performant data access and especially when user-defined FDR filters are possible, the in-memory computing capabilities of ProteomicsDB can be used to its full potential.

An important aspect of repositories and databases storing results is to provide access to the underlying data to support the re-analysis of data. For this, ProteomicsDB provides programmatic access to the data via different OData services. However, these are currently bound to rather specific questions a user might have and thus only provide limited functionality. With the capabilities of the OData protocol in combination with the in-memory storage, unlimited and comprehensive access to the entire content in ProteomicsDB is possible, but not fully implemented yet. The implementation of such a model would provide data to many labs which might not have direct access to large amounts proteomics data. This is of special importance to enable researchers in the fields of computational proteomics, computational biology and related bioinformatics to test and refine algorithms on large heterogeneous datasets.

While SAP HANA enables the creation of database-side R-procedure which are passed to an R-server, due to technical limitations, this feature was not used to its full potential. With many packages enabling data comparison, analysis and interpretation[50], R provides a rich environment to supplement and extends the capabilities of ProteomicsDB. Current use-cases, such as model fitting, can be performed in real-time and could enable a more immersive data exploration.

In addition to different interfaces to other programming languages like C++ and python, SAP HANA supports the integration of other data sources as so called data providers. A potentially very powerful extension of ProteomicsDB is the access to raw MS data via Hadoop[51]. In contrast to many other "-omics" technologies, proteomics data is largely comprised of unidentified and thus unused features[52]. Maxquant already enables the matching of similar but not reproducibly identified features to reduce the number of missing values for protein quantification to increase the accuracy of expression estimation. However, this is only performed within experiments and does not allow the matching of features across thousands of files in an iterative and online manner. Storing all identified isotope features within ProteomicsDB is beyond the scope, but Hadoop might enable the integration of these features to enable the retrospective quantification of features based on what has been observed and identified in previous runs. Similarly, only a

fraction of the acquired MS/MS spectra is confidently identified in a single run. As previous studies have shown, spectral clustering enables the identification of these[53]. Here again, previous analyses might enable a more thorough interpretation of the acquired data due to the implementation of, for example, a dependent peptide search or the use of spectra libraries to enable the retrospective or additional identification of PTMs or variant peptides.

ProteomicsDB already enables the storage of reference spectra. Recently, a study reported the generation of reference spectra on a large collection of synthetically generated peptides[54]. However, since the raw MS data are not publicly available, reference spectra generated from high quality matches stored in databases such as ProteomicsDB can be used as an alternative source. As almost any FDR filter applied to PSMs, peptides and proteins will results in the elimination of true-positives, resources such as ProteomicsDB could provide means to cross-validate acquired MS/MS spectra. Since the correlation of consensus spectra to experimental spectra largely depends on the relative intensities of reproducibly found peaks, a feature widely ignored by common search engines, no fragment peak heuristics are necessary and thus should enable orthogonal validation.

One major shortcoming of current bottom-up shotgun proteomics approach is our limited understanding of its basic underlying mechanisms. For example, the exact peptide properties which lead to for example good ionization and downstream fragmentation are not known and thus hamper *in silico* prediction. However, these are of particularly importance for targeted proteomics approaches such as multiple and parallel reaction monitoring or data independent acquisition methods. Given the large number of spectra and identifications recorded in different biological background, machine learning approaches applied to those will not only deepen our understanding of the principle mechanisms, but also enable the generation of background-dependent interference maps. With this, a data-driven automatic generation of targeted assays for most previously identified proteins is in principle possible[55].

6 Abbreviations

OData	Open data protocol
REST	Representational state transfer
HTTP	Hypertext transfer protocol
HTPPS	Secure HTTP
API	Application programming interface
JSON	JavaScript object notation
XS	Extended application services
PRIDE	Proteomics identification archive
EC_{50}	Half maximal effective concentration
IC50	Half maximal inhibitory concentrations
CETSA	Cellular thermal shift assay
ITDR	Isothermal dose response
MassIVE	Mass spectrometry interactive virtual environment
PSM	Peptide spectrum match
FDR	False discovery rate
LC/MS-MS	Liquid chromatography tandem mass spectrometry
CID	Collision-induced dissociation
HCD	Higher-energy collisional dissociation
Ajax	Asynchronous JavaScript and XML
XML	Extensible markup language
SQL	Structured query language
SILAC	Stable isotope labeling with amino acids in cell culture
TMT	Tandem mass tag
PTM	Post translational modification
GO	Gene ontology
SRM/MRM/PRM	Single/multiple/parallel reaction monitoring
RNA	Ribonucleic acid
RNASeq	RNA sequencing
BIC	Bayesian information criterion
pEC_{50}	$-\log_{10} EC_{50}$
K_d	Dissociation constant
HTML	HyperText markup language
DIA	data-independent acquisition
PEP	posterior error probability
PSM	peptide spectrum match

7 References

1 Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H. & Vizcaino, J. A. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 15, 930-949, doi:10.1002/pmic.201400302 (2015).

2 Vizcaino, J. A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research* 41, D1063-1069, doi:10.1093/nar/gks1262 (2013).

3 Farrah, T. *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *Journal of proteome research* 13, 60-75, doi:10.1021/pr4010037 (2014).

4 Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Molecular & cellular proteomics : MCP* 11, 492-500, doi:10.1074/mcp.O111.014704 (2012).

5 Choudhary, C. & Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews. Molecular cell biology* 11, 427-439, doi:10.1038/nrm2900 (2010).

6 Welle, K. A. *et al.* Time-resolved analysis of proteome dynamics by TMT-SILAC hyperplexing. *Molecular & cellular proteomics : MCP*, doi:10.1074/mcp.M116.063230 (2016).

7 Koch, H. *et al.* Phosphoproteome profiling reveals molecular mechanisms of growth factor mediated kinase inhibitor resistance in EGFR overexpressing cancer cells. *Journal of proteome research*, doi:10.1021/acs.jproteome.6b00621 (2016).

8 Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature methods* 10, 730-736, doi:10.1038/nmeth.2557 (2013).

9 Schirle, M., Bantscheff, M. & Kuster, B. Mass spectrometry-based proteomics in preclinical drug discovery. *Chemistry & biology* 19, 72-84, doi:10.1016/j.chembiol.2012.01.002 (2012).

10 Csordas, A. *et al.* PRIDE: quality control in a proteomics data repository. *Database : the journal of biological databases and curation* 2012, bas004, doi:10.1093/database/bas004 (2012).

11 Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature communications* 7, 11778, doi:10.1038/ncomms11778 (2016).

12 Faerber, F. *et al.* SAP HANA database: data management for modern business applications. *SIGMOD Rec.* 40, 45-51, doi:10.1145/2094114.2094126 (2012).

13 Melnikova, I. & Golden, J. Targeting protein kinases. *Nature reviews. Drug discovery* 3, 993-994, doi:10.1038/nrd1600 (2004).

14 Fedorov, O., Muller, S. & Knapp, S. The (un)targeted cancer kinome. *Nature chemical biology* 6, 166-169, doi:10.1038/nchembio.297 (2010).

15 Page, T. H., Smolinska, M., Gillespie, J., Urbaniak, A. M. & Foxwell, B. M. Tyrosine kinases and inflammatory signalling. *Current molecular medicine* 9, 69-85 (2009).

16 Hainaut, P. & Plymoth, A. Targeting the hallmarks of cancer: towards a rational approach to next-generation cancer therapy. *Current opinion in oncology* 25, 50-51, doi:10.1097/CCO.0b013e32835b651e (2013).

17 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* 144, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

18 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* 100, 57-70 (2000).

19 Liu, Y. & Gray, N. S. Rational design of inhibitors that bind to inactive kinase conformations. *Nature chemical biology* 2, 358-364, doi:10.1038/nchembio799 (2006).

20 Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* 298, 1912-1934, doi:10.1126/science.1075762 (2002).

21 Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nature biotechnology* 29, 1039-1045, doi:10.1038/nbt.2017 (2011).

22 Zinn, N., Hopf, C., Drewes, G. & Bantscheff, M. Mass spectrometry approaches to monitor protein-drug interactions. *Methods* 57, 430-440, doi:10.1016/j.ymeth.2012.05.008 (2012).

23 Bantscheff, M. *et al.* Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature biotechnology* 25, 1035-1044, doi:10.1038/nbt1328 (2007).

24 Medard, G. *et al.* Optimized chemical proteomics assay for kinase inhibitor profiling. *Journal of proteome research* 14, 1574-1586, doi:10.1021/pr5012608 (2015).

25 Martinez Molina, D. & Nordlund, P. The Cellular Thermal Shift Assay: A Novel Biophysical Assay for In Situ Drug Target Engagement and Mechanistic Biomarker Studies. *Annual review of pharmacology and toxicology* 56, 141-161, doi:10.1146/annurev-pharmtox-010715-103715 (2016).

26 Jafari, R. *et al.* The cellular thermal shift assay for evaluating drug target interactions in cells. *Nature protocols* 9, 2100-2122, doi:10.1038/nprot.2014.138 (2014).

27 Martinez Molina, D. *et al.* Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* 341, 84-87, doi:10.1126/science.1233606 (2013).

28 Scheuermann, T. H., Padrick, S. B., Gardner, K. H. & Brautigam, C. A. On the acquisition and analysis of microscale thermophoresis data. *Analytical biochemistry* 496, 79-93, doi:10.1016/j.ab.2015.12.013 (2016).

29 Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* 346, 1255784, doi:10.1126/science.1255784 (2014).

30 Franken, H. *et al.* Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature protocols* 10, 1567-1593, doi:10.1038/nprot.2015.101 (2015).

31 Neuhauser, N., Michalski, A., Cox, J. & Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Molecular & cellular proteomics : MCP* 11, 1500-1509, doi:10.1074/mcp.M112.020271 (2012).

32 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, doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (1999).

33 Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* 4, 923-925, doi:10.1038/nmeth1113 (2007).

34 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, doi:10.1038/nbt.1511 (2008).

35 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* 10, 1794-1805, doi:10.1021/pr101065j (2011).

36 Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* 30, 918-920, doi:10.1038/nbt.2377 (2012).

37 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* 473, 337-342, doi:10.1038/nature10098 (2011).

38 Ahrne, E., Molzahn, L., Glatter, T. & Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* 13, 2567-2578, doi:10.1002/pmic.201300135 (2013).

39 Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Molecular & cellular proteomics : MCP* 5, 144-156, doi:10.1074/mcp.M500230-MCP200 (2006).

40 Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-Response Analysis Using R. *PLoS ONE* 10, e0146021, doi:10.1371/journal.pone.0146021 (2015).

41 Carpenter, R. G. Principles and procedures of statistics, with special reference to the biological sciences. *The Eugenics Review* 52, 172-173 (1960).

42 Schwarz, G. Estimating the Dimension of a Model. 461-464, doi:10.1214/aos/1176344136 (1978).

43 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419, doi:10.1126/science.1260419 (2015).

44 Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* 42, D1091-1097, doi:10.1093/nar/gkt1068 (2014).

45 Piccaluga, P. P., Paolini, S. & Martinelli, G. Tyrosine kinase inhibitors for the treatment of Philadelphia chromosome-positive adult acute lymphoblastic leukemia. *Cancer* 110, 1178-1186, doi:10.1002/cncr.22881 (2007).

46 Huang da, W. *et al.* Extracting biological meaning from large gene lists with DAVID. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 13, Unit 13 11, doi:10.1002/0471250953.bi1311s27 (2009).

47 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4, 44-57, doi:10.1038/nprot.2008.211 (2009).

48 Weisberg, E., Manley, P. W., Cowan-Jacob, S. W., Hochhaus, A. & Griffin, J. D. Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. *Nature reviews. Cancer* 7, 345-356, doi:10.1038/nrc2126 (2007).

49 Santos, F. P., Kantarjian, H., Cortes, J. & Quintas-Cardama, A. Bafetinib, a dual Bcr-Abl/Lyn tyrosine kinase inhibitor for the potential treatment of leukemia. *Current opinion in investigational drugs* 11, 1450-1465 (2010).

50 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* 5, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

51 Hillman, C., Ahmad, Y., Whitehorn, M. & Cobley, A. Near Real-Time Processing of Proteomics Data Using Hadoop. *Big data* 2, 44-49, doi:10.1089/big.2013.0036 (2014).

52 Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of proteome research* 10, 1785-1793, doi:10.1021/pr101060v (2011).

53 Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature methods* 13, 651-656, doi:10.1038/nmeth.3902 (2016).

54 Kusebauch, U. *et al.* Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* 166, 766-778, doi:10.1016/j.cell.2016.06.041 (2016).

55 Lawrence, R. T., Searle, B. C., Llovet, A. & Villen, J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nature methods* 13, 431-434, doi:10.1038/nmeth.3811 (2016).

Chapter 3

Mass spectrometry based draft of the human proteome

Contents

“The first draft of anything is shit.”
- Ernest Hemingway

„No passion in the world is equal to the passion to alter someone else's draft.”
- H. G. Wells

1 Introduction

The human genome represents a rather static definition of the potential protein inventory of every cell in the human body (in health and disease). In contrast, the proteomes of cells are very divers and highly complex. Protein expression typically spans 4-5 orders of magnitude[1]) in cell lines (and presumably tissues) and more than 10 orders of magnitude in body fluids[2]. The molecular complexity is even higher as proteins are often expressed as splice variants and/or processed and chemically modified on the co- or post-translational level.

The large-scale interrogation of biological systems by mass spectrometry based proteomics provides insights into protein abundance, cell type and time dependent expression patterns, post-translational modifications (PTMs) and protein-protein interactions, all of which carry biological information that is best investigated at the protein level. Perhaps surprisingly, it is still not clear which of the 19,629/20,493 human genes annotated in UniProt Swiss-Prot/TrEMBL[3] are translated into proteins. Therefore, major efforts are underway including the human proteome project (HPP) that aims to broadly characterize the human proteome[4], the human protein atlas project (HPA) which seeks to generate antibodies for all human proteins[5] and the ProteomeXchange consortium which facilitates the gathering and sharing of proteomic data[6]. Despite the fact that a plethora of individual human proteomic studies exist, only few systematic efforts in assembling and characterizing human proteomes have been reported thus far[7-10]. In part this is because most mass spectrometry-based proteomic data does not reside in public repositories, its annotation is often sketchy and the data generation and processing platforms are of varying capability, performance and maturity. Importantly, there also is a significant challenge in making such 'big data' more widely accessible to the scientific community because the development of scalable analysis tools is only in its infancy.

This chapter presents a mass spectrometry-based draft of the human proteome available via ProteomicsDB, a public in-memory database for real time analysis of big data. In order to identify at least one protein per protein coding gene, the strategy towards a mass spectrometry based draft of the human proteome focused on the assumption that the selection of datasets has to cover the breadth and depth of the protein repertoire in human samples (Figure 3.37a). First, studies on (mostly cancer) cell lines which analyzed in depth a single cell line or a large number of divers cell lines (e.g.[11,12]) were selected. To capture tissue and body fluid specific proteins, 31 tissues and body fluids were analyzed in-house and similar datasets from the public domain, often from rarely analyzed or not readily accessible anatomical components, were added. In order to address low abundant proteins such as kinases or transcription factors, ~1,300 affinity purifications including kinome enrichments[12], HDAC enrichments[13] and protein-protein-interaction studies[14] were added. Similarly, numerous PTM studies, which increased the coverage of proteins that are highly modified and/or of low abundance (e.g.[15,16]) were included. However, with the progression of assembling data in ProteomicsDB, it became clear that numerous protein groups were underrepresented (e.g. membrane proteins, keratin associated proteins, cytokines), and the focus shifted towards datasets which may fill these gaps. For instance, to increase the chance of identifying keratin-associated proteins as well as other proteins with only few tryptic peptides, numerous tissues (including hair follicles) were analyzed using proteases different from trypsin (namely chymotrypsin and/or LysC). To further increase the coverage of the human proteome, a data-driven approach was selected to systematically assemble publically available

and in-house data from experiments in which missing protein coding genes have been identified in previous studies.

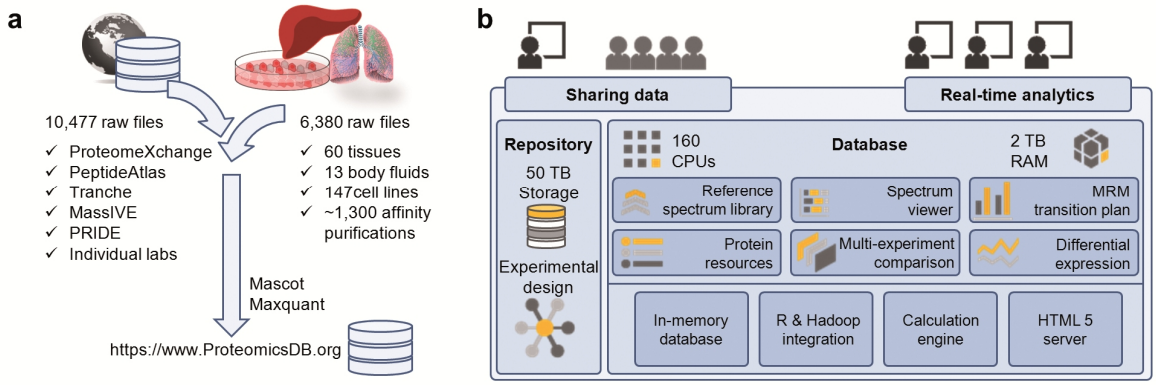


Figure 3.37 | Strategy for the assembly of the human proteome. a, Experimental workflow for the identification and quantification of proteins. b, Data storage and processing is performed in ProteomicsDB, an in-memory database enabling fast computation on large data sets backed by 160 CPUs and 2TB of RAM.

The information assembled from human tissues, cell lines and body fluids allowed estimating the size of the protein coding genome, identified organ-specific proteins and a large number of translated lincRNAs. Analysis of mRNA and protein expression profiles of human tissues revealed conserved control of protein abundance, integration of drug sensitivity data allowed the identification of proteins predicting resistance or sensitivity and proteome profiles also hold considerable promise for analyzing the composition and stoichiometry of protein complexes. The unique high performance in-memory platform ProteomicsDB (Figure 3.37b) enables navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

2 Methods

Briefly, Proteomic data were downloaded from public repositories, contributed by individual laboratories and specifically generated for this study by the authors' laboratories. For the latter, human tissue specimens were obtained from the bio bank of the TU München following approval of the study by the local ethics committee. Samples were collected within the first 30 minutes after resection, macroscopically resected by an experienced pathologist, snap frozen and stored in liquid nitrogen until use. Body fluids requiring no invasive procedures were provided by volunteers. Proteins were extracted under denaturing conditions and either separated by LDS-PAGE followed by in-gel protease digestion or digested in solution in the presence of chaotropic agents. Synthetic peptides were produced by solid phase chemistry following the standard Fmoc strategy and used without purification. Peptides were separated by ultra-high pressure liquid chromatography and analyzed on Orbitrap mass spectrometers using either resonance-type or beam-type collision induced dissociation. For peptide identification, tandem mass spectra were processed in parallel using Mascot Distiller and MaxQuant/Andromeda[17] and searched against Uniprot and/or a custom build fasta formatted sequence file containing lincRNA sequences Search results and tandem mass spectra were imported into ProteomicsDB (https://www.proteomicsdb.org) and filtered at 1% PSM FDR and 5% local peptide length dependent FDR. For bioinformatic analysis, data was extracted from ProteomicsDB using HANA Studio and further processed using custom python and R scripts. Gene ontology analysis was performed using David (http://david.abcc.ncifcrf.gov) and REViGO (http://revigo.irb.hr/).

2.1 Sample preparation

2.1.1 Source of human tissues and fluids

Tissue samples were obtained from the biobank of the Klinikum rechts der Isar and Faculty of Medicine of the TU München (BBTU). Healthy tissue samples were obtained by surgery and macroscopically dissected by an experienced pathologist. Tissue was snap frozen within the first 20 minutes after resection and stored in liquid nitrogen (-196°C) until usage. Before analysis, tissue type and tissue quality (no necrosis) was confirmed by a pathologist. For this purpose an HE stained slide was prepared from the actual sample to be used. Body fluids requiring no invasive procedures were provided by healthy volunteers. All patients/donors provided informed written consent, and the study was approved by the Ethics Committee of the Technische Universität München.

2.1.2 Preparation of protein extracts from tissues and body fluids

In order to generate a map of protein expression of all major tissues and organs in the human body ('human body map' project in ProteomicsDB), 31 different tissues and body fluids from healthy volunteers (see above) were analyzed. Fresh frozen tissue was thawed, cut into small pieces and extensively washed with precooled phosphate buffered saline. The tissue was homogenized in cold lysis buffer (50 mM Tris/HCl pH 7.5, 5% glycerol, 1.5 mM MgCl2, 150 mM NaCl, 0.8% Nonident P-40, 1 mM dithiothreitol [DTT] and 25 mM NaF) with freshly added protease (SIGMAFAST, Sigma-Aldrich, Germany) and phosphatase inhibitors (20 mM NaF, 1 mM sodium orthovanadate, 5 mM calyculin A; Sigma-Aldrich, Germany) using ceramic beads and an automated homogenizer (Precellys 24 Homogenisator, Peqlab, Germany) with 2x 10 second

pulses at 5000 rpm. Homogenates were clarified by centrifugation for 15 minutes at 10,000 xg at 4°C and protein concentration was determined by the Bradford method. The supernatant was stored at -80 °C. Approximately 30 µl of body fluids (cerumen, saliva, ascites etc.) were boiled with 2x NuPAGE sample buffer (Life Technologies, Germany) prior to LDS-PAGE and in-gel digestion.

2.1.3 Protein separation and in-gel digestion

100 µg protein extract from tissues or 30 µl from body fluid extracts were reduced and alkylated by 10 mM DTT at 55 °C for 10 min followed by the addition to 55 mM chloroacetamide and incubation at room temperature for 30 min and subsequently denatured at 95 °C for 10 min. Samples were then separated via a 4–12% NuPAGE gel (Life Technologies, Germany) and cut into 24 slices prior to in-gel digestion using trypsin (Promega, Germany), LysC (Wako Chemicals, Japan) or chymotrypsin (Roche, Germany). In-gel digestion was performed according to standard procedures[18] and peptides were analyzed by LC-MS/MS.

2.1.4 Chymotrypsin in-solution digestion and hydrophilic strong anion exchange chromatography (hSAX) peptide separation

In-solution digestion of testis protein extract was performed according to standard high urea procedures. Briefly, the lysate was denatured in 8 M urea, 0.1 M Tris/HCl, subsequently diluted to 2 M urea followed by protein digestion with chymotrypsin. After overnight digestion at 30 °C, peptides were concentrated and purified on C18 StageTips as described[19]. Peptide separation using hSAX was essentially performed as described[20] and the resulting 48 fractions (1 minute collection time per fraction) were dried down and subsequently analysed by LC-MS/MS.

2.2 Sample acquisition and retrieval

2.2.1 Liquid chromatography tandem mass spectrometry (LC-MS/MS)

LC-MS/MS was performed by coupling an Eksigent nanoLC-Ultra 1D+ (Eksigent, Dublin, CA) to an Orbitrap Elite instrument (Thermo Scientific, Bremen, Germany). Peptides from an in-gel digest were delivered to a trap column (100 µm×2 cm, packed in-house with Reprosil-Pur C18-GOLD 5 µm resin, Dr. Maisch, Ammerbuch, Germany) at a flow rate of 5 µL/min in 100% solvent A (0.1% formic acid, FA, in HPLC grade water). After 10 min of loading and washing, peptides were transferred to an analytical column (75 µm×40 cm, packed in-house with Reprosil-Gold C18, 3 µm resin, Dr. Maisch, Ammerbuch, Germany) and separated at a flow rate of 300 nL/min using a 110 min gradient from 4% to 32% solvent B (solvent A: 0.1% FA, 5% DMSO in HPLC grade water; solvent B: 0.1% FA, 5% DMSO in acetonitrile). Both solvent A and B contained 5% DMSO to boost the nESI response and the MS signal of peptides[21]. The eluent was sprayed via stainless steel emitters (Proxeon) at a spray voltage of 2.2 kV and a heated capillary temperature of 275°C. The Orbitrap Elite instrument was operated in data-dependent mode, automatically switching between MS and MS2. Full scan MS spectra (m/z 360 – 1300) were acquired in the Orbitrap at 30,000 (m/z 400) resolution using an automatic gain control (AGC) target value of 1e6 charges. Tandem mass spectra of up to 15 precursors were generated in the multipole collision cell by using higher energy collisional dissociation (HCD) (AGC target value 2×104, normalized collision energy of 30%) and analyzed in the Orbitrap at a resolution of 15,000. Precursor ion isolation width was set to 2.0 Th, the maximum injection time for MS/MS was 100 ms and dynamic exclusion was set to 30 s. Internal calibration was performed on-the-fly using a DMSO-related lock mass (m/z 401.922718, $[C_6H_{10}O_{14}S_3]^+$)[21].

Measurements using an Orbitrap Velos used the same LC system, column and gradient as described and similar data acquisition parameters. The main differences are that the Orbitrap Velos used HCD of the top 10 precursor ions at an AGC target value of 3×104 and Orbitrap readout at a resolution of 7,500. Measurements using a LTQ Orbitrap XL used the same LC system, column and gradient as described and similar data acquisition parameters. The main differences are that the LTQ Orbitrap XL used CID fragmentation with normalized collision energy of 35% of the top 10 precursor ions at an AGC target value of 5×103 and ion trap readout.

2.2.2 Retrieval of MS-based proteomics data from public repositories and other sources

Alongside published and unpublished data generated at the TU München, and datasets obtained from individual laboratories (Parag Mallick, Stanford School of Medicine; Hanno Steen, Harvard Medical School, Boston; Tamar Geiger and Mathias Mann, Max-Planck-Institute of Biochemistry, Martinsried; Shabaz Mohammed, Javier Munoz and Albert Heck, Utrecht University; Magnus Berle, Haukeland University Hospital, Oslo; Andrew Emili, University of Toronto; Roman Zubarev, Karolinska Insitute, Stockholm), MS-based proteomics datasets were downloaded from public repositories and servers including ProteomeXchange[22,23], PeptideAtlas[24,25], Tranche[26,27], MassIVE[28], CPTAC data portal[29,30], Broad Institute's proteomics FTP server[31], SCOR[32,33], and PRIDE[22,34]. Raw data files of all published studies and obtained from the public domain are made available through ProteomicsDB[1,12-16,21,32,35-81]. Two exceptions are the data sets from the CPTAC consortium because CPTAC requires permission to download/use their data. In addition, there is a small amount of data (10% of the total) that originate from (ongoing, unpublished) projects in the TUM lab and which are outside the scope of this work. These will also be made available for download once the respective manuscripts have been accepted for publication. Still, even for these projects, ProteomicsDB shows the peptides/spectra for the purpose of protein identification (without disclosing experimental details and sample annotations).

2.3 Data analysis

2.3.1 Peptide identification using Mascot and Maxquant/Andromeda

Raw MS data files from Orbitrap-type instruments (including LTQ-FT) were processed in parallel with two different data processing pipelines and search engines, Mascot[82] and Maxquant/Andromeda[17,83].

The following protein sequence databases were used for peptide identification: (1) the UniProtKB complete human proteome (download date: 05 Sep 2012; 86,725 sequences) contains the UniProtKB/Swiss-Prot complete human proteome (20,225 canonical sequences of protein-coding genes and 16,624 manually curated isoform sequences) and 49,876 UniProtKB/TrEMBL sequences, the latter representing all predicted protein sequences from Ensembl (except fragments) that were found to be absent from the UniProtKB/Swiss-Prot complete proteome (for a detailed description, see also[84,85]); (2) the cRAP database (common Repository of Adventitious Proteins; download date: 05 Sep 2012; 113 sequences) contains common laboratory proteins, proteins added by accident through dust or physical contact; and proteins used as molecular weight or mass spectrometry quantitation standards[86]. The UniProtKB complete human proteome and cRAP database were concatenated and every database search was performed against both databases.

In the Mascot workflow, raw MS data files were processed using Mascot Distiller (version 2.3.2, Matrix Science, UK) with processing parameters separately optimized for high and low resolution

tandem mass spectra. Data processing comprised peak picking, de-isotoping and charge deconvolution of fragment ions. The resulting peaklist files were searched using the Mascot search engine (version 2.4.1, Matrix Science, UK) against the UniProtKB complete human proteome and the cRAP database. The target-decoy option of Mascot was enabled (on-the-fly search against a decoy database with reversed protein sequences) and search parameters included a precursor tolerance of 10 ppm and a fragment tolerance of 0.5 Da for CID spectra and 0.05 Da for HCD spectra. Enzyme specificity was set to trypsin, LysC, GluC, or chymotrypsin (as applicable), and up to two missed cleavage sites were allowed. The Mascot ^{13}C option, which accounts for the mis-assignment of the monoisotopic precursor peak, was set to 1 and the following variable modifications were included by default: oxidation of Met as well as acetylation of the protein amino-terminus. Other variable and fixed modifications (such as phosphorylation of Ser, Thr and Tyr or acetylation of Lys as well as iTRAQ, TMT, SILAC or carbamidomethylation of Cys) were set as provided in the corresponding publications. Mascot search results were processed using the Mascot Percolator stand-alone software to obtain posterior error probabilities (PEPs)[87,88].
In the Maxquant workflow, raw MS data files were processed by Maxquant (version 1.3.0.3) for peak detection and quantification[17]. MS/MS spectra were searched against the same set of target and decoy databases as described for Mascot using the Andromeda search engine[83]. Proteases, variable and fixed modifications were specified as above. Mass accuracy of the precursor ions was determined by the time-dependent recalibration algorithm of Maxquant, and fragment ion mass tolerance was set to of 0.6 Da and 20 ppm for CID and HCD, respectively. No FDR cut-off was specified (see below), but a minimum peptide length of 6 amino acids was required. In few cases, where Maxquant (version 1.3.03) was unable to process the raw files, a newer version of Maxquant (version 1.4.05) was used.
In two cases, where raw MS data were unavailable (Burkhart_Blood_2012, Didangelos_MCP_2011), the available data was processed only through the Mascot pipeline, and, if required, peaklist files were converted into Mascot generic format (mgf) prior to database search. A special case represents the PeptideAtlas spectrum library (release December 2012) containing spectra of 253,693 distinct peptides from high and low resolution tandem mass spectra[71]. In order to make this resource available in ProteomicsDB, all spectra representing peptide identifications from high resolution MS and MS/MS measurements (including their respective modifications) were parsed and converted into mgf format and subjected to Mascot database search.

2.3.2 Identification of peptides of lincRNAs and transcripts of unknown coding potential (TUCPs)
In order to assess the presence and abundance of peptides representing translation products of lincRNAs and TUCPs, 23 tissue datasets of the human body map (adrenal gland, anus, cervix uteri, esophagus, gall bladder, kidney, liver, lung, nasopharynx, oral cavity, ovary, pancreas, placenta, prostate, salivary glands, seminal vesicle, spleen, stomach, testis, thyroid gland, tonsils, uterus [post-menopause], uterus [pre-menopause]) as well as a deep HeLa proteome dataset[66] (excluding GluC data) were searched separately against two lincRNA/TUCP sequence databases each appended to the human UniProtKB and cRAP database using Mascot with default parameters. The following two lincRNA/TUCP reference catalogues were used to search for translation products: (1) The Ensembl human non-coding map (GRCh37; download date: 02 Nov. 2013) comprises 13,564 non coding genes and (2) the Broad Institute's human body map of 21,487

lincRNAs and TUCPs (14,281 non coding transcripts as well as 7,206 novel transcripts with potential coding capacity, TUCPs)[89]. The Broad Institute utilized RNA sequencing and transcript abundance estimations to identify and characterise lincRNAs across 24 tissues and cell types. The BED files of Broad Institute's lincRNA transcripts were converted to FASTA using BEDTools[90]. The FASTA nucleotide sequences were translated to protein sequences by translating all three reading frames using custom python scripts.

To minimize the level of uncertainty in assigning identified peptides to lincRNAs/TUCPs, the identified peptides were as rigorously filtered as peptide identifications from standard database searches (see above) and, in addition, a Mascot delta score of >10 was required to assure that the best match for the spectrum is significantly better (10x) than the second best match[91]. Identified lincRNA peptides were further subjected to a BLAST search against all sequences in the UniProtKB complete human proteome database using parameters for short sequences (blastp -word_size 2 -matrix PAM30 -seg "no" -evalue 20000 -comp_based_stats 0). Peptides with identical sequence matches (and isobaric sequence variants thereof) in the UniProtKB complete human proteome were categorically rejected. Identified peptides representing translation products of lincRNAs or TUCPs were imported into ProteomicsDB and the corresponding links are provided in[92]. To reduce redundancy on peptide and transcript level, identified transcripts from the Ensembl and Broad Institute's database search were grouped together according to their matched peptides as same-sets and sub-sets of peptide identifications.

2.3.3 Generation of reference peptide spectrum libraries

Reference peptides for uncertain protein-coding genes[93] were manually selected based on MS/MS evidence for these genes. Reference peptides for cytokines were selected for 361 proteins associated with the keyword 'cytokine' in UniProtKKB/Swiss-Prot using peptides present in SRMAtlas with lengths from 7 to 25 amino acids. Reference peptides for proteins with weak or no evidence in ProteomicsDB (at the time of synthesis) were selected based on the availability of PSMs or, for proteins not observed in ProteomicsDB, a set of peptides with good MS properties was derived by bioinformatic prediction using PeptideSieve[94].

Reference peptides for uncertain genes were synthesized by resin based solid-phase peptide synthesis at the authors labs at the TU München[95] or by SPOT synthesis technology[96,97] at the authors labs at JPT. After synthesis, the reference peptide sets were pooled (up to 1000 peptides per pool) and appropriately diluted in LC solvent A (1:100-1:10,000) before LC-MS/MS analyses using an Orbitrap Velos or Elite (see above). To generate reference spectrum libraries for both CID and HCD spectra, the peptide pools were analyzed at least once with each fragmentation type. Reference peptide spectra were processed using Mascot and Maxquant as described above and imported into ProteomicsDB. Available reference spectra and their links into ProteomicsDB are provided in[92].

2.3.4 Data analysis in ProteomicsDB

ProteomicsDB utilize the main memory as the primary data storage. This reduces disk seek when querying data and, ultimately, results in faster data retrieval. Additionally, SAP HANA is specifically optimized for in memory operations which reduces the need of indexing tables and on top allows direct operations on compressed columns. This allows ProteomicsDB to show data and run queries using the entire database in real-time without the requirement of pre-assembled builds.

All relevant information from search result files of Mascot (dat file) and Maxquant ("combined/txt" folder, apl and res files) were imported into ProteomicsDB. In particular, ProteomicsDB comprises the up to top 10 best matches for each experimental spectrum both for Mascot and for Andromeda. Storing these PSMs allows to compute delta scores between the top two (or n) best matches which is useful, for instance, for computing false localization rates for phosphopeptides.
All projects imported into ProteomicsDB were manually annotated in order to provide high quality meta information such as project, experiment and sample descriptions in a computer-readable format. The annotations comprise experimental design of imported studies as well as biological and biochemical workflows. Where available, experiment annotations were translated into controlled vocabulary of open ontologies, remaining annotations were incorporated into a ProteomicsDB controlled vocabulary. The following ontologies are used in ProteomicsDB: Brenda tissue ontology (release date: 20 Dec 2012)[98], PSI-MS (version 3.44.0), UO (version 1.2) and sepCV (version 1.0). Identifiers, gene names, sequences etc. in ProteomicsDB are based on UniprotKB (download date: 05 Sep 2012), which, for instance, allows the localization of identified proteins to their chromosomal location (Figure 3.38a).

2.3.5 False discovery rate (FDR) control using a global and a local FDR filter

Target and decoy peptide spectrum matches (PSMs; including lower ranking hits) were imported into ProteomicsDB irrespective of their scores or posterior error probability. A two-step approach was chosen to control for spectrum and peptide level FDR. First, each LC-MS/MS experiment was filtered to 1% PSM level FDR using a global target-decoy approach[99]. Because this can still result in the retention of a considerable number of false matches, peptide identifications had to pass a second length dependent Mascot or Andromeda score threshold of 5% local FDR (Extended Data Figure B1 and B2 in the Appendix). To this end, peptide identifications of the same length from all experiments in ProteomicsDB were sorted in bins of 1 score points and the target-decoy FDR was calculated for each bin individually ('local score and length dependent FDR'). As an example, an Andromeda score of 100 (35 for Mascot) is required for peptides of length 7. This threshold is a high hurdles for a 7 amino acid peptide and, as a result, the majority of all PSMs of length 7 are in fact rejected.
In this study, no protein FDR measure was applied as the concept of a protein FDR is problematic (see below and next chapter). A comparison to 27 published high-throughput studies shows that this filtering scheme is in line with the often-used 1% protein FDR criterion and avoids the (at this time) unsolved issue of target-decoy searching that artificially high protein FDRs are generated when analyzing very large data sets[100].

2.3.6 Analysis of protein expression

Protein abundance was estimated for UniProtKB/Swiss-Prot sequences using the iBAQ approach (intensity based absolute protein quantification)[101]. To this end, peptide intensities for a given protein in a sample were obtained from Maxquant, summed up and divided by the number of observable peptides (length 6 to 30, no missed cleavage). In order to be able to compare protein abundances across multiple samples, experiments and projects, the iBAQ protein intensities were normalized based on the total sum of all protein intensities. Unless otherwise stated, the displayed protein abundance values were $\log_{10}$ transformed and right-shifted by 10 $\log_{10}$ units into positive

numerical space. It is important to note that ProteomicsDB aggregates any isoform specific abundance information at the gene locus level. Isoform-specific abundance information is currently not calculated as isoforms are usually under-sampled in MS-based proteomics. Protein expression analysis including normalization, hierarchical cluster analysis and principle component analysis was performed using R (v 2.12.1;[102]). Cluster analyses using a variety of common algorithms and metrics were performed to group the tissues and body fluids on the basis of protein expression pattern.

Principle component analysis (PCA) of proteome profiles of tissues and cell lines (Figure 3.39b, upper panel) was performed on 6094 proteins showing significant expression differences (ANOVA, Benjamini-Hochberg adjusted *p*-value of 0.05) between the tissue groups (ovary: 3 tissues, 4 cell lines; colon: 3 tissues, 8 cell lines; kidney: 1 tissue, 8 cell lines; lung: 4 tissues, 11 cell lines). PCA of proteome profiles of tissues samples (colon: 3; breast: 2; liver: 2; kidney: 1; ovary: 3; lung: 4; prostate: 2) was performed on all 10,710 proteins (Figure 3.39b, lower panel).

Hierarchical clustering of the top100 proteins per tissue (Figure 3.39c) as well as protein kinases and transcription factors (TF; Figure 3.40a) was performed on log transformed normalized iBAQ intensities using euclidean distance and complete linkage. For the analysis of tissue-specific kinase and TF expression, human protein kinases and TFs were retrieved from UniProtKB, resulting in protein expression profiles for 310 protein kinases (Pfam protein kinase domain), 557 transcription factors (GO term 'transcription factor') and 39 proteins annotated as kinases as well as transcription factors.

Tissue-specific proteins (Figure 3.40b) were calculated based on normalized iBAQ values for 47 tissues. Median protein expression values were used in all cases where multiple datasets for the same tissue were available. Proteins were considered as tissue-specific if their expression in a given tissue was at least ten-fold higher than the average expression in the remaining tissues. Classification and functional enrichment analysis of protein lists (such as the core proteome, missing proteins or tissue specific proteins) were performed using the DAVID Bioinformatics Database[103] as well as ReviGO[104].

The stoichiometry of protein complexes (nuclear pore complex [NPC], core proteasome; Figure 3.43, Extended Data Figure B9 in the Appendix) was reconstructed using normalized iBAQ values, and complex stoichiometry was calculated relative to either a particular subunit (nuclear pore complex: subunit Nup43) or the median subunit expression across all subunits (core proteasome). NPC compositions are based on data of triplicate shotgun experiments of HeLa nuclear extracts from a recent study of Ori *et al.*[67], and the proteasome composition was determined for 109 different samples (29 tissues from the human body map, 80 different cell line samples from two recent studies profiling the proteomes of 11 and 59 cell lines, respectively[11,12].

2.3.7 Genome-wide comparison of protein and transcript abundance levels across twelve tissues

For the systematic genome-wide comparison of protein and mRNA abundances across multiple tissues (

Figure 3.41), quantitative transcriptomics data (RNASeq) were downloaded from the Human Protein Atlas[105] (download date: Nov 11, 2013). Transcript expression data (abundances expressed as fragments per kilobase per million, FPKM) was extracted from the RNA-Seq data for 12 tissues (adrenal gland, esophagus, kidney, ovary, pancreas, prostate, salivary gland, spleen, stomach, testis, thyroid gland, uterus). Normalized iBAQ values were exported from ProteomicsDB for

theses tissues. In cases where multiple full proteomes were available for a single tissue or organ, the median protein expression values were used. Proteins were mapped to transcripts using BioMart[106] resulting in 6104 transcripts/proteins with expression data. For the comparison of protein and transcript abundances, protein and mRNA expression was re-scaled using z-score transformation (Extended Data Figure B7a in the Appendix). Translation rate and transcript abundance are the major determinants of protein expression, while transcript and protein half-life only play a minor role[101]. The ratio between mRNA and protein abundance is a proxy for the amount of protein which can be synthesized from a given mRNA and was calculated for all 12 tissues. The median ratio across all 12 tissues was then used to predict protein abundance from mRNA expression.

2.3.8 Elastic net analysis

In order to predict proteins involved in drug resistance and sensitivity, proteome and drug-response profiles for 24 FDA-approved drugs and 35 cell lines were used to identify known and potential protein markers for drug sensitivity and resistance. As described previously[12], an elastic net regression[107] was employed using full cellular proteome data of 135 experiments corresponding to 35 cell lines and 10,825 proteins (Supporting Table 7). In cases where multiple full cellular proteomes were available for a single cell line, the median protein expression values were used. Drug activity levels were obtained from the Cancer Cell Line Encyclopedia (CCLE) resource[108]. All features were regressed to fit a Gaussian model of drug activity area for each drug.

3 Results

3.1 Assembly of the proteome in ProteomicsDB

This draft of the human proteome was assembled from 16,857 liquid chromatography tandem mass spectrometry (LC-MS/MS) experiments using human tissues, cell lines, body fluids and PTM and affinity purifications and its analysis in ProteomicsDB, an in-memory database designed for the real time analysis of big data (https://www.proteomicsdb.org). The strategy (Figure 3.37a) was to combine data available from repositories and otherwise contributed by colleagues (60% of total) with published as well as new data from the authors' laboratories (40% of total). All datasets were re-processed using Mascot and MaxQuant[17,82] and the resulting 1.1 billion peptide spectrum matches (PSMs) were imported into ProteomicsDB. The database (Figure 3.37b) comprises a public repository, a web interface featuring several data views and analysis tools and an application programming interface (API). At the heart of ProteomicsDB is an 'in-memory' computational resource commanding 2 TB of RAM and 160 central processor units (CPUs) which allows keeping all data in the main memory all of the time. This makes computational tasks very efficient illustrated by the capability to display and annotate any of the currently ~71 million identified peptide mass spectra in real time (Extended Data Figure B1 in the Appendix). Controlling the quality of peptide and protein identifications is important but exactly how this is best accomplished is still debated in the community[109,110].

To control false-discovery rate across all samples in ProteomicsDB and find a reasonable compromise between sensitivity and specificity of the filtered data, two-stage filtering approach based on a classical target-decoy search strategy was applied. The results of the two-stage filtering approach for the two separate search engines Mascot and Maxquant/Andromeda are depicted in Extended Data Figure B1 in the Appendix.

3.2 Proteomic annotation of the genome

At the time of writing, ProteomicsDB held protein evidence for 18,097 of the 19,629 human genes annotated in Swiss-Prot (92%) as well as 19,376 out of 86,771 protein isoforms listed in Uniprot (22%). Chromosomes were evenly covered with the notable exceptions of chromosome 21 and the Y-chromosome (Figure 3.38a). The former contains many proteins with few MS-compatible tryptic peptides.257 human proteins (gene level) do not produce any such peptides rendering trypsin as the most frequently used protease in proteomics ineffective. As a result, alternative proteases or top down sequencing approaches will have a role to play in the eventual completion of the human proteome (Extended Data Figure B3a in the Appendix)[66,111]. To facilitate this, ProteomicsDB provides a tool predicting the best protease or combinations thereof for any protein which can also be valuable when systematically mapping PTMs.

ProteomicsDB covers 97% of the 13,378 genes with annotated evidence on protein and 84% (of 5,531) with evidence on transcript level. The overlap with proteins detected by antibodies in the HPA project is 93% (of 15,156 HPA proteins) providing independent evidence that these genes exist as proteins. Conversely, proteomic coverage of genes inferred from homology (52% of 159), genes marked as predicted (64% of 72) or uncertain (56% of 489) was considerably lower suggesting that the protein coding human genome may be several hundred genes smaller than previously anticipated. Of the 44 tried uncertain genes[93], 36 could be manually validated by comparing the experimental spectra against its reference spectra generated from synthetic

peptides. Among the identified uncertain genes were three long intergenic non-coding RNAs (lincRNAs, Extended Data Figure B3 in the Appendix). This surprising result initiated a search of ~9 million tandem MS spectra from tissues and cell lines against 13,564 lincRNA sequences from Ensembl and 21,487 lincRNAs and TUCPs (transcripts of uncertain coding potential) from the Broad Institute[89]. This returned 430 high quality peptides (no homology to Uniprot sequences) from 404 lincRNAs/TUCPs. There was no apparent bias in chromosomal location or biological source and the abundance distribution of translated lincRNA peptides was broadly similar to that of peptides from ordinary proteins (Extended Data Figure B3 in the Appendix). This is the largest number of lincRNA/TUCP translation products with direct peptide evidence reported to date[112] arguing that translation of such transcripts is more common than previously anticipated[113-115]. The biological significance of translated lincRNAs and TUCPs is not clear at present. These may constitute proteins 'in evolution' representing hitherto undiscovered biology[116] or arise by stochastic chance marking such proteins as 'biological noise'.

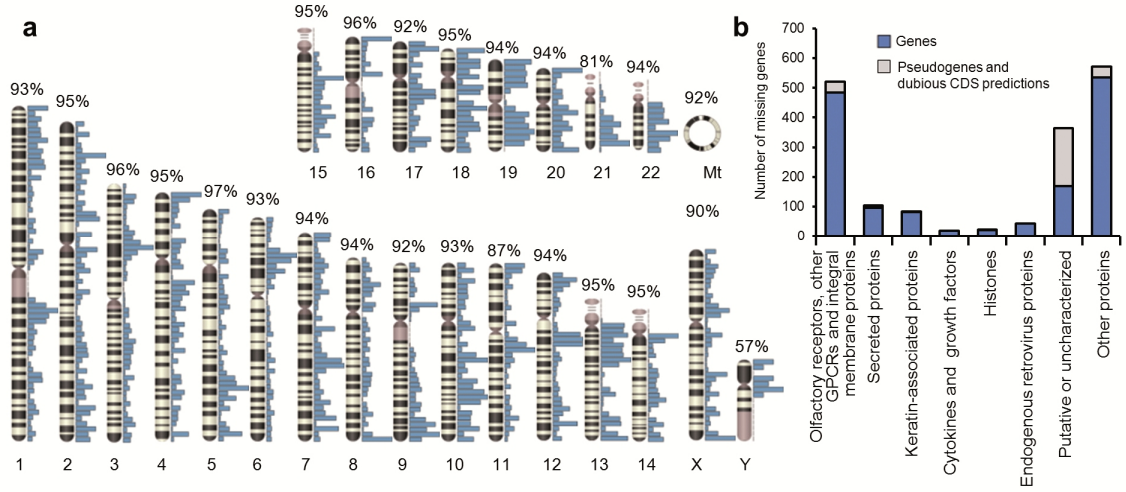


Figure 3.38 | Characterization of the human proteome. a, Chromosomal coverage of the 18,097 proteins identified in this study exceeds 90% in all but three cases. Blue bars indicate the density of proteins in a particular chromosomal region. b, Gene ontology analysis of the 'missing' proteome' identifies GPCRs, secreted and keratin-associated proteins as the major protein classes underrepresented in proteomic experiments.

3.3 Estimating and normalizing protein abundance in ProteomicsDB

Protein abundance in cells is often expressed as an approximated measure of concentration, such as copies per cell. However, this metric requires detailed knowledge about a couple of parameters including but not limited to cell number, size and morphology, and cannot be readily translated to tissues, extracellular structures or body fluids for which other concentration measures are required. One pragmatic approach to overcome this issue is to express the abundance of a protein as a fraction of all proteins detected and quantified in a sample. This normalization on the sum of all protein abundances avoids the difficulties associated with concentration measures. However, this approach is not particularly sensitive to the scaling of copy numbers by different parameters such as cell volume, shape, number, and potentially many others. This may lead to some inaccuracies, i.e. overestimating/underestimating proteins that have some relationship with a particular parameter (e.g. histones scaling with the number of genomes, ribosomes scaling with

the cell volume). Nonetheless, this pragmatic approach enables the straightforward comparison of protein expression across a wide range of different types of samples.

Five intensity-based approaches for the estimation of protein abundances were systematically investigated, in order to provide a consistent and reasonably accurate estimation of protein abundance enabling the comparison of protein expression between samples of an experiment or project as well as across multiple different projects. Spectral-counting based approaches were not considered as these are associated with less accurate protein abundance estimations[117,118].

The metrics investigated in detail are top3 intensity[119], top3 divided by the number of observable peptides, iBAQ[101], average intensity[120] and sum of all peptide intensities. The performance of these protein abundance metrics was compared with accurate copy number estimates from a large-scale absolute protein abundance data set (based on the AQUA technology). For this purpose, reference protein copy numbers for U2-OS cells from Beck et al.[1], and calculated copy numbers for the five abundance metrics using a large-scale U2-OS cell line data set acquired by Geiger et al.[11] (Extended Data Figure B5a in the Appendix) were retrieved. While all five metrics perform reasonably well (median fold error <2.5), the two approaches which take the number of observable peptides into account (iBAQ, top3 intensity divided by the number of observable peptides) are in better agreement with the original copy number estimates (Extended Data Figure B5e in the Appendix). Note that the observed median fold errors are in line with the quantitative accuracy estimated by Beck et al. using a bootstrapping approach for their AQUA quantified proteins. When comparing the quantitative accuracy of the two most frequently employed approaches, iBAQ and top3, as a function of the number of quantified peptides (Extended Data Figure B5f in the Appendix) as well as protein length (Extended Data Figure B5g in the Appendix), it becomes obvious that iBAQ is slightly less biased for low abundant proteins and small proteins than the top3 approach.

To further investigate the effect of the normalization based on the sum of all protein abundance measures, two data sets with considerably different proteome coverage (e. g. samples measured on different instruments) were compared. For this purpose, the protein expression values of nine different cell lines analyzed acquired on two different mass spectrometers[12] were selected. As exemplified for the Colo-205 cell line dataset in Extended Data Figure B5b in the Appendix for all five protein abundance metrics, the differences in intensity distribution before normalization primarily reflect the differences in sensitivity of the instruments. The chosen normalization ensures that what is an abundant protein in one data set is also an abundant protein in the other (by virtue of each protein being expressed as the fraction of the total; Extended Data Figure B5c in the Appendix). The main difference after normalization is thus the number of identified proteins rather than the resulting (relative) abundances. The histograms show that the data is very well aligned on the high abundance side of the histogram. For the low abundance proteins this visualization is somewhat misleading because the x-axis is in log10 scale, so the right half of the histogram actually covers >99% of the total intensity. The Q-Q plots for the same data are a more appropriate visualization for this purpose and show that the data is well aligned over approximately 4.5 orders of magnitude (Extended Data Figure B5d in the Appendix). As a consequence, one should confine the interpretation of the proteome profiles to abundances of >500-1,000 copies per cell (or an appropriate equivalent measure for body fluids), which is in line with e.g. the study of Beck et al.[1]. Extended Data Figure B6 a-c in the Appendix depicts the remaining eight cell line proteomes using iBAQ quantification before and after normalization (a) and the corresponding Q-Q plots (b).

Following the rationale that normalization should enable the comparison of disparate data sets, including the comparison of samples, experiments and projects based on label-free and stable isotope labeling quantification, this analysis was extended to investigate how the normalization affects the comparison of label-free and label-based approaches. Note that the MS1 intensity of light, medium or heavy labeled peptides was utilized to calculate the protein abundance estimate for every SILAC or dimethyl labeling channel separately. Similarly, for isobaric quantification, the corresponding MS2 reporter ion intensity (for iTRAQ or TMT) was used to calculate protein abundance estimates for the quantification channels separately. As exemplified in Extended Data Figure B6c in the Appendix based on the comparison of data from SILAC-labeled and label-free analyzed MCF-7 cell digests from two different studies[11,49], the protein abundance of MS1 based label-free and label-based quantification approaches exhibit a similar dynamic range of 4-5 orders of magnitude and can be as comparably re-scaled. In contrast, as exemplified using data from MCF-7 cell lines based on label-free[11] and iTRAQ quantification[56], quantification based on MS2 reporter ion signals shows drastically different intensity distribution characteristics with respect to dynamic range and symmetry of the distribution (Extended Data Figure B6d in the Appendix). Although desirable, the quantitative comparison of protein abundances between MS1 (label-free, SILAC and dimethyl labeling) and MS2 based approaches (iTRAQ, TMT) is currently not possible without introducing gross errors (or re-scaling artefacts). In light of these results, MS1 and MS2 quantification results are kept separate throughout this study as well as in ProteomicsDB. Furthermore, affinity data are, if not mentioned otherwise, not included in any quantitative comparison.

The distribution of 347 samples depicted in Extended Data Figure B6e in the Appendix underscores that the re-scaling using total sum normalization actually results in very similar abundance distributions for samples with MS1-level quantification.

Last, but not least, a commonly employed approach to illustrate that derived protein abundance metrics actually reflect protein copy numbers is to compare the abundance of proteins with a known stoichiometry within complexes. To this end, normalized iBAQ values were compared with the reported composition of nuclear pore complexes from a recent study[67] (NPC). As depicted in Extended Data Figure B9a in the Appendix, the stoichiometry of the NPC components derived from normalized iBAQ values of triplicate measurements of HeLa nuclear extracts from the same study are in very good agreement with the reported stoichiometries from a corresponding AQUA experiment (median fold error < 32%). Note that the components of the stable Nup107 sub-complex even show a median fold error of less than 10%.

In summary, the results of the technical analysis as well as the biological data shown in Figure 3.39, Figure 3.40 and Figure 3.41 indicate that the overall approach is appropriate and enables a reasonable biological interpretation of the protein expression profiles.

3.4 Core proteome and missing proteome

Aggregating the data used for building the draft proteome shows that proteome coverage rapidly saturates at ~16-17,000 proteins, which is similar to transcriptome coverage obtained by RNAseq. Addition of human tissue and body fluid data each led to small but noticeable contributions not provided by cell lines. The same is true when adding PTM or affinity data to shotgun proteomic data (Extended Data Figure B4 in the Appendix). When comparing five of the largest data sets in ProteomicsDB[11,12,15,76], the existence of a human core proteome[121] of ~10-12,000 ubiquitously

expressed proteins can be postulated the primary function of which is the general control and maintenance of cells but the low abundance range of which is, enriched in proteins with regulatory functions (Extended Data Figure B4 in the Appendix). The observed proteome saturation implies that adding more shotgun data will not considerably increase coverage albeit increasing confidence in individual proteins. Instead, it is likely, that the 'missing proteome' (Figure 3.38b) will have to be identified by more focused experimentation. It is also possible that a considerable part of the missing proteome constitutes (pseudo)genes that are no longer expressed. G-protein coupled receptors are underrepresented in ProteomicsDB and the respective transcripts are also notoriously absent in RNAseq data[105]. Earlier work suggests that more than half of the 853 human GPCRs have lost their function over the course of human evolution and may be considered obsolete[122] Similarly, a large number of functionally uncategorized proteins are annotated pseudogenes potentially further reducing the number of (actual) protein coding genes. Cytokines may be underrepresented because of experimental issues as small, secreted proteins can still be difficult to obtain from the supernatants of cells, the intercellular space of tissues or from body fluids. In order to fill the remaining gaps in the human proteome, ProteomicsDB provides a facility to engage the community by 'adopting' a missing protein, i.e. to provide mass spectrometric evidence for its existence.

In addition, reference spectra acquired from synthetic peptides for 435 peptides for all 273 cytokines as well as 3,539 further peptides for proteins not yet well covered are available in ProteomicsDB so that any identification of such proteins in the future may be validated using the synthetic reference standard.

3.5 Functional proteome expression analysis

Profiles of 27 human tissues and body fluids (human body map) complemented with publically available data were generated to begin to analyze human proteomes in functional terms. A simple common task is to compare the expression level of a single protein across many biological sources (Figure 3.39a). While housekeeping proteins such as GAPDH show high (and sometimes extreme) expression throughout, high levels of the proto-oncogene EGFR are more confined to e.g. breast (cancer) tissue. Similarly, β-catenin, a member of the Wnt pathway, is highly expressed in colon cancer cells where the protein participates in the development of the malignancy. Principle component analysis (PCA) of protein abundances in 42 proteomes shows that protein expression in a particular tissue and its corresponding cell lines is broadly similar and that there are more substantial differences between tissues of different organs (Figure 3.39b). This result is important for the interpretation of data presented further below and also contributes to the ongoing discussion regarding the suitability of cell lines as model systems for studying human biology. A comparative analysis of the 100 most highly expressed proteins in each of 47 human organs and body fluids (Figure 3.39c) revealed that ~70% of these proteins are found in common but show expression differences of up to 5 orders of magnitude. Interestingly, even the most highly abundant proteins in a tissue or fluid often point to molecular processes associated with the respective biological specialization; myofibrillar proteins including troponins are abundant in the heart, proteases in the pancreas and neuronal proteins in cerebrospinal fluid.

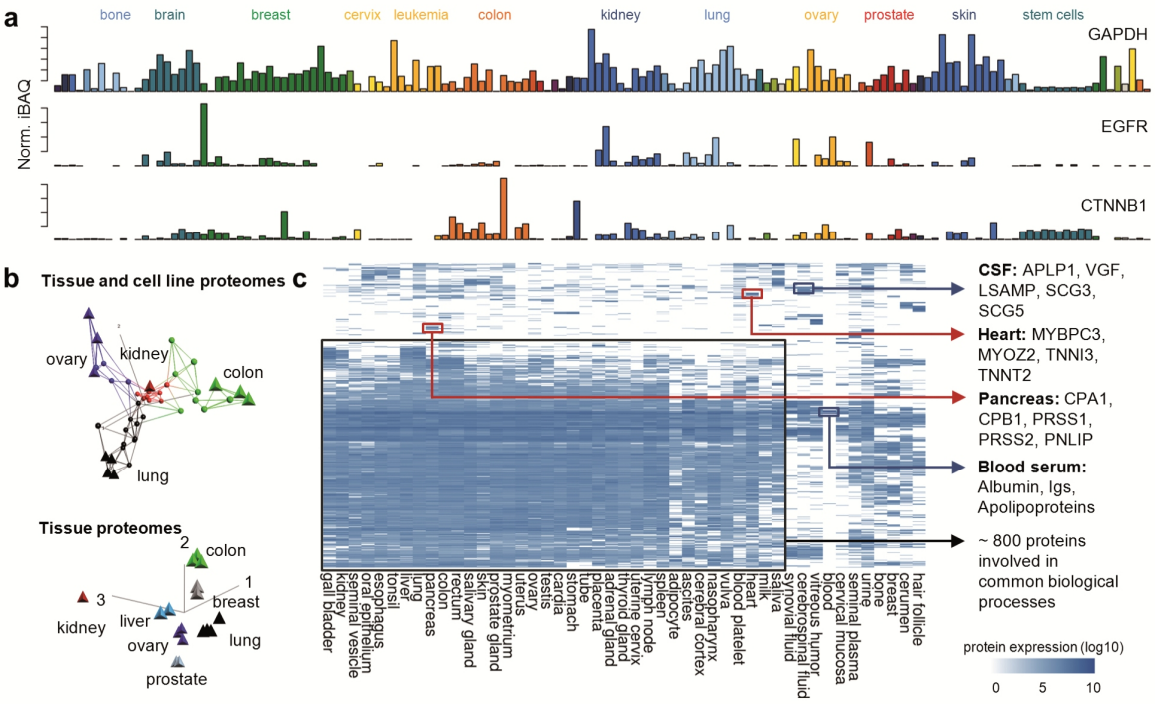


Figure 3.39 | Global protein expression analysis. a, Protein expression in different tissues and cell lines showing that levels of housekeeping (GAPDH), signaling (EGFR) and tumor-associated (CTNNB1) proteins can vary substantially between tissues (grouped by color). b, Principle component analysis showing that cell lines (circles) retain protein expression characteristics of their respective primary tissue (triangles) and that proteomes of different organs are more diverse. c, Hierarchical clustering of the 100 most highly expressed proteins from each of 47 tissues and body fluids. Despite the presence of a large group of common proteins, clusters of organ/fluid selective proteins with respective biological functions can readily be identified.

Similar observations can be made when investigating proteins forming functional classes such as protein kinases or transcription factors (TFs) (Figure 3.40). Akin to core proteomes, some of the 349 detected kinases and 557 TFs are broadly expressed, but others appear to be confined to few organs where they drive more specific processes. For instance, the kinases HCK, ZAP70, LCK, JAK3, TXK and FGR are found in a tight cluster of kinases in the spleen and all play important roles in the biology of immune cells. This is 'mirrored' by transcription factors in the same cluster with strong ties to immunity including the NFkB system (REL, PRKCH, NFKBIE) and Toll-like receptor signaling (SIGIRR, IRF5, ARRB2, NLRC4). It is noteworthy that many of the proteins in the spleen cluster are also highly expressed in the lung, a primary entry point for human pathogens. The number of proteins that are exclusively or preferentially detected in a particular organ is surprisingly small and gene ontology analysis invariably highlights organ-specific biology (Extended Data Figure B6 in the Appendix). For instance, adipocytes are rich in proteins involved in lipid storage, platelets in growth factors and placenta in proteins relating to hormonal regulation and pregnancy. The above shows that even disparate, but high-quality proteomic data can be used to construct protein expression maps across an entire complex organism. A recent report has shown that this is feasible in mice[123] but, organism wide proteome expression profiling has not been described in humans before. In addition, the identification of a considerable number of proteins with no ascribed function but exclusive (or high) expression in particular organs implies a functional role. The contextual information provided in ProteomicsDB may thus provide guidance for the eventual identification of the biological role of these orphan proteins.

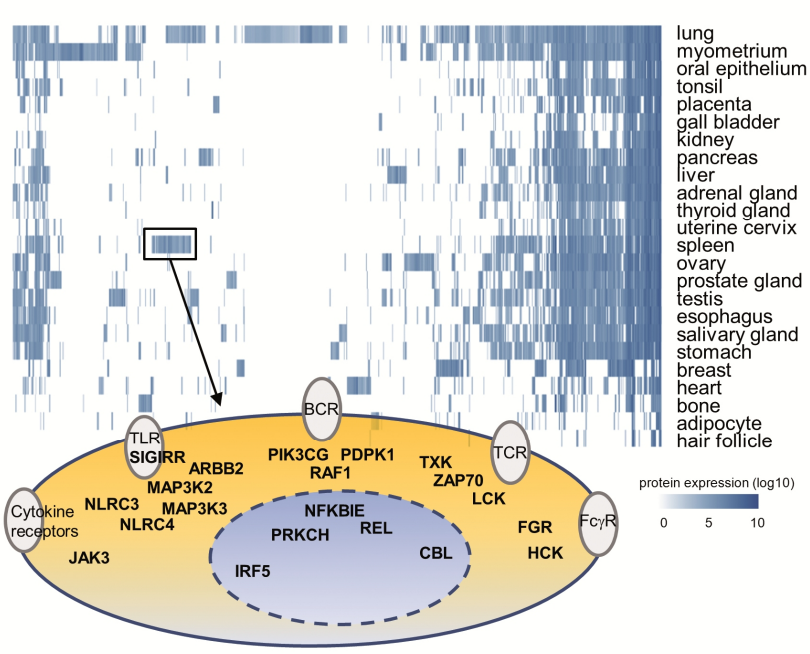


Figure 3.40 | Functional protein expression analysis. Quantitative expression analysis of 906 kinases and transcription factors (TF) across 24 tissues (top panel) identifies organ-selective signatures indicative of the underlying biology. The highlighted cluster in spleen contains the kinases LCK, ZAP70 and JAK and the TFs SIGIRR, NFKBIE and NLRC3 with strong links to the immune system (bottom panel).

3.6 Proteomic and transcriptomic correlation

The comparison of mRNA (RNAseq)[105] and protein expression profiles for 12 human tissues (Extended Data Figure B7 in the Appendix) shows clear correlations, although in all cases the Spearman's rank correlation coefficients are rather moderate and somewhat poorer than that previously reported for cell lines. This is likely due to the fact that tissues generally comprise a mixture of cell types, connective tissue and blood. Both mRNA and protein levels vary a lot between tissues as one might expect; however, the ratio of protein and mRNA levels is remarkably conserved between tissues for any given protein (Figure 3.41a)[124]. Schwanhausser *et al.*[101] have previously shown that the translation rate constant is one dominant factor determining protein abundance in cell lines. Using the ratio of protein-to-mRNA levels as a proxy for translation rates, the data indicates that this is also true for human tissues and that the ratio is similar in every tissue (Figure 3.41b). It therefore appears that the translation rate is a fundamental, encoded (constant) characteristic of a transcript suggesting that the actual amount of protein in a given cell is primarily controlled by regulating mRNA levels. Having learned the protein/mRNA ratio for every protein and transcript, it now becomes possible to predict protein abundance in any given tissue with good accuracy from the measured mRNA abundance (Figure 3.41c, Extended Data Figure B7 in the Appendix).

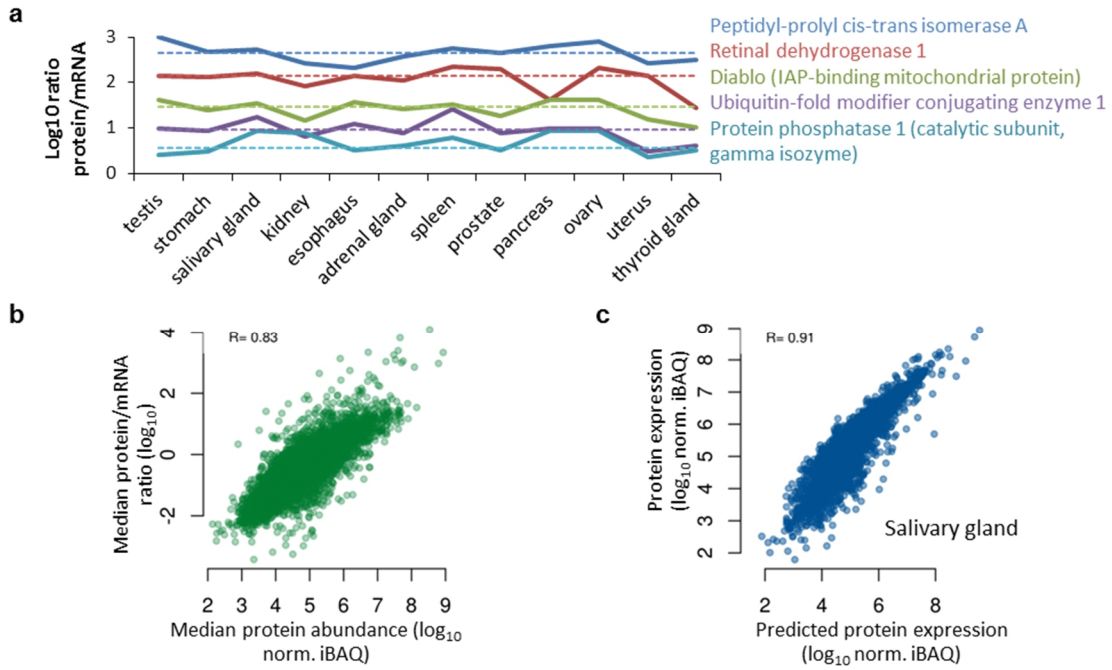


Figure 3.41 | Correlation analysis of mRNA and protein expression levels. a, Analysis of mRNA and protein levels across 12 organs shows that the protein/mRNA ratio is largely conserved. b, The median translation rates of all transcripts across all tissues correlate well with protein abundance c, leading to the ability to predict individual protein levels from the respective mRNA levels.

3.7 Analysis of protein expression and drug sensitivity

It was previously shown that protein expression can be correlated to drug sensitivity[12]. Similarly, here this method was applied to discover sensitivity/resistance markers for 24 drugs in 35 human cancer cell lines using drug sensitivity data provided by the cancer cell line encyclopedia (CCLE)[108]. For instance, analysis of the EGFR kinase inhibitors Erlotinib and Lapatinib identified a number of common proteins associated with drug sensitivity or resistance (Figure 3.42, Extended Data Figure B8 in the Appendix). The primary target (EGFR) as well as annexin A1 (ANXA1, a direct EGFR substrate), and EGFR interacting proteins at stress fibers (PDLIM, KRT5, KRT14) all indicate drug sensitivity while high expression of ANXA6 or S100A4 renders cells less responsive. Consequently, knock-down of ANXA6 in BT549 cells has been shown to sensitize cells for lapatinib[125] and addition of S100A4 to cells in culture has been shown to stimulate EGFR and to promote metastasis[126]. High expression of S100 proteins is often associated with resistance against kinase inhibitors, suggesting that S100 overexpression may be a general molecular resistance mechanism. The data further suggest that similar effects can be postulated for the Zn-finger protein THAP2, the NAD(P) transhydrogenase NNT and Dermcidin (DCD). Likewise, high expression of MED11 (part of the RNA Pol II mediator complex), IFI35 (an interferon induced protein of unknown function), HECTD1 (a E3 ubiquitin ligase) and CDRT1 (an orphan F-box protein) should promote drug sensitivity but their molecular connections to EGFR are not clear at present. In light of a recent report showing increased phosphorylation of HECTD1 upon EGF treatment[127], it is tempting to speculate that a HECTD1/CDRT1 complex may be involved in regulating the stability of EGFR via the ubiquitin/proteasome system.

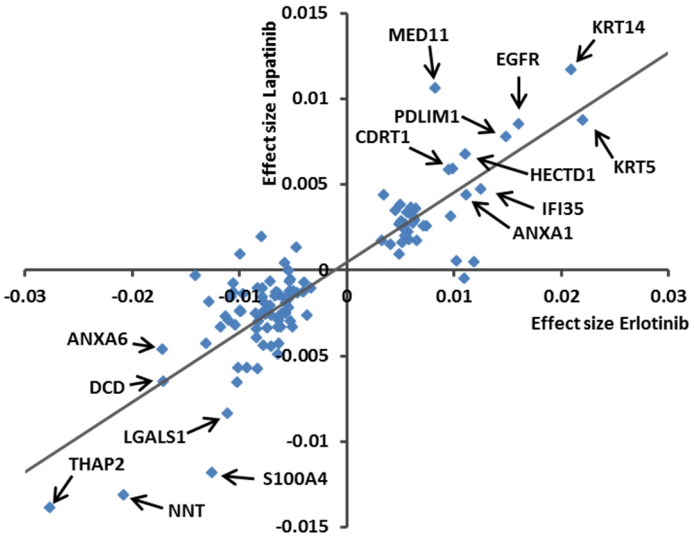


Figure 3.42 | Drug sensitivity and resistance analysis. Elastic net analysis for the identification of drug sensitivity (positive effect size) or resistance (negative effect size) markers against the EGFR kinase inhibitors Erlotinib and Lapatinib in cancer cell lines.

3.8 Composition and stoichiometry of protein complexes

The composition and stoichiometry of protein complexes is typically analyzed by affinity purification coupled to MS-based protein analysis and it emerges that protein expression profiling may also have potential for this purpose[52]. Stoichiometries measured by iBAQ for the nuclear pore complex agreed well with a prior study using absolute protein quantification by spiked peptide standards (Extended Data Figure B9 in the Appendix)[67]. Using the proteasome as an example, its composition and stoichiometry was explored across cell lines and tissues (Figure 3.43). The constitutive core proteasome consists of 2x7 non-catalytic alpha and 2x7 catalytic beta subunits but e.g. an 'immunoproteasome' has been identified in which the β1,2 and 5 subunits are replaced by homologous proteins (β1i, β2i and β5i) in immune cells[128,129]. This analysis shows that the proteasome in the salivary gland is primarily of the constitutive type and that lymph nodes almost exclusively contain the immunoproteasome (Figure 3.43a and b). The same analysis across >100 cell line and tissue samples (Figure 3.43c) reveals that the immunoproteasome is surprisingly widely expressed including in tissues for which no primary immunological function would be expected. In addition, the data implies that the molecular composition and stoichiometry of proteasomes is heterogeneous and cell type dependent. Correlation analysis of the expression of all beta subunits (Figure 3.43d) strongly suggests that the β 1, 2 and 5 subunits and their respective immunoproteasome counterparts are expressed independently (no correlation). In contrast, it appears that the remaining subunits (β 3, 4, 6, 7) are co-expressed with either group.

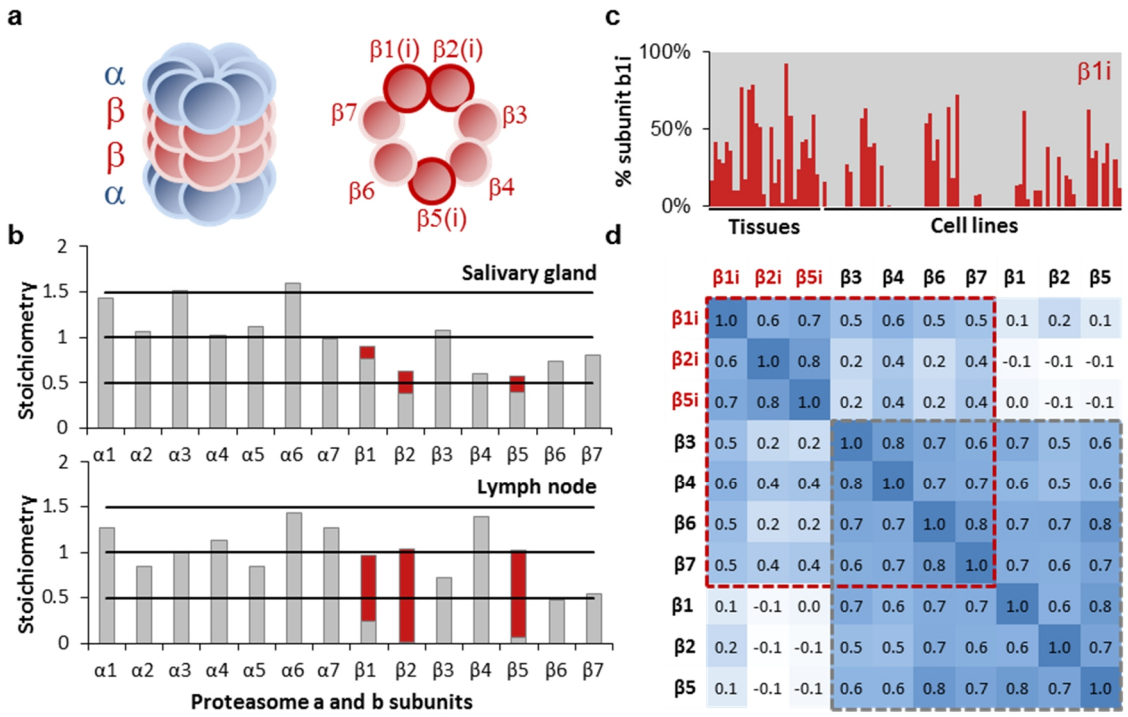


Figure 3.43 | Analysis of the composition and stoichiometry of the proteasome. a, Schematic structure of the 'constitutive' proteasome and the 'immunoproteasome' (marked by the suffix 'i'). b, Stoichiometry derived by iBAQ quantification of the constitutive proteasome (grey) and the immunoproteasome (red) in the salivary gland and the lymph node. c, Expression analysis of the ***β***1 subunit across >100 tissue and cell line proteomes reveals that many cells express both forms of the proteasome. d, Expression correlation analysis of all ***α***/***β*** subunits across the said tissues and cell lines showing strong co-expression of the ***β***1i, ***β***2i and ***β***5i subunits as well as all other ***α***/***β*** subunits but no correlation with the expression of the corresponding ***β***1, ***β***2 and ***β***5 subunits.

3.9 Post-translational modifications

Proteomic data collections can be valuable data mines for post-translational modification analysis or developing proteome technology. ProteomicsDB currently contains 81,721 unique phosphorylated peptides representing 11,025 human genes, demonstrating that more than half of all human proteins are substrates of kinases. Similarly, there are 29,031 unique ubiquitinylated peptides from 5,769 proteins representing substrates of ubiquitin ligases as well as 16,693 acetylated peptides from 7,098 proteins that are substrates of acetylases. This analysis also detected N-terminal peptides for 7,977 proteins and C-terminal peptides for 6,778 proteins confirming a large number of translation start and stop sites (Extended Data Figure B10a in the Appendix).

3.10 Proteotypic peptides and targeted proteomics

So-called 'proteotypic' peptides[94] have proven useful as quantification standards in targeted proteomic measurements which are increasingly employed to develop e.g. clinical biomarker assays[130]. Briefly, experimental proteotypicity of a single peptide counts how often this particular peptide was observed when the protein was identified. The cumulative proteotypicity describes how often a protein was identified using either one of the top most frequently identified peptides. Example: one particular peptide may have been observed in 50% of all cases that the protein was identified. Another peptide may have also been observed in 50% of all cases. However, these two

peptides are not necessarily observed together. The cumulative proteotypicity of the two peptides may therefore vary from 50% (both peptides observed together every time a protein is identified) to 100% (both peptides never observed together when a protein is identified). The experimental proteotypicity of a single peptide is very useful e.g. for selecting peptides for SRM/MRM experiments. The cumulative proteotypicity can be used in a similar fashion (i.e. deciding about how many AQUA peptides to synthesize for a given protein) but also offers an explanation why e.g. the top3 intensity method works well for quantifying a protein, since it turns out that very few peptides are responsible for almost all identifications of a particular protein. An additional use is that the detection of a proteotypic peptide gives confidence in protein identifications based on single/few peptides (either because a particular protein may not produce more than one/few MS compatible peptides upon protease digestion or the protein is of low abundance in which case the detection of a proteotypic peptide is much more likely than the detection of any other peptide).
ProteomicsDB enabled the determination of the proteotypicity of ~500,000 peptides and expanded the concept to chemically labeled peptides (Extended Data Figure B10b in the Appendix). The 71 million peptide precursor ion and 18 billion peptide fragment ion measurements allow the computational assessment of the specificity of targeted measurements ahead of the actual experiment. Exemplified by the peptide LHYGLPVVVK of the proto-oncogene ***β***-catenin (Figure 3.44a, Extended Data Figure B10c in the Appendix), mining of ProteomicsDB revealed a large number of potentially interfering peptides that may distort the quantification of the target peptide. Interference can be substantially reduced by high resolution instruments (Figure 3.44b)[131] and by limiting the allowed interferences to the tissue in question (Figure 3.44c). Likely, the combination of experimental proteotypicity, interference estimation and high resolution instrumentation will provide for more robust targeted proteomic assays in the future.

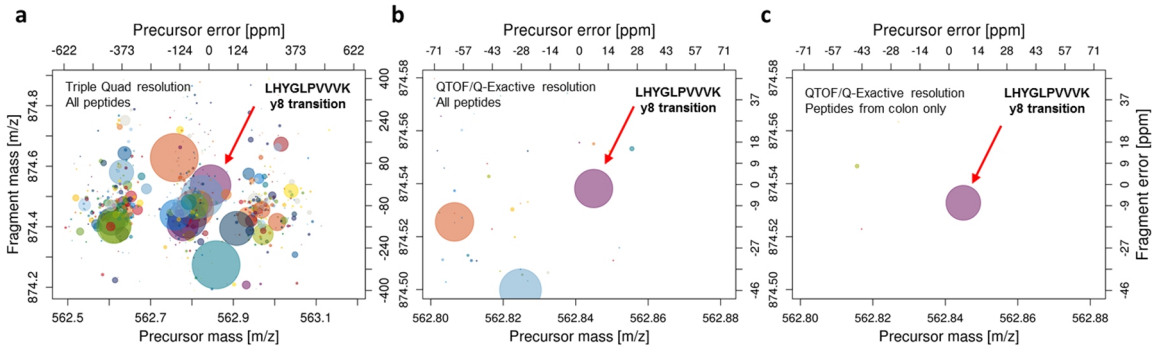


Figure 3.44 | Computation of molecular interferences in targeted experiments. The transition of the target peptide LHYGLPVVVK (y8 fragment ion, ***β***-catenin) is marked with an arrow. All other circles in the plot are interfering SRM transitions of other peptides found in ProteomicsDB that fall within the same mass tolerance. a, The y8 transition of the peptide LHYGLPVVVK (b-catenin, marked with an arrow) in a 0.7/0.7 Da slice of the precursor and fragment ion window typically employed on triple quadrupole mass spectrometers. The size of the circle represents the relative intensity of the y8 fragment in a full tandem mass spectrum of this peptide. All other circles are interfering peptides (extracted from the entire ProteomicsDB) that have precursor and fragment ions in the same m/z window and with varying intensities (circle size). b, Interference can be reduced by using high resolution MS and confining the analysis to the tissue in question (here colon, c). Such interference plots in conjunction with the proteotypicity of peptides can be valuable for the design of targeted proteomic experiments.

4 Discussion

This study showed that an extensive draft of the human proteome can be assembled from disparate but high quality proteomic data. However, similar to the evolution of the human genome projects, the eventual completion of the human proteome will take further time and effort but will also lead to substantial improvements in technology which is still needed. One main issue to address is proteome coverage and resolution. While DNA/RNA sequencing technologies have attained single nucleotide resolution, the amino acid coverage of proteins is still poor, which currently impairs our ability to detect e.g. splice variants, PTMs, mutations or isoforms in a systematic fashion. A related challenge is to improve the ability to sample a proteome comprehensively i.e. 'all proteins, all the time'. Both challenges suffer from the large dynamic range of expression of both proteins and proteoforms, the lack of a technique similar to PCR and the destructive measurement process.

While some of the many applications that can be envisaged for the use of this collection and some of the biological insights that may be generated by mining the proteome were shown, another important area of future research concerns overcoming the uncertainties associated with peptide/protein identification by sequence database searching[132]. As shown for some proteins, ProteomicsDB offers the functionality to store and compare experimentally observed spectra against reference spectra from synthetic peptides. Expanding this to all human proteins not only allows the systematic comparison of such spectra but also allows a more unbiased guidance when generating targeted assays. ProteomicsDB and similar resources have a role to play in these challenges as the data assembled will enable the development of computational tools and lab reagents facilitating proteome wide discovery experiments, multiplexed quantitative protein assays as well as the general exploration of the human proteome.

The estimation of protein false discovery rates in very large proteomics experiments is a challenge for which no satisfactory solution has been found yet. The rapidly saturating number of true positive identifications and the slow but steady growth of false positive identifications when aggregating more and more experiments leads to the paradoxical situation that, eventually, all true and false positive proteins will have been observed resulting in a protein FDR of up to 100%. This is almost independent of the previous cutoffs applied. One view is that protein FDRs are actually not very meaningful because proteomics measures peptides not proteins and the definition of a 'decoy protein' is quite problematic. Another view is that one should lower the peptide/PSM FDR to the degree necessary to reach a desired protein FDR[71]. The downside of this approach is that many valid identifications are removed rendering this approach more conservative than necessary. Then there is the 'middle ground' represented by e. g. the MAYU approach that attempts to overcome the scaling issues associated with the classical protein FDR approach. Unfortunately, these approaches also suffer from scaling issues and are generally not designed to allow an online and real time adjustment since they require a complete re-computation when more data is added.

5 Outlook

ProteomicsDB has been proven to facilitate the exploration and analysis of the first draft of the human proteome. However at its current state, it covers mostly classical full proteomes. As technologies and sample preparation methods become more advanced and provide access to specific sub-proteomes and other omics data, additional data types will provide additional insights. For example PTMs provide an orthogonal view, as they directly manipulate activation, degradation and stability of proteins and thus are of high interest when understanding e.g. resistance or sensitivity mechanisms to drug. Even though the function of many modification sites is not yet known, even a comparatively simple catalogue of occurrences and their abundances will already broaden our understanding of their function as ubiquitous sites are likely not of functional importance. The integration of e.g. phosphorylation-, ubiquitinylation- and acetylation-datasets will help to build a future version of the human proteome that provides a more direct link between protein expression and activity.

Of similar importance is the combination of multi omics technologies. Each of the four main omics areas, namely genomics, transcriptomics, proteomics and metabolomics, provides unique and orthogonal information not accessible with a single method alone. Therefore, the integration of data across all omics levels is expected to give deeper biological insight. Many challenges lie ahead of this, as for most combinations it is not yet established how to combine them properly. For example, only some metabolites are annotated and assigned to a process and proteins. Quantitative co-occurrence analysis could provide first clues which biological entities influence each other.

As shown here, the correlation of phenotypic data, such as drug sensitivity data, to protein expression estimates can be used to find potential markers for resistance and sensitivity. The incorporation of the activation status of proteins or the presence of specific mutations will likely increase the accuracy of such models. This could have a direct impact on medicine as personalized treatments depending on the molecular footprint of cancer cells can be used to predict the best combination of drugs. Additionally, the integrated analysis of multiple omics data is expected to provide better (multivariate) biomarkers which are able to differentiate cancer or even different subtypes of cancer. This assumption is driven by the fact that each omics technology excels others at specific tasks, e.g. genomics is very good at measuring mutations and proteomics is good at measuring expression. Furthermore, the integration of longitudinal studies and more replicates of tissues, cell lines and fluids will help to quantify the variance between or within one biological source.

Last but not least, the extension to other organisms, such as mice, rats, mini pigs and other model organisms used in the pharmacological industry, will provide unique opportunities to compare proteomes of different organisms to judge and quantify the comparability of tox and disease models. This could enable a translation of the pharmacokinetics and pharmacodynamics into humans. However, this relies on the integration of additional data such as signaling pathways, metabolic networks and activation status of proteins.

6 Author Contribution

This chapter is based on[92].
Mathias Wilhelm developed and implemented the database model, designed and implemented the backend, designed and developed the frontend of ProteomicsDB, performed data analysis and data interpretation.

7 Abbreviations

API	Application programming interface
CID	Collision induced dissociation
ESI	Electrospray ionization
FDR	False discovery rate
HCD	Higher energy CID
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
lincRNA	Long intergenic non-coding RNA
m/z	mass to charge ratio
MS	Mass Spectrometry
nESI	nano-ESI
NPC	Nuclear pore complex
PCA	Principle component analysis
PEP	Posterior error probability
PRM	Parallel reaction monitoring
PSM	Peptide spectrum match
PTM	Post translational modification
QTOF	Quadrupole time-of-flight
SRM	Selected reaction monitoring
TF	Transcription factor
TMT	Tandem mass tag
TUCP	Transcript of uncertain coding potential

8 References

1 Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol Syst Biol* 7, 549, doi:10.1038/msb.2011.82 (2011).

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

3 UniProt, C. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* 41, D43-47, doi:10.1093/nar/gks1068 (2013).

4 Paik, Y. K. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotechnol* 30, 221-223, doi:10.1038/nbt.2152 (2012).

5 Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28, 1248-1250, doi:10.1038/nbt1210-1248 (2010).

6 Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32, 223-226, doi:10.1038/nbt.2839 (2014).

7 Kim, M. S. *et al.* A draft map of the human proteome. *Nature* 509, 575-581, doi:10.1038/nature13302 (2014).

8 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419, doi:10.1126/science.1260419 (2015).

9 Farrah, T. *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J Proteome Res* 13, 60-75, doi:10.1021/pr4010037 (2014).

10 Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* 11, 492-500, doi:10.1074/mcp.O111.014704 (2012).

11 Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* 11, M111 014050, doi:10.1074/mcp.M111.014050 (2012).

12 Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* 4, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).

13 Bantscheff, M. *et al.* Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nat Biotechnol* 29, 255-265, doi:10.1038/nbt.1759 (2011).

14 Joshi, P. *et al.* The functional interactome landscape of the human histone deacetylase family. *Mol Syst Biol* 9, 672, doi:10.1038/msb.2013.26 (2013).

15 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat Methods* 10, 634-637, doi:10.1038/nmeth.2518 (2013).

16 Hahne, H. *et al.* Proteome wide purification and identification of O-GlcNAc-modified proteins using click chemistry and mass spectrometry. *J Proteome Res* 12, 927-936, doi:10.1021/pr300967y (2013).

17 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, doi:10.1038/nbt.1511 (2008).

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

19 Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* 2, 1896-1906, doi:10.1038/nprot.2007.261 (2007).

20 Ritorto, M. S., Cook, K., Tyagi, K., Pedrioli, P. G. & Trost, M. Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes. *J Proteome Res* 12, 2449-2457, doi:10.1021/pr301011r (2013).

21 Hahne, H. *et al.* DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat Methods* 10, 989-991, doi:10.1038/nmeth.2610 (2013).

22 Hermjakob, H. & Apweiler, R. The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev Proteomics* 3, 1-3, doi:10.1586/14789450.3.1.1 (2006).

23 *ProteomeXchange*, <http://www.proteomexchange.org/>

24 Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Res* 34, D655-658, doi:10.1093/nar/gkj040 (2006).

25 *PeptideAtlas*, <http://www.peptideatlas.org/>

26 Hill, J. A., Smith, B. E., Papoulias, P. G. & Andrews, P. C. ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. *J Proteome Res* 9, 2809-2811, doi:10.1021/pr1000972 (2010).

27 *Proteome Commons (Tranche)*, <https://www.proteomecommons.org/tranche/downloads.jsp>

28 *MassIVE*, <http://proteomics.ucsd.edu/ProteoSAFe/datasets.jsp>

29 Tao, F. 1st NCI annual meeting on Clinical Proteomic Technologies for Cancer. *Expert Rev Proteomics* 5, 17-20, doi:10.1586/14789450.5.1.17 (2008).

30 *Clinical Proteomic Tumor Analysis Consortium Data Portal*, <https://cptac-data-portal.georgetown.edu/cptacPublic/>
31 *Broad Institute Proteomics FTP server*, <ftp://ftp.broadinstitute.org/distribution/proteomics/public_datasets/>
32 Phanstiel, D. H. *et al.* Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat Methods* 8, 821-827, doi:10.1038/nmeth.1699 (2011).
33 *Stem Cell Omics Repository*, <http://scor.chem.wisc.edu/>
34 *PRIDE PRoteomics IDEntifications database*, <http://www.ebi.ac.uk/pride/>
35 Addona, T. A. *et al.* A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. *Nat Biotechnol* 29, 635-643, doi:10.1038/nbt.1899 (2011).
36 Arabi, A. *et al.* Proteomic screen reveals Fbw7 as a modulator of the NF-kappaB pathway. *Nat Commun* 3, 976, doi:10.1038/ncomms1975 (2012).
37 Bergamini, G. *et al.* A selective inhibitor reveals PI3Kgamma dependence of T(H)17 cell differentiation. *Nat Chem Biol* 8, 576-582, doi:10.1038/nchembio.957 (2012).
38 Berle, M. *et al.* Quantitative proteomics comparison of arachnoid cyst fluid and cerebrospinal fluid collected perioperatively from arachnoid cyst patients. *Fluids Barriers CNS* 10, 17, doi:10.1186/2045-8118-10-17 (2013).
39 Bhattacharjee, M. *et al.* A multilectin affinity approach for comparative glycoprotein profiling of rheumatoid arthritis and spondyloarthropathy. *Clin Proteomics* 10, 11, doi:10.1186/1559-0275-10-11 (2013).
40 Burgener, A. *et al.* Comprehensive proteomic study identifies serpin and cystatin antiproteases as novel correlates of HIV-1 resistance in the cervicovaginal mucosa of female sex workers. *J Proteome Res* 10, 5139-5149, doi:10.1021/pr200596r (2011).
41 Burkhart, J. M. *et al.* The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* 120, e73-82, doi:10.1182/blood-2012-04-416594 (2012).
42 Casado, P. *et al.* Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci Signal* 6, rs6, doi:10.1126/scisignal.2003573 (2013).
43 Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293-1307, doi:10.1016/j.cell.2012.02.009 (2012).
44 Cho, C. K. *et al.* Quantitative proteomic analysis of amniocytes reveals potentially dysregulated molecular networks in Down syndrome. *Clin Proteomics* 10, 2, doi:10.1186/1559-0275-10-2 (2013).
45 Dawson, M. A. *et al.* Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* 478, 529-533, doi:10.1038/nature10509 (2011).
46 Didangelos, A. *et al.* Extracellular matrix composition and remodeling in human abdominal aortic aneurysms: a proteomics approach. *Mol Cell Proteomics* 10, M111 008128, doi:10.1074/mcp.M111.008128 (2011).
47 Farina, A. *et al.* Bile carcinoembryonic cell adhesion molecule 6 (CEAM6) as a biomarker of malignant biliary stenoses. *Biochim Biophys Acta*, doi:10.1016/j.bbapap.2013.06.010 (2013).
48 Fernandez-Saiz, V. *et al.* SCFFbxo9 and CK2 direct the cellular response to growth factor withdrawal via Tel2/Tti1 degradation and promote survival in multiple myeloma. *Nat Cell Biol* 15, 72-81, doi:10.1038/ncb2651 (2013).
49 Geiger, T., Madden, S. F., Gallagher, W. M., Cox, J. & Mann, M. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer Res* 72, 2428-2439, doi:10.1158/0008-5472.CAN-11-3711 (2012).
50 Giansanti, P., Stokes, M. P., Silva, J. C., Scholten, A. & Heck, A. J. Interrogating cAMP-dependent kinase signaling in jurkat T cells via a protein kinase A targeted immune-precipitation phosphoproteomics approach. *Mol Cell Proteomics* 12, 3350-3359, doi:10.1074/mcp.O113.028456 (2013).
51 Guo, H. *et al.* Integrative network analysis of signaling in human CD34(+) hematopoietic progenitor cells by global phosphoproteomic profiling using TiO2 enrichment combined with 2D LC-MS/MS and pathway mapping. *Proteomics* 13, 1325-1333, doi:10.1002/pmic.201200369 (2013).
52 Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* 150, 1068-1081, doi:10.1016/j.cell.2012.08.011 (2012).
53 Hennrich, M. L., Groenewold, V., Kops, G. J., Heck, A. J. & Mohammed, S. Improving depth in phosphoproteomics by using a strong cation exchange-weak anion exchange-reversed phase multidimensional separation approach. *Anal Chem* 83, 7137-7143, doi:10.1021/ac2015068 (2011).
54 Hennrich, M. L., van den Toorn, H. W., Groenewold, V., Heck, A. J. & Mohammed, S. Ultra acidic strong cation exchange enabling the efficient enrichment of basic phosphopeptides. *Anal Chem* 84, 1804-1808, doi:10.1021/ac203303t (2012).
55 Iwata, K. *et al.* The human oligodendrocyte proteome. *Proteomics* 13, 3548-3553, doi:10.1002/pmic.201300201 (2013).
56 Johansson, H. J. *et al.* Retinoic acid receptor alpha is associated with tamoxifen resistance in breast cancer. *Nat Commun* 4, 2175, doi:10.1038/ncomms3175 (2013).
57 Kentsis, A. *et al.* Urine proteomics for profiling of human disease using high accuracy mass spectrometry. *Proteomics Clin Appl* 3, 1052-1061, doi:10.1002/prca.200900008 (2009).

58 Kliemt, S. *et al.* Sulfated hyaluronan containing collagen matrices enhance cell-matrix-interaction, endocytosis, and osteogenic differentiation of human mesenchymal stromal cells. *J Proteome Res* 12, 378-389, doi:10.1021/pr300640h (2013).

59 Kruse, U. *et al.* Chemoproteomics-based kinome profiling and target deconvolution of clinical multi-kinase inhibitors in primary chronic lymphocytic leukemia cells. *Leukemia* 25, 89-100, doi:10.1038/leu.2010.233 (2011).

60 Larance, M., Ahmad, Y., Kirkwood, K. J., Ly, T. & Lamond, A. I. Global subcellular characterization of protein degradation using quantitative proteomics. *Mol Cell Proteomics* 12, 638-650, doi:10.1074/mcp.M112.024547 (2013).

61 Maier, S. K. *et al.* Comprehensive identification of proteins from MALDI imaging. *Mol Cell Proteomics* 12, 2901-2910, doi:10.1074/mcp.M113.027599 (2013).

62 Munoz, J. *et al.* The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *EMBO J* 31, 3079-3091, doi:10.1038/emboj.2012.166 (2012).

63 Munoz, J. *et al.* The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol Syst Biol* 7, 550, doi:10.1038/msb.2011.84 (2011).

64 Muraoka, S. *et al.* In-depth membrane proteomic study of breast cancer tissues for the generation of a chromosome-based protein list. *J Proteome Res* 12, 208-213, doi:10.1021/pr300824m (2013).

65 Nagaprashantha, L. D. *et al.* Proteomic analysis of signaling network regulation in renal cell carcinomas with differential hypoxia-inducible factor-2alpha expression. *PLoS One* 8, e71654, doi:10.1371/journal.pone.0071654 (2013).

66 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7, 548, doi:10.1038/msb.2011.81 (2011).

67 Ori, A. *et al.* Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol Syst Biol* 9, 648, doi:10.1038/msb.2013.4 (2013).

68 Marimuthu, A. *et al.* SILAC-based quantitative proteomic analysis of gastric cancer secretome. *Proteomics Clin Appl* 7, 355-366, doi:10.1002/prca.201200069 (2013).

69 Papachristou, E. K. *et al.* The shotgun proteomic study of the human ThinPrep cervical smear using iTRAQ mass-tagging and 2D LC-FT-Orbitrap-MS: the detection of the human papillomavirus at the protein level. *J Proteome Res* 12, 2078-2089, doi:10.1021/pr301067r (2013).

70 Paulo, J. A. *et al.* Proteomic analysis (GeLC-MS/MS) of ePFT-collected pancreatic fluid in chronic pancreatitis. *J Proteome Res* 11, 1897-1912, doi:10.1021/pr2011022 (2012).

71 Farrah, T. *et al.* The state of the human proteome in 2012 as viewed through PeptideAtlas. *J Proteome Res* 12, 162-171, doi:10.1021/pr301012j (2013).

72 Pirmoradian, M. *et al.* Rapid and deep human proteome analysis by single-dimension shotgun proteomics. *Mol Cell Proteomics* 12, 3330-3338, doi:10.1074/mcp.O113.028787 (2013).

73 Prewitz, M. C. *et al.* Tightly anchored tissue-mimetic matrices as instructive stem cell microenvironments. *Nat Methods* 10, 788-794, doi:10.1038/nmeth.2523 (2013).

74 Rhee, H. W. *et al.* Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* 339, 1328-1331, doi:10.1126/science.1230593 (2013).

75 Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics* 12, 2341-2353, doi:10.1074/mcp.O113.028142 (2013).

76 Shiromizu, T. *et al.* Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project. *J Proteome Res* 12, 2414-2421, doi:10.1021/pr300825v (2013).

77 Werner, T. *et al.* High-resolution enabled TMT 8-plexing. *Anal Chem* 84, 7188-7194, doi:10.1021/ac301553x (2012).

78 Wu, Z. *et al.* Quantitative chemical proteomics reveals new potential drug targets in head and neck cancer. *Mol Cell Proteomics* 10, M111 011635, doi:10.1074/mcp.M111.011635 (2011).

79 Wu, Z., Moghaddas Gholami, A. & Kuster, B. Systematic identification of the HSP90 candidate regulated proteome. *Mol Cell Proteomics* 11, M111 016675, doi:10.1074/mcp.M111.016675 (2012).

80 Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* 499, 79-82, doi:10.1038/nature12223 (2013).

81 Farrah, T. *et al.* A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics* 10, M110 006353, doi:10.1074/mcp.M110.006353 (2011).

82 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, doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (1999).

83 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10, 1794-1805, doi:10.1021/pr101065j (2011).

84 *UniProt release 2011_05*, <http://www.uniprot.org/news/2011/05/03/release>

85 *UniProtKB FAQs: What are complete proteomes?*, <http://www.uniprot.org/faq/15>

86 *The Global Proteome Machine > common Repository of Adventitious Proteins*, <http://www.thegpm.org/cRAP/index.html>

87 Brosch, M., Yu, L., Hubbard, T. & Choudhary, J. Accurate and sensitive peptide identification with Mascot Percolator. *J Proteome Res* 8, 3176-3181, doi:10.1021/pr800982s (2009).

88 Kall, 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, doi:10.1038/nmeth1113 (2007).

89 Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25, 1915-1927, doi:10.1101/gad.17446611 (2011).

90 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

91 Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell. Proteomics* 10, M110 003830, doi:M110.003830 [pii] 10.1074/mcp.M110.003830 (2011).

92 Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582-587, doi:10.1038/nature13319 (2014).

93 Lane, L. *et al.* Metrics for the Human Proteome Project 2013-2014 and strategies for finding missing proteins. *J Proteome Res* 13, 15-20, doi:10.1021/pr401144x (2014).

94 Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 25, 125-131, doi:nbt1275 [pii] 10.1038/nbt1275 (2007).

95 Hahne, H. & Kuster, B. A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J. Am. Soc. Mass Spectrom.* 22, 931-942, doi:10.1007/s13361-011-0107-y (2011).

96 Wenschuh, H. *et al.* Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Biopolymers* 55, 188-206, doi:10.1002/1097-0282(2000)55:3<188::AID-BIP20>3.0.CO;2-T (2000).

97 Frank, R. & Overwin, H. SPOT synthesis. Epitope analysis with arrays of synthetic peptides prepared on cellulose membranes. *Methods Mol Biol* 66, 149-169, doi:10.1385/0-89603-375-9:149 (1996).

98 Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* 39, D507-513, doi:10.1093/nar/gkq968 (2011).

99 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, doi:10.1038/nmeth1019 (2007).

100 Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* 8, 2405-2417, doi:10.1074/mcp.M900317-MCP200 (2009).

101 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* 473, 337-342, doi:10.1038/nature10098 (2011).

102 Team, R. D. C. A Language and Environment for Statistical Computing. (2012).

103 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44-57, doi:10.1038/nprot.2008.211 (2009).

104 Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6, e21800, doi:10.1371/journal.pone.0021800 (2011).

105 Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Mol Cell Proteomics* 13, 397-406, doi:10.1074/mcp.M113.035600 (2014).

106 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).

107 Zou & Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* 67, 199-320 (2005).

108 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603-607, doi:10.1038/nature11003 (2012).

109 Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J Am Soc Mass Spectrom* 22, 1111-1120, doi:10.1007/s13361-011-0139-3 (2011).

110 Higdon, R. *et al.* IPM: An integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *J Proteomics* 75, 116-121, doi:10.1016/j.jprot.2011.06.003 (2011).

111 Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480, 254-258, doi:10.1038/nature10575 (2011).

112 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* 489, 101-108, doi:10.1038/nature11233 (2012).

113 Banfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* 22, 1646-1657, doi:10.1101/gr.134767.111 (2012).

114 Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240-251, doi:10.1016/j.cell.2013.06.009 (2013).

115 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, 789-802, doi:10.1016/j.cell.2011.10.002 (2011).

116 Flintoft, L. Non-coding RNA: Ribosomes, but no translation, for lincRNAs. *Nat Rev Genet* 14, 520, doi:10.1038/nrg3534 (2013).

117 Malmstrom, J. *et al.* Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* 460, 762-765, doi:10.1038/nature08184 (2009).

118 Ahrne, E., Molzahn, L., Glatter, T. & Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* 13, 2567-2578, doi:10.1002/pmic.201300135 (2013).

119 Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* 5, 144-156, doi:10.1074/mcp.M500230-MCP200 (2006).

120 Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* 4, 222, doi:10.1038/msb.2008.61 (2008).

121 Schirle, M., Heurtier, M. A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* 2, 1297-1305, doi:10.1074/mcp.M300087-MCP200 (2003).

122 Hughes, G. M., Teeling, E. C. & Higgins, D. G. Loss of olfactory receptor function in hominin evolution. *PLoS One* 9, e84714, doi:10.1371/journal.pone.0084714 (2014).

123 Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol Cell Proteomics* 12, 1709-1722, doi:10.1074/mcp.M112.024919 (2013).

124 Low, T. Y. *et al.* Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep* 5, 1469-1478, doi:10.1016/j.celrep.2013.10.041 (2013).

125 Koumangoye, R. B. *et al.* Reduced annexin A6 expression promotes the degradation of activated epidermal growth factor receptor and sensitizes invasive breast cancer cells to EGFR-targeted tyrosine kinase inhibitors. *Mol Cancer* 12, 167, doi:10.1186/1476-4598-12-167 (2013).

126 Klingelhofer, J. *et al.* Epidermal growth factor receptor ligands as new extracellular targets for the metastasis-promoting S100A4 protein. *FEBS J* 276, 5936-5948, doi:10.1111/j.1742-4658.2009.07274.x (2009).

127 Argenzio, E. *et al.* Proteomic snapshot of the EGF-induced ubiquitin network. *Mol Syst Biol* 7, 462, doi:10.1038/msb.2010.118 (2011).

128 Hisamatsu, H. *et al.* Newly identified pair of proteasomal subunits regulated reciprocally by interferon gamma. *J Exp Med* 183, 1807-1816 (1996).

129 Nandi, D., Jiang, H. & Monaco, J. J. Identification of MECL-1 (LMP-10) as the third IFN-gamma-inducible proteasome subunit. *J Immunol* 156, 2361-2364 (1996).

130 Domon, B. Considerations on selected reaction monitoring experiments: implications for the selectivity and accuracy of measurements. *Proteomics Clin Appl* 6, 609-614, doi:10.1002/prca.201200111 (2012).

131 Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol Cell Proteomics* 11, 1709-1723, doi:10.1074/mcp.O112.019802 (2012).

132 Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol* 31, 557-564, doi:10.1038/nbt.2585 (2013).

Chapter 4

A scalable approach for protein false discovery rate estimation

Contents

"The discovery of truth is prevented more effectively, not by the false appearance things present and which mislead into error, not directly by weakness of the reasoning powers, but by preconceived opinion, by prejudice."

- Arthur Schopenhauer

"All generalizations are false, including this one."

- Mark Twain

1 Introduction

Shotgun proteomics is the most popular approach for large-scale identification and quantification of proteins. The rapid evolution of high-end mass spectrometers in recent years[1-5] has made proteomic studies feasible that identify and quantify as many as 10,000 proteins in a sample[6-8]. This enables many lines of new scientific research including, for example, the analysis of many human proteomes, and proteome-wide protein-drug interaction studies[9-11]. One fundamental step in most proteomic experiments is the identification of proteins in the biological system under investigation. To achieve this, proteins are digested into peptides, analyzed by LC-MS/MS, and tandem mass spectra are used to interrogate protein sequence databases using search engines that match experimental data to data generated *in silico*[12,13]. Peptide spectrum matches (PSMs) are commonly assigned by a search engine using either a heuristic or a probabilistic scoring scheme[14-18]. Proteins are then inferred from identified peptides and a protein score or a probability derived as a measure for the confidence in the identification[13,19].

Estimating the proportion of false matches (false discovery rate; FDR) in an experiment is important to assess and maintain the quality of protein identifications. Owing to its conceptual and practical simplicity, the most widely used strategy to estimate FDR in proteomics is the target-decoy database search strategy (target-decoy strategy; TDS)[20]. The main assumption underlying this idea is that random matches (false positives) should occur with similar likelihood in the target database and the decoy (reversed, shuffled, or otherwise randomized) version of the same database[21,22]. The number of matches to the decoy database, therefore, provides an estimate of the number of random matches one should expect to obtain in the target database. The number of target and decoy hits can then be used to calculate either a local or global FDR for a given data set[21-26]. This general idea can be applied to control the FDR at the level of PSMs, peptides and proteins, typically by counting the number of target and decoy observations above a specified score.

Despite the significant practical impact of the TDS, it has been observed that a peptide FDR which results in an acceptable protein FDR (of say 1%) for a small or medium sized dataset, turns into an unacceptably high protein FDR when the dataset grows larger[22,27]. This is because the basic assumption of the classical TDS is compromised when a large proportion of the true positive proteins have already been identified. In small data sets, containing say only a few hundred to a few thousand proteins, random peptide matches will be distributed roughly equally over all decoy and 'leftover' target proteins, allowing for a reasonably accurate estimation of false positive target identifications by using the number of decoy identifications. However, in large experiments comprising hundreds to thousands of LC-MS/MS runs, 10,000 or more target proteins may be genuinely and repeatedly identified, leaving an ever smaller number of (target) proteins to be hit by new false positive peptide matches. In contrast, decoy proteins are only hit by the occasional random peptide match but fully count towards the number of false positive protein identifications estimated from the decoy hits. The higher the number of genuinely identified target proteins gets, the larger this imbalance becomes. If this is not corrected for in the decoy space, an overestimation of false positives will occur.

This problem has been recognized and e. g. Reiter and colleagues suggested a way for correcting for the over-estimation of false positive protein hits termed MAYU[27]. Following the main assumption that protein identifications containing false positive PSMs are uniformly distributed

over the target database, MAYU models the number of false positive protein identifications using a hypergeometric distribution. Its parameters are estimated from the number of protein database entries and the total number of target and decoy protein identifications. The protein FDR is then estimated by dividing the number of expected false positive identifications (expectation value of the hypergeometric distribution) by the total number of target identifications. While this approach was specifically designed for large datasets (tested on ~1,300 LC-MS/MS runs from digests of C. elegans proteins), it is not clear how far the approach actually scales. Another correction strategy for over-estimation of false positive rates, the R factor, was suggested initially for peptides[28] and more recently for proteins[29]. A ratio, R, of forward and decoy hits in the low probability range is calculated, where the number of true peptide or protein identifications is expected to be close to zero, and hence, R should approximate one. The number of decoy hits is then multiplied (corrected) by the R factor when performing FDR calculations. The approach is conceptually simpler than the MAYU strategy and easy to implement, but is also based on the assumption that the inflation of the decoy hits intrinsic in the classic target decoy strategy occurs to the same extent in all probability ranges.

In the context of the above, it is interesting to note that there is currently no consensus in the community regarding if and how protein FDR should be calculated for data of any size. One perhaps extreme view is that, owing to issues and assumptions related to the peptide to protein inference step and ways of constructing decoy protein sequences, protein level FDRs cannot be meaningfully estimated at all[30]. This is somewhat unsatisfactory as an estimate of protein level error in proteomic experiments is highly desirable. Others have argued that target-decoy searches are not even needed when accurate p-values of individual PSMs are available[31] while yet others choose to tighten the PSM or peptide FDRs obtained from TDS analysis to whatever threshold necessary to obtain a desired protein FDR[32]. This is likely too conservative.

This chapter characterizes the picked TDS for protein FDR estimation and investigates its scalability compared to that of the classic TDS FDR method in datasets of increasing size up to ~19,000 LC-MS/MS runs. The results show that the picked TDS is effective in preventing decoy protein over-representation, identifies more true positive hits and works equally well for small and large proteomic data sets.

2 Methods

2.1 Datasets and data processing

The data basis for this study was a large collection of LC-MS/MS runs along with the derived human protein identification data deposited in ProteomicsDB (https://www.proteomicsdb.org). At the time of writing, this comprised 19,013 LC-MS/MS runs, the majority of which represent two recently published drafts of the human proteome[9,10]. In ProteomicsDB, biological samples are grouped into experiments of varying number of LC-MS/MS runs. Raw MS files from each experiment were searched in parallel using Mascot[16] and Maxquant/Andromeda[15,33] against a concatenated protein sequence database containing the UniProtKB complete human proteome (download date: 05 Sep 2012; 86,725 sequences) and cRAP (common Repository of Adventitious Proteins; download date: 05 Sep 2012; 113 sequences) as described[9]. Briefly, in the Mascot workflow, MS files were processed using Mascot Distiller using peak picking, de-isotoping and charge deconvolution. The resulting peaklist files were searched with the target-decoy option enabled (on-the-fly search against a decoy database with reversed protein sequences), a precursor tolerance of 10 ppm and a fragment tolerance of 0.5 Da for CID spectra and 0.05 Da for HCD spectra, an enzyme specificity of trypsin, LysC, GluC, or chymotrypsin (as appropriate), a maximum of two missed cleavages sites, the Mascot 13C option of 1 and oxidation of Met as well as acetylation of protein amino-terminus as variable modifications. Additional variable and fixed modifications were set as appropriate for individual experiments (e.g. SILAC or TMT or phosphorylation etc.). In the Maxquant workflow, MS files were searched against the same target-decoy protein sequence database as described above but using the Andromeda search engine. Proteases, variable and fixed modifications were specified as above. Mass accuracy of the precursor ions was determined by the time-dependent recalibration algorithm of Maxquant, and fragment ion mass tolerance was set to 0.6 Da and 20 ppm for CID and HCD, respectively. Further details regarding sample handling and data acquisition can be found in[9].

2.2 Procedure for peptide length dependent score normalization

Search engine-specific local peptide length-dependent score cutoffs as reported in Wilhelm et al.[9] were calculated as follows. All peptide spectrum matches (PSMs) of the same length were binned separately for Mascot and Andromeda in intervals of 1 score point and smoothed by a moving average with a window size of 5 to account for fluctuations likely introduced by the scoring algorithm. The local false discovery rates in each score bin were calculated by dividing the number of decoy PSMs by the number of target PSMs and the resulting distribution was smoothed using a moving average with a window size of 5 to account for small fluctuations. The minimum score over all bins with a local false discovery rate less than 0.05 was defined to be the local peptide length-dependent cutoff. Normalized scores of PSMs were calculated by dividing the Mascot ion score or Andromeda score by the corresponding search engine specific local peptide length-dependent cutoff.

2.3 PCM q-value calculation

For the purpose of this study, a q-value is defined to be the minimum FDR at which a PSM, peptide or protein will appear in the filtered output list. Such q-values are commonly used to filter a list of

observations to obtain a particular FDR. Instead of using all PSMs for this purpose, only the PSM with the highest normalized search engine score was chosen that represents one peptide sequence detected at one charge state and carrying a particular peptide modification (termed PCM). PCMs for each LC-MS/MS run were then sorted in decreasing order by their normalized Mascot or Andromeda scores. Empirical q-values were calculated by traversing the list from top to bottom and dividing the cumulative number of decoys by the number of cumulative targets. To assure monotonicity a second traversal from bottom to top changes the empirical q-value from the top to bottom traversal to the minimum q-value observed so far. Next, the relationship between logarithmic q-values and normalized scores was modeled by a linear regression using the highest and lowest scoring PCMs with an empirical q-value below 0.01 as fulcrums. Then, all q-values were recalculated using the predicted slope a and intercept b of the model: $-log_{10}q - value = a \cdot normalized\ score + b$, by multiplying the normalized score with the predicted slope a and adding the predicted intercept b. Last, the resulting list of PCMs was filtered at 1% FDR.

2.4 Protein inference

Peptides matching to either one particular protein isoform (protein unique) or to multiple protein isoforms originating from the same gene (gene unique) are classified as unique peptides (gene-centric uniqueness). All other peptides are classified as shared (Figure 4.45). Shared peptides were discarded from protein inference. For the purpose of this study, it is not differentiated between the identification of a specific protein isoform and the identification of at least one protein isoform of a gene, thus proteins/genes colored in blue and green are used.

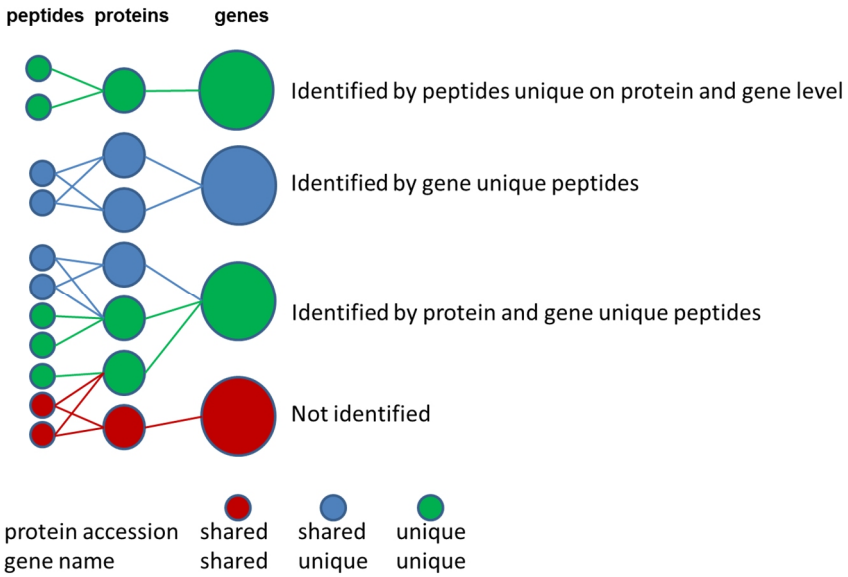


Figure 4.45 | Gene-centric uniqueness calculation. Peptides matching to either one particular protein isoform (green circles, protein unique) or to multiple protein isoform but with the same gene name (blue circles, gene unique) are classified as unique peptides. All others, namely peptides matching to multiple protein isoforms with different gene names (red circles), are classified as shared. Shared peptides were discarded during the protein inference whereas both protein unique and gene unique peptides give rise to the identification of gene products.

2.5 Protein score calculation

For data presented in Figure 4.47a, protein scores were calculated as the sum of Mascot ion scores of the best scoring peptide matches below 1% PSM FDR. For all other analyses, protein scores were calculated either as the sum of the Q-scores ($-\log_{10}$ transformed q-values) of all matched PCMs which passed a defined q-value threshold or by the maximum Q-score of all PCMs. Again, all methods only considered unique peptides.

2.6 Protein q-value calculation

To estimate protein q-values, proteins were sorted in decreasing order by their score. Empirical protein q-values were calculated by traversing the list from top to bottom and dividing the cumulative number of decoys by the number of cumulative targets. To assure monotonicity, a second traversal from bottom to top changes the empirical q-value to the minimum q-value observed so far. This step was repeated each time a new dataset was introduced. For Figure 4.47a, the experiment containing the largest number of IDs was selected first, followed by the experiment with the second largest number of IDs and so forth. This was necessary to illustrate that the number of protein IDs at 1% FDR initially rises, reaches a maximum and then decreases again. For data shown in Figure 4.51, data were aggregated in random order.

2.7 Picked protein FDR approach

In contrast to the classic TDS, the picked TDS treats target and decoy sequences of the same protein as a pair. If the protein score for the target (forward) amino acid sequence is higher than that of the respective decoy (reversed) sequence, the target sequence is counted as a hit and the decoy sequence is discarded. Conversely, if the decoy sequence scores higher than the target sequence, it counts as a decoy hit and the target sequence is discarded. This way, no bias is introduced with respect to how target and decoy proteins contribute to the protein FDR. The protein FDR was estimated using the target and decoy hits in the same way as in the classic approach.

2.8 Protein FDR simulation

A simulated dataset $\boldsymbol{S}$ consists of a set of true positive $\boldsymbol{P}^i$ (present proteins), false positive $\boldsymbol{A}^i$ (absent proteins) and decoy identifications $\boldsymbol{B}^i$. For each dataset $\boldsymbol{S}^i$ during the aggregation, the number of identified proteins $k^{\boldsymbol{i}} = \left|\boldsymbol{P}^{\boldsymbol{i}}\right| + \left|\boldsymbol{A}^{\boldsymbol{i}}\right|$ was drawn from a normal distribution. The dataset specific protein FDR fdr^i was drawn from a modified exponential distribution which ensures that the protein FDR is worse than a user specified value. Next, the set of true positive identifications containing $k^i \cdot (1 - fdr^i)$ proteins is sampled from a set of observable proteins $\boldsymbol{O}$ using a prior, which reflects the relative observation rates of proteins as seen in ProteomicsDB. This ensures that certain proteins are present in almost every single dataset and only a couple of datasets add rare proteins (if at all). False positive protein identifications were sampled uniformly from the set of proteins which are left after sampling the true positive identifications ($\boldsymbol{O} \setminus \boldsymbol{P}^i$) to ensure that a protein is either a true positive or false positive. In contrast, decoy identifications are uniformly sampled from the set of all observable proteins $\boldsymbol{O}$. The assumption that the number of false

positive protein identifications is equal to the number of decoy identifications $\boldsymbol{B}^i = \boldsymbol{A}^i$ holds and thus the estimated and actual protein FDR using the classic model are equal.

The aggregated dataset $\boldsymbol{S}^j_{aggregated} = \{\boldsymbol{S}^1, \dots, \boldsymbol{S}^j\}$is composed of true positive identifications $\boldsymbol{TP}^j = \cup_{k=1}^{j} \boldsymbol{P}^k$, false positive identifications $\boldsymbol{FP}^j = \cup_{k=1}^{j} \boldsymbol{A}^k \setminus \boldsymbol{TP}^j$ and decoy identifications $\boldsymbol{D}^j = \cup_{k=1}^{j} \boldsymbol{B}^k$. Notably, the set of false positive identifications cannot contain proteins which are present in the union of all true positive identifications. However, with an increasing number of aggregated datasets, the underlying assumption of $|\boldsymbol{FP}^j| = |\boldsymbol{D}^j|$does not hold since the set of false identifications will decrease as $|\boldsymbol{TP}^j|$ increases.

3 Results

3.1 Breakdown of the classic target-decoy protein FDR model

In large proteomic studies, identifying tens or hundreds of thousands of peptides, the classic target-decoy strategy (TDS) model overestimates protein FDR due to the fact that the higher the number of genuinely identified target proteins gets, the more imbalanced the ratio of potential new target and decoy protein identifications becomes. This inevitably leads to an accumulation of decoy proteins and overestimated protein FDR. To illustrate this problem, a simulation study using empirical probabilities was performed followed by a validation using real data.

3.1.1 Simulation of protein FDR overestimation in large datasets

Generally, the aggregation of multiple datasets will lead to an inflated protein FDR. However, *a priori* it is unclear which proteins are truly present or absent, thus the accumulation effect can only be studied using the observed numbers of target and decoy identifications while the effect on true and false positive identifications is hidden. Figure 4.46a shows the result of a simulated aggregation of 500 distinct datasets. In contrast to real data, it is known which dataset (for instance a single raw file, the results of an experiment or an entire study) introduces which protein as a true or false positive and thus allows the assessment of the classical TDS estimated protein FDR. Each dataset is filtered to a minimum of 0.01 protein FDR and contains on average 5000 target protein identifications. In order to reflect real data, both the number of proteins as well as the individual protein FDR is drawn from a distribution (Figure 4.46b and c, see methods for details) including which proteins are likely to be present in a single dataset (Figure 4.46d). As expected, both the number of target (depicted in solid blue, containing true and false positive identifications) and decoy (solid red line) identifications rise constantly and reach almost the same total number of proteins after adding 500 datasets. The estimated protein FDR using the classic TDS (solid purple line) is close to 1 (100%). However, the number of true positive identifications (dashed blue line) actually totals at 15,940 protein identifications. This shows that the number of false positive identifications cannot be used without correction since the number of decoy protein identifications significantly overestimates the number of false positive target identifications (dashed red line). Notably, the actual or real protein FDR (dashed purple line) reaches its maximum after adding roughly 100 datasets and starts to decrease after that, indicating that new datasets provide new and unique evidence for proteins formerly being unidentified or were the result of false matching. After aggregating hundreds or even thousands of datasets, most of the true positive target identification will contain some false positive peptide spectrum matches but this does not undermine their presence. This effect is not model by the classical TDS and thus the number of decoy identifications rises irrespectively of the protein FDR of an individual dataset.

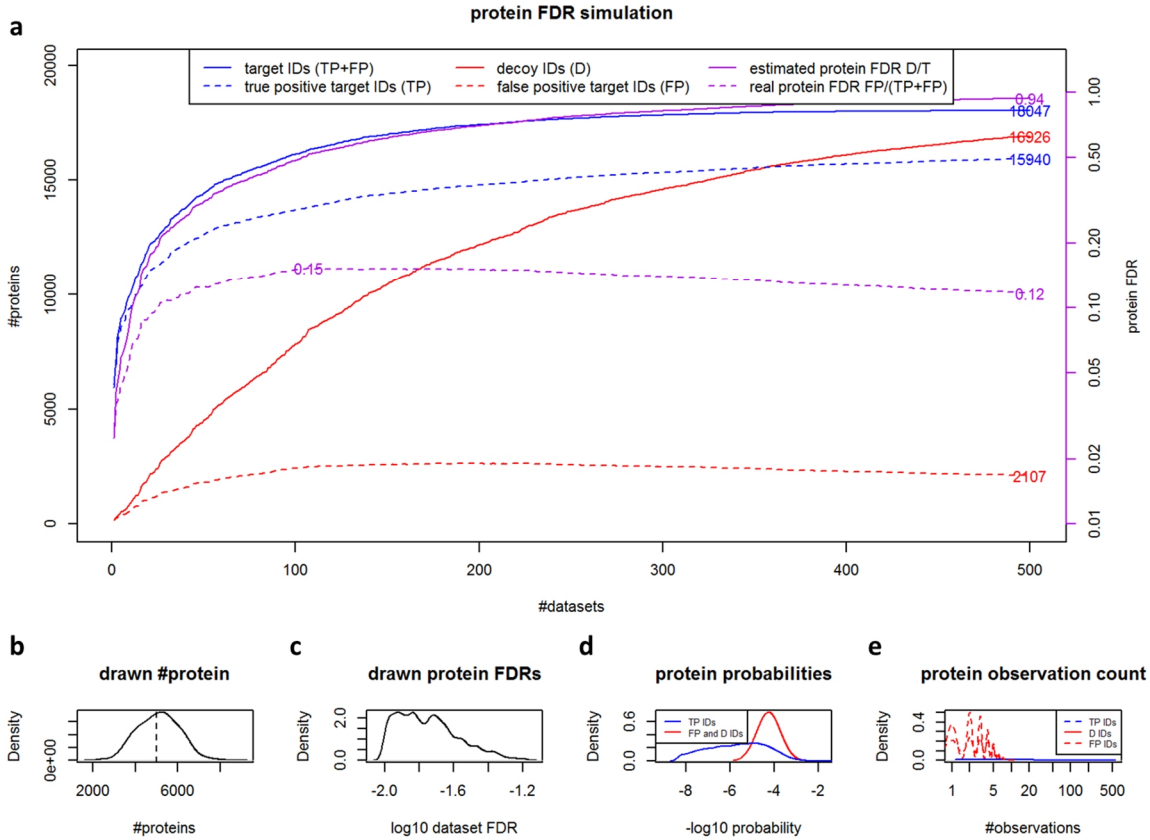


Figure 4.46 | Protein FDR simulation. a, Simulation of 500 datasets each containing evidence for about 5000 proteins each at varying protein FDR (minimum 1%). While the number of target protein identifications (solid blue) containing true and false positive identifications saturates quickly, the number of decoy proteins (solid red) increases almost linearly, reaching almost the same total number. In contrast, the number of true positive target protein identifications (dashed blue) reaches saturation much quicker and plateaus at 15940 proteins. The number of false positive protein identifications (dashed red) reaches its maximum at around 100 datasets but then starts to decrease as more and more datasets are added which provide new and correct evidence for the presence of a protein. The overall trend is mirrored by the classic protein FDR estimation (solid purple), increasing constantly and reaching almost 1.0, and the real protein FDR (dashed purple), which increases first, but reaches its maximum at around 100 datasets (15% protein FDR) but then decreases. b, shows the distribution of the number of proteins drawn for a dataset. c, shows the distribution of individual dataset protein FDR levels. d, shows the probability distribution of selecting a protein being a true positive (green) and false positive or decoy (red) identification. e, shows the distribution of observations for target (blue) and decoys (red) identifications after aggregation of 500 datasets.

3.1.2 Experimental validation of simulation study

To validate the effects shown in Figure 4.46a, the protein identification results from 1,974 aggregated Mascot searches (representing a total of >18,000 distinct LC-MS/MS runs) were analyzed (Figure 4.47a). Search results of each LC-MS/MS run were filtered at 0.01 PSM FDR threshold and all search results were subsequently ranked in descending order according to the number of proteins identified. Individual protein scores were calculated by summing up Mascot ion scores of the best PSM for all unique peptides of that protein. Based on these criteria, the largest search result contained 8,255 target proteins and 321 decoy proteins with 7,250 identified proteins at <1% protein FDR. Subsequently the second, third and so on largest search result was added and the calculation of the protein FDR was repeated each time. Figure 4.47a shows that the number of identified target proteins quickly rose when adding further search results and that considerable saturation occurred by the time 100-150 search results had been combined. Decoy

protein identifications rose at a slower rate but nevertheless approached the number of target hits as the number of aggregated search results reached completion. As an example, 14,137 target proteins were identified when aggregating the first 50 search results but the protein FDR had meanwhile reached 35%. Adding another 1,924 search results increased the target protein IDs by 4,137 proteins but also increased the classic TDS FDR to ~89% implying that only 1,936 of all proteins were true. It is obvious, that the latter figure cannot be correct if the first search results alone already contained 7,250 proteins at 1% FDR.

The situation could only be partially remedied by introducing a protein FDR filter. When forcing a 1% protein FDR at each aggregation step, protein coverage peaked at the 110th search result (10,433 proteins) but then dropped to 6,511 proteins when 1,860 further search results were added. Given this clear breakdown of the classic TDS protein FDR approach, this study proposes an alternative idea, which is refer to as the 'picked' target decoy strategy (picked TDS). Before introducing this concept, the heterogeneous nature of the data in ProteomicsDB required data harmonization that is described in the following section.

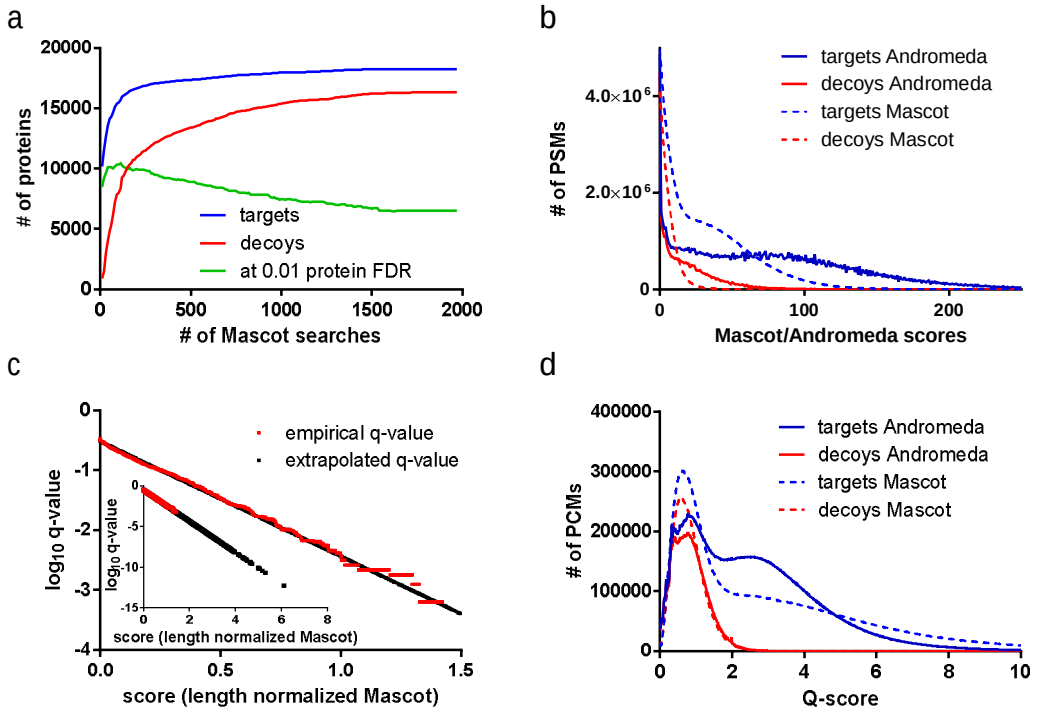


Figure 4.47 | Breakdown of the classic TDS and q-value calculation for data harmonization. a, To illustrate the breakdown of the classic TDS 1970 Mascot search results (18754 raw files) filtered at 1% PSM FDR, were aggregated while re-computing the number of proteins at 1% protein FDR at each step. Protein scores were derived by summing Mascot ion scores of the best peptide matches. The number of target (blue) and decoy (red) proteins saturated quickly while the number of proteins at 1% protein FDR (green) reached its maximum at an early stage but then continuously decreased and stopped at less proteins than in the beginning. This indicates that the classic TDS is not working when dealing with large data. b, The Mascot (dashed) and Andromeda (solid) target (blue) and decoy (red) PSM score distributions show vast differences in the scoring scheme precluding their combination without prior normalization. c, To obtain continuous PCM q-values, a linear extrapolation model (black) trained on the empirically calculated PCM q-values (orange) was used. The inset shows that after extrapolation, meaningful q-values can be assigned to PCMs which have a higher score than the best decoy. d, Following q-value extrapolation (Q-score is defined as $-\log_{10}$(q-value)), Mascot (dashed) and Andromeda (solid) target (blue) and decoy (red) q-value distributions align well, particularly in the q-value range where most false positive identifications are expected and thus allow the combination of the search results.

3.2 Data harmonization using extrapolated q-values

The human proteome data deposited in ProteomicsDB comes from a wide variety of biological samples and biochemical experiments and was acquired on different generations of Thermo Orbitrap instruments and using different fragmentation methods as well as resolution settings. Therefore, the data needed to be aggregated and harmonized in a way that allows a consistent and unified treatment of the results. At the time of writing ProteomicsDB contained 18,754 Thermo Orbitrap raw files for which Mascot was used as a search engine and 17,471 raw files for which Andromeda was used. Figure 4.47b illustrates profound differences in the score distribution of the two search engines which is rooted in the differences in the underlying scoring schemes. Both Mascot and Andromeda assign significantly higher scores to longer peptides. To correct for that, both scores are normalized using the length dependent thresholds (see methods section for details and Figure C1a and b in the Appendix)[9,21,33,34].

The target decoy score distributions are strongly dependent on the type of sample analyzed and the type of fragmentation method used (high or low resolution CID, HCD). For instance, dimethyl labeled tryptic digests of human embryonic stem cells measured by low resolution CID yielded very different target-decoy distributions compared to unlabeled tryptic digests of the melanoma cell line A375 measured by HCD (Figure C1c and d in the Appendix). Thus, it is not sensible to use a single threshold value to achieve say 1% PSM FDR in heterogeneous and large data sets. Instead, these thresholds should be derived for each LC-MS/MS run separately (Figure C1e and f in the Appendix). This is achieved by calculating q-values or posterior error probabilities, e.g. using routines implemented in Maxquant[33] for Andromeda results and Percolator[24] or PeptideProphet[35], for Mascot results. In order to be consistent for both Andromeda and Mascot datasets, ProteomicsDB offers a simple procedure to calculate q-value compatible with both search engines. Instead of using all PSMs for this purpose, only the highest scoring PSM is chosen. This PSM represents one peptide sequence that can carry modifications and is detected with a certain charge state (termed PCM, the best PSM so to speak) because reducing the redundancy of the PSM information (i.e. many spectra hitting the same peptide) into the best PCMs (here) or the best peptide[21,36] results in more robust significance threshold estimates and which are less affected by the oversampling of high abundance peptides compared to using PSMs[37].
In order to appropriately deal with the fact that the number of decoy hits is very small for high scoring PCMs, q-value are linearly extrapolated using the empirical q-values (Figure 4.47c). Without extrapolation, there would be no difference in q-value between say a PCM of Mascot score of 70 and 150 even though the PCM with the higher score should carry more weight than the lower scoring PCM (or peptide for that matter). This procedure allows the combination of results from the two search engines because the distributions of $-\log_{10}$ transformed q-values (referred to as Q-scores) aligned very well (Figure 4.47d), particularly at low q-values where most of the false positives are expected. Target and decoy PCMs that passed the q-value requirement of 0.01 showed only a weak saturation trend as a function of the size of the dataset and consequently lead to only a minimal increase in global PCM FDR (Figure C2 in the Appendix).

3.3 The 'picked' TDS to estimate protein FDR

After data harmonization, the overestimation of false positive protein identifications with the classic TDS was investigated. The PCM q-value cut-off was set to 0.01. Proteins scores were derived from the best scoring unique peptide for every protein (the peptide with the best PCM Q-score, see above). While other more sophisticated strategies exist for calculating protein scores[22,28], using the best peptide hit or the sum of peptide scores are common practice in the proteomics community. The resulting Q-score distribution of target and decoy proteins according to the classical TDS is shown in Figure 4.48a. As one might expect, the bimodal appearance suggests that the lower score range mainly contains false positive protein identifications[35]. At the same time, the number of decoy proteins in that score range is massively higher than that of the target proteins clearly illustrating the aforementioned overestimation of false positive proteins.

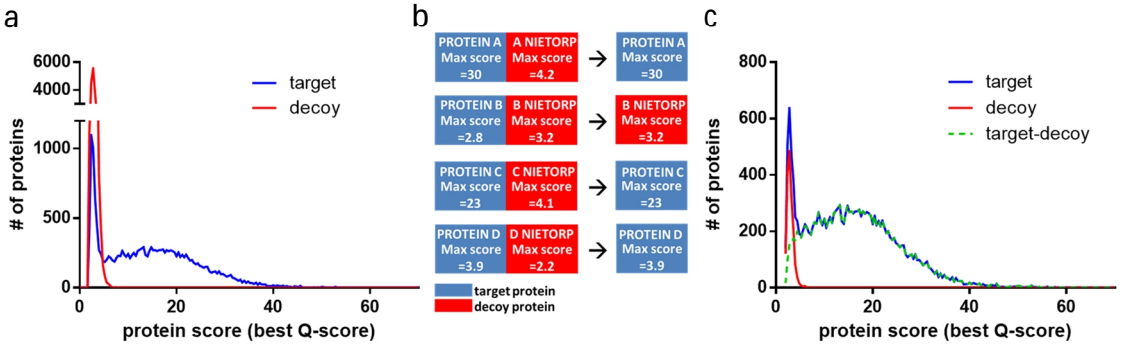


Figure 4.48 | Protein FDR estimation using the classic and picked target-decoy strategy. A PCM q-value cut-off of below 0.01 was used. a, Using the number of decoy proteins from the classic TDS massively overestimates the number of false positive protein identifications. This is apparent by the almost 6-fold higher amplitude of the decoy (red) protein distribution in the low scoring region compared to that of the target proteins (blue). b, The picked TDS treats target and decoy sequences of the same protein as a pair. If the protein score of the target (blue) amino acid sequence is higher than that of the respective decoy (red) sequence, the target sequence is counted as a hit and the decoy sequence is discarded. Conversely, if the decoy sequence scores higher than the target sequence, it counts as a decoy hit and the target sequence is discarded. c, After applying the picked approach, the decoy (red) protein distribution superimposes with the target (blue) protein distribution which allows proper protein FDR estimation using the number decoy proteins, and yields a reasonable distribution of true protein hits (green dashed line), calculated as the difference between the distributions of target and decoy hits.

In contrast to the classic TDS, the 'picked' TDS (Figure 4.48b) treats target and decoy sequences of the same protein as a pair rather than as individual entities. If the protein score (Q-score) for the target sequence is higher than that of the respective decoy sequence, the target sequence is counted as a hit and the decoy sequence is discarded. Conversely, if the decoy sequence scores higher than the target sequence, it counts as a decoy hit and the target protein is discarded. This idea was in part inspired by the decoy fusion approach used for peptides[38], and in part by the established practice in the field of using a concatenated target and decoy database in order to select only for those PSMs that have the best score in either the target or the decoy space, rather than selecting hits which pass a score threshold in both target and decoy space[20,36].

Figure 4.49 depicts an exaggerated example of how the picking is performed and its impact on the global protein FDR estimation. In large single or combined datasets, the major proportion of both target and decoy proteins are matched by PSMs. In this example (left hand side) the database consists of 9 proteins (listed from A-I) and all, except 1 target and 2 decoy proteins are supposedly

identified. The classic TDS (without picking) estimates the protein FDR to be 87% since 7 decoy proteins are present. However, the real FDR is only 37% since only 3 of the estimated 7 false positive protein identifications are present. The picking approach (right panel) selects either the target or the decoy protein from a target-decoy pair based on their scores and discards the other. For instance, for protein A the target was selected since the decoy was not identified, for protein B the decoy was selected since its score is higher than that of its target identification, and so on. This process is performed on all proteins separately leaving a list of 9 proteins in the final result set. The effect is a superior (here perfect) estimation of the protein FDR. Note that all true positive target identifications are still present, however, 1 (namely that of protein B) of the false positive identifications was discarded in favor of its decoy version.

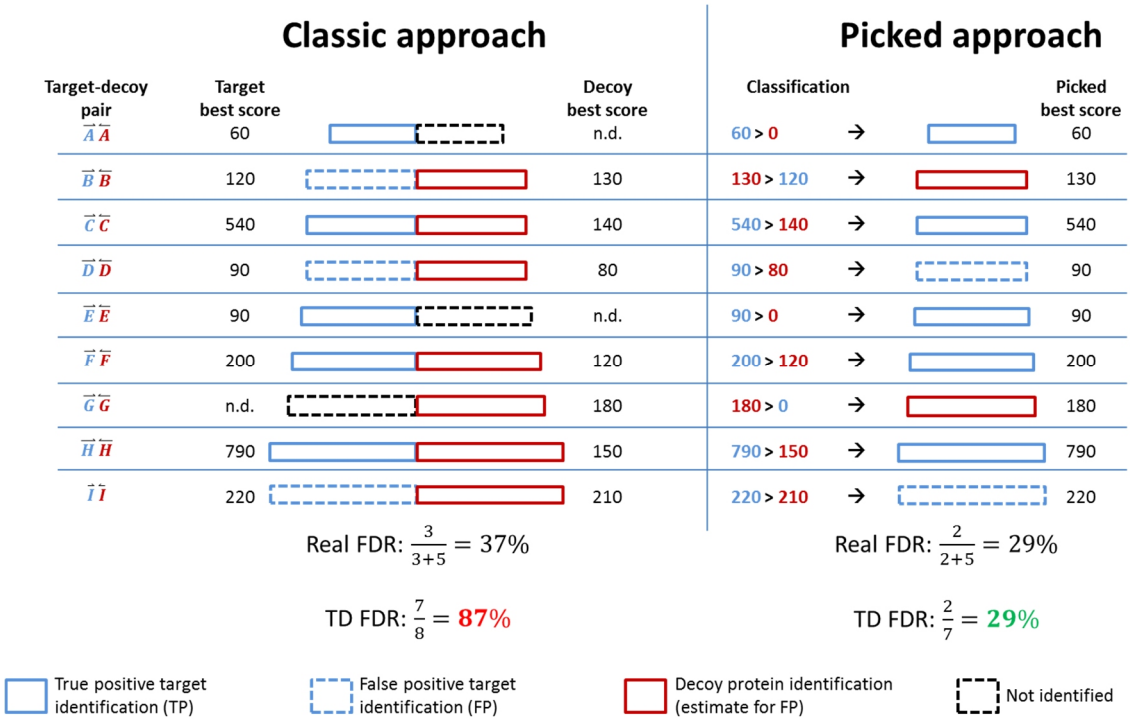


Figure 4.49 | Example of the overestimation of protein FDR on large datasets. Most of the decoys are identified a large proteomics experiments (or repository). As illustrated, the classic approach overestimates the number of false positive identifications (here) by almost 3-fold. The picked approach controls for the overestimation by "picking" either the target or the decoy and precisely estimates the true protein FDR.

This picking approach is symmetric, as it has no built-in bias for the selection of either a decoy or a false positive target protein. For the picked TDS, the target distribution is again bimodal (Figure 4.48c), but now, the decoy distribution is nearly identical to the low Q-score range of the target distribution as would be expected for well-functioning FDR approach[35]. As a result, the distribution of true positive protein identifications (i.e. the difference between the target and decoy hits, dotted green line in Figure 4.48c) approaches zero for very low protein scores indicating that the estimation of false positive IDs is accurate. Results also compared favorably to the recently described R-factor correction approach[29] that addresses over-estimation of decoy hits by an empirically derived correction factor. Performed on the same data, the R-factor corrected decoy distribution alternates between positive and negative values for true positive protein identification at low protein scores, but still provides a much more sensible overall picture than the uncorrected TDS (Figure 4.50).

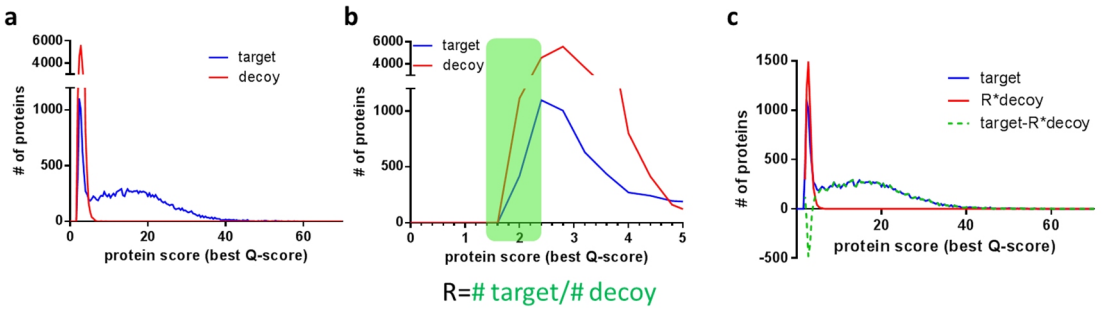


Figure 4.50 | R factor correction. a, Using the number of decoy proteins from the classic TDS massively overestimates the number of false positive protein identifications, decoy proteins (red), target proteins (blue). b, The R factor is calculated as the ratio between the number of target and decoy hits with a score below the 2.4 (score at which the FDR passes the 0.8 value as suggested in Shanmugam et al) c, After applying the R factor correction, the decoy (red) protein distribution agrees better with the target (blue) protein distribution which yields more reasonable protein FDR estimation using the adjusted number of decoy proteins. The distribution of true protein hits (green dashed line), calculated as the difference between the distributions of target and decoy hits is more sensible than for the standard decoy approach, although negative values are observed for low scoring proteins.

Interestingly, when using the sum of Q-scores of all PCMs of a given protein as a score, a much poorer separation between the false positive and true positive distribution can be observed (Extended Data Figure C3a and b in the Appendix), which might be attributed to the fact that large decoy proteins can accumulate high Q-scores by way of many low scoring peptides.

3.4 Performance evaluation of the picked target decoy strategy

As one might expect, when comparing the differences between target and decoy protein identifications at different PCM q-value cut-offs (Figure 4.51a), very similar numbers of proteins were observed for low PCM q-values. However, at higher q-value cut-offs (starting at approximately 10^{-4}), the number of true positive identifications approaches zero for the classic TDS. Conversely, the number of true positive protein identifications for the picked TDS reaches a stable plateau at 15,817 proteins. The non-decreasing true positive trend as a function of more permissive q-value cut-offs is a hallmark of a well-functioning FDR estimation method[35]. When examining protein FDR in the same way (Figure 4.51b), the classic TDS protein FDR approaches 1.0 for q-values of 0.001 and higher. Instead, the picked TDS protein FDR plateaus at a maximum of 10%. Interestingly, the picked TDS protein FDR using summed Q-scores showed similar performance suggesting that the picked TDS is a more reliable and generally applicable protein FDR estimation method (Extended Data Figure C3c and d in the Appendix).

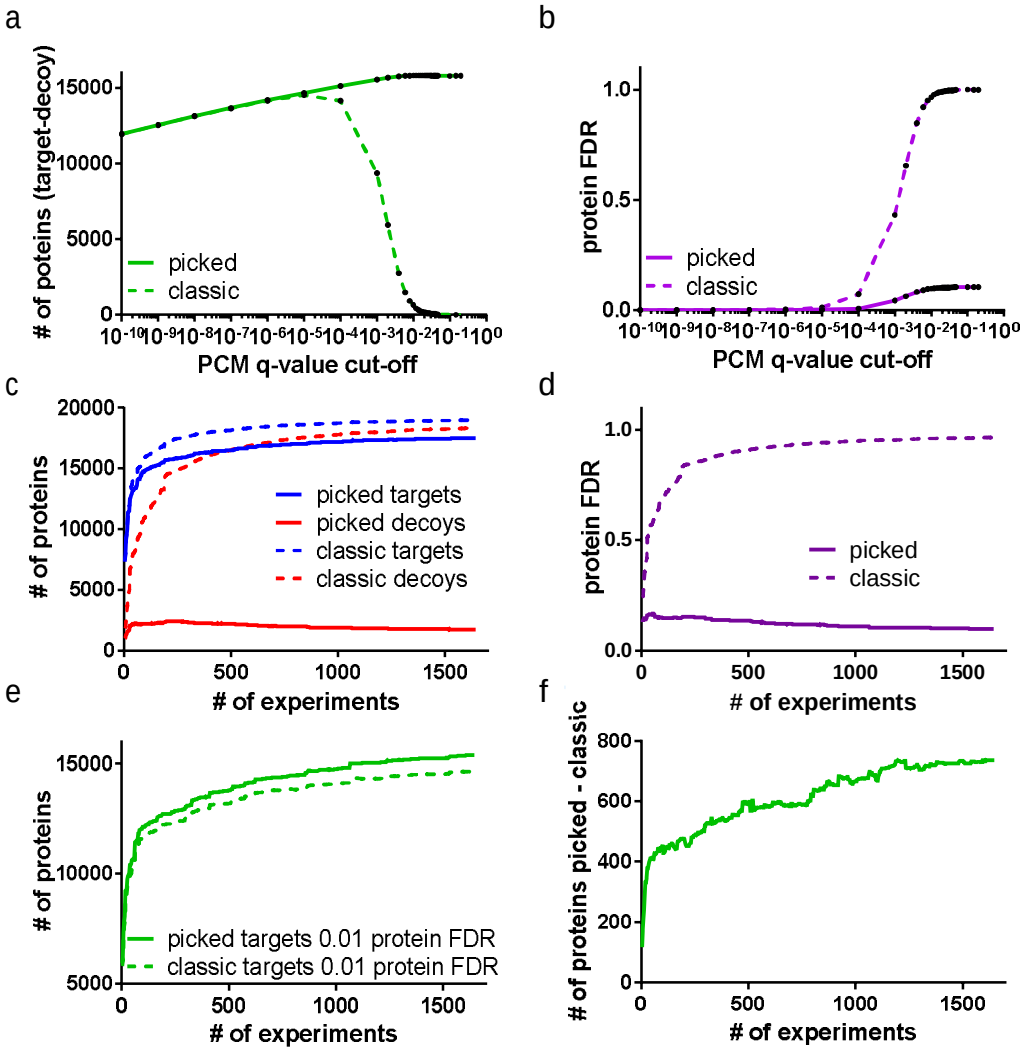


Figure 4.51 | Comparison of the classic TDS to the picked TDS. First, the performance of the picked (solid) and classic (dashed) approach when filtering the PCMs on various FDR cutoffs using the best PCM q-value as protein score was compared. a, With increasing PCM q-value cutoffs, the number of true positive protein identifications (number of target proteins – number of decoy proteins) increases and is comparable between the picked and classic approach. At roughly 10^{-4} PCM q-value cutoff, the number of true positive proteins starts to decrease and quickly drops to almost 0 for the classic approach, whereas true positive proteins IDs increase further and converges at stable plateau of 15,817 proteins in the picked approach. b, The estimated protein FDR of the classic and picked approach mirrors the trend seen in panel A. While the estimated protein FDR increases constantly when increasing the PCM q-value cutoff and eventually reaches 100%, the picked approach starts to rise much later and plateaus at roughly 10%. c, Second, the classic and picked approach when accumulating experiments was compared. The cumulative number of target (blue) protein identifications of the classic and picked approach increases with increasing amount of data whereas the classic approach saturates more rapidly and reports higher numbers of proteins. Conversely, while the number of decoy (red) protein identifications reported by the classic approach saturate and approach the number of target proteins, the number of decoy proteins reported by the picked approach quickly reaches a maximum and decreases when adding more experiments. d, This is again mirrored in the estimated overall protein FDR of the picked and classic approach (see Figure 4.46a for comparison with the simulation data). e, The number of proteins identified at 1% proteins FDR is increasing in both picked and classic approach, but the picked approach consistently reports higher numbers of proteins. f, The difference between the number of proteins reported at 1% proteins FDR between the picked and classic approach increases with increasing number of experiments reaching close to 800 proteins.

Repeating the analysis shown in Figure 4.47a, this time using a PCM q-value cut-off of 0.01, the best PCM for protein scoring and a random aggregation of the experiments (both Mascot and Andromeda results), shows a significant difference between the classical and the picked TDS. It is apparent, that the picked TDS identifies fewer target proteins than the classic TDS but, importantly, shows a massively lower number of decoy protein identifications too (Figure 4.51c and Extended Data Figure C4a in the Appendix). At some point, the decoys increase faster than the targets when using the classic TDS (Extended Data Figure C4b in the Appendix). For the picked TDS, the decoy protein hits show the opposite trend: after an initial very mild increase, the number of decoys actually decreases (Extended Data Figure C4c in the Appendix) implying that addition of new data holds the potential that a protein previously assigned as a false positive (or not identified at all) is supported by a high quality PCM in the new data (see Figure 4.46a for comparison with the simulation data). The above trends are mirrored in the respective protein FDR calculations (Figure 4.51d and Extended Data Figure C4d in the Appendix): while the protein FDR increases for the classic TDS as the dataset grows larger, it steadily decreases for the picked TDS. When filtering the data at 0.01 protein FDR, the number of confidently identified proteins increases for both the classic and the picked TDS as the analyzed dataset grows larger (Figure 4.51e). However, the picked TDS is consistently more sensitive and the absolute difference of identified proteins also steadily increases as the dataset grows larger (Figure 4.51f). In the complete dataset, the classic approach detects 14,638 proteins at 1% protein FDR whereas 15,375 proteins are found with the picked TDS. It is worth noting that the before mentioned R-factor correction approach only partially compensates for this difference (Extended Data Figure C5 in the Appendix).
An interesting detail in the described analysis is the observation that using the best PCM for a protein is very robust with respect to which PCM q-value threshold is applied, whereas the results of protein identification using the sum of Q-scores of PCMs for a protein are much more sensitive to picking an optimal PCM q-value threshold and may completely collapse at high PCM q-values (Extended Data Figure C6 in the Appendix). The picked TDS using the sum of Q-scores does however perform as well as the best Q-score approach at a PCM FDR of 0.0001. It is important to note though, that permissive FDR thresholds (e.g. FDR > 0.01) lead to accumulation of false peptide identifications in the data set and might impair other aspects of data analysis such as quantification, identification of post translational modifications, and protein isoforms and therefore should be avoided in practice. Applying too stringent PCM FDR criteria, however, can also impair subsequent analyses, e.g. quantification, since a lot of good peptide data are excluded. Filtering the PCMs for each LC-MS/MS run in the data set to $q < 0.01$ and applying 1% protein FDR yields 0.13% PCM FDR using the classic TDS and 0.086% PCM FDR with the picked TDS and provides a good balance between peptide coverage and FDR.

This study has shown that the picked TDS outperforms the classical TDS for very large datasets. Its utility is, however, already evident for small or medium sized individual studies (Figure 4.52a). In no case does the picked TDS result in less protein identifications and, interestingly, the gain in protein IDs becomes larger as the number of protein identifications in a particular study increases. Finally, the picked TDS was applied to a number of published large scale protein identification projects[9,39-45]. Figure 4.52b shows that the picked TDS consistently identified a larger number of proteins than the classic TDS. However, the differences in data processing, the search engine and database used and other parameters might also contribute to the observed differences to published protein identifications.

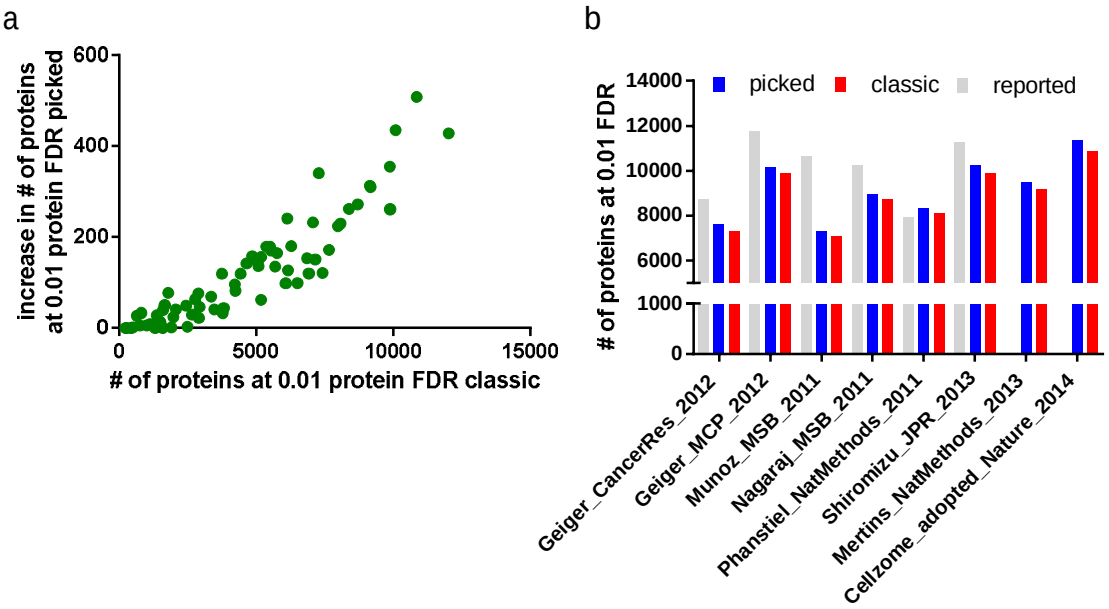


Figure 4.52 | Effects of the picked approach on focused datasets. a, This shows the increase of confidently identified proteins using the picked approach versus the number of proteins reported by the classic approach for 76 data sets (green dots) to illustrate the effect of the picked approach on studies of varying size. The picked approach invariably identifies more proteins than the classic approach and the difference increases with the number of proteins identified in a given data set. b, Re-assessment of the number of proteins reported in a number of publications showed that the picked approach (blue) identified more proteins than the classic approach (red). It is also evident that the picked TDS is more conservative than the number of proteins reported in many of these publications (grey).

4 Discussion

This study investigated the scalability and performance of the "picked" target decoy strategy for estimating protein false discovery rates in large proteomics data sets. The picked TDS addresses decoy protein overestimation typically observed for the classic TDS and takes into account that the probability of creating a false positive PSM is not equal for all proteins. For example, large target and decoy proteins are more prone to accumulating high scoring random matches and are likely to accrue higher protein scores than small proteins both of which artificially inflates the protein FDR. Other parameters that may give rise to similar or related effects are amino acid composition, the number of measurable proteotypic peptides, the type of protease used, the number of tolerated missed cleavages sites, type of mass spectrometer and fragmentation technique used and so on. All of these can be at least partially addressed by simple data harmonizing steps and conceptually extending the line of reasoning from the commonly employed approach of concatenating target and decoy sequences for database searching, to treating target and decoy versions of a given protein sequence as a pair. For proteins that have PSMs/PCMs in both their respective target and decoy sequences, this algorithm will only "pick" the one with the highest score and discard the other. As demonstrated above, this approach does not create the excess of decoy hits observed for the classic TDS FDR but does not alter the target protein distribution. The almost perfect overlap of target and decoy distributions in the low-scoring region suggests little or no bias and, therefore, explains the superior performance of the picked TDS, in line with prior work on the theoretical treatment of the matter[35]. The obtained results also compare favorably to the previously described R-factor approach that corrects for over-representation of decoy hits by normalizing the distribution with an empirically derived factor[29]. The superior performance of the picked TDS is likely due to the fact that it avoids a bias intrinsic in traditional target decoy strategies, whereas the R-factor approach aims at compensating for this bias using a simple but assumption-based model.

A major shortcoming of any decoy generation method is the uncertainty regarding whether or not a decoy peptide is in fact a decoy peptide. While this is easily checked by comparing all peptide sequences to the limited target space that is typically used for protein identification (e.g. Uniprot), it is quite difficult to exclude the possibility that a decoy sequence may actually represent a genuine variant of a known peptide sequence or indeed a genuine but so far undetected or modified peptide. Even if this number of peptides may be fairly low, each might contribute one high scoring decoy protein identification and thus increase the protein FDR. If there are more such cases, it may even substantially limit the number of proteins that can be identified in a complete proteome because the control of protein FDR may create a glass ceiling, a barrier that cannot be breached no matter how good the mass spectrometric data may be.

Even though the picked strategy is unbiased with respect to the compared sets (here sets of highest scoring PCMs) and can thus be used for both a gene-centric (all peptides matching uniquely to a single gene model) or isoform-centric (all peptides matching uniquely to an isoform) analysis, pairing sets is only allowed if each property of the paired sets follow the same distribution. In proteomics, these are mainly features that affect the matching and quality of the spectra, such as peptide length, number of missed cleavages and number of peptides theoretically achievable. However, comparing protein groups is not possible, since it is unlikely that a target and decoy protein group will contain the exact same proteins and thus are comparable in terms

of number of peptides. While this is not an issue in ProteomicsDB, this will affect the usability of this method in smaller studies, unless proteins are grouped in e.g. a gene-centric fashion, since both the quantification as well as interpretation of groups of proteins which do not share a common biology (e.g. same gene) is difficult and may lead to wrong conclusions.

The analysis further revealed that protein scoring using the best PCM score for a given protein performed better than summing up all PCM scores for a protein. At least in part this is due to the fact that the latter is more susceptible to protein length bias, and that the inevitable accumulation of low-scoring peptide matches observed in large data sets has a stronger impact on sum-based protein scoring be it the number of PSMs, the search engine score or posterior error probabilities. Similar observations have lead researchers to adopt the 'best peptide' approach which is conceptually similar to the PCM scoring[22]. Applying extremely stringent peptide filters might improve scalability of sum-based protein scoring, however this will come at the loss of protein and peptide coverage.
For both protein scoring approaches (best Q-score or sum of Q-scores), and in contrast to the classic TDS FDR estimate that approaches 100% protein FDR as the dataset grows larger, the number of decoy hits is actually reduced upon adding new experimental data when using the picked TDS. This is an entirely expected behavior because a false positive protein identification represented by a low scoring target or decoy hit might 'switch' to become a true positive, high scoring target hit when new high quality experimental evidence (i.e. a good tandem MS spectrum) is added to the dataset. It is often assumed that adding more data to an already large data set will only add more false positives. This is a misconception, at least as far as whole proteome identification is concerned because the quality of the extra data will determine if a novel protein can or cannot be identified (Extended Data Figure C4c in the Appendix).

An important conclusion from this analysis is that it should be possible, at least in principle, to confidently identify all proteins in a proteome by accumulating large quantities of high quality LC-MS/MS data provided that all the relevant biological protein sources of an organism have been sampled with sufficient depth.

5 Outlook

While the best PCM approach showed the strongest separation between false and true positives, this method does not take multiple or reproducible evidence into account. As indicated, the generation of the decoy space is thus critical to the success of this method. To circumvent this, further investigations into alternative approaches have to be conducted. Furthermore, this method does not allow the assignment of proper p-values. However, it is possible to use the protein score distributions to calculate posterior error probabilities (local FDR).

One possible improvement is to use all highest scoring spectra per PCM (multiple evidence) to generate a distribution of all pairwise PCM ratios. The resulting distributions can be used for multiple purposes. First, all pairwise ratios within the target space should in theory only generate a unimodal distribution centered on 1. If the set of best PCMs contains a significant number of false positive matches, two mirrored modes to the left and right of the center should appear since all ratios generated with the false positives will be higher or lower. More interestingly, generating the same distribution for the decoy space might provide hints to which peptides within the decoy space might be unknown but present since these should likewise generate the same effect as described in the target space. PCMs exclusively (or enriched) present in the distribution which is centered at a ratio larger than 1 are likely “true positive” decoy matches. Second, generating all pairwise PCM ratios between the target and decoy space will result in a bimodal distribution. False positive matches should be located around 1 while true positive matches appear as outliers to the right. Using this approach, an individual PCM FDR can be computed which can be used to filter false positive PCMs within each target-decoy pair. This should further decrease the number of false matches and thus increase quantification accuracies.
Another approach is the direct comparison of the target and decoy score distribution of all highest scoring spectra per PCM. Here, known statistical methods for comparing distributions can be used such as the Kolmogorov–Smirnov test or a modified Bayesian test[46]. In contrast to the sum of best PCMs for scoring proteins, this approach should not be affected by the dilution of true positives. The dilution effect mainly originates from the fact that long protein, generating a lot of peptides, can accumulate the same total score as short proteins matched with only a couple but high scoring PCMs.

6 Author Contribution

This chapter is based on[47].

Mathias Wilhelm (including others) conceptualized the study, implemented and tested the approach, performed data analysis, refined the model and wrote the manuscript.

7 Abbreviations

CID	Collision-induced dissociation
FDR	False discovery rate
HCD	Higher energy collision induced dissociation
IDs	Identifications
PSM	Peptide spectrum match
PCM	Best scoring PSM per peptide charge modification combination
Q-score	$-\log_{10}$ q-value
SILAC	Stable isotope labeling with amino acids in cell culture
TDS	Target-decoy strategy
TMT	Tandem mass tag

8 References

1 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 & cellular proteomics : MCP*, doi:10.1074/mcp.M114.043489 (2014).

2 Kelstrup, C. D. *et al.* Rapid and Deep Proteomes by Faster Sequencing on a Benchtop Quadrupole Ultra-High-Field Orbitrap Mass Spectrometer. *Journal of proteome research*, doi:10.1021/pr500985w (2014).

3 Helm, D. *et al.* Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Molecular & cellular proteomics : MCP*, doi:10.1074/mcp.M114.041038 (2014).

4 Yamana, R. *et al.* Rapid and deep profiling of human induced pluripotent stem cell proteome by one-shot NanoLC-MS/MS analysis with meter-scale monolithic silica columns. *Journal of proteome research* 12, 214-221, doi:10.1021/pr300837u (2013).

5 Hebert, A. S. *et al.* The one hour yeast proteome. *Molecular & cellular proteomics : MCP* 13, 339-347, doi:10.1074/mcp.M113.034769 (2014).

6 Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell reports* 4, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).

7 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, doi:10.1038/nmeth.2834 (2014).

8 Ritorto, M. S., Cook, K., Tyagi, K., Pedrioli, P. G. & Trost, M. Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes. *Journal of proteome research* 12, 2449-2457, doi:10.1021/pr301011r (2013).

9 Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582-587, doi:10.1038/nature13319 (2014).

10 Kim, M. S. *et al.* A draft map of the human proteome. *Nature* 509, 575-581, doi:10.1038/nature13302 (2014).

11 Savitski, M. M. *et al.* Proteomics. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* 346, 1255784, doi:10.1126/science.1255784 (2014).

12 Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods* 4, 787-797, doi:10.1038/nmeth1088 (2007).

13 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* 4, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).

14 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. *Journal of the American Society for Mass Spectrometry* 5, 976-989, doi:10.1016/1044-0305(94)80016-2 (1994).

15 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* 10, 1794-1805, doi:10.1021/pr101065j (2011).

16 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, doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (1999).

17 Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467, doi:10.1093/bioinformatics/bth092 (2004).

18 Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *Journal of proteome research* 3, 958-964, doi:10.1021/pr0499491 (2004).

19 Serang, O. & Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface* 5, 3-20 (2012).

20 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* 4, 207-214, doi:10.1038/nmeth1019 (2007).

21 Jeong, K., Kim, S. & Bandeira, N. False discovery rates in spectral identification. *BMC bioinformatics* 13 Suppl 16, S2, doi:10.1186/1471-2105-13-S16-S2 (2012).

22 Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* 73, 2092-2123, doi:10.1016/j.jprot.2010.08.009 (2010).

23 Choi, H. & Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of proteome research* 7, 47-50, doi:10.1021/pr700747q (2008).

24 Kall, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of proteome research* 7, 40-44, doi:10.1021/pr700739d (2008).

25 Blanco, L., Mead, J. A. & Bessant, C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *Journal of proteome research* 8, 1782-1791 (2009).

26 Wang, G., Wu, W. W., Zhang, Z., Masilamani, S. & Shen, R. F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical chemistry* 81, 146-159, doi:10.1021/ac801664q (2009).

27 Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* 8, 2405-2417, doi:10.1074/mcp.M900317-MCP200 (2009).

28 Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & cellular proteomics : MCP* 10, M111 007690, doi:10.1074/mcp.M111.007690 (2011).

29 Shanmugam, A. K., Yocum, A. K. & Nesvizhskii, A. I. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *Journal of proteome research* 13, 4113-4119, doi:10.1021/pr500496p (2014).

30 Cottrell, J. *Does protein FDR have any meaning?*, 2013).

31 Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry* 22, 1111-1120, doi:10.1007/s13361-011-0139-3 (2011).

32 Farrah, T. *et al.* State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *Journal of proteome research* 13, 60-75, doi:10.1021/pr4010037 (2014).

33 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, doi:10.1038/nbt.1511 (2008).

34 Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. Combining results of multiple search engines in proteomics. *Molecular & cellular proteomics : MCP* 12, 2383-2393, doi:10.1074/mcp.R113.027797 (2013).

35 Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* 74, 5383-5392 (2002).

36 Granholm, V., Navarro, J. F., Noble, W. S. & Kall, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of proteomics* 80, 123-131, doi:10.1016/j.jprot.2012.12.007 (2013).

37 Savitski, M. M., Scholten, A., Sweetman, G., Mathieson, T. & Bantscheff, M. Evaluation of data analysis strategies for improved mass spectrometry-based phosphoproteomics. *Analytical chemistry* 82, 9843-9849, doi:10.1021/ac102083q (2010).

38 Zhang, J. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics : MCP* 11, M111 010587, doi:10.1074/mcp.M111.010587 (2012).

39 Geiger, T., Madden, S. F., Gallagher, W. M., Cox, J. & Mann, M. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer research* 72, 2428-2439, doi:10.1158/0008-5472.CAN-11-3711 (2012).

40 Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & cellular proteomics : MCP* 11, M111 014050, doi:10.1074/mcp.M111.014050 (2012).

41 Munoz, J. *et al.* The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Molecular systems biology* 7, 550, doi:10.1038/msb.2011.84 (2011).

42 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* 7, 548, doi:10.1038/msb.2011.81 (2011).

43 Phanstiel, D. H. *et al.* Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nature methods* 8, 821-827, doi:10.1038/nmeth.1699 (2011).

44 Shiromizu, T. *et al.* Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project. *Journal of proteome research* 12, 2414-2421, doi:10.1021/pr300825v (2013).

45 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods* 10, 634-637, doi:10.1038/nmeth.2518 (2013).

46 Serang, O., Paulo, J., Steen, H. & Steen, J. A. A non-parametric cutout index for robust evaluation of identified proteins. *Molecular & cellular proteomics : MCP* 12, 807-812, doi:10.1074/mcp.O112.022863 (2013).

47 Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Molecular & cellular proteomics : MCP* 14, 2394-2404, doi:10.1074/mcp.M114.046995 (2015).

Chapter 5

General discussion and outlook

Contents

“The Guide says there is an art to flying", said Ford, "or rather a knack.
The knack lies in learning how to throw yourself at the ground and miss.
... Clearly, it is this second part, the missing, that presents the difficulties.”
- Douglas Adams; The Hitchhiker's Guide to the Galaxy

1 Big biological data

With advances in genomics, transcriptomics, proteomics and imaging, researchers are now generating data at unprecedented speeds. For example, Gene Expression Omnibus[1] and ArrayExpress[2] provide microarray and next-generation sequence datasets on more than one million samples. Ultimately, the goal of such efforts is to generate new hypothesis from big data. Yet, scientists have to be aware that these are not experimentally verified and typically only describe correlations of factors. Less ambitious but equally important goals are to provide means to validate and expand hypotheses using for example larger cohorts or to support the design of experiments by the wealth of data that is available[3]. In all these applications, knowing the precise context of the experiment is essential. Both the experimental conditions as well as the characterization of the environment under which the data was collected have to be monitored and recorded in order to allow cross-experiment comparisons. This in itself is a challenge as our understanding of which and how many factors interact and influence the outcome of an experiment is limited[4].

Especially in proteomic experiments, the wealth of techniques regarding cell culture, sample preparation, multiplexing, measurement technique, data analysis and the combination thereof increases the complexity of cross-experiment data analysis. Sampling artifacts, protein extraction protocols and dynamic range issues impair even the comparison of protein expression values between cell lines, body fluids and tissues. In contrast, genomics datasets did not suffer from this to the same extend. An example is the analysis of the mutation status of genes, which is comparatively static information with rather binary readout. Driven by this, genomics and also transcriptomics matured into the prime data providers for big biological resources and enabled the analysis of large cohorts of cell lines and also patients[5]. However, the complexity of living organisms is largely determined by the dynamic and versatile nature of the products of its genome. Proteins are often directly responsible for the abnormal behavior of cells and thus are the primary target of almost all drugs. Most treatments neither directly nor immediately alter the genome of a system, but rather modify the complex interactions of protein and enzymes.

To be able to describe and quantify biological effects in terms of their variability, many replicates or longitudinal studies are necessary. While earlier efforts largely focused on the acquisition of disease states[6], today's research is shifting towards understanding what can be expected to be normal[7]. This is only economically feasible if multiplexing technologies enable the parallel identification and quantification of biological entities. Proteomics provides such methods, yet lags behind genomics in terms of throughput and depth, thus falling short in its ability to comprehensively characterize the variability of biological systems[8,9]. However, the ability to differentiate repeating signals from technical or biological artifacts as biologically significant and functionally relevant is a common problem in each omics measurement. The combination of different experimental conditions, platforms and acquisition approaches can be used to avoid the misclassification of such signals as functionally relevant. The analysis of multiple omics datasets is expected to provide insight into the underlying mechanisms not achievable from individual studies alone. If the results of different studies are stored including their (complete) experimental annotation, the continuous integration of new studies allows us to broaden our understanding of living systems.

Beyond the experimental challenges in big biological data, many statistical challenges exist as well. A typical problem is that measurement depth typically surpasses the number of independent samples[10]. This impairs statistical analysis as it increases the number of false positive correlations by generating correlations just by chance alone. Additionally, the reliability of models decreases with each dimension added to the system. Dimensionality reduction can reduce this problem, however, this is problematic as key mechanistic information can be lost. So far, most big data efforts have almost invariably generated data that is much more complex than anticipated.

1.1 Re-analysis versus re-measurement

One of the hopes of big data is that not all questions have to be defined beforehand but permit the validation or invalidation of hypotheses based on data mining or re-analysis. For this to work, the acquired data, their annotation and the obtained results of studies have to be as complete and comprehensive as possible in order to enable their unbiased integration. However, most analyses are driven towards a specific hypothesis, which influences data processing decisions. For example, the identification of peptides in a proteomics measurement is typically based on a database containing known (expected) protein sequences. While this approach generally increases the selectivity[11], researchers interested in the occurrence of peptides resulting from e.g. untranslated regions cannot readily use such processing results. With the implementation of repositories providing raw data through standardized services[12], an alternative is to retrieve and re-process data under different hypotheses using newly developed and specialized software. The results could be re-imported and result in the piecemeal annotation of data.
With the constant upward trend in terms of storage and computational demand in all omics technologies, local reprocessing of large amounts of data is likely not feasible in the long term. While the cost of most omics technologies has fallen fast enough to enable individual labs to operate their own machines, storage and processing infrastructure has not followed suit. A possible alternative are cloud-based applications, enabling the processing of big data using custom pipelines in large data centers[13].

However, to date even whole genome next-generation sequencing data is criticized for being incomplete. As described by Alkan et al.[14], limitations of the next-generation shotgun sequencing approach and properties of the data itself, rather than algorithmic inefficiencies, lead to the loss of more than 20% of the genome during assembly of raw data. This incompleteness of data is certainly also the case for proteomic measurements. Data acquired only a couple of years ago is generally viewed outdated and corresponding samples are re-measured using new mass spectrometers to increase coverage and depth. While this creates a conflict for big data efforts, the storage and integration of multi-omics datasets might be able to depict a (more) complete picture of the complexity of living systems than any technology alone. Additionally, providing unified and easy to use access to previous research also enables researchers, which lack access to or knowledge of how to process specific omics technologies, to cross compare and analyze their data together with other data.

2 Bottom-up proteomics

Ultimately, proteomics seeks to consistently and reproducibly identify and quantify all proteoforms, a specific molecular form of a product of a single gene including changes due to genetic variations, alternative splicing and post-translational modifications[15], in any biological system. A particularly interesting technology is using pores of nanometer diameter. The analyte of interest is passed through a protein pore and the sequence is deduced by measuring the drop in current[16]. While it has been shown to work with single un-digested proteins, this technology is still in its infancy and does not yet allow the high-throughput analysis of samples[17].

Bottom-up mass spectrometry-based proteomics is receiving more and more attention in the scientific community[18] because the characterization of thousands of proteins is possible. This enables the precise monitoring of the complex interplay between proteins and their function with regard to e.g. post translational modifications, protein-protein or protein-drug interactions. Yet, many challenges lie ahead to transform this technology into a method that can be used as reliably as modern genomics technologies[19]. This is in large driven by two reasons. First, the peptide-centric data acquisition requires the digestion of proteins using site-specific proteases. This not only disconnects the entities we measure (peptides) from the entities we are primarily interested in (proteins), but also introduces a significant source of variation due to the digestion and fractionation (offline and online) of the sample. Many approaches exist to control and monitor each step, yet studies have shown that differences between labs exists when multiple labs analyzed the same sample[20]. However, data generated within one lab does not seem to suffer from the same issues[21]. Second, the semi-stochastic nature of the data-dependent acquisition impairs the consistent and reproducible measurement of samples. With increasing sample heterogeneity, the number of missing values increases rapidly and, because the lack of identification does not imply absence, significantly impairs subsequent statistical analysis. While many alternative acquisition approaches exist which try to circumvent some of these issues, their implementation is often complicated, not readily available in many labs or only enables the measurement of a limited set of entities[22]. Post-acquisition data processing is contributing to this, as many different search engines, quantification methods and statistical tests exist to identify differentially regulated or otherwise interesting proteins. Each possible combination of these tools, often a choice based on personal experience, typically generates different but equally significant lists of features.

Bottom-up proteomics has been transformed into a high-throughput technology, yet only a miniscule fraction of the information generated by an LC-MS measurement is being used for data analysis. This is because only a comparatively small percentage of eluting features are selected for fragmentation, and only up to half of these generate interpretable MS/MS spectra. While some efforts exist to help decrease the number of unmatched MS/MS spectra[23,24], it is largely unknown how many potentially productive spectra are typically acquired.

These “unidentified” features hold a lot of potential and their integration into current bioinformatics workflows could allow a more comprehensive data analysis or enable retrospective data analysis and integration. However, their use is currently limited by the lack of specialized software, databases and models, especially since most experiments are being analyzed in isolation of previous measurements.

Because of the data-dependent acquisition and the lack of a persistent memory of the mass spectrometer, many features picked for fragmentation were already identified in previous measurements. With an estimated core proteome of 10,000-12,000 proteins, the number of redundant MS/MS spectra being acquired in each experiment is considerable, leading to comparatively low entropy (information theory)[25]. The following subsections highlight potential scenarios in which both data acquisition and analysis of bottom-up proteomics measurements could make use of prior knowledge in order to increase the information content of new MS runs. This can be realized by utilizing performant databases, machine learning and smarter data acquisition, all of which exploit the reproducible characteristics of the underlying measurements (*m/z*, retention time and fragmentation).

2.1 Machine learning

Machine learning is becoming a popular tool in proteomics[26] to model complex processes covering nearly the entire workflow from wet- to dry-lab[27]. With the ever increasing amount of data and the wealth of problems in proteomics, many applications are available which model complex relationships to predict for example proteotypicity[28], retention time[29-31] and fragmentation[32] of peptides. However, the development typically relies on the selection of features which are likely associated with the problem. While this approach works in most cases, it is not farfetched to assume that, because this approach relies on our (limited) understanding of the underlying mechanism, the accuracy of the resulting models is limited. Neuronal networks provide an alternative approach as these systems automatically infer rules. Culminating in deep learning, neural networks have proven to be able to learn complex problems without prior selection of specific features on games like Mario and Go[33,34].

Due to the lack of specific models, many processes in proteomics data analysis still rely on (simple) heuristics. For example, the interpretation of experimental fragmentation spectra is largely driven by the number of peaks matching to all possible fragment peaks without considering their relative intensity. With modern machine learning frameworks[35] and the large number of spectra stored in public repositories, the development of applications for fragmentation prediction is possible. While algorithms such as deep learning will not directly result in the generation of new knowledge, since the extraction of the learned rules is often not possible, once learned, such fragmentation models could greatly increase the sensitivity and accuracy of search engines. By generating *in silico* spectra comprised of only expected observable fragment peaks associated with a predicted intensity, the number of false positive matches generated by chance will be decreased.

ProteomicsDB already enables the storage of reference spectra, including the acquisition parameters such as fragmentation method, collision energy and retention time. At the time of writing, reference data on 4000 synthetically generated peptides are already available and more than 1 million are being generated. Being able to trains models for experimental properties on hundred thousands of data points will result in models that are more accurate. This could be of high importance in targeted proteomics, since retention time and fragmentation of peptides is crucial as it defines their accessibility. Models which accurately predict both will not only simplify the generation of such assays, but can also be used to optimize the chromatographic gradient and fragmentation by choosing parameters which favor the generation of specific fragment peaks and reduce the expected interference.

This concept can be expanded to store measurements of synthetically generated proteins. Having such data allows the modeling of other important peptide properties, such as digestion and ionization efficiencies. This could enable the prediction of expected visible features based on the presence of proteins, again with possible applications in targeted assays. Being able to predict the potential search space will decrease the search space in database searching and thus likely decrease the number of false positive matches.

2.2 Feature matching

The reproducibility of precursor *m/z*, retention time and fragmentation of peptides are key to enable their targeted measurement and machine learning. However, these characteristics can also be exploited in data-dependent acquisition approaches by matching eluting isotope features across multiple aligned LC-MS runs[36]. This processing step, implemented as "match-between-runs" within Maxquant[37,38], increases the reproducible quantification of features as long as it was identified in at least one run and thus reduces the number of missing values. Taken one step further, a recent study has shown that this feature can be used to quantify peptides and thus proteins to a depth generally not possible without fractionation by adding a deep-coverage sample during data processing[39]. This reduced measurement time while keeping the number of quantified features high.

The concept of providing identifications from other (potentially unrelated) LC-MS runs for feature matching during data analysis can be further extended to a database which stores previously identified features[40]. This would not only reduce computation time, since Maxquant performs a database search every time a sample is analyzed, but feature matching would also benefit from all previously conducted measurements. With the in-memory computing capabilities of ProteomicsDB, an online match-between-runs service based on all identifications stored in ProteomicsDB is technically conceivable. This could increase the number of features used for quantification significantly and circumvent the necessity of an in-depth measurement of each specimen. Further supplemented by specific models enabling for example the conversion of retention times between different gradients or stationary phases using retention time indices, new identifications added to such a system could be matched onto old data, ultimately enabling the post hoc quantification of unidentified features in previous experiments.
Such a retrospective match-between-runs system would benefit most when MS1 features are recorded at maximum depth. However, the sampling depth of mass-spectrometers is best when the abundance distribution of features is narrow (homogenous), because the limited dynamic range creates an artificial lower bound on the limit of detection. Due to the large dynamic range of protein expression, only the most abundant features within an MS1 spectrum are visible, leaving low abundant features such as peptides from low abundant proteins or peptides carrying PTMs largely unseen. In order to increase the dynamic range in an experiment, offline fractionation or the depletion of high abundant proteins can be performed. However, this increases the necessary acquisition time and potentially introduces unreliable sample preparation steps, leading to larger technical variation. In order to circumvent this, the typical MS1 scan range could be acquired in multiple smaller windows while decreasing the number of MS/MS spectra. With this, only a small proportion of the acquired *m/z* range will be affected by highly abundant feature. This should increase the apparent sensitivity and thus also the number of visible features

in most windows. However, the identification of such low abundant features is challenging since current acquisition methods are optimized for high throughput (high duty cycle) and thus high abundant peptide species (low automatic gain control and maximum injection time).

2.3 Smart data acquisition

As previously discussed, many MS1 features can potentially be deduced and matched from previous experiments. When doing so, the number of necessary MS/MS scans can be reduced and thus increase the accuracy of quantification and visible features due to a decreased duty cycle. In order to increase entropy of proteomic measurements, one approach could be to spend the available time on characterizing unexpected or unseen features by selecting precursors based on real-time, prior MS data or external knowledge[41,42]. Features visible in MS1 can be classified into four categories: (I) features which were successfully identified in previous runs, (II) features which were picked but did not lead to an identification, (III) features which were not yet selected for fragmentation, and (IV) features which should not be selected for fragmentation. High performant resources such as ProteomicsDB could play an important role in guiding data acquisition by enabling fast queries on large collections of both identified and unidentified features from previous runs. Interfaced with the acquisition software of a mass spectrometer, the prioritization and acquisition parameters of fragmentation spectra of visible features can be controlled and adjusted. The highest priority is on the identification of class II and III features. For example, previously picked but unidentified features can be re-measured under different acquisition conditions by adjusting the fragmentation method, collision energy or maximum injection time. At least important are confirmatory scans to validate previously identified features.

Smart acquisition methods will benefit from machine learning approaches for example by predicting optimal fragmentation energies of phosphorylated peptides or injection times of low abundant features. Furthermore, this concept can be extended to specifically prioritize predicted peptides of proteins, which have only been observed with a few peptides. Predicted proteotypic peptides of these proteins, including retention time and acquisition parameters, can be preferentially selected for fragmentation, confirming or supporting the identification of (thus far) poorly identified proteins.

2.4 Protein fingerprints

The protein inference problem is one of the most challenging issues in proteomics[43]. While recent studies suggest that only one predominant isoform per gene is expressed per cell type[44], the unambiguous identification of specific isoforms using proteomics is still a challenge due to shared peptides. In part, this can only be tackled by increasing the depth at which complex mixtures of peptides are analyzed. This will not only generate more peptide identifications, potentially identifying isoform specific peptides, but also enable the systematic integration of quantification information of peptides during protein inference[45].

In theory, a protein present in a sample should leave a unique fingerprint of multiple *m/z* and retention time features. Limited by sampling depth and dynamic range, most of them are likely currently not visible or not identified by commonly employed acquisition approaches and thus cannot be used in bioinformatics workflows. However, the combination of the approaches

proposed here might enable the systematic identification and quantification of such fingerprints by increasing depth, coverage and specificity. With increased dynamic range and match-between-runs, these fingerprints can be extracted and could enable the differentiation of protein isoforms. Additionally, the models learned from synthetically generated peptides and proteins can be used to predict expected fingerprints under the hypothesis that a specific protein is present. If visible but unidentified, these features could be used as confirmatory information during protein inference by supporting the presence of a protein via additional features, boosting q-values of proteins which were only observed with e.g. one peptide in a sample. In theory, the lack of specific features can also be used to deduce the presence of truncates protein isoforms. A model requiring the presence of such fingerprints should decrease the number of false matches by requiring many features to be present simultaneously. The presence of two or more features at expected or predicted retention times is much less likely for decoy proteins than for targets and thus likely increases specificity.

3 References

1 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* 41, D991-995, doi:10.1093/nar/gks1193 (2013).

2 Parkinson, H. *et al.* ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research* 39, D1002-1004, doi:10.1093/nar/gkq1040 (2011).

3 Dolinski, K. & Troyanskaya, O. G. Implications of Big Data for cell biology. *Molecular biology of the cell* 26, 2575-2578, doi:10.1091/mbc.E13-12-0756 (2015).

4 Reality check on reproducibility. *Nature* 533, 437, doi:10.1038/533437a (2016).

5 Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576-582, doi:10.1038/nature14129 (2015).

6 Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics* 47, 373-380, doi:10.1038/ng.3242 (2015).

7 Zheng-Bradley, X. & Flicek, P. Applications of the 1000 Genomes Project resources. *Briefings in functional genomics*, doi:10.1093/bfgp/elw027 (2016).

8 Tyanova, S. *et al.* Proteomic maps of breast cancer subtypes. *Nature communications* 7, 10259, doi:10.1038/ncomms10259 (2016).

9 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346-352, doi:10.1038/nature10983 (2012).

10 Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics* 8, 33, doi:10.1186/s12920-015-0108-y (2015).

11 Shanmugam, A. K. & Nesvizhskii, A. I. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. *Journal of proteome research* 14, 5169-5178, doi:10.1021/acs.jproteome.5b00504 (2015).

12 Vizcaino, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* 32, 223-226, doi:10.1038/nbt.2839 (2014).

13 Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nature reviews. Genetics* 14, 333-346, doi:10.1038/nrg3433 (2013).

14 Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nature methods* 8, 61-65, doi:10.1038/nmeth.1527 (2011).

15 Smith, L. M., Kelleher, N. L. & Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nature methods* 10, 186-187, doi:10.1038/nmeth.2369 (2013).

16 Acharya, S., Edwards, S. & Schmidt, J. Research highlights: nanopore protein detection and analysis. *Lab on a chip* 15, 3424-3427, doi:10.1039/c5lc90076j (2015).

17 Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. *Nat Nano*, doi:10.1038/nnano.2016.120 (2016).

18 Larance, M. & Lamond, A. I. Multidimensional proteomics for cell biology. *Nature reviews. Molecular cell biology* 16, 269-280, doi:10.1038/nrm3970 (2015).

19 Manolio, T. A. *et al.* Implementing genomic medicine in the clinic: the future is here. *Genetics in medicine : official journal of the American College of Medical Genetics* 15, 258-267, doi:10.1038/gim.2012.157 (2013).

20 Aebersold, R. A stress test for mass spectrometry-based proteomics. *Nature methods* 6, 411-412, doi:10.1038/nmeth.f.255 (2009).

21 Tabb, D. L. *et al.* Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *Journal of proteome research* 15, 691-706, doi:10.1021/acs.jproteome.5b00859 (2016).

22 Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods* 9, 555-566, doi:10.1038/nmeth.2015 (2012).

23 Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* 33, 743-749, doi:10.1038/nbt.3267 (2015).

24 Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature methods* 13, 651-656, doi:10.1038/nmeth.3902 (2016).

25 Shannon, C. E. The mathematical theory of communication. 1963. *M.D. computing : computers in medical practice* 14, 306-317 (1997).

26 Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. & Bacardit, J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics : a journal of integrative biology* 17, 595-610, doi:10.1089/omi.2013.0017 (2013).

27 Kelchtermans, P. *et al.* Machine learning applications in proteomics research: how the past can boost the future. *Proteomics* 14, 353-366, doi:10.1002/pmic.201300289 (2014).

28 Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nature biotechnology* 25, 125-131, doi:10.1038/nbt1275 (2007).

29 Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 12, 1111-1121, doi:10.1002/pmic.201100463 (2012).

30 Moruz, L. *et al.* Chromatographic retention time prediction for posttranslationally modified peptides. *Proteomics* 12, 1151-1159, doi:10.1002/pmic.201100386 (2012).
31 Spicer, V. *et al.* N-capping motifs promote interaction of amphipathic helical peptides with hydrophobic surfaces and drastically alter hydrophobicity values of individual amino acids. *Analytical chemistry* 86, 11498-11502, doi:10.1021/ac503352h (2014).
32 Dong, N. P. *et al.* Prediction of peptide fragment ion mass spectra by data mining techniques. *Analytical chemistry* 86, 7446-7454, doi:10.1021/ac501094m (2014).
33 Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* 518, 529-533, doi:10.1038/nature14236 (2015).
34 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484-489, doi:10.1038/nature16961 (2016).
35 Bahrampour, S., Ramakrishnan, N., Schott, L. & Shah, M. Comparative Study of Deep Learning Software Frameworks. *CoRR* abs/1511.06435 (2015).
36 Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* 16, 104-117, doi:10.1093/bib/bbt080 (2015).
37 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* 13, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).
38 Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & cellular proteomics : MCP* 11, M111 014050, doi:10.1074/mcp.M111.014050 (2012).
39 Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nature neuroscience* 18, 1819-1831, doi:10.1038/nn.4160 (2015).
40 Pasa-Tolic, L., Masselon, C., Barry, R. C., Shen, Y. & Smith, R. D. Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques* 37, 621-624, 626-633, 636 passim (2004).
41 Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J. & Mann, M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Molecular & cellular proteomics : MCP* 11, M111 013185, doi:10.1074/mcp.M111.013185 (2012).
42 Kreimer, S. *et al.* Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling. *Journal of proteome research*, doi:10.1021/acs.jproteome.6b00312 (2016).
43 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* 4, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).
44 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419, doi:10.1126/science.1260419 (2015).
45 He, Z. *et al.* Protein inference: A protein quantification perspective. *Computational biology and chemistry* 63, 21-29, doi:10.1016/j.compbiolchem.2016.02.006 (2016).

Acknowledgement

Es ist vermessen zu glauben man hat sich den Doktor selbst verdient. Eine gehörige Menge Glück durch die richtigen Situationen und Menschen immer und immer wieder in die richtige Richtung gestoßen und gezogen zu werden gehört dazu. Und gerade mein Weg war gepflastert mit damals richtig geglaubten Entscheidungen. Meine Familie, und ganz besonders meine Eltern, sind hier wohl zuerst zu nennen. Ohne sie wäre ich heute nicht in der Lage diese Arbeit einzureichen, da ich schon längest vom Weg abgekommen wäre. Im übertragenen Sinne, Danke für alle Fesseln, Prellungen und Streitereien. Natürlich geht mein Danke hier auch an Lotti, die mich in der ganzen Zeit ertragen hat und mir den Rücken freigehalten hat. Wahrscheinlich wäre ich ohne dich in dieser Zeit sowohl mental als auch physisch verhungert.

Steine sind nicht gerade die besten Gesprächspartner. Zum Glück habe ich jedoch hervorragende am Lehrstuhl für Proteomik und Bioanalytik unter der Leitung von Prof. Bernhard Küster kennenlernen dürfen. Über die Zeit sind viele Kollegen enge Freunde geworden mit denen ich gern jede Minute am verbracht habe und werde. Für jeden Topf gibt es einen Deckel und so gibt es am Lehrstuhl für jedes Problem eine Person die helfen konnte. Vielen Danke an alle für die großartige Zeit, die offenen Ohren, die Diskussionen und die Hilfe. Nicht nur wissenschaftlich sondern auch privat. Nicht zuletzt Bernhard selbst hat hier einen großen Beitrag zu meiner wissenschaftlichen und persönlichen Weiterentwicklung geliefert. An der Stelle möchte ich mich auch bei alle meinen Studenten bedanken. Jeder von euch hat einen Beitrag zu dieser Arbeit geliefert.

Auch bei allen Kollegen und Freunden bei SAP in Walldorf und Potsdam, die mich während der Zeit begleitet und aktiv an ProteomicsDB mitgearbeitet haben, möchte ich mich bedanken. Durch euer Engagement hat die Arbeit nicht nur Spaß gemacht, sondern hat maßgeblich mit zum Erfolg geführt.

Des Weiteren gilt mein Dank auch Prof. Dr. Hans-Werner Mewes und Prof. Dr. Oliver Kohlbacher für die Bereitschaft als weitere Gutachter, sowie Prof. Dr. Dmitrij Frishman für die Übernahme des Prüfungsvorsitzt.

Zu guter Letzt bleiben viele andere Freunde und Mentoren die mich im Laufe meiner Ausbildung ein Stück gestoßen und gezogen haben. Besonders Bielefeld und Boston sind hier als wichtige Abschnitte in meinem Leben zu nennen. Auch wenn euer Beitrag vielleicht nicht direkt offensichtlich ist, ihr alle habt daran mitgewirkt. Vor allem in Bielefeld wurden viele Grundsteine gelegt und ich bin jedem einzelnen von euch dankbar dafür.

Publication record

Main Publications of this thesis:

- 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 Proteomics. doi: 10.1074/mcp.M114.046995
- Wilhelm M*, Schlegl J*, Hahne H*, Moghaddas Gholami A*, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B (2014) Mass-spectrometry-based draft of the human proteome. Nature. doi: 10.1038/nature13319

Additional publications submitted:

- Zolg D*, Wilhelm M*, Schnatbaum K, Zerweck J, Knaute T, Delanghe B, Bailey DJ, Gessulat S, Ehrlich HC, Weininger M, Yu P, Schlegl J, Kramer K, Schmidt T, Moritz RL, Aebersold R, Wenschuh H, Moehring T, Aiche S, Huhmer A, Reimer U, Kuster B (2016; accepted for publication) Building ProteomeTools based on a complete synthetic human proteome.
- Klaeger S*, Heinzlmeir S*, Wilhelm M*, Qiao H, Helm D, Polzer H, Vick B, Reiter K, Reinecke M, Ruprecht B, Petzoldt S, Koch H, Schoof M, Canevari G, Casale E, Re Depaolini S, Feuchtinger A, Meng C, Wu Z, Zecha J, Schmidt T, Rueckert L, Becker W, Huenges J, Gohlke B, Garz A, König P, Hahne H, Ruland J, Preissner R, Götze K, Kayser G, Tönisson N, Greif P, Schlegl J, Ehrlich H, Aiche S, Felder ER, Kramer K, Schneider S, Walch A, Médard G, Jeremias I, Spiekermann K, Kuster B (2016; submitted) The target landscape of clinical kinase drugs.
- Frejno M, Zenezini CR, Wilhelm M, Koch K, Zheng R, Klaeger S, Meng C, Jarzab A, Heinzlmeir S, Johnstone E, Domingo E, Kerr D, Jesinghaus M, Slotta-Huspenina J, Knapp S, Feller SM, Kuster B (2016; submitted) Pharmacoproteomic characterisation of human colon and rectal cancer.

Additional publications during PhD:

- Koch H, Wilhelm M, Ruprecht B, Beck S, Frejno M, Klaeger S, Kuster B (2016) Phosphoproteome profiling reveals molecular mechanisms of growth factor mediated kinase inhibitor resistance in EGFR overexpressing cancer cells. J Proteome Res. doi: 10.1021/acs.jproteome.6b00621
- Heinzlmeir S*, Kudlinzki D*, Sreeramulu S, Klaeger S, Gande S, Linhard V, Qiao H, Helm D, Wilhelm M, Ruprecht B, Saxena K, Médard G, Schwalbe H, Kuster B (2016) Chemical proteomics and structural biology define EPHA2 inhibition by clinical kinase drugs, ACS Chem Biol. doi: 10.1021/acschembio.6b00709
- Yu P, Hahne H, Wilhelm M, Kuster B (2016) Ethylene glycol improves electrospray ionization efficiency in bottom-up proteomics. Anal Bioanal Chem. doi: 10.1007/s00216-016-0023-x
- Hao Y, Colak R, Teyra J, Corbi-Verge C, Ignatchenko A, Hahne H, Wilhelm M, Kuster B, Braun P, Kaida D, Kislinger T, Kim P (2015) Semi-supervised learning predicts approximately a third of the alternative splicing isoforms as functional proteins. Cell Rep. doi: 0.1016/j.celrep.2015.06.031
- Médard G, Pachl F, Ruprecht B, Klaeger S, Heinzlmeir S, Helm D, Qiao H, Ku X, Wilhelm M, Kuehne T, Wu Z, Dittmann A, Hopf C, Kramer K, Kuster B (2015) Optimized chemical proteomics assay for kinase inhibitor profiling. J Proteome Res. doi: 10.1021/pr5012608

- Helm D, Vissers JP, Hughes CJ, Hahne H, Ruprecht B, Pachl F, Grzyb A, Richardson K, Wildgoose J, Maier SK, Marx H, Wilhelm M, Becher I, Lemeer S, Bantscheff M, Langridge JI, Kuster B (2014) Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics, Mol Cell Proteomics. doi: 10.1074/mcp.M114.041038
- Moghaddas GA*, Hahne H*, Wu Z*, Auer FJ, Meng C, Wilhelm M, Kuster B (2013) Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Rep. doi: 10.1016/j.celrep.2013.07.018

Earlier publications:

- Hoffmann N, Wilhelm M, Doebbe A, Niehaus K, Stoye J (2014) BiPACE 2D - Graph-based multiple alignment for comprehensive two-dimensional gas chromatography-mass spectrometry. Bioinformatics. doi: 10.1093/bioinformatics/btt738
- Froehlich JW, Dodds ED, Wilhelm M, Serang O, Steen JA, Lee RS (2013) A classifier based on accurate mass measurements to aid large scale, unbiased glycoproteomics. Mol Cell Proteomics. doi: 10.1074/mcp.M112.025494
- Hoffmann N, Keck M, Neuweger H, Wilhelm M, Högy P, Niehaus K, Stoye J (2012) Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. BMC Bioinformatics. doi: 10.1186/1471-2105-13-214
- Wilhelm M*, Kirchner M*, Steen JA, Steen H (2012) mz5: Space- and Time-efficient Storage of Mass Spectrometry Data Sets. Mol Cell Proteomics. doi: 10.1074/mcp.O111.011379

* Authors contributed equally to this work (underlined in case Mathias Wilhelm was co-first author). Authors appear in the order of the original publication.

Curriculum vitae

PERSONAL DETAILS	Name:	Mathias Wilhelm
	Date of birth:	06.10.1984
	Address:	Technical University of Munich Chair of Proteomics and Bioanalytics Emil-Erlenmeyer-Forum 5, 85354 Freising, Germany
	Email:	mathias.wilhelm@tum.de

EDUCATION	Doctoral candidate, Bioinformatics At Technical University of Munich, Munich, Germany In collaboration with SAP SE, Walldorf, Germany Thesis: "An in-memory platform for the exploration and analysis of big data in biology"	08/2012 to present
	Master of Science, Informatics in the Natural Science At Bielefeld University, Bielefeld, Germany In collaboration with Boston Children's Hospital/Harvard Medical School, Boston, MA, USA Thesis: "Computational Mass Spectrometry Method Development and Clinical Application"	10/2009 to 10/2011
	Bachelor of Science, Bioinformatics and Genome Research At Bielefeld University, Bielefeld, Germany Thesis: "Analysis and visualization of comprehensive 2D gas chromatography coupled to time-of-flight mass spectrometry (GCxGC-TOF-MS) datasets"	10/2006 to 10/2009
WORK AND RESEARCH EXPERIENCE	Research assistant II at Boston Children's Hospital/Harvard Medical School, Boston, MA, USA	02-07/2012
	Research assistant at the Technical University Munich, Chair of Proteomics and Bioanalytics, Munich, Bavaria, Germany	11/2011 to 01/2012
	Research assistant at the Bielefeld University, Faculty of Technology, Bielefeld, North Rhine-Westphalia, Germany	04-09/2010
TEACHING	Courses:	
	Intensive Course Proteomics, Technical University of Munich	2014/15/16
	Advanced Java Programming, Bielefeld University	2009/10/11
	Computer Science Preparation Course, Bielefeld University	2008/09/10
	Master projects:	
	Assessing variability in large scale protein-drug interaction studies using proteomics, Technical University of Munich	2015
	Improving spectra identification of isobaric labeled peptides and its application in assay data, Technical University of Munich	2015
	Bachelor projects:	

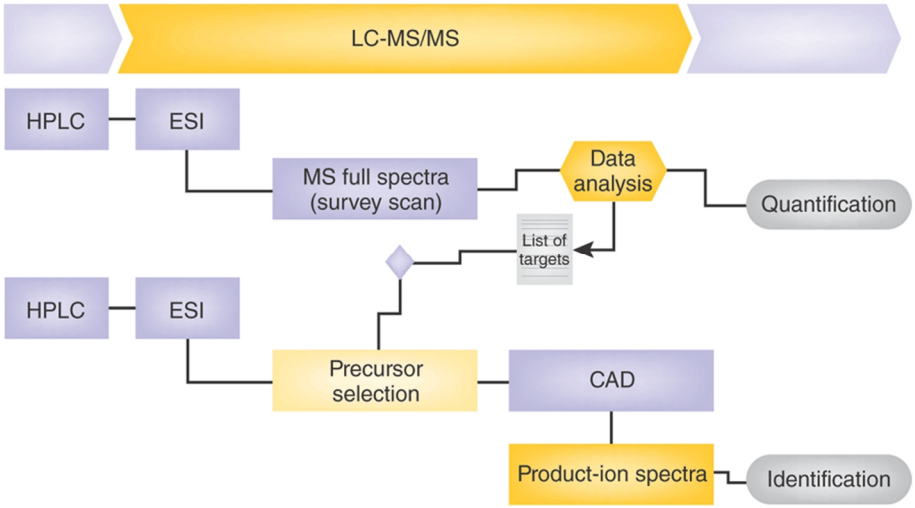
	Enabling rapid and easy data access to ProteomicsDB via a stand-alone Java application, Technical University of Munich	2015
	Web-based interactive visualization of multidimensional datasets using parallel coordinates, Technical University of Munich	2015
	Retention time prediction of phosphopeptides, Technical University of Munich	2014
	Improving of human gene models by mass spectrometry based proteomics, Technical University of Munich	2013
INVITED TALKS	Mass-spectrometry-based draft of the human proteome Thermo Scientific User Meeting, Munich, Germany	07/2015
	Mass-spectrometry-based draft of the human proteome Thermo Scientific User Meeting, Berlin, Germany	08/2014
	Mass-spectrometry-based draft of the human proteome 10th Siena Meeting – From Genome To Proteome: 20th years of Proteomics, Siena, Italy	08/2014
TALKS AND WORKSHOPS	Mass spectrometry centric analysis of public proteomic data in ProteomicsDB 63rd ASMS Conference on Mass Spectrometry and Allied Topics, St Louis, MO, USA	06/2015
	Picked protein FDR, a scalable approach for protein false discovery rate estimation in large proteomic data sets 63rd ASMS Conference on Mass Spectrometry and Allied Topics, St Louis, MO, USA	06/2015
	Workshop: ProteomicsDB 63rd ASMS Conference on Mass Spectrometry and Allied Topics, St Louis, MO, USA	06/2015
	ProteomicsDB: An in-memory data platform to explore and analyse the human proteome Big Data and Predictive Computational Modeling, Garching, Germany	05/2015
	A refined mass spectrometry based draft of the human proteome US HUPO - 11th Annual Conference, Tempe, AZ, USA	03/2015
	Workshop: ProteomicsDB 62nd ASMS Conference on Mass Spectrometry and Allied Topics, Baltimore, MD, USA	06/2014
	ProteomicsDB 4th ProteomeXchange meeting, Rüdesheim, Germany	04/2014

Appendix

Contents

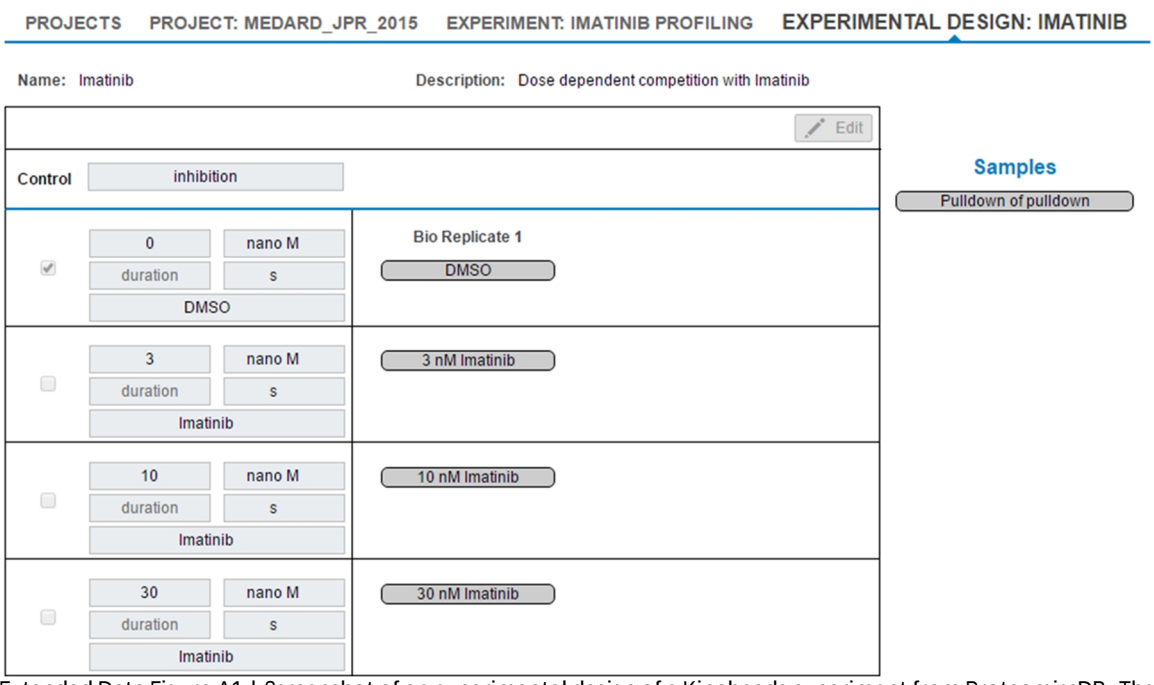
"... in rabbits, hares, and some other herbivores
it is involved in the digestion of cellulose."
- Unknown

0 - Introduction

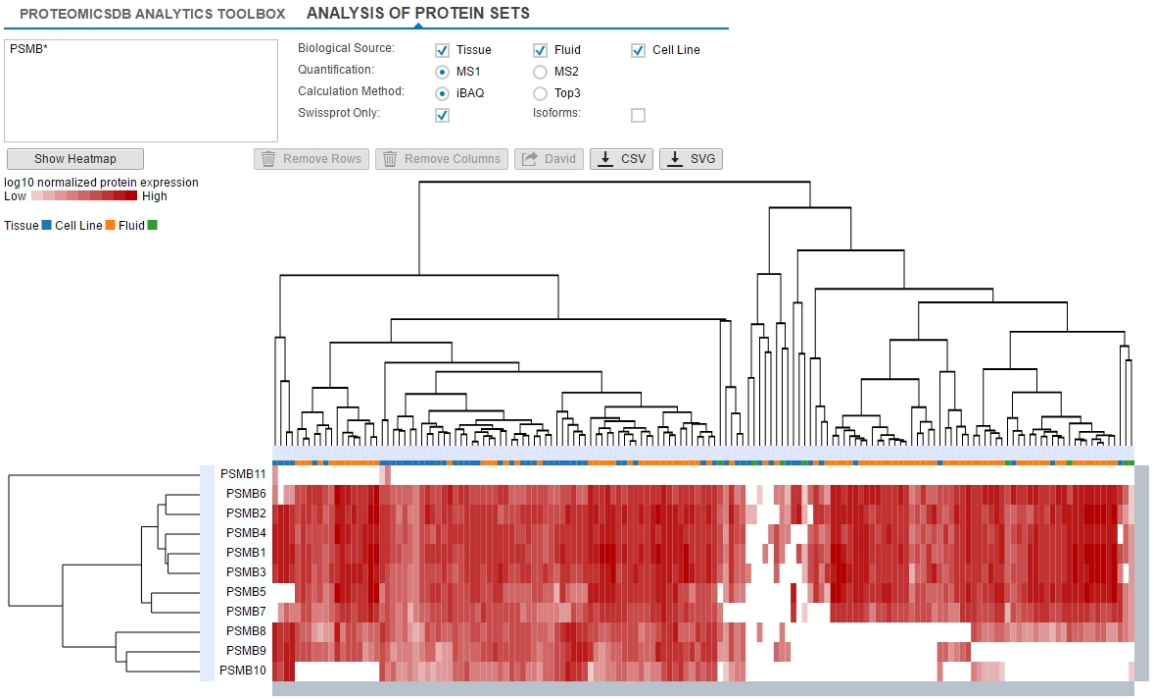


Extended Figure 01 | Acquisition schema of a directed bottom-up shotgun experiments. The sample is first analyzed in LC-MS mode, and the results are analyzed using a suite of bioinformatic tools to quantify the peptides. Typically, peptides that are of particular interest (e.g., those that are regulated by comparing multiple samples) are included in a list of targets for MS/MS sequencing. In a second step, the sample is reanalyzed to sequence exclusively the peptide ions present on the target list. The resulting MS/MS spectra enable the amino acid sequence to be determined.

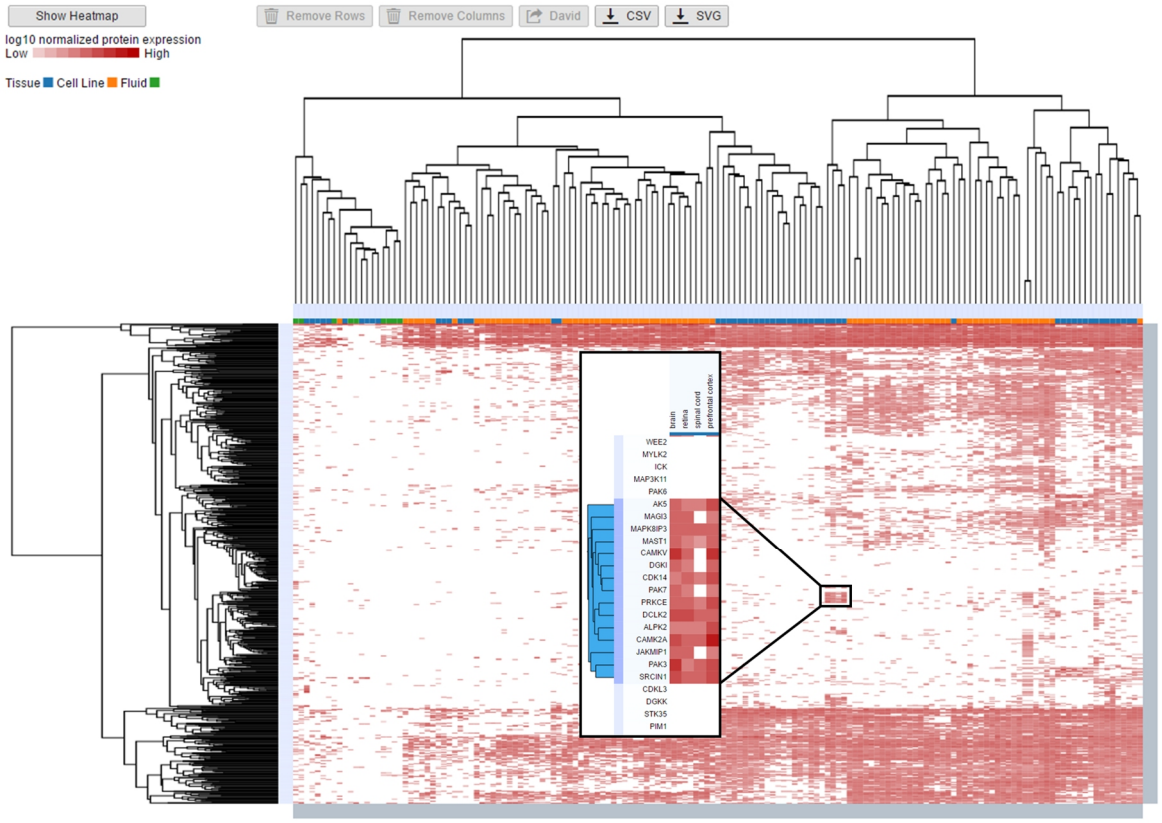
A - ProteomicsDB



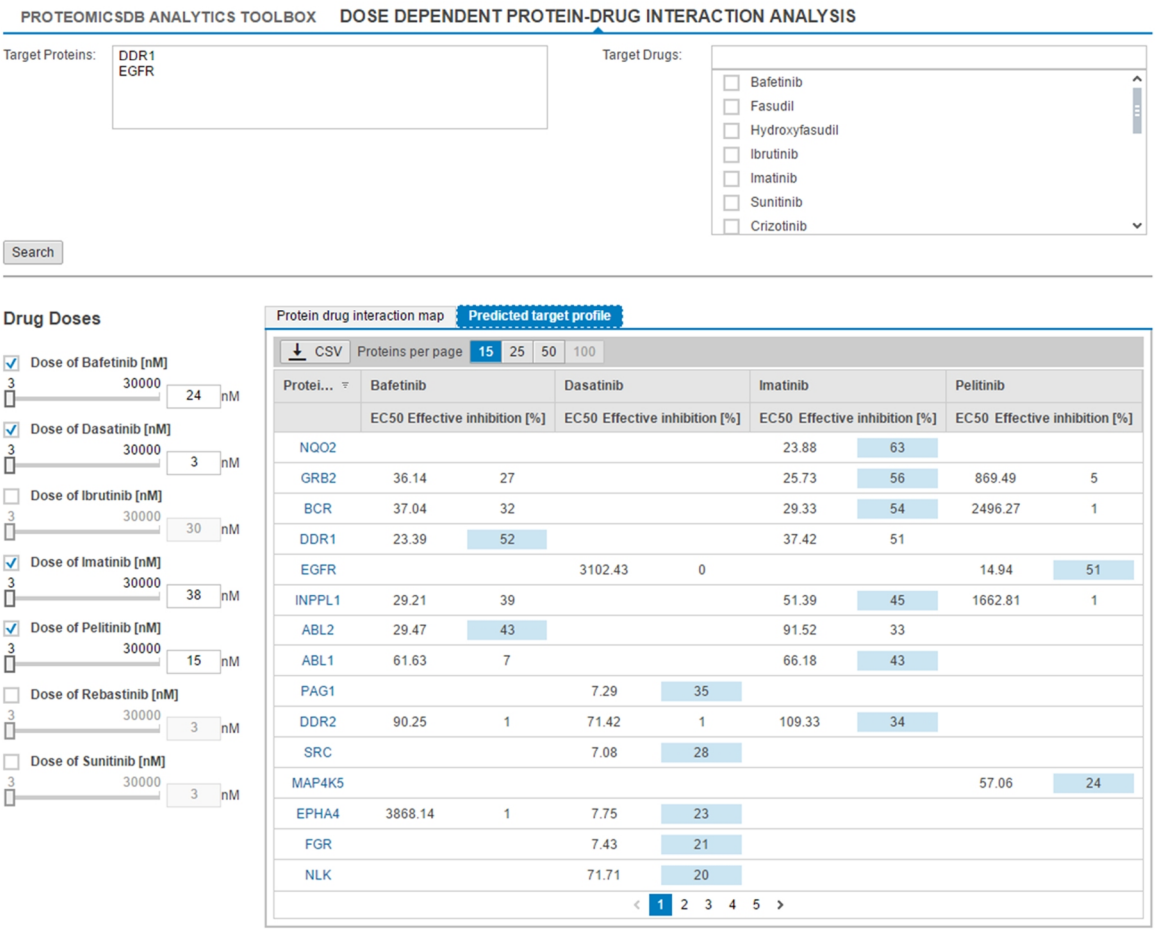
Extended Data Figure A1 | Screenshot of an experimental design of a Kinobeads experiment from ProteomicsDB. The experiment consists of one treatment (column, inhibition). Each condition (e.g. 3 mM Imatinib) was measured with one biological replicate. All samples assigned to the experiment, but not yet used in the experimental design, are listed on the right-hand side and can be moved by drag-and-drop to specific conditions.



Extended Data Figure A2 | Screenshot of the protein expression heatmap of all major components of the proteasome complex from ProteomicsDB. The online heatmap allows to search for a custom set of proteins (here PSMB*). This figure is an extension to Figure 2.15 and shows the expression of the major components of the proteasome across tissues, fluids and cell lines.

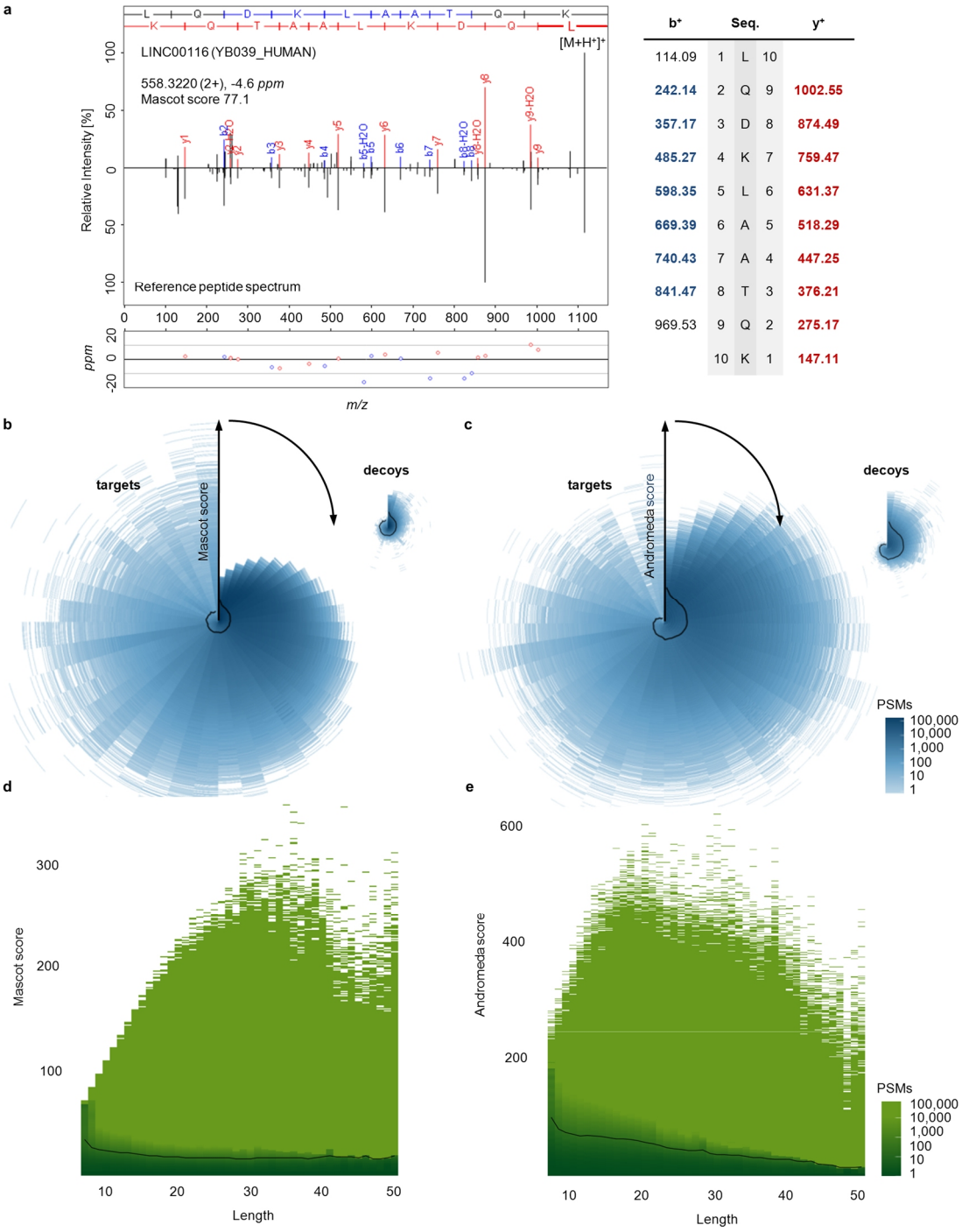


Extended Data Figure A3 | Screenshot of the protein expression heatmap of all proteins which have the term “kinase” in their description. The highlighted cluster contains exclusively neuronal tissues (brain, retina, spinal cord, prefrontal cortex). The 15 marked proteins, such as the protein kinases CAMK2A, included in this cluster are associated with learning and memory.

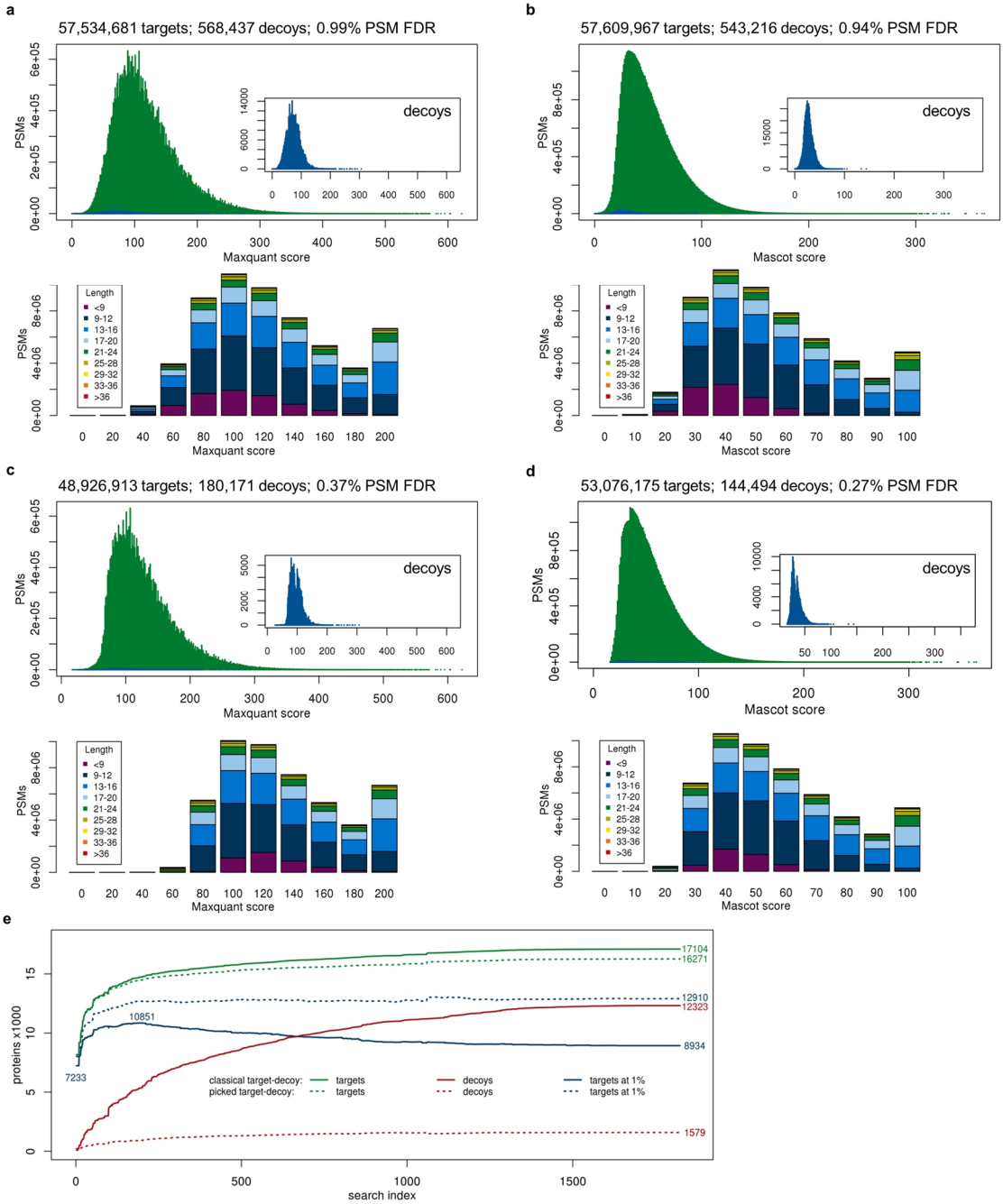


Extended Data Figure A4 | Screenshot of a dose-dependent protein-drug interaction map from ProteomicsDB. The protein-drug interaction landscape of four selected (checkbox) kinase inhibitors which target either DDR1 or EGFR enables the visual inspection of inhibitors and their combination. This figure is the extension to Figure 2.18 and shows the predicted effects of drugs (columns) on proteins (rows) based on the stored curve fits.

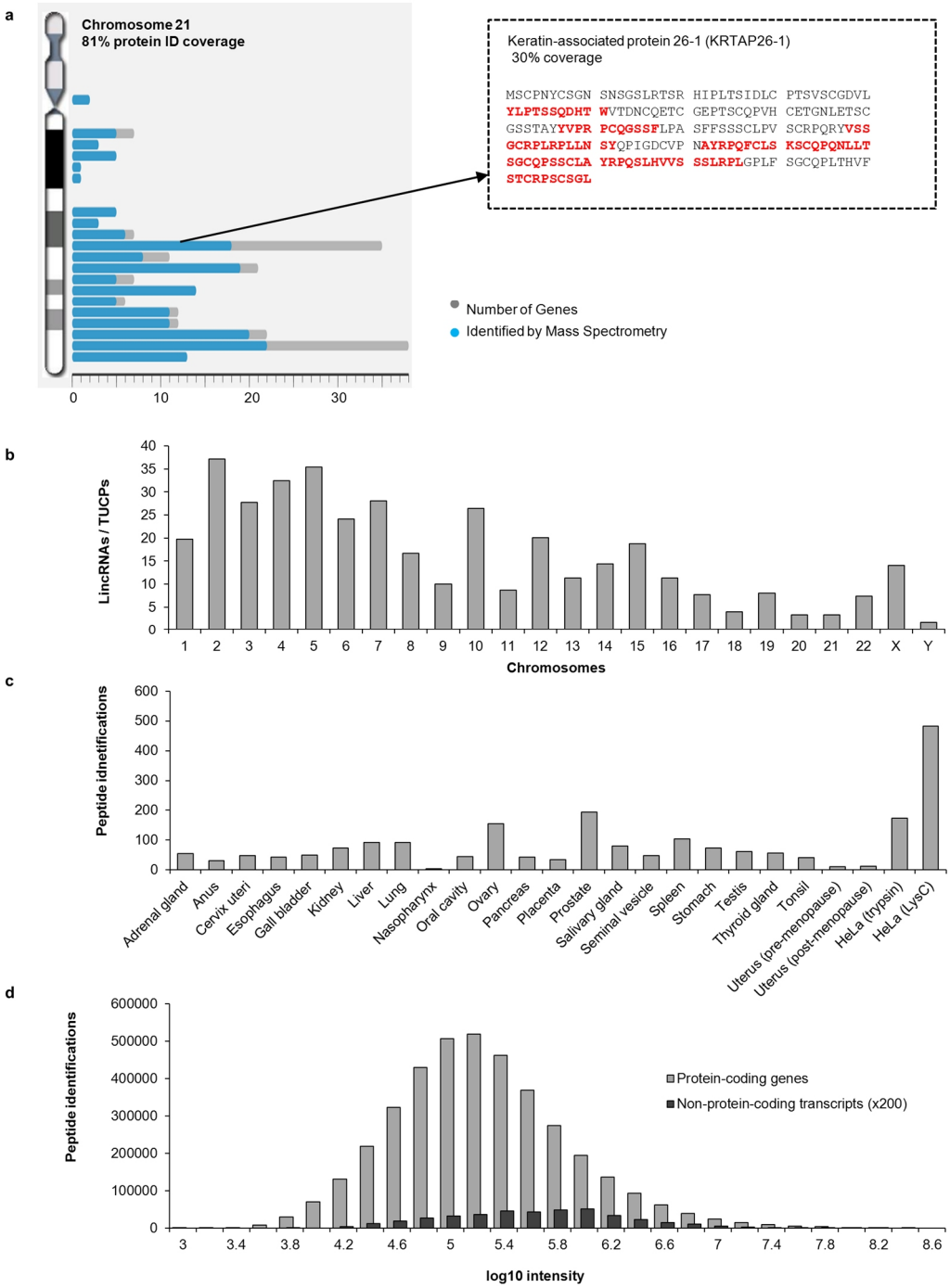
B - Mass spectrometry based draft of the human proteome



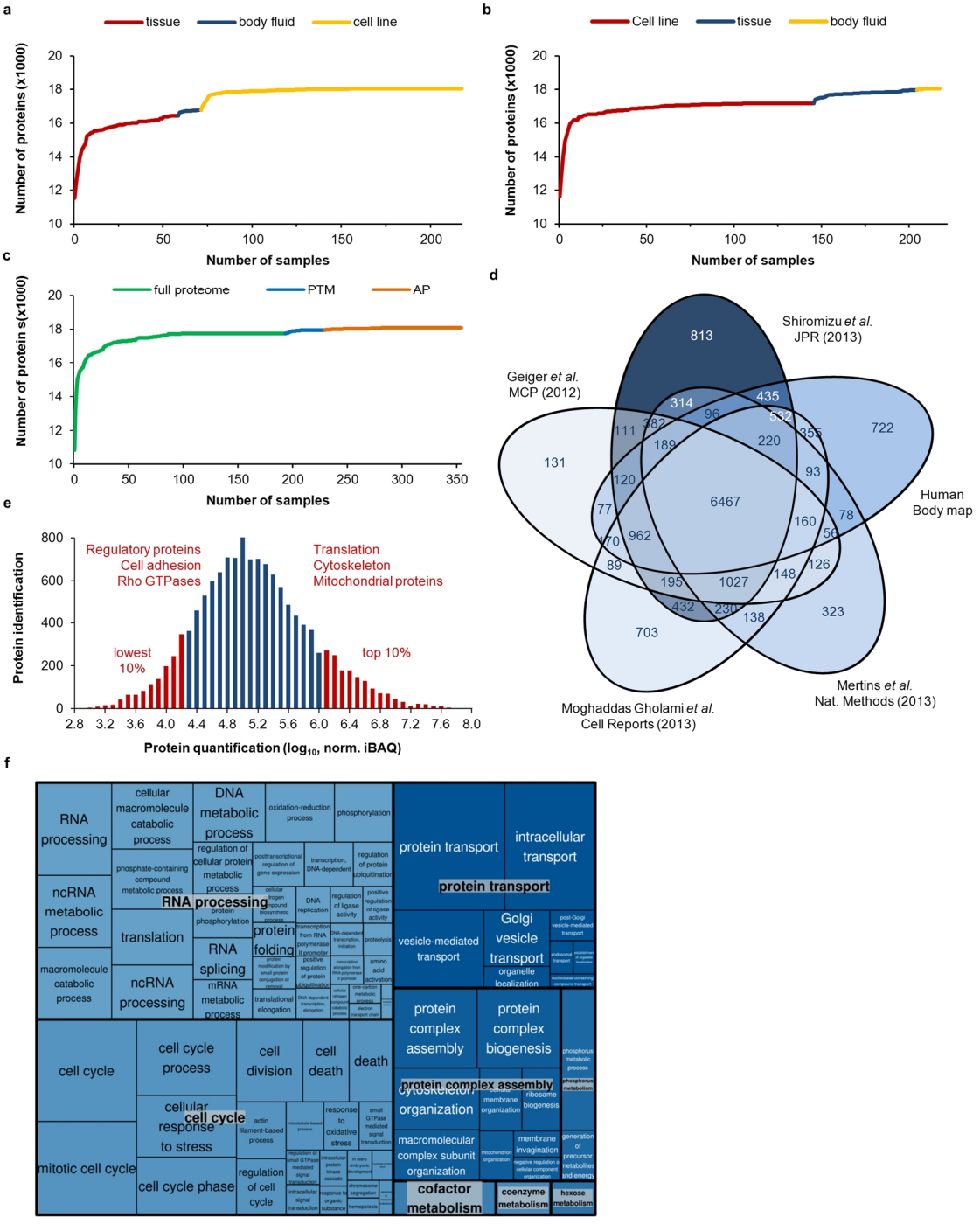
Extended Data Figure B1 | Peptide and protein identifications. a, Spectrum viewer enabling access to >70 million annotated tandem mass spectra of endogenous peptides and synthetic reference standards in real time. b, Peptide and protein identification criteria followed a two-step process. 1. For each LC-MS/MS run, we applied a global 1% target-decoy False Discovery Rate (FDR) cut on the level of peptide spectrum matches (PSMs, not shown). 2. In addition, we applied a peptide length dependent local FDR cut of 5% for all PSMs (prior to 1% filtering using all PSMs). Peptide length and score distribution for targets and decoys for the search engine Mascot. c, same as in a but for Andromeda. d, e Heat maps showing false discovery rates as a function of search engine score and peptide length. Solid lines indicate the 5% local FDR.



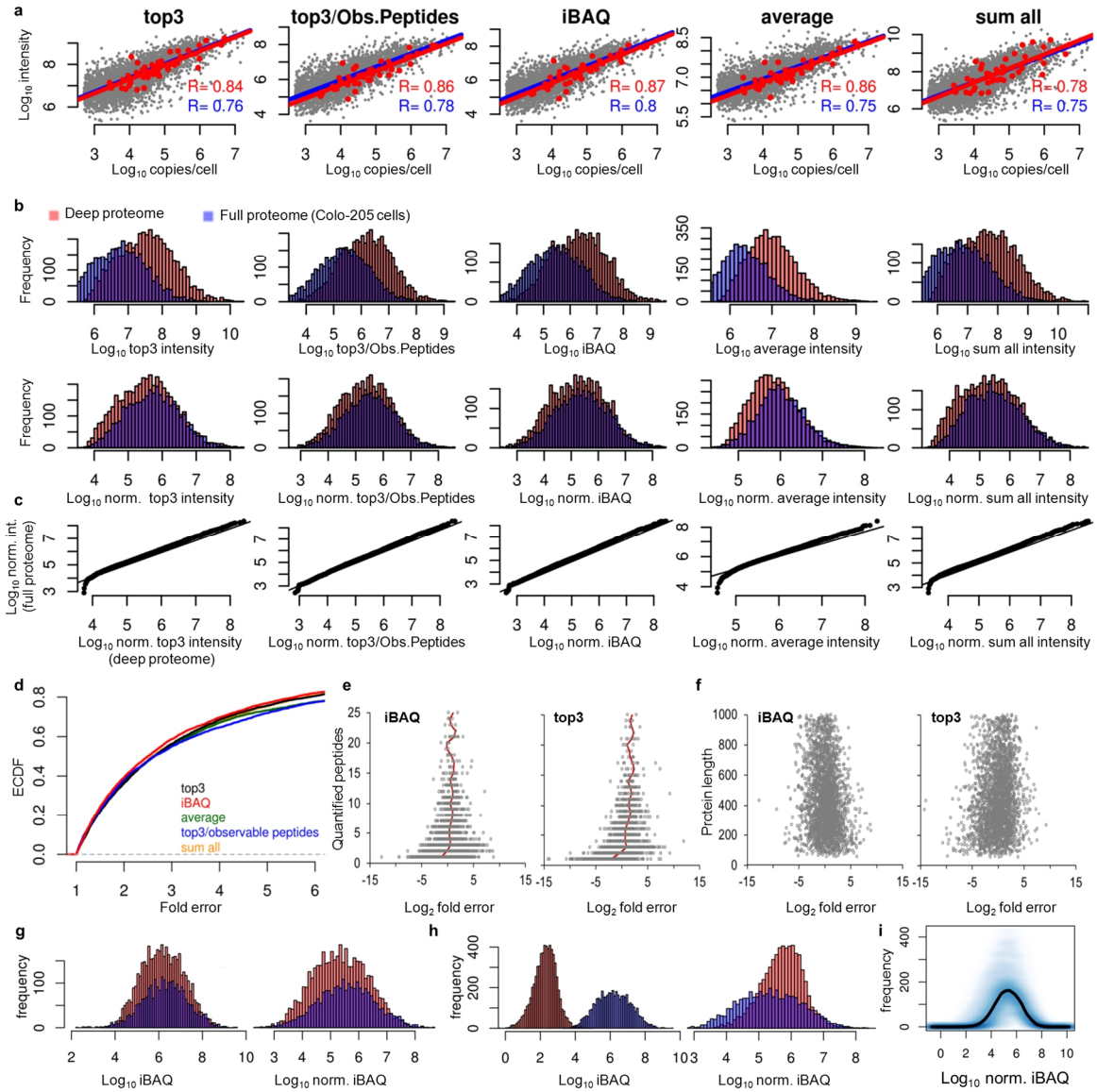
Extended Data Figure B2 | Protein identification quality in very large data sets. a, The first step filters every LC-MS/MS run at 1% PSM FDR. Upper panel: Score distribution for target and decoy PSMs following 1% PSM FDR filtering for Maxquant identifications. Lower panel: shows the binned peptide length distribution for target PSMs. b, Same as a, but for Mascot identifications. c, Second filtering step. Same as a, but this time applying an additional 5% local length and score dependent FDR on the total aggregated data for Maxquant identifications in ProteomicsDB. It is apparent, that the second filtering step improves the FDR about 3-fold and removes most PSMs shorter than 9 amino acids. d, Same as c, but for Mascot identifications. e, Comparative analysis of protein FDR characteristics of two different approaches based on Mascot analysis. In the classical target-decoy approach, aggregation of large quantities of data leads to accumulation of large numbers of decoy proteins and a concomitant loss of true target proteins when filtering the data at 1% protein FDR. The alternative 'picked' target-decoy method does not suffer from this scaling problem and maintains a constant decoy rate (and therefore lower protein FDR) but at the expense of lower sensitivity of target protein detection compared to the classical target-decoy approach.



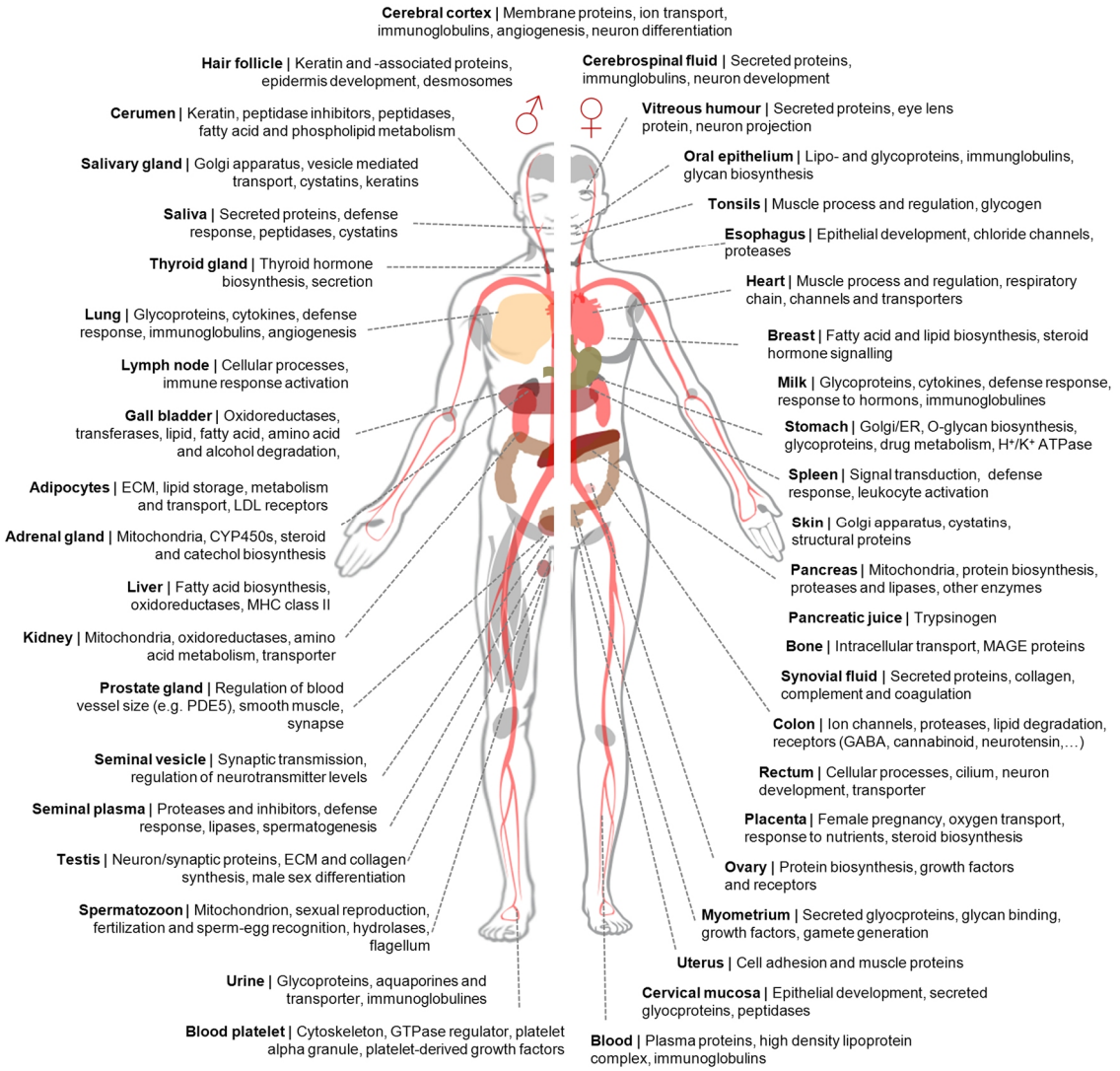
Extended Data Figure B3 | Further characterization of the proteome. a, Some proteins are refractory to identification using tryptic digestion because they do not generate sufficient/any peptides that are within the productive mass range of a mass spectrometer typically used for bottom up proteomics. This can be improved by the use of alternative proteases e.g. chymotrypsin as shown here for one of the many keratin associated proteins localized on chromosome 21 (detected chymotryptic peptides in red). b and c, Translation of lincRNAs is rare but does exist and can be identified a across all chromosomes as well as b in many tissues and in HeLa cells. d, Peptide intensity distribution of protein-coding genes and non-coding transcripts. Interestingly, the abundance of translated lincRNAs is broadly similar to that of classical proteins.



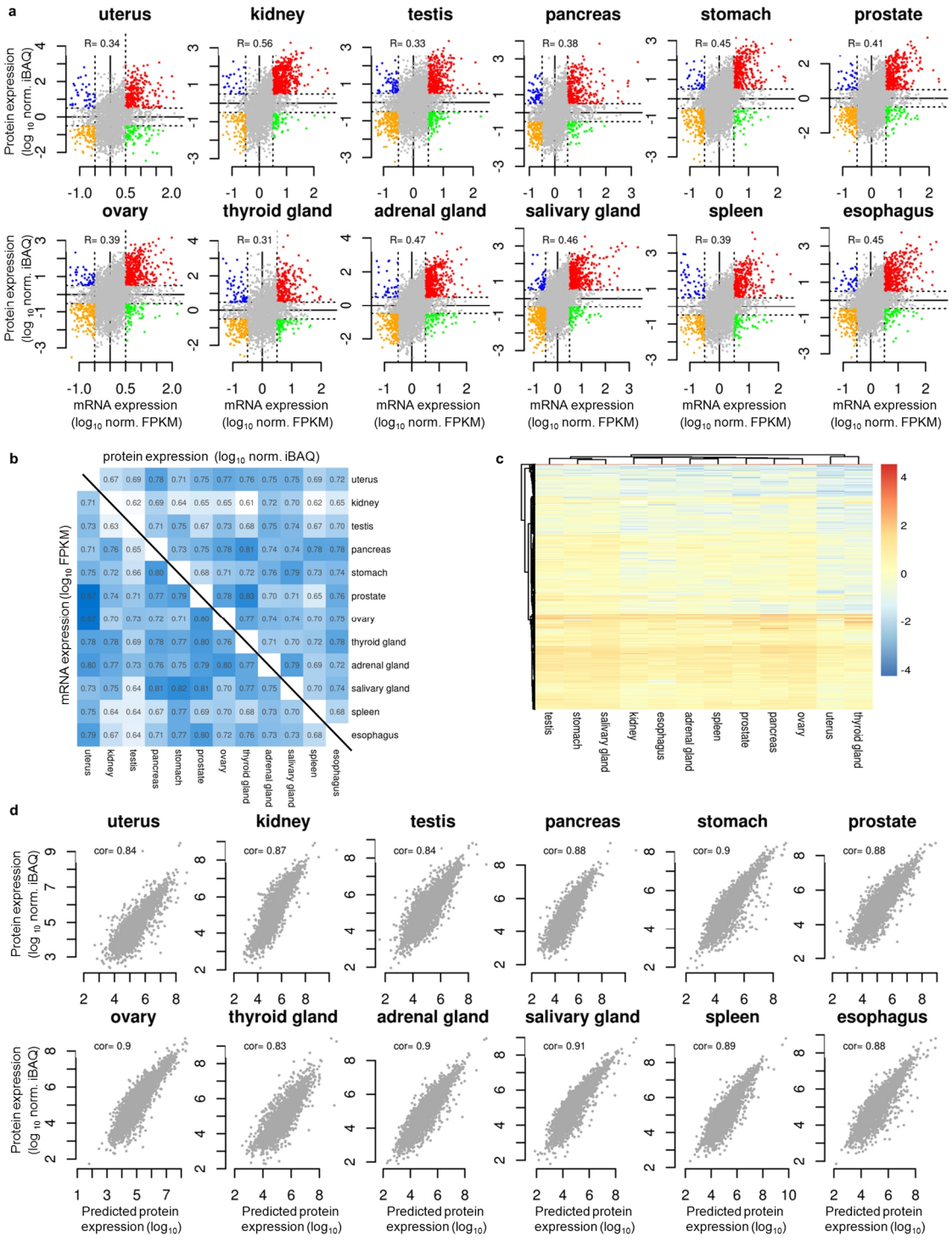
Extended Data Figure B4 | Further characterization of the proteome. a, Proteome coverage rapidly saturates with the addition of shotgun proteomic data. Tissue proteomes saturate at ~16,000 proteins but both body fluids and cell lines add small but noticeable numbers of proteins not covered in the tissues (see also b and c for a different ordering of samples). b, Same plot as a, but different ordering of samples. c, Saturation plots showing that PTMs and affinity purifications each contribute distinctly to the coverage of the proteome. d, Comparison of five large-scale projects suggesting that a 'core proteome' of 10-12,000 ubiquitously expressed proteins exists. e, Abundance distribution of the 'core proteome' based on the normalized iBAQ method. The 10% most highly expressed proteins are dominated by proteins relating to energy production and protein synthesis. The 10% lowest abundant proteins are enriched in proteins with regulatory functions. f, Tree view summary of GO term analysis for the proteins constituting the 'core proteome' showing that the core proteome is mainly concerned with biological processes relating to the homeostasis and life cycle of cells.



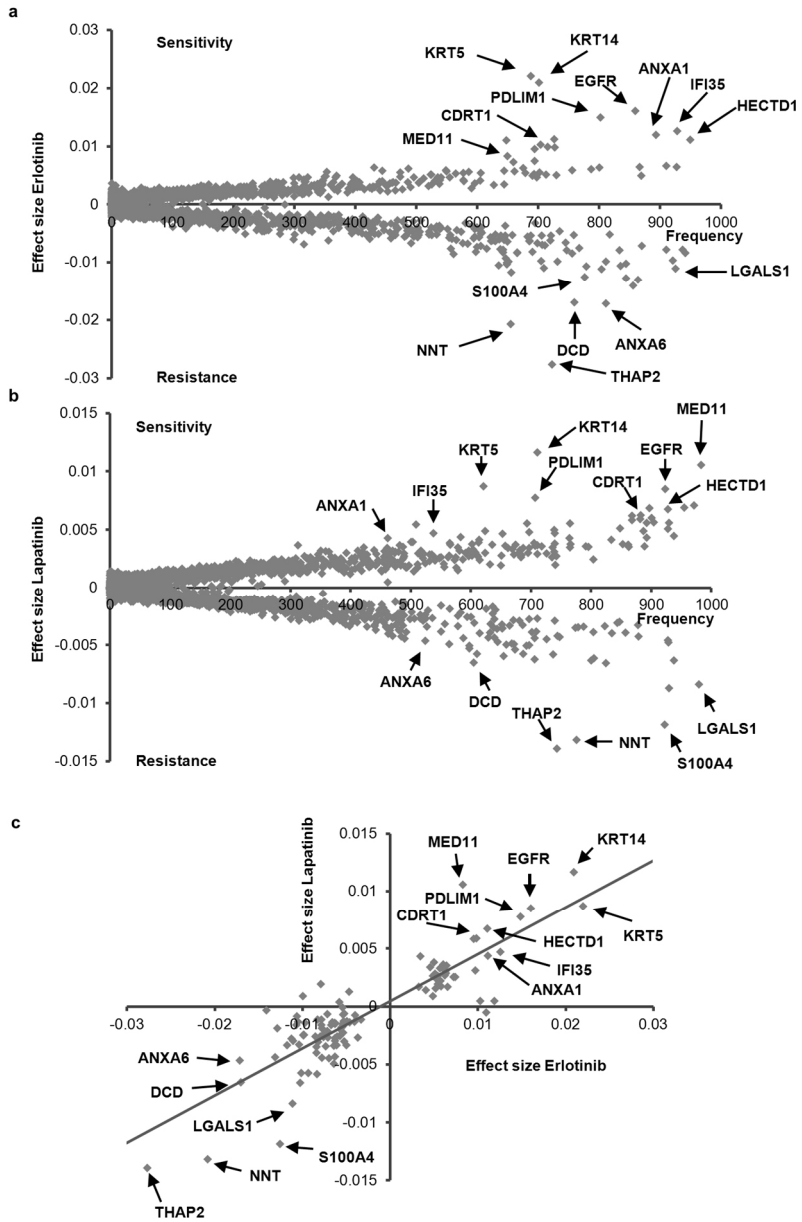
Extended Data Figure B5 | Comparative analysis of five intensity based label-free absolute quantification approaches. a, Linearity of intensity (U2-OS cell line data from Geiger *et al.*, Mol. Cell. Proteomics, 2012) and copies per cell for AQUA quantified proteins (red dots, red regression line; same cell line; Beck *et al.*, Mol. Syst. Biol. 2012) and derived copy number estimates (grey dots, blue regression line; from the same study). b, Total sum normalization re-scaled intensity distributions of Colo-205 cell digests measured on two different mass spectrometers (Orbitrap Elite data in red, LTQ Orbitrap XL data in blue; Moghaddas Gholami *et al.*, Cell Rep. 2013). c, Q-Q plots of the normalized data presented in (b) illustrating good alignment of data across 4.5 orders of magnitude. d, Empirical cumulative density function (ECDF) of error distributions derived from (a). e, Comparison of the fold error of iBAQ and top3 as a function of the number of quantified peptides. f, Same as (e) but for protein length. iBAQ shows slightly smaller errors from low peptide numbers compared to the top3 method. g, Comparison of iBAQ and total sum normalized iBAQ for heavy SILAC-labeled MCF-7 cell digests (red bars; Geiger *et al.*, Cancer Res., 2012) and label-free quantified MCF-7cell digests (blue bars) before (left panel) and after normalization (right panel) showing no influence of the presence of the SILAC label on quantification results. h, Comparison of iBAQ and total sum normalized iBAQ for iTRAQ reporter ion intensity based quantification (red bars; MCF-7 cell digest; Johansson *et al.*, Nat. Commun., 2013) and label-free quantified MCF-7 cell digests (blue bars; same as (a) and (c)) before (left panel) and after normalization (right panel). The intensity distribution characteristics of iTRAQ and label-free measurements are too different to allow for comparative analyses of MS1 and MS2 based quantification data. i, Normalized iBAQ distributions of 347 cell line and tissue proteomes (all MS1 quantified) available in ProteomicsDB showing the general applicability of MS1 based quantification across many sources of biological material.



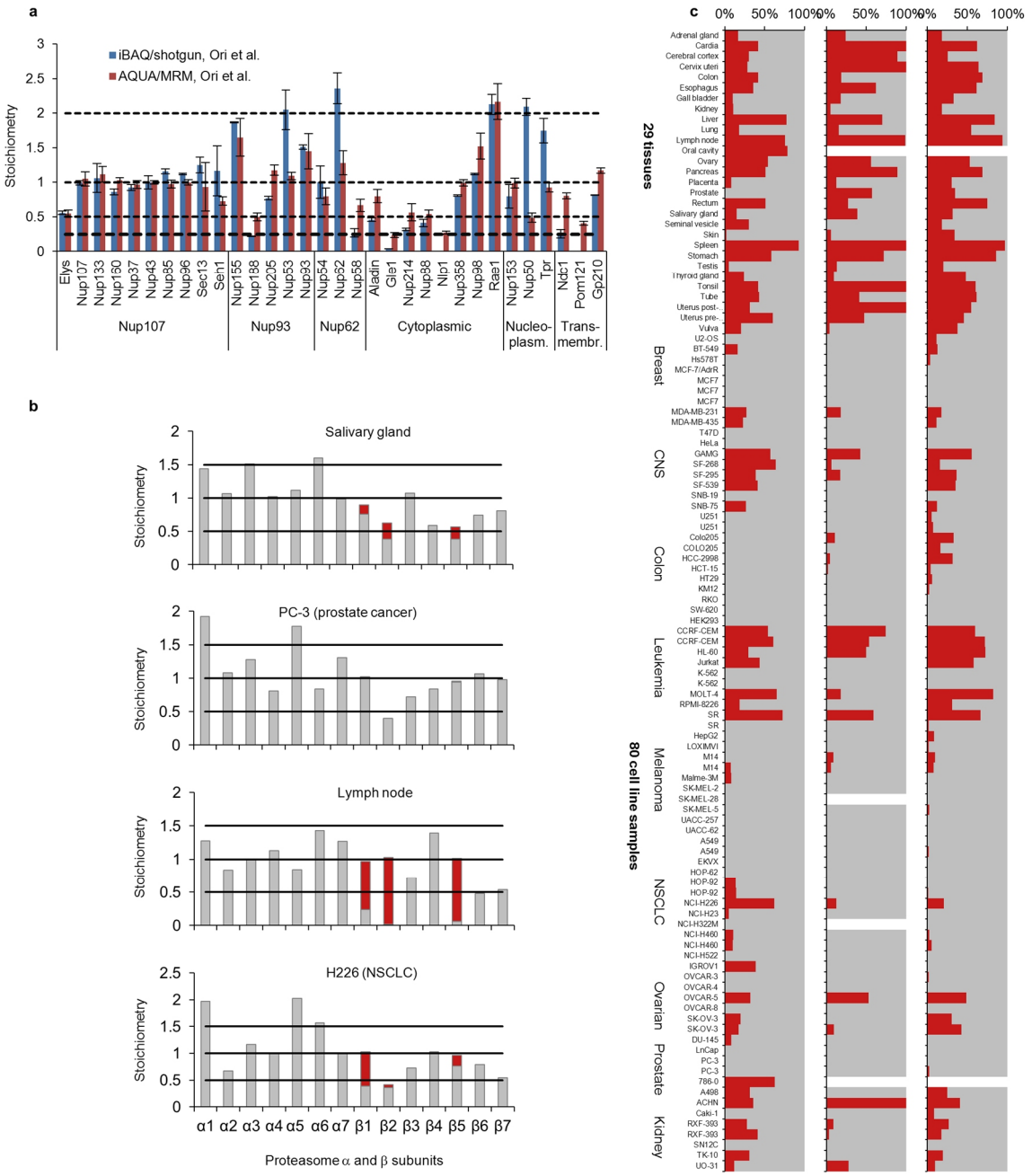
Extended Data Figure B6 | Functional protein expression analysis. Gene ontology analysis of proteins with 10x above average expression levels in a particular organ/body fluid invariably highlights protein signatures with direct organ related functional significance.



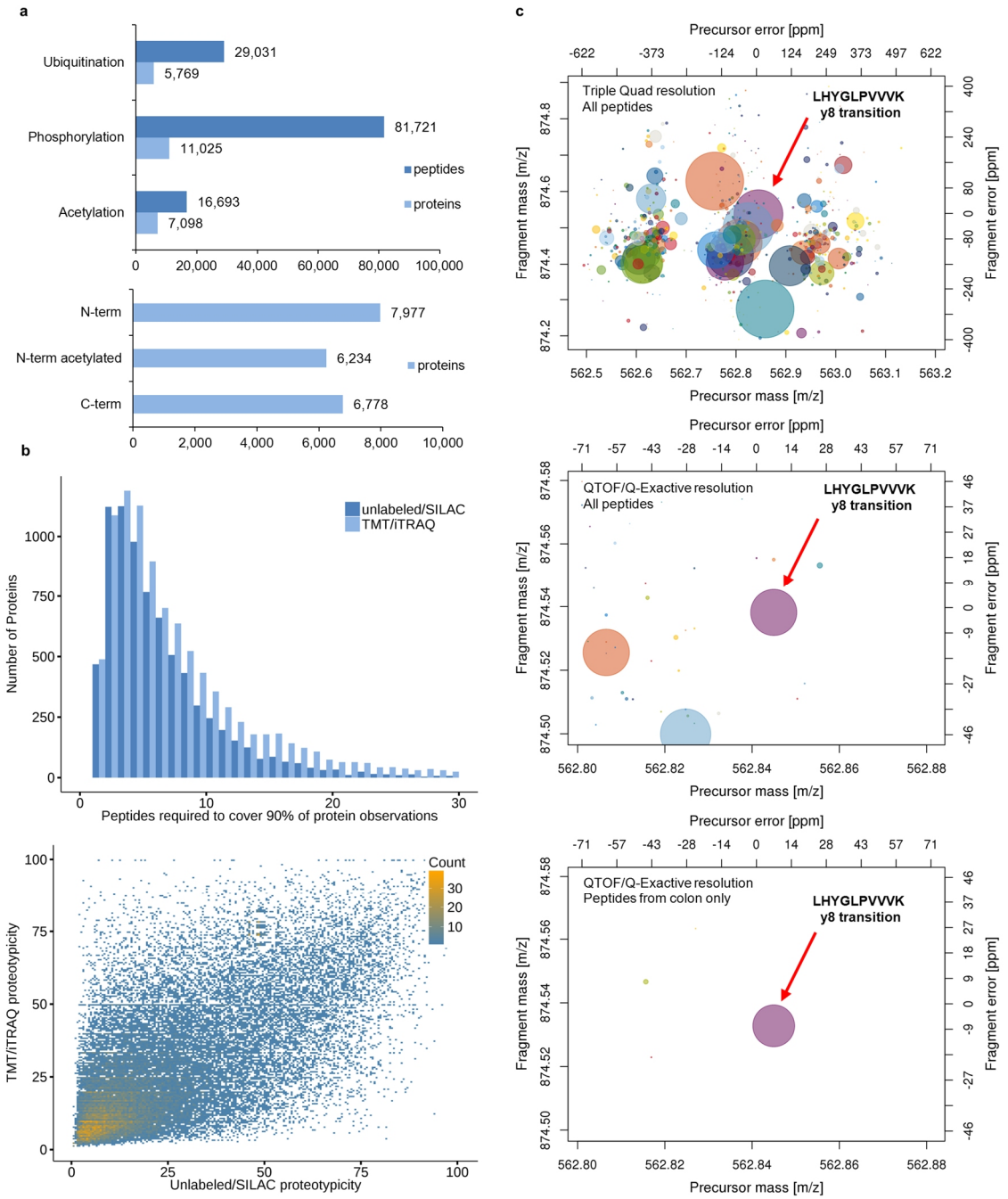
Extended Data Figure B7 | Protein vs mRNA expression analysis. a, Comparison of mRNA and protein expression of 12 human tissues showing the general rather poor correlation of protein and mRNA levels implying the widespread application of transcriptional, translational and post-translational control mechanisms of protein abundance regulation. Spearman correlation coefficients vary from 0.41 (thyroid gland) to 0.55 (kidney). 'Corner proteins' (0.5 logs to either side of zero) are marked in colors. b, Clustering of mRNA expression (left triangle) and protein expression (right triangle) across the 12 tissues does not reveal tissues with common profiles suggesting that the transcriptomes and proteomes of human tissues are quite different from each other. c, The ratio of protein and mRNA level for a protein is approximately constant across many tissues. The heatmap shows proteins and tissues clustered according to their protein/mRNA ratio. d, Using the median ratio of protein/mRNA across 12 tissues, it is possible to predict protein levels from mRNA levels for every tissue with a good correlation coefficient.



Extended Data Figure B8 | Protein markers for drug sensitivity and resistance. a, Elastic net analysis of protein expression and drug sensitivity for the EGFR kinase inhibitor Erlotinib. Positive effect size values indicate that high protein expression is associated with drug sensitivity. Negative effect size values indicate that high protein expression is associated with drug resistance. b, Same as in (a) but for the EGFR kinase inhibitor Lapatinib. c, Correlation analysis of the elastic net effect sizes for Erlotinib and Lapatinib (proteins with elastic net frequencies of <600 are not shown for clarity). Proteins in the upper right quadrant are common markers for drug sensitivity (including EGFR as the primary target of both drugs). Proteins in the bottom left quadrant are common markers for drug resistance (including S100A4, a known resistance marker for Lapatinib). Proteins that are strong markers for sensitivity or resistance are annotated in each plot and most proteins can be easily placed into EGFR signaling and regulation pathways.

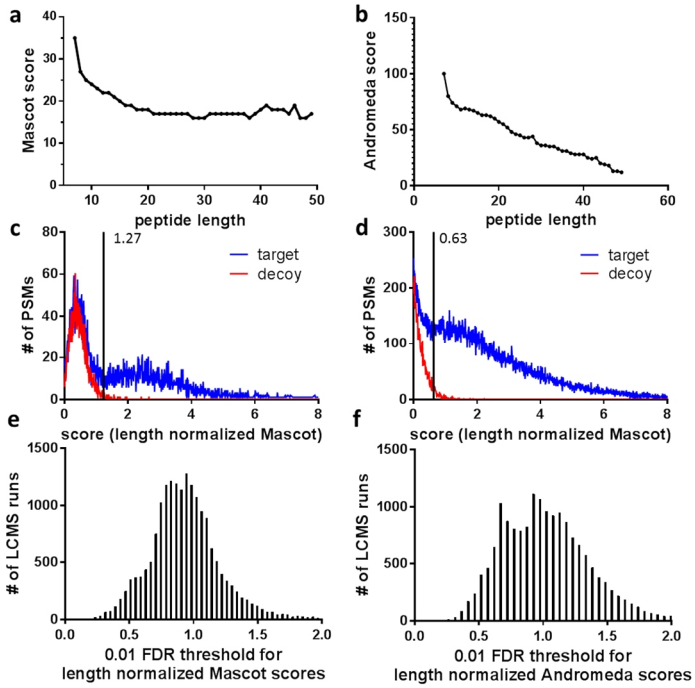


Extended Data Figure B9 | Protein complex composition and stoichiometry from shotgun proteomic data. a, Stoichiometry of the nuclear pore complex (NPC) reconstructed from shotgun proteomics data. To illustrate that normalized iBAQ values from shotgun experiments actually reflect protein copy numbers, we reconstructed the stoichiometry of the NPC (data from nuclear extracts of HeLa cells; Ori *et al.*, Mol. Syst. Biol., 2013; blue bars; error bars indicate standard deviation from triplicate experiments) and compared it to the stoichiometry determined in the same study using AQUA peptides and MRM experiments (red bars). Note that most of the time the stoichiometries are in very good agreement between the methods and the stoichiometries reported in the literature. b, Stoichiometry of the **α** and ***β*** subunits of the proteasome reconstructed from shotgun proteomics data (examples). ***β*** subunits of the constitutive proteasome are indicated in grey, immunoproteasome subunits (***β***i) are indicated in red. Note that PC-3 cells are devoid of the immunoproteasome while cells in the lymph node almost exclusively express this version of the molecular machine. c, Systematic assessment of the fraction of ***β***i subunits (red bars) and ***β*** subunits (grey bars) across 29 tissue samples and 80 cell line samples. Note that many cell lines and tissues contain both versions of the proteasome and the data also suggests that further forms of the proteasome with different subunit composition may exist.

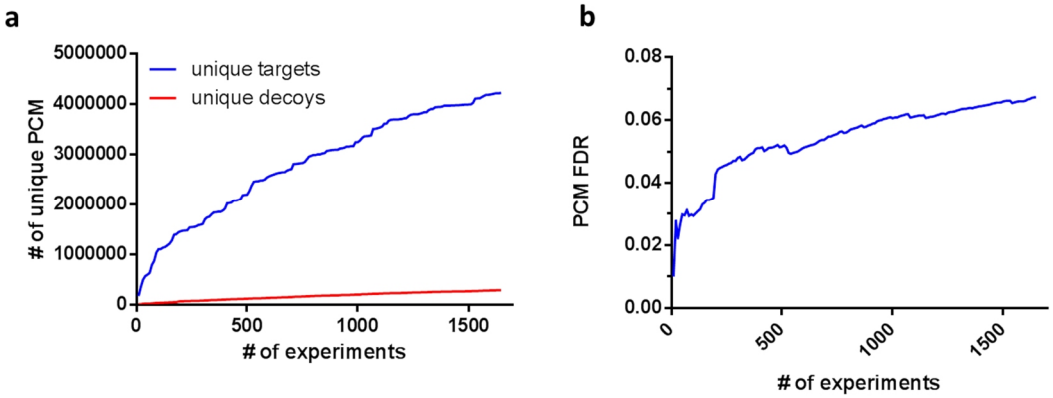


Extended Data Figure B10 | Examples for the analytical utility of large MS-based data collected in ProteomicsDB. a, Enumeration of post-translational modifications and protein termini. b, Computation of proteotypic peptides. Generally the same 1-5 peptides are identified every time a protein is identified (left panel) making proteotypic peptides useful for assessing protein identification and as reagents for targeted MS measurements. We note that the proteotypicity of a peptide strongly depends on the presence/absence of a chemical modification (right panel, here TMT or iTRAQ). c, Analysis of the selectivity of MRM transitions. The left panel shows the y8 transition of the peptide LHYGLPVVVK (b-catenin, marked with an arrow) in a 0.7/0.7 Da slice of the precursor and fragment ion window typically employed on triple quadrupole mass spectrometers. The size of the circle represents the relative intensity of the y8 fragment in a full tandem mass spectrum of this peptide. All other circles are interfering peptides (extracted from the entire ProteomicsDB) that have precursor and fragment ions in the same m/z window and with varying intensities (circle size). Interference can be reduced by using high resolution MS (middle panel) and confining the analysis to the tissue in question (here colon, right panel). Such interference plots in conjunction with the proteotypicity of peptides can be valuable for the design of targeted proteomic experiments.

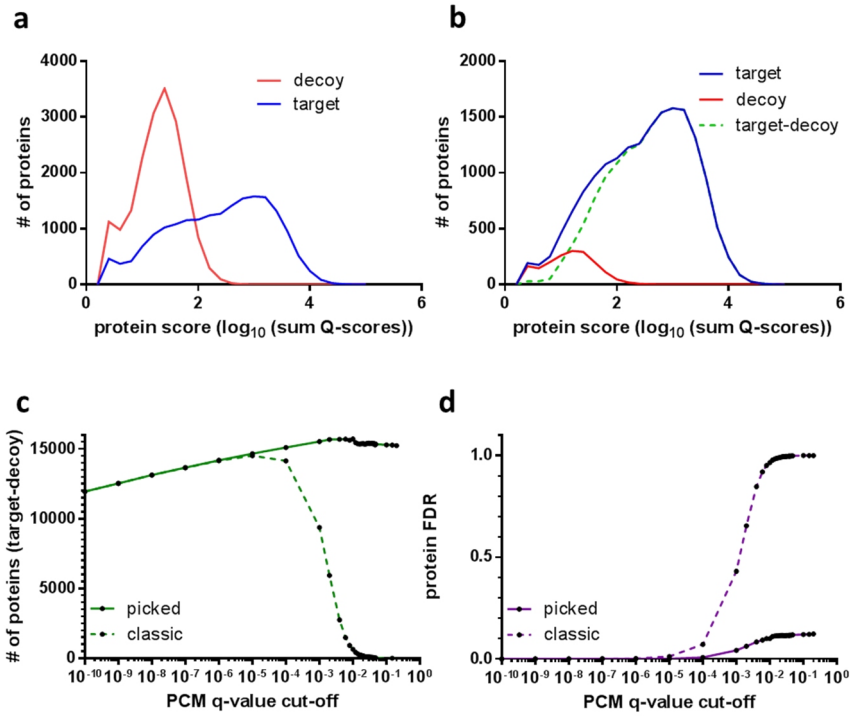
C - Picked protein FDR



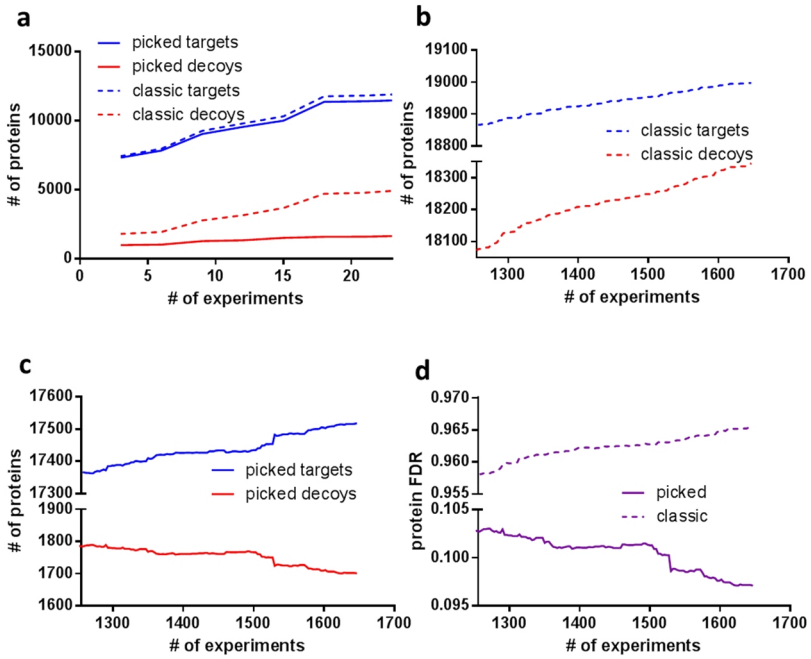
Extended Data Figure C1 | Search engine score normalization and variations in score cutoffs to reach 1% PCM FDR. a, b The local peptide length dependent score cutoffs at 5% PSM FDR between Mascot (a) and Andromeda (b) used for the score normalization are vastly different. While the cutoffs determined for Mascot decrease at the beginning and converge at ~17, the cutoffs used for Andromeda decrease constantly. c, d To illustrate vast differences in data quality dependent on technical and biological differences we plotted the score histograms of length normalized Mascot ion scores for (c) a dimethyl labeled tryptic digests of human embryonic stem cells measured by low resolution CID and (d) an unlabeled tryptic digests of the melanoma cell line A375 measured by HCD. To reach 1% PCM FDR, the labeled dataset had to be cut at 0.63 whereas the unlabeled dataset at 1.27. e,f Differences in data quality require different length normalized score cutoffs to reach 1% PCM FDR. While the range of length normalized score cutoffs is similar for Mascot (e) and Andromeda (f), the shape of the distribution varies.



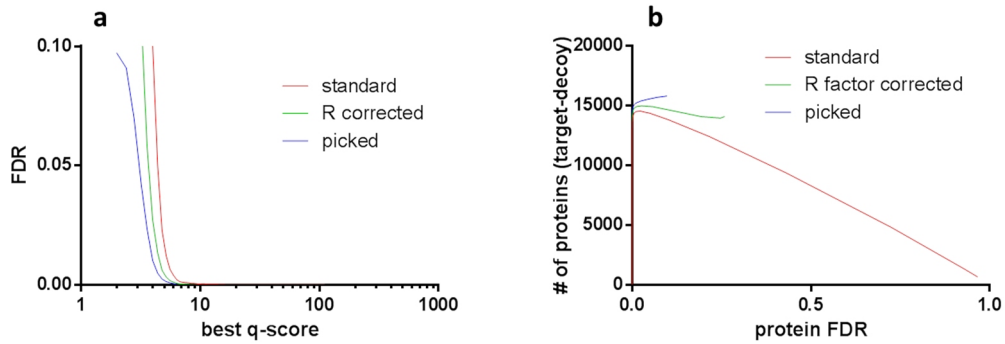
Extended Data Figure C2 | Target and decoy PCM saturation. a, In contrast to the saturation of proteins when accumulating multiple experiments, the number of unique target PCMs (blue) only shows a slight saturation effect. Furthermore, the numbers of unique decoy PCMs (red) increases linearly with increasing amount of data. b, This is mirrored by the global PCM FDR. The sharp increase at ~250 experiments in the PCM FDR is due to an experiment containing multiple LC-MS/MS raw files acquired while optimizing an acquisition method and thus contains highly redundant target PCMs but many random decoy PCMs.



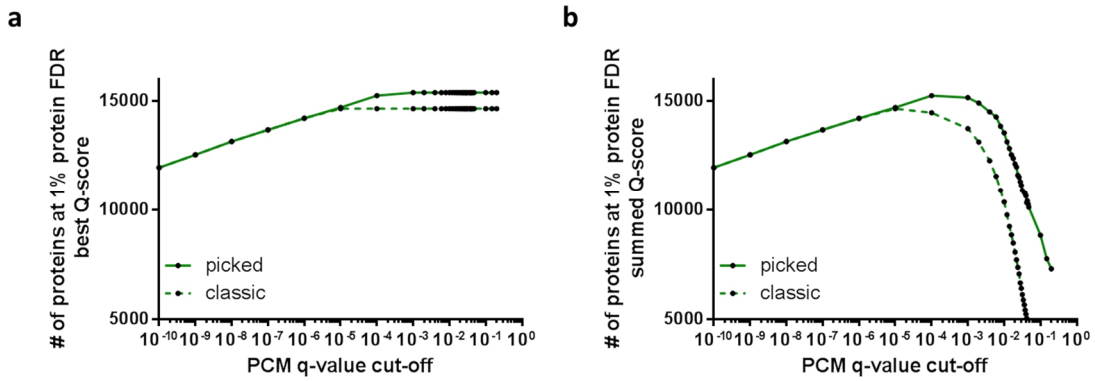
Extended Data Figure C3 | Protein FDR estimation using the classic and picked TDS using the sum of best Q-scores as protein score. a, Using the sum of best Q-scores of all PCMs matching to a protein as protein score, the number of decoy proteins (red) of the classic TDS massively overestimates the number of false positive protein identifications. Furthermore, the target distribution (blue) shows no bimodal shape and is not well separated from the decoy distribution. b, After applying the picked approach, the decoy (red) protein distribution superimposes with the target (blue) protein distribution which allows a more accurate protein FDR estimation. c, Comparing the performance of the picked (solid) and classic (dashed) approach when filtering the PCMs on various FDR shows a similar trend as in Figure 4.51. With increasing PCM q-value cutoffs, the number of true positive protein identifications (number of target proteins – number of decoy proteins) increases and is comparable between the picked and classic approach. At roughly 10^{-4} PCM q-value cutoff, the number of true positive proteins starts to decrease and quickly drops to 0 for the classic approach, whereas true positive proteins IDs increase further and converge at a rather stable plateau in the picked approach. The slight decrease at the end is likely due to accumulation of false positive PCMs which further deteriorates the separation of decoy and target proteins. d, The estimated protein FDR of the picked (solid) and classic (dashed) approach mirrors the trend seen in panel c. While the estimated protein FDR increases constantly when increasing the PCM q-value cutoff and eventually reaches 100% in the classic approach, the picked approach starts to rise much later and plateaus at roughly 10%.



Extended Data Figure C4 | Enlarged illustrations of the comparison of the classic and picked TDS from Figure 4.51. a, Even when aggregating small numbers of experiments, the picked (solid) TDS outperforms the classic (dashed) TDS. While the numbers of target proteins (blue) is comparable (marginally higher number of the classic approach) the difference between the number of decoy proteins (red) reported by the classic and picked approach is starting to increase. b, The overestimation of false positive proteins by the classic approach is particularly apparent when comparing the number of target (dashed blue) and decoy (dashed red) proteins at the end of the aggregation process. The number of decoy proteins is increasing more rapidly than the number of target proteins and is approaching the same limit. c, The picked approach shows a complete opposite effect. The number of decoy proteins reported by the picked approach (solid red) is decreasing because of new evidence (especially at ~1540 experiments) introduced by additional experiments. d, The trend explained in panel b and c is mirrored by the estimated protein FDR in the picked (solid) and classic (dashed) TDS. While the protein FDR increases and approaches 100% in the classic approach, the picked approach sows a decrease, potentially reaching close to 0% when adding more data.



Extended Data Figure C5 | R factor FDR. a, R factor correction produces more reasonable protein FDR curves than the standard decoy strategy, the agreement between the picked and R factor approach is not perfect, but better than between either of the approaches and the standard approach. b, Number of true protein hits as a function of FDR for the standard, picked and R factor approach. Both the R factor and picked approaches perform better than the standard strategy, with the picked TDS consistently yielding higher coverage.



Extended Data Figure C6 | Comparison of best and sum Q-score protein scoring of the classic and picked TDS. a, When using the best Q-score to score proteins, the number of proteins identified at 1% proteins FDR is increasing in both picked (solid) and classic (dashed) approach, but the picked approach consistently reports higher numbers of proteins. b, Using the sum of best Q-scores of all PCMs matching to a protein, the number of proteins identified at 1% protein FDR is first increasing in both picked (solid) and classic (dashed) approach, but is starting to decrease and breaks down at high PCM q-value cutoffs. The picked approach shows a delayed behavior but also overestimates the number of false positive proteins IDs using the decoy proteins. Especially at high PCM q-value cutoffs, the decoy and target protein distribution start to blend into each other (data not shown) and shows almost no separation any more.