Amino Acids (2012) 43:1075–1085
DOI 10.1007/s00726-012-1287-x

INVITED REVIEW

Key issues in the acquisition and analysis of qualitative and quantitative mass spectrometry data for peptide-centric proteomic experiments

Andrew J. Thompson · Mika Abu · Diane P. Hanger

Received: 9 December 2010 / Accepted: 3 April 2012 / Published online: 22 July 2012

Abstract Proteomic technologies have matured to a level enabling accurate and reproducible quantitation of peptides and proteins from complex biological matrices. Analysis of samples as diverse as assembled protein complexes, whole cell lysates or sub-cellular proteomes from cell cultures, and direct analysis of animal and human tissues and fluids demonstrate the incredible versatility of the fundamental nature of the technique that forms the basis of most proteomic applications today (mass spectrometry). Determining the mass of biomolecules and their fragments or related products with high accuracy can convey a highly specific assay for detection and identification. Importantly, ion currents representative of these specifically identified analytes can be accurately quantified with the correct application of smart isobaric tagging chemistries, heavy and light isotopically derivatised samples or standards, or by careful application of workflows to compare unlabelled samples in so-called 'label-free' and targeted selected reaction monitoring experiments. In terms of exploring biology, a myriad of protein changes and modifications are being increasingly probed and quantified, including diverse chemical changes from relatively decisive modifications such as protein splicing and truncation, to more transient dynamic modifications such as phosphorylation, acetylation and ubiquitination. Proteomic workflows can be complex beasts and several key considerations to ensure effective applications have been outlined in the recent literature. The past year has witnessed the publication of several excellent reviews that thoroughly describe the fundamental principles underlying the state of the art. This review further elaborates on specific critical issues introduced by these publications and raises other important unaddressed considerations and new developments that directly impact on the effectiveness of proteomic technologies, in particular for, but not necessarily exclusive to peptide-centric experiments. These factors are discussed both in terms of qualitative analyses, including dynamic range and sampling issues, and developments to improve the translation of peptide fragmentation data into peptide and protein identities, as well as quantitative analyses, including data normalisation and the utility of ontology or functional annotation, the effects of modified peptides, and considered experimental design to facilitate the use of robust statistical methods.

Keywords Proteomics · Peptide-centric · Mass spectrometry · Quantitation · Data analysis

A. J. Thompson (✉)
The Institute of Cancer Research, Proteomics Core Facility,
237 Fulham rd, Chelsea, London SW3 6JB, UK
e-mail: Andrew.thompson@icr.ac.uk

M. Abu · D. P. Hanger
MRC Centre for Neurodegeneration Research,
Institute of Psychiatry, King's College London, London, UK

Introduction

Several excellent manuscripts have recently described the extraordinary biological analyses enabled by considered application of the maturing proteomic technologies, including quantitation of protein–protein interactions and protein complex components and modifications (Gavin 2010; Przbylski 2010; Wepf 2009), direct quantitation of biomarkers for disease in human fluids (Dayon 2008; Glen 2010), quantitation of protein signalling responses between different interacting cell types (Jorgensen 2009), and more recently quantitation of proteins exchanged between different organisms (Rechavi 2010). The technology and

applications are now maturing to the extent that comprehensive characterisation of the human proteome is being considered (Nilsson 2010). The recent achievements have been made possible by smart application of mass spectrometry in an arguably non-intuitive fashion. Ideally, the direct mass analysis of an exact molecule of interest, usually proteins in proteomic studies, would be expected to be most relevant and accurate. However, using the more common mass spectrometry platforms today—time-of-flight (TOF), triple quadrupole (QQQ), and ion trap (IT) instruments—the sensitivity of analysis, and to a lesser degree the accuracy of analysis, decreases exponentially with increasing mass; it is not uncommon to require picomole quantities of material for direct experiments on intact proteins (van Duijn 2010). This compromises our present ability to comprehensively analyse and quantify intact proteins directly from biological matrices. In contrast, current technology can very sensitively analyse smaller molecules, for instance routine detection of low attomole quantities of peptides with high accuracy can be easily achieved (routine part-per-million now verging on part-per-billion mass accuracy; Cox 2008). Hence, most mass spectrometry-based proteomic strategies are 'bottom-up' in the sense that peptide sets are analysed that are derived from the parent intact proteins by incubation with selected proteases, as opposed to the 'top-down' approach of directly analysing the masses of the intact proteins themselves. The 'bottom-up' proteomics approach is therefore peptide-centric, in that peptides are analysed as proxies for their parent proteins and the protein-level interpretation is inferred from the peptide-level results. The most advanced applications of mass spectrometry in proteomics to date employ peptide-centric strategies to quantify peptides and infer quantitative information relating to the originating parent protein(s). The approach is particularly applicable for the identification, site localisation, and quantitation of chemical changes to proteins including post-translational modifications (PTMs), since the mass difference due to incorporation of the more common smaller modifications, such as methylation, acetylation and phosphorylation, is more easily discriminated against the peptide mass than the intact protein mass. Furthermore, fragmentation data can be obtained more completely and with far higher sensitivity at the peptide level than the protein level, enabling confident sequencing of the modified peptide and often localisation of the precise site of modification. The resulting PTM identification and localisation information from these experiments can then be coupled with quantitation of the ion current to determine changes in abundance of the precise modification(s) between samples.

The general strengths and weaknesses of peptide-centric approaches for mass spectrometry-based protein analysis have been recently well described by Duncan (2010). Key weaknesses presented include the disconnection of analytical information from the intact protein level, loss of the combinatorial relationships of modifications, and inconsistency with the reproducibility and predictability of the peptides generated from protein digestion, and the reader is directed to this manuscript for further detail. Similarly, three frequently applied proteomic strategies for peptide-centric analyses (shotgun/discovery proteomics, directed mass spectrometry, and targeted proteomics) have also been recently described (Domon 2010). These strategies are increasingly focussed in their application and are thus suited to specific purposes, from lower-throughput hypothesis-generating discovery screens through to higher-throughput, highly specific hypothesis-driven monitoring of selected biological products or markers. The reader is directed to this article for discussion of general considerations regarding selectivity, limit of detection, dynamic range, data density, and repeatability and reproducibility, as well as technical points focussed around the function and performance of selected mass spectrometry instrumentation for these applications. Since the above points have been recently described elsewhere they will not be further elaborated here and the reader is referred to the above articles for further details. The present review will instead build on the recent literature to present further key challenges in terms of both the qualitative and the quantitative aspects of proteomic analyses.

Key challenges in qualitative proteomic analyses

Peptide-centric proteomic analysis is reliant on identification of peptides from the tandem mass spectrometry peptide fragmentation spectra. Despite the wealth of biological information generated from the experiments, the mass spectral data acquired is often not comprehensively representative of all peptides and proteins present in the sample, and further, the majority of acquired data remains uninterpreted. In a backhanded sense this compliments the utility of the technology, but clearly more efficient and effective acquisition and use of the data would yield greater scientific rewards. Two significant challenges that contribute to these issues are the challenge of generating sufficiently high-quality fragmentation spectra that comprehensively represent the protein population analysed, and the challenge of translating the fragmentation mass spectra obtained into biological observations. Neither of these challenges can be distilled to a single root issue to resolve, but some key factors can be addressed. These factors include sample crowding effects and dynamic range of protein abundance issues, and the translation of fragmentation spectra into peptide and protein information, which are discussed in the following sections.

Sample 'crowding' and dynamic range of protein abundance issues

Most mass spectrometry applications in proteomics employ online elution of peptides for real-time pre- and post-fragmentation mass analysis in the mass spectrometer. In this style of experiment, the instrument duty cycle—the rate at which the instrument can analyse peptides—becomes a limiting factor. This is especially so for analyses of complex biological matrices where proteolytic digestion can yield hundreds of thousands of peptides resulting in 'crowded' samples for real-time on-the-fly analysis. 'Crowding' of samples can have two significant effects: there can be too many peptides eluting within the duty cycle timeframe for the mass spectrometer to effectively analyse leading to under sampling and stochastic sampling effects, as illustrated by Liu (2004) in the replicate analysis of the yeast proteome and further described in detail by Duncan (2010); and co-elution of peptides of near-isobaric masses can result in the mass spectrometer acquiring fragmentation spectra of mixed precursor peptides that are difficult to interpret confidently. Introducing chemical modifications or isotopes can further exacerbate the 'crowding' effect, for instance SILAC quantitative experiments can effectively double or triple the number of peptide ions compared to a non-SILAC analysis of the same sample type. An additional issue is the large dynamic range of abundances of proteins, particularly observed in the more complex biological samples and often exemplified by the analysis of serum (Anderson 2002). Since the acquisition of fragmentation spectra is commonly prioritised towards the more intense peptide ions, the results are biased towards the more abundant proteins causing preferential under sampling of the least abundant components. The net effect of the 'crowding' and dynamic range issues is incomplete representation of peptides and proteins, especially in more complex samples, and a concomitant variation in the repeatability and reproducibility of the experiments due to stochastic sampling. Solutions that at least partly address these issues include: prior fractionation of samples, for instance by multi-dimensional chromatography (Washburn 2001), SDS-PAGE separation (Schirle 2003) or application of selected affinity techniques (phosphorylation, ubiquitination enrichment or immunoprecipitation of target analytes) to reduce sample complexity to more manageable levels; immunodepletion of abundant proteins (Pieper 2003) to partly address issues with the dynamic range of protein abundance and allow sampling of less abundant components that may be otherwise missed; and directed use of the mass spectrometer to acquire data for (inclusion list), or avoid (exclusion list) specific peptide ions (Scherl 2004; Zerck 2009). Further improvements to mass spectrometry instrumentation also continue to address these issues. For example, only a few years previously typical duty cycle rates were in the order of 3–4 fragmentation spectra per second. More recent instruments operate routinely at 6–8 fragmentation spectra per second, and newer systems claim over 50 fragmentation spectra per second. Further, the improvement in ion transmission and packaging in recent mass spectrometry platforms increases the sensitivity of analyses. Technological advances of this kind should improve the quality of mass spectral data and greatly alleviate the 'crowding' issue, and to a lesser extent the dynamic range of abundance issue, which should improve the depth of proteomic coverage, and repeatability and reproducibility of analyses. Indeed, the focus for improving the efficiency and effective use of mass spectrometry in proteomics may be shifting away from physical sampling issues and towards downstream informatics. This is evident in an inter-laboratory investigation reported by Bell (2009) to assess the reproducibility of proteomic analysis using a relatively simple mixture of 20 proteins designed to test analysis of 'crowded' samples. In this assessment, although only one out of 27 laboratories correctly and comprehensively identified a set of specific analytes, inspection of the raw datasets revealed that most laboratories did detect the targets but failed to report them. This highlighted informatics data processing and reporting also as a significant barrier to analytical completeness, in addition to instrument limitations.

Improving the translation of peptide fragmentation mass spectra into peptide identities

In typical peptide-centric experiments, the mass spectra are assigned to specific peptide sequences using probabilistic, stochastic or descriptive model-based algorithms, such as in the Sequest (Eng 1994), Mascot (Perkins 1999), Phenyx (Colinge 2003) and X-Tandem! (Craig 2004) search engines recently compared by Dagda (2010). The assigned peptide subsets are grouped under protein identities for which the peptide sets are most conceivably derived from the given experimental conditions. Duncan (2010) thoroughly described critical assumptions that affect the outcome and effectiveness of this process. One of the key considerations is that the fragmentation mass spectra are not interpreted by the algorithms, but are instead assigned an identity by pattern matching of the fragment ions against in silico-derived theoretical fragmentation patterns of peptide sequences. The theoretical patterns are calculated based on the experimental conditions applied to a pre-programmed protein set or database, such as the human proteome. These conditions are often expected to generate a specific character of peptide, for instance tryptic digestion generates peptide produced by cleavage of the protein sequence C-terminal to lysine and arginine residues, and

alkylation of cysteine residues introduces a specific mass adduct at this amino acid. The data interrogation strategy is therefore essentially hypothesis-driven, even though the proteomic experiment itself may not be, and it is assumed that the protein database and modification set used suitably represent the sequences and modifications present in the sample. However, this is rarely the case, and incompleteness of the database used, the presence of errors, discrepancies or variability in the sequences, and presence of unanticipated proteins and modifications in the biological sample analysed contribute to a failure to comprehensively interpret the mass spectra dataset. The mass spectral data are therefore under-utilised and the extent of both qualitative and quantitative biological information recorded in the experiments is under-represented in the final interpretations.

Several alternative strategies to better interpret mass spectral data have been emerging in recent years to attempt less rigidly hypothesis-driven data interrogation, and to sequence peptide fragmentation spectra without reference to theoretical spectral patterns derived from protein sequence databases. Consequently, these more discovery type approaches seek to bypass the assumptions of protein content, and modification, of samples inherent in the more hypothesis-driven search strategies used by the more common spectral analysis algorithms. Eliminating some, if not all, of the assumptions in these essentially de novo sequencing strategies may potentially improve the effective application of the peptide-centric techniques, for example in personalised medicine applications where genomes and proteomes of individuals can incorporate unpredictable variations that are unaccounted for in existing protein databases.

While de novo tools are not new, their effectiveness at interpreting mass spectra presently falls short for routine proteomics applications (Kim 2009a; Chi 2010). However, their utility is markedly improving, for example the recent demonstration of challenging de novo sequencing of antibodies by combining spectral alignments from multi-enzyme digests into spectral contigs and ultimately protein contigs, using a similar strategy to the assembly of DNA fragments for sequencing (Bandeira 2008). Other new informatic approaches include the recent spectral network for identification of mass-shifted spectral pairs (Bandeira 2007), spectral dictionary construction of a spectral library of possible de novo sequence (and modification) solutions to mass spectral data (Kim 2009b) and development of the new Vonode algorithm to derive 'consensus sequence tags' to match and score de novo mass spectra data against (Pan 2010). Tools of these kinds may not only improve the sequence assignment of spectra, in general, but may also more precisely locate sites of modifications, which are poorly determined by commonly used database interrogation algorithms and must frequently be validated by manual inspection of the tandem mass spectra. Further developments include improving the extent of correct de novo sequence assignments by analysing high resolution data obtained using different fragmentation modes (Chi 2010), as well as the development of approaches to predict features of peptide fragmentation spectra, such as relative intensities of fragment ions, to improve spectral matching and scoring (Frank 2009a, b). Maximising the use of higher resolution and higher mass accuracy data, and building in predictive features into mass spectral interpretation, would benefit both the current database-search strategies as well as the developing de novo sequencing strategies. However, continuing to develop all the above aspects of informatics processing of mass spectral data is essential to maximise the accuracy and depth of information obtained from the interpretation of increasingly higher-throughput analyses of chemically dynamic proteomes.

Deriving protein identifications from peptide-centric data: problems with protein inference and unanticipated variation in protein sequence and modification

Peptide-centric experiments suffer from an inherent disconnection of the protein level information from the peptides analysed. This manifests as a protein inference problem (Nesvizhskii 2005) exemplified by the observation that some peptide sequences may be present in several different proteins, and therefore the qualitative and quantitative interpretations based on these sequences are not unambiguously representative for a specific gene product. The problem not only complicates analysis of proteins with homologous domains but can also extend more subtly to functional interpretations of single gene products for which peptide sequences may be actually unique. For example, in whole cell lysates, peptides uniquely representative of a single gene product may be a combined pool of peptides from differentially translocated and/or modified forms of the protein that may be undertaking different functions. Consequently, this may impact on the biological interpretation of the data. Conceptually, the problem would be resolved by analysing unique discriminants for each protein residing in specific locations or in specific modification states, but clearly this cannot be easily achieved.

The use of 'proteotypic peptides' is an important first step towards achieving the analysis of unique protein discriminants. The term 'proteotypic peptide' refers to the relatively few but preferentially detected peptides in the pool of peptides generated from proteolytic digestion of each analyte protein in a sample (Mallick 2007). The preferential detection is related to the individual physicochemical properties of each peptide produced by proteolysis. Identification of the key properties that contribute to a

peptide's observability during mass spectrometry-based proteomic experiments can be used to predict the subset of proteotypic peptides in a known protein sample, and unsurprisingly peptide charge properties have been used to this effect. Several recent publications further describe new methods to improve upon the prediction and utility of proteotypic peptides. Support vector machine models (Webb-Robertson 2010b) and Random forest classifiers (Fusaro 2009) have both been applied to identify proteotypic peptide characteristics. Recurring themes reported include the recognition of increased positive charge peptide characteristics as an indicator for proteotypic peptides, and conversely identification of cysteine as a classifier for 'non-proteotypic' peptides. Information of this kind can then be further used to improve the efficiency and accuracy of peptide identification from fragmentation spectra by data interrogation against databases distilled to primarily contain theoretical or predicted proteotypic peptide sequences (Webb-Robertson 2010a).

Proteotypic peptides are potentially very valuable discriminants for specific proteins as long as they are also unique 'signature' peptides for their gene products. However, the possibility of modifications occurring within the proteotypic peptide sequences must be considered as this can impact particularly on the utility of these sequences for quantifying the parent protein. Ideally, proteotypic peptides should not be modified, either in vivo or during sample preparation, and the assumed non-influence of modifications on the use of proteotypic peptides will become more certain as our depth of knowledge of particularly in vivo modifications increases. Also, these peptides alone do not necessarily discriminate the functional or spatial information of the gene product, and it is unclear whether proteotypic peptides can be identified or generated that will specifically do so. Importantly, it should also be noted that specific analysis of proteotypic peptides necessarily only analyses a fraction of each full protein sequence in question, for instance in selected reaction monitoring (SRM) experiments. This strategy therefore precludes comprehensive proteomic analysis since many dynamic modifications occurring in non-proteotypic sequences will not be monitored. However, the strategy does provide a baseline representation of the gene product population essential for quantitative experiments.

Other approaches to address aspects of the protein inference problem include an elegant informatic approach to more confidently infer protein interpretations of the peptide data by considering genetic models relevant to the sample analysed (Qeli 2010), and a Bayesian approach to improve upon the commonly used ProteinProphet algorithm (Nesvizhskii 2003) to return a higher true-positive and lower false-positive protein identifications (Li 2009). Although such informatic approaches may improve the data interpretation, they cannot entirely resolve the ambiguous origin of some peptides in peptide-centric experiments. However, appropriate experimental design, for instance using sample fractionation to specifically focus the analyses on cellular locations, e.g., vesicles, nuclei, mitochondria, the plasma membrane, etc., or enrichment for specific modification subsets, e.g., phosphorylation or ubiquitination, can also reduce some of the complications of protein inference and enable more relevant and contextually focussed interpretation of the peptide mass spectral data.

Focussing the experiments from the outset also has utility in enabling a more thorough informatic assessment of the consequently focussed data using the currently more common database interrogation search strategies. For instance, in typical experiments to elucidate protein interaction partners, analysis of the affinity-purified target protein together with putative interaction partner co-eluents provides a much simpler, and focussed, dataset to interpret, which can be interrogated more effectively than say a dataset representing an entire cell lysate. For example, error tolerant searches of key proteins can be used to probe for unexpected modifications and amino acid substitutions/database errors by including a myriad of potential modifications as variables in the search parameters, for example the unimod list of modifications referred to by the Mascot search engine. This strategy can identify unanticipated modifications in the samples, including true biological modifications, and also accounting for amino acid mutations or polymorphisms, protein sequence truncations and unexpected cleavages during sample proteolysis. Although increasing the number of sequence and modification combinations and permutations under consideration also increases the number of false-positive identifications and assignments returned; this effect should become reduced with the higher resolution and higher accuracy mass spectral data promised by the latest developments in instrumentation. Further, the strategy is not easily applied to large datasets or high numbers of proteins selected for error tolerant searching, because of the high level of computational power needed to compute and probe for the myriad of modification combinations. Clearly, the approach becomes freely applicable as increasing computational power becomes generally available. However, de novo search strategies may be able to more efficiently achieve similar, or potentially more accurate, results as the methods mature.

Key challenges in quantitative analyses

The general approaches in quantitative interpretation of proteomics data and the software tools available have been recently comprehensively reviewed by Mueller (2008).

The strengths and weaknesses of label-free quantitative approaches such as spectral counting or direct comparison of ion current measurements, as well as the differential stable isotope labelling approaches including isobaric tagging, SILAC, ICAT and heavy oxygen incorporation are well described and will not be further discussed here. This section of the review will instead discuss some additional specific aspects pertinent to discovery style quantitative experiments, in particular SILAC (Ong 2002) and isobaric tagging strategies (Thompson 2003; Ross 2004), although some considerations are also relevant to targeted analytical approaches as well.

In SILAC and isobaric tagging quantitative experiments, samples are differentially labelled in groups, frequently in pairs of condition versus control. For example, SILAC heavy versus SILAC light isotopic labels, or isobaric tagging with paired or grouped tagging arrangements. Ion currents derived from peptides from the sample states are then compared to determine relative changes, and the peptide results grouped and collated to infer changes to the proteins to which each peptide set is attributed. However, variations exist in the data processing methods used. This can include calculating intensity (abundance) ratios for individual peptides to be grouped under a protein identification followed by calculating a mean, median or weighted average ratio from the peptide set to infer quantitative change, or summing the intensity values for individual peptides similarly grouped under a protein identification to infer a quantitative protein measure followed by calculating the ratio of change at the protein level. Sample normalisation methods can also vary, and are frequently performed by calculating global correction factors using the global mean, median or weighted average of the dataset ratios. Differences in the effectiveness of these methods have been reported, for instance calculating the quantitative protein ratios based on the mean of the protein-assigned peptide ratios can under-represent the magnitude of the protein ratio. Calculation of a weighted average can afford more accurate protein quantitation resulting from decreasing the impact of lower quality measurements on the protein quantitation calculation (Park 2008). Alternatively, using the median of the peptide ratios may afford a similarly improved result. It is important to note that there is no consensus on a 'correct' or 'best' strategy to perform the quantitative analysis, and the most appropriate approach can be entirely dependent on the quantitative technology used, experimental design and sample setup. However, all these approaches are peptide-centric in nature, and several common factors come into play that can affect the quantitative data interpretation. Key factors include influences on sample normalisation, effects of peptide modifications, and how data can be statistically analysed, which are discussed, in turn, below.

Sample normalisation

Before data can be quantitatively interpreted, data are usually normalised to account for differences due to loading variations and other similar global effects. In SILAC and isobaric tagging experiments, this is frequently performed by calculating the global mean or median ratio of change in log space for an entire peptide dataset. The deviation of the ratio from unity (log space zero) represents a normalisation factor to re-align the data and correct for any global offset. In other types of quantitative experiments, for instance label-free approaches directly comparing raw ion currents, a similarly effective global normalisation can be derived from the ratio of the total ion current (TIC) of acquisition runs, or from summation of intensities from specific peptide sets, between the samples for relative comparison. In all these cases, application of a global normalisation parameter assumes any protein changes between samples do not markedly affect the global mean/median log ratio or TIC ratio. This does not necessarily hold true. For example, if an entire sub-proteome has altered in response to a condition, the shift in log ratios or TIC may skew the normalisation factor applied to the samples. This is experiment dependent, and may be due to real biological responses or artefacts of preparation. Indeed, effective enrichment of sub-proteomes has been used to identify proteins segregating with different organelles (reviewed in Walther and Mann 2010). If the enrichment is unanticipated, this can result in the application of a misleading normalising factor and subsequently biased quantitation. We have observed such an effect previously in experiments quantifying protein changes in membrane rafts (Fig. 1). In some of these datasets, mitochondrial and cytoskeletal-associated sub-proteomes were differentially represented in the samples. These sub-proteomes constituted a high enough proportion of the entire peptide dataset to introduce a 20 % skew in the global dataset ratio in some cases if included for normalisation. Identification of the differential representation of the sub-proteomes by ontological annotation of the peptide data enabled a more appropriate normalisation of the datasets by excluding these sub-proteomes from the calculation of a normalisation factor, which was therefore based on the remaining proteins that constituted a largely unchanging baseline protein subset. In general, the analysis of sub-fractionated samples would be expected to be most susceptible to these effects, since translocation and/or recruitment of vesicles and other cellular compartments may be differentiated between samples. However, it is also possible that this effect may be observed in the comparison of whole cell lysates, for instance if morphological differences are observed between samples there may exist sufficient differences in the expression of abundant

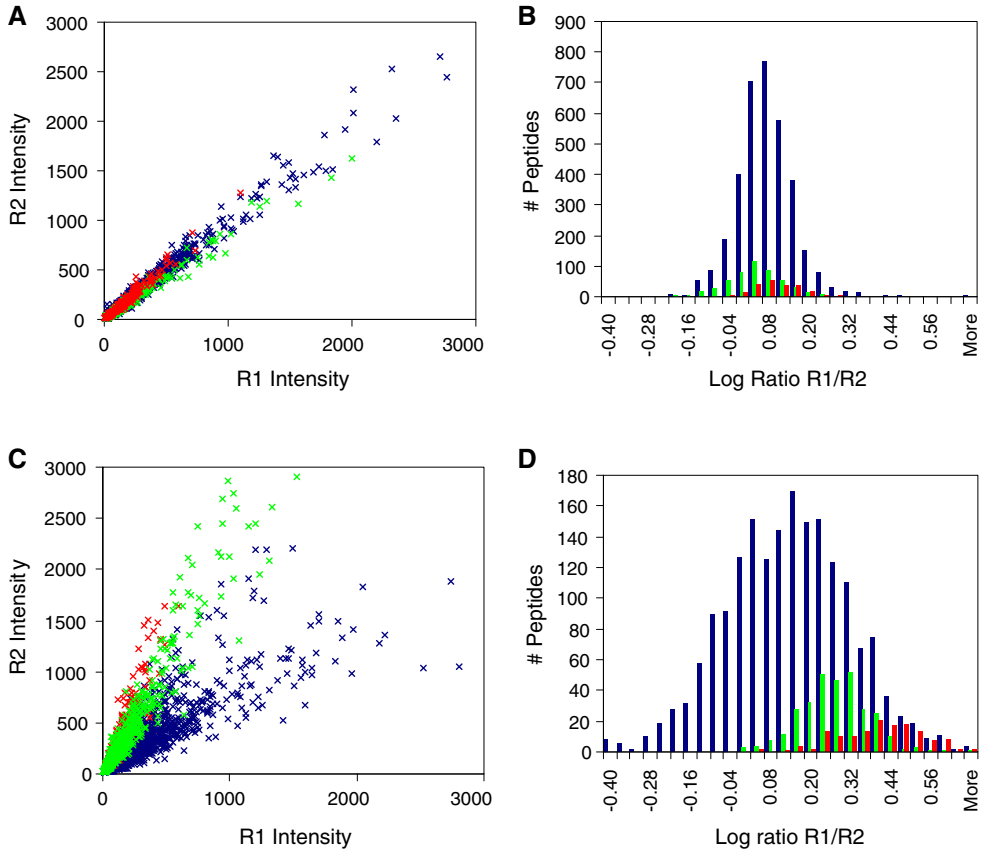

Fig. 1 Demonstration of shifts in sub-proteomes within a quantitative dataset. In this example of iTRAQ quantitation of membrane raft samples, mitochondrial and cytoskeletal-associated proteins were observed to occasionally exhibit markedly different behaviours relative to the remaining (bulk) population of proteins. Proteins were annotated using information from the Swiss-Prot database. **a** Plot of reporter ion intensities (R1 and R2) for a sample not displaying significant sub-proteome shifts. **b** Histogram plot of the frequency of peptides with log intensity ratio (R1/R2) values in 0.04 width bins. Data are shown before normalisation against the global dataset ratio median. Normalised data should centre on 0.00 (unity ratio). **c** Reporter ion intensity plot for a sample displaying sub-proteome shifts. **d** Histogram plot of the data exhibiting shifts in sub-proteomes. Cytoskeletal and associated and mitochondrial proteomes are clearly right-shifted compared to the global dataset, indicating much higher abundance of these proteins in sample R1 compared to sample R2. *Colours* represent data attributed to cytoskeletal and associated proteins (*green*), mitochondrial proteins (*red*), or all other categories (*blue*)

proteins, such as cytoskeletal elements, that a blanket global normalisation approach may also become biased. Analysis of samples at different phases of the cell cycle may similarly be affected. A possible solution to address this issue is to annotate the quantitative data with functional and/or spatial biological information, and to compare the quantitative ratios of these annotated data subsets to identify potential influences. This may enable the identification of an 'unchanging' baseline protein population for the purposes of sample normalisation.

The normalisation issues discussed thus far concern variations integral to the samples being analysed. Extraneous variations also affect normalisation processes, especially with regard to inter-run quantitative comparisons. Multiplex-style experiments have an advantage in that peptides to be relatively quantified can be combined together in a single LC–MS/MS acquisition, which can reduce, or even eliminate, issues with inter-run variation. However, inter-run variation can be particularly problematic for label-free quantitative mass spectrometry strategies, which by their nature require the comparison of data acquired in discrete LC–MS/MS runs. Inter-run variations can include systematic biases inducing global drifts in data quality, as well as non-systematic variations such as sporadic appearance of contaminants between runs. Normalising data affected by these different biases and variations require further solutions. Kultima (2009) rigorously investigated ten different normalisation methods for label-free peptide quantitation and concluded that a novel procedure (RegrRun), combining linear regression with further normalisation to account for a bias observed for the analytical sample run order, produced the most precise results. This was in addition to pre-filtering data to only analyse MS peaks matched across at least 50 % of the data set,

which would be expected to account for interference from sporadic contaminants. Similarly, Callister (2006) only analysed peptides common to all LC–MS runs in a critical assessment of four different normalisation approaches, concluding that linear regression was most effective at reducing systematic bias in their data. Other sources of variation can arise from the nature of the quantitative measures themselves. For instance, the considerable differences observed in the intensity of different peptide ion currents, which is largely dependent on the physico-chemical properties of the peptides themselves, result in widely varying heterogeneity of quantitative variance for different peptides detected within the same experiment. This subsequently and adversely affects the precision of protein quantitation. Clearly, lower intensity signals suffer the most variance, and thresholds have been applied to filter out high variance signals that may adversely affect quantitative calculations (Thompson 2009). More recently, Karp et al. (2010) applied a variance stabilising normalisation (VSN) method to improve the precision of the quantitative peptide measurements and to avoid imposing ad hoc filtering rules. The VSN approach presents a significant advance in quantitative methods as it aims to maximally retain data for analysis by stabilising the variance over a wide range of signal intensities. Importantly, the authors of these publications stress the need to further improve upon current normalisation techniques, which should be more effectively achieved as the sources of experimental variation are identified and their influences are accounted for. An important step towards this is the recent publication of 46 metrics recommended to assess the performance of LC–MS platforms (Rudnick 2010). These metrics functionally monitor liquid chromatography, mass spectrometry (both ionisation and software-controlled dynamic sampling) and informatic data analysis. The further development and use of comprehensive metrics of this kind should identify significant sources of variation to be subsequently controlled. Mass spectrometry-based quantitative proteomics applications are clearly powerful, but if they are to fulfil a potential for routine general use, such as in robust clinical diagnostic and prognostic assays, identifying and addressing the sources of variation, including further improving data normalisation, will be required.

Peptide modifications

A critical assumption in the peptide-centric approach is that the relative changes calculated for each peptide is representative of change in the relative abundance of the parent protein. However, this may not always be the case, especially when considering that peptide levels may quantify differently if a modification has changed between the sample states, irrespective of the parent protein abundance has changed or not. Peptide modifications may be true biological changes such as in phosphorylation sites, or may be artifactually introduced during sample processing, such as by deliberate alkylation of cysteine residues or by undesired, but often tolerated, oxidation of methionine. Modifications like these not only increase sample complexity by diluting peptide sequences across differentially modified isoforms but can also affect quantitative accuracy if the extent of the peptide modifications varies between sample preparations. This can occur despite parallel preparation of samples for quantitation using identically sourced reagents, which may have been expected to maintain consistency in the reactions and processing of the samples. For example, we have observed differential oxidation of methionine in the parallel processing of simple standard protein mixtures used to assess quantitation by iTRAQ (Fig. 2). In this experiment, principle component analysis for selected variables highlighted methionine-containing peptides as exhibiting a distinct behaviour, which was subsequently identified as a difference in the extent of methionine oxidation and was observable as a significantly lower quantitative ratio of 2.45 for the

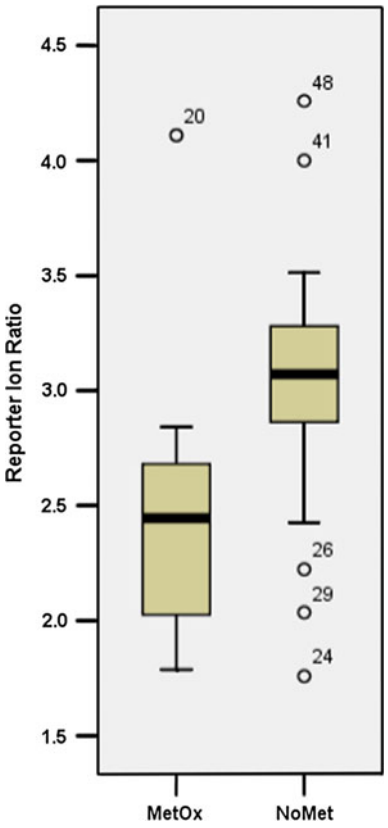

Fig. 2 Demonstration of distinct differential quantitation of modified peptides. In this example, analysis of transferrin protein in a test iTRAQ 4-plex experiment, the subset of methionine-containing peptides quantified with a significantly lower ratio compared to the greater non-methionine-containing subset of peptides ($p < 0.001$). The expected protein ratio for this experiment was 3.3

methionine oxide peptide subset compared to a ratio of 3.2 for all non-methionine-containing peptides. Therefore, inclusion of the methionine peptide data to calculate the protein ratio resulted in an under-representation of the true protein difference. If this effect can be observed for methionine oxidation, similar effects could be expected to also occur for other common side-reactions such as pyro-glutamisation of peptide N-terminal glutamine, deamidation of asparagine/glutamine, as well as the digestion process itself including retarded cleavage at lysine/arginine when adjacent to aspartic acid/glutamic acid. While this in no way implies that peptide-centric quantitation is faulty, it does suggest that further checks on the integrity of the data before quantitative interpretation, such as by principle component analysis to investigate the behaviour of peptides incorporating amino acids susceptible to modification, may help improve the quality and accuracy of the results. Importantly, such approaches may also identify true, though unanticipated, biological modifications by identifying peptide sequences that exhibit a different quantitative behaviour to other peptides attributed to the same protein precursor.

'Missing data' and statistical analysis

Robust statistical analysis of SILAC and iTRAQ quantitative strategies can suffer from a 'missing data' problem. This arises because replicate sample analyses are required to generate sufficient datasets for robust statistical assessments by *t* test or ANOVA since the multiplexing capabilities of the experiments are low—SILAC experiments are commonly employed in a 2-plex (heavy versus light) or 3-plex (heavy vs. medium vs. light) internal experiment arrangement, and isobaric tags were originally released as a 4-plex labelling set. As discussed previously, proteomic experiments, particularly of the more complex biological matrices, exhibit stochastic sampling effects. As a result, replicate runs of samples do not detect identical peptide sets, and proteins identified based on fewer peptides also may not be reproducibly detected, so the replicate experiments therefore suffer from 'missing data' that undermines application of statistical tests such as ANOVA. Recently, Webb-Robertson (2010a) sought to address this issue by combining a *G* test filter with the ANOVA calculation, and in earlier work, Oberg (2008) suggested the use of iterative regression to facilitate the application of ANOVA. However, the release of 6-plex and 8-plex isobaric tagging reagents enables another significant but simpler solution to the issue. The higher multiplexing capability allows at least three replicates of two sample states to be internally analysed in an experiment. The impact of this is threefold, as reported recently (Engmann 2010). First, since the same peptide data are acquired for all the isobaric tagged samples, quantitation is based on the observation of the same peptides for all the samples to be compared, avoiding inter-run technical variance and also eliminating the 'missing data' problem caused by stochastic sampling issues. Secondly, since measurements of the same peptides are recorded for all the samples, quantitative intensity measurements can be summed to represent a quantitative protein value to be directly and statistically compared between the samples. In the more usual experiments at present, stochastic sampling can complicate the summing approach since different peptides ionise with different efficiencies, and therefore quantitative results should be carefully considered if calculated based on different peptide sets observed for the same protein between independent acquisitions. Thirdly, since at least a three versus three sample tagging arrangement can be incorporated, statistical tests such as *t* test and ANOVA can be used in a straightforward manner to determine both the variance within each triplicate set as well as the significance of changes between the triplicate sets. Again, this is made possible since the measurements are made for the same peptides in all the samples in a single acquisition run. These factors have been described in relation to a simple 2-way sample comparison. In the event that the analysis of more than two sample types in an experiment is desired, therefore necessitating independent acquisition runs, three channels in each experiment can be reserved for a common control. This type of experimental setup enables as many samples to be compared and analysed by robust statistical tests that assess both the internal variance of each sample type as well as the significance of differences between samples, via the internal common control, and limited only by the capacity of the instrument platform to perform the multiple independent acquisitions required. In essence, the higher multiplexing capacities for the more recently available isobaric tags enable statistically robust quantitation based on a consistent dataset, avoiding issues related to both stochastic sampling and circumventing the essentially arbitrary pairing of samples commonly used to calculate peptide and protein ratios in typical quantitative experiments.

Conclusions

Proteomic technologies are maturing to an extent to enable accurate quantitation of both proteins and modifications from complex biological matrices. Routine applications uncover a wealth of qualitative and quantitative biological information, but despite this the remarkable results could be potentially much more informative. The multidisciplinary nature of applied proteomics unsurprisingly necessitates multidisciplinary solutions: physical advances

in instrumentation enable complex biological matrices to be more comprehensively sampled; the development of new informatic algorithms promise to more comprehensively translate mass spectra into biological information; and smart chemistry advances enable robust experimental design for confident quantitative analyses. Continuing these developments, and others, to improve upon the already impressive applications of proteomic technologies can only increase the utility of especially the quantitative aspects of the technique. Not only may whole proteome characterisation soon be possible but personalised temporal characterisation of individual proteomes may also be subsequently enabled, to facilitate specific tailoring of medical treatments.

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson NL, Anderson NG (2002) The human plasma proteome: history, character and diagnostic prospects. Mol Cell Proteomics 1:845–867

Bandeira N, Tsur D, Frank A, Pevzner PA (2007) Protein identification by spectral networks analysis. Proc Natl Acad Sci 104:6140–6145

Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR (2008) Beyond Edman degradation: automated de novo protein sequencing of monoclonal antibodies. Nat Biotechnol 26:1336–1338

Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJM, HUPO Test Sample Working Group (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods 6:423–430

Callister SJ, Barry RC, Adkins JN, Johnson ET, Qian W-J, Webb-Robertson B-JM, Smith RD, Lipton MS (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. J Proteome Res 5:277–286

Chi H, Sun R-X, Yang B, Song C-Q, Wang L-H, Liu C, Fu Y, Yuan Z-F, Wang H-P, He S-M, Dong M-Q (2010) pNovo: de novo sequencing and identification using HCD spectra. J Proteome Res 9:2713–2724

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3:1454–1463

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualised p.p.b-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26:1367–1372

Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467

Dagda RK, Sultana T, Lyons-Weiler J (2010) Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics. J Proteomics Bioinform 3:39–47

Dayon L, Hainard A, Licker V, Turck N, Kuhn K, Hochstrasser DF, Burkhard PR, Sanchez JC (2008) Relative quantitation of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. Anal Chem 80:3372–3378

Domon B, Aebersold R (2010) Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol 28:710–721

Duncan MW, Aebersold R, Caprioli RM (2010) The pros and cons of peptide-centric proteomics (2010). Nat Biotechnol 28:659–664

Eng JK, McCormack AL, Yates JR III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5:976–989

Engmann O, Campbell J, Ward M, Giese KP, Thompson AJ (2010) Comparison of a protein-level and peptide-level labelling strategy for quantitative proteomics of synaptosomes using isobaric tags. J Proteome Res 9:2725–2733

Frank AM (2009a) Predicting intensity ranks of peptide fragment ions. J Proteome Res 8:2226–2240

Frank AM (2009b) A ranking-based scoring function for peptide-spectrum matches. J Proteome Res 8:2241–2252

Fusaro VA, Mani DR, Mesirov JP, Carr SA (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol 27:190–198

Gavin A-C, Maeda K, Kuhner S (2010) Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. Curr Opin Biotechnol 22:1–8

Glen A, Evans CA, Gan CS, Cross SS, Hamdy FC, Gibbins J, Lippitt J, Eaton CL, Noirel J, Wright PC, Rehman I (2010) Eight-plex iTRAQ analysis of variant metastatic human prostate cancer cells identifies candidate biomarkers of progression: an exploratory study. Prostate 70:131–1332

Jorgensen C, Sherman A, Chen GI, Pasculescu A, Poliakov A, Hsiung M, Wilkinson DG, Linding R, Pawson T (2009) Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. Science 326:1502–1509

Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS (2010) Addressing accuracy and precision issues in iTRAQ quantitation. Mol Cell Proteomics 9:1885–1897

Kim S, Bandeira N, Pevzner PA (2009a) Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. Mol Cell Proteomics 8:1391–1400

Kim S, Gupta N, Bandeira N, Pevzner PA (2009b) Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. Mol Cell Proteomics 8:53–69

Kultima K, Nilsson A, Scholz B, Rossbach UL, Falth M, Andren PE (2009) Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. Mol Cell Proteomics 8:2285–2295

Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, Tang H (2009) A Bayesian approach to protein inference problem in shotgun proteomics. J Comput Biol 16:1183–1193

Liu H, Sadygov RG, Yates III JR (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201

Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R (2007) Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25:125–131

Mueller LN, Brusniak M-Y, Mani DR, Aebersold R (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 7:51–61

Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomics data: the protein inference problem. Mol Cell Proteomics 4:1419–1440

Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying protein by tandem mass spectrometry. Anal Chem 75:4646–4658

Nilsson T, Mann M, Aebersold R, Yates III JR, Bairoch A, Bergeron JMJ (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods 7:681–685

Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, Cooper LT, Onuma OK, Spiro C, Therneau TM, Bergen HR III (2008) Statistical analysis of relative labelled mass spectrometry data from complex samples using ANOVA. J Proteome Res 7:225–233

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labelling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1:376–386

Pan C, Park BH, McDonald WH, Carey PA, Banfield JF, VerBerkmoes NC, Hettich RL, Samatova NF (2010) A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. BMC Bioinform 11:118

Park SK, Venable JD, Xu T, Yates III JR (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods 5:319–322

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3568

Pieper R, Su Q, Gatlin CL, Huang S-T, Anderson NL, Steiner S (2003) Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. Proteomics 3:422–432

Przbylski C, Junger MA, Aubertin J, Radvanyi F, Aebersold R, Pflieger D (2010) Quantitative analysis of protein complex constituents and their phosphorylation states on a LTQ-orbitrap instrument. J Proteome Res 9:5118–5132

Qeli E, Ahrens CH (2010) PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol 28:647–650

Rechavi O, Kalman M, Fang Y, Vernitsky H, Jacob-Hirsch J, Foster L, Kloog Y, Goldstein I (2010) Trans-SILAC: sorting out the non-cell-autonomous proteome. Nat Methods 7:923–927

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3:1154–1169

Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham A-JL, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-Montoto L, Variyath AM, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Carr SA, Fisher SJ, Gibson BW, Paulovich AG, Regnier FE, Rodriguez H, Spiegelman C, Tempst P, Liebler DC, Stein SE (2010) Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. Mol Cell Proteomics 9:225–241

Scherl A, Francois P, Converset V, Bento M, Burgess JA, Sanchez JC, Hochstrasser DF, Schrenzel J, Corthais GL (2004) Nonredundant mass spectrometry: a strategy to integrate mass spectrometry acquisition and analysis. Proteomics 4:917–927

Schirle M, Heurtier MA, Kuster B (2003) Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 2:1297–1305

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Hamon C (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75:1895–1904

Thompson AJ, Williamson R, Schofield E, Stephenson J, Hanger D, Anderton B (2009) Quantitation of glycogen synthase kinase-3 sensitive proteins in neuronal membrane rafts. Proteomics 9:3022–3035

van Duijn E (2010) Current limitations in native mass spectrometry based structural biology. J Am Soc Mass Spectrom 21:971–978

Walther TC, Mann M (2010) Mass spectrometry-based proteomics in cell biology. J Cell Biol 190:491–500

Washburn MP, Wolters D, Yates JR III (2001) Large scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242–247

Webb-Robertson B-JM, McCue LA, Waters KM, Matzke MM, Jacobs JM, Metz TO, Varnum SM, Pounds JG (2010a) Combined statistical analysis of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data. J Proteome Res 9:5748–5756

Webb-Robertson B-JM, Cannon WR, Oehmen CS, Shah AR, Gurumoorthi V, Waters KM (2010b) A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. Bioinformatics 26:1677–1683

Wepf A, Glatter T, Schmidt A, Aebersold R, Gstaiger M (2009) Quantitative interaction proteomics using mass spectrometry. Nat Methods 6:203–205

Zerck A, Nordhoff E, Resemann A, Mirgorodskaya E, Suckau D, Reinert K, Lehrach H, Gobom J (2009) An iterative strategy for precursor ion selection for LC-MS/MS based shotgun proteomics. J Proteome Res 8:3239–3251