



RESEARCH ARTICLE

Design and Application of a Data-Independent Precursor and Product Ion Repository

Konstantinos Thalassinos,^{1,10} Johannes P. C. Vissers,² Stefan Tenzer,³ Yishai Levin,^{4,11} J. Will Thompson,⁵ David Daniel,⁵ Darrin Mann,⁵ Mark R. DeLong,⁵ M. Arthur Moseley,⁵ Antoine H. America,⁶ Andrew K. Ottens,⁷ Greg S. Cavey,⁸ Georgios Efstathiou,¹ James H. Scrivens,¹ James I. Langridge,² Scott J. Geromanos⁹

¹Biological Sciences, University of Warwick, Warwick, UK

²Waters Corporation, Manchester, UK

³Institute for Immunology, University Medical Center of the Johannes Gutenberg University, Mainz, Germany

⁴Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK

⁵Institute for Genome Sciences & Policy, Duke University School of Medicine, Durham, NC, USA

⁶Plant Research International, Wageningen, The Netherlands

⁷Departments of Anatomy and Neurobiology and Biochemistry, Virginia Commonwealth University, Richmond, VA, USA

⁸Southwest Michigan Innovation Center, Kalamazoo, MI, USA

⁹Waters Corporation, Milford, MA, USA

¹⁰*Present Address:* Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, UK

¹¹*Present Address:* Department of Biological Services, Weizmann Institute of Science, Rehovot, Israel

Abstract

The functional design and application of a data-independent LC-MS precursor and product ion repository for protein identification, quantification, and validation is conceptually described. The ion repository was constructed from the sequence search results of a broad range of discovery experiments investigating various tissue types of two closely related mammalian species. The relative high degree of similarity in protein complement, ion detection, and peptide and protein identification allows for the analysis of normalized precursor and product ion intensity values, as well as standardized retention times, creating a multidimensional/orthogonal queryable, qualitative, and quantitative space. Peptide ion map selection for identification and quantification is primarily based on replication and limited variation. The information is stored in a relational database and is used to create peptide- and protein-specific fragment ion maps that can be queried in a targeted fashion against the raw or time aligned ion detections. These queries can be conducted either individually or as groups, where the latter affords pathway and molecular machinery analysis of the protein complement. The presented results also suggest that peptide ionization and fragmentation efficiencies are highly conserved between experiments and practically independent of the analyzed biological sample when using similar instrumentation. Moreover, the data illustrate only minor variation in ionization efficiency with amino acid sequence substitutions occurring between species.

Konstantinos Thalassinos, Johannes P.C. Vissers and Scott J. Geromanos contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s13361-012-0416-9) contains supplementary material, which is available to authorized users.

Correspondence to: Johannes Vissers; e-mail: hans_vissers@waters.com

Received: 29 December 2011
Revised: 9 May 2012
Accepted: 13 May 2012
Published online: 31 July 2012

Finally, the data and the presented results illustrate how LC-MS performance metrics can be extracted and utilized to ensure optimal performance of the employed analytical workflows.

Key words: Protein and peptide identification, Quantification and validation, LC-MS, Label-free, Data-independent analysis, Fragment ion identification repository, Targeted and non-targeted proteomics

Introduction

The controversy surrounding the quality of discovery data generated from shotgun LC-MS proteomic experiments continues without resolution. For example, Ptolemy and Rifai [1] suggest a serious review of both the terminology and validation schema utilized in biomarker discovery experiments. The basis of their report is the apparent disconnect between the level of funding and effort associated with biomarker discovery and the limited number of protein biomarkers actually in use in routine clinical management. In addition, Ransohoff [2] details how the strong claims from both the genomic and proteomic biomarker initiatives suffer from relatively poor experimental design, reproducibility, and applicability. Likewise, White [3] discusses the potential cost of high-throughput proteomics, describing a culture that motivates laboratories to generate large lists of protein, peptide, and post-translational modification biomarker candidates, typically at the expense of accuracy and reproducibility. Fundamental data acquisition and data processing changes may be required to address accuracy and reproducibility issues [4–6]. The protein complement of comparable biological samples is known to be qualitatively and quantitatively similar. However, measurement efforts based on shotgun LC-MS experiments generally lack sufficient reproducibility. Stochastic and serendipitous data sampling arguments have been advanced as an explanation for why these experiments are not reproducible [7, 8], despite the protein complement and hydrophobicity and ionization efficiency of their proteolytic peptides being vastly the same. There is, however, a growing body of evidence that suggests that the selectivity of LC-MS/MS-based strategies may be insufficient to deal with the complexity of a proteolytically digested (sub)proteome [9–11].

A significant source of error in proteomic experiments results from the algorithmic interpretation of product ion spectra derived from chimeric and composite MS spectra. To date, in the instance of complex mixture experiments, most search engines do not acknowledge the fact that a typical data-dependent analysis (DDA) product ion spectrum is most likely to arise from co-fragmented peptides. Approximately two-thirds of all precursor ion detections in a complex protein digest mixture are at least two-and-a-half orders of magnitude lower in intensity than the most abundant ions [10, 12]. Consequently, the incidence of overlapping isotopic clusters of similar m/z and intensity is significant. The specificity of DDA acquisitions is challenged under such conditions, especially when the search

engine peptide score is primarily based on the intensity of the matched product ions relative to the unmatched. Acquiring DDA data faster or with higher sensitivity hardly reduces these sources of error. An increase in speed and sensitivity without a concurrent increase in specificity will generally produce compromised information by generating more low abundant mixed spectra. Overloading the separation column can also exacerbate the chimeric and composite challenge since this will produce peak broadening and tailing of higher abundance peptides and enhances the incidence of interference. On the other hand, improving the overall separation capacity of the LC-MS method can have a positive impact on error rates. This is however only achieved if column flow rate, gradient and sample loading are harmonized to reduce the incidence of composite and chimeric spectra. A secondary benefit of increased chromatographic resolution is that chromatographic peak widths are reduced, which improves electrospray sensitivity by presenting a higher peptide concentration per unit time to the mass spectrometer. Multidimensional chromatography, when properly implemented, should have a positive impact on overall separation capacity [13, 14]. Alternatively, an additional dimension of separation such as ion mobility (IM) can be very effective in reducing chimeric and composite interferences and the benefits of this approach have been demonstrated in data-independent analysis (DIA) strategies [14, 15]. The application and combination of IM with DDA is less common. The technical advantages and limitations of DDA and DIA methods, including their main differences, have been discussed in detail [9, 10, 16]. Lastly, data processing errors associated with charge state assignment, de-isotoping and centroiding can be especially problematic when processing low abundance, overlapping isotopic cluster data [11, 17–19].

Different statistical approaches have been used to estimate the contribution to error from peptide sequence database search algorithms. The most widely used method to date has been the use of a decoy database strategy, whereby the decoy database is concatenated to the database of interest to infer a false positive rate (FPR) or false discovery rate (FDR) at the peptide and/or protein level [20–22]. This approach is based on the assumption that peptides from a random or reverse decoy database can be identified at a rate similar to that of the peptides from the original database [20]. The amino acids sequences of proteins are, however, not organized randomly or in a reversed manner. More specifically, the frequency of various sequence motifs commonly found in a given proteome may not be correctly represented in

a decoy version of the database, resulting in an apparent low number of hits to the decoy database, in turn leading to an underestimation of identification error rate [3]. Generic statistical tools have been employed and their merits demonstrated [23–25] to calculate peptide and protein FDR and FPRs as well. Which peptides are identified and how they are scored varies significantly between the various employed methods.

Assuming that high resolution exact mass measurement of peptide precursor and product ions, peptide fragmentation efficiency, relative retention time, and drift time are similar on comparable instruments, operated in a comparable manner, querying these metrics, as well as the relative intensity values of peptide precursor and product ions, should provide means to identify and quantify proteins in complex mixtures. To that end, spectral library searching has been suggested as an alternative to the more traditional sequence database search approach [26]. In this strategy, an unknown spectrum is compared with a library of known spectra and a match achieved based on the similarity of physicochemical properties. Spectral searching and the use of libraries have been the premise of GC-MS for the interpretation of unknown spectra for some time [27–29]. Its utility in proteomics research has been further explored by a number of research groups, and it is moving gradually into more mainstream use and acceptance [30–34]. Spectral libraries must contain correctly identified spectra to have value. This generally requires the accumulation of replicate spectra, which results in challenging data storage and computational requirements, forming the motivation to cluster spectra and develop so-called spectral archives [35–37]. Currently, the sharing of experimental MS/MS data between laboratories to more effectively use spectral libraries and archives is not widespread. Several identification and data repositories such as PRIDE [38], Tranche [39], PeptideAtlas [40], and Peptidome [41] facilitate spectral upload, viewing, and comparison, but generally do not offer the ability to build validated composite MS/MS spectra, or conduct spectral or fragment ion searches against the validated composite spectra.

The work described in this paper demonstrates that the construction of a fragment ion repository using high-specificity product ion spectra, in combination with appropriate aggregation and query of the repository, provides promise as a strategy for characterizing complex protein digests both qualitatively and quantitatively. It will be shown how the strategy can be utilized in a targeted fashion to monitor the presence of a single or a number of proteins in a complex mixture as well as determining the stoichiometry of proteins in biological pathways. The strategy is based on maximizing the selectivity and specificity of the analytical workflow and on the use of signal replication. The method relies on technical and biological replication of DIA acquired precursor and product ion information of similar and dissimilar samples from a multitude of tissues and species, prepared, processed, and acquired in multiple laboratories. The quality of the fragment ion repository relies on the fact that no two datasets, either technical or biological, are likely to be fully identical in all analytical dimensions. The peptides and

associated product ions detected in these mixtures illustrate however reproducible behavior and their physicochemical properties can be confirmed. These concepts and their application will be disclosed and discussed.

Materials and Methods

Relational Database Repository

A development repository derived from data-independent CID spectra of 740,278 redundant, 100,434 non-redundant peptide ions was created from 207 DIA LC-MS data sets of tryptic digest of various *Rattus norvegicus* and *Mus musculus* tissue and body fluid samples. The digestion methods and experimental conditions were generic and described in more detail elsewhere [12, 42–45]. ProteinLynx Global SERVER ver. 2.4 was used as the database search algorithm for the preliminary data-independent identifications using either the reviewed entries of *Rattus norvegicus* (release 2010_11, 7,551 entries) or *Mus musculus* (release 2010_11, 16,320 entries) UniProtKB databases. Sequence information of internal standard proteins was added to the databases to normalize the data sets or to conduct quantification [46]. Guideline identification criteria were applied throughout [47]. In addition to the information provided by the search algorithm, including identification score and FDR [48], normalized fragment ion intensities f_1 and f_2 and a normalized peptide intensity p_1 are calculated. Their definition and an explanation are provided in Results section. In addition, the peptide retention times are standardized as described by Tarasova et al. [49], by obtaining linear fit parameters based on hydrophobicity [50] and standardizing to a reference. This approach affords initial population of the repository with orthogonal identification information from different instruments and laboratories. A variant, based on repository content based retention time normalization is applied in this study, which requires the upload of a sufficient number of identification results and in silico information to construct a reliable and robust reference retention normalization mechanism. Grouping normalized parameters f_1 , f_2 , and p_1 creates a so-called ion map of which the utility will be explained in detail in the section 3. Finally, a normalized protein molar P_1 amount is expressed. Parameters f_1 , f_2 , p_1 , and P_1 are related to unique repository identifiers. In addition, accurate precursor and product ion mass information is uploaded into the relational database, as well as, if applied, their associated ion mobility values. Direct comparison of the experimental drift time with a standard mobility database is likely to enhance peptide ion identification [51]. As for mass, drift time values are not normalized or standardized. The relational database and queries rely on the native accuracy and precision of the mass and mobility measurements.

Taxonomy (<ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>) and tissue (http://www.brenda-enzymes.info/ontology/tissue/tree/update/update_files/BrendaTissueOBO) identifiers are appended as ancillary, queryable information to the content of the repository during upload of the data. In addition, the protein accession numbers are mapped to the UniParc database [52].

Isoform/Homology Filtering

The protein concentrations are estimated as described [46]. Briefly, the average ion intensity of the three most abundant peptides identified to a protein is standardized to that of an internal standard spiked into the sample at known concentration. However, the observed signal intensity of sequence common peptides can be a summed value arising from redundant identifications. This is advantageous from a qualitative perspective since the intensity of the redundant peptides is cumulative. From a quantitative perspective, it hampers data analysis, especially if the contribution of the individual protein isoform cannot be addressed or accessed. Certain quantification schemes therefore disregard or down-weight these peptides to express a quantitative value, which could be problematic for highly homologous proteins since the number of proteotypic peptides could be small. An extension to the earlier presented absolute quantification schema is discussed.

The average intensity, in contrast to calculating the average intensity of the n best ionizing peptides, is calculated from the n most abundant proteotypic peptides. These averaged intensities

are subsequently used to segment the total observed intensity of the common peptide belonging to each parent protein. In instances where no proteotypic peptides can be identified, the identified proteins will be grouped and an absolute amount assigned to the group as a whole. Next, the peptides are re-ordered based on their segmented intensities for the sequence common and non-segmented intensities of the proteotypic peptides and the molar amounts calculated. The segmentation process is illustrated in Supplementary Figure 1. This leads to improved estimation of the amount and concentration of protein isoforms and homologues. In the instance of comparative analysis, the method also provides a better estimate of the relative amounts or fold changes for different homologous proteins between two or more conditions, since the information content obtained from non-proteotypic peptides is more detailed and comprehensive.

Fragment Ion Repository/Relational Database

The fragment ion repository is a multi-user application, with the information stored in a mySQL relational database server and the website run on Apache. Data entry and queries are

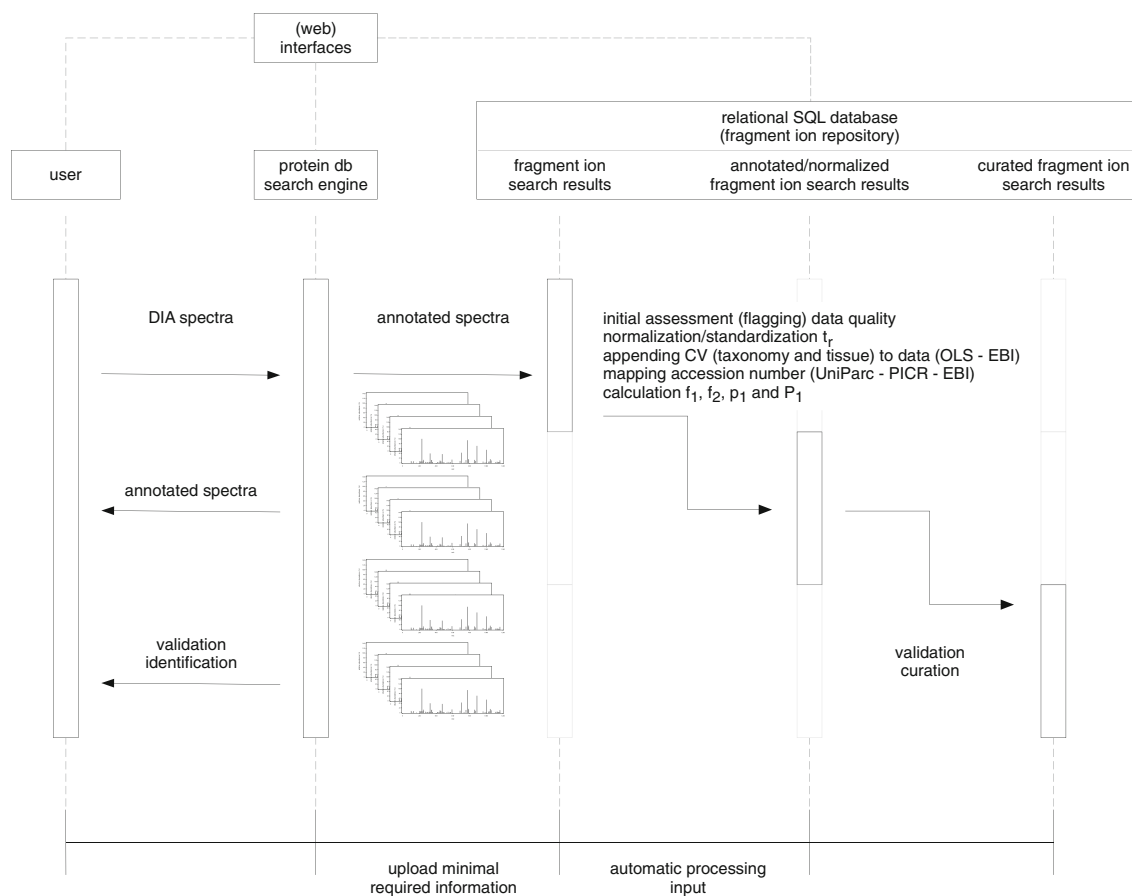


Figure 1. Workflow/activity diagram of the upload, annotation, processing, and creation of a multi-user relational database-based, data-independent fragment ion repository. Data are uploaded and species and tissue taxonomy meta data entered. The calculation of fragment, precursor, protein relationships, and standardized elution times is conducted automatically and appended to the database, including species-independent protein accession number mapping. A web front-end is used to query the database

achieved using a combination of PHP, Perl, and JavaScript. A single or multiple zip archives are uploaded to the server. Each zip file consists of the search results in comma-separated values format and the search criteria in text file format. The latter also holds instrumental performance and acquisitions settings and is used as a first pass assessment of the quality of the uploaded data. Equivalent and additional proteomics LC-MS performance metrics [53, 54] can be readily retrieved from the repository content for one-dimensional LC-MS experiments. An example is provided in Supplementary Figure 2, illustrating the interquartile retention time range, median retention time, interquartile

retention time ratio, average chromatographic peak width at half height and number of identified peptides/min within the interquartile retention time range prior to retention time standardization, excluding in-source fragment, losses, and variable modifications. Performance metrics have intrinsic value but are not discussed in detail.

Results

The upload, annotation, and processing workflow of data-independent analysis (DIA) LC-MS data into the fragment ion repository is shown in Figure 1. A more detailed

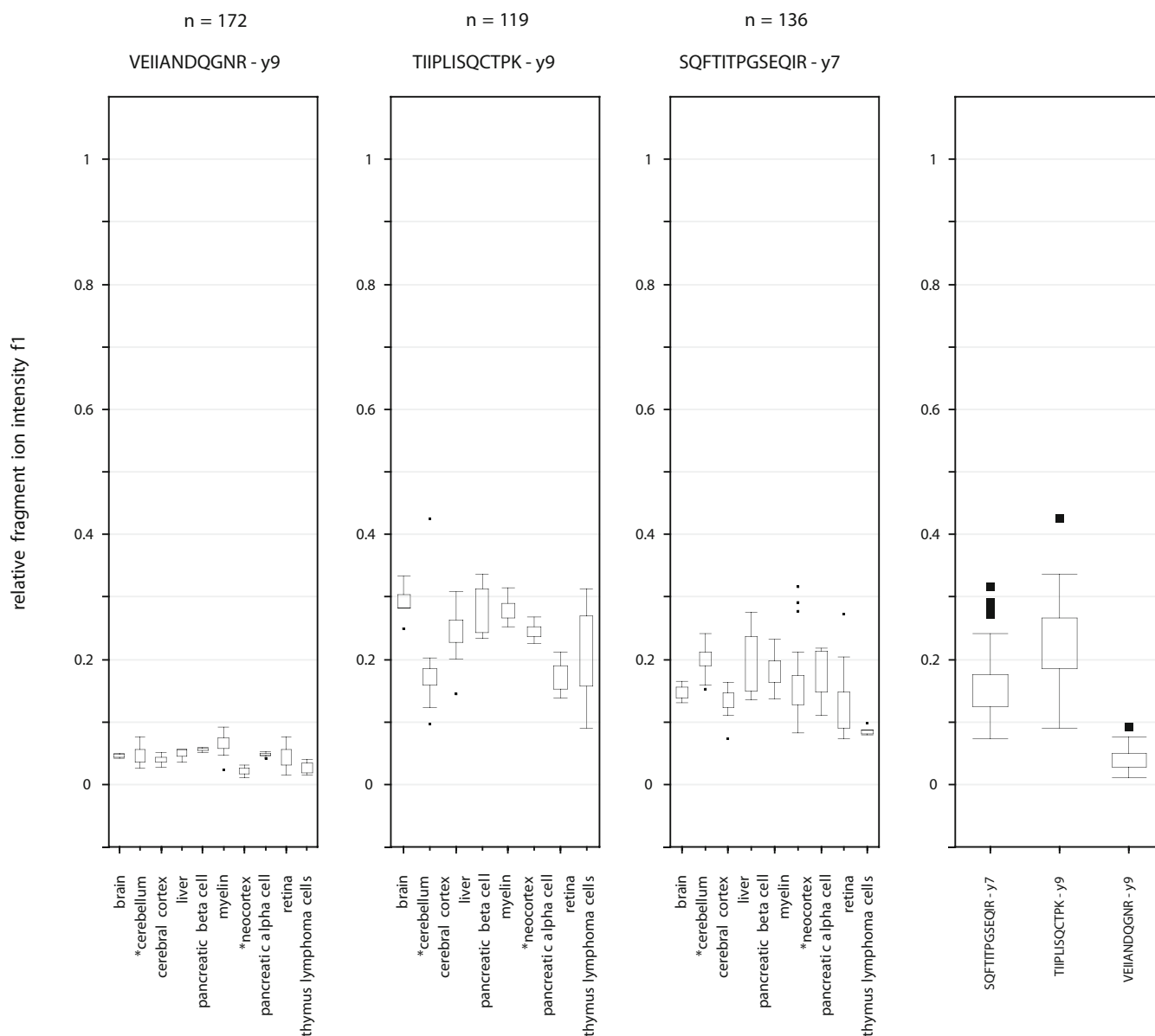


Figure 2. Intra- (tissue/experiment) and inter- (sequence) relative fragment ion intensity f_1 distribution examples for one- and two-* dimensional LC-DIA-MS experiments of different mouse and rat proteomes. The statistical distribution of ratio f_1 values for three fragment ions originating from three different peptides are shown as a box plot for each of the experiments deposited in the database. The overall distribution of the ratio f_1 values for each of the fragment ion is shown in the right-hand panel. Outlier data are marked as black squares

description of this process and the developed software is provided in the [Materials and Methods section](#). Briefly, the uploaded results are initially quality flagged on the basis of the automatically derived precursor/product ion search tolerances and data resolution as calculated by the utilized search engine. This information is currently stored in the database and provides the possibility to reject data in case the quality is suspicious from an MS perspective. Currently, no rejection criteria are applied during the database population process. LC quality metrics are under study and their significance discussed in the [Materials and Methods section](#). and Supplementary Figure 2, respectively. The protein accession numbers are mapped to a universal identifier using user provided species taxonomy as a filter. During this process, tissue information is appended to the processed data as well. Next, the data is processed (i.e., normalized on retention time, fragment ion intensity, precursor ion intensity, estimated molar amounts, fragment ion to multiple parent peptide, and protein relationships determined. These normalized parameters describe fragment and precursor ions relationships and are calculated as follows:

$$I_{\text{fragment ion } f, p} / I_{\text{peptide ion } p}(f1)$$

and

$$I_{\text{fragment ion } f, p} / \sum_i I_{\text{fragment ion } f(i), p}(f2)$$

where $I_{\text{fragment ion } f, p}$ =intensity fragment ion of peptide p and $I_{\text{peptide ion } p}$ the intensity of peptide p. Normalized ratio f1 describes the fragmentation efficiency for a given fragment ion in relation to the precursor intensity and normalized ratio f2 the preferred fragmentation pathway of a particular fragment ion for a given sequence. Alternatively, the intensities of y''_{max} and b_{max} could be subtracted from the precursor and summed fragment ion intensity to account for unfragmented peptide precursor. Relative, normalized peptide intensities are calculated as follows:

$$I_{\text{peptide } p, P} / \sum_i I_{\text{peptide } p(i), P}(P1)$$

where $I_{\text{peptide } p, P}$ =intensity peptide precursor ion of protein P. Finally, a normalized protein molar amount is expressed:

$$n_P / \sum_i n_{P(i)}(P1)$$

where n_P =estimated (molar) amount of protein P1 [47]. Relative and normalized molar amounts can be used for stoichiometry and pathway analyses purposes. Already foreseen is a final, automated validation and curation process, based on variation converging with either technical or biological experiment increment, before final upload of the normalized spectral information into the fragment ion repository/relational database. In order to demonstrate the importance of normalized DIA spectrum intensity values and their use to create a fragment ion repository, no attempts were made to remove outlier data at this stage.

The results of 207 DIA LC-MS experiments were uploaded to the relational database, representing the experimental results from six different laboratories and 10 tissue types. Currently, the database holds in total the identification results from 69,907 proteins and 1,032,110 peptides. Note that these results do not represent unique protein identifications, but unique protein-sample and peptide-sample combinations. Peptide identifications include in-source fragments and the losses of water and ammonia from precursor ions. These peptides can be readily mapped to multiple proteins of which the relationships can be retrieved by means of standard SQL language queries. The total number of redundant fragment ion identifications equals 7,480,798. This number of identifications is reduced to 2,181,901 by excluding the y and b ion losses of water and ammonia, y_{max} , b_{max} , in source fragments and variable modifications. More than 85.6% of these fragment ion identifications replicated at least twice across the database content. With decoy entries excluded this amounts to 86.5%. This highlights the quality of the current content of the fragment ion identification repository, which is a primary prerequisite, as will be illustrated later, for successful and confident identifications. In contrast, only 8.3% of the decoy entries were identified in at least two LC-MS experiments. This number readily decreases to 0.7% when a replication rate of at least five is applied. With the same criteria applied, the replication rate of the non-decoy fragment ion identifications equals 72.5%. This is expected, as decoy signals are not amplified during the process of identification replication. Interestingly, two fragment ions, associated with two particular data sets, account for the majority of the replicating decoy identifications. Excluding these two ions, reduced the number of replicating decoy fragment ion identifications further down to 0.4%. The same logic was applied throughout the analysis of the database content unless mentioned otherwise. These basic database entry statistics suggest however that accurate mass product ion signal replication is a very strong metric and that the value of product ion mass accuracy with respect to improving specificity is most likely underestimated by most sequence search algorithms.

The fragment ion relationships f1 and f2 describe the fragmentation efficiency for a given fragment ion in relation to the precursor intensity and the preferred fragmentation pathway of a particular fragment ion for a given sequence. Examples for f1 and f2 for three highly replicating fragment ions are shown in Figures 2 and 3, respectively. Outlier data are marked by a black square and are indicative of incorrect identifications or interfered ion intensity measurements. The value and specificity of f1 and f2 increases substantially by supplementing the fragment ion signature with additional orthogonal information, such as standardized peptide retention time [49], drift time [51, 55] or collision cross section [56]. Supplementary Figure 3 illustrates the addition of the standardized retention time to f1 and f2, thereby creating a subset ion map for the three fragment ions of interest, indicating substantially improved specificity compared with the results shown in Figs. 2 and 3. Panel (a) of Supplementary Figure 3 illustrates that experimental retention time

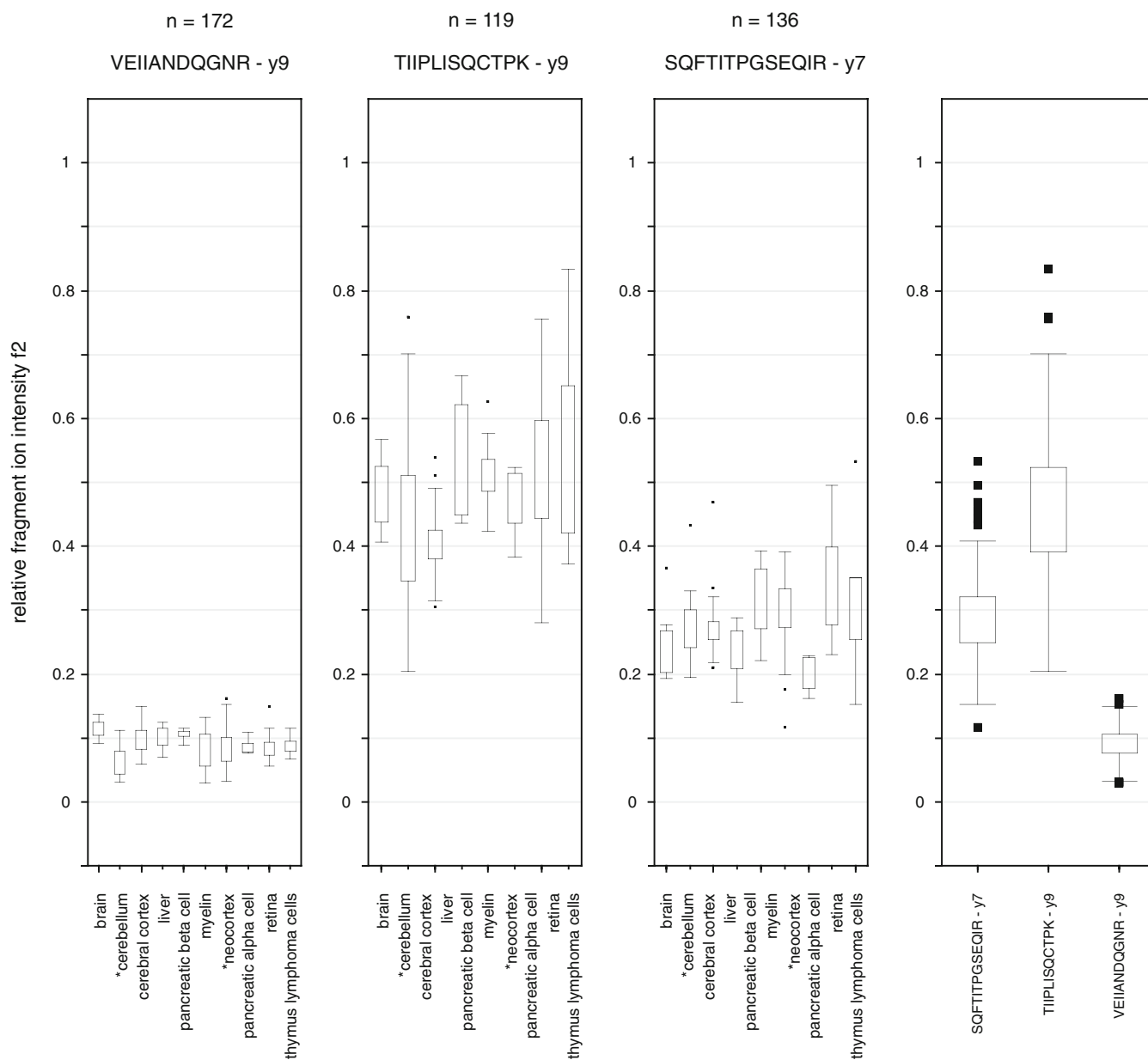


Figure 3. Intra- (tissue/experiment) and inter- (sequence) relative fragment ion intensity f_2 distribution examples for one and two* dimensional LC-DIA-MS experiments of different mouse and rat proteomes. The statistical distribution of ratio f_2 values for three fragment ions originating from three different peptides are shown as a box plot for each of the experiments deposited in the database. The overall distribution of the ratio f_2 values for each of the fragment ion is shown in the right-hand panel. Outlier data are marked as black squares

correlate linearly with predicted normalized hydrophobicity. Panel (b) shows f_1 and f_2 as a function of non-standardized (raw) retention time, and panel (c) f_1 and f_2 as a function of retention time after normalization and standardization [49]. Drift time and collision cross section information were not acquired and/or available for the results described. Drift time is expected to further increase specificity [51], whereas cross section information is believed to be useful for the analysis of post-translationally modified peptides [56]. Median values for f_1 and f_2 were 31.8% and 28.5%, respectively. Fragment ion relationships f_1 and f_2 can be utilized to create

queryable library-like database spectra. An example is shown in Fig. 4. The circles represent curated average database f_2 values with the database frequency in parentheses. A normalized experimental spectrum is superimposed. The relative intensity of the majority of the observed fragment ions is in agreement with the database entries, contradicting a recent study [57]. Moreover, the spectrum indicates remarkable similarity with previously presented data from other species and sample types [46]. For this particular example, the absolute normalized fragment ion intensity standard deviation was as high as 50% to 70% for

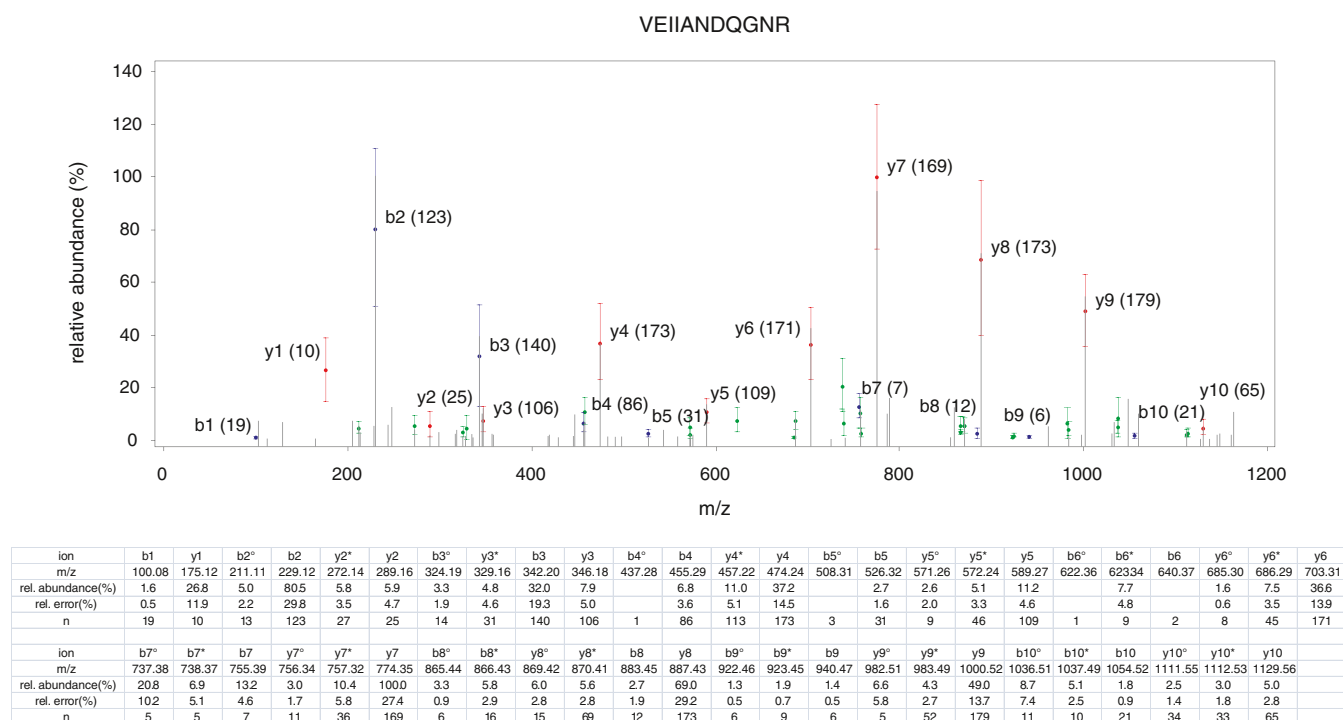


Figure 4. Experimental DIA fragment ion spectrum VEIANDQGNR overlaid with the normalized f2 fragment ion intensities of Hspa5 (78 kDa glucose-regulated protein) from mouse and rat with replication rate in parenthesis. Color legend: grey=experimental spectrum, red=y ion, blue=b ion and green=neutral loss; marker legend: *=neutral loss of ammonia, and °=neutral loss of water

the lower abundant product ions with lower ion statistics, whereas for the more abundant fragment ions, this value is closer to 35%.

An additional benefit of spectral library-like searches is demonstrated in Supplementary Figure 4 for various cancer cell line samples of mammalian species similar to the described organisms under study. The left hand side of Supplementary Figure 4 illustrates the absolute increase in number of identified proteins, corresponding to relative increases of 18%, 28%, and 4% for PC-3, MDA-MB-231, and Hep-G2, respectively. The right hand side shows the identification distributions of the sample for both approaches, i.e., database and repository centric. For the latter, an ion match tolerance of ± 10 ppm was used. In addition, the retention and drift times, as discussed in the Materials and Methods section, were normalized and standardized as the experiments were conducted on different instruments in different laboratories. The match tolerances were ± 1 min and ± 1 drift time bin, respectively. As can be noticed from the presented results, the sample common number of identified proteins substantially increased, whereas the number of sample unique proteins decreased, by using a spectrum library-like search approach. More importantly, this protein identification increment was primarily achieved through ion detections from the lower concentration ranges of the dynamic range of the studied cancer cell line proteomes.

The median calculated value for p1 equaled 44.2% ($n=150$), ranging from 23.7% ($n=43$) to 45.6% ($n=50$) for individual experiments. The results from the two-dimensional LC-MS experiments were excluded from p1 value trend analysis since different, first dimension-dependent, non-linear second dimension gradients were applied in order to optimally utilize the available chromatographic space/increase system peak capacity. Precursor intensities arising from multiple fractions are typically summed in the instance of two-dimensional data-independent LC-MS experiment and error measurement can, therefore, be slightly higher than expected. This could be overcome by extracting the precursor and product ion intensity from only one second dimension gradient separation, preferably the more dominating contributing one in terms of identification confidence and ion statistics, in order to calculate fragment ion relationships f1 and f2. This was, however, not considered at this moment of time. As previously mentioned, no attempts were made to remove outlier data, which could be part of the earlier proposed automatic digital curation process with the upload of new results. Alternatively, machine learning algorithms could be implemented and employed [58]. Examples for p1 for three highly replicating peptides are shown in Supplementary Figure 5. In this instance, the two-dimensional LC-MS data are included to illustrate that they exhibit somewhat more scatter. Peptides are expected to have similar ionization efficiencies, regardless of the sample and protein origin [12, 46]. This could go readily unnoticed in a DDA experiment, especially in the case of



Figure 5. Relative fragment ion intensity p1 isoform distribution and commonality. Isoform-independent tryptic fragment number annotation is provided in Supplementary Figure 6. Marker legend: number = replication rate, size=average relative precursor ion intensity p1, angle arrow=variance relative precursor ion intensity p1, and color=tryptic fragment number

chimeric or composite spectrum instances. Some of the problems associated with chimeric events were emphasized already in the [Introduction section](#). Undoubtedly, amino acid sequence-ionization efficiency relationships cannot be established when fragment ion spectra are incorrectly annotated. Moreover, DDA experiments are duty cycle limited, thereby limiting the opportunity to detect and identify all peptides of interest [10]. This would be especially the case for high in-spectrum dynamic range occurrences in combination with automatic gain control, as applied with trap based mass analyzers, whereby low abundant peptides are unnoticed [59]. DIA acquisition methods are, therefore, arguably more suited to

quantify ionization efficiencies by means of electrospray LC-MS as they are only detection limited. Superimposed chimeric DIA fragmentation spectra are searched with dedicated search engines since they, by default, arise from co-eluting, non-isolated peptides [46]. Species- and tissue-independent peptide ionization efficiency consistency is illustrated in Fig. 5, where the average p1 value, the related coefficient of variation, and replication rate are summarized for Aldolase A, B, and C for both rat and mouse using the amino acid sequence of Aldolase A from mouse as the alignment reference. The tryptic fragment number annotation of the latter is shown in Supplementary Figure 6. Despite sequence differences between the protein

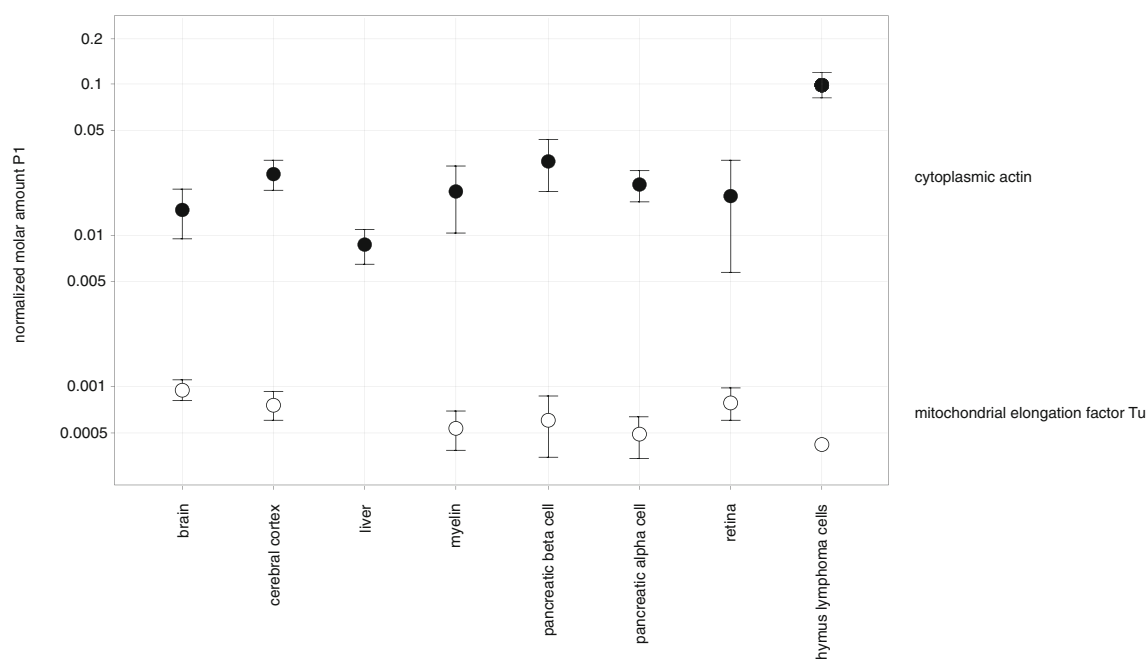


Figure 6. Relative distribution estimated molar within-sample amounts P1 examples for one dimensional LC-DIA-MS experiments of different mouse and rat proteomes

isoforms and species, similar precursor intensities can be observed. Normalized peptide p1 intensity values can be used for example for pathway analysis since peptides identified to associated proteins are also expected to show relative ionization distribution efficiency similarity [60].

Proteotypic information and estimated relative within-sample molar amounts P1 were retrieved from the repository for 62,760 proteins, representing 2971 non-redundant, replicating identifications from non-fractionated mammalian samples. The actual total number of protein uploads equaled 69,907. In other words, proteotypic, quantifiable information was obtained for 89.8% of the identified proteins. Moreover, the repository currently holds 2266 replicating, species-independent genes, whereas to date 16,106 non-redundant and reviewed (curated) primary gene names can be retrieved from the utilized protein sequence databases, equaling a 14.1% depth of genome coverage. This very high volume of genomic and proteotypic content stems from the use of high-quality, curated databases and, equally important, a data-independent scanning approach and search algorithm that can extract the required type of information from complex samples. As an example, Fig. 6 illustrates normalized molar amounts of mitochondrial elongation factor Tu and cytoplasmic actin identified in both rat and mouse in various tissues and indicates that the protein abundance level of both proteins is similar across the investigated samples. The relative abundance between the two proteins of interest is approximately 30-fold and is relatively consistent across all tissue types and species. Mitochondrial elongation factor Tu was consistently low in abundance and not quantified in one of the samples. Cytoplasmic actin is more abundant in one of the investigated tissue samples, which may be biologi-

cally relevant for the sample and/or perturbation under study or sample preparation procedure related. The relational database captures, however, meta data during upload, which could be used to investigate discrepancies in more detail or filter the results. Relative molar amount comparison between tissues would benefit from geometric normalization as would be more typically applied in microarray analysis [61]. It has been recently demonstrated that this normalization technique can also be applied to the data obtained from label-free, data-independent experiments [62]. This information can be utilized for both inter- and intra-sample stoichiometry analysis [62–64], as shown in Supplementary Figure 7, where the within-tissue consistency for a well-described multi-enzyme complex is shown. The E1 α (PDHA1), E1 β (PDHB), and E2 (DLAT) subunits of the pyruvate dehydrogenase complex were normalized to E3 (DLD), which is associated with other protein complexes. A good within-tissue agreement was observed vs. the expected 1:1:1 subunit ratio for the majority of the samples in which the proteins of interest were identified and quantified. The E1 α subunit did not follow the commonly observed trend for the cerebral cortex samples and was identified at an approximate 1:1.5 ratio vs. the two other subunits.

Concept and Outlook

Recent instrument developments and LC-MS based proteomics techniques have considerably improved the speed of analysis, depth of protein coverage, and information content that can be obtained from complex biological sample mixtures. Despite these impressive developments, identification and quantification variation is still a concern and,

thus, alternative and complementary methods are even required to date. The value and use of a data-independent fragment ion repository has therefore been explored. The required sensitivity and selectivity for the purpose of protein identification and quantification has been demonstrated in previous paragraphs. More conceptually, the following schema can be considered for validation and or conformation, hypothesis driven studies or selected reaction monitoring (SRM)/multiple reaction monitoring (MRM) method development [65, 66].

1. A minimum of three peptides with three product ions are selected from any given protein. The fragment selection is based upon highest replication rate and smallest signature product ion variation across both experiments and samples.
2. In addition, similar precursor and product ions, three times three for the complete 'protein set', are selected from second and third protein that are consistently present with the protein of interest. Together, these proteins outline a fragment ion signature (i.e., ion map).
3. Unknown samples can be subsequently mapped against the fragmentation database signature to validate the presence of the target protein, with the additional ions and their associated intensity ratios acting as an internal validation mechanism.

Various statistical and computational tools and methods are currently considered and implemented for the analysis of the content of the fragment ion repository in order to facilitate the above and more mathematical accounting of the information that resides in the relational database [67]. These developments, query tools, and the public section of the repository will become open source and can be followed at: <http://sites.duke.edu/ionmap/>. The fragment ion database can currently only be populated with qualitative results obtained through DIA experiments, also known as LC-MS^E [68]. DDA experiments generally do not afford precursor and product ion intensity measurements across the complete chromatographic peak or MS and MS/MS intensity recording for the same amount of time with the same gain applied. Hence, calculating normalized fragment and precursor ion intensities such as f1 and p1 could be more challenging. However, it has been demonstrated previously that DDA and DIA product ion spectra share great similarity [9], arguing that normalized f2 values and aggregate MS/MS spectra originating and derived from DDA spectra could be used to complement the content of the repository, which may hold great value in the instance of the more targeted analysis of fractionated or enriched samples. The presented concepts can be easily transferred to other application areas, including lipodomics [69] or metabolomics [70], facilitating the characterization and quantification of other molecule types. As spectral libraries and fragment ion repositories find more widespread use in proteomics, some of the remaining objection will be solved. In addition to the identification of fragment ions, peptides, and proteins, data-independent fragment ion repositories have great potential with regard to the quantification of protein abundances,

stoichiometry, and the reliable quantification of post-translational modifications. In conclusion, repositories are a valuable addition to the requirements of systems biology, not only allowing quantitative analysis of low-abundant proteins, but also delivering reliably quantitative data when proteins are analyzed across multiple samples in multiple laboratories.

Acknowledgments

Timothy Riley is kindly acknowledged for his valuable contribution throughout the development of this work. S.T. is supported by Deutsche Forschungsgemeinschaft (SFB490, Z3) and Forschungszentrum Immunologie Mainz.

References

1. Ptolemy, A.S., Rifai, N.: What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. *Scand. J. Clin. Lab. Invest. Suppl* **242**, 6–14 (2010)
2. Ransohoff, D.F.: How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J. Clin. Epidemiol.* **60**(12), 1205–1219 (2007)
3. White, F.M.: The potential cost of high-throughput proteomics. *Sci. Signal.* **4**(160), pe8 (2011)
4. Kaur, P., O'Connor, P.B.: Algorithms for automatic interpretation of high resolution mass spectra. *J. Am. Soc. Mass Spectrom.* **17**(3), 459–468 (2006)
5. Hoopmann, M.R., Finney, G.L., MacCoss, M.J.: High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* **79**(15), 5620–32 (2007)
6. Sun, Y., Zhang, J., Braga-Neto, U., Dougherty, E.R.: BPDA - a Bayesian peptide detection algorithm for mass spectrometry. *BMC Bioinformatics* **11**, 490 (2010)
7. Hoopmann, M.R., Merrihew, G.E., von Haller, P.D., MacCoss, M.J.: Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. *J. Proteome Res.* **8**(4), 1870–1875 (2009)
8. Montoya, A., Beltran, L., Casado, P., Rodriguez-Prados, J.-C., Cutillas, P.R.: Characterization of a TiO₂ enrichment method for label-free quantitative phosphoproteomics. *Methods* **54**(4), 370–378 (2011)
9. Geromanos, S.J., Vissers, J.P., Silva, J.C., Dorschel, C.A., Li, G.Z., Gorenstein, M.V., Bateman, R.H., Langridge, J.I.: The detection, correlation, and comparison of peptide precursor and product ions from data-independent LC-MS with data dependant LC-MS/MS. *Proteomics* **9**(6), 1683–1695 (2009)
10. Michalski, A., Cox, J., Mann, M.: More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10**(4), 1785–1793 (2011)
11. Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N., Old, W.M.: Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies. *J. Proteome Res* **9**(8), 4152–4160 (2010)
12. Vissers, J.P., Pons, S., Hulin, A., Tissier, R., Berdeaux, A., Connolly, J.B., Langridge, J.I., Geromanos, S.J., Ghaleh, B.: The use of proteome similarity for the qualitative and quantitative profiling of reperfused myocardium. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **877**(13), 1317–1326 (2009)
13. Frahm, J.L., Howard, B.E., Heber, S., Muddiman, D.C.: Accessible proteomics space and its implications for peak capacity for zero-, one-, and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. *J. Mass Spectrom.* **41**(3), 281–288 (2006)
14. Rodríguez-Suárez, E., Hughes, C., Gethings, L., Giles, K., Wildgoose, J., Stapels, M., Fadgen, K.E., Geromanos, S.J., Vissers, J.P., Elortza, F., Langridge, J.I.: An ion mobility assisted data-independent LC-MS strategy for the analysis of complex biological samples. *Current. Anal. Chem.* Special Issue: Ion Mobility Spectrometry: Using Size and Shape

- to Understand Real-World Systems at the Molecular Level, HT-SBJ-CAC-0005
15. Geromanos, S.J., Hughes, C., Golick, D., Ciavarini, S., Gorenstein, M.V., Richardson, K., Hoyes, J.B., Vissers, J.P., Langridge, J.I.: Simulating and validating proteomics data and search results. *Proteomics* **11**(6), 1189–1211 (2011)
 16. Blackburn, K., Mbeunkui, F., Mitra, S.K., Mentzel, T., Goshe, M.B.: Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. *J. Proteome Res.* **9**(7), 3621–3637 (2010)
 17. Wang, J., Pérez-Santiago, J., Katz, J.E., Mallick, P., Bandeira, N.: Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteom.* **9**(7), 1476–1485 (2010)
 18. Bern, M., Finney, G., Hoopmann, M.R., Merrihew, G., Toth, M.J., MacCoss, M.J.: Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal. Chem.* **82**(3), 833–841 (2010)
 19. Bell, A.W., Deutsch, E.W., Au, C.E., Kearney, R.E., Beavis, R., Sechi, S., Nilsson, T., Bergeron, J.J.: A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6**(6), 423–430 (2009)
 20. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**(3), 207–214 (2007)
 21. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J.: Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**(11), 923–925 (2007)
 22. Reiter, L., Claassen, M., Schimpf, S.P., Jovanovic, M., Schmidt, A., Buhmann, J.M., Hengartner, M.O., Aebersold, R.: Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteom.* **8**(11), 2405–2417 (2009)
 23. Cargile, B.J., Bundy, J.L., Stephenson Jr., J.L.: Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **3**(5), 1082–1085 (2004)
 24. Zhu, P., Bowden, P., Tucholska, M., Zhang, D., Marshall, J.G.: Peptide-to-protein distribution versus a competition for significance to estimate error rate in blood protein identification. *Anal. Biochem.* **411**(2), 241–253 (2011)
 25. Bowden, P., Beavis, R., Marshall, J.: Tandem mass spectrometry of human tryptic blood peptides calculated by a statistical algorithm and captured by a relational database with exploration by a general statistical analysis system. *J. Proteom.* **73**(1), 103–111 (2009)
 26. Craig, R., Cortens, J.P., Beavis, R.C.: The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **19**(13), 1844–1850 (2005)
 27. Domokos, L., Hennberg, D., Weimann, B.: Computer-aided identification of compounds by comparison of mass spectra. *Anal. Chim. Acta* **165**, 61–74 (1985)
 28. Stein, S.E., Scott, D.R.: Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**(9), 859–866 (1994)
 29. Ausloos, P., Clifton, C.L., Lias, S.G., Mikaya, A.I., Stein, S.E., Tchekhovskoi, D.V., Sparkman, O.D., Zaikin, V., Zhu, D.: The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* **10**(4), 287–299 (1999)
 30. Craig, R., Cortens, J.C., Fenyo, D., Beavis, R.C.: Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **5**(8), 1843–1849 (2006)
 31. Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S., MacCoss, M.J.: Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**(16), 5678–5684 (2006)
 32. Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., Aebersold, R.: Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**(5), 655–667 (2007)
 33. Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., Stein, S.E., Aebersold, R.: Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **5**(10), 873–875 (2008)
 34. Yen, C.Y., Houel, S., Ahn, N.G., Old, W.M.: Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol. Cell. Proteom.* **10**(7), M111.007666 (2011)
 35. Frank, A.M., Monroe, M.E., Shah, A.R., Carver, J.J., Bandeira, N., Moore, R.J., Anderson, G.A., Smith, R.D., Pevzner, P.A.: Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8**(7), 587–591 (2011)
 36. Frank, A.M.: Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* **8**(5), 2226–2240 (2009)
 37. Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S.P., Smith, R.D., Pevzner, P.A.: Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**(1), 113–122 (2008)
 38. Vizcaino, J.A., Côté, R., Reisinger, F., Foster, J.M., Mueller, M., Rameseder, J., Hermjakob, H., Martens, L.: A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **9**(18), 4276–4283 (2009)
 39. Smith, B.E., Hill, J.A., Gjukich, M.A., Andrews, P.C.: Tranche distributed repository and ProteomeCommons.org. *Methods Mol. Biol.* **696**, 123–145 (2011)
 40. Deutsch, E.W., Lam, H., Aebersold, R.: PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9**(5), 429–434 (2008)
 41. Ji, L., Barrett, T., Ayanbule, O., Troup, D.B., Rudnev, D., Muertter, R.N., Tomashevsky, M., Soboleva, A., Slotta, D.J.: NCBI Peptidome: a new repository for mass spectrometry proteomics data. *Nucleic Acids Res.* **38**, D731–D735 (2010)
 42. Ma, D., Chan, M.K., Lockstone, H.E., Pietsch, S.R., Jones, D.N., Cilia, J., Hill, M.D., Robbins, M.J., Benzel, I.M., Umrana, Y., Guest, P.C., Levin, Y., Maycox, P.R., Bahn, S.: Antipsychotic treatment alters protein expression associated with presynaptic function and nervous system development in rat frontal cortex. *J. Proteome Res.* **8**(7), 3284–3297 (2009)
 43. Krämer-Albers, E.M., Bretz, N., Tenzer, S., Winterstein, C., Möbius, W., Berger, H., Nave, K.A., Schild, H., Trotter, J.: Oligodendrocytes secrete exosomes containing major myelin and stress-protective proteins: Trophic support for axons? *Proteomics Clin. Appl.* **1**(11), 1446–1461 (2007)
 44. Reidel, B., Thompson, J.W., Farsiu, S., Moseley, M.A., Skiba, N.P., Arshavsky, V.Y.: Proteomic profiling of a layered tissue reveals unique glycolytic specializations of photoreceptor cells. *Mol. Cell. Proteom.* **10**(3), M1110.002469 (2011)
 45. Fuller, B.F., Lewis, M.H., Ottens, A.K.: Systems biology reveals altered metabolism and mitochondrial proliferation in developing brain consequent to environmental tobacco smoke exposure. manuscript submitted (2012)
 46. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P., Geromanos, S.J.: Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteom.* **5**(1), 144–156 (2006)
 47. Bradshaw, R.A., Burlingame, A.L., Carr, S., Aebersold, R.: Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteom.* **5**(5), 787–788 (2006)
 48. Li, G.Z., Vissers, J.P., Silva, J.C., Golick, D., Gorenstein, M.V., Geromanos, S.J.: Database searching and accounting of multiplexed precursor and product ion spectra from the data-independent analysis of simple and complex peptide mixtures. *Proteomics* **9**(6), 1696–1719 (2009)
 49. Tarasova, I.A., Guryca, V., Pridatchenko, M.L., Gorshkov, A.V., Kieffer-Jaquinod, S., Evreinov, V.V., Masselon, C.D., Gorshkov, M.V.: Standardization of retention time data for AMT tag proteomics database generation. *J. Chromatogr. B Analyt. Technol. Biomed Life Sci.* **877**(4), 433–440 (2009)
 50. Krokhin, O.V., Craig, R., Spicer, V., Ens, W., Standing, K.G., Beavis, R.C., Wilkins, J.A.: An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell. Proteom.* **3**(9), 908–919 (2004)
 51. Valentine, S.J., Ewing, M.A., Dilger, J.M., Glover, M.S., Geromanos, S., Hughes, C., Clemmer, D.E.: Using ion mobility data to improve peptide identification: intrinsic amino acid size parameters. *J. Proteome Res.* **10**(5), 2318–2329 (2011)
 52. Côté, R.G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., Hermjakob, H.: The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* **8**, 401 (2007)
 53. Rudnick, P.A., Clauser, K.R., Kilpatrick, L.E., Tchekhovskoi, D.V., Neta, P., Blonder, N., Billheimer, D.D., Blackman, R.K., Bunk, D.M., Cardasis, H.L., Ham, A.J., Jaffe, J.D., Kinsinger, C.R., Mesri, M., Neubert, T.A., Schilling, B., Tabb, D.L., Tegeler, T.J., Vega-Montoto, L., Variyath, A.M., Wang, M., Wang, P., Whiteaker, J.R., Zimmerman, L.J., Carr, S.A., Fisher, S.J., Gibson, B.W., Paulovich, A.G., Regnier,

- F.E., Rodriguez, H., Spiegelman, C., Tempst, P., Liebler, D.C., Stein, S.E.: Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteom* **9**(2), 225–241 (2010)
54. Paulovich, A.G., Billheimer, D., Ham, A.J., Vega-Montoto, L., Rudnick, P.A., Tabb, D.L., Wang, P., Blackman, R.K., Bunk, D.M., Cardasis, H.L., Clauser, K.R., Kinsinger, C.R., Schilling, B., Tegeler, T.J., Varyath, A.M., Wang, M., Whiteaker, J.R., Zimmerman, L.J., Fenyo, D., Carr, S.A., Fisher, S.J., Gibson, B.W., Mesri, M., Neubert, T.A., Regnier, F.E., Rodriguez, H., Spiegelman, C., Stein, S.E., Tempst, P., Liebler, D.C.: Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteom.* **9**(2), 242–254 (2010)
55. Giles, K., Williams, J.P., Campuzano, I.: Enhancements in traveling wave ion mobility resolution. *Rapid Commun. Mass Spectrom.* **25**(11), 1559–1566 (2011)
56. Thalassinos, K., Grabenauer, M., Slade, S.E., Hilton, G.R., Bowers, M.T., Scrivens, J.H.: Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry. *Anal. Chem.* **81**(1), 248–254 (2009)
57. Barsnes, H., Eidhammer, I., Martens, L.: A global analysis of peptide fragmentation variability. *Proteomics* **11**(6), 1181–1188 (2011)
58. Li, Y.F., Arnold, R.J., Tang, H., Radivojac, P.: The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.* **9**(12), 6288–6297 (2010)
59. Baker, E.S., Livesay, E.A., Orton, D.J., Moore, R.J., Danielson 3rd, W.F., Prior, D.C., Ibrahim, Y.M., LaMarche, B.L., Mayampurath, A.M., Schepmoes, A.A., Hopkins, D.F., Tang, K., Smith, R.D., Belov, M.E.: An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies. *J. Proteome Res.* **9**(2), 997–1006 (2010)
60. Vissers, J.P., Langridge, J.I., Aerts, J.M.: Analysis and quantification of diagnostic serum markers and protein signatures for Gaucher disease. *Mol. Cell. Proteom.* **6**(5), 755–766 (2007)
61. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., Speleman, F.: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, RESEARCH0034 (2002)
62. Martens, G.A., Jiang, L., Verhaeghen, K., Connolly, J.B., Geromanos, S.G., Stangé, G., Van Oudenhove, L., Devreese, B., Hellemans, K.H., Ling, Z., Van Schravendijk, C., Pipeleers, D.G., Vissers, J.P., Gorus, F.K.: Protein markers for insulin-producing beta cells with higher glucose sensitivity. *PLoS One* **5**(12), e14214 (2010)
63. Röwer, C., Vissers, J.P., Koy, C., Kipping, M., Hecker, M., Reimer, T., Gerber, B., Thiesen, H.J., Glocker, M.O.: Towards a proteome signature for invasive ductal breast carcinoma derived from label-free nanoscale LC-MS protein expression profiling of tumors and glandular tissue. *Anal. BioAnal. Chem.* **395**(8), 2443–2456 (2009)
64. Muñoz, I.G., Yébenes, H., Zhou, M., Mesa, P., Serna, M., Park, A.Y., Bragado-Nilsson, E., Beloso, A., de Cárcer, G., Malumbres, M., Robinson, C.V., Valpuesta, J.M., Montoya, G.: Crystal structure of the open conformation of the mammalian chaperonin CCT in complex with tubulin. *Nat. Struct. Mol. Biol.* **18**(1), 14–19 (2011)
65. Arnott, D., Kishiyama, A., Luis, E.A., Ludlum, S.G., Marsters Jr., J.C., Stults, J.T.: Selective detection of membrane proteins without antibodies: a mass spectrometric version of the Western blot. *Mol. Cell. Proteomics* **1**(2), 148–156 (2002)
66. Anderson, L., Hunter, C.L.: Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteom.* **5**(4), 573–588 (2006)
67. Lucas, J.E., Thompson, J.W., Dubois, L.G., McCarthy, J., Tillmann, H., Thompson, A., Shire, N., Hendrickson, R., Dieguez, F., Goldman, P., Schwarz, K., Patel, K., McHutchison, J., Moseley, M.A.: Metaprotein expression modeling for label-free quantitative proteomics. *BMC Bioinformatics.* **13**(1):74 (2012)
68. Silva, J.C., Denny, R., Dorschel, C.A., Gorenstein, M., Kass, I.J., Li, G.Z., McKenna, T., Nold, M.J., Richardson, K., Young, P., Geromanos, S.: Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **77**(7), 2187–2200 (2005)
69. Castro-Perez, J.M., Kamphorst, J., Degroot, J., Lafeber, F., Goshawk, J., Yu, K., Shockcor, J.P., Vreeken, R.J., Hankemeier, T.: Comprehensive LC-MS(E) lipidomic analysis using a shotgun approach and its application to biomarker detection and identification in osteoarthritis patients. *J. Proteome Res.* **10**(7), 3303–3308 (2011)
70. Crockford, D.J., Maher, A.D., Ahmadi, K.R., Barrett, A., Plumb, R.S., Wilson, I.D., Nicholson, J.K.: ¹H NMR and UPLC-MS(E) statistical heterospectroscopy: characterization of drug metabolites (xenometabolome) in epidemiological studies. *Anal. Chem.* **80**(18), 6835–6844 (2008)