Published in final edited form as:
Expert Rev Proteomics. 2022 December 18; 19(7-12): 297–310. doi:10.1080/14789450.2022.2160324.

Proteomic repository data submission, dissemination, and reuse: key messages

Yasset Perez-Riverol[1]

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

Abstract

The creation of ProteomeXchange data workflows in 2012 transformed the field of proteomics, consisting of the standardization of data submission and dissemination, and enabling the widespread reanalysis of public MS proteomics data worldwide. ProteomeXchange has triggered a growing trend toward public dissemination of proteomics data, facilitating the assessment, reuse, comparative analyses, and extraction of new findings from public datasets. By 2022, the consortium is integrated by PRIDE, PeptideAtlas, MassIVE, jPOST, iProX, and Panorama Public. More than 37,000 datasets have been submitted to ProteomeXchange and almost 70% are now publicly available. The success of ProteomeXchange and the amount of proteomics data available in the public domain have triggered the creation and/or growth of other protein knowledgebase resources such as ProteomicsDB, GPMDB, and MassIVE.quant, Expression Atlas, PeptideAtlas, Scop3P and others. This manuscript reviews the current ecosystem of resources, guidelines, and file formats for proteomics data dissemination and reanalysis. Special attention is drawn to new exciting quantitative and post-translational modification-oriented resources. Finally, the challenges and future directions on data depositions including the lack of metadata, and cloud-based and high-performance software solutions for fast and reproducible reanalysis of the available data are discussed.

Keywords

Proteomics databases; public proteomics data; protein expression; standard file formats; data reuse and reanalysis

Correspondence to: Yasset Perez-Riverol.

Contact: Yasset Perez-Riverol (yperez@ebi.ac.uk).

Geolocation information
Data sharing and reuse have become more common and standard for the proteomics community. This manuscript highlights major databases for storing proteomics data and the major challenges for data submission, dissemination, and reuse.

Disclosure statement
The authors report there are no competing interests to declare.

Data deposition
This manuscript does not contain any new data, no data was generated.

Europe PMC Funders Author Manuscripts

1 Introduction

Data sharing in proteomics has prompted a new relationship between the proteomics community and data. Researchers not only made public the data associated with their research manuscript but also complemented their research with the reanalysis of public data [1,2]. In a typical mass spectrometry (MS) proteomics experiment, the data generated is organized into multiple levels: (i) experimental design and sample metadata; (ii) the raw data acquired by the instruments; (iii) processed results, including peptide/protein identification and quantification values; and (iv) the resulting biological conclusions. In addition to the variety of analytical methods available in proteomics, the diversity of metadata that must be captured to understand each analytical method, a convoluted set of file formats are available in proteomics to store all the generated data (e.g., spectra, peptides, proteins quantification values, post-translational modifications).

In 2011, the ProteomeXchange (PX - http://www.proteomexchange.org) consortium [3] was founded to coordinate data dissemination, publication, and replication across multiple resources and databases which guarantees the sustainability of public proteomics data in the long term. In coordination with the HUPO Proteomics Standard Initiative (HUPO-PSI) [4], ProteomeXchange not only provides an infrastructure for data deposition but also guidelines for data dissemination, standardization, and the development of standard file formats to exchange proteomics data. By 2022 the consortium is integrated by the PRIDE database (EMBL-EBI, Hinxton, UK) [5], PeptideAtlas/PASSEL (Institute for Systems Biology, Seattle, USA) [6,7], MassIVE (University of California, USA), jPOST (the jPOST project, Japan) [8], iProX (Beijing Proteome Research Center, China) [9] and Panorama Public (University of Washington, Seattle, USA) [10]. ProteomeCentral (http://proteomecentral.proteomexchange.org) provides a central resource to search for datasets in all PX partners and a group of services and APIs that enable to retrieve the information from each resource partner (e.g., the universal spectrum identifier service - http://proteomecentral.proteomexchange.org/usi/) [11].

ProteomeXchange resources for almost 10 years have been focused on the dissemination of the original results associated with each publication. Recently, new resources emerging from the ProteomeXchange partners such as MassIVE.quant [12], or the ExpressionAtlas proteomics [13,14] has been started to systematically reanalyse the publicly deposited data with a special interest in new biological questions like protein differential expression or post-translational modifications. In addition, novel protein knowledgebase resources have been created or consolidated in recent years because of the immense number of high-quality datasets available through ProteomeXchange including ProteomicsDB [15], the Global Proteome Machine Database (GPMDB) [16], Scop3P [17], OpenProt [18], and the sORFs [19]. Some of these resources perform systematic reanalysis of public proteomics data, providing a standardized view of peptides and protein identifications, including post-translational modifications (PTMs) and single amino acid variants (SAAVs). However, in recent years, quantitative proteomics data has become more reproducible, accurate, and reliable, prompting the creation of multiple quantitative resources such as ProteomicsDB [15], and MassIVE.quant [12] and Expression Atlas [13,14], MatrisomeDB [20].

This manuscript aims 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. A review of all proteomics data archives, including the members of ProteomeXchange, is described in summary. ProteomeXchange update manuscripts [3,21,22] described in detail the new improvements and features on each resource and a previous review published in 2015 [2] explained the based guidelines for data submission and dissemination through ProteomeXchange. Instead, more details are provided about databases that reanalyse public proteomics to deliver protein expression profiles or information about posttranslational modifications or novel non-canonical peptides. In addition, a list of challenges related to integration, dissemination, and data mining/reuse are discussed; including the progress and challenges in systematic reanalysis of public proteomics data and how data should be deposited for better findability, reproducibility, and reuse.

2 ProteomeXchange and public proteomics data

The current ecosystem of proteomics resources can be classified into three different groups (Figure 1): (i) ProteomeXchange archives responsible for data submission guidelines, file format standardization and submission systems; (ii) peptide, protein identification resources including information about variants, mutations, or post-translational modifications; (iii) and protein expression resources including differential expression and baseline (absolute) expression profiles. Table 1 shows different proteomics databases organized by these categories. Most of the proteomics archive databases are part of ProteomeXchange including the PRIDE database (PRIDE Archive), MassIVE, PeptideAtlas/PASSEL, JPOST, Panorama Public and iProX. The only major archive of proteomics data outside ProteomeXchange is the Proteomics Data Commons (PDC) as part of the CPTAC Consortium [23].

Peptide and Protein sequence databases include peptide and protein evidence from MS-based proteomics experiments, and other sequence events such as posttranslational modifications, single amino acid variants or other non-canonical peptides. PeptideAtlas [7] and GPMDB [16] are the two leading resources in that group, which also includes MassIVE-KB [24], PRIDE Peptidome [5], sORFs [19], Scope3P [17], OpenProt [25] or HLA Ligand Atlas [26]. The number of protein expression databases is increasing with three major resources ProteomicsDB [15],MassIVE.quant [12] and ExpressionAtlas [14] followed by some more specific databases such as MatrisomeDB [20] and Immunological Proteomic Resource [27].

2.1 ProteomeXchange archives

Since 2012, the number of resources and members of PX has increased from 2 databases (PRIDE and PeptideAtlas/PASSEL) [3] to 6 resources [22]: the PRIDE database [5] and the PASSEL resource within PeptideAtlas [6,7], MassIVE, jPOST [8], iProX [9] and Panorama Public [10]. PX aims to provide a common framework for the cooperation of proteomics, with a special focus on the archival of original findings and data from submitters and proteomics researchers. The consortium’s members have agreed to provide a sufficient set of common experimental and technical metadata for each submitted dataset. All the submitted

datasets get a unique and universal identifier (PXD identifier) and by 2022, more than 22,000 datasets have been made public through ProteomeXchange (Figure 2A). For most of the following sections in this manuscript, the discussion around challenges and future directions for data deposition and dissemination will be focused on the PRIDE Archive rather than all PX databases because it remains the major archive for data deposition for the proteomics community (Figure 2A). A short description of the other resources is also included in this manuscript and more detailed information could be found in other reviews and publications [2,22].

2.1.1 PRIDE Archive—The PRIDE database [28] (EMBL-EBI, Hinxton, UK - https://www.ebi.ac.uk/pride/) is the major archive for public proteomics data (Figure 2A-B). The number of submissions to PRIDE Archive continues to increase at an extraordinary rate and since 2018, the average number of submissions has always been more than 250 per month [5]. The main resource of the PRIDE ecosystem is the PRIDE Archive which stores the data from submitters associated with their manuscripts and research. Proteomics data can be submitted to PRIDE as partial or complete submissions [29]. For partial submissions, the submitter can deposit the original results in the format exported by the analysis tool (e.g., MSF from ProteomeDiscoverer). In complete submissions, the results should also be provided in one of the standard HUPO-PSI file formats mzIdentML [30] or mzTab [31,32]. If mzIdentML and mzTab are used, the corresponding peak list and RAW files referenced by the RESULT files also need to be included. A “complete” submission ensures that the processed results data can be integrated into PRIDE but also that the community can read the reported results without the need to install software used by the submitters of the dataset.

Box 1 highlights the list of files that should be provided during the submission process to PRIDE Archive for better dissemination of the results of the experiment. Apart from the RAW files, the PRIDE Archive guidelines recommended providing the output results for the software used in the analysis; the FASTA or spectral library used to perform the identification and quantification analysis; configuration files (if available) for the software tools; the link to other relevant omics information should be provided (e.g., accession of RNA data deposited in GEO); and finally the sample to data relationship file format including the experimental design and the variables and conditions under study.

MassIVE (https://massive.ucsd.edu/) is the second-largest proteomics archive in ProteomeXchange. More importantly, MassIVE systematically replicates datasets from other PX partners such as PRIDE, iProX, or jPOST. By 2022, MassIVE reported over 10 thousand datasets between original submissions, replicates of datasets in other repositories, and reanalysis of existing public data. The MassIVE submission process is Web-based and enables the user to organize the files into the following categories: (i) license files; (ii) spectrum files; (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; (vi) methods and protocols (https://ccmsucsd.github.io/MassIVEDocumentation/#submit_data/).

jPOST (https://repository.jpostdb.org/) is the Japanese repository for proteomics data including mass spectrometry and antibodies data. By 2022, it stores 1582 projects including

224 species (around 55 Terabytes of data). jPOST has successfully implemented a web interface and easy-to-use submission system. The submission protocol has two major steps: *Project* - a group of files for one paper grouped into a project; and *Preset* - an experimental protocol (without files associated) -. Instead of using Aspera or FTP like other resources like MassIVE or PRIDE; jPOST has implemented a high-speed web transfer method comparable with existing commercial options (see Figure 2 in jPOST original publication [33]). All metadata submitted to jPOST is submitted to Resource Description Framework (RDF - https://integbio.jp/rdf/) data model, the proteome data could also be linked to a wide variety of other data, such as genomes and transcriptomes, under the 'linked-open data' concept. Like PRIDE Archive and MassIVE, jPOST has created a group of web interfaces to enable users to review and visualize the 'Complete' submissions. For example, if the dataset is a complete submission and the results are provided in mzIdentML or mzTab, users visualize in the 'peptide details' panel all the PSMs identified in the project including the protein accession, peptide sequence, search engine score and the link to the mass spectra use to identify the corresponding sequence (e.g., PEEVHHGEEEVETFAFQAEIAQLMSLIINTFYSNK - https://repository.jpostdb.org/peakview/?fileid=f_0000564405&projectid=JPST000814.0&id=1).

The Chinese Human Proteome Project (CNHPP) was launched to conduct human proteome research focusing on China's ten most prevalent cancers. To facilitate open access to proteome data worldwide, and to support proteomics research projects in China and beyond, the integrated proteome resource iProX (http://www.iProX.org) was created in [9]. Users can submit their proteomics datasets to iProX as public or private datasets. By 2022, 1060 proteomics datasets have been submitted to iProX and 635 are public. Web-based and fast Aspera (https://asperasoft.com/) based transfer protocols are offered by the iProX database for submitter data upload. The web interface is used to straightforwardly upload small files, whereas the fast Aspera-based transfer method is recommended for multiple large files, which is common in proteomic studies.

Panorama Public [10] (https://panoramaweb.org/project/Panorama%20Public/begin.view) is the first ProteomeXchange mainly focused on quantitative proteomics data deposition and sharing. Panorama Public is built on Panorama, an open-source data management system for mass spectrometry data processed with the Skyline [34] targeted mass spectrometry environment. The main file format contained in the Panorama Public submission is the corresponding Skyline output (.sky.zip - https://panoramaweb.org/Panorama%20Public/2020/PNNL%20%20EGFR%20pathway%20peptide%20standards%20at%20different%20DDM%20concentrations/project-begin.view). In addition to the main result file format, the submission should contain the RAW files used to identify perform the data analysis or a link to the RAW data files in other databases such as PRIDE Archive. One of the key features of Panorama Public is the possibility to explore in detail the quantitative results of a proteomics experiment (e.g., chromatograms). Within Panorama, a user can view a list of the targeted analytes along with chromatograms and marked peak boundaries for each analyte in all the replicates (e.g., https://panoramaweb.org/Panorama%20Public/2020/PNNL%20%20EGFR%20pathway%20peptide%20standards%20at%20different%20DDM%20concentrations/targetedms-showPeptide.view?id=25679913). In addition, for each

experiment resources present the list of proteins, peptides, precursors, transitions, replicates, and calibration curves.

2.2 Sample and experimental metadata

A special mention should be made of how experimental design and sample metadata has been stored in ProteomeXchange archives and the new file format and data model SDRF (Sample to Data Relationship format) [35]. In 2015, *Griss et. al.* [36] highlighted that the lack of metadata in ProteomeXchange repositories - especially in quantitative experiments - made difficult the reuse of public proteomics data. While ProteomeXchange captures metadata like species, instruments or tissues, the lack of a sample metadata model and experimental design format make it difficult to associate the sample to the specific raw file in a dataset.

The MAGE-TAB for proteomics (including the SDRF format) was developed by the ProteomeXchange databases and HUPO-PSI to enable submitters of proteomics data and users of those databases to exchange the sample metadata for each dataset. Sample to Data Relationship format (SDRF - https://github.com/bigbio/proteomics-metadata-standard) captures the sample metadata, the relation of each sample and the data file acquired in the experiment. The SDRF is a tab-delimited file format where each column is the sample or the data properties, or the factor values (the condition or variables under study). Each row of the file corresponds to the relation between the sample and the data file. SDRF is an ontology-based format, that uses controlled vocabularies to describe the properties and the values for each sample and data file. The ontology lookup service (OLS) [37] can be used to annotate the SDRF files and a set of libraries are available in Python and Java languages to validate the submitted files to ProteomeXchange. By 2022, more than 150 datasets in PRIDE Archive were annotated using SDRF files and a list of experiments from different MS analytical approaches can be found in the GitHub repo (https://github.com/bigbio/proteomics-metadata-standard/tree/master/annotated-projects).

2.3 Data submission challenges

Since 2019, PRIDE supports the mzTab file format, which contains protein expression/quantification values in addition to the identification information. Since the mzTab creation in 2014, three major tools have supported the file format: Mascot [38], the protein inference toolkit (PIA) [39,40], OpenMS [41] and MaxQuant/MaxDIA [42,43]. Because mzTab is the gold standard for quantitative proteomics, more software tools and bioinformatics frameworks should support the format as an output of the tool. While proteomics has been focussing on the quantitative profile of proteins since the first analytical methods for MS-based analysis, still most results of complete submissions are only identification (peptide and protein identification). MzIdentML file format (only including peptide/protein identification information) has been more widely adopted by major software packages and tools (https://www.psidev.info/tools-implementing-mzidentml). In the meantime, it is recommended that the users submit to the PRIDE Archive, the input and output of quantitative software packages such as MSstats [44] or Perseus [45].

Europe PMC Funders Author Manuscripts

Figure 2B shows that the number of submissions in PRIDE Archive is growing every year but also the number of large datasets (more than 100 MS runs) submissions. With the increase in data size, multiple problems arise including data transfer (uploading/downloading) performance, data processing (validation, submission), and data consistency validation (e.g., file checksum check). All major archives in ProteomeXchange support multiple protocols for data transferring including FTP, HTTPS or Aspera; however, the volume of the data may cause delays during the submission. The data submitter must start working with the PX resource with sufficient time that does not delay the publication process. In PRIDE Archive, for example, the user can start loading the data to the database incrementally by requesting to PRIDE team and FTP account and perform the final submission when the data is already transferred. This mechanism has been used to submit datasets with more than 1000 MS run files. The PRIDE Archive database is exploring new and faster ways for data transfer of large datasets, including new protocols such as Globus (https://www.globus.org/).

3 Reanalysis of public proteomics data

The mission of ProteomeXchange is to make the proteomics datasets and the final research: findable, reproducible, and reusable. The data deposition guidelines and file formats required to perform a submission (Box 1) are designed to facilitate these three pillars of open data. Figure 3A shows the distribution of reanalysis performed by independent researchers in different types of omics data including transcriptomics (GEO, ArrayExpress), proteomics (PRIDE Archive, MassIVE, jPOST) and genomics (EGA - European Genotype-phenome Archive). On average, public transcriptomics datasets are reused 3 times more than proteomics datasets (average number of reanalyses by omics type: 8.6 Transcriptomics, 8 Genomics, 3 Proteomics). For example, the GEO dataset GSE14333 (Expression data from 290 primary colorectal cancers [46]) has been analysed by the community more than 300 times, while the most reanalysed dataset in proteomics, PXD000561 (A draft map of the human proteome [47]) has been reanalysed 34 times.

In Figure 3B, all the ProteomeXchange datasets that have been reanalysed at least once by an independent researcher (direct citation of the dataset accession in a manuscript) or one of the proteomics databases that systematically performs reanalysis of public proteomics data (GPMDB, ProteomicsDB, Expression Atlas, PeptideAtlas, Scope3P, MassIVE). More than 4200 (20% of the publicly available) datasets in ProteomeXchange have been reanalysed at least once. GPMDB and PeptideAtlas together have reanalysed more than 94% of the datasets, while the others only have 6% of the data. GPMDB, PeptideAtlas (as other repositories like MassIVE-KB [24] or PRIDE Peptidome [5]) are mainly focused on protein/peptide identification results including post-translational modifications and sequence amino acid variants (SAAVs) and not on quantitative results. However, if the numbers do not look promising, the amount of data reprocessed and the number of resources providing identification and quantitative proteomics data are growing.

3.1 Proteomics and peptide identification resources

Protein and peptide identifications databases have been focused on providing information about protein sequences including peptide sequence evidence, post-translational modifications, and amino acid variants [2,7,16,28]. GPMDB [16,48] and PeptideAtlas [7] are the most established and well-known databases that systematically reanalyse public proteomics data using standard pipelines. However, two other resources, MassIVE-KB and PRIDE Peptidome have also started providing peptide identifications from public proteomics datasets. By 2022, PeptideAtlas contains 2,954,162, 485,199, and 535,340 unique peptides for Human, Yeast and Arabidopsis (among other species and builds), while GPMDB provided more than 22 million peptides for multiple species. In general, PeptideAtlas contains less evidence than GPMDB, however, PeptideAtlas pipelines and workflows uses more stringent quality control and statistical threshold for every piece of evidence including sequence amino acid length thresholds, false positive rate depending on the build size [49].

For almost 10 years, the focus of PeptideAtlas and GPMDB has been providing peptide evidence for most of the proteins of the human proteome including isoforms [2,50]. Recently both resources have been pushing the frontiers towards post-translation modifications (PTMs) including more data for phospho-proteomics experiments and other PTMs [51] (http://psyt.thegpm.org/psyt/index.html). One of the most important features of the GPMDB workflow, based on the X!Tandem search engine [52], is the possibility to accurately identify unexpected post-translational modifications (e.g., phosphorylation – Figure 4A). At the same time, PeptideAtlas (Figure 4B) has increased the number of phospho-proteomics reanalyses and the development of new tools for quality control of phospho-sites including the release of PTMPhrophet [51,53,54].

In addition to GPMDB and PeptideAtlas, other novel resources have been recently created with a special focus on PTM evidence from MS-based public data: Scop3P [17], MassIVE-KB [24] or PRIDE Peptidome [5]. Scop3P (https://iomics.ugent.be/scop3p/) integrates sequences (UniProtKB/Swiss-Prot), sequence structures (PDB), and uniformly reprocessed phosphoproteomics data (PRIDE) to annotate all known human phosphosites. The ionbot search engine, which uses MS2Rescore and MS2PIP to boost the number of peptide identifications, was used to reanalyse more than 30 phosphoproteomics datasets from the PRIDE database. The Scope3P workflow uses the PhosphoRS algorithm [55] to compute the localization probability for the P-sites. If there are multiple peptide spectrum matches (PSMs) or multiple peptides for a given P-site, then the P-site was only included if a site probability of at least 0.5 was found in at least one of these identifications. The Scop3P web interface allows queries by protein accession or PX accession to navigate the phospho evidence for each protein. Figure 3C presents a Venn diagram of the number of phospho-sites present in PeptideAtlas compared with Scop3P. While both resources shared 36,546 phospho peptides, PeptideAtlas contains 45’908 unique phosphopeptides, while Scop3P stores 36,885 unique evidence.

MassIVE-KB (https://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp) [24] and PRIDE Peptidome (https://www.ebi.ac.uk/pride/peptidome) [5] are two new resources from ProteomeXchange partners MassIVE and PRIDE to provide peptide and protein

identification evidence from original submitted datasets or reanalyses. MassIVE-KB uses clustering to analyze 227 human datasets, and 27'404 LC/MS runs to finally obtain 2.1 million high-quality precursors (peptide + modifications + charge state) representing 19,610 human proteins. PRIDE Peptidome uses a combination of clustering and the protein inference tool (PIA) to access the quality of each PSM. The best peptide per project is selected based on three rules: (i) the peptidoform passes the peptide FDR threshold for the assay; (ii) the cluster where the peptidoform belongs only contains that peptidoform; and (ii) the peptide sequence is longer than seven amino acids. The sparkMS (https://github.com/bigbio/sparkms) used Spark (https://spark.apache.org/) and PySpark to group millions of PSMs in less than 6 hours, which enabled the data analysis of such a large-scale amount of data. PRIDE Peptidome is not only focused on human datasets by containing other species such as Mouse, Arabidopsis, Gallus, etc.

Proteogenomics is the study of how genetic information encoded in the genome influences the expression and function of proteins [56], while immunopeptidomics is the collective identification and quantification using MS-based proteomics of sample-specific repertoires of HLA-presented [57]. In addition to the study of posttranslational modifications, public data reuse has been focused on the study of non-canonical peptides including small open reading frames (ORFs), inmunopeptides or single amino acid variants (SAAVs) [56,58]. OpenProt (https://www.openprot.org/) [18] is a resource that reuses public proteomics data to annotate many non-annotated open reading frames (ORFs) in eukaryotic genomes. In summary, the OpenProt identification workflow uses the SearchGUI tool [59] and a FASTA database created from all predicted open reading frames from ENSEMBL genomes [60,61]. By 2022, the OpenProt database had reanalysed around 170 ProteomeXchange datasets for 10 different species. An intuitive web interface is provided for every peptide from MS evidence that supports an ORFs (e.g. https://openprot.org/p/altorfDbView/79/43095155/596702/ENSP00000365998.2/2/msDetectionDetail). However, no link to the spectra visualization is provided which decreases the possibility to validate and visual check novel non-canonical peptides, a mechanism that has been suggested by multiple authors for novel peptides [11,56].

Similarly, sORFs (http://www.sorfs.org/) [19] developed the concept of PRIDE-reSpin, a pipeline that uses SearchGUI [59] and PeptideShaker [62] to reanalyse public proteomics data from PRIDE database to identify small open reading frames (sORFs). The protein FASTA database is created from ribosome profiling (RIBO-seq) and a novel noise filtering algorithm is used to distinguish sORFs with true ribosomal activity from simulated noise, consequently reducing the false identification rate. Importantly, the sORFs web interface provides for every peptide evidence the corresponding link to the spectra making it easier for the user to check the quality of the MS supporting the evidence.

Interestingly, while the number of resources is growing, the interoperability efforts and standards are also consolidating. For example, for many years proteomics resources and proteomics datasets use to employ different Protein names and identifier systems including Uniprot, ENSEMBL, RefSeq, IPI, UniRef100, etc [63]. Currently, all major identification resources use two major reference systems Uniprot (Accession or Protein Name), and ENSEMBL (Gene Accession or Gene Name). All the databases previously mentioned

including PeptideAtlas, GPMDB, proteomicsDB, PRIDE Peptidome or MassIVE-KB employ one of these two annotations.

3.2 Bioinformatic solutions for proteomics data reanalysis

In recent years multiple bioinformatics tools and workflows have been used to systematically reanalyse public proteomics data. In 2010, Trans-proteomic pipeline [64] was the first distributed system that enables the reanalysis of public datasets and has been used for more than 10 years to build PeptideAtlas. Trans-proteomic pipeline has been predominantly used in peptide and protein identifications and PTM localization analyses [53,54] by the PeptideAtlas team. Similarly, PeptideShaker [62] and the MassIVE-KB reanalysis interface (https://ccms-ucsd.github.io/MassIVEDocumentation/#reanalyze_spectra/) enables users of the PRIDE database and MassIVE to perform peptide/protein identification reanalysis in a semi-automatic approach. Both platforms use different search engines such as MS-GF+ [65], Comet [66] or PepNovo [67] to perform the peptide identification and finally transform the results into standard file formats such as mzIdentML. Different to identification approaches, reanalysis of quantification data has been dominated by MaxQuant [13,15]. ProteomicsDB uses MaxQuant to compute the IBAQ values, while PRIDE/ExpressionAtlas has been using MaxQuant to get the iBAQ values and then compute a normalised iBAQ value for each protein [13]. While these three software solutions are available for the community, quantitative proteomics reanalysis is still not a common practice like other fields such as transcriptomics [68,69]. In order to be able to large-scale reanalyse public proteomics data new solutions should be available that enable the parallelization of each independent step of the proteomics data analysis in cloud and distributed computers; and new algorithms and file formats must be developed to enable fast reprocessing of huge volumes of data [70,71].

3.3 Challenges to systematic identification-based reanalysis

Large-scale reanalysis of proteomics performing only the identification of peptides, proteins and PTMs (no quantitative analysis), presents two big challenges. First, peptide/protein identification is computationally expensive algorithmically, most of the algorithms and search engines for peptide identification, protein inference and PTM localization demand CPU and memory resources [65,66]. Only a few solutions like Trans-proteomic Pipeline [64] enable to analyse of large-scale datasets using high-performance computing or cloud options, enabling the analysis of thousands of datasets [70]. However, the panorama has been changing considerably in recent years with the emergence of faster peptide identification search engines [72,73] and new cloud-based workflows that enable the analysis of large datasets [58,74–76]. The second challenge around large-scale reanalysis of public proteomics datasets – including the aggregation of millions of peptide-spectrum matches (PSMs) – is the statistical significance and quality control of the identification results [77–79]. The Human Proteome Project (HPP) every year release a set of guidelines and metrics to access the quality and reliability of public proteome evidences (peptides and proteins), the latest version is from 2019-2022 [80,81]. The most established metrics are 1% FDR at the protein level, 2 unique peptides per protein and more than 9 amino acids for each peptide identified [50,79,81].

Four major approaches are now well-established to control the false positives in large-scale proteomics data integration projects: (i) stringent FDR control at the level of the dataset that produces low FDR at the total integration level combined with quality control rules such as peptide amino acid length, number of unique peptides per proteins [6,49,50,79], this method is implemented by PeptideAtlas; (ii) FDR calculation after integrating all PSMs from the datasets integrated [77] (implemented by ProteomicsDB [15]); (iii) Spectra clustering of the PSM identifications and stringent filtering of the low-quality clusters [24,82,83], used by MassIVE-KB and PRIDE Peptidome; (iv) control the FDR at the project level and provide visualization tools for exploring the data quality. All the methods have their strengths and weaknesses; while high-quality and well-controlled FDR is desired in database integrations, sometimes it is not technically possible because the scale of the data or even the statistical methods are not available. For example, GPMDB stores billions of PSMs across multiple experiments making challenging to scale FDR calculations and some of those statistical methods may affect some relevant biological peptides from unique experiments that are extremely important in exploratory research. In the same way, new algorithms have been scaled for independent dataset analysis, and more tools and architectures should be built/developed to tackle these challenges. Box 2 provides a group of guidelines that can facilitate individual users to evaluate the quality and statistical significance of independent or integrated reanalyses.

While different rules are in place for quality control on different databases and resources; some consensuses exist across all MS-based evidence resources which enable users to validate the findings: the provenance of the data must be available – PX accession for each reanalysis-; FDR must be controlled at the level of the PSM at least on each independent dataset; visualization tools to inspect the spectra associated with each peptide evidence should be provided. All major resources including PeptideAtlas, MassIVE-KB, PRIDE Peptidome, and ProteomicsDB include visualization components for the spectra. However, efforts should be made by individual resources to increase interoperability and cross references among resources. For example, some of the databases (e.g., sORFs.org) do not contain APIs to programmatically query the information of the resource, and others such as OpenProt do not provide the spectrum reference for each novel reference. Recently ProteomeXchange released a new format and service, the Universal Spectrum Identifier (USI) to enable the exchange of mass spectra across all resources [11]. In addition, the ProteomeXchange consortium is working on ProXI (ProteomeXchange Interface - http://proteomecentral.proteomexchange.org/PROXI.php), a common language representation and API definition to exchange proteomics datasets.

4 Protein expression resources and reanalysis

Figure 3B shows that more than 96% of the reanalyses are performed on the identification of peptides, proteins and PTMs. While most public datasets are quantitative studies, only 6% of the total reanalyses performed in the field are quantitative (baseline/absolute or differential expression). However, this panorama is changing with the rise of novel protein expression resources [12,14,15] and new file formats that enable to submit the more sample metadata and the experimental design for each experiment [35]. MS-based protein expression resources are not new in proteomics [2]. Multiple databases have been developed to present

expression from public datasets or the data generated by independent labs including MaxQB [84], PaxDB [85], MOPED [86] and others [2]. These resources have been mainly focusing on two different types of expression profiles: protein differential expression; and absolute/baseline expression (Figure 1). Differential expression in all resources has been mainly based on intensity-based label-free or TMT log-fold change ratios; while baseline/absolute expression is more dissimilar and has been mainly based on different methods like spectral counting, IBAQ intensity values [87]; spectral counting [88]; or the proteomic ruler [89].

MaxQB (http://maxqb.biochem.mpg.de/mxdb/) and PaxDB (https://pax-db.org/) are pioneering databases providing baseline protein expression data. MaxQB uses in-house data previously processed with the MaxQuant tool [43] and provides for every protein the IBAQ values on each specific sample/condition (e.g. http://maxqb.biochem.mpg.de/mxdb/protein/show?sourceId=P04439 – Protein Expression Tab). By 2022, MaxQB contains more than 880 humans, and 500 mouse experiments covering more than 30,000 proteins. Differently, PaxDB uses the data previously reanalysed by PeptideAtlas and uses spectral counting as the value for the absolute/baseline expression of each reported protein. PaxDB provides protein expression data for more species than any resource available. However, the method used has been proven (spectral counting) to be less accurate than the intensity-based methods such as IBAQ, or proteomics ruler. These two resources haven't been updating in since 2015, then more information about their pipelines and submission systems can be read in a previous work [2].

Three different databases are leading the current efforts for protein expression reanalysis: ProteomicsDB, MassIVE.quant and Expression Atlas. ProteomicsDB (https://www.proteomicsdb.org) was the first database that started reanalysing proteomics data from the public domain and visualising it in a common interface. ProteomicsDB is mainly focused on baseline/absolute expression based on IBAQ values and currently stores data for four species (Human, Mouse, *Arabidopsis thaliana*, and *Oryza sativa*). ProteomicsDB has analysed almost 50 experiments from PX archives complementing other experiments from their instruments. ExpressionAtlas and PRIDE teams have started to integrate reanalyses from PRIDE Archive public datasets into other resources - Open Targets [90] - that also contain other omics types like genomics or gene expression data. In ExpressionAtlas (https://www.ebi.ac.uk/gxa), two types of proteomics profiles are provided; differentially express profiles and baseline profiles based on the IBAQ values [13]. Datasets are reanalysed with a standard pipeline using MaxQuant [13] or OpenSWATH [75] and submitted to the ExpressionAtlas database. By 2022, the PRIDE team in collaboration with ExpressionAtlas had managed to analyse more than 85 experiments.

Different to ProteomicsDB and PRIDE/ExpressionAtlas, where a standard data analysis pipeline has been used for all datasets, MassIVE.quant (https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp) has implemented a different strategy, where MassIVE users can reanalyse existing data on MassIVE or submit their reanalysis from public proteomics data. Reanalyses can be submitted from different analysis tools including MaxQuant or Spectronaut, and MSstats is used for the differential expression analysis and statistical assessment of the quantitative results [44]. MSstats has been integrated with a

lot of tools and quantification workflows [34,91,92] enabling MassIVE.quant to grip more users, data pipelines and reanalyses.

Finally, it is important to mention two new resources Matrisomedb [20] (http://matrisomedb.pepchem.org/) and the Immunological Proteome Resource [27] (http://immpres.co.uk/). The two protein expression resources, different to ProteomicsDB, MassIVE.quant and Expression Atlas, are focused on specific protein families or diseases. Matrisomedb is a searchable collection of curated proteomic datasets from 17 studies on the extracellular matrix (ECM) of 15 different normal tissue types, six cancer types (different grades of breast cancers, colorectal cancer, melanoma, and insulinoma) and other diseases including vascular defects and lung and liver fibrosis. The original datasets in the PRIDE Archive were reanalysed using the ProLuCID search engine [93] and the quantitative value for each protein is the sum of all MS1 signal intensities of all its identified peptides. Immunological Proteome Resource is a database focused on immunoproteins from in-house (no public proteomics data) datasets. The workflow contains datasets from TMT and Label-free analyses, and MaxQuant is used to estimate the MS1 intensities for each protein on each sample and the final quantification value is based on the proteomics ruler algorithm [89].

5 Deep learning and public data

Deep learning is having a massive impact on the biomedical research [94], particularly in proteomics [95,96]. In 2019, two major works triggered the use of identified spectra from public datasets or synthetic libraries to predict the theoretical spectra from peptide sequences using deep learning models Prosit [97] and DeepMass [98]. Both models focused on the prediction of only b/year fragment ions and were mostly applied to tryptic peptides. Prosit used was on the ProteomeTools synthetic peptide resource of 550,000 tryptic peptides measured by 21 million tandem mass spectra at various collision energies. Prosit and DeepMass were originally designed to generate spectral libraries for data-independent acquisition (DIA) proteomics experiments for any organisms based only on protein sequences, triggering a group of novel algorithms for library-free DIA analysis [96]. Library-free algorithms and tools such as DIA-NN [99], DeepDIA [100] and MaxDIA [42] have transformed DIA data analysis to make it more scalable, easier to perform by non-experts and less expensive because the DDA is not needed. All Prosit and ProteomeTools datasets have been deposited on PRIDE for other developers and algorithms that would like to develop their models (e.g., https://www.ebi.ac.uk/pride/archive?filter=project_keywords_facet%3D%3DProsit).

In addition to spectra prediction algorithms, a set of novel algorithms to predict the retention time of peptides like AutoRT [101], and DeepLC [102] and to boost the number of identification in proteogenomics and inmunopeptidomics studies such DeepRescore [103] and MS2Rescore [104] has been recently released. The DeepLC model for example was developed using multiple ProteomeXchange datasets (PXD000953, PXD000954, PXD006932). Clustering algorithms have benefited from deep-learning models enabling the clustering of millions of spectra in the MassIVE and PRIDE database [105,106].

Deep learning and public proteomics data continue to grow together. Recently, a group of researchers created the proteomicsML community (https://www.proteomicsml.org/) [107] to standardize and collect different datasets used for machine learning in proteomics (https://www.ebi.ac.uk/pride/archive?filter=project_tags_facet%3D%3DBenchmarking,project tags facet%3D%3DMachine%20learning) and the models associated with them. A set of tutorials are provided to researchers to learn how to build and deploy machine learning models and can be easily executed in Google Colab (e.g. https://www.proteomicsml.org/tutorials/fragmentation/nist-1-parsing-spectral-library.html).

6 Conclusions and future directions

The growing amount of publicly available proteomics data is triggering the creation of more proteomics databases and the integration between MS-based proteomics and other knowledgebase resources like UniProt, Open Targets and Ensembl. The success and continued growth of ProteomeXchange guidelines and resources have made possible the availability of thousands of datasets that can respond to various biological questions from protein expression to protein structural variations and posttranslational modifications. All these data are systematically reanalysed and reused by independent researchers and other databases and resources that use mainly different tools and workflows to query the data with different biological questions.

Most of the reanalysis and reuse of public proteomics has been focused on peptide and protein sequence identification. Databases such as GPMDB and PeptideAtlas have been systematically reanalysing public proteomics data for over 10 years with their pipelines covering more than 20% of the publicly available data (3972 PX datasets). Apart from peptide identification, both resources and others like Scop3P have switched attention to other protein sequence questions such as posttranslational modifications and single amino acid variants. The development of the Universal Spectrum Identifier standard and its implementation by multiple PX resources enables checking and visualising the spectral evidence in support of key findings in publications and public data, which can mitigate the challenges behind the statistical assessment of large-scale datasets.

Sample to Data relationship format (SDRF) within the MAGE-TAB for Proteomics will change how ProteomeXchange repositories capture the sample metadata and the experimental design. Especially in quantitative data reuse, the lack of metadata and the correct experimental design plays a major role, making the reanalysis difficult and sometimes impossible. Three major efforts, ProteomicsDB, ExpressionAtlas/PRIDE and MassIVE.quant, are rea-annotating public datasets or enabling users to submit quantification reanalysis following a set of specific guidelines. Differential expression profiles and baseline expression across different tissues can be found in some of these resources leading the field from the identification to the quantitative world.

Finally, large-scale reprocessing workflows, tools and algorithms must grow and flourish to perform massively large-scale reanalyses. Such capabilities currently remain limited to a couple of dedicated proteomics bioinformatics groups. However, as the data have been generated by the community, and thus belong to the community, large-scale reprocessing

should also be made available to the general community. Only then can we start to realize the full potential of the publicly shared proteomics data.

Acknowledgements

I would like to thank Lennart Martens' team, Eric Deutsch, and Ronald Beavis for providing the number of datasets reanalysed by the resources and list of phospho peptides. Thanks to Juan A. Vizcaino for the feedback and discussions about ProteomeXchange repositories.

Abbreviations

IBAQ	Intensity-Based Absolute Quantitation
MS	Mass spectrometry
PRIDE	Proteomics Identification Database
PTM	Post-translational modification
PX	ProteomeXchange
SDRF	Sample to Data relationship format
TMT	Tandem mass tag

References

1. Martens L, Vizcaino JA. A Golden Age for Working with Public Proteomics Data. Trends Biochem Sci. 2017; 42 (5) 333–341. [PubMed: 28118949]
2. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015; 15 (5-6) 930–949. [PubMed: 25158685]
3. Vizcaino JA, Deutsch EW, Wang R, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014; 32 (3) 223–226. [PubMed: 24727771]
4. Deutsch EW, Orchard S, Binz PA, et al. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. J Proteome Res. 2017; 16 (12) 4288–4298. [PubMed: 28849660]
5. Perez-Riverol Y, Bai J, Bandla C, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 2022; 50 (D1) D543–D552. [PubMed: 34723319]
6. Desiere F, Deutsch EW, King NL, et al. The PeptideAtlas project. Nucleic Acids Res. 2006; 34: D655–658. [PubMed: 16381952]
7. Deutsch EW. The PeptideAtlas Project. Methods Mol Biol. 2010; 604: 285–296. [PubMed: 20013378]
8. Moriya Y, Kawano S, Okuda S, et al. The jPOST environment: an integrated proteomics data repository and database. Nucleic Acids Res. 2019; 47 (D1) D1218–D1224. [PubMed: 30295851]
9. Chen T, Ma J, Liu Y, et al. iProX in 2021: connecting proteomics data sharing with big data. Nucleic Acids Res. 2022; 50 (D1) D1522–D1527. [PubMed: 34871441]
10. Sharma V, Eckels J, Schilling B, et al. Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. Mol Cell Proteomics. 2018; 17 (6) 1239–1244. [PubMed: 29487113]
11. Deutsch EW, Perez-Riverol Y, Carver J, et al. Universal Spectrum Identifier for mass spectra. Nat Methods. 2021; 18 (7) 768–770. [PubMed: 34183830]

12. Choi M, Carver J, Chiva C, et al. MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. Nat Methods. 2020; 17 (10) 981–984. [PubMed: 32929271]
13. Jarnuczak AF, Najgebauer H, Barzine M, et al. An integrated landscape of protein expression in human cancer. Sci Data. 2021; 8 (1) 115. [PubMed: 33893311]
14. Moreno P, Fexova S, George N, et al. Expression Atlas update: gene and protein expression in multiple species. Nucleic Acids Res. 2022; 50 (D1) D129–D140. [PubMed: 34850121]
15. Samaras P, Schmidt T, Frejno M, et al. ProteomicsDB: a multi-omics and multiorganism resource for life science research. Nucleic Acids Res. 2020; 48 (D1) D1153–D1163. [PubMed: 31665479]
16. Fenyo D, Beavis RC. The GPMDB REST interface. Bioinformatics. 2015; 31 (12) 2056–2058. [PubMed: 25697819]
17. Ramasamy P, Turan D, Tichshenko N, et al. Scop3P: A Comprehensive Resource of Human Phosphosites within Their Full Context. J Proteome Res. 2020; 19 (8) 3478–3486. [PubMed: 32508104]
18. Brunet MA, Brunelle M, Lucier JF, et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res. 2019; 47 (D1) D403–D410. [PubMed: 30299502]
19. Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2018; 46 (D1) D497–D502. [PubMed: 29140531]
20. Shao X, Taha IN, Clauser KR, Gao YT, Naba A. MatrisomeDB: the ECM-protein knowledge database. Nucleic Acids Res. 2020; 48 (D1) D1136–D1144. [PubMed: 31586405]
21. Deutsch EW, Csordas A, Sun Z, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 2017; 45 (D1) D1100–D1106. [PubMed: 27924013]
22. Deutsch EW, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: enabling ‘big data’ approaches in proteomics. Nucleic Acids Res. 2020; 48 (D1) D1145–D1152. [PubMed: 31686107]
23. Wu P, Heins ZJ, Muller JT, et al. Integration and Analysis of CPTAC Proteomics Data in the Context of Cancer Genomics in the cBioPortal. Mol Cell Proteomics. 2019; 18 (9) 1893–1898. [PubMed: 31308250]
24. Wang M, Wang J, Carver J, Pullman BS, Cha SW, Bandeira N. Assembling the Community-Scale Discoverable Human Proteome. Cell Syst. 2018; 7 (4) 412–421. e415 [PubMed: 30172843]
25. Brunet MA, Lucier JF, Levesque M, et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res. 2021; 49 (D1) D380–D388. [PubMed: 33179748]
26. Marcu A, Bichmann L, Kuchenbecker L, et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. J Immunother Cancer. 2021; 9 (4)
27. Brenes AJ, Hukelmann JL, Spinelli L, et al. The Immunological Proteome Resource. bioRxiv. 2022. 2022.2008.2029.505666
28. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019; 47 (D1) D442–D450. [PubMed: 30395289]
29. Ternent T, Csordas A, Qi D, et al. How to submit MS proteomics data to ProteomeXchange via the PRIDE database. Proteomics. 2014; 14 (20) 2233–2241. [PubMed: 25047258]
30. Vizcaino JA, Mayer G, Perkins S, et al. The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics. Mol Cell Proteomics. 2017; 16 (7) 1275–1285. [PubMed: 28515314]
31. Hoffmann N, Rein J, Sachsenberg T, et al. mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. Anal Chem. 2019; 91 (5) 3302–3310. [PubMed: 30688441]

Europe PMC Funders Author Manuscripts

32. Griss J, Jones AR, Sachsenberg T, et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics. 2014; 13 (10) 2765–2775. [PubMed: 24980485]
33. Okuda S, Watanabe Y, Moriya Y, et al. jPOSTrepo: an international standard data repository for proteomes. Nucleic Acids Res. 2017; 45 (D1) D1107–D1111. [PubMed: 27899654]
34. Pino LK, Searle BC, Bollinger JG, Nunn B, MacLean B, MacCoss MJ. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. Mass Spectrom Rev. 2020; 39 (3) 229–244. [PubMed: 28691345]
35. Dai C, Fullgrabe A, Pfeuffer J, et al. A proteomics sample metadata representation for multiomics integration and big data analysis. Nat Commun. 2021; 12 (1) 5854 [PubMed: 34615866]
36. Griss J, Perez-Riverol Y, Hermjakob H, Vizcaino JA. Identifying novel biomarkers through data mining-a realistic scenario? Proteomics Clin Appl. 2015; 9 (3-4) 437–443. [PubMed: 25347964]
37. Perez-Riverol Y, Ternent T, Koch M, et al. OLS Client and OLS Dialog: Open Source Tools to Annotate Public Omics Datasets. Proteomics. 2017; 17 (19)
38. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20 (18) 3551–3567. [PubMed: 10612281]
39. Uszkoreit J, Perez-Riverol Y, Eggers B, Marcus K, Eisenacher M. Protein Inference Using PIA Workflows and PSI Standard File Formats. J Proteome Res. 2019; 18 (2) 741–747. [PubMed: 30474983]
40. Uszkoreit J, Maerkens A, Perez-Riverol Y, et al. PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. J Proteome Res. 2015; 14 (7) 2988–2997. [PubMed: 25938255]
41. Pfeuffer J, Sachsenberg T, Alka O, et al. OpenMS - A platform for reproducible analysis of mass spectrometry data. J Biotechnol. 2017; 261: 142–148. [PubMed: 28559010]
42. Sinitcyn P, Hamzeiy H, Salinas Soto F, et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. Nat Biotechnol. 2021; 39 (12) 1563–1573. [PubMed: 34239088]
43. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26 (12) 1367–1372. [PubMed: 19029910]
44. Choi M, Chang CY, Clough T, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014; 30 (17) 2524–2526. [PubMed: 24794931]
45. Tyanova S, Temu T, Sinitcyn P, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016; 13 (9) 731–740. [PubMed: 27348712]
46. Jorissen RN, Gibbs P, Christie M, et al. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. Clin Cancer Res. 2009; 15 (24) 7642–7651. [PubMed: 19996206]
47. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature. 2014; 509 (7502) 575–581. [PubMed: 24870542]
48. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004; 3 (6) 1234–1242. [PubMed: 15595733]
49. van Wijk KJ, Leppert T, Sun Q, et al. The Arabidopsis PeptideAtlas: Harnessing worldwide proteomics data to create a comprehensive community proteomics resource. Plant Cell. 2021; 33 (11) 3421–3453. [PubMed: 34411258]
50. Omenn GS, Lane L, Lundberg EK, Beavis RC, Overall CM, Deutsch EW. Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications. J Proteome Res. 2016; 15 (11) 3951–3960. [PubMed: 27487407]
51. Kalyuzhnyy A, Eyers PA, Eyers CE, et al. Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation. J Proteome Res. 2022; 21 (6) 1510–1524. [PubMed: 35532924]
52. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004; 20 (9) 1466–1467. [PubMed: 14976030]

53. Shteynberg DD, Deutsch EW, Campbell DS, et al. PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. J Proteome Res. 2019; 18 (12) 4262–4272. [PubMed: 31290668]
54. Ramsbottom KA, Prakash A, Riverol YP, et al. Method for Independent Estimation of the False Localization Rate for Phosphoproteomics. J Proteome Res. 2022; 21 (7) 1603–1615. [PubMed: 35640880]
55. Taus T, Kocher T, Pichler P, et al. Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res. 2011; 10 (12) 5354–5362. [PubMed: 22073976]
56. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014; 11 (11) 1114–1125. [PubMed: 25357241]
57. Chong C, Coukos G, Bassani-Sternberg M. Identification of tumor antigens with immunopeptidomics. Nat Biotechnol. 2022; 40 (2) 175–188. [PubMed: 34635837]
58. Umer HM, Audain E, Zhu Y, et al. Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. Bioinformatics. 2021.
59. Barsnes H, Vaudel M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. J Proteome Res. 2018; 17 (7) 2552–2555. [PubMed: 29774740]
60. Yates AD, Allen J, Amode RM, et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022; 50 (D1) D996–D1003. [PubMed: 34791415]
61. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res. 2022; 50 (D1) D988–D995. [PubMed: 34791404]
62. Vaudel M, Burkhart JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol. 2015; 33 (1) 22–24. [PubMed: 25574629]
63. Cote RG, Jones P, Martens L, et al. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinformatics. 2007; 8: 401. [PubMed: 17945017]
64. Deutsch EW, Mendoza L, Shteynberg D, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010; 10 (6) 1150–1159. [PubMed: 20101611]
65. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014; 5 5277 [PubMed: 25358478]
66. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013; 13 (1) 22–24. [PubMed: 23148064]
67. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem. 2005; 77 (4) 964–973. [PubMed: 15858974]
68. Perez-Riverol Y, Bai M, da Veiga Leprevost F, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol. 2017; 35 (5) 406–409. [PubMed: 28486464]
69. Perez-Riverol Y, Zorin A, Dass G, et al. Quantifying the impact of public omics data. Nat Commun. 2019; 10 (1) 3512 [PubMed: 31383865]
70. Perez-Riverol Y, Moreno P. Scalable Data Analysis in Proteomics and Metabolomics Using BioContainers and Workflows Engines. Proteomics. 2020; 20 (9) e1900147 [PubMed: 31657527]
71. Neely BA. Cloudy with a Chance of Peptides: Accessibility, Scalability, and Reproducibility with Cloud-Hosted Environments. J Proteome Res. 2021; 20 (4) 2076–2082. [PubMed: 33513299]
72. Solntsev SK, Shortreed MR, Frey BL, Smith LM. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. J Proteome Res. 2018; 17 (5) 1844–1851. [PubMed: 29578715]
73. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods. 2017; 14 (5) 513–520. [PubMed: 28394336]
74. Fahrner M, Foll MC, Gruning BA, Bernt M, Rost H, Schilling O. Democratizing data-independent acquisition proteomics analysis on public cloud infrastructures via the Galaxy framework. Gigascience. 2022; 11
75. Walzer M, Garcia-Seisdedos D, Prakash A, et al. Implementing the reuse of public DIA proteomics datasets: from the PRIDE database to Expression Atlas. Sci Data. 2022; 9 (1) 335. [PubMed: 35701420]

76. Bichmann L, Gupta S, Rosenberger G, et al. DIAproteomics: A Multifunctional Data Analysis Pipeline for Data-Independent Acquisition Proteomics and Peptidomics. J Proteome Res. 2021; 20 (7) 3758–3766. [PubMed: 34153189]
77. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Mol Cell Proteomics. 2015; 14 (9) 2394–2404. [PubMed: 25987413]
78. Serang O, Kall L. Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. J Proteome Res. 2015; 14 (10) 4099–4103. [PubMed: 26257019]
79. Omenn GS, Lane L, Lundberg EK, Beavis RC, Nesvizhskii AI, Deutsch EW. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. J Proteome Res. 2015; 14 (9) 3452–3460. [PubMed: 26155816]
80. Deutsch EW, Lane L, Overall CM, et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. J Proteome Res. 2019; 18 (12) 4108–4116. [PubMed: 31599596]
81. Omenn GS, Lane L, Overall CM, et al. The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. J Proteome Res. 2022.
82. Perez-Riverol Y, Vizcaino JA, Griss J. Future Prospects of Spectral Clustering Approaches in Proteomics. Proteomics. 2018; 18 (14) e1700454 [PubMed: 29882266]
83. Griss J, Perez-Riverol Y, Lewis S, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat Methods. 2016; 13 (8) 651–656. [PubMed: 27493588]
84. 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 (3) M111 014068
85. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015; 15 (18) 3163–3168. [PubMed: 25656970]
86. Montague E, Janko I, Stanberry L, et al. Beyond protein expression, MOPED goes multi-omics. Nucleic Acids Res. 2015; 43: D1145–1151. [PubMed: 25404128]
87. Schwanhausser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473 (7347) 337–342. [PubMed: 21593866]
88. Lundgren DH, Hwang SI, Wu L, Han DK. Role of spectral counting in quantitative proteomics. Expert Rev Proteomics. 2010; 7 (1) 39–53. [PubMed: 20121475]
89. Wisniewski JR, Hein MY, Cox J, Mann M. A “proteomic ruler” for protein copy number and concentration estimation without spike-in standards. Mol Cell Proteomics. 2014; 13 (12) 3497–3506. [PubMed: 25225357]
90. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Res. 2019; 47 (D1) D1056–D1065. [PubMed: 30462303]
91. Pinter N, Glatzer D, Fahrner M, et al. MaxQuant and MSstats in Galaxy Enable Reproducible Cloud-Based Analysis of Quantitative Proteomics Experiments for Everyone. J Proteome Res. 2022; 21 (6) 1558–1565. [PubMed: 35503992]
92. Bai M, Deng J, Dai C, Pfeuffer J, Perez-Riverol Y. LFQ-based peptide and protein intensity downstream analysis. 2022.
93. Xu T, Park SK, Venable JD, et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J Proteomics. 2015; 129: 16–24. [PubMed: 26171723]
94. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596 (7873) 583–589. [PubMed: 34265844]
95. Wen B, Zeng WF, Liao Y, et al. Deep Learning in Proteomics. Proteomics. 2020; 20 (21-22) e1900335 [PubMed: 32939979]
96. Meyer JG. Deep learning neural network tools for proteomics. Cell Rep Methods. 2021; 1 (2) 100003 [PubMed: 35475237]
97. Gessulat S, Schmidt T, Zolg DP, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods. 2019; 16 (6) 509–518. [PubMed: 31133760]

98. Tiwary S, Levy R, Gutenbrunner P, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods. 2019; 16 (6) 519–525. [PubMed: 31133761]

99. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods. 2020; 17 (1) 41–44. [PubMed: 31768060]

100. Yang Y, Liu X, Shen C, Lin Y, Yang P, Qiao L. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun. 2020; 11 (1) 146. [PubMed: 31919359]

101. Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun. 2020; 11 (1) 1759 [PubMed: 32273506]

102. Bouwmeester R, Gabriels R, Hulstaert N, Martens L, Degroeve S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods. 2021; 18 (11) 1363–1369. [PubMed: 34711972]

103. Li K, Jain A, Malovannaya A, Wen B, Zhang B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. Proteomics. 2020; 20 (21-22) e1900334 [PubMed: 32864883]

104. Declercq A, Bouwmeester R, Hirschler A, et al. MS(2)Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. Mol Cell Proteomics. 2022; 21 (8) 100266 [PubMed: 35803561]

105. Qin C, Luo X, Deng C, et al. Deep learning embedder method and tool for mass spectra similarity search. J Proteomics. 2021; 232 104070 [PubMed: 33307250]

106. Bittremieux W, May DH, Bilmes J, Noble WS. A learned embedding for efficient joint analysis of millions of mass spectra. Nat Methods. 2022; 19 (6) 675–678. [PubMed: 35637305]

107. Rehfeldt T, Gabriels R, Bouwmeester R, et al. ProteomicsML: An Online Platform for Community-Curated Datasets and Tutorials for Machine Learning in Proteomics. 2022.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Data submission guidelines and formats recommended to perform a submission to PRIDE Archive or another ProteomeXchange database.

ProteomeXchange data submission guidelines describe the files and metadata that must be provided for every submission in ProteomeXchange databases (http://www.proteomexchange.org/). However, these guidelines are the minimum information required for a dataset to be reproducible and reusable. Depending on the type of study and analytical method more data could be provided to improve the findability, reproducibility, and reusability.

For all the submissions the following data must be provided

- Instrument files with the acquired spectra (RAW files)
- Identification/quantification software output (MaxQuant peptide/protein outputs).
- FASTA database used for the identification of spectral library (DIA experiments).
- Configuration parameters for all software used in the experiment (e.g., MaxQuant parameters file).

The Sample and Data Relationship Format (SDRF)

The Sample and Data Relationship Format for proteomics were developed in 2021 (https://github.com/bigbio/proteomics-metadata-standard) to capture the experimental design and sample metadata in proteomics experiments. It is recommended to provide an SDRF file format for each analysis performed in the experiment including the description of the variables (factor values) and conditions under study.

Standard file formats for complete submissions

- mzTab or mzIdentML files: If the software used in the data analyses supports one of the HUPO-PSI standard file formats it should be provided. This will enable other users and tools to reuse the submitted results.
- Include the target and decoy peptides in your peptide identifications: Most of the tools and workflows that reuse data from PRIDE mzIdentMLs and mzTab files will perform the target/decoy quality assessment themselves with specific thresholds. If the decoy peptides are not included in the files, they can’t be used.
- PTM localization scores and PSM with the corresponding accession in the PSI-MS ontology vocabulary.

Protein quantification results

The input and output results files of differential expression data analysis tool, for example, MSstats. Currently, MSstats is supported by multiple tools including MaxQuant, Skyline, OpenMS and Spectronaut.

Quality Control reports

Every software provides multiple quality control reports as pdfs, text files or images as part of the experiment data analysis. Those files could be provided to the community for a better understanding of the quality of the experiment, and the statistical validation of the final biological discoveries.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Recommended guidelines for resources and groups that reanalyse and reuse public proteomics data

While data submission and dissemination associated with original publication and research has been standardized in ProteomeXchange archives, reuse identification and protein expression resources lack of guidelines and standardization. The following rules and guidelines may help to understand and validate reused datasets:

Dataset provenance

- The accession on ProteomeXchange of each reanalysed dataset must be provided.
- Detailed description of the workflow, analysis software including parameters must be provided and findable by the users.
- The results of the reanalyses could be provided using one of the standard file formats for users to check the quality of the reanalysis results.

Quality control on individual datasets analyses

While resource level FDR and quality control metrics are desired, in some cases it is technically complex, or the resource may not want to be lost unique and biologically relevant evidence during the integration:

- False discovery rates (FDR) applied at PSM, and protein level must be available for individual datasets.
- Additional quality control rules such as amino acid length for peptides, the number of miss-cleavages or number of unique peptides per protein should be available for users.
- Statistical scores from search engines to posterior error probabilities must be available for each evidence PSM, peptide or protein, making easy for users of the resource to check the quality of each evidence.
- For single amino acid variants (SAAVs) and PTMs the corresponding method and statistical score used to validate the position and event must be provided.

Data visualization

- For peptide/protein sequence identifications including posttranslational modifications, single amino acid variants (SAAVs) or complete sequence; the spectrum visualization should be available and more reliable if is provided using USIs.
- Protein coverage, number of unique peptides per protein, peptide intensities (in quantitative resources) must be provided.
- Quality control reports for each individual datasets should be provided, including precursor delta masses distributions, coefficient of variations or distribution of missing values in quantitative analysis.

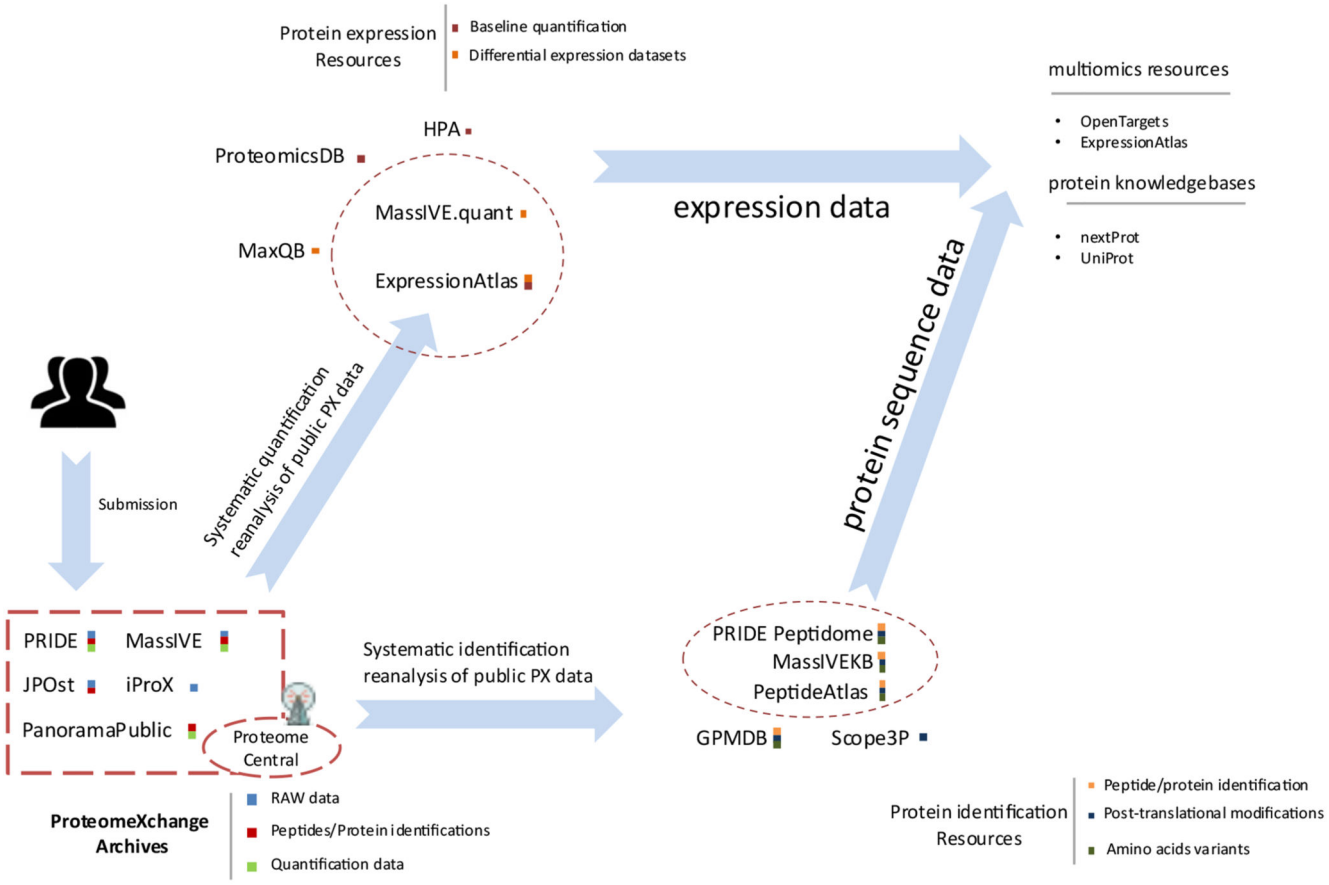


Figure 1. Proteomics resources ecosystem from data archives to knowledgebase protein and multiomics databases.

Proteomics data dissemination can be divided into four major categories: (i) ProteomeXchange archives responsible for data submission and dissemination of the original results associated with publications; (ii) peptide, protein identification resources including information about variants, mutations, or posttranslational modifications; (iii) protein expression resources including differential expression and baseline (absolute) expression profiles; (iv) and finally multiomics and protein knowledgebase resources that aggregate the protein sequence and expression information from the previous categories with other omics data like gene expression and variant information.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

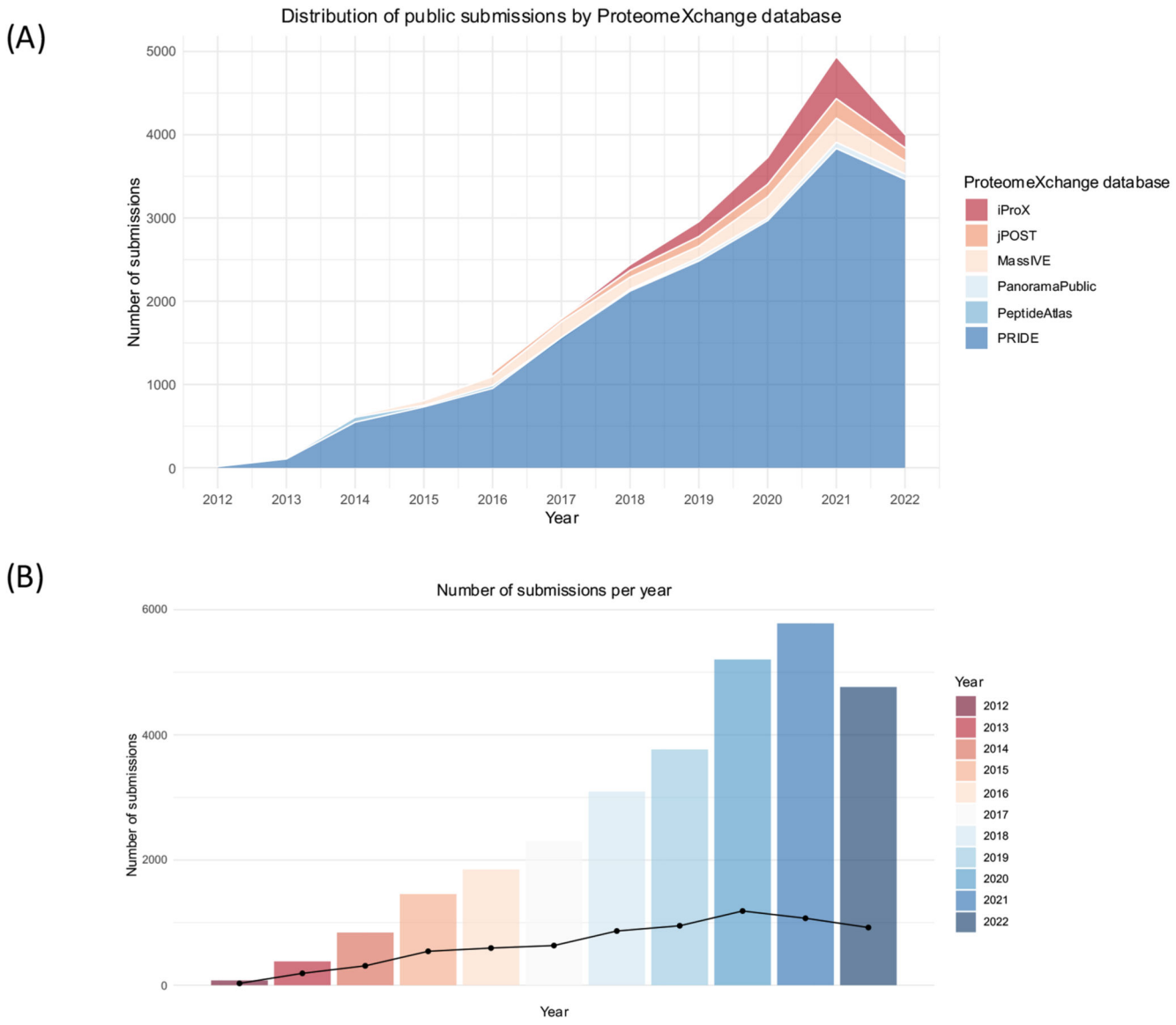


Figure 2.

(A) Distribution of the number of public submissions per ProteomeXchange archive (October 2022). **(B)** Number of submissions in PRIDE Archive per year (bars); and the number of submissions with more than 100 MS runs per year (lines).

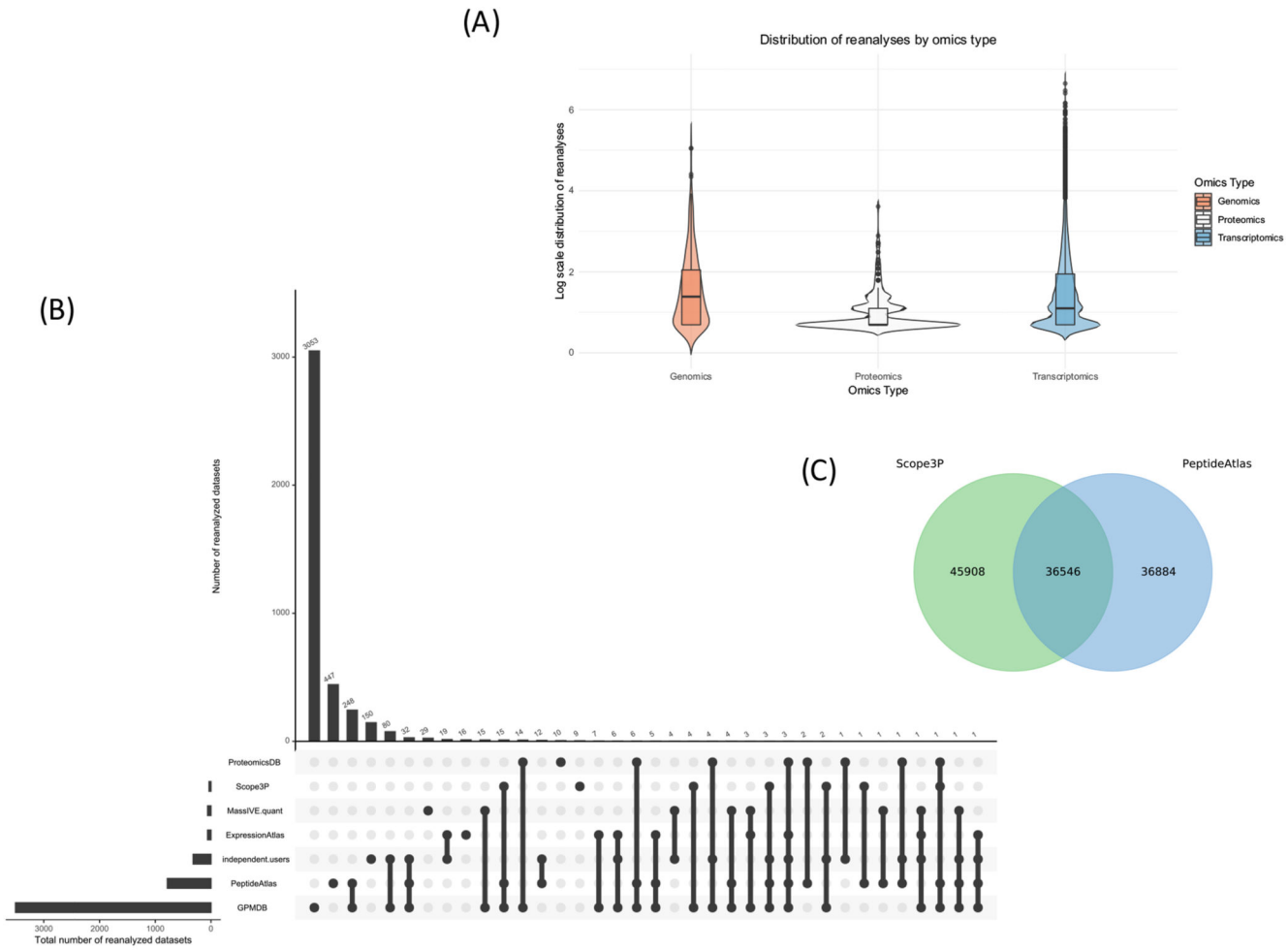


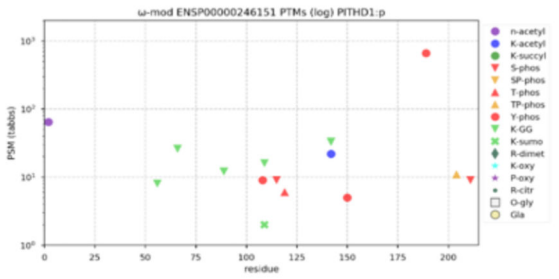
Figure 3.

(A) Distribution of datasets reanalyses by independent researchers per omics type. A reanalysis is counted if a PubMed publication cites a dataset directly by the dataset accession. **(B)** Overlap of the reanalyses by different databases (PeptideAtlas, GPMDB, ProteomicsDB, ExpressionAtlas and MassIVE) and reanalyses performed by independent researchers. **(C)** The number of distinct human phospho-peptides identified by PeptideAtlas and Scop3P.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

(A)



PITHD1:p, PITH domain containing 1
GRCh38.p13; HGNC:25022; TSL:1; UP:Q9GZP4; GlyGen: Q9GZP4
alt: HT014, TXNL1CL, C1orf128

This Modification-Abundance (M-A) diagram shows the number of times a residue has been observed with a particular PTM in a peptide-to-spectrum match in GPMDB, as a function of the residue's position in the corresponding protein sequence. Note that the Y-axis has a log scale. The particular residues and observation numbers are detailed in the following table:

https://gpmdb.thegpm.org/_/ptm_png/l=ENSP00000246151

(B)

https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetPeptide?_tab=3&atlas_build_id=461&searchWithinThis=Peptide+Sequence&searchForThis=GSGGGSSGGSIGGR&action=QUERY

Figure 4.

(A) Visualization of the modifications per amino acid position (protein ENSP00000246151 - https://gpmdb.thegpm.org/_/ptmpng/l=ENSP00000246151). **(B)** Visualization of phospho-sites for peptides in PeptideAtlas (GSGGGSSGGSIGGR - https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetPeptide?_tab=3&atlasbuild_id=461&searchWithinThis=Peptide+Sequence&searchForThis=GSGGGSSGGSIGGR&action=QUERY).

Table 1

Proteomics databases and resources that provide peptide and protein evidence from Mass spectrometry data.

Database Name	URL	Protein data types	Proteome Xchange
	Proteomics data archives		
PRIDE	https://www.ebi.ac.uk/pride	RAW, Identification Results, others	X
MassIVE	https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp	RAW, Identification Results, others	X
JPOST	https://repository.jpostdb.org/	RAW, Identification Results, others	X
iProX	https://www.iprox.cn/	RAW, Identification Results, others	X
Panorama Public	https://panoramaweb.org	RAW, Identification Results, others (Skyline analyses)	X
PeptideAtlas PASSEL	http://www.peptideatlas.org/passel/	RAW, Identification Results, others (SRM)	X
PDC	https://proteomic.datacommons.cancer.gov/	CPTAC Consortium Archive	
	Protein sequence databases		
PeptideAtlas	http://www.peptideatlas.org/	Peptide/Protein identifications, Posttranslational modifications	X
GPMDB	https://gpmdb.thegpm.org/	Peptide/Protein identifications, Posttranslational modifications, Single amino acid variants (SAAVs)	
Scope3P	https://iomics.ugent.be/scop3p	Posttranslational modifications	
OpenProt	https://www.openprot.org/	Proteogenomics, Open Reading Frames	
sORFs	http://www.sorfs.org/	Proteogenomics, Small open reading frames	
MassIVE-KB	https://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp	Peptide/Protein identifications	X
PRIDE Peptidome	https://www.ebi.ac.uk/pride/peptidome	Peptide/Protein identifications	X
	Protein expression databases		
MaxQB	http://maxqb.biochem.mpg.de/mxdb/	Protein expression Peptide/Protein identifications	
PaxDB	https://pax-db.org/	Protein expression	
ExpressionAtlas	https://www.ebi.ac.uk/gxa/home	Protein expression	
MassIVE.quant	https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp	Protein expression Peptide/Protein identifications	
ProteomicsDB	https://www.proteomicsdb.org/	Protein expression Peptide/Protein identification	
Matrisomedb	http://matrisomedb.pepchem.org/	Protein expression	
Immunological Proteome	http://immpres.co.uk/	Protein Expression	