PeptideShaker enables reanalysis of MS-derived proteomics data sets

To the Editor:

Mass spectrometry (MS)-based proteomics is commonly used to identify and quantify the hundreds to thousands of proteins that are present in complex biological samples. 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[5]. Importantly, repository data sharing has enabled the first high-profile studies in which the repurposing of publicly available proteomics data has revealed new biological insights[6,7].

However, proteomics data processing and (re-)analysis currently remain far from routine practice. There is a pressing need for a user-friendly, open source tool that empowers users to carry out state-of-the-art proteomics data analysis at any stage in the data life cycle[8,9]. To maximize the value of public proteomics data, reuse and repurposing must become straightforward, allowing the completion of the proteomics data cycle. Here we describe PeptideShaker (http://peptide-shaker.googlecode.com), a proteomics informatics software that can be used at any stage in the proteomics data cycle for the analysis and interpretation of primary data, enabling data sharing and dissemination and re-analysis of publicly available proteomics data. Importantly, PeptideShaker can work with the combined output of multiple identification algorithms (**Fig. 1**).

To identify peptides and proteins PeptideShaker uses the target-decoy search strategy[10] to estimate posterior error probabilities and uses these to unify the peptide-to-spectrum match (PSM) lists of different search engines, thus increasing the confidence and sensitivity of hits compared with single-search-engine processing[11,12]. PeptideShaker provides statistical confidence estimates for each peptide and protein, taking into account protein inference issues[13]. Furthermore, as well as providing false discovery rates (FDRs) at the PSM, peptide and protein levels, PeptideShaker calculates reliable false negative rates (FNRs), providing the user with a novel and highly useful interface to filter results according to an FDR-versus-FNR cost-benefit rationale that has so far been absent from proteomics. This filter for specificity and sensitivity includes interactive graphs providing immediate feedback on the values of FDR and FNR for PSMs, peptides and proteins at any chosen threshold.

In addition to these identification reliability measures, PeptideShaker also provides confident modification site inference using the latest localization methods (**Supplementary Note 1**). The reliability of all statistical metrics has been validated in detail using the complex *Pyrococcus furiosus* standard with an entrapment database for FDR accuracy verification[14] (**Supplementary Note 1**).

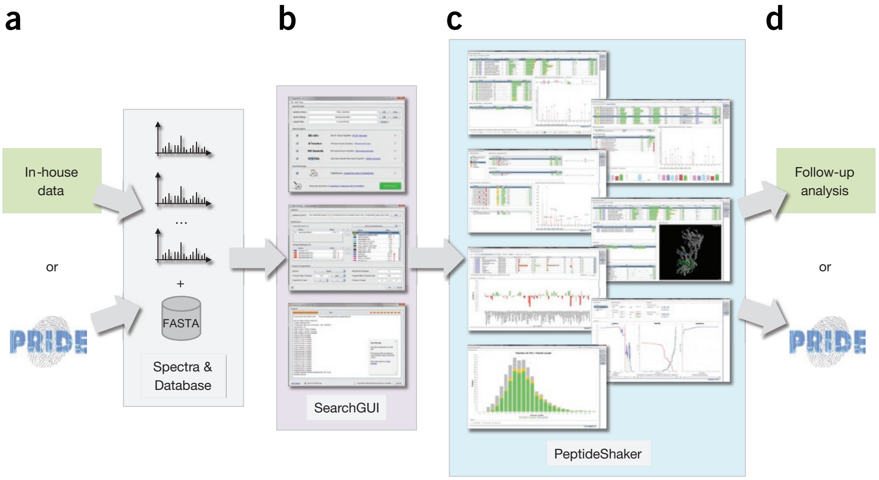
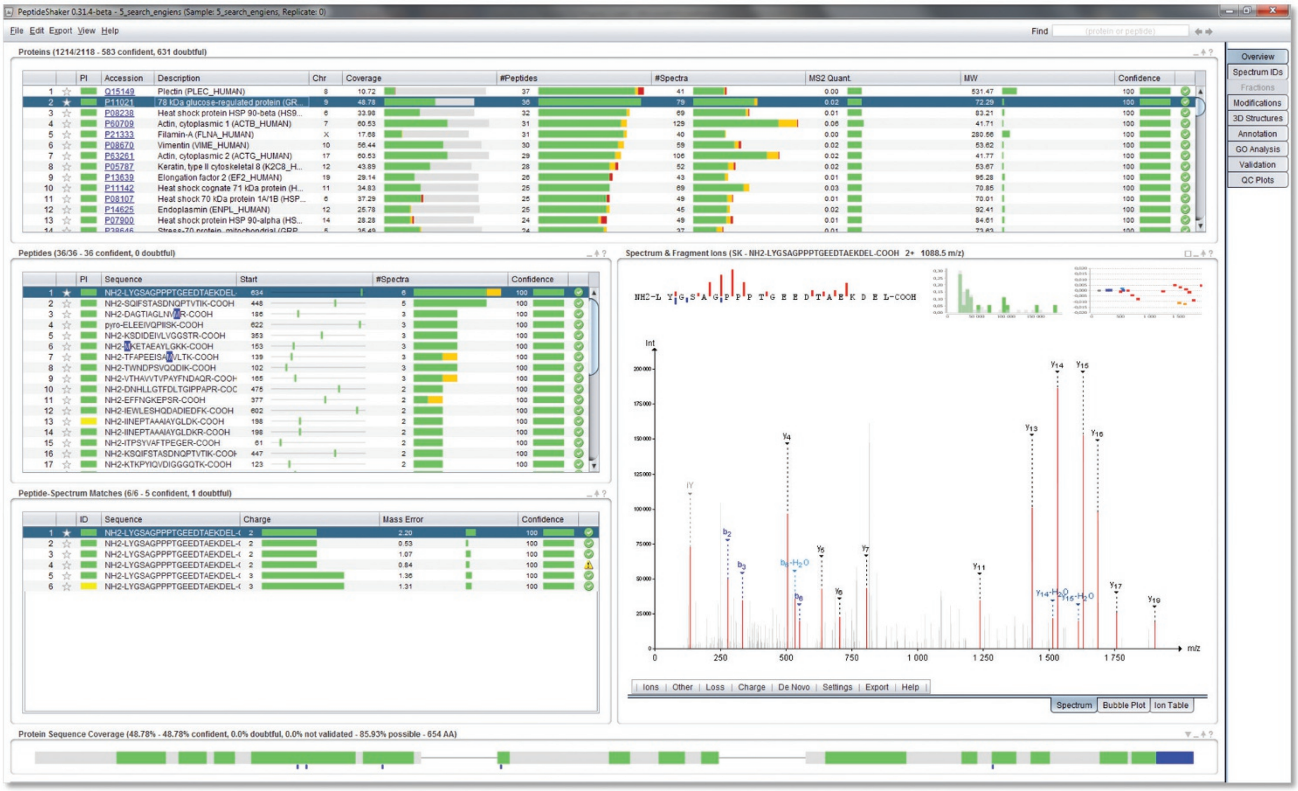


Figure 1 PeptideShaker data-analysis cycle. (**a**) Data generated in-house or publicly available from PRIDE can be used as input. (**b**) The sequence database is selected in SearchGUI, and search parameters are configured, allowing data to be searched with multiple identification software algorithms. (**c**) Search results are processed, combined, interpreted and displayed in PeptideShaker. (**d**) After the analysis, the results can either be submitted directly to PRIDE and ProteomeXchange or exported for follow-up analysis.

PeptideShaker's user-oriented interface is divided into nine linked tabs, such that selections in any one tab are automatically propagated to the other tabs. The initial display is the 'Overview' tab (**Fig. 2**), which shows a single interactive view that includes 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 three-dimensional structures (with mapped modifications), link to functional annotation resources, gene ontology analysis, identification validation and quality control (**Supplementary Note 1**).

Numerous visualizations are provided in PeptideShaker to help the user understand the significance of the underlying data. For example, the software takes advantage of the identification multiplicity typical of proteomics experiments by visualizing multiple recorded PSMs[15] for a given peptide or by displaying posttranslational modification localization both within and across spectra (**Supplementary Note 1**). Moreover, PeptideShaker provides chromosome and gene mapping, modification analysis and intuitive coverage annotation on the sequence of every identified protein, meeting the goals of the Human Proteome Project[16].

Results from PeptideShaker can be readily submitted to PRIDE and ProteomeXchange using the built-in PRIDE and mzIdentML[17] exports; at the time of writing, this has already resulted in 52 publicly available PeptideShaker-derived ProteomeXchange assays. Other export options include 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 Nonlinear (http://www.nonlinear.com) Progenesis liquid chromatography (LC)-MS package for label-free quantification.



PeptideShaker 0.31.4-beta - 5_search_engiens (Sample: 5_search_engiens, Replicate: 0)

Proteins (1214/2118 - 583 confident, 631 doubtful)

		PI	Accession	Description	Chr	Coverage	#Peptides	#Spectra	MS2 Quant.	MW	Confidence	
1	☆	🟩	Q15149	Plectin (PLEC_HUMAN)	8	10.72	37	41	0.00	531.47	100	✓
2	★	🟩	P11021	78 kDa glucose-regulated protein (GR...	9	48.78	36	79	0.02	72.29	100	✓
3	☆	🟩	P08238	Heat shock protein HSP 90-beta (HS9...	6	33.98	32	69	0.01	83.21	100	✓
4	☆	🟩	P60709	Actin, cytoplasmic 1 (ACTB_HUMAN)	7	60.53	31	129	0.06	41.71	100	✓
5	☆	🟩	P21333	Filamin-A (FLNA_HUMAN)	X	17.68	31	40	0.00	280.56	100	✓
6	☆	🟩	P08670	Vimentin (VIME_HUMAN)	10	56.44	30	59	0.02	53.62	100	✓
7	☆	🟩	P63261	Actin, cytoplasmic 2 (ACTG_HUMAN)	17	60.53	29	106	0.02	41.77	100	✓
8	☆	🟩	P05787	Keratin, type II cytoskeletal 8 (K2C8_H...	12	43.89	28	52	0.02	53.67	100	✓
9	☆	🟩	P13639	Elongation factor 2 (EF2_HUMAN)	19	29.14	26	43	0.01	95.28	100	✓
10	☆	🟩	P11142	Heat shock cognate 71 kDa protein (H...	11	34.83	25	69	0.03	70.85	100	✓
11	☆	🟩	P08107	Heat shock 70 kDa protein 1A/1B (HSP...	6	37.29	25	49	0.01	70.01	100	✓
12	☆	🟩	P14625	Endoplasmin (ENPL_HUMAN)	12	25.78	25	45	0.02	92.41	100	✓
13	☆	🟩	P07900	Heat shock protein HSP 90-alpha (HS...	14	28.28	24	49	0.01	84.61	100	✓
14	☆	🟩	P38646	Stress-70 protein, mitochondrial (GRP...	5	35.49	24	37	0.01	73.63	100	✓

Peptides (36/36 - 36 confident, 0 doubtful)

		PI	Sequence	Start	#Spectra	Confidence	
1	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-	634	6	100	✓
2	☆	🟩	NH2-SQIFSTASDNQPTVTIK-COOH	448	5	100	✓
3	☆	🟩	NH2-DAGTIAGLNVMR-COOH	186	3	100	✓
4	☆	🟩	pyro-ELEEIVQPIISK-COOH	622	3	100	✓
5	☆	🟩	NH2-KSDIDEIVLVGGSTR-COOH	353	3	100	✓
6	☆	🟩	NH2-MKETAEAYLGKK-COOH	153	3	100	✓
7	☆	🟩	NH2-TFAPEEISAMVLTK-COOH	139	3	100	✓
8	☆	🟩	NH2-TWNDPSVQQDIK-COOH	102	3	100	✓
9	☆	🟩	NH2-VTHAVVTVPAYFNDAQR-COOH	165	3	100	✓
10	☆	🟩	NH2-DNHLLGTFDLTGIPPAPR-COC	475	2	100	✓
11	☆	🟩	NH2-EFFNGKEPSR-COOH	377	2	100	✓
12	☆	🟩	NH2-IEWLESHQDADIEDFK-COOH	602	2	100	✓
13	☆	🟨	NH2-IINEPTAAAIAYGLDK-COOH	198	2	100	✓
14	☆	🟩	NH2-IINEPTAAAIAYGLDKR-COOH	198	2	100	✓
15	☆	🟩	NH2-ITPSYVAFTPEGER-COOH	61	2	100	✓
16	☆	🟩	NH2-KSQIFSTASDNQPTVTIK-COOH	447	2	100	✓
17	☆	🟩	NH2-KTKPYIQVDIGGGQTK-COOH	123	2	100	✓

Peptide-Spectrum Matches (6/6 - 5 confident, 1 doubtful)

		ID	Sequence	Charge	Mass Error	Confidence	
1	★	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	2.20	100	✓
2	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	0.53	100	✓
3	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	1.07	100	✓
4	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	2	0.84	100	⚠
5	☆	🟩	NH2-LYGSAGPPPTGEEDTAEKDEL-(	3	1.36	100	✓
6	☆	🟨	NH2-LYGSAGPPPTGEEDTAEKDEL-(	3	1.31	100	✓

Spectrum & Fragment Ions (SK - NH2-LYGSAGPPPTGEEDTAEKDEL-COOH 2+ 1088.5 m/z)

Protein Sequence Coverage (48.78% - 48.78% confident, 0.0% doubtful, 0.0% not validated - 85.93% possible - 654 AA)

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. A graphical representation of the currently selected spectrum match (bottom right) and a visual representation of the protein sequence coverage (bottom) are shown.

Furthermore, unidentified spectra can be specifically exported for further processing, such as *de novo* sequencing. PeptideShaker can also recalibrate spectrum files and, on the basis of the obtained identifications, generate inclusion or exclusion lists according to MS instrument software specifications. Interestingly, owing to the distributed efforts of the open-source community, PeptideShaker has recently been incorporated into the Galaxy[18] framework by a third party (https://bitbucket.org/galaxyp/peptideshaker).

PeptideShaker is also the first tool to allow the automated and straightforward reanalysis of any public PRIDE data set of interest. At application start-up, users can choose to run PeptideShaker in 'reshake mode', which provides direct access to publicly available PRIDE experiments (**Supplementary Note 1**). PeptideShaker then extracts the spectra and search settings for the selected PRIDE experiment(s) and presents these to the user before searching again using SearchGUI[19]. At this stage, the user can modify all parameters—for example, to add modifications or to specify an updated or different sequence database. With a single click the reanalysis of the data can be started and the results open directly in PeptideShaker upon completion of the search.

PeptideShaker is developed in Java and is open source under the Apache2 license, which is very permissive. It works on Windows, OS X and Linux, and can handle proteomics projects of sizes ranging from a few thousand spectra to several million spectra. In addition to the graphical front end, PeptideShaker also provides comprehensive command line support for scripting, cross-platform executable binaries, source code, documentation, support, example files and additional information on its website (http://peptide-shaker.googlecode.com). Additionally, a snapshot of the source code used for **Supplementary Note 1** is available in **Supplementary Software**. Comprehensive tutorial material including demonstration data sets is available online (http://compomics.com/bioinformatics-for-proteomics).

PeptideShaker not only allows processing, analysis and interpretation of freshly acquired data, but also automates the conversion of these data into comprehensively annotated mzIdentML files, enabling straightforward submission to PRIDE and ProteomeXchange. It also supports direct reprocessing and repurposing of any publicly available data set in PRIDE. The availability of this software should facilitate completion of the proteomics data cycle.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3109).

ACKNOWLEDGMENTS
M.V., R.P.Z., J.M.B. and A.S. gratefully acknowledge the financial support by the Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen and by the Bundesministerium für Bildung und Forschung. 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 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 the PRIDE team for help regarding the use of the PRIDE archive web services and for validation of the PRIDE export formats. Finally, our thanks to 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 testing and participated in writing the manuscript. L.M. supervised the programming, provided ideas and participated in writing the manuscript. All authors participated in the preparation of the manuscript.

Marc Vaudel[1,2], Julia M Burkhart[1], René P Zahedi[1], Eystein Oveland[2–4], Frode S Berven[2,4,5], Albert Sickmann[1], Lennart Martens[6,7] & 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, Bergen, Norway.

e-mail: lennart.martens@vib-ugent.be

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Martens, L. *et al. Proteomics* **5**, 3537–3545 (2005).
2. Côté, 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. Vizcaíno, J.A. *et al. Nat. Biotechnol.* **32**, 223–226 (2014).
6. Matic, I., Ahel, I. & Hay, R.T. *Nat. Methods* **9**, 771–772 (2012).
7. Hahne, H., Moghaddas Gholami, A. & Kuster, B. *Mol. Cell. Proteomics* **11**, 843–850 (2012).
8. Editors. *Nat. Biotechnol.* **31**, 857 (2013).
9. Martin, S.F. *et al. J. Proteomics* **88**, 41–46 (2013).
10. Elias, J.E. & Gygi, S.P. *Nat. Methods* **4**, 207–214 (2007).
11. Shteynberg, D., Nesvizhskii, A.I., Moritz, R.L. & Deutsch, E.W. *Mol. Cell. Proteomics* **12**, 2383–2393 (2013).
12. Nesvizhskii, A.I. *J. Proteomics* **73**, 2092–2123 (2010).
13. Nesvizhskii, A.I. & Aebersold, R. *Mol. Cell Proteomics* **4**, 1419–1440 (2005).
14. Vaudel, M. *et al. J. Proteome Res.* **11**, 5065–5071 (2012).
15. Barsnes, H., Eidhammer, I. & Martens, L. *Proteomics* **11**, 1181–1188 (2011).
16. Paik, Y.K. *et al. Nat. Biotechnol.* **30**, 221–223 (2012).
17. Jones, A.R. *et al. Mol. Cell. Proteomics* **11**, M111.014381 (2012).
18. Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy, T. *Genome Biol.* **11**, R86 (2010).
19. Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A. & Martens, L. *Proteomics* **11**, 996–999 (2011).

Intellectual property issues and synthetic biology standards

To the Editor:

We commend Galdzicki *et al.*[1] on their development of the Synthetic Biology Open Language (SBOL), as described in the June 2014 issue. The technical standards being developed by such groups are much-needed mechanisms for enabling collaboration by diverse research organizations, for accelerating scientific progress in synthetic biology and for the eventual commercialization of resulting technologies. We were surprised, however, to see no mention of intellectual property (IP) or other legal issues pertaining to the use and deployment of SBOL or other synthetic biology standards. Although we understand that the focus of Galdzicki's article was primarily technical, legal issues today are inextricably entwined with standards that enable product and process interoperability. Certainly, the authors acknowledged the prevalence of standards in "every engineering field," but they overlooked the legal issues that have bedeviled standards developers over the past two decades in industries ranging from wireless telecommunications to computer networking to semiconductor memory.

Patents, in particular, have posed challenges to standards developers and implementers in the information and communications technology (ICT) sector. Disputes can arise when a participant in the standards-development process holds patents covering the standard and then seeks to charge unanticipated royalties on products that conform to the standard. To address these risks, the rules of most ICT standards development organizations require that such patents be disclosed and/or licensed to implementers of their standards[2]. These rules may require that licenses be either royalty-free or royalty-bearing on terms that are "fair, reasonable and nondiscriminatory" (FRAND). Although significant disagreement remains regarding the interpretation of FRAND and other standards-related commitments[3,4], a vast amount of standardization occurs in the ICT sector under such policies.

nature
biotechnology
THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY
VOLUME 32 NUMBER 6 JUNE 2014
www.nature.com/naturebiotechnology
High-throughput RNA-protein affinity mapping
Definitive hematopoiesis in a dish
CRISPR-Cas correction of disease mutation

Although the biological sciences have remained largely unaffected by the standards litigation that has plagued the ICT sector[4], there is no unique structural feature that immunizes biomedical and other bioscience standards from such disputes[5]. It is well-known that universities and other research institutions have acquired substantial patent assets, as have private sector biotech and pharmaceutical companies[6]. Patents have pervaded synthetic biology from the outset[7], and they are playing an increasingly important role in its development[8]. As such, it is not hard to envision a scenario in which patents may cover one or more key aspects of synthetic biology standards, such as SBOL, immediately confronting the field with the prospect of unanticipated royalty payments. Indeed, it is possible that biotech "patent assertion entities" may emerge to enforce synthetic biology patent portfolios.

We consider it vital for standards-setting initiatives in synthetic biology, like SBOL, to address several questions. Has any effort been made to determine whether such patents exist? Are participants in SBOL standardization activities required or encouraged to disclose patents that they may obtain on standardized technologies? If such patents are obtained, what rules will govern the terms on which such patents will be made available to the community?

We are concerned that the synthetic biology scientific community is not prepared for the eventual emergence of patents affecting synthetic biology standards, and that this young community may thus be vulnerable to opportunism by unscrupulous actors. But all is not lost. Indeed, the synthetic biology community has enthusiastically adopted an ethos of contributing standardized biological components, such as BioBrick parts, to the community free from patent encumbrances