Published online 29 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D1063–D1069***
doi:10.1093/nar/gks1262

The Proteomics Identifications (PRIDE) database and associated tools: status in 2013

Juan Antonio Vizcaíno[1,*], Richard G. Côté[1], Attila Csordas[1], José A. Dianes[1], Antonio Fabregat[1], Joseph M. Foster[1], Johannes Griss[1], Emanuele Alpi[1], Melih Birim[1], Javier Contell[1], Gavin O'Kelly[1], Andreas Schoenegger[1,3], David Ovelleiro[1], Yasset Pérez-Riverol[1,2], Florian Reisinger[1], Daniel Ríos[1], Rui Wang[1] and Henning Hermjakob[1]

[1]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]Department of Proteomics, Center for Genetic Engineering and Biotechnology, Havana, Cuba and [3]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria

Received September 21, 2012; Revised October 30, 2012; Accepted November 5, 2012

ABSTRACT

The PRoteomics IDEntifications (PRIDE, http://www.ebi.ac.uk/pride) database at the European Bioinformatics Institute is one of the most prominent data repositories of mass spectrometry (MS)-based proteomics data. Here, we summarize recent developments in the PRIDE database and related tools. First, we provide up-to-date statistics in data content, splitting the figures by groups of organisms and species, including peptide and protein identifications, and post-translational modifications. We then describe the tools that are part of the PRIDE submission pipeline, especially the recently developed PRIDE Converter 2 (new submission tool) and PRIDE Inspector (visualization and analysis tool). We also give an update about the integration of PRIDE with other MS proteomics resources in the context of the ProteomeXchange consortium. Finally, we briefly review the quality control efforts that are ongoing at present and outline our future plans.

INTRODUCTION

Mass spectrometry (MS)-based proteomics approaches are widely used in the life sciences. There are three main workflows, with bottom-up proteomics the most extensively used technique (also known as shot-gun proteomics) (1). In this experimental set up, the proteins to be analysed are enzymatically digested by a protease (most often trypsin) into potentially highly complex peptide mixtures, which are then subjected to fractionation by multidimensional liquid chromatography steps before they are measured in the mass spectrometer. Other main approaches are top-down, where intact proteins are measured (2), and targeted proteomics (such as Selected Reaction Monitoring, SRM), where the researcher tries to detect specific proteins in a given sample (3).

The PRoteomics IDEntifications (PRIDE, http://www.ebi.ac.uk/pride) database was originally set up in 2004 (4–8) to enable public data deposition in the MS proteomics field, and to support the experimental data described in publications during the manuscript review process. The main data types stored in PRIDE are protein and peptide identifications (IDs) and quantitative values (including post-translational modifications, PTMs), the analysed mass spectra and the related technical and biological metadata. PRIDE supports bottom-up proteomics approaches, mainly tandem MS (MS/MS) data, but also Peptide Mass Fingerprinting datasets, and presents the data as originally analysed by the researchers, with several popular search engines/analysis workflows fully supported. Unlike other MS proteomics resources, such as PeptideAtlas (9) and the Global Proteome Machine Database (GPMDB) (10), no reprocessing of the data is performed because PRIDE aims to reflect the author's analysis view on the experimental data. In fact, PRIDE remains as the unique generic resource of this kind since National Center for Biotechnology Information (NCBI) the repository Peptidome (11), its sibling resource in the USA, was discontinued in April 2011. Other MS data repositories, such as MaxQB (12), are more specialized [for an extensive review, see (13)] or are restricted to one particular analysis workflow.

For SRM data, the new PeptideAtlaS SRM Experiment Library (PASSEL) (14) is the main available resource.

*To whom correspondence should be addressed. Tel: +44 1223 492686; Fax: +44 1223 494468; Email: pst@ebi.ac.uk

At present, there is no widely used resource devoted to top-down proteomics approaches. In addition to the 'pure' MS proteomics resources, there are other databases that can present an extra layer of information on top of the MS experiments without storing the underlying mass spectra. Some recently developed databases of this kind are the Model Organism Protein Expression Database (MOPED) (15), PaxDB (16) (both of them focused on protein expression information) and neXtProt (17).

Several services have been developed by the PRIDE team, which are heavily used by external users but also by PRIDE itself, especially the 'Protein Identifier Cross-Reference' (PICR) service (a protein identifier mapping resource) (18) and the 'Ontology Lookup Service' (OLS) (to query, browse and navigate biomedical ontologies) (19). In addition, 'Database on Demand' is a service to generate tailored databases for performing proteomics searches (20). As a key point, to improve and make the data submission process easier, several tools have also been made available to the proteomics community, such as the popular PRIDE Converter (21), PRIDE Inspector (22) and the new PRIDE Converter 2 (23). It is important to highlight that all the softwares, including the PRIDE core and web modules (http://code.google.com/p/ebi-pride/), are developed in Java and are open source. PRIDE is a recommended submission site of key journals such as *Proteomics*, *Molecular and Cellular Proteomics* and *Nature Biotechnology*. Currently, scientific journals and funding agencies alike are increasingly mandating public deposition of MS data to support the publication of related proteomics manuscripts.

In this manuscript, we summarize developments in the PRIDE database and associated tools since the previous *Nucleic Acids Research* (NAR) database update (8). We will also outline the PRIDE data deposition process, introduce the ProteomeXchange (PX) consortium and quality control (QC) efforts and highlight future developments.

DATA CONTENT IN PRIDE AND HOW TO ACCESS IT

There has been a substantial increase in the amount of stored data in PRIDE during the past years. By September 2012, PRIDE contains 25 853 MS-based proteomics experiments (compared with 9908 when the last NAR manuscript was submitted, in September 2009), around 11.1 million identified proteins (2.5 million in September 2009), 61.9 million identified peptides (11.5 million in September 2009) and 324 million spectra (50.3 million in September 2009). Note that these data holdings are absolute figures, not distinguishing public and pre-publication data. At the moment of writing, 66.7% (17 219) of the experiments were publicly available.

The complete set of data in PRIDE comprises 323 taxonomy identifiers (compared with 60, in September 2009), including human and many model organisms (Table 1). In comparison with figures from 3 years ago, animal species still provide the majority of the data. A total of 89 animal species are represented, contributing 62.2% and 61.9% of all protein and peptide IDs in PRIDE, respectively. Human continues to be the most represented species (28.1% and 37.2% of protein and peptide IDs, respectively). As a matter of fact, human and mouse alone account for almost as many IDs as the other species together: 44.2 % and 49.4 % of the protein and peptide IDs, respectively.

Table 1. Data content in PRIDE split by taxonomic divisions

Group of organisms (number of species)	% Protein IDs	% Peptide IDs
Animals (89)	62.2	61.9
Plants (46)	19.3	14.8
Fungi (22)	2.7	3.0
Bacteria (122)	12.7	17.4
Others (44)	3.1	2.9
Species		
Homo sapiens	28.1	37.2
Mus musculus	16.1	12.2
Zea mays	9.5	2.2
Arabidopsis thaliana	6.8	10.1
B. subtilis	5.5	4.1
Sus scrofa	5.5	1.7
Rattus norvegicus	3.7	2.8
Drosophila melanogaster	3.3	2.3
Danio rerio	1.7	0.9
Puniceispirillum marinum	1.5	1.5
E. coli	1.2	1.7
S. cerevisiae	1.2	1.8

Only the top 12 species in terms of protein and peptide identifications are shown.

However, the relative proportion of other groups of organisms has increased, especially in the case of plants (46 taxonomy identifiers, 19.3% and 14.8%, respectively) and bacteria (12.7% and 17.4%, respectively). Bacteria are again by far the group of organisms with the highest number of taxonomy identifiers (122). Fungi are also represented (22 taxonomy identifiers, 2.7 and 3.0%, respectively). Apart from human, the most represented organisms in PRIDE are (in this order) mouse, maize, Arabidopsis, *Bacillus subtilis*, pig, rat, Drosophila, zebrafish, *Puniceispirillum marinum*, *Escherichia coli* and *Saccharomyces cerevisiae* (Table 1).

Table 2 includes the most abundant PTMs present in the database. Not surprisingly, the most often found PTM is oxidation (5.7 million modified sites), mainly due to the high amount of methionine oxidation, a modification that can be biologically relevant (24) but that, in most cases for MS proteomics experiments, is an artifact. Formylation is the second most abundant PTM (around 1.3 million sites), mainly owing to the data present from just one organism (maize). Phosphorylation comes in third place (around 1.1 million sites), with the highest proportion of data coming from human experiments. There is also a considerable amount of other PTMs such as dioxidation, deamidation, acetylation or dehydration, among others (Table 2).

This wealth of data can be accessed in different ways (Figure 1):

(a) PRIDE web interface (http://www.ebi.ac.uk/pride). The home page was updated earlier in 2012, but no other major changes have been done to the current

Table 2. Protein modification content in PRIDE as a whole, and split by species (human and main model organisms represented in PRIDE)

Modification type	Total	Human	Mouse	Maize	Arabidopsis	Drosophila	*E. coli*	*S. cerevisiae*
Oxidation	5 707 426	2 291 925	883 599	25 907	449 530	2400	58 921	97 803
Deamidation	663 884	333 091	48 120	30	6964	83	17 335	10 859
Phosphorylation	1 143 766	741 619	112 910	171 539	35 521	239	0	33 670
Acetylation	626 510	380 124	22 662	1419	10 614	290	1356	18 762
Dioxidation	1 123 206	31 788	1625	928 614	0	0	0	0
Deamination	132 145	87 016	29 374	462	0	262	272	0
Dehydration	202 436	15 492	32 863	148 686	0	30	0	0
Methylthio	110 727	74 177	0	0	2425	0	82	0
Formylation	1 297 285	5286	629	1 289 660	0	0	0	0
Monomethylation	299 246	261 941	48	0	277	0	0	0

web interface. However, it is possible to access all experiments launching the PRIDE Inspector tool using Java Web Start (see below).

(b) PRIDE BioMart (http://www.ebi.ac.uk/pride/prideMart.do). The BioMart interface is useful for batch data retrieval (25). In the current version of the PRIDE BioMart (running on BioMart version 0.7), data integration with Reactome (26) has been extended (27), by enabling the link between phosphorylated proteins present in Reactome pathways and phosphorylated proteins detected by MS approaches stored in PRIDE. The PRIDE BioMart data can also be accessed using a Representational State Transfer web service, which is heavily used. In addition, it is also possible to access the PRIDE BioMart at www.biomart.org, together with many others. In that case, apart from Reactome, it is possible to perform common data searches involving PRIDE and other resources such as UniProt (28), InterPro (29), Ensembl (30) and the Catalogue of Somatic Mutations in Cancer (COSMIC) (31).

(c) PRIDE FTP file server (ftp.ebi.ac.uk/pub/databases/pride/). At present, XML files corresponding to all public experiments in PRIDE can be downloaded in the mzData and PRIDE XML formats.

(d) PRIDE Distributed Annotation System (DAS) server (http://www.ebi.ac.uk/pride-das/das/PrideDataSource/). A PRIDE DAS server (32) was set up following the new specification 1.6. This service is publicly available and can be accessed through DAS clients such as Dasty (33). The PRIDE DAS server has been designed for visualizing protein sequence and annotation, to display the identified peptides for the protein specified in the DAS request, together with the associated PTMs, and total peptide coverage.

(e) A public PRIDE MySQL instance is now available and used, for instance, by the PRIDE Inspector, making this tool be the ideal way to access PRIDE data for most use cases (see next section).

Among the new datasets present in PRIDE, it is important to highlight that by July 2012, all data originally stored in the discontinued NCBI Peptidome (11) had been reannotated, converted into a PRIDE compatible format and made publicly available under experiment accessions 17900-18271.

PRIDE SUBMISSION PROCESS AND RELATED TOOLS

Submission tools

At the moment of writing, submissions to PRIDE are performed using a publicly available XML data format called PRIDE XML, which is derived from mzData. The first step to perform a submission to PRIDE is then to generate PRIDE XML files. Several tools are available to make that process feasible and as straightforward as possible. The new PRIDE Converter 2 (23) is now the recommended submission tool. It can convert a variety of popular proteomics data formats (e.g. Mascot.dat, X!Tandem.xml, OMSSA.csv, Proteome Discoverer.msf, plus all the mass spectral formats, among others) into well-annotated PRIDE XML files.

PRIDE Converter 2 can be used in two modes: (i) a Graphical User Interface mode, suited for most users; and (ii) a Command Line interface mode that makes possible the integration of the conversion process in external pipelines. Batch conversion of files is supported in both modes. Importantly, quantification results for the most popular techniques and 2D gel spot information can now be integrated in PRIDE XML files with PRIDE Converter 2, by providing that information in a new Proteomics Standards Initiative (PSI) tab-delimited standard format called mzTab (http://code.google.com/p/mztab/). Detailed documentation for general users and developers is available at http://code.google.com/p/pride-converter-2/. The PRIDE Converter 2 framework also includes the *PRIDE mzTab Generator, PRIDE XML Merger* and *PRIDE XML Filter* (23). A mechanism to make a combined submission to PRIDE and IntAct (34), the molecular interactions resource at the European Bioinformatics Institute, is also present in the PRIDE Converter 2.

The original PRIDE Converter tool (21), one of the main reasons behind the large increase in data contents in PRIDE, is no longer recommended as the main submission tool and will not be maintained any more. Apart from the tools provided by the PRIDE team, there are several existing third-party pipelines/tools that produce PRIDE XML files, such as ProteinLynx Global Server

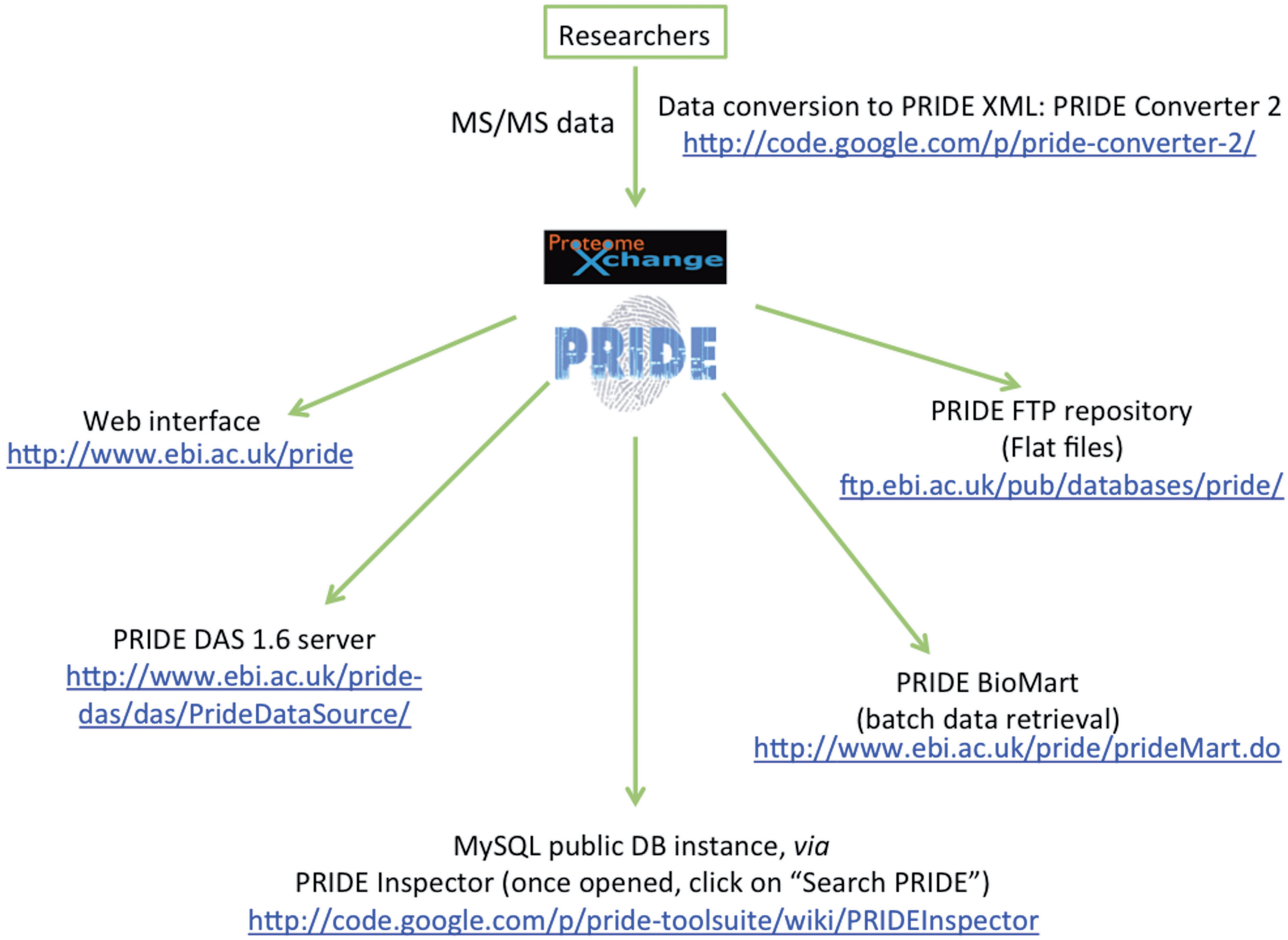

Figure 1. Summary of the ways the user can access and retrieve data from PRIDE. The web links to the existing PRIDE tools are also highlighted, including the PRIDE Converter 2 (needed for data submission).

(Waters), hEIDI (http://biodev.extra.cea.fr/docs/heidi), OmicsHub Proteomics (Integromics), PeptideShaker (http://peptide-shaker.googlecode.com) and Proteios (35).

At the moment, we are implementing full support in PRIDE for the new PSI standard formats mzML v1.1 (for MS data) (36) and mzIdentML v1.1 (for protein and peptide IDs) (37), based on the Java libraries jmzML (38), jmzIdentML (39) and jmzReader (40). When this work is complete, data submissions in these formats will be natively supported, rather than having to be run through the PRIDE Converter 2. This will enable us to support in a much better way the reporting of protein inference (41) and ambiguity in modification position.

PRIDE Inspector

PRIDE Inspector is a popular tool introduced in 2011 (>4500 downloads by September 2012), which can be used to visualize and perform an initial quality assessment of the submitted data to PRIDE (22) (http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector). PRIDE Inspector provides different views on the available data, each focusing on a different aspect: experimental details (biological and technical metadata), protein, peptide, quantification values (if available) and summary charts. This last view is one of the major strengths of the tool because it is possible to perform an initial assessment of data quality, using a variety of simple charts that are generated automatically (22). Using PRIDE Inspector, proteomics researchers can examine their own data sets before the actual submission to PRIDE is performed, and journal editors and reviewers can perform a thorough review of submitted and private data at the pre-publication stage.

In addition, through the 'Search PRIDE Database' option, PRIDE Inspector can access data already in PRIDE for data mining purposes using the PRIDE public MySQL instance, which is updated regularly. Apart from being a stand-alone tool, as mentioned before, PRIDE Inspector can also be accessed using Java Web Start at the PRIDE web page.

PRIDE and the ProteomeXchange consortium

PRIDE is a founding member of the PX consortium (http://www.proteomexchange.org) (42). The members of the consortium, led by PRIDE and PeptideAtlas (9), are implementing a system to enable the automated and standardized sharing of MS-based proteomics data between the main existing MS proteomics repositories.

In the first implementation of the data workflow, PRIDE acts as the initial submission point for MS/MS data. At the moment of writing, around 50 PX datasets have been submitted (see updated list of publicly available datasets at http://proteomecentral.proteomexchange.org). As a result of the PX efforts, PRIDE has started to accept raw data (in either binary or an open XML format) because it is a mandatory component of a PX submission. The files are accessible through FTP and are stored at the EBI raw file repository (43).

QUALITY CONTROL IN PRIDE

A major focus of PRIDE development in the past 2 years was to ensure at the very least minimal annotation of experiments and to perform basic quality checks of the submitted data to PRIDE. The development of PRIDE Inspector was the key step forward in that direction, and many of its components have been used in the internal PRIDE submission pipeline. The automated pipeline allows the detection of clear errors in the submitted data that are notified to the submitter and can then be corrected (44). Finally, the development of the PRIDE Converter 2 as the new submission tool has improved consistency at the level of experimental annotation.

In 2013, we will finalize and release a new resource called PRIDE-Q, as a quality-controlled subset of PRIDE, which can fulfil both a minimum level of annotation and Peptide Spectrum Match quality standards.

DISCUSSION

In the past 3 years, PRIDE has worked on two main tasks: (i) development of robust data submission pipelines (such as the development of the PRIDE Converter 2), including the initial implementation of the PX consortium data workflow, and the possibility to capture quantification information in a standardized way; and (ii) establishment of QC checks, including the development of PRIDE Inspector and an internal data submission pipeline, able to flag obvious errors that can be then communicated to the submitters.

However, many more future efforts will be needed in both directions. We are working on the development of the new PRIDE system (including a new database schema and web interface) that will fully support the PSI standards mzML and mzIdentML. This will be a gradual process, and support for these formats will be added sequentially to the PRIDE system (also as submission formats) and tools, while we will also keep supporting PRIDE XML.

On the other hand, the release of PRIDE-Q, envisioned to help non-expert proteomics biologists to 'digest' the potentially complex information coming from MS data, will happen in 2013. However, the quality requirements will need to be refined and will evolve dynamically over time.

It is worth highlighting that PRIDE has been used for research purposes in several recent studies involving the meta-analysis of combined data coming from very different proteomics experimental setups (45–47), the improvement and assessment of existing protein sequence databases (48,49) or to support genomics-related findings (50,51). Some research has also been performed to demonstrate the usefulness of data in PRIDE to perform *a posteriori* QC of the stored data (52). We expect that this trend will continue to grow in the near future, and that PRIDE continues on a trajectory from a publication-centric repository to an integrative resource for MS-based protein expression data. PRIDE will keep playing an important role for the community, also in the context of the nascent Human Proteome Project (53). We invite interested parties in PRIDE developments (including the associated software and tools) to follow the PRIDE Twitter account (@pride_ebi).

ACKNOWLEDGEMENTS

The PRIDE team would like to thank all data submitters for their contributions.

FUNDING

The PRIDE team is funded by the Wellcome Trust [WT085949MA]; EU FP7 grants 'Sling' [226073]; 'ProteomeXchange' [260558]; 'PRIME-XS' [262067]; 'LipidomicNet' [202272]; BBSRC grant 'PRIDE Converter' [reference BB/I024204/1] and EMBL core funding. Funding for open access charge: Wellcome Trust [WT085949MA].

Conflict of interest statement. None declared.

REFERENCES

1. Mallick,P. and Kuster,B. (2010) Proteomics: a pragmatic perspective. *Nat. Biotechnol.*, **28**, 695–709.
2. Cui,W., Rohrs,H.W. and Gross,M.L. (2011) Top-down mass spectrometry: recent developments, applications and perspectives. *Analyst*, **136**, 3854–3864.
3. Picotti,P. and Aebersold,R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods*, **9**, 555–566.
4. 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.
5. 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.
6. 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.
7. 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.
8. Vizcaino,J.A., Foster,J.M. and Martens,L. (2010) Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. *J Proteomics*, **73**, 2136–2146.
9. 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.

10. 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.
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. Schaab,C., Geiger,T., Stoehr,G., Cox,J. and Mann,M. (2012) Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell Proteomics*, **11**, M111.014068.
13. Mead,J.A., Bianco,L. and Bessant,C. (2009) Recent developments in public proteomic MS repositories and pipelines. *Proteomics*, **9**, 861–881.
14. Farrah,T., Deutsch,E.W., Kreisberg,R., Sun,Z., Campbell,D.S., Mendoza,L., Kusebauch,U., Brusniak,M.Y., Huttenhain,R., Schiess,R. *et al.* (2012) PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics*, **12**, 1170–1175.
15. Kolker,E., Higdon,R., Haynes,W., Welch,D., Broomall,W., Lancet,D., Stanberry,L. and Kolker,N. (2012) MOPED: model organism protein expression database. *Nucleic Acids Res.*, **40**, D1093–D1099.
16. Wang,M., Weiss,M., Simonovic,M., Haertinger,G., Schrimpf,S.P., Hengartner,M.O. and von Mering,C. (2012) PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell Proteomics*, **11**, 492–500.
17. Lane,L., Argoud-Puy,G., Britan,A., Cusin,I., Duek,P.D., Evalet,O., Gateau,A., Gaudet,P., Gleizes,A., Masselot,A. *et al.* (2012) neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.*, **40**, D76–D83.
18. Wein,S.P., Cote,R.G., Dumousseau,M., Reisinger,F., Hermjakob,H. and Vizcaino,J.A. (2012) Improvements in the protein identifier cross-reference service. *Nucleic Acids Res.*, **40**, W276–W280.
19. Cote,R., Reisinger,F., Martens,L., Barsnes,H., Vizcaino,J.A. and Hermjakob,H. (2010) The Ontology Lookup Service: bigger and better. *Nucleic Acids Res.*, **38**, W155–W160.
20. 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.
21. Barsnes,H., Vizcaino,J.A., Eidhammer,I. and Martens,L. (2009) PRIDE Converter: making proteomics data-sharing easy. *Nat. Biotechnol.*, **27**, 598–599.
22. Wang,R., Fabregat,A., Rios,D., Ovelleiro,D., Foster,J.M., Cote,R.G., Griss,J., Csordas,A., Perez-Riverol,Y., Reisinger,F. *et al.* (2012) PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat. Biotechnol.*, **30**, 135–137.
23. Cote,R.G., Griss,J., Dianes,J.A., Wang,R., Wright,J.C., van den Toorn,H.W., van Breukelen,B., Heck,A.J., Hulstaert,N., Martens,L. *et al.* (2012) The PRIDE Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol. Cell Proteomics*, **11**, 1682–1689.
24. Stadtman,E.R., Van Remmen,H., Richardson,A., Wehr,N.B. and Levine,R.L. (2005) Methionine oxidation and aging. *Biochim. Biophys. Acta*, **1703**, 135–140.
25. Zhang,J., Haider,S., Baran,J., Cros,A., Guberman,J.M., Hsu,J., Liang,Y., Yao,L. and Kasprzyk,A. (2011) BioMart: a data federation framework for large collaborative projects. *Database*, **2011**, bar038.
26. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
27. Ndegwa,N., Cote,R.G., Ovelleiro,D., D'Eustachio,P., Hermjakob,H., Vizcaino,J.A. and Croft,D. (2011) Critical amino acid residues in proteins: a BioMart integration of Reactome protein annotations with PRIDE mass spectrometry data and COSMIC somatic mutations. *Database (Oxford)*, **2011**, bar047.
28. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
29. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
30. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
31. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.
32. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
33. Villaveces,J.M., Jimenez,R.C., Garcia,L.J., Salazar,G.A., Gel,B., Mulder,N., Martin,M., Garcia,A. and Hermjakob,H. (2011) Dasty3, a WEB framework for DAS. *Bioinformatics*, **27**, 2616–2617.
34. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
35. Hakkinen,J., Vincic,G., Mansson,O., Warell,K. and Levander,F. (2009) The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.*, **8**, 3037–3043.
36. Martens,L., Chambers,M., Sturm,M., Kessner,D., Levander,F., Shofstahl,J., Tang,W.H., Rompp,A., Neumann,S., Pizarro,A.D. *et al.* (2011) mzML—a community standard for mass spectrometry data. *Mol. Cell Proteomics*, **10**, R110 000133.
37. Jones,A.R., Eisenacher,M., Mayer,G., Kohlbacher,O., Siepen,J., Hubbard,S.J., Selley,J.N., Searle,B.C., Shofstahl,J., Seymour,S.L. *et al.* (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell Proteomics*, **11**, M111.014381.
38. Cote,R.G., Reisinger,F. and Martens,L. (2010) jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics*, **10**, 1332–1335.
39. Reisinger,F., Krishna,R., Ghali,F., Rios,D., Hermjakob,H., Vizcaino,J.A. and Jones,A.R. (2012) jmzIdentML API: a Java interface to the mzIdentML standard for peptide and protein identification data. *Proteomics*, **12**, 790–794.
40. Griss,J., Reisinger,F., Hermjakob,H. and Vizcaino,J.A. (2012) jmzReader: a Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics*, **12**, 795–798.
41. Nesvizhskii,A.I. and Aebersold,R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics*, **4**, 1419–1440.
42. 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.
43. Editorial. (2012) A home for raw proteomics data. *Nat. Methods*, **9**, 419.
44. Csordas,A., Ovelleiro,D., Wang,R., Foster,J.M., Rios,D., Vizcaino,J.A. and Hermjakob,H. (2012) PRIDE: quality control in a proteomics data repository. *Database (Oxford)*, **2012**, bas004.
45. Mueller,M., Vizcaino,J.A., Jones,P., Cote,R., Thorneycroft,D., Apweiler,R., Hermjakob,H. and Martens,L. (2008) Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. *Proteomics*, **8**, 1138–1148.
46. 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.
47. Gonnelli,G., Hulstaert,N., Degroeve,S. and Martens,L. (2012) Towards a human proteomics atlas. *Anal Bioanal Chem*, **404**, 1069–1077.
48. Griss,J., Cote,R.G., Gerner,C., Hermjakob,H. and Vizcaino,J.A. (2011) Published and perished? The influence of the searched protein database on the long-term storage of proteomics data. *Mol. Cell Proteomics*, **10**, M111.008490.
49. Griss,J., Martin,M., O'Donovan,C., Apweiler,R., Hermjakob,H. and Vizcaino,J.A. (2011) Consequences of the discontinuation of

the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. *Proteomics*, **11**, 4434–4438.
50. Knowles,D.G. and McLysaght,A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res.*, **19**, 1752–1759.
51. Panchin,A.Y., Gelfand,M.S., Ramensky,V.E. and Artamonova,I.I. (2010) Asymmetric and non-uniform evolution of recently duplicated human genes. *Biol. Direct.*, **5**, 54.
52. Foster,J.M., Degroeve,S., Gatto,L., Visser,M., Wang,R., Griss,J., Apweiler,R. and Martens,L. (2011) A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics*, **11**, 2182–2194.
53. Paik,Y.K., Omenn,G.S., Uhlen,M., Hanash,S., Marko-Varga,G., Aebersold,R., Bairoch,A., Yamamoto,T., Legrain,P., Lee,H.J. *et al.* (2012) Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.*, **11**, 2005–2013.