Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing

Ludovic C. Gillet,[1,*] Alexander Leitner,[1,*] and Ruedi Aebersold[1,2]

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, 8093 Zürich, Switzerland; email: aebersold@imsb.biol.ethz.ch

[2]Faculty of Science, University of Zürich, 8057 Zürich, Switzerland

Annu. Rev. Anal. Chem. 2016. 9:449–72

First published online as a Review in Advance on March 30, 2016

The *Annual Review of Analytical Chemistry* is online at anchem.annualreviews.org

This article's doi: 10.1146/annurev-anchem-071015-041535

*These authors contributed equally to the preparation of this review.

Keywords

data-dependent acquisition, data-independent acquisition, peptide-centric, shotgun proteomics, spectrum-centric, targeted proteomics

Abstract

Proteins constitute a key class of molecular components that perform essential biochemical reactions in living cells. Whether the aim is to extensively characterize a given protein or to perform high-throughput qualitative and quantitative analysis of the proteome content of a sample, liquid chromatography coupled to tandem mass spectrometry has become the technology of choice. In this review, we summarize the current state of mass spectrometry applied to bottom-up proteomics, the approach that focuses on analyzing peptides obtained from proteolytic digestion of proteins. With the recent advances in instrumentation and methodology, we show that the field is moving away from providing qualitative identification of long lists of proteins to delivering highly consistent and accurate quantification values for large numbers of proteins across large numbers of samples. We believe that this shift will have a profound impact for the field of proteomics and life science research in general.

1. INTRODUCTION

Data-independent acquisition (DIA): LC-MS/MS mode of acquisition in which a user-predefined set of mass ranges is repeatedly selected by the MS instrument for fragmentation

Proteomics aims at characterizing the entire protein content present in a cell, tissue, or bodily fluid at a given point in time. Depending on the scope of the biological question, a proteomic analysis may consist of one or a combination of the following steps: identification of the proteins, including the nature and position of any posttranslational modifications (PTMs; e.g., phosphorylation or glycosylation); measurement of proteins' dynamic quantitative changes between conditions (e.g., normal versus disease, or control versus treated samples); and study of the protein conformations or interactions within larger protein complexes or in the context of broader biological networks or pathways. The overarching goal of these analyses is to acquire a better understanding of the biological processes in play in cells or organs, identify new drug targets, advance the understanding of mechanisms of drug actions, improve diagnosis and prognosis of diseases (through biomarker discovery), or perform better patient stratification for therapeutic treatments.

For most of these analyses, liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) has become an indispensable technology. Based on the method of protein/proteome characterization, the proteomic field can be subdivided into two main analytical streams: the top-down and the bottom-up approaches. The top-down approach relies on the analysis of intact proteins by MS and their extensive characterization through the fragmentation of intact proteins within the mass spectrometer, followed by the measurement of these fragment ions. In contrast, the bottom-up approach proceeds through a peptide-to-protein inference logic. The proteins present in a sample first undergo digestion into smaller peptides through the use of specific proteolytic enzymes. The mass spectrometer is then used to identify the sequence of these peptides, usually through the generation of a diagnostic fragment ion spectrum. The identified peptide sequences must then be (re)assigned to the proteins they originate from, through a nontrivial process called protein inference (1). Although a priori more convoluted than the top-down workflow, the bottom-up approach has been increasingly adopted by the community, because the mass range and fragmentation characteristics of peptides match more closely the MS instrumentation capabilities than that of large intact proteins. Over the past two decades, several implementations of data acquisition and data query strategies have been developed to allow fast, high-throughput qualitative and quantitative characterization of the protein content of cells or organs in a relatively straightforward manner.

In this review, we focus mainly on summarizing the current state of bottom-up MS technology applied to proteomics. We show that even within the bottom-up proteomic field, an already overwhelming set of MS acquisition modes and data query strategies is now available. We summarize the principles and limitations of the three main acquisition modes used in the field: data-dependent acquisition (DDA; frequently termed shotgun, or discovery, proteomics); targeted data acquisition, carried out mostly by selected reaction monitoring (SRM; also referred to as multiple reaction monitoring, or MRM); and data-independent acquisition (DIA). We discuss the dimensionality and properties of the data structure resulting from each acquisition mode and describe the different data query strategies (spectrum-centric or peptide-centric) that can be used to explore these data sets. We show that each combination of a particular acquisition mode and data query strategy delivers qualitative and quantitative information that differs significantly in the quality and completeness of the resultant protein quantification matrices. Finally, we show that a new trend in proteomics research is emerging that aims to use MS as a tool for biological exploration by delivering highly consistent protein quantification matrices across large sample cohorts rather than just generating long lists of identified proteins for a limited number of samples.

2. APPLICATIONS OF MS-BASED PROTEOMICS TO BIOLOGICAL QUESTIONS AND EXPERIMENTAL DESIGN

2.1. MS-Based Proteomics in the Omics Era

The emergence of modern technologies for the comprehensive measurement of biomolecules, also referred to as omics technologies, has been reshaping the scope and type of biological experimentation. Whereas an array of well-established protocols and methods—including, for example, site-directed mutagenesis, Western blotting, in vitro protein assays, and in vivo cell imaging—are used in classical experimental biology to assess the presence, sequence, function, and biochemical mode of action of one or a few specific proteins, omics technologies focus on producing large-scale data sets at high throughput, providing broad insights into the molecular makeup of cells and tissues. Omics approaches generally rely on gathering "big data," for instance, from controls and cases in large (clinical) sample cohorts or perturbation series in systems biology studies (2); they rely on external resources such as databases for data mining, gene ontology maps for network analyses, orthogonal data sets for association studies, and mathematical modeling to decipher a posteriori the molecular foundations that explain biological or phenotypic observations. The rise of these new methods has been provokingly presented by some as the end of classical hypothesis-driven science whereby correlation would supersede causality (3), opening passionate discussions regarding the future of science in the big data era (4, 5).

Both small-scale mechanistic studies and data-driven omics approaches have made extensive use of MS technology. A first set of techniques has focused on the analysis of a small number of proteins, with the aim of extensively characterizing them. This is exemplified by LC-MS/MS top-down strategies that can distinguish different protein variants (6) by analyzing intact proteins [see (7) and the review by Kelleher and colleagues (8) in this volume]. For the characterization of secondary (protein domain folds) and tertiary (global folding) protein structures and identification of interacting partners or interfaces of protein subunits in the context of protein complexes, a suite of structural proteomics techniques has been developed (9). Dedicated workflows have also been devised to provide insight into subcellular protein localization and turnover (10). The interaction of proteins with other proteins or small molecules such as drugs can be revealed by various forms of interaction proteomics, such as affinity purification coupled to MS (11, 12). Although very valuable, these techniques require a high level of expertise and optimization on an almost per protein level, and their detailed description falls beyond the scope of this review.

In contrast, a range of bottom-up proteomic techniques has been developed to perform rapid, robust, large-scale, and high-throughput qualitative and quantitative analyses of the protein composition of complex samples, thereby producing large inventory lists of proteins for the respective samples. Each of these techniques has different performance characteristics that optimally match specific applications, as described in more detail below. Moreover, as discussed throughout this review, high-throughput, bottom-up proteomics techniques are increasingly being used to generate highly reproducible quantitative measurements of specific proteins across large sample cohorts via targeted acquisition schemes or peptide-centric data query strategies. Ultimately, MS-based proteomics may now enable biologists to perform hypothesis testing on hundreds or thousands of hypotheses concurrently from suitably acquired data sets.

2.2. Challenges Underlying Bottom-Up Proteomics Studies

As mentioned above, most proteomic studies are currently geared toward the discovery or validation of differential protein regulation on a large scale in response to biological perturbations.

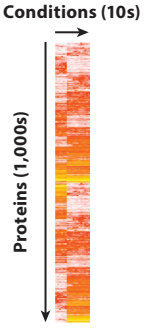
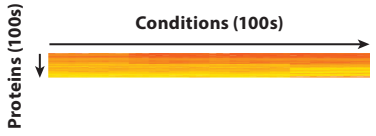
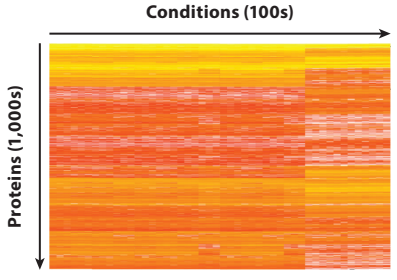
Experiment type	Aim	Analytical challenges	Typical protein quantification matrix
Discovery project	Identify proteins up/downregulated upon perturbation or disease state	Sample complexity proteome subfractionation Use of incomplete genomic databases Identification of protein variants Selection of quantotypic peptides for protein quantification	
Validation project	Systematically monitor biomarkers of interest across large sample cohorts	Biological sampling reproducibility Sample preparation reproducibility LC-MS/MS condition control	
Systems biology **Pathway perturbation** **Treatment time course**	Discover proteins up/downregulated upon perturbation Understand/model the proteomic changes on a systemic level	Combination of above Going beyond assay libraries for peptide identification	

Figure 1

Biological experimental design, analytical challenges, and related protein quantification matrix considerations. The heatmap representations have been generated and modified from in-house data and are shown here only for illustration purposes. Abbreviation: LC-MS/MS, liquid chromatography–tandem mass spectrometry.

Regulation might be reflected by changes in global protein abundance, modifications of the protein PTM stoichiometry, or changes in protein activity. These types of biological questions are best approached through large-scale, high-throughput characterization of the protein content of biological samples using one or several implementations of the bottom-up proteomics MS acquisition modes (DDA, SRM, DIA). The different MS acquisition strategies and the type of information that can be generated thereby are described more extensively in the next sections. Here, we focus more generally on the strategies and challenges underlying discovery/validation proteomic analyses using the example of biomarker studies (**Figure 1**), which aim at identifying diagnostic factors.

In biomarker studies, an initial discovery step usually involves the analysis of a relatively small number of well-defined samples characteristic of a given disease state (e.g., tumor versus control, invasive versus noninvasive tumor), and the analytical emphasis is usually put on identifying as many significantly differentially abundant candidates as possible. The first analytical challenge at this stage concerns the complexity of and the wide range of protein concentration in these types

of samples. This challenge can be addressed by performing depletion of high-abundance proteins and/or by using fractionation or enrichment techniques that target a specific subproteome (e.g., N-glycosylated proteins expressed at the cell surface and/or shed into the extracellular environment).

Quantotypic peptide: an amino acid sequence obtained from proteolytic digestion used as a proxy for quantification of its protein of origin in a biological sample

A second challenge concerns the identification of the natural peptide variants. This is becoming of particular relevance for biomarker projects that aim at analyzing samples from large, genetically heterogeneous human populations. As discussed more extensively in Section 3.3, the prototypical protein identification strategy in bottom-up discovery proteomics relies on spectrum-centric database searching to assign peptide sequences to the precursors ionized and fragmented in the mass spectrometer. These protein sequence databases are usually derived from genomic sequencing information. With current genomic sequencing community efforts and the advent of new RNA sequencing technology, we expect to see a vast improvement in the annotation of natural single nucleotide polymorphism (SNP) variants in protein databases. However, genomic/transcriptomic annotations will remain of little use for the identification of PTMs of proteins, and therefore efforts in the analytical preparation, mass spectrometric analysis, and bioinformatic strategies to identify these PTMs will remain important for the field.

A third challenge may concern the way protein quantification is inferred from peptide intensities in bottom-up proteomic approaches. Until now, the selection of which peptides and how many peptides per protein to use to infer the protein quantity has been mostly based on empirical rules. This decision may be more rationally streamlined in the future by using the recently introduced concept of quantotypic peptides (13), which can faithfully be used to infer protein quantification, and by applying algorithms for the automated selection of sibling peptides originating from the same protein and with highly correlated quantification profiles across sample sets (14, 15).

From a purely statistical perspective, though, the accuracy and consistency of protein quantification at the discovery stage has remained a relatively minor concern. Because the small number of samples involved in this phase does not allow the predictive power of the candidates for larger populations to be assessed in a statistically sound manner, the resulting protein quantification matrices may tolerate a relatively large number of missing values (**Figure 1**). The typical result of such discovery projects is thus a list of putative candidates that require further validation through the analysis of larger sample cohorts to assess their predictive prognostic power in an extended population (16). All the challenges mentioned above are generic to bottom-up approaches and are therefore common to all MS-based proteomics acquisition and data analysis strategies that are presented in Section 3. Until now, for the purposes of primary biomarker discovery, DDA acquisition coupled to spectrum-centric searching strategies remains the method of choice.

The validation phase of such a project aims at obtaining the most accurate and complete data matrix possible in which the proteins of interest are accurately and reproducibly quantified across a large number of samples (**Figure 1**). Therefore, emphasis needs to be put on the analytical parameters that may affect the consistency and reproducibility of the measurements, including the quality of the biological sampling, the reproducibility of the proteomic sample preparation, and the overall consistency and reproducibility of LC-MS/MS measurements. The latter requirements match more closely the performance characteristics of targeted data acquisition schemes or those of DIA methods combined with peptide-centric data query strategies. With the improvement in completeness and accuracy of these protein quantification matrices, biologically significant discoveries will be achievable.

The considerations given here in the context of biomarker discovery studies are equally applicable to other types of biological projects involving the systemic analysis of peptide and protein responses to various types of cellular perturbations (**Figure 1**). Bottom-up proteomics thus offers a rich array of methods that are particularly powerful when performance profiles are chosen to match specific output and application requirements.

3. TECHNICAL OVERVIEW OF EXPERIMENTAL STRATEGIES IN MS-BASED PROTEOMICS

3.1. Peptide Preparation and Separation

As introduced above, most large-scale, high-throughput proteomic strategies use the bottom-up approach, which is based on the mass spectrometric measurement of peptides. Proteins first undergo digestion into peptides, usually with the help of a specific protease, most commonly trypsin (17). Trypsin is favored because it is robust, cheap, and relatively specific, although special care is necessary in quantitative workflows to ensure reproducibility. For the most part, trypsin also generates peptides in a mass range of 500 to 3,000 Da, which is optimal for chromatographic separation and yields peptides that ionize and fragment well due to the presence of a C-terminal lysine or arginine residue that efficiently protonates under acidic conditions. Proteases with complementary cleavage specificities can be used to increase proteome coverage (18). Regardless of the enzymatic cleavage strategy, bottom-up proteomics increases sample complexity considerably (each protein generates many peptides) and complicates data analysis. Also, as the species identified by MS are peptides and not proteins, the amino acid sequence of the identified peptides must be (re)assigned to the proteins from which they originate (1, 19).

Following enzymatic digestion and purification, the resulting peptide mixtures are typically separated according to their hydrophobicity by reversed-phase, high-performance liquid chromatography (LC), and peptides eluting from the column are directly ionized by electrospray ionization before entering the mass spectrometer. Higher-efficiency peptide separation has been increasingly achieved via higher LC operating pressures, longer columns, and reduced particle sizes of the chromatographic material (20). However, even the most highly resolving single-dimension separations hardly reach a chromatographic peak capacity of 1,000 (21) and are thus insufficient to separate the tens to hundreds of thousands of different peptide species generated by the digestion of a complex proteome. This means that many peptides will coelute from the column and coionize simultaneously, which has profound implications on the performance characteristics of the MS data acquisition strategies as discussed in Sections 4–6.

3.2. Mass Spectrometric Analysis

The (partially) separated and ionized peptides are finally analyzed in the mass spectrometer. It determines the mass-to-charge (m/z) ratio of intact ions (precursors) or fragments thereof using different analyzers or combinations of analyzers (**Figure 2**). In contemporary proteomics research, different instrument configurations may be used depending on the primary operating modes (DDA, SRM, DIA). These different acquisition modes differ fundamentally in the way precursor and fragment m/z information is registered, and in the way the resulting data structure is analyzed (**Figure 3**).

In DDA, or shotgun, mode (22) an instrument is operated in iterative acquisition cycles of intact precursor-level spectra (MS, or MS^1) and fragment ion spectra (MS/MS, or MS^2) (**Figure 3*a***). Decisions about which precursors to select for fragmentation are made in real time by the instrument software, according to predefined criteria. The strengths and weaknesses of this acquisition mode are discussed in Section 4. Typical instrument configurations for DDA workflows include different Orbitrap hybrids (23) or quadrupole time-of-flight (TOF) (24, 25) designs (**Figure 2**). Peptide identity is usually derived from the combined information of the precursor mass and the corresponding fragment ion masses recorded in the MS/MS spectra, usually following untargeted, spectrum-centric searching strategies (26) (Section 3.3).

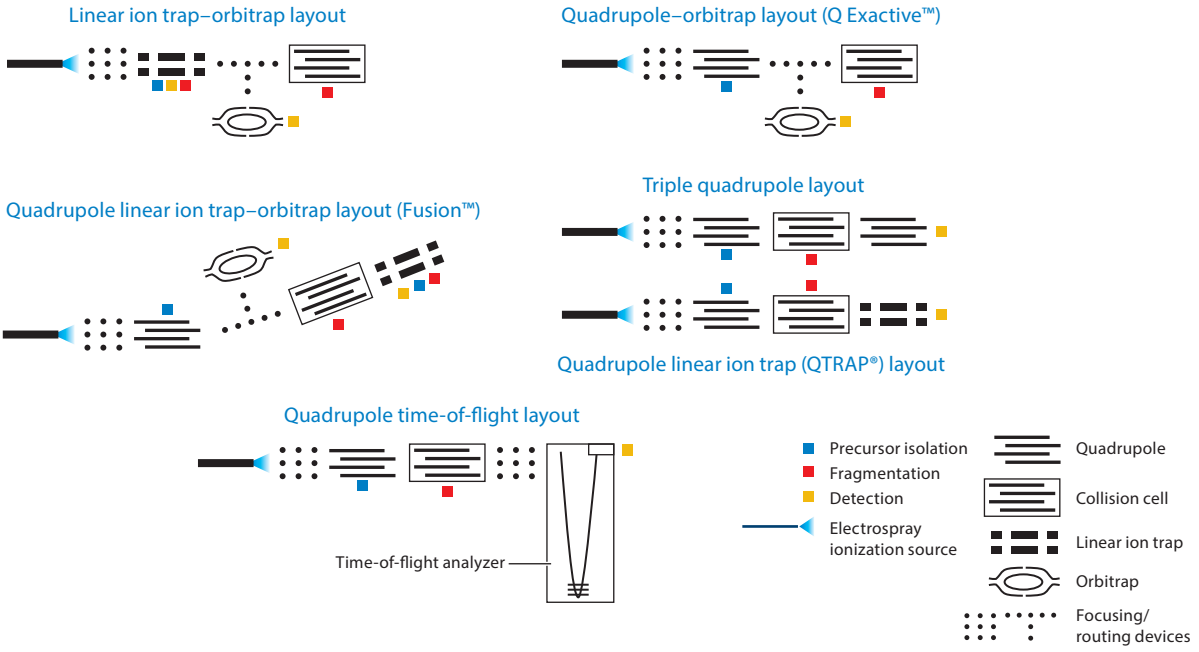
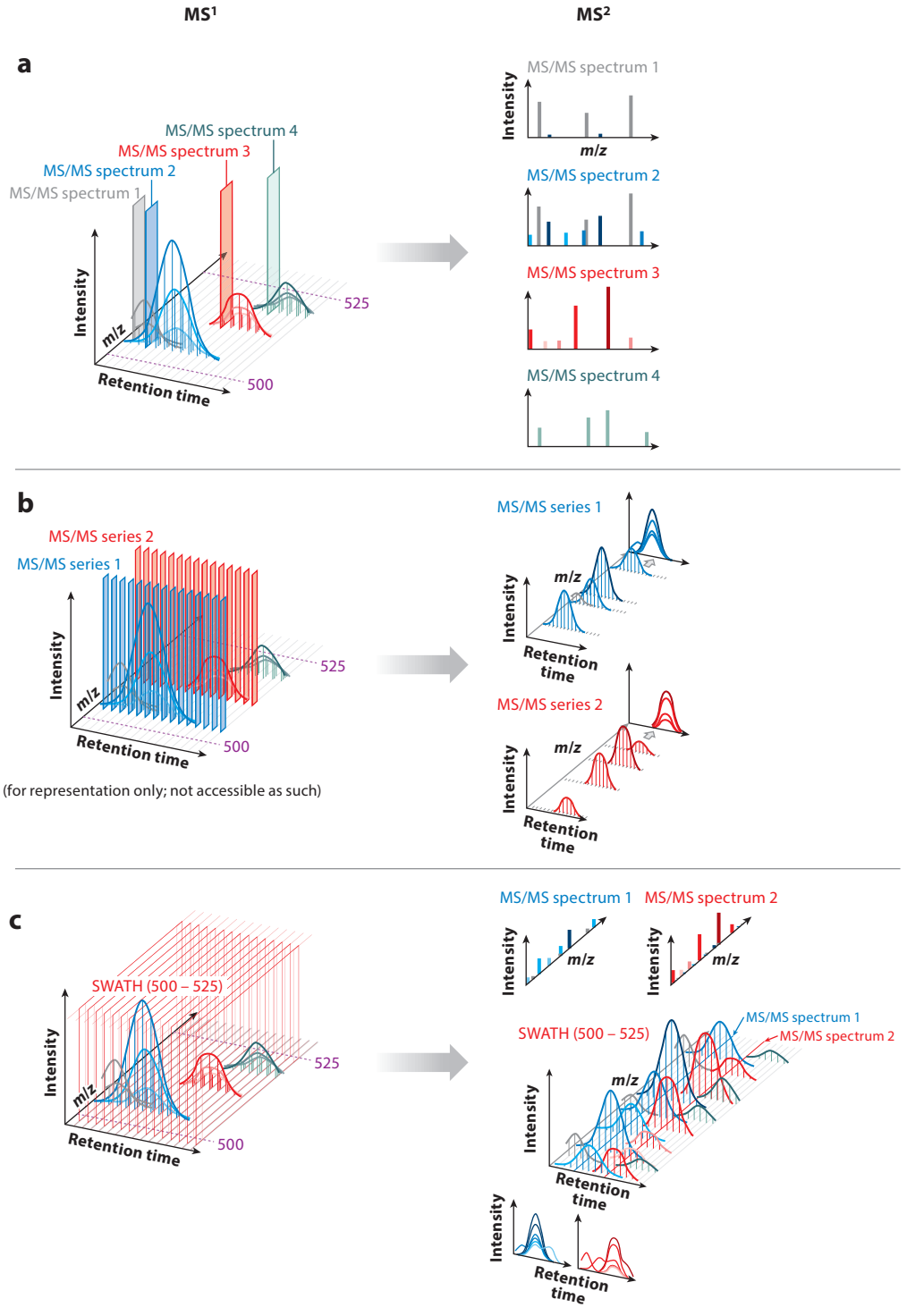


Figure 2

Schematic overview of different mass spectrometer designs used for contemporary proteomics research. Colored squares specify where the three most important experimental steps are executed: precursor isolation (*blue*), fragmentation (*red*), and detection (*yellow*).

For SRM (27), the most widely used targeted proteomic acquisition method, no complete MS/MS spectra are acquired. Rather, a set of discrete, predetermined fragment ion signals is repeatedly recorded for each predefined peptide over time (**Figure 3*b***). The combinations of precursor *m/z* and fragment ion *m/z* pairs (typically three to five per peptide) are termed transitions. The lists of transitions are predefined in the acquisition method. Therefore, prior knowledge about the identity of the targeted peptides and of their fragmentation characteristics is required. The technique is best applied for consistently quantifying or validating the presence of targeted peptides, rather than discovering new peptides/proteins. Both qualitative and quantitative information is directly derived from the resulting fragment ion chromatographic signals of individual transitions. For SRM, different instrument types have been used. For optimal combination of sensitivity and quantitative accuracy, this workflow is commonly carried out on triple quadrupole instruments or quadrupole linear ion traps (QTRAP) operated as triple quadrupoles (**Figure 2**). Recently, to improve resolution at the fragment ion level, instruments in which the third quadrupole is replaced with a high-resolution accurate mass analyzer have also been used in peptide-targeted acquisition. These methods, called parallel reaction monitoring (PRM) (28), or MS/MS^{ALL}, now allow recording of time-resolved MS/MS signals of predefined sets of peptides at high resolution.

DIA methods (29) combine certain aspects of both DDA and SRM methods. As with the targeted acquisition mode, the instrument performs continuous time-resolved acquisition of MS/MS signals for each precursor mass region (**Figure 3*c***). In contrast to DDA and SRM methods, however, the aim of the DIA mode is to exhaustively acquire MS/MS spectra for any possible precursor mass, regardless of the actual detection (DDA) or the presumed presence (SRM) of peptides/precursors. Because the current instrumentation is too slow to cover the complete tryptic peptide *m/z* range (typically 300–1,200) with a small precursor isolation window in the time frame of



a typical LC separation, the compromise of the DIA method is to use a much wider precursor isolation mass window for precursor fragmentation as compared with the 1–3 Da used in SRM or DDA. Thus, a mixture of several precursor (peptide) ions is deliberately isolated and fragmented, creating multiplexed MS/MS spectra consisting of fragment ion signals from different peptides. Given sufficient acquisition speed of the instrument and appropriate window sizes, it has become possible to cover the entire informative *m/z* range of a peptide mixture and to generate exhaustive maps of all observable peptide fragments with continuous MS/MS acquisition and sufficient time resolution of the data points (**Figure 3*c***).

3.3. MS/MS Data Structure and Analysis Strategies

Although quite different in their MS acquisition principles and performance characteristics (Sections 4–6), the three main bottom-up proteomic acquisition methods presented above essentially rely on the same type of information for peptide identification: the sequential recording of fragment ion signals (MS/MS) for the fragmented peptides. However, because of the difference in periodicity of these MS/MS recordings per peptide precursor, each acquisition scheme produces a significantly different type of MS/MS data structure (**Figure 3**).

Figure 3

Data structure of the LC-MS and LC-MS/MS signals acquired in the three main bottom-up proteomic acquisition strategies: (*a*) data-dependent acquisition (DDA), (*b*) targeted selected reaction monitoring (SRM), and (*c*) data-independent acquisition (DIA). Consecutive precursor MS spectra (*m/z* versus intensity, also referred to as MS^1 scans) can be displayed side-by-side across the LC separation (time), resulting in a three-dimensional MS^1 map of the LC-MS data. At the left of each panel, an excerpt is depicted of 17 consecutive MS^1 scans (e.g., spaced by 3.5 s) spanning approximately 1 min of LC separation and a mass range of 500–525 *m/z*. In this limited time-and-mass space, four precursor ions are represented, with ion signals in gray, blue, red, and green, respectively, and different color shading for the monoisotopic peak and the first two ^{13}C isotopes of the precursors. Note that the chromatographic elution profiles of the peptides span approximately 7–10 MS^1 scans, which supports the reconstruction of the chromatographic elution profiles of the precursors with sufficient resolution in the time dimension. The precursor isolation schemes are represented for each species by colored "windows." In DDA (panel *a*, *left*), all four peptides are selected for fragmentation with 1–3 Da isolation windows, but only once and only when the precursor signal is detected above a certain threshold (i.e., at the beginning of the chromatographic peak). In SRM (panel *b*, *left*), only the mass ranges of two peptides of interest (*blue* and *red*) are selected for fragmentation with 0.7–1 Da isolation, repeatedly and with a given periodicity during the elution. In DIA (panel *c*, *left*), the complete precursor mass range is selected for fragmentation in consecutive and contiguous steps (e.g., of 25 Da as shown here for region 500–525 *m/z* and adjacent regions), repeatedly during the LC separation, regardless of the detection or presumed presence of peptides in that mass-and-time space. Note that two peptides (*blue* and *gray*) are represented with similar precursor masses (e.g., within less than 0.5-Da difference), which results in acquisition of MS/MS spectra for the blue peptide that contain interfering fragment ion signals from the gray peptide upon typical DDA (MS/MS spectrum 2), SRM (MS/MS series 1), and DIA precursor isolation schemes (MS/MS spectrum 1 and *blue* chromatogram traces in the bottom panel). Other interfering signals are shown in a similar manner. The corresponding MS/MS signals recorded for each acquisition mode are shown at the right of each panel. Only the monoisotopic peak of the fragment ions is depicted (again using *gray*, *blue*, *red*, and *green*). For DDA, the time-discontinuous MS/MS spectra recordings (also referred to as MS^2 scans) can only be comprehended using the spectrum-based (mass-intensity) dimension. For SRM, the mass-discontinuous MS/MS recordings are most easily visualized using the chromatogram-based (time-intensity) dimension. For DIA, the time-and-mass-continuous MS/MS recordings can be displayed side-by-side across the LC separation (time), resulting in a three-dimensional MS^2 map of the LC-MS/MS data. The DIA MS/MS data structure can be visualized and explored either by the LC-MS/MS acquisition sequence in the spectrum-based dimension or by compiling the fragment ions' elution profiles in the chromatogram-based (time-intensity) dimension. Note that the chromatographic elution profiles of the fragment ion recordings in SRM (panel *b*, *right*) and DIA (panel *c*, *right*) match exactly those of their corresponding precursors shown at the left of the panels and span the same 7–10 MS^2 scans, again allowing the reconstruction of the chromatographic elution profiles for the fragment ions with sufficient resolution in the time dimension. The color shading is representative of the ion intensities of the peptide fragments as reported in the assay library used for the data acquisition or extraction. Abbreviations: LC-MS, liquid chromatography–mass spectrometry; LC-MS/MS, liquid chromatography–tandem mass spectrometry; MS or MS^1, precursor mass spectrum; MS/MS or MS^2, fragment ion mass spectrum; SWATH, sequential windowed acquisition of all theoretical fragment ion spectra.

Table 1 A unified view of LC-MS/MS data acquisition schemes and data query strategies[a]

Data acquisition scheme	Data query strategy (26)[b]	
	Spectrum centric	Peptide centric
Data-dependent acquisition	Database searching (91) Spectral library matching (40, 92) De novo sequencing (93, 94) Accurate mass and time (MS^1) (34, 35) Match between runs (MS^1) (33)	SALSA (MS^2) (46)
Targeted data acquisition (SRM, PRM)	SRM: not applicable PRM: database searching	LC-peak coelution scoring (75, 76)
Data-independent acquisition	Blunt/raw MS^2 searching (41–43, 80) Demultiplexed MS^2 searching (30, 37, 44)	FT-ARM (MS^2) (47) SWATH MS targeted extraction (38)

[a]Abbreviations: FT-ARM, Fourier transform-all reaction monitoring; LC-MS/MS, liquid chromatography-tandem mass spectrometry; MS^1, precursor mass spectrum; MS^2, fragment ion mass spectrum; PRM, parallel reaction monitoring; SALSA, scoring algorithm for spectral analysis; SRM, selected reaction monitoring.
[b]Reference numbers are shown in parentheses.

For example, the very extensive but discontinuous-in-time nature of the MS/MS in DDA mode results in data where the MS/MS spectra mainly serve as a central scoring dimension and query unit for the peptide identifications (**Figure 3*a***). Indeed, the vast majority of the community using the DDA acquisition mode analyzes their data via untargeted spectrum-centric database searching strategies. Alternatively, the continuous-in-time precursor isotopic patterns present in DDA MS^1 data can also be used to infer the identity of precursors for which MS/MS spectra have not been acquired or identified using database searching. This inference can be performed using accurate mass and time (AMT) tags, coordinates of the precursor chromatographic traces to transfer the peptide identification across LC-MS/MS runs upon MS^1 map alignments (30–33), or extensive AMT coordinate libraries (34, 35). It is important to note, however, that those strategies that solely rely on retention time and precursor mass coordinates in MS^1 maps may not provide the level of identification confidence achievable by peptide-centric query strategies that score multiple, independent coeluting MS/MS fragment ion traces from targeted SRM/PRM or DIA data, and therefore require special care when handling false discovery rate (FDR) estimation (36).

Conversely, the continuous-in-time signal acquisition scheme of SRM that does not generate complete MS/MS spectra can almost exclusively be analyzed using peptide-centric query strategies in the chromatogram dimensionality (**Figure 3*b***). DIA schemes that are continuous-in-time and in the MS/MS dimension are equally amenable to spectrum- and peptide-centric query strategies in any dimensionality (**Figure 3*c*** and **Table 1**). As a side note, it should be mentioned that the MS/MS chromatogram dimensionality in SRM and DIA naturally also embeds MS-related information such as the fragment ion intensity (in SRM) or the monoisotopic fragment ion mass accuracy and fragment charge state (in DIA).

On top of the dimensionalities used to represent the data, it is important to distinguish two different types of data query strategies that can be applied, depending on whether the data is directly used as a source of query to identify peptides or whether biological hypotheses (i.e., the existence of a given peptide species) are used for interrogating the data. In the literature, these approaches were sometimes referred to as untargeted (37) or targeted data analysis (38) strategies. However, to avoid confusion of the latter with the "targeted" MS data acquisition schemes (SRM and PRM), we have adopted the terminology of peptide-centric used by MacCoss and colleagues (26) for the rest of this review, whereby the peptide is the starting query unit used to interrogate

the data. In contrast, strategies that start from mass spectra as a query to infer peptide sequences are referred to as spectrum-centric (26). With this formalism of data acquisition and data query strategy, essentially all MS and MS/MS data analysis methods can be presented in a unified view (**Table 1**).

MS/MS spectrum: unit recording of the mass spectrometer acquiring fragments of analyte (peptide) precursors following a fragmentation event (e.g., collision-induced dissociation); also termed MS^2, or fragment ion, spectrum

In essence, the spectrum-centric data query strategy usually considers MS/MS signals to be the central unit used to identify the peptide sequences uniquely associated with each spectrum. The peptide identity can be inferred by (*a*) matching the experimental fragment ion signals against in silico generated MS/MS spectra from peptide sequence databases (**Figure 4*a***), (*b*) comparing the experimental fragmentation patterns to those of spectral libraries containing MS/MS information of previously identified peptides, or (*c*) trying to decipher de novo the experimental fragment ion series without the help of databases or libraries—or by any combination of the three approaches. In spectrum-centric strategies, MS/MS signals may be raw MS/MS spectra from DDA (39, 40) or DIA (41–43) data sets used directly for querying databases, or heavily preprocessed MS/MS signals originating from demultiplexed DIA MS/MS data spectra (37, 44, 45). Note that, in contrast to DIA, the discontinuous-in-time nature of MS/MS spectra acquired in DDA mode does not allow reconstitution of the chromatographic elution dimension of the MS/MS signals. In the absence of orthogonal information, it is therefore almost impossible with DDA data to confidently determine whether the fragment ion signals recorded in a MS/MS spectrum are indeed derived from the intended peptide precursors or from interfering fragments from coisolated precursor species, or whether they are just noise (**Figures 3*a*** and **4*a***).

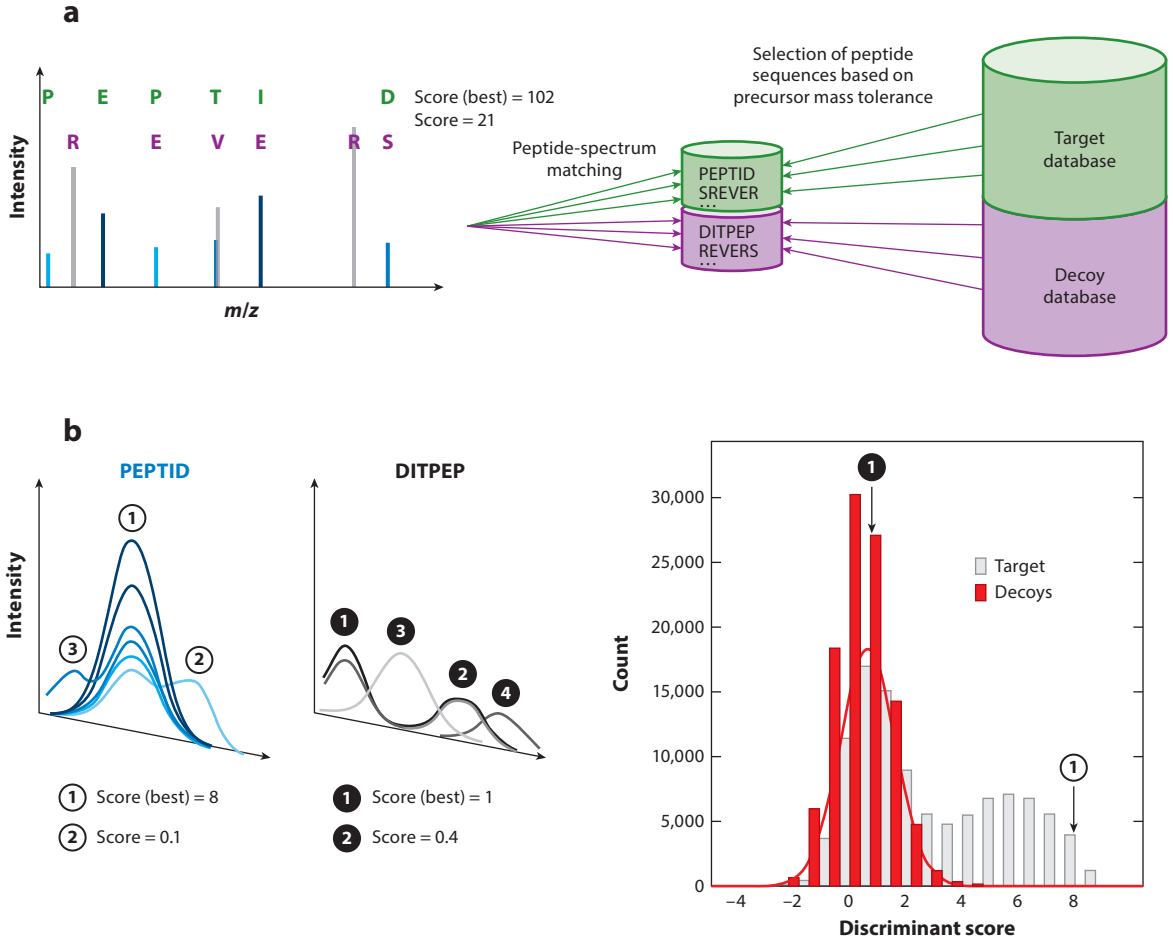
Conversely, the peptide-centric query strategy constitutes a complete change of paradigm for the data analysis. In essence, this strategy starts from a list of hypotheses (Is this peptide present in my data?) and interrogates the acquired data sets for the presence of the peptides of interest. Peptide-centric query strategies have been attempted on individual MS/MS spectra with DDA (46) or DIA data sets (47) but are probably most clearly exemplified by the use of fragment ion chromatographic traces extracted from MS/MS DIA maps (38) (**Figures 3*c*** and **4*b***). For the latter method, the MS/MS signals have to be acquired repeatedly for each precursor, and with a frequency that permits reconstitution of the chromatographic peak shape with sufficient resolution (typically 8–10 MS/MS scans are acquired across a chromatographic peak). The peptide identification can then be inferred by assessing whether a series of chromatographic traces of fragment ions expected for a given peptide (e.g., known from prior experimental observations) do indeed exactly coelute with the expected relative intensities and at the expected retention time (**Figure 4*b***). This scenario is very similar to the way that the presence of a peptide is confirmed in SRM targeted data acquisition. This approach warrants intrinsically consistent identification characteristics for the analysis, and in many ways a more biologically oriented manner of querying LC-MS/MS data structures.

3.4. Quantitative Proteomics

Beyond identification, it is important in most biological studies to know the quantity of a protein (and by proxy, a peptide) present in a sample. In bottom-up proteomics, peptide quantities are typically assessed from the signals of the corresponding precursors, either in their intact form using MS^1 signals or from one or several of their fragments using MS/MS data. Unfortunately, equal amounts of different peptides do not yield equal signals in the mass spectrometer as a result of sequence-dependent differences in ionization efficiencies. Moreover, because multiple peptides are concurrently ionized in LC-MS/MS measurements, peptides that enter the ion source at the same time might mutually influence their ionization efficiencies through a process called ion suppression. To overcome or at least minimize these intrinsic limitations, different quantitative proteomic

workflows have been established to compare peptide levels across samples of interest (48). In one approach, samples are labeled either metabolically or chemically (49) with different stable isotopes or spiked with differentially isotopically labeled forms of the peptides of interest. Because the various isotopomers have negligible differences in ionization and are subject to near-identical ion suppression effects, these labeling strategies provide the highest accuracy for peptide quantification. Assuming that the labeled references are present in the same amount in all the samples, endogenous peptide abundances can be assessed as the ratios of endogenous peptides to their labeled isotopomers. Alternatively, peptide amounts in different samples can also be compared in the absence of dedicated labeled references (label-free quantitation), by directly comparing their MS^1 or MS^2 signals across samples. This requires more sophisticated normalization and chromatographic alignment procedures to compensate for retention time shifts and to avoid mismatching peptide abundances across runs but offers higher versatility when comparing large numbers of samples.

In addition to quantitative accuracy, quantification consistency and completeness are other fundamental aspects to consider. Also, in this case, isotope-labeling strategies provide the highest performance, but the number of isotopomers that can be mixed and measured in a single injection is limited. For label-free approaches or when sample numbers are beyond the limit of multiplexing, consistent peptide identification is the key to complete results, making the choice of data



acquisition method and the query strategy across large sample sets particularly important. Both quantification accuracy and quantification consistency must be taken into consideration depending on the specific aim or stage of a biological project, with an emphasis on accuracy at the discovery phase and consistency at the validation phase (**Figure 1**).

4. DATA-DEPENDENT ACQUISITION (SHOTGUN) WORKFLOWS

The DDA method continues to dominate the proteomics field as the best established strategy to generate information about the composition of a complex (peptide) mixture in a relatively straightforward manner, and on diverse instrument platforms. The motivation to develop such an approach is clear: For a reasonably complex sample, a large number of compounds will be simultaneously ionized as they elute from the LC column, in many cases overwhelming the instrument's sequencing capabilities. Therefore, the capacity to select precursors for fragmentation and acquire MS/MS spectra (**Figure 3*a***) should be directed to the analytes that display the strongest signals and thus have the highest chance to yield successful identification.

The foundations of the DDA method lie in the 1990s, when it became possible to program mass spectrometers to execute specific selections in real time (50). Since then, the capabilities of instrument control software have expanded substantially, most noticeably due to the possibility of determining charge states of ions in real time in high-resolution instruments. Typically, the instrument operator can now select ions for fragmentation based on the abundance rank in the MS^1 spectrum, the absolute signal intensity, and the charge state of the precursor.

Charge state preselection can be used to exclude singly charged ions that are frequently contaminants or, if they are peptides, are more difficult to fragment than their multiply charged counterparts. Similarly, highly charged precursors that yield spectra that are difficult to interpret are also frequently excluded. Furthermore, ions that have been selected for fragmentation can be placed on a dynamic exclusion list to prevent immediate reselection of the same ion, or the sequencing of ions with a particular *m/z* can be generally prevented by inclusion or exclusion

Figure 4

Spectrum-centric (untargeted) and peptide-centric (targeted) query modes of fragment ion (MS/MS) data and related decoy scoring considerations. Two strategies of MS/MS data query are illustrated: (*a*) the spectrum-centric query strategy of data-dependent acquisition and (*b*) the peptide-centric query strategy of data-independent acquisition, using the chromatogram-based dimensionality. With the (*a*) spectrum-centric query approach, each spectrum is assessed independently against a subset of peptide sequences retrieved from the database within the user-defined precursor mass tolerance (e.g., the MS^2 spectrum 2 recorded for the blue peptide of **Figure 3*a*** will trigger a search at its illustrative precursor mass 505.555 ± 0.05 Da, retrieving the amino acid sequences PEPTID and SREVER from the target database and the corresponding reversed amino acid sequences DITPEP and REVERS from the decoy database). In general, only the highest-scoring match (e.g., PEPTID) is reported as peptide identification. With the (*b*) peptide-centric query approach, each precursor of interest from the assay library triggers the extraction of an independent series of fragment ion chromatograms within the expected peptide elution (time) window. The extraction of the fragment ion chromatograms of PEPTID yields three possible peaks within the time window: one peak encompassing the coelution of the six fragment ion traces (①), and two peaks having only one fragment ion trace each (② and ③). Each instance of coeluting traces is scored as a potential identification candidate. In general, only the highest scoring match (here, the six-member fragment ion peak group ①) is reported as peptide identification. To assess the likelihood that six random chromatographic ion traces would coelute by chance in that MS/MS-and-time space, one decoy assay for each peptide assay is extracted. The reversed decoy amino acid sequence DITPEP yields four peak group candidates (❶–❹); the one scoring highest (❶) is reported as the assay identification. The distribution of the highest-scoring target and decoy peak groups (*b*, *right*) then provides an estimate of the false discovery rate, after proper correction for multiple testing and upon rescaling the decoy scoring distribution to that of the low-scoring (false) targets (*red curve*). PEPTID and SREVER are two illustrative amino acid sequences that would be retrieved from the targeted database for a mass of 505.555 ± 0.05 Da; DITPEP and REVERS are the corresponding reversed (decoy) amino acid sequences. Abbreviations: MS or MS^1, precursor mass spectrum; MS/MS or MS^2, fragment ion mass spectrum.

lists. The instrument is able to adjust fragmentation conditions (activation energy or even the fragmentation method) as a function of these parameters (51).

Overall, the most decisive parameter that determines the information content of a shotgun proteomics experiment is the abundance rank. It became apparent early on that the automated precursor selection introduces an element of stochasticity (randomness), because the appearance of a precursor ion spectrum at a given chromatographic retention time will never be absolutely identical between each run, due to variability in the chromatographic separation (52). Therefore, it cannot be guaranteed that the same precursors will be selected for fragmentation, even when the same sample is analyzed repeatedly. The increasing acquisition or sequencing speed of new instruments, now reaching up to 20 Hz in some cases (53, 54) has mitigated this undersampling problem to some extent, but the overlap between replicates still does not reach 100%.

Simply increasing the sequencing speed of instruments alone will not solve the problem. As pointed out several years ago (55), the seemingly straightforward step of precursor isolation causes complications in complex peptide mixtures, such as unfractionated cell lysate, when more than one peptide may frequently be isolated at the same time. Precursor isolation in modern mass spectrometers used for proteomics is performed in either a quadrupole (or similar multipole) device or a linear ion trap. Although the isolation window can be adjusted for both devices, it cannot be set infinitely small, because of isolation efficiency and potential sensitivity issues. In a typical DDA setup, isolation widths may range from 1 to 3 *m/z*, which results in the isolation of the monoisotopic peak of the target ion and (some of) its isotope peaks, but potentially other coeluting peptides with a similar *m/z* (**Figure 3*a***, blue and gray peptides) as well. If such coeluting peptides are isolated and cofragmented, this results in a chimeric spectrum with contributions from several precursors (**Figure 3*a***, MS/MS spectra 1 and 2). This effect can be seen as an involuntary multiplexing, in contrast to the deliberate multiplexing in the DIA methods discussed in Section 6.

The implications of this effect are twofold: First, the identification of peptides based on their MS/MS spectra becomes more difficult because, for example, many database search engines use the number of unassigned fragment ion peaks (which increases for chimeric spectra) for scoring. In current DDA workflows, typically fewer than 50% of all acquired MS/MS spectra are assigned to a peptide sequence. This attrition rate may be caused by such coisolation events but is, of course, also the result of insufficient spectral quality due to low signal intensity and the presence of unexpected modifications. Recently, some free (56, 57) and commercial [e.g., Mascot (58) since version 2.5] search engines have introduced features to identify multiple peptides from a single MS/MS spectrum, although this feature may have implications for error models. At a fundamental level, it challenges the "one MS/MS spectrum = one peptide" dogma of conventional DDA and will require further adaptation of computational workflows.

The second implication concerns quantification methods based on isobaric tags, such as iTRAQ (isobaric tags for relative and absolute quantification) or TMT (tandem mass tags) (59), for which reporter ions observed in the MS/MS spectra are used for relative quantification. In such a case, coisolation of unrelated precursors may result in a distortion of the quantitative ratios because of the contribution of more than one peptide to the reporter ion abundances (60). Because most protein levels are expected to remain unchanged in a typical quantification project, cofragmentation of one peptide that changes in abundance is likely to occur with another peptide that is unchanged under the same conditions, which will result in a dynamic range compression, that is, larger changes will be attenuated. Various approaches to overcome this problem have been proposed, such as extended fractionation (61) or dedicated instrument acquisition modes (62–64), but they come with the penalty of increased measurement times or losses in sensitivity.

Although identification of the same peptides in all runs is not necessarily a prerequisite for protein-level quantitation, because quantitative information of many peptides is combined, the

problem becomes more severe for peptides carrying PTMs or resulting from SNPs. In such cases, the relevant information is obtained at the peptide level and variability cannot be compensated for by the combination of sibling peptides originating from the same protein. For PTMs that can occur on multiple residues within the same peptide, such as phosphorylation, the situation becomes especially complicated because consistency on the level of the sequence identity and the modification site localization (65) is required.

The abovementioned limitations of shotgun MS pose considerable challenges for proteomics applications in medicine and systems biology, where large sample cohorts need to be analyzed qualitatively and quantitatively in a robust manner. To achieve consistency in such large data sets, a further refinement of data acquisition and, particularly, data analysis strategies is required. The emergence of alternatives to the DDA method may provide a preferential solution to these research questions, as discussed in Sections 5 and 6.

5. TARGETED-ACQUISITION WORKFLOWS

In 2012, targeted proteomics was named method of the year by the journal *Nature Methods*. This recognizes several achievements of the technology, including the demonstration of unprecedented quantification accuracy and reproducibility for proteomic samples analyzed by SRM in a cross-laboratory study (66), a result that DDA has not managed to achieve so far (67). As such, targeted data acquisition has indeed become the method of choice for the validation of protein candidates in independent, potentially large sample cohorts that were originally detected by high-throughput, discovery-type DDA analyses (68, 69). To achieve the required accuracy and reproducibility, the mass spectrometer does not select precursors for MS/MS fragmentation in real time during the chromatographic separation. Rather, the instrument is preprogrammed to acquire consistently and repeatedly the same set of fragment ion signals for a predefined list of target peptides (**Figure 3*b***). This systematic acquisition of transitions over the whole peptide elution and the resulting chromatogram-based data structure (**Figure 3*b***) are the key concepts that underlie the consistent peptide quantification across large sample sets in targeted proteomics. The fact that relatively few predetermined peptides are targeted per analysis constitutes the biggest strength (consistent quantification) but also the biggest limitation of the method. First, the design of transition lists requires extensive preliminary work in terms of assay generation before the SRM measurement can even be started (70). Parameters to be optimized include selection of both the most promising peptides per protein and the most intense transitions, optimization of instrument parameters such as collision energy and transmission settings to yield maximum sensitivity, and determination of the chromatographic retention time (71). The latter can be exploited in so-called scheduled SRM methods in which transitions for particular peptides are only monitored close to their expected elution time and not during the whole run, thereby maximizing the list of candidates to be monitored in one analysis (72). The benefit of this elaborate optimization procedure is the excellent sensitivity of SRM, which will likely continue to be the most sensitive acquisition method in the coming years.

During assay generation, one of the most significant (and probably most widely overlooked) problems arises upon deciding which peptides of a particular protein to select to obtain a faithful representation of protein abundance. Although common to all other bottom-up proteomic approaches, this question is particularly critical for targeted acquisition strategies that can typically only monitor a few peptides per LC-MS/MS run. In practice, most targeted acquisition proteomic workflows therefore focus on monitoring only a few proteotypic peptides for the protein of interest. These peptides are usually selected based on empirical rules such as being fully tryptic, doubly charged precursors of medium length that do not contain amino acids prone to modification

Proteotypic peptide: an amino acid sequence originating from one unique protein form during proteolytic digestion of a biological sample, confirming the protein's existence and identification

(e.g., oxidation) (73). These criteria, however, do not preclude the possibility that selected protein regions will unexpectedly yield peptides with missed cleavage sites, peptides resulting from unspecific cleavage events, or peptides with unanticipated natural or artifactual modifications. SRM of all possible proteolytic forms of the peptides selected for a given protein has been used in specific biological applications (74), but it simply remains impractical for targeted acquisition workflows on a large scale. In such cases, targeted acquisition should select peptides based on quantotypic criteria rather than proteotypicity alone (where quantotypic qualifies peptides that can faithfully be used for reliable protein quantification purposes) (13).

The two-stage filtering operated on the precursor and fragment ion levels by the first and third quadrupole, respectively (in the widely used triple quadrupole mass spectrometer), is often presented as granting a very high level of specificity for the transition measurements in SRM acquisition. However, in complex sample matrices, interfering transition signals may still be observed at relatively high frequency (blue peptide in **Figure 3*b***). As mentioned above, quadrupoles usually cannot transmit ions with isolation widths below 0.7–1 Da without losing sensitivity. Recently, new acquisition strategies have replaced the third quadrupole with a high-resolution accurate mass analyzer (either Orbitrap or TOF) to improve mass resolution and accuracy for fragment ion detection, while retaining the targeted acquisition concept. These methods (PRM or MS/MS^{ALL}) also allow users to record time-resolved MS/MS signals of predefined sets of peptides. Contrary to SRM, however, the methods record complete MS/MS spectra of the precursors that are monitored consistently over the elution profile of the peptide. In practical terms, the width of the precursor isolation window is the major difference between these new methods and the DIA acquisition scheme presented in Section 6. However, even in recently published applications of PRM, interferences are still observed despite the high-resolution accurate mass detection of fragment ions. Analogous to the MS/MS specificity frontier faced by DDA in complex sample matrices, the faster and more sensitive instrumentation of SRM, PRM, and related methods will challenge the "one assay = one peptide" dogma of targeted data acquisition by producing ultimately more and more multiplexed MS/MS data signals.

For all these methods and in the large majority of cases, peptide identification is still carried out by visual inspection of the SRM/PRM data using tools such as Skyline (75). Tools have also been developed to automate chromatogram peak scoring of fragment ion groups in a high-throughput manner when manual inspection becomes impractical, such as beyond dozens or hundreds of runs (76). The high-throughput identification of the targeted peptides and their confidence assessment using decoys and FDR estimation are discussed more extensively in the context of the peptide-centric, chromatogram-based analysis of DIA data sets in Section 6.

In summary, although targeted acquisition strategies can now offer the peptide identification and quantification consistency that the DDA mode was lacking, the relatively small number of peptides that can be monitored per run has encouraged the development of other strategies that would combine the consistency in quantification of SRM with the throughput of DDA identification.

6. DATA-INDEPENDENT ACQUISITION WORKFLOWS

By definition, DIA aims at acquiring the time-resolved complete MS/MS sequencing data for all possible mass ranges, independently of the presumption or detection of precursors. Because of the undersampling limitations of DDA discussed above and the limited number of peptides that can be monitored by SRM, a vast majority of peptides from biological samples may still evade MS/MS acquisition and identification, even on the fastest instrumentation in either mode. In that sense, DIA is the only acquisition method that can be considered to generate permanent digital maps or MS/MS records of all the analytes in a biological sample that are above the detection

limit of the respective instrument. To achieve this performance, however, DIA requires opening up the precursor isolation windows to yield a cycle time (the time required for the instrument to cover the whole intended mass range with adjacent precursor ion selection windows) suitable to provide sufficient MS/MS data points in the time domain (e.g., with eight sampling points per chromatographic peak). Increasing the width of the precursor isolation window is a fundamental violation of the "one MS/MS spectrum = one peptide" dogma of DDA and SRM. Therefore, to ensure the meaningful analysis of DIA data sets, many concepts originally endorsed by the proteomics community needed to be reconsidered.

Early implementations of DIA were proposed several years ago (41, 42, 77–79) when fast-scanning instruments capable of delivering fragment ion data with high mass accuracy were not readily available. High-accuracy fragment ion data are important for DIA methods because they greatly improve the specificity for deconvoluting highly multiplexed fragment ion spectra (38, 47). Another equally instrumental factor for the successful implementation of the method was the introduction of rectangular precursor isolation windows (38). In DDA and SRM, a Gaussian isolation function centered on the precursor is beneficial for improving the specificity of the precursor selection, but box-shaped isolation windows are preferred in DIA to limit the quantitative attrition in precursor filtering at the edges of the isolation windows.

Compared with DDA and targeted methods, improvements in instrumentation regarding scan speed (to cycle faster through smaller isolation windows) and dynamic range (to detect signals of cofragmenting peptides of vastly different intensity) will likely have the highest impact for DIA. It is foreseeable that at some point, DIA isolation widths may become as small as a few daltons, which will bring DIA in the same precursor selection and fragment ion specificity range as DDA, SRM, or PRM, but without missed precursors. Eventually, this will result in the ultimate convergence of the DDA, targeted, and DIA acquisition modes, and eventually only the data dimensionality and the data query strategy that is applied (spectrum-centric or peptide-centric) will make a difference, because all MS/MS spectra will have to be considered as chimeric or multiplexed at some level anyway. And, because complex samples will generate a high fraction of multiplexed peptides even at that stage of instrument performance, it can be expected that such data sets will be most efficiently analyzed by peptide-centric querying strategies.

Despite the highly convoluted multiplexed nature of the MS/MS spectra acquired in DIA mode, the data sets were initially analyzed via spectrum-centric approaches using database searching or spectral library matching strategies (41–43, 80) (**Table 1**). Taking advantage of the continuous-in-time dimensionality of these MS/MS data sets and clustering fragment and precursor ion signals based on their chromatographic peak shapes, the first breakthrough in DIA analysis was achieved (44, 45) (**Table 1**). Although producing highly preprocessed, demultiplexed pseudo-MS/MS spectra, the analysis was still performed in a spectrum-centric manner using database searching. In 2012, the Aebersold group (38) introduced targeted data extraction as a prototypic peptide–centric, chromatogram-based strategy to analyze DIA data sets [sequential windowed acquisition of all theoretical mass spectra (SWATH) MS] (**Table 1**). In short, the approach relies on libraries of assays based on bona fide MS/MS spectra that uniquely identify the query peptide (81) to extract fragment ion traces for precursors of interest in the DIA data sets, and on assessing the quality of their coelution to infer the peptide identification. This reduces the highly convoluted DIA MS/MS maps to a SRM-like data structure (**Figure 3*c***) with essentially the same specificity and reproducibility characteristics (38), but with the added advantage that basically any peptide of interest can be queried in a peptide-centric manner ad libitum from the DIA data sets after the data have been acquired.

It is important to emphasize that the SWATH MS peptide-centric approach differs significantly from the DDA spectrum-centric query strategies in several aspects (26). First, the knowledge of

the accurate mass of the precursor becomes almost irrelevant for the method and is only used to determine the swath of isolation windows from which fragments are extracted. Second, the peptide identification does not rely on MS/MS-database matches but is solely achieved by assessing the chromatographic characteristics of the extracted fragment ion signals (coelution, peak shape, relative intensity correlation with the assay library or spiked-in isotopically labeled standard), as in SRM, plus a few other fragment ion scores exclusive to the MS/MS-continuous nature of DIA (monoisotopic fragment *m/z*, charge state, mass accuracy). In other words, it does not matter if several precursors coelute and cofragment in a large isolation window into chimeric MS/MS spectra. As long as each query precursor has a unique set of fragment ions near that retention time, the qualitative assessment of their coeluting fragment ion chromatographic profiles is sufficient for their identification.

Although visual inspection can still be used for peptide identification upon SWATH MS peptide-centric extraction, the large numbers of assays that can be queried from a single data file (several tens of thousands at present) are best analyzed using automated algorithms that perform fragment ion chromatogram peak scoring and assess the confidence of the identification in an objective and high-throughput manner. Retrospectively, the mProphet algorithm developed for the high-throughput analysis of large SRM data sets (76) can be seen as the first purely chromatogram-based, peptide-centric identification tool to have paved the way for the DIA targeted extraction strategy. Since then, other automated tools have been developed (37, 82, 83) that show impressive levels of agreement in terms of peptide identification and quantification (P. Navarro, J. Kuharev, L.C. Gillet, O.M. Bernhardt, B. MacLean, H.L. Röst, S.A. Tate, C. Tsou, L. Reiter, G. Rosenberger, Y. Perez-Riverol, A.I. Nesvizhskii, R. Aebersold, S. Tenzer, manuscript submitted).

As high-throughput, peptide-centric data query strategies are relatively new, we reemphasize here some key differences as compared with shotgun database search strategies, especially regarding the use of decoys for FDR estimation in SWATH MS DIA peptide-centric extraction mode (**Figure 4**). In essence, decoys are nonnatural peptides generated, for example, by reversing the amino acid sequence of all entries in a sequence database. During the identification step, the search tools are agnostic about whether they are testing a true or a decoy peptide sequence. Decoys were originally introduced in bottom-up proteomics because the raw scores of peptide-spectra matches to genuine (target) protein sequences alone were not sufficient to control confidence in peptide identification during database searches (84).

In spectrum-centric approaches, the peptide confidence is derived from scoring the quality of the match between a measured MS/MS spectrum (experimental data) and individual corresponding, theoretically computed target or decoy spectra (**Figure 4*a***). Because the search engines are agnostic, whether they match the experimental spectrum using an actual target or a computed decoy spectrum, the frequency with which a decoy spectrum is assigned under specific search conditions is used to calculate the FDR of a particular data set. Importantly, with the spectrum-centric strategy, the number of retrieved forward and reverse database sequences for a MS/MS spectrum match is always equal for each precursor-MS/MS spectrum searched and increases with the size of the database and the search space (affected by the choice of search parameters such as variable modifications or precursor mass tolerance), but not with the number of MS/MS spectra that are searched. The decoys thus typically provide an estimate of the quality of a spectrum or the chance of a random match between a measured MS/MS spectrum and an in silico database-derived spectrum within the defined search space (**Figure 4*a***).

In contrast, the peptide-centric approach assesses the likelihood that a particular peptide is present in the sample, for example, by identifying and scoring the chromatographic peak that is most representative of the peptide fragment assay used for the targeted extraction. As with

spectrum-centric strategies, the score of the raw chromatogram peak group is not sufficient to define a threshold to assign a given FDR to peptide identifications. To achieve this, nonnatural decoy assays are generated (e.g., by reversing the amino acid sequence of natural peptides) and processed, in our case, by targeted extraction and chromatographic peak group scoring exactly like the target assays in the same data file (**Figure 4*b***). Because it is not expected that all transitions of a decoy assay coexist in an actual sample, the decoys provide an estimate of the local sample background complexity, that is, of the likelihood of getting coeluting fragment ion traces by chance in a given matrix. Importantly, with this approach the number of decoys increases with the number of tests/targets queried. Therefore, with peptide-centric approaches, it is not statistically correct to estimate the FDR in targeted extraction approaches by naive counting of decoys as in the spectrum-centric strategy, unless appropriate corrections for multiple testing or proper decoy distribution scaling are made (**Figure 4**). This can be done by using the Storey & Tibshirani (85) approach, as was suggested in the original mProphet paper (76), or by correcting for the fraction of false targets, similar to correcting for the percentage of incorrect targets as described by Käll et al. (86) for DDA data (**Figure 4*b***).

From the quantification perspective, peptide-centric chromatographic extraction of DIA data sets was shown to achieve peptide quantification consistency and accuracy similar to that of SRM, intermediary sensitivity between SRM and MS^1 label-free quantification, and 4 logs of intrascan dynamic range (38). This enables the consistent identification and quantification of several thousands of proteins in complex samples (87, 88). Because it does not suffer from the undersampling issues of DDA or SRM, the method enables production of quantification matrices for a large number of proteins and across a large number of samples with minimal missing values (**Figure 1**), enabling experimental biology to interrogate big data by testing relevant hypotheses with unprecedented accuracy.

Lastly, it should be noted that the large-scale, peptide-centric query of DIA data is a quite recent strategy compared with DDA spectrum-centric database searching, which has evolved over more than 20 years. The strategy is now facilitated by optimized protocols for spectral library generation (81), the availability of extensive organism-specific assay libraries (87, 89, 90), and automated data extraction and scoring tools (37, 82, 83).

7. CONCLUDING REMARKS

The past few years have seen several transitioning phases for MS-based proteomics. First, with improved sequencing speed and mass accuracy of recent instrument generations, DDA has emerged as a rapid and relatively straightforward method to generate large lists of protein identifications. Second, with the advent of SRM targeted acquisition, proteomics has gained visibility with regard to quantification consistency and reproducibility. Now, with the combination of near-complete MS/MS maps acquired by DIA methods together with peptide-centric data query strategies, the field has the opportunity to bridge the gap between the number of protein identifications and the consistency and accuracy of those quantifications. By producing highly consistent and near-complete protein quantification matrices, the technology can now enable experimental biology to explore the depth of the proteome with unprecedented consistency. As commented above, it is probably in DIA mode that improvements in instrumentation will have the highest immediate impact, and in the foreseeable future, it is likely that high-resolution accurate mass MS/MS maps will be acquired with small precursor isolation windows in DIA mode. The time- and mass-continuous data structure of the MS/MS maps from DIA data sets will then offer unprecedented opportunities to query the data in ways that were just not possible before. For example, moving away from the spectrum-centric and database-searching dogma, the biologist can now query for the existence

of pretty much any possible peptide of interest (including splice forms, SNP variants, or those carrying a PTM) in DIA maps. In the context of personalized medicine where protein database information will remain de facto incomplete, this opportunity will potentially expand the identification of new personalized markers beyond what was possible using current standard methods. Combined with the quantification consistency offered by the DIA peptide-centric strategy, this will hopefully open MS-based proteomics to a new era and may ultimately reconcile big data and hypothesis-testing biology.

DISCLOSURE STATEMENT

The research group of R.A. is supported by Sciex by allowing access to prototype instrumentation.

ACKNOWLEDGMENTS

We thank Michael MacCoss, Ying Sonia Ting, and Jarrett D. Egertson (University of Washington, Seattle) for discussions and clarifications regarding the spectrum- and peptide-centric terminology. We thank George Rosenberger, Ben Collins, and Moritz Heusel (ETH Zürich, Switzerland) for general discussions and comments. This work was supported by ETH Zürich and European Union (EU) grants (European Research Council advanced grant 233226 "Proteomics v3.0" and EU grant 262067 "PrimeXS") to R.A.

LITERATURE CITED

1. Nesvizhskii AI, Aebersold R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteom.* 4:1419–40
2. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929–34
3. Anderson C. 2008. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, June 23. **http://www.wired.com/print/science/discoveries/magazine/16-07/pb_theory**
4. Pigliucci M. 2009. The end of theory in science? *EMBO Rep.* 10:534
5. Mazzocchi F. 2015. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Rep.* 16:1250–55
6. Smith LM, Kelleher NL, Consort. Top Down Proteom. 2013. Proteoform: a single term describing protein complexity. *Nat. Methods* 10:186–87
7. Catherman AD, Skinner OS, Kelleher NL. 2014. Top down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* 445:683–93
8. Toby TK, Fornelli L, Kelleher NL. 2016. Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* 9:499–519
9. Walzthoeni T, Leitner A, Stengel F, Aebersold R. 2013. Mass spectrometry supported determination of protein complex structure. *Curr. Opin. Struct. Biol.* 23:252–60
10. Larance M, Lamond AI. 2015. Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* 16:269–80
11. Choudhary C, Mann M. 2010. Decoding signalling networks by mass spectrometry–based proteomics. *Nat. Rev. Mol. Cell Biol.* 11:427–39
12. Bensimon A, Heck AJ, Aebersold R. 2012. Mass spectrometry–based proteomics and network biology. *Annu. Rev. Biochem.* 81:379–405
13. Worboys JD, Sinclair J, Yuan Y, Jorgensen C. 2014. Systematic evaluation of quantotypic peptides for targeted analysis of the human kinome. *Nat. Methods* 11:1041–44
14. Chang CY, Picotti P, Huttenhain R, Heinzelmann-Schwarz V, Jovanovic M, et al. 2012. Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell Proteom.* 11:M111.014662

15. Teo G, Kim S, Tsou CC, Collins B, Gingras AC, et al. 2015. mapDIA: preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J. Proteom.* 129:108–20
16. Oberg AL, Vitek O. 2009. Statistical design of quantitative mass spectrometry–based proteomic experiments. *J. Proteome Res.* 8:2144–56
17. Vandermarliere E, Mueller M, Martens L. 2013. Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom. Rev.* 32:453–65
18. Tsiatsiani L, Heck AJ. 2015. Proteomics beyond trypsin. *FEBS J.* 282:2612–26
19. Claassen M, Reiter L, Hengartner MO, Buhmann JM, Aebersold R. 2012. Generic comparison of protein inference engines. *Mol. Cell Proteom.* 11:O110.007088
20. Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, et al. 2012. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell Proteom.* 11:M111.013722
21. Han J, Ye L, Xu L, Zhou Z, Gao F, et al. 2014. Towards high peak capacity separations in normal pressure nanoflow liquid chromatography using meter long packed capillary columns. *Anal. Chim. Acta* 852:267–73
22. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR 3rd. 2013. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* 113:2343–94
23. Eliuk S, Makarov A. 2015. Evolution of Orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem.* 8:61–80
24. Andrews GL, Simons BL, Young JB, Hawkridge AM, Muddiman DC. 2011. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Anal. Chem.* 83:5442–46
25. Beck S, Michalski A, Raether O, Lubeck M, Kaspar S, et al. 2015. The Impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol. Cell Proteom.* 14:2014–29
26. Ting YS, Egertson JD, Payne SH, Kim S, MacLean B, et al. 2015. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol. Cell Proteom.* 14:2301–7
27. Picotti P, Aebersold R. 2012. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* 9:555–66
28. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. 2012. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell Proteom.* 11:1475–88
29. Chapman JD, Goodlett DR, Masselon CD. 2013. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* 33:452–70
30. Hoopmann MR, Finney GL, MacCoss MJ. 2007. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* 79:5620–32
31. 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* 7:3470–80
32. Weisser H, Nahnsen S, Grossmann J, Nilse L, Quandt A, et al. 2013. An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res.* 12:1628–44
33. 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.* 13:2513–26
34. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, et al. 2002. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2:513–23
35. Zimmer JS, Monroe ME, Qian WJ, Smith RD. 2006. Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.* 25:450–82
36. Stanley JR, Adkins JN, Slysz GW, Monroe ME, Purvine SO, et al. 2011. A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics. *Anal. Chem.* 83:6135–40
37. Tsou CC, Avtonomov D, Larsen B, Tucholska M, Choi H, et al. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* 12:258–64

38. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, et al. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteom.* 11:O111.016717
39. Eng JK, Searle BC, Clauser KR, Tabb DL. 2011. A face in the crowd: recognizing peptides through database search. *Mol. Cell Proteom.* 10:R111.009522
40. Lam H. 2011. Building and searching tandem mass spectral libraries for peptide identification. *Mol. Cell Proteom.* 10:R111.008565
41. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* 1:39–45
42. Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, et al. 2009. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* 81:6481–88
43. Carvalho PC, Han X, Xu T, Cociorva D, da Gloria Carvalho M, et al. 2010. XDIA: improving on the label-free data-independent analysis. *Bioinformatics* 26:847–48
44. Geromanos SJ, Vissers JP, Silva JC, Dorschel CA, Li GZ, et al. 2009. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* 9:1683–95
45. Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ. 2010. Deconvolution of mixture spectra from ion-trap data-independent–acquisition tandem mass spectrometry. *Anal. Chem.* 82:833–41
46. Liebler DC, Hansen BT, Davey SW, Tiscareno L, Mason DE. 2002. Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Anal. Chem.* 74:203–10
47. Weisbrod CR, Eng JK, Hoopmann MR, Baker T, Bruce JE. 2012. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J. Proteome Res.* 11:1621–32
48. Aebersold R, Mann M. 2003. Mass spectrometry–based proteomics. *Nature* 422:198–207
49. Gingras AC, Gstaiger M, Raught B, Aebersold R. 2007. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 8:645–54
50. Stahl DC, Swiderek KM, Davis MT, Lee TD. 1996. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J. Am. Soc. Mass Spectrom.* 7:532–40
51. Swaney DL, McAlister GC, Coon JJ. 2008. Decision tree–driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* 5:959–64
52. Liu H, Sadygov RG, Yates JR 3rd. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76:4193–201
53. Kelstrup CD, Jersie-Christensen RR, Batth TS, Arrey TN, Kuehn A, et al. 2014. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J. Proteome Res.* 13:6187–95
54. Scheltema RA, Hauschild JP, Lange O, Hornburg D, Denisov E, et al. 2014. The Q exactive HF, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell Proteom.* 13:3698–708
55. 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.* 10:1785–93
56. Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. 2005. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* 5:4096–106
57. 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.* 10:1794–805
58. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–67
59. Rauniyar N, Yates JR 3rd. 2014. Isobaric labeling–based relative quantification in shotgun proteomics. *J. Proteome Res.* 13:5293–309
60. 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.* 8:5347–55

61. Ow SY, Salim M, Noirel J, Evans C, Wright PC. 2011. Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. *Proteomics* 11:2341–46
62. 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* 8:933–35
63. Ting L, Rad R, Gygi SP, Haas W. 2011. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* 8:937–40
64. 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.* 86:7150–58
65. Chalkley RJ, Clauser KR. 2012. Modification site localization scoring: strategies and performance. *Mol. Cell Proteom.* 11:3–14
66. Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, et al. 2009. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring–based measurements of proteins in plasma. *Nat. Biotechnol.* 27:633–41
67. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, et al. 2009. A HUPO test sample study reveals common problems in mass spectrometry–based proteomics. *Nat. Methods* 6:423–30
68. Surinova S, Radova L, Choi M, Srovnal J, Brenner H, et al. 2015. Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. *EMBO Mol. Med.* 7:1153–65
69. Surinova S, Choi M, Tao S, Schuffler PJ, Chang CY, et al. 2015. Prediction of colorectal cancer diagnosis based on circulating plasma proteins. *EMBO Mol. Med.* 7:1166–78
70. Carr SA, Abbatiello SE, Ackermann BL, Borchers C, Domon B, et al. 2014. Targeted peptide measurements in biology and medicine: best practices for mass spectrometry–based assay development using a fit-for-purpose approach. *Mol. Cell Proteom.* 13:907–17
71. Lange V, Picotti P, Domon B, Aebersold R. 2008. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* 4:222
72. Stahl-Zeng J, Lange V, Ossola R, Eckhardt K, Krek W, et al. 2007. High sensitivity detection of plasma proteins by multiple reaction monitoring of *N*-glycosites. *Mol. Cell Proteom.* 6:1809–17
73. Ludwig C, Aebersold R. 2014. Getting absolute: determining absolute protein quantities via selected reaction monitoring mass spectrometry. In *Quantitative Proteomics*, ed. CE Eyers, SJ Gaskell, pp. 80–109. Cambridge, UK: R. Soc. Chem.
74. Feng Y, De Franceschi G, Kahraman A, Soste M, Melnik A, et al. 2014. Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* 32:1036–44
75. 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* 26:966–68
76. Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, et al. 2011. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* 8:430–35
77. Purvine S, Eppel JT, Yi EC, Goodlett DR. 2003. Shotgun collision–induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* 3:847–50
78. Plumb RS, Johnson KA, Rainville P, Smith BW, Wilson ID, et al. 2006. UPLC/MS^{E}; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun. Mass Spectrom.* 20:1989–94
79. Geiger T, Cox J, Mann M. 2010. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell Proteom.* 9:2252–61
80. Panchaud A, Jung S, Shaffer SA, Aitchison JD, Goodlett DR. 2011. Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal. Chem.* 83:2250–57
81. Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, et al. 2015. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* 10:426–41
82. Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, et al. 2014. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* 32:219–23
83. Bruderer R, Bernhardt OM, Gandhi T, Miladinovic SM, Cheng LY, et al. 2015. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell Proteom.* 14:1400–10

84. Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4:207–14
85. Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *PNAS* 100:9440–45
86. Käll L, Storey JD, MacCoss MJ, Noble WS. 2008. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* 7:29–34
87. Selevsek N, Chang CY, Gillet LC, Navarro P, Bernhardt OM, et al. 2015. Reproducible and consistent quantification of the *Saccharomyces cerevisiae* proteome by SWATH-MS. *Mol. Cell Proteom.* 14:739–49
88. Huang Q, Yang L, Luo J, Guo L, Wang Z, et al. 2015. SWATH enables precise label-free quantification on proteome scale. *Proteomics* 15:1215–23
89. Rosenberger G, Koh CC, Guo T, Röst HL, Kouvonen P, et al. 2014. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* 1:140031
90. Schubert OT, Ludwig C, Kogadeeva M, Zimmermann M, Rosenberger G, et al. 2015. Absolute proteome composition and dynamics during dormancy and resuscitation of *Mycobacterium tuberculosis*. *Cell Host Microbe* 18:96–108
91. Nesvizhskii A. 2007. Protein identification by tandem mass spectrometry and sequence database searching. In *Mass Spectrometry Data Analysis in Proteomics*, ed. R Matthiesen, pp. 87–119. New York: Humana
92. Frank AM, Monroe ME, Shah AR, Carver JJ, Bandeira N, et al. 2011. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* 8:587–91
93. Kim S, Gupta N, Bandeira N, Pevzner PA. 2009. Spectral dictionaries: integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell Proteom.* 8:53–69
94. Allmer J. 2011. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev. Proteom.* 8:645–57

Annual Review of
Analytical Chemistry

Volume 9, 2016

Contents