Journal of proteome research

Technical Note

Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for seamless end-to-end metaproteomics data analysis

Tim Van Den Bossche, Pieter Verschaffelt, Kay Schallert, Harald Barsnes, Peter Dawyndt, Dirk Benndorf, Bernhard Y. Renard, Bart Mesuere, Lennart Martens, and Thilo Muth

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.0c00136 • Publication Date (Web): 20 May 2020

Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for seamless end-to-end metaproteomics data analysis

Tim Van Den Bossche [1,2], Pieter Verschaffelt [1,3], Kay Schallert [4,5], Harald Barsnes [6,7], Peter Dawyndt [3], Dirk Benndorf [4,5,8], Bernhard Y. Renard [9,10], Bart Mesuere [1,2,3], Lennart Martens[1,2], Thilo Muth [9,11]

[1] VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium

[2] Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium

[3] Department of Applied Mathematics, Computer Science, and Statistics, Ghent University, 9000 Ghent, Belgium

[4] Bioprocess Engineering, Faculty for Process and Systems Engineering, Otto von Guericke University, Magdeburg, Germany

[5] Microbiology, Department of Applied Biosciences and Process Technology, Anhalt University of Applied Sciences, Köthen, Germany

[6] Proteomics Unit (PROBE), Department of Biomedicine, University of Bergen, Norway

[7] Computational Biology Unit (CBU), Department of Informatics, University of Bergen, Norway

[8] Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

[9] Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany

[10] Hasso-Plattner-Institute, Faculty of Digital Engineering, University of Potsdam, 14482 Potsdam, Germany

[11] eScience Division (S.3), Federal Institute for Materials Research and Testing, 12205 Berlin, Germany

Corresponding Author

*Prof. Dr. Lennart Martens, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. E-mail: lennart.martens@vib-ugent.be , Tel: +3292649358

ABSTRACT: Although metaproteomics, the study of the collective proteome of microbial communities, has become increasingly powerful and popular over the past few years, the field has lagged behind on the availability of user-friendly, end-to-end pipelines for data analysis. We therefore describe the connection from two commonly used metaproteomics data processing tools in the field, MetaProteomeAnalyzer and PeptideShaker, to Unipept for downstream analysis. Through these connections, direct end-to-end pipelines are built from database searching to taxonomic and functional annotation.

KEYWORDS: metaproteomics, software, pipelines

INTRODUCTION

In the past few years, the study of microbial communities, or microbiomes, has become an important field, with a wide variety of applications in medicine, ecology, wastewater treatment, and biogas plants, amongst others[1,2]. The growing popularity of microbiome studies has been driven by technological and methodological advances in the respective omics fields[3]. Indeed, the most commonly used methods to study microbiomes are metagenomics and metatranscriptomics, which describe the genome and transcriptome of the microbial community, respectively. These methods provide insights into taxonomic composition and functional potential of the microbiome. However, to do an in-depth study of the actual function of the microbiome, and to gain insights into the host-environment interaction, it is important to have information on the protein level[4]. This information can be obtained through metaproteomics, the study of the collective proteome of microbial communities[5]. Metaproteomics thus provides important complementary information to metagenomics and metatranscriptomics analyses[6].

A typical metaproteomics workflow is very similar to shotgun proteomics and consists of sample preparation, protein extraction, tryptic digest and peptide analysis using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)[7]. The resulting MS/MS spectra are analyzed using database search engines, resulting in peptide-to-spectrum matches (PSMs) that can then be mapped back to proteins, taxa and functions. These database search engines can be used separately, or combined to increase the number of PSMs and proteins[8,9]. Combined search strategies are usually encapsulated in user-friendly software such as the well-established MetaProteomeAnalyzer (MPA)[10] and the SearchGUI/PeptideShaker pipeline[11–13].

MPA is built to manage, process, and interpret complex metaproteomics data, and currently has two versions available, MPA Portable version 2.0[14] and MPA Server version 3.0 (http://www.mpa.ovgu.de)[15]. MPA Portable is optimized to run on desktop computers or compute cluster environments, either with a user interface or as command line tool. MPA Server is optimized to still run with larger datasets, adding optimized database and memory management.

PeptideShaker is also meant for the analysis and interpretation of (meta)proteomics data, enabling data sharing and dissemination and re-analysis of publicly available (meta)proteomics data in the ProteomeXchange Consortium[16] partner PRIDE[17].

Downstream taxonomic and functional analysis of these peptide identifications is provided by Unipept (https://unipept.ugent.be), a web application that features highly interactive data visualizations for the comprehensive downstream analysis of identified peptides[18,19]. Importantly, Unipept also performs a metaproteomics-specific type of protein matching. For each identified tryptic peptide, Unipept calculates the lowest common ancestor (LCA) based on the mapping of known tryptic peptides from UniProtKB[20] to the complete taxonomic lineage of the NCBI Taxonomy Database[21].

However, even though MPA, PeptideShaker and Unipept are well-established and user-friendly tools in the field of metaproteomics, connecting output from MPA or PeptideShaker to Unipept has so far relied on a manual export and import operation by the user, a process that had to be repeated each time for any desired false discovery rate (FDR) level. This arbitrary process does not only requires valuable time, but is prone to errors made by the users. We therefore implemented an intuitive and automated connection from both MPA and PeptideShaker to Unipept, allowing

identified peptides (filtered on the chosen FDR threshold within MPA or PeptideShaker) to be uploaded directly to Unipept.

MATERIALS AND METHODS

Implementation

We have developed two dedicated, end-to-end metaproteomics data analysis pipelines by seamlessly integrating two popular metaproteomics data processing tools, MPA and PeptideShaker, with Unipept for downstream data processing. The code is available on the GitHub pages of the MetaProteomeAnalyzer (https://github.com/compomics/meta-proteome-analyzer) and PeptideShaker (https://github.com/compomics/peptide-shaker). This connection has been implemented in MPA Portable version 2.0, MPA Server version 3.0, PeptideShaker version 1.44, and PeptideShaker version 2.0 (beta), and all later versions. The pipelines were tested on Windows 10 and various Linux systems with an Oracle Java version >1.8 installed. MPA, PeptideShaker and Unipept (https://unipept.ugent.be) are freely available and licensed under a permissive open source license (Apache 2.0, Apache 2.0 and MIT license, respectively).

Data availability

The dataset used to illustrate the power of the pipelines is publicly available in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with dataset identifier PXD017035. For this article we reprocessed 45 raw files (over 2.1 million MS/MS spectra) from this dataset together, more specifically the first replicate

of each file. The data has been analyzed on a virtual machine (Ubuntu 18.04 LTS) with 32 cores and 300 GB RAM available.

Protein identification

The raw files were converted using the ThermoRawFileParserGUI[22] (version 1.2.1) to peak lists (.mgf files) using the "native Thermo library peak picking" as peak picking option and "Ignore missing instrument properties" as error option.

The peak lists (.mgf files) obtained from MS/MS spectra were identified using X! Tandem version X! Tandem (Vengeance version 2015.12.1)[23], MS Amanda (version 2.0.0.9695)[24], MS-GF+ (version v2018.04.09)[25], and Comet (version 2018.01 rev. 3)[26].The searches were conducted using SearchGUI version 3.3.17[12].

Protein identification was conducted against a concatenated target/decoy database of all the reference proteomes of the species present in the extended simplified human intestinal microbiota (SIHUMIx) sample[27], concatenated with a cRAP database of contaminants (https://thegpm.org/cRAP). The decoy sequences were created by reversing the target sequences in SearchGUI, and the identification settings were as follows: specific cleavage with trypsin with a maximum of two missed cleavages; 10.0 ppm as MS1 tolerance and 0.02 Da as MS2 tolerance; Carbamidomethylation of C as fixed modification; Oxidation of M as variable modification; Acetylation of protein N-termini, Pyrrolidone from E and Q as variable modifications during the refinement procedure of X! Tandem.

Peptides and proteins were inferred from the spectrum identification results using PeptideShaker version 1.16.43[13]. PSMs, peptides and proteins were validated at a 1% FDR estimated using the decoy hit distribution.

RESULTS AND DISCUSSION

To illustrate the user-friendliness of the pipeline, we reprocessed 45 RAW-files of the SIHUMIx dataset with the SearchGUI/PeptideShaker pipeline. This dataset reflects the majority of known metabolic activities typically found in the human intestine and consists of eight bacterial species (*Anaerostipes caccae, Bifidobacterium longum, Bacteroides thetaiotaomicron, Blautia producta, Clostridium butyricum, Clostridium ramosum, Escherichia coli* and *Lactobacillus plantarum*) covering the dominant genera Firmicutes, Bacteroidetes and Proteobacteria in human faeces[27]. We reprocessed the dataset with SearchGUI and imported the identification files in PeptideShaker (for more details, see the Methods section). In total, 1.097.782 of the 2.156.648 PSMs were identified (50.9% PSM identification rate), leading to 67905 uniquely identified peptides (Supporting.

These validated peptides were exported via the Export menu > Follow Up Analysis > Export to Unipept [Figure 1]. A similar approach in MPA is described here: https://github.com/compomics/meta-proteome-analyzer/blob/master/docu/Suppl_MPA_Unipept.pdf.

In the Unipept web application, we visualize the taxonomies via a treemap, sunburst plot, treeview and hierarchical outline [Figure 2]. Moreover the user can determine the function of the proteins identified in the sample by browsing through their EC (Enzyme Commission) numbers or GO

(gene ontology) annotations[28,29]. Moreover, all of the results can be easily exported to a comma-separated, semi-colon-separated and tab-separated file.

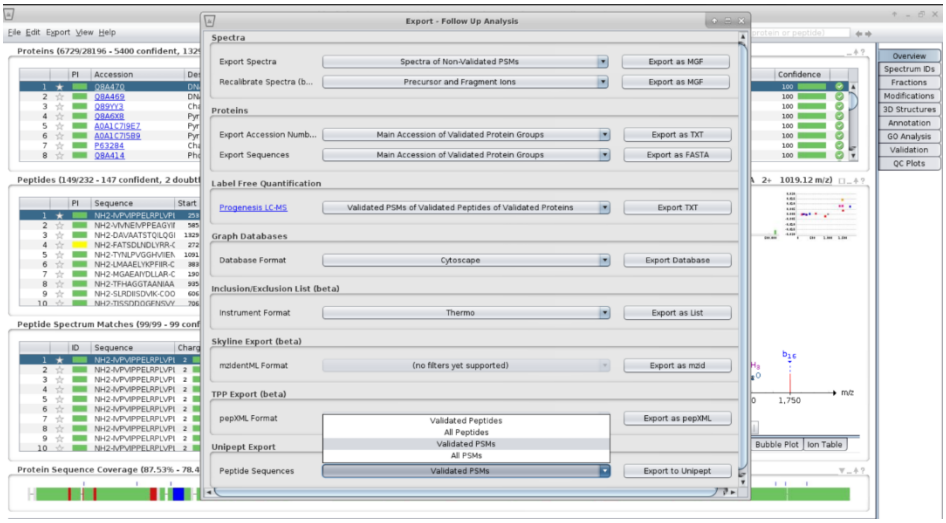


Figure 1. Screenshot of PeptideShaker version 1.44 providing the user with options to export only the validated peptides, all peptides, the validated PSMs or all PSMs to Unipept.

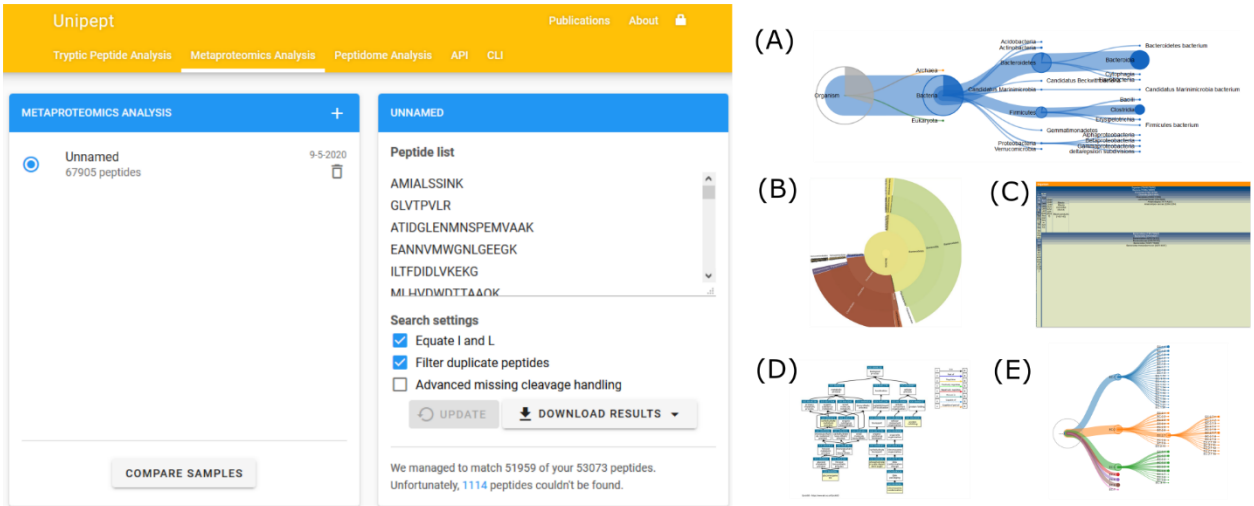


Figure 2. The exported peptides from MPA or PeptideShaker are immediately visible in Unipept (left). Several visualizations are instantly visible on the Unipept webbrowser: a treemap (A), sunburst (B) and treeview (C) for taxonomic analysis. For functional analyses the GO trees (D) for

biological processes, cellular components and molecular functions are available, as well as the trees for EC numbers (E).

CONCLUSIONS

Here, we have presented two end-to-end pipelines for metaproteomics data analysis. We therefore have combined two powerful and commonly used metaproteomics data analysis tools in the field, MetaProteomeAnalyzer and PeptideShaker, with the user-friendly Unipept web interface for taxonomic and functional downstream analysis.

AVAILABILITY

The code is available on the GitHub pages of the MetaProteomeAnalyzer (https://github.com/compomics/meta-proteome-analyzer) and PeptideShaker (https://github.com/compomics/peptide-shaker) and are licensed under the Apache License, version 2.0.

We reprocessed publicly available data of the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with dataset identifier PXD017035.

SUPPORTING INFORMATION

The Default Peptide report exported from PeptideShaker version 1.44 is available in the Supporting information.

ACKNOWLEDGMENTS

This work was supported by the Research Foundation - Flanders (FWO) [grant no. 1S90918N (SB) to TVDB; 1164420N to PV; 12I5217N to BM; G042518N to LM, I002819N to PD]; by a FEBS Summer Fellowship [to TVDB]; by the European Union's Horizon 2020 Program (H2020-INFRAIA-2018-1) [823839 to LM]; by the German Federal Ministry of Education and Research (BMBF) of the project 'MetaProteomanalyzer Service' within the German Network for Bioinformatics Infrastructure (de.NBI) [031L103 to DB and KS]; the Trond Mohn Foundation and the Research Council of Norway [to HB], and by the Deutsche Forschungsgemeinschaft (DFG) [RE 3474/2–2, to BYR]. The authors declare no conflict of interest.

REFERENCES

(1) Blackburn, J. M.; Martens, L. The Challenge of Metaproteomic Analysis in Human Samples. Expert Review of Proteomics. 2016, pp 135–138. https://doi.org/10.1586/14789450.2016.1135058.

(2) Kleiner, M.; Thorson, E.; Sharp, C. E.; Dong, X.; Liu, D.; Li, C.; Strous, M. Assessing Species Biomass Contributions in Microbial Communities via Metaproteomics. Nature Communications. 2017. https://doi.org/10.1038/s41467-017-01544-x.

(3) Zhang, X.; Li, L.; Butcher, J.; Stintzi, A.; Figeys, D. Advancing Functional and Translational Microbiome Research Using Meta-Omics Approaches. Microbiome. 2019. https://doi.org/10.1186/s40168-019-0767-6.

(4) Kleiner, M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. mSystems. 2019. https://doi.org/10.1128/msystems.00115-19.

(5) Wilmes, P.; Bond, P. L. The Application of Two-Dimensional Polyacrylamide Gel Electrophoresis and Downstream Analyses to a Mixed Community of Prokaryotic Microorganisms. Environmental Microbiology. 2004, pp 911–920. https://doi.org/10.1111/j.1462-2920.2004.00687.x.

(6) Schiebenhoefer, H.; Van Den Bossche, T.; Fuchs, S.; Renard, B. Y.; Muth, T.; Martens, L. Challenges and Promise at the Interface of Metaproteomics and Genomics: An Overview of Recent Progress in Metaproteogenomic Data Analysis. Expert Review of Proteomics. 2019, pp 375–390. https://doi.org/10.1080/14789450.2019.1609944.

(7) Muth, T.; Benndorf, D.; Reichl, U.; Rapp, E.; Martens, L. Searching for a Needle in a Stack of Needles: Challenges in Metaproteomics Data Analysis. Mol. BioSyst. 2013, pp 578–585. https://doi.org/10.1039/c2mb25415h.

(8) Muth, T.; Kolmeder, C. A.; Salojärvi, J.; Keskitalo, S.; Varjosalo, M.; Verdam, F. J.; Rensen, S. S.; Reichl, U.; de Vos, W. M.; Rapp, E.; et al. Navigating through Metaproteomics Data: A Logbook of Database Searching. PROTEOMICS. 2015, pp 3439–3453. https://doi.org/10.1002/pmic.201400560.

(9) Verheggen, K.; Raeder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and Evolution of Database Search Engines-a Central Component of Mass Spectrometry Based Proteomic Workflows. Mass Spectrometry Reviews. 2017. https://doi.org/10.1002/mas.21543.

(10) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehtevä, M.; Reichl, U.; Martens, L.; Rapp, E. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. Journal of Proteome Research. 2015, pp 1557–1565. https://doi.org/10.1021/pr501246w.

(11) 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, pp 996–999. https://doi.org/10.1002/pmic.201000595.

(12) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. Journal of Proteome Research. 2018, pp 2552–2555. https://doi.org/10.1021/acs.jproteome.8b00175.

(13) 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. Nature Biotechnology. 2015, pp 22–24. https://doi.org/10.1038/nbt.3109.

(14) Muth, T.; Kohrs, F.; Heyer, R.; Benndorf, D.; Rapp, E.; Reichl, U.; Martens, L.; Renard, B. Y. MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go. Analytical Chemistry. 2018, pp 685–689. https://doi.org/10.1021/acs.analchem.7b03544.

(15) Heyer, R.; Schallert, K.; Büdel, A.; Zoun, R.; Dorl, S.; Behne, A.; Kohrs, F.; Püttker, S.; Siewert, C.; Muth, T.; et al. A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 H Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer. Frontiers in Microbiology. 2019. https://doi.org/10.3389/fmicb.2019.01883.

(16) Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; et al. The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition. Nucleic Acids Research. 2017, pp D1100–D1106. https://doi.org/10.1093/nar/gkw936.

(17) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; et al. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. Nucleic Acids Research. 2019, pp D442–D450. https://doi.org/10.1093/nar/gky1106.

(18) Mesuere, B.; Devreese, B.; Debyser, G.; Aerts, M.; Vandamme, P.; Dawyndt, P. Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples. Journal of Proteome Research. 2012, pp 5773–5780. https://doi.org/10.1021/pr300576s.

(19) Singh, R. G.; Tanca, A.; Palomba, A.; Van der Jeugt, F.; Verschaffelt, P.; Uzzau, S.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept 4.0: Functional Analysis of Metaproteome Data. Journal of Proteome Research. 2019, pp 606–615. https://doi.org/10.1021/acs.jproteome.8b00716.

(20) Consortium, T. U.; The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Research. 2019, pp D506–D515. https://doi.org/10.1093/nar/gky1049.

(21) Federhen, S. The NCBI Taxonomy Database. Nucleic Acids Research. 2012, pp D136–D143. https://doi.org/10.1093/nar/gkr1178.

(22) Craig, R.; Beavis, R. C. TANDEM: Matching Proteins with Tandem Mass Spectra. Bioinformatics. 2004, pp 1466–1467. https://doi.org/10.1093/bioinformatics/bth092.

(23) 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. Journal of Proteome Research. 2014, pp 3679–3684. https://doi.org/10.1021/pr500202e.

(24) Kim, S.; Pevzner, P. A. MS-GF Makes Progress towards a Universal Database Search Tool for Proteomics. Nature Communications. 2014. https://doi.org/10.1038/ncomms6277.

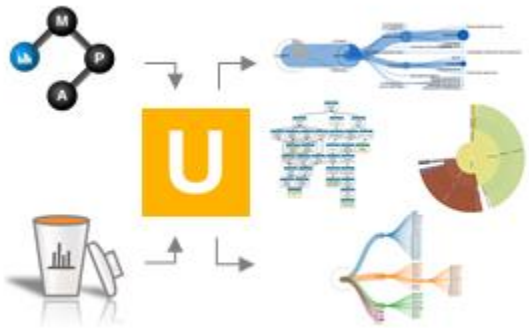
(25) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. PROTEOMICS. 2013, pp 22–24. https://doi.org/10.1002/pmic.201200439.

(26) Schäpe, S. S.; Krause, J. L.; Engelmann, B.; Fritz-Wallace, K.; Schattenberg, F.; Liu, Z.; Müller, S.; Jehmlich, N.; Rolle-Kampczyk, U.; Herberth, G.; et al. The Simplified Human Intestinal Microbiota (SIHUMIx) Shows High Structural and Functional Resistance against Changing Transit Times in In Vitro Bioreactors. Microorganisms. 2019, p 641. https://doi.org/10.3390/microorganisms7120641.

(27) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Michael Cherry, J.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. Gene Ontology: Tool for the Unification of Biology. Nature Genetics. 2000, pp 25–29. https://doi.org/10.1038/75556.

(28) The Gene Ontology Consortium; The Gene Ontology Consortium. The Gene Ontology Resource: 20 Years and Still GOing Strong. Nucleic Acids Research. 2019, pp D330–D338. https://doi.org/10.1093/nar/gky1055.

For TOC Only



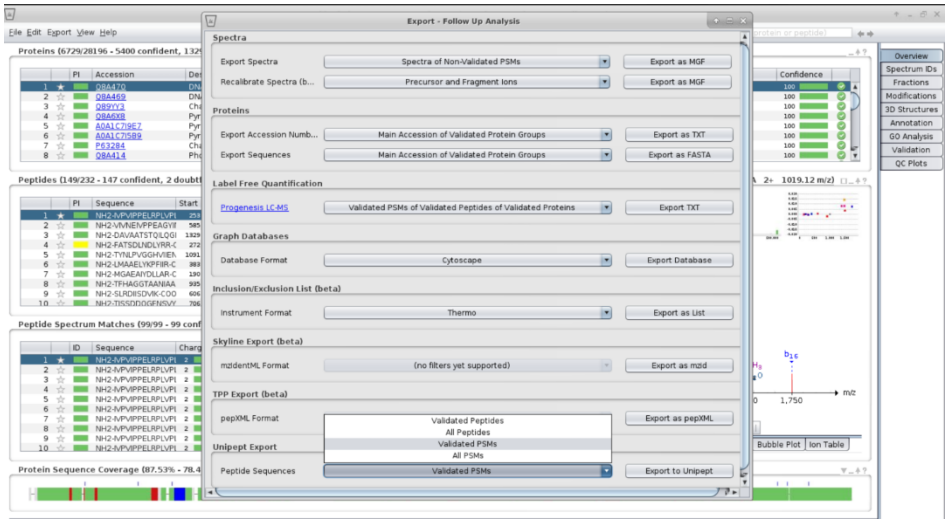


Figure 1. Screenshot of PeptideShaker version 1.44 providing the user with options to export only the validated peptides, all peptides, the validated PSMs or all PSMs to Unipept.

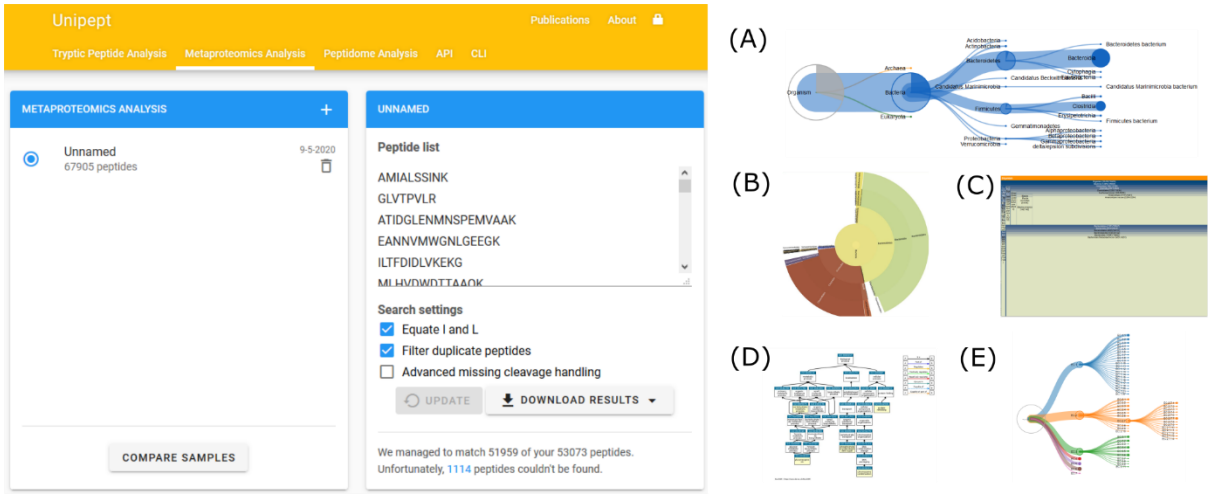


Figure 2. The exported peptides from MPA or PeptideShaker are immediately visible in Unipept (left). Several visualizations are instantly visible on the Unipept webbrowser: a treemap (A), sunburst (B) and treeview (C) for taxonomic analysis. For functional analyses the GO trees (D) for biological processes, cellular components and molecular functions are available, as well as the trees for EC numbers (E).

For TOC Only