STANDARDISATION & GUIDELINES

Shedding light on black boxes in protein identification

Marc Vaudel[1,2], A. Saskia Venne[2], Frode S. Berven[1,3,4], René P. Zahedi[2], Lennart Martens[5,6] and Harald Barsnes[1]

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Norway
[2] Leibniz—Institut für Analytische Wissenschaften—ISAS—e.V., Dortmund, Germany
[3] Department of Clinical Medicine, The KG Jebsen Centre for MS-research, University of Bergen, Bergen, Norway
[4] Department of Neurology, The Norwegian Multiple Sclerosis Competence Centre, Haukeland University Hospital, Bergen, Norway
[5] Department of Medical Protein Research, VIB, Ghent, Belgium
[6] Department of Biochemistry, Ghent University, Ghent, Belgium

Performing a well thought-out proteomics data analysis can be a daunting task, especially for newcomers to the field. Even researchers experienced in the proteomics field can find it challenging to follow existing publication guidelines for MS-based protein identification and characterization in detail. One of the primary goals of bioinformatics is to enable any researcher to interpret the vast amounts of data generated in modern biology, by providing user-friendly and robust end-user applications, clear documentation, and corresponding teaching materials. In that spirit, we here present an extensive tutorial for peptide and protein identification, available at http://compomics.com/bioinformatics-for-proteomics. The material is completely based on freely available and open-source tools, and has already been used and refined at numerous international courses over the past 3 years. During this time, it has demonstrated its ability to allow even complete beginners to intuitively conduct advanced bioinformatics workflows, interpret the results, and understand their context. This tutorial is thus aimed at fully empowering users, by removing black boxes in the proteomics informatics pipeline.

Received: November 5, 2013
Revised: January 10, 2014
Accepted: January 22, 2014

Keywords:
Bioinformatics / Open Source / Protein Identification / Publication Guidelines / Tutorial

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

Proteomics aims at answering complex biological questions using advanced technology [1] and workflows include multiple intricate steps such as: (1) sample preparation [2–5]; (2) biological compound separation [6] and ionization [7, 8]; (3) electromagnetic transport, trapping, and fragmentation of complex, ionized, gas phase molecules [9]; (4) their high-accuracy mass measurement [10]; and finally, (5) the interpretation and dissemination of the often vast amounts of data produced [11–14]. Proteomics informatics, positioned at the interface between the experimental raw results and the biological interpretation, has the potential of bringing a detailed understanding of the experimental results to the scientists, empowering them to deduce the most correct interpretation.

However, a common pitfall in this scenario is to consider bioinformatics tools as black boxes that "automagically" retrieve lists of protein accession numbers from spectrum files (Fig. 1). Such an approach does not only disregard the outstanding capabilities of information technology in biology, but can also lead scientists to draw inappropriate conclusions based on experimental or computational artifacts [15–17]. It is therefore the scientific responsibility of the proteomics community as a whole to move toward fully transparent workflows. Two aspects are critical to avoid black boxes: (1) the methods and their implementation details have to be freely available; and (2) the software has to support intuitive

Correspondence: Dr. Harald Barsnes, Proteomics Unit, Department of Biomedicine, University of Bergen, Jonas Liesvei 91, N-5009 Bergen, Norway
E-mail: harald.barsnes@biomed.uib.no
Fax: +47-55-58-63-60

Colour Online: See the article online to view Figs. 2 and 3 in colour.

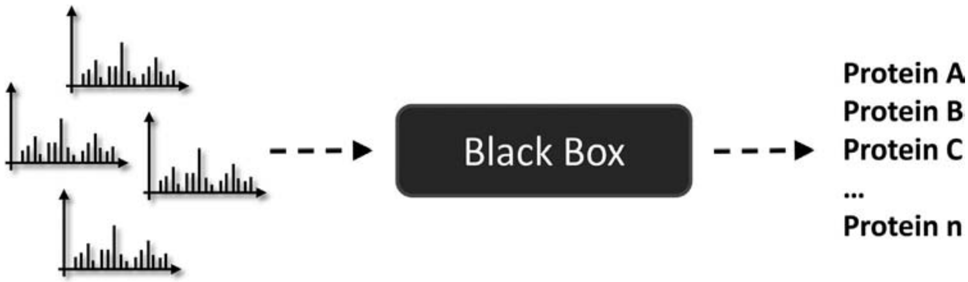

Figure 1. Block boxes represent an all too common way of thinking about bioinformatics tools in proteomics, where the spectra are used as input and a list of proteins "automagically" comes out at the other end.

interpretation, inspection, and validation of the results by any user.

The first objective requires the scientist to be familiar with bioinformatic and statistical methods [18]. However, these methods and their vocabulary—with its numerous cryptic acronyms (FDR, FNR, PEP, GO, KEGG, A-, D-, MD-scores, etc.)—present a first challenge when trying to understand proteomics informatics. Moreover, a transparent implementation requires the development of more high-quality open-source software [19, 20]. Too often, tools are mainly meant to be used in-house or have been developed to tackle a very specific issue that may not be relevant to other labs. And while it is one thing to make tools that do their job in close contact with the developers and in a specific environment, it is a very different (and much more demanding!) task to develop and maintain tools meant to be used by the proteomics community at large. As a result, labs without in-house bioinformatics support face a wide gap between the listed publication requirements for protein identification and characterization [21] and the ability to achieve this level of reporting detail using only open-source tools.

The second objective, the intuitive interaction with the results, is achieved by putting the user at the center of the development focus: the demands for user-friendliness, documentation, and support cannot be stressed enough. The installation and execution of proteomics software should ideally not require advanced computer skills or specific hardware. (Although with the growing size of modern proteomics datasets, better hardware usually means quicker processing.) Moreover, user-friendly tools ought to allow (1) visual inspection of the data; (2) interaction with the results; and (3) validation of the final output—even on large datasets. This enables highly useful quality control [22, 23], and provides a crucial link between the experiment and the biological conclusion. Documentation and support can take on many forms, from simple text files to rich interactive web pages and discussion forums, and can be applied at many levels, from how to install and start the tool to point-and-click guides for important features. A challenge here is to make sophisticated bioinformatics methods easily accessible to nonexpert users, while at the same time showing how best to use the tools to meet the quality requirements of the field, and get valid and confident results.

In order to shed light on the many black boxes in proteomics informatics, it is thus crucial to combine open-source software with user-friendliness and extensive documentation and teaching material. There are examples of such material, but it is most often focused on a specific tool or software package, e.g. TPP (http://tools.proteomecenter.org/wiki/index.php?title=TPP_Tutorial) or OpenMS (http://ftp.mi.fu-berlin.de/OpenMS/release-documentation/OpenMS_tutorial.pdf), or limited to a given subject, e.g. selected reaction monitoring [24]. To complement these efforts, we have created extensive, freely available online tutorial material for protein identification and characterization. It covers a complete workflow from sequence database generation to the sharing of the results, and relies entirely on user-oriented, community-developed open-source software. The tutorials have been developed over a 4-year period and have already been used and evaluated at numerous international courses and workshops, allowing us to validate and further improve the quality of the material. All the material is available under the permissive Creative Commons Attribution-Share Alike 3.0 license, and is freely available at http://compomics.com/bioinformatics-for-proteomics (Fig. 2).

The tutorial details the main bioinformatics steps of protein identification in sequence: (1) database generation with a focus on UniProt databases [25]; (2) peak list generation using the standard ProteoWizard library [26]; (3) peptide to spectrum matching using the freely available OMSSA [27] and X!Tandem [28] search engines via SearchGUI [29]; (4) detailed processing and inspection of identification results using PeptideShaker (http://peptide-shaker.googlecode.com); and (5) peptide and protein identification validation using the target/decoy approach [30]. The complete workflow is illustrated using a dataset obtained from an LC-MS analysis of a HeLa lysate on a Q Exactive mass spectrometer (see Supporting Information for details). Notably, the tools and methods presented are applicable to any fragmentation and shotgun MS technique, independently from the manufacturer.

Following the identification-related tutorials, a functional analysis using various online resources is conducted to further annotate the identified proteins and show the reader how to enrich a list of protein identifications with existing biological knowledge. Databases and tools introduced in this context include: UniProt [25], Ensembl [31], Gene Ontology [32], Dasty3 [33], STRING [34], Reactome [35], and the Protein Data Bank [36]. The purpose here is not to give an in-depth introduction to each resource, but rather to make the reader

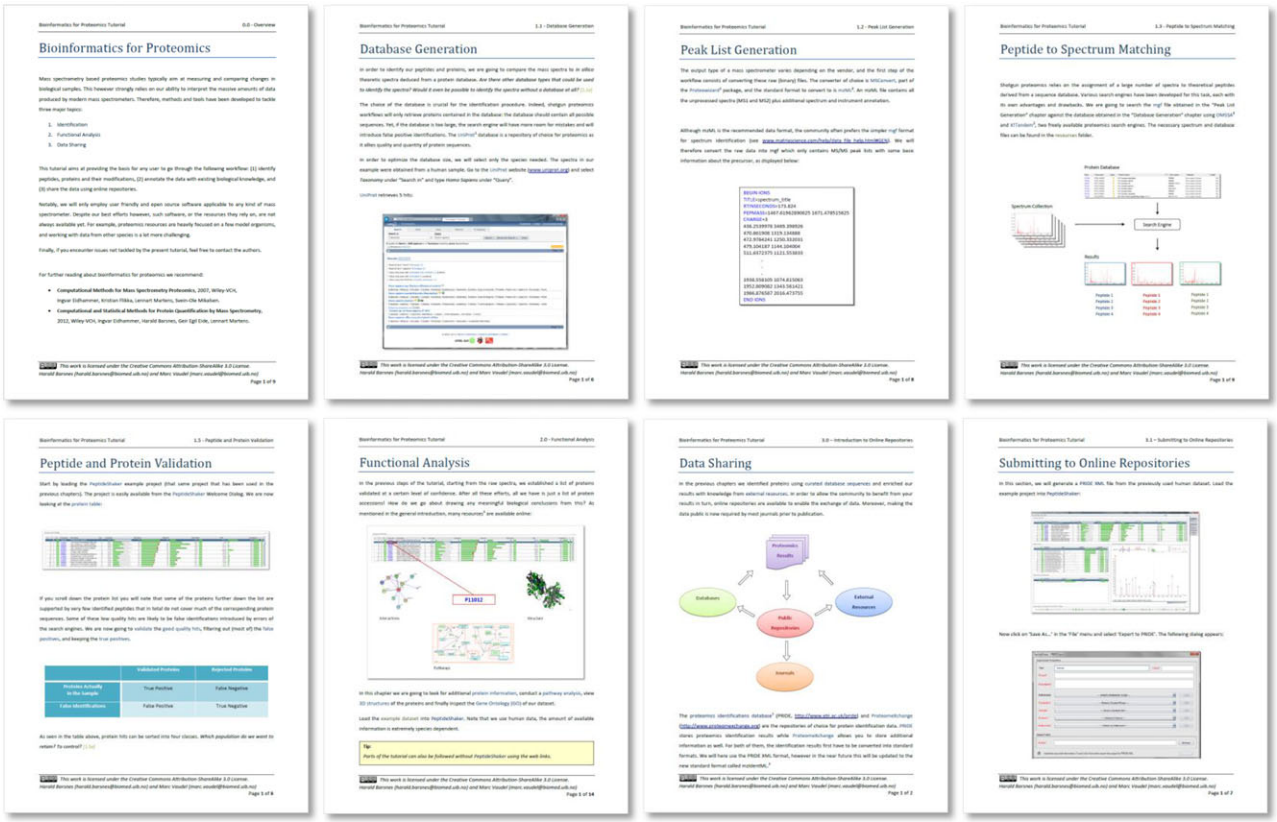

Figure 2. Overview of the main topics covered in the tutorial: (1) Bioinformatics for Proteomics (introduction); (2) Database Generation; (3) Peak List Generation; (4) Peptide to Spectrum Matching; (5) Peptide and Protein Validation; (6) Functional Analysis; (7) Data Sharing; and (8) Submitting to Online Repositories. The full content is available at http://compomics.com/bioinformatics-for-proteomics.

aware of the numerous resources that can be used to increase the understanding of the obtained proteomics data [15, 37].

Finally, the tutorial covers the increasingly important step of submitting the analysis results to PRIDE [38] according to the ProteomeXchange guidelines (http://www.proteomexchange.org). Additionally, the use of PRIDE Inspector [39] to view publicly available datasets is demonstrated, and a novel way to perform simple reanalysis of such data with only a couple of mouse clicks is introduced. An overview of all the tools and how they interact is shown in Fig. 3.

All the material is written with the novice proteomics user in mind, clarifying all concepts and acronyms commonly found in protein identification. But the tutorials also highlight aspects that even experienced proteomics researchers may not have considered in detail. The individual sections are independent, enabling the reader to focus on specific subjects without the need to go through the entire tutorial. Screenshots and illustrations are used extensively throughout the text, both to ensure readers that they are on the right path as well as to emphasize important details.

Finally, the text contains numerous questions and tips helping the reader throughout the analysis and drawing attention to crucial details in a proteomics analysis pipeline. The questions range from *What is the difference between using a mass tolerance in ppm or Dalton?* to *Would you rather use an FDR, PEP, or FNR validation threshold?* Detailed answers to all questions can be found at the tutorial web page. Feedback on the tutorial material can be provided at the same location using an online evaluation form and the provided feedback will be used as part of the ongoing process of maintaining, improving, and extending the material. We are already planning additional sections on PTM localization analysis, de novo sequencing, and various approaches for quantitative proteomics. New sections will be made available on the web site as soon as the content has been tested and validated.

In conclusion, by covering a clear and complete protein identification and characterization workflow, from sequence database generation to the sharing of the resulting proteomics identifications, our tutorial material will allow any researcher to perform high-quality proteomics data analysis and reach the standards of the publication guidelines. Moreover, by only relying on universal open-source, user-friendly, and well-documented tools, we here present a fully transparent workflow that empowers the scientists to master and understand every detail of the process. With this, we hope to contribute to shedding much needed light on the remaining proteomics bioinformatics black boxes.

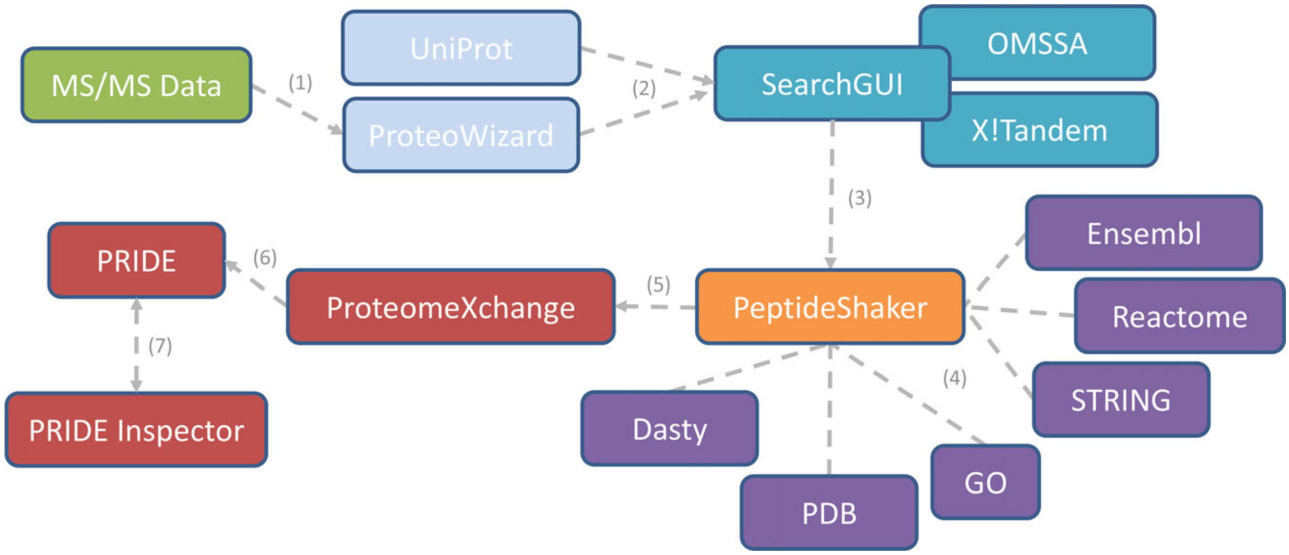

Figure 3. Overview of the proteomics workflow covered in the tutorial with a focus on the (freely available) tools employed and how they interact. (1) Raw MS/MS data is the starting point, and these data are then (2) converted to peak lists used as input to the search engines along with a sequence database. (3) A search is performed and identified peptides and proteins validated and (4) annotated with existing biological knowledge. (5) The results are converted and (6) made publicly available. (7) Finally, the public data can be inspected.

The financial support by the Ministerium für Innovation, Wissenschaft, Forschung und Technologie des Landes Nordrhein-Westfalen and by the Bundesministerium für Bildung und Forschung is gratefully acknowledged by M.V., A.S.V., and R.P.Z. L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), the PRIME-XS project, grant agreement number 262067, and "ProteomeXchange" project, grant agreement number 260558, both funded by the European Union 7th Framework Program. H.B. is supported by the Research Council of Norway. Finally, the authors would like to thank all the course participants through the years for testing and improving the tutorial material.

The authors have declared no conflict of interest.

References

[1] Nilsson, T., Mann, M., Aebersold, R., Yates, J. R., 3rd et al., Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* 2010, *7*, 681–685.

[2] Bodzon-Kulakowska, A., Bierczynska-Krzysik, A., Dylag, T., Drabik, A. et al., Methods for samples preparation in proteomic research. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, *849*, 1–31.

[3] Canas, B., Pineiro, C., Calvo, E., Lopez-Ferrer, D., Gallardo, J. M., Trends in sample preparation for classical and second generation proteomics. *J. Chromatogr. A* 2007, *1153*, 235–258.

[4] Granvogl, B., Ploscher, M., Eichacker, L. A., Sample preparation by in-gel digestion for mass spectrometry-based proteomics. *Anal. Bioanal. Chem.* 2007, *389*, 991–1002.

[5] Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A., Zahedi, R. P., Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* 2012, *75*, 1454–62.

[6] Gevaert, K., Van Damme, P., Ghesquiere, B., Impens, F. et al., A la carte proteomics with an emphasis on gel-free techniques. *Proteomics* 2007, *7*, 2698–2718.

[7] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989, *246*, 64–71.

[8] Hillenkamp, F., Karas, M., Beavis, R. C., Chait, B. T., Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* 1991, *63*, 1193A–1203A.

[9] Barsnes, H., Eidhammer, I., Martens, L., A global analysis of peptide fragmentation variability. *Proteomics* 2011, *11*, 1181–1188.

[10] Makarov, A., Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* 2000, *72*, 1156–1162.

[11] Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 2010, *73*, 2092–2123.

[12] Barsnes, H., Martens, L., Crowdsourcing in proteomics: public resources lead to better experiments. *Amino Acids* 2013, *44*, 1129–1137.

[13] Vaudel, M., Sickmann, A., Martens, L., Current methods for global proteome identification. *Expert Rev. Proteomics* 2012, *9*, 519–532.

[14] Martens, L., Hermjakob, H., Proteomics data validation: why all must provide data. *Mol. Biosyst.* 2007, *3*, 518–522.

[15] Vaudel, M., Sickmann, A., Martens, L., Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics. *Biochim. Biophys. Acta* 2014, *1844*, 12–20.

[16] Everett, L. J., Bierl, C., Master, S. R., Unbiased statistical analysis for multi-stage proteomic search strategies. *J. Proteome Res.* 2010, *9*, 700–707.

[17] Foster, L. J., Interpretation of data underlying the link between colony collapse disorder (CCD) and an invertebrate iridescent virus. *Mol. Cell Proteomics* 2011, *10*, M110.006387.

[18] Editors, Matters of significance. *Nat. Methods* 2013, *10*, 805.

[19] Martin, S. F., Falkenberg, H., Dyrlund, T. F., Khoudoli, G. A. et al., PROTEINCHALLENGE: crowd sourcing in proteomics analysis and software development. *J. Proteomics* 2013, *88*, 41–46.

[20] Editors, In need of an upgrade. *Nat. Biotechnol.* 2013, *31*, 857.

[21] Kinsinger, C. R., Apffel, J., Baker, M., Bian, X. et al., Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam Principles). *J. Proteome Res.* 2012, *11*, 1412–1419.

[22] Martens, L., Vizcaino, J. A., Banks, R., Quality control in proteomics. *Proteomics* 2011, *11*, 1015–1016.

[23] Pichler, P., Mazanek, M., Dusberger, F., Weilnbock, L. et al., SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *J. Proteome Res.* 2012, *11*, 5540–5547.

[24] Lange, V., Picotti, P., Domon, B., Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* 2008, *4*, 222.

[25] UniProt Consortium, The universal protein resource (UniProt). *Nucleic Acids Res.* 2008, *36*, D190–D195.

[26] Chambers, M. C., Maclean, B., Burke, R., Amodei, D. et al., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 2012, *30*, 918–920.

[27] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. et al., Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, *3*, 958–964.

[28] Fenyo, D., Beavis, R. C., A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* 2003, *75*, 768–774.

[29] Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., Martens, L., SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 2011, *11*, 996–999.

[30] Elias, J. E., Gygi, S. P., Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* 2010, *604*, 55–71.

[31] Flicek, P., Ahmed, I., Amode, M. R., Barrell, D. et al., Ensembl 2013. *Nucleic Acids Res.* 2013, *41*, D48–D55.

[32] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, *25*, 25–29.

[33] Villaveces, J. M., Jimenez, R. C., Garcia, L. J., Salazar, G. A. et al., Dasty3, a WEB framework for DAS. *Bioinformatics* 2011, *27*, 2616–2617.

[34] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M. et al., STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2013, *41*, D808–D815.

[35] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M. et al., Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 2009, *37*, D619–D622.

[36] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G. et al., The Protein Data Bank. *Nucleic Acids Res.* 2000, *28*, 235–242.

[37] Vizcaino, J. A., Mueller, M., Hermjakob, H., Martens, L., Charting online OMICS resources: a navigational chart for clinical researchers. *Proteomics Clin. Appl.* 2009, *3*, 18–29.

[38] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537–3545.

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

