Mass Spectrometric Analysis of Histone Proteoforms

Zuo-Fei Yuan,[1] Anna M. Arnaudo,[1,2]
and Benjamin A. Garcia[1]

[1]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: bgarci@mail.med.upenn.edu

[2]Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544

Annu. Rev. Anal. Chem. 2014. 7:113–28

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

This article's doi:
10.1146/annurev-anchem-071213-015959

Keywords

histone proteoforms, posttranslational modifications, variants, mass spectrometry, identification, quantification

Abstract

Histones play important roles in chromatin, in the forms of various posttranslational modifications (PTMs) and sequence variants, which are called histone proteoforms. Investigating modifications and variants is an ongoing challenge. Previous methods are based on antibodies, and because they usually detect only one modification at a time, they are not suitable for studying the various combinations of modifications on histones. Fortunately, mass spectrometry (MS) has emerged as a high-throughput technology for histone analysis and does not require prior knowledge about any modifications. From the data generated by mass spectrometers, both identification and quantification of modifications, as well as variants, can be obtained easily. On the basis of this information, the functions of histones in various cellular contexts can be revealed. Therefore, MS continues to play an important role in the study of histone proteoforms. In this review, we discuss the analysis strategies of MS, their applications on histones, and some key remaining challenges.

1. INTRODUCTION

Histones play important roles in chromatin, in the forms of the number of different histone proteoforms, e.g., various posttranslational modifications (PTMs) and sequence variants (1). First, there are numerous PTMs on histones, which include methylation, acetylation, phosphorylation, ubiquitination, and SUMOylation, etc. Each PTM is related to many distinct protein functions. Moreover, some PTMs have cross talk (i.e. combination of modifications) with one another and function synergistically to regulate gene expression. Reference 2 provides examples of histone PTMs.

Histones have five families: H1, H2A, H2B, H3, and H4. Each family has a canonical sequence and different sequence variants. The H1 variants include H1.0–H1.5, H1.t, H1.x, HILS1, H1oo, etc. The H2A variants include H2A.J, H2A.V, H2A.X, H2A.Z, H2A.Bbd, macroH2A, etc. The H2B variants include H2B1A, H2B1B, H2B1C, etc. The H3 variants include H3.1–H3.3, H3.1t, CENP-A, etc. Reference 3 provides examples of histone variants. The diversity of histone proteoforms makes them a challenge to identify and characterize.

Traditionally, antibody-based methods (e.g., western blot) are used to analyze histone modifications (4). These methods have multiple disadvantages. First, antibodies are not available for every new PTM discovered. Second, PTMs on neighboring amino acids may prevent antibody binding, a phenomenon called epitope occlusion. Third, the quantification of PTMs via antibody-based methods is inaccurate at best. Fortunately, all these disadvantages can be overcome using mass spectrometry (MS). MS is a sensitive and efficient way to detect both previously identified and novel PTMs. Moreover, there are various MS-based methods to accurately quantify PTMs. MS methods also allow for identification and quantification of histone variants, which may be too similar in sequence to study using antibodies. Thus, MS is the key technology to analyze histone proteoforms. MS applications on histone proteoforms can be found in the review (5).

Although MS is an important technology, it still faces some challenges. In this review, we cover the fundamentals of mass spectrometers, three MS strategies (i.e., bottom-up, top-down, and middle-down) for studying histones, and some remaining challenges of MS.

2. MASS SPECTROMETRY FOR HISTONE ANALYSIS

MS emerged more than a century ago and its application to biology, especially proteins, started as far back as 1958 (6). Since then, many methods have been developed to analyze proteins, including improvements in sample preparation, ionization, fragmentation, and detection. In this section, the fundamental methods and three strategies (i.e., bottom-up, top-down, middle-down) of MS are introduced.

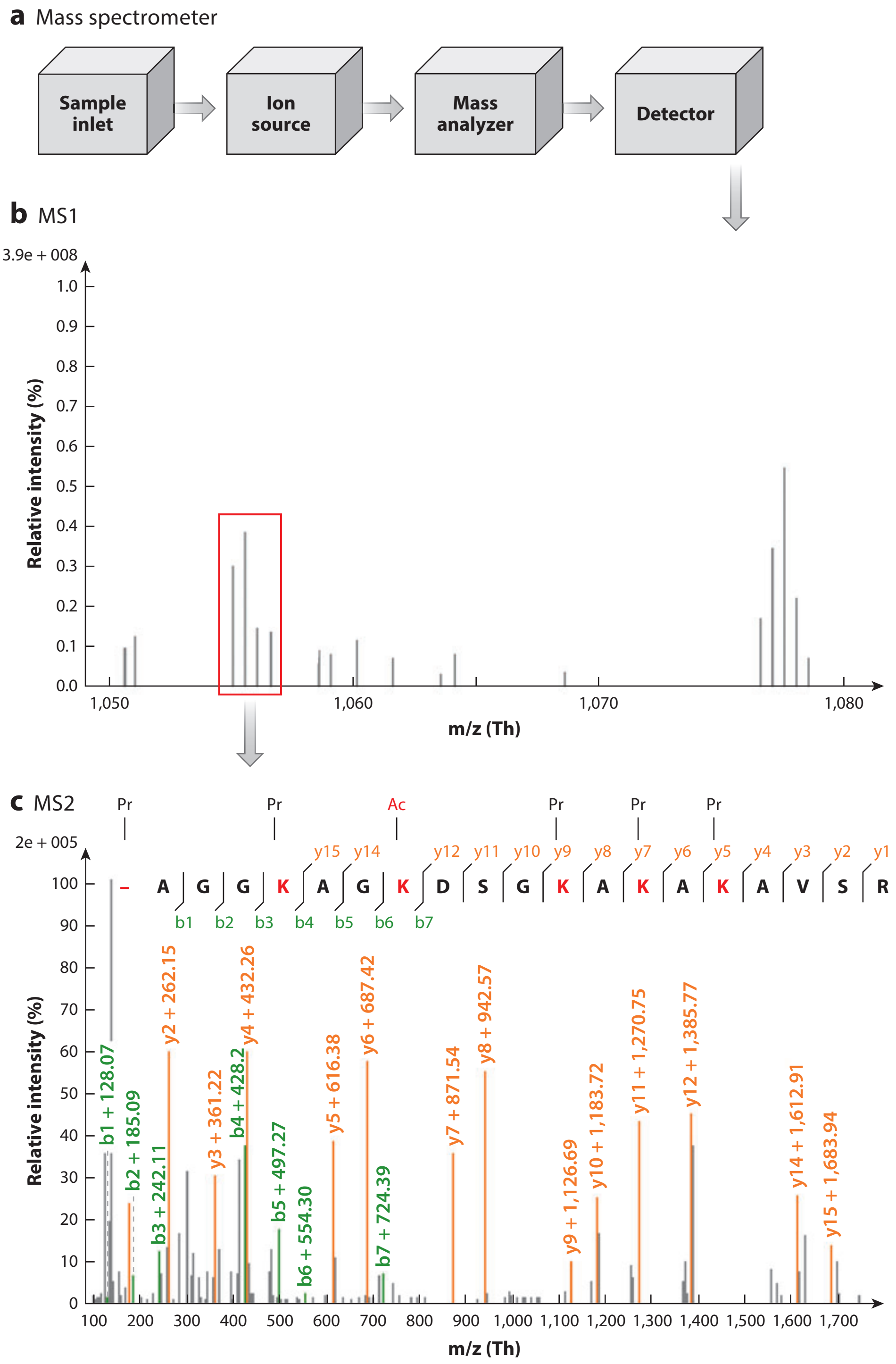
2.1. Fundamentals of Mass Spectrometry

A typical mass spectrometer consists of four components: a sample inlet, an ion source, a mass analyzer, and a detector (7). **Figure 1*a*** shows a layout for these components. Samples undergoing mass spectrometric analysis go through numerous steps. First, they have to be introduced into the

⟶

Figure 1

Fundamentals of mass spectrometry. (*a*) A mass spectrometer consists of four components: a sample inlet, an ion source, a mass analyzer, and a detector. (*b*) Precursor ions are scanned in MS1. (*c*) Some precursor ions are selected, fragmented, and scanned in MS2. Abbreviations: Ac, Acetylation; MS1 and MS2, first-level mass spectrum and second-level mass spectrum (respectively); Pr, propionylation.


a Mass spectrometer
Sample inlet
Ion source
Mass analyzer
Detector
b MS1
3.9e + 008
Relative intensity (%)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1,050
1,060
1,070
1,080
m/z (Th)
c MS2
2e + 005
Pr
Pr
Ac
Pr
Pr
Pr
y15 y14 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y2 y1
– A G G K A G K D S G K A K A K A V S R
b1 b2 b3 b4 b5 b6 b7
Relative intensity (%)
100
90
80
70
60
50
40
30
20
10
0
b1 + 128.07
b2 + 185.09
y2 + 262.15
b3 + 242.11
y3 + 361.22
b4 + 428.2
y4 + 432.26
b5 + 497.27
b6 + 554.30
y5 + 616.38
y6 + 687.42
b7 + 724.39
y7 + 871.54
y8 + 942.57
y9 + 1,126.69
y10 + 1,183.72
y11 + 1,270.75
y12 + 1,385.77
y14 + 1,612.91
y15 + 1,683.94
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300
1,400
1,500
1,600
1,700
m/z (Th)

instrument. They can be eluted through liquid chromatography (LC) into the mass spectrometer or embedded in a matrix on a target plate. Then the ion source converts sample molecules to ions, using electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (8, 9). In the magnetic or electric field of the mass analyzer, ions can fly with different rates or circulate with different frequencies depending on their mass-to-charge ratio (m/z). When ions fly to a detector or circulate with stable frequency, detection of the ions and digitization of the resultant signal yields a first-level mass spectrum (MS1), which contains m/z values and intensities for sample ions (commonly called the precursor ions). **Figure 1*b*** shows an example of MS1.

However, MS1 is not enough to distinguish some precursor ions. For example, PEPTIDE and PEPDITE have the same m/z values. To distinguish these precursor ions, we have to fragment them and obtain m/z values for their components. For example, PEPT and PEPD have different m/z values. Therefore, the second level of mass spectrum (MS2) is essential if other information is lacking (e.g., retention time in LC). To obtain an MS2, a precursor ion is selected from the MS1, isolated, broken into fragment ions, analyzed, and detected. **Figure 1*c*** shows an example of MS2.

To understand the complexity of the MS2 spectra, it is necessary to understand what those peaks represent. Although a single precursor mass is selected, there are many copies of the precursor ion available for fragmentation (shown in **Figure 1*b***). Due to different fragmentation efficiencies, different copies of the precursor ion are fragmented at each site along the amino acid sequence. Thus, the types and intensities of fragment ions are different (shown in **Figure 1*c***). Additionally, one sequence can be fragmented into N-terminal and C-terminal parts. For example, -NH-CRH-CO- is the backbone of one residue (C, H, N, O are the basic elements, with R representing a specific residue), and the linear connection of residues consists of a sequence. Breaking the bond -CO-NH-, collision-induced dissociation (CID) produces b and y ions from the N-terminal and C-terminal parts, respectively; breaking the bond -NH-CRH-, electron transfer dissociation (ETD) and electron capture dissociation (ECD) produce c and z ions from the N-terminal and C-terminal parts, respectively (10, 11). The end result is an MS2 that contains the m/z values and intensities for different N- and C-terminal fragments of the precursor ion.

From the peaks in an MS2, the amino acid sequence can be obtained. The mass difference of two adjacent peaks is equal to the mass of the amino acid that separates them (e.g., y_7 and y_8 in **Figure 1*c***). If all the fragment ions from one terminus are detected in the MS2, the sequence can be inferred by their mass differences. This is the basis for de novo sequencing, which infers the sequence from the MS2 directly (12–14). However, typically some fragment ions are missing or buried in the noise peaks due to lower fragmentation efficiencies at those positions. In this case, only parts of the sequence can be obtained. Moreover, if the N- and C-terminal fragment ions are missing (i.e., b1, y1, c1, and z1), the sequence terminus cannot be determined. To overcome these problems, database searching methods have been developed. A database search matches peaks from an MS2 with the theoretical fragment ions produced from sequences in a protein database to obtain the most likely sequence (15–19). If sequences in a protein database are correctly annotated, database searching is generally more effective than de novo sequencing.

The presence of PTMs further complicates MS2 analysis. All the fragment ions containing a PTM have a mass shift (equal to the mass of the PTM) compared to the unmodified fragment ions. If the PTM is not specified in the database search, the modified fragment ions and the MS2 cannot be matched correctly. When the PTM is entered into the database, all potential sites are considered (e.g., acetylation can occur on any lysine in the sequence). The best matched result from such a search gives the peptide sequence and assigns the PTM to a single amino acid in that sequence. **Figure 1*c*** shows a spectrum for a histone peptide containing an acetylation site.

However, the types or sites of PTMs are usually unknown. Fortunately, there are two methods that can resolve this problem. The first method uses the mass difference of the modified and

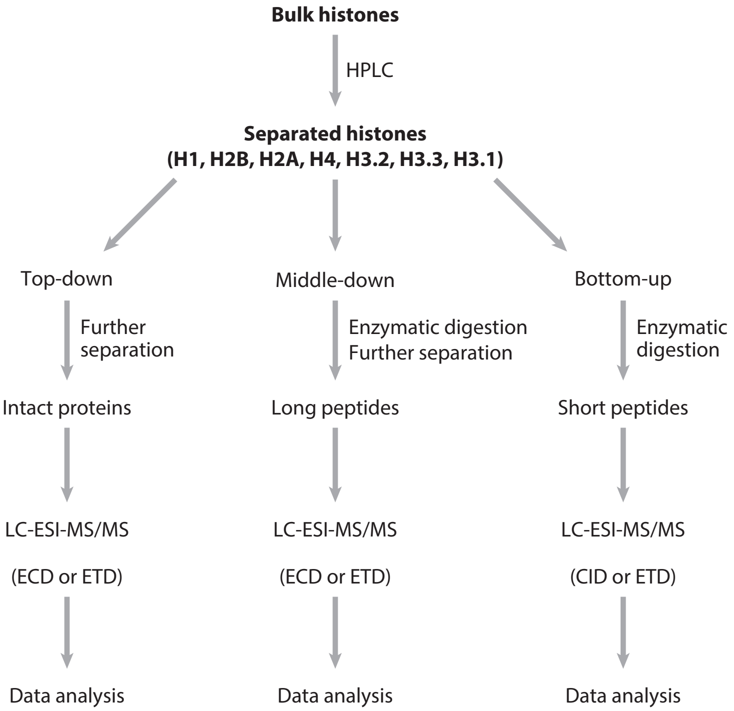


Figure 2

Mass spectrometry methods that can be used to analyze histone proteins. Abbreviations: CID, collision-induced dissociation; ECD, electron capture dissociation; ETD, electron transfer dissociation; ESI, electrospray ionization; HPLC, high-performance liquid chromatography; LC, liquid chromatography; MS/MS, tandem mass spectrometry.

unmodified precursor ions (20). This method is fast because it uses only MS1 and LC retention time information. This analysis requires that both unmodified and modified precursor ions be in the sample and that at least one of them is present in high abundance. Therefore, this method is not sensitive for low-level PTMs. The second method is open searching, which opens the precursor mass tolerance to 300 Da and considers all amino acid sites within the sequence (21). This method is slow but sensitive. Therefore, once MS2s are collected with modified peptides, both restricted (known) and unrestricted (unknown) PTMs can be identified and assigned to a specific site.

The above principles are about the collection and interpretation of mass spectra. There are three MS strategies for analyzing proteins: bottom-up, top-down, and middle-down (shown in **Figure 2**).

2.2. Bottom-Up Mass Spectrometry

In bottom-up MS, peptides are analyzed and used for protein identification. The procedure for bottom-up is as follows. First, proteins are extracted from cells. To reduce sample complexity, proteins can be separated by different techniques such as two-dimensional gel electrophoresis or high-performance liquid chromatography (HPLC). Second, proteins are digested into peptides with proteases. Third, peptides are eluted by LC, ionized, and scanned to generate MS1s. Some peptides are selected and fragmented to generate MS2s. After the mass spectra are generated, peptide sequences are assigned by database searching. The identified peptides can be assembled into proteins. Therefore, bottom-up is a peptide-centric MS technology.

To digest proteins into peptides, different enzymes can be used. The most commonly used enzyme is trypsin, which cleaves at the C-terminal of lysine (K) and arginine (R) residues. However, the direct use of trypsin on histones is problematic. The N-terminal tails of histones are lysine and arginine rich, so trypsin digestion results in small pieces that cannot be detected by MS. Although there are miscleavages, the number of miscleavages is small (e.g., one or two), and miscleavages are not reproducible. Other enzymes can be used, but they are much less specific than trypsin and also result in nonreproducible digests. Therefore, histones should be derivatized before trypsin digestion.

There are two derivatization methods reported for histones. The first method uses acetic anhydride, which reacts with lysines and blocks trypsin digestion (22, 23). The second method uses propionic anhydride (24–27). Both chemical acetylation and propionylation occur on unmodified and monomethylated lysines as well as the N-terminal amino acid. The propionylation not only blocks trypsin digestion at lysines, but also enhances hydrophobicity of histone peptides, thereby increasing the chromatographic resolution of different peptides. After the trypsin digestion, another propionyl group is added to the N-terminal of each peptide, further enhancing hydrophobicity. The increase in hydrophobicity and reproducible digestions resulting from propionylation make this the preferred derivatization method for histones.

2.3. Top-Down Mass Spectrometry

In contrast to bottom-up, top-down analyzes whole proteins. The top-down procedure is as follows. First, protein mixtures are separated. Second, they are eluted and introduced into the instrument. Third, proteins are scanned, and MS1s are generated. Some proteins are selected and fragmented to generate MS2s. After the mass spectra are generated, they can be identified as proteins by database searching. Therefore, top-down is a protein-centric MS technology.

There are several differences between top-down and bottom-up methods. First, the molecular weights of precursor ions are significantly different. In bottom-up, the lengths of tryptic peptides are often between 6 and 20 amino acids. Thus, most peptides are ~2 kDa. In top-down, proteins are often longer than 100 amino acids and larger than 10 kDa. For example, the molecular weights of canonical histones are between 11 kDa and 15 kDa. The different molecular weights lead to different properties and challenges. Peptides are generally soluble, whereas proteins are often insoluble. Peptides and proteins require different types of separation. Peptides typically have low charge states (+2 or +3 charge states), whereas proteins are highly charged (+10 to +100 charge states).

Second, the fragmentation type is different. In CID, the fragmentation energy is high, and labile modifications will be lost (e.g., phosphorylation). In ETD, the fragmentation method results in electron transfer, and it is better for highly charged ions and labile modifications. In bottom-up, there are only a few PTMs on one peptide, so both CID and ETD can be used. In top-down, histones are highly charged and there are multiple PTMs, so ETD or related ECD is essential for studying PTMs on whole proteins.

Third, the data complexity is different. In bottom-up, it is easy to identify the monoisotopic peak and charge state for precursor ions in high-resolution MS. Sometimes there are two or more peptides coeluting, which can make analysis more complex. Fortunately, several methods have been developed to determine the monoisotopic peak and charge state for peptides (28–32). In top-down, it is difficult to detect the monoisotopic peak and charge state, even if there is only one precursor ion and the MS is high resolution. Usually, the monoisotopic peak is buried in the noise, and the resolution is not high enough to distinguish charge states. Manual interpretation of these data is difficult for these reasons. Although several computational methods have been developed to resolve these problems, their performance needs to be improved (33–35). Because the fragment ions in top-down are also highly charged, their monoisotopic peaks and charge states also need be determined. PTMs lead to similar or the same m/z values for precursor ions, which further complicates data analysis. The high number of PTMs on histones makes top-down analysis of histone proteoforms very complex.

For these reasons, top-down is more difficult than bottom-up. However, top-down has the advantage of viewing all the PTMs at the protein level. The whole protein can also be sequenced, making histone variants easy to study. To combine the advantages of top-down (global view of PTMs and variants) and bottom-up (easy to operate and sensitive), middle-down was developed.

2.4. Middle-Down Mass Spectrometry

As implied by the name, middle-down is in between top-down and bottom-up. The procedure is similar to bottom-up. The difference is that specific enzymes are used to obtain much longer peptides. For example, Glu-C is used to cleave histone H3 at the C-terminal of glutamic acid and to obtain the 1–50 peptide. Similarly, Asp-N is used to cleave histone H4 at the N-terminal of aspartic acid and to obtain the 1–23 peptide. Middle-down histone peptides have multiple PTMs and can be used for combinatorial PTM analysis. Because the peptides in middle-down are much longer than those in bottom-up, middle-down has similar problems as top-down and is less sensitive than bottom-up. Middle-down peptides have much wider charge-state distributions than bottom-up peptides. Only one charge state is selected at a time for fragmentation. Thus, the signal in each charge state is lower in middle-down than in bottom-up.

The above three strategies all have their own advantages and disadvantages. Bottom-up is easy to operate and the most sensitive. However, it lacks the global view of PTMs and loses the relationship between peptides and proteins (i.e., it is difficult to distinguish proteins and their variants by peptides). Top-down is able to view all PTMs and protein variants but is difficult to operate and less sensitive. Middle-down is in the middle. Nowadays, bottom-up is mature and has become the workhorse in mass spectrometric analysis, whereas top-down and middle-down are promising but require expertise.

3. APPLICATIONS

Because histones are heavily modified and have several sequence variants, analysis of histone proteoforms is very difficult. Fortunately, MS can provide a vast amount of information in a high-throughput way and without prior knowledge. With different techniques, identification and quantification are the basic information obtained from MS. More information can be obtained through, for example, distinguishing histone variants, the discovery of histone-binding proteins, and the analysis of combinatorial histone PTMs. Thus, MS is a powerful tool for uncovering histone function. With the further development of MS, more functions and applications will be found in the future.

3.1. Identification of Novel Sites or Types of Histone Posttranslational Modifications

In contrast to antibody-based methods, MS does not need any prior knowledge of PTMs. One application of MS is to find more sites for the known types of PTMs (e.g., new acetylation sites within histones). In these cases, the PTMs can be identified by MS2 and database searches. From the identification results, the spectra of the specific PTM can be separated. Generally, the identification results need to be manually checked because the site assignment of PTMs will lead to a combinatorial explosion of candidate sequences and a high false discovery rate (FDR). Thus, to confirm a PTM site, several steps need be done. First, the match score should be high enough. Second, the sequence and the spectrum should match well, and there should be fragment ions to support the PTM site assignment. Third, the chromatography of the precursor ion should be checked. Only after all of these properties are checked can the site be considered reliable. These tools are useful for newly discovered PTMs as well, e.g., citrullination of arginine on histones (36). If the site appears unreliable, the sample preparation should be improved.

Another application is to find unexpected PTMs. As mentioned in Section 2.1, open searching is usually used for novel PTM identification. Similar to restricted searching, the identification results from open searching also need to be checked. At the beginning, only the PTM mass and potential site are known. The results can be compared with the PTM database Unimod (**http://www.unimod.org**). If the mass and site match well, this PTM is already known. Otherwise, this PTM may be novel. If the combinational mass of known PTMs is still not equal to this PTM (e.g., the sum of two modifications, or one modification minus another modification), more techniques are needed to validate it. Crotonylation of lysine on histones was found in this way (37).

If there is no enrichment, the concentration of novel PTM sites and types can be low. In these cases, the most sensitive bottom-up should be used to discover novel PTM sites or types.

3.2. Quantification of Histone Posttranslational Modifications

Knowing the type and site of a PTM is important. More important is to know the level of PTMs, because the level can be related to function, e.g., gene regulation. There are two kinds of quantification methods: label-based methods and label-free methods. In the label-free method, no extra experiments are needed. However, the reproducibility is very important for reliable analysis. One label-free method is spectral counting, in which the number of identified spectra is used to quantify precursor ions (38–40). However, there are factors that affect the identification, including ionization efficiency and precursor selection. In a word, spectral counting is simple, but it is inaccurate.

Another label-free method relies on calculating the area under the precursor peak (41). For one sequence with different modifications, such as unmodified, mono-, di-, tri-, methylation, the area under each precursor peak is calculated. Then the proportion of each form is calculated by dividing by the total area. The reproducibility of this method is also important. In bottom-up MS, miscleavages will affect the calculation of the total area and lead to the inaccurate assignment of proportions to each form.

There are two kinds of labeling methods: in vitro and in vivo. Chemical derivatization, such as isobaric tags for relative and absolute quantitation (iTRAQ), is an in vitro method (42–44). In iTRAQ, samples are treated separately. After trypsin digestion, each sample interacts with different reagents. Each reagent contains a reporter group and a balance group. In the 4-plex reagents, the masses of reporter groups are 114, 115, 116, and 117 Da, and the corresponding masses of balance groups are 31, 30, 29, and 28 Da. Then the four samples are equally mixed. In the MS1s, the same peptides from different samples elute at the same time and are detected as the same m/z. All the peptides are selected and fragmented, during which the balance groups are lost and the reporter groups are detected. The relative ratios of the reporter groups in the MS2s represent the relative quantification of the peptide from the four different samples.

Metabolic stable isotope labeling, such as stable isotope labeling by amino acids in cell culture (SILAC), is an in vivo labeling method (45). In SILAC, one sample is cultured in normal media (light), while the other is cultured in heavily labeled media (heavy), e.g., $^{13}C_6{}^{15}N_2$ on lysine and $^{13}C_6{}^{15}N_4$ on arginine. Because trypsin digests at the C-terminal of lysine and arginine, all tryptic peptides will be labeled heavily. After the cells are harvested, the light and heavy samples are mixed equally. In MS1s, the heavy peptide and the light peptide have a mass difference of 8 Da (one heavy lysine) or 10 Da (one heavy arginine), or some other combination. The heavy and light pairs are determined and the relative ratio can be calculated by the intensity of each peptide. After one of them is fragmented and an MS2 spectrum is generated, the peptide sequence will be identified.

In SILAC, the heavy and light isotopes are mixed early during sample preparation, whereas in iTRAQ, the heavy and light isotopes are mixed after trypsin digestion. Therefore, it is easy to introduce biases with iTRAQ. Nevertheless, both SILAC and iTRAQ need be calibrated to ensure proper relative quantification.

For bottom-up, the quantification of peptides can be accurate in label-based methods. However, it is difficult to obtain accurate protein quantification because of shared peptides between protein variants and the varying ionization efficiencies of different peptides. For top-down, there are no such problems.

3.3. Distinguishing Histone Variants

Each histone has its own family. In some families, the variants can have as little difference as a few amino acids. To distinguish them is a challenge that requires both separation and MS. First, bulk histones can be separated into each family (as shown in **Figure 2**) by reversed-phase HPLC. Second, each family can be separated further by other methods, such as weak cation exchange hydrophilic interaction LC. Third, mass spectrometers are used to identify the family members. For bottom-up, the unique peptides help to distinguish the variants. However, these unique peptides may not be identified for many reasons having to do with, for example, ionization efficiency, precursor selection, fragmentation, and the algorithms for identification. Thus, it is difficult to study most variants using bottom-up. Instead, top-down is usually used to investigate histone variants. Because top-down is not as sensitive as bottom-up, larger amounts of sample are needed for top-down than bottom-up. After the data are generated, computational programs may be utilized to identify the variants and their PTMs. In summary, distinguishing histone variants is a big challenge for MS-based technologies and involves many steps including separation, MS, and data analysis. There are several papers that delve into analyzing histone variants using top-down methods (see, e.g., 46–48).

3.4. Discovery of Histone-Binding Proteins

Histone PTMs can bind proteins to regulate genes (49–51). To discover histone-binding proteins, histone peptides are used as baits to essentially pull down histone binders from cell lysates. The binders are both identified and quantified using mass spectrometric methods, named forward and backward experiments. In the forward experiment, the light lysate with the unmodified histone peptide serves as the control, and the heavy lysate with the modified histone peptide serves as the experiment. The control and the experiment are then mixed equally. If there is a protein binding to the modified histone peptide, the heavy-to-light ratio should be high. For further confirmation, the backward experiment needs be done by switching the histone peptide baits so that the modified histone peptide is used with a light lysate. In this experiment, the heavy-to-light ratio should be low. The two samples require preparation for MS analysis. The proteins eluted from the histone peptides can be separated by gel electrophoresis. The gels are cut into bands. In-gel digestion is carried out, and bottom-up data are obtained for each band. Computational methods are used to identify and quantify the histone-binding proteins. These experiments generate long lists of candidate proteins; therefore, candidate proteins need to be studied to assess biological function.

3.5. Analysis of Combinatorial Histone Posttranslational Modifications

In general, a single PTM can have various functions, such as acetylation, methylation, and phosphorylation (52–54). Recently, it has been found that several histone PTMs work together, e.g., K27me1–3/S28ph on H3, which means mono-, di-, or trimethylation occurs on lysine 27 while phosphorylation simultaneously occurs on serine 28 (55). Because histones are heavily modified, there are many potential PTM combinations to be discovered, which poses challenges for current technologies. One potential approach is as follows. As mentioned above, the histone-binding proteins can be discovered by one PTM. Then other PTMs near the binding site can also be identified. When one of the latter PTMs is changed to the unmodified form, it can be determined whether the abundance of the binding protein has been changed. If so, the latter PTM is the combinatorial one with the PTM on the binding site. Therefore, this analysis is the most complex.

4. CHALLENGES

Despite recent advances in MS techniques, there are still many challenges, including those in sample preparation, MS, and data analysis. When these challenges are overcome, MS will become even more powerful for probing histone PTMs and their functions.

4.1. Sample Preparation

As mentioned above, there are two significant differences between peptides and proteins: Peptides are generally soluble, whereas proteins can precipitate. Also, peptides are separated relatively easily, whereas proteins can be more difficult to separate. For these reasons, bottom-up has become a common approach, and there are several protocols for sample preparation. Top-down needs to be improved before it can become a high-throughput and widely applied technology.

4.2. Mass Spectrometry

The mass spectrometer is also not perfect. At least five aspects of MS can present problems: ionization efficiency, precursor ion selection, fragmentation, detection, and resolution. First, the ionization efficiency is different for all precursor ions. Some precursor ions ionize well and can be detected easily, whereas others ionize poorly. In one protein, the quantification of peptides is different due to different ionization efficiencies. Therefore, ionization efficiency can cause problems for identification and quantification.

Second, precursor ion selection is another problem. In data-dependent acquisition mode (DDA), the top n most intense peaks are selected for fragmentation. In DDA, an isolation window of 2 Da is used and allows for one or more precursor ions to be selected for fragmentation. To prevent repeated selection of the same precursor ions, the dynamic exclusion windows can be set such that several seconds pass between selecting precursor ions with the same mass. However, this mode is not suitable for low-abundance precursor ions because they may never be selected for fragmentation. Some low-abundance precursor ions are important. For example, when some precursor ions contain PTMs, their abundances are often low if no enrichment or purification has been done. To overcome this problem, data-independent acquisition mode (DIA) is used to fragment all precursor ions in a wider window (56–58). For example, the MS1 can be partitioned into windows of 25 m/z. With this approach, many precursor ions, including the low-abundance precursor ions, can be fragmented at the same time. The large number of precursor ions fragmented makes it difficult to identify the MS2s. Thus, DDA and DIA both have their own pros and cons.

Third, fragmentation is not well understood, and this has implications for identification and quantification. Some mechanisms have been identified, such as charge-remote fragmentation (59); however, this is just the tip of the iceberg. The lack of accurate intensity predictions when generating theoretical spectra means that only the m/z values of theoretical ions are used in database searching. Alternatively, the fragmentation patterns can be archived in a spectral library and will then carry more information than m/z values alone. From this point of view, spectral library searching is more accurate than database searching (60).

Fourth, there are MS detection limitations. In sample mixtures, some proteins and peptides are abundant, whereas others are scarce. If the peaks are weak or buried in noise, they will be difficult to detect. Therefore, it is important to separate the high- and low-abundance proteins or peptides. Another challenge can come from the isotopic distribution, especially in top-down. Because the molecular weights of proteins are much larger than those of peptides, there are many more isotopic peaks, and the peak intensities are in a normal distribution. The large number of isotopic components causes the monoisotopic peak (the peak corresponding to ions with molecular mass calculated using only the most common isotope of each element in the molecule) to be much lower in intensity than the middle peaks. When the difference is as large as four orders of magnitude, such as is the case for 15-kDa proteins, it is difficult to detect the monoisotopic peak. Therefore, detection is important for correctly assigning identity.

Lastly, MS resolution can be a limitation. Resolution is a measure of the ability to separate adjacent peaks. For bottom-up, the resolution is high enough to separate isotopic peaks. The charge state is determined easily by the m/z intervals between the isotopic peaks and is usually +2 or +3. However, for top-down, the resolution is not high enough. For example, the m/z interval of +20 is 0.05016 (the mass difference of ^{13}C and ^{12}C divided by 20), and the m/z interval of +21 is 0.04777. The mass difference is 0.00239, which is 0.12 ppm at 20 kDa. It is difficult to have such high resolution and mass accuracy. To obtain high resolution, scan speed is low and this decreases the speed of data acquisition. Therefore, there is a balance between resolution and speed.

4.3. Data Analysis

Several computational problems impact data analysis. First, the mass spectra need to be preprocessed. Preprocessing includes noise deletion, monoisotopic peak and charge state detection, spectra filtration, and spectra clustering (61). The most important is monoisotopic peak and charge state detection. As mentioned above, it is easy to determine the monoisotopic peak and charge state for precursor ions in bottom-up. However, in middle-down and top-down, the detection of the monoisotopic peak and charge state is much more difficult due to the limitations in MS detection and resolution. Moreover, the coeluted precursor ions in DDA and the cofragmented precursor ions in DIA make the MS2s more complex. For the former, each precursor ion that is detected can be identified by the MS2. For the latter, the fragment ions can be correlated to the corresponding precursor ions by their similar chromatography.

Second, identification needs to be improved. Although several algorithms have been developed for bottom-up, the identification rate is still low—only 10–40% of the spectra can be identified.

There are many reasons for the low identification rates, such as sample loss, MS loss, and imperfectness of algorithms (62). When the sample preparation and MS are done well, the identification rate can be as high as 80%. One example comes from the ABRF iPRG 2013 study (63). The challenge is to increase the identification rate at a certain FDR. Some exceptions should be considered in spectral identification, such as unexpected PTMs and semi- or nonspecific digestion. These lead to a combinatorial explosion of candidates and more false positives. How to process them quickly and accurately is key for accurate identification. Furthermore, there are methods to control the FDR. For each spectrum, the p-value can be calculated from the score distribution of candidates. For all spectra, FDR can be calculated by the mixture model of correct and incorrect matches or by target-decoy database searching (64, 65).

The site localization of PTMs is a special identification problem. If there is more than one potential PTM site on a peptide, this will cause problems for site localization (66). The fragment ions between the two sites should be checked, and they can distinguish the sites. The probability of each potential site can be calculated from these distinguishable ions. If there is a significant difference in the probability of two sites, then the most probable site is the correct one. Otherwise, the site localization cannot be determined. This assumes that there is only one correct site. In some situations, several precursor ions with one sequence but different sites coelute and are cofragmented. In this case, all potential sites should be detected.

Third, quantification should be checked carefully because problems can arise from both experimental and computational work. When carrying out experiments, samples should be mixed equally. However, this is difficult to control, so the distribution of quantification needs to be checked. The initial quantification can be normalized to the center of distribution. Moreover, interferences in quantification, such as noise peaks and coeluted precursor ions, can cause issues. Even in DDA, ~50% of MS2s are mixed spectra (67). When interference happens, the quantification may be inaccurate. Only isotopes that do not have interfering peaks can be used for accurate quantification. Therefore, detecting non- or less-interfered isotopes or removing overlapping peaks is key to improving the accuracy of quantification.

Lastly, data analysis pipelines should be available. In a single experiment, considerable raw data will be generated, including technical replicates. Identification and quantification are the basic analyses. However, several steps to run these basic analyses must be implemented, such as format conversion, database indexing, database searching, result filtering, and quantification (61). It is easy to make mistakes when setting parameters, and it is time-consuming to run each step manually. The ideal pipeline is to prepare the raw data, database file, and parameter file once and then run each step automatically to obtain the results. This will decrease the amount of work associated with basic analyses, allowing researchers to focus on deeper analysis, such as checking protein function.

5. CONCLUSIONS

A mass spectrometer detects the m/z values of ions. From the MS2s, the sequences of peptides or proteins can be determined. Using label-free or label-based methods, quantification can be obtained—lending insight into protein function. The basic analyses of identification and quantification can be implemented with three strategies: bottom-up, top-down, and middle-down. Each strategy has its pros and cons. When applied to histones, these strategies provide a wealth of information. In contrast to antibody-based methods, no prior knowledge is needed for MS analysis. Therefore, MS is a powerful technology for studying histone proteoforms and their functions. MS can provide information about novel types or sites of PTMs, including combinational PTMs; can distinguish histone variants; and can identify histone-binding proteins. To complete these tasks, techniques need be optimized, via sample preparation, MS methods, data analysis platforms, etc. Although these techniques are imperfect, they have helped to resolve many practical problems. In the future, these techniques will improve further, and more discoveries will follow.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

B.A.G. acknowledges funding from an NIH Innovator grant (DP2OD007447) from the Office of the Director and from the National Science Foundation Early Faculty CAREER award.

LITERATURE CITED

1. Smith LM, Kelleher NL. 2013. Proteoform: a single term describing protein complexity. *Nat. Methods* 10:186–87
2. Arnaudo AM, Garcia BA. 2013. Proteomic characterization of novel histone post-translational modifications. *Epigenet. Chromatin* 6:24–30
3. Arnaudo AM, Molden RC, Garcia BA. 2011. Revealing histone variant induced changes via quantitative proteomics. *Crit. Rev. Biochem. Mol. Biol.* 46:284–94
4. Cheung P. 2004. Generation and characterization of antibodies directed against di-modified histones, and comments on antibody and epitope recognition. *Methods Enzymol.* 376:221–34
5. Britton LM, Gonzales-Cope M, Zee BM, Garcia BA. 2011. Breaking the histone code with quantitative mass spectrometry. *Expert Rev. Proteomics* 8:631–43
6. Andersson C-O. 1958. Mass spectrometric studies on amino acid and peptide derivatives. *Acta Chem. Scand.* 12:1353
7. Dass C. 2007. *Fundamentals of Contemporary Mass Spectrometry*. Hoboken, NJ: Wiley
8. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246:64–71
9. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, et al. 1988. Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of flight mass spectrometry. *Rapid Commun. Mass Spectrom.* 2:151–53
10. Wells JM, McLuckey SA. 2005. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* 402:148–85
11. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. 2004. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. USA* 101:9528–33
12. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. 1999. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 6:327–42
13. Ma B, Zhang K, Hendrie C, Liang C, Li M, et al. 2003. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17:2337–42
14. Taylor JA, Johnson RS. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 11:1067–75
15. Eng JK, McCormack AL, John R, Yates I. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5:976–89
16. 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
17. Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–67

18. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. 2004. Open mass spectrometry search algorithm. *J. Proteome Res.* 3:958–64
19. Bern M, Cai Y, Goldberg D. 2007. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* 79:1393–400
20. Fu Y, Xiu LY, Jia W, Ye D, Sun RX, et al. 2011. DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. *Mol. Cell Proteomics* 10:M110.000455
21. Ye D, Fu Y, Sun RX, Wang HP, Yuan ZF, et al. 2010. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* 26:i399–i406
22. Smith CM, Haimberger ZW, Johnson CO, Wolf AJ, Gafken PR, et al. 2002. Heritable chromatin structure: mapping "memory" in histones H3 and H4. *Proc. Natl. Acad. Sci. USA* 99 (Suppl. 4):16454–61
23. Smith CM, Gafken PR, Zhang Z, Gottschling DE, Smith JB, Smith DL. 2003. Mass spectrometric quantification of acetylation at specific lysines within the amino-terminal tail of histone H4. *Anal. Biochem.* 316:23–33
24. Bonaldi T, Imhof A, Regula JT. 2004. A combination of different mass spectroscopic techniques for the analysis of dynamic changes of histone modifications. *Proteomics* 4:1382–96
25. Peters AH, Kubicek S, Mechtler K, O'Sullivan RJ, Derijck AA, et al. 2003. Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. *Mol. Cell* 12:1577–89
26. Syka JE, Marto JA, Bai DL, Horning S, Senko MW, et al. 2004. Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J. Proteome Res.* 3:621–26
27. Garcia BA, Mollah S, Ueberheide BM, Busby SA, Muratore TL, et al. 2007. Chemical derivatization of histones for facilitated analysis by mass spectrometry. *Nat. Protoc.* 2:933–38
28. 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
29. Park K, Yoon JY, Lee S, Paek E, Park H, et al. 2008. Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. *Anal. Chem.* 80:7294–303
30. Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26:1367–72
31. Yuan ZF, Liu C, Wang HP, Sun RX, Fu Y, et al. 2012. pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics* 12:226–35
32. Niu M, Mao X, Ying W, Qin W, Zhang Y, Qian X. 2012. Determination of monoisotopic masses of chimera spectra from high-resolution mass spectrometric data by use of isotopic peak intensity ratio modeling. *Rapid Commun. Mass Spectrom.* 26:1875–86
33. Senko MW, Beu SC, McLafferty FW. 1995. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 6:229–33
34. Senko MW, Beu SC, McLafferty FW. 1995. Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *J. Am. Soc. Mass Spectrom.* 6:52–56
35. Horn DM, Zubarev RA, McLafferty FW. 2000. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 11:320–32
36. Anzilotti C, Pratesi F, Tommasi C, Migliorini P. 2010. Peptidylarginine deiminase 4 and citrullination in health and disease. *Autoimmun. Rev.* 9:158–60
37. Tan M, Luo H, Lee S, Jin F, Yang JS, et al. 2011. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* 146:1016–28
38. Lundgren DH, Hwang SI, Wu L, Han DK. 2010. Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* 7:39–53
39. Heinecke NL, Pratt BS, Vaisar T, Becker L. 2010. PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. *Bioinformatics* 26:1574–75
40. Carvalho PC, Hewel J, Barbosa VC, Yates JR 3rd. 2008. Identifying differences in protein expression levels by spectral counting and feature selection. *Genet. Mol. Res.* 7:342–56

41. DiMaggio PA Jr, Young NL, Baliban RC, Garcia BA, Floudas CA. 2009. A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Mol. Cell Proteomics* 8:2527–43
42. Gan CS, Chong PK, Pham TK, Wright PC. 2007. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J. Proteome Res.* 6:821–27
43. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. 2004. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics* 3:1154–69
44. Aggarwal K, Choe LH, Lee KH. 2006. Shotgun proteomics using the iTRAQ isobaric tags. *Brief Funct. Genomics Proteomics* 5:112–20
45. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics* 1:376–86
46. Boyne MT 2nd, Pesavento JJ, Mizzen CA, Kelleher NL. 2006. Precise characterization of human histones in the H2A gene family by top down mass spectrometry. *J. Proteome Res.* 5:248–53
47. Siuti N, Roth MJ, Mizzen CA, Kelleher NL, Pesavento JJ. 2006. Gene-specific characterization of human histone H2B by electron capture dissociation. *J. Proteome Res.* 5:233–39
48. Thomas CE, Kelleher NL, Mizzen CA. 2006. Mass spectrometric characterization of human histone H3: a bird's eye view. *J. Proteome Res.* 5:240–47
49. Brownell JE, Zhou J, Ranalli T, Kobayashi R, Edmondson DG, et al. 1996. Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* 84:843–51
50. Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 410:116–20
51. LeRoy G, Rickards B, Flint SJ. 2008. The double bromodomain proteins Brd2 and Brd3 couple histone acetylation to transcription. *Mol. Cell* 30:51–60
52. Jenuwein T, Allis CD. 2001. Translating the histone code. *Science* 293:1074–80
53. Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* 403:41–45
54. Kouzarides T. 2007. Chromatin modifications and their function. *Cell* 128:693–705
55. Garcia BA, Barber CM, Hake SB, Ptak C, Turner FB, et al. 2005. Modifications of human histone H3 variants during mitosis. *Biochemistry* 44:13202–13
56. 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
57. 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 Proteomics* 11:O111.016717
58. 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
59. Cheng C, Gross ML. 2000. Applications and mechanisms of charge-remote fragmentation. *Mass Spectrom. Rev.* 19:398–420
60. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, et al. 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7:655–67
61. Domon B, Aebersold R. 2006. Challenges and opportunities in proteomics data analysis. *Mol. Cell Proteomics* 5:1921–26
62. Nesvizhskii AI, Vitek O, Aebersold R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 4:787–97
63. Bandeira N, Chambers MC, Cottrell JS, Deutsch EW, Kapp EA, et al. 2013. iPRG-2013: Proteome Informatics Research Group study: using RNA-Seq data to refine proteomic data analysis. *J. Biomol. Tech.* 24:S23–S24
64. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74:5383–92

65. 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
66. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* 24:1285–92
67. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. 2010. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* 9:4152–60