Review

Introduction to Opportunities and Pitfalls in Functional Mass Spectrometry Based Proteomics

Marc Vaudel[1,2 $], Albert Sickmann[1,3], Lennart Martens[4,5]

[1] Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund, Germany

[2] Proteomics Unit (PROBE), Department of Biomedicine, University of Bergen, Bergen, Norway

[3] Medizinisches Proteom-Center (MPC), Ruhr-Universität, Bochum, Germany

[4] Department of Medical Protein Research, VIB, Ghent, Belgium

[5] Department of Biochemistry, Ghent University, Ghent, Belgium

[$] Correspondence: Marc Vaudel

PROBE

Building for Basic Biology

University of Bergen

Jonas Liesvei 91

N-5009 Bergen

Norway

phone: +47 55 58 63 68

fax: +47 55 58 63 60

E-mail: marc.vaudel@biomed.uib.no

Abstract

With the advent of mass spectrometry based proteomics, the identification of thousands of proteins has become commonplace in biology nowadays. Increasingly, efforts have also been invested toward the detection and localization of posttranslational modifications. It is furthermore common practice to quantify the identified entities, a task supported by a panel of different methods. Finally, the results can also be enriched with functional knowledge gained on the proteins, detecting for instance differentially expressed gene ontology terms or biological pathways.

In this study, we review the resources, methods and tools available for the researcher to achieve such a quantitative functional analysis. These include statistics for the post-processing of identification and quantification results, online resources and public repositories. With a focus on free but user-friendly software, preferably also open-source, we provide a list of tools designed to help the researcher manage the vast amount of data generated. We also indicate where such applications currently remain lacking. Moreover, we stress the eventual pitfalls of every step of such studies.

Keywords

Proteomics, data interpretation, online resources

Abbreviations

PTM: posttranslational modification; FDR: false discovery rate; IEF: isoelectric focusing; OMSSA: Open Mass Spectrometry Search Algorithm; PEP: Posterior Error Probability; FNR:

False Negative Rate; PRIDE: PRoteomics IDEntifications database; PICR: Protein Identifier Cross-Referencing; GO: Gene Ontology

1. Introduction

The progress in the application of mass spectrometry to biological compounds has revolutionized the field of biology: the large scale identification of proteins provides a unique snapshot of a biological system of interest at a given time point.[1, 2] The study of proteins and their modifications in a single sample, or differentially between samples dramatically increased our understanding of living cells and allowed the setup of ambitious experiments[3-5] opening new opportunities for biomedical research.[6] The canonical example of the latter is the comparison of the proteomes of a disease affected population against those of a control population.[7] Such studies aim to identify biomarkers – an easily detectable indicator of a biological state – for the targeted disease.[8] However, the efficiency of statistical comparison between metrics associated to a biological entity is questioned in the literature.[9-12] Indeed, such studies suffer from the high variance inherently found in biological systems,[9] from the low number of replicates typically analyzed,[10] and from experimental artifacts, errors and missing values.[13-15] As a result, the fine nuances of the proteomic variations are often not statistically significant when compared to the global variance of the system.

In order to tackle these issues, the proteomics community has started an ambitious systematic sharing of resources.[16] The rationale is the following: when bringing knowledge from previous experiments and other fields like genomics and transcriptomics together, one will have a better understanding of the results and might be able to extract patterns of interest from the crowd.[17] As a result, the community saw the emergence of quantitative biological pathway or protein interaction analyses. Such systemic approaches aim at

providing a fine grained picture of the biological features of interest, hoping at identifying pathology specific disturbances undetectable otherwise.

This process can typically be subdivided into four main tasks: (1) the identification of the biological entities, (2) their absolute or relative quantification, (3) the functional analysis of these entities, and (4) the public dissemination of the results in standardized formats. Starting from the identified and quantified peptides and proteins results canonically obtained from a shotgun proteomics experiment,[18] we thus detail in the present review the current follow-up resources available in proteomics. Focusing on free and user-friendly tools, we list the applications – when existing – allowing researchers to reach these objectives. We also list the potential pitfalls involved in these post-identification steps.

2. Global PTM, Peptides and Protein identification

The typical outcome of a proteomic identification process is a list of identified peptides and proteins with posttranslational modifications (PTMs) mapped onto the sequence. However, these identifications typically contain a certain proportion of false positives.[19] Tremendous progress has been achieved in the monitoring of error rates in proteomics, notably with the use of target-decoy databases[20] – as comprehensively reviewed by Nesvizhskii AI.[21] It has therefore become possible to filter a dataset of interest at a desired False Discovery Rate[22] (FDR) independently from the specific scoring used by the search engine – an demarche common to other scientific fields.[23] For example, in a previous study,[24] where three isoelectric focusing (IEF) fractions of the same sample were analyzed using OMSSA,[25] a canonical FDR threshold set at 1% required the filtering of all hits with an e-value higher than 0.15, 0.20 and 0.19 for fractions 3, 9 and 20, respectively.

When comparing the target and decoy distribution of hits at these scores, it is possible to estimate an unbiased quality metric, the Posterior Error Probability (PEP).[26] Concretely, among hundred hits with a PEP of 25%, one expects 25 false positives, 75 true positives. The complement of the PEP (100%-25%=75%) hence indicates a confidence in the identification. Note that the threshold scores used for the IEF fractions example, although very similar, correspond to a confidence of 72%, 96% and 84%, respectively. This variability in confidence between fractions at fixed FDR shows the heterogeneity found in proteomic results and highlights the necessity for thorough statistical post-processing of the identification matches.

As schematized in Figure 1A, proteomic experiments typically consist of several samples that are measured in replicates (technical and biological) and that may each be further fractionized. Peptides are inferred from the obtained mass spectra, and proteins are then in turn inferred from the peptides. In the example of the IEF fractions above, we illustrate the importance of processing sets of spectra specifically: the PEP at a given OMSSA score differs from one fraction to the other. Using the OMSSA score for the merged PSM set hence results in a lower identification rate (6,084 PSMs at 1% FDR, in orange Figure 1A) compared to the same set scored using a fraction specific PEP (in black Figure 1A, 6,247 PSMs at 1% FDR: +2.6%). Such specific processing is comparable to charge and modification specific scoring[27], also mandatory when different mass spectrometers or experimental workflows are used on the same sample. A critical point is then to ensure a statistically relevant size of the subgroups of PSMs retained for scoring.[28] Statistical processing hence makes it possible to filter identification matches at a given quality threshold with a high accuracy[29] and merge the results *a posteriori*.

However, as illustrated in Figure 1B with the concatenation of three hypothetical datasets, merging results obtained on different replicates substantially increases the share of false positives since false positive identifications are more likely to differ between replicates than true identifications – a problem well known to affect searches using multiple search engines[30] and peptide and protein inference approaches.[31] In this simple example, where every dataset was filtered to 1% FDR, 25% of the correct matches were unique to a particular dataset, while all false positives were unique. The final FDR therefore reached 1.7% across the datasets: it is hence vital to monitor the quality level of the final result set. Crucially, as illustrated in Figure 1C, a peptide or a protein can score moderately (in orange) in each of the replicates (like protein D), preventing it from being validated at a quality driven FDR within that replicate. However, its presence among all replicates may make it more confident than another protein scoring well in only one replicate (like protein C). Keeping all identifications from all datasets when creating the merged results and subsequently filtering the combined set thus allows rescuing such peptides and proteins, reducing the False Negative Rate (FNR).

Finally, as illustrated in Figure 1D, when a peptide is shared between different proteins (e.g., peptide 2 that is shared between proteins A and B), it is not always possible to resolve the correct protein identification; this is the well-described but often underestimated protein inference problem,[32] which has particularly strong incidence on the quantification of proteins. In the illustrative example Figure 1D, the uniquely matched peptide 1 scores well in the first replicate and gives evidence for the presence of protein A. In replicate 2 however, this peptide now receives a poor score and would not pass a stringent quality threshold, thus impairing the protein inference within this replicate. This kind of situation

typically occurs when proteins are identified using different fractionation methods or different mass spectrometers. It is hence crucial to consider all peptide candidates across all replicates for protein inference as well.

In summary, an ideal post-identification workflow for proteomics treats identification results accounting for specificities of fractions and replicates (technical and biological) while taking advantage of the study design in its whole to reduce the FNR at a controlled FDR. However, such a statistical analysis is complicated by the fact that identification results between fractions and replicates are not independent. Moreover, there is no guarantee that the target/decoy strategy holds when merging replicates. On the contrary, it can lead to similar issues as multi-stage search strategies.[33] Finally, since such experimental designs can easily lead to the generation of several millions of spectra – as evidenced by some of the biggest datasets[34-37] submitted to the PRoteomics IDEntifications (PRIDE) repository,[38] global analysis of complex proteomic studies are also challenging in terms of processing time, computational space, and data management. Several free packages exist that allow the processing of large sets of spectra[39-43], these offer different solutions for the final compilation of the results between replicates, most notably, the MaxQuant/Perseus[39] tool combination allows combination of large datasets, statistical analysis and interaction with external resources.

3. PTM, Peptides and Protein quantification

The identified peptides and proteins are subsequently quantified – taking into account their modification status. Various non-targeted methods exist for this purpose, with the quantification output being either absolute or relative. Absolute quantification is obtained by spectrum counting[44-46] or by comparing intensities of peptides of interest with a spiked-in labeled version of known quantity.[47] Relative quantification is usually achieved by comparing the intensities of peptides measured in parallel or multiplexed after chemical labeling.[48] The quantification is then directly obtained by comparing peptide intensities[49] or by comparing intensities of reporter ions released by the label upon fragmentation.[50, 51] Several reviews cover the advantages and shortcomings of these methods.[52-56]

The obtained intensities are typically normalized prior to processing. As demonstrated in Figure 2A with different spectra of a regulated peptide monitored by iTRAQ quantification (courtesy of F Beck and RP Zahedi, unpublished data), when several quantification channels are available, a reference channel is typically used for normalization (here the reporter at m/z 114). The result after this normalization step shows high variance for the 115/114, 116/114 and 117/114 ratios, however. As shown in Figure 2B, after normalizing using the average of intensities at m/z 116 and 117, it actually appears that the high variance was due to the intensities at m/z 114 and this variance was propagated to all other channels by normalization. It is thus preferable to normalize using the median of the intensities or on a set of intensities which are known from the experimental design to be a reliable reference. A final ratio is estimated (shown in red Figure 2A and B) for the channel using an estimator that combines robustness and accuracy: maximum likelihood estimators were shown to

perform best.[13] This type of processing can be achieved using user friendly tools such as Rover[57] or Perseus (http://maxquant.org/).

The overall aim of statistical studies based on quantification results is typically to assert the confidence in a hypothesis, for instance the significance of a protein regulation regarding a hypothesis of stability: *is the protein significantly regulated when compared to the stable background?* Given a distribution of ratios, as illustrated in blue in Figure 2C for the distribution of iTRAQ ratios of two biological replicates from a publicly available dataset,[37] the confidence in a regulation will thus be obtained by comparing potentially regulated ratios to the global distribution. Assuming that the total complement of ratios is normally distributed, such a test can for instance be performed using thresholds calculated from the standard deviation of the ratios after logarithmic transformation. However, as illustrated by the dotted red line in Figure 2C, the mean and standard deviation based normal distribution is a poor model for the experimental data; here, the model fails the Kolmogorov-Smirnov test with a statistic of 0.0759 and a corresponding p-value of $8.686*10^{-10}$. It is possible to address this issue by reverting to the use of a robust distribution based on the median and quartile values – implicitly correcting for non-symmetrical distributions. As displayed by the green dashed line in Figure 2C, the match with the actual distribution is improved in this case. Yet the Kolmogorov-Smirnov test continues to fail (albeit with a much less dramatic statistic of 0.0317 and p-value of 0.04698). This example shows the crucial impact of the choice of the statistical model, and it illustrates that the significance of regulation must be computed and interpreted with care.

In more advanced experimental designs, the question to answer is typically more complex, for example: *is a protein significantly regulated in patient samples when compared to control*

samples? This question is typically assessed by analysis of variance (ANOVA). The design of the experiment plays a crucial role here,[9] especially regarding the number of replicates that will determine the statistical power of the experiment. The power of an experiment describes its ability to discriminate the actually regulated compounds from the experimental artifacts.[10] The subject of replication is essential in proteomics due to the low number of samples available for clinical studies. Low number of samples also make the study more sensitive toward outliers thus creating important issues regarding pooling strategies as often used with gel-based proteomics.[58] Perseus (http://maxquant.org/) provides several analysis strategies that support the statistical analysis of quantitative experiments in a user friendly fashion, as illustrated in Figure 2D with the display of four iTRAQ ratios obtained from biological replicates,[37] where the color represents the abundance of the protein.

4. External resources for functional studies

The emergence of resources for functional proteomics has broadened the scope of quantitative proteomics from the monitoring of single peptides or proteins to entire systems. The rationale behind such analyses is that biological deregulations do not only affect single proteins but rather disturb or activate entire systems. The comparison of the sub-proteome of interest between the healthy and disease state thus provides a fine grained picture of the regulations involved. In order to enrich the resulting large amount of identification and quantification results with functional knowledge, the community can fortunately already make use of various well-established resources, as reviewed in detail by Vizcaíno JA *et al*..[16] These external resources can be most readily queried using the

accession numbers of the identified and quantified proteins. Note that the portability of such accession numbers across databases and through time (as some accession numbers may not be stable over time) is ensured by a very powerful and user friendly application: the Protein Identifier Cross-Referencing (PICR) service.[59]

Proteins are also associated with so-called Gene Ontology (GO) terms,[60] that are organized into three categories: cellular component, molecular function and biological process. These GO terms can be obtained and analyzed for a given list of protein accession numbers using web interfaces[61, 62] or using dedicated software like Ontologizer.[63] It is moreover possible to compare the frequency of occurrence of a GO term in a dataset to the frequency of the same term in a given background. Similar to the analysis performed in quantification studies, the GO term expression analysis comes with a question, most notably: *is my term significantly differentially expressed in diseased samples?* Although quite some effort has been expended towards the correction of potential biases in the evaluation of GO term expression,[64] it is crucial to note that the type of sample used might influence the answer. For instance, when studying a given cell type, the detected regulated expression might be due to the cell type and not the disease.

Another source of potentially useful external information comes from interaction databases[65-67] that list known interacting proteins, as displayed in Figure 3A for data obtained from STRING[68] on protein P04114. Note that PSIMEX,[69] an integrated version of these resources, is available. Further information about protein function is available *via* pathway analyses[70, 71] as displayed in Figure 3B with a view on the pathway *platelet homeostasis* in Reactome.[71] Interestingly, it is possible to study the expression levels of a pathway in a dataset.[72] It is however important to consider the sensitivity of such an

approach in light of protein detection and protein inference issues. Indeed, if unique peptides for particular isoforms occurring in the pathway are not found, the coverage might easily be falsely underestimated. The pathways obviously also contain entities that are not detected in typical proteomic studies, such as metabolites, yet precisely because of this reason, pathways provide an ideal means to link proteomics identifications to metabolomics data in multi-omics analyses. Finally, pathway coverage is obviously also highly dependent on the FDR thresholding used during identification processing.

Further information on proteins that can be quite relevant to the interpretation of proteomics results are provided by their three-dimensional structure,[73] as illustrated in Figure 3C where a protein structure is displayed using jmol[74] in the PeptideShaker (http://peptide-shaker.googlecode.com) tool. Identified peptides and their modifications are mapped onto the structure, immediately revealing the identified peptides and the sites of modification from a structural perspective. The latter can provide powerful orthogonal information for modification site localization.[75]

It is thus clear that a complete cloud of heterogeneous resources are available to the researcher interested in placing proteomics results in a biological context. Navigating a variety of resources one by one can be very cumbersome and error-prone however, and simplified methods to access all this information in single analysis sweeps have therefore been constructed. Of these, Biomarts and the DAS protocol stand out as useful resources. Originally developed for the Ensembl database,[76-79] Biomarts allow dynamic, easily user-defined querying of available resources, offering a uniform interface regardless of the underlying resource. Moreover, Biomarts for two different resources can be linked, allowing across-resource queries. As a result, it is quite easy to perform exhaustive and up

to date protein annotation for a large dataset from a single (across-)Biomart query. The Distributed Annotation System[80] (DAS) on the other hand, is meant to aggregate positional annotations for a given protein or gene sequence from a very broad set of compatible annotation servers. Web interfaces like Dasty[81] provide user-friendly tools to aggregate these resources, and present the annotations in a global overview (see Figure 3D). Despite the capacity of these web interfaces to gather functional information, mapping vast quantitative proteomic studies on these external resources and navigating the results rapidly becomes impossible. Notably, no free tool allows taking into account the experimental design in the statistical processing of the external information while allowing its intuitive navigation, although open source network navigation interfaces already exist.[82]

5. Public repositories

As demonstrated in Section 4 above, proteomics research is not anymore limited to obtaining a dataset in the lab, but takes increasing advantage of the global knowledge on proteins, their interactions, features and roles as determined by the collective actions of the community. In order to improve the outcome of this global process of integrating knowledge, public repositories were set up to explicitly enable the sharing of proteomics data and associated results; for a detailed review see[83]. By making their data available, scientists can actively contribute to an overall collaborative effort, enabling their findings to act in feedback loops by for instance aiding in the annotation of protein databases that are in turn so useful in proteomics research. Recently, substantial progress in this area has been made to facilitate the submission to repositories with applications like PRIDE Converter[84, 85] and the ProteomeXchange (http://proteomexchange.org) submission tool. At the same time, inspecting and retrieving data from repositories such as PRIDE[38] has been made much easier thanks to the powerful PRIDE Inspector application[86] that greatly simplifies the interaction between the user and the repository. In fact, user friendliness is a vital feature for repositories, as the ultimate potential of the data will depend on the user's ability to deposit their findings in the repository. An important yet often overlooked aspect here is the annotation of the data, which is of fundamental importance for the efficient re-use of the information. Here, user-friendly interfaces such as the OLS Dialog[87] can aid substantially in lowering the required effort to correctly annotate a dataset. Automated checks on the semantic validity of the annotation can then be used to ensure that the user reports the correct type of information.[88] High quality public datasets are not only of interest for the

simple sharing of the obtained results, but also for their potential value when reused in research with a novel focus.[89-92]

6. Discussion and overview

Thanks to the advent of shotgun proteomics, protein analysis went from the laborious identification of a few entities to the automated identification and quantification of thousands of proteins with their mapped PTMs. However, for both identification as well as quantification, current workflows contain some unseen pitfalls that are not always adequately handled by established methods. This is especially the case for advanced, multi-sample experiments using several replicates. This is unfortunate, since such experiments are actually of the highest biological interest. The community is thus in need for renewed efforts to develop powerful, user friendly software able to handle such experiments easily and correctly. This task is however complicated by the amount of data involved. Indeed, in order to resolve protein inference issues, precisely map PTMs, and achieve global proteome quantification, it is necessary to navigate and process several millions of spectra in a practical time scale with reasonable computational requirements. To successfully tackle such a challenge, new, more advanced computational techniques may be required,[93] while reducing algorithmic complexity without sacrificing performance should be a high priority for developers in the field.

These techniques should no longer solely focus on delivering lists of peptides and proteins, but aim at identifying entire biological systems, mapping the experimental results on them and allow their intuitive navigation. This task is made particularly difficult by the complex experimental design of many proteomic experiments. Notably, the combination of shotgun and targeted studies for the study of functional changes in proteomics is crucial for the identification and validation of biomarkers. However, such investigations currently rely heavily on inefficient manual data manipulation.

Since functional proteomics relies on acquired knowledge, the quality and quantity of information available is strongly species dependent. Moreover, the study of certain samples breaks standard statistical assumptions in proteomics, notably in terms of database size and identification reliability. Despite the availability of resources for some organisms, [94] these problems are commonplace in plant proteomics and metaproteomics[95] for instance. There is therefore a strong need for increased species coverage for functional proteomics, which may be answered at least in part by cross-species portability of the existing annotation, as for instance performed by Reactome and Ensembl [REFS!!].

The second breakthrough that has become evident in proteomics over the past few years is the generalization of data and knowledge sharing. This collective effort of the community has enriched the field with global access to large amounts of high quality information. Here again, despite the progress in data sharing, current resources remain mainly populated with data from established methods and well-studied model organisms. Species with a partially or non-sequenced genome are simply not yet part of this dynamic. Similarly, innovative experimental designs do not always fit in the predefined data structures of external resources and repositories. This, together with the sometimes insufficient annotation of the datasets currently impairs the mapping of experimental data to external resources, hence blocking their use for functional proteomics. This issue can be solved by the establishment of efficient curation systems, and the development of robust, automated reprocessing methods.

Yet despite these caveats, the accurate and quantitative systematic proteomics analysis of model organisms is now within reach, by bringing together cutting edge competence in

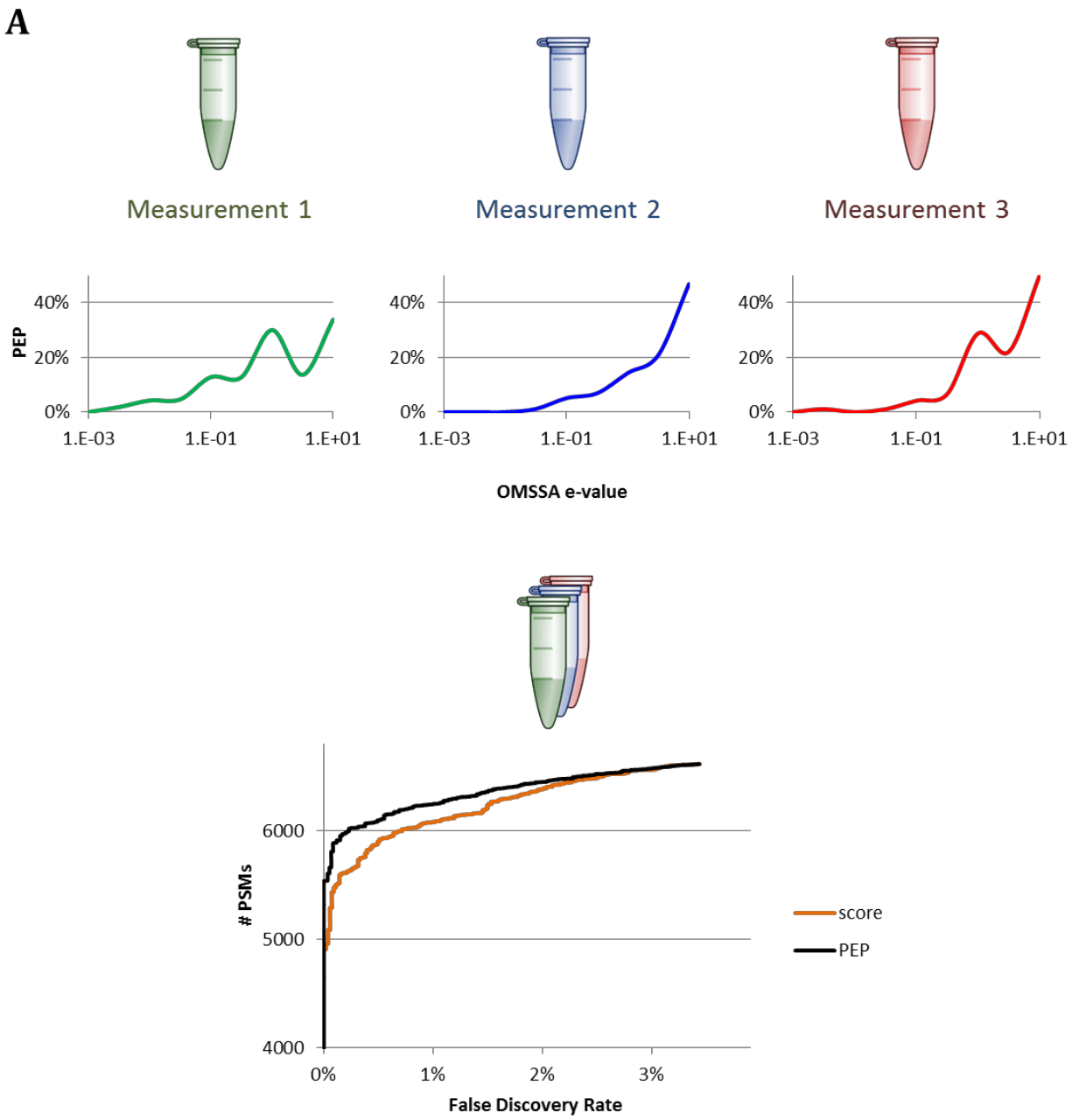
various fields comprising biology, sample preparation, chromatography, mass spectrometry, data interpretation and statistics.

Acknowledgments

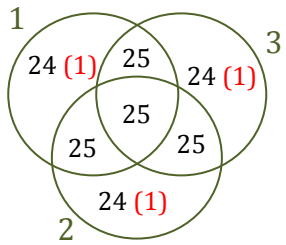
The financial support by the Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen and by the Bundesministerium für Bildung und Forschung (SARA, DYNAMO) is gratefully acknowledged. L.M. acknowledges the financial support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks") and the PRIME-XS and ProteomeXchange projects funded by the European Union 7th Framework Program under grant agreement numbers 262067 and 260558.

Figures

Figure 1:



B



FDR: 1% -> 1.7%

C

	1	2	3	Id
Protein A				
Protein B				
Protein C				
Protein D				
Protein E				
Protein F				
Total	**2**	**1**	**1**	**4**

FNR: 50% -> 33%

D

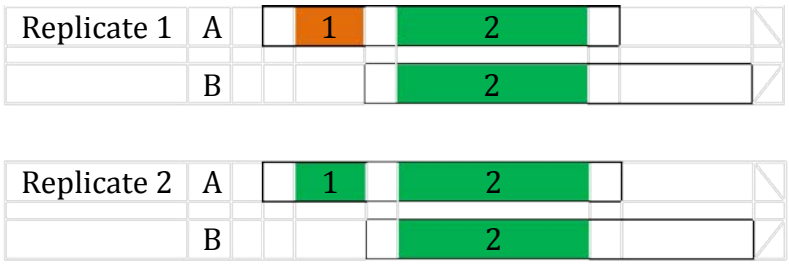


Figure 2:

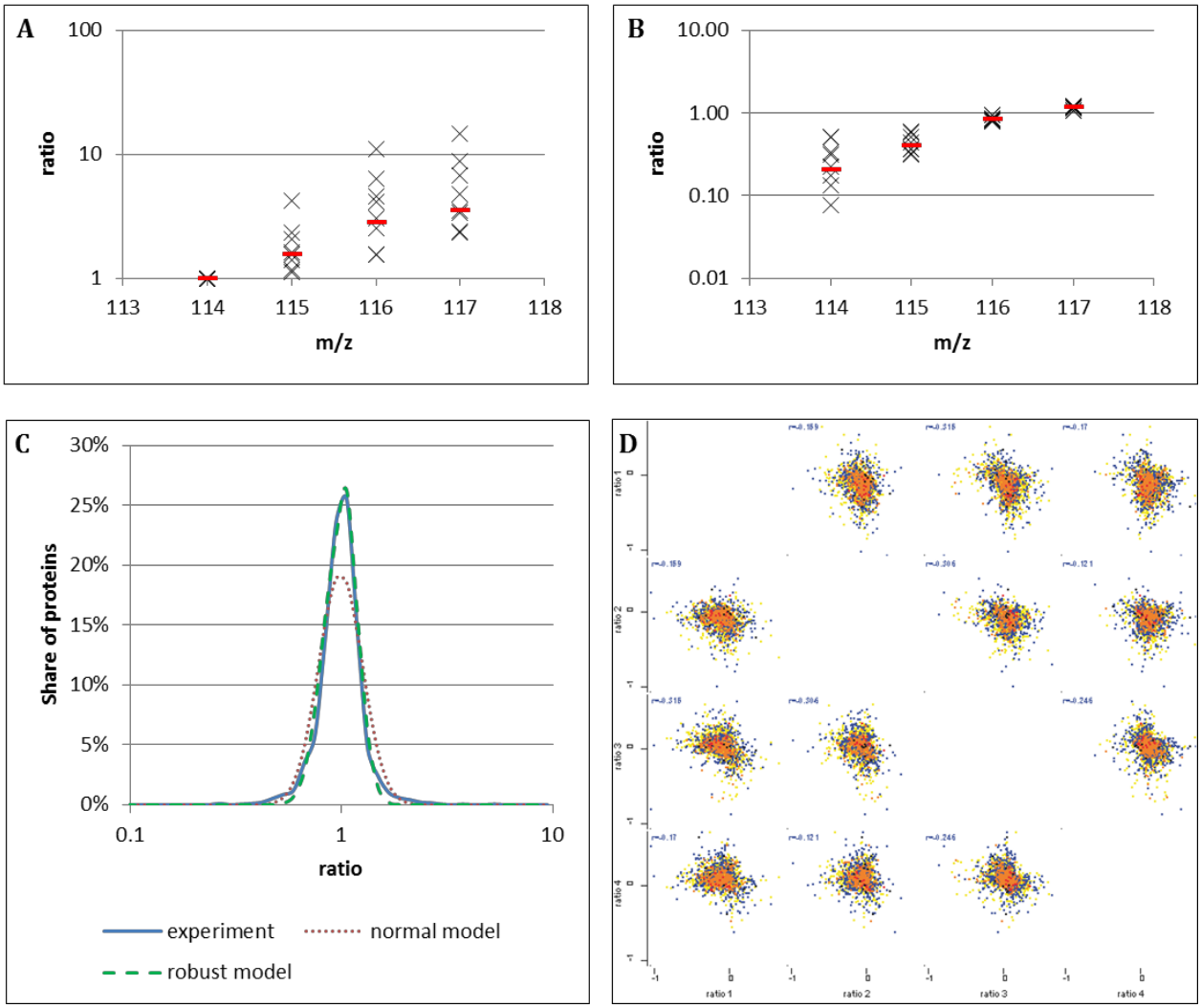
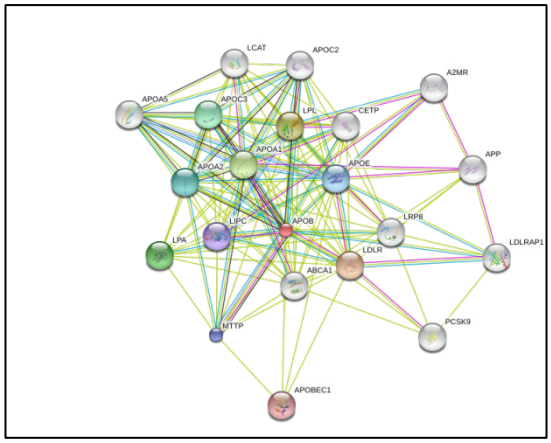
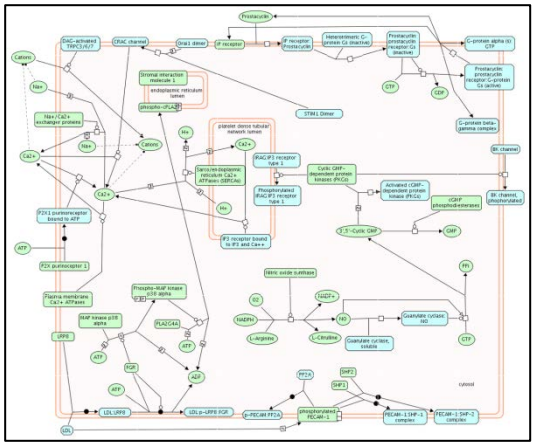

A
100
10
1
ratio
113
114
115
116
117
118
m/z
B
10.00
1.00
0.10
0.01
ratio
113
114
115
116
117
118
m/z
C
30%
25%
20%
15%
10%
5%
0%
Share of proteins
0.1
1
10
ratio
experiment
normal model
robust model
D
ratio 1
ratio 2
ratio 3
ratio 4

Figure 3:

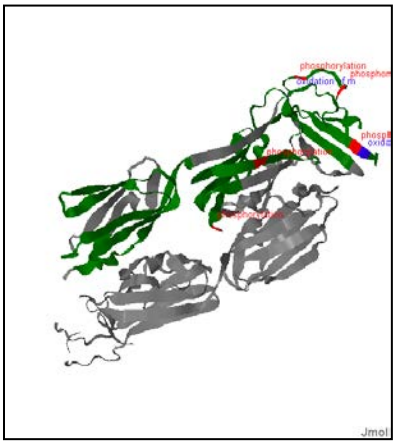
A



B



C



D

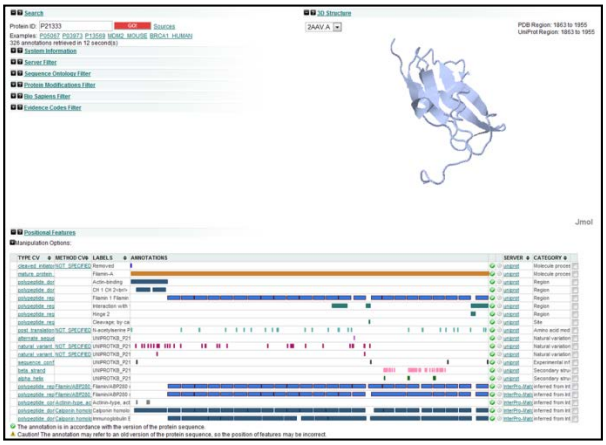


Figure legends

Figure 1: **Taking advantage of the experimental design**. (A) A typical proteomics experiment consists of several samples analyzed in replicates. Here, we take the simple example of three measurements of Isoelectric Focusing fractions from which we want to infer peptide and protein identifications. For every fraction, we represent the target/decoy derived Posterior Error Probability (PEP) at a given score. For the merged result set of Peptide-Spectrum Matches (PSMs), the number of PSMs is plotted at a given False Discovery Rate (FDR) when sorted against the OMSSA score (orange) and against the inferred PEP (black). (B) Processing all samples separately and merging the results increases the FDR substantially: considering an example where 25% of the proteins identified in a sample are unique to that sample, and this includes all false positives (numbers in red). When merging these three datasets (that are each filtered at 1% FDR), a final dataset is obtained with an FDR equal to 1.7%. (C) When considering six proteins identified in the three datasets at different confidence levels (indicated by red, orange and green for bad, medium and good confidence, respectively), it can for instance be seen that protein D is found in all samples yet is not validated due to its moderate score in each sample. The fact that it is found in all samples however, makes it quite likely that it should in fact be included in the global set of identifications. Indeed, although a false negative in all datasets, this protein could be rescued by scoring the identifications globally. Similarly, protein B is not validated in sample 3 but its presence in the global identification suggests that the peptides found in sample 3 should be used for quantification. (D) In proteomics it is sometimes impossible to infer the presence of a protein due to the absence of an identified unique peptide as illustrated here for replicate 1. While protein sequences A and B can be

distinguished by peptide 1, this peptide does not receive a high enough score (orange) for identification, and will therefore not be used for protein inference. In replicate 2 however, peptide 1 receives a higher score (green) allowing the unambiguous identification of protein A. Yet if protein A is confidently identified in the second replicate, it is likely to have been in the first replicate as well. A suitable study design can thus help resolve protein inference by analyzing the data globally.

Figure 2: **Quantification post-processing**. (A) An example of peptide iTRAQ 4-plex ratios measured in different spectra (black crosses). When normalizing on the 114 channel, the obtained peptide ratio using a maximal likelihood estimator (red) for the different channels, 114:115:116:117 are 1:1.56:2.79:3:50, respectively. Note how important the variance on every channel is: regulation between 116 and 117 cannot be confidently assessed here. (B) when normalizing on the average intensity of channels 116 and 117, the peptide ratios are 0.2:0.4:0.82:1.18, equivalent to 1:2.00:4.10:5.85. The regulation is hence more pronounced and the variability on the channels is reduced, except for the low intensities. In (B) significant regulation can be inferred whereas in (A) no significance is detected; this simple pitfall can be avoided by normalizing on the most reproducible measurements. (C) When asserting the regulation of proteins, as performed here with 1,869 iTRAQ protein ratios obtained between two replicates (blue distribution), one typically compares the regulation to the standard deviation of the distribution under the assumption that 1.96 times the standard deviation yields 95% confidence for a two-sided test. The assumption is of course that the data is normally distributed. As shown by the red dotted line, the assumed normal distribution based on the mean and the standard deviation of the data, does not fit the

experimental distribution at all however, failing the Kolmogorov-Smirnov test (see main text). A normal distribution calibrated on the quartiles of the experimental distribution (dashed lines in green) fits the experimental distribution better but still does not perform very well on the Kolmogorov-Smirnov test (see main text). The choice of the reference statistical model is therefore crucial to assess regulation, yet this topic is rarely touched upon in quantitative proteomics studies. (D) Statistical analysis and display can be performed on any kind of data in the Perseus software. Here, the four ratios obtained from a set of 3,253 proteins quantified with iTRAQ between four biological replicates are displayed in correlation plots. The color gradient from black to red indicates increasing absolute quantification. Note that highly abundant proteins are less subject to relative variation than low abundant proteins.

Figure 3: **External resources for proteomics. (**A) A network of protein-protein interactions as rendered by STRING for Apolipoprotein B-100 (accession P04114). (B) Pathway *platelet homeostasis*, as shown in Reactome. (C) 3D structure of a protein with the identified peptides and post-translational modifications mapped and highlighted, as shown by PeptideShaker (http://peptide-shaker.googlecode.com). (D) Annotation view of a protein in Dasty, using combined information from a variety of annotation servers.

References

1. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** 422, (6928), 198-207.
2. Vidal, M.; Chan, D. W.; Gerstein, M.; Mann, M.; Omenn, G. S.; Tagle, D.; Sechi, S., The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. *Clin Proteomics* **2012,** 9, (1), 6.
3. Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M., System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol Cell Proteomics* **2012,** 11, (3), M111 013722.
4. Boersema, P. J.; Geiger, T.; Wisniewski, J. R.; Mann, M., Quantification of the N-glycosylated secretome by super-SILAC during breast cancer progression and in human blood samples. *Mol Cell Proteomics* **2012**.
5. Wisniewski, J. R.; Dus, K.; Mann, M., Proteomic workflow for analysis of archival formalin fixed and paraffin embedded clinical samples to a depth of 10,000 proteins. *Proteomics Clin Appl* **2012**.
6. Mann, M., Proteomics for biomedicine: a half-completed journey. *EMBO Mol Med* **2012,** 4, (2), 75-7.
7. Wisniewski, J. R.; Ostasiewicz, P.; Dus, K.; Zielinska, D. F.; Gnad, F.; Mann, M., Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol Syst Biol* **2012,** 8, 611.
8. Hanash, S.; Taguchi, A., The grand challenge to decipher the cancer proteome. *Nat Rev Cancer* **2010,** 10, (9), 652-60.
9. Oberg, A. L.; Vitek, O., Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res* **2009,** 8, (5), 2144-56.
10. Levin, Y., The role of statistical power analysis in quantitative proteomics. *Proteomics* **2011,** 11, (12), 2565-7.
11. Clough, T.; Thaminy, S.; Ragg, S.; Aebersold, R.; Vitek, O., Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics* **2012,** 13 Suppl 16, S6.
12. Bessarabova, M.; Ishkin, A.; Jebailey, L.; Nikolskaya, T.; Nikolsky, Y., Knowledge-based analysis of proteomics data. *BMC Bioinformatics* **2012,** 13 Suppl 16, S13.
13. Burkhart, J. M.; Vaudel, M.; Zahedi, R. P.; Martens, L.; Sickmann, A., iTRAQ protein quantification: a quality-controlled workflow. *Proteomics* **2011,** 11, (6), 1125-34.
14. Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D., Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **2012,** 13 Suppl 16, S5.
15. Ow, S. Y.; Salim, M.; Noirel, J.; Evans, C.; Rehman, I.; Wright, P. C., iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J Proteome Res* **2009,** 8, (11), 5347-55.
16. Vizcaino, J. A.; Mueller, M.; Hermjakob, H.; Martens, L., Charting online OMICS resources: A navigational chart for clinical researchers. *Proteomics Clin Appl* **2009,** 3, (1), 18-29.
17. Cox, J.; Mann, M., 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **2012,** 13 Suppl 16, S12.

18. Vaudel, M.; Sickmann, A.; Martens, L., Current methods for global proteome identification. *Expert Review of Proteomics* **2012,** 9, (5), 519-532.
19. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002,** 74, (20), 5383-92.
20. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
21. Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **2010,** 73, (11), 2092-123.
22. Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995,** 57, (1), 289-300.
23. Storey, J. D.; Tibshirani, R., Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **2003,** 100, (16), 9440-5.
24. Vaudel, M.; Burkhart, J. M.; Radau, S.; Zahedi, R. P.; Martens, L.; Sickmann, A., Integral Quantification Accuracy estimation for Reporter Ion based quantitative proteomics (iQuARI). *J Proteome Res* **2012**.
25. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *J Proteome Res* **2004,** 3, (5), 958-64.
26. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **2007,** 4, (10), 787-97.
27. Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology* **2006,** 24, (10), 1285-92.
28. Vaudel, M.; Burkhart, J. M.; Sickmann, A.; Martens, L.; Zahedi, R. P., Peptide identification quality control. *Proteomics* **2011,** 11, (10), 2105-14.
29. Ma, K.; Vitek, O.; Nesvizhskii, A. I., A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC bioinformatics* **2012,** 13 Suppl 16, S1.
30. Searle, B. C.; Turner, M.; Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* **2008,** 7, (1), 245-53.
31. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** 75, (17), 4646-58.
32. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005,** 4, (10), 1419-40.
33. Everett, L. J.; Bierl, C.; Master, S. R., Unbiased statistical analysis for multi-stage proteomic search strategies. *J Proteome Res* **2010,** 9, (2), 700-7.
34. Schrimpf, S. P.; Weiss, M.; Reiter, L.; Ahrens, C. H.; Jovanovic, M.; Malmstrom, J.; Brunner, E.; Mohanty, S.; Lercher, M. J.; Hunziker, P. E.; Aebersold, R.; von Mering, C.; Hengartner, M. O., Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol* **2009,** 7, (3), e48.
35. Cox, B.; Kotlyar, M.; Evangelou, A. I.; Ignatchenko, V.; Ignatchenko, A.; Whiteley, K.; Jurisica, I.; Adamson, S. L.; Rossant, J.; Kislinger, T., Comparative systems biology of human

and mouse as a tool to guide the modeling of human placental pathology. *Mol Syst Biol* **2009,** 5, 279.
36. Merrihew, G. E.; Davis, C.; Ewing, B.; Williams, G.; Kall, L.; Frewen, B. E.; Noble, W. S.; Green, P.; Thomas, J. H.; MacCoss, M. J., Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res* **2008,** 18, (10), 1660-9.
37. Burkhart, J. M.; Vaudel, M.; Gambaryan, S.; Radau, S.; Walter, U.; Martens, L.; Geiger, J.; Sickmann, A.; Zahedi, R. P., The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* **2012**.
38. Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R., PRIDE: the proteomics identifications database. *Proteomics* **2005,** 5, (13), 3537-45.
39. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** 26, (12), 1367-72.
40. Bertsch, A.; Gropl, C.; Reinert, K.; Kohlbacher, O., OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **2011,** 696, 353-67.
41. Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J., COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. *Proteomics* **2011,** 11, (6), 1064-74.
42. Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R., A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010,** 10, (6), 1150-9.
43. Slotta, D. J.; McFarland, M. A.; Markey, S. P., MassSieve: panning MS/MS peptide data for proteins. *Proteomics* **2010,** 10, (16), 3035-9.
44. Powell, D. W.; Weaver, C. M.; Jennings, J. L.; McAfee, K. J.; He, Y.; Weil, P. A.; Link, A. J., Cluster analysis of mass spectrometry data reveals a novel component of SAGA. *Mol Cell Biol* **2004,** 24, (16), 7249-59.
45. Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **2005,** 4, (9), 1265-72.
46. Paoletti, A. C.; Parmely, T. J.; Tomomori-Sato, C.; Sato, S.; Zhu, D.; Conaway, R. C.; Conaway, J. W.; Florens, L.; Washburn, M. P., Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc Natl Acad Sci U S A* **2006,** 103, (50), 18928-33.
47. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **2003,** 100, (12), 6940-5.
48. de Leenheer, A. P.; Thienpont, L. M., Applications of isotope dilution-mass spectrometry in clinical chemistry, pharmacokinetics, and toxicology. *Mass Spectrometry Reviews* **1992,** 11, (4), 249-307.
49. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **2002,** 1, (5), 376-86.

50. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003,** 75, (8), 1895-904.
51. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** 3, (12), 1154-69.
52. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **2007,** 389, (4), 1017-31.
53. Vaudel, M.; Sickmann, A.; Martens, L., Peptide and protein quantification: a map of the minefield. *Proteomics* **2010,** 10, (4), 650-70.
54. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* **2012,** 404, (4), 939-65.
55. Ong, S. E.; Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **2005,** 1, (5), 252-62.
56. Domon, B.; Aebersold, R., Mass spectrometry and protein analysis. *Science* **2006,** 312, (5771), 212-7.
57. Colaert, N.; Van Huele, C.; Degroeve, S.; Staes, A.; Vandekerckhove, J.; Gevaert, K.; Martens, L., Combining quantitative proteomics data processing workflows for greater sensitivity. *Nat Methods* **2011,** 8, (6), 481-3.
58. Karp, N. A.; Lilley, K. S., Investigating sample pooling strategies for DIGE experiments to address biological variability. *Proteomics* **2009,** 9, (2), 388-97.
59. Cote, 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* **2007,** 8, 401.
60. The Gene Ontology project in 2008. *Nucleic Acids Res* **2008,** 36, (Database issue), D440-4.
61. Binns, D.; Dimmer, E.; Huntley, R.; Barrell, D.; O'Donovan, C.; Apweiler, R., QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **2009,** 25, (22), 3045-6.
62. Sherman, B. T.; Huang da, W.; Tan, Q.; Guo, Y.; Bour, S.; Liu, D.; Stephens, R.; Baseler, M. W.; Lane, H. C.; Lempicki, R. A., DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **2007,** 8, 426.
63. Bauer, S.; Grossmann, S.; Vingron, M.; Robinson, P. N., Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **2008,** 24, (14), 1650-1.
64. Grossmann, S.; Bauer, S.; Robinson, P. N.; Vingron, M., Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* **2007,** 23, (22), 3024-31.
65. Kerrien, S.; Aranda, B.; Breuza, L.; Bridge, A.; Broackes-Carter, F.; Chen, C.; Duesbury, M.; Dumousseau, M.; Feuermann, M.; Hinz, U.; Jandrasits, C.; Jimenez, R. C.; Khadake, J.; Mahadevan, U.; Masson, P.; Pedruzzi, I.; Pfeiffenberger, E.; Porras, P.; Raghunath, A.;

Roechert, B.; Orchard, S.; Hermjakob, H., The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **2012,** 40, (Database issue), D841-6.
66. Bader, G. D.; Betel, D.; Hogue, C. W., BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **2003,** 31, (1), 248-50.
67. Xenarios, I.; Salwinski, L.; Duan, X. J.; Higney, P.; Kim, S. M.; Eisenberg, D., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **2002,** 30, (1), 303-5.
68. Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering, C., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **2011,** 39, (Database issue), D561-8.
69. Orchard, S.; Kerrien, S.; Abbani, S.; Aranda, B.; Bhate, J.; Bidwell, S.; Bridge, A.; Briganti, L.; Brinkman, F. S.; Cesareni, G.; Chatr-aryamontri, A.; Chautard, E.; Chen, C.; Dumousseau, M.; Goll, J.; Hancock, R. E.; Hannick, L. I.; Jurisica, I.; Khadake, J.; Lynn, D. J.; Mahadevan, U.; Perfetto, L.; Raghunath, A.; Ricard-Blum, S.; Roechert, B.; Salwinski, L.; Stumpflen, V.; Tyers, M.; Uetz, P.; Xenarios, I.; Hermjakob, H., Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* **2012,** 9, (4), 345-50.
70. Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y., KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **2008,** 36, (Database issue), D480-4.
71. Haw, R.; Hermjakob, H.; D'Eustachio, P.; Stein, L., Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **2011,** 11, (18), 3598-613.
72. Good, D. M.; Zubarev, R. A., Drug target identification from protein dynamics using quantitative pathway analysis. *J Proteome Res* **2011,** 10, (5), 2679-83.
73. Sussman, J. L.; Lin, D.; Jiang, J.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E., Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* **1998,** 54, (Pt 6 Pt 1), 1078-84.
74. Hanson, R., Jmol - a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography* **2010,** 43, (5 Part 2), 1250-1260.
75. Vandermarliere, E.; Martens, L., Protein structure as a means to triage proposed post-translational modification sites. *Proteomics* **2012**.
76. Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Brent, S.; Chen, Y.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; Gordon, L.; Hendrix, M.; Hourlier, T.; Johnson, N.; Kähäri, A.; Keefe, D.; Keenan, S.; Kinsella, R.; Kokocinski, F.; Kulesha, E.; Larsson, P.; Longden, I.; McLaren, W.; Overduin, B.; Pritchard, B.; Riat, H. S.; Rios, D.; Ritchie, G. R. S.; Ruffier, M.; Schuster, M.; Sobral, D.; Spudich, G.; Tang, Y. A.; Trevanion, S.; Vandrovcova, J.; Vilella, A. J.; White, S.; Wilder, S. P.; Zadissa, A.; Zamora, J.; Aken, B. L.; Birney, E.; Cunningham, F.; Dunham, I.; Durbin, R.; Fernández-Suarez, X. M.; Herrero, J.; Hubbard, T. J. P.; Parker, A.; Proctor, G.; Vogel, J.; Searle, S. M. J., Ensembl 2011. *Nucleic Acids Research* **2011,** 39, (suppl 1), D800-D806.
77. Kasprzyk, A.; Keefe, D.; Smedley, D.; London, D.; Spooner, W.; Melsopp, C.; Hammond, M.; Rocca-Serra, P.; Cox, T.; Birney, E., EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* **2004,** 14, (1), 160-9.

78. Smedley, D.; Haider, S.; Ballester, B.; Holland, R.; London, D.; Thorisson, G.; Kasprzyk, A., BioMart--biological queries made easy. *BMC Genomics* **2009,** 10, 22.
79. Kasprzyk, A., BioMart: driving a paradigm change in biological data management. *Database (Oxford)* **2011,** 2011, bar049.
80. Dowell, R. D.; Jokerst, R. M.; Day, A.; Eddy, S. R.; Stein, L., The distributed annotation system. *BMC Bioinformatics* **2001,** 2, 7.
81. Villaveces, J. M.; Jimenez, R. C.; Garcia, L. J.; Salazar, G. A.; Gel, B.; Mulder, N.; Martin, M.; Garcia, A.; Hermjakob, H., Dasty3, a WEB framework for DAS. *Bioinformatics* **2011,** 27, (18), 2616-7.
82. Bastian, M.; Heymann, S.; Jacomy, M., *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009.
83. Mead, J. A.; Bianco, L.; Bessant, C., Recent developments in public proteomic MS repositories and pipelines. *Proteomics* **2009,** 9, (4), 861-81.
84. Barsnes, H.; Vizcaino, J. A.; Eidhammer, I.; Martens, L., PRIDE Converter: making proteomics data-sharing easy. *Nat Biotechnol* **2009,** 27, (7), 598-9.
85. Cote, R. G.; Griss, J.; Dianes, J. A.; Wang, R.; Wright, J. C.; van den Toorn, H. W.; van Breukelen, B.; Heck, A. J.; Hulstaert, N.; Martens, L.; Reisinger, F.; Csordas, A.; Ovelleiro, D.; Perez-Riverol, Y.; Barsnes, H.; Hermjakob, H.; Vizcaino, J. A., The PRIDE Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol Cell Proteomics* **2012**.
86. Wang, R.; Fabregat, A.; Rios, D.; Ovelleiro, D.; Foster, J. M.; Cote, R. G.; Griss, J.; Csordas, A.; Perez-Riverol, Y.; Reisinger, F.; Hermjakob, H.; Martens, L.; Vizcaino, J. A., PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol* **2012,** 30, (2), 135-7.
87. Barsnes, H.; Cote, R. G.; Eidhammer, I.; Martens, L., OLS dialog: an open-source front end to the ontology lookup service. *BMC Bioinformatics* **2010,** 11, 34.
88. Montecchi-Palazzi, L.; Kerrien, S.; Reisinger, F.; Aranda, B.; Jones, A. R.; Martens, L.; Hermjakob, H., The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics* **2009,** 9, (22), 5112-9.
89. Matic, I.; Ahel, I.; Hay, R. T., Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites. *Nat Methods* **2012,** 9, (8), 771-2.
90. Foster, J. M.; Degroeve, S.; Gatto, L.; Visser, M.; Wang, R.; Griss, J.; Apweiler, R.; Martens, L., A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* **2011,** 11, (11), 2182-94.
91. Volders, P. J.; Helsens, K.; Wang, X.; Menten, B.; Martens, L.; Gevaert, K.; Vandesompele, J.; Mestdagh, P., LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* **2012**.
92. Gascoigne, D. K.; Cheetham, S. W.; Cattenoz, P. B.; Clark, M. B.; Amaral, P. P.; Taft, R. J.; Wilhelm, D.; Dinger, M. E.; Mattick, J. S., Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **2012,** 28, (23), 3042-50.
93. Halligan, B. D.; Geiger, J. F.; Vallejos, A. K.; Greene, A. S.; Twigger, S. N., Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms. *J Proteome Res* **2009,** 8, (6), 3148-53.

94. Tsesmetzis, N.; Couchman, M.; Higgins, J.; Smith, A.; Doonan, J. H.; Seifert, G. J.; Schmidt, E. E.; Vastrik, I.; Birney, E.; Wu, G.; D'Eustachio, P.; Stein, L. D.; Morris, R. J.; Bevan, M. W.; Walsh, S. V., Arabidopsis reactome: a foundation knowledgebase for plant systems biology. *The Plant cell* **2008,** 20, (6), 1426-36.
95. Muth, T.; Benndorf, D.; Reichl, U.; Rapp, E.; Martens, L., Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular bioSystems* **2013,** 9, (4), 578-85.