HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2014 September 10.

Published in final edited form as:

Nat Biotechnol. 2014 March ; 32(3): 223–226. doi:10.1038/nbt.2839.

ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination

Juan A. Vizcaíno[1,17], **Eric W. Deutsch**[2,17], **Rui Wang**[1], **Attila Csordas**[1], **Florian Reisinger**[1], **Daniel Ríos**[1], **José A. Dianes**[1], **Zhi Sun**[2], **Terry Farrah**[2], **Nuno Bandeira**[3], **Pierre-Alain Binz**[4], **Ioannis Xenarios**[4,5,6], **Martin Eisenacher**[7], **Gerhard Mayer**[7], **Laurent Gatto**[8], **Alex Campos**[9], **Robert J. Chalkley**[10], **Hans-Joachim Kraus**[11], **Juan Pablo Albar**[12], **Salvador Martinez-Bartolomé**[12], **Rolf Apweiler**[1], **Gilbert S. Omenn**[2,13], **Lennart Martens**[14,15], **Andrew R. Jones**[16], and **Henning Hermjakob**[1]

Juan A. Vizcaíno: juan@ebi.ac.uk

[1]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK [2]Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA [3]Center for Computational Mass Spectrometry, University of California, San Diego, La Jolla, CA, USA [4]Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1211 Geneva 4, Switzerland [5]University of Lausanne, Lausanne, Switzerland and Center for Integrative Genomics, University of Lausanne, 1005 Lausanne, Switzerland [6]Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland [7]Medizinisches Proteom-Center, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, Germany [8]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QR, United Kingdom [9]Integromics SL, Santiago Grisolia, 2, Tres Cantos, 28760, Madrid, Spain [10]Department of Pharmaceutical Chemistry, University of California San Francisco, CA 94158, USA [11]Wiley-VCH Verlag, Boschstraße 12, 69469 Weinheim, Germany [12]ProteoRed-ISCIII, National National Center for Biotechnology-CSIC, Madrid, Spain [13]Center for Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, 48109-2218, USA [14]Department of Medical Protein Research, VIB, A. Baertsoenkaai 3 B-9000 Ghent, Belgium [15]Department of Biochemistry, Ghent University, A. Baertsenkaai 3 B-9000 Ghent, Belgium [16]Institute of Integrative Biology, University of Liverpool, UK, L697ZB

[17]Both authors contributed equally to the manuscript.

Competing financial interests

The authors have no competing financial or commercial interests.

Author contributions

JAV, HH, and EWD led the current implementation of the ProteomeXchange data workflow, guidelines and related software. RW developed the 'ProteomeXchange submission tool'. Further authors contributed to the development of the ProteomeXchange consortium in different ways, e.g. contributing to the initial ProteomeXchange prototypes in the past, developing software and data standards, or contributing in different aspects to the implementation of the guidelines and the data workflow. JAV, EWD and HH wrote the manuscript.

All authors have agreed to all the content in the manuscript, including the data as presented.

To the Editor

There is a growing trend towards public dissemination of proteomics data, which is facilitating the assessment, reuse, comparative analyses and extraction of new findings from published data[1, 2]. This process has been mainly driven by journal publication guidelines and funding agencies. However, there is a need for better integration of public repositories and coordinated sharing of all the pieces of information needed to represent a full mass spectrometry (MS)–based proteomics experiment. Your July 2009 editorial "Credit where credit is overdue"[3] exposed the situation in the proteomics field, where full data disclosure is still not common practise. Olsen and Mann[4] identified different levels of information in the typical experiment, starting from raw data and going through peptide identification and quantification, protein identifications and ratios and the resulting biological conclusions. All of these levels should be captured and properly annotated in public databases, using the existing MS proteomics repositories for the MS data (raw data, identification and quantification results) and metadata, whereas the resulting biological information should be integrated in protein knowledgebases, such as UniProt[5]. A recent editorial in *Nature Methods*[6] again highlighted the need for a stable repository for raw MS proteomics data. In this Correspondence, we report on the first implementation of the ProteomeXchange consortium, an integrated framework for submission and dissemination of MS-based proteomics data.

Among the existing MS proteomics repositories with a broad target audience, the PRIDE (PRoteomics IDEntifications) database[7] (European Bioinformatics Institute, EBI, Cambridge, UK; http://www.ebi.ac.uk/pride) and PeptideAtlas[8] (Institute for Systems Biology, ISB, Seattle, USA; http://www.peptideatlas.org) are two of the most prominent. Both are mainly focused on tandem MS (MS/MS) data storage. Whereas PRIDE represents the information as originally analysed by the researcher (thus constituting a primary resource), data in PeptideAtlas are reprocessed through a common pipeline (the Trans-Proteomic Pipeline) to provide a uniformly analyzed view on the data with a focus on low protein false discovery rates (constituting a secondary resource). In addition, ISB has set up the first repository for SRM data, PASSEL[9] (PeptideAtlas SRM Experiment Library, http://www.peptideatlas.org/passel/). There are other resources dedicated to storing MS proteomics data, each of them with different focuses and functionalities, for instance GPMDB (where data are reprocessed using the search engine X!Tandem)[10]. At a higher abstraction level, resources like UniProt and neXtProt are integrating proteomics results into a wider context of functional annotation from many different sources, including antibody-based methods.

Although most of the proteomics resources mentioned have existed for a long time, they have acted independently with limited coordination of their activities. As a result, data providers were unclear to which repository they should submit their dataset, and in what form, with choices ranging from full raw data to highly processed identifications and quantifications. In addition, no repository could store both raw data and results. Similar issues arose for data consumers, who could not always find the data supporting a protein modification in UniProt, or know whether a particular dataset from PRIDE had been integrated into PeptideAtlas.

The ProteomeXchange (PX) consortium (http://www.proteomexchange.org) was formed in 2006 (ref. 11) to overcome these challenges, developing from a loose collaboration into an international consortium of major stakeholders in the domain, comprising, among others, primary (PRIDE, PASSEL) and secondary resources (PeptideAtlas, UniProt), proteomics bioinformaticians, investigators (including some involved in the HUPO Human Proteome Project), and representatives from journals regularly publishing proteomics data (Supplementary Notes, section 7). The aim of the ProteomeXchange consortium is to provide a common framework and infrastructure for the cooperation of proteomics resources by defining and implementing consistent, harmonised, user-friendly data deposition and exchange procedures among the major public proteomics repositories.

ProteomeXchange provides unified data submission for multiple MS data types and delivers different 'views' of the deposited data, such as the raw data suitable for reprocessing, the author-generated identifications and highly filtered composite results in resources like UniProt, all linked by a universal shared identifier. Authors are able to cite the resulting ProteomeXchange accession number for datasets reported in their publications. As such, a dataset (with appropriate metadata) is becoming publishable *per se* and can be tracked if used by various consumers in different publications.

Individual resources can join ProteomeXchange by implementing the ProteomeXchange data submission and dissemination guidelines, and metadata requirements. In the current version (http://www.proteomexchange.org/concept), the mandatory information comprises: (i) mass spectrometer output files (raw data, either in a binary format, or in a standard open format such as mzML); (ii) processed identification results (two submission modes are available, see below); and (iii) sufficient metadata to provide a suitable biological and technological background, including method information such as transition lists in the case of SRM data. Other types of information, such as peak list files (processed versions of mass spectra most often used in the identification process) and quantification results can also be provided.

Two main MS proteomics workflows are now fully supported: tandem MS and SRM data (Fig. 1 and Supplementary Fig. 1). PRIDE acts as the initial submission point for MS/MS data, whereas PASSEL is the initial submission point for SRM data. It is expected that in most cases, one ProteomeXchange dataset will correspond to data from one publication, and it will be clearly linked to it. However, this concept is flexible and a mechanism for grouping different ProteomeXchange datasets is also available, for example for large-scale collaborative studies. At present, two different submission modes are available for MS/MS data:

1. - 'Complete submission': this requires peptide and protein identification results to be fully supported and integrated in the receiving repository (PRIDE at present). The search engine output files (plus the associated spectra) must therefore first be converted to PRIDE XML or mzIdentML format (a process supported by several popular and user-friendly tools, Supplementary Notes, section 5). Complete submissions make the data fully available for querying, and thus maximise the potential for data re-use in MS. This in turn increases the visibility of the associated

publication. A DOI (Digital Object Identifier) is assigned to each dataset, allowing formalized credit to be given to submitters and their principal investigators, through a citation index, as proposed in your editorial[3].

2. - 'Partial submission': For these submissions, peptide or protein identification results cannot be integrated in PRIDE because data converters and exporters to the supported formats are not yet available. In this case, search engine output files can be directly provided in their original format. Although partial submissions are searchable by their metadata, they are not fully searchable by results such as protein identifiers, and will not receive a DOI. However, partial submissions are important as they allow data from novel experimental approaches to be deposited into the ProteomeXchange resources, rather than having to reject these until the workflows have been mapped into a representation in PRIDE or another ProteomeXchange partner.

For the submission of MS/MS datasets, a stand-alone, open-source Java tool has been made available, the 'ProteomeXchange submission tool' (http://www.proteomexchange.org/submission) (Supplementary Notes section 5, Supplementary Figs. 2–10). The tool allows interactive submission of small datasets as well as large- scale batch submissions.

For SRM datasets, a web form (http://www.peptideatlas.org/submit) can be used for submission to PASSEL. Similar to the guidelines stated above for MS/MS datasets, PASSEL submissions require mass spectrometer output files, study metadata, peptide reagents, analysis result files and the actual SRM transition lists, the information that drives the instrument data acquisition. Once datasets are submitted, they are checked by a curator and then loaded into the main PASSEL database, which facilitates interactive exploration of the data and results.

The submitted information and files can selectively be made available to journal editors and reviewers during manuscript peer review. Once the manuscript is accepted for publication or the submitter informs the receiving repository directly, the data will be publicly released (Fig. 1). At this point, the availability of the dataset, as well as basic metadata, will be disseminated through a public RSS feed (http://groups.google.com/group/proteomexchange/feed/rss_v2_0_msgs.xml). The RSS feed includes a link to an XML message (ProteomeXchange XML), which is created by the receiving repository (Supplementary Notes, section 3), and made available from ProteomeCentral, the portal for all public ProteomeXchange datasets (http://proteomecentral.proteomexchange.org) (Supplementary Notes, section 2). Repositories such as PeptideAtlas or GPMDB as well as any interested end users can subscribe to this RSS feed and trigger actions, including incorporation of the data into local resources, re-processing or biological analysis. This reprocessing is already occurring in practice. For example, two ProteomeXchange datasets (PXD000134 and PXD000157) have been used in the latest build of the human proteome in PeptideAtlas, and PXD000013 (ref. 12) was reprocessed and nominated as technical dataset of the year 2012 by GPMDB (http://www.thegpm.org/dsotw_2012.html - 201210071).

ProteomeXchange started to accept regular submissions in June 2012. By the beginning of August 2013, 373 ProteomeXchange datasets have been submitted (consisting of 341

tandem MS and 32 SRM datasets, Fig. 2), a total of ~25 TB of data. The largest submission so far (currently still private) comprised 5 TB of data. For a current list of the publicly available datasets, see http://proteomecentral.proteomexchange.org/.

In summary, ProteomeXchange provides an infrastructure for efficient and reliable public dissemination of proteomics data, supporting crucial validation, analysis and reuse. By providing and linking different interpretations of the data we aim to maximise dataset visibility as well as their potential benefit to different communities. Citability and traceability are addressed through the assignment of DOIs and a common identifier space. The consortium is open to the participation of additional resources (Supplementary Notes, Section 9). Although all repositories depend on continuous funding for continuous operation, the ProteomeXchange core repositories PRIDE and PeptideAtlas are well established, with first publications in 2005 (ref. 7,8), and have strong institutional backing (Supplementary Notes, section 8), ensuring that the data will remain reliably available for the foreseeable future. We are confident that the ProteomeXchange infrastructure will support the growing trend towards public availability of proteomics data, maximising its benefit to the scientific community through increased ease of access, greater ability to re-assess interpretations and extract further biological insight, and greater citation rates for the submitters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all the members of the community who participated as stakeholders in the ProteomeXchange meetings. This work was supported by the EU FP7 grant ProteomeXchange [grant number 260558]. JAV, AC, FR and DR were also funded by the Wellcome Trust [grant number WT085949MA]. EWD, ZS and TF are also funded in part by NIH/NIGMS grant No. R01 GM087221, NSF MRI [grant number 0923536], and the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg. ME is funded by P.U.R.E. (http://www.pure.rub.de, Protein Unit for Research in Europe), a project of Nordrhein-Westfalen (Germany). LG was supported by the EU FP7 PRIME-XS project [grant number 262067]. RW was supported by the BBSRC ‘PRIDE Converter’ grant [reference BB/I024204/1].

Abbreviations

DOI	Digital Object Identifier
EBI	European Bioinformatics Institute
ISB	Institute for Systems Biology
PASSEL	PeptideAtlaS SRM Experiment Library
PRIDE	PRoteomics IDEntifications
PX	ProteomeXchange
RSS	Rich Site Summary
SRM	Selected Reaction Monitoring

References

1. Hahne H, Kuster B. Mol Cell Proteomics. 2012; 11:1063–1069. [PubMed: 22826440]
2. Matic I, Ahel I, Hay RT. Nat Methods. 2012; 9:771–772. [PubMed: 22847107]
3. Nat Biotechnol. 2009; 27:579. Editors. [PubMed: 19587644]
4. Olsen JV, Mann M. Sci Signal. 2011; 4:pe7. [PubMed: 21325203]
5. The UniProt Consortium. Nucleic Acids Res. 2012; 40:D71–75. [PubMed: 22102590]
6. Nat Methods. 2012; 9:419. Editors. [PubMed: 22803195]
7. Martens L, et al. Proteomics. 2005; 5:3537–3545. [PubMed: 16041671]
8. Deutsch EW, et al. Proteomics. 2005; 5:3497–3500. [PubMed: 16052627]
9. Farrah T, et al. Proteomics. 2012; 12:1170–1175. [PubMed: 22318887]
10. Craig R, Cortens JP, Beavis RC. J Proteome Res. 2004; 3:1234–1242. [PubMed: 15595733]
11. Hermjakob H, Apweiler R. Expert Rev Proteomics. 2006; 3:1–3. [PubMed: 16445344]
12. Vaudel M, et al. J Proteome Res. 2012; 11:5072–5080. [PubMed: 22874012]

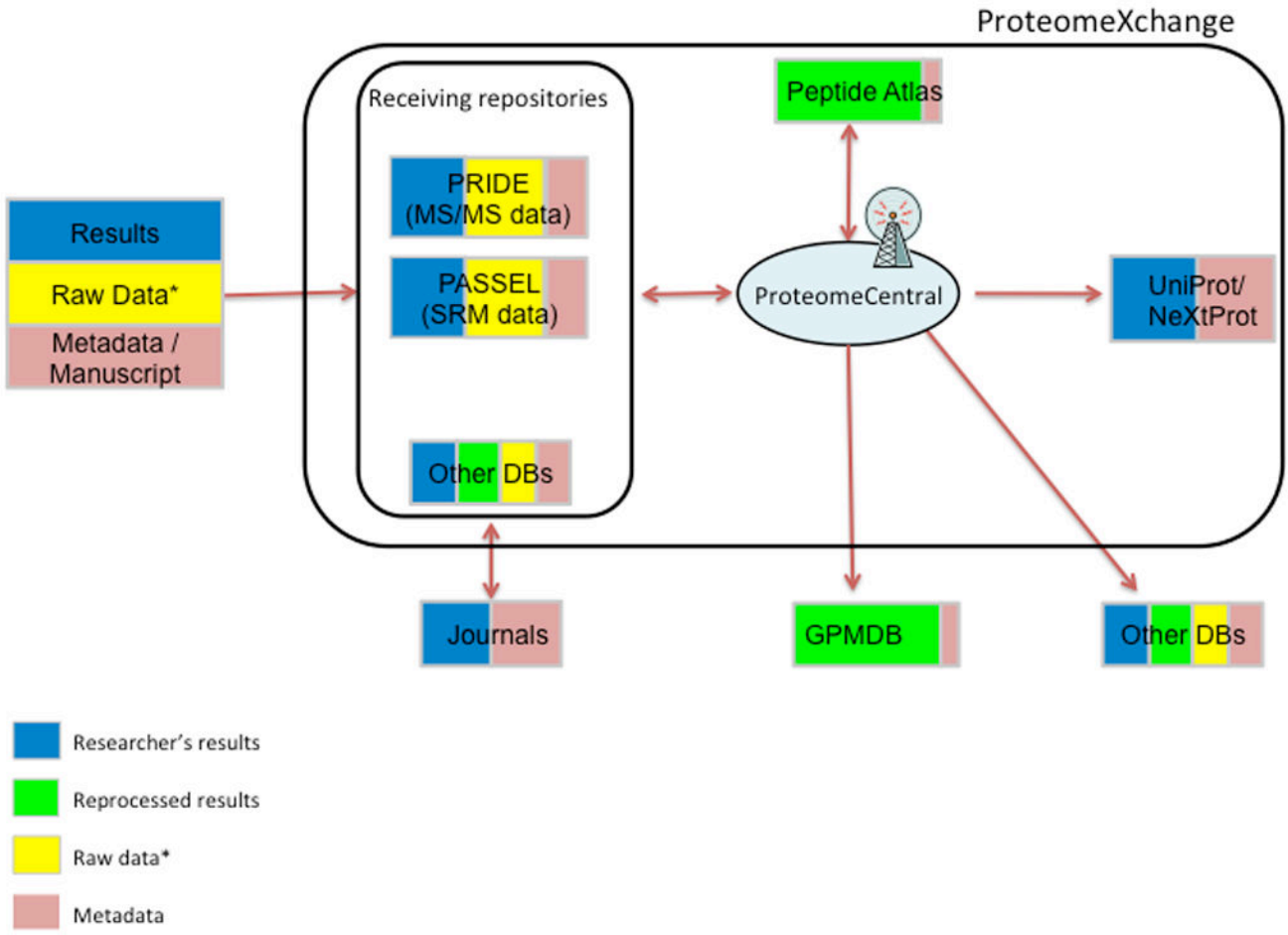

Figure 1.

Representation of the ProteomeXchange workflow for MS/MS and SRM data. *Raw data represents mass spectrometer output files.

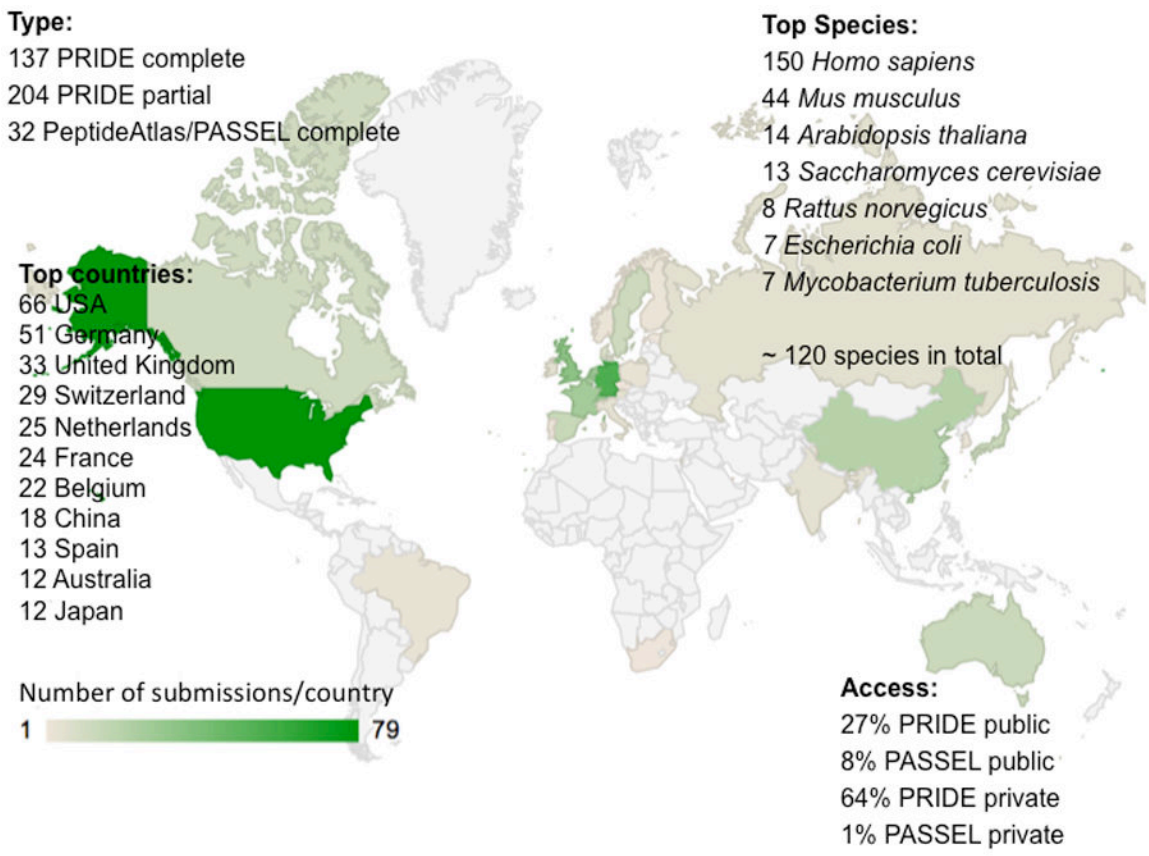

Figure 2.

Summary of the main metrics of ProteomeXchange submissions (by August 2013). The number of data sets is indicated for submission type, data access status and for the top species and countries represented.