Using the PRIDE database and ProteomeXchange for submitting and accessing public proteomics datasets

Author(s):

Andrew F. Jarnuczak[1] and Juan Antonio Vizcaíno[1, *]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

* Corresponding author: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Phone: + 44 (0) 1223 492686. Email: juan@ebi.ac.uk

Significance Statement

Availability of biological data in the public domain, usually accompanying the publication of a scientific paper, is generally considered a good scientific practice since it enables the assessment of the reported results. Additionally, data can often be reused with different purposes to the objectives of the original study. While public data sharing has been commonplace in other disciplines (e.g. genomics and transcriptomics), it has only recently become popular in proteomics. One of the main reasons behind that has been the establishment of the ProteomeXchange Consortium, standardizing data submission and dissemination between the main proteomics resources worldwide, including PRIDE, PeptideAtlas/PASSEL, MassIVE and jPOST. Here we first explain how to submit data to the PRIDE database, the world-leading proteomics data repository. Second, we describe how to access data in any of the ProteomeXchange resources.

ABSTRACT

The ProteomeXchange (PX) Consortium is the unifying framework for world-leading mass spectrometry (MS)-based proteomics repositories. Current members include the PRIDE database (UK), PeptideAtlas/PASSEL (UNIT 13.25) and MassIVE (USA), and jPOST (Japan). The Consortium standardizes submission and dissemination of public proteomics data worldwide. This is achieved through implementing common data submission guidelines and enforcing metadata requirements by each of the members. Furthermore, the members use a common identifier space. Each dataset receives a unique (PXD) accession number and is publicly accessible as soon as the associated scientific publications are released. This protocol provides a step by step guide how to submit data to the PRIDE database, and describes how to access the PX portal (called ProteomeCentral), which can be used to search datasets available in any of the PX members.

Keywords: PRIDE database, proteomics, mass spectrometry, data, data repository, how to

INTRODUCTION

The use of mass spectrometry (MS) and associated technologies to investigate proteomes at a large scale has historically been restricted to a handful of specialized laboratories. However, as the cost of instrumentation becomes lower and bioinformatics solutions are more readily available, increasing numbers of life science researchers have at present the ability to generate and analyze proteomics data. Data standardization and open access dissemination through public repositories such as those included in the ProteomeXchange (PX) Consortium (Deutsch et al., 2017; Vizcaíno et al., 2014) have become an important part of the research cycle in proteomics. This abundance of data in the public domain, only a recent phenomenon, presents a wealth of opportunities for researchers, for instance to extract new knowledge, develop computational methods, to learn from others and ultimately speed up their own research.

In fact, free availability of primary data is an integral part of reproducible research. This is particularly reflected by the requirement of an increasing number of scientific journals to make experimental data supporting the corresponding publication accessible to others. Proteomics databases like those included in PX enable this requirement.

The PX Consortium (Deutsch et al., 2017; Vizcaíno et al., 2014), was initiated to provide a unifying framework for proteomics resources, and enable its members to standardise data deposition and data dissemination procedures. The Consortium, formally established in 2011, currently contains the world-leading proteomics repositories. The members are PRIDE (EMBL-European Bioinformatics Institute, Cambridge, UK) (Vizcaíno et al., 2016), PeptideAtlas (UNIT 13.25; Institute for Systems Biology, Seattle, USA) (Desiere et al., 2006), MassIVE (Center for Computational Mass Spectrometry, University of California, San Diego, USA), and jPOST (various institutions, Japan) (Okuda et al., 2017). Datasets available in any of the PX repositories can be accessed through the ProteomeCentral common portal (http://proteomecentral.proteomexchange.org/). To facilitate data reuse and reanalysis, the PX members fully support open data formats developed under the umbrella of the Proteomics Standards Initiative (PSI) (Deutsch et al., 2015). The PRIDE (PRoteomics IDEntifications) database (now renamed as PRIDE Archive) is one of PX founding members and stores approximately ~90% of all PX datasets at the moment of writing (Deutsch et al., 2017).

Here we provide two basic protocols involving the use of PX resources. Basic Protocol 1 explains how to perform data submission to PX *via* the PRIDE database. Basic Protocol 2 describes the ProteomeCentral common interface and the type of searchable information within the portal. We expect that both protocols will help familiarize the user with these proteomics resources and their functionality, and will offer a starting point for

those interested in submitting and/or mining public proteomics datasets to gain new knowledge or supplement their own research.

PROTOCOL 1

PERFORMING A COMPLETE SUBMISSION TO ProteomeXchange *via* the PRIDE database

Introductory paragraph

The PRIDE database (http://www.ebi.ac.uk/pride/archive/) was originally established in EMBL-EBI (Cambridge, UK) in 2004 (Martens et al., 2005). PRIDE can store, among other data types, peptide and protein identifications and related quantification values, the corresponding mass spectra (both as processed peak lists and raw data), gel images, the searched protein sequence databases or spectral libraries, programming scripts and any other technical and/or biological metadata provided by the data submitters. It is important to highlight that PRIDE stores all processed results as originally analyzed by the authors (Vizcaíno et al., 2016). As it is the case of the other PX resources, PRIDE fully supports and promotes the use of several PSI data standards.

Due to the great diversity in the field, and to accommodate as many techniques and workflows as possible, two types of data submissions are available to the users: 'Complete' and 'Partial'. Both submission types must always contain the raw files (mass spectrometer output files), and some agreed technical and biological metadata information. The preferred submission type is 'Complete', which requires that the peptide/protein identification results are provided ideally in the PSI mzIdentML standard file format (Jones et al., 2012), enabling the connection between the identifications and the corresponding mass spectra. As mzIdentML files do not contain the underlying MS/MS spectra information, the corresponding peak list files (for example, mgf files) are also required for 'Complete' submissions. Datasets can then be visualized with various tools, including PRIDE Inspector – an open source standalone application for the visualization and validation of proteomics data encoded in PSI standard formats as well as other popular mass spectra data files (Perez-Riverol et al., 2016), and can be parsed and integrated into other PRIDE related resources, like e.g. PRIDE Cluster (Griss et al., 2016). Support for an additional open data format, the standard mzTab (Griss et al., 2014), which accommodates both identification and quantification results in a simpler tabular format, is being finalised at the moment of writing.

The alternative 'Partial' submission option is available for those cases where the processed results are not available in the supported open data formats (for instance, due to the lack of data exports and/or converters). A 'Partial' submission then requires search result output files containing peptide and protein identification information but the file format is not enforced. The main drawback is that data submitted using this option

cannot be visualized using the standard open source tools and libraries provided by PRIDE, although external tools may be available.

The following protocol describes how to perform a data submission to PRIDE, fulfilling PX requirements. For simplicity purposes, we mostly assume that users have all the necessary files ready. However, one can follow the protocol using any public 'Complete' dataset that can be obtained from PRIDE. In that case, we suggest to use as example data those files contained in the dataset PXD004612 (https://www.ebi.ac.uk/pride/archive/projects/PXD004612/files). A screenshot of the dataset is shown as **Figure 1**.

Necessary Resources

Hardware: A computer with Internet connection. Operating System: Windows, Mac OS or Linux.

Software: Any web browser, a supported version of Java JRE (Java SE 8 at the time of writing) (https://java.com/en/download/).

1. Prepare the data to be submitted. The following file types can be submitted to PRIDE:

    i. 'RAW' (required) –Files generated directly (output) by the mass spectrometer. The raw files might have different extensions or be contained in a folder structure, depending on the vendor. For example Thermo Scientific Orbitrap output files have the extension '.raw' (an example can be found at https://www.ebi.ac.uk/pride/archive/projects/PXD001334/files). Waters Synapt series or Bruker Impcat II mass spectrometers write their raw output into a folder (see https://www.ebi.ac.uk/pride/archive/projects/PXD000347/files and https://www.ebi.ac.uk/pride/archive/projects/PXD001592/files, respectively). Raw files are compulsory in both 'Partial' and 'Complete' submissions.

    ii. 'RESULT' or 'SEARCH' (required) – Files containing peptide and protein identification data, and are usually produced and exported by a search engine. For 'Complete' submissions, they are labelled as 'RESULT' and must be provided by the users in a standard format: either mzIdentML or (in the near future) mzTab. For 'Partial' submissions, the results are labelled as 'SEARCH', and can be in any relevant file format.

iii. ‘PEAK’ (required for ‘Complete’ submissions, optional for ‘Partial’ submissions) –Processed peak list files. Mascot Generic Format (mgf) (http://www.matrixscience.com/help/data_file_help.html) is one such file type, but there are others (e.g. dta, ms2, pkl and apl).

iv. ‘QUANT’ (optional) –Files containing the peptide/protein quantification results, for example the output of statistical software packages or a list of calculated fold changes.

v. ‘GEL’ (optional) – gel images.

vi. ‘FASTA’/’SP_LIBRARY’ (optional) – Files used for performing the mass spectral searches. First of all, FASTA is a text based file containing protein and polypeptide sequences. Second, SP_LIBRARY can be any file format containing a spectral library.

vii. METADATA (required) – Metadata related to biological or technical aspects of the performed experiment. This section is particularly important, for other researchers attempting to reproduce the data analysis workflow. We encourage users to provide as much information as possible, including a detailed list of mass spectrometer data acquisition settings, the experimental design and detailed information about study factors. The current guidelines are summarized in the detailed PRIDE submission tutorial (available at http://www.proteomexchange.org/docs/Submission_Tutorial.pdf).

viii. ‘OTHER’ (optional) - Any other data type, like programming scripts, pdf files, etc.

2. Create your PRIDE account by filling in the form in the following page: http://www.ebi.ac.uk/pride/archive/register.

3. Go to https://github.com/proteomexchange/px-submission-tool, and scroll down to the ‘Quick Download’ button to download the PX Submission Tool. This Java-based desktop application can be used to select the files for submission, group them together (e.g. raw and the corresponding search engine output files), annotate the study with the required biological and technical metadata, and upload the data to the EMBL-EBI servers.

4. Decompress the ‘px-submission-tool.zip’ folder. To extract files on a Mac computer, double-click it. To extract files on a Windows machine, right-click it and select ‘Extract all’. Note that double-clicking a zipped file on Windows will just display the file contents and will not unzip it. The extracted folder contains a

number of subfolders. One can find useful resources and documentation covering topics not mentioned here in the 'help' folder.

5. Open the application by double clicking the 'px-submission-tool-x.x.x.jar' file ('x' represents any integer number included in the present version of the tool). **Figure 2** shows the initial screen of the tool.

6. Select the 'Complete Submission' option. As indicated above, the other available option is a 'Partial' submission, but here we assume that a 'Complete' submission can be performed. If this is not the case, one can download example files from the recommended dataset PXD004612 (https://www.ebi.ac.uk/pride/archive/projects/PXD004612/files) (**Figure 1)**, and follow this protocol.

7. Click on 'Next'. A second screen containing the information required to start the data submission will appear. Once you have collected all of the information click on 'Next'.

8. Enter your PRIDE login details (created in step 1).

9. Enter submission details. Here one should provide the required metadata about the dataset. The mandatory fields include:

    i. 'Project Title'. Ideally, it should be a descriptive title representing the content of the dataset, rather than being the title of the associated publication.

    ii. 'Keywords'. One should provide a list of words and/or concepts that best categorize the relevant dataset. This will help other researchers to find suitable datasets. A good practice when picking keywords is to use different words to those already present in the title, and select specific rather than very general terms (for example, 'HT55 Human colon carcinoma' versus 'cell line').

    iii. 'Project description'. One should briefly describe the project, in a similar fashion to the abstract of a scientific publication.

    iv. 'Sample processing protocol'. One should describe the experimental methods used for processing the biological samples, and the MS acquisition methods.

    v. 'Data processing protocol'. One should describe here the main data analysis steps, going from the generation of the raw instrument files to the final reported results.

vi. ‘Experiment type’. One should select the terms that can best describe the experimental approach. More than one term can be used.

Click on ‘Next’.

10. Click on the ‘Add Files’ box and select the dataset files to upload. The PX Submission tool will automatically try to guess the file types (‘RAW’, ‘RESULT’, etc.) using their file extension. If that is incorrect, one can manually change the file type from the drop-down menu. For ‘Complete’ submissions, one should provide at least the ‘PEAK’ (peak list files), ‘RESULT’ and ‘RAW’ files. For ‘Partial‘ submissions, search result output files (containing peptide and protein identifications) should be labelled as ‘SEARCH’ rather than ‘RESULT’.

11. Click on ‘Next’. The tool will now check the validity of the file formats supplied (mzIdentML, mzTab (in the near future), the corresponding peak list files). In case of any issues, a warning message will appear.

12. In the next window, one should define the relationships between the files to be uploaded. Click on the ‘+Annotate’ button next to each ‘RESULT’ file. In the resulting window (**Figure 3**), when one clicks on the drop-down menu in each of the boxes, this brings up a set of the most commonly used annotations for different categories: species (mandatory), tissue (mandatory), instrument (mandatory), cell type, disease and quantification method (optional). If the terms applicable to the dataset are not initially present in the most frequently used terms, one has to scroll to the bottom of the list and click on ‘Other’. This will bring up a dialog (connecting to the Ontology Lookup Service (Côté et al., 2010)) where one can search, browse and select the appropriate ontology or controlled vocabulary (CV) terms (**Figure 3**). Furthermore, ‘Experimental factor’ information should be provided here. Experimental factors can be used to describe the biological samples and other experimental variables, for example: control/treatment groups, biological and technical replicates, time-points in a time-series study, etc. This information is often crucial when one wants to perform a reanalysis of the data.

    Once all files are annotated, click on ‘Next’.

13. Fill in the contact details for the principal investigator (or lab head) in this step. Click on ‘Next’.

14. In the next window, one can add additional details. First of all, if the submitted dataset is part of a larger project or Consortium (one example would be the ‘Human Proteome Project’), one can add that information here. Importantly, if the project is based on reprocessing a dataset previously submitted to any PX

repository, or the study generated other 'omics' datasets that are also publicly available, one should enter the details here, to facilitate users to access related genomics, transcriptomics and/or metabolomics datasets. Finally, if the corresponding publication is already published, the corresponding PubMed identifier can be added here as well. Click on 'Next'.

15. In the following window, one can check the contents of the dataset before the actual file submission. As a key point, here one can also export the summary file (a plain text tab-delimited file containing all the submission related information), by clicking the 'Export summary file' box.

16. Click on 'Submit'. The files will then be uploaded to the EMBL-EBI servers. The default upload protocol is the Aspera file transfer (http://asperasoft.com/). However, if that protocol is not possible in your IT environment the tool will automatically switch to the slower but more widely supported FTP file transfer protocol.

17. Provide feedback about the submission process and click on 'Finish'. A screenshot of the feedback window is shown in **Figure 4**.

18. Initially users will receive immediately a confirmation e-mail in response to the data submission. This will contain a submission reference (which is not the PX accession number), and an internal ticket number to facilitate the communication.

19. Once the PRIDE team processes the data successfully, users will receive a further e-mail with the full dataset submission details (dataset PXD accession number) and reviewer account details (including a username and password), for providing access to journal reviewers and editors. It is important to highlight that all datasets are private by default.

PROTOCOL 2

EXPLORING PUBLIC DATASETS IN ProteomeXchange RESOURCES using ProteomeCentral

A large number of independent repositories accommodating various proteomics data are currently available (Perez-Riverol, Alpi, Wang, Hermjakob, & Vizcaíno, 2015). The PX portal (called Proteome Central) website (available at http://proteomecentral.proteomexchange.org/) provides a simple gateway for users willing to search and retrieve public MS-based proteomics data in any of the PX repositories. Although the original data is stored in any of the current four member databases (PRIDE, PeptideAtlas/PASSEL, MassIVE and jPOST), the portal provides a single access point to all PX datasets. This protocol describes how to search for relevant public datasets.

Necessary Resources

Hardware: A computer with Internet connection.

Software: Any web browser.

Searching PX for public datasets

1. Access the PX home page at http://www.proteomexchange.org (**Figure 5**). The home page displays a Twitter timeline for the PX account (@proteomexchange), highlighting the most recently published datasets. Here one can also select links to a) 'Access Data', b) 'Submit Data', and c) 'Access member repositories'.

2. Click on 'Access Data'. This redirects to ProteomeCentral website (http://proteomecentral.proteomexchange.org/), where users can search for datasets using free-text queries or clicking on the interactive graphics to filter the default list of datasets: by 'species', 'title', 'keywords', and 'instrument name' (**Figure 6**).

3. Enter a free-text query in the search box, for example 'metaproteomics'. Users can query public existing datasets by keywords, submitter and/or principal investigator, biological information about the samples (e.g. species, tissue, cell type), protocols, etc. Note that currently it is not possible to query by protein identifier or protein accession number.

4. Once the search has finished, click on one of the links in the dataset Identifier column (PXD identifier). This will redirect to the corresponding dataset page in ProteomeCentral, which contains information such as the dataset title, description, other metadata as well as direct links to the primary PX repository where the project files included in the dataset can be downloaded (e.g. website URL and FTP location).

5. Go back to the search results page and click on one of the links in the publication column. Where available, users will be directed to a PubMed page corresponding to the literature citation associated with the dataset, where all details about the study and the dataset can be found.

Building advanced queries

6. To access the advanced search options, click 'Go to Advanced Search'. The drop-down menu enables a more fine-grained search by combining multiple queries using a Boolean logic.

7. Enter a query in the text boxes. For example, to discover 'what benchmarking datasets are available for the two popular Orbitrap instruments', from the drop-down menus select: 'Keywords' – Contains – 'benchmark', 'Instrument' – Contains – 'orbitrap'. Click on the 'Advanced Search' box to execute the query.

8. The results of the search can be further sorted by species, instrument, etc, by clicking on the corresponding column. The example query and search results are shown in **Figure 7**.

9. At the moment of writing, the first dataset displayed as a result of this search is PXD003472. Go to the corresponding dataset page, by clicking on the dataset identifier. One can see from the 'Hosting Repository' information that this dataset is stored in PRIDE. Scroll down to the bottom of the page and follow the 'PRIDE project URI' link.

10. On the corresponding PRIDE page (http://www.ebi.ac.uk/pride/archive/projects/PXD003472), one can find more information about the dataset by reading the general description, and the sample and data processing protocol fields, among others. One can also download the project files by following the links at the top right side of the page.

Accessing individual PX partner repositories

11. Finally, go back to the main PX portal at http://www.proteomexchange.org. From the 'Member repositories' section one can access all the PX partner repositories: PRIDE, PeptideAtlas/PASSEL, MassIVE and jPOST.

COMMENT

Data submission is naturally the first critical step in the PX data workflow. As illustrated by the Basic Protocol 1, the submission to PX *via* the PRIDE database requires users to provide the appropriate files and information. These requirements, particularly the quality of experimental metadata, are important for ensuring the reproducibility of the studies but also necessary for the data to be reusable. As highlighted at different steps in the submission process, datasets should be annotated with as much information as possible, including ideally the experimental design and information about study factors.

Basic Protocol 2 is devoted to the use of the ProteomeCentral web interface, which enables users to search and retrieve public proteomics datasets available in any of the PX repositories, providing a single access point to the Consortium. Otherwise, users would need to search in each of the PX resources individually.

FUTURE DIRECTIONS

The field of proteomics is still rapidly evolving. New MS techniques and data analysis pipelines are being continuously developed. The complexity and volume of data keep growing in parallel. The PX Consortium and the PRIDE database in particular respond to these challenges by engaging with the proteomics community, and when possible, by adding new functionality and improving the existing submission system. For example, as already mentioned, full support for an additional tab-delimited PSI standard file format, mzTab, is being finalized at present.

PRIDE is also working closely with commercial and free-to-use/open source software providers (e.g. ProteomeDiscoverer, MaxQuant, OpenMS, to name a few) helping to achieve that 'Complete' submissions can be supported from those tools. Support for 'Complete' submissions from Data Independent Acquisition approaches, like the Waters MS^E or Sciex SWATH-MS technology, is anticipated in the medium-term future. Finally, for the advanced users, PRIDE provides a RESTful web service API (Application Programming Interface) that can be used for programmatic access to the data.

With regards to the ProteomeCentral PX portal, in the future we aim to support protein-based (and even peptide-based) queries across all PX resources. Funding is being sought at present with that overall goal in mind.

ACKNOWLEDGEMENTS

The authors want to acknowledge funding provided by the Wellcome Trust [grant number WT101477MA], BBSRC [grant numbers BB/L024225/1 and BB/N022440/1] and EMBL core funding.

REFERENCES

Center for Computational Mass Spectrometry, University of California, S. D. (n.d.). MassIVE. Retrieved February 16, 2017, from http://massive.ucsd.edu/

Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., & Hermjakob, H. (2010). The Ontology Lookup Service: bigger and better. *Nucleic Acids Research*, *38*(Web Server issue), W155-60. https://doi.org/10.1093/nar/gkq331

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., … Aebersold, R. (2006). The PeptideAtlas project. *Nucleic Acids Research*, *34*(90001), D655–D658. https://doi.org/10.1093/nar/gkj040

Deutsch, E. W., Albar, J. P., Binz, P.-A., Eisenacher, M., Jones, A. R., Mayer, G., … Hermjakob, H. (2015). Development of data representation standards by the human proteome organization proteomics standards initiative. *Journal of the American Medical Informatics Association : JAMIA*, *22*(3), 495–506. https://doi.org/10.1093/jamia/ocv001

Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., … Vizcaíno, J. A. (2017). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, *45*(D1), D1100–D1106. https://doi.org/10.1093/nar/gkw936

Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., … Hermjakob, H. (2014). The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics : MCP*, *13*(10), 2765–75. https://doi.org/10.1074/mcp.O113.036681

Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D. L., Dianes, J. A., Del-Toro, N., … Vizcaíno, J. A. (2016). Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods*, *13*(8), 651–656. https://doi.org/10.1038/nmeth.3902

Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., … Creasy, D. (2012). The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & Cellular Proteomics : MCP*, *11*(7), M111.014381. https://doi.org/10.1074/mcp.M111.014381

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., … Apweiler, R. (2005). PRIDE: The proteomics identifications database. *PROTEOMICS*, *5*(13), 3537–3545. https://doi.org/10.1002/pmic.200401303

Okuda, S., Watanabe, Y., Moriya, Y., Kawano, S., Yamamoto, T., Matsumoto, M., … Ishihama, Y. (2017). jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research*, *45*. https://doi.org/10.1093/nar/gkw1080

Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H., & Vizcaíno, J. A. (2015). Making

proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics*, *15*(5–6), 930–49. https://doi.org/10.1002/pmic.201400302

Perez-Riverol, Y., Xu, Q.-W., Wang, R., Uszkoreit, J., Griss, J., Sanchez, A., … Vizcaíno, J. A. (2016). PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. *Molecular & Cellular Proteomics : MCP*, *15*(1), 305–17. https://doi.org/10.1074/mcp.O115.050229

Vizcaíno, J., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., … Hermjakob, H. (2016). 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, *44*(D1), D447–D456. https://doi.org/10.1093/nar/gkv1145

Vizcaíno, J., Deutsch, E. E. W., Wang, R., Vizcaino, J. A., Deutsch, E. E. W., Wang, R., … Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotech*, *32*(3), 223–226. https://doi.org/10.1038/nbt.2839

INTERNET RESOURCES (optional)

1. http://www.proteomexchange.org/

The ProteomeXchange web page contains all the information about the current members of the Consortium, linking to the submission guidelines, and ways to access data.

2. http://proteomecentral.proteomexchange.org/cgi/GetDataset

ProteomeXchange portal (ProteomeCentral). Web portal that enables access to public datasets in any of the PX resources.

3. http://www.ebi.ac.uk/pride/

PRIDE database. Home page of the PRIDE database.

4. http://www.peptideatlas.org/ (PeptideAtlas), http://massive.ucsd.edu (MassIVE), http://jpostdb.org (jPOST).

Home pages of the remaining PX partner repositories: PeptideAtlas, MassIVE and jPOST, respectively.

5. http://www.proteomexchange.org/docs/Submission_Tutorial.pdf

A more detailed tutorial about how to submit datasets to PRIDE.

FIGURE LEGENDS

Figure 1. Screenshot of the PRIDE web interface corresponding to the files included in dataset PXD004280. It contains three types of files: 'RESULT', 'PEAK' and 'RAW'. The files included in this dataset can be used to perform the submission steps described in Basic Protocol 1.

Figure 2. Screenshot of the main graphical user interface in the PX Submission Tool.

Figure 3. Screenshot of the 'Experimental Annotation window' (left) and 'Ontology Lookup Service' based dialog (right) in the PX Submission Tool. Clicking on the 'Other…' option in the scroll down menu for each annotation field brings up the relevant ontology terms.

Figure 4. Screenshot of the window displayed upon successful submission to PRIDE in the PX submission Tool.

Figure 5. Screenshot of the ProteomeXchange home page.

Figure 6. Screenshot of the ProteomeCentral web interface. It provides an overview of the most recently published datasets.

Figure 7. Screenshot of the advanced search window in ProteomeCentral including an example query.

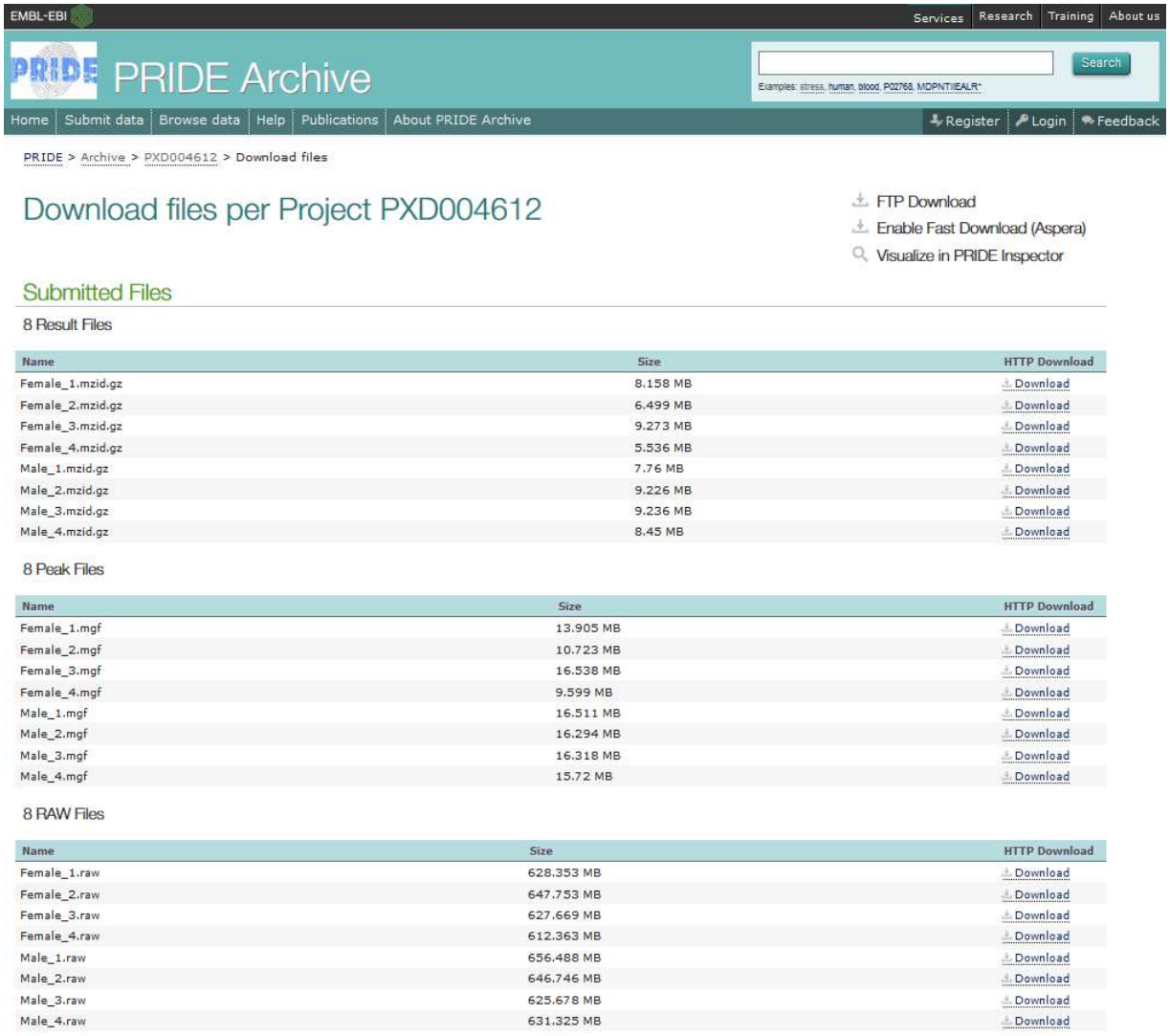
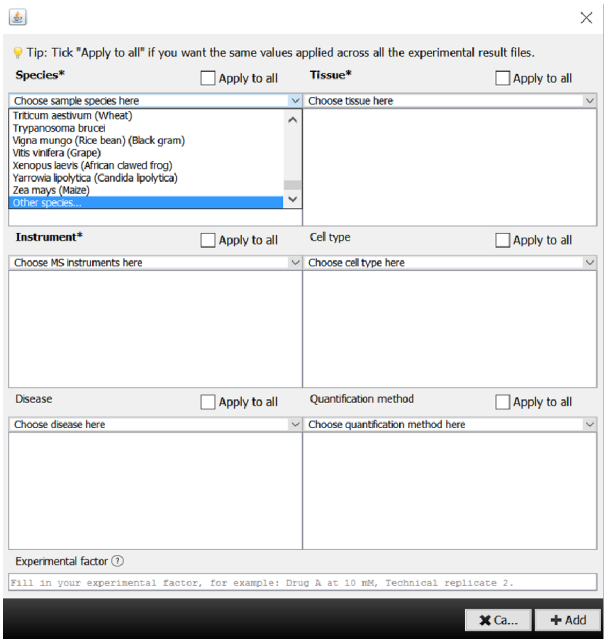

EMBL-EBI
Services
Research
Training
About us
PRIDE Archive
Search
Examples: stress, human, blood, P02768, MDPNTIEALR*
Home
Submit data
Browse data
Help
Publications
About PRIDE Archive
Register
Login
Feedback
PRIDE > Archive > PXD004612 > Download files
Download files per Project PXD004612
FTP Download
Enable Fast Download (Aspera)
Visualize in PRIDE Inspector
Submitted Files
8 Result Files
Name
Size
HTTP Download
Female_1.mzid.gz 8.158 MB Download
Female_2.mzid.gz 6.499 MB Download
Female_3.mzid.gz 9.273 MB Download
Female_4.mzid.gz 5.536 MB Download
Male_1.mzid.gz 7.76 MB Download
Male_2.mzid.gz 9.226 MB Download
Male_3.mzid.gz 9.236 MB Download
Male_4.mzid.gz 8.45 MB Download
8 Peak Files
Name
Size
HTTP Download
Female_1.mgf 13.905 MB Download
Female_2.mgf 10.723 MB Download
Female_3.mgf 16.538 MB Download
Female_4.mgf 9.599 MB Download
Male_1.mgf 16.511 MB Download
Male_2.mgf 16.294 MB Download
Male_3.mgf 16.318 MB Download
Male_4.mgf 15.72 MB Download
8 RAW Files
Name
Size
HTTP Download
Female_1.raw 628.353 MB Download
Female_2.raw 647.753 MB Download
Female_3.raw 627.669 MB Download
Female_4.raw 612.363 MB Download
Male_1.raw 656.488 MB Download
Male_2.raw 646.746 MB Download
Male_3.raw 625.678 MB Download
Male_4.raw 631.325 MB Download

Figure 1.


Welcome
ProteomeXchange Submission Tool (version 2.3.6)
Choose submission option below
Complete Submission (recommended)
Use this option if you can provide your identification results in either mzIdentML or PRIDE XML format. It will then be possible to fully integrate the results in PRIDE and visualise them (e.g. as required by MCP).
In addition to a PXD identifier, a permanent Digital Object Identifier (DOI) will be provided to uniquely identify the dataset.
Partial Submission
You should only choose this option if your search results cannot be converted to mzIdentML or PRIDE XML. Identifications will not be integrated in PRIDE. However, files will be made available to download (and maybe visualised with other external tools).
A PXD identifier will be provided to uniquely identify the dataset, but not a DOI.
Resubmission
Bulk submission
Submission guidelines
Ca...
Back
Next

Figure 2.


Tip: Tick "Apply to all" if you want the same values applied across all the experimental result files.
Species*
Apply to all
Tissue*
Apply to all
Choose sample species here
Choose tissue here
Triticum aestivum (Wheat)
Trypanosoma brucei
Vigna mungo (Rice bean) (Black gram)
Vitis vinifera (Grape)
Xenopus laevis (African clawed frog)
Yarrowia lipolytica (Candida lipolytica)
Zea mays (Maize)
Other species...
Instrument*
Apply to all
Cell type
Apply to all
Choose MS instruments here
Choose cell type here
Disease
Apply to all
Quantification method
Apply to all
Choose disease here
Choose quantification method here
Experimental factor
Fill in your experimental factor, for example: Drug A at 10 mM, Technical replicate 2.
Ca...
Add

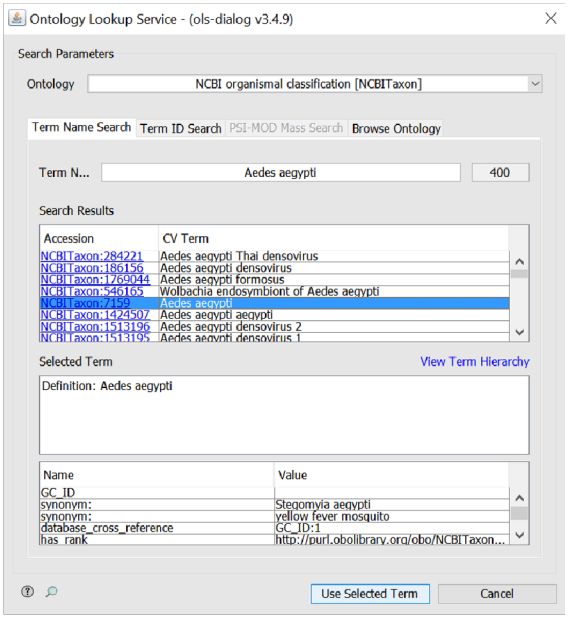

Ontology Lookup Service - (ols-dialog v3.4.9)
Search Parameters
Ontology
NCBI organismal classification [NCBITaxon]
Term Name Search
Term ID Search
PSI-MOD Mass Search
Browse Ontology
Term N...
Aedes aegypti
400
Search Results
Accession
CV Term
NCBITaxon:284221
Aedes aegypti Thai densovirus
NCBITaxon:186156
Aedes aegypti densovirus
NCBITaxon:1769044
Aedes aegypti formosus
NCBITaxon:546165
Wolbachia endosymbiont of Aedes aegypti
NCBITaxon:7159
Aedes aegypti
NCBITaxon:1424507
Aedes aegypti aegypti
NCBITaxon:1513196
Aedes aegypti densovirus 2
Selected Term
View Term Hierarchy
Definition: Aedes aegypti
Name
Value
GC_ID
synonym:
Stegomyia aegypti
synonym:
yellow fever mosquito
database_cross_reference
GC_ID:1
has_rank
http://purl.obolibrary.org/obo/NCBITaxon...
Use Selected Term
Cancel

Figure 3.

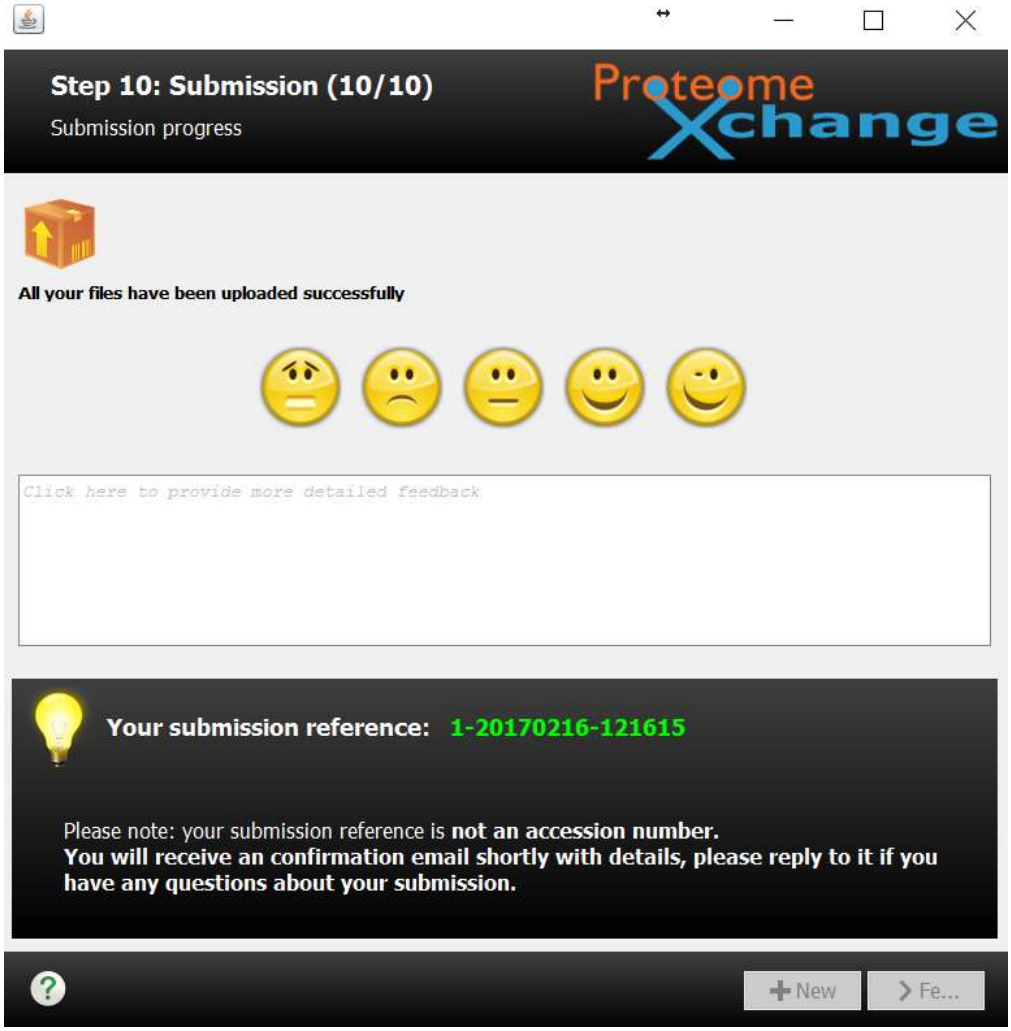

Step 10: Submission (10/10)
Submission progress
ProteomeXchange
All your files have been uploaded successfully
Click here to provide more detailed feedback
Your submission reference: 1-20170216-121615
Please note: your submission reference is not an accession number.
You will receive an confirmation email shortly with details, please reply to it if you have any questions about your submission.
New
Fe...

Figure 4.

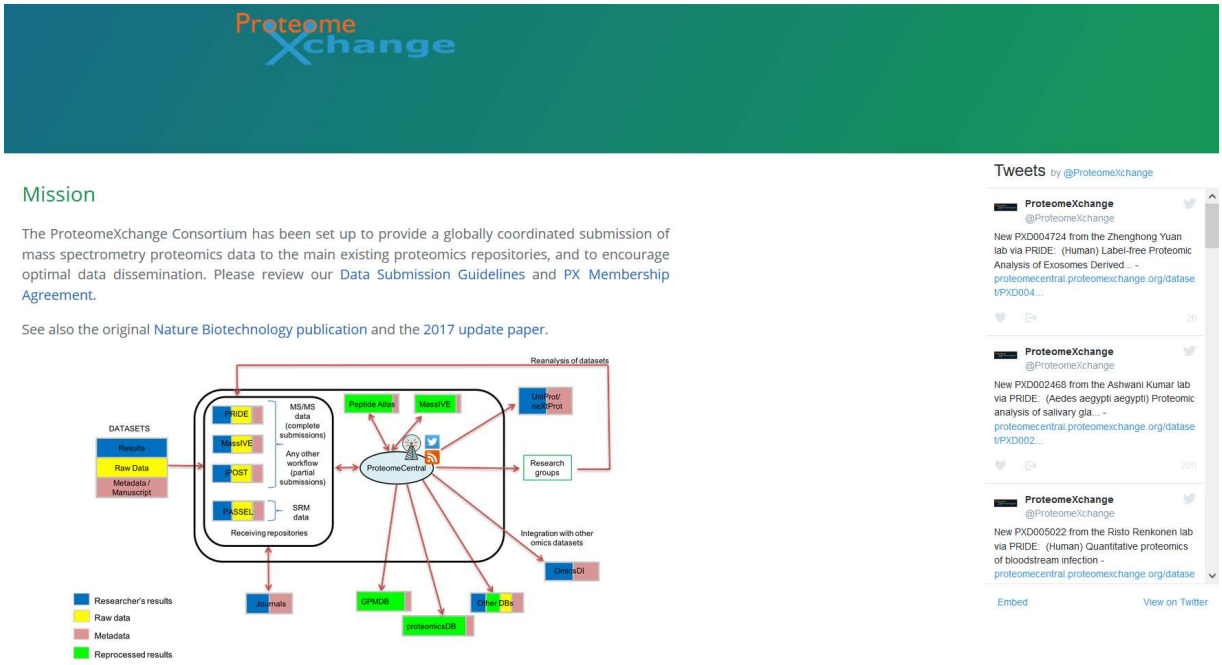
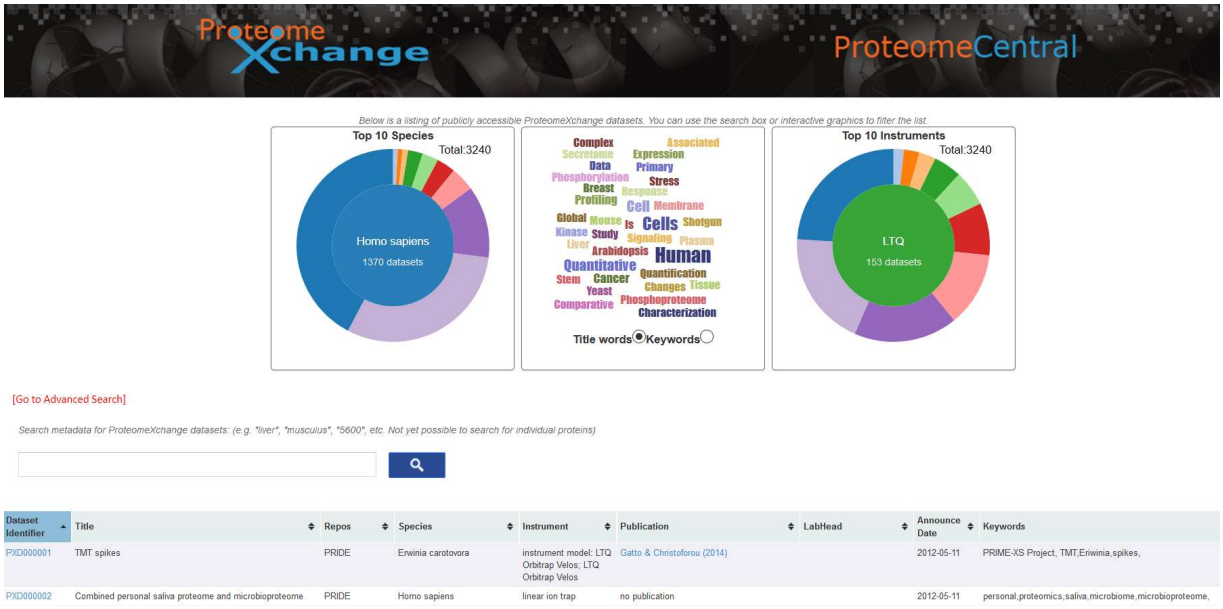

ProteomeXchange
Mission
The ProteomeXchange Consortium has been set up to provide a globally coordinated submission of mass spectrometry proteomics data to the main existing proteomics repositories, and to encourage optimal data dissemination. Please review our Data Submission Guidelines and PX Membership Agreement.
See also the original Nature Biotechnology publication and the 2017 update paper.
Reanalysis of datasets
DATASETS
Results
Raw Data
Metadata / Manuscript
PRIDE
MassIVE
PROST
PASSEL
MS/MS data (complete submissions)
Any other workflow (partial submissions)
SRM data
Receiving repositories
Peptide Atlas
MassIVE
UniProt/ neXtProt
ProteomeCentral
Research groups
Integration with other omics datasets
OmicsDI
Journals
GPMDB
Other DBs
proteomicsDB
Researcher's results
Raw data
Metadata
Reprocessed results
Public Data
Access Data
Public PXD datasets can be browsed over at ProteomeCentral. An RSS feed is also available.
Tweets by @ProteomeXchange
ProteomeXchange
@ProteomeXchange
New PXD004724 from the Zhenghong Yuan lab via PRIDE: (Human) Label-free Proteomic Analysis of Exosomes Derived... - proteomecentral.proteomexchange.org/dataset/PXD004...
ProteomeXchange
@ProteomeXchange
New PXD002468 from the Ashwani Kumar lab via PRIDE: (Aedes aegypti aegypti) Proteomic analysis of salivary gla... - proteomecentral.proteomexchange.org/dataset/PXD002...
ProteomeXchange
@ProteomeXchange
New PXD005022 from the Risto Renkonen lab via PRIDE: (Human) Quantitative proteomics of bloodstream infection - proteomecentral.proteomexchange.org/datase
Embed
View on Twitter

Figure 5.



Dataset Identifier	Title	Repos	Species	Instrument	Publication	LabHead	Announce Date	Keywords
PXD000001	TMT spikes	PRIDE	Erwinia carotovora	instrument model: LTQ Orbitrap Velos; LTQ Orbitrap Velos	Gatto & Christoforou (2014)		2012-05-11	PRIME-XS Project, TMT,Erwinia,spikes,
PXD000002	Combined personal saliva proteome and microbioproteome	PRIDE	Homo sapiens	linear ion trap	no publication		2012-05-11	personal,proteomics,saliva,microbiome,microbioproteome,

Figure 6.

Dataset Identifier	Title	Repos	Species	Instrument	Publication	LabHead	Announce Date	Keywords
PXD003472	S. cerevisiae DDA orbitrap-based benchmark datasets	PRIDE	Saccharomyces cerevisiae (Baker's yeast)	LTQ Orbitrap Velos; Q Exactive	Jarnuczak et al. (2016)	Professor Simon Hubbard	2017-01-30	Technical, benchmark dataset, peptide ionization, peptide detectability, hydrophobicity, coelution,
PXD003659	Probing the protein inventory of the human pathogen Clostridium difficile	PRIDE	Peptoclostridium difficile (strain 630) (Clostridium difficile)	LTQ Orbitrap Elite	Dataset with its publication pending	Andreas Otto	2016-08-12	Biomedical, Clostridium difficile, benchmark proteome, comprehensive peptide library, label-free protein quantification,
PXD000790	PIA - Mouse Benchmark Dataset	PRIDE	Mus musculus	LTQ Orbitrap Elite	Uszkoreit et al. (2015)	Katrin Marcus	2015-05-08	mouse, benchmark,

Figure 7.