Journal of Proteomics 134 (2016) 25–36

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Genomic variability and protein species — Improving sequence coverage for proteogenomics

Rainer Bischoff [a,*], Hjalmar Permentier [a,b], Victor Guryev [c], Peter Horvatovich [a]

[a] Department of Analytical Biochemistry, Research Institute of Pharmacy, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands
[b] Interfaculty Mass Spectrometry Centre, Research Institute of Pharmacy, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands
[c] European Research Institute for the Biology of Ageing, University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 4 June 2015
Received in revised form 6 September 2015
Accepted 14 September 2015
Available online 21 September 2015

Keywords:
Genomics
Proteomics
Mass spectrometry
Proteogenomics
DNA/RNA sequencing

ABSTRACT

Protein heterogeneity may result from many factors often closely related to the regulation of biological mechanisms. This review addresses one source of protein heterogeneity, the translation of genetic variability and transcriptional modulation to the protein level. We provide an overview how customized protein sequence databases generated using genomic and transcriptomic sequence information in conjunction with approaches to increase protein sequence coverage can aid in gaining a deeper insight into variability at the protein level. Modern approaches of DNA/RNA sequencing open the possibility to obtain detailed sequence information from individual genomes and transcriptomes at single nucleotide resolution. Further studies tried to correlate genetic variability with important biological consequences such as the risk for developing a disease or defining a personalized approach towards therapy (also called "personalized or precision medicine"). Linking genomic and transcriptomic information to complex biological mechanisms has, however, remained elusive due to the fact that there is no direct cause and effect relationship between changes at the DNA/RNA level and downstream consequences. In this review we give an overview of the challenges of integrating genomics and transcriptomics data with proteomics data and link variability at the DNA/RNA level to protein variability and protein species.

Biological significance: The manuscript focuses on a recent trend in proteomics, namely the integration of genomic and proteomic data. Genetic and transcriptomic variability accounts for a considerable part of protein variability and is at the basis of many protein species, many of which not yet described at the protein level but many also identified as proteins or peptides with unknown function. The review highlights the challenges of current proteomics methodology, notably incomplete sequence coverage, which make it difficult to appreciate the full complexity of any proteome and leads to the fact that much variability at the DNA/RNA level is not captured at the protein level. We outline a few strategies to ameliorate this situation.

1. Introduction

Proteomics attempts to study the complete protein complement that is present in a cell, tissue, body fluid or another biological sample. Recent advances in liquid phase separations, mass spectrometry and bioinformatics enabled researchers to perform deep proteome analyses covering almost every protein as predicted by the corresponding genomes and transcriptomes [1,2]. However, proteomes contain more sequence variation than is captured in the canonical sequences of public sequence databases, which are widely used for protein identification. Canonical sequences provide the longest sequence reads that show the strongest homology to proteins in other species while providing the most complete description of protein domains and covering the most prevalent protein forms in a given species [3]. Genetic variability between individuals in a population, transcript variants stemming from alternative splicing and post-transcriptional RNA modification events, all introduce new protein species. One gene may thus give rise to multiple proteins each representing a different protein form. According to this definition, each modification of a protein leads to a distinct protein species. While this review focuses on human biology, the same general principles apply to other organisms. We will not cover the vast area of post-translational protein modifications due to limitations in scope and the many excellent overviews that have appeared on the topic (see [4,5] as examples).

Proteogenomics has originally been used in studies where the goal was to confirm gene annotation. An example of such a study is presented by Bringans et al., where the predicted genes of the fungal wheat pathogen *Stagonospora nodorum* were corroborated using 2D-LC MALDI-TOF/TOF proteomics data [6]. In this study, two versions of genome annotation predicted a total of 16,116 genes from which 2134 were confirmed

* Corresponding author.
E-mail address: r.p.h.bischoff@rug.nl (R. Bischoff).

http://dx.doi.org/10.1016/j.jprot.2015.09.021
1874-3919/

by proteomics data. More recently the term proteogenomics was also applied to studies where DNA and RNA-Seq data are used to generate sample-specific protein sequence databases for peptide and protein identification. RNA-Seq data predicts protein sequences contributed by events like RNA editing or splicing, which reflect the protein product more closely than protein sequences deduced from DNA. For this reason, we will mainly focus on studies using the RNA-Seq approach in this review. There is a clear trend towards generating individual transcriptome references to guide proteomics data analysis and protein identification through database search (DBS). The RNA-Seq-based proteogenomics approach accounts for sample-specific, genomic variability and hence offers the best option of matching in silico generated peptide mass spectra with experimental data. Since DBS can only identify peptides that are present in the database, it is necessary to generate RNA-Seq-based databases to match experimental MS/MS spectra to sample-specific protein species. This advance is due to significant improvements in nucleic acid sequencing technology allowing for fast and cost-efficient transcriptome reconstruction by de novo assembly or by alignment sequence reads to the reference genome [7,8]. It is to be expected that most future proteomics work will no longer make use of generic, species-wide but rather of individualized sequence databases. We can envision that this represents the beginning of the 'personal -omics era' that will be further extended to encompass personal glycomics, metabolomics as well as other data modalities. An example of this development is the Chromosome-Centric Human Proteome Project (C-HPP), which has the goal to catalog all identified human protein species in a 'gene-centric' manner and enrich it with metadata such as MS/MS spectra of unique peptides. This and other resources will become very relevant to address biological questions by relating genetic variability to protein species and function [9,10].

2. Genomics and protein variability

2.1. Personalized genomics and proteomics data interpretation

Advances in DNA/RNA sequencing technologies enable a more targeted strategy for the definition of a reference proteome. Personal and population genomics provides a viable approach to constructing sample- or group-specific reference databases. The value of whole genome-sequencing can be exemplified by the recent sequencing of the genomes of 250 Dutch families [11], that delivered many new transcript isoforms not represented in public databases but inherent to genetic variability in this population. Thus, while an average Dutch genome has over two thousand common non-synonymous variants (the allele frequency in this population exceeds 5%), it also contains about 300 personal or rare variants affecting the content of individual or group-specific proteomes. Using a public reference database might thus not only result in lower identification rates but also bias proteome analysis by missing rare protein species that are present but cannot be matched to experimental MS data. These variants may be particularly interesting when trying to link a phenotype (e.g. susceptibility for a disease) to the proteome profile.

Another, underappreciated source of genetic variation is structural changes in genomes [12]. Small insertions/deletions (indels) and larger differences such as deletions, duplications and (retro-) transpositions, collectively involve more bases in genomes than single-base variants. It is therefore expected that they might have a more profound influence on the gene and transcript content of an individual. While methodologies for large-scale genome structure analysis are still under active development, we expect that progress in this area will further improve the identification of personal and rare protein species.

2.2. Transcriptome sequencing approaches

RNA sequencing represents a whole class of transcriptome reading techniques under a general au term of RNA-Seq approach. Multiple modifications of RNA isolation, fragmentation and library preparation protocols as well as sequencing methods have been developed, addressing various combinations of types of RNA molecules. These protocols offer different utilities for a proteogenomics approach.

- Classic transcriptome profiling aims at determining the full complement of cellular RNA. As ribosomal RNA (rRNA) constitutes a large fraction of cellular RNA, but does not encode for protein sequences, an rRNA depletion step retaining all coding and non-coding RNA species improves the cost-efficiency of sequencing.
- Construction of polyA-selected transcriptome libraries by using an oligo-dT primer used to selectively enrich for polyadenylated mRNA species. The latter represent RNA molecules transcribed by RNA polymerase II and, to our current understanding, represent virtually all protein coding RNA messengers.
- Ribosome fractionation followed by RNA sequencing addressing transcripts associated with ribosomes corresponding to the actively translated part of the transcriptome [13]. This method not only highlights which RNA molecules currently undergoing translation, but can also be used for inferring the translational efficiency for each molecule, depending on the frequency with which that RNA is found in monosomal and polysomal ribo-fractions.
- Ribosome profiling (Ribo-Seq) or sequencing of ribosome-protected RNA fragments allowing for detection of ribosome density at nucleotide level resolution and revealing translation efficiency [14]. Despite a short, ~30 base sequence tag protected by a ribosome, this method is very informative for characterizing translation dynamics.

While it is commonly recognized that polyA-selected RNA-Seq libraries provide the most cost-efficient and comprehensive characterization of the protein-coding component of a transcriptome, other methods extend and enrich our understanding of transcriptome dynamics. Thus the Ribo-Seq method has the potential to offer a method to detect small (30 and fewer amino acids) open reading frames (ORF) such as upstream ORFs — an ORF within the 5′ region of the mRNA, that is commonly considered as untranslated, unless there is an evidence that these regions are bound by ribosomes. The uORFs have been neglected because of their size, but may be potent regulatory molecules [15].

We must also note that recent studies reported ribosome binding to RNA that was previously considered non-coding [16]. It is thus possible that a larger variety of transcripts participate in translation. However, functional significance of such translational events remains unclear and it has to be seen how many of them contribute to observable phenotypic changes.

The aforementioned experimental approaches can be used to define the current transcription state (see Fig. 1 for an overview of genome- and transcriptome-associated processes that generate transcriptome variation), namely:

1) the transcription level of each gene
2) the variety and abundance of transcript isoforms
3) novel open reading frames (ORFs) (e.g. short peptides missed in gene annotations)
4) non-synonymous RNA-editing events
5) rearranged transcripts (e.g. gene fusion events due to carcinogenesis or inherited genomic defects)
6) transcripts coming from disease agents or contaminations (e.g. RNA from bacteria or parasites).

For organisms with complex transcriptomes, featuring many transcript isoforms per gene, inferring and quantifying transcript variants from short sequence reads is challenging. In this case, third-generation single-molecule sequencing (e.g. using Pacific Biosciences RS instrument or other emerging technologies featuring sequence reads that can span thousands of bases) or a combination of the former with short-read technologies might hold the key for characterizing the complete set of full-length transcripts [17].

In respect to experimental and computational challenges that transcriptome reconstruction is facing, the proteogenomics approach has a

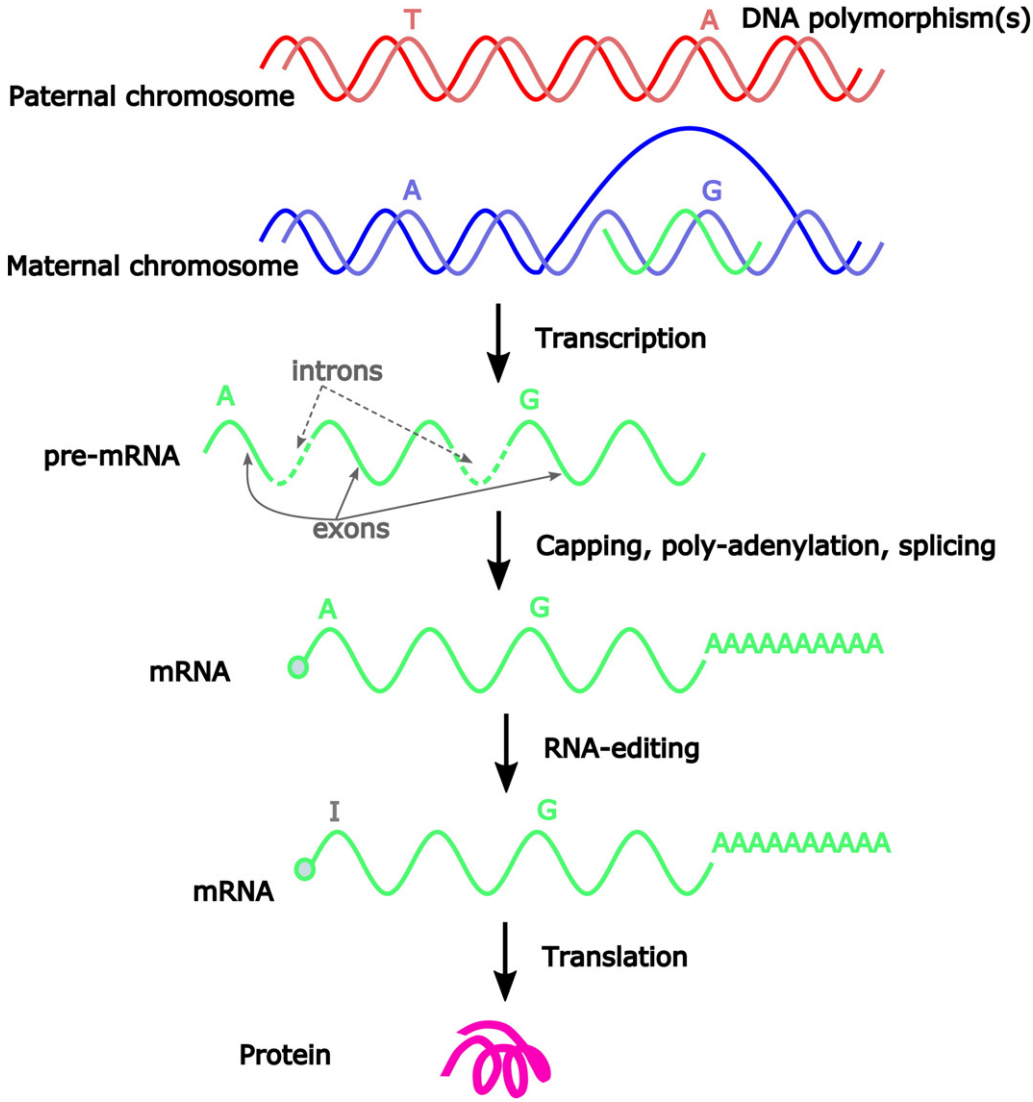


Fig. 1. Genetic and transcriptome-associated sources of proteome variation. Genome variants, differences in gene transcription, splicing, and editing all result in protein species that may vary from sample to sample. Note that RNA-editing may introduce non-canonical RNA bases like inosine(I) in the most abundant A-to-I editing type. Inosine is recognized as a different base (G) during the translation process.

potential to help in assessing the performance of different platforms and experimental approaches as well as the best combination of bioinformatics tools for read alignment and de novo assembly.

Transcriptome profiling provides a rich source of data to guide proteomics analyses. However, there are potential caveats of these approaches that may result in data misinterpretation. A few examples are given below:

1) proteins that are synthesized elsewhere and transported to the studied tissue (examples: hemoglobin, albumin) may be not included in the protein sequence database made exclusively from RNA-Seq transcripts and therefore missed.
2) proteins whose half-life exceeds that of the corresponding messenger RNA significantly may be underrepresented in a protein sequence database made exclusively from RNA-Seq transcripts.
3) the inadvertent selection of non-coding transcripts with internal polyA stretches may lead to an increased chance for false positive protein matches.
4) the presence of sequencing reads that originate from immature transcripts and those corresponding to introns (the number of intronic reads is often comparable to the number of reads mapped to annotated exons) may further inflate the protein sequence database.

These potential drawbacks make it challenging to marry genomics and proteomics data in a joint proteogenomics approach. However, the prospect of identifying a nearly-complete and highly representative sample-specific protein species with a link to disease development or important biological mechanisms is worth taking this challenge.

3. Proteomics — approaches and limitations

Current proteomics approaches result in partial sequence coverage with large gaps. The presence of a protein may be confidently inferred from the identification of only a few peptides, which together may cover only 5% of the protein sequence. This rather incomplete view of the proteome is in stark contrast to the complete sequence coverage obtainable by DNA/RNA sequencing. Genetic variability may occur at specific locations in the protein sequence leading to highly homologous protein isoforms such as splice junction or single amino acid variants. It is thus critical to look at the limitations of current proteomics approaches and to define ways how to improve them with respect to

sequence coverage. We first consider what the current state of the art is in high-throughput, bottom-up proteomics, and later discuss methods for improving sequence coverage.

3.1. The mainstream shotgun bottom-up approach

Fragmentation of peptides in a tandem mass spectrometer occurs by a mass selection and an activation step. The activation step imparts energy and leads to peptide fragmentation preferentially at the peptide bond. The two main activation methods are Collision-Induced Dissociation (CID) and Electron Transfer Dissociation (ETD). The first is available in most tandem MS instruments while ETD requires specific instrumentation, although it is gaining rapid adoption in the proteomics field. Other fragmentation methods exist (e.g. photon excitation, see review by Oh [18]), but CID and ETD are most widely used as they are more readily accessible in commercial instrumentation and provide highly complementary information (Fig. 2). A combination of CID and ETD has been recently described (EThcD) where peptides are first subjected to ETD and the fragments subsequently to CID. This approach alleviates the need to perform separate runs or run alternate scans using CID and ETD. EThcD has been shown to significantly increase coverage of the highly variable HLA class I-presented peptides [19].

For large scale proteomics, single-stage fragmentation (MS/MS or MS^2) is the norm, but multi-stage fragmentation (MS^n) is possible with some MS instrumentation (notably quadrupole ion traps). MS^n can increase structural information considerably and increase sequence coverage, but at the expense of longer duty cycles, which renders these approaches incompatible with high-resolution LC separations of complex samples having peak widths of a few seconds. MS^n is, however, a viable option for low–complexity samples, for example, in-gel digested proteins, where fewer peptides co-elute in a given time window.

The current mainstream proteomics approach to identifying proteins in complex biological samples is based on protein digestion followed by LC–MS/MS (bottom-up proteomics). This approach relies on the use of proteases having clearly defined cleavage sites to digest all proteins in the mixture (typically trypsin) followed by analysis of the resulting highly complex peptide mixture by data dependent (DDA) or independent (DIA) LC–MS/MS. In DDA the mass spectrometer works in a sequential mode acquiring first mass spectra without fragmentation (the so-called survey scan) followed by fragmentation of the most abundant peptides in the order of decreasing intensity. Already selected peptide ions are subsequently excluded from further fragmentation for a certain time, typically corresponding to twice the peak width at half height. Depending on the duty cycle, sample complexity and sensitivity of the mass spectrometer, up to 20 peptide ions are selected for fragmentation during each MS cycle. Generally peptides with a charge of +2, +3 or +4 and with ion intensities above a user-defined threshold are automatically chosen for fragmentation. While in DDA the mass of the intact peptide prior to fragmentation is generally known within a narrow mass window, sampling across the chromatographic peak is stochastic. In DIA, such as SWATH [20] and MS^E [21], a survey scan is followed by fragmentation of peptide ions across a range of m/z values resulting in MS/MS spectra from multiple precursors simultaneously. A critical point in DIA is to match the fragment ion spectra to the corresponding precursor ions. This can be done by superimposing the chromatographic elution profiles of peptide precursor and fragment ions (so-called spectral deconvolution) or by matching fragment ion spectra to spectral libraries [22–24]. Multiple peptides may co-elute within the selection window of DDA resulting in mixed or "chimeric" fragment ion spectra notably when analyzing highly complex peptide mixtures that exceed the separation capacity of the chromatographic column. Failing to assign fragment ions unambiguously to a given peptide precursor ion renders identifications ambiguous resulting in potential false positives or false negatives. Since proteogenomics approaches use the same data acquisition strategies, the problem of chimeric fragment ion spectra is not alleviated even though they use sample-specific databases build from RNA-Seq or exon DNA data.

3.2. Fragmentation mechanisms and peptide spectrum matching

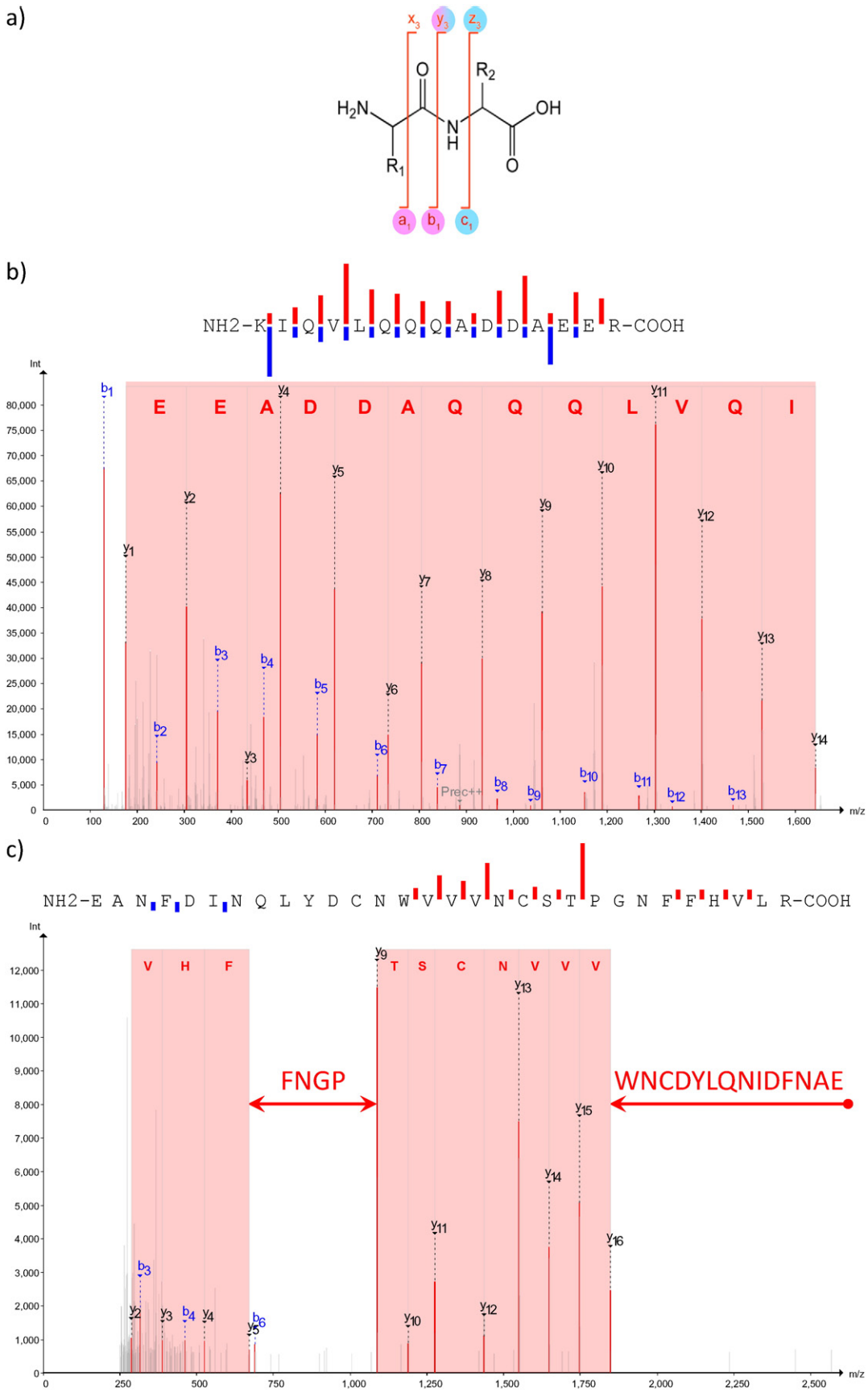
The sequence information content of a tandem mass spectrum depends primarily on the physicochemical properties of the fragmented peptide. The extent to which the activation method can be tuned by adjusting parameters is rather limited. In CID the amount of activation energy is primarily determined by the adjustable collision energy. However, a higher energy does usually not lead to more primary fragments (i.e. y- and b-ions generated by cleavage of a single peptide bond), but rather results in the secondary break-up of primary fragments. Therefore, collision energies in CID are optimized to obtain the highest yield of primary fragments over precursor ions while avoiding the generation of secondary and higher order fragments. The optimal energy is strongly correlated with two characteristics of the peptides, mass and charge state.

The most important factor that determines whether complete sequence information can be derived from an MS/MS spectrum is the peptide sequence itself. The CID fragmentation mechanism of peptides has been extensively studied (reviewed by Paizs and Suhai [25]). The complex details reveal that it should not be viewed as a simple breaking of the peptide bond and explain why certain members of an ion series (e.g. b_1 ions) cannot be detected since there is no conceivable mechanism for their formation.

Peptide fragmentation by ETD proceeds through a different mechanism than CID, as it makes use of ion-ion reactions to impart energy on the peptide leading to fragmentation. The reagent ion is a negatively charged aromatic radical anion derived, for example, from fluoranthene or azobenzene [26,27] and capable of single-electron transfer to positively charged peptide ions. In ETD, the bonds adjacent to the backbone amide break most easily, but unlike in CID, this takes place at the C-terminal side of the amide nitrogen, leading to c- and z-ion formation. ETD works best with multiply charged cations [28] and is therefore an attractive method for fragmentation of larger peptides and even intact proteins, which have many positive charges and are incidentally also difficult to fragment by CID. Fragmentation efficiency may vary significantly between peptides resulting in few or no fragment ions as one extreme and the generation of many, non-informative small fragment ions as the other extreme.

In DDA not all peptides are submitted to fragmentation due to 'undersampling' (when too many peptides elute at a similar retention time, the MS instrument does not have sufficient time to select all of them for fragmentation), due to ion abundance not passing a preset threshold considered necessary to produce high quality MS/MS spectra or due to the exclusion of peptide ions with too low or too high charge states. In addition to fragments originating from the target peptides, MS/MS spectra contain noise peaks, which must be discriminated from peptide-derived peaks. The ideal case, where a fragment ion spectrum

Fig. 2. Principles of peptide fragmentation leading to incomplete sequence reads in proteomics. a) Schematic representation of a peptide fragment ion series (a, b, c for N-terminal and x, y, z for C terminal fragments) obtained with collision induced dissociation (CID; purple) or electron transfer dissociation (ETD; blue). b) Fragment ion (MS/MS) spectra with a complete ion series for the peptide KIQVLQQQADDAEER that is suitable to establish the complete amino acid sequence either through database search (DBS) or de novo sequencing. c) Fragment ion series with gaps of the peptide EANFDINQLYDCNWVVVNCSTPGNFFHVLR. Red arrows show the gaps corresponding to the FNGP and WNCDYLQNIDFNAE sequence tags in the y-ion series. These gaps are only partially covered by the corresponding b ions for the larger gap making de novo interpretation of this MS/MS spectrum impossible. Non-identified signals are shown in gray. They may correspond to noise, non-interpreted fragment ions or fragment ions from coeluting peptides that fell into the precursor ion selection window. (visualization made with PeptideShaker [92]).


a)
b)
NH2-K I Q V L Q Q Q A D D A E E R-COOH
E E A D D A Q Q Q L V Q I
Int
m/z
c)
NH2-E A N F D I N Q L Y D C N W V V V N C S T P G N F F H V L R-COOH
V H F
T S C N V V V
FNGP
WNCDYLQNIDFNAE
Int
m/z

allows an unambiguous de novo interpretation to establish the entire amino acid sequence of a peptide is rare. Recent efforts in the development of novel fragmentation methods as well the combination of such methods with chemical derivatization promise to overcome the problem of incomplete fragment ion spectra.

The most successful and widely used strategy to analyze MS/MS data is the database search approach (DBS). This approach requires high-quality reference databases for matching peptide mass spectra against a set of known proteins. Most commonly, public, species-wide databases such as Ensembl [29], UniProt [30] or NCBI RefSeq [31] are used. These databases contain sets of curated and predicted protein sequences for a wide range of organisms. When using a public repository as reference database, it is necessary to make a choice between the use of a well-defined set of proteins leading to more reliable identifications or including entries with little or no experimental support for a more complete identification of protein species.

DBS uses a target list of protein sequences that is hypothesized to be present in the sample and calculates a similarity score between the list of measured fragment ion masses and all possible theoretical fragment ion masses that can be predicted for the peptide calculated from a target list of protein sequences. The 'hit' with the highest similarity score is considered the peptide sequence match (PSM) for a given MS/MS spectrum. PSM scores require a statistical interpretation, which estimates the false discovery rate (FDR) of the resulting peptide identifications. There are two main statistical approaches to estimating the FDR of a set of identified peptides. PeptideProphet estimates the FDR by fitting different mathematical distributions to the correct and incorrect identifications as part of a mixture model to estimate the number of correct and incorrect PSMs [32]. This empirical Bayesian approach uses an Expectation-Maximization algorithm to calculate the parameters of the two sampling distributions. The target-decoy approach uses a reversed or scrambled protein or peptide sequence list to estimate the incorrect distribution. PeptideProphet can be applied to a decoy sequence list to estimate the correct and incorrect distributions more accurately or use complex semiparametric modeling of the two distributions. Many other tools based on similar approaches have been developed to estimate the FDR such as Percolator [33–35]. Recently, an additional validation level was introduced, which calculates a peptide-level FDR by taking the best PSM score for each peptide into account [36]. The size of the protein sequence database used for DBS has a strong influence on the distribution of correct and incorrect PSMs. Trying to match measured MS/MS spectra to a large database containing many protein sequences that are absent in the sample, increases the likelihood of erroneous PSMs. This situation is challenging for empirical Bayesian approaches to discriminate between correct and incorrect PSM score distributions. If precursor ion selection is not unambiguous due to peptides with similar or identical masses eluting close to each other, mixed fragment ion spectra are obtained, which give a low score with automated identification tools [37,38]. In this case it is better to submit mixed MS/MS spectra to search engines that can deal with multiple identifications as implemented in Andromeda [39], ProbIDtree [40] or JUMP [41]. It should be noted that the application of target-decoy approaches has been criticized because it may provide inaccurate FDR estimates for small datasets and multi-step database searches [36,42]. It is thus pivotal to use a list of protein sequences that corresponds as closely as possible to those in the measured proteome. Increasing the reliability of PSMs and peptide identifications is of prime importance in proteogenomics, since many splice isoforms and genetic variants such as single amino acid variants can only be identified through a single peptide. One way to improve the reliability of identifications is to achieve a more distinct separation of the correct and incorrect PSM score distributions. Besides the use of high-resolution mass spectrometry this can be achieved by adding physicochemical parameters measured in the separation step, such as retention time [43] and isoelectric point [44], to the identification procedure. The ultimate proof of a correct identification remains however a match between the MS/MS spectrum of a given peptide (isoform- or genetic-variant-related) and the MS/MS spectrum of the corresponding synthetic peptide as shown by Kim et al. [1].

De novo sequencing tools may provide unambiguous identifications but only for a small fraction of spectra with high-intensity and complete fragment ion series. For MS/MS spectra containing a limited number of short gaps, de novo sequencing tools [45,46] may provide a list of all possible peptide sequences, which can be filtered by comparison with a target list of protein sequences, for example using a public protein sequence database such as UniProt. Another popular approach is using short sequential sequence-tags, which are deduced from the successive fragments of an ion series without a gap [47]. The identified sequence tag is then used together with the precursor ion mass to determine the PSM score. The use of spectral libraries using consensus spectra assembled from high-quality MS/MS spectra of identified peptides and using a spectral similarity score, such as a dot product that takes the intensity of fragment ion signals into account, is gaining acceptance in the proteomics community since most detected peptides have already been identified in other studies. It is thus critical that rigorous quality control procedures be applied before admitting any MS/MS spectrum into a peptide spectral library [48,49]. The main identification approaches based on PSMs are summarized in Fig. 3 and have been reviewed by Nesvizhskii [50].

3.3. Identified proteins and proteins in the sample

There are several caveats when using public databases with respect to identifying protein species:

1) the data posted in public databases might be incomplete and not all protein species are reported, especially for non-model organisms
2) the reported protein species may not be reliable, depending on the quality of the evidence supporting their presence ranging from direct experimental evidence at the protein level to in silico predictions. neXtProt categorizes protein evidence in five levels: evidence at the protein level, evidence at the transcript level, evidence inferred from homology (derived by sequence similarity of known proteins in related species), evidence predicted from gene sequence and uncertain [51].
3) the presence of "new" protein species in a sample due to genetic rearrangements, for example, during cancer development

A major question is thus how to construct the database that contains the protein sequences that we expect to be present in the analyzed sample. By lack of any other source of information, best practice is to use a high-quality manually curated protein sequence database like Swiss-Prot or another database which contains a "best guess" of the protein composition of the analyzed sample. This approach leads to the generation of enormous sequence databases representing canonical sequences, the longest sequences explaining the most common forms of gene products. Using such "average" protein sequence databases leaves a considerable number of high-quality MS/MS spectra unidentified, part of which can be attributed to sequence variants that are absent from the consensus database. The peptide identification rate is an important quality parameter of the DDA LC–MS/MS dataset, since low rates may be attributed to poor quality and/or chimeric MS/MS spectra, unexpected post-translational modifications or absence of the protein sequence that contains the measured peptide from the protein database. Fig. 4 shows the intensity distribution of isotope clusters in single-stage MS data that, based on the isotope distribution, are assumed to correspond to peptides (in gray), the intensity distribution of clusters that were selected for fragmentation (in red) and those clusters leading to PSM-based identifications after statistical validation (in green) illustrating the low rate at which peptide-related isotope clusters lead to peptide identifications in a typical DDA LC–MS/MS experiment [52].

The ultimate goal, reconstitution of the list of all proteins that are present in the actual sample, is particularly difficult with shotgun

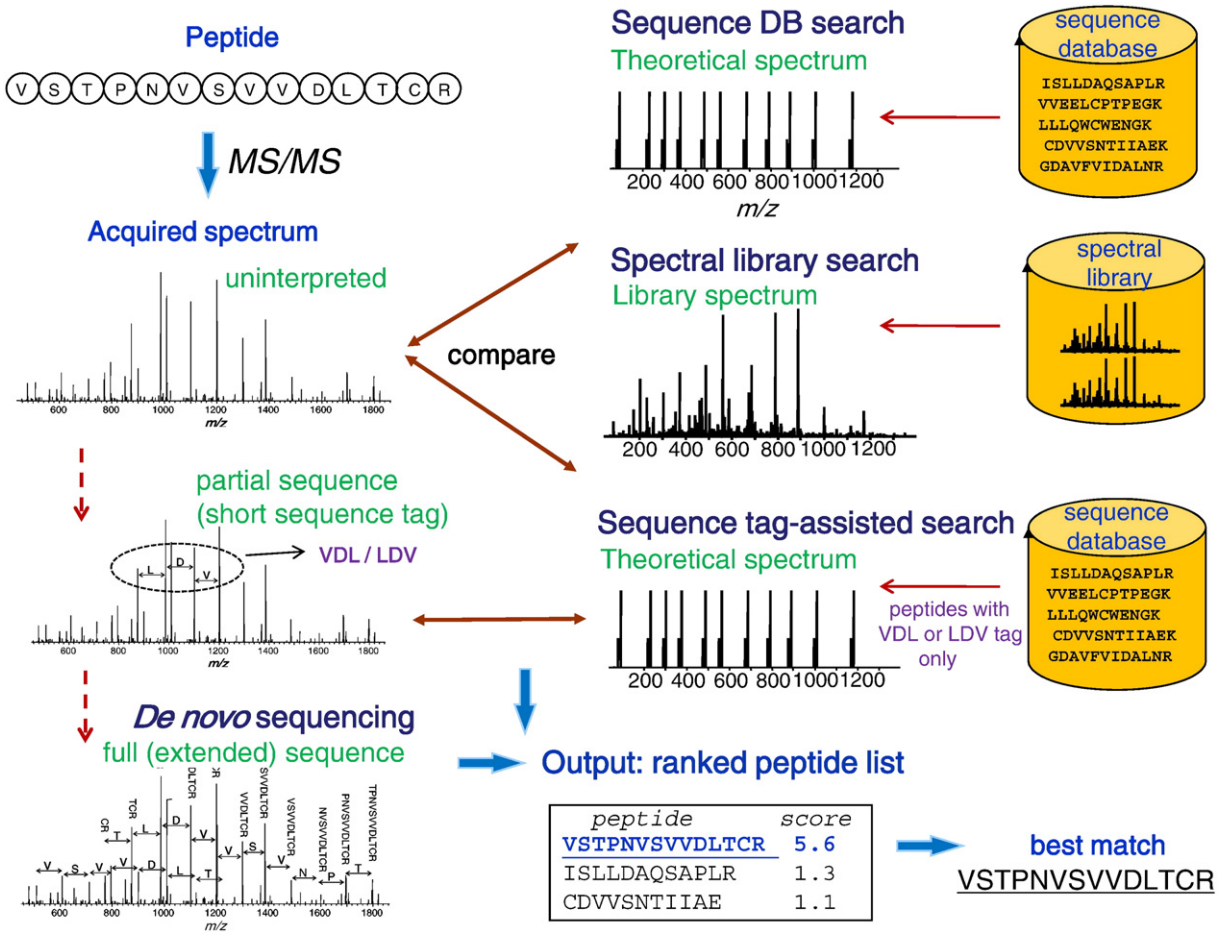


Fig. 3. Overview of the principal MS/MS spectrum matching strategies for peptide identification. Identification of the primary sequence of a peptide based on an MS/MS spectrum can be performed by correlating it with theoretical spectra predicted from a protein sequence database for precursor ions within a given m/z window (DBS approach) or against high-quality consensus spectra from a spectral library (spectral library search). The de novo sequencing approach does not rely on a reference protein sequence database to determine the amino acid sequence of a peptide. Hybrid peptide spectral matching (PSM) approaches such as the short-sequence-tag-assisted database search start with identification of short sequence tags (length 3 amino acids in this illustration), which is followed by DBS in which the list of candidate peptides is restricted to those peptides that contain the identified sequence tag, have the same precursor ion and mass tags that correspond to the mass values before and after the short identified sequence tag (adapted from Fig. 2 in [50]).

proteomics, since only a small fraction of the peptides is identified based on PSMs (see Fig. 4). Another difficulty is that many peptides do not map uniquely to one protein species. One reason for the low sequence coverage and the large number of shared peptides is the widespread use of trypsin as the only protease determining the length distribution of peptides resulting from a particular proteome. A limitation of short peptides is that they are very often not specific enough to match uniquely to one particular protein. Taken together, this leads to the so-called protein inference problem (PIP). In this procedure a list of peptides, that were identified confidently with a fixed FDR, is used to reconstruct a minimal set of protein groups from the original set of protein sequences that formed the basis of the PSM-based database search. Protein groups contain proteins with different sequences that cannot be discriminated from each other based on the reconstituted PSMs. A classification hierarchy with respect to protein/peptide sequence redundancy has been proposed by Farrah et al. [53]. PSMs used for protein assembly contain false positives and the fact that erroneous PSMs are distributed randomly throughout the proteome, while multiple correctly identified peptides tend to map to the same protein, leads to higher FDRs for proteins compared to FDRs at the PSM level. It is therefore necessary to estimate the FDR also at the protein level. This limitation in combination with the PIP currently prevents us from determining the real protein composition in biological samples using a bottom-up, shotgun proteomics approach and specifically to identify or differentiate protein species in a biological sample. A top-down proteomics approach, providing sequence information on intact proteins, overcomes these specific limitations and is an attractive alternative if it can be employed on a proteome-wide scale, for which a number of technical hurdles still have to be taken. An intermediate approach, where larger peptides are analyzed than in the typical tryptic digests, may proof to be the best compromise between data quality and technical feasibility. This so-called middle-down proteomics approach is described in more detail in Section 4.2.

It is clear from the above that current large-scale proteomics efforts fail to capture the major part of protein variability and that novel approaches are needed. One way forward is making use of genomics and transcriptomics data to generate dedicated databases to guide proteomics data analysis, the so-called proteogenomics approach. Aggregate databases or multistep database searches in a subset of large sequence databases are widely used for this purpose. Aggregating multiple databases or performing PSM-based identifications in multiple steps explodes, however, the search space leading to high error rates, which are difficult to estimate with current approaches such as the target-decoy FDR calculation. For example, Kim et al. followed a 7-step identification procedure using SEQUEST and Mascot in a combination of 7 databases [1]. The first search was performed using the human Reference Sequence (RefSeq) database (the RefSeq database contains well annotated non-redundant DNA, transcript and protein sequences)

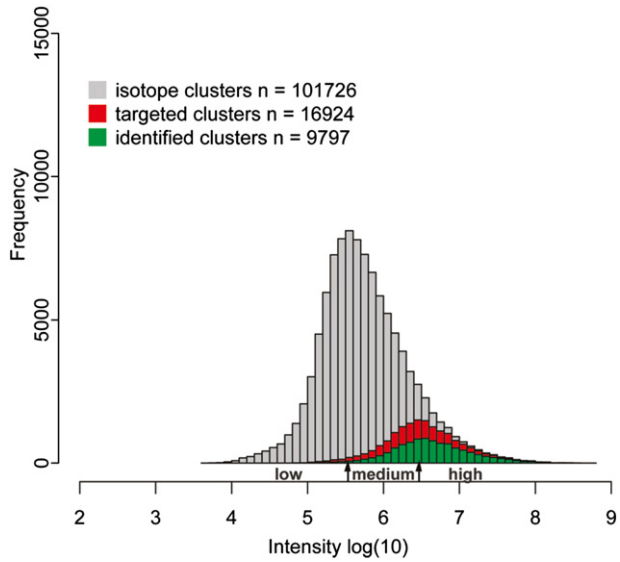


Fig. 4. Histogram of the frequency of the detected peptide-related isotope clusters in an LC-MS/MS run for single-stage MS (gray), isotope clusters submitted for fragmentation (red) and identified peptides (green). Data was filtered for peptides eluting between 5 and 30% of organic phase and having a charge of +2 or more. The histogram is split into low, medium and high abundance peptides (adapted from Fig. 2 in [52]).

[31]. All unmatched spectra were submitted for DBS against (1) a 6-frame-translated hg19 human genome database from NCBI, (2) 3-frame-translated RefSeq mRNA sequences, (3) a 3-frame-translated pseudogene database with sequences derived from NCBI and Gerstein's pseudogene database, (4) 3-frame-translated non-coding RNAs from NONCODE [54], (5) an N-terminal peptide database derived from RefSeq mRNA sequences from NCBI and (6) a signal peptide database from SignalP and HPRD. Using this PSM identification approach the authors identified protein products for 17 294 genes, 223 385 exons, 4 105 N-termini, 66 947 exon-exon junctions, 329 signal peptide cleavage sites, 193 novel protein coding regions (pseudogenes, non-coding RNAs, upstream ORFs, other ORFs), 106 novel exons, 198 novel N-termini, 70 gene/protein extensions and 40 exon extensions. While impressive, these results should be taken with care, as FDR estimation of multistep DBS with large sequence databases is inaccurate [42,55] notwithstanding the fact that 98 PSM-based identifications of novel peptide sequences from various proteogenomics categories were confirmed based on MS/MS spectra of the corresponding synthetic peptides. Problems with FDR calculation were demonstrated by Ezkurdia et al. showing that in this study 108 olfactory receptors were identified despite the fact that nasal tissue was not included in the analysis [56]. The authors listed 3 main reasons for the false identification of olfactory receptors: (1) no proper distinction between discriminating and non-discriminating peptides, resulting in the identification of olfactory receptors by peptides that map to multiple genes, (2) misidentification of peptides having a glutamine to pyroglutamic acid modification in non N-terminal positions and (3) identifications that were based on low-quality spectra. In general multistep DBS identification approaches using different databases for different types of genetic variants and an adjusted FDR for each database are preferred over a one-step large DBS identification procedure [42,85]. However the computational mass spectrometry community is currently searching for the new approaches to select the most appropriate identification and FDR calculation approach for the integration of large genome-wide data datasets with proteomics data.

Other proteogenomics approaches did not use aggregated databases or complex multistep identification approaches but databases derived by RNA-Seq of mRNA from the same sample. This approach results in much smaller protein sequence databases, which nevertheless include a large number of highly homologous protein sequences. For example, in a study by Low et al. 2D-LC–MS/MS proteomics data were acquired from samples of liver tissue from hypertensive and control rats that were digested with multiple proteases [57]. The data were analyzed using a sample-specific database from genome and RNA-seq data. From 2 million PSMs, 175,000 non-redundant peptides from 26,463 protein species were identified. In this study 1195 predicted genes were validated, and evidence was found for 83 splice events, 126 single amino acid variants and 20 isoforms with non-synonymous RNA editing events (Table 1).

There is no agreement yet on the best approach to perform proteogenomics analyses, but it is evident that smaller protein sequence databases based on RNA-seq and genomics data are preferable due to the more accurate search space. However, this approach is still not without problems as protein sequences may be missed (see Section 2.2). The conventional target-decoy estimation of FDR is not accurate in this case and the best identification strategy is still under debate in the bioinformatics community [55].

4. Matching genomics and proteomics — approaches to improving protein sequence coverage

4.1. Predicting peptide fragmentation

Integration of current proteomics data with genomics or transcriptomics data is hampered by the fact that the experimental evidence for peptide sequences is almost always incomplete and protein sequence coverage is often low. To get unequivocal evidence of the entire peptide sequence, without relying on a database, a complete fragment ion series is required to allow de novo sequencing. Both DBS and de novo sequencing algorithms mainly use the mass-to-charge ratio information of the predicted fragment ions because prediction of fragment ion intensity is rather inaccurate. While the study of peptide fragmentation has provided some insight in the general mechanisms [25], prediction of the relative intensity of peptide fragments is a complex issue. Some progress has, however, been made based on purely physicochemical parameters [58–60]. Prediction rules have also been empirically derived for electrospray [61,62] and MALDI MS/MS data [63], which differ primarily in the charge state of the generated peptide ions. Lack of standardization of fragmentation conditions complicates predictions further. Different ionization techniques (ESI vs MALDI) and tandem mass analyzers (e.g. ion trap vs Q-ToF) produce different MS and MS/MS data, although the same main peptide fragment ion series are typically detected. However, in all cases inhomogeneous fragmentation is observed with peptide bond cleavage adjacent to certain amino acid residues being more prevalent. For instance, proline uniquely has an imine-group that strongly favors fragmentation of the peptide bond at its N-terminal side [64]. Other prevalent residue- or sequence-dependent fragmentation mechanisms have been reported, specifically providing fragments at the C-terminal side of acidic residues [65]. In addition to residue-specific fragmentation, the position of a residue within a peptide influences the fragmentation behavior. The large dynamic range in fragment ion intensity leads to the observation

Table 1
Example of protein predictions constructed by a proteogenomics approach [57].

Isoform source	Predicted by DNA and RNA-Seq	Confirmed by MS/MS
Public gene annotation	32,971	13,088
Non-synonymous DNA polymorphisms	6187	126[a]
Predicted gene models supported by RNA-Seq	2903	1187
Splicing events	2545	83[a]
Non-synonymous RNA-editing	196	20[a]

[a] The proportion is small, since peptide identification has to overlap the variable sites.

that MS/MS spectra of intense precursor ions are more likely to show complete ion series than low-intensity precursors. The greatly disparate fragment ion intensities can only be controlled to a limited extent by adjusting instrumental parameters, and the most viable solution is to change peptide properties.

4.2. Protein digestion methods

The residue specificity of the digestion method has two key effects on peptide fragmentation and sequence coverage. First, the average length and mass of the generated peptides are strongly affected. Larger peptides are more difficult to fragment, particularly in CID, due to the distribution of the collision energy across multiple chemical bonds in large molecules. The second effect concerns residue distribution, in particular of those with distinct proton affinities due to acidic and basic side chain moieties [66]. Tryptic peptides will always have a basic Arg or Lys at the C-terminus (except for the C-terminal peptide of the protein), and an indeterminate number and distribution of acidic residues, while peptides of a GluC protein digest likewise have Glu and sometimes Asp at the C-terminus and a variable composition of basic residues. None of the available digestion methods can provide full peptide and protein sequence coverage, but the combination of information from multiple digestions yields complementary data [67] and is often used to cover the complete sequence of, for example, biopharmaceutical proteins, for which some new, unusual proteases are being evaluated [68].

Trypsin is the protease of choice in bottom-up proteomics since it produces peptides of lengths and composition that are amenable to ionization and fragmentation. For example, the average and median peptide mass upon full tryptic cleavage of the human proteome is around 1100 Da and 800 Da, respectively. There are on average 1.2 acidic (Asp and Glu) and 1.3 basic residues (Lys, Arg and His) per tryptic peptide. The relatively short peptides produced by trypsin have drawbacks for complete protein sequence coverage, and alternative cleavage methods deserve consideration [60]. While nonspecific proteases such as elastase and pepsin, producing even shorter peptides, are sometimes used, a more promising approach for high sequence coverage, while at the same time reducing the peptide interference problem, is to target larger peptides. Instrumental improvements in CID methods and the development of ETD allow fragmentation of larger peptides and open the way to so-called middle-down proteomics [69], situated between standard, bottom-up proteomics targeting peptides of 1–3 kDa and top-down proteomics, targeting intact proteins using somewhat specialized equipment [70]. The mass range between 3 and 10 kDa may well prove to be optimal for reaching full protein coverage, and peptides in this range can be produced by targeting rare residues, such as Met, Cys, Tyr and Trp, which can be achieved by chemical or electrochemical digestion methods.

A particularly interesting chemical cleavage reagent is cyanogen bromide (CNBr). This reagent has found use in membrane protein analysis due to its specificity for Met, which is overrepresented in transmembrane helices (TMHs). CNBr reacts well within detergent-solubilized TMH protein regions, unlike the more bulky proteases. As an added benefit CNBr cleavage produces a homoserine lactone upon cleavage, which is a form of active ester and allows amidation at basic pH [71], a derivatization step which may also be employed to improve fragmentation (see next section). The average and median peptide mass upon full CNBr cleavage of the human proteome is around 5000 Da and 3300 Da, respectively, with on average of 5.4 acidic (Asp and Glu) and 6.4 basic residues (Lys, Arg and His) per CNBr peptide. Peptides in this size range are still amenable to reversed phase LC–MS and their expected high charge state combined with their high mass make them ideal candidates for ETD fragmentation.

An alternative cleavage method studied in our group uses electrochemical (EC) oxidation of Tyr and Trp [72–74] and shares the advantages of CNBr cleavage of higher average peptide mass than trypsin (average and median of 2800 Da and 1810 Da, respectively) and a C-terminal activated spirolactone suitable for amidation [75]. There are on average 3.0 acidic (Asp and Glu) and 3.5 basic residues (Lys, Arg and His) per EC-generated peptide.

4.3. Derivatization methods

Another strategy to attain more complete sequence coverage involves chemical derivatization to introduce groups that affect fragmentation by changing the chemical properties of the peptide. Due to its reactivity, by far the most common target for derivatization of peptides is the N-terminal amine, but depending on the experimental conditions, the ε-amine group of lysine may also be derivatized. Guanidination of the ε-amine group by its selective reaction with *o*-methylisourea transforms it into homoarginine with a pK_a similar to that of arginine making it unreactive during subsequent labeling of the N-terminal amine [76]. C-terminal labeling is less frequently used in proteomics, since the reactivity of the carboxylic acid group in aqueous buffers is low. Chemical activation is possible and certain cleavage reactions activate the C-terminus of a peptide by lactone formation (see previous section).

Derivatization with groups carrying a permanent charge such as quaternary ammonium or phosphonium ions (e.g. tris(2,4,6-trimethoxyphenyl)-phosphonium or TMPP) has been used for a long time in MS to improve ionization. The presence of a permanent charge at the N- or C-terminus in a peptide greatly favors the formation and prominence of b- and y-ions, respectively [77]. While most commonly used for N-terminal derivatization [78], C-terminal derivatization with TMPP has been shown to improve peptide fragmentation [79], in particular of the C-terminal peptide of proteins and GluC-derived peptides.

Tertiary amines have a high proton affinity, and when used for N-terminal and lysine amine derivatization in combination with tryptic digestion, lead to peptides with comparably basic groups at both the C- and N-terminal side. Tertiary amines are commonly used in isobaric labeling reagents, such as TMT and iTRAQ, and have been shown to lead to more homogeneous fragmentation and the presence of abundant b- and y-ions [80,81]. As in most derivatization methods, fragments derived from the reagent (including, but not limited to TMT and iTRAQ isotopically-labeled reporter ions) contribute to the MS/MS spectra, but do not seem to negatively affect sequence ion yields. Interestingly, acetylation of the N-terminus has also been reported to enhance the b-ion series and suppress the y-ion series [77]. Another benefit of N-terminal modification is that the b_1 ion is commonly observed, unlike in underivatized peptides. Finally, the mass increase imparted by the derivatization reagent should be taken into account for selection of peptide precursor ions, since it can be considerable. For example, 8-plex iTRAQ used on a single Lys-containing peptide increases the mass by two times at 304 Da.

The location of the proton(s) along the peptide backbone or side chains has a direct effect on fragmentation by CID. The mobile proton model has been proposed to explain fragmentation behavior of peptides with basic residues at different locations [25,47,82]. This is most strikingly observed in singly-charged peptides, which give rise to mainly y-type ions derived from the C-terminus while b-ions, derived from the N-terminus are less frequently observed because of the tendency of charge retention on the more basic C-terminus.

In tryptic peptides, where there are no or few mobile protons due to the basic C-terminal residue, peptide fragmentation can be strongly influenced by negative charge derivatization. Introduction of a negative charge at the N-terminus of peptides has been shown to drastically improve MS/MS spectral quality by mitigating the effect of basic residues, and ensuring the presence of a mobile proton. Derivatization with sulfonic acid groups at the N-terminus is the most commonly employed method, in the form of 4-sulphophenyl isothiocyanate (SPITC), sulfobenzoic acid or cysteic acid [83,84]. The presence of a (nearly) complete y-ion series greatly facilitates de novo sequencing (Fig. 5).

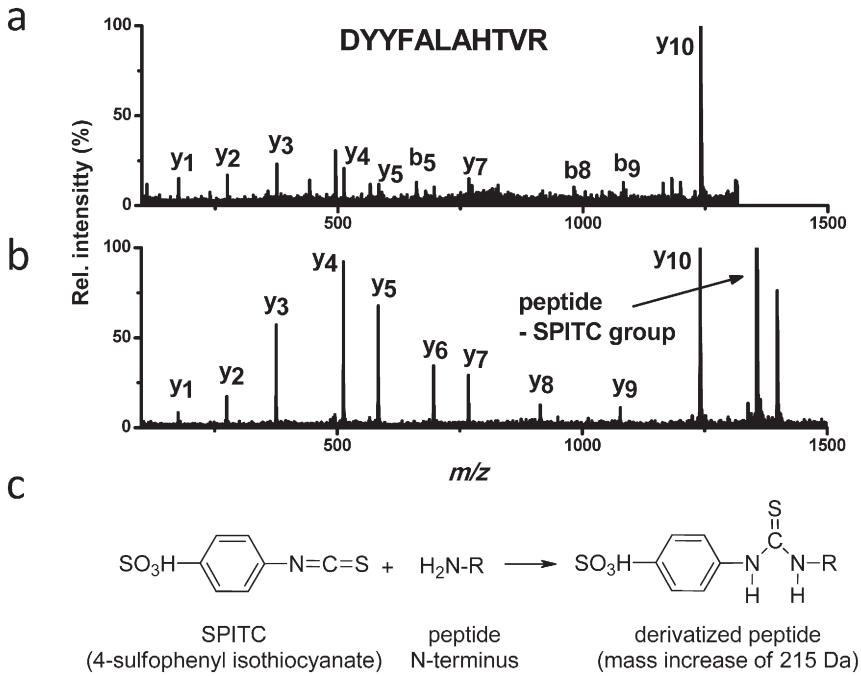


Fig. 5. Improvement in fragmentation upon charge-derivatization of a peptide. a) MALDI-TOF/TOF fragmentation spectrum of the underivatized peptide DYYFALAHTVR (charge state +1); b) MALDI-TOF/TOF of the same peptide after derivatization of the N-terminus with SPTIC, showing a complete y-ion series and the absence of b-ions; c) reaction of SPITC with a peptide N-terminus.

The loss of peptides during chemical derivatization and the potentially detrimental effect on MS sensitivity prevented the wider acceptance of derivatization for large-scale proteomics experiments. A notable exception is the use of isotopically labeled iTRAQ or TMT for quantitative proteomics, which shows that with careful optimization of reaction conditions, peptide and signal loss can be minimized. In general, derivatization of peptides opens new routes to influencing their fragmentation and complements the use of different fragmentation methods. In the ideal situation every peptide would be equally well ionized and each of its peptide bonds would fragment with the same probability resulting in fragments from all expected ion series with the same abundance. An acceptable compromise with respect to peptide sensitivity and fragment intensity using a single combination of digestion method, labeling chemistry, ionization method, and fragmentation method is most likely not feasible for most proteins and the use of complementary methods will remain necessary.

5. Proteogenomics data integration — an example

The value of proteogenomics data integration was shown in several recent studies [85]. One of these studies, by Low et al., which was already briefly discussed in Section 3.3, set out to investigate differences between spontaneously hypertensive and normotensive rats, common models for hypertension research [57]. Genome sequences of these inbred rats identified over ten thousand non-synonymous variants in four-and-a-half thousand protein-coding genes. The liver transcriptome constructed for these strains further enriched the assortment of protein isoforms due to splicing (over one thousand genes), confirmed gene predictions (almost three thousand) and RNA editing (two hundred non-synonymous changes), many of which were also detected at the peptide level (see Section 3.3 and Table 1).

Previous genomic and physiological studies generated large lists of candidate genes [86–88]. The inherited hypertension phenotype in this model suggests that disease is caused by a genetic defect that should manifest itself at the transcriptome and proteome levels. While hundreds of genes showed differential expression at the RNA and proteome level, only 41 genes showed consistent changes at both levels, including a prominent hit from a human genome wide association study, the *Cyp17a1* gene [89]. Despite initial failure to identify a functional variant causing this deregulation based on public gene annotation, transcriptome data highlighted an additional exon (missing in the public gene build) and an adjacent DNA polymorphism. This change in the *Cyp17a1* core promoter is hypothesized to alter a forkhead-box DNA binding site and is likely responsible for abolition of transcription and the virtual absence of the protein product in hypertensive rats.

Although the effect of this DNA polymorphism still has to be shown, modern genome editing tools allow direct testing of this hypothesis, e.g. by editing the forkhead-box DNA binding site in normotensive rats or by recovering its function in hypertensive rats. This study represents an example how an integrated proteogenomics approach identified a potential causative DNA variant for a gene that might prove to be a common cross-specific modulator of hypertension.

6. Concluding remarks

Next generation DNA/RNA sequencing has opened the possibility to study inter-individual variability at the genome and transcriptome level revealing thousands of changes in coding and non-coding regions. Trying to translate these changes into personalized or precision medicine to predict disease susceptibility or prevent disease development has, however, remained difficult. To aid in this transition, it is critical to investigate how genetic variability and environmental factors reflect on the proteome by integrating genomic/transcriptomic and proteomic data in what is called proteogenomics. Integrative analysis of genome, transcriptome and proteome data can deliver a better insight into interactions between 'Omics layers' and may pinpoint potential disease factors that escape compensatory regulation and therefore manifest themselves at multiple levels [57]. Several recent studies aim at streamlining proteogenomics data analysis by integrating ribosome profiling and MS data (e.g. the PROTEOFORMER tool) [90] or by analyzing and visualizing proteogenomics data (e.g. with PGTools) [91].

This review provides an overview of current limitations in proteomics methodology notably with respect to obtaining complete fragment ion mass spectra that allow reading the entire amino acid sequence of a

peptide and ultimately of the corresponding protein, as this forms the basis of detecting protein species resulting from genetic variability. Indeed current proteomics methodology captures only a small part of proteomic variability resulting from genetic variability. A first step is to use individualized sequence databases as reference for proteomic data analysis. It is clear that, even then, much of the proteomic landscape will remain invisible but tools such as derivatisation or middle-down proteomics are being developed to improve the situation by increasing the protein sequence coverage and it is conceivable that our view on how genetic variability influences biological networks and phenotypes will become more complete.

Conflict of interest

The authors declare no conflict of interests.

References

[1] M.S. Kim, S.M. Pinto, D. Getnet, R.S. Nirujogi, S.S. Manda, R. Chaerkady, et al., A draft map of the human proteome, Nature 509 (2014) 575–581.
[2] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M.M. Savitski, et al., Mass-spectrometry-based draft of the human proteome, Nature 509 (2014) 582–587.
[3] H. Schluter, R. Apweiler, H.G. Holzhutter, P.R. Jungblut, Finding one's way in proteomics: a protein species nomenclature, Chem. Central J. 3 (2009) 11.
[4] C.T. Walsh, Posttranslational Modifications of Proteins: Expanding Nature's Inventory, Roberts and Company Publishers, 2006.
[5] C.T. Walsh, S. Garneau-Tsodikova, J. Gatto, Protein posttranslational modifications: the chemistry of proteome diversifications, Angew. Chem. Int. Ed. 44 (2005) 7342–7372.
[6] S. Bringans, J.K. Hane, T. Casey, K.C. Tan, R. Lipscombe, P.S. Solomon, et al., Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*, BMC Bioinform. 10 (2009) 301.
[7] M. Guttman, M. Garber, J.Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, et al., Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, Nat. Biotechnol. 28 (2010) 503–510.
[8] S. Harrer, S.C. Kim, C. Schieber, S. Kannam, N. Gunn, S. Moore, et al., Label-free screening of single biomolecules through resistive pulse sensing technology for precision medicine applications, Nanotechnology 26 (2015) 182502.
[9] W. Hancock, G. Omenn, P. Legrain, Y.K. Paik, Proteomics, human proteome project, and chromosomes, J. Proteome Res. 10 (2011) 210.
[10] Y.K. Paik, S.K. Jeong, G.S. Omenn, M. Uhlen, S. Hanash, S.Y. Cho, et al., The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome, Nat. Biotechnol. 30 (2012) 221–223.
[11] Consortium GotN, Whole-genome sequence variation, population structure and demographic history of the Dutch population, Nat. Genet. 46 (2014) 818–825.
[12] J. Weischenfeldt, O. Symmons, F. Spitz, J.O. Korbel, Phenotypic impact of genomic structural variation: insights from and for human disease, Nat. Rev. Genet. 14 (2013) 125–138.
[13] M.J. del Prete, R. Vernal, H. Dolznig, E.W. Mullner, J.A. Garcia-Sanz, Isolation of polysome-bound mRNA from solid tissues amenable for RT-PCR and profiling experiments, RNA 13 (2007) 414–421.
[14] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, J.S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling, Science 324 (2009) 218–223.
[15] C. Barbosa, I. Peixeiro, L. Romao, Gene expression regulation by upstream open reading frames and human disease, PLoS Genet. 9 (2013), e1003529.
[16] N.T. Ingolia, L.F. Lareau, J.S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes, Cell 147 (2011) 789–802.
[17] K.F. Au, V. Sebastiano, P.T. Afshar, J.D. Durruthy, L. Lee, B.A. Williams, et al., Characterization of the human ESC transcriptome by hybrid sequencing, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) E4821–E4830.
[18] H.B. Oh, B. Moon, Radical-driven peptide backbone dissociation tandem mass spectrometry, Mass Spectrom. Rev. 34 (2015) 116–132.
[19] G.P. Mommen, C.K. Frese, H.D. Meiring, J. van Gaans-van den Brink, A.P. de Jong, C.A. van Els, et al., Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD), Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 4507–4512.
[20] L.C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, et al., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, Mol. Cell Proteomics 11 (2012)http://dx.doi.org/10.1074/mcp.O111.016717.
[21] G. Hopfgartner, D. Tonoli, E. Varesio, High-resolution mass spectrometry for integrated qualitative and quantitative analysis of pharmaceuticals in biological matrices, Anal. Bioanal. Chem. 402 (2012) 2587–2596.
[22] A. Bilbao, E. Varesio, J. Luban, C. Strambio-De-Castillia, G. Hopfgartner, M. Muller, et al., Processing strategies and software solutions for data-independent acquisition in mass spectrometry, Proteomics 15 (2015) 964–980.
[23] O.T. Schubert, L.C. Gillet, B.C. Collins, P. Navarro, G. Rosenberger, W.E. Wolski, et al., Building high-quality assay libraries for targeted analysis of SWATH MS data, Nat. Protoc. 10 (2015) 426–441.
[24] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, et al., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nat. Methods 12 (2015) 523–526.
[25] B. Paizs, S. Suhai, Fragmentation pathways of protonated peptides, Mass Spectrom. Rev. 24 (2005) 508–548.
[26] D.M. Crizer, S.A. McLuckey, Electron transfer dissociation of amide nitrogen methylated polypeptide cations, J. Am. Soc. Mass Spectrom. 20 (2009) 1349–1354.
[27] J.E.P. Syka, J.J. Coon, M.J. Schroeder, J. Shabanowitz, D.F. Hunt, Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, Proc. Natl. Acad. Sci. 101 (2004) 9528–9533.
[28] Y. Xia, H.P. Gunawardena, D.E. Erickson, S.A. McLuckey, Effects of cation charge-site identity and position on electron-transfer dissociation of polypeptide cations, J. Am. Chem. Soc. 129 (2007) 12232–12243.
[29] F. Cunningham, M.R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, et al., Ensembl 2015, Nucleic Acids Res. 43 (2015) D662–D669.
[30] Uniprot Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.
[31] K.D. Pruitt, G.R. Brown, S.M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, et al., RefSeq: an update on mammalian reference sequences, Nucleic Acids Res. 42 (2014) D756–D763.
[32] K. Ma, O. Vitek, A.I. Nesvizhskii, A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet, BMC Bioinform. 13 (Suppl. 16) (2012) S1.
[33] L. Käll, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, Nat. Methods 4 (2007) 923–925.
[34] L. Käll, J.D. Storey, M.J. MacCoss, W.S. Noble, Assigning significance to peptides identified by tandem mass spectrometry using decoy databases, J. Proteome Res. 7 (2008) 29–34.
[35] L. Käll, J.D. Storey, W.S. Noble, Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry, Bioinformatics 24 (2008) i42–i48.
[36] K. Jeong, S. Kim, N. Bandeira, False discovery rates in spectral identification, BMC Bioinforma. 13 (Suppl. 16) (2012) S2.
[37] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N.G. Ahn, W.M. Old, Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies, J. Proteome Res. 9 (2010) 4152–4160.
[38] H. Li, K.B. Hwang, D.G. Mun, H. Kim, H. Lee, S.W. Lee, et al., Estimating influence of cofragmentation on peptide quantification and identification in iTRAQ experiments by simulating multiplexed spectra, J. Proteome Res. 13 (2014) 3488–3497.
[39] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (2011) 1794–1805.
[40] N. Zhang, X.J. Li, M. Ye, S. Pan, B. Schwikowski, R. Aebersold, ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer, Proteomics 5 (2005) 4096–4106.
[41] X. Wang, Y. Li, Z. Wu, H. Wang, H. Tan, J. Peng, JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy, Mol. Cell Proteomics 13 (2014) 3663–3673.
[42] N. Gupta, N. Bandeira, U. Keich, P.A. Pevzner, Target-decoy approach and false discovery rate: when things may go wrong, J. Am. Soc. Mass Spectrom. 22 (2011) 1111–1120.
[43] T. Baczek, R. Kaliszan, Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics, Proteomics 9 (2009) 835–847.
[44] R.M. Branca, L.M. Orre, H.J. Johansson, V. Granholm, M. Huss, A. Perez-Bercoff, et al., HiRIEF LC–MS enables deep proteome coverage and unbiased proteogenomics, Nat. Methods 11 (2014) 59–62.
[45] K.F. Medzihradszky, R.J. Chalkley, Lessons in de novo peptide sequencing by tandem mass spectrometry, Mass Spectrom. Rev. 34 (2015) 43–63.
[46] J. Seidler, N. Zinn, M.E. Boehm, W.D. Lehmann, De novo sequencing of peptides by MS/MS, Proteomics 10 (2010) 634–649.
[47] D.L. Tabb, Z.Q. Ma, D.B. Martin, A.J. Ham, M.C. Chambers, DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring, J. Proteome Res. 7 (2008) 3838–3846.
[48] Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, J.A. Vizcaíno, Making proteomics data accessible and reusable: current state of proteomics databases and repositories, Proteomics 15 (2015) 930–950.
[49] O.T. Schubert, L.C. Gillet, B.C. Collins, P. Navarro, G. Rosenberger, W.E. Wolski, et al., Building high-quality assay libraries for targeted analysis of SWATH MS dta, Nat. Protoc. 10 (2015) 426–441.
[50] A.I. Nesvizhskii, A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics, J. Proteomics. 73 (2010) 2092–2123.
[51] L. Lane, A. Bairoch, R.C. Beavis, E.W. Deutsch, P. Gaudet, E. Lundberg, et al., Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins, J. Proteome Res. 13 (2014) 15–20.
[52] A. Michalski, J. Cox, M. Mann, More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC–MS/MS, J. Proteome Res. 10 (2011) 1785–1793.
[53] T. Farrah, E.W. Deutsch, G.S. Omenn, D.S. Campbell, Z. Sun, J.A. Bletz, et al., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas, Mol. Cell Proteomics 10 (2011) (M110.006353).

[54] D. Bu, K. Yu, S. Sun, C. Xie, G. Skogerbo, R. Miao, et al., NONCODE v3.0: integrative annotation of long noncoding RNAs, Nucleic Acids Res. 40 (2012) D210–D215.
[55] Omenn GS, Lane L, Lundberg EK, Beavis RC, Nesvizhskii AI, Deutsch EW. Metrics for the Human Proteome Project 2015: Progress on the human proteome and guidelines for high-confidence protein identification. J. Proteome Res. 2015.
[56] I. Ezkurdia, J. Vazquez, A. Valencia, M. Tress, Analyzing the first drafts of the human proteome, J. Proteome Res. 13 (2014) 3854–3855.
[57] T.Y. Low, S. van Heesch, H. van den Toorn, P. Giansanti, A. Cristobal, P. Toonen, et al., Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis, Cell Rep. 5 (2013) 1469–1478.
[58] S. Sun, K. Meyer-Arendt, B. Eichelberger, R. Brown, C.Y. Yen, W.M. Old, et al., Improved validation of peptide MS/MS assignments using spectral intensity prediction, Mol. Cell. Proteomics 6 (2007) 1–17.
[59] Z. Zhang, Prediction of low-energy collision-induced dissociation spectra of peptides, Anal. Chem. 76 (2004) 3908–3922.
[60] Z. Zhang, Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges, Anal. Chem. 77 (2005) 6364–6373.
[61] J.E. Elias, F.D. Gibbons, O.D. King, F.P. Roth, S.P. Gygi, Intensity-based protein identification by machine learning from a library of tandem mass spectra, Nat. Biotechnol. 22 (2004) 214–219.
[62] E.A. Kapp, F. Schutz, G.E. Reid, J.S. Eddes, R.L. Moritz, R.A. O'Hair, et al., Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation, Anal. Chem. 75 (2003) 6251–6264.
[63] J. Khatun, K. Ramkissoon, M.C. Giddings, Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry, Anal. Chem. 79 (2007) 3032–3040.
[64] L.A. Breci, D.L. Tabb, J.R. Yates 3rd, V.H. Wysocki, Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra, Anal. Chem. 75 (2003) 1963–1971.
[65] A.G. Sullivan, F.L. Brancia, R. Tyldesley, R. Bateman, K. Sidhu, S.J. Hubbard, et al., The exploitation of selective cleavage of singly protonated peptide ions adjacent to aspartic acid residues using a quadrupole orthogonal time-of-flight mass spectrometer equipped with a matrix-assisted laser desorption/ionization source, Int. J. Mass Spectrom. 210–211 (2001) 76–665.
[66] D.L. Tabb, Y. Huang, V.H. Wysocki, J.R. Yates 3rd., Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides, Anal. Chem. 76 (2004) 1243–1248.
[67] B. Meyer, D.G. Papasotiriou, M. Karas, 100% protein sequence coverage: a modern form of surrealism in proteomics, Amino Acids 41 (2011) 291–310.
[68] K. Srzentić, L. Fornelli, Ü.A. Laskay, M. Monod, A. Beck, D. Ayoub, et al., Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies, Anal. Chem. 86 (2014) 9945–9953.
[69] C. Wu, J.C. Tran, L. Zamdborg, K.R. Durbin, M. Li, D.R. Ahlf, et al., A protease for 'middle-down' proteomics, Nat. Methods 9 (2012) 822–824.
[70] A.J. Forbes, M.T. Mazur, H.M. Patel, C.T. Walsh, N.L. Kelleher, Toward efficient analysis of >70 kDa proteins with 100% sequence coverage, Proteomics 1 (2001) 927–933.
[71] A. Compagnini, V. Cunsolo, S. Foti, R. Saletti, Improved accuracy in the matrix-assisted laser desorption/ionization-mass spectrometry determination of the molecular mass of cyanogen bromide fragments of proteins by post-cleavage reaction with tris(hydroxymethyl)aminomethane, Proteomics 1 (2001) 967–974.
[72] H.P. Permentier, A.P. Bruins, Electrochemical oxidation and cleavage of proteins with on-line mass spectrometric detection: development of an instrumental alternative to enzymatic protein digestion, J. Am. Soc. Mass Spectrom. 15 (2004) 1707–1716.
[73] J. Roeser, N.F.A. Alting, H.P. Permentier, A.P. Bruins, R. Bischoff, Boron-doped diamond electrodes for the electrochemical oxidation and cleavage of peptides, Anal. Chem. 85 (2013) 6626–6632.
[74] J. Roeser, H.P. Permentier, A.P. Bruins, R. Bischoff, Electrochemical oxidation and cleavage of tyrosine- and tryptophan-containing tripeptides, Anal. Chem. 82 (2010) 7556–7565.
[75] J. Roeser, N.F.A. Alting, H.P. Permentier, A.P. Bruins, R.P.H. Bischoff, Chemical labeling of electrochemically cleaved peptides, Rapid Commun. Mass Spectrom. 27 (2013) 546–552.
[76] F.L. Brancia, H. Montgomery, K. Tanaka, S. Kumashiro, Guanidino labeling derivatization strategy for global characterization of peptide mixtures by liquid chromatography matrix-assisted laser desorption/ionization mass spectrometry, Anal. Chem. 76 (2004) 2748–2755.
[77] G. Sonsmann, A. Romer, D. Schomburg, Investigation of the influence of charge derivatization on the fragmentation of multiply protonated peptides, J. Am. Soc. Mass Spectrom. 13 (2002) 47–58.
[78] Z.H. Huang, J. Wu, K.D. Roth, Y. Yang, D.A. Gage, J.T. Watson, A picomole-scale method for charge derivatization of peptides for sequence analysis by mass spectrometry, Anal. Chem. 69 (1997) 137–144.
[79] C. Nakajima, H. Kuyama, T. Nakazawa, O. Nishimura, C-terminal sequencing of protein by MALDI mass spectrometry through the specific derivatization of the alpha-carboxyl group with 3-aminopropyltris-(2,4,6-trimethoxyphenyl)phosphonium bromide, Anal. Bioanal. Chem. 404 (2012) 125–132.
[80] T. Gandhi, P. Puri, F. Fusetti, R. Breitling, B. Poolman, H.P. Permentier, Effect of iTRAQ labeling on the relative abundance of peptide fragment ions produced by MALDI-MS/MS, J. Proteome Res. 11 (2012) 4044–4051.
[81] P. Pichler, T. Kocher, J. Holzmann, M. Mazanek, T. Taus, G. Ammerer, et al., Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap, Anal. Chem. 82 (2010) 6549–6558.
[82] V.H. Wysocki, G. Tsaprailis, L.L. Smith, L.A. Breci, Mobile and localized protons: a framework for understanding peptide dissociation, J. Mass Spectrom. 35 (2000) 1399–1406.
[83] T. Keough, R.S. Youngquist, M.P. Lacey, Sulfonic acid derivatives for peptide sequencing by MALDI MS, Anal. Chem. 75 (2003) 156a–165a.
[84] M. Yamaguchi, T. Nakazawa, H. Kuyama, T. Obama, E. Ando, T.A. Okamura, et al., High-throughput method for N-terminal sequencing of proteins by MALDI mass spectrometry, Anal. Chem. 77 (2005) 645–651.
[85] A.I. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat. Methods 11 (2014) 1114–1125.
[86] S.S. Atanur, I. Birol, V. Guryev, M. Hirst, O. Hummel, C. Morrissey, et al., The genome sequence of the spontaneously hypertensive rat: analysis and functional significance, Genome Res. 20 (2010) 791–803.
[87] C. Morrissey, I.C. Grieve, M. Heinig, S. Atanur, E. Petretto, M. Pravenec, et al., Integrated genomic approaches to identification of candidate genes underlying metabolic and cardiovascular phenotypes in the spontaneously hypertensive rat, Physiol. Genomics 43 (2011) 1207–1218.
[88] M. Simonis, S.S. Atanur, S. Linsen, V. Guryev, F.P. Ruzius, L. Game, et al., Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel, Genome Biol. 13 (2012) r31.
[89] X. Li, Y. Ling, D. Lu, Z. Lu, Y. Liu, H. Chen, et al., Common polymorphism rs11191548 near the CYP17A1 gene is associated with hypertension and systolic blood pressure in the Han Chinese population, Am. J. Hypertens. 26 (2013) 465–472.
[90] J. Crappe, E. Ndah, A. Koch, S. Steyaert, D. Gawron, S. De Keulenaer, et al., PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration, Nucleic Acids Res. 43 (2015), e29.
[91] S.H. Nagaraj, N. Waddell, A.K. Madugundu, S. Wood, A. Jones, R.A. Mandyam, et al., PGTools: a software suite for proteogenomic data analysis and visualization, J. Proteome Res. 14 (2015) 2255–2266.
[92] M. Vaudel, J.M. Burkhart, R.P. Zahedi, E. Oveland, F.S. Berven, A. Sickmann, et al., PeptideShaker enables reanalysis of MS-derived proteomics data sets, Nat. Biotechnol. 33 (2015) 22–24.