Journal of proteome research

pubs.acs.org/jpr

Technical Note

PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data

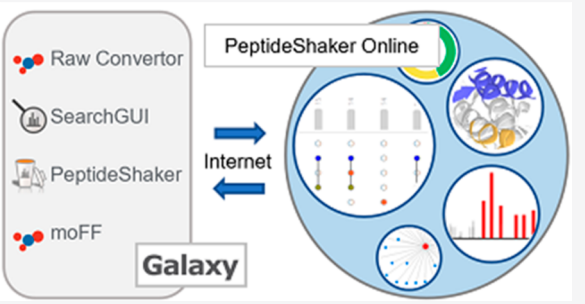
Yehia Mokhtar Farag,* Carlos Horro, Marc Vaudel, and Harald Barsnes

Cite This: https://doi.org/10.1021/acs.jproteome.1c00678

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Mass spectrometry-based proteomics is a high-throughput technology generating ever-larger amounts of data per project. However, storing, processing, and interpreting these data can be a challenge. A key element in simplifying this process is the development of interactive frameworks focusing on visualization that can greatly simplify both the interpretation of data and the generation of new knowledge. Here we present PeptideShaker Online, a user-friendly web-based framework for the identification of mass spectrometry-based proteomics data, from raw file conversion to interactive visualization of the resulting data. Storage and processing of the data are performed via the versatile Galaxy platform (through SearchGUI, PeptideShaker, and moFF), while the interaction with the results happens via a locally installed web server, thus enabling researchers to process and interpret their own data without requiring advanced bioinformatics skills or direct access to compute-intensive infrastructures. The source code, additional documentation, and a fully functional demo is available at https://github.com/barsnes-group/peptide-shaker-online.



KEYWORDS: *mass spectrometry, data processing, Galaxy, interaction, visualization*

INTRODUCTION

Mass spectrometry-based proteomics generates large amounts of data,[1] and it is essential that the data can be processed and analyzed in such a way that the researcher generating the data can interpret its biological meaning correctly. In addition to biological knowledge, this often requires direct access to significant computational resources and advanced computational skills. The overall challenge can be split into three main categories: (i) access to computational resources; (ii) availability of user-friendly bioinformatics software; and (iii) having the biological understanding to translate the data into useful knowledge.

The first category can be addressed by high-performance computing environments that provide the required resources through powerful servers instead of the more limited personal computers,[2] while at the same time making the stored data more portable and accessible; i.e., there is no need to download or move the data.[3] Adding interactive visualization to such setups can help with the second category of the need for user-friendly bioinformatics software, and can also play a key role in the data processing and simplify the interpretation of the results.[4] Interactive visualization can furthermore greatly reduce the complexity of interpretation by providing direct interaction with the data and by dividing it into distinct levels, thus enabling the biological researcher to focus on interpreting the data and extracting biological knowledge.[5]

One way to port bioinformatic pipelines to remote servers is to take advantage of Galaxy, a web-based scientific analysis platform including more than 5500 specialized tools, in addition to workflow support and data storage management, thus providing the required infrastructure for large-scale proteomics data analysis.[6] Galaxy is however limited when it comes to advanced interactive visualization of the results and may not be straightforward to use for nonprogrammers.

As a response to these challenges, we here present PeptideShaker Online, a user-friendly web-based framework for the identification of mass spectrometry-based proteomics data, from raw file conversion to interactive visualization of the search results.

METHODS

PeptideShaker Online consists of two main components; a Galaxy-based backend where the data is stored and the search and data processing are performed,[7] and a locally installed web-based frontend supporting SearchGUI[8] search and interactive visualization of PeptideShaker[9] projects. The data

Received: August 20, 2021

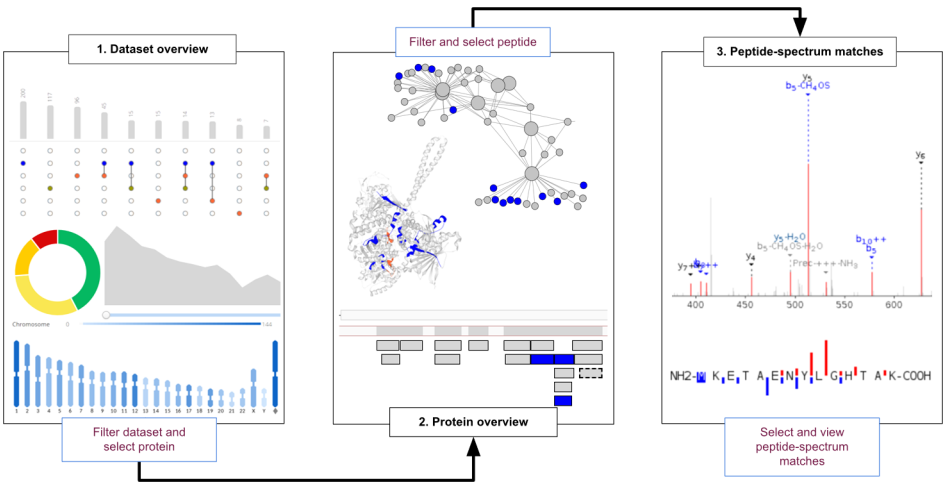


Figure 1. Overview of the three data set levels and the filter-and-select approach. The user filters the data set at the Data Set Overview level to find and select a protein (group) for closer inspection at the Protein Overview level. Here the user selects a peptide from the protein−peptide network or protein coverage table in order to see the peptide and spectrum details at the Peptide-Spectrum Matches level.

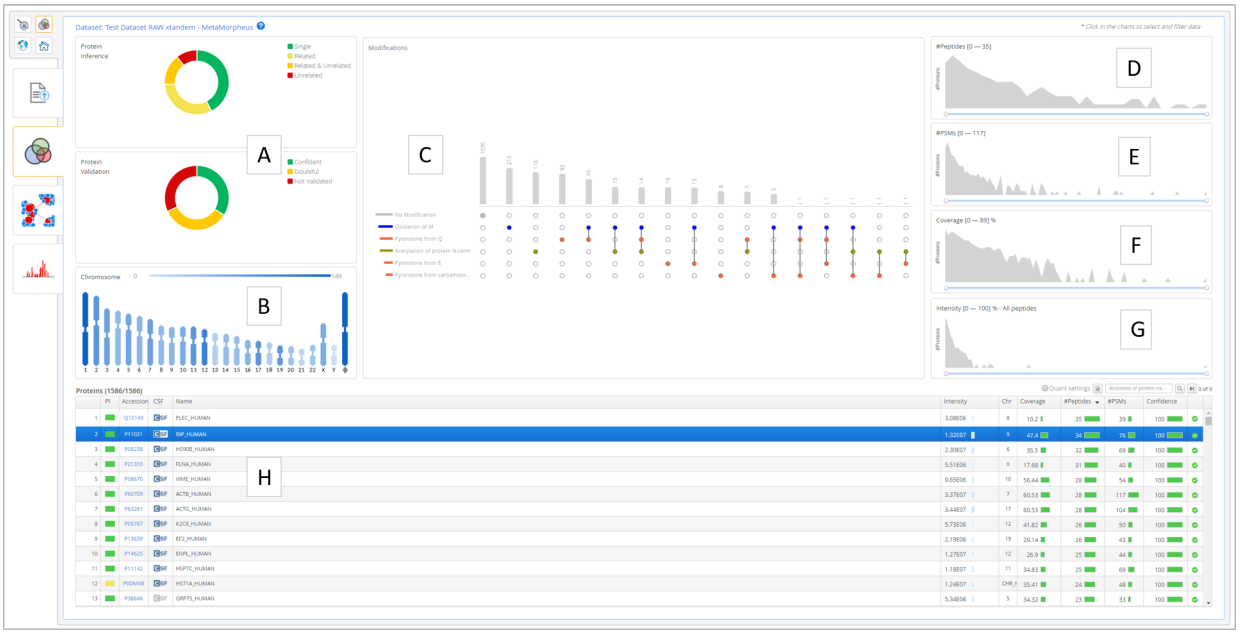


Figure 2. Data set overview. (A) Protein inference and validation filters, (B) chromosome filter, (C) post-translational modifications filter, (D) number of peptides filter, (E) number of peptide-spectrum filter, (F) protein coverage filter, (G) protein intensity filter, and (H) protein table with the currently filtered results.

processing is done via the Galaxy platform using (i) ThermoRawFileParser[10] for converting Thermo raw files into mzML[11] or mgf; (ii) SearchGUI for protein identification based on ten proteomics search and de novo engines, namely OMSSA,[12] X! Tandem,[13] MyriMatch,[14] MS Amanda,[15] MS-GF+,[16] Comet,[17] Tide,[18] MetaMorpheus,[19] DirectTag,[20] and Novor;[21] (iii) PeptideShaker for interpretation of the peptide identification data from SearchGUI; and finally (iv) moFF for extracting MS1 intensities from the mass spectra.[22] Spectrum input is supported as either mgf or mzML for identification, and Thermo raw files for both identification and quantification.

Vaadin 7.26 (https://vaadin.com) and Java 8 are used for the frontend implementation, and Tomcat server version 9 (https://tomcat.apache.org) is used to host the demo web application. Reactome[23] is used for the proteoform network data, Lite-Mol[24] for the protein 3D structures, compomics-utilities 5.0.15[25] to produce the main spectrum charts, and Jersey 2.34 (https://eclipse-ee4j.github.io/jersey) for managing the connections between Galaxy and PeptideShaker Online.

For a complete list of libraries, the full source code, additional documentation, and step-by-step instructions on how to deploy PeptideShaker Online on your own web server, please see https://github.com/barsnes-group/peptide-shaker-online.

RESULTS

As a web-based interactive proteomics framework, PeptideShaker Online can be deployed in proteomics laboratories and facilities. It aims to simplify the mass spectrometry-based proteomics data identification through providing the users with an intuitive easy-to-use interface, thus removing the need for

https://doi.org/10.1021/acs.jproteome.1c00678
J. Proteome Res. XXXX, XXX, XXX−XXX

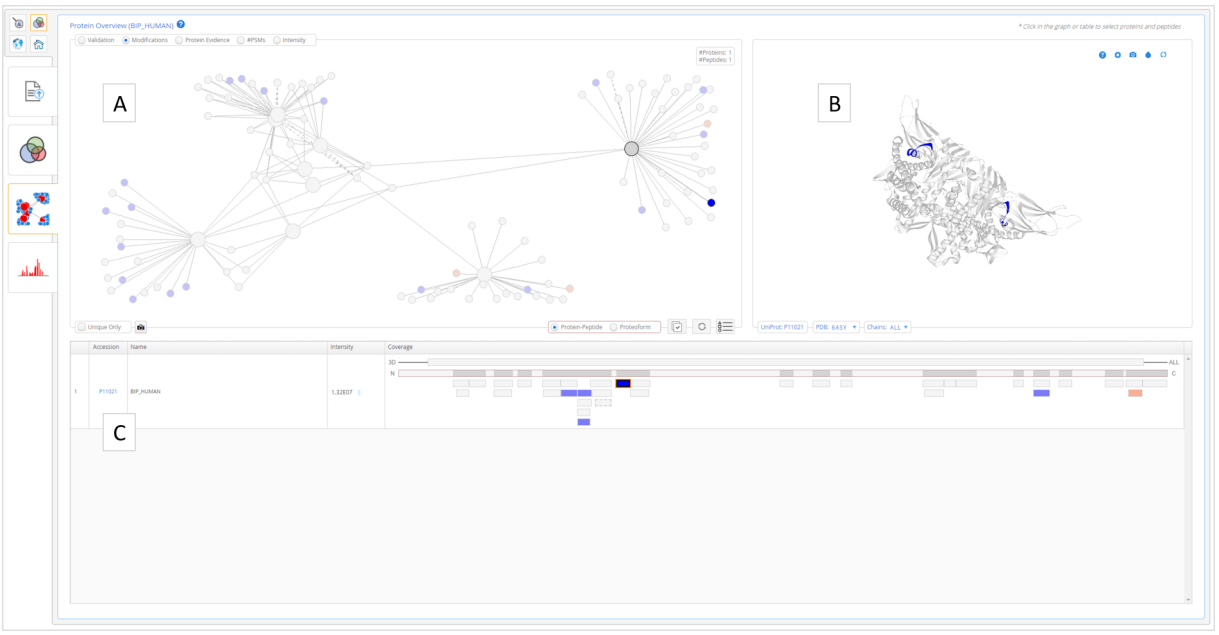


Figure 3. Protein overview. (A) Protein-peptide (or proteoform) network, (B) protein 3D structure, and (C) protein coverage table.

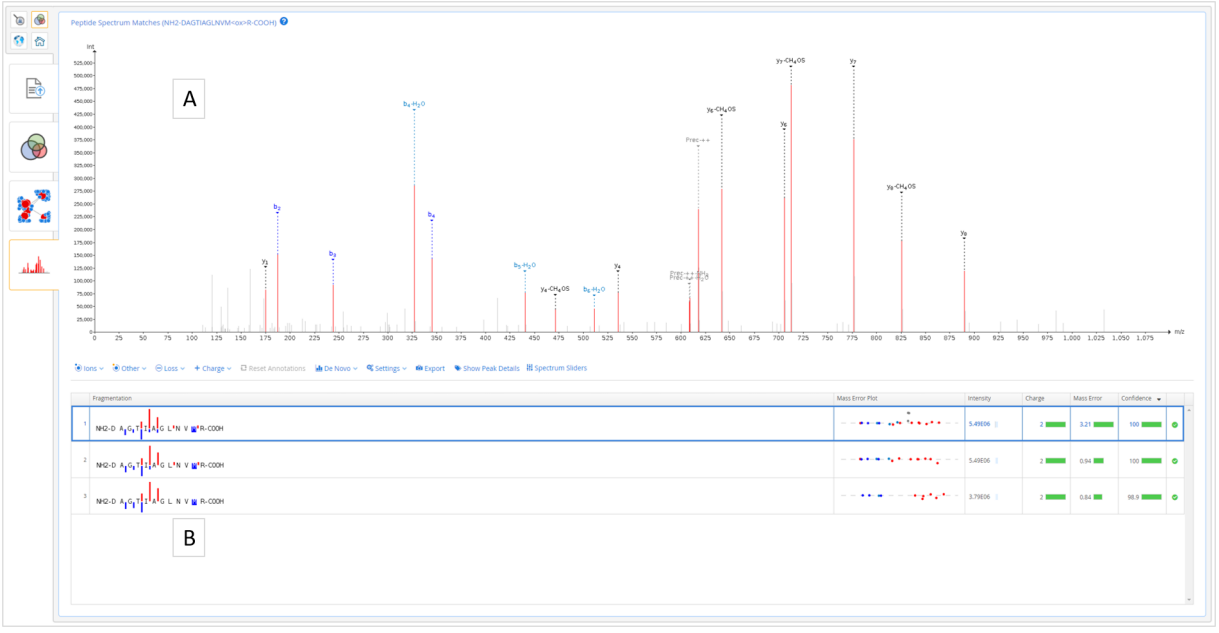


Figure 4. Peptide-spectrum matches. (A) Interactive spectrum viewer, and (B) peptide-spectrum matches table.

advanced programming skills or the use of command lines. Additionally, the result interpretation is enhanced through supporting coordinated interactive visualization combined with splitting the data into three separate levels: (i) Data Set Overview, (ii) Protein Overview, and (iii) Peptide-Spectrum Matches. Thus, instead of dealing with the data as one large and complex unit, the user can inspect the data at three separate levels through a filter-and-select approach (Figure 1).

The top level is the Data Set Overview (Figure 2), providing a summary of the currently selected data set and searching for the proteins of interest through interactive filtering of the protein identifications, including: (i) post-translational modifications; (ii) protein inference and validation categories; (iii) number of peptides and peptide-spectrum matches, in addition to protein coverage and quantification intensity; (iv) chromosome mapping; and finally, (v) a protein table showing the currently filtered proteins.

Next is the Protein Overview (Figure 3), showing the details for the currently selected protein, including: (i) a protein-peptide network (with related protein groups and peptides) or a proteoform network for the related proteoforms interactions; (ii) the protein 3D structure; and (iii) the protein coverage. For all three visualizations, the peptides are color-coded according to a user-selected property, e.g., the number of peptide-spectrum matches or the post-translational modification type.

Finally, there is the Peptide-Spectrum Matches level (Figure 4), including the annotated spectrum matches for the selected peptide. Notably, the spectrum viewer is fully interactive and supports both manual and automatic de novo sequencing in

https://doi.org/10.1021/acs.jproteome.1c00678
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 1. Details for the Example Datasets Included in the Demo Version of PeptideShaker Online

name	spectrum format	FASTA	search parameters	search engines	id	quant
Sample 1	mzML	The reviewed sequences for human from UniProt[26]	Modifications: Oxidation of M, (variable) and Carbamidomethylation of C (fixed). Enzyme: Trypsin with max two missed cleavages. Tolerances: 10 ppm (precursors) and 0.02 Da (fragment ions).	X! Tandem, MS-GF+, OMSSA, Comet, Tide, MyriMatch, MetaMorpheus, MS Amanda, DirectTag and Novor	yes	
Sample 2	raw	The reviewed sequences for human from UniProt	Modifications: Oxidation of M, (variable) and Carbamidomethylation of C (fixed). Enzyme: Trypsin with max two missed cleavages. Tolerances: 10 ppm (precursors) and 0.02 Da (fragment ions).	X! Tandem, MS-GF+, OMSSA, Comet, Tide, MyriMatch, MetaMorpheus, MS Amanda, DirectTag and Novor	yes	yes

addition to customizable peak annotation. This level also includes a peptide-spectrum matches table with sequence fragmentation charts and mass error plots.

Furthermore, PeptideShaker Online makes it possible for users to share their own processed results using project-specific links. Besides saving time, this feature makes the data more secure and portable given that there is no transfer of the underlying data files between the users. The users can also export the data directly as either an Excel spreadsheet or images. Finally, PeptideShaker Online also supports the uploading and visualization of locally processed data files in tab-delimited file formats, making it possible to use the framework without having to reprocess the data with the full pipeline, for example, when the desired spectrum files are not available.

To test the framework and its main features, a fully functional demo is available, which includes two processed data sets (Table 1) and supporting new searches based on the example data provided. Due to local resource limitations, the maximum number of concurrent users for the demo is set to five. Note that the demo uses a public Galaxy user key by default. When installing their own version of PeptideShaker Online, users should rather use personal Galaxy API keys to control access to the data. For information about how to set up your own PeptideShaker Online web server, please visit the project's GitHub page: https://github.com/barsnes-group/peptide-shaker-online.

CONCLUSION

In summary, PeptideShaker Online is a user-friendly web-based framework for the identification of mass spectrometry-based proteomics data, from raw file conversion to interactive visualization of the results. The framework is easily expandable by either including additional tools from the Galaxy platform or introducing new data visualization levels in the web-based frontend. PeptideShaker Online makes the identification of proteomics data more accessible to researchers lacking advanced computational skills, thus moving the data interpretation closer to the biologists in charge of generating the data. Furthermore, the coordinated interactive visualizations combined with splitting the data into distinct levels allows for intuitive data exploration and thus contributes to a better understanding of proteomics data and its inherent complexity.

AUTHOR INFORMATION

Corresponding Author

Yehia Mokhtar Farag − *Proteomics Unit, Department of Biomedicine, University of Bergen, 5020 Bergen, Norway; Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway;* orcid.org/0000-0003-2573-7538; Email: yehia.farag@uib.no

Authors

Carlos Horro − *Proteomics Unit, Department of Biomedicine, University of Bergen, 5020 Bergen, Norway; Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway*

Marc Vaudel − *Department of Clinical Sciences, University of Bergen, 5020 Bergen, Norway;* orcid.org/0000-0003-1179-9578

https://doi.org/10.1021/acs.jproteome.1c00678
J. Proteome Res. XXXX, XXX, XXX−XXX

Harald Barsnes − *Proteomics Unit, Department of Biomedicine, University of Bergen, 5020 Bergen, Norway; Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00678

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

YMF, CH, and HB are supported by the Bergen Research Foundation (BFS2016REK02). MV is supported by the Research Council of Norway (Project #301178).

■ REFERENCES

(1) Noble, W. S.; MacCoss, M. J. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.* **2012**, *8* (1), e1002296.
(2) Farag, Y.; Berven, F. S.; Jonassen, I.; Petersen, K.; Barsnes, H. Distributed and interactive visual analysis of omics data. *J. Proteomics* **2015**, *129*, 78−82.
(3) Grimm, D. J. The Dark Data Quandary. *Am. Univ. Law Rev.* **2019**, *68* (3), 761−821.
(4) Vaudel, M.; Venne, A. S.; Berven, F. S.; Zahedi, R. P.; Martens, L.; Barsnes, H. Shedding light on black boxes in protein identification. *Proteomics* **2014**, *14* (9), 1001−5.
(5) Oveland, E.; Muth, T.; Rapp, E.; Martens, L.; Berven, F. S.; Barsnes, H. Viewing the proteome: how to visualize proteomics data? *Proteomics* **2015**, *15* (8), 1341−55.
(6) Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Cech, M.; Chilton, J.; Clements, D.; Coraor, N.; Gruning, B. A.; Guerler, A.; Hillman-Jackson, J.; Hiltemann, S.; Jalili, V.; Rasche, H.; Soranzo, N.; Goecks, J.; Taylor, J.; Nekrutenko, A.; Blankenberg, D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46* (W1), W537−W544.
(7) Jalili, V.; Afgan, E.; Gu, Q.; Clements, D.; Blankenberg, D.; Goecks, J.; Taylor, J.; Nekrutenko, A. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **2020**, *48* (W1), W395−W402.
(8) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**, *17* (7), 2552−2555.
(9) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33* (1), 22−4.
(10) Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J. Proteome Res.* **2020**, *19* (1), 537−542.
(11) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W. mzML−a community standard for mass spectrometry data. *Mol. Cell Proteomics* **2011**, *10* (1), R110.000133.
(12) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958−64.
(13) 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* (4), 768−74.
(14) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6* (2), 654−61.
(15) Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res.* **2014**, *13* (8), 3679−84.
(16) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(17) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22−4.
(18) Diament, B. J.; Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **2011**, *10* (9), 3871−9.
(19) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17* (5), 1844−1851.
(20) Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J.; Chambers, M. C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* **2008**, *7* (9), 3838−46.
(21) Ma, B. Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1885−94.
(22) Argentini, A.; Goeminne, L. J.; Verheggen, K.; Hulstaert, N.; Staes, A.; Clement, L.; Martens, L. moFF: a robust and automated approach to extract peptide ion intensities. *Nat. Methods* **2016**, *13* (12), 964−966.
(23) Griss, J.; Viteri, G.; Sidiropoulos, K.; Nguyen, V.; Fabregat, A.; Hermjakob, H. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell Proteomics* **2020**, *19* (12), 2115−2125.
(24) Sehnal, D.; Deshpande, M.; Varekova, R. S.; Mir, S.; Berka, K.; Midlik, A.; Pravda, L.; Velankar, S.; Koca, J. LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods* **2017**, *14* (12), 1121−1122.
(25) Barsnes, H.; Vaudel, M.; Colaert, N.; Helsens, K.; Sickmann, A.; Berven, F. S.; Martens, L. compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinf.* **2011**, *12*, 70.
(26) UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489.