Annual Review of Biomedical Data Science

Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data

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

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

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

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

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

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

*These authors contributed equally to this article

Keywords

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

Abstract

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

INTRODUCTION

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

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

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

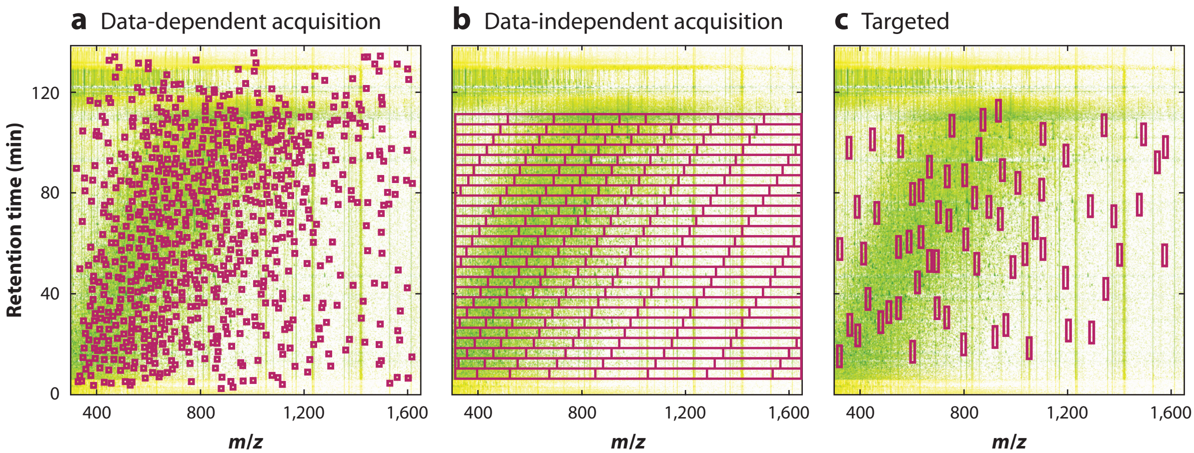


Figure 1

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

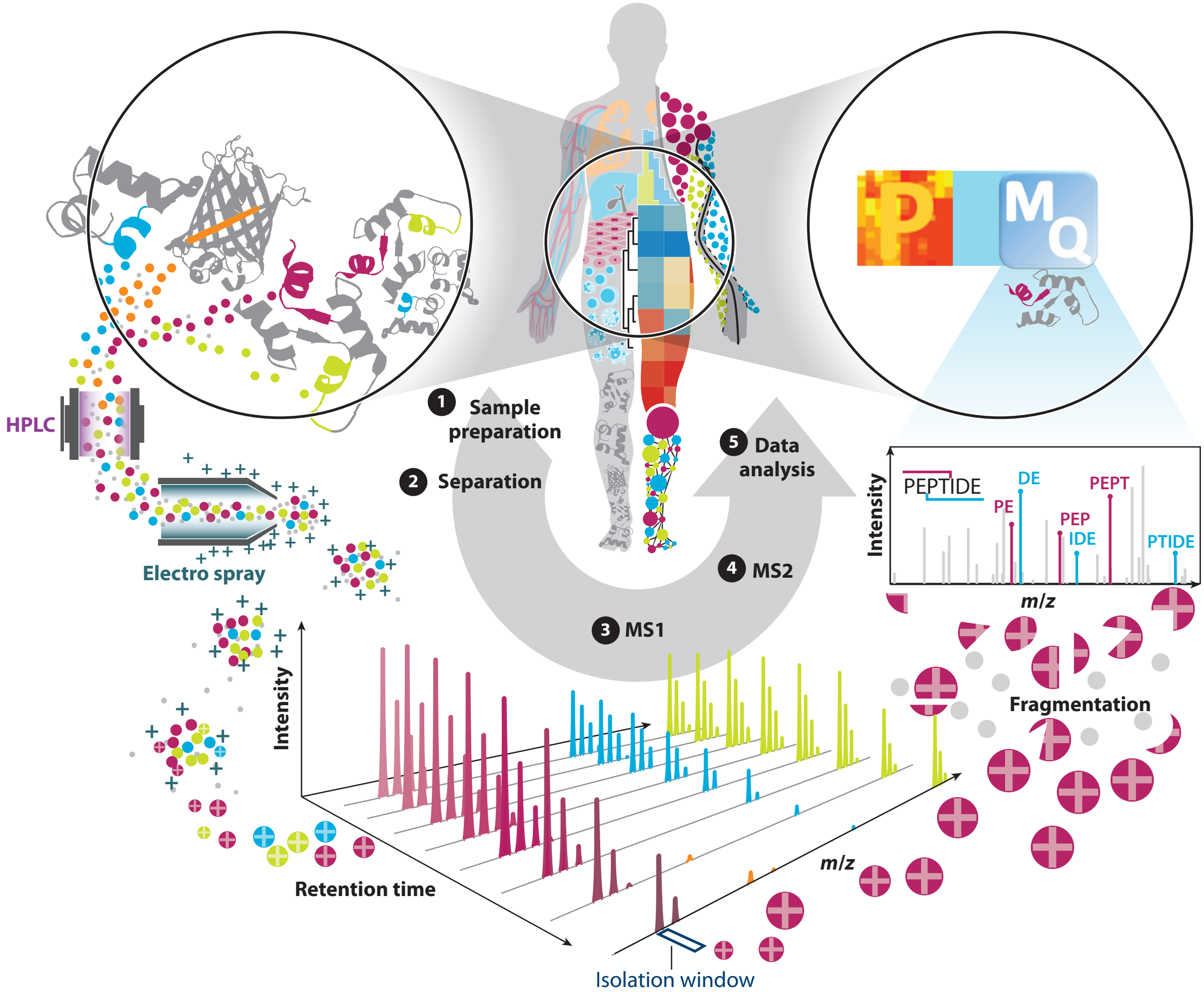


Figure 2

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

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

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

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

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

IDENTIFICATION AND QUANTIFICATION OF PEPTIDES, PROTEINS, AND POSTTRANSLATIONAL MODIFICATIONS

Liquid Chromatography-Mass Spectrometry Features

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

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

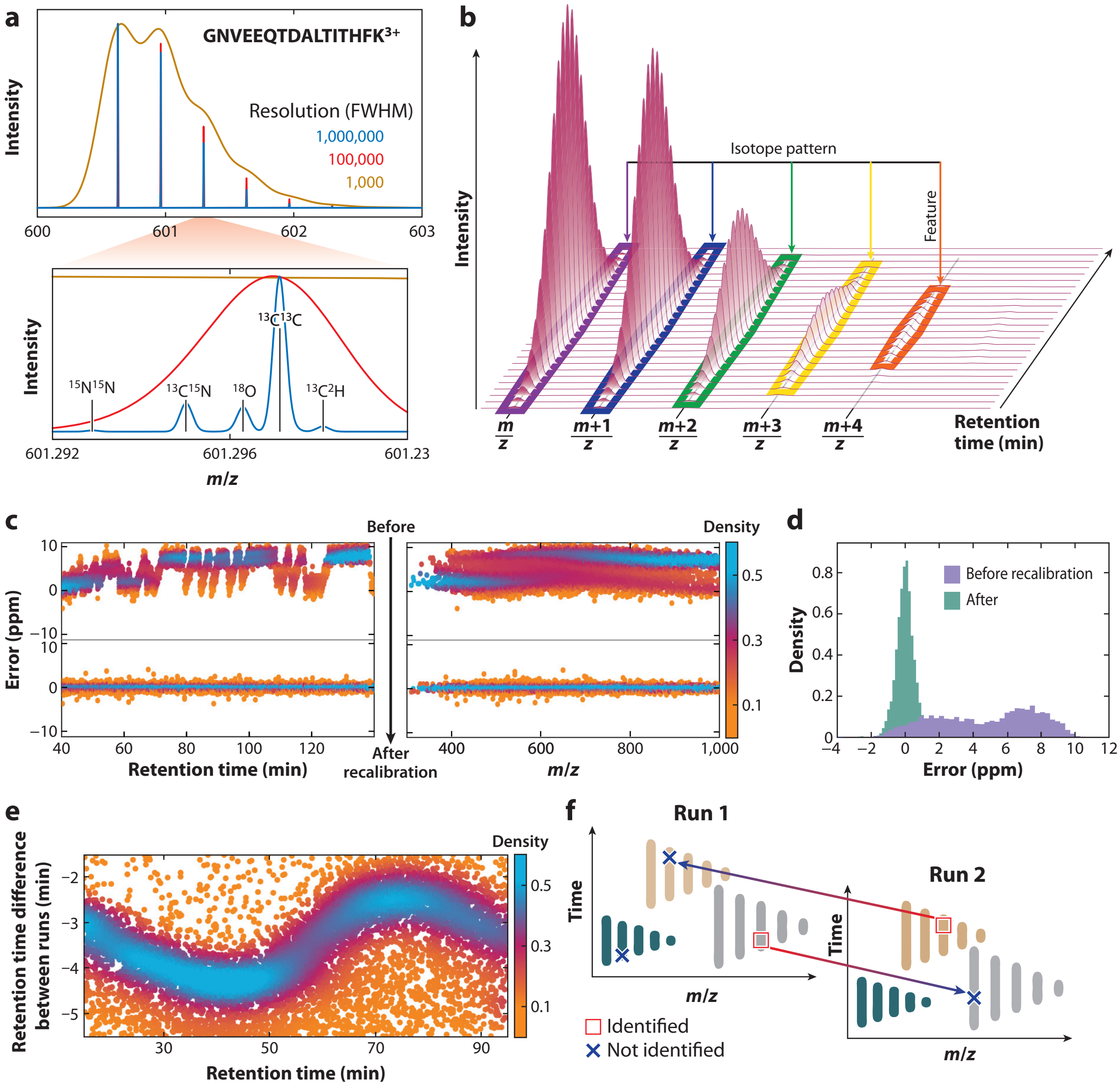


Figure 3

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

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

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

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

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

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

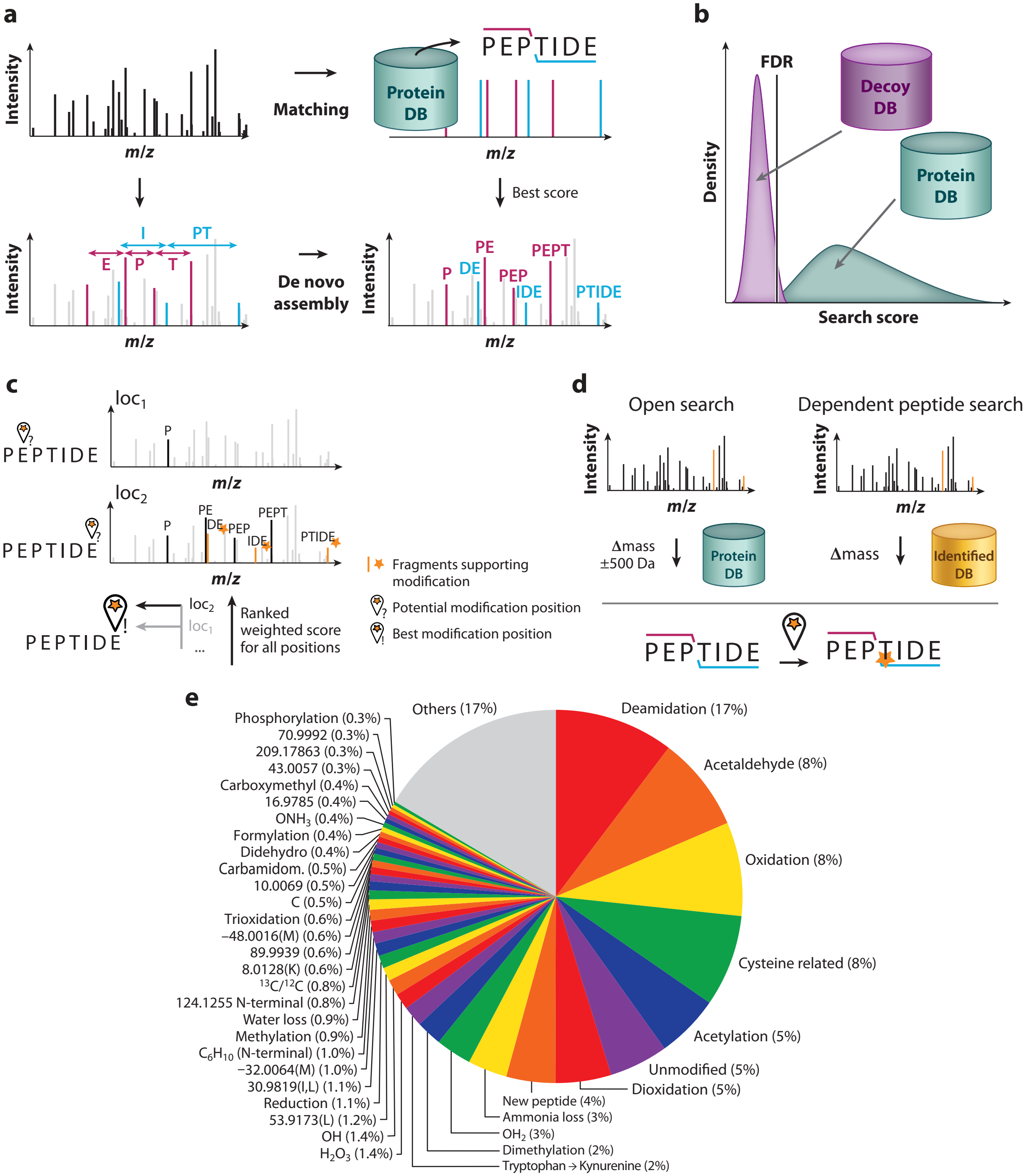
(**Figure 3*f***). Following this strategy, the quantification profiles across many samples become more complete, which partially removes the stochastic behavior of the data-dependent acquisition in bottom-up proteomics. Determining and controlling false discovery rates (FDRs) for these kind of matching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

Peptide Identification

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

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

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



(Caption appears on following page)

Figure 4 (*Figure appears on preceding page*)

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

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

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

There are a number of special topics in peptide identification, starting with dipeptides resulting from cross-linked proteins (90, 91), which have the challenge of a vastly increased search space due to pairing of peptides, for which several popular software packages are available (92–97). In proteogenomics searches (98), peptides are identified based on customized protein sequence databases generated from genomic or transcriptomic information. Search spaces for proteogenomics searches are typically larger than in conventional searches since they often involve three- or six-frame translations of genomic sequences. Furthermore, these search spaces are heterogeneous, since the sequence content ranges from clearly existing, manually validated protein sequences to in silico–translated genomic regions without any prior evidence for their expression. Hence, extra measures need to be taken in the identification process to account for this heterogeneity. Proteomics of species without sequenced genome requires tools to integrate incomplete sequencing data with homologous sequence data from closely related species (99).

Protein Inference and False Discovery Rate

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

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

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

The phenomenon of error expansion from peptide to protein identification in large data sets is well known in the field (106, 108). Even if the FDR is thoroughly controlled at the PSM level, if no additional measures are taken, the FDR on protein level can become arbitrarily large. Hence, it is highly important to use workflows that control FDR on the protein level (25, 106, 108, 109) to limit the number of proteins falsely claimed to be present in the sample, particularly if the number of identified proteins is a relevant outcome of the study.

Quantification

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

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

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

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

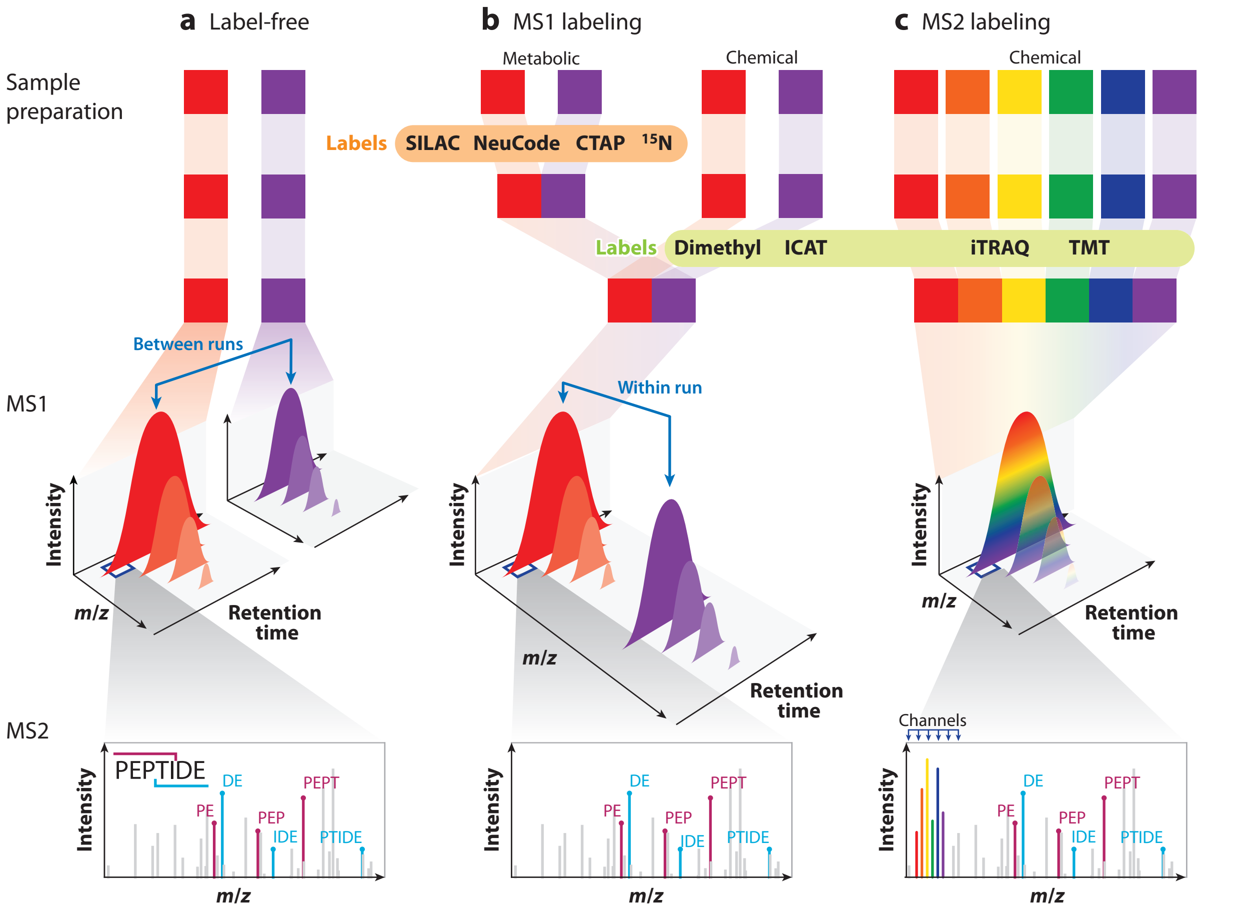


Figure 5

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

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

In isobaric labeling (114–116), peptides in different samples are labeled with different molecules per sample that have the same mass but that eject different reporter ions upon fragmentation. The biggest advantage of isobaric labeling is its multiplexing capacity. Up to 11 samples can be measured simultaneously with the currently available tandem mass tag reagents. The downside is that the presence of coeluting peptides in the isolation window for fragmentation leads to ratio compression (117). To be precise, cofragmentation makes ratios wrong in arbitrary and individual ways. However, since it is often a valid assumption that most of the proteins are not changing between samples, the cofragmented peptides are likely to have 1:1 ratios, thus compressing the ratios of changing proteins. There are several experimental strategies to reduce or remove the cofragmentation problem, such as gas-phase purification (118), MultiNotch MS3 (119), and use of complementary ions (120). There are several computational methods that reduce ratio compression. Reporter ions of low intensity are prone to carry more noise and be more affected by cofragmentation signals. Hence, peptides with higher reporter ion intensities should be given higher weights when calculating protein intensities. Another approach is to calculate the fraction of precursor signal divided by the total MS1 signal observed in the isolation window (121, 122), which can be used for filtering peptides used for quantification. To some extent, this quantity can also be used to correct for ratio compression (123).

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

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

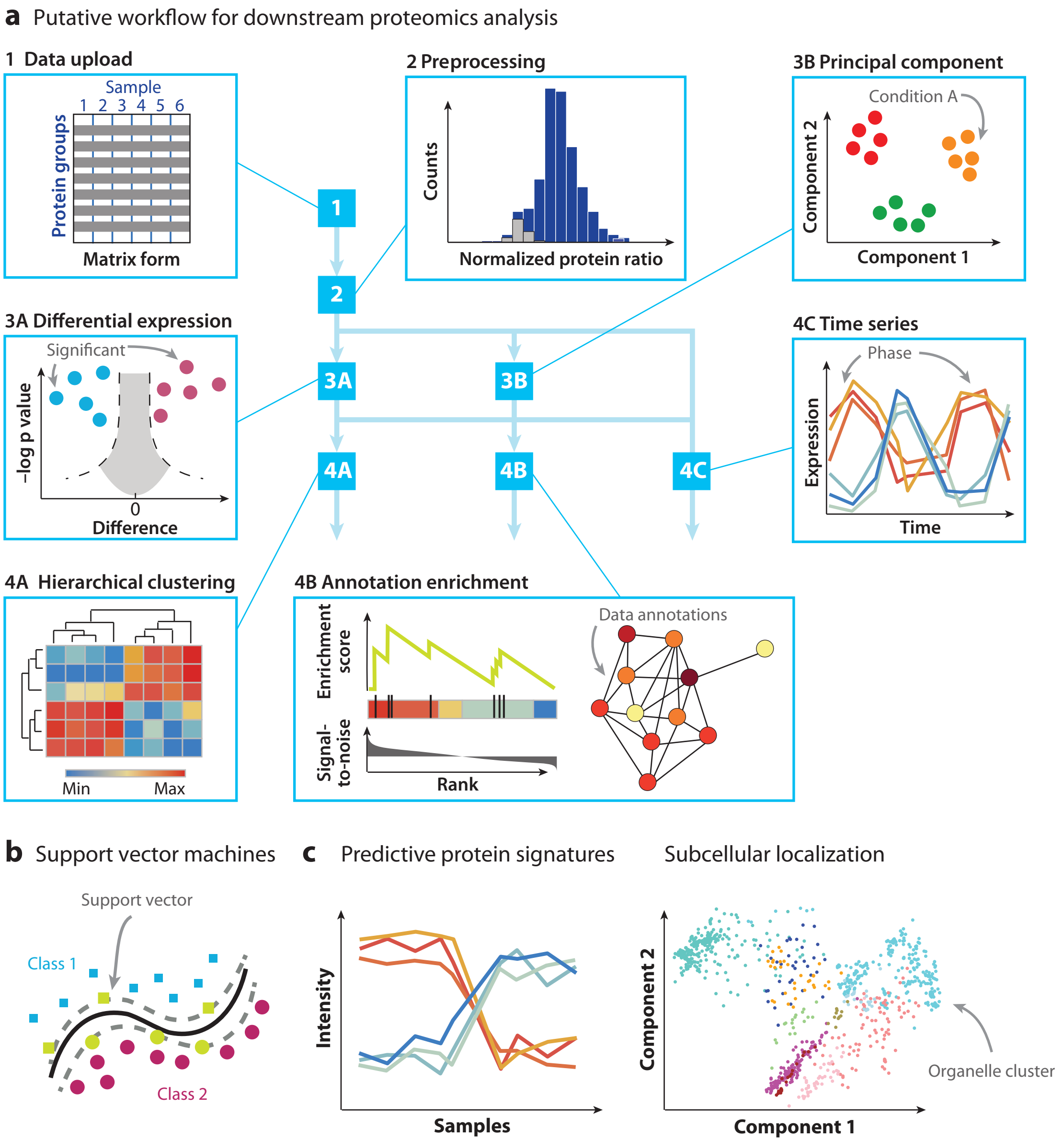
DOWNSTREAM DATA ANALYSIS

Exploratory Statistics

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

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

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



(Caption appears on following page)

Figure 6 (*Figure appears on preceding page*)

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

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

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

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

Posttranslational Modifications

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

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

Machine Learning

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

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

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

Network Biology

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

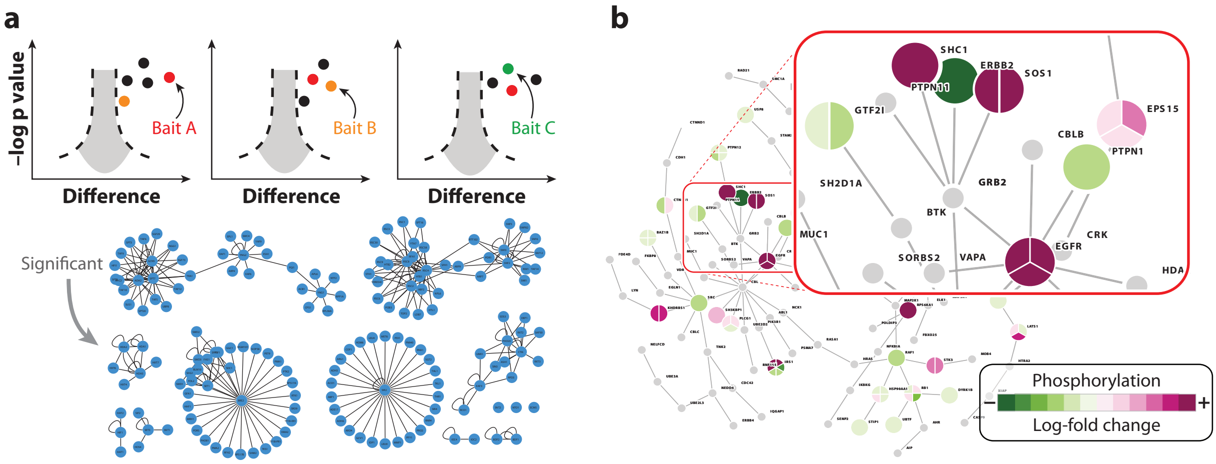


Figure 7

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

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

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

For general-purpose network analysis, Cytoscape (161) has emerged as the de facto standard. Through its plugin infrastructure, it provides a wealth of analyses and visualizations, often integrating expression-omics technologies with interaction networks. Cytoscape reads networks from various standard formats and can extend them with interactions and pathways from various databases. Tools such as BiNGO (162) can identify significantly enriched gene ontology terms in these networks. Large-scale networks can be clustered into modules, either by topology (MCODE; see Reference 163) or by differential expression (jActiveModules; see Reference 164). Alternatively, network reconstruction tools, such as ANAT (165), identify a subset of interactions connecting, for example, differentially expressed proteins to their signaling stimulus.

Multiomics Data Analysis

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

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

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

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

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

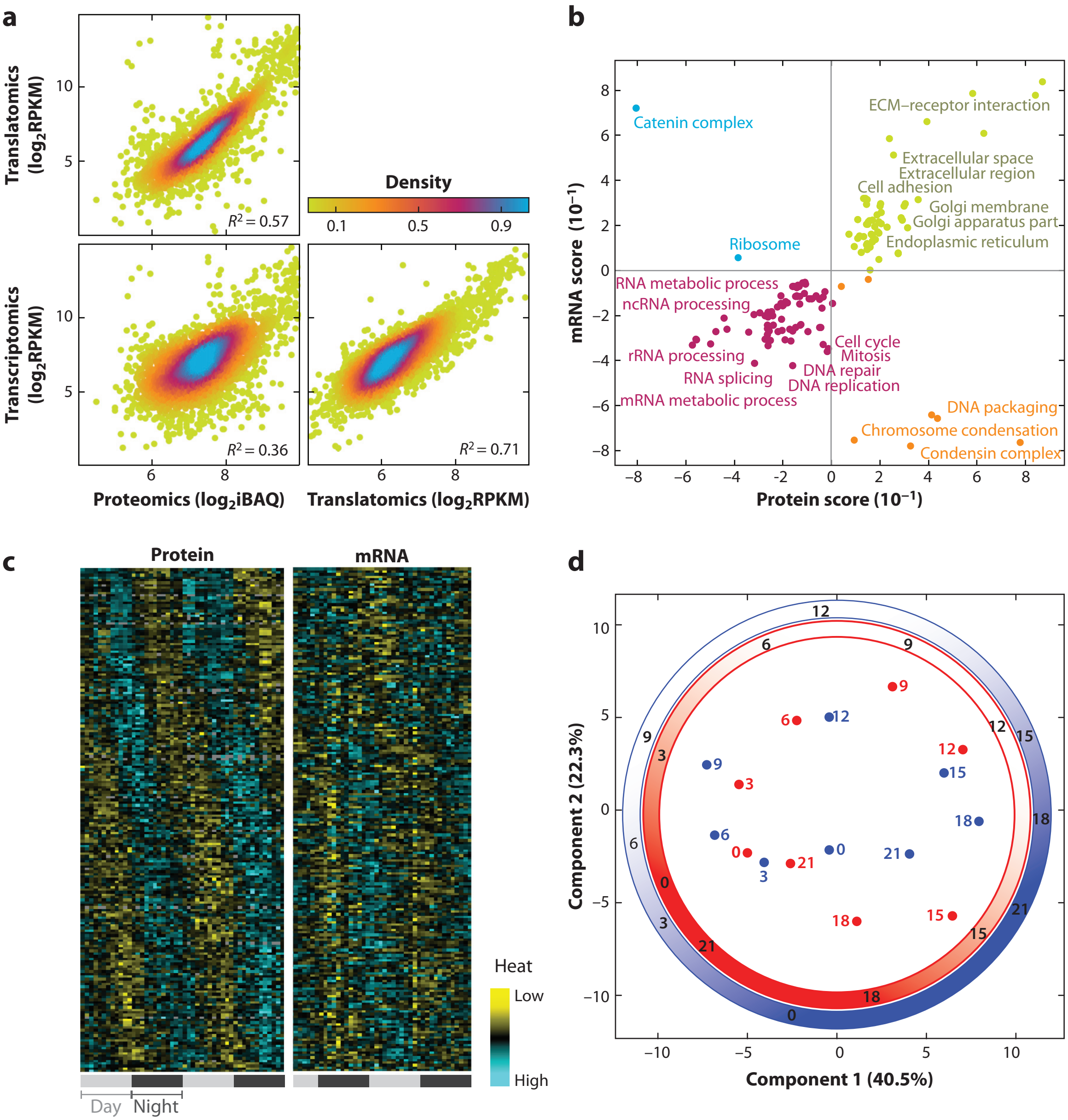


Figure 8

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

DISCUSSION AND OUTLOOK

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

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

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

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

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

LITERATURE CITED

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

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

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

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

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

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

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

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

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