DOI 10.1002/pmic.201500295
Proteomics 2016, *16*, 214–225

REVIEW

Exploring the potential of public proteomics data

Marc Vaudel[1], Kenneth Verheggen[2,3,4], Attila Csordas[5], Helge Ræder[6], Frode S. Berven[1,7], Lennart Martens[2,3,4], Juan A. Vizcaíno[5] and Harald Barsnes[1,6]*

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway
[2] Medical Biotechnology Center, VIB, Ghent, Belgium
[3] Department of Biochemistry, Ghent University, Ghent, Belgium
[4] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium
[5] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[6] Department of Clinical Science, KG Jebsen Center for Diabetes Research, University of Bergen, Bergen, Norway
[7] Department of Clinical Medicine, KG Jebsen Centre for Multiple Sclerosis Research, University of Bergen, Bergen, Norway

In a global effort for scientific transparency, it has become feasible and good practice to share experimental data supporting novel findings. Consequently, the amount of publicly available MS-based proteomics data has grown substantially in recent years. With some notable exceptions, this extensive material has however largely been left untouched. The time has now come for the proteomics community to utilize this potential gold mine for new discoveries, and uncover its untapped potential. In this review, we provide a brief history of the sharing of proteomics data, showing ways in which publicly available proteomics data are already being (re-)used, and outline potential future opportunities based on four different usage types: use, reuse, reprocess, and repurpose. We thus aim to assist the proteomics community in stepping up to the challenge, and to make the most of the rapidly increasing amount of public proteomics data.

Received: July 13, 2015
Revised: August 25, 2015
Accepted: September 28, 2015

Keywords:
Bioinformatics / Computational proteomics / Data analysis / Databases / Data standards

1 Introduction

1.1 Background

Historically, a large proportion of the proteomics community was reticent to openly share the data they produced. However, the sharing of not only the knowledge obtained through proteomics experiments (through scientific publications), but also of the underlying data, has increasingly become standard practice, and is now even mandatory or strongly advised in many of the relevant scientific journals [1–3]. In addition, a number of funders (e.g. the Wellcome Trust and the NIH) are enforcing the public deposition of data from projects they fund as a way to maximize the value of the funds provided. As a result, the amount of publicly shared MS-based proteomics data has grown substantially, both in terms of number of submission and total data amount, as illustrated in Fig. 1.

Two key factors strongly contributed to this success: first, the sharing of the data has become much easier with the development of user-friendly tools and infrastructure; and second, the proteomics community, triggered by scientific journals and funders, has now agreed that it is good scientific practice to make the underlying data available when publishing novel findings.

There were several challenges to overcome in order to get to this point, see Fig. 2. The first of these challenges was the need for central and long-term public repositories to store the generated data. Several such generic repositories are now

Correspondence: Dr. Harald Barsnes, Proteomics Unit, Department of Biomedicine, University of Bergen, Jonas Liesvei 91, N-5009 Bergen, Norway
E-mail: harald.barsnes@uib.no
Fax: +47-55-58-63-60

Abbreviation: **PSM**, peptide to spectrum match

*Additional corresponding author: Dr. Juan A. Vizcaíno
E-mail: juan@ebi.ac.uk
Colour Online: See the article online to view Figs. 1–4 in colour.

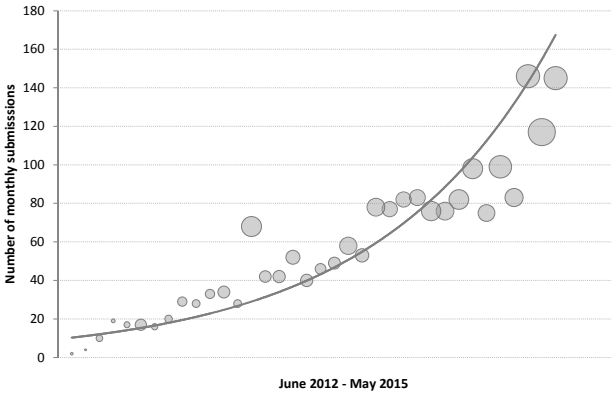

Figure 1. The amount of publicly available proteomics data is increasing, here indicated by the monthly submission statistics for PRIDE from June 2012 to May 2015. The *x*-axis represents the months and the *y*-axis the monthly number of submissions. The size of the bubbles indicate the data amount submitted each month. Note that the cumulative size of PRIDE data reached the 100 TB milestone in April 2015.

available, for example PRIDE [4], GPMDB [5], PeptideAtlas [6], and MassIVE (http://massive.ucsd.edu/ProteoSAFe) for shotgun results; and PASSEL [7], SRMAtlas (http://www.srmatlas.org), and Panorama [8] for targeted proteomics quantification data. More specific databases have also been established, related to: diseases, for example TBDB for tuberculosis [9]; organisms, for example ProteomicsDB [10] and the Human Proteome Map [11] for the human proteome, and pep2pro for *Arabidopsis* [12]; or subproteomes, for example CSF-PR [13] for cerebrospinal fluid or TOPPR [14] and TopFIND [15] for in vivo cleaved proteins. For a comprehensive overview of the current proteomics databases and repositories, please see Perez-Riverol et al. [16].

The next milestone was the development of data-sharing standards and associated software libraries, allowing ready access to otherwise proprietary data formats [17]. This ongoing endeavor, led by the HUPO-PSI (Human Proteome Organization—Proteomics Standards Initiative—http://www.psidev.info), has resulted in key data standards for the field, including mzML (for MS data), mzIdentML (for peptide/protein identification data), mzTab (for peptide/protein identification and quantification data), mzQuantML (for peptide/protein quantification data), and TraML (for transition lists in targeted proteomics approaches) [18–22]. Importantly, support for these standards is provided through software libraries or tools such as ProteoWizard [23], PRIDE Converter [24, 25], mzidLibrary [26], and PRIDE Inspector [27]. Successful adoption of these standards is moreover demonstrated by the existence of import and/or export capabilities in many of the most popular software in the field.

The final piece of the puzzle was the creation of an overarching system to share submitted data between repositories, and to develop a single, user-friendly submission workflow. This goal was obtained with the establishment of the ProteomeXchange consortium [28], which connects some of the most used proteomics databases (at present PRIDE, MassIVE, PASSEL, and PeptideAtlas) with a single submission system and the use of unique identifiers that can be tracked across these databases and over time.

However, while publicly available proteomics data represent an invaluable resource for extracting new knowledge [29], they have so far, with a few notable exceptions, remained largely unused. At the same time, data reprocessing has become the standard in related fields, such as genomics, see Rung et al. [30]. The time has now come for the field of proteomics to also start utilizing its public data as a test bed for novel ways of interpreting proteomics data, and as a potential goldmine for new discoveries. The heterogeneous nature of the accumulated data also provides a global view on the state

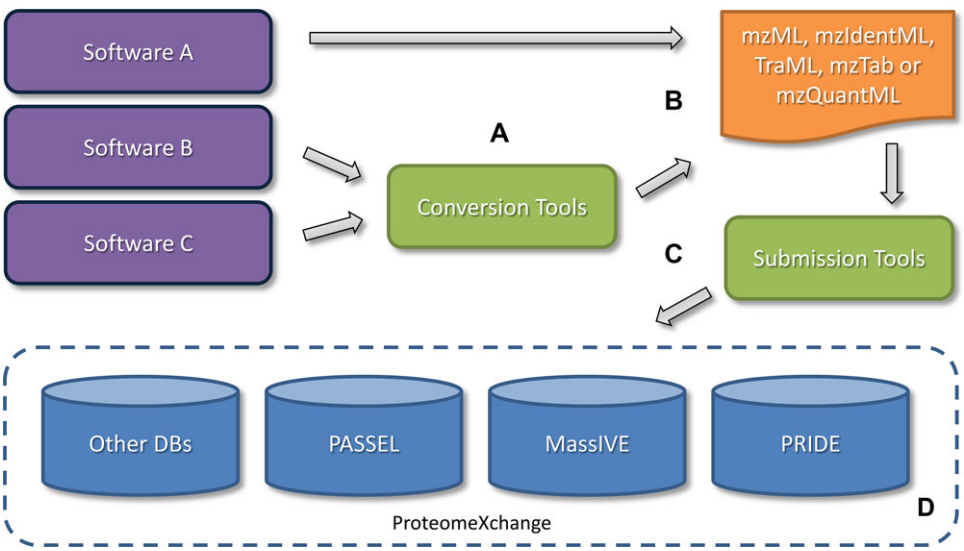

Figure 2. The major milestones that enabled efficient proteomics data sharing: (A) standard data formats for sharing proteomics data, (B) data format converters and software exporters able to generate output in the standard formats, (C) tools for simplifying the submission of proteomics data to central proteomics repositories, and (D) central proteomics repositories that store and disseminate public proteomics data, here indicated by the main ProteomeXchange member repositories.

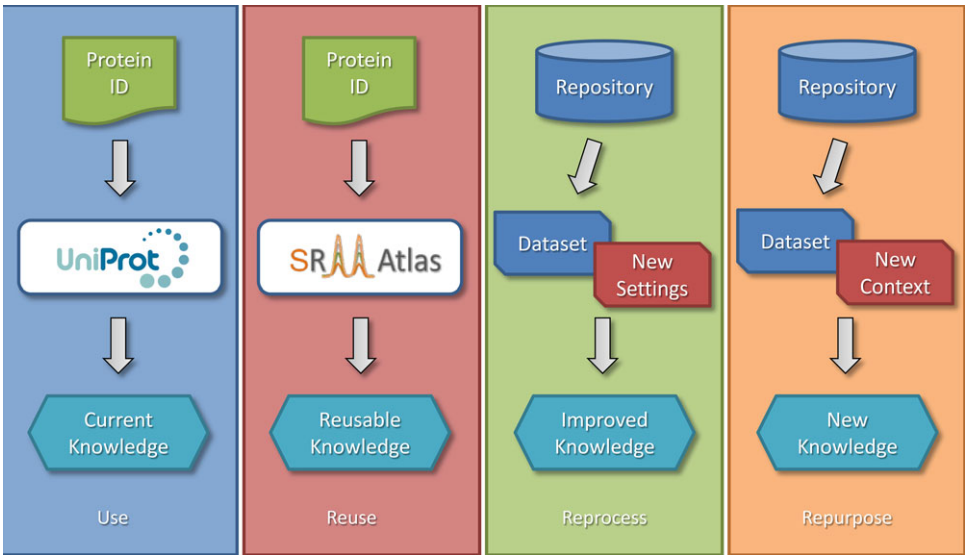

Figure 3. The four ways in which public proteomics data can be utilized: (i) use, (ii) reuse, (iii) reprocess, and (iv) repurpose. See main text for details.

of the art and the evolution of the field as a whole, and reduces bias toward specific protocols or instruments.

There are four ways in which these shared proteomics data can be utilized: (i) *use*, (ii) *reuse*, (iii) *reprocess*, and (iv) *repurpose* (Fig. 3), each of which will be described in detail in the following sections.

1.2 Data use through protein knowledge-bases

An example of the direct *use* of proteomics data is by looking up information about a given protein as indexed in an online protein knowledge-base, such as UniProt [31] or neXtProt [32]. This does not result in knowledge beyond what has already been published, but does provide the means to understand the current context of the protein(s) in question. For example, MS proteomics data deposited in public repositories is used by UniProt and neXtProt to enrich sequence annotations at the level of the evidence that supports protein existence (isoforms and variant sequences included). This information is provided to users in two ways: (i) via the protein evidence values, or (ii) through cross-references to proteomics resources (e.g. PRIDE and PeptideAtlas, among others). The next step will be the incorporation of PTMs based on the information available in proteomics repositories, as is already done in databases such as PhosphoSitePlus [33]. Currently, the main integration of this information occurs via manual curation of relevant publications.

1.3 Reusing data to improve proteomics approaches

In the case of *reuse*, information is not only extracted, but also reused in new experiments with the potential of generating new knowledge. One of the best examples is the reuse of SRM transitions generated by others, via SRMAtlas (http://www.srmatlas.org) or Panorama [8], where existing transitions for specific proteins in a given instrumental setup can be found. Note that it is also possible to develop tools to look for novel transitions in publicly available shotgun datasets. For example, MRMaid [34], PeptidePicker [35, 36], and ProteomicsDB [10] do this by reusing identification data coming from PRIDE and other sources.

One particular type of data reuse, already popular in other disciplines, is to analyze data from a large number of publications/datasets in a combined way, so-called *meta*-analysis studies. Indeed, the availability of large amounts of proteomics data has the advantage that it can be used for data mining purposes, that is extracting aggregated knowledge from the data provided by the community worldwide. The principle being: the more data, the better the understanding.

In fact, *meta*-analysis studies can indeed provide new information that can be directly applied in proteomic analytic workflows. One example is a study aimed at improving the understanding of the cleavage mechanism and performance of trypsin [37, 38], a crucial parameter in proteomic workflows. By inspecting the cleavage profile of all peptide identifications deposited in PRIDE, it was possible to train an algorithm that predicts trypsin cleavage sites, a functionality that is available through a web interface [39]. Similarly, the study of deposited data was used to monitor peptide elution during LC, and enabled the optimization of gradients in silico [40]. Public MS data have also been mined to study the fragmentation pattern of different fragmentation methods [41], and to predict peptide fragmentation patterns [42].

PRIDE data have also been reused through the combination of data from significantly different experimental setups. For example, Klie et al. [43] used a noise-tolerant algorithm to extract new knowledge from the datasets that comprise the

HUPO Plasma Proteome Project [44]. Another example can be found in Müller et al. [45], where two proteomics datasets related to the CNS were remapped against a more recent version of the protein sequence database used in the original studies. This enabled the authors to look for the expression of specific splice isoforms from CNS-related genes. Finally, in another example of PRIDE data reuse, UniProtKB was determined to be the most suitable reference database for long-term proteomics data storage [46].

Large-scale biological results can also be reused because of their indexing in databases, notably via so-called BioMarts [47] or more recently, web services [48, 49]. Mining such data in their biological context may allow the extraction of novel biomarkers, as discussed in Griss et al. [50].

1.4 Reusing data via spectral libraries and spectral archives

Additional spectrum interpretation strategies such as de novo sequencing or spectral databases are also promising approaches to increase the identification rate of spectra in MS-based proteomics. The creation of spectral libraries most strongly benefits from the growing amount of shared data [51, 52]. Several repositories, including PeptideAtlas, GPMDB, and PRIDE, and research groups such as the one at NIST (National Institute of Standards and Technology), provide spectral libraries for different species, which can in turn be used to perform spectral searches.

When assessing the similarity of spectra, spectral clustering can be performed [53–56]. While transitive identifications and consensus or representative spectra have been reported in all of these studies, the concept was further developed in the creation of spectral archives [57]. Spectral clustering has since been adopted by PRIDE to make quality assessments on the submitted data at the peptide to spectrum match (PSM) level [58]. After clustering, a representative spectrum is built for all peptides consistently identified across different datasets. The accuracy of this representative spectrum thus improves with every new dataset submitted to PRIDE, allowing an automated quality assessment of the PSM data. The key role of proteomics repositories in the further development of spectral archives was highlighted by H. Lam, who envisioned a future where it would be possible to perform a centralized data analysis by performing spectral searches [59].

1.5 Data reprocessing through improved bioinformatic approaches

In the case of *reprocess*, the data are reprocessed with the intention of obtaining new knowledge or to provide an updated view on the results. This can result in novel findings, but mainly serves the same purpose as the original experiment. For example, a shotgun dataset can be reprocessed with a different algorithm or an updated sequence database.

Perhaps, the simplest step one can take when reprocessing a dataset is to analyze the potential effect of adding common contaminants if these were not included in the original search, as this makes it possible to rule out common false positive findings. For example, it could potentially turn out that an important finding could be better explained by a match with a common contaminant such as human keratin or trypsin [60]. For instance, a standard list of contaminants can be found in the common Repository of Adventitious Proteins (cRAP—http://www.thegpm.org/crap), provided by the GPM team.

The gene and protein sequence databases that identification depends on are constantly evolving and improving [46]. This means that reprocessing a proteomics dataset with an updated version of the gene or protein database can result in improved findings. This is particularly true for poorly annotated species. In addition, updating a database to include known isoforms and/or mutations will provide a different view of the dataset.

Analogously, the software used to process proteomics data is also constantly improving, either by the further development of existing algorithms or by the establishment of new analysis approaches. The use of up-to-date techniques for the reprocessing of older datasets allows valuable information to be extracted from the acquired data without the need to repeat the experiment. This is particularly important for data from valuable or unique samples, where it ensures that as much information as possible can be obtained from these samples.

It should be noted that some of the existing proteomics databases, most notably GPMDB [5] and PeptideAtlas [6], routinely reprocess their data using dedicated bioinformatics tools and pipelines. GPMDB makes use of the X!Tandem search engine [61], whereas PeptideAtlas employs the Trans Proteomic Pipeline [62]. The data reprocessed by PeptideAtlas is organized into different builds, each including data from a single proteome (e.g. human) or subproteome (e.g. human plasma). Each build is generated based on the raw MS/MS spectra submitted to PeptideAtlas over the years, or from data deposited in other public repositories, for example PRIDE. In addition to human, many species now have specific PeptideAtlas builds, including, for example *Candida albicans* [63] and horse [64], among many others.

The GPMDB pipeline reprocesses the MS/MS data provided by users or raw data stored in other repositories, such as those from ProteomeXchange. Till the end of 2014, some of the reprocessed datasets were highlighted on a weekly basis on the GPM website (http://www.thegpm.org/news.html).

Both resources, PeptideAtlas and GPMDB, are also joining efforts in the context of the Chromosome-based (C-) and Biology/Disease (B/D) Human Proteome Projects (HPP) [65, 66], together with neXtProt and the antibody-based resource Human Proteome Atlas [67]. This is a clear example of the utility of large-scale and centralized (re-)processing, as it can ensure consistent processing and thus comparable results. The C-HPP team provides regular updates on the status of

completion of the human proteome and on the enumeration of the so-called "missing" proteins, that is proteins that have never been reliably detected experimentally [68].

1.6 Reusing and reprocessing enables scientific discussion

Perhaps, the most common current use case for shared proteomics data is the evaluation of existing results, often as part of the manuscript review process. This can be achieved by inspecting the data as provided by the authors, or by reprocessing the raw data by mimicking the original processing and then assessing the reproducibility of the results. The evaluation can be carried out at two levels: at the level of the individual PSMs, or at the level of the entire dataset. An example of the former is the checking of spectrum annotation quality, for example for post-translationally modified peptides. This can, for example, be achieved via the use of visualization tools such as MS-viewer [69], Scaffold Viewer, Thermo MSF Viewer, Peptizer [70], ProteoIDViewer [26], or TOPPView [71], among others.

For validation at the dataset level, tools such as PRIDE Inspector [27] and PeptideShaker [72], can be used to inspect and reprocess the data, respectively. Note that PeptideShaker provides a direct connection to PRIDE datasets to enable their streamlined reprocessing. The need for visual and interactive solutions should be noted here, as this can dramatically improve the validation procedure compared to looking at static images or tables [73].

One of the most famous examples of data discussion, involving both visual inspection and reprocessing, is related to the proteomics investigations of *Tyrannosaurus rex* fossil bone samples. The initial publications by Asara et al. [74, 75] proved controversial in the proteomics community (see, e.g. [76, 77]). As a consequence, the authors decided to make their data publicly available (PRD000074 in PRIDE), such that other researchers could inspect and reprocess the data themselves. Among others, this resulted in Bern et al. concluding that the original data did not contain any *T. rex* proteins [78]. The debate remains to be definitively settled, but the spirited scientific discussion highlights the importance of making the underlying data for published work available so that all sides can scientifically and reasonably discuss the findings based on the same evidence.

Another example is a study by Bromenshenk et al., which claimed to have found a link between viral and fungal contamination and the ongoing honey bee colony collapse disorder [79], a study that sparked global public interest. However, after the authors shared the data with others (available on request only), it became clear that this too could be a false positive outcome due to the systematic misidentification of bee-derived spectra as viral or fungal sequences, due to searching against a protein sequence database that lacked all honey bee sequences [80–82]. This discussion too still continues; however, as the same dataset was recently used to illustrate the opinion that, in order to improve statistical power, researchers should remove irrelevant peptides from the database before searching [83]. Here again, the inspection and reprocessing of the original experimental data enabled a scientific discussion and made it possible to collectively improve the scientific output, and paved the way for new discoveries [84].

More recently, there is an ongoing debate about the two drafts of the human proteome published in *Nature* in 2014 [10, 11]. Both studies provided an exemplary precedent by sharing all generated data (available as datasets PXD000561 and PXD000865 in PRIDE). This has enabled the community to start a discussion about the reliability of the results, see for instance Ezkurdia et al. [85].

1.7 Data repurposing in proteogenomics studies

Finally, when *repurposing* public data, these data are considered in light of a question or a context that is entirely different from the original study. It should be noted that repurposing thus often involves reprocessing as well. One example is the reprocessing of proteomics datasets to improve genome annotation in so-called proteogenomics approaches. For example, Brosch et al. reprocessed shotgun proteomics data from PeptideAtlas to discover novel protein-coding genes and to improve gene annotation in the mouse genome [86]. At the time, they found alternatively spliced translations from 53 genes along with ten entirely novel protein-coding genes. Another example is provided by LNCipedia [87], a resource for human long noncoding RNAs. PRIDE-based reanalysis of human proteomics data has provided evidence that some long-noncoding RNAs in LNCipedia are potentially translated to proteins [87].

In another proteogenomics study, Ezkurdia et al. reprocessed public proteomics data available in GPMDB and PeptideAtlas to identify peptides covering 35% of the genes annotated by the GENCODE consortium for the human genome [88]. Among other findings, they found that 150 genes expressed multiple alternative protein isoforms. Additionally, in a second analogous study, they concluded that the human proteome was composed of around 19 000 protein-coding genes [89], a lower number by around 1000 genes than the canonical assumption. In a related recent third study, they also reused public proteomics data from the same resources to suggest that most genes had a single dominant isoform at the protein level [90].

Existing proteomics data can also be reused in proteogenomics approaches. In a recent study devoted to psoriasis [91], the generated data were integrated with public data available in PRIDE (dataset PRD000053), proteomics data from other studies, and gene expression data available in the GEO (Gene Expression Omnibus) database [92]. As a final example in this section, Zhu et al. employed public proteomics data to develop a tool that can identify differentially regulated splice variants [93].

Because of the massive amounts of publicly available data and their inherent heterogeneity, the chances of reliably detecting protein expression evidence is higher in such reprocessing and repurposing approaches. However, due to the unconventional sequence population of the databases in proteogenomics, and their often extensive size, the estimation of false positive rates by traditional approaches can be impaired [83,94]. In the near future, it is therefore expected that the creation of such sequence databases will be coupled to ribosome profiling data, to discern the exact start of translation of putative proteins [95]. Indeed, tools such as ProteoFormer can already be used to generate proteomics-compatible protein sequence databases from such ribosome profiling data [96].

1.8 Reprocessing for better PTM localization and repurposing to find new PTMs

Finding and localizing PTMs are essential tasks in proteomics data analysis [97], and for this purpose multiple PTM localization scores have been developed [98], for example A-score [99], PTM score [100], MD-score [101], phosphoRS [102], and D-score [103]. Setting a threshold for these scores is, however, challenging, and solutions have only recently been established [104,105]. If such approaches were not applied in the original analysis, it is worth reprocessing the data, as this can dramatically improve the quality of the PTM annotation on the protein sequences. The reported location of specific PTMs can furthermore be refined using additional techniques, for example by considering the three-dimensional structure of the protein as indicated by Vandermarliere et al. [106].

It is also possible to repurpose existing datasets to look for PTMs that were not considered in the original analysis, for example via mass-tolerant database searches [107]. This task is made difficult by the substoichiometric nature of modified proteins, thus usually requiring experimental enrichment techniques to enable detection [108–110]. It is therefore often not straightforward to simply reprocess a dataset to find such modifications, but here again, the large amount of public data increases the probability to uncover modified peptides. Successful studies have therefore used enriched phosphoproteomics datasets to find peptides with unusual modifications that had a high probability of being co-enriched. Matic et al. [111] reanalyzed a mouse dataset to identify a total of 88 mono–ADP-ribosylation sites in 79 different proteins, with eight sites found modified also by ribose phosphate, a modification derived from ADP-ribose. In the reanalysis of another mouse dataset, Hahne and Küster [112] discovered an O-GlcNAc-6-phosphate modification on 23 peptides corresponding to 11 proteins.

1.9 Toward quantitative, across-source reprocessing

At the moment few repositories contain quantitative proteomics data, though it is possible to include quantitative information in data submissions to proteomics resources such as PRIDE. However, it is not yet possible to visualize and inspect this information properly due to a lack of suitable tools. Such tool development will most likely hinge on more widespread adoption of the PSI standards for quantitative information, namely mzQuantML and mzTab.

There are, however, several protein expression databases, most notably MOPED [113] and PaxDb [114], which can be used to extract information about the expression levels of individual proteins. Both resources routinely make use of publicly available data in PRIDE and PeptideAtlas, among others. In PaxDb, identification data from filtered datasets are first mapped onto a common namespace, and quantification values are then derived after reprocessing with a standardized spectral counting pipeline. PaxDb is a meta-resource in which protein expression is estimated across a number of species (more than 50 at the time of writing), and recently even across cell lines [115]. MOPED presents a multiomics resource for human and model organisms, including at present gene, protein, and pathway expression information [116].

Another resource to highlight in this context is ProteomicsDB, which provides abundance estimates according to the label-free intensity-based iBAQ method [117]. ProteomicsDB is one of the main outputs of the draft human proteome by Wilhelm et al. [10], and represents a nice example of data reprocessing. For their analysis, they combined their own generated experimental results with publicly available data. In fact, around 40% of the data used to generate this draft of the human proteome were obtained from public resources such as PRIDE, MassIVE, and PeptideAtlas (see Supporting Information Table 1 in [10] for the complete list). However, new datasets are reprocessed regularly and incorporated into ProteomicsDB, including also RNAseq data and phospho-proteomics experiments.

The ability to compare protein abundances among datasets across public repositories would provide the possibility to virtually create new quantitative experiments, paving the way for in silico proteomics (Fig. 4). However, accurate absolute quantification of peptides and proteins in datasets is made challenging by the need for internal standards. Relative quantification is impaired by the heterogeneity of the data present in repositories, and their often suboptimal annotation [118]. It is therefore worth mentioning that in-depth annotation of the experimental design is essential in order to correctly interpret quantitative information from public proteomics data.

The development of bioinformatics and statistics tools for the robust and accurate interpretation of such heterogeneous data will allow the setup of creative designs where datasets from different sources can be repurposed and compared. This could, for example, enable the in silico comparison of large patient cohorts based on the aggregation of multiple smaller cohorts. Such approaches can, however, be made impossible if significant sample variability is introduced during sample extraction and preparation, for example when PTM enrichment is conducted.

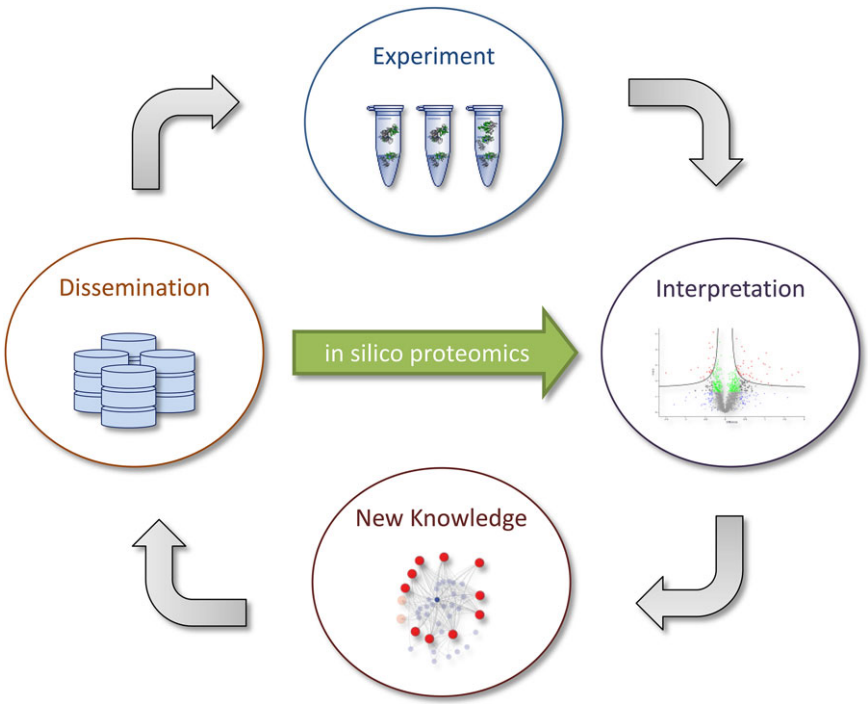

Figure 4. The rapidly growing amount of publicly available proteomics data opens up the opportunity for in silico proteomics, that is using bioinformatics to test hypotheses directly through the available data, instead of going via the generation of new experimental data.

2 Discussion

The growing amount of publicly available proteomics data has already been put to great use, both as a means to validate published results and to generate new knowledge via reprocessing and repurposing. With the achievement of the required milestones for data sharing (i.e. data standards, user-friendly software, and public databases) and the push toward public data from journals and funders, the amount of shared data will only continue to grow rapidly.

There are, however, still some limiting factors that ought to be addressed. The first of these is the need for proper annotation, especially regarding experimental design. Indeed, even though minimal reporting standards have been developed for proteomics data (the so-called MIAPE (Minimum Information About a Proteomics Experiment) guidelines [119]), there remains a gap between what is reported and what ought to be reported. While it is possible to attempt to infer the missing information as, for example, done by the pride-asap pipeline [120], this is often far from straightforward and may result in incorrect assumptions. The only real solution is to make it easier for submitters to provide additional information, or to annotate this information automatically in the standard file formats. This work has already started, notable in LIMS systems such as MASPECTRAS [121], ms_lims [122], and Proteios [123], but it will still take some time before it is straightforward to capture all the desired information.

A related challenge is the provision of easy access to public data while catering to the need for visual and interactive analysis [124]. There are already several tools, including PRIDE Inspector [27] and PeptideShaker [72] that support this concept, but more are certainly needed. This is especially true for tools that link and display information from multiple resources in a meaningful way. Easy access for developers is also vital, for example, via systems such as BioMart [125], or more recently, via web services [48, 49].

It is also crucial that scientists get credit for sharing their data, especially when these data are reused in new contexts. The ProteomeXchange accession number should therefore always be used when a dataset is reused and the corresponding publication(s) should be cited. ProteomeXchange also issues a DOI (Digital Object Identifier) to "Complete" submissions (i.e. submissions where data are provided in accordance with public standards, so they are easier to access and reuse), as a way to improve dataset tracking and to give credit to authors [126]. It will also be useful if resources provide dataset access statistics, given the current trend of putting increased value on so-called "altmetrics" methods [127] to capture the impact of scientists' work.

Moving forward, data-independent acquisition approaches such as MS^E and SWATH-MS will become more popular in the field [128]. And even though some public data for these approaches already exist, it is expected that public deposition of this type of data will significantly increase in the coming years. In fact, there are already dedicated resources in place such as SWATH-Atlas (http://www.swathatlas.org) that can be used for planning SWATH experiments, for depositing experiments, and for exploring the results of deposited datasets. A particular characteristic of SWATH-MS data is that, once generated, these can potentially be reanalyzed multiple times using different spectral libraries, which are set to improve over time as public data increase. These developments open

up numerous novel possibilities for the reanalysis of public proteomics data.

Another very interesting upcoming opportunity is the reprocessing of datasets generated in "multi-omics" studies. At present, these type of studies pose a challenge for both traditional repositories, which are most often field-specific (e.g. proteomics, genomics, or transcriptomics), as well as for researchers, given that at present it is not straightforward to link public data coming from paired samples located in different resources (e.g. MS proteomics and RNAseq data obtained in the same study). There are, however, ongoing efforts to link different studies performed on the same sample [129]. Over time, the existence of personalized sequence databases (from DNA exome sequencing), or the existence of public data containing both gene and protein expression data for a given sample will become commonplace, opening up yet more opportunities for data analysts.

Many of the approaches highlighted in this review can also be exploited in the metabolomics field, where the first stable data repositories and data standards are now starting to be established [130]. For example, spectral libraries have been used for the analysis of MS metabolomics data already, many years before the same approach was applied to the proteomics field, and we can expect to see more examples of techniques adopted from related fields in the future.

Finally, the need for customizable, large-scale reprocessing systems should be highlighted. Such capabilities currently remain limited to a couple of dedicated proteomics bioinformatics groups. However, as the data have been generated by the community, and thus belong to the community as a whole, large-scale reprocessing should also be made available to the general community. Only then can we start to realize the full potential of the publicly shared proteomics data.

K.V. and L.M. acknowledge support from Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), Ghent University grant BOF12/GOA/014, and the IWT SBO grant "INSPECTOR" (120025). A.C. is supported by EMBL core funding. J.A.V. is supported by the Wellcome Trust (grant number WT101477MA) and the BBSRC (grant number BB/L024225/1). H.B. and H.R. are supported by Bergen Forskningsstifelse, and H.R. is further supported by Novo Nordisk Fonden and Western Norway Regional Health Authority. F.B. is supported by the Kristian Gerhard Jebsen foundation.

The authors have declared no conflict of interest.

3 References

[1] Editors, Democratizing proteomics data. *Nat. Biotechnol.* 2007, *25*, 262.

[2] Editors, Thou shalt share your data. *Nat. Methods* 2008, *5*, 209–209.

[3] Burlingame, A. L., Carr, S. A., Bradshaw, R. A., Chalkley, R. J., On credibility, clarity and compliance. *Mol. Cell Proteomics* 2015, *7*, 1731–1733.

[4] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537–3545.

[5] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2004, *3*, 1234–1242.

[6] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I. et al., The PeptideAtlas project. *Nucleic Acids Res.* 2006, *34*, D655–D658.

[7] Farrah, T., Deutsch, E. W., Kreisberg, R., Sun, Z. et al., PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* 2012, *12*, 1170–1175.

[8] Sharma, V., Eckels, J., Taylor, G. K., Shulman, N. J. et al., Panorama: a targeted proteomics knowledge base. *J. Proteome Res.* 2014, *13*, 4205–4210.

[9] Reddy, T. B., Riley, R., Wymore, F., Montgomery, P. et al., TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res.* 2009, *37*, D499–D508.

[10] Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A. et al., Mass-spectrometry-based draft of the human proteome. *Nature* 2014, *509*, 582–587.

[11] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S. et al., A draft map of the human proteome. *Nature* 2014, *509*, 575–581.

[12] Hirsch-Hoffmann, M., Gruissem, W., Baerenfaller, K., pep2pro: the high-throughput proteomics data processing, analysis, and visualization tool. *Front Plant Sci.* 2012, *3*, 123.

[13] Guldbrandsen, A., Vethe, H., Farag, Y., Oveland, E. et al., In-depth characterization of the cerebrospinal fluid proteome displayed through the CSF Proteome Resource (CSF-PR). *Mol. Cell Proteomics* 2014, *11*, 3152–3163.

[14] Colaert, N., Maddelein, D., Impens, F., Van Damme, P. et al., The Online Protein Processing Resource (TOPPR): a database and analysis platform for protein processing events. *Nucleic Acids Res.* 2013, *41*, D333–D337.

[15] Lange, P. F., Overall, C. M., TopFIND, a knowledgebase linking protein termini with function. *Nat. Methods* 2011, *8*, 703–704.

[16] Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H., Vizcaino, J. A., Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 2014, *15*, 930–949.

[17] Martens, L., Nesvizhskii, A. I., Hermjakob, H., Adamski, M. et al., Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* 2005, *5*, 3501–3505.

[18] Martens, L., Chambers, M., Sturm, M., Kessner, D. et al., mzML—a community standard for mass spectrometry data. *Mol. Cell Proteomics* 2011, *10*, R110 000133.

[19] Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O. et al., The mzIdentML data standard for mass

spectrometry-based proteomics results. *Mol. Cell Proteomics* 2012, *11*, M111 014381.

[20] Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M. et al., The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell Proteomics* 2014, *13*, 2765–2775.

[21] Walzer, M., Qi, D., Mayer, G., Uszkoreit, J. et al., The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol. Cell Proteomics* 2013, *12*, 2332–2340.

[22] Deutsch, E. W., Chambers, M., Neumann, S., Levander, F. et al., TraML: a standard format for exchange of selected reaction monitoring transition lists. *Mol. Cell Proteomics* 2011, *11*, R111.015040.

[23] Chambers, M. C., Maclean, B., Burke, R., Amodei, D. et al., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 2012, *30*, 918–920.

[24] Cote, R. G., Griss, J., Dianes, J. A., Wang, R. et al., The PRoteomics IDEntification (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, *11*, 1682–1689.

[25] Barsnes, H., Vizcaíno, J. A., Eidhammer, I., Martens, L., PRIDE Converter: making proteomics data-sharing easy. *Nat. Biotechnol.* 2009, *27*, 598–599.

[26] Ghali, F., Krishna, R., Lukasse, P., Martinez-Bartolome, S. et al., Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. *Mol. Cell Proteomics* 2013, *12*, 3026–3035.

[27] Wang, R., Fabregat, A., Rios, D., Ovelleiro, D. et al., PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat. Biotechnol.* 2012, *30*, 135–137.

[28] Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A. et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32*, 223–226.

[29] Barsnes, H., Martens, L., Crowdsourcing in proteomics: public resources lead to better experiments. *Amino Acids* 2013, *44*, 1129–1137.

[30] Rung, J., Brazma, A., Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* 2013, *14*, 89–99.

[31] UniProt Consortium, The universal protein resource (UniProt). *Nucleic Acids Res.* 2008, *36*, D190–D195.

[32] Lane, L., Argoud-Puy, G., Britan, A., Cusin, I. et al., neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.* 2012, *40*, D76–D83.

[33] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M. et al., PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 2015, *43*, D512–D520.

[34] Fan, J., Mohareb, F., Bond, N. J., Lilley, K. S., Bessant, C., MRMaid 2.0: mining PRIDE for evidence-based SRM transitions. *OMICS* 2012, *16*, 483–488.

[35] Mohammed, Y., Domanski, D., Jackson, A. M., Smith, D. S. et al., PeptidePicker: a scientific workflow with web interface for selecting appropriate peptides for targeted proteomics experiments. *J. Proteomics* 2014, *106*, 151–161.

[36] Mohammed, Y., Borchers, C. H., An extensive library of surrogate peptides for all human proteins. *J. Proteomics* 2015, pii: S1874-3919(15)30079-8. doi: 10.1016/j.jprot.2015.07.025. [Epub ahead of print].

[37] Vandermarliere, E., Mueller, M., Martens, L., Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom. Rev.* 2013, *32*, 453–465.

[38] Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A., Zahedi, R. P., Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* 2012, *75*, 1454–1462.

[39] Fannes, T., Vandermarliere, E., Schietgat, L., Degroeve, S. et al., Predicting tryptic cleavage from proteomics data using decision tree ensembles. *J. Proteome Res.* 2013, *12*, 2253–2259.

[40] Moruz, L., Pichler, P., Stranzl, T., Mechtler, K., Kall, L., Optimized nonlinear gradients for reversed-phase liquid chromatography in shotgun proteomics. *Anal. Chem.* 2013, *85*, 7777–7785.

[41] Barsnes, H., Eidhammer, I., Martens, L., A global analysis of peptide fragmentation variability. *Proteomics* 2011, *11*, 1181–118.

[42] Degroeve, S., Martens, L., MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* 2013, *29*, 3199–3203.

[43] Klie, S., Martens, L., Vizcaino, J. A., Cote, R. et al., Analyzing large-scale proteomics projects with latent semantic indexing. *J. Proteome Res.* 2008, *7*, 182–191.

[44] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W. et al., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226–3245.

[45] Mueller, M., Vizcaino, J. A., Jones, P., Cote, R. et al., Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. *Proteomics* 2008, *8*, 1138–1148.

[46] Griss, J., Cote, R. G., Gerner, C., Hermjakob, H., Vizcaino, J. A., Published and perished? The influence of the searched protein database on the long-term storage of proteomics data. *Mol. Cell Proteomics* 2011, *10*, M111 008490.

[47] Kasprzyk, A., BioMart: driving a paradigm change in biological data management. *Database* 2011, *2011*, bar049.

[48] Reisinger, F., Del-Toro, N., Ternent, T., Hermjakob, H., Vizcaino, J. A., Introducing the PRIDE Archive RESTful web services. *Nucleic Acids Res.* 2015, *43*, W599–604.

[49] Fenyo, D., Beavis, R. C., The GPMDB REST interface. *Bioinformatics* 2015, *31*, 2056–2058.

[50] Griss, J., Perez-Riverol, Y., Hermjakob, H., Vizcaino, J. A., Identifying novel biomarkers through data mining—a realistic scenario? *Proteomics Clin. Appl.* 2014, *9*, 437–443.

[51] 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.* 2006, *78*, 5678–5684.

[52] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. et al., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007, *7*(5), 655–667.

[53] Tabb, D. L., Thompson, M. R., Khalsa-Moyers, G., VerBerkmoes, N. C., McDonald, W. H., MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J. Am. Soc. Mass Spectrom.* 2005, *16*(8), 1250–1261.

[54] Flikka, K., Meukens, J., Helsens, K., Vandekerckhove, J. et al., Implementation and application of a versatile clustering tool for tandem mass spectrometry data. *Proteomics* 2007, *7*(18), 3245–3258.

[55] Falkner, J. A., Falkner, J. W., Yocum, A. K., Andrews, P. C., A spectral clustering approach to MS/MS identification of post-translational modifications. *J. Proteome Res.* 2008, *7*(11), 4614–4622.

[56] Frank, A. M., Bandeira, N., Shen, Z., Tanner, S. et al., Clustering millions of tandem mass spectra. *J. Proteome Res.* 2008, *7*(1), 113–122.

[57] Frank, A. M., Monroe, M. E., Shah, A. R., Carver, J. J. et al., Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* 2011, *8*(7), 587–591.

[58] Griss, J., Foster, J. M., Hermjakob, H., Vizcaino, J. A., PRIDE Cluster: building a consensus of proteomics data. *Nat. Methods* 2013, *10*(2), 95–96.

[59] Lam, H., Spectral archives: a vision for future proteomics data repositories. *Nat. Methods* 2011, *8*(7), 546–548.

[60] Ghesquiere, B., Helsens, K., Vandekerckhove, J., Gevaert, K., A stringent approach to improve the quality of nitrotyrosine peptide identifications. *Proteomics* 2011, *11*(6), 1094–1098.

[61] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*(9), 1466–1467.

[62] Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J. et al., Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.* 2015, *9*, 745–754.

[63] Vialas, V., Sun, Z., Loureiro y Penha, C. V., Carrascal, M. et al., A *Candida albicans* PeptideAtlas. *J. Proteomics* 2014, *97*, 62–68.

[64] Bundgaard, L., Jacobsen, S., Sorensen, M. A., Sun, Z. et al., The Equine Peptideatlas: a resource for developing proteomics-based veterinary research. *Proteomics* 2014, *14*(6), 763–773.

[65] Aebersold, R., Bader, G. D., Edwards, A. M., van Eyk, J. E. et al., The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* 2013, *12*(1), 23–27.

[66] Paik, Y. K., Jeong, S. K., Omenn, G. S., Uhlen, M. et al., The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* 2012, *30*(3), 221–223.

[67] Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C. et al., Proteomics. Tissue-based map of the human proteome. *Science* 2015, *347*(6220), 1260419.

[68] Horvatovich, P., Lundberg, E. K., Chen, Y. J., Sung, T. Y. et al., A quest for missing proteins: update 2015 on Chromosome-Centric Human Proteome Project. *J. Proteome Res.* 2015, *14*, 3415–3431.

[69] Baker, P. R., Chalkley, R. J., MS-viewer: a web-based spectral viewer for proteomics results. *Mol. Cell Proteomics* 2014, *13*(5), 1392–1396.

[70] Helsens, K., Timmerman, E., Vandekerckhove, J., Gevaert, K., Martens, L., Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol. Cell Proteomics* 2008, *7*(12), 2364–2372.

[71] Sturm, M., Kohlbacher, O., TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.* 2009, *8*(7), 3760–3763.

[72] Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E. et al., PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* 2015, *33*(1), 22–24.

[73] Farag, Y., Berven, F. S., Jonassen, I., Petersen, K., Barsnes, H., Distributed and interactive visual analysis of omics data. *J. Proteomics* 2015, pii: S1874-3919(15)30030-0. doi: 10.1016/j.jprot.2015.05.029. [Epub ahead of print].

[74] Asara, J. M., Schweitzer, M. H., Freimark, L. M., Phillips, M., Cantley, L. C., Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science* 2007, *316*(5822), 280–285.

[75] Asara, J. M., Garavelli, J. S., Slatter, D. A., Schweitzer, M. H. et al., Interpreting sequences from mastodon and *T. rex*. *Science* 2007, *317*(5843), 1324–1325.

[76] Buckley, M., Walker, A., Ho, S. Y., Yang, Y. et al., Comment on "Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry". *Science* 2008, *319*(5859), 33.

[77] Pevzner, P. A., Kim, S., Ng, J., Comment on Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry". *Science* 2008, *321*(5892), 1040.

[78] Bern, M., Phinney, B. S., Goldberg, D., Reanalysis of *Tyrannosaurus rex* mass spectra. *J. Proteome Res.* 2009, *8*(9), 4328–4332.

[79] Bromenshenk, J. J., Henderson, C. B., Wick, C. H., Stanford, M. F. et al., Iridovirus and microsporidian linked to honey bee colony decline. *PLoS One* 2010, *5*(10), e13181.

[80] Knudsen, G. M., Chalkley, R. J., The effect of using an inappropriate protein database for proteomic data analysis. *PLoS One* 2011, *6*(6), e20873.

[81] Foster, L. J., Bromenshenk et al., (*PLoS One*, 2011, *5*(10):e13181) have claimed to have found peptides from an invertebrate iridovirus in bees. *Mol. Cell Proteomics* 2012, *11*(1), A110 0063871.

[82] Foster, L. J., Interpretation of data underlying the link between colony collapse disorder (CCD) and an invertebrate iridescent virus. *Mol. Cell Proteomics* 2011, *10*(3), M110 006387.

[83] Noble, W. S., Mass spectrometrists should search only for peptides they care about. *Nat. Methods* 2015, *12*(7), 605–608.

[84] Daughenbaugh, K. F., Martin, M., Brutscher, L. M., Cavigli, I. et al., Honey bee infecting Lake Sinai viruses. *Viruses* 2015, *7*(6), 3285–3309.

[85] Ezkurdia, I., Vazquez, J., Valencia, A., Tress, M., Analyzing the first drafts of the human proteome. *J. Proteome Res.* 2014, *13*, 3854–3855.

[86] Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O. et al., Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res.* 2011, *21*(5), 756–767.

[87] Volders, P. J., Verheggen, K., Menschaert, G., Vandepoele, K. et al., An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* 2015, *43*(8), 4363–4364.

[88] Ezkurdia, I., del Pozo, A., Frankish, A., Rodriguez, J. M. et al., Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol. Biol. Evol.* 2012, *29*(9), 2265–2283.

[89] Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A. et al., Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* 2014, *23*(22), 5866–5878.

[90] Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M. et al., Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS Comput. Biol.* 2015, *11*(6), e1004325.

[91] Swindell, W. R., Remmer, H. A., Sarkar, M. K., Xing, X. et al., Proteogenomic analysis of psoriasis reveals discordant and concordant changes in mRNA and protein abundance. *Genome Med.* 2015, *7*(1), 86.

[92] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C. et al., NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013, *41*(Database issue), D991–D995.

[93] Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R. M. et al., SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol. Cell Proteomics* 2014, *13*(6), 1552–1562.

[94] Nesvizhskii, A. I., Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 2014, *11*(11), 1114–1125.

[95] Koch, A., Gawron, D., Steyaert, S., Ndah, E. et al., A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* 2014, *14*(23–24), 2688–2698.

[96] Crappe, J., Ndah, E., Koch, A., Steyaert, S. et al., PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* 2015, *43*(5), e29.

[97] Vaudel, M., Sickmann, A., Martens, L., Current methods for global proteome identification. *Expert Rev. Proteomics* 2012, *9*(5), 519–532.

[98] Chalkley, R. J., Clauser, K. R., Modification site localization scoring: strategies and performance. *Mol. Cell Proteomics* 2012, *11*(5), 3–14.

[99] 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. *Nat. Biotechnol.* 2006, *24*(10), 1285–1292.

[100] Olsen, J. V., Blagoev, B., Gnad, F., Macek, B. et al., Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 2006, *127*(3), 635–648.

[101] Savitski, M. M., Lemeer, S., Boesche, M., Lang, M. et al., Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell Proteomics* 2011, *10*(2), M110 003830.

[102] Taus, T., Kocher, T., Pichler, P., Paschke, C. et al., Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* 2011, *10*(12), 5354–5362.

[103] Vaudel, M., Breiter, D., Beck, F., Rahnenfuhrer, J. et al., D-score: a search engine independent MD-score. *Proteomics* 2013, *13*(6), 1036–1041.

[104] Fermin, D., Walmsley, S. J., Gingras, A. C., Choi, H., Nesvizhskii, A. I., LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol. Cell Proteomics* 2013, *12*(11), 3409–3419.

[105] Fermin, D., Avtonomov, D., Choi, H., Nesvizhskii, A. I., LuciPHOr2: site localization of generic post-translational modifications from tandem mass spectrometry data. *Bioinformatics* 2015, *31*(7), 1141–1143.

[106] Vandermarliere, E., Martens, L., Protein structure as a means to triage proposed PTM sites. *Proteomics* 2013, *13*(6), 1028–1035.

[107] Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B. et al., A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* 2015, *33*(7), 743–749.

[108] Loroch, S., Dickhut, C., Zahedi, R. P., Sickmann, A., Phosphoproteomics—more than meets the eye. *Electrophoresis* 2013, *34*(11), 1483–1492.

[109] Olsen, J. V., Mann, M., Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell Proteomics* 2013, *12*(12), 3444–3452.

[110] Solari, F. A., Dell'Aica, M., Sickmann, A., Zahedi, R. P., Why phosphoproteomics is still a challenge. *Mol. Biosyst.* 2015, *11*, 1487–1493.

[111] Matic, I., Ahel, I., Hay, R. T., Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites. *Nat. Methods* 2012, *9*(8), 771–772.

[112] Hahne, H., Kuster, B., Discovery of O-GlcNAc-6-phosphate modified proteins in large-scale phosphoproteomics data. *Mol. Cell Proteomics* 2012, *11*(10), 1063–1069.

[113] Kolker, E., Higdon, R., Haynes, W., Welch, D. et al., MOPED: model organism protein expression database. *Nucleic Acids Res.* 2012, *40*(Database issue), D1093–D1099.

[114] Wang, M., Weiss, M., Simonovic, M., Haertinger, G. et al., PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell Proteomics* 2012, *11*(8), 492–500.

[115] Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., von Mering, C., Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 2015, *15*, 3163–3168.

[116] Montague, E., Janko, I., Stanberry, L., Lee, E. et al., Beyond protein expression, MOPED goes multi-omics. *Nucleic Acids Res.* 2015, *43*(Database issue), D1145–D1151.

[117] Schwanhausser, B., Busse, D., Li, N., Dittmar, G. et al., Global quantification of mammalian gene expression control. *Nature* 2011, *473*(7347), 337–342.

[118] Gonnelli, G., Hulstaert, N., Degroeve, S., Martens, L., Towards a human proteomics atlas. *Anal. Bioanal. Chem.* 2012, *404*(4), 1069–1077.

[119] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A. et al., The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 2007, *25*(8), 887–893.

[120] Hulstaert, N., Reisinger, F., Rameseder, J., Barsnes, H. et al., Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. *J. Proteomics* 2013, *95*, 89–92.

[121] Hartler, J., Thallinger, G. G., Stocker, G., Sturn, A. et al., MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinform.* 2007, *8*, 197.

[122] Helsens, K., Colaert, N., Barsnes, H., Muth, T. et al., ms_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. *Proteomics* 2010, *10*(6), 1261–1264.

[123] Hakkinen, J., Vincic, G., Mansson, O., Warell, K., Levander, F., The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.* 2009, *8*(6), 3037–3043.

[124] Oveland, E., Muth, T., Rapp, E., Martens, L. et al., Viewing the proteome: how to visualize proteomics data? *Proteomics* 2014, *15*, 1341–1355.

[125] Smedley, D., Haider, S., Ballester, B., Holland, R. et al., BioMart—biological queries made easy. *BMC Genomics* 2009, *10*, 22.

[126] Credit where credit is overdue. *Nat. Biotechnol.* 2009, *27*(7), 579.

[127] Priem, J., Scholarship: beyond the paper. *Nature* 2013, *495*(7442), 437–440.

[128] Law, K. P., Lim, Y. P., Recent advances in mass spectrometry: data independent analysis and hyper reaction monitoring. *Expert Rev. Proteomics* 2013, *10*(6), 551–566.

[129] Faulconbridge, A., Burdett, T., Brandizi, M., Gostev, M. et al., Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Res.* 2014, *42*(Database issue), D50–D52.

[130] Salek, R. M., Haug, K., Conesa, P., Hastings, J. et al., The MetaboLights repository: curation challenges in metabolomics. *Database* 2013, *2013*, bat029.

