DOI 10.1002/pmic.201400302 *Proteomics* 2015, *15*, 930–949

REVIEW

Making proteomics data accessible and reusable: Current state of proteomics databases and repositories

Yasset Perez-Riverol, Emanuele Alpi, Rui Wang, Henning Hermjakob and Juan Antonio Vizcaíno

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Received: July 1, 2014
Revised: August 6, 2014
Accepted: August 22, 2014

Compared to other data-intensive disciplines such as genomics, public deposition and storage of MS-based proteomics, data are still less developed due to, among other reasons, the inherent complexity of the data and the variety of data types and experimental workflows. In order to address this need, several public repositories for MS proteomics experiments have been developed, each with different purposes in mind. The most established resources are the Global Proteome Machine Database (GPMDB), PeptideAtlas, and the PRIDE database. Additionally, there are other useful (in many cases recently developed) resources such as ProteomicsDB, Mass Spectrometry Interactive Virtual Environment (MassIVE), Chorus, MaxQB, PeptideAtlas SRM Experiment Library (PASSEL), Model Organism Protein Expression Database (MOPED), and the Human Proteinpedia. In addition, the ProteomeXchange consortium has been recently developed to enable better integration of public repositories and the coordinated sharing of proteomics information, maximizing its benefit to the scientific community. Here, we will review each of the major proteomics resources independently and some tools that enable the integration, mining and reuse of the data. We will also discuss some of the major challenges and current pitfalls in the integration and sharing of the data.

Keywords:
Bioinformatics / Databases / MS / Repositories

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

In the age of systems biology and data integration, proteomics data represent a crucial component to understand the "whole picture" of life. Proteomics technologies—particularly MS-based protein identification and quantification approaches—have matured immensely through cumulative advances in high-throughput analytical methodologies [1–5], sample preparation [6], improved instrumentation [7], and the availability of protein sequence databases [8] and computational analysis tools [1, 9]. Therefore, with the development of more powerful and sensitive analytical methods and instrumentation, the identification and quantification of a high proportion of the expressed proteins in a given condition is now achievable in an average experiment [10, 11]. In parallel, as a result,

Correspondence: Dr. Juan Antonio Vizcaíno, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
E-mail: juan@ebi.ac.uk
Fax: +44 1223 494 484

Abbreviations: **C-HPP**, chromosome-based Human Proteome Project; **COPaKB**, Cardiac Organellar Protein Atlas Knowledgebase; **FDR**, false discovery rate; **HPM**, human proteome map; **HPRD**, human protein reference database; **MassIVE**, Mass Spectrometry Interactive Virtual Environment; **MOPED**, Model Organism Protein Expression Database; **NCBI**, National Center for Biotechnology Information; **PASSEL**, PeptideAtlas SRM Experiment Library; **PSM**, peptide spectrum match; **PX**, ProteomeXchange; **TIQAM**, targeted identification for quantitative analysis by MRM; **TPP**, trans-proteomic pipeline

Colour Online: See the article online to view Figs. 1–3 in colour.

the size of data produced in proteomics laboratories has increased by several orders of magnitude [12].

Compared to other data-intensive fields such as genomics, deposition and storage of original proteomics, data in public resources have been less common [13]. This is regrettable since proteome studies are usually more complex than its counterpart genomics ones. In fact, data interpretation in proteomics can be considerably more complex than in genomics due to the wide variety of analytical approaches [14, 15], bioinformatics tools and pipelines [16, 17], and the related statistical analysis [18, 19]. However, thanks to the guidelines promoted by several scientific journals and funding agencies [20], there is a growing consensus in the community about the need for the public dissemination of proteomics data, which is already facilitating the assessment, reuse, comparative analyses, and extraction of new findings from published data [13, 21].

The complexity of proteomics data is heightened by alternative splicing, PTMs, and protein degradation events, and is further amplified by the interconnectivity of proteins into complexes and signaling networks that are highly divergent in time and space [1]. In order to address this complexity, new analytical and bioinformatics methodologies are developed every year [22, 23], which complicate the data standardization and deposition. Additionally, the audience interested in proteomics data is very heterogeneous. It includes, biologists elucidating the mechanisms of regulation of specific proteins, MS researchers improving the current analytical methods, or computational biologists developing new software tools for the analysis and interpretation of the data [24].

Data sharing in proteomics requires substantial investment and infrastructure. Several public repositories have been developed, each with different purposes in mind. Well-established databases for proteomics data are the Global Proteome Machine Database (GPMDB) [25], PeptideAtlas [26], and the PRIDE database [27]. Additionally, at present there are other resources (many of them recently developed) such as ProteomicsDB [28], MassIVE (Mass Spectrometry Interactive Virtual Environment), Chorus, MaxQB [29], PASSEL (PeptideAtlas SRM Experiment Library) [30], MOPED (Model Organism Protein Expression Database) [31], PaxDb [32], Human Proteinpedia [33], and the human proteome map (HPM) [34]. Furthermore, there are several more specialized resources that will only be cited briefly in this review. It is important to mention here that no single proteomics data resource will be ideally suited to all possible use cases and all potential users. Regrettably, two widely used resources were discontinued due to lack of funding: Peptidome [35] and Tranche [36]. This had a negative impact on the efforts promoting data sharing in the field, as it was perceived by the community that effort invested in data sharing was lost.

Recently, the ProteomeXchange (PX) consortium [37] has been formed to enable a better integration of public repositories, maximizing its benefits to the scientific community through the implementation of standardized submission and dissemination pipelines for proteomics information. By August 2014, PRIDE, PeptideAtlas, PASSEL, and MassIVE are the active members of the consortium.

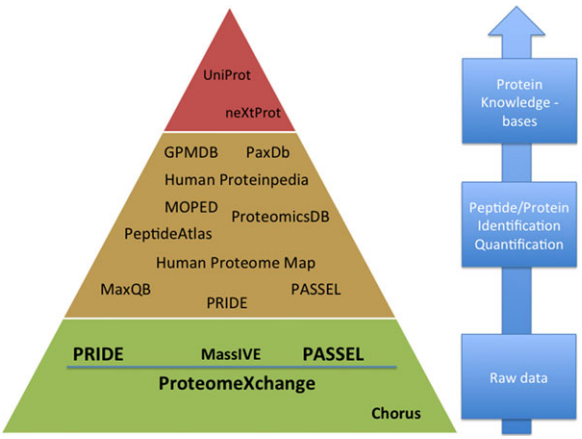

Figure 1. Hierarchy of proteomics data repositories and databases according to the different data types stored: raw MS data repositories, resources that store peptide/protein identification and quantification results, and protein knowledge bases. Some resources are duplicated in different levels because they can be included in more than one category.

The aim of this review is to provide an up-to-date overview of the current state of proteomics data repositories and databases, providing a solid starting point for those who want to perform data submission and/or data mining. There are a few comparable reviews available in the literature [24, 38–41], but there is a need for an update since this has been quite a dynamic field over the past few years. In this manuscript, we will not include a thorough review about protein knowledge bases, such as the Universal Protein resource (UniProt) [42] and neXtProt [43], but we will explain how MS proteomics information is made available in these resources.

2 Organization of proteomics repositories and databases

The information generated in a typical proteomics experiment can be organized in three different levels [44]: (i) raw data; (ii) processed results, including peptide/protein identification and quantification values; and (iii) the resulting biological conclusions. Technical and/or biological metadata can be provided for each level independently.

These three categories enable the classification of the existing MS proteomics repositories according to their level of specialization (Fig. 1). In our view, these three levels of information should be captured and properly annotated in public databases and repositories, ideally using data standards, when available. In fact, the development of proteomics resources is more feasible due the maturity of some data standard formats [45–47] and open source tools [9], which facilitate public data deposition.

Some of the first open formats developed included mzXML (for MS data) [48], pepXML, and protXML (for

peptide/protein identifications) that were developed as part of the trans-proteomic pipeline (TPP) [49]. In the context of the Proteomics Standards Initiative (PSI), several standard data formats have been developed over the last few years, which reflect the variety in data types within the field, and therefore those that can be supported by proteomics resources. The main formats developed have been mzML (for MS data) [45], mzIdentML (for peptide and protein identification results) [46], mzQuantML (for capturing a detailed trace of each stage of quantitative analysis) [47], TraML (for representing input transitions in SRM approaches) [50], and the recent mzTab (for capturing a simpler summary of the final results) [51]. The aims and functionalities of the existing resources will be explored in detail in the following sections (Table 1).

3 Resources

3.1 The PX consortium

The PX consortium [37] (http://www.proteomexchange.org) was created to promote the collaboration and integration of major stakeholders in the domain of MS proteomics repositories comprising, among others, primary (PRIDE [27] and PASSEL [30]) and secondary resources (PeptideAtlas), proteomics researchers, and representatives from journals regularly publishing proteomics data. Recently, in June 2014, MassIVE joined the consortium. The aim of PX is to provide a common framework for the cooperation of proteomics resources by defining and implementing consistent, harmonized, user-friendly data deposition and dissemination procedures. In addition, another important goal is to enable and provide "mutual backup" if one of the resources has funding issues.

The consortium's members have agreed in providing a sufficient set of common experimental and technical metadata. This information is stored using the PX XML format [37]. Finally, all the submitted datasets get a unique and universal identifier (PXD identifier).

3.1.1 Data submission and format support

By August 2014, two major workflows are fully supported in PX: MS/MS and SRM approaches. In the first stable implementation of the data workflow, PRIDE acts as the initial submission point for MS/MS data, whereas PASSEL has the equivalent role for SRM data. Both workflows will be explained in detail below. At the moment of writing, MassIVE has just joined PX aiming to have an equivalent role to PRIDE.

There are two different PX MS/MS submission modes: "Complete" and "Partial." To perform a "complete" submission means that after all the files have been submitted, it is possible for the receiving repository to connect directly the processed identification results with the mass spectra. This can be achieved if the processed identification results are available in a format supported by the receiving repository (e.g., mzIdentML) and if peak list files are included in the submission. "Complete" submissions get a Digital Object Identifier (DOI) to facilitate its traceability.

On the other hand, after performing a "partial" submission the connection between the spectra and the identification results cannot be done in a straightforward way. In this case, the processed results are not available in a supported format by the receiving repository and the corresponding search engine output files (in heterogeneous formats) are made available for download. For both types of submissions, metadata and raw data are always stored for each dataset.

Although "partial" submissions are searchable by their metadata, peptide and protein identifications cannot be captured by the receiving repository, which decreases the ability of reviewers to check the data and can make data reuse by third parties challenging. For instance, "partial" submissions do not qualify for the requirements on spectra annotation from the journal MCP (*Molecular and Cellular Proteomics*—http://www.mcponline.org/site/misc/ParisReport_Final.xhtml).

Finally, it needs to be highlighted that all PX members support private review of the data during the manuscript review process. The submitted data remains private before manuscript publication and login details are provided to facilitate access for reviewers and journal editors during the manuscript review process.

3.1.2 Data mining and visualization

ProteomeCentral (http://proteomecentral.proteomexchange.org) is the centralized portal for accessing all PX datasets, independently from the original resource where data were stored. ProteomeCentral provides the ability to search metadata associated with datasets in the participating repositories (PRIDE and MassIVE for MS/MS data, PASSEL for SRM data, or reprocessed original PX datasets in PeptideAtlas). It is then possible to query the archive and identify datasets of interest using biological and technical metadata, keywords, tags, or publication information. To monitor the release of new public PX datasets, researchers can subscribe to a Rich Site Summary feed (http://groups.google.com/group/proteomexchange/feed/rss_v2_0_msgs.xml). Next, the main characteristics of the individual members of the consortium will be explained.

3.2 PRIDE

The PRIDE database [27] (http://www.ebi.ac.uk/pride/) was initially developed at the European Bioinformatics Institute (EBI, Cambridge, UK) to store the experimental data included in publications, supporting the manuscript review process. The main data types stored in PRIDE are peptide/protein identifications (including PTMs), peptide/protein expression

Table 1. Main characteristics of the major MS-based proteomics repositories and databases

Repositories	Raw data	Support for targeted approaches	Metadata	Human protein expression information	Species	Quantification data	Related stand-alone tools	Web services URL	URL
PRIDE (June 2014)	X	–	High level	41 835 Protein accessions 269 806 Unique peptide sequences Approximately 101 million spectra	Approximately 450 species	X	PRIDE Inspector, PRIDE Converter 2, PeptideShaker	http://www.ebi.ac.uk/pride/ws/archive/	http://www.ebi.ac.uk/pride/
PeptideAtlas (Human, August 2013)	X	X	Medium level	14 018 Proteins 338 013 Peptides Approximately 258 million spectra	*Mus musculus, Candida albicans, Candida, Caenorhabditis elegans, Drosophila melanogaster, Halobacterium, Equuscaballus, Rattus norvegicus, Saccharomycescerevisiae, Danio rerio, Susscrofa, Mycobacterium tuberculosis*	–	TIQAM, TIQAM–Digestor, TIQAM–PeptideAtlas, TIQAM–Viewer, ATAQS, PIPE2, PABST		http://www.peptideatlas.org/
GPMDB (May 2014)	–	X	High level	136 373 Protein accessions 1 786 698 Peptides Approximately 1020 million spectra	*Bos taurus, Canis familiaris, Homo sapiens, M. musculus, R. norvegicus, Gallus gallus, D. rerio, Xenopus tropicalis, Anopheles gambiae, Apis mellifera, D. melanogaster, C. elegans, Oryza sativa, Arabidopsis thaliana, Saccharomyces cerevisiae, Bacillus anthracis* Ames, *Escherichia coli* (K12), *Lactococcus lactis* (Il1403), *M. tuberculosis, Shigella dysenteriae, Salmonella typhi, Salmonella typhimurium* (LT2)	–	–	http://rest.thegpm.org/1	http://gpmdb.thegpm.org/
MassIVE (May 2014)	X	–	Low level	*	*	–	–	–	http://massive.ucsd.edu/
Chorus (May 2014)	X	–	Low level	*	*	–	–	–	https://chorusproject.org/
ProteomicsDB (May 2014)	X	–	High level	18 097 Proteins 739 406 Peptides Approximately 70 million spectra	*H. sapiens*	X	–	–	https://www.proteomicsdb.org/
MOPED (May 2014)	–	–	Medium level	17 141 Proteins 250 000 Unique peptides Approximately 15 million spectra	*H. sapiens, M. musculus, C. elegans, S. cerevisiae*	X	–	–	https://www.proteinspire.org/MOPED/
Human Proteinpedia (May 2014)	X	–	Low level	15 231 Proteins 1 960 352 Peptides Approximately 5 million spectra	*H. sapiens*	–	–	–	http://www.humanproteinpedia.org/
MaxQB (May 2014)	–	–	Medium level	14 732 Proteins 370 551 Peptides Approximately 20 million spectra	*M. musculus, H. sapiens S. cerevisiae*	X	–	–	http://maxqb.biochem.mpg.de/mxdb/

Table 1. Continued

Repositories	Raw data	Support for targeted approaches	Metadata	Human protein expression information	Species	Quantification data	Related stand-alone tools	Web services URL	URL
PaxDb (May 2014)	–	–	Low level	10 482 Proteins 143 456 Peptides Approximately 24 million spectra	*A. thaliana, M. musculus, H. sapiens, S. cerevisiae, D. melanogaster, E. coli* (K12), *Microcystis aeruginosa, C. elegans, Leptospira interrogans* serovar *Copenhageni, M. tuberculosis* (H37Rv), *Streptococcus pyogenes M1 GAS, Schizosaccharomyces pombe, B. taurus, S. dysenteriae, B. subtilis, G. gallus, Thermococcus gammatolerans* (EJ3), *Mycoplasma pneumoniae, Halobacterium* (NRC–1), *S. typhimurium* LT2, *R. norvegicus, Deinococcus deserti* (VCD115), *Shigella flexneri, Synechocystis, A. mellifera, C. familiaris, Sus scrofa, X. tropicalis, O. sativa*	X	–	http://pax-db.org/api/search?q = human	http://pax-db.org/
HPM (June 2014)	–	–	Low level	10 482 Proteins 293 000 Unique peptides Approximately 25 million spectra	*H. sapiens*	X	–	–	http://humanproteomemap.org

X, the feature or characteristic is supported; -, the feature is not supported; *, it was not possible to retrieve the corresponding information.

values, the analyzed mass spectra (both as raw data and peak lists), and the related technical/biological metadata. In PRIDE, data are stored as originally analyzed by the researchers (the author's analysis view on the data), supporting many popular search engines/analysis workflows. Most of the terms, information and metadata supporting the PRIDE data, are based on ontologies or controlled vocabularies [52]. In this context, the "Ontology Lookup Service" [53] was developed as a spin-off of PRIDE to enable the querying, browsing, and navigation of biomedical ontologies. Since the inception of PX, the number of datasets submitted to PRIDE has grown considerably (Fig. 2). In parallel, the size of the experiments in terms of spectrum numbers has also increased significantly (Fig. 2).

3.2.1 Data submission and format support

As a member of the PX consortium, PRIDE supports both "complete" and "partial" submissions. For "complete" submissions, processed identification results need to be provided in PRIDE XML (PRIDE original data format) or the PSI standard mzIdentML (version 1.1) format. If mzIdentML is used, the corresponding peak list files referenced by the mzIdentML files need to be included as well. A "complete" submission ensures that the processed results data can be integrated in PRIDE and visualized using tools, such as PRIDE Inspector [54] (see below). Therefore, for performing a "complete" submission, the output files from the analysis software need to be converted or exported to either mzIdentML (see http://www.psidev.info/tools-implementing-mzidentml and http://www.ebi.ac.uk/pride/help/archive/submission/mzidentml) or PRIDE XML (using PRIDE Converter 2 [55] or other external tools). It is important to highlight that the recently implemented support for mzIdentML makes possible that data from a growing number of tools/analysis software (that were never supported via PRIDE XML) can now be fully supported by PRIDE.

The PRIDE Converter 2 tool suite (http://pride-converter-2.googlecode.com) [55] is an open source and platform-independent software that allows users to convert several search engine output files to PRIDE XML. The tool suite consists of four different applications: PRIDE Converter 2, PRIDE mzTab generator, PRIDE XML merger, and PRIDE XML filter. All of these tools can be launched using the graphical user interface or from the command line. PRIDE Converter 2 currently supports the output from MASCOT (.dat) [56], X!Tandem (.xml) [57], OMSSA (.csv), among others. It also supports different mass spectra file formats (mzML, dta, mgf, mzData, mzXML, and pkl). As mentioned above, there are other tools not developed by the PRIDE team that can also export to PRIDE XML [37]. It is expected that, as the mzIdentML format becomes more and more popular, the use of PRIDE XML will decrease in time.

The "partial" submission route is aimed at situations where processed identification results cannot be generated in

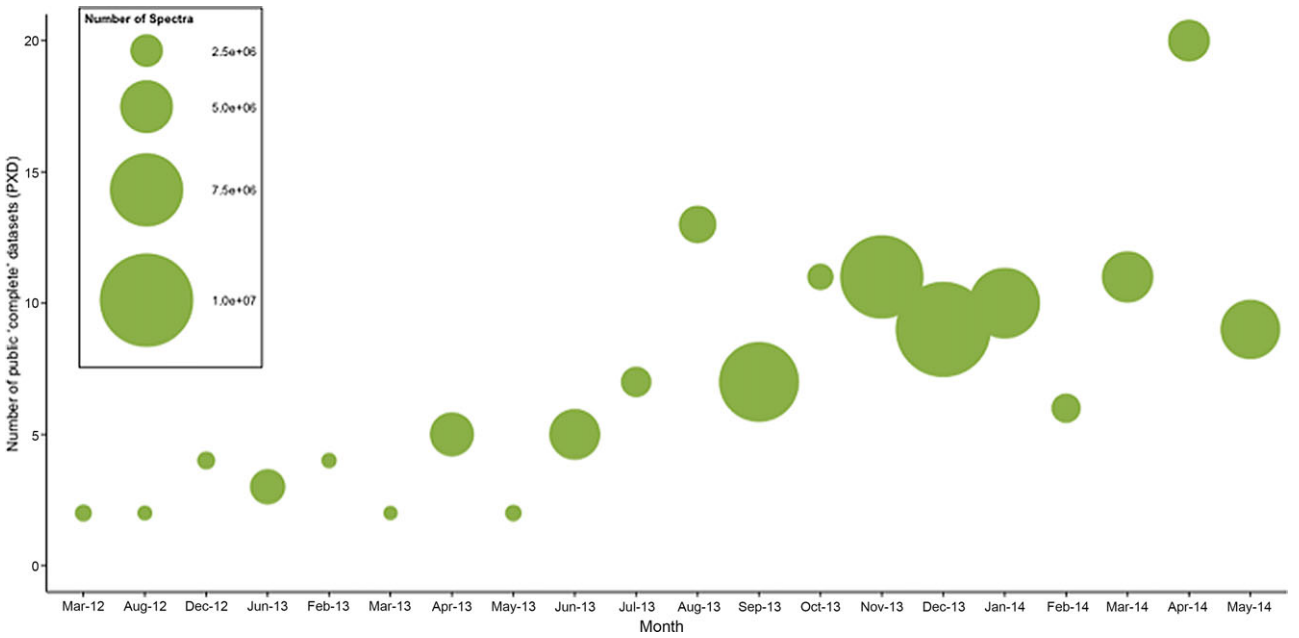

Figure 2. Bubble chart representation of the size of the PX complete submissions to PRIDE (until May 2014). The *x*-axis includes months with at least one submission, since PX submissions started (from March 2012). The *y*-axis corresponds to the number of PX "complete" public datasets submitted to PRIDE in each specific month. The size of each bubble represents the total number of mass spectra included in all the datasets in a given month.

mzIdentML/PRIDE XML if there is no converter/exporter available. However, the "partial" submission mechanism also enables any data from any proteomics workflow to be stored in PRIDE. As a consequence, some datasets are already available coming from workflows, such as top-down proteomics, MS imaging, or SWATH-MS, among others.

The PRIDE/PX submission process has been recently described in detail [58]. The PX submission tool (http://www.proteomexchange.org/submission) [37] is an open-source standalone tool that provides a user-friendly graphical user interface for performing the actual data submission, through a series of steps:

(i) Select all the files needed for submission.
(ii) Interactively group-related different types of files (e.g., the corresponding raw and processed results files).
(iii) Ensure a minimum level of metadata (according to the PX guidelines).
(iv) Transfer the files via the Aspera (from version 2.1) or FTP file transfer protocols. Aspera (http://asperasoft.com/) can perform up to 50 times faster than FTP, enabling a convenient way of transferring large datasets.

In addition to the PX submission tool, datasets containing a high number of files can also be submitted using a command line based alternative (http://www.ebi.ac.uk/pride/help/archive/aspera) [58]. Each dataset becomes publicly available on acceptance or publication of the corresponding manuscript, or when the authors tell PRIDE to do so.

3.2.2 Data mining and visualization

PRIDE Inspector (http://pride-toolsuite.googlecode.com) [54] is an open source standalone tool that can be used to efficiently browse and visualize MS proteomics data. PRIDE Inspector can be used by researchers before submission and also by journal editors and reviewers during the manuscript review process. The latest version available at the moment of writing (version 2.1) supports identification results in PRIDE XML and mzIdentML (used in PX "complete" submissions). It also supports spectra files in a variety of formats (mgf, pkl, ms2, mzXML, mzData, and mzML). PRIDE Inspector enables users to visualize and check the data at different levels. It has different panels devoted to experimental metadata, protein, peptide, and spectrum-centric information. Finally, the "summary charts" tab provides eight different charts that can be used to evaluate some aspects of the quality of the dataset. Apart from the visualization functionality, it can access some of the most popular protein databases (UniProt knowledgebase (UniProtKB), Ensembl [59], and the National Center for Biotechnology Information (NCBI) nonredundant database) to retrieve the most up-to-date protein sequences and names for the reported protein identifiers.

The PRIDE Archive website (http://www.ebi.ac.uk/pride/archive/) provides the web interface to query and retrieve the information in PRIDE. It was launched on January 2014 and at the moment of writing, is still under iterative development, so new features are being added constantly. By August 2014, the current version allows querying PRIDE

using keywords, publication, species, tissues, diseases, modifications, instruments, peptide sequences, and protein identifiers. When a specific dataset is selected, the users are directed to the dataset summary page, which also lists the assays (equivalent to the old PRIDE experiment numbers) related to the project, in the case of "complete" submissions. Users can download all the files via FTP or Aspera, and/or visualize the results using PRIDE Inspector. In the coming months, visualization for peptide/protein identifications will be available in the PRIDE web. At present the BioMart interface (http://www.ebi.ac.uk/pride/legacy/prideMart.do) is the easiest way to access this information. However, it is planned that it will be soon replaced by new PRIDE web services.

Additionally, PRIDE data can be accessed using the "PRIDE Cluster" webpage (http://www.ebi.ac.uk/pride/cluster/) [60]. It includes public identified spectra in PRIDE that have been clustered using the "PRIDE Cluster" algorithm (https://code.google.com/p/pride-spectra-clustering/). The PRIDE Cluster resource currently provides two main methods for accessing its data: (i) retrieve all clusters that contain a given peptide identification and (ii) retrieve all clusters with a consensus spectrum similar to a queried spectrum. In addition, spectral libraries for several species are also provided. "PRIDE Cluster" is the first quality control step attempted in a highly heterogeneous MS proteomics repository, such as PRIDE.

3.3 PASSEL

PASSEL [30] (http://www.peptideatlas.org/passel/) supports the submission of datasets generated by SRM approaches by storing the experimental results and the corresponding raw data. The submitted raw data are automatically reprocessed in a uniform manner using mQuest, a component of the mProphet software suite [61], and the results are loaded into the database [62]. The original files and the corresponding reprocessed results are made available to the community. Additionally, the measured transitions are incorporated into the SRMAtlas [62] catalog of transitions. Detailed metadata and structured sample information are in accordance with the PX guidelines.

3.3.1 Data submission and format support

PASSEL uses a web interface for the data submission process (http://www.peptideatlas.org/submit) and is now supporting the following data types (http://www.peptideatlas.org/upload/): (i) study metadata, such as dataset title, submitter contact information, sample source, sample preparation, and instrument used; (ii) transition lists describing which transitions were measured for each peptide and targeted ions, along with optional supporting information (collision energy, expected retention time, and expected relative intensities). This information is available in a tab-separated file or in the standard TraML format [50]; and (iii) mass spectrometer output files in mzML [45] or mzXML [48] format. If these are not available, the vendor formats .wiff (AB SCIEX), .raw (Thermo), or .d (Agilent) are also supported. All the files supplied by the users are uploaded to a specially created FTP account and they can be browsed using the "PASSEL Experiment Browser".

3.3.2 Data mining and visualization

The "PASSEL Experiment Browser" enables filtering based on fields, such as research contact information, organism, sample, and instrument type. As mentioned above, it is also possible to download the original files provided (http://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetSELExperiments).

The "PASSEL Data Browser" provides a description of each selected transition group and the data collected for it, a link to visualize the trace group using the "Chromavis" chromatogram viewer and further links to the information available in SRMAtlas [62] and PeptideAtlas [26], in order to extract or compare spectral features of the targeted peptides (http://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetSELTransitions).

3.4 PeptideAtlas

The PeptideAtlas project (http://www.peptideatlas.org) [26, 63, 64] was originally created to serve as the end-point for the TPP processing software [49]. In recent years, PeptideAtlas has grown as a data reprocessing resource and it has served as a research database for the development of spectral libraries [65] and SRM-related tools [66, 67]. For instance, PeptideAtlas provides information about proteotypic peptides using detectability scores [68].

Nowadays, PeptideAtlas is one of the biggest and well-curated protein expression data resources. The initial PeptideAtlas publication reported the identification of 27% of the human genes in Ensembl with a protein false discovery rate (FDR) likely 10% or higher [63]. Over the years, the PeptideAtlas team has developed different tools to control the assignment of incorrect identifications, such as PeptideProphet [69] and ProteinProphet [70], and more recently MAYU [71], to control the protein FDR when different datasets are combined. This platform and the statistically accurate protocol used to curate the protein/peptide identification data have turned PeptideAtlas into a very reliable protein expression database. In 2013, the generated 1% protein-level FDR Human PeptideAtlas had at least one peptide for around 14 000 different UniProtKB/Swiss-Prot entries [26].

In the context of PX, partners have recently started to track the reanalysis of original PX datasets by providing RPXD identifiers to those, and linking them to the original reprocessed datasets. By August 2014, several PX datasets reanalyzed by PeptideAtlas are already publicly available in ProteomeCentral.

3.4.1 Data submission and format support

PeptideAtlas reprocessed data are organized into different builds [72], each includes data from a single proteome or subproteome (see species in Table 1). Each build is generated with raw MS/MS spectra submitted using the "submission form" (http://www.peptideatlas.org/upload/) or the data deposited into another public repository, such as PRIDE. These spectra are searched against a sequence database, a spectral library or both. Peptide and protein identifications are mapped to a comprehensive reference protein database (for the latest human builds, the searched database is a combination of UniProtKB/Swiss-Prot, Ensembl, and sequences from the International Protein Index (IPI)), and postprocessed using the TPP [72]. It also annotates each protein and peptide with supporting data, such as genome mappings, sequence alignments, links to different databases, such as GPMDB or the Human Protein Atlas [73], uniqueness of peptide–protein mappings, observability of peptides, predicted observable peptides, estimated protein abundances and cross-references to other databases, such as RefSeq, UniGene and UniProt. All the processed results are loaded into SBEAMS (systems biology experiment analysis management system) proteomics that is a proteomics analysis database built as a module under the SBEAMS framework.

3.4.2 Data mining and visualization

The PeptideAtlas web search interface can be used to search for proteins by protein accession, peptide sequence, gene name, keyword, or phrase. If a specific protein is requested, the "protein view page" summarizes all the information available for that protein. The top section provides basic information about the protein, including alternative names as well as the total number of corresponding spectra (observations) and distinct peptides. The following two sections, "sequence motifs" and "sequence," summarize the peptide coverage of the protein. Finally, a similar diagram to a genome browser view summarizes all the peptides that map either uniquely or redundantly to the protein, including information on segments unlikely to be observed by MS. Information about signal peptides and transmembrane domains is also provided, where available. One of the best features of PeptideAtlas is the Cytoscape [74] plug-in that allows the user to view the distinct peptides for a particular protein as a network with associated proteins.

Recently, PeptideAtlas implemented the "PeptideAtlas Chromosome Explorer" (http://www.peptideatlas.org/peptideatlasExplorer/) to summarize and classify the identified human proteome using a chromosome-oriented view. The present version shows the number of protein observations and the UniProt entries by chromosome using a circular histogram plot. PeptideAtlas is actively involved in the HUPO chromosome-based Human Proteome Project (C-HPP) [75] and provides a dedicated website to access protein identifications per chromosome (http://www.peptideatlas.org/hupo/c-hpp/).

PeptideAtlas heavily supports targeted proteomics workflows in different ways: (i) SRMAtlas (http://www.srmatlas.org/) is a compendium of targeted proteomics assays to detect and quantify proteins using SRM/MRM-based proteomics workflows; (ii) TIQAM (targeted identification for quantitative analysis by MRM) [76], which is a desktop application to facilitate the selection of peptide and transitions. It consists of three applications: "TIQAM-Digestor," "TIQAM-PeptideAtlas," and "TIQAM-Viewer"; (iii) the automated and targeted analysis with quantitative SRM tool (ATAQS) [77] is a software pipeline tool that contains modules to design, manage, analyze, and validate an MRM assay. ATAQS uses FireGoose [78] to connect to various web services, such as PeptideAtlas (used to select spectra), TIQAM (to generate in silico peptides for a given protein [76]), PIPE2 (to generate a list of proteins, to design an MRM assay, and for other various analysis tasks), and PABST (peptide atlas best SRM transition, used to generate optimal transitions).

3.5 MassIVE

The MassIVE data repository (http://massive.ucsd.edu) is a community resource developed by the Center for Computational Mass Spectrometry (University of California, San Diego) to promote the global-free exchange of MS data. MassIVE provides a location for researchers to access public raw datasets and accompanying files, often alongside publication. One of the key features of MassIVE (still under development) compared with other resources is that its functionality is aimed at networking and providing a social platform for researchers. MassIVE users are not only able to browse, download, but also comment on datasets. These comments can be accompanied with new data or new analyses that enrich the original dataset.

3.5.1 Data submission and format support

The submission to MassIVE is based on two main steps: (i) upload the data files to ProteoSAFe (http://massive.ucsd.edu/ProteoSAFe/); and (ii) invoke the MassIVE dataset submission workflow for those files. The MassIVE team strongly recommends that submitters use FTP to upload and organize their dataset files as opposed to the ProteoSAFe file upload web interface.

MassIVE dataset files are organized into the following categories: (i) license files—specifying how and under which conditions the dataset files may be downloaded and used; (ii) spectrum files—any mass spectrum files constituting the main body of a dataset (in mzXML and/or raw binary format); (iii) result files—the output of any search engine; (iv) sequence databases—any protein or other sequence databases that were searched against (.fasta format); (v) spectral libraries—any spectral library files that were searched,

if applicable; (vi) methods and protocols, for any open-format files containing explanations or discussions of the experimental procedures used to obtain or analyze a given dataset.

The MassIVE submission web presents an input form including all the files to be uploaded and relevant metadata. The metadata fields include the species, instruments, PTMs, and contact information. When a dataset is submitted, the ProteoSAFe validation is performed.

3.5.2 Data mining and visualization

The "MassIVE datasets" browser (http://massive.ucsd.edu/ProteoSAFe/datasets.jsp) presents a list of all public datasets in a tabular format. Users can sort and filter by any column. Alternatively, users can select specific datasets by checking the boxes next to them. MassIVE tried to rescue as many datasets as possible from the defunct Tranche repository. For this imported data, the Tranche hash is also displayed, enabling the users to search for specific Tranche datasets.

Once data are submitted to the repository, it can be shared with others (e.g., reviewers) using password-protected access to the datasets, or be made publicly available by the submitter. In the near future, it is planned that anyone with access to a given dataset will be able to reanalyze it using the many workflows available in ProteoSAFe. After the reanalysis, the results can then be used to either enrich existing datasets or to create new ones. MassIVE enables on-site reanalysis using a variety of data workflows: standard database searches (MS-GF+) [79], proteogenomics searches against genomics/transcriptomics sequences (ENOSI) [80–82], discovery of unexpected modifications (MODa) [83], identification of mixture spectra using spectral library (MSPLIT) [84] and database (MixDB) search [85], de novo sequencing of peptides (PepNovo) [86] and proteins (Meta-SPS) [87], molecular spectral networks (including both peptides and metabolites), and top-down protein identification (MS-Align+) [88].

3.6 Chorus

Chorus (http://chorusproject.org) is a cloud-based application that provides scientists the ability to store, analyze, and share their MS data regardless of the original raw file format generated. The Chorus team's aim is to create a complete catalogue of the mass spectrometric data in a way that can be openly accessed, and make it freely accessible to both the scientific community and the general public. Chorus was originally announced in 2013 and is enabled by Amazon Web Services, Inc. (http://aws.amazon.com). It is a global cloud-based computing environment that provides customized data analysis tools that convert proprietary data formats to a common "Map Reduce" format for processing large datasets using parallel and distributed algorithms on the cloud. The initial data analysis tools include chromatographic and mass spectral viewers as well as a database search engine for protein sequence identification.

The registered users should first request laboratory membership or create their own laboratory. After that, they can create a project, an instrument, and upload raw files. Once this is done, the users can create experiments and organize them into the created projects.

3.7 GPMDB

GPMDB [25] (http://gpmdb.thegpm.org) is one of the most well-known protein expression databases. The GPMDB pipeline reprocesses the MS/MS data provided by users or raw data stored in other repositories, such as PX, using the popular open source search engine X!Tandem [57]. Peptide and protein identifications are generated and stored in XML files, which are indexed in a MySQL database.

3.7.1 Data submission and format support

The data submission process supports MS data in different formats (.dta, .pkl, .mgf, mzXML, and mzData) and it can take place via the "simple search page." Once the data have been processed using X!Tandem, users can choose whether to submit their data or not to GPMDB. In addition to X!Tandem, users can also use a spectral search engine called X!Hunter (http://xhunter.thegpm.org/) [89] or the proteotypic peptide profiler X!P3 (http://p3.thegpm.org/) [90] to analyze their data. All the identified peptides are matched to the Ensembl [59] genome database.

3.7.2 Data mining and visualization

The GPMDB main entry point is a web-based search interface, where users can search by keywords, data sources, protein identifiers, or gene names. Also, the current GPMDB implementation provides another three alternative pages to quickly access and search the data: (i) dataset search by accession (also called gpm #, http://gpmdb.thegpm.org/gpmnum.html); (ii) sequence search (also called sequence, http://gpmdb.thegpm.org/seq.html); (iii) access based on GO terms [91], chromosome, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [92], BRaunschweig ENzyme DAtabase (BRENDA) [93] tissue ontology (http://gpmdb.thegpm.org/go/index.html); and (iv) the amino acid polymorphisms and PTMs search for proteins.

When a search is performed, the main "protein table" lists all the matched proteins and different features are shown, such as protein accession, number of observations in GPMDB, log(e), and an evidence code that rates the current evidence for the observation of each protein. The "protein view" shows all the peptide identifications for a specific protein and the corresponding sequence coverage to specify the reliability of the identification.

More recently, GPMDB has been actively involved in the C-HPP project [94]. In this context, human protein identification information in GPMDB is now summarized into a collection of spreadsheets called "Guide to the Human Proteome." This guide contains the information organized into separate spreadsheets for each chromosome as well as for the mitochondrial DNA. It also contains protein accession numbers, HGNC (HUGO (Human Genome Organization) Genome Nomenclature Committee) [95] gene names, and chromosomal coordinates taken from Emsembl.

3.8 ProteomicsDB

ProteomicsDB (http://www.proteomicsdb.org/) is a human protein expression database that stores protein and peptide identifications and quantification values. The resource was jointly developed by the Technical University of Munich, the company SAP (Walldorf, Germany), and the SAP Innovation Center. It was announced in 2013 and it has been recently highlighted as the main output of a study drafting the human proteome [28]. It contains information of more than 62 projects and more than 300 experiments. The proteins identified in the resource map to over 18 000 human genes, representing around 90% of the human proteome. However, the FDR calculations when different datasets are combined were not taken into account in the analysis [18], as PeptideAtlas does [63, 71]. It currently contains approximately 70 million spectra from human cancer cell lines, tissues, and body fluids. ProteomicsDB enables real-time analysis and is based on the SAP HANA platform (http://www.saphana.com) for rapid data mining and visualization. It has been built to enable public sharing of datasets as well as to enable users to access and review data prior to publication.

3.8.1 Data submission and format support

The submission pipeline is well structured in three different levels: (i) projects that contain the general information, such as publication, title, and an experiment summary; (ii) experiments that include a name, description, and scope (the scope represents a definition of the experiment type, e.g., PTM or full tissue proteome); and (iii) experiment files—when the users upload a set of files, each file is sent to a verification process before it is definitely accepted or rejected. A large number of the experiments in the database were downloaded from repositories such as Tranche, PRIDE/PX, and PeptideAtlas [37], and later reprocessed using two parallel pipelines based on Andromeda/MaxQuant [96] and MASCOT, and quantified using MaxQuant [97].

3.8.2 Data mining and visualization

A user-friendly web interface allows users browse the human proteome, including protein-level information, such as protein function and expression. Protein expression can be visualized across the complete human body. This novel feature is integrated into the "Protein expression tab" and comprises the human body map showing the expression of a given protein in more than 30 tissues, organs and body fluids. Also, the protein expression of cell lines can be projected onto the tissue of origin and experimental details for any sample of interest can be readily obtained. The "chromosome view" shows those proteins identified in each chromosome region, including the description of the proteins, length, unique peptides, unique PSMs (peptide spectrum match), shared PSMs and sequence coverage.

3.9 MaxQB

MaxQB (http://maxqb.biochem.mpg.de/mxdb/) [29] was released in 2012 as a database that stores and displays collections of large proteomics projects and allows joint analysis and comparison. MaxQB serves as a generic repository and analysis platform for high-resolution bottom-up experiments. It stores details about protein and peptide identifications together with the corresponding high- or low-resolution fragment spectra and quantitative information, such as protein ratios or label-free derived intensities.

3.9.1 Data submission and format support

Differently to other databases, MaxQB only stores experiment data generated locally by the Mann group. This feature reduces greatly the data heterogeneity and facilitates the data analysis. To enable smooth upload of data, MaxQB is tightly integrated with MaxQuant [97]. When a MaxQuant analysis is performed, the user of MaxQuant is asked whether she/he wants to upload the data or not. This new feature will allow the integration of new datasets from others into MaxQB. Alternatively, the data can be manually uploaded through the web interface.

Additional metadata information should be provided in the submission process, such as the project name, experiment name, and workflow parameters. All the data are stored in a relational database. Several human protein sequence databases (UniProt including variants, Ensembl and IPI) were uploaded to MaxQB to build a reference protein database.

3.9.2 Data mining and visualization

The MaxQB web interface allows the users query the database by different fields, such as gene name, organism, and source database. Alternatively, an advance query builder can be used if the user is not familiar with the query syntax. Apart of protein detail information and sequence coverage, MaxQB can also display expression information within any of the proteomes compared with all other quantified proteins in that proteome. The expression of a given protein is estimated by the sum of its peptide signals after normalization of the

total proteome signals to each other in MaxQuant. The iBAQ algorithm [98, 99], based on spectral counting, is now implemented in MaxQuant and can also be used to estimate protein quantification values. One of the major conclusions that MaxQB data analysis provides is that the peptide rank order can be used as a component of the protein identification score.

3.10 MOPED

MOPED (http://moped.proteinspire.org) [31, 100] is an expanding resource that enables rapid browsing of protein expression information coming from humans and several model organisms. MOPED provides protein-level expression data, meta-analysis capabilities, and quantitative data from standard analyses. In order to address the metadata diversity, the MOPED database developed a multi-"omics" metadata checklist that has been used for collecting metadata, which has been made available to the community through DELSA (www.delsaglobal.org).

3.10.1 Data submission and format support

In order to start a dataset submission, MOPED requires a minimum set of metadata that must be included at the experiment level. Users must then supply a brief experimental description, the source organism, and the journal reference. The MOPED pipeline [101] is based on reanalyzing MS data from other public repositories using SPIRE (Systematic Protein Investigative Research Environment http://proteinspire.org) [102]. SPIRE integrates open source search tools, such X!Tandem, OMSSA, and various statistical models into its pipeline. MOPED estimates protein absolute expression and concentration values using spectral counting. Since the probability of identifying peptides depends on peptide properties, such as the peptide sequence, the APEX quantitative approach weights spectral counts using estimates of these probabilities to improve the accuracy of the provided absolute expression values [101]. Each identified protein links to various protein and pathway databases, including GeneCards [103], UniProt, KEGG and Reactome [104]. The current MOPED database contains data from four of the most studied organisms: human, mouse, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*.

3.10.2 Data mining and visualization

The web interface contains three main panels: "protein absolute expression" "protein relative expression" and "gene relative expression." Each of the panels contains the "MOPED search box" that supports queries by keywords, tissues, conditions, and pathways. After a search is performed, the "protein ID and expression summary" section displays expression data. Each protein row in the expression summary table displays the protein accession, description, concentration, organism, localization, sequence coverage, spectral count, and gene name.

The interface also features a "visualization panel" known as the chord diagram. It can break down proteins in experiments by organism, tissue, localization, and condition. The absolute and relative expression matrix in these panels shows the expression of the identified proteins per condition. In the future, MOPED plans to integrate proteomics with transcriptomics and metabolomics data through biological pathways and networks.

3.11 PaxDb

The first version of the PaxDb resource (http://pax-db.org/) [32] was published and released in 2012. PaxDb is a database dedicated to integrate information on absolute protein abundance levels, based on a deep coverage of the proteome, consistent data postprocessing, and enabling comparability across different organisms. PaxDb is a meta-resource since it takes the information previously published in data repositories, such as PRIDE and PeptideAtlas, and does not accept any direct data submissions.

3.11.1 Data submission and format support

The current PaxDb pipeline analyzes the spectral counting information from PeptideAtlas builds—that is, which peptides have been identified and how often over the whole build. This part of PaxDb's data import is entirely based on the original scoring and quality cutoffs implemented by PeptideAtlas. In the case of identified peptide sequences reported from MS/MS approaches (e.g., PRIDE datasets), the current pipeline remaps each peptide to the corresponding protein, based on sequence matches using reference genomes from the STRING database [105]. Protein abundance values are converted for each dataset into protein abundance estimates, using a consistent value across different techniques. Instead of using "molar concentration" or "molecules per cell," the system expresses all abundances in "parts per million." This means that each protein is listed relatively to all other protein molecules in the same sample. In the case of biochemical, biophysical, or label-free MS experiments, parts per million values are directly computed by rescaling the provided abundance estimates by their total sum. In the case of spectral counting data, the method is based on the likelihood of detectability, as described earlier [106].

3.11.2 Data mining and visualization

The PaxDb website allows an ad hoc query of a protein family of interest, allowing multiple proteins requests simultaneously, as well as browsing and comparing complete datasets.

In addition, the user queries are searched against the annotations of all proteins in PaxDb using a full-text search. For each organism, a distinct summary page provides information on the data origin, its coverage and estimated quality, and the distribution of abundance values of each dataset as a histogram. Additionally, the most abundant proteins in the organism are also listed. The PaxDb protein view shows a brief description of the protein functional role as annotated by UniProt and/or by other model organism databases. The species browser allows the navigation through all the proteins for a specific model organism. Furthermore, it shows all the protein abundances in a specific dataset and also for the whole of the PaxDb datasets (addition of all datasets for a given specific species). All PaxDb data are freely available.

3.12 Human Proteinpedia

Human Proteinpedia (http://www.humanproteinpedia.org) [33] is a public human proteome repository for sharing protein expression data derived from multiple experimental platforms. It incorporates diverse features of the human proteome, including protein–protein interactions, enzyme–substrate relationships, PTMs, subcellular localization, and expression of proteins in various human tissues and cell lines, in diverse biological conditions (including disease states). The Human Proteinpedia database falls neatly in the annotation-oriented database category [38]. It complements the curated human protein reference database (HPRD) [107] with community-provided annotation. It collects submitted proteomics data and uses the findings to directly annotate the protein entries in HPRD with observed PTMs and subcellular localization [108]. In contrast with the resources previously described, Human Proteinpedia does not only contain MS-based experiments, but also other approaches such as coimmunoprecipitation, Western blotting, fluorescence, immunohistochemistry, and protein and peptide microarrays. All the data should be annotated by the users (http://pdas.hprd.org/) and it is not reprocessed.

3.12.1 Data submission and format support

Users can provide data annotations in four different ways, after registering in the system: (i) data on individual proteins along with experimental evidence through the use of web forms, (ii) upload data via the web in a batch mode, (iii) sending data through FTP/e-mail to the support team, and (iv) distributed annotation service servers setup by the contributing laboratories for the data upload [109]. By June 2014, the current version includes data from more than 249 laboratories, including 2710 distinct experiments, more than 15 000 proteins, almost 2 million peptides, around 5 million spectra, and 2906 annotations related to subcellular localization (Table 1).

The major differences with other repositories are as follows: (i) it does not exclusively contain MS-derived data, as mentioned already; (ii) data from proteomics experiments are viewed in the context of a protein–protein interaction resource (HPRD); (iii) it restricts the data to that derived from human tissues or cell lines; and (iv) data annotation related to various protein features can be done manually.

3.12.2 Data mining and visualization

The query page in Human Proteinpedia (http://www.humanproteinpedia.org/query) enables the search by gene symbol, protein name, accession number, type of protein feature, and type of experimental platform annotated. After a query, the "protein detail view" includes a table where each row shows a "platform" (the type of the experiment, e.g., MS or immunohistochemistry) and related metadata. In the case of MS experiments, each platform contains a table containing all the peptides identified for each protein. All the data are freely available (http://www.humanproteinpedia.org/download).

3.13 HPM

The HPM (http://www.humanproteomemap.org) [34] has just been developed as an output of a recent draft study of the human proteome. The original proteomics study [34] was carried out on 30 histologically normal human tissues and primary cells using high-resolution MS. The generated tandem mass spectra correspond to proteins encoded by 17 294 genes, accounting for approximately 84% of the annotated protein-coding genes in the human genome. However, FDR calculations when different datasets are combined are not discussed in enough detail in the original manuscript [18]. The aim of the HPM is to make possible to review, navigate and visualize the protein expression information evidences of gene families, protein complexes, signaling pathways and biomarkers.

3.13.1 Data submission and format support

MS/MS data obtained from all different experiments were searched against the Human RefSeq database using SEQUEST [110] and MASCOT [56] through the Proteome Discoverer platform (Thermo Fisher Scientific). Then, q values were estimated using the Percolator algorithm within the Proteome Discoverer suite. Protein and peptide identifications obtained from SEQUEST and MASCOT were converted into MySQL tables. NCBI RefSeq annotations were used as additional information about the genes from various public resources. Normalized spectral counts [34] were used to represent expression of proteins and peptides. The

resource was developed as a protein expression database and do not support the submission of new data from external users.

3.13.2 Data mining and visualization

The web portal was developed using a three-tier web architecture with presentation, application, and persistence layers. Users are able to search using gene or protein identifiers and the information is presented in a tabular format. For each peptide, a high-resolution MS/MS spectrum from the best scoring identification is shown on the spectrum viewer page using the "Lorikeet" JQuery plugin (https://code.google.com/p/lorikeet).

3.14 Other proteomics resources

There are other less widely used resources that will not be explained here in detail. First of all, the Cardiac Organellar Protein Atlas Knowledgebase (COPaKB; http://www.heartproteome.org/copa/) [111] is a centralized platform of high-quality cardiac proteomics data, bioinformatics tools, and relevant cardiovascular phenotypes. Currently, COPaKB features eight organellar modules, comprising 4203 MS/MS experiments from human, mouse, *Drosophila* and *C. elegans* as well as expression data of 10 924 proteins in the human myocardium. COPaKB has an attractive web interface which provides a number of effective workflows to guide cardiovascular investigators from the actual proteomics data to the systematic biomedical interpretation. The COPaKB team continues to develop innovative workflows to help providing a better understanding of protein functions in cardiovascular diseases. COPaKB will cover additional modules on organelles and cells from cardiovascular-relevant model systems. The content of each module will expand with the available public data.

Pep2pro (http://fgcz-pep2pro.uzh.ch/) [112, 113] is a comprehensive analysis database specifically suitable for performing flexible data analysis. Pep2pro is a further development of the "AtProteome" resource and provides data from *Arabidopsis thaliana*. The current pipeline employs PepSplice [114] and TPP tools for peptide/protein identification, the characterization of whole genome hits and the PTMs. The database is organized in assemblies, similarly to the PeptideAtlas builds. The *Arabidopsis* assembly from 2011 contained more than 14 522 protein identifications, 141 235 identified peptides, and around 2 million spectra.

iProX (Integrated proteome resources, http://www.iprox.org/) is a repository jointly developed by Beijing Proteome Research Center and other institutions in China. It has been recently developed reusing part of the PRIDE source code. At the moment of writing, it is not fully operational but contains already some stored datasets.

3.15 Proteomics information available through UniProt and neXtProt

UniProt (http://www.uniprot.org) [115] is among the most used of the protein sequence and functional annotation providers. Among the UniProt databases is the UniProtKB, which provides a broad range of protein sequence datasets for a large number of species, specifically tailored for an effective coverage of the sequence space while maintaining a high-quality level of sequence annotations and mappings to the genomics and proteomics information.

MS proteomics data deposited in the main public repositories is flowing into UniProtKB to enrich protein sequence annotations at the level of the evidence, supporting the existence of a protein (isoforms and variant-containing sequences included). This information is thus provided to users mainly in two different ways: (i) indirectly via the UniProtKB protein existence values (http://www.uniprot.org/manual/protein_existence) that are starting to be assigned also on the PSM-level content publicly provided by proteomics repositories (together with the accompanying statistical assessment for each PSM) and (ii) directly through explicit links to relevant cross-referenced resources (http://www.uniprot.org/database), which cover PRIDE, PeptideAtlas, MaxQB and PaxDb, but also others such as PhosphoSitePlus [116].

Organism-specific mappings of the peptides reported in the repositories to the UniProtKB sequences are also shared with groups producing genome builds to improve the corresponding gene annotations. In the future, it is planned that PTMs [117] from proteomics repositories will also be integrated in UniProtKB.

neXtProt (http://www.nextprot.org) [43] is a web-based protein knowledge platform to support research uniquely on human proteins. The set of manually curated annotations extracted from UniProtKB/Swiss-Prot for human, which constitutes the heart of the resource, is constantly complemented with quality-filtered carefully selected high-throughput experiments from different scientific research areas concerning abundance, distribution, subcellular localization, interactions, and cellular functions. neXtProt is part of the consortium, which is driving the C-HPP project. Thus, neXtProt labels as "missing proteins" those ones, which so far lack experimental evidence [43].

All the relevant proteomics MS-related information (like for instance PTM-related MS-based papers on *N*-glycosylation, phosphorylation, S-nitrosylation, ubiquitination, and sumoylation) has been integrated into neXtProt and is available via the web interface in the "Proteomics view" of the "Protein perspective" layout.

4 Data reuse from public resources

Since the current volume of proteomics data deposition is rapidly increasing, new approaches based on the reanalysis

of the data and/or new uses of the stored data are being developed. The same public dataset can be analyzed using different pipelines to discover, confirm, or highlight new biological evidences. Resources such as GPMDB and PeptideAtlas have been doing this for many years already, emphasizing control of the number of false-positives at both peptide and protein level.

Data availability per se in repositories has enabled data reanalysis by third parties triggering a discussion in the field about controversial datasets [118–122]. Furthermore, targeted reanalysis of public data by individual groups is now starting to flourish. Remarkably, the draft of the human proteome reported recently and available in ProteomicsDB includes a big proportion of reanalyzed public datasets (roughly 40% of the MS runs) [28]. In addition, new PTMs have also been described as a result of reanalysis of public datasets with a different purpose in mind. For instance, two recent studies described new PTMs after reanalyzing available phosphoproteomics studies [123,124]. Some proteogenomics studies have also made use of public datasets [125].

Data from repositories is often used in the design of SRM/MRM transitions for targeted proteomics approaches [126]. A mentioned already, GPMDB and PeptideAtlas provide already this functionality. Data from PRIDE is being used by the MRMAID resource, with the same purpose in mind [127]. Another popular data reuse is the building of spectral libraries. Several repositories build their own libraries (e.g., PeptideAtlas, PRIDE) that can be used in spectral searches. In addition, meta-analysis of data in PRIDE (without reprocessing the data) has already happened, enabling the extraction of new knowledge [128–132]. Finally, data in proteomics repositories can also be used to improve the content of the protein sequence databases [133].

In this context, PeptideShaker (http://peptide-shaker.googlecode.com/) is a recently developed Java application [134, 135] that allows the postprocessing and visualization of protein identification experiments and it greatly facilitates the reprocessing of PRIDE datasets by using the "PRIDE Reshake" functionality. One of the key features of PeptideShaker is the validation of the results using different values, such as FDR and false-negative rate [136]. These values are displayed in the "FDR/false-negative rate plot" and the availability of these metrics make possible the generation of a "cost/benefit" curve, also known as receiver operating characteristic curve, which enables the users to optimize the quality thresholds.

Finally, it is also important to highlight that not only the publicly available raw data files are used for reanalysis purposes. The availability of output files from pipelines/software can also help third parties to develop new tools and demonstrate its utility. This was the case for the recently developed SpliceVista visualization tool [137].

We strongly encourage proteomics researchers to upload raw data and processed results to public repositories. Before starting a new project, public data can be used to guide new research. This is already common practice in the case of targeted proteomics workflows.

5 Pitfalls and future challenges

Data sharing and dissemination is a nontrivial task since it requires substantial investment in infrastructure and software development [138,139]. As mentioned before, one of the aims of the PX consortium is to provide backup when one of the resources has funding problems. As a proof of concept, PRIDE and the defunct Peptidome joined forces to transfer all data from Peptidome to PRIDE [140]. In parallel, as mentioned already, MassIVE has tried to rescue as many datasets originally stored in Tranche as possible, but it has been quite a challenging task.

In June 2012, PRIDE started to accept and handle raw data as part of the PX data workflow. PX and other resources have proven that public data dissemination in a decentralized and well-structured mode is possible and actually important for the proteomics community. In our opinion, the next step would be to foster a higher integration for these resources. At present, the main layer of integration among PX resources happens at the level of the metadata.

The alternative is to access knowledge bases, such as UniProt or neXtProt, but the amount of data coming from MS proteomics repositories in these resources is still limited. Ideally it should be possible for users to look for all available data for a given protein or peptide (including PTMs) in a more straightforward way. At present, scientists need to access the different proteomics resources when they want to get all existing information about a given protein.

To illustrate this idea, we studied the protein expression evidences for the UniProtKB/Swiss-Prot human proteome, including only canonical sequences (release 2014_05, 20,265 entries) in seven resources that have a uniform data processing pipeline: GPMDB, PeptideAtlas, MaxQB, Human Proteinpedia, PaxDb, ProteomicsDB and the HPM. This proved to be a far from trivial task (see all the retrieved data in Supporting Information, including the protein lists taken from each resource). A total of 7880 proteins represent the "core" proteome since they are stored in all the resources (Fig. 3A). In addition, 4244, 3074, 2008, 1135, 815 and 824 proteins were identified in six, five, four, three, two and only one resource, respectively (Fig. 3A). Only 285 proteins were not found in any of these seven resources. In fact, most of the described resources shared more than 60% of the protein identifications among them (Supporting Information). In our opinion, this redundancy can be used as a "reliability" measure of the evidence of protein expression. For instance, the 13 016 proteins identified in GPMDB, PeptideAtlas, and ProteomicsDB are more likely to be true-positives than the 1171 proteins identified only by ProteomicsDB (Fig. 3B). At the same time, the "exclusive" identifications in ProteomicsDB are new valuable evidences, but probably need further investigation.

As mentioned already, the identification of false-positives in big resources and/or when different datasets are combined is still a problem [18, 26] and protein expression resources should be as thorough as possible in the statistical analysis.

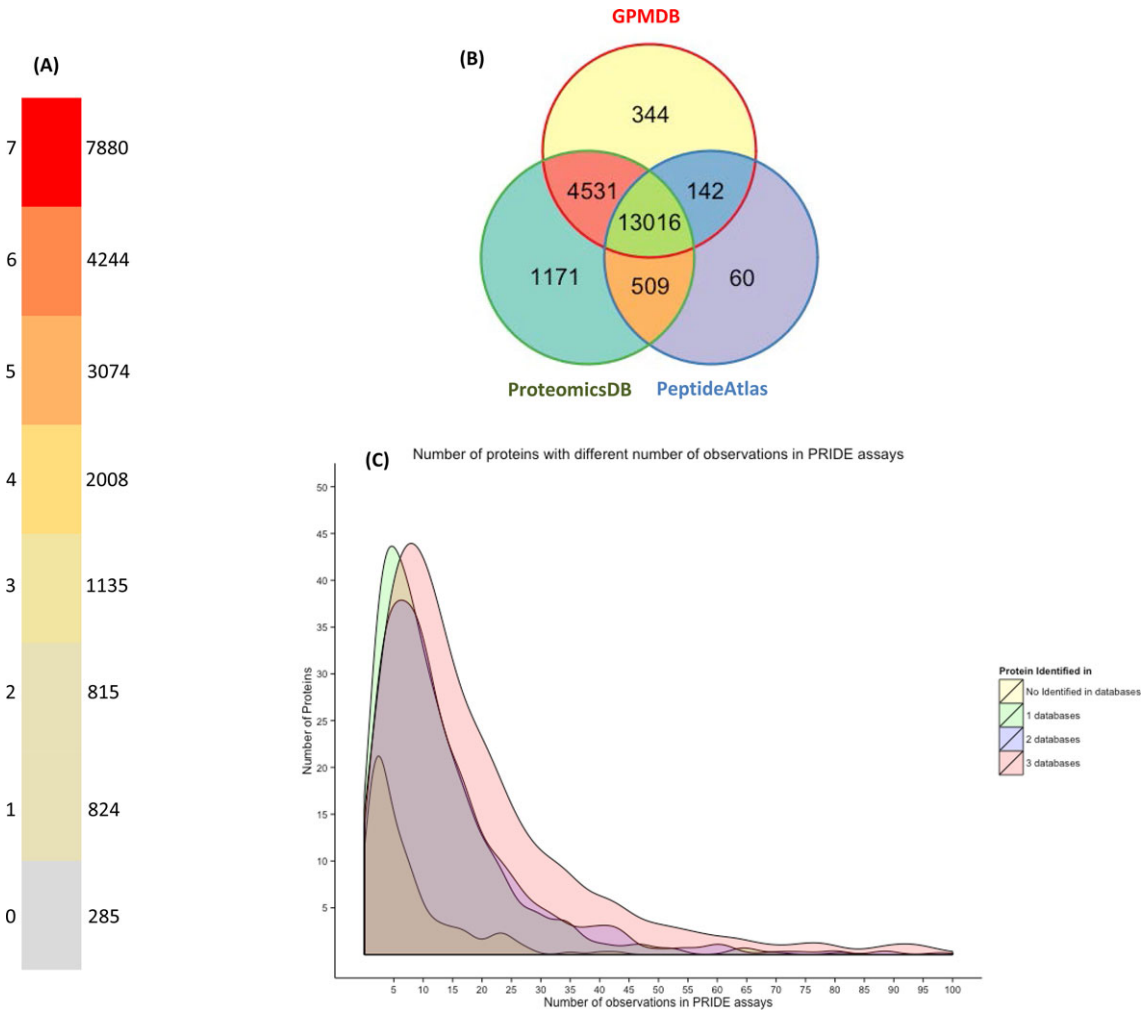

Figure 3. (A) Number of UniProtKB/Swiss-Prot human proteins (release 2014_05, 20,265 entries) observed in different proteomics resources that have a uniform data processing pipeline (GPMDB, ProteomicsDB, PeptideAtlas, HPM, PaxDb, MaxQB, Human Proteinpedia; PRIDE is not included); (B) Venn diagram representing the human protein identifications observed in GPMDB, PeptideAtlas, and ProteomicsDB; (C) Area chart showing the distribution of the number of PRIDE assays for those proteins present in three, two, and one proteomics resources, or for those proteins not identified at all.

With the identification of a high proportion of the proteins in the human proteome [28, 34, 63], new analysis tools should be implemented to retrieve protein expression evidences from different resources and to detect/highlight those identifications considered as the most and/or least reliable.

In our opinion it is also needed to increase the reuse and reanalysis of the available data, since they can provide new valuable scientific knowledge. To illustrate this idea, Fig. 3C shows the distribution of the number of PRIDE assays for those proteins present in three, two, and one proteomics resources (Fig. 3A), or for those proteins not identified at all. It can be observed that, when a protein is identified in three resources, those proteins are likely to be identified, in average, in 13 different PRIDE assays. More interesting is the case of the least observed proteins. For example, if a protein is observed in only one resource, it is observed in nine different PRIDE assays in average. In addition, those 285 proteins without any evidence can be observed in an average of five different PRIDE assays. Therefore, the number of times one given protein is reported in PRIDE (especially if not present in other resources) could be used to select datasets to be reanalyzed.

Finally, another challenge is that at present, studies integrating different "omics" technologies are becoming more and more popular. This type of studies poses a challenge for traditional repositories (which are usually field-specific) and researchers since it is not straightforward to link data from different approaches, for instance MS proteomics and RNAseq data obtained in the same study. Some big institutes such as the EBI and NCBI have implemented specific databases for BioSamples [141] to enable the linking between different studies performed using the same sample. The use of sample

identifiers is now starting to facilitate the connection between different "omics" datasets.

6 Conclusions

In recent years, there has been a big progress in the development of proteomics repositories and protein expression databases. However, more integration between those databases is required. More tools, such as PRIDE Inspector and PeptideShaker, should be developed to make easier the visualization, reuse, and reanalysis of the data in repositories. Simultaneously, in order to engage users, it seems advisable to monitor download activity of public datasets and measure the community interest on them. Since proteomics has become such a fundamental part of biological research, it is expected that the amount of information available in proteomics resources will keep growing in the coming years. We encourage researchers to make use of this plethora of data.

Y.P.R is supported by the BBSRC "PROCESS" grant (reference BB/K01997X/1). R.W. is supported by the BBSRC "Quantitative Proteomics" grant (reference BB/I00095X/1). J.A.V. is supported by the Wellcome Trust (grant number WT101477MA) and the EU FP7 grants "ProteomeXchange" (grant number 260558) and PRIME-XS (grant number 262067).

The authors thank Michael MacCoss (Chorus), Christoph Schaab (MaxQB), Eugene Kolker and Roger Higdon (MOPED), Advani Jayshree (Human Proteinpedia), Ronald Beavis (GPMDB), Mingcong Wang (PaxDb), and Nuno Bandeira (MassIVE) for providing the original data used in Fig. 3 and in the Supporting Information, and for the general feedback about their resources.

7 References

[1] Altelaar, A. F., Munoz, J., Heck, A. J., Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 2013, *14*, 35–48.

[2] Betancourt, L. H., De Bock, P. J., Staes, A., Timmerman, E. et al., SCX charge state selective separation of tryptic peptides combined with 2D-RP-HPLC allows for detailed proteome mapping. *J. Proteomics* 2013, *91*, 164–171.

[3] Branca, R. M., Orre, L. M., Johansson, H. J., Granholm, V. et al., HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* 2014, *11*, 59–62.

[4] Ramos, Y., Gutierrez, E., Machado, Y., Sanchez, A. et al., Proteomics based on peptide fractionation by SDS-free PAGE. *J. Proteome Res.* 2008, *7*, 2427–2434.

[5] Ramos, Y., Garcia, Y., Perez-Riverol, Y., Leyva, A. et al., Peptide fractionation by acid pH SDS-free electrophoresis. *Electrophoresis* 2011, *32*, 1323–1326.

[6] Wisniewski, J. R., Zougman, A., Nagaraj, N., Mann, M., Universal sample preparation method for proteome analysis. *Nat. Methods* 2009, *6*, 359–362.

[7] Michalski, A., Damoc, E., Hauschild, J. P., Lange, O. et al., Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* 2011, *10*, M111 011015.

[8] UniProt, C., Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucl. Acids Res.* 2013, *41*, D43–D47.

[9] Perez-Riverol, Y., Wang, R., Hermjakob, H., Muller, M. et al., Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim. Biophys. Acta* 2014, *1844*, 63–76.

[10] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M. et al., The quantitative proteome of a human cell line. *Mol. Syst. Biol.* 2011, *7*, 549.

[11] Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J. et al., Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* 2011, *7*, 548.

[12] Csordas, A., Ovelleiro, D., Wang, R., Foster, J. M. et al., PRIDE: quality control in a proteomics data repository. *Database* 2012, *2012*, bas004.

[13] Credit where credit is overdue. *Nat. Biotechnol.* 2009, *27*, 579.

[14] DeSouza, L. V., Siu, K. W., Mass spectrometry-based quantification. *Clin. Biochem.* 2013, *46*, 421–431.

[15] Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M. et al., Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* 2011, *11*, 535–553.

[16] Allmer, J., Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev. Proteomics* 2011, *8*, 645–657.

[17] Hoopmann, M. R., Moritz, R. L., Current algorithmic solutions for peptide-based proteomics data generation and identification. *Curr. Opin. Biotechnol.* 2013, *24*, 31–38.

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

[19] Gupta, N., Pevzner, P. A., False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* 2009, *8*, 4173–4181.

[20] Kinsinger, C. R., Apffel, J., Baker, M., Bian, X. et al., Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles). *Proteomics* 2012, *12*, 11–20.

[21] Anonymous, A home for raw proteomics data. *Nat. Methods* 2012, *9*, 419.

[22] Distler, U., Kuharev, J., Navarro, P., Levin, Y. et al., Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* 2014, *11*, 167–170.

[23] Lanucara, F., Holman, S. W., Gray, C. J., Eyers, C. E., The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nat. Chem.* 2014, *6*, 281–294.

[24] Riffle, M., Eng, J. K., Proteomics data repositories. *Proteomics* 2009, *9*, 4653–4663.

[25] 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.

[26] Farrah, T., Deutsch, E. W., Omenn, G. S., Sun, Z. et al., State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J. Proteome Res.* 2014, *13*, 60–75.

[27] Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A. et al., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucl. Acids Res.* 2013, *41*, D1063–D1069.

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

[29] Schaab, C., Geiger, T., Stoehr, G., Cox, J., Mann, M., Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell. Proteomics* 2012, *11*, M111.014068.

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

[31] Montague, E., Stanberry, L., Higdon, R., Janko, I. et al., MOPED 2.5—an integrated multi-omics resource: multi-omics profiling expression database now includes transcriptomics data. *Omics J. Integr. Biol.* 2014, *18*, 335–343.

[32] 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*, 492–500.

[33] Kandasamy, K., Keerthikumar, S., Goel, R., Mathivanan, S. et al., Human Proteinpedia: a unified discovery resource for proteomics research. *Nucl. Acids Res.* 2009, *37*, D773–D781.

[34] 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.

[35] Slotta, D. J., Barrett, T., Edgar, R., NCBI peptidome: a new public repository for mass spectrometry peptide identifications. *Nat. Biotechnol.* 2009, *27*, 600–601.

[36] Smith, B. E., Hill, J. A., Gjukich, M. A., Andrews, P. C., Tranche distributed repository and ProteomeCommons.org. *Methods Mol. Biol.* 2011, *696*, 123–145.

[37] Vizcaino, 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.

[38] Martens, L., Proteomics databases and repositories. *Methods Mol. Biol.* 2011, *694*, 213–227.

[39] Vizcaino, J. A., Foster, J. M., Martens, L., Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. *J. Proteomics* 2010, *73*, 2136–2146.

[40] Mead, J. A., Bianco, L., Bessant, C., Recent developments in public proteomic MS repositories and pipelines. *Proteomics* 2009, *9*, 861–881.

[41] Mead, J. A., Shadforth, I. P., Bessant, C., Public proteomic MS repositories and pipelines: available tools and biological applications. *Proteomics* 2007, *7*, 2769–2786.

[42] The UniProtConsortium, Activities at the Universal Protein Resource (UniProt). *Nucl. Acids Res.* 2014, *42*, D191–D198.

[43] Gaudet, P., Argoud-Puy, G., Cusin, I., Duek, P. et al., neXtProt: organizing protein knowledge in the context of Human Proteome Projects. *J. Proteome Res.* 2013, *12*, 293–298.

[44] Olsen, J. V., Mann, M., Effective representation and storage of mass spectrometry-based proteomic data sets for the scientific community. *Sci. Signal.* 2011, *4*, pe7.

[45] 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.

[46] 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.

[47] 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.

[48] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M. et al., A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[49] Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T. et al., A guided tour of the trans-proteomic pipeline. *Proteomics* 2010, *10*, 1150–1159.

[50] 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* 2012, *11*, R111.015040.

[51] Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M. et al., The mzTab data exchange format: communicating MS-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteomics* 2014, *pii*, mcp.O113.036681.

[52] Mayer, G., Jones, A. R., Binz, P. A., Deutsch, E. W. et al., Controlled vocabularies and ontologies in proteomics: overview, principles and practice. *Biochim. Biophys. Acta* 2014, *1844*, 98–107.

[53] Cote, R., Reisinger, F., Martens, L., Barsnes, H. et al., The ontology lookup service: bigger and better. *Nucl. Acids Res.* 2010, *38*, W155–W160.

[54] 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.

[55] 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.

[56] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching

sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

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

[58] Ternent, T., Csordas, A., Qi, D., Gomez-Baena, G. et al., Standardization and guidelines: how to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* 2014, in press.

[59] Flicek, P., Amode, M. R., Barrell, D., Beal, K. et al., Ensembl 2014. *Nucl. Acids Res.* 2014, *42*, D749–D755.

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

[61] Reiter, L., Rinner, O., Picotti, P., Huttenhain, R. et al., mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* 2011, *8*, 430–435.

[62] Picotti, P., Lam, H., Campbell, D., Deutsch, E. W. et al., A database of mass spectrometric assays for the yeast proteome. *Nat. Methods* 2008, *5*, 913–914.

[63] Farrah, T., Deutsch, E. W., Hoopmann, M. R., Hallows, J. L. et al., The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* 2013, *12*, 162–171.

[64] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. et al., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* 2005, *6*, R9.

[65] Lam, H., Aebersold, R., Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. *Methods* 2011, *54*, 424–431.

[66] Picotti, P., Rinner, O., Stallmach, R., Dautel, F. et al., High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* 2010, *7*, 43–46.

[67] Deutsch, E. W., Lam, H., Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* 2008, *9*, 429–434.

[68] Mallick, P., Schirle, M., Chen, S. S., Flory, M. R. et al., Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* 2007, *25*, 125–131.

[69] 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*, 5383–5392.

[70] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[71] Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M. et al., Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* 2009, *8*, 2405–2417.

[72] Farrah, T., Deutsch, E. W., Aebersold, R., Using the Human Plasma PeptideAtlas to study human plasma proteins. *Methods Mol. Biol.* 2011, *728*, 349–374.

[73] Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A. et al., A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1920–1932.

[74] Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J. et al., A travel guide to cytoscape plugins. *Nat. Methods* 2012, *9*, 1069–1076.

[75] 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*, 221–223.

[76] Lange, V., Malmstrom, J. A., Didion, J., King, N. L. et al., Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. *Mol. Cell. Proteomics* 2008, *7*, 1489–1500.

[77] Brusniak, M. Y., Kwok, S. T., Christiansen, M., Campbell, D. et al., ATAQS: a computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. *BMC Bioinform.* 2011, *12*, 78.

[78] Shannon, P. T., Reiss, D. J., Bonneau, R., Baliga, N. S., The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinform.* 2006, *7*, 176.

[79] Kim, S., Gupta, N., Pevzner, P. A., Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* 2008, *7*, 3354–3363.

[80] Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M. et al., Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 21034–21038.

[81] Woo, S., Cha, S. W., Merrihew, G., He, Y. et al., Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* 2014, *13*, 21–28.

[82] Castellana, N. E., Shen, Z., He, Y., Walley, J. W. et al., An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*. *Mol. Cell. Proteomics* 2014, *13*, 157–167.

[83] Na, S., Bandeira, N., Paek, E., Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* 2012, *11*, M111.010199.

[84] Wang, J., Perez-Santiago, J., Katz, J. E., Mallick, P., Bandeira, N., Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* 2010, *9*, 1476–1485.

[85] Wang, J., Bourne, P. E., Bandeira, N., Peptide identification by database search of mixture tandem mass spectra. *Mol. Cell. Proteomics* 2011, *10*, M111.010017.

[86] Frank, A., Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, *77*, 964–973.

[87] Guthals, A., Clauser, K. R., Bandeira, N., Shotgun protein sequencing with meta-contig assembly. *Mol. Cell. Proteomics* 2012, *11*, 1084–1096.

[88] Liu, X., Sirotkin, Y., Shen, Y., Anderson, G. et al., Protein identification using top-down. *Mol. Cell. Proteomics* 2012, *11*, M111.008524.

[89] Craig, R., Cortens, J. C., Fenyo, D., Beavis, R. C., Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* 2006, *5*, 1843–1849.

[90] Craig, R., Cortens, J. P., Beavis, R. C., The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* 2005, *19*, 1844–1850.

[91] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, *25*, 25–29.

[92] Kanehisa, M., Goto, S., KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* 2000, *28*, 27–30.

[93] Gremse, M., Chang, A., Schomburg, I., Grote, A. et al., The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucl. Acids Res.* 2011, *39*, D507–D513.

[94] Lane, L., Bairoch, A., Beavis, R. C., Deutsch, E. W. et al., Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.* 2014, *13*, 15–20.

[95] Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J. et al., Guidelines for human gene nomenclature. *Genomics* 2002, *79*, 464–470.

[96] Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A. et al., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 2011, *10*, 1794–1805.

[97] 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*, 1367–1372.

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

[99] Schwanhausser, B., Busse, D., Li, N., Dittmar, G. et al., Corrigendum: global quantification of mammalian gene expression control. *Nature* 2013, *495*, 126–127.

[100] Kolker, E., Higdon, R., Haynes, W., Welch, D. et al., MOPED: model organism protein expression database. *Nucl. Acids Res.* 2012, *40*, D1093–D1099.

[101] Higdon, R., Stewart, E., Stanberry, L., Haynes, W. et al., MOPED enables discoveries through consistently processed proteomics data. *J. Proteome Res.* 2014, *13*, 107–113.

[102] Kolker, E., Higdon, R., Morgan, P., Sedensky, M. et al., SPIRE: systematic protein investigative research environment. *J. Proteomics* 2011, *75*, 122–126.

[103] Safran, M., Dalah, I., Alexander, J., Rosen, N. et al., GeneCards Version 3: the human gene integrator. *Database* 2010, *2010*, baq020.

[104] D'Eustachio, P., Reactome knowledgebase of human biological pathways and processes. *Methods Mol. Biol.* 2011, *694*, 49–61.

[105] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M. et al., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucl. Acids Res.* 2011, *39*, D561–D568.

[106] Weiss, M., Schrimpf, S., Hengartner, M. O., Lercher, M. J., von Mering, C., Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* 2010, *10*, 1297–1306.

[107] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S. et al., Human protein reference database—2009 update. *Nucl. Acids Res.* 2009, *37*, D767–D772.

[108] Goel, R., Harsha, H. C., Pandey, A., Prasad, T. S., Human protein reference database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol. BioSyst.* 2012, *8*, 453–463.

[109] Muthusamy, B., Thomas, J. K., Prasad, T. S. K., Pandey, A., Access guide to human proteinpedia. *Curr. Protoc. Bioinformatics*, 2013, *Chapter 1*, Unit 1.21.

[110] Eng, J. K., McCormack, A. L., Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[111] Zong, N. C., Li, H., Li, H., Lam, M. P. et al., Integration of cardiac proteome biology and medicine by a specialized knowledgebase. *Circ. Res.* 2013, *113*, 1043–1053.

[112] Baerenfaller, K., Hirsch-Hoffmann, M., Svozil, J., Hull, R. et al., pep2pro: a new tool for comprehensive proteome data analysis to reveal information about organ-specific proteomes in *Arabidopsis thaliana*. *Integr. Biol.* 2011, *3*, 225–237.

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

[114] Roos, F. F., Jacob, R., Grossmann, J., Fischer, B. et al., PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. *Bioinformatics* 2007, *23*, 3016–3023.

[115] UniProt, C., Activities at the Universal Protein Resource (UniProt). *Nucl. Acids Res.* 2014, *42*, D191–D198.

[116] Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B. et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucl. Acids Res.* 2012, *40*, D261–D270.

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

[118] 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*, 280–285.

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

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

[121] Bromenshenk, J. J., Henderson, C. B., Wick, C. H., Stanford, M. F. et al., Iridovirus and microsporidian linked to honey bee colony decline. *PloSOne* 2010, *5*, e13181.

[122] 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*, A110.0063871.

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

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

[125] 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*, 756–767.

[126] 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, *106C*, 151–161.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

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

[132] Foster, J. M., Degroeve, S., Gatto, L., Visser, M. et al., A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* 2011, *11*, 2182–2194.

[133] Griss, J., Martin, M., O'Donovan, C., Apweiler, R. et al., Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. *Proteomics* 2011, *11*, 4434–4438.

[134] Barsnes, H., Vaudel, M., Colaert, N., Helsens, K. et al., Compomics utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* 2011, *12*, 70.

[135] Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., Martens, L., SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 2011, *11*, 996–999.

[136] Vaudel, M., Burkhart, J. M., Sickmann, A., Martens, L., Zahedi, R. P., Peptide identification quality control. *Proteomics* 2011, *11*, 2105–2114.

[137] 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*, 1552–1562.

[138] Martens, L., Resilience in the proteomics data ecosystem: how the field cares for its data. *Proteomics* 2013, *13*, 1548–1550.

[139] Perez-Riverol, Y., Hermjakob, H., Kohlbacher, O., Martens, L. et al., Computational proteomics pitfalls and challenges: HavanaBioinfo 2012 workshop report. *J. Proteomics* 2013, *87*, 134–138.

[140] Csordas, A., Wang, R., Rios, D., Reisinger, F. et al., From Peptidome to PRIDE: public proteomics data migration at a large scale. *Proteomics* 2013, *13*, 1692–1695.

[141] Gostev, M., Faulconbridge, A., Brandizi, M., Fernandez-Banet, J. et al., The BioSample database (BioSD) at the European Bioinformatics Institute. *Nucl. Acids Res.* 2012, *40*, D64–D70.

