# Data Submission Guidelines for ProteomeXchange

## 1. Preamble

This document aims to provide detailed guidelines for the submission of mass spectrometry (MS) derived proteomics data to the ProteomeXchange (PX) consortium (http://www.proteomexchange.org). An early version of this policy was first adopted by the Human Proteome Organization (HUPO) Plasma Proteome Project for the collection of its phase II data (1) and recently by the HUPO Human Proteome Project (2).

## 2. Definitions

### 2.1. Data types

Proteomics data come in a variety of forms, which are defined here:

- **Mass spectrometer output files:** the data and metadata generated by mass spectrometers, usually one file per run (although some instruments put multiple runs per file). The data may be the original profile mode scans or may already have had some basic processing like centroiding applied. They may be:
    - i) raw data (as described below).
    - ii) peak list spectra in a standardized format such as mzML, mzXML or mzData (see below), but they cannot be 'processed peak lists' (see below).

  However, it is important that all of the scans that were generated are included with applicable metadata.

- **Raw data:** the binary, vendor-specific output files directly created by the instrument software. These files are typically large (several gigabytes) and require specialized software in order to be read.

- **Standardized MS data formats:** There are currently three widely known mass spectrometry data formats in Proteomics: mzXML (3) (developed at the Institute of Systems Biology (ISB), Seattle, USA), mzData (now made obsolete, originally developed by the HUPO Proteomics Standards Initiative (PSI)), and the successor to both of the above: mzML (4) (currently v1.1, jointly developed by the ISB and PSI, http://www.psidev.info/mzml). These data formats can be used to represent processed peak lists, as well as raw data. In addition to the mass spectra, they contain detailed metadata that provide context to the measurements.

- **Processed peak lists:** Heavily processed form of mass spectrometry data, usually derived from the raw data files through various (semi-)automatic steps, e.g.: centroiding, deisotoping, and charge deconvolution. These files are formatted in plain text, with typical formats like dta, pkl, ms2 or mgf. They usually contain only a subset of only the MS2 scans (MS1 scans are excluded), and are missing significant amounts of metadata that were present in the source format.

- **Protein/peptide identifications:** Proteomics mass spectra can be matched to peptides or proteins, resulting in identifications for those spectra. Typically a

spectrum is considered identified if the score attributed to a peptide or protein match qualifies against an *a priori* or *a posteriori* defined threshold. In the case of fragmentation spectra, the initial identification will consist of a peptide sequence; subsequent steps will derive a list of proteins from the identified peptides. The protein assembly step can be a discernible process with its own input and output files, or it can be implicit in the overall identification software. This information can be represented by a variety of data formats called search engine output files (see below).

- **Protein/peptide quantification:** Protein/peptide expression values can also be obtained from a MS-based proteomics experiment. There is a high diversity of approaches that result in the existence of very heterogeneous software and data analysis pipelines. Some search engines are able to perform both identification and quantification, and produce 'search engine output files' containing both types of data. However, if there is software that only performs the quantification part of the analysis, the generated data is represented in 'quantification software output files' (see below).

- **Search engine output files**: They contain the data and metadata generated by the software (usually called search engines) used for performing the identification and often the quantification of peptides and proteins. Each search engine has its own specific output file. The formats are typically formatted in either plain text or XML, with typical formats like Mascot .dat, OMSSA xml, etc.
  In addition to each specific format, a data standard format called mzIdentML (currently v1.1, http://www.psidev.info/mzidentml) (5) has been developed by the PSI to represent this kind of information. Some search engine output files can represent as well quantification results, but this is not the case of mzIdentML. A second standard data format called mzTab (http://code.google.com/p/mztab/), currently under development, can represent both identification and quantification results.

- **Supported protein/peptide identification results:** This definition includes all protein/peptide identification processed data that can be fully represented by the receiving repository in ProteomeXchange. For the PRIDE database, as the PX submission point for tandem MS/MS datasets, the data formats supported are PRIDE XML and mzIdentML (version 1.1, see above). mzIdentML can now be exported from a variety of tools (see updated list at http://www.psidev.info/tools-implementing-mzidentml).
  PRIDE XML can represent both mass spectra data and protein/peptide identifications, and for some basic use cases, quantification information as well. 'Search engine output files' can be converted to PRIDE XML using PRIDE Converter 2 (6) (http://code.google.com/p/pride-converter-2/) and a few other available tools.

- **Quantification software output files**: the data and metadata generated by the software used for performing exclusively the quantification analysis of peptides and proteins. In addition to each specific format from each software tool, a data standard format called mzQuantML (currently v1.0, http://www.psidev.info/mzquantml) has been released by the PSI to represent this kind of information (7). As mentioned before, a second data format called mzTab

(http://code.google.com/p/mztab/) can also represent quantification results, although is currently not yet finished.

- **Metadata:** Whereas mass spectra present the core output of any mass spectrometer, a simple collection of spectra does not provide sufficient information for confident interpretation. Something similar happens for the peptide and protein identifications and their expression values. This lack of context can be solved by providing relevant metadata along with the spectra and/or the identifications and quantification data. Mass spectrometer, search engine, and quantification software output files (see above) typically accommodate this information.

### 2.2. Proteomics data resources

In the ProteomeXchange consortium there are currently two kinds of proteomics data resources (Figure 1, Appendix I):

- **Archival resources**: They should contain processed data as published by the authors as well as the raw source files that generated the results. At the moment PRIDE is the representative of this type in the consortium for tandem MS/MS datasets, and the PASSEL component of PeptideAtlas is the representative for Selected Reaction Monitoring (SRM) datasets. However, others could be added at the future to the consortium (see section 7).

- **Secondary data resources**: These ones build upon the primary data provided by submitters, which are stored in the archival resources. The representative resources in the consortium are at the moment PeptideAtlas and UniProt, but potentially there could be others outside the consortium, like for instance GPMDB or neXtProt.

### 2.3. Other definitions

- **ProteomeCentral**: It is the resource that generates a unique identifier for each ProteomeXchange dataset and also, it constitutes a registry for all ProteomeXchange submissions. This queryable archive provides the users with an efficient way to identify datasets of interest. For instance, it is a way to monitor the re-use of particular datasets, and give an efficient way to monitor the volume and impact of the ProteomeXchange data exchange. It is available at http://proteomecentral.proteomexchange.org.

**- PX XML message**: Once the data is published or made publicly available by the submitter, the data is disseminated from the archival repository (PRIDE for MS/MS data, PASSEL for SRM data) to the rest of the consortium by an RSS message, that links to a PX XML file. This file, with a defined XML schema (available at the ProteomeXchange Google SVN), contains the essential metadata information about the datasets, and how to retrieve all the files that are part of a PX submission. The PX XML messages can be versioned and all of them are available in ProteomeCentral. This XML message allows PeptideAtlas, UniProt and other resources to evaluate and integrate the data.

The overall data workflow is summarized in Figure 1, for both MS/MS data and SRM datasets.

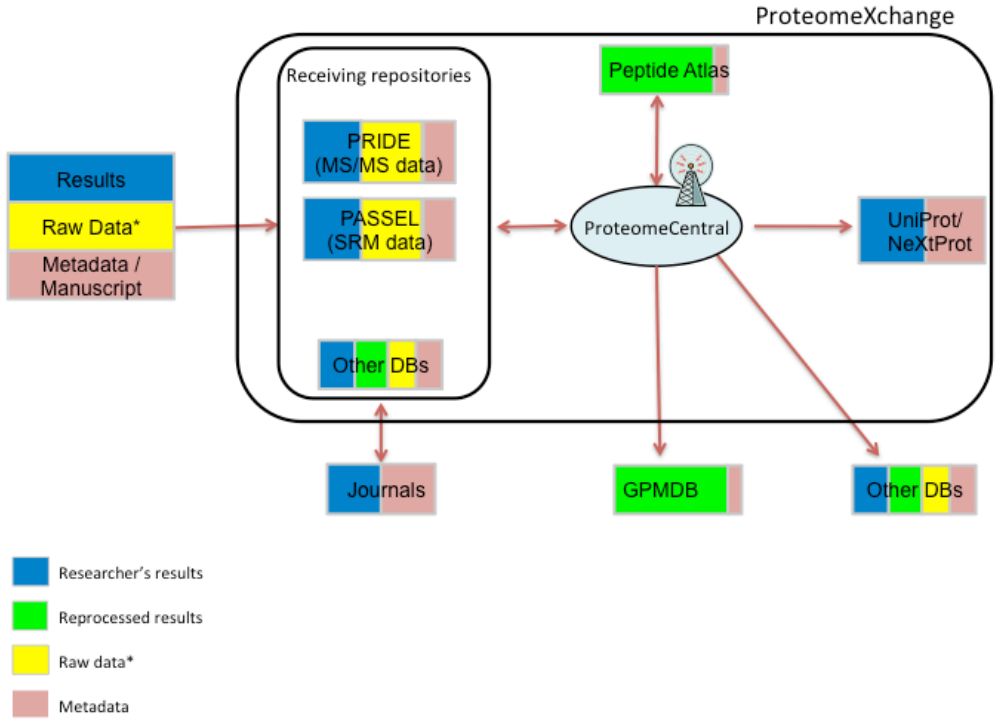


**Figure 1:** Overview of the ProteomeXchange data flow

## 3. Submission workflow for MS/MS datasets

### 3.1. New datasets

At present, PRIDE is the submission point for this type of data in PX. Each ProteomeXchange data submission coming from MS/MS proteomics workflows shall consist of three major components: mass spectrometer output files, study metadata, and peptide/protein identifications (see definitions in section 2). Other components are optional at present: quantitative information, peak list files (depending on the data format used for the submission it can be mandatory), gel images and any other file type.

There are two main different case scenarios for this type of data submissions ("Complete" and "Partial", see below). Datasets composed of only raw data (plus related metadata) will not be considered compliant to fulfil the requirements of the consortium and will not get a PXD identifier.
By default, all the data will remain private during the manuscript review process. The secondary data resources (such as PeptideAtlas and UniProt) will only have access to the data once it has been made publicly available.

Each submission is expected to contain the data supporting one manuscript (although there could be exceptions), and will be assigned a unique ProteomeXchange identifier (PXD_accession number) through the ProteomeCentral service (Figure 1).

Experimental metadata will be made available at PRIDE and in ProteomeCentral (the information contained in the PX XML message, see Appendix II). This represents a subset of the Minimal Information About a Proteomics Experiment (MIAPE) recommendations for mass spectrometry data and proteomics identifications (8,9). However, providing richer metadata is possible and encouraged.

A detailed tutorial about how to submit MS/MS datasets to ProteomeXchange *via* PRIDE including the supported formats can be found at http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/px_submission_tutorial.pdf.

Two main submission modes are supported (Figure 2):
A)- "Complete" submissions. This is aimed for datasets containing supported identification results and raw data, plus the related metadata. Such submissions will be assigned a Digital Object Identifier (DOI).
Datasets supporting a submitted manuscript will represent the main use case. In addition, there can be interesting datasets that are not generated for publication purposes. There are two main sub-categories within this category:
- Data labelled with an organization 'stamp', like ABRF or HPP, etc.
- Data produced in an individual lab. These can be used for testing or educational purposes.

All the required data (including raw files and processed identification result files) should be submitted to the PRIDE database, as the interaction with a curator is considered essential to ensure complete and high quality data submissions. This is the complete workflow (Figure 2):
1- All the 'search engine output files' need to be converted to either the mzIdentML (which can be exported from different search engines) or PRIDE XML formats (using PRIDE Converter 2 or other available tools).
2- The PX submission tool (available at the URL http://www.proteomexchange.org/submission) needs to be used for doing the actual submission.

As an example, see dataset in ProteomeCentral PXD000500 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD000500).

B)- "Partial" submissions. This mechanism is aimed for datasets containing raw data and search engine output files, plus the related metadata. Supported identification results cannot be generated since there is no a converter/exporter available to PRIDE XML or mzIdentML. Data will go to the raw file archive resource at the EBI and these datasets will not get a DOI assigned.

The PX submission tool (available at http://www.proteomexchange.org/submission) needs to be used for doing the actual submission. As mentioned before, metadata will be made available in the PX XML message, and it should include at least the sufficient information needed to create the PX XML message (see Appendix II).

As an example, see dataset in ProteomeCentral PXD000711 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD000711).

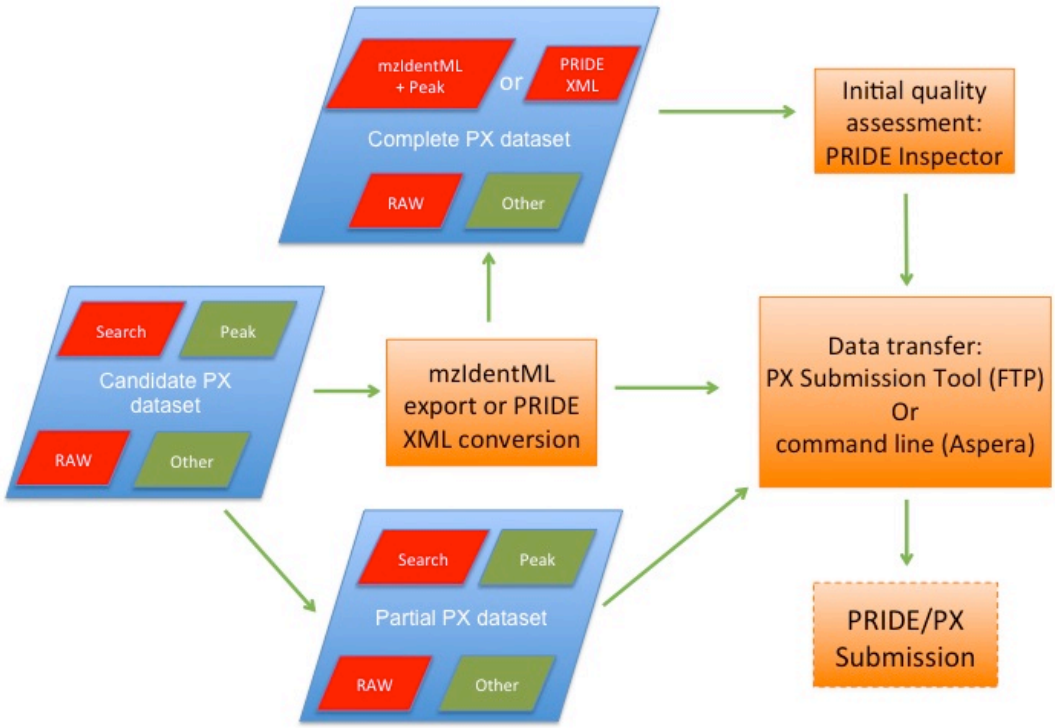


**Figure 2:** Overview of the PRIDE/PX submission process with the two default submission types: Complete and Partial.

All the files that are part of a PX MS/MS submission are labelled for simplification purposes in the following categories:

- 'RAW'. Label for 'raw data' (mass spectrometer output files).
- 'RESULT'. This label will be attributed to 'supported identification results' (PRIDE XML, mzIdentML version 1.1) for "Complete" submissions. In some cases 'RESULT' files (depending on the data format used) could also include quantification results (in some cases PRIDE XML can contain quantitative information), in addition to identification results.
- 'PEAK'. Label for 'processed peak list' spectra formats. For "Complete" submissions performed with mzIdentML, 'PEAK' files are mandatory (since peak lists are not included in mzIdentML). Otherwise, they are optional.
- 'SEARCH'. Label for 'search engine output files'. They are only optional for "Complete" submissions, but as mentioned earlier, mandatory for "Partial" ones, since they represent the results of the study. They can include identification and also include quantification results, if the same software performed identification and quantification.
- 'QUANT'. Quantification output files (they can be used if quantitation information was not included in the 'RESULT' files). The use of the data

standards mzQuantML and mzTab is recommended but due to the limited number of existing implementations, it cannot be enforced. The situation will improve as these data standards gain in popularity.
- 'GEL': Gel image files.
- 'OTHER'. Label for any other type of file that can be included in the submission.

In any case, each dataset becomes publicly available on acceptance or publication of the manuscript supported by the dataset (See Section 6), or when the authors tells PRIDE to do so). When a submission becomes publicly available, a short summary will be released though a public announcement system, as a RSS feed with a link to the corresponding PX XML message file. All the PX XML messages will be stored in ProteomeCentral.
Again, a detailed tutorial about how to submit MS/MS datasets to ProteomeXchange *via* PRIDE can be found at http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/px_submission_tutorial.pdf.

### 3.2. Reprocessed datasets

The data workflow for reprocessed datasets starts when any member of the ProteomeXchange consortium (at present PeptideAtlas) makes a reinterpretation of existing data in any of the archival repositories (at present PRIDE for MS/MS data). A new ProteomeXchange identifier will be obtained from ProteomeCentral (it will be a RPXD identifier instead of the standard PXD). As an example, see dataset RPXD000665 in ProteomeCentral: http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=RPXD000665).

However, the original ProteomeXchange accession number is retained in the PX XML message to allow coordinated search for different views of data from one given submission. This ensures that a simple one-time submission from a contributor is automatically distributed to all ProteomeXchange repositories with sufficient information. When the reanalysis is done by a ProteomeXchange member a XML broadcast will be produced, which will include the new PXD identifier, but also the old one. All the relevant information about the connection between the datasets will be stored in ProteomeCentral. Three main situations may arise when a PX dataset is reanalysed:

a) If the data reinterpretation gets published in a separate publication as 'independent' findings:
- Data must go to PRIDE (as any other new MS/MS dataset).
- Ideally, it should not be needed to re-upload the raw data (references to URLs are allowed).

b) The reinterpretation does not get published as 'independent' new findings. In this case, data can be kept in a reprocessing repository (e.g. PeptideAtlas). For instance, this applies to all new PeptideAtlas builds that get published.

c) In the case of a mixture of new and reprocessed data in one given dataset, they should be considered to be a new dataset, so the dataset should be submitted to PRIDE.

## 4. Submission workflows for Selected Reaction Monitoring (SRM) datasets

### 4.1. New data

New datasets acquired *via* SRM should be submitted to ProteomeXchange repositories designed for this kind of data *via* http://www.proteomexchange.org/submit/. At this time the PASSEL component of PeptideAtlas is the only ProteomeXchange repository for SRM data, although this may change in the future.

For such submissions, 3 main items are required:

1. Mass spectrometer output files, preferably raw files.
2. Transition list describing the peptides that the instrument targeted.
3. Analysis results.

Once submissions are received, they are checked by a curator, run through the PASSEL pipeline, and then loaded into the PASSEL database. The ProteomeXchange workflow is very similar to that for MS/MS data, as shown in Figure 3.

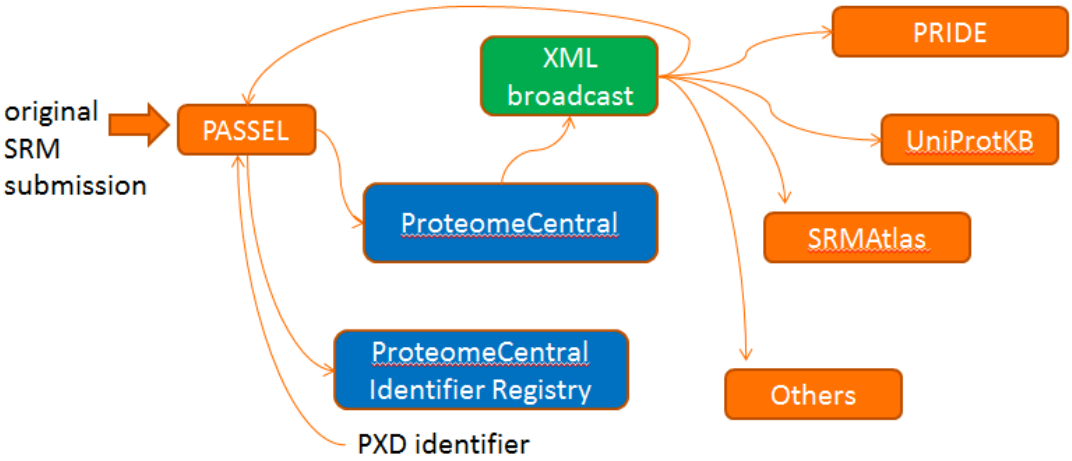


**Figure 3.** Workflow for original SRM submissions to PASSEL. After processing by PASSEL, an identifier is obtained from ProteomeCentral, and then an announcement message is transmitted to ProteomeCentral describing the submission. ProteomeCentral then broadcasts the submission to all subscribers.

## 5. Data ownership

The PRIDE database does not assume editorial control or ownership over the submitted data; it maintains the original submitter as owner of these data. PRIDE requires that a submitter is explicitly identified for each dataset. This is done by associating a user account on the PRIDE system with a submitted (set of) experiment(s). Upon public availability of the data, the original data ownership is maintained in the database, although obviously dissemination and reuse of the released data are no longer restricted at that point. PASSEL also does not assume editorial control of the submitted data. Users specify at submission time the date on which the data become publicly accessible. On this date, the data are automatically released. The data owner has the option of adjusting this date in case of review delays, etc.

## 6. Data privacy

PRIDE allows data to be kept private for any duration of time, until the owner of the data (as identified by the associated PRIDE user account, see above) gives explicit permission to release the data. A variant occurs when privately submitted data are associated with a manuscript submitted to a journal. The public availability of the submitted data will then be coordinated with the publication of the associated article in correspondence with the journal editor. For PASSEL, data become automatically available on the date that the submitter specifies.

PRIDE and PASSSEL can automatically provide reviewer accounts for each submitted experiment, which can be communicated to journal editors and referees in a submitted manuscript, thus allowing confidential reviewing of the privately submitted data.

The date of submission, as well as the date of public release, is archived in the PRIDE database system. After public release of the data, all the files and information available in the dataset will be made available to the general public without further reservations. The original ownership of the data will remain asserted in the PRIDE database, however. Any restrictions on data dissemination or reuse are obviously removed upon public availability of the data.

## 7. Membership in the ProteomeXchange consortium

As of February 2014, PRIDE and PeptideAtlas (PASSEL component) are the only archival resources in the ProteomeXchange consortium. Applications for recognition as archival resources are welcome, and will be decided upon by the ProteomeXchange consortium based on the following key criteria:

1. Experience and funding level of resource.
2. Stability.
3. Availability of dedicated curation staff.
4. Ability to store and make accessible raw data, metadata, and interpretations.
5. Worldwide unrestrained availability of stored datasets for download.

The last version of the ProteomeXchange collaborative agreement (which can be found at http://www.proteomexchange.org/documents/proteomexchange-collaborative-agreement) describes the steps needed to become a member of the consortium.

## 8. References


1. Omenn, G.S., Aebersold, R. and Paik, Y.K. (2009) 7(th) HUPO World Congress of Proteomics: launching the second phase of the HUPOPlasma Proteome Project (PPP-2) 16-20 August 2008, Amsterdam, The Netherlands. *Proteomics*, **9**, 4-6.
2. Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C.H., Corthals, G.L., Costello, C.E. *et al.* (2011) The human proteome project: current state and future direction. *Mol Cell Proteomics*, **10**, M111 009993.
3. Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R. *et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*, **22**, 1459-1466.
4. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Rompp, A., Neumann, S., Pizarro, A.D. *et al.* (2011) mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics*, **10**, R110 000133.
5. Jones, A.R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S.J., Selley, J.N., Searle, B.C., Shofstahl, J., Seymour, S.L. *et al.* (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*, **11**, M111 014381.
6. Cote, R.G., Griss, J., Dianes, J.A., Wang, R., Wright, J.C., van den Toorn, H.W., van Breukelen, B., Heck, A.J., Hulstaert, N., Martens, L. *et al.* (2012) 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*, **11**, 1682-1689.
7. Walzer, M., Qi, D., Mayer, G., Uszkoreit, J., Eisenacher, M., Sachsenberg, T., Gonzalez-Galarza, F.F., Fan, J., Bessant, C., Deutsch, E.W. *et al.* (2013) The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol Cell Proteomics*, **12**, 2332-2340.
8. Taylor, C.F., Binz, P.A., Aebersold, R., Affolter, M., Barkovich, R., Deutsch, E.W., Horn, D.M., Huhmer, A., Kussmann, M., Lilley, K. *et al.* (2008) Guidelines for reporting the use of mass spectrometry in proteomics. *Nat Biotechnol*, **26**, 860-861.
9. Binz, P.A., Barkovich, R., Beavis, R.C., Creasy, D., Horn, D.M., Julian, R.K., Jr., Seymour, S.L., Taylor, C.F. and Vandenbrouck, Y. (2008) Guidelines for reporting the use of mass spectrometry informatics in proteomics. *Nat Biotechnol*, **26**, 862.
10. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J. and Apweiler, R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537-3545.
11. Vizcaino, J.A., Cote, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J. *et al.* (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res*, **41**, D1063-1069.
12. Deutsch, E.W., Lam, H. and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep*, **9**, 429-434.

13. Farrah, T., Deutsch, E.W., Kreisberg, R., Sun, Z., Campbell, D.S., Mendoza, L., Kusebauch, U., Brusniak, M.Y., Huttenhain, R., Schiess, R. *et al.* (2012) PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics*, **12**, 1170-1175.
14. Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B. *et al.* (2010) A guided tour of the Trans-Proteomic Pipeline. *Proteomics*, **10**, 1150-1159.
15. Farrah, T., Deutsch, E.W., Omenn, G.S., Campbell, D.S., Sun, Z., Bletz, J.A., Mallick, P., Katz, J.E., Malmstrom, J., Ossola, R. *et al.* (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics*, **10**, M110 006353.

## 9. Appendix I: General information about PRIDE and PeptideAtlas

There are currently several proteomics data repositories, each with a different major focus. Here we discuss the PRoteomics IDEntifications Database (PRIDE) at the EBI (10,11), and PeptideAtlas and its PASSEL (PeptideAtlas SRM Experiment Library) component at ISB (12,13).

PRIDE is focused on submissions of identifications, usually presented in correlation with a manuscript or published paper. Mass spectrometer output files were routinely made available by PRIDE in mzData format only. However, PRIDE now also stores raw binary data as well. Metadata are required as part of the submission to PRIDE (see Appendix II). The datasets included in PRIDE can be searched by the experimental metadata. Data submissions can be held privately within PRIDE, albeit allowing reviewers and journal editors access if desired, until an optional preset date is reached, or until the submitter chooses to release the data to the public.

PeptideAtlas/PASSEL allows researchers to easily submit proteomic data sets generated by SRM. The raw data are automatically processed in a uniform manner and the results are stored in a database, where they may be downloaded or browsed *via* a web interface that includes a chromatogram viewer.

Submitted files to PeptideAtlas are reprocessed using multiple search strategies and the Trans Proteomic Pipeline (TPP) (14). All experiments are then combined to form an inclusive view of all peptides and proteins observed for each species across all contributed data. Atlases for organ and biofluid proteomes of a given species are well-developed, as well (15). PeptideAtlas contains associated software tools that support data analysis and mining, including spectral searching, proteotypic peptide selection for targeted proteomics approaches, and a general estimate of protein abundance.

Each of these repositories may thus present different views on an experiment, or contain different data components of an experiment. The ProteomeXchange consortium was formed to enhance communication and automated exchange of data among these and other proteomics repositories to ensure a more transparent and more efficient access for the community to the available data.

Finally, it is important to note that PRIDE and PeptideAtlas support and actively promote the use of the Proteomics Standards Initiative (PSI) standard data formats.

## 10. Appendix II: Metadata requirements for tandem MS/MS submissions

Proteomics data are substantially enriched when sufficient metadata are provided. The presence of the metadata required in this Appendix will be enforced for any PX MS/MS data submission (they are mandatory in the PX Summary File format, see http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/proteomexchange_submission_summary_version_2.0.pdf). Most of this information is included in the PX XML file. The user will need to provide:

- Contact name and e-mail for the submission. The contact details of the data submitters need to be provided, allowing interested users to contact the original authors if desired.
- Lab Head or Principal Investigator.
- Name of the PX dataset.
- Project description: it could be considered as the abstract information of the dataset (provided as free text).
- Summary of the Sample Protocol (provided as free text).
- Summary of the Data analysis Protocol (provided as free text).
- Experiment type. Chosen from a drop-down menu.
- Keywords: A list of keywords that describe the content and type of the experiment being submitted. Multiple entries should be comma separated.
- Sample annotation: species. At least one NEWT Controlled Vocabulary (CV) term is mandatory per dataset.
- Sample annotation: tissue. Using the BRENDA Tissue ontology (BTO), accessible at http://obo.cvs.sourceforge.net/obo/obo/ontology/anatomy/BrendaTissue.obo)
- Instrument details. Using the PSI-MS CV. It is accessible at http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo.
- Quantification method (if applicable).
- Protein post-transcriptional modifications (PTMs). They are reported using the PSI-MOD ontology (accessible at http://psidev.cvs.sourceforge.net/psidev/psi/mod/data/PSI-MOD.obo).

**Optional information:**

- Sample annotation: cell type. Use the “Cell Type” ontology.
- Sample annotation: Disease. Use the “Human Disease” ontology (DOID).
- Dataset optional details:
    - your dataset is part of a bigger project/effort (for instance the Human Proteome Project or ‘PRIME-XS’). It is a way to tag your dataset to enable grouping this way.
    - there is already a PubMed ID associated with it (the data has been already published).
    - your dataset represents a reanalysis of an earlier public PX dataset
    - there are other related “omics” datasets (for instance transcriptomics, metabolomics data present in other repositories) that can be associated

with it. In this case, please provide the accession number of the dataset in the corresponding repository.