Brief Communication

PeptideShaker: Completing the proteomics data cycle.

Marc Vaudel[1,2], Julia M. Burkhart[1], René P. Zahedi[1], Eystein Oveland[2,3,4], Frode S. Berven[2,4,5], Albert Sickmann[1], Lennart Martens[6,7,*] and Harald Barsnes[2,8]

[1] Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany

[2] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

[3] Department of Clinical Medicine, University of Bergen, Bergen, Norway

[4] The KG Jebsen Centre for MS-research, Department of Clinical Medicine, University of Bergen, Bergen, Norway

[5] The Norwegian Multiple Sclerosis Competence Centre, Department of Neurology, Haukeland University Hospital, Bergen, Norway

[6] Department of Medical Protein Research, VIB, Ghent, Belgium

[7] Department of Biochemistry, Ghent University, Ghent, Belgium

[8] Computational Biology Unit, University of Bergen, Norway

[*] Corresponding author

Abstract

Mass spectrometry-based proteomics data, obtained experimentally or *via* public repositories, remains challenging to analyze. This challenge is particularly felt when re-analysing or re-purposing public data sets, endeavours that otherwise hold great potential. Here, we present PeptideShaker, an open source system that greatly simplifies the interpretation and dissemination of proteomics data, and that automates the re-analysis of public data, thus completing the proteomics data cycle for the first time (http://peptide-shaker.googlecode.com).

Main Text

Mass spectrometry-based proteomics has become a common tool in the life sciences to identify and quantify hundreds to thousands of proteins in complex biological samples. Additionally, widespread data sharing *via* publicly accessible repositories such as PRIDE[1] has now become standard practice, aided by robust user-oriented tools for data submission[2-3] and inspection[4], and bolstered by the advent of the ProteomeXchange initiative (http://proteomexchange.org). Importantly, repository data sharing has enabled the first high-profile studies where the re-purposing of publicly available proteomics data has revealed novel insights into biology[5-6].

Yet, proteomics data processing and (re-)analysis currently remain far from routine practice, prompting an urgent and tangible need for a user-friendly, open source tool that empowers any user to perform state-of-the-art proteomics data analysis at any stage in the data life cycle[7-8]. Indeed, in order to maximize the immense value of public proteomics data, their re-use and re-purposing must become straightforward, allowing the completion of the proteomics data cycle.

We therefore developed PeptideShaker, a uniquely powerful proteomics informatics tool that covers the entire proteomics data cycle: from the analysis and interpretation of primary data, *via* data sharing and dissemination, to the straightforward re-analysis of publicly available proteomics data. In accordance with the proven importance of open source software for the advancement of related fields[9-10], PeptideShaker is explicitly designed to work with the combined output of the popular, freely available and open source search engines OMSSA[11] and X!Tandem[12] *via* the SearchGUI[13] interface (**Figure 1**).

In order to deliver state-of-the-art identification performance, PeptideShaker utilizes the target/decoy search strategy[14] for estimating posterior error probabilities, and employs these to unify the peptide to spectrum match (PSM) lists of different search engines (see **Supplementary Material** for details), thus increasing the confidence and sensitivity of hits compared to single search engine processing[15-16]. Thus, PeptideShaker accurately estimates statistical confidence for every peptide and protein, taking into account protein inference issues[17]. Furthermore, rather than only providing false discovery rates (FDR) at the PSM, peptide and protein levels, PeptideShaker also calculates reliable false negative rates (FNR), providing the user with a novel and highly useful interface to filter results according to a FDR *versus* FNR cost-benefit rationale that has so far been missing from proteomics. Indeed, PeptideShaker allows the user to choose the desired balance between specificity and sensitivity, with interactive graphs providing immediate feedback on the actual FDR and FNR for PSMs, peptides and proteins at any chosen threshold. In addition to these identification reliability measures, PeptideShaker also provides both A-score[18] and D-score[19-20] for confident modification site localization. The reliability of all statistical metrics has been validated in detail using complex standards[21] as demonstrated in the **Supplementary Material**.

PeptideShaker's user-oriented interface is divided into nine constantly linked tabs, such that selections in any one tab are automatically propagated to the other tabs. The initial display is the Overview tab (**Figure 2**), showing a single interactive view that encompasses identified proteins, peptides, and spectra. Additional tabs feature specific aspects of a typical proteomics analysis pipeline, including spectrum identification details (with an emphasis on multiple search engine comparison); protein fractionation analysis; modification site localization analysis; protein 3D

structures (with mapped modifications); functional annotation; gene ontology analysis; identification validation; and quality control (see **Supplementary Material**).

Numerous novel visualization approaches are included in PeptideShaker, to allow the user to readily pick up on the meaning and significance of the underlying data. PeptideShaker for instance takes maximum advantage of the identification multiplicity typical of proteomics experiments by visualizing multiple recorded PSMs[22] for a given peptide, or by displaying post-translational modification localization both within and across spectra (see **Supplementary Material**). Moreover, PeptideShaker provides intuitive chromosome, gene, modification and sequence coverage annotation for every identified protein, meeting the goals of the Human Proteome Project[23].

Great care has been taken in PeptideShaker to ensure that results are easily submitted to PRIDE/ProteomeXchange using the built-in PRIDE export; at the time of writing this already resulted in 93 submitted PeptideShaker-derived ProteomeXchange datasets. Additional export options are also available, including a variety of spreadsheet-compatible text files; a descriptive certificate of analysis; high resolution image formats for all displayed graphics; exports to common graph database formats, including Cytoscape (http://cytoscape.org); and export to the widely used Progenesis LC-MS package (http://www.nonlinear.com) for label free quantification. Furthermore, unidentified spectra can be exported specifically for further processing such as *de novo* sequencing. PeptideShaker can also recalibrate spectrum files and, based on the obtained identifications, generate inclusion or exclusion lists which are fully compatible with standard MS instrument controlling software. Interestingly, owing to the distributed efforts of the open source community,

PeptideShaker has recently been incorporated into the Galaxy[24] framework by a third party (https://bitbucket.org/galaxyp/galaxyp-toolshed-peptideshaker).

Besides the ability to submit to PRIDE, PeptideShaker is the first tool to allow the automated and straightforward re-analysis of any public PRIDE dataset of interest. At application start-up, users can choose to run PeptideShaker in so-called Reshake mode that directly provides access to publicly available PRIDE experiments (see **Supplementary Figure 3**). PeptideShaker then extracts the spectra and search settings for the selected PRIDE experiment(s) and presents these to the user before re-searching using SearchGUI. At this stage, the user can modify all parameters, e.g., to add modifications or to specify an updated or different sequence database. With a single click the re-analysis of the data is then started and the results open directly in PeptideShaker upon completion of the search.

Hence, PeptideShaker not only allows processing, analysis and interpretation of freshly acquired data, but also automates the conversion of these data into comprehensively annotated PRIDE files, enabling straightforward submission to PRIDE and ProteomeXchange (http://www.proteomexchange.org). Even more importantly, it is also the first tool that supports direct re-processing and re-purposing of any publicly available dataset in PRIDE. PeptideShaker thus is the first software to fully complete the proteomics data cycle.

PeptideShaker is developed in Java and is open source under the very permissive Apache2 license. It works on Windows, Mac OS X and Linux, and can handle proteomics projects ranging in size from a few thousands to several millions of spectra. In addition to the graphical frontend, PeptideShaker also provides comprehensive command line support for scripting, as detailed on the website. Cross-platform executable binaries, source code, documentation, support, example files and

additional information are available at http://peptide-shaker.googlecode.com. Comprehensive tutorial material including demonstration datasets is available online at http://compomics.com/bioinformatics-for-proteomics.

Acknowledgments

M.V., R.P.Z., J.M.B. and A.S gratefully acknowledges the financial support by the Ministerium für Innovation, Wissenschaft, Forschung und Technologie des Landes Nordrhein-Westfalen and by the Bundesministerium für Bildung und Forschung. L.M. acknowledge the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), the PRIME-XS project, grant agreement number 262067, and the '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. The study was supported by the Kristian Gerhard Jebsen Foundation. The authors would like to thank all PeptideShaker users for thorough testing of the software and for coming up with valuable suggestions for improvements.

Author contributions

M.V. and H.B. did all of the programming and participated in writing the manuscript. R.P.Z., J.M.B., E.O., F.S.B. and A.S. contributed ideas, performed thorough testing, and participated in writing the manuscript. L.M. supervised the programming, provided ideas, and participated in writing the manuscript.

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

Competing financial interests

The authors declare no competing financial interests.

References

1. Martens, L. et al., *Proteomics* **5**, 3537-3545 (2005).
2. Cote, R.G. et al., *Mol Cell Proteomics* **11**, 1682-1689 (2012).
3. Barsnes, H., Vizcaíno, J.A., Eidhammer, I. & Martens, L., *Nat Biotechnol* **27**, 598-599 (2009).
4. Wang, R. et al., *Nat Biotechnol* **30**, 135-137 (2012).
5. Matic, I., Ahel, I. & Hay, R.T., *Nat Methods* **9**, 771-772 (2012).
6. Hahne, H., Moghaddas Gholami, A. & Kuster, B., *Mol Cell Proteomics* **11**, 843-850 (2012).
7. Editors, *Nature Biotechnology* **31**, 857 (2013).
8. Martin, S.F. et al., *J Proteomics* **88**, 41-46 (2013).
9. Reich, M. et al., *Nat Genet* **38**, 500-501 (2006).
10. Gentleman, R.C. et al., *Genome Biol* **5**, R80 (2004).
11. Geer, L.Y. et al., *J Proteome Res* **3**, 958-964 (2004).
12. Craig, R., Cortens, J.C., Fenyo, D. & Beavis, R.C., *J Proteome Res* **5**, 1843-1849 (2006).
13. Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A. & Martens, L., *Proteomics* **11**, 996-999 (2011).
14. Elias, J.E. & Gygi, S.P., *Nat Methods* **4**, 207-214 (2007).
15. Shteynberg, D., Nesvizhskii, A.I., Moritz, R.L. & Deutsch, E.W., *Mol Cell Proteomics* (2013).
16. Nesvizhskii, A.I., *J Proteomics* **73**, 2092-2123 (2010).
17. Nesvizhskii, A.I. & Aebersold, R., *Mol Cell Proteomics* **4**, 1419-1440 (2005).
18. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P., *Nat Biotechnol* **24**, 1285-1292 (2006).
19. Savitski, M.M. et al., *Mol Cell Proteomics* **10**, M110 003830 (2011).
20. Vaudel, M. et al., *Proteomics* **13**, 1036-1041 (2013).
21. Vaudel, M. et al., *J Proteome Res* **11**, 5065-5071 (2012).
22. Barsnes, H., Eidhammer, I. & Martens, L., *Proteomics* **11**, 1181-1188 (2011).
23. Paik, Y.K. et al., *Nat Biotechnol* **30**, 221-223 (2012).
24. Goecks, J., Nekrutenko, A. & Taylor, J., *Genome Biol* **11**, R86 (2010).

Figure legends

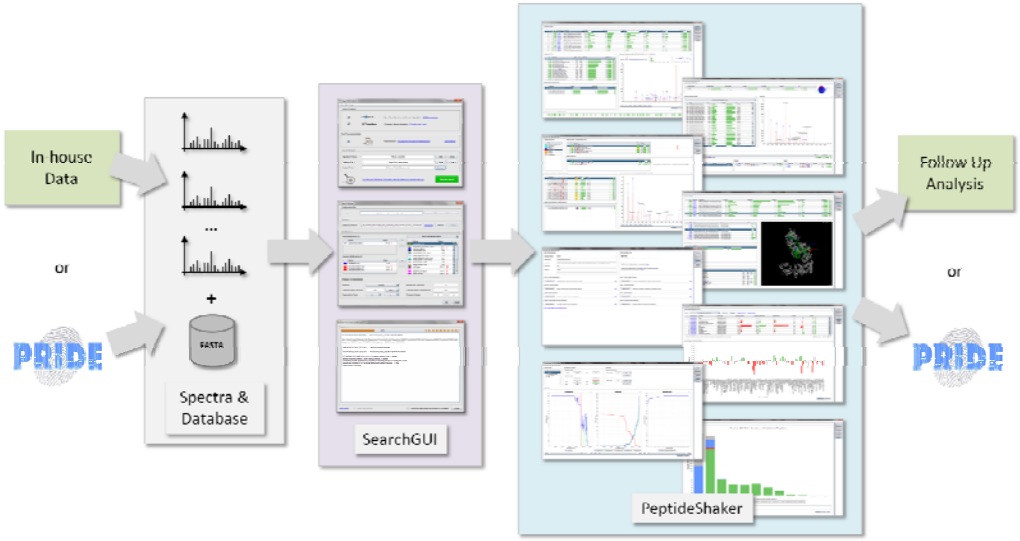


Figure 1: PeptideShaker data analysis cycle. (1) Either in-house generated data or publicly available data from PRIDE can be used as input. (2) The sequence database is selected in SearchGUI and search parameters are configured, allowing the data to be (re-)searched with OMSSA and X!Tandem simultaneously at the press of a button. (3) Search results are then processed, combined, interpreted and displayed in PeptideShaker. (4) After the analysis, the results can either be directly submitted to PRIDE/ProteomeXchange, or they can be exported in ways that support a variety of follow up analyses.

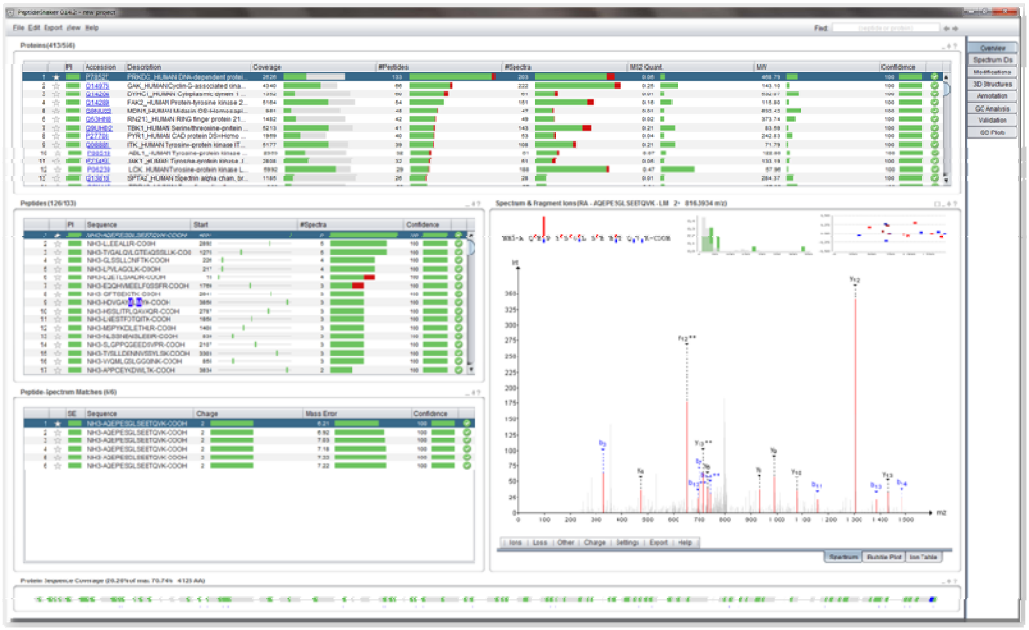
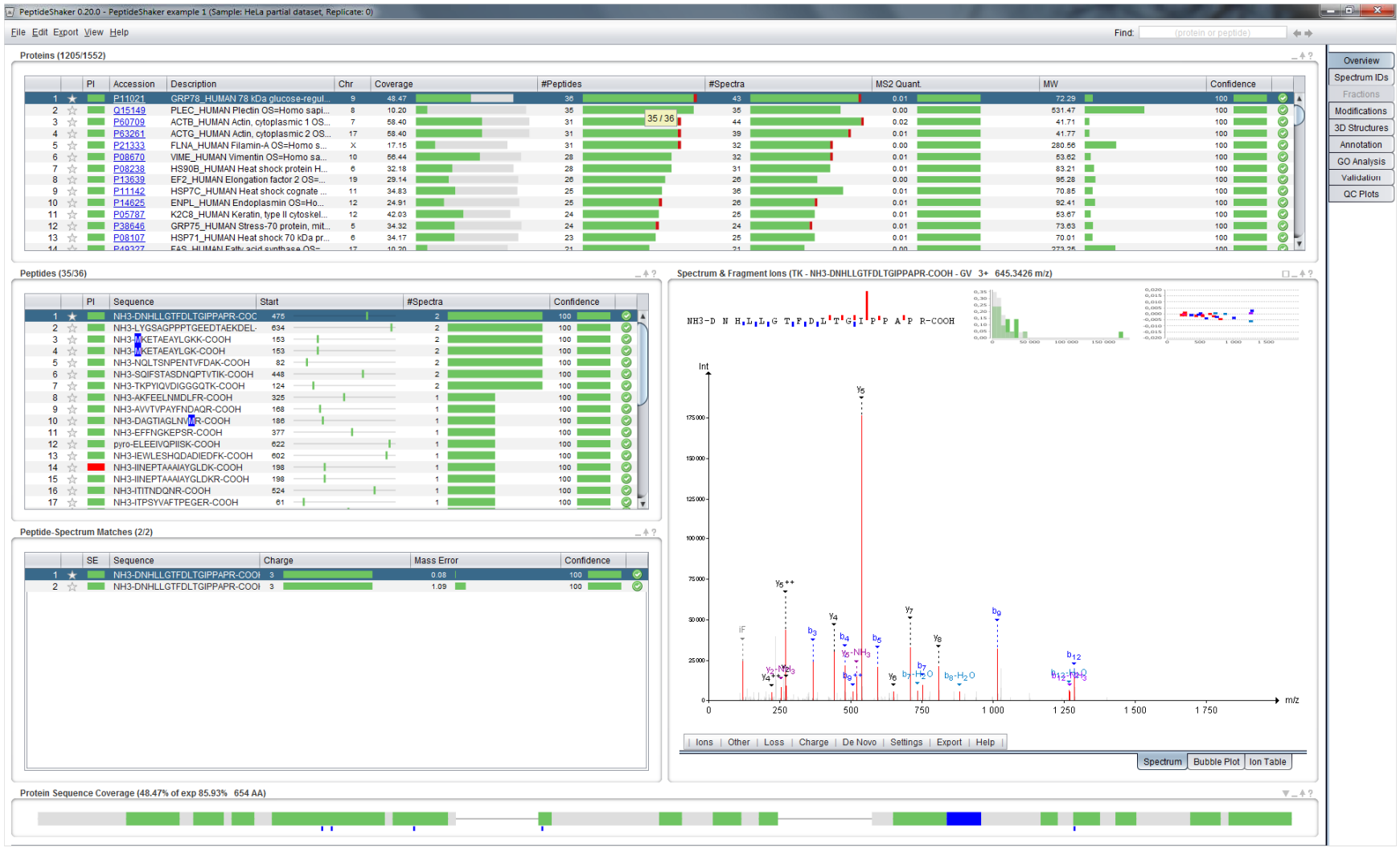


Figure 2: The PeptideShaker Overview tab, showing all proteins in the data set, along with the peptides and peptide to spectrum matches for the selected protein and peptide. The currently selected spectrum match is shown graphically at the lower right, and a visual representation of the protein sequence coverage is shown at the bottom.


...
+
FASTA
Spectra & Database
SearchGUI
PeptideShaker


PeptideShaker 0.20.0 - PeptideShaker example 1 (Sample: HeLa partial dataset, Replicate: 0)
File Edit Export View Help
Find: (protein or peptide)
Proteins (1205/1552)
Peptides (35/36)
Peptide-Spectrum Matches (2/2)
Spectrum & Fragment Ions (TK - NH3-DNHLLGTFDLTGIPPAPR-COOH - GV 3+ 645.3426 m/z)
Protein Sequence Coverage (48.47% of exp 85.93% 654 AA)
Overview
Spectrum IDs
Fractions
Modifications
3D Structures
Annotation
GO Analysis
Validation
QC Plots
Ions | Other | Loss | Charge | De Novo | Settings | Export | Help
Spectrum
Bubble Plot
Ion Table