Talanta 83 (2011) 1209–1224

Contents lists available at ScienceDirect

Talanta

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

Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC–MS for biomarker discovery

Christin Christin[a,b], Rainer Bischoff[a,b], Péter Horvatovich[a,b,*]

[a] Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands
[b] Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Available online 11 November 2010

Keywords:
Computational biology
Quantitative data processing
Statistical analysis and validation
Label-free quantification
Biomarker discovery
Comparative LC–MS profiling

ABSTRACT

Label-free quantitative LC–MS profiling of complex body fluids has become an important analytical tool for biomarker and biological knowledge discovery in the past decade. Accurate processing, statistical analysis and validation of acquired data diversified by the different types of mass spectrometers, mass spectrometer parameter settings and applied sample preparation steps are essential to answer complex life science research questions and understand the molecular mechanism of disease onset and developments. This review provides insight into the main modules of label-free data processing pipelines with statistical analysis and validation and discusses recent developments. Special emphasis is devoted to quality control methods, performance assessment of complete workflows and algorithms of individual modules. Finally, the review discusses the current state and trends in high throughput data processing and analysis solutions for users with little bioinformatics knowledge.

1. Introduction

The recent widespread application of mass spectrometry to quantify and identify large numbers of compounds in biological matrices leads to an explosion of acquired data. The goals of these measurements are to explore the underlying molecular mechanism of disease, to identify compounds (biomarkers) strongly related to the stage of the disease, its onset or progression for diagnostic purposes, to identify novel drug targets, and to follow the efficiency of treatment. The dynamic behavior of multifactorial diseases requires a systems biology approach to find reliable biomarkers taking molecular regulatory mechanisms, compound flux and concentration changes into account [1]. To explore robust changes in molecular systems related to disease, it is necessary to analyze a large number of samples from different biological entities, for example from different, clinically well characterized patient groups. Generally biomarker research is based on complex biological samples containing a large number of diverse compounds such as proteins, peptides and metabolites. Liquid chromatography coupled to mass spectrometry (LC–MS) is one of the most widely used comprehensive profiling techniques to measure compounds in biological materials. A single comprehensive LC–MS analysis cannot cover all types of compounds in the samples. Instead, it measures one class of compounds such as metabolites, lipids, and proteins leading to biomarker discovery in this class of molecules. Even with a technique targeting one of the above mentioned classes of compounds, not all types of molecules can be measured due to ionization limitations of the electrospray interface. Another challenging problem is the wide dynamic concentration range of the compounds, which can reach 9–11 orders of magnitude in the case of body fluids such as blood [2,3]. From this wide dynamic concentration range, modern mass spectrometers are only able to cover 2–4 orders of magnitude. The gap between the existing and

Abbreviations: 2D-LC–MS, two dimensional liquid chromatography coupled to mass spectrometry; AMT, accurate mass and retention time tag; APEX, absolute protein expression; APML, annotated putative peptide markup language; Cap-LC–MS, LC–MS system equipped with capillary LC column (1 mm internal diameter) and using ionspray for ionization; Chip LC, LC–MS system equipped with nano-LC column (75 μm internal diameter) integrated in a microfluidic device and using electrospray for ionization; CID, collision induced dissociation; DDA, data dependent acquisition; emPAI, exponentially modified protein abundance; ETD, electron transfer dissociation; FTMS, Fourier transform ion cyclotron resonance mass spectrometry; HDSS, high dimensionality small sample size problem; HUPO PSI, Human Proteome Organization Proteomics Standard Initiative; LC, liquid chromatography; LC–MS, liquid chromatography coupled to mass spectrometry; MEND, matched filtration with experimental noise determination; MRM, multiple reaction monitoring; MS, mass spectrometry; MS/MS, fragment ion mass spectra of selected precursor ions; MS-1, single stage mass spectrometry; nano-LC, liquid chromatography using chromatographic column of internal diameter smaller than 100 μm; PTM, post translational modification; RSD, relative standard deviation; XML, extensible markup language.

* Corresponding author at: Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands. Tel.: +31 50 363 3341; fax: +31 50 363 7582.
E-mail address: p.l.horvatovich@rug.nl (P. Horvatovich).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.10.029

measurable dynamic concentration range can be reduced by using comprehensive fractionation (4–6 orders of magnitude), multidimensional chromatography (up to 8 orders of magnitude) [4] or targeting a specific subclass of compounds, *e.g.* by using an affinity enrichment step of a certain type of glycoproteins on a lectin column (up to 5–7 orders of magnitude) [5]. Another challenging factor is that although proteins and protein complexes are directly involved in the molecular processes of biological phenomena, their peptide constituents obtained after enzymatic cleavage are measured since they are more suitable for liquid chromatography analysis and have better ionization properties than intact proteins or protein complexes. The most widely used endopeptidases cut proteins at well-defined sequence positions, resulting in non-overlapping peptides mixtures, from which only a fraction of theoretical possible peptides are detected. In this peptide centric approach also called as "bottom-up", or "shotgun" strategy, the quantity of initial proteins is determined indirectly based on few or more peptides, which leads to misleading quantification and identification in the presence of multiple highly homologous proteins having one or few peptides in common, proteins with multiple splice variants, proteins presenting different degrees of post-translation modifications (PTMs) or in the presence of various truncated forms of the same protein [6,7].

Biomarker discovery requires close collaboration between medical researchers, analytical chemists and bioinformaticians in order to obtain the relevant molecular information related to different aspects of disease [8,9]. This includes patient cohort selection, sampling of the biological material, sample storage, sample preparation, choice and optimization of LC–MS profiling platform, data analysis providing protein identifications, quantification, statistical analysis and experimental validation of the results. Several review papers describe the various techniques and steps of the protein profiling for biomarker discovery in detail [9,10].

Bioinformatics plays an important role in this process as it has the goal to extract quantitative and qualitative information for a large number of compounds (proteins and metabolites) that are present in complex biological samples and to select the discriminatory compounds between predefined sample sets. Recent advances in sample preparation methods, liquid chromatography and mass spectrometry instrumentation resulted in a large diversity of acquired data. This results in a huge challenge for bioinformatics to provide reliable information extraction and knowledge generation approaches. The computational tools must evolve continuously to keep up with the different types of generated data. Besides direct information extraction and knowledge discovery from raw data, bioinformatics plays an important role in experimental design, quality assessment of the profiling platform, sampling methods, sample handling, storage and preparation methods, or quality control of data pre-processing, statistical analysis and statistical validation.

This review focuses on fundamental data processing and current challenges in supporting biomarker discovery research in proteomics for diagnosis and treatment follow-up using LC–MS of label-free, shotgun proteomics data, highlighting significant innovations in the bioinformatics field such as new algorithms, data integration, high throughput automatic data preprocessing solutions, quality control of different data processing modules and complete workflows, including assessment of the quality of sample preparation steps and LC–MS profiling platforms [9,11–19]. We will also investigate how insights from analytical chemistry contribute to parameter optimization leading to the development of novel bioinformatics applications that provide more accurate and reliable information extraction from the raw data. Alternative approaches based on differential labeling of samples with reagents having the same chemical but different stable isotope constitution have been covered in other reviews [20–27] and will not be treated here. This review limits the discussion further to biomarker discovery aiming to determine comprehensively the identity and quantity of sample constituting proteins using analytical methods with low sample throughput. Biomarker validation using analytical methods with high sample throughput providing quantitative information on preselected list of proteins by using analytical methods such as multiple reaction monitoring, antibody arrays and ELISA will not covered here. Recommendation on analytical, clinical and informatics aspects of biomarker discovery and validation as well their limitations was discussed recently by several reviews [28–34].

2. Data processing pipelines in LC–MS

LC–MS has become the major platform for analyzing samples in biomarker discovery research due to its relatively high throughput (60–90 min for analysis of one sample), sensitivity, selectivity and coverage of many peptides and proteins [9,35,36]. In label-free LC–MS experiments, proteins or produced tryptic peptides are not modified chemically and their isotope constitution is unchanged. In label-free experiments, a large number of samples are analyzed independently by LC–MS resulting in corresponding raw data files. The quantitative and compound identity information is extracted using dedicated data processing pipelines. This is followed by matching compound quantity and identity across several chromatograms resulting in a matrix containing quantitative information about a large number of compounds in the different samples. In shotgun proteomics approach the target compounds are proteins, therefore methods are required to determine the original protein composition of samples and their quantities based on incomplete set of measured constituting peptides. Compounds discriminating between predefined classes of samples are obtained from this matrix using dedicated statistical analysis and validation pipelines. When a systems biology approach is involved in the biomarker discovery process, it is necessary to couple the list of discriminating proteins to protein interaction (*e.g.* STRING, BIND) or pathway (*e.g.* KEGG) databases [21,37] to elucidate the disease mechanism. Fig. 1 shows the main parts of a generic proteomics pipeline for biomarker discovery.

Most of the signals measured by LC–MS are not related to real compounds but are part of white noise, background ions or simply chemical noise. Different mass analyzers generate data of different structure due to differences in scanning speed, mass resolution, measured dynamic concentration range, changes in peak width and resolution across the m/z domain and varying mass accuracy [38]. The most common mass analyzers applied in proteomics biomarker research are quadrupole, 3-dimensional quadrupole iontrap, 2-dimensional linear iontrap, time of flight, and inductively-coupled resonance (ICR) trap family of mass spectrometers such as Orbitrap and Fourier transform ion cyclotron resonance mass spectrometers (FTMS) [39]. Besides mass spectrometers may dispose different number of mass analyzers, and could use different ionization method such as electrospray, ionspray, matrix assisted laser desorption ionization (MALDI) to name the most frequently used method to analyze proteomics samples. In label-free LC–MS proteomics experiments, there are two types of widely used mass spectrometry data. The first data type contains mass spectra obtained with one mass analyzers and is referred to as single stage mass spectrometry data (MS-1) in the literature. The second data type is heterogeneous and contains cyclic series of MS-1 and precursor ion fragmented spectra (MS/MS). Each cycle begins with MS-1 spectra, then it is followed by a defined number (generally 1–10) of MS/MS spectra obtained from the most abundant ions of the MS-1 spectra. This acquisition mode is referred to as data dependent acquisition (DDA) and abbreviated as DDA MS/MS data. The reader is referred to dedicated books [38,40,41] and reviews [39,42,43] for further reading on the main character-

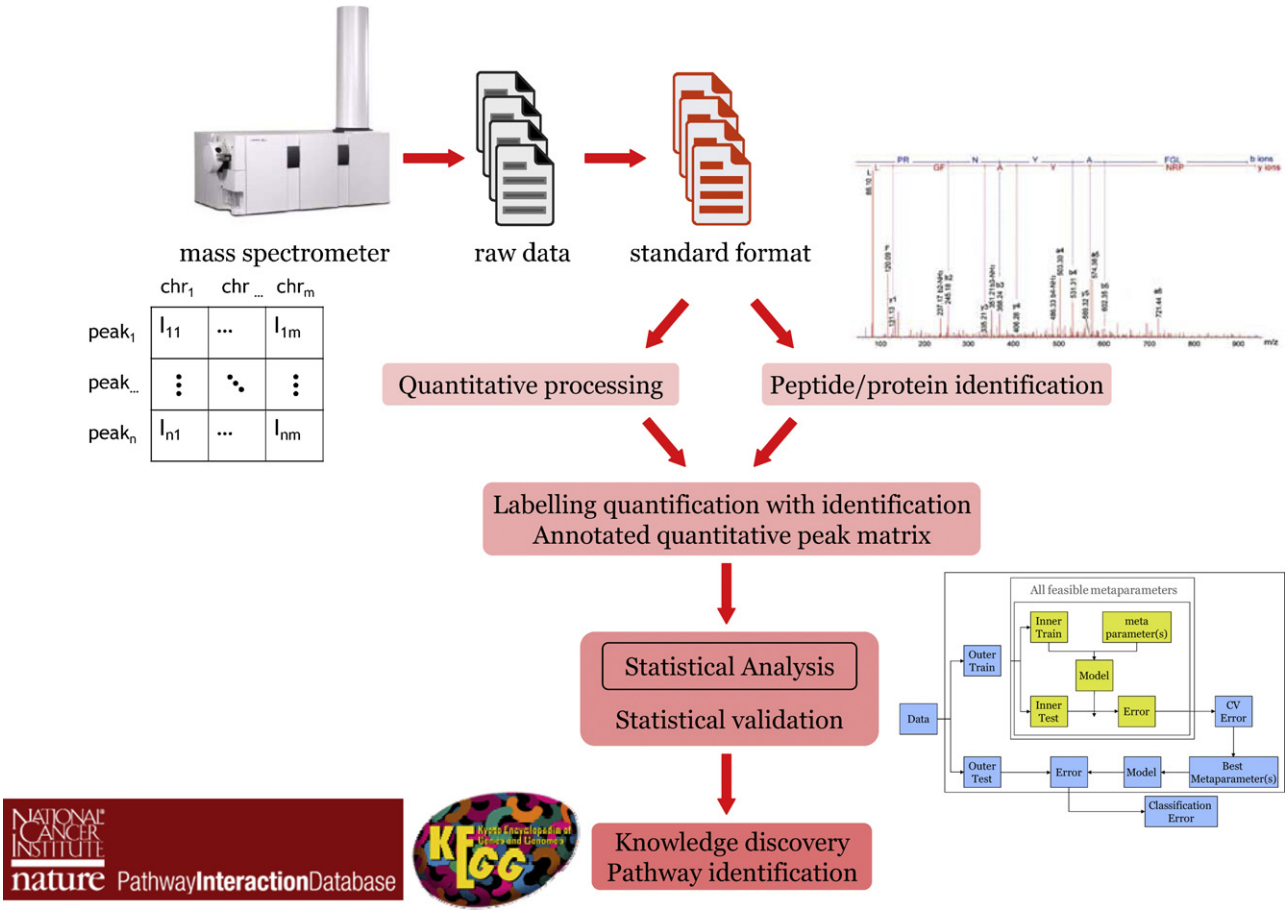

Fig. 1. Main modules of a generic biomarker data processing workflow. Raw data from the mass spectrometer are converted into one of the standard data formats such as mzXML, mzData or mzML. Quantitative information and identification of proteins and peptides are performed separately from the same file or from a different data file. This is followed by labeling of the quantitative information with identifications. The statistical analysis and validation is performed on the labeled quantitative data and provides a list of discriminatory proteins that can be used for knowledge discovery with pathway analysis tools using for example KEGG (http://www.genome.jp/kegg/) or the Pathway Interaction Database (http://pid.nci.nih.gov/).

istics of different type of mass analyzers, ionization methods and acquisition modes. Label-free quantification is a semi-quantitative method and provides information on relative quantity changes of the same compounds in different samples. For most applications such as biomarker discovery, detection of relative protein changes is sufficient information, but in system biology type of studies, the use of stable isotope labeled standard is necessary to provide absolute quantity of proteins in samples [44].

Quantitative information can be obtained from both MS-1 and DDA MS/MS data. Quantitative methods using DDA MS/MS data are based on spectral counting, and use the number of MS/MS spectra that are acquired per peptide ion(s) for the quantification of a given protein. Abundant proteins generate abundant peptide fragments that have a higher probability to be selected as precursor ions for DDA MS/MS analysis. Nevertheless, in spite that spectral counts shows good linearity with analyzed protein amount [45,46], the number of MS/MS spectra per protein suffer from saturation effect, undersampling, and from the limited linear concentration range compared to MS1 quantification methods [47]. Spectral counting methods enable both absolute and relative quantification of proteins. Several bioinformatics methods use the spectral counting approach [46,48–50]. Exponentially modified protein abundance index (emPAI) [51,52] uses the number of identified peptides to calculate the relative molar or weight fraction of a given protein in the respective sample. Absolute protein expression (APEX) [53,54] uses the measured and predicted peptide counts for quantification of peptides and proteins by considering the influence of the recovery of peptides from the cation-exchange and reversed-phase LC dimensions as well as the predicted ionization efficiency of the peptide in the ion source of a particular mass spectrometer. Recently a new method, which combines the quantification of MS-1 and MS/MS spectra by taking the ion count in MS-1 of the three most abundant peptides provides better quantification for proteins than spectral counting and gives the absolute protein quantity by using a single protein standard [55]. DDA MS/MS measurement is subjected to large variability regarding the identified peptide and proteins [56], therefore more precise quantification providing a larger dynamic concentration range than spectral counting can be obtained using peptide ion counts in MS-1 data. Recently, a modified version of the MS/MS acquisition strategy called directed MS was introduced with modern high resolution Q-TOF and Orbitrap instruments. Directed MS differs from DDA MS/MS in using different strategy to select precursor ion for fragmentation. Instead of using the most abundant signal intensity for the precursor-ion selection, it performs an MS1 analysis first and obtains an inclusion list of precursor ions with retention time window after data processing. The second MS/MS analysis is performed on precursors, which are present in the inclusion list obtained previously. This method prevents multiple reanalysis of the same peptide, and allows identification of low abundance components and peptides with interesting features such as distinctive isotopic pattern, mass defect or differently modified peptides [44,57].

Multiple reaction monitoring (MRM) is gaining popularity in targeted quantitative analysis for small proteomes and has the

advantage to cover a large dynamic concentration range across 5 orders of magnitude [58–60]. MRM has relatively high sample throughput (30–60 min for analysis of one sample), is able to measure few hundreds of proteins in one experiment and requires monitoring of 5 peptides per protein selected with the help of PeptideAtlas [61] or with prediction using bioinformatics tools such as PeptideSieve [62]. Monitoring of each proteolytic peptide requires at least 3 optimized MRM transitions selected with use of a spectral library [63,64]. However experimental validation of the MRM transitions and their selectivity for a given problem is required to conduct reliable analysis, which can be performed by synthesis and analysis of synthetic peptide standards. Synthetic, stable isotope labeled peptide standards may be used for absolute quantification. Due to their wide dynamic concentration range, MRM-based methods can be successfully applied for validation of multiple biomarker candidates [65,66]. Recent perspective paper describes and compares the DDA MS/MS, directed MS and MRM based proteomics analysis strategies facilitating the methodological choice for experimental researchers [44].

A large variety of raw data formats from different mass spectrometer vendors were recently standardized using several alternatives of extensible markup language (XML) formats. Widely used formats are mzXML [67] (developed at the Institute for Systems Biology in Seattle Proteome Center) and mzData [68] (developed by the Human Proteome Organization Proteomics Standard Initiative or HUPO-PSI). These two formats were lately merged by the HUPO-PSI [69,70] into a new standard called mzML [71]. Several standardization attempts mainly by HUPO-PSI were made recently to standardize other types of proteomics data format such as peak list, proteomics experiments, however these formats are less widely used by the proteomics community [72–74].

Label-free LC–MS data pre-processing pipelines convert the raw data into a matrix containing quantitative information on the characterized and preferably identified compounds in each of the samples amenable for statistical analysis. The main modules of such pipelines with the data flow during this conversion are presented in Fig. 2. This procedure begins with raw data pre-filtering (such as noise reduction, data reduction etc.), and is followed by detection and quantification of compound-related features, and results in feature lists characterized among other things by quantity, retention time and m/z. These feature lists can be further reduced by deisotoping and summing up the intensity of compound-derived ions with different charge states. However, these steps can be also performed after the features have been matched across multiple chromatograms. Peptide-related features in different chromatograms have to be aligned or corrected in all three dimensions of MS-1 data: time alignment in the retention time dimension, mass calibration in the m/z dimension and normalization in the intensity dimension. The final step is peak matching, which has the goal to find the same peaks in multiple chromatograms and provide the quantitative feature/peak matrix characterized by m/z and retention time values. Data processing pipeline should be flexible enough to adapt to the characteristics of the datasets that are dependent on pre-analytical factors, the type of mass spectrometer and experimental design of the sample preparation and sample profiling platform. Many data processing applications and workflows consisting of multiple modules, which are interconnected by input and output parameters and data, are available free of charge or commercially. Work has been dedicated to construct optimized data analysis pipelines for label-free LC–MS [27,48], such as Viper [75], OpenMS [76–79], mzMine [80,81], XPRESS [82], SIEVE, Superhirn [83], Census [84], MapQuant [85], SpecArray [86], MsMetrix [87], PEPPeR [88] or XCMS [89] originally developed for metabolomics but also applicable to the analysis of proteomics data.

2.1. Data reduction

MS-1 data is three dimensional in nature with retention time, m/z and ion count dimensions. This information is generally stored with succeeding mass spectra storing information in mass–intensity pairs. This raw data is often converted into a two-dimensional regular matrix, with a procedure called meshing, resulting in an intensity matrix, where the columns and rows correspond to a given mass and retention time. Two types of raw mass spectrometry files are provided by the mass spectrometers. Profile data contains all acquired data points, and centroid data is pre-processed by the acquisition software generally with algorithms operating on single MS spectra. Storing data in centroid mode may result in loss of information for certain data processing algorithms, which perform peak detection in both dimensions, but reduces considerably the size of the acquired data. Data processing algorithms, especially those that are written in interpreted, complex high-level programming language such as R or Matlab, generally load all data into the computer memory and are thus limited by the available memory. These algorithms apply data reduction to fit the amount of data to the available memory. This is most frequently done by binning [18,80,89] which sums intensities between predefined consecutive and disjoint mass domains. This works well when most of the data points of the Gaussian peaks are within the mass borders of the bin, but leads to fluctuating saw tooth type splitting of the peak for centroid data when the bin borders fall in the fluctuation domain of the peak maxima along the consecutive m/z traces. This problem can be avoided by using a two-dimensional Gaussian filter that smoothes fluctuations in both retention time and mass dimensions thus avoiding the sawtooth splitting of peaks (for details see Fig. 3). Other approaches to reduce the intensity fluctuation of binning were reported recently, however each of them is computationally intensive and results in varying bin widths [90–92]. The quality of the LC–MS data determines the accuracy of feature detection and quantification. Choosing between binning or 2-dimensional Gaussian smoothing of the data has a dramatic effect on quantification when data reduction is applied. Data processing pipelines using programming languages with the possibility to allow user-defined memory management are advised as is the use of streaming to overcome memory limitations in the case of profile data at their original resolution. Streaming is a programming technique which reads and processes only part of the data in one time, and after processing, the results of each part are written on a continuously growing file. The algorithm goes over all data parts resulting in the complete processing of the file, Streaming allows to process large files independently of the available amount of RAM. Data reduction should be avoided if possible due to information loss.

2.2. Noise characterization, feature detection and extraction

A chemical compound with a given charge and isotope distribution is represented as a three-dimensional Gaussian peak in MS-1 and is often denominated as ‘feature’ in the data processing world. Due to the natural isotope distribution and to the occurrence of multiple charge states, one chemical compound results in multiple Gaussian peaks with the same retention time. These features must, at the first level, be discriminated from noise to determine their main characteristics such as quantity represented by the peak volume, area or height, retention time and mass to charge ratio of the center of the peaks, as well as the extension of the Gaussian peaks in the m/z and retention time dimensions. The second level is the extraction of compound characteristics related to charge state determination and the identification of isotope peak clusters. First level feature characteristics are obtained by all data processing pipelines while extraction of second level characteristics is optional and can be performed at a later stage after matching the same

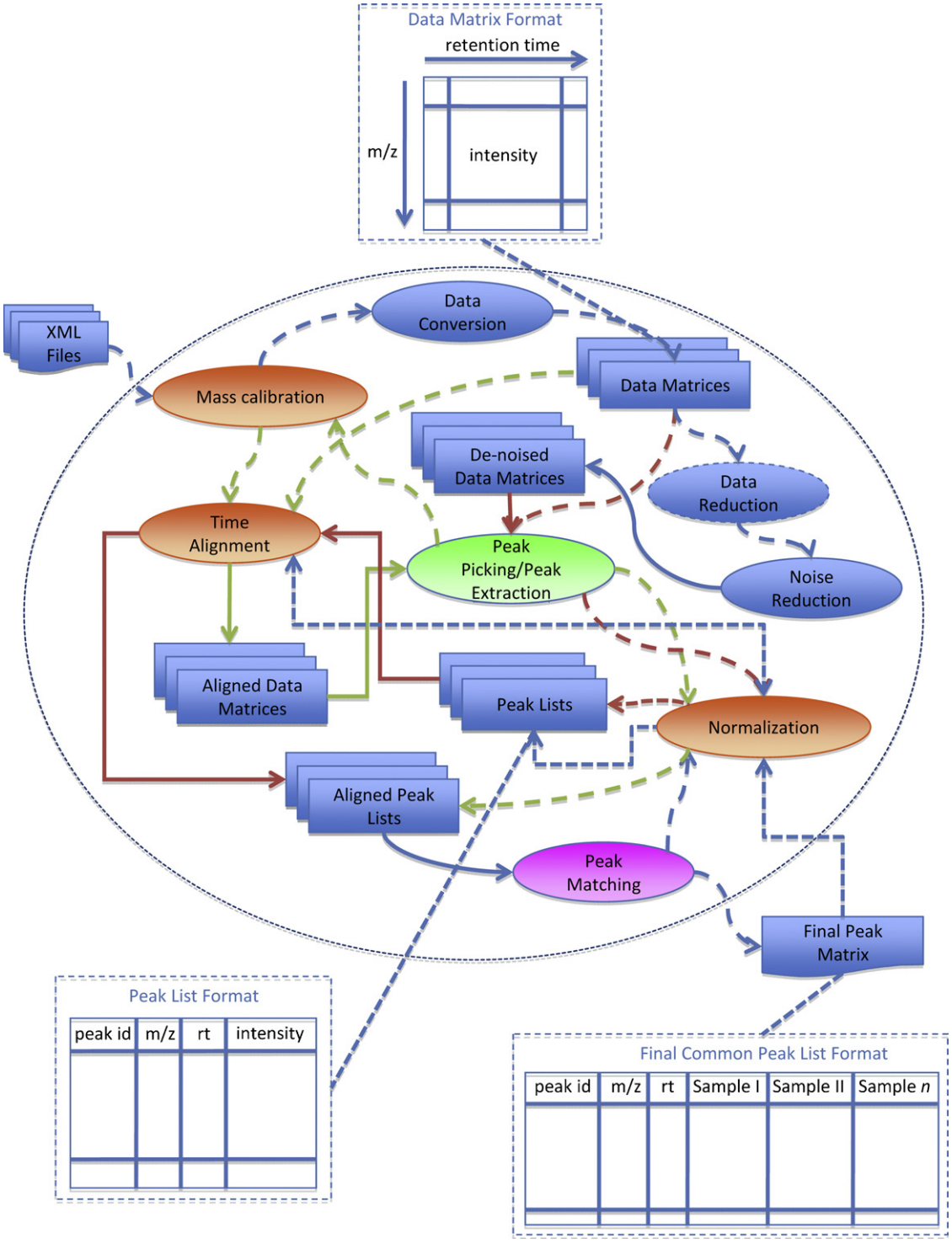

Fig. 2. Main modules of quantitative data processing pipelines. Raw data in standard format (mzXML, mzdata.xml or mzML) may be recalibrated for increased mass accuracy and converted into a resampled matrix, where intensity values are in the matrix, rows (or columns) correspond to *m*/*z* and columns (or rows) to retention time values (exact values are stored in separate vectors). This matrix may be optionally subjected to data reduction and noise filtering. The obtained data matrix is subjected to peak picking (ellipse in green) and peak quantification providing a peak list containing the most important characteristics of the identified peaks, usually retention time and *m*/*z* values of the peak centroid, peak quantity expressed in peak height, area or volume, and optionally peak extension in the *m*/*z* and retention time dimensions, as well as charge state and an index that is used to couple peaks of the same isotope cluster or different charge states of the isotope clusters to the same compound. Alignment in the three available dimensions (time alignment, mass calibration and intensity normalization) can be performed either at the peak list or at the raw data file level (ellipses in orange). The aligned and normalized peak lists of different samples are then matched (ellipse in purple), resulting in a quantitative peak matrix containing information about the matched peaks in different samples, where columns (or rows) correspond to different samples and rows (or columns) to different peaks, which are later coupled to identities at the peptide and protein level. This quantitative peak matrix is used for statistical analysis to identify discriminating peaks between predefined classes of samples. The figure indicates the two most common data flows with the order of modules using blue and red arrows. Dashed arrows indicate optional linkage of modules. Another order of the modules and data flows is possible as well the integration of different modules, such as time alignment with peak matching. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

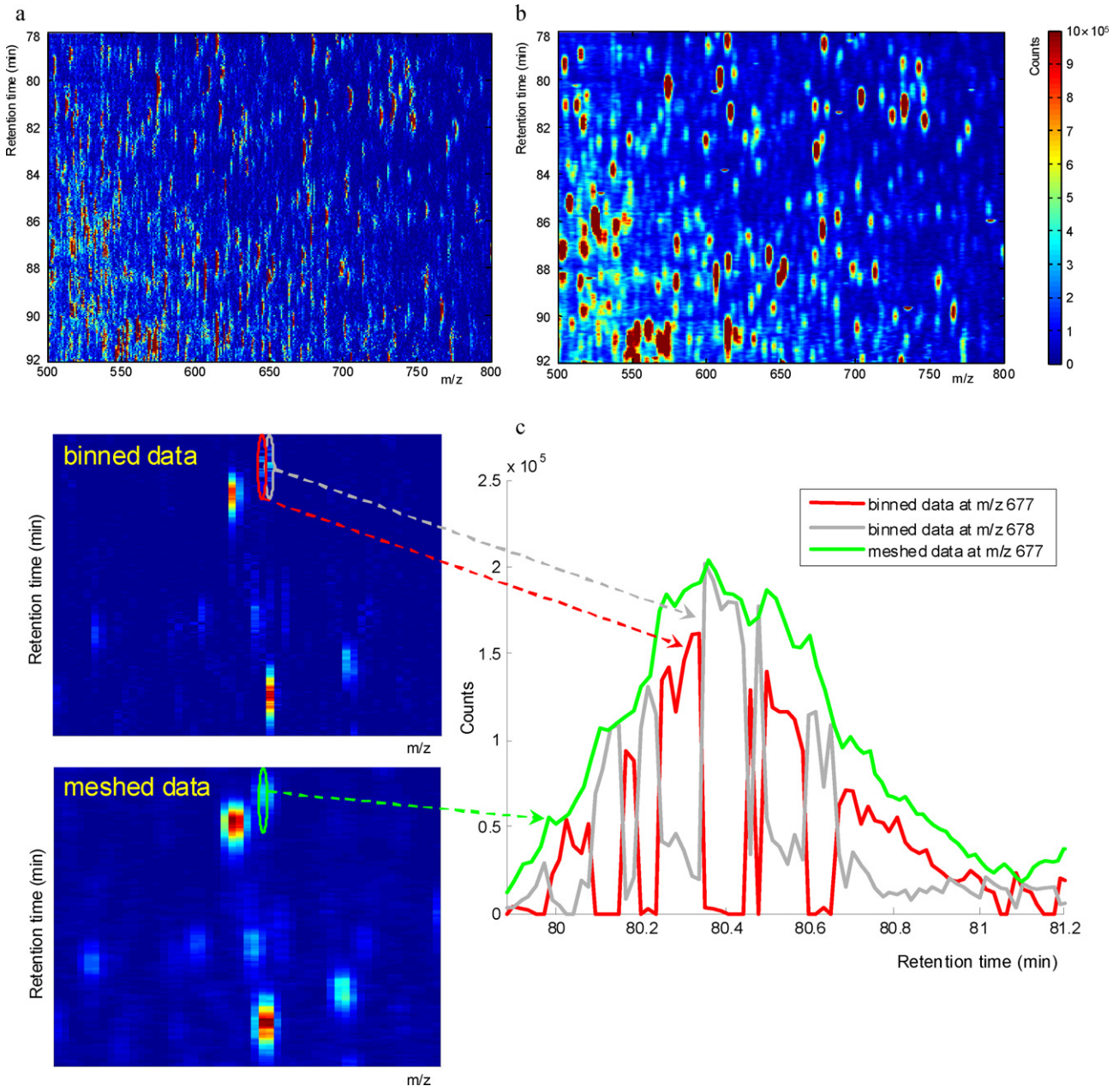

Fig. 3. Raw centroid ion trap MS-1 LC–MS image of depleted, trypsin-digested human serum obtained after binning (summing up intensity across 1 amu intervals having borders at fractional decimal of 0.5 *m*/*z* for each integer *m*/*z* value), (a) and after applying a two-dimensional Gaussian filter using the same degree of data reduction in the *m*/*z* dimension as for binning (b). Binning results in noisy data, which leads to a poor feature detection and quantification efficiency, in contrast to the data obtained after Gaussian smoothing. Peak detection becomes extremely difficult in case of fluctuating peak maxima in between mass spectra in two adjacent bins, (c) as represented in a part of an LC–MS image highlighting a peak with extracted ion chromatograms, where the highest intensity of the peak in centroid data fluctuates between the border of the bins. This fluctuation results in a saw tooth peak shape in adjacent extracted ion chromatograms, which will lead to poor performance in case of use of feature detection algorithms recognizing Gaussian peak shapes in individual mass traces.

feature of the same compounds in multiple chromatograms. If low resolution mass spectrometry data or considerable data reduction in the mass dimension is used for high-resolution mass spectrometry data, isotope peaks may collapse into a single Gaussian peak or into series of strongly overlapping Gaussian peaks. In this type of data, peak detection algorithms will detect the isotope peaks cluster as one Gaussian peak and provide the average mass of the peak cluster.

Noise characterization is important and can be regarded as a part of the peak detection step, which tries to discriminate noise from compound-related features signal. Noise in LC–MS data originates from different sources [93–95]. To discriminate noise from features, it is useful to take the noise model of the different mass analyzers into account in the peak detection algorithm [96]. Mass analyzers and detectors define the background white noise. Another type of noise of chemical origin is called chemical noise. Chemical noise originates from molecule clusters formed during electrospray ionization (*e.g.* solvent clusters), from chemical contaminations inside the mass spectrometer (*e.g.* in the ion source), from the chromatographic column or from the ambient air such as polydimethylcyclosiloxanes, phthalate or the plasticizers di-*n*-butyl-phthalate and (di-(2-ethylhexyl)-adipate) [9,97]. Ion suppression effect [98] distorts the compound-related signals and is dependent on sample composition and therefore also on the upstream sample preparation steps. Formation of eluent ion clusters during electrospray ionization and elution of contaminants

from the chromatographic columns are highly influenced by the water/acetonitrile ratio of the eluent, which changes gradually during reverse-phase LC–MS. This results in varying chemical background noise in both the retention time and m/z dimensions. Contaminants in the LC eluent or in the ambient air result in stripes of constant m/z across a large part of the retention time axis.

Numerous denoising and baseline (average noise level) subtraction algorithms exist in the literature, such as moving average [18,99], Savitzky–Golay filters [100] or entropy-based noise reduction [101] to name a few examples. As these algorithms will not be covered in this review, the reader is referred to reviews and publications on this topic [102–107]. It is important to choose baseline removal and denoising algorithms, which do not alter the quantitative information of the peaks. Many feature detection algorithms were developed to discriminate compound-related peaks from noise since the introduction of LC–MS. These algorithms match the isotope pattern of compounds in the m/z dimension (1D peak picking) or the peak shape based on extracted ion chromatograms (2D peak picking) or on the full 3-dimensional LC–MS data (3D peak picking) to detect Gaussian-shaped peaks. Algorithms that match isotope patterns have the disadvantage that they do not use the full 3-dimensional structure of MS-1 data, but apply peak detection on individual MS spectra, which are subject to a high noise content. VIPER [75], SpecArray [86] and SuperHirn [83] are examples of data processing workflows using peak picking algorithms based on isotope pattern matching. Examples of 2D peak matching algorithms based on extracted ion chromatograms are the *M–N* rule [18], and the matched filtration with experimental noise determination (MEND) algorithms [104]. *M–N* rules detect an LC–MS feature in extracted ion chromatograms when the intensity exceeds the local baseline *N* times for *M* consecutive points. MEND matches Gaussian peak profiles on noise defiltered extracted ion chromatograms. Another example for a 2D peak shape matching algorithm is based on wavelet decomposition developed by Coppadona et al. [105,106] to remove noise and define the baseline of extracted ion chromatograms. Finally the algorithm uses the difference of baseline and denoised data to detect peaks. Other peak detection methods use modified version of the binning algorithms adapting the bin size to the peak width in mass dimension. This type of algorithm detects peaks either by locating regions in centroid data with large amount of missing noise in mass dimensions and having high intensity signal with close mass value in consecutive retention time points [92], or by peak detection using Kalman filter to extract Pure Ion Chromatograms containing only information on peaks without noise [91] and by centWave method detecting peaks containing regions based on analysis of mass variation of centroid intensities in the retention time dimensions and identifying the peaks using continuous wavelet transformation and optionally Gauss-fitting in the chromatographic domain [90]. Three-dimensional peak detection methods using shape matching are applied in MapQuant [85], which fits a 3-dimensional Gaussian curve on local maxima. The apLCMS pipeline [103] uses a two-dimensional density kernel function to identify groups of peaks, while MZmine [80,81], msInspect [108] and LCMS-2D [107] use peak shape in the retention time dimension and the isotopic pattern in the m/z dimension. OpenMS [76–79] uses a three-dimensional wavelet function taking the average of the peptide isotope composition into account by constructing a mixture Gaussian model. There are many other peak detection methods and the reader is directed to a recent review on the topic [109].

To compare the performance of different peak quantification algorithms, the peak picking methods of msInspect and mzMine were compared by analyzing a tryptic digest of a mixture of 48 recombinant proteins resulting in ~800 peptides by FTMS in MS-1 and MS/MS mode with the help of a receiver operating characteristics (ROC) curve. The comparison showed that the isotope pattern matching algorithm of msInspect was superior in performance to mzMine using predefined peak shape template for peak detection [109]. Peak tailing or fronting and saturation of the detector lead to peak splitting for some features. The occurrence of peak splitting depends on the peak detection method and should be evaluated for each algorithm using different types of data. The algorithm developed by de Groot et al. [110] uses *K*-mean clustering to correct for split peaks and to correct peaks that were incorrectly aligned in the retention time dimension. Other quality control criteria applicable only to high resolution data such as Orbitrap and FTMS use mass deviance to assess if a detected compound correspond to real peptides [111]. Mass deviance is the difference of the decimal fraction of the monoisotopic peak of the detected compound and the nearest theoretical tryptic peptide. Overlaying on a part of a 2 or 3 dimensional MS-1 raw or pre-processed data with the location and extent of detected features as it possible in OpenMS framework enable to assess visually the accuracy of peak picking method [76,78,79].

2.3. Alignment of LC–MS images

Three-dimensional MS-1 LC–MS images are prone to nonlinear shifts in all of the three dimensions. In the mass dimension alignment is based on proper calibration of the mass analyzer preferably with internal standards. In the intensity dimension, normalization may be used, and in the retention time dimension, time alignment is necessary. Mass calibration should provide alignment to the exact mass. Intensity normalization could be relative to compare relative intensity of the peptides and proteins, but may provide exact alignment if absolute quantification is required. Retention time alignment is relative, in spite of the fact that retention time indices may be used for identification [112]. After successful alignment of all LC–MS chromatograms common peaks in different chromatograms are matched and their relative or absolute quantities are reported in the form of a matrix that is amenable to statistical analysis. In this matrix rows (or columns) correspond to samples and columns (or rows) to features or peptide peak identities. Alignment of LC–MS images can be performed with different goals in mind depending on the experimental design, such as to transfer peak identity information from separate MS/MS datasets to MS-1, or to combine data from several chromatograms corresponding to the fractions of a 2D-LC–MS analysis of a single sample.

2.3.1. Mass calibration

The m/z dimension is the most stable dimension toward shifts. The absence of shifts does, however, not mean that the measured values are accurate. This requires calibration of the mass spectrometer, preferably with internals standards that are present in each spectrum. Mass analyzers are, however, measuring instruments that are prone to small nonlinear shifts requiring automated algorithms to compensate for inaccurate mass calibration or to enhance mass accuracy, especially for high resolution mass spectrometers. Ions of chemical background noise originating from eluents or from the ambient air such as polysiloxanes or continuously added calibration standards can be used for mass calibration and to increase mass accuracy. A polynomial mass calibration function was used by Scheltema et al. [113] to increase mass accuracy of metabolites measured with an Orbitrap mass spectrometer. The algorithm improved mass accuracy from 1–2 ppm to 0.21 ppm using background ions, such as polysiloxanes, as internal standards. Haas et al. [114] used polydimethylcyclosiloxanes to enhance mass accuracy of MS/MS spectra and reported a higher identification rate of peptides. Another strategy involves the use of already identified peptides as calibration standards, a strategy that was successfully applied to improve peptide identification based on MS/MS spectra [115]. An interesting approach developed by Dijkstra and Jansen

[116] superimposes isotope clusters of the same peptide at different charge states in SELDI–TOF–MS spectra to improve mass accuracy. This approach can be easily adapted to spectra obtained with other instruments and electrospray ionization as multiple charging of peptides is a common phenomenon.

2.3.2. Intensity normalization

High throughput LC–MS data comes with nonlinear and systematic bias in recorded peptide ion intensity, affected mostly by differences in injected sample amount, differences or drifts in ionization efficiency, differences in ion transmission efficiency or detector saturation, and carryover between LC runs. The resulting bias should be corrected in order to enhance statistical classification accuracy. Systematic bias due to a difference in injected sample amount should be minimized, *e.g.* by determining the injected amount with a total protein assay or by taking the area under the curve of the UV trace of a previous analysis into account. Sometimes normalization can be based on a single factor such as the average or median abundance of peptides derived from so-called 'housekeeping proteins' or other compounds that are known not to be affected by the investigated disease or sample dilution factor, such as creatinin in urine samples [117]. However, all intensity normalization approaches have drawbacks, *e.g.* normalization based on housekeeping proteins should not be applied for samples changing in constitution of these proteins considerably. Remaining nonlinear bias can be removed using normalization methods applied in microarray analysis [118] *e.g.* by performing a nonlinear regression of matched peak intensities in two samples with the same or close composition. Various normalization methods were developed and assessed for label-free quantification using LC–MS [100,102,119–121]. Linear regression can be applied to bias that has a linear pattern across analyses such as sample carryover in the trapping column. Nonlinear bias caused *e.g.* by detector saturation can be resolved by non-linear or local regression techniques, and quantile normalization may be used to transform peptide quantity distribution of all sample to the same distribution using quantile plots. Callister et al. [120] compared different normalization methods using LC–FTICR–MS datasets and concluded that global or linear regression worked best in most cases when applied iteratively. A recent study by Kultima et al. [119] compared 10 different normalization methods using datasets from mouse, rat and quail that were analyzed by a nano-LC–MS system coupled either to a Q-TOF or to an LTQ mass spectrometer. Karpievitch et al. [121] developed a normalization method based on singular value decomposition to remove systematic and nonlinear bias to avoid overfitting by dimension reduction for label-free LC–MS proteomics samples. In general, normalization algorithms use matched peak matrices. Therefore normalization procedures are implemented after peak matching and prior to the statistical analysis.

2.4. Time alignment

The retention time of compounds is subjected to considerable non-linear shifts between LC–MS experiments and requires particular attention and more sophisticated alignment algorithms than the two other domains. Complex proteomics samples, such as serum and urine, contain several tens of thousands of peptides, so that even small retention time shifts may result in serious peak mismatching, if peak matching is only based on the retention time and m/z coordinates across multiple samples. Retention time shifts are due to parameters of liquid chromatography that are hard to control, such as small changes in eluent composition, pH, column ageing or temperature changes and have a highly non-linear behavior, especially when the combination of a trapping and analytical column system is used. Accurate algorithms to correct retention time shifts is one of the most critical points of data processing to provide accurately matched peaks and quantitative data that are suitable for statistical analysis.

The goal of time alignment is to find the maximum overlap between similar peaks in different chromatograms and provide a retention time transformation function, which can be used either to change the retention times of all peaks in a peak list, or to change the associated retention time of mass scans in the raw data. The major difference between time alignments methods using MS-1 data is how many data dimensions they use in their benefit function to drive the time alignment procedure. Earlier developments considered only 1-dimensional data next to the retention time dimension in their benefit function (*e.g.* the TIC or BPC) [122–130]. Recent algorithms use two-dimensional profiles that take the separation of compounds in the mass and the retention time dimension into account. The latter approaches provide more accurate time alignment of highly complex LC–MS 'omics' data. Two-dimensional alignment algorithms differ in terms of whether they use the raw data, pre-processed data obtained after noise filtering and data reduction [94,95,100,126,127,131–133] or peak lists after the peak detection step [77,134,135].

A large number of algorithms were developed to define the optimal search space for non-linear retention time correction, such as Correlation Optimized Warping [95,122,126,128,129,135], Parametric Time Warping [94,124], Dynamic Time Warping [94,100,125,127,129–133,136], a geometric approach based on pose clustering [77,137], Loess regression on matched compound pairs [134], the Continuous Profile Model combined with a Hidden Markov Model [138] to list a few. Time alignment based on DDA MS/MS data uses the correlation between MS/MS information of the same compound [88,139]. Other types of algorithms create retention time and mass tags by normalization of retention time and accurate mass. These tags are subsequently used to align multiple LC–MS datasets in both the m/z and retention time dimensions or through comparison with a database [136].

Generally time alignment is performed by selecting one chromatogram as reference and aligning all others to that reference pair wise. This approach requires the *a priori* selection of a reference chromatogram and must assess how selecting different reference chromatograms affects the quality of time alignment. Robust time alignment methods should not depend on the choice of the reference chromatogram [94,95]. The Continuous Profile Model developed by Listgarten et al. [138] does not use a reference chromatogram, but performs the alignment of all chromatograms in one step. The performance of different time alignment algorithms depends on many parameters, such as the number of common peaks shared between chromatograms, the complexity of the samples, the compound distribution in retention time–m/z space, the compound concentration variability and noise distribution. A comparison of different algorithms with different characteristics shows that time alignment methods that take the three-dimensional nature of MS-1 data into account perform better for complex proteomics samples with large compound concentration variability [94,95,135].

In most studies the same chromatographic columns and strict standard operating procedures are applied in order to lower analytical variability. It is rare that the elution order of compounds changes under these conditions and a monotonic time alignment function is appropriate [140]. However the elution order may change during extensive studies over a long period of time or when different types of columns are used (*e.g.* when using different types of *n*-octadecyl bonded silica reverse phase stationary phases) [141]. It is also known that the pH of the eluent has a dramatic influence on the selectivity of RP columns [142,143], and this can lead to a changing elution order when analysing complex proteomics samples. Inversion of the elution order of peptides or metabolites is not commonly reported and thus probably not recognized. It is important in

the future to explore this phenomenon in greater detail, especially when different types of LC RP-C18 columns are used within one study or when large-scale studies are performed in different laboratories. If peak elution order changes are frequently observed, this would require novel time alignment algorithms, which can deal with peak elution order changes adequately. For further reading on time alignment the reader is referred to specialized reviews [140,144] and articles presenting results from performance comparisons of different time alignment methods [94,129,137].

The quality of time alignment is generally evaluated by visualization of the entire chromatogram, or by visualization of common peaks to assess the local time alignment accuracy. Co-injected standard peptides or peptides derived from highly abundant housekeeping proteins can also be used for this purpose. Comparing the quality of different time alignment algorithms in such a way is a daunting task and visualization of entire chromatograms does not always allow a proper quality control of the time alignment results. A quality assessment method based on the sum of the overlapping peak area between pairs of chromatograms provides a global readout of alignment quality and permits comparison of the relative performance of various types of algorithms [94,95]. Similarity score calculated after time alignment for assessing the quality of time alignment, may also be used to assess eventual bias in experimental design and to detect whether there is a systematic difference between sample replicates or injection order [145].

2.5. Peak matching

The peak matching process identifies common peaks in different chromatograms based on proximity of aligned retention time and mass or peptide identity based on MS/MS data when this information is linked to the peak list. Numerous clustering algorithms have been applied to match peaks such as *K*-mean, hierarchical or pose clustering [102,144,146]. The procedure provides a list of clusters of the same compounds in different chromatograms, from which a quantitative feature matrix is constructed. The matrix contains a quantitative measure of the feature and rows (or columns) that correspond to the samples and columns (or rows) to the features (*e.g.* peptide ions) characterized by retention time and m/z value. Features in this matrix can be further processed by deisotoping and by integration of different charge states of the same compound resulting in a quantitative peptide matrix, which further may be combined with identification results. Deisotoping and decharging can also be performed at peak list level prior to the peak matching procedure. MS-1 data processing pipelines must assign a quantitative value to peaks that do not have a correspondence in all chromatograms. Some pipelines extract noise at the corresponding location, while others filter out peaks found only in minority of samples to avoid bias for single and rare events in the subsequent statistical analysis.

2.6. Peptide and protein identification

Peptide and protein identifications are generally performed using information from MS/MS spectra. Before identification, MS/MS spectra are filtered to remove noise and the cleaned spectra are used for the identification process. The most widely used protein identification approach is database search, where lists of peaks in MS/MS spectra are compared with molecule fragments obtained by *in-silico* fragmentation of sequences stored in the database [147–149]. This comparison may be performed by calculating a similarity score between the *in-silico* fragments and the measured fragments, and the peptide with the best match receives the highest score. A threshold is used to limit the number of false positive identifications while at the same avoiding to penalize true positives. It is thus noteworthy that so-called identified peptides and proteins always contain a chance that they are false positives. This matching approach is used by the Sequest algorithm [150]. Interpretative models use the assumption that MS/MS spectra consist of a continuous series of fragment ions that can be interpreted as a partial short amino acid sequence tag of the intact peptide. PeptideSearch [151], MS-Seq [152] and GutenTag [153] use this strategy. Stochastic model-based algorithms use probability estimates for peptide fragmentation and subsequent predictions of the resulting mass spectra that are compared with the measured MS/MS spectra. SCOPE [154] and Olav [155] (the basis of the Phenyx search engine) are examples of this category. Finally programs such as Mascot [156] and the open source OMSSA [157] apply statistical and probabilistic models using empirically generated ion probabilities of peptide sequences stored in the database. Spectral library search algorithms represent another category of peptide identification algorithms. These algorithms compare noise-filtered MS/MS spectra with databases containing high-quality, experimental MS/MS spectra using similarity scores [158,159]. Clustering MS/MS spectra of the same peptide can enhance the probability for successful peptide and protein identifications significantly while at the same time decreasing the number of spectra that are sent for database search by one order of magnitude [160]. Database search and spectral library search algorithms have limited capability to identify peptides with PTM, since the peptide with a given PTM should be either present in the database or defined by the user prior to the search. Open modification search programs such as Popitam [161] and Inspect [162] use MS/MS spectra of already identified peptides and allow unexpected mass shifts in the fragmentation pattern of the peptides due to PTM. Open modification search algorithms are, however, computationally more intensive than database search algorithms and therefore they generally use a limited number of peptide sequences for identification. *De novo* sequencing algorithms such as PEAKS [163,164], PepNovo [165], EigenMS[166], Lutefisk [167], Sherenga [168], MSNovo [169], PILOT [170], NovoHMM [171], and AUDENS [172] use only information from the experimentally acquired MS/MS spectra and basic constants such as the mass of the amino acids to elucidate the most probable sequence of the fragmented peptide.

A simple peptide identification method using label-free MS-1 data exploits so-called accurate mass and time tag (AMT) information that is calculated from accurate mass and retention times with or without normalization to match pre-identified peptide sequences in a database to the newly acquired data [112,173,174]. This identification strategy has the advantage to perform MS/MS based peptide and protein identifications on pooled or representative samples using time-consuming profiling techniques with a large peak capacity (*e.g.* 2D-LC–MS with DDA MS/MS data acquisition) followed by the quantitative analysis of a large number of samples with faster LC–MS platforms operating in MS-1 mode, which cover a larger measured concentration range. Basically this technique can be considered as a generic peptide and protein identification transfer system that uses mass and retention time information for the matching, and has the disadvantage that the identification transfer may be sensitive to the LC parameters and that high-resolution mass spectrometers with high mass accuracy (FTMS or Orbitrap) are required to reduce the chance for incorrect matching at an acceptable rate [173,174]. In most case normalization of the retention time is performed through regression of the observed and predicted retention times using training datasets and a neural network for retention time prediction. The databases containing peptide identifications with AMT tags are generally obtained from different analyses by generating reference maps [112,175].

To decrease false positive identifications, Scaffold [176] combines identification results of different database search programs such as Sequest [150], Mascot [156] or X! Tandem [177] (version

performing parallelized processing is called X!! Tandem [178]) and calculates a composite probability score, providing more reliable protein identification compared to the single score of one program. A protein identification score is constructed from identified peptide scores and relates to the probability that a given identification is a true positive. With its combined scores, Scaffold also provides a more reliable grouping of peptides for protein identification. The probability for false positives can be further decreased by comparing the measured and predicted retention times of the identified peptides. Retention time prediction algorithms use statistical methods based on quantitative structure–retention time relationships [179], which are in turn based on a large number of molecular descriptors or training datasets and regression methods taking the amino acid composition of the peptide into account [180,181].

Database search algorithms are biased to the protein present in the database and are poor to detect splice variants and proteins with PTMs. In order to be able to identify proteins with splice variants and PTM, the Swisspit [182,183] workflow combines the results of identifications obtained with the Phenyx [155] and X! Tandem [177] database search type algorithms with the Popitam and Inspect open modification search type algorithms. This is a combination of first assigning identifications using database search programs and subsequently submitting unassigned MS/MS spectra to open modification search algorithms with restriction to use peptide sequences identified previously by both database search tools. The Swisspit [182,183] workflow resulted in a higher identification rate of 77% for small well annotated PTM-rich lens datasets compared to 21% obtained with the combination of the two database search programs only.

Another ingenious approach use spectral network analysis for peptide identification [184]. This method finds pairs of MS/MS spectra that differ only in one modification or amino acid by searching for corresponding *b*- and *y*-type fragment ions. From multiple paired spectra, a network is constructed and is used to propagate peptide identification from peptide without PTM to the same peptide with different numbers of PTM or amino acids changes. Spectral network analysis was further adopted to include data from multistage MS/MS such as MS^3 or MS^4 in the interpretation [185]. Spectral dictionaries extend the sequence tag approach by generating sets of full-length peptide *de novo* reconstructions. These spectral dictionaries are then searched in a database equipped with hash table or suffix tree providing a fast identification algorithm, with high true positive identification rates [186].

Since database search algorithms always provide a list of identified peptide and proteins with given scores, it is important to test the statistical significance of the obtained score against a decoy database containing incorrect protein sequences obtained, for example, by reversing or randomizing existing protein sequences [187–190]. The presence of highly homologous proteins represent an actual challenge for the protein identification software, therefore in case of list of proteins with high sequence homology the results should be taken with precaution specially, when only peptides shared with other proteins are identified. The reader is referred to dedicated book and reviews for further reading on the peptide and protein identification algorithms [147,148,191–193], quality control methods [194] and influence of parameters affecting the quality of MS/MS [195].

3. Statistical analysis and validation

3.1. Coupling feature quantification with peptide and protein identity

Quantitative feature matrices should be first transformed to quantitative compound matrices by summing up the quantity of isotope clusters and different charge states of the same compound (one signal per compound matrix). Peptide quantity should be further matched with compound identity at the peptide and ultimately the protein level. Regarding protein quantity, different methods can be used starting from summing up the intensity of constituting peptides to taking the sum of the three most abundant peptides of each protein [55]. Mapping peptide identity to the quantity of extracted features involves several steps. Precursor ion mass of a given charge state should be matched to the isotope cluster of the same charge and the corresponding quantity (*e.g.* represented by the sum of all isotopomer peak height or the peak height of the monoisotopic peak only) of the isotope cluster should be combined with the quantity of the other isotope clusters of the same peptide with different charge state. This step is followed by the determination of all identified peptides constituting individual proteins. The peptide centric nature of the shotgun proteomics approach makes quantification of original protein mixture challenging in the presence of multiple proteins with high sequence homology, truncated protein forms, proteins having different PTMs or multiple splice variants [6,7]. For accurate protein quantification, with either MS1 or spectral counting methods, precise peptide and protein identification including detection of all protein variants is necessary, because the identified peptides will provide the list of peptides unique for the protein and peptides that are shared between several proteins, and which will allow accurate quantification of all protein variants. Recently Zhang et al. [196] evaluated different strategies for spectral counting quantification and found that the most accurate quantification was obtained by adding to the spectral counts of peptides unique for the protein, the corresponding molar proportion of the spectral counts of peptides shared between different proteins. The identification of exact protein form is also important for the development of accurate targeted MRM assays, while the presence of peptides shared between proteins could bias considerably the measured protein quantities [6,7,60]. A recent review by Podwojski et al. [197] deals with this problem in detail.

Annotated quantitative peptide matrices can be obtained by other methods [198], than those based on MS-1 data, such as the already mentioned MRM-based methods, spectral counting algorithms and by quantifying spot intensity in 2-dimensional gel electrophoresis coupled with peptide fingerprinting or LC–MS/MS identification [199–201]. Immunochemical techniques based on antibody arrays are especially interesting for the targeted profiling and validation of proteins in complex biological samples [202–204]. Quantification methods provide absolute or relative protein quantities or other type of biodescriptors if the goal of the analysis is to compare biological pattern for sample classification [205].

3.2. Sample/feature dimensions and feature selection/transformation methods

The main application of statistical analysis, also called post-processing methods, is to find peptides and proteins that discriminate between different groups of preclassified samples. Discriminating peptides or proteins are selected from the common peaks after data processing, which means that the validity of the ultimate statistical result depends on the quality of data processing. Statistical analysis of quantitative proteomics experiments suffer from the high dimensionality given by the large number of identified peptides and proteins accompanied by a much lower sample size. Due to this characteristic, processed proteomics data are often referred as megavariate data [206,207] leading to a high dimensionality small sample size (HDSS) problem [146]. HDSS is the main reason, why most of the widely used classification methods such as linear discriminant analysis cannot be directly applied to analyze quantitative proteomics datasets. In datasets with HDSS properties, a large number of compounds may be found to differ significantly between predefined classes of samples using, *e.g.* the univariate

t-test, but they may not be confirmed in other independent set of samples. Disease-related changes generally affect a small portion of proteins and peptides in living organisms, which represent the truly discriminating molecules between predefined groups of samples, and which stay true when measuring new sample sets. To find a small number of truly discriminating proteins among a very large number of other non-disease related proteins in datasets with HDSS, it is necessary to either use statistical methods that are insensitive to uninformative features (noise) or to reduce the number of features (dimensions) prior to the actual statistical analysis to a number that does not exceed the number of samples (independent observations). Only very few methods such as Support Vector Machines [208–210] or Learning Vector Quantizations [211,212], claim to be insensitive to a large amount of noise peaks contained in HDSS datasets and generally do not require upfront feature selection. Other methods require dimension reduction, which can be performed by removing uninformative peaks and selecting statistically relevant discriminative peaks (so-called feature selection) or to perform data transformation to accentuate class differences. Feature selection methods are most widely used [102,213–216], since these methods not only help to remedy the HDSS problem, but also provide a list of discriminatory peaks. Selected features corresponding to a limited number of biomarker candidates must be validated by measuring a larger sample set with fast and targeted analytical methods such as LC–MS/MS in MRM mode [3]. The results can also be used as input for databases and algorithms to link them to biochemical, secretion, molecular interaction or signaling pathways that may be involved in disease-related biological processes.

Feature selection can be performed in a supervised manner using univariate or multivariate selection algorithms. Univariate methods assume that features are mutually independent, so that each feature is evaluated individually based on its individual relevance to discriminate between predefined classes of samples. The simplest method for feature selection is the univariate Student's *t*-test, which must be corrected for multiple testing. In context feature selection methods (multivariate methods) take the interdependency between features into account, when evaluating the individual strength or rank of a given feature. Collective assessment and selection of variable subsets is another type of feature selection. This method selects a feature subset by evaluating all possible correlations or other forms of dependencies between features. Since the number of subsets increases exponentially with the number of features, it becomes an exhaustive task to evaluate all possible subset combinations in the feature space. Therefore, most collective feature selection methods are based on heuristic search strategies, such as forward selection and backward elimination [217]. Forward selection methods start with an empty feature subset and add features step-by-step to maximize a predefined scoring function. The procedure is stopped when newly selected features have a small contribution to the value of the scoring function. Backward elimination proceeds inversely and starts from the full feature set and eliminates features until a given scoring function reaches its maximum. An example of such an approach is the Reduce Feature Elimination method that can be combined with a classifier such as a Support Vector Machine [217]. Feature transformation methods construct new features from the original features while maintaining the initial data structure as accurately as possible. Typical feature transformation methods create a supervised or unsupervised mapping function that changes the initial feature space into a transformed variable space. The most widely used methods for feature transformation is principal component analysis (PCA), Fourier and the wavelet transformations. The most popular feature transformation methods coupled to statistical classification are Principal Component Linear Discriminant Analysis and Partial Least Square Linear Discriminant Analysis. Many feature selection methods are available and it is difficult to predict, which of them perform the best with respect to the others. Since many of the features in biological samples are correlated, in context, collective feature selection and feature transformation methods taking account of this correlation are preferable.

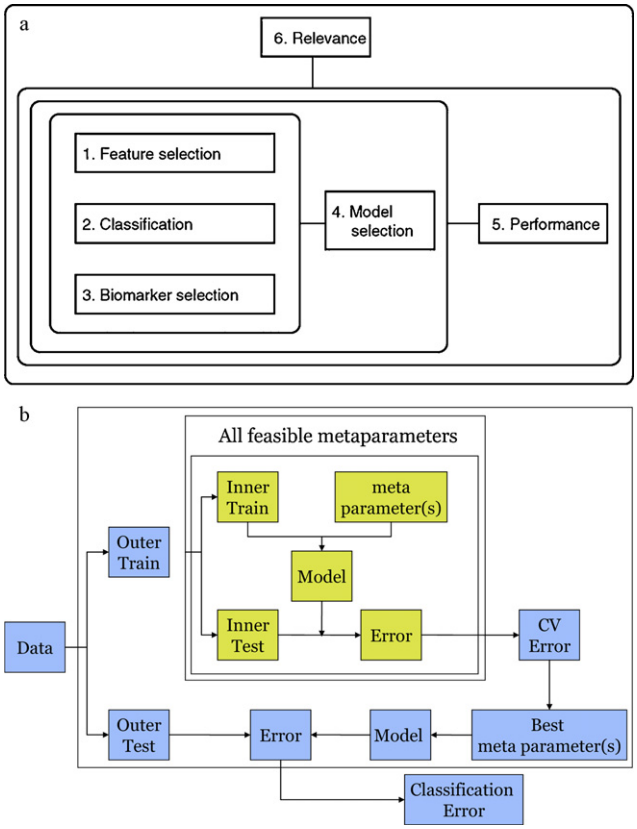

Fig. 4. Main steps of the statistical analysis and validation strategy for proteomics data suffering from the HDSS problem. Panel (a) gives a schematic overview of the modules for supervised statistical analysis and validation. The core (modules 1–3) represents the statistical analysis modules for feature selection (module 1; optional), classification (module 2) and biomarker selection (module 3; optional). The outer part represents modules for statistical validation comprised of a model selection module (4), a module to assess performance of the classifier (5) and a module assessing the relevance of the selected biomarkers by permutation tests (6). Panel (b) gives an overview over the generic double cross validation strategy to measure the performance of the feature selection-classifier modules (modules 4–6).
Figures taken with reprint permission from Smit et al. [214].

3.3. *Classification methods and statistical validation*

Fig. 4 shows the main modules of statistical data analysis and validation [214,216]. Feature selection and statistical analysis form the core, which is surrounded by different stages of validation layers to ensure that the resulting classification models are robust with respect to new sample sets. From the core modules, only the classification module is compulsory while feature selection and ultimately biomarker selection are optional if the classification method is not sensitive to HDSS and if sample classification is the only goal without identifying the underling molecular determinants. The first validation layer serves to select the optimal classification model (module 4) and the second layer measures classifier performance (module 5) by providing an error rate when classifying new samples, which were not used for building and selection of the classification model. Finally the relevance of the discrimination of the model is determined by comparing its performance to models obtained by chance using *e.g.* permutation tests (module 6). Permutation tests randomly reassign sample group

labels thus generating random models and the performance of a large number of random models is compared to the performance of the model obtained with the correct sample labels [216]. Generally validation modules 4 and 5 are based on a double cross validation strategy, with model selection occurring in the inner loop and classification performance being determined in the outer loop. Double cross validation strategies provide an unbiased way of evaluating model selection criteria and classification performance. This is achieved by dividing samples in each group into a training set, which is used for model building, and a test set that is used for performance measurement in the inner and outer loops, respectively. Another way to assess model performance is to determine the sensitivity and specificity of the statistical model by calculating the so-called ROC curves [214,218]. Current developments in statistical analysis focus on a combination of several methods resulting in ensemble classifiers. The reader is referred to more comprehensive reviews for details about feature reduction, dimensional reduction, statistical analysis and validation methods [146,216,219].

Most biomarker discovery studies are performed using samples from different groups, which are obtained from different individuals (*e.g.* patients or animals) not related to each other (cross-sectional study design). A study design where samples from the same individual are analyzed at different time points (longitudinal study design) enables to lower the biological variability. It is possible to further decrease the effect of biological variability by matching the different clinical parameters such as age, sex, smoking habits or living style. In this case statistical analysis should take the relation between samples into account by using adapted methods (*e.g.* time series analysis) to correlate compound concentration to time or other parameters such as drug dosage or disease development status [220–223].

Quantitative peak matrices can be used for different purposes than identifying class specific discriminating compounds. Fig. 5a shows the histogram of the relative standard deviation (RSD) of the quantity of all compounds in a given dataset, where variability was only subjected to analytical variance due to the use of two different LC–MS analysis platforms. The histograms show that the two platforms perform equally well to quantify peptides in serum samples depleted of the six most abundant proteins [99]. The aligned peak matrix also contains information about the global concentration variability of compounds in the different datasets. This comparison can be performed similarly using histograms of the RSD as presented in Fig. 5b for three different types of sample sets [95]. The figure shows that there are large differences in concentration variability between samples sets of different origin, such as acid-precipitated urine, serum depleted of the 6 most abundant proteins and a serum sample from one patient that was subjected to varied pre-analytical factors in a factorial design study. While the compounds of the first two datasets were subjected to biological and analytical variability in addition to errors during data processing, compounds of the last dataset were only subjected to analytical variability and data processing errors. The histograms indicate that well regulated body fluids, such as serum, show less concentration variability than the same type of sample measured with different preanalytical factors in a factorial design study. An excreted, non-regulated body fluid, such as urine, showed the largest concentration variability. Finally, quantitative peak matrices can be used to evaluate and compare the quality of different data processing pipelines using datasets containing compounds that were added in known concentrations by spiking. Grossmann et al. [47] compared the quantification performance of two relative spectral count methods (emPAI and APEX), absolution protein quantification method using the abundances of the three most abundant peptides developed by Silva et al. [55] and four different protein quantification methods using Superhirn (MS1 based quantification) by assessing the robustness and dynamic range of the spiked-in protein as well as non-altered proteins detected in the mixture of spiked yeast samples. The protein quantification method of Silva

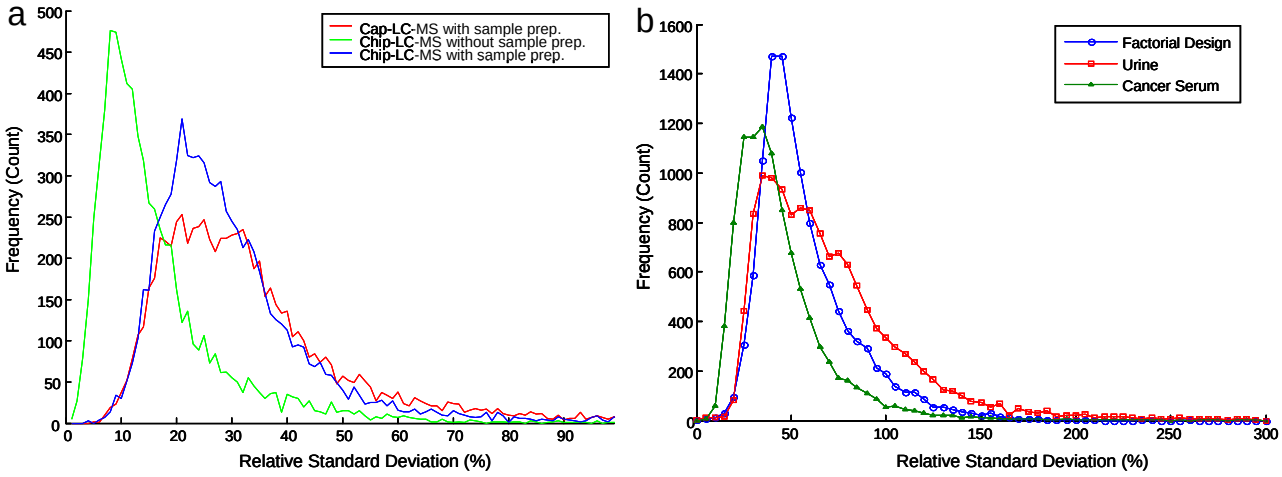

Fig. 5. Histograms of the relative standard deviation (RSD) of compound concentrations calculated based on results of the corresponding quantitative peak matrices. (a) Assessment of two LC–MS profiling platforms used for comparative proteomics studies. The Chip-LC–MS platform was equipped with a reverse phase nano-LC column (75 μm internal diameter) integrated in a microfluidic device and coupled to the mass spectrometer via an electrospray interface, while the Cap-LC–MS platform used a 1 mm internal diameter reverse phase column coupled to the mass spectrometer via an electrospray interface using a nebulisation gas (ionspray). The histograms show that the Chip-LC–MS (in blue) and the Cap-LC–MS (in red) platforms result in similar compound concentration variabilities and can thus be considered as equivalent quantitative profiling platforms. In both cases 10 serum samples from the same patient were depleted of the 6 most abundant proteins and underwent individual sample preparation procedures. These data thus contains only the analytical variability. Five injections of the same sample in Chip-LC–MS (green) resulted in histograms with a lower RSD indicating that most of the analytical variability is not caused by the Chip-LC–MS profiling platform itself but originates from the sample preparation steps (depletion or trypsin digestion). (b) Assessment and comparison of compound concentration variability after accurate data processing of different types of body fluid analyses from ongoing biomarker research projects. Twenty chromatograms of trypsin-digested human serum samples obtained from 10 different patients at two time points and depleted of the 6 most abundant proteins give a narrower RSD histogram with respect to compound concentration (green histogram) than the same type of samples obtained from one patient and subjected to an experimental design study (19 chromatograms) obtained by varying pre-analytical parameters (blue histogram). Fifty acid-precipitated human urine samples result in the widest RSD distribution (red histogram), which may be explained by the fact that serum is a well-regulated body fluid while urine is excreted and thus devoid of homeostatic control. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)
Figures taken with reprint permission from Horvatovich et al. [99] and Christin et al. [95].

et al. [55] with own implementation or using Superhirn [83] provided the best performance compared to emPAI or APEX methods, however their results should be taken carefully as they have used one bovine protein (Fetuin-A of 38.4 kDA) for spiking, which has very different composition (low homology) than that of the yeast proteome. Another recent paper compared the quantification performance of two commercial label-free LC–MS data processing software (Eluciator and Progenesis) based on spiked samples, and reported considerable differences in the performance [224].

4. Conclusion and future trends

Improvement and development of new data processing pipelines and individual modules will continue in the future as mass spectrometry-based molecular profiling is gaining momentum in life science research and development of new mass spectrometers and sample preparation methods is on the agenda of numerous research groups and companies. In current bioinformatics literature, there is increasing emphasis on the development of algorithms which assess and compare the performance of data processing methods. These algorithms also provide substantial support for parameter optimization and troubleshooting of algorithms. Performance assessment and quality control can be only performed with high quality standard data, where compound composition and quantity are precisely known [225]. To assess the performance of protein identification workflows, which report enhanced performance to detect known and unknown post translation modifications, there is a need for open access well annotated MS/MS data [182]. Similarly, for assessing quantitative label-free data processing workflows, it is necessary to provide reference data from different biological origins such as urine, serum treated with different depletion techniques or different cell lines spiked with known compounds in known concentrations. Development and access to standardized samples such as the recently introduced yeast standard sample help this procedure [225]. Well-characterized datasets also help to evaluate the performance of statistical analysis and validation strategies. Raw and processed data simulation software could, in certain circumstances, replace real datasets or create datasets with particular properties, which are difficult to obtain experimentally, and which may reduce assessment time considerably. For this reason, well characterized and documented datasets stored in free access databases such as Human Proteinpedia [226–228], and the development of accurate data simulation tools will, in the future, enhance the comparison and accurate performance assessment of different modules or/and complete data processing workflows. An example of quality assessment and parameter optimization of time alignment algorithms using well-defined samples is provided by Peters et al. in a recent publication [229].

Another trend is to develop data processing solutions to integrate highly diverse data such as data obtained with different instruments or in different laboratories. An example of such diverse data, which cannot be processed with the actual data processing pipelines, is presented in Fig. 6 showing a representation of two raw LC–MS datasets that were obtained form the same serum sample with the same ion trap mass spectrometer, but using two different ionization methods and two different HPLC techniques. While the two samples contain exactly the same compounds at the same concentration and have the same measured dynamic concentration range, the differences caused by the different chromatographic methods and the different ionization modes results in different peak distributions in the retention time and m/z space. Current LC–MS data processing workflows are not able to accurately combine these types of data sets, leaving this challenge for future developments.

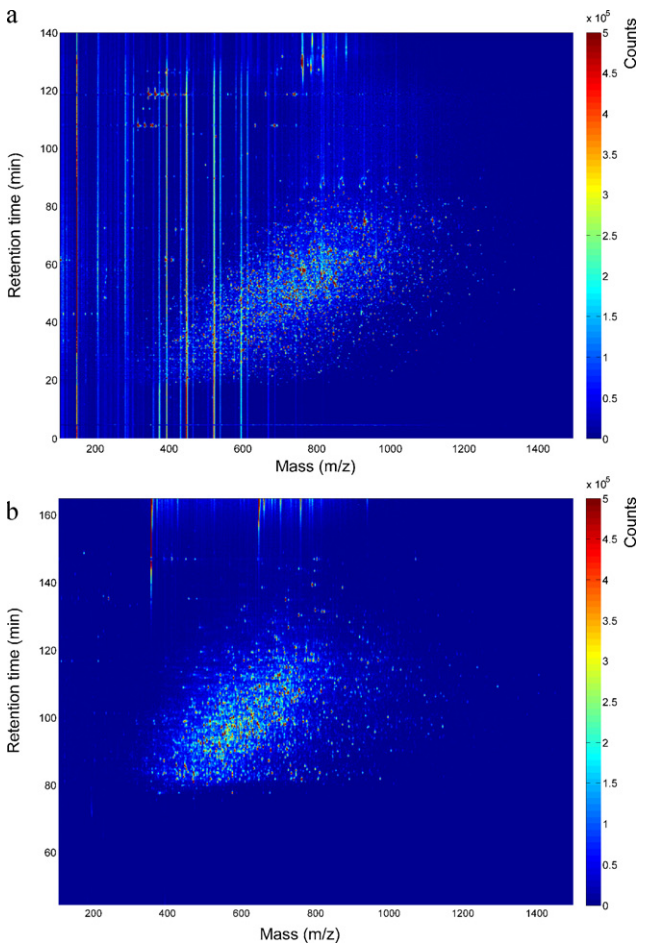

Fig. 6. LC–MS images of trypsin-digested human serum samples depleted of the 6 most abundant proteins obtained with the Chip-LC–MS (electrospray) (a) or the Cap-LC–MS platform (ionspray) (b). See Horvatovich et al. [99] and the caption of Fig. 5 for a description of the different LC–MS systems.

Newly developed or enhanced algorithms are emerging rapidly within bioinformatics research groups. However these new algorithms, with a high potential to ameliorate information extraction accuracy from raw data and biological knowledge discovery, are not used by the majority of the data producing, application-oriented proteomics laboratories. The main reason for the low penetration of new bioinformatics solutions is that mass spectrometer vendors generally provide user-friendly data processing and evaluation pipelines supported by training sessions favoring the application of their own software packages, even if the performance of these software packages is not assessed and compared with others. The newly developed algorithms, even if the open source program code is available, require on the other hand extensive bioinformatics expertise, which is not present in most data producing proteomics laboratories. In order to allow a breakthrough for the widespread application of newly developed algorithms and software tools, it is necessary to develop infrastructure programs, which provide data processing services using integrated tools with access to high-capacity, parallel computing facilities, such as large local clusters or grid. Indeed biologists planning proteomics or in general life science experiments to answer relevant biological questions may work more efficiently if they have to concentrate only on the experimental design of the biological study, on production of high quality data and interpretation of the obtained data using easy-to-use, user friendly data processing services. To facilitate data interpretation, the complexity of the software and hardware operations should be hidden, and the end user should be only exposed to raw

data management, parameter setting of data processing and other bioinformatics tools, to monitor the data processing status and to visualize the processed data in a user-friendly way. An example for such a framework software is Galaxy [230,231] or Genepattern [232,233], which is extensively used to analyze new generation DNA sequencing data. The framework should make the integration of new bioinformatics tool easy and allow to modify complex data processing workflows to adapt to the large diversity of mass spectrometers and sample preparation methods that generate highly diverse data. A key element for the efficient and easy integration of diverse bioinformatics tools in such a software framework is to use a standard format, which serves to interconnect the input and output files of the integrated tools.

Attempts have been made in the field of proteomics with the framework program CORRA [234], which integrates the SuperHirn [83] and SpecArray [86] label-free quantitative data processing pipelines, and includes MS/MS identification based on Sequest [150] with a range of R-based statistical tools. CORRA uses the Annotated Putative peptide Markup Language (APML) format to integrate the different modules of the quantitative data processing pipelines with protein identification and statistical analysis. CORRA provides a user-friendly web interface and executes the different processing tasks on a local cluster. However CORRA cannot manage large amounts of diverse metadata, as needed for effective project management, and the integrated tools are limited to bioinformatics modules developed in a closely collaborating bioinformatics community. Developing further CORRA and APML or other similar initiatives should provide a breakthrough in using newly developed bioinformatics tools and therefore accelerate life science research.

Acknowledgements

This work was part of the Bioassist and BioRange programs of the Netherlands Bioinformatics Centre (NBIC) and Gaining Momentum Initiatives of NBIC and the Netherlands Proteomics Center (NPC). NBIC and NPC programs are supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

[1] J. van der Greef, P. Stroobant, R. van der Heijden, Curr. Opin. Chem. Biol. 8 (2004) 559–565.
[2] N.L. Anderson, N.G. Anderson, Mol. Cell. Proteomics 1 (2002) 845–867.
[3] R. Schiess, B. Wollscheid, R. Aebersold, Mol. Oncol. 3 (2009) 33–44.
[4] Y. Shen, J.M. Jacobs, D.G. Camp II, R. Fang, R.J. Moore, R.D. Smith, W. Xiao, R.W. Davis, R.G. Tompkins, Anal. Chem. 76 (2004) 1134–1144.
[5] M. Pernemalm, R. Lewensohn, J. Lehtio, Proteomics 9 (2009) 1420–1427.
[6] M.W. Duncan, R. Aebersold, R.M. Caprioli, Nat. Biotechnol. 28 (2010) 659–664.
[7] M.W. Duncan, A.L. Yergey, S.D. Patterson, Proteomics 9 (2009) 1124–1127.
[8] P. Horvatovich, N. Govorukhina, R. Bischoff, Analyst 131 (2006) 1193–1196.
[9] P.L. Horvatovich, R. Bischoff, Eur. J. Mass Spectrom. (Chichester, England) 16 (2010) 101–121.
[10] Y. Guo, Z. Fu, J.E. Van Eyk, Proc. Am. Thorac Soc. 4 (2007) 9–17.
[11] A. Negishi, M. Ono, Y. Handa, H. Kato, K. Yamashita, K. Honda, M. Shitashige, R. Satow, T. Sakuma, H. Kuwabara, K. Omura, S. Hirohashi, T. Yamada, Cancer Sci. 100 (2009) 514–519.
[12] C.I. Balog, P.J. Hensbergen, R. Derks, J.J. Verweij, G.J. van Dam, B.J. Vennervald, A.M. Deelder, O.A. Mayboroda, Clin. Chem. 55 (2009) 117–125.
[13] R. Ralhan, L.V. Desouza, A. Matta, S. Chandra Tripathi, S. Ghanny, S. Datta Gupta, S. Bahadur, K.W. Siu, Mol. Cell. Proteomics 7 (2008) 1162–1173.
[14] R.E. Higgs, M.D. Knierman, V. Gelfanova, J.P. Butler, J.E. Hale, Methods Mol. Biol. 428 (2008) 209–230.
[15] Y. Levin, E. Schwarz, L. Wang, F.M. Leweke, S. Bahn, J. Sep. Sci 30 (2007) 2198–2203.
[16] S. Pan, H. Zhang, J. Rush, J. Eng, N. Zhang, D. Patterson, M.J. Comb, R. Aebersold, Mol. Cell. Proteomics 4 (2005) 182–190.
[17] J. Gao, L.A. Garulacan, S.M. Storm, G.J. Opiteck, Y. Dubaquie, S.A. Hefta, D.M. Dambach, A.R. Dongre, Methods 35 (2005) 291–302.
[18] D. Radulovic, S. Jelveh, S. Ryu, T.G. Hamilton, E. Foss, Y. Mao, A. Emili, Mol. Cell. Proteomics 3 (2004) 984–997.
[19] S. Bodovitz, T. Joos, Trends Biotechnol. 22 (2004) 4–7.
[20] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, B. Kuster, Anal. Bioanal. Chem. 389 (2007) 1017–1031.
[21] M. Gstaiger, R. Aebersold, Nat. Rev. Genet. 10 (2009) 617–627.
[22] G.W. Becker, Brief Funct. Genomic Proteomic 7 (2008) 371–382.
[23] K. Gevaert, F. Impens, B. Ghesquiere, P. Van Damme, A. Lambrechts, J. Vandekerckhove, Proteomics 8 (2008) 4873–4885.
[24] A. Leitner, W. Lindner, Proteomics 6 (2006) 5418–5434.
[25] I.C. Guerrera, O. Kleiner, Biosci. Rep. 25 (2005) 71–93.
[26] W. Yan, S.S. Chen, Brief Funct. Genomic Proteomic 4 (2005) 27–38.
[27] A. Panchaud, M. Affolter, P. Moreillon, M. Kussmann, J. Proteomics 71 (2008) 19–33.
[28] M. Buyse, D.J. Sargent, A. Grothey, A. Matheson, A. de Gramont, Nat. Rev. Clin. Oncol. 7 (2010) 309–317.
[29] T. Nilsson, M. Mann, R. Aebersold, J.R. Yates III, A. Bairoch, J.J. Bergeron, Nat. Methods 7 (2010) 681–685.
[30] U. Scherf, R. Becker, M. Chan, S. Hojvat, Scand. J. Clin. Lab. Invest. Suppl. 242 (2010) 96–102.
[31] D.S. Tan, G.V. Thomas, M.D. Garrett, U. Banerji, J.S. de Bono, S.B. Kaye, P. Workman, Cancer J. 15 (2009) 406–420.
[32] D. Lin, Z. Hollander, A. Meredith, B.M. McManus, Can. J. Cardiol. 25 (Suppl. A) (2009) 9A–14A.
[33] B.K. Dunn, P.D. Wagner, D. Anderson, P. Greenwald, Semin. Oncol. 37 (2010) 224–242.
[34] H. Mischak, G. Allmaier, R. Apweiler, T. Attwood, M. Baumann, A. Benigni, S.E. Bennett, R. Bischoff, E. Bongcam-Rudloff, G. Capasso, J.J. Coon, P. D'Haese, A.F. Dominiczak, M. Dakna, H. Dihazi, J.H. Ehrich, P. Fernandez-Llama, D. Fliser, J. Frokiaer, J. Garin, M. Girolami, W.S. Hancock, M. Haubitz, D. Hochstrasser, R.R. Holman, J.P. Ioannidis, J. Jankowski, B.A. Julian, J.B. Klein, W. Kolch, T. Luider, Z. Massy, W.B. Mattes, F. Molina, B. Monsarrat, J. Novak, K. Peter, P. Rossing, M. Sanchez-Carbayo, J.P. Schanstra, O.J. Semmes, G. Spasovski, D. Theodorescu, V. Thongboonkerd, R. Vanholder, T.D. Veenstra, E. Weissinger, T. Yamamoto, A. Vlahou, Sci. Transl. Med. 2 (2010) 46ps42.
[35] G. Chen, B.N. Pramanik, Y.H. Liu, U.A. Mirza, J. Mass Spectrom. 42 (2007) 279–287.
[36] G. Chen, B.N. Pramanik, Drug Discov. Today 14 (2009) 465–471.
[37] S. Rho, S. You, Y. Kim, D. Hwang, BMB Rep. 41 (2008) 184–193.
[38] E.D. Hoffmann, Mass Spectrometry: Principles and Applications, 3rd ed., J. Wiley, Chichester, West Sussex, England, 2007.
[39] J.R. Yates, C.I. Ruse, A. Nakorchevsky, Annu. Rev. Biomed. Eng. 11 (2009) 49–79.
[40] R.S. Ekman, J. Westman-Brinkmalm, A. Kraj, Mass Spectrometry: Instrumentation, Interpretation, and Applications, John Wiley & Sons, Inc., New Jersey, 2009.
[41] J.T. Watson, O.D. Sparkman, Introduction to Mass Spectrometry: Instrumentation, Applications and Strategies for Data Interpretation, fourth ed., John Wiley & Sons, Ltd., Chichester, 2007.
[42] B. Domon, R. Aebersold, Science 312 (2006) 212–217.
[43] X. Han, A. Aslanian, J.R. Yates III, Curr. Opin. Chem. Biol. 12 (2008) 483–490.
[44] B. Domon, R. Aebersold, Nat. Biotechnol. 28 (2010) 710–721.
[45] H. Liu, R.G. Sadygov, J.R. Yates III, Anal. Chem. 76 (2004) 4193–4201.
[46] W.M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K.G. Pierce, A. Mendoza, J.R. Sevinsky, K.A. Resing, N.G. Ahn, Mol. Cell. Proteomics 4 (2005) 1487–1502.
[47] J. Grossmann, B. Roschitzki, C. Panse, C. Fortes, S. Barkow-Oesterreicher, D. Rutishauser, R. Schlapbach, J. Proteomics 73 (2010) 1740–1746.
[48] L.N. Mueller, M.Y. Brusniak, D.R. Mani, R. Aebersold, J. Proteome Res. 7 (2008) 51–61.
[49] M. Li, W. Gray, H. Zhang, C.H. Chung, D. Billheimer, W.G. Yarbrough, D.C. Liebler, Y. Shyr, R.J. Slebos, J. Proteome Res. 9 (2010) 4295–4305.
[50] W. Zhu, J.W. Smith, C.M. Huang, J. Biomed. Biotechnol. (2010) 840518.
[51] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, M. Mann, Mol. Cell. Proteomics 4 (2005) 1265–1272.
[52] K. Shinoda, M. Tomita, Y. Ishihama, Bioinformatics 26 (2010) 576–577.
[53] J.C. Braisted, S. Kuntumalla, C. Vogel, E.M. Marcotte, A.R. Rodrigues, R. Wang, S.T. Huang, E.S. Ferlanti, A.I. Saeed, R.D. Fleischmann, S.N. Peterson, R. Pieper, BMC Bioinform. 9 (2008) 529.
[54] P. Lu, C. Vogel, R. Wang, X. Yao, E.M. Marcotte, Nat. Biotechnol. 25 (2007) 117–124.
[55] J.C. Silva, M.V. Gorenstein, G.Z. Li, J.P. Vissers, S.J. Geromanos, Mol. Cell. Proteomics 5 (2006) 144–156.
[56] D.L. Tabb, L. Vega-Montoto, P.A. Rudnick, A.M. Variyath, A.J. Ham, D.M. Bunk, L.E. Kilpatrick, D.D. Billheimer, R.K. Blackman, H.L. Cardasis, S.A. Carr, K.R. Clauser, J.D. Jaffe, K.A. Kowalski, T.A. Neubert, F.E. Regnier, B. Schilling, T.J. Tegeler, M. Wang, P. Wang, J.R. Whiteaker, L.J. Zimmerman, S.J. Fisher, B.W. Gibson, C.R. Kinsinger, M. Mesri, H. Rodriguez, S.E. Stein, P. Tempst, A.G. Paulovich, D.C. Liebler, C. Spiegelman, J. Proteome Res. 9 (2010) 761–776.
[57] A. Schmidt, M. Claassen, R. Aebersold, Curr. Opin. Chem. Biol. 13 (2009) 510–517.
[58] P. Picotti, B. Bodenmiller, L.N. Mueller, B. Domon, R. Aebersold, Cell 138 (2009) 795–806.
[59] P. Picotti, O. Rinner, R. Stallmach, F. Dautel, T. Farrah, B. Domon, H. Wenschuh, R. Aebersold, Nat. Methods 7 (2010) 43–46.
[60] S. Pan, R. Aebersold, R. Chen, J. Rush, D.R. Goodlett, M.W. McIntosh, J. Zhang, T.A. Brentnall, J. Proteome Res. 8 (2009) 787–797.
[61] E.W. Deutsch, H. Lam, R. Aebersold, EMBO Rep. 9 (2008) 429–434.
[62] P. Mallick, M. Schirle, S.S. Chen, M.R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, R. Aebersold, Nat. Biotechnol. 25 (2007) 125–131.
[63] C.A. Sherwood, A. Eastham, L.W. Lee, A. Peterson, J.K. Eng, D. Shteynberg, L. Mendoza, E.W. Deutsch, J. Risler, N. Tasman, R. Aebersold, H. Lam, D.B. Martin, J. Proteome Res. 8 (2009) 4396–4405.

[64] X. Yang, I.M. Lazar, BMC Cancer 9 (2009) 96.
[65] R. Huttenhain, J. Malmstrom, P. Picotti, R. Aebersold, Curr. Opin. Chem. Biol. 13 (2009) 518–525.
[66] V. Lange, P. Picotti, B. Domon, R. Aebersold, Mol. Syst. Biol. 4 (2008) 222.
[67] S. Orchard, P. Jones, C. Taylor, W. Zhu, R.K. Julian Jr., H. Hermjakob, R. Apweiler, Methods Mol. Biol. 367 (2007) 261–270.
[68] S. Orchard, C. Taylor, H. Hermjakob, W. Zhu, R. Julian, R. Apweiler, Expert Rev. Proteomics 1 (2004) 179–183.
[69] R.G. Cote, F. Reisinger, L. Martens, Proteomics 10 (2010) 1332–1335.
[70] E. Deutsch, Proteomics 8 (2008) 2776–2777.
[71] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W.H. Tang, A. Rompp, S. Neumann, A.D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.A. Binz, E.W. Deutsch, Mol. Cell. Proteomics (2010).
[72] C.F. Taylor, Proteomics 6 (Suppl. 2) (2006) 39–44.
[73] C.F. Taylor, Drug Discov. Today 12 (2007) 527–533.
[74] C.F. Taylor, N.W. Paton, K.S. Lilley, P.A. Binz, R.K. Julian Jr., A.R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E.W. Deutsch, M.J. Dunn, A.J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T.A. Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T.M. Vondriska, J.P. Whitelegge, M.R. Wilkins, I. Xenarios, J.R. Yates III, H. Hermjakob, Nat. Biotechnol. 25 (2007) 887–893.
[75] M.E. Monroe, N. Tolic, N. Jaitly, J.L. Shaw, J.N. Adkins, R.D. Smith, Bioinformatics 23 (2007) 2021–2023.
[76] O. Kohlbacher, K. Reinert, C. Gropl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, M. Sturm, Bioinformatics 23 (2007) e191–e197.
[77] E. Lange, C. Gropl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber, K. Reinert, Bioinformatics 23 (2007) i273–i281.
[78] K. Reinert, O. Kohlbacher, Methods Mol. Biol. 604 (2010) 201–211.
[79] M. Sturm, A. Bertsch, C. Gropl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, O. Kohlbacher, BMC Bioinform. 9 (2008) 163.
[80] M. Katajamaa, J. Miettinen, M. Oresic, Bioinformatics 22 (2006) 634–636.
[81] M. Katajamaa, M. Oresic, BMC Bioinform. 6 (2005) 179.
[82] D.K. Han, J. Eng, H. Zhou, R. Aebersold, Nat. Biotechnol. 19 (2001) 946–951.
[83] L.N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.Y. Brusniak, O. Vitek, R. Aebersold, M. Muller, Proteomics 7 (2007) 3470–3480.
[84] S.K. Park, J.D. Venable, T. Xu, J.R. Yates III, Nat. Methods 5 (2008) 319–322.
[85] K.C. Leptos, D.A. Sarracino, J.D. Jaffe, B. Krastins, G.M. Church, Proteomics 6 (2006) 1770–1782.
[86] X.J. Li, E.C. Yi, C.J. Kemp, H. Zhang, R. Aebersold, Mol. Cell. Proteomics 4 (2005) 1328–1340.
[87] H.D. Meiring, E.C. Soethout, A.P. de Jong, C.A. van Els, Curr. Protoc. Immunol. 77 (2007), 16.3.1–16.3.20.
[88] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M.A. Gillette, S.A. Carr, Mol. Cell. Proteomics 5 (2006) 1927–1941.
[89] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779–787.
[90] R. Tautenhahn, C. Bottcher, S. Neumann, BMC Bioinform. 9 (2008) 504.
[91] K.M. Aberg, R.J. Torgrip, J. Kolmert, I. Schuppe-Koistinen, J. Lindberg, J. Chromatogr. A 1192 (2008) 139–146.
[92] R. Stolt, R.J. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chem. 78 (2006) 975–983.
[93] W. Windig, J.M. Phalp, A.W. Payne, Anal. Chem. 68 (1996) 3602–3606.
[94] C. Christin, H.C. Hoefsloot, A.K. Smilde, F. Suits, R. Bischoff, P.L. Horvatovich, J. Proteome Res. (2010).
[95] C. Christin, A.K. Smilde, H.C. Hoefsloot, F. Suits, R. Bischoff, P.L. Horvatovich, Anal. Chem. 80 (2008) 7012–7021.
[96] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, F. Suits, Bioinformatics 24 (2008) 1070–1077.
[97] K.L. Busch, Spectroscopy 17 (2002) 6.
[98] T.M. Annesley, Clin. Chem. 49 (2003) 1041–1044.
[99] P. Horvatovich, N.I. Govorukhina, T.H. Reijmers, A.G. van der Zee, F. Suits, R. Bischoff, Electrophoresis 28 (2007) 4493–4505.
[100] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, C.H. Becker, Anal. Chem. 75 (2003) 4818–4826.
[101] Y. Li, H. Qu, Y. Cheng, Anal. Chim. Acta 612 (2008) 19–22.
[102] J. Listgarten, A. Emili, Mol. Cell. Proteomics 4 (2005) 419–434.
[103] T. Yu, Y. Park, J.M. Johnson, D.P. Jones, Bioinformatics 25 (2009) 1930–1936.
[104] V.P. Andreev, T. Rejtar, H.S. Chen, E.V. Moskovets, A.R. Ivanov, B.L. Karger, Anal. Chem. 75 (2003) 6314–6326.
[105] S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti, L. Pattini, Anal. Chem. 80 (2008) 4960–4968.
[106] S. Cappadona, P. Nanni, M. Benevento, F. Levander, P. Versura, A. Roda, S. Cerutti, L. Pattini, J. Biomed. Biotechnol. 2010 (2010) 131505.
[107] P. Du, R. Sudha, M.B. Prystowsky, R.H. Angeletti, Bioinformatics 23 (2007) 1394–1400.
[108] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, M. McIntosh, Bioinformatics 22 (2006) 1902–1909.
[109] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, Y. Huang, Curr. Genomics 10 (2009) 388–401.
[110] J.C. de Groot, M.W. Fiers, R.C. van Ham, A.H. America, Proteomics 8 (2008) 32–36.
[111] B.D. Piening, P. Wang, C.S. Bangur, J. Whiteaker, H. Zhang, L.C. Feng, J.F. Keane, J.K. Eng, H. Tang, A. Prakash, M.W. McIntosh, A. Paulovich, J. Proteome Res. 5 (2006) 1527–1534.
[112] J.S. Zimmer, M.E. Monroe, W.J. Qian, R.D. Smith, Mass Spectrom. Rev. 25 (2006) 450–482.
[113] R.A. Scheltema, A. Kamleh, D. Wildridge, C. Ebikeme, D.G. Watson, M.P. Barrett, R.C. Jansen, R. Breitling, Proteomics 8 (2008) 4647–4656.
[114] W. Haas, B.K. Faherty, S.A. Gerber, J.E. Elias, S.A. Beausoleil, C.E. Bakalarski, X. Li, J. Villen, S.P. Gygi, Mol. Cell. Proteomics 5 (2006) 1326–1337.
[115] R.M. Danell, S.A. Ouvry-Patat, C.O. Scarlett, J.P. Speir, C.H. Borchers, J. Am. Soc. Mass Spectrom. 19 (2008) 1914–1925.
[116] M. Dijkstra, R.C. Jansen, Proteomics 9 (2009) 3869–3876.
[117] R.F. Kemperman, P.L. Horvatovich, B. Hoekman, T.H. Reijmers, F.A. Muskiet, R. Bischoff, J. Proteome Res. 6 (2007) 194–206.
[118] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed, Nucleic Acids Res. 30 (2002) e15.
[119] K. Kultima, A. Nilsson, B. Scholz, U.L. Rossbach, M. Falth, P.E. Andren, Mol. Cell. Proteomics 8 (2009) 2285–2295.
[120] S.J. Callister, R.C. Barry, J.N. Adkins, E.T. Johnson, W.J. Qian, B.J. Webb-Robertson, R.D. Smith, M.S. Lipton, J. Proteome Res. 5 (2006) 277–286.
[121] Y.V. Karpievitch, T. Taverner, J.N. Adkins, S.J. Callister, G.A. Anderson, R.D. Smith, A.R. Dabney, Bioinformatics 25 (2009) 2573–2580.
[122] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, J. Chromatogr. A 961 (2002) 237–244.
[123] D. Clifford, G. Stone, I. Montoliu, S. Rezzi, F.o.-P. Martin, P. Guy, S. Bruce, S. Kochhar, Anal. Chem. (2009).
[124] P.H. Eilers, Anal. Chem. 76 (2004) 404–411.
[125] A. Kassidas, J.F. MacGregor, P.A. Taylor, AIChe 44 (1998) 864.
[126] N.-P.V. Nielsen, J.M. Carstensen, J.R. Smedsgaard, J. Chromatogr. A 805 (1998) 17–35.
[127] H.-J. Ramaker, E.N.M. van Sprang, J.A. Westerhuis, A.K. Smilde, Anal. Chim. Acta 498 (2003) 133–153.
[128] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231–241.
[129] A.M. van Nederkassel, M. Daszykowski, P.H. Eilers, Y.V. Heyden, J. Chromatogr. A 1118 (2006) 199–210.
[130] J. Vial, H. Nocairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiebaut, B. Teillet, D.N. Rutledge, J. Chromatogr. A 1216 (2009) 2866–2872.
[131] A. Prakash, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold, B. Schwikowski, Mol. Cell. Proteomics 5 (2006) 423–432.
[132] J.T. Prince, E.M. Marcotte, Anal. Chem. 78 (2006) 6140–6152.
[133] R.G. Sadygov, F.M. Maroto, A.F. Huhmer, Anal. Chem. 78 (2006) 8207–8217.
[134] K. Podwojski, A. Fritsch, D.C. Chamrad, W. Paul, B. Sitek, K. Stuhler, P. Mutzel, C. Stephan, H.E. Meyer, W. Urfer, K. Ickstadt, J. Rahnenfuhrer, Bioinformatics 25 (2009) 758–764.
[135] F. Suits, J. Lepre, P. Du, R. Bischoff, P. Horvatovich, Anal. Chem. 80 (2008) 3095–3104.
[136] N. Jaitly, M.E. Monroe, V.A. Petyuk, T.R. Clauss, J.N. Adkins, R.D. Smith, Anal. Chem. 78 (2006) 7397–7409.
[137] E. Lange, R. Tautenhahn, S. Neumann, C. Gropl, BMC Bioinform. 9 (2008) 375.
[138] J. Listgarten, R.M. Neal, S.T. Roweis, P. Wong, A. Emili, Bioinformatics 23 (2007) e198–e204.
[139] C.C. Tsou, C.F. Tsai, Y.H. Tsui, P.R. Sudhir, Y.T. Wang, Y.J. Chen, J.Y. Chen, T.Y. Sung, W.L. Hsu, Mol. Cell. Proteomics 9 (2010) 131–144.
[140] K.M. Aberg, E. Alm, R.J. Torgrip, Anal. Bioanal. Chem. 394 (2009) 151–162.
[141] M. De Beer, F. Lynen, K. Chen, P. Ferguson, M. Hanna-Brown, P. Sandra, Anal. Chem. 82 (2010) 1733–1743.
[142] M. Gilar, P. Olivova, A.E. Daly, J.C. Gebler, J. Sep. Sci. 28 (2005) 1694–1703.
[143] M. Gilar, P. Olivova, A.E. Daly, J.C. Gebler, Anal. Chem. 77 (2005) 6426–6434.
[144] M. Vandenbogaert, S. Li-Thiao-Te, H.M. Kaltenbach, R. Zhang, T. Aittokallio, B. Schwikowski, Proteomics 8 (2008) 650–672.
[145] A. Prakash, B. Piening, J. Whiteaker, H. Zhang, S.A. Shaffer, D. Martin, L. Hohmann, K. Cooke, J.M. Olson, S. Hansen, M.R. Flory, H. Lee, J. Watts, D.R. Goodlett, R. Aebersold, A. Paulovich, B. Schwikowski, Mol. Cell. Proteomics 6 (2007) 1741–1748.
[146] M. Hilario, A. Kalousis, C. Pellegrini, M. Muller, Mass Spectrom. Rev. 25 (2006) 409–449.
[147] E. Kapp, F. Schutz, Curr. Protoc. Protein Sci. 49 (2007), 25.2.1–25.2.19.
[148] A.I. Nesvizhskii, Methods Mol. Biol. 367 (2007) 87–119.
[149] R.G. Sadygov, D. Cociorva, J.R. Yates III, Nat. Methods 1 (2004) 195–202.
[150] J.K. Eng, A.L. McCormack, J.R. Yates Iii, J. Am. Soc. Mass Spectrom. 5 (1994) 976–989.
[151] M. Mann, M. Wilm, Anal. Chem. 66 (1994) 4390–4399.
[152] K.R. Clauser, P. Baker, A.L. Burlingame, Anal. Chem. 71 (1999) 2871–2882.
[153] D.L. Tabb, A. Saraf, J.R. Yates III, Anal. Chem. 75 (2003) 6415–6421.
[154] V. Bafna, N. Edwards, Bioinformatics 17 (Suppl. 1) (2001) S13–S21.
[155] J. Colinge, A. Masselot, M. Giron, T. Dessingy, J. Magnin, Proteomics 3 (2003) 1454–1463.
[156] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Electrophoresis 20 (1999) 3551–3567.
[157] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant, J. Proteome Res. 3 (2004) 958–964.
[158] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, R. Aebersold, Proteomics 7 (2007) 655–667.
[159] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, S.E. Stein, R. Aebersold, Nat. Methods 5 (2008) 873–875.

[160] A.M. Frank, N. Bandeira, Z. Shen, S. Tanner, S.P. Briggs, R.D. Smith, P.A. Pevzner, J. Proteome Res. 7 (2008) 113–122.
[161] P. Hernandez, R. Gras, J. Frey, R.D. Appel, Proteomics 3 (2003) 870–878.
[162] S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, V. Bafna, Anal. Chem. 77 (2005) 4626–4639.
[163] B. Ma, G. Lajoie, Curr. Protoc. Bioinform. 25 (2009), 13.10.1–13.10.8.
[164] N.S. Tannu, S.E. Hemby, BMC Genomics 8 (2007) 270.
[165] A. Frank, P. Pevzner, Anal. Chem. 77 (2005) 964–973.
[166] M. Bern, D. Goldberg, J. Comput. Biol. 13 (2006) 364–378.
[167] J.A. Taylor, R.S. Johnson, Anal. Chem. 73 (2001) 2594–2604.
[168] V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, P.A. Pevzner, J. Comput. Biol. 6 (1999) 327–342.
[169] L. Mo, D. Dutta, Y. Wan, T. Chen, Anal. Chem. 79 (2007) 4870–4878.
[170] P.A. DiMaggio Jr., C.A. Floudas, Anal. Chem. 79 (2007) 1433–1446.
[171] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, J.M. Buhmann, Anal. Chem. 77 (2005) 7265–7273.
[172] J. Grossmann, F.F. Roos, M. Cieliebak, Z. Liptak, L.K. Mathis, M. Muller, W. Gruissem, S. Baginsky, J. Proteome Res. 4 (2005) 1768–1774.
[173] A.D. Norbeck, M.E. Monroe, J.N. Adkins, K.K. Anderson, D.S. Daly, R.D. Smith, J. Am. Soc. Mass Spectrom. 16 (2005) 1239–1249.
[174] C.D. Masselon, S. Kieffer-Jaquinod, S. Brugiere, V. Dupierris, J. Garin, Rapid Commun. Mass Spectrom. 22 (2008) 986–992.
[175] Y.J. Kim, B. Feild, W. Fitzhugh, J.L. Heidbrink, J.W. Duff, J. Heil, S.M. Ruben, T. He, Anal. Biochem. 393 (2009) 155–162.
[176] B.C. Searle, Proteomics 10 (2010) 1265–1269.
[177] R. Craig, R.C. Beavis, Bioinformatics 20 (2004) 1466–1467.
[178] R.D. Bjornson, N.J. Carriero, C. Colangelo, M. Shifman, K.H. Cheung, P.L. Miller, K. Williams, J. Proteome Res. 7 (2008) 293–299.
[179] T. Baczek, R. Kaliszan, Proteomics 9 (2009) 835–847.
[180] M. Gilar, A. Jaworski, P. Olivova, J.C. Gebler, Rapid Commun. Mass Spectrom. 21 (2007) 2813–2821.
[181] K. Shinoda, M. Sugimoto, M. Tomita, Y. Ishihama, Proteomics 8 (2008) 787–798.
[182] A. Quandt, A. Masselot, P. Hernandez, C. Hernandez, S. Maffioletti, R.D. Appel, F. Lisacek, Proteomics 9 (2009) 2648–2655.
[183] A. Quandt, P. Hernandez, A. Masselot, C. Hernandez, S. Maffioletti, C. Pautasso, R.D. Appel, F. Lisacek, Bioinformatics 24 (2008) 1416–1417.
[184] N. Bandeira, D. Tsur, A. Frank, P.A. Pevzner, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 6140–6145.
[185] N. Bandeira, J.V. Olsen, J.V. Mann, M. Mann, P.A. Pevzner, Bioinformatics 24 (2008) i416–i423.
[186] S. Kim, N. Gupta, N. Bandeira, P.A. Pevzner, Mol. Cell. Proteomics 8 (2009) 53–69.
[187] L. Bianco, J. Mead, C. Bessant, J. Proteome Res. (2009).
[188] M. Fitzgibbon, Q. Li, M. McIntosh, J. Proteome Res. 7 (2008) 35–39.
[189] J.E. Elias, S.P. Gygi, Methods Mol. Biol. 604 (2010) 55–71.
[190] K. Pendarvis, R. Kumar, S.C. Burgess, B. Nanduri, BMC Bioinform. 10 (Suppl. 11) (2009) S17.
[191] C. Hughes, B. Ma, G.A. Lajoie, Methods Mol. Biol. 604 (2010) 105–121.
[192] S.J.H.A.R. Jones, Proteome Bioinformatics, Humana Press, Hatfield, 2010.
[193] C. Xu, B. Ma, Drug Discov. Today 11 (2006) 595–600.
[194] J. Salmi, T.A. Nyman, O.S. Nevalainen, T. Aittokallio, Proteomics 9 (2009) 848–860.
[195] S.J. Barton, J.C. Whittaker, Mass Spectrom. Rev. 28 (2009) 177–187.
[196] Y. Zhang, Z. Wen, M.P. Washburn, L. Florens, Anal. Chem. 82 (2010) 2272–2281.
[197] K. Podwojski, M. Eisenacher, M. Kohl, M. Turewicz, H.E. Meyer, J. Rahnenfuhrer, C. Stephan, Expert Rev. Proteomics 7 (2010) 249–261.
[198] K.W. Lau, A.R. Jones, N. Swainston, J.A. Siepen, S.J. Hubbard, Proteomics 7 (2007) 2787–2799.
[199] H. Kim, S. Eliuk, J. Deshane, S. Meleth, T. Sanderson, A. Pinner, G. Robinson, L. Wilson, M. Kirk, S. Barnes, Methods Mol. Biol. 371 (2007) 349–391.
[200] J.L. Lopez, J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 849 (2007) 190–202.
[201] E. Marengo, E. Robotti, M. Bobba, Methods Mol. Biol. 428 (2008) 291–325.
[202] L.L. Lv, B.C. Liu, Expert Rev. Proteomics 4 (2007) 505–513.
[203] J.D. Reid, C.E. Parker, C.H. Borchers, Curr. Opin. Mol. Ther. 9 (2007) 216–221.
[204] A. VanMeter, M. Signore, M. Pierobon, V. Espina, L.A. Liotta, E.F. Petricoin III, Expert Rev. Mol. Diagn. 7 (2007) 625–633.
[205] S.C. Basak, B.D. Gute, Curr. Opin. Drug Discov. Dev. 11 (2008) 320–326.
[206] L. Eriksson, E. Johansson, F. Lindgren, M. Sjostrom, S. Wold, J. Comput. Aided Mol. Des. 16 (2002) 711–726.
[207] E.S. Ronald, J. Chemometrics 16 (2002) 261–262.
[208] E. Byvatov, G. Schneider, Appl. Bioinform. 2 (2003) 67–77.
[209] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, R.A. Clark, Artif. Intell. Med. 32 (2004) 71–83.
[210] J.C. Rajapakse, K.B. Duan, W.K. Yeo, Am. J. Pharmacogenomics 5 (2005) 281–292.
[211] J.S. Torrecilla, E. Rojo, M. Oliet, J.C. Dominguez, F. Rodriguez, J. Agric. Food Chem. 57 (2009) 2763–2769.
[212] T. Villmann, F.M. Schleif, M. Kostrzewa, A. Walch, B. Hammer, Brief. Bioinform. 9 (2008) 129–143.
[213] H.J. Issaq, Q.N. Van, T.J. Waybright, G.M. Muschik, T.D. Veenstra, J. Sep. Sci. 32 (2009) 2183–2199.
[214] S. Smit, H.C. Hoefsloot, A.K. Smilde, J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 866 (2008) 77–88.
[215] M.M. Hendriks, S. Smit, W.L. Akkermans, T.H. Reijmers, P.H. Eilers, H.C. Hoefsloot, C.M. Rubingh, C.G. de Koster, J.M. Aerts, A.K. Smilde, Proteomics 7 (2007) 3672–3680.
[216] S. Smit, M.J. van Breemen, H.C. Hoefsloot, A.K. Smilde, J.M. Aerts, C.G. de Koster, Anal. Chim. Acta 592 (2007) 210–217.
[217] J. Reunanen, Feature Extraction, 2006, pp. 119–136.
[218] P. Silcocks, in: X.-H. Zou, N. Obuchowski, D. McClish (Eds.), Statistical Methods in Diagnostic Medicine, Wiley & Sons Interscience, New York, 2002.
[219] D.A. Stead, N.W. Paton, P. Missier, S.M. Embury, C. Hedeler, B. Jin, A.J. Brown, A. Preece, Brief. Bioinform. 9 (2008) 174–188.
[220] H. Roelofsen, G. Alvarez-Llamas, M. Dijkstra, R. Breitling, K. Havenga, J. Bijzet, W. Zandbergen, M.P. de Vries, R.J. Ploeg, R.J. Vonk, Proteomics 7 (2007) 3219–3228.
[221] D.J. Vis, J.A. Westerhuis, A.K. Smilde, J. van der Greef, BMC Bioinform. 8 (2007) 322.
[222] A.K. Smilde, J.J. Jansen, H.C. Hoefsloot, R.J. Lamers, J. van der Greef, M.E. Timmerman, Bioinformatics 21 (2005) 3043–3048.
[223] M.J. Nueda, A. Conesa, J.A. Westerhuis, H.C. Hoefsloot, A.K. Smilde, M. Talon, A. Ferrer, Bioinformatics 23 (2007) 1792–1800.
[224] R. Zhang, A. Barton, J. Brittenden, J.T.J. Huang, D. Crowther, J. Proteomics Bioinform. 3 (2010) 6.
[225] A.G. Paulovich, D. Billheimer, A.J. Ham, L. Vega-Montoto, P.A. Rudnick, D.L. Tabb, P. Wang, R.K. Blackman, D.M. Bunk, H.L. Cardasis, K.R. Clauser, C.R. Kinsinger, B. Schilling, T.J. Tegeler, A.M. Variyath, M. Wang, J.R. Whiteaker, L.J. Zimmerman, D. Fenyo, S.A. Carr, S.J. Fisher, B.W. Gibson, M. Mesri, T.A. Neubert, F.E. Regnier, H. Rodriguez, C. Spiegelman, S.E. Stein, P. Tempst, D.C. Liebler, Mol. Cell. Proteomics 9 (2010) 242–254.
[226] K. Kandasamy, S. Keerthikumar, R. Goel, S. Mathivanan, N. Patankar, B. Shafreen, S. Renuse, H. Pawar, Y.L. Ramachandra, P.K. Acharya, P. Ranganathan, R. Chaerkady, T.S. Keshava Prasad, A. Pandey, Nucleic Acids Res. 37 (2009) D773–D781.
[227] S. Mathivanan, M. Ahmed, N.G. Ahn, H. Alexandre, R. Amanchy, P.C. Andrews, J.S. Bader, B.M. Balgley, M. Bantscheff, K.L. Bennett, E. Bjorling, B. Blagoev, R. Bose, S.K. Brahmachari, A.S. Burlingame, X.R. Bustelo, G. Cagney, G.T. Cantin, H.L. Cardasis, J.E. Celis, R. Chaerkady, F. Chu, P.A. Cole, C.E. Costello, R.J. Cotter, D. Crockett, J.P. DeLany, A.M. De Marzo, L.V. DeSouza, E.W. Deutsch, E. Dransfield, G. Drewes, A. Droit, M.J. Dunn, K. Elenitoba-Johnson, R.M. Ewing, J. Van Eyk, V. Faca, J. Falkner, X. Fang, C. Fenselau, D. Figeys, P. Gagne, C. Gelfi, K. Gevaert, J.M. Gimble, F. Gnad, R. Goel, P. Gromov, S.M. Hanash, W.S. Hancock, H.C. Harsha, G. Hart, F. Hays, F. He, P. Hebbar, K. Helsens, H. Hermeking, W. Hide, K. Hjerno, D.F. Hochstrasser, O. Hofmann, D.M. Horn, R.H. Hruban, N. Ibarrola, P. James, O.N. Jensen, P.H. Jensen, P. Jung, K. Kandasamy, I. Kheterpal, R.F. Kikuno, U. Korf, R. Korner, B. Kuster, M.S. Kwon, H.J. Lee, Y.J. Lee, M. Lefevre, M. Lehvaslaiho, P. Lescuyer, F. Levander, M.S. Lim, C. Lobke, J.A. Loo, M. Mann, L. Martens, J. Martinez-Heredia, M. McComb, J. McRedmond, A. Mehrle, R. Menon, C.A. Miller, H. Mischak, S.S. Mohan, R. Mohmood, H. Molina, M.F. Moran, J.D. Morgan, R. Moritz, M. Morzel, D.C. Muddiman, A. Nalli, J.D. Navarro, T.A. Neubert, O. Ohara, R. Oliva, G.S. Omenn, M. Oyama, Y.K. Paik, K. Pennington, R. Pepperkok, B. Periaswamy, E.F. Petricoin, G.G. Poirier, T.S. Prasad, S.O. Purvine, B.A. Rahiman, P. Ramachandran, Y.L. Ramachandra, R.H. Rice, J. Rick, R.H. Ronnholm, J. Salonen, J.C. Sanchez, T. Sayd, B. Seshi, K. Shankari, S.J. Sheng, V. Shetty, K. Shivakumar, R.J. Simpson, R. Sirdeshmukh, K.W. Siu, J.C. Smith, R.D. Smith, D.J. States, S. Sugano, M. Sullivan, G. Superti-Furga, M. Takatalo, V. Thongboonkerd, J.C. Trinidad, M. Uhlen, J. Vandekerckhove, J. Vasilescu, T.D. Veenstra, J.M. Vidal-Taboada, M. Vihinen, R. Wait, X. Wang, S. Wiemann, B. Wu, T. Xu, J.R. Yates, J. Zhong, M. Zhou, Y. Zhu, P. Zurbig, A. Pandey, Nat. Biotechnol. 26 (2008) 164–167.
[228] S. Mathivanan, A. Pandey, Mol. Cell. Proteomics 7 (2008) 2038–2047.
[229] S. Peters, E. van Velzen, H.G. Janssen, Anal Bioanal Chem 394 (2009) 1273–1281.
[230] J. Taylor, I. Schenck, D. Blankenberg, A. Nekrutenko, Curr. Protoc. Bioinform. 19 (2007), 10.5.1–10.5.25.
[231] D. Blankenberg, J. Taylor, I. Schenck, J. He, Y. Zhang, M. Ghent, N. Veeraraghavan, I. Albert, W. Miller, K.D. Makova, R.C. Hardison, A. Nekrutenko, Genome Res. 17 (2007) 960–964.
[232] H. Kuehn, A. Liberzon, M. Reich, J.P. Mesirov, Curr. Protoc. Bioinform. 22 (2008), 7.12.1–7.12.39.
[233] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, J.P. Mesirov, Nat. Genet. 38 (2006) 500–501.
[234] M.Y. Brusniak, B. Bodenmiller, D. Campbell, K. Cooke, J. Eddes, A. Garbutt, H. Lau, S. Letarte, L.N. Mueller, V. Sharma, O. Vitek, N. Zhang, R. Aebersold, J.D. Watts, BMC Bioinform. 9 (2008) 542.