How to submit MS proteomics data to ProteomeXchange via the PRIDE database

Tobias Ternent[1]*, *Attila Csordas*[1]*, *Da Qi*[2], *Guadalupe Gómez-Baena*[2], *Robert J. Beynon*[2], *Andrew R. Jones*[2], *Henning Hermjakob*[1] and *Juan Antonio Vizcaíno*[1]

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

The ProteomeXchange (PX) consortium has been established to standardize and facilitate submission and dissemination of MS-based proteomics data in the public domain. In the consortium, the PRIDE database at the European Bioinformatics Institute, acts as the initial submission point of MS/MS data sets. In this manuscript, we explain step by step the submission process of MS/MS data sets to PX via PRIDE. We describe in detail the two available workflows: 'complete' and 'partial' submissions, together with the available tools to streamline the process. Throughout the manuscript, we will use one example data set containing identification and quantification data, which has been deposited in PRIDE/ProteomeXchange with the accession number PXD000764 (http://proteomecentral.proteomexchange.org/dataset/PXD000764).

Received: April 2, 2014
Revised: June 11, 2014
Accepted: July 17, 2014

Keywords:
Bioinformatics / Data standard / Data submission / Data visualization / Proteomics repositories

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

1 Introduction

The availability of MS-based proteomics data in the public domain is still low when compared with other 'omics' disciplines such as genomics and transcriptomics. However, due to the guidelines promoted by several scientific journals and funding agencies [1], and the general perception that sharing data is a good scientific practise and beneficial for the field, the culture in the proteomics community is evolving in that direction. Several MS proteomics repositories have been established to address the demand for storage and availability of proteomics data in the public domain. Two of the most prominent resources, the PRIDE database (European Bioinformatics Institute (EBI), Cambridge, UK) [2] and PeptideAtlas (Institute for Systems Biology, ISB, Seattle, USA) [3] have led to the development of the ProteomeXchange (PX) consortium (http://www.proteomexchange.org). The goal of PX is to provide a common framework and infrastructure for the cooperation of proteomics resources by defining and implementing standard and user-friendly data deposition and dissemination procedures [4]. Furthermore, the main objective is to provide the scientific community with an easier and unified way to submit and access MS proteomics data.

In the first stable implementation of the PX data workflow [4], PRIDE acts as the initial submission point of MS/MS data whereas PASSEL (PeptideAtlas Selected Reaction Monitoring (SRM) Experiment Library) [5] at ISB has the equivalent role for SRM data. The PRIDE database

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

Abbreviations: DOI, digital object identifier; **EBI**, European Bioinformatics Institute; **ISB**, Institute for Systems Biology; **mgf**, Mascot generic file; **OLS**, Ontology Lookup Service; **PASSEL**, PeptideAtlas SRM Experiment Library; **PSI**, Proteomics Standards Initiative; **PX**, ProteomeXchange; **SRM**, Selected Reaction Monitoring

*These authors have contributed equally to this work.
Colour Online: See the article online to view Figs. 1–3 in colour.

is a data repository including protein/peptide identification and expression information (including PTMs), the supporting spectral evidence (both peak lists and raw data) and the related biological and technical metadata [2]. Data submitted to PRIDE remain private during the manuscript review process. Once the manuscript is published or the submitters give their permission, data in PRIDE are disseminated through the ProteomeCentral, the portal of all PX submissions (http://proteomecentral.proteomexchange.org/cgi/GetDataset) [4]. Data in PRIDE are linked from ProteomeCentral and can be accessed directly through the new PRIDE Archive web interface. By June 2014, MS/MS data sets in PRIDE have accounted for ~95% of all the PX data sets. In this manuscript we will describe in detail how to perform submissions of MS/MS data to PX via the PRIDE database.

2 Before performing a submission

2.1 Definitions of data types and files

There are a variety of data types in proteomics that can be submitted to PX/PRIDE. For complete definitions of the different data types and the corresponding data formats, see Supporting Information, Section 1. The data types and corresponding file tags are as follows:

(i) Mass spectrometer output files, labelled as 'RAW'.
(ii) Processed peak lists, labelled as 'PEAK'.
(iii) Search engine output files: Processed identification results are labelled either as 'RESULT' (if they are available in a standard format: either mzIdentML [6] or PRIDE XML) or 'SEARCH' (any other file format). They contain peptide/protein identification data and in some cases quantification information also, if identification and quantification are performed at the same time.
(iv) Quantification software output files: Quantification results, labelled as 'QUANT'.
(v) Metadata: Related biological or technological metadata provide the experimental context.
(vi) Gel images, labelled as 'GEL'.
(vii) Files used to perform the mass spectral search, either sequence database files (labelled as 'FASTA') or spectral library files (labelled as 'SP_LIBRARY').
(viii) Any other data type (e.g. scripts, pdf files, etc.): They are labelled as 'OTHER'.

2.2 Submission types to PX via PRIDE

Two different submission types are available: 'complete' and 'partial'. In both cases, 'RAW' files and metadata are mandatory. Also in both cases, processed identification results are also required, but the difference occurs with the file format in which these results are provided:

(i) 'Complete' submission: Processed identification results are provided as either PRIDE XML or mzIdentML ('RESULT') files. If mzIdentML is used, the corresponding 'PEAK' files referenced from the mzIdentML files are also mandatory. A 'complete' submission ensures that the processed results data can be integrated in the PRIDE database, visualized using the PRIDE Inspector tool (see Section 5), and that the identification information is made fully searchable.
(ii) 'Partial' submission: Processed identification results are provided in other formats ('SEARCH' files). The processed results cannot be integrated and made searchable in PRIDE, or visualized using PRIDE Inspector. However, all the files are available to download. This mechanism allows data generated from software that cannot export to standard formats, or from novel experimental approaches to be deposited into PRIDE.

In both cases, other data types can be provided optionally such as 'QUANT', 'GEL', 'FASTA', 'SP_LIBRARY' or 'OTHER'.

2.3 Example data set

The title of the example data set used is 'Discovery of new cerebrospinal fluid biomarkers for meningitis in children'. The data set consists of 12 runs: four of them are non-infected samples (controls) and the other eight are infected samples (positive for bacterial meningitis). It was deposited in PRIDE/ProteomeXchange as a 'complete' submission (accession number PXD000764, DOI 10.6019/PXD000764 (where DOI is digital object identifier)). It can be accessed at http://www.ebi.ac.uk/pride/archive/projects/PXD000764. Complete details about how the data set was generated are available in Supporting Information, Section 2.

3 PRIDE submission overview

A general overview of the submission process for a 'complete' submission can be seen in Fig. 1. The process can be split in four stages (see also the submission 'Cheat Sheet' in Supporting Information):

(i) Get the files ready for the submission: This involves the conversion or export of the processed identification results and corresponding spectra files into mzIdentML or PRIDE XML ('RESULT') files.
(ii) Check the files before submission: This involves using a visualization tool such as PRIDE Inspector.
(iii) Perform the submission: It mainly involves the annotation and upload of the data set into PRIDE using the PX submission tool or via command line.
(iv) Post-submission stage: Refinement of the data set may be required, in communication with the PRIDE team.

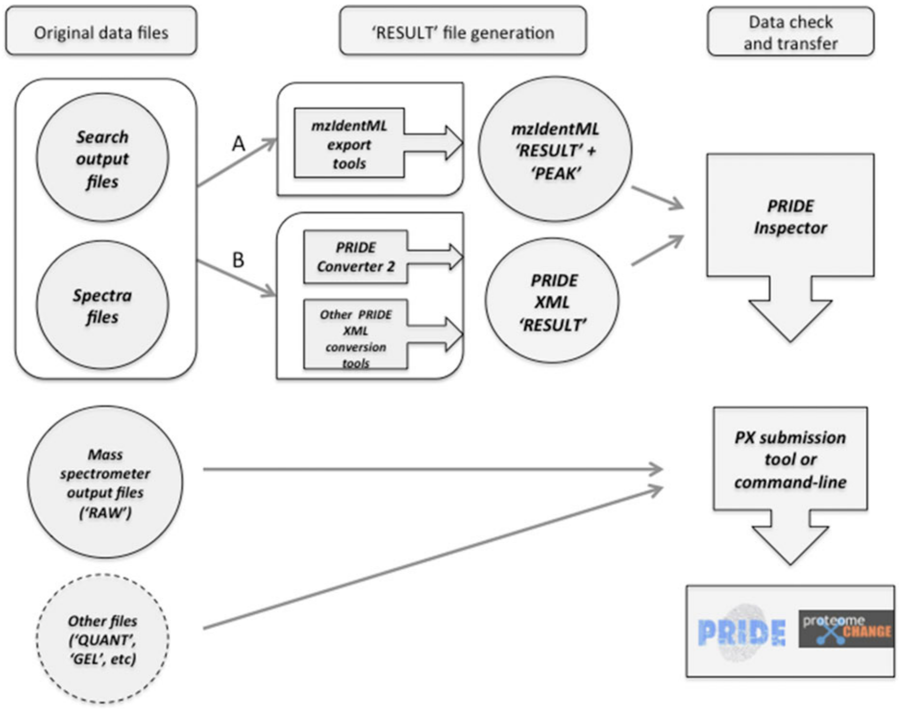


Figure 1. Overview of the PRIDE/ProteomeXchange 'complete' submission workflow. First data files associated are collected and the search output (processed results) and spectra files are converted into 'RESULT' files (A: mzIdentML + 'PEAK' files; or B: PRIDE XML), which can then be checked with a visualisation tool such as PRIDE Inspector. The data set is then transferred using the PX submission tool or via command line, as explained in the main text.

Although generally there are different options available to generate 'RESULT' files for performing a 'complete' submission, it might happen that the actual pipeline or analysis software used by the submitter cannot provide such files currently. In these cases, a 'partial' submission is the only available option (see an overview figure for 'partial' submissions in Supporting Information Fig. 1). For 'partial' submissions, stages (i) and (ii) cannot be performed, so the submitters will start in stage (iii).

4 Get the files ready for the submission

Once the submitter has all the necessary files, the first thing that needs to be checked is whether the original search engine output files plus the corresponding spectra files can be converted or exported into either mzIdentML or PRIDE XML ('RESULT') files. The user can then decide whether to perform a 'complete' or 'partial' submission.

Table 1 lists the available tools that implement export to mzIdentML (see http://www.psidev.info/tools-implementing-mzIdentML for an updated version). Table 2 lists the tools implementing conversion or export to PRIDE XML. Each tool will have its own way of generating the files so we recommend users to check the manuals available for each software. Over time, we are working with the producers of the main proteomics software packages to enable export to the accepted data standards.

PRIDE Converter 2 (http://code.google.com/p/pride-converter-2/) can be used to convert a variety of popular proteomics data formats (search engine output files, e.g. Mascot .dat, X!Tandem .xml, etc.), into well-annotated PRIDE XML files [7]. PRIDE Converter 2 can be used in two modes: as a graphical user interface or command line interface. Tutorials are available at https://code.google.com/p/pride-converter-2/downloads/list.

The mzIdentML 'RESULT' files present in the example data set were generated using Mascot Server version 2.4 (Matrix Science, http://www.matrixscience.com/help/export_help.html#MZIDENTML). The accompanying 'PEAK' files were mgf (Mascot generic files, see full details in Supporting Information, Section 2).

5 Check the files before submission

It is advisable to check the data in detail before performing the data submission. These checks can ensure that data are annotated correctly and that there are no obvious issues or inconsistencies [8]. PRIDE Inspector (http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector) can be used to visualize and perform an initial quality assessment of the submitted data [9]. It can support the PRIDE XML and mzIdentML formats ('RESULT' files), together with a number of different spectra file formats.

For instance, one of the valuable checks that can be performed is to observe the 'delta *m/z*' outlier values for the reported peptide identifications, which are calculated as the difference between the experimental *m/z* value and the theoretical mass of the identified peptide [10]. If the resulting

Table 1. List of the tools that implement export to the mzIdentML format (version 1.1), by June 2014

Tool	Formats	URL
idConvert (ProteoWizard) [15]	pep.xml, prot.xml (Trans Proteomic Pipeline)	http://proteowizard.sourceforge.net/
IDPicker [16]	Native support	http://fenchurch.mc.vanderbilt.edu/software.php
Mascot (Matrix Science)	Native support	http://www.matrixscience.com/
MS-GF+	Native support	http://proteomics.ucsd.edu/Software/MSGFPlus.html#pubs
mzIdLibrary [11]	OMSSA .xml, X!Tandem .xml	https://code.google.com/p/jmzidentml
Myrimatch [17]	Native support	http://fenchurch.mc.vanderbilt.edu/software.php
OpenMS [18]	Native support	http://open-ms.sourceforge.net/
PAnalyzer [19]	Native support	https://code.google.com/p/ehu-bio/wiki/PAnalyzer
Peaks (Bioinformatics Solutions Inc)	Native support	http://www.bioinfor.com/
Phenyx (GeneBio)	Native support	http://www.genebio.com/products/phenyx/
ProCon	Sequest .out ProteomeDiscoverer (v1.2/1.3/1.4) .msf ProteinScape 2.1 (Bruker) database content	http://www.medizinisches-proteom-center.de/procon
Pepitome [20]	Native support	http://fenchurch.mc.vanderbilt.edu/software.php
ProteinPilot (AB SCIEX)	Native support	From ProteinPilot 5.0 (to be available by the end of 2014)
Scaffold (Proteome Software)	Native support	http://www.proteomesoftware.com/products/scaffold/
TagRecon [21]	Native support	http://fenchurch.mc.vanderbilt.edu/software.php
PeptideShaker	Native support	https://code.google.com/p/peptide-shaker/

Updated information is available at http://www.psidev.info/tools-implementing-mzIdentML#.

value is outside of a normal range (depending on the accuracy of the mass spectrometer used), this constitutes a good indication that something has gone wrong in either the annotation or in the data generation, the former being the most likely option. Supporting Information Fig. 2 is a screenshot taken from one of the mzIdentML files (plus the corresponding mgf file) from the example data set.

There are other free tools available for visualizing 'RESULT' files. For mzIdentML files, 'ProteoIDViewer' [11] has some extra features not supported currently by PRIDE Inspector, for example calculating identification statistics if a decoy database search has been performed.

In the case of 'partial' submissions, processed results output in other formats ('SEARCH' files) are submitted. In this case, the assessment and careful investigation of the data is often not possible with freely available tools. Some journals such as *Molecular and Cellular Proteomics* mandate to provide annotated spectra in several scenarios. 'Complete'

Table 2. List of the tools that implement the conversion or export to the PRIDE XML format, by June 2014

Tool	Formats	URL
EasyProt [22]	Native support	http://easyprot.unige.ch/
hEIDI	Native support	http://biodev.extra.cea.fr/docs/heidi
PeptideShaker	Native support	https://code.google.com/p/peptide-shaker/
PRIDE Converter 2 [7]	Mascot .dat, X!Tandem .xml, OMSSA .csv, Crux .txt, ProteomeDiscoverer .msf (plus the corresponding spectra files)	https://code.google.com/p/pride-converter-2/
OmicsHub Proteomics (Integromics)	Native support	https://www.integromics.com/products/proteomics/
ProteinLynx Global Server (PLGS, Waters Corporation).	Native support	http://www.waters.com/waters/en_US/ProteinLynx-Global-SERVER-(PLGS)/nav.htm?cid=513821&locale=en_US
Proteios [23]	Native support	http://www.proteios.org/
ProteoRed MIAPE Extractor tool [24]	Native support	http://www.proteored.org/MIAPEExtractor

MIAPE, minimum information about a proteomics experiment.

submissions can fulfil this requirement. However, 'partial' ones can only meet this requirement if a free spectral viewer (e.g. [12]) is available (see http://www.ebi.ac.uk/pride/help/archive/faq-journal-MCP).

6 Perform the data submission

Before starting the process, users must first register at PRIDE (http://www.ebi.ac.uk/pride/archive/register). The default assumption is that all of the files belonging to one study or manuscript will be uploaded at once and handled as a unit (corresponding to one PX identifier). However in practice, there is some flexibility for the submitter about how to organize the submission. Splitting the data set associated with one manuscript into different sub-data sets can be acceptable if there are sensible reasons to do it. There are two alternatives available for actually performing the submission:

(i) The PX submission tool: From version 2.1 (available from June 2014), it makes use of the Aspera file transfer protocol (http://asperasoft.com/) by default. Aspera functionality usually provides much faster file transfer speeds than FTP (up to 50 times), but this depends on the location where the submission is done. However, the tool can also provide FTP transfer functionality for those cases where there are issues with Aspera.

(ii) Via command-line (using the Aspera file transfer protocol). This option is available for submitters with bioinformatics support who prefer not to use the PX submission tool, due to the manual work involved (e.g. if the submission contains a large number of files). Some scripting knowledge is needed to follow this approach. All the details about this alternative are available in Supporting Information, Section 3.

6.1 Submission using the PX submission tool

The PX submission tool (http://www.proteomexchange.org/submission) is a stand-alone tool that can be used to perform the data submission [4]. It can (i) select all the files to be submitted; (ii) group related different file types (e.g. the corresponding 'RAW' and 'RESULT' files); (iii) ensure a minimum level of experimental annotation and (iv) transfer the files to the EBI. We will describe briefly the steps involved in the submission using the example data set. Figs. 2 (steps 1–4) and 3 (steps 5–8) display an overview of the whole process.

More details are available in the web tutorial available in the EBI e-learning platform (http://www.ebi.ac.uk/training/online/course/proteomexchange-submissions-pride) or at the PX/PRIDE submission manual (http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/px_submission_tutorial.pdf).

6.1.1 Step 1 – Submission type

After the PX submission tool is launched, the type of submission must be chosen: 'complete' or 'partial'. In this case, we will follow the 'complete' submission route (Fig. 2, panel 1).

6.1.2 Step 2 – Data set details

Basic metadata are provided to describe the overall study, such as title, description, sample processing and data processing protocols, keywords and experiment type (this one is selected from a pre-defined list; Fig. 2, panel 2).

6.1.3 Step 3 – Adding files

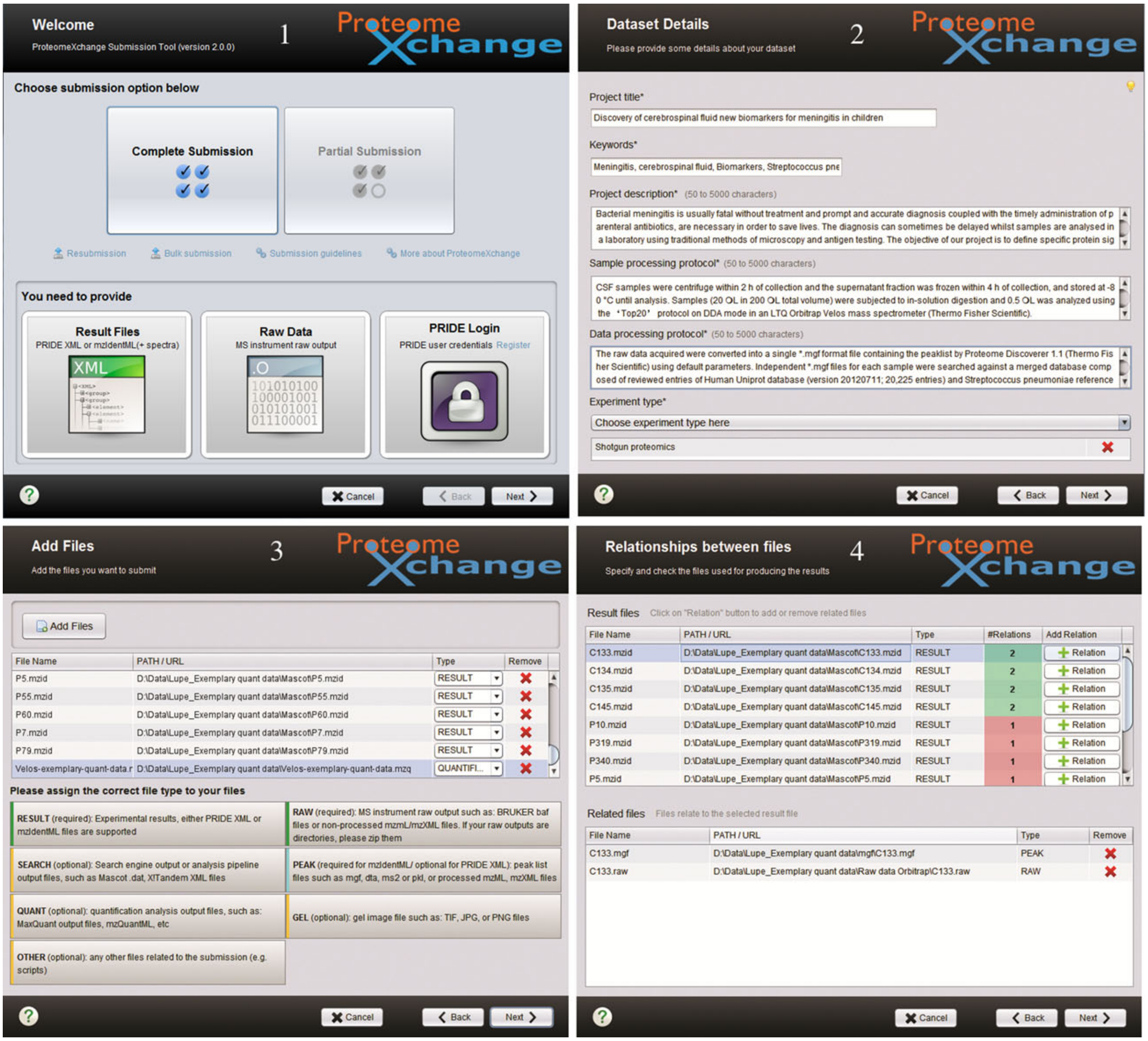
It involves the selection and tagging of the files to be submitted. There are two variants of 'complete' submission depending on the type of 'RESULT' files used: PRIDE XML or mzIdentML. The difference between these two subtypes is that PRIDE XML files do not require additional 'PEAK' files to be included, but mzIdentML files do. The example data set contained 12 raw files ('RAW'), 12 mzIdentML files ('RESULT') and the corresponding 12 mgf files (PEAK), and one mzQuantML file ('QUANT') [13] (Fig. 2, panel 3). Once the files are selected, an appropriate file tag is assigned automatically by the tool ('RAW', 'RESULT', etc.).

6.1.4 Step 4 – Mapping files

In this step, the relationships between the different files can be captured. For the example data set, each 'RESULT' file was related to exactly one 'PEAK' and one 'RAW' file. All 'RESULT' files were related to the same, single 'QUANT' file (Fig. 2, panel 4). The PX submission tool attempts to automatically map the relationships based upon similar file names, which can be edited manually. For this reason, sensible file names are encouraged as this step will take significantly less time in case of many files.

6.1.5 Step 5 – Annotation

In this step, each 'RESULT' file needs to be annotated with sample-related metadata. The following information is required per file: species, tissue and the mass spectrometer. Optionally, information about the cell type, disease or quantification method (if relevant) can also be entered. The annotations are provided as controlled vocabulary terms from drop-down menus (Fig. 3, panel 1). If the appropriate terms are not present in these drop-down menus, the users can search for them through the Ontology Lookup Service (OLS) [14]. Finally, experimental factor related information can also be provided.



Welcome
ProteomeXchange Submission Tool (version 2.0.0)

Choose submission option below

Complete Submission

Partial Submission

You need to provide

Result Files
PRIDE XML or mzIdentML(+ spectra)

Raw Data
MS instrument raw output

PRIDE Login
PRIDE user credentials Register

Dataset Details
Please provide some details about your dataset

Project title*
Discovery of cerebrospinal fluid new biomarkers for meningitis in children

Keywords*
Meningitis, cerebrospinal fluid, Biomarkers, Streptococcus pne

Project description* (50 to 5000 characters)
Bacterial meningitis is usually fatal without treatment and prompt and accurate diagnosis coupled with the timely administration of p arenteral antibiotics, are necessary in order to save lives. The diagnosis can sometimes be delayed whilst samples are analysed in a laboratory using traditional methods of microscopy and antigen testing. The objective of our project is to define specific protein sig

Sample processing protocol* (50 to 5000 characters)
CSF samples were centrifuge within 2 h of collection and the supernatant fraction was frozen within 4 h of collection, and stored at -8 0 °C until analysis. Samples (20 OL in 200 OL total volume) were subjected to in-solution digestion and 0.5 OL was analyzed using the 'Top20' protocol on DDA mode in an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific).

Data processing protocol* (50 to 5000 characters)
The raw data acquired were converted into a single *.mgf format file containing the peaklist by Proteome Discoverer 1.1 (Thermo Fis her Scientific) using default parameters. Independent *.mgf files for each sample were searched against a merged database comp osed of reviewed entries of Human Uniprot database (version 20120711; 20,225 entries) and Streptococcus pneumoniae reference

Experiment type*
Choose experiment type here
Shotgun proteomics ✖

Add Files
Add the files you want to submit

File Name	PATH / URL	Type	Remove
P5.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P5.mzid	RESULT	✖
P55.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P55.mzid	RESULT	✖
P60.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P60.mzid	RESULT	✖
P7.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P7.mzid	RESULT	✖
P79.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P79.mzid	RESULT	✖
Velos-exemplary-quant-data.r	D:\Data\Lupe_Exemplary quant data\Velos-exemplary-quant-data.mzq	QUANTIFI...	✖

Please assign the correct file type to your files

RESULT (required): Experimental results, either PRIDE XML or mzIdentML files are supported

RAW (required): MS instrument raw output such as: BRUKER baf files or non-processed mzmL/mzXML files. If your raw outputs are directories, please zip them

SEARCH (optional): Search engine output or analysis pipeline output files, such as Mascot .dat, X!Tandem XML files

PEAK (required for mzIdentML/ optional for PRIDE XML): peak list files such as mgf, dta, ms2 or pkl, or processed mzML, mzXML files

QUANT (optional): quantification analysis output files, such as: MaxQuant output files, mzQuantML, etc

GEL (optional): gel image file such as: TIF, JPG, or PNG files

OTHER (optional): any other files related to the submission (e.g. scripts)

Relationships between files
Specify and check the files used for producing the results

Result files Click on "Relation" button to add or remove related files

File Name	PATH / URL	Type	#Relations	Add Relation
C133.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C133.mzid	RESULT	2	Relation
C134.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C134.mzid	RESULT	2	Relation
C135.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C135.mzid	RESULT	2	Relation
C145.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C145.mzid	RESULT	2	Relation
P10.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P10.mzid	RESULT	1	Relation
P319.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P319.mzid	RESULT	1	Relation
P340.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P340.mzid	RESULT	1	Relation
P5.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P5.mzid	RESULT	1	Relation

Related files Files relate to the selected result file

File Name	PATH / URL	Type	Remove
C133.mgf	D:\Data\Lupe_Exemplary quant data\mgf\C133.mgf	PEAK	✖
C133.raw	D:\Data\Lupe_Exemplary quant data\Raw data Orbitrap\C133.raw	RAW	✖

Figure 2. Screenshots of the submission of the example data set using the PX submission tool: steps 1–4, as explained in the main text.

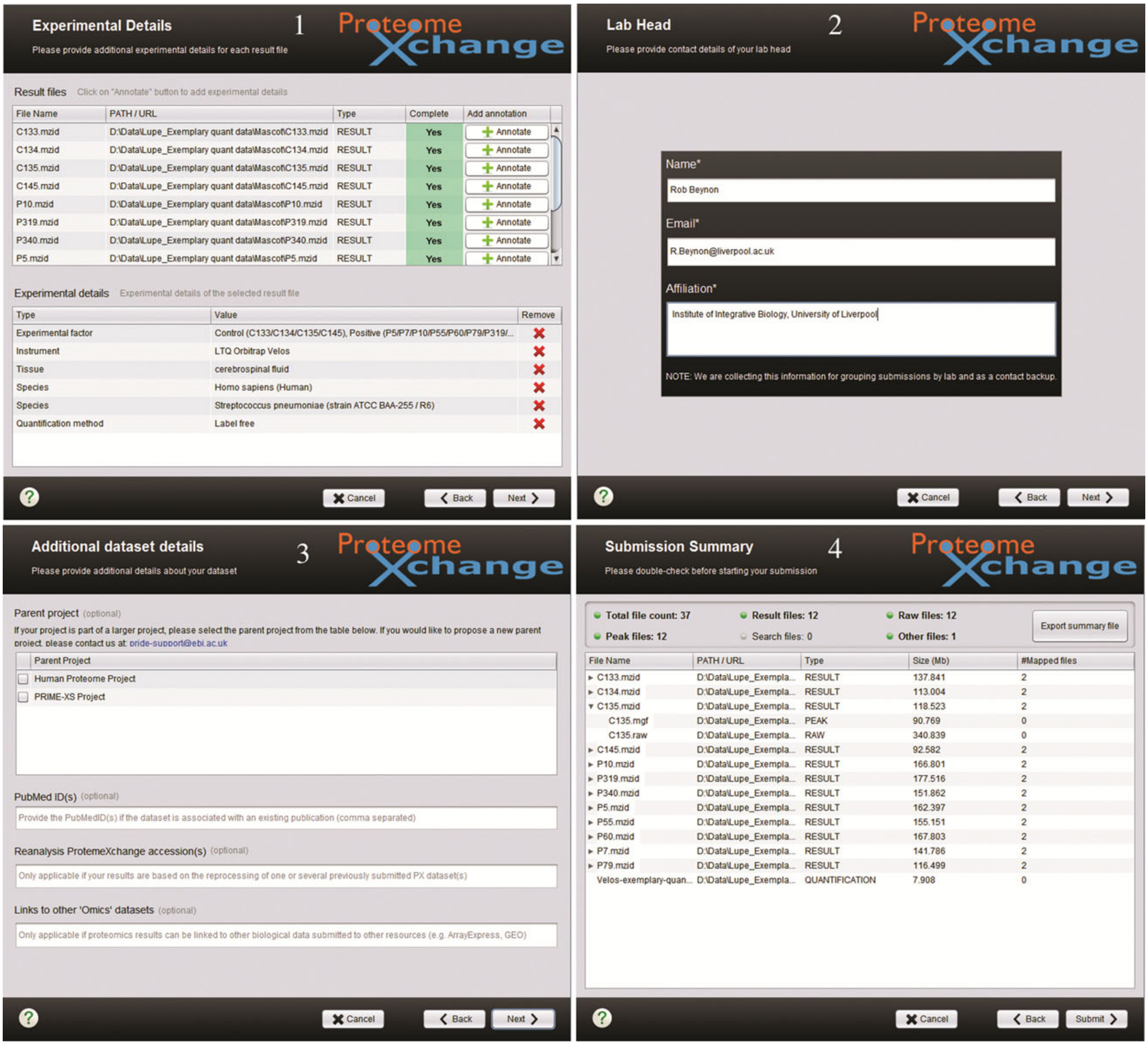
6.1.6 Step 6 – Lab head

The lab head or principal investigator contact details need to be provided here (Fig. 3, panel 2). PRIDE is keeping track of this information to facilitate the attribution of data sets.

6.1.7 Step 7 – Additional details

In this optional step, additional metadata can be provided, if relevant (Fig. 3, panel 3). In the case of the example data set, this step was skipped since none of the extra annotations were needed. First, 'parent project' tags can be added that can be used to group data sets (e.g. 'Human Proteome Project'). The PRIDE team needs to be contacted in advance if new 'parent projects' are needed (pride-support@ebi.ac.uk). Furthermore, if the same biological sample has been investigated using experimental approaches other than proteomics (e.g. transcriptomics, metabolomics, etc.) and the corresponding data are available in other public resources, then it is encouraged to provide those external identifiers (see http://www.ebi.ac.uk/pride/help/archive/faq - multi-omics).

Additionally, it is also possible to provide a PubMed identifier if the corresponding manuscript is already published at the submission time. Finally, it is also encouraged to provide a PX identifier, if the submitted data set constitutes a reanalysis of a previously submitted data set to PX. This allows to link different analyses performed over the same original data.



Experimental Details

Please provide additional experimental details for each result file

Result files Click on "Annotate" button to add experimental details

File Name	PATH / URL	Type	Complete	Add annotation
C133.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C133.mzid	RESULT	Yes	Annotate
C134.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C134.mzid	RESULT	Yes	Annotate
C135.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C135.mzid	RESULT	Yes	Annotate
C145.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\C145.mzid	RESULT	Yes	Annotate
P10.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P10.mzid	RESULT	Yes	Annotate
P319.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P319.mzid	RESULT	Yes	Annotate
P340.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P340.mzid	RESULT	Yes	Annotate
P5.mzid	D:\Data\Lupe_Exemplary quant data\Mascot\P5.mzid	RESULT	Yes	Annotate

Experimental details Experimental details of the selected result file

Type	Value	Remove
Experimental factor	Control (C133/C134/C135/C145), Positive (P5/P7/P10/P55/P60/P79/P319/...	✗
Instrument	LTQ Orbitrap Velos	✗
Tissue	cerebrospinal fluid	✗
Species	Homo sapiens (Human)	✗
Species	Streptococcus pneumoniae (strain ATCC BAA-255 / R6)	✗
Quantification method	Label free	✗

Lab Head

Please provide contact details of your lab head

Name*: Rob Beynon

Email*: R.Beynon@liverpool.ac.uk

Affiliation*: Institute of Integrative Biology, University of Liverpool

NOTE: We are collecting this information for grouping submissions by lab and as a contact backup.

Additional dataset details

Please provide additional details about your dataset

Parent project (optional)

If your project is part of a larger project, please select the parent project from the table below. If you would like to propose a new parent project, please contact us at: pride-support@ebi.ac.uk

	Parent Project
☐	Human Proteome Project
☐	PRIME-XS Project

PubMed ID(s) (optional)

Provide the PubMedID(s) if the dataset is associated with an existing publication (comma separated)

Reanalysis ProtemeXchange accession(s) (optional)

Only applicable if your results are based on the reprocessing of one or several previously submitted PX dataset(s)

Links to other 'Omics' datasets (optional)

Only applicable if proteomics results can be linked to other biological data submitted to other resources (e.g. ArrayExpress, GEO)

Submission Summary

Please double-check before starting your submission

● Total file count: 37 ● Result files: 12 ● Raw files: 12
● Peak files: 12 ○ Search files: 0 ● Other files: 1

File Name	PATH / URL	Type	Size (Mb)	#Mapped files
▸ C133.mzid	D:\Data\Lupe_Exempla...	RESULT	137.841	2
▸ C134.mzid	D:\Data\Lupe_Exempla...	RESULT	113.004	2
▾ C135.mzid	D:\Data\Lupe_Exempla...	RESULT	118.523	2
C135.mgf	D:\Data\Lupe_Exempla...	PEAK	90.769	0
C135.raw	D:\Data\Lupe_Exempla...	RAW	340.839	0
▸ C145.mzid	D:\Data\Lupe_Exempla...	RESULT	92.582	2
▸ P10.mzid	D:\Data\Lupe_Exempla...	RESULT	166.801	2
▸ P319.mzid	D:\Data\Lupe_Exempla...	RESULT	177.516	2
▸ P340.mzid	D:\Data\Lupe_Exempla...	RESULT	151.862	2
▸ P5.mzid	D:\Data\Lupe_Exempla...	RESULT	162.397	2
▸ P55.mzid	D:\Data\Lupe_Exempla...	RESULT	155.151	2
▸ P60.mzid	D:\Data\Lupe_Exempla...	RESULT	167.803	2
▸ P7.mzid	D:\Data\Lupe_Exempla...	RESULT	141.786	2
▸ P79.mzid	D:\Data\Lupe_Exempla...	RESULT	116.499	2
Velos-exemplary-quan...	D:\Data\Lupe_Exempla...	QUANTIFICATION	7.908	0

Figure 3. Screenshots of the submission of the example data set using the PX submission tool: steps 5–8, as explained in the main text.

6.1.8 Step 8 – Summary screen

It provides an overview of the whole submission process, including information about all the files (including tags and sizes) and file mappings (Fig. 3, panel 4). The idea is to enable users to perform a final review before the actual file upload takes place.

6.1.9 Step 9 – Upload of files

The submitters can proceed to the final step, where the tool uploads the files using Aspera file transfer functionality by default. In addition, the tool also uploads a 'PX summary file' (see Supporting Information, Sections 1 and 3) that is created by the tool in the background, which summarizes all the information submitted. When the transfer has finished, the submitter will get a confirmation e-mail. The actual time required to upload a data set logically depends on the size of the data set and bandwidth available.

6.1.10 Differences for 'partial' submissions

The steps involved in a 'partial' submission are almost identical. Obviously in step 1, the 'partial' submission option needs to be selected. Additionally, processed results ('SEARCH') files in different formats are uploaded, together with the 'RAW' and optionally, other file types ('QUANT', 'PEAK', etc.). The other main difference is that each data set is annotated as a whole (e.g. sample metadata), instead of annotating each individual file.

6.1.11 Bulk submissions

The mechanism to perform bulk submissions (aimed for very large data sets) using the PX submission tool or via command-line is explained in Supporting Information, Section 4. The 'PX summary file' needs to be created independently before performing the submission, so some scripting experience is required.

7 Post-submission steps

7.1 Internal checks

The submitted files will be checked automatically by the PRIDE internal pipelines. The output of these checks will be first reviewed by a curator [10] and if some inconsistency or missing information is detected, the curators will contact back the submitter in an iterative manner until the data set is considered to be correct. A PX PXD identifier will then be issued for each data set. Additionally, for 'complete' submissions, a DOI and PRIDE assay accession numbers will also be generated. The submitter will also receive a username and password for providing private access to the data. Information about how to access a private data set is available at Supporting Information, Section 5.

7.2 How to modify an already submitted data set

A given data set can be modified while it remains private. This can be done through the 'Resubmission' option using the PX submission tool (available in step 1, Fig. 2, panel 1 and Supporting Information Figs. S4 and S5). The whole data set needs to be submitted again.

7.3 How to make a data set public or add the corresponding published reference

By default, a data set will be made publicly available after the related manuscript has been accepted, or when PRIDE staff is notified to do so by the original submitter. There are two ways to do it: (i) contacting the PRIDE team by e-mail, or (ii) using the PRIDE Archive website (http://www.ebi.ac.uk/pride/archive). To use the web option, the user will need to be logged in and click on the 'Publish' button located next to each unpublished data set.

The corresponding reference associated with a given data set can also be provided in both ways. It is encouraged that the final version of the reference is always provided. This could potentially be available quite some time after the actual acceptance of the manuscript.

8 Future perspectives

In this manuscript, we have explained in detail the submission process of MS/MS data sets to PX via PRIDE. It is important to highlight that at present, experimental approaches other than MS/MS and SRM (which should be submitted to PX via PASSEL) can also be submitted to PX via PRIDE, using the 'partial' submission mechanism. Different 'experiment types' can be selected in the PX submission tool (step 2 described above, Fig. 2, panel 2). So, PRIDE can also store data sets from other approaches, as top-down or data-independent acquisition (e.g. SWATH-MS) experiments.

The current overall procedure explained is only expected to undergo minor modifications in the medium term. However, some of the details may change with regard to metadata annotation or the structure of the 'PX summary file' format. It is also planned that additional file tags will be added in the future to accommodate new data types. Updated documentation is always available at the PRIDE and ProteomeXchage websites (e.g. at http://www.proteomexchange.org/submission). A frequently asked questions section (including 'Troubleshooting') is available at http://www.ebi.ac.uk/pride/help/archive/faq.

It is also planned that in the near future, the new PSI (Proteomics Standards Initiative) standard format mzTab (https://code.google.com/p/mztab/, containing both identification and quantification results) will also be supported as a 'RESULT' file for performing 'complete' submissions. Finally, we encourage the proteomics community to take advantage of this infrastructure and tools, and to get familiarised with the process. It is expected that new resources will join PX so new ways of submitting MS/MS data to PX will become available. At the moment of writing the MassIVE repository (University of California, San Diego) has just formally joined PX, although its role in the overall PX data workflow has not yet been fully clarified. However, MassIVE is already taking submissions of MS/MS data sets.

The MS proteomics data in this paper have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [2]: data set identifier PXD000764.

The authors want to acknowledge funding from the EU FP7 grant 'ProteomeXchange' (grant number 260558). J.A.V. also wants to acknowledge the EU FP7 grant 'PRIME-XS' (grant number 262067) and Wellcome Trust (grant number WT101477MA). H.H. wants to acknowledge funding from BBSRC (BB/I00095X/1). A.R.J. would also like to acknowledge funding from BBSRC (BB/K01997X/1, BB/I00095X/1). G.G-B. was in receipt of a Marie Curie Intra-European Fellowship for Career Development (FP7-PEOPLE-IEF-2008). We also want to acknowledge Prof. Enitan Carrol for providing the samples, and Dr. Philip Brownridge and Dr. Duncan Robertson for instrument support.

The authors have declared no conflict of interest.

9 References

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

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

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

[4] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A. et al., ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32*, 223–226.

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

[6] Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O. et al., The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* 2012, *11*, M111 014381.

[7] Cote, R. G., Griss, J., Dianes, J. A., Wang, R. et al., The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol. Cell. Proteomics* 2012, *11*, 1682–1689.

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

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

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

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

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

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

[14] Cote, R., Reisinger, F., Martens, L., Barsnes, H. et al., The Ontology Lookup Service: bigger and better. *Nucleic Acids Res.* 2010, *38*, W155–W160.

[15] Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 2008, *24*, 2534–2536.

[16] Ma, Z. Q., Dasari, S., Chambers, M. C., Litton, M. D. et al., IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* 2009, *8*, 3872–3881.

[17] Tabb, D. L., Fernando, C. G., Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* 2007, *6*, 654–661.

[18] Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A. et al., OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinformatics* 2008, *9*, 163.

[19] Prieto, G., Aloria, K., Osinalde, N., Fullaondo, A. et al., PAnalyzer: a software tool for protein inference in shotgun proteomics. *BMC Bioinformatics* 2012, *13*, 288.

[20] Dasari, S., Chambers, M. C., Martinez, M. A., Carpenter, K. L. et al., Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J. Proteome Res.* 2012, *11*, 1686–1695.

[21] Dasari, S., Chambers, M. C., Slebos, R. J., Zimmerman, L. J. et al., TagRecon: high-throughput mutation identification through sequence tagging. *J. Proteome Res.* 2010, *9*, 1716–1726.

[22] Gluck, F., Hoogland, C., Antinori, P., Robin, X. et al., EasyProt—an easy-to-use graphical platform for proteomics data analysis. *J. Proteomics* 2013, *79*, 146–160.

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

[24] Medina-Aunon, J. A., Martinez-Bartolome, S., Lopez-Garcia, M. A., Salazar, E. et al., The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards. *Mol. Cell. Proteomics* 2011, *10*, M111.008334.