# *The Proteomics Identifications database: 2010 update*

*D736–D742* *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
*doi:10.1093/nar/gkp964*

# The Proteomics Identifications database: 2010 update

**Juan Antonio Vizcaíno[1], Richard Côté[1], Florian Reisinger[1], Harald Barsnes[2], Joseph M. Foster[1], Jonathan Rameseder[1,3], Henning Hermjakob[1] and Lennart Martens[1,*]**

[1]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]Department of Informatics, University of Bergen, Norway and [3]Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received September 10, 2009; Revised October 6, 2009; Accepted October 13, 2009

## ABSTRACT

**The Proteomics Identifications database (PRIDE, http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute has become one of the main repositories of mass spectrometry-derived proteomics data. For the last 2 years, PRIDE data holdings have grown substantially, comprising 60 different species, more than 2.5 million protein identifications, 11.5 million peptides and over 50 million spectra by September 2009. We here describe several new and improved features in PRIDE, including the revised submission process, which now includes direct submission of fragment ion annotations. Correspondingly, it is now possible to visualize spectrum fragmentation annotations on tandem mass spectra, a key feature for compliance with journal data submission requirements. We also describe recent developments in the PRIDE BioMart interface, which now allows integrative queries that can join PRIDE data to a growing number of biological resources such as Reactome, Ensembl, InterPro and UniProt. This ability to perform extremely powerful across-domain queries will certainly be a cornerstone of future bioinformatics analyses. Finally, we highlight the importance of data sharing in the proteomics field, and the corresponding integration of PRIDE with other databases in the ProteomExchange consortium.**

## INTRODUCTION

Mass spectrometry (MS) is currently the most commonly used technology for the identification and quantification of proteins. Like in any other 'omics' field, the amount of data generated by MS-based proteomics has increased exponentially in the last few years, which has prompted the development of several data repositories. The Proteomics Identifications database (PRIDE) (http://www.ebi.ac.uk/pride) was developed at the European Bioinformatics Institute (EBI), as a repository for the results of MS-based proteomics experiments, allowing data from a vast range of approaches, instruments and analysis platforms to be stored and disseminated in a common structured and queryable format. Originally established as a production service in 2005, PRIDE has previously been described (1–3) along with guidelines on using the database and its associated tools (4–6).

PRIDE does not stand alone in this field, however, as several other proteomics databases have been established over the past few years. GPMDB (7), PeptideAtlas (8) and Proteinpedia (9) are among the most important representatives of these (10). Additionally, the Tranche system (http://tranche.proteomecommons.org) provides a data transfer layer relying on peer-to-peer Internet protocol technology. Finally, the most recently launched proteomics repository is the NCBI Peptidome (11), a centralized, public proteomics data repository not dissimilar from PRIDE in its objectives. For an up-to-date review covering the capabilities of a comprehensive selection of proteomics MS repositories see Mead *et al.* (12).

PRIDE stores three different kinds of information: MS and MS/MS mass spectra as peak lists, the derived peptide and protein identifications (IDs) and any associated metadata. Indeed, one of the advantages PRIDE offers over other proteomics databases lies in the amount of structured metadata it contains, which is a key requirement to put the stored data in biological and/or technical context. Furthermore, together with the newly released NCBI Peptidome, the established PRIDE database constitutes an actual structured data repository, and does not assume any editorial control over submitted data.

Another important feature of PRIDE is that it allows data to remain private while anonymously sharing it with journal editors and reviewers through so-called 'reviewer log-in accounts'. As a result, PRIDE is now the recommended submission point for proteomics data for several journals such as *Nature Biotechnology* (13), *Nature*

*To whom correspondence should be addressed. Tel: +44 1223 492 610; Fax: +44 1223 494 484; Email: lennart.martens@ebi.ac.uk

*Methods* (14) and *Proteomics* (http://www3.interscience.wiley.com/cgi-bin/jabout/76510741/2120_instruc.pdf).

Two highly influential informatics tools have been developed in support of the PRIDE database over the years: the Ontology Lookup Service (OLS) (15) and the Protein Identifier Cross-Referencing system (PICR) (16). PICR is used to map all submitted protein identifications in PRIDE to all known accession numbers for these proteins in the most important protein databases, as well as some genomic ones (16). PICR mappings are performed on the entire PRIDE database at regular intervals in order to keep all accession numbers up-to-date. The mappings allow accurate experiment-to-experiment comparison, even if the experiments relied on different sequence databases for the identifications, and since the mappings include an historical archive of identifiers, they also readily translate now defunctional accession numbers through time. In addition to these two established tools, a new application called Database on Demand (DoD, http://www.ebi.ac.uk/pride/dod) has recently been added to the PRIDE toolkit (17). This tool allows custom sequence databases to be built in order to optimize the results from search engines for gel-free proteomics experiments (18). DoD allows users to process and combine the UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and IPI databases using *in silico* protein maturation filters, serial enzymatic digests to reflect both *in vivo* and *in vitro* cleavages, amino- or carboxy-terminal ragging of sequences and mass and composition-based output filters, amongst others.

In this article, we will, however, focus on the improvements made to the PRIDE system over the last 2 years, and we will also highlight recently submitted datasets of interest.

## PRIDE DATABASE CONTENT

Data content in PRIDE has increased substantially since the last PRIDE NAR database issue was published (3). By 1 September 2009, PRIDE contained 9908 experiments (compared with only 3185 when the last NAR manuscript was submitted on September 2007), more than 2.5 million identified proteins (in contrast to 330 000 on September 2007; a 7.5-fold increase), 11.5 million identified peptides (versus 2.1 million on September 2007; a 5.5-fold increase) and 50.3 million spectra (2.6 million on September 2007; a 19-fold increase). This dramatic growth of data content in PRIDE is visually represented in Figure 1. The increase in the total number of peptide IDs (5.5-fold) is reflected in the growth of the number of unique peptide sequences in the database, which have seen a 6.1-fold increase. This latter observation is particularly interesting, as the identification of unique sequences typically levels off when applying repeat analyses in proteomics (18,19). It seems that it is primarily the diversity of data flowing into PRIDE (in terms of sample, experimental technique, instrument and search engine) that allows the number of unique sequences to grow in lockstep with the total number of submitted peptide IDs (20,21).

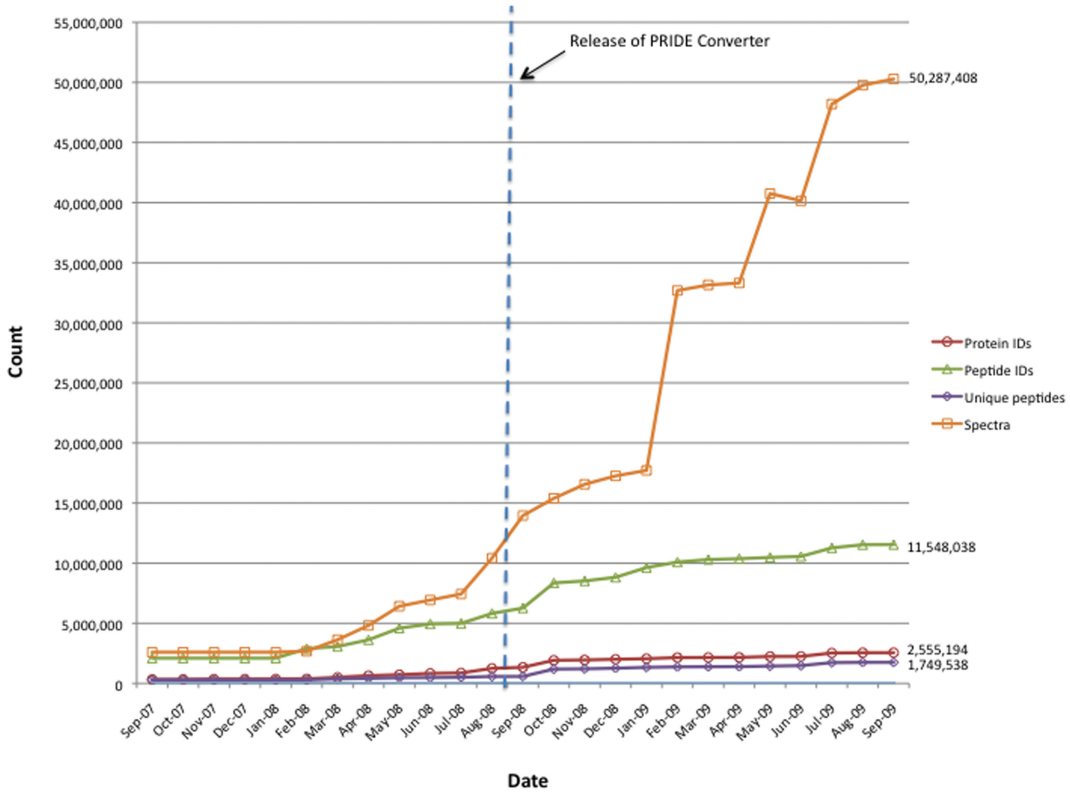


**Figure 1.** Growth of data content in PRIDE from September 2007 until September 2009. The data included in the graph are number of spectra, protein identifications, peptide identifications and unique peptides.

**Table 1.** Data content in PRIDE split by taxonomic divisions

| | Protein IDs (%) | Peptide IDs (%) |
|---|---|---|
| Group of organisms (number of species) | | |
| Animals (17) | 84.4 | 74.3 |
| Plants (8) | 7.6 | 11.5 |
| Fungi (8) | 3.9 | 2.2 |
| Bacteria (20) | 2.6 | 11.6 |
| Others (7) | 1.5 | 0.4 |
| Species | | |
| *Homo sapiens* | 38.1 | 35.7 |
| *D. melanogaster* | 16.2 | 12.6 |
| *Mus musculus* | 14.8 | 9.0 |
| *A. thaliana* | 6.7 | 11.0 |
| *Danio rerio* | 6.3 | 4.0 |
| *Rattus norvegicus* | 3.6 | 3.1 |
| *C. elegans* | 3.3 | 8.9 |
| *Aspergillus niger* | 1.5 | 0.3 |
| *S. typhimurium* | 1.3 | 7.3 |
| *Chlamydomonas reinhardtii* | 1.3 | 0 |

Only the top 10 species in terms of protein and peptide identifications are shown.

Note that these data holdings are absolute figures, not distinguishing public and private data. By 1 September 2009, 8570 experiments in PRIDE were publicly available (86.5% of the total number of experiments). The rest of the data are private since, as mentioned previously, PRIDE allows data to remain confidential during manuscript review.

The complete set of data in PRIDE comprises 60 species, including several model organisms (Table 1). Animal species provide the majority of these data, since a total of 17 animal species are represented, contributing 84.4 and 74.3% of all protein and peptide IDs in PRIDE, respectively. Perhaps unsurprisingly, the largest amount of protein and peptide IDs comes from human samples (38.1 and 35.7% of all protein and peptide IDs in PRIDE, respectively). *Drosophila melanogaster* (16.2 and 12.6%) and mouse (14.8 and 9.0%) are the next most popular data sources.

Bacteria are the most often represented group of organisms with 20 different species, among which *Salmonella typhimurium* provides the largest number of protein and peptide IDs (1.3 and 7.3% of all protein and peptide IDs in PRIDE, respectively). Additionally, eight different plant and fungal species are represented, thus constituting the second and third most popular group of organisms in PRIDE (Table 1).

Apart from the overall association of data in experiments, PRIDE also has the concept of projects, which provide a way to organize several related experiments together in a hierarchical structure. Some of the most relevant datasets that have recently been made publicly available in PRIDE are organized under such a project structure. First of all, datasets belonging to the second phase of the HUPO Plasma Proteome Project (PPP2) (22) are now present in PRIDE. The most complete and most thoroughly annotated dataset so far within the PPP2 project comes from Richard Smith's lab at PNNL (PRIDE accession numbers 8172–8544, both inclusive) (23,24). This set of experiments also constituted the first test case for a full ProteomExchange (25) submission, which is explained in more detail in the last section of this article.

As already shown, PRIDE is increasingly receiving submissions from species other than human and mouse. In this context, PRIDE now stores two large datasets (accession number 9776; and accession numbers 9777–9794, both inclusive) that have been used to improve the annotation of the *Caenorhabditis elegans* genome (26,27). Interestingly, experiment 9776 is also the largest single experiment in PRIDE, comprising >85 GB of uncompressed information. Currently, the largest set of experiments in PRIDE that belongs to the same project (accession numbers 3866–7955, both inclusive) constitutes a quantitative analysis of the secretory pathway in rat (28). Another very interesting dataset (accession numbers 3321–3354, both inclusive) (29) describes a high-density, organ-specific proteome catalog generated from different organs, developmental stages and undifferentiated cultured cells from *Arabidopsis thaliana*.

Perhaps most strikingly, PRIDE now also stores proteomics data from several extinct animals, including one dataset from *Tyrannosaurus rex* (accession number 8633) (30), which is one of the most widely discussed proteomics datasets published to date (see http://pubs.acs.org/action/showStoryContent?doi=10.1021%2Fon.2008.11.21.172568).

## IMPROVEMENTS IN THE PRIDE WEB INTERFACE: SPECTRUM FRAGMENTATION ANNOTATION

The main improvement in the PRIDE web interface is the ability to store and display fragment ion annotations on tandem mass spectra. As mentioned before, and described in detail in the next section, it has recently become possible for PRIDE users to submit files containing fragment ion annotation directly, and subsequently visualize these in the online 'PRIDE Spectrum Viewer' (Figure 2). This feature has important implications for journal requirements relating to publication-associated proteomics data.

At present, each journal essentially develops custom guidelines for data submission, which differ in scope and stringency. The journal *Molecular and Cellular Proteomics* (MCP), an early adopter in terms of guidelines, has developed the so-called Paris guidelines for reporting proteomics data (31) that include the requirement to provide annotated fragmentation spectra in several defined cases. The fact that PRIDE can now handle the submission and visualization of this type of spectral annotation ensures that submitters can achieve MCP compliance with ease.

Another important development in the web interface is the seamless integration of PICR mappings into PRIDE queries, the Venn diagram comparison tool and the PRIDE BioMart interface. As a result, PRIDE has a 'memory' of all identifiers ever used for a given protein in the majority of proteomics databases, enabling users to query by whichever accession number or identifier is most convenient for them. Furthermore, the Venn

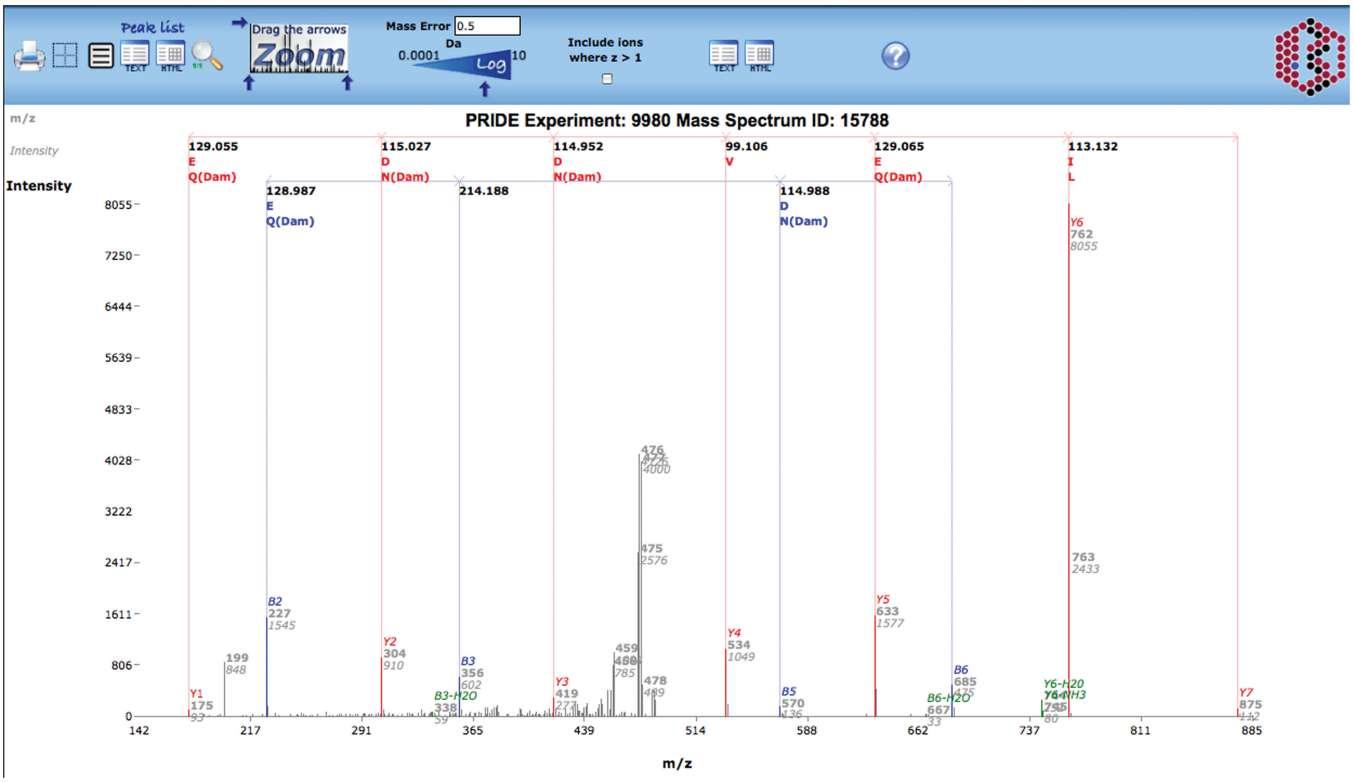


**Figure 2.** Screenshot showing spectrum fragmentation annotation on tandem mass spectra as visualized in the 'PRIDE Spectrum Viewer'. Y ions are highlighted in red, B ions in blue and neutral loss ions in green.

diagram-based experiment comparison tool provides the most comprehensive tool available to the community to date to compare two protein lists. Finally, the PRIDE 'Identification Detail View' page has been thoroughly revised to reflect the PICR mappings, along with a new view on the protein sequence that allows identified peptides and even post-translational modifications to be highlighted.

## THE SUBMISSION PROCESS MADE EASIER: PRIDE CONVERTER

It is important to highlight here that PRIDE is entirely dependent on data submissions by researchers, as detailed proteomics data cannot be curated from existing literature. The development of a new submission tool called PRIDE Converter (32) (http://pride-converter.googlecode.com) has been key in the large growth of data content in PRIDE during the last year (Figure 1). PRIDE Converter has made submission to PRIDE a simple and efficient process, since a submitter can now convert a wide variety of the most common proteomics data formats (Figure 3) directly to PRIDE XML in eight easy steps in a user-friendly, wizard-like graphical user interface. The resulting PRIDE XML file is then automatically schema-verified, and can subsequently be submitted directly to the PRIDE database. One of the most recent features added to PRIDE Converter (available from version 2.0 onwards, released in July 2009) is the handling of spectrum fragmentation annotation from Mascot .dat files, OMSSA .omx files and ms_lims (33). As a result, it is now trivial for users of these formats to submit and visualize spectrum annotations in PRIDE, which also results in immediate compliance with the requirements of journals such as MCP. It is important to note here that compliance in the past essentially comprised submitting hundreds or even thousands of PDF renderings of spectra to the journal, causing several logistical problems, as well as constituting a downright data loss, since spectra in this form are of course no longer machine-readable.

Another very useful improvement in the PRIDE submission process is the possibility to upload (typically quite large) PRIDE XML files to an EBI FTP server (6). Finally, it is important to highlight that there are no longer any file size limitations for submissions to PRIDE; indeed, as mentioned above, the largest single submission to PRIDE currently stands at a PRIDE XML file of >85 GB.

## IMPROVEMENTS IN THE PRIDE BIOMART INTERFACE: ACROSS-RESOURCE QUERIES

PRIDE is of course also a very interesting tool for large-scale data mining. Currently, the easiest way to do this is by using the PRIDE BioMart interface (34). In the current BioMart interface (http://www.biomart.org/biomart/martview/), it is possible to retrieve data from PRIDE alone, but also to integrate information from PRIDE with other resources. By September 2009, PRIDE data

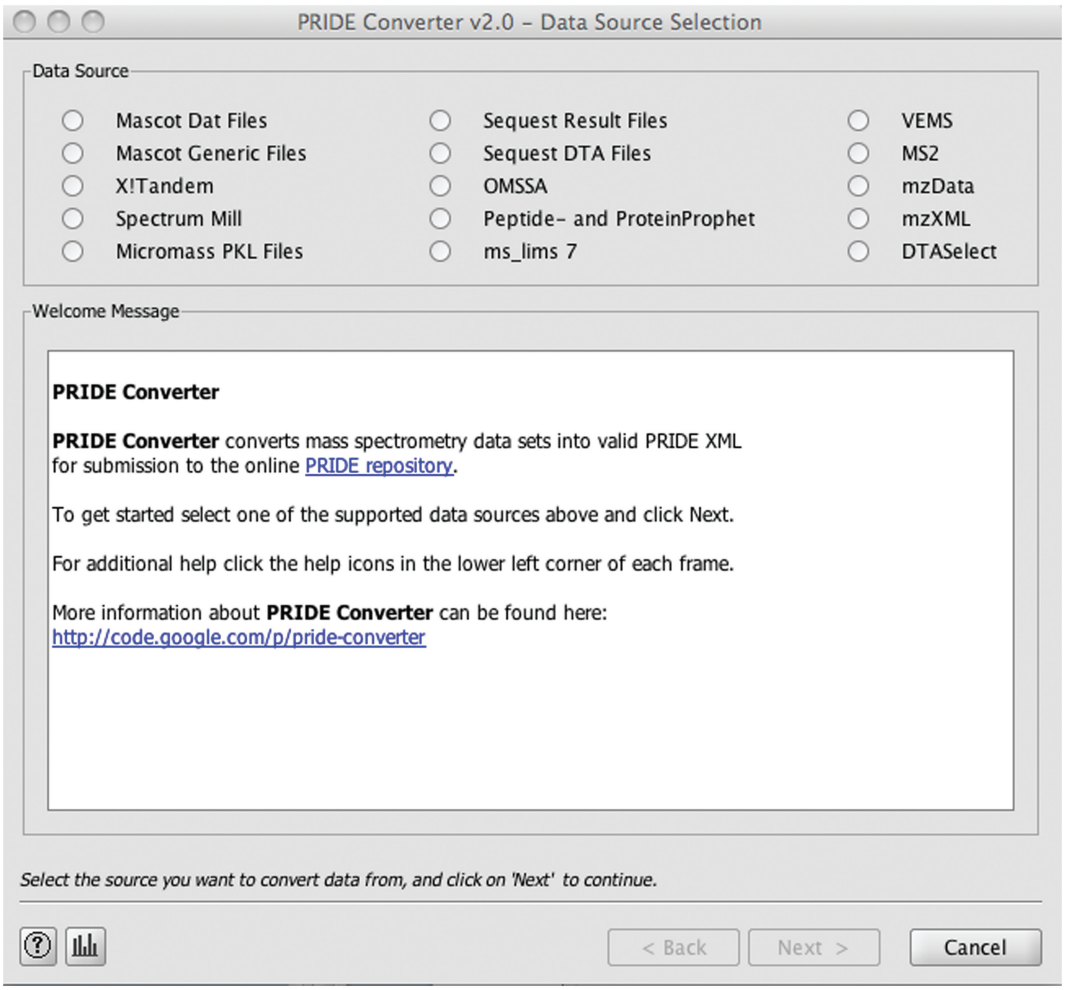


**Figure 3.** Screenshot showing the opening frame of PRIDE Converter. It allows the user to choose the appropriate format from the list of supported software-specific formats that can be converted into PRIDE XML.

could be combined with data resources such as Ensembl (35), UniProt (36), Reactome (37), InterPro (38), the Macromolecular Structure Database (MSD) (39), the IPI database (40) and the Rat Genome Database (RGD) (41), but the number of resources is continuously growing. As such, a complex query combining MS data from PRIDE with genomic, structural or pathway data has become straightforward and very fast. This integrative way to access bioinformatics data will undoubtedly become even more powerful and increasingly important in the life sciences as information obtained through multiple omics technologies is pieced together for a more complete picture of the underlying biology (42,43).

For even more sophisticated data mining purposes, users can also download the corresponding XML files for each experiment from the EBI FTP server (ftp://ftp.ebi.ac.uk/pub/databases/pride), where these files are now available as gzipped files rather than the original zipped files. The reason for this transition is that many PRIDE XML files had grown larger than the maximal allowed file size for generic zip files (which is 4 GB). A final important improvement for the FTP files is that they now also contain all the protein mapping information from PICR.

## INTEGRATING PRIDE WITH OTHER PROTEOMICS DATABASES

PRIDE is a founding partner of the ProteomExchange consortium, together with other important proteomics databases (25). The core members of this consortium (PRIDE, NCBI Peptidome, Tranche, PeptideAtlas and GPMDB) are currently working on the implementation of a system that will allow proteomics data sharing between all the members, with PRIDE and NCBI Peptidome as the initial ProteomExchange submission points. Draft guidelines for ProteomExchange submissions are available (http://www.proteomexchange.org), and a large-scale ProteomExchange pilot submission has already been performed (6). It must be noted here that it is not trivial to fulfil all the requirements for a ProteomExchange submission. This is the reason why, in addition to the ProteomExchange initiative, PRIDE and NCBI Peptidome have formally agreed that they will

replicate and share all their public data, again ensuring that data become optimally visible to the scientific community.

Finally, at a different level of integration, PRIDE data are now used in cross-references from UniProt (available from UniProt release 14.6, on December 2008). This allows PRIDE submitters to dramatically improve the exposure of their data, and provides a first point of entry in using PRIDE data to annotate UniProt protein entries.

## DISCUSSION

We have here described how the PRIDE database has evolved from its original role as a repository of proteomics identifications arising from MS data, to a knowledgebase that provides tools for complex queries and data retrieval, dataset comparison and access to additional automated annotation of submitted datasets. There has been a huge growth in data content, which can be traced for two key developments: the new PRIDE Converter submission tool has made data submission much easier and more straightforward, and the fact that various journals in the field are now strongly supporting and even mandating deposition of proteomics data in proteomics repositories in general, and PRIDE in particular.

Since the last PRIDE NAR database issue (3), there have been other significant developments in the PRIDE system. One of the most important improvements is that it is now possible to submit files containing fragment ion annotations on tandem mass spectra and visualize these annotations in the 'PRIDE Spectrum Viewer'. In the same context, we are currently developing a pipeline to derive this information in a generic, automatic way for all the experiments already present in PRIDE. At present, this automatic annotation feature is already visible for three experiments in PRIDE (accession numbers 1–3, both inclusive) and this form of automatic annotation will be extended to the rest of PRIDE in the near future. Such a pipeline will be especially useful as there are several output file formats from search engines or proteomics pipelines that do not explicitly contain such information, for instance SEQUEST and Trans-Proteomic Pipeline (TPP) output files.

The current lack of generally practiced data sharing in the proteomics field has recently been addressed by an editorial in *Nature Biotechnology* (44). In this context, the value of PRIDE as a data repository was recently clearly proven as the public availability of a dataset from *T. rex* (30), and enabled an ensuing healthy discussion in the proteomics community.

Initiatives such as ProteomExchange and the public data replication agreement between PRIDE and NCBI Peptidome are expected to help overcome the community's reticence about data disclosure. However, it must also be taken into account that, mainly due to the existence of different data formats and the inherent complexity of the data to be exchanged, this data-sharing process is very resource-intensive and time-consuming for the data repositories. In order to overcome this situation, PRIDE has always been supportive of community data standards. Therefore, in the next version of the PRIDE system (version 3), compliance with the HUPO Proteomics Standards Initiative's data standards mzML and mzIdentML (previously known as analysisXML) will be ensured (http://www.psidev.info).

Another ongoing key development in PRIDE that will benefit the proteomics community is the creation of a new database called PRIDE-Q (for 'Q-rated') that will contain only the highest quality data from the PRIDE repository. The relationship between the current PRIDE repository and the planned PRIDE-Q resource is very similar to the existing relationship between UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. Whereas UniProtKB/Swiss-Prot aims for assured quality and annotation, the UniProtKB/TrEMBL aims primarily at fully capturing all available data. This 5-year project, funded by the Wellcome Trust, started at the beginning of 2009 and will represent the first concerted effort to perform quality control analysis on proteomics data of such heterogeneity.

## ACKNOWLEDGEMENTS

The PRIDE team would like to thank all data submitters for their contributions. The authors would also like to thank Rolf Apweiler for his support.

## FUNDING

The ProDaC grant of the European Union (Grant LSHG-CT-2006-036814); Wellcome Trust (Grant WT085949MA). Funding for open access charge: Wellcome Trust.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
2. Jones,P., Cote,R.G., Martens,L., Quinn,A.F., Taylor,C.F., Derache,W., Hermjakob,H. and Apweiler,R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **34**, D659–D663.
3. Jones,P., Cote,R.G., Cho,S.Y., Klie,S., Martens,L., Quinn,A.F., Thorneycroft,D. and Hermjakob,H. (2008) PRIDE: new developments and new datasets. *Nucleic Acids Res.*, **36**, D878–D883.
4. Martens,L., Jones,P. and Cote,R. (2008) Using the Proteomics Identifications Database (PRIDE), Chapter 13, Unit 13.8. In: *Current Protocols in Bioinformatics*. John Willey & Sons, Somerset, NJ, USA.
5. Jones,P. and Cote,R. (2008) The PRIDE proteomics identifications database: data submission, query, and dataset comparison. *Methods Mol. Biol.*, **484**, 287–303.
6. Vizcaino,J.A., Côté,R., Reisinger,F., Foster,J., Mueller,M., Rameseder,J., Hermjakob,H. and Martens,L. (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, **9**, 4276–4283.
7. Craig,R., Cortens,J.P. and Beavis,R.C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.*, **3**, 1234–1242.

8. Deutsch,E.W., Lam,H. and Aebersold,R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.*, **9**, 429–434.
9. Kandasamy,K., Keerthikumar,S., Goel,R., Mathivanan,S., Patankar,N., Shafreen,B., Renuse,S., Pawar,H., Ramachandra,Y.L., Acharya,P.K. *et al.* (2009) Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res.*, **37**, D773–D781.
10. Mead,J.A., Shadforth,I.P. and Bessant,C. (2007) Public proteomic MS repositories and pipelines: available tools and biological applications. *Proteomics*, **7**, 2769–2786.
11. Slotta,D.J., Barrett,T. and Edgar,R. (2009) NCBI peptidome: a new public repository for mass spectrometry peptide identifications. *Nat. Biotechnol.*, **27**, 600–601.
12. Mead,J.A., Bianco,L. and Bessant,C. (2009) Recent developments in public proteomic MS repositories and pipelines. *Proteomics*, **9**, 861–881.
13. Editors. (2007) Democratizing proteomics data. *Nat. Biotechnol.*, **25**, 262.
14. Editors. (2008) Thou shalt share your data. *Nat. Methods*, **5**, 209.
15. Cote,R.G., Jones,P., Martens,L., Apweiler,R. and Hermjakob,H. (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, **36**, W372–W376.
16. Cote,R.G., Jones,P., Martens,L., Kerrien,S., Reisinger,F., Lin,Q., Leinonen,R., Apweiler,R. and Hermjakob,H. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, **8**, 401.
17. Reisinger,F. and Martens,L. (2009) Database on Demand – an online tool for the custom generation of FASTA formatted sequence databases. *Proteomics*, **9**, 4421–4424.
18. Gevaert,K., Van Damme,P., Ghesquiere,B., Impens,F., Martens,L., Helsens,K. and Vandekerckhove,J. (2007) A la carte proteomics with an emphasis on gel-free techniques. *Proteomics*, **7**, 2698–2718.
19. Sickmann,A., Reinders,J., Wagner,Y., Joppich,C., Zahedi,R., Meyer,H.E., Schonfisch,B., Perschil,I., Chacinska,A., Guiard,B. *et al.* (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl Acad. Sci. USA*, **100**, 13207–13212.
20. Klie,S., Martens,L., Vizcaino,J.A., Cote,R., Jones,P., Apweiler,R., Hinneburg,A. and Hermjakob,H. (2008) Analyzing large-scale proteomics projects with latent semantic indexing. *J. Proteome Res.*, **7**, 182–191.
21. Martens,L., Muller,M., Stephan,C., Hamacher,M., Reidegeld,K.A., Meyer,H.E., Bluggel,M., Vandekerckhove,J., Gevaert,K. and Apweiler,R. (2006) A comparison of the HUPO Brain Proteome Project pilot with other proteomics studies. *Proteomics*, **6**, 5076–5086.
22. Omenn,G.S., Aebersold,R. and Paik,Y.K. (2009) 7(th) HUPO World Congress of Proteomics: launching the second phase of the HUPOPlasma Proteome Project (PPP-2) 16-20 August 2008, Amsterdam, The Netherlands. *Proteomics*, **9**, 4–6.
23. Liu,T., Qian,W.J., Gritsenko,M.A., Xiao,W., Moldawer,L.L., Kaushal,A., Monroe,M.E., Varnum,S.M., Moore,R.J., Purvine,S.O. *et al.* (2006) High dynamic range characterization of the trauma patient plasma proteome. *Mol. Cell. Proteomics*, **5**, 1899–1913.
24. Liu,T., Qian,W.J., Gritsenko,M.A., Camp,D.G. 2nd, Monroe,M.E., Moore,R.J. and Smith,R.D. (2005) Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J. Proteome Res.*, **4**, 2070–2080.
25. Hermjakob,H. and Apweiler,R. (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev. Proteomics*, **3**, 1–3.
26. Merrihew,G.E., Davis,C., Ewing,B., Williams,G., Kall,L., Frewen,B.E., Noble,W.S., Green,P., Thomas,J.H. and MacCoss,M.J. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome Res.*, **18**, 1660–1669.
27. Schrimpf,S.P., Weiss,M., Reiter,L., Ahrens,C.H., Jovanovic,M., Malmstrom,J., Brunner,E., Mohanty,S., Lercher,M.J., Hunziker,P.E. *et al.* (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.*, **7**, e48.
28. Gilchrist,A., Au,C.E., Hiding,J., Bell,A.W., Fernandez-Rodriguez,J., Lesimple,S., Nagaya,H., Roy,L., Gosline,S.J., Hallett,M. *et al.* (2006) Quantitative proteomics analysis of the secretory pathway. *Cell*, **127**, 1265–1281.
29. Baerenfaller,K., Grossmann,J., Grobei,M.A., Hull,R., Hirsch-Hoffmann,M., Yalovsky,S., Zimmermann,P., Grossniklaus,U., Gruissem,W. and Baginsky,S. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, **320**, 938–941.
30. Asara,J.M., Schweitzer,M.H., Freimark,L.M., Phillips,M. and Cantley,L.C. (2007) Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science*, **316**, 280–285.
31. Bradshaw,R.A., Burlingame,A.L., Carr,S. and Aebersold,R. (2006) Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics*, **5**, 787–788.
32. Barsnes,H., Vizcaino,J.A., Eidhammer,I. and Martens,L. (2009) PRIDE Converter: making proteomics data-sharing easy. *Nat. Biotechnol.*, **27**, 598–599.
33. Helsens,K., Colaert,N., Barsnes,H., Muth,T., Flikka,K., Staes,A., Timmerman,E., Wortelkamp,S., Sickmann,A. and Vandekerckhove,J. ms_lims, a simple yet powerful open source LIMS for mass spectrometry-driven proteomics. *Proteomics*, in press.
34. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
35. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
36. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
37. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
38. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
39. Tagari,M., Tate,J., Swaminathan,G.J., Newman,R., Naim,A., Vranken,W., Kapopoulou,A., Hussain,A., Fillon,J., Henrick,K. *et al.* (2006) E-MSD: improving data deposition and structure quality. *Nucleic Acids Res.*, **34**, D287–D290.
40. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
41. Dwinell,M.R., Worthey,E.A., Shimoyama,M., Bakir-Gungor,B., DePons,J., Laulederkind,S., Lowry,T., Nigram,R., Petri,V., Smith,J. *et al.* (2009) The Rat Genome Database 2009: variation, ontologies and pathways. *Nucleic Acids Res.*, **37**, D744–D749.
42. Chicurel,M. (2002) Bioinformatics: bringing it all together. *Nature*, **419**, 751, 753, 755 passim.
43. Mueller,M., Martens,L. and Apweiler,R. (2007) Annotating the human proteome: beyond establishing a parts list. *Biochim. Biophys. Acta*, **1774**, 175–191.
44. Editors. (2009) Credit where credit is overdue. *Nat. Biotechnol.*, **27**, 579.