

PRIDE Inspector: a tool to visualize and validate MS proteomics data

To the Editor:

Your editorial “Credit where credit is overdue”¹ aptly summarized the existing situation in the proteomics field, where full data disclosure remains very much a work in progress. Importantly, it also correctly pointed out that “the software provided by the public repositories for searching and analyzing proteomics data is not as efficient and as user friendly as it could be.” In this context, we introduce to readers PRIDE Inspector (<http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector>), a user-friendly, freely available open-source software tool that allows the user to efficiently browse and visualize mass spectrometry (MS) proteomics data. One of the key features of PRIDE Inspector is that it allows the user to carry out an initial assessment on data quality and reliability. PRIDE Inspector can thus be used by researchers before they submit their data, by journal editors and peer reviewers during the manuscript review process and by any interested user in the field after public release of the data in PRIDE (the PRoteomics IDentifications database; **Fig. 1**).

Despite the increasing popularity of MS-based proteomics and the overall tendency in the life sciences toward open sharing of biological data, relatively little proteomics data are currently available in the public domain. This situation is changing, however, thanks to stricter data-sharing guidelines by scientific journals and funding agencies. Some proteomics journals (e.g., *Proteomics* and *Molecular and Cellular Proteomics*; MCP) recommend, and in some concrete cases mandate, public deposition of MS data in support of manuscripts. Journals from the *Nature* group also strongly recommend submission of proteomics data to repositories like PRIDE², PeptideAtlas³ and Tranche⁴ (<http://www.nature.com/authors/policies/availability.html>).

Nevertheless, in practical terms, this public data-sharing policy can succeed only if reliable and user-friendly software tools exist to streamline the submission task.

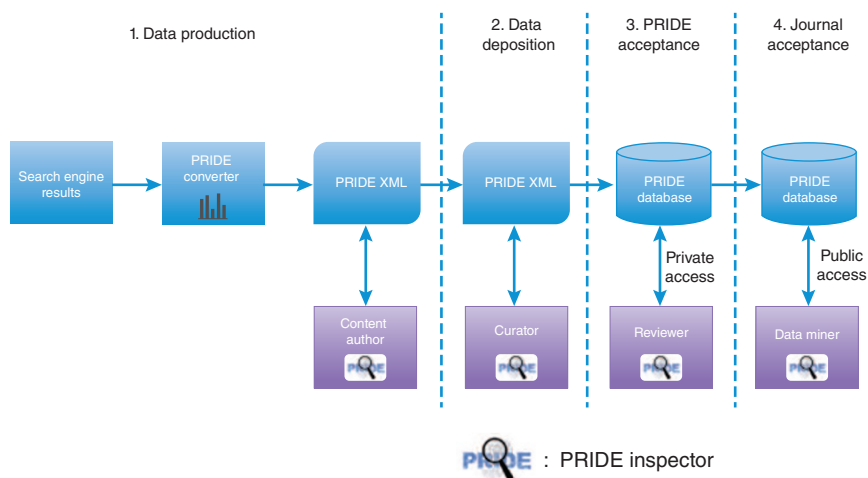


Figure 1 PRIDE Inspector helps to perform every stage of the PRIDE submission workflow. The workflow consists of four stages. (1) Data production. Search engine output results are converted into PRIDE XML files using PRIDE Converter. Authors can then use the PRIDE Inspector to perform an initial assessment on data quality and check metadata annotation before submission to PRIDE. (2) Data deposition. The submitted PRIDE XML files are reviewed by PRIDE's in-house curators using the PRIDE Inspector. (3) PRIDE acceptance. The submission is accepted by PRIDE and the data are kept private. Journal reviewers and editors can access these private PRIDE experiments using the PRIDE Inspector. (4) Journal acceptance. The submission is made public in the PRIDE database after journal acceptance. Data miners can extract, download or view PRIDE experiments using the PRIDE Inspector.

Therefore, the PRIDE Converter⁵ application (<http://code.google.com/p/pride-converter>) was developed for data submissions to the PRIDE database². Not only has PRIDE Converter rapidly become the most popular data submission path for PRIDE (accounting for 77% of all PRIDE experiments submitted since January 2009), its release also corresponded to the start of a very substantial increase in the amount of deposited data in PRIDE (**Supplementary Fig. 1**). Of course, the availability of data in public repositories is only a first step. The interpretation and validation of proteomics data remain controversial, especially for cases where proteins have been identified on the basis of one unique peptide-to-spectrum match, or if post-translational modifications (PTMs) are reported. The ability to inspect and validate reported results during the review process, as well as after publication, is therefore of

paramount importance. Because of the amount of data involved, such inspections can be undertaken efficiently only with the help of suitable software tools that combine ease of access with effective visualizations.

Although viewers for MS proteomics data are already available^{6,7}, they tend to suffer from different types of limitations. They may have been developed around a single proprietary and/or unique data format, fail to properly handle the very large files that are routinely produced, have only limited visualization and analysis functionality or be costly to license for smaller groups or individuals. We therefore developed PRIDE Inspector as a very user-friendly, freely available tool to browse, inspect and analyze proteomics data from the PRIDE repository or other data presented in standard formats.

PRIDE Inspector is a stand-alone Graphical User Interface (GUI) written in

Java. It is released under the Apache2 open-source license and can be freely downloaded. Furthermore, PRIDE Inspector can also be started through a direct web link from the PRIDE homepage (<http://www.ebi.ac.uk/pride>). The main features of PRIDE Inspector are listed in the **Supplementary Notes**, along with a description of its overall software architecture and other technical details.

PRIDE Inspector supports fast loading of PRIDE XML and mzML⁸ (the community data standard for MS data) files, and it provides direct access to all public PRIDE data through a direct MySQL database connection. Moreover, this software includes an automated data download capability for private PRIDE experiments that allows journal editors and peer reviewers with the correct log-in credentials to assess the relevant experiment(s) during peer review. In addition, the Web Start version available at the PRIDE homepage adds the ability to start the application and access a particular data set through a simple URL.

PRIDE Inspector presents different views to the users, each focusing on a specific aspect of the data (**Fig. 2**). Depending on the type of information available for a file format or PRIDE data set, some views can remain inactive (**Supplementary Fig. 2**). For that reason, an 'Experiment Summary' overview window is available in the bottom left part of the GUI. A context-sensitive 'Help' function is also included, providing tailored documentation for the current view. Currently, there are six views available in PRIDE Inspector. First, the 'Overview' tab, which

includes easily readable, uniform experimental metadata. The precise information displayed can vary slightly depending on the file format used and is split into three different views: 'Experiment General', 'Sample and Protocol' and 'Instrument and Processing' (**Supplementary Figs. 3–6**).

The second view concerns proteins (**Supplementary Figs. 7 and 8**) and is possibly the most interesting view for biologists. For each identified protein, peptides, PTMs and corresponding spectra are displayed in a concise manner. Metadata related to protein identification (e.g., as search engine or search database) are also provided here. A powerful spectrum viewer is available as well, including an automatic annotation of the spectra based on submitted fragment ions. Combinations of up to three amino acids are indicated next to the mass differences between consecutive peaks (**Supplementary Figs. 7 and 9**).

PRIDE Inspector also accesses some of the most popular protein databases (UniProtKB, UniParc, IPI (International Protein Index), Ensembl and NCBI nr database) by means of a web service to retrieve the most up-to-date protein sequences and names for the reported identifiers. Using the PRIDE Inspector sequence viewer (**Supplementary Figs. 8 and 11**), it is possible to highlight different features in the protein sequence, such as identified peptides and PTMs. The updated status of the protein identifier in the database (active, deleted, changed, unknown, merged or demerged;

see **Supplementary Notes**) is also provided, which can affect the reliability of the protein identification. In fact, it is then possible to find peptides that originally matched the sequence of the identified protein, but that no longer match the most recent version of the sequence in the database.

The third view then focuses on the peptide identifications themselves. Metadata, such as peptide score (adapted for the search engine used) and observed PTMs, are displayed for each peptide (**Supplementary Figs. 10 and 11**). In both protein and peptide views, the difference between experimental and theoretical mass-over-charge ratio (delta m/z) is calculated for each peptide precursor and highlighted in the application, which can be useful as an indication for errors or inconsistencies. For both views, it is also possible to filter out the decoy matches and, as such, a straightforward estimation of the peptide false-discovery rate is also provided.

The fourth view is aimed at accessing and visualizing all spectra in the data set, not only the identified ones (**Supplementary Fig. 12**). For mzML files, chromatograms are displayed here as well (**Supplementary Fig. 13**). Submitted metadata (e.g., precursor m/z and intensity) are shown for each entry, along with calculated information, such as the number of peaks or the total peak intensity. Manual annotation of spectra is supported as well for quick *de novo* sequencing.

In its fifth view, PRIDE Inspector provides a collection of summary charts for assessing the overall properties of the data set. At the time of writing, up to eight different charts can be generated per data set, depending on the information available (**Supplementary Figs. 14–18**). These simple and easily understandable charts can provide a quick overview on data quality and reliability. Importantly, information in the spectrum-related charts can be shown for identified, unidentified or all spectra. Each chart is documented thoroughly in the supplementary information.

Finally, a sixth tab focuses on the quantification information, where available (**Supplementary Fig. 19**). This kind of data is currently only present in a small number of PRIDE submissions, but it is expected to become more and more popular. Apart from visualizing the quantification values for both protein and peptides, it is also possible to generate histograms where the expression values of up to ten proteins can be compared. Sample metadata for each reagent can also be easily visualized. Ratios can always be recalculated if the user decides to change the control sample.

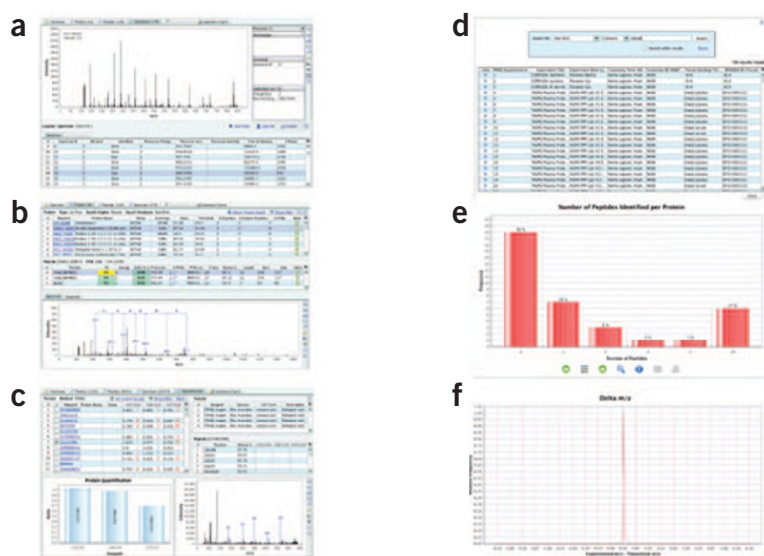


Figure 2 Screenshots showing some of the graphical features of PRIDE Inspector. (a) Section of the spectrum view tab. (b) Protein view tab, including the spectrum viewer showing MS/MS fragment ion annotations (only b ion annotations are shown). (c) Quantification view. (d) 'Search PRIDE' panel. (e) Number of peptides identified per protein chart. (f) 'Delta m/z ' chart.

Apart from the six main tabs, the 'Search PRIDE' panel gives access to all public data in PRIDE. It is then easy to search for particular experiments filtering by different types of metadata (**Supplementary Figs. 20 and 21**). In addition to data visualization and analysis functionality, PRIDE Inspector also provides various data export options (**Supplementary Fig. 22**). First of all, all spectra can be exported to Mascot Generic Format (mgf) files. In addition, details for all protein and/or peptide identifications (including PTMs), and the peptide-to-protein mappings can be output as tables in tab-delimited format. Finally, spectra and chromatograms (including annotations) can be saved as images in various formats.

PRIDE Inspector is fully supported and maintained by the PRIDE team. Moreover, it provides extra application programming interfaces (APIs) and libraries, which can be reused independently by the scientific community: the PRIDE XML JAXB (Java Architecture for XML Binding) library (for rapid and memory-efficient reading of PRIDE XML files) and the PRIDE mzGraph Browser library (for the visualization and annotation of spectra and chromatograms). These libraries are described in the **Supplementary Notes**. In addition, new features can be easily added to PRIDE Inspector thanks to its modular software architecture and permissive open-source licensing. Currently ongoing extensions include full support of the version 1.1 of the mzIdentML community standard for peptide and protein identifications⁹ because this format has only just reached stability (v1.1 was released on September 2011). Once mzIdentML is fully supported, it will also be possible to check thoroughly the issues related to protein inference¹⁰. This means that researchers need to be aware of this limitation when interpreting protein identifications reported by ambiguous (or shared) peptides. The PRIDE XML format is limited for that aim in the sense that only one of the possible peptide-protein mappings is usually reported.

PRIDE Inspector thus provides a user-friendly, comprehensive tool for the browsing, inspection and evaluation of data in the PRIDE database, or in a compatible standard file format. As such, we believe that PRIDE Inspector will substantially increase the ability of researchers, editors and peer reviewers to explore, review, evaluate and reuse proteomics data.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS

R.W. did most of the programming of the core components and the GUI. A.F. was mainly responsible for the chart component. D.R. was the main developer behind the access component of the PRIDE MySQL instance. D.O., J.M.F., R.G.C., J.G., A.C., Y.P.-R. and F.R. contributed to multiple areas during the development of the tool and also participated in the writing of the documentation and testing process. L.M. had the original idea and started the project. H.H. and J.A.V. supervised the whole process. J.A.V. and L.M. wrote the manuscript. All authors have agreed to all the content in the manuscript, including the data as presented.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust (grant number WT085949MA) and EMBL core funding. R.G.C. is supported by EU FP7 grant SLING (grant number 226073). J.A.V. is supported by the EU FP7 grants LipidomicNet (grant number 202272) and ProteomeXchange (grant number 260558). A.F. was partially supported by the Spanish network COMBIOMED (RD07/0067/0006, ISCIII-FIS). L.M. would like to acknowledge support from the EU FP7 PRIME-XS grant (grant number 262067).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Rui Wang¹, Antonio Fabregat¹, Daniel Ríos¹, David Ovelleiro¹, Joseph M Foster¹, Richard G Côté¹, Johannes Griss^{1,2}, Attila Csordas¹, Yasset Perez-Riverol^{1,3},

Florian Reisinger¹, Henning Hermjakob¹, Lennart Martens^{4,5} & Juan Antonio Vizcaino¹

¹EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

²Department of Medicine I, Medical University of Vienna, Vienna, Austria. ³Department of Proteomics, Center for Genetic Engineering and Biotechnology, Cubanacán, Playa, Ciudad de la Habana, Cuba. ⁴Department of Medical Protein Research, Ghent, Belgium. ⁵Department of Biochemistry, Ghent University, Ghent, Belgium. e-mail: juan@ebi.ac.uk

1. Anonymous. *Nat. Biotechnol.* **27**, 579 (2009).
2. Vizcaino, J.A. *et al. Nucleic Acids Res.* **38**, D736–D742 (2010).
3. Deutsch, E.W., Lam, H. & Aebersold, R. *EMBO Rep.* **9**, 429–434 (2008).
4. Hill, J.A., Smith, B.E., Papoulias, P.G. & Andrews, P.C. *J. Proteome Res.* **9**, 2809–2811 (2010).
5. Barsnes, H., Vizcaino, J.A., Eidhammer, I. & Martens, L. *Nat. Biotechnol.* **27**, 598–599 (2009).
6. Searle, B.C. *Proteomics* **10**, 1265–1269 (2010).
7. Medina-Aunon, J.A., Carazo, J.M. & Albar, J.P. *Proteomics* **11**, 334–337 (2011).
8. Martens, L. *et al. Mol. Cell Proteomics* **10**, R11000133 (2011).
9. Eisenacher, M. *Methods Mol. Biol.* **696**, 161–177 (2011).
10. Nesvizhskii, A.I. & Aebersold, R. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).

Corrected after print 8 February 2012.

OpenVigil—free eyeballs on AERS pharmacovigilance data

To the Editor:

The US Food and Drug Administration's (FDA) Adverse Event Reporting System (AERS) makes adverse drug reaction (ADR) data available to the community through the US Freedom of Information Act. Analysis of these data, however, is complicated by the file format and the huge size of the database. An editorial in this journal urged the FDA to update the database and make access "available to all [...] FDA staffers as well as experts from outside the agency"¹. Indeed, in a subsequent article, Pratt and Danese² announced their implementation of an AERS query engine, which they dubbed 'FDable'. Unfortunately, like other commercial search engines, FDable is not free.

As cost is clearly an impediment to investigators at academic institutions

with no or low financial budgets available for pharmacovigilance research, we have developed an AERS search engine named



OpenVigil that is publicly available free of charge through the webserver of our university (<http://www.uni-kiel.de/pharmacology/pvt>). To guarantee that this search option will remain free and available, we have additionally published the source code of the current version on Sourceforge.net under the GNU General Public License (GPL) (<http://openvigil.sourceforge.net/>).

The usage of raw pharmacovigilance data is limited by several factors, including under- and overreporting, missing denominator (that is, number of doses for a drug), wrong, duplicate and/or missing data in the database. These problems have been discussed and summarized by

Erratum: PRIDE Inspector: a tool to visualize and validate MS proteomics data

Rui Wang, Antonio Fabregat, Daniel Ríos, David Ovelleiro, Joseph M Foster, Richard G Côté, Johannes Griss, Attila Csordas, Yasset Perez-Riverol, Florian Reisinger, Henning Hermjakob, Lennart Martens & Juan Antonio Vizcaíno
Nat. Biotechnol. **30**, 135–137 (2012); published online 8 February 2012; corrected after print 8 February 2012.

In the print version of this article, “Supplementary” appears alone on the last two pages. The citation should have been for “Supplementary Notes.” The error has been corrected in the HTML and PDF versions of the article before online publication.