Received: 28 March 2022 | Revised: 15 April 2022 | Accepted: 2 May 2022

DOI: 10.1002/imt2.25

SHORT COMMUNICATION

iMeta WILEY

iMetaLab Suite: A one-stop toolset for metaproteomics

Leyuan Li[1,2] | **Zhibin Ning**[1,2] | **Kai Cheng**[1,2] | **Xu Zhang**[1,2] | **Caitlin M. A. Simopoulos**[1,2] | **Daniel Figeys**[1,2]

[1]School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[2]Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Ontario, Canada

Correspondence
Daniel Figeys, School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada.
Email: dfigeys@uottawa.ca

Funding information
Ontario Ministry of Economic Development and Innovation, Grant/Award Number: REG1-4450; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: 210034; Genome Canada and the Ontario Genomics Institute, Grant/Award Number: OGI-114; Canadian Institutes of Health Research, Grant/Award Number: ECD-144627

Abstract

Metaproteomics is a recently thriving technique that studies the collection of proteins in complex microbiomes of the human, animal, plant, and environment. The bioinformatics workflow required for metaproteomics research, from the database search and protein quantification to downstream functional and taxonomic analysis has been challenging and thus limiting the accessibility of metaproteomics to microbiome researchers. To overcome these challenges, we have developed a set of tools named iMetaLab Suite. iMetaLab Suite includes the following components: (1) MetaLab Desktop, an automated database search software that facilities proteins identification and quantitation from microbiomes; (2) the automated iMetaReport that allows users to quickly access database search results and data set profiles; and (3) an interactive online toolset, iMetaShiny, covering most frequently used functional, taxonomic, and statistical analysis in metaproteomics. iMetaLab Suite is a free, easily accessible, and actively updated toolset available to assist researchers to explore metaproteomic data.

KEYWORDS
bioinformatics, database search, metaproteomics, microbiome, statistical analysis, visualization

Highlights

- A one-stop solution for metaproteomics data analysis for nonexpert.
- Database search and result reports that include taxonomy and function.
- Interactive tools for frequently used metaproteomics data analysis tools.

INTRODUCTION

Proteins make up roughly 50% of the dry mass of microbial cells and play various roles in the microbes. Therefore, it is important to use proper techniques to understand the composition of proteins and the functional units of microbiomes. Metaproteomics is such a technique. Briefly, peptides derived from a protein extraction and digestion workflow are subjected to LC-MS/MS analysis, and the resulting

Leyuan Li and Zhibin Ning contributed equally to the manuscript.

MS/MS spectra are compared with in silico generated theoretical spectra for peptide identification. This approach is easy to conduct for single-species proteomics studies since the database is species-specific and the size is ideal. For example, *Escherichia coli* strain K12 has a protein FASTA sequence database of 4375 protein sequences (1845 kB in size) from UniProt. However, when it comes to microbiome reference protein catalogs, database size increases dramatically to capture as many potential species as possible. As an example, the integrated gene catalog (IGC) database of the human gut microbiome has 9.9 million sequences and a size of 3.17 GB [1], that is, around 2000 times bigger than the *E. coli* strain K12 database. Using these large reference protein catalogs as databases, not only challenges computational capability but most importantly, negatively impacts the false-discovery rate (FDR) modeling of the target-decoy approach. To overcome this challenge, we previously developed the MetaPro-IQ workflow that uses an iterative database search strategy to generate a reduced data set-specific database for a MaxQuant search [2]. A conventional MaxQuant search output provides quantified peptide and protein group tables. Under the complex microbiome context, it is necessary, but challenging, to derive accurate taxonomic matches and comprehensive functional annotations from these search outputs. In addition, downstream data analysis and visualization of microbiome data adds an additional dimension of complexity compared to conventional proteomics, as both taxonomic and functional information are associated with the proteins. These challenges altogether make metaproteomics not easily accessible to scientists who are not experts in bioinformatics.

To overcome this challenge, we developed the iMetaLab Suite, which includes the entire framework of database search (MetaLab Desktop) for protein identification and quantification [3], an automated report (iMetaReport), and a variety of interactive tools for data analysis and visualization (iMetaShiny). iMetaLab was rooted from our previous MetaPro-IQ workflow, the implementation of which required computational knowledges. Upon rising requests from scientists, we wrapped up the workflow into a desktop standalone version in which we eventually involved features of spectra clustering [3], post-translational modification analysis [4], and built-in iMetaReport modules. We share the toolset with the microbiome research community. iMetaLab Suite now has registered users from over 160 different institutions around the world. We aim to make iMetaLab Suite a free and one-stop toolset for metaproteomics, with increasing amounts of tools under active development.

RESULTS

Overview of iMetaLab Suite

The iMetaLab Suite tools (Figure 1) are accessible through https://iMetaLab.ca. The MetaLab Desktop software can be freely downloaded from the website or through email requests sent to techteam.metalab@gmail.com to access the latest version. The software takes user input of LC-MS/MS raw files, experimental design meta table (optional), workflow, and parameter settings. Detailed documentation of the MetaLab Desktop is accessible at https://wiki.imetalab.ca/. Under default settings, MetaLab will execute a database search and automatically generate result tables, including Summary, Peptide, ProteinGroup, Taxonomy, and Function tables that are frequently used in downstream analysis. Different formats of the taxonomy and functional results are generated to meet different data visualization requirements.

The iMetaReport is generated automatically following MetaLab database search. A pop-up notice will be sent to the user to navigate to an html report. The iMetaReport contains five major tabs that statistically and visually summarize the Summary, Peptide, ProteinGroup, Taxonomy, and Function outputs, respectively. Optimal visualization is achieved when users input their experimental design (meta table) at the database search step. A sample iMetaReport can be accessed at: https://report.imetalab.ca.

The iMetaShiny apps are a collection of data analysis and visualization Shiny apps that are frequently applied in metaproteomics data analysis. The apps are divided into three subclasses based on their purposes. The first subclass of apps are for data set and sample overview, including Quick summary of LC-MS/MS identification for quality checks, Enrichment analysis that includes both taxonomic and functional enrichment based on user-input protein list or table, and Tree view based on user-input NCBI taxonomic IDs. The second subclass of apps is for multivariate data analysis, including dimensionality reduction tools (PCA, PLS-DA, and t-SNE), differential protein expression analysis, Sets analysis, pepFunk [5], Batch effect explorer, and co-occurrence analysis. The third subclass of apps is for data visualizations based on user-input tables preformatted to meet the plotting requirements. For each of the Shiny apps, a sample data set is given to demonstrate the workflow and to guide users to prepare their input data table. We also provide a 96-well plate randomizer and a Sample scrambler to aid users in their metaproteomics

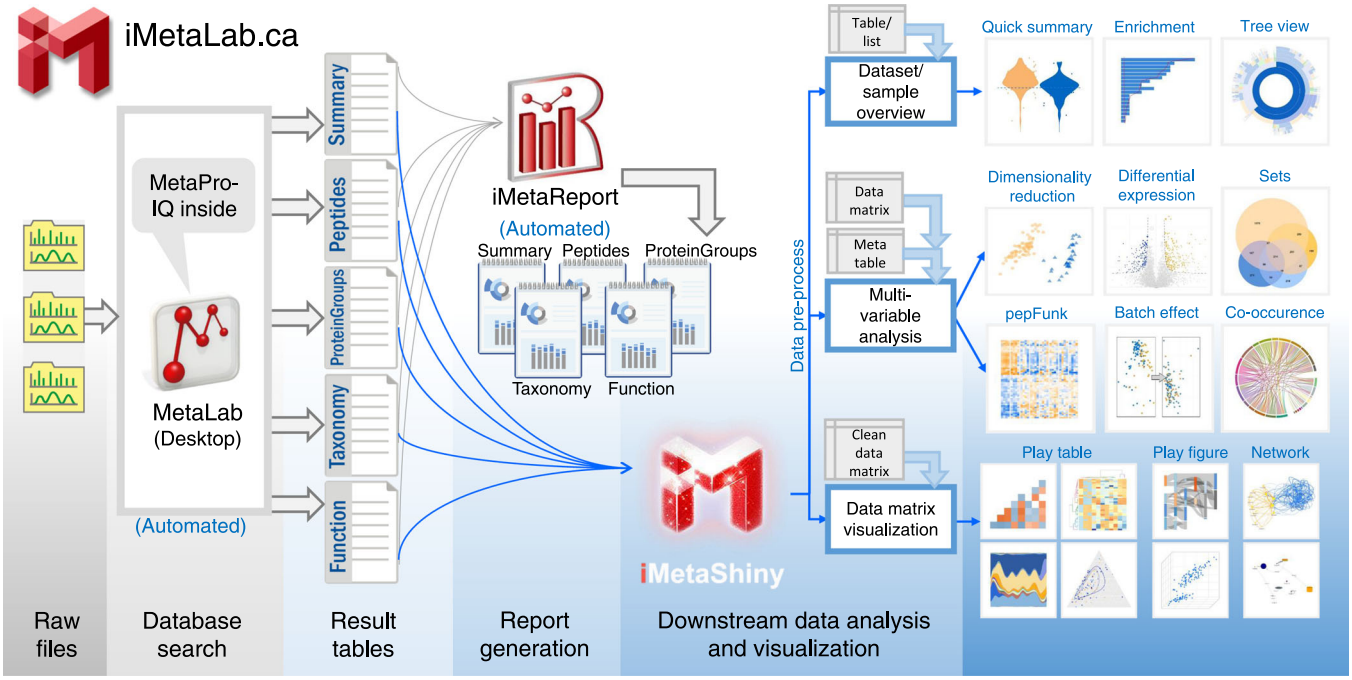

FIGURE 1 Framework of the iMetaLab Suite. Users load raw files to the MetaLab Desktop software to perform an automated metaproteomics database search. After the search, a series of result tables will be generated. Based on the search results, iMetaReport will be automatically generated, covering quick views of identification summary, peptides, proteinGroups, taxonomy, and function of the data set. Using the result tables, users can go to iMetaShiny for various types of downstream data analysis and visualization.

experimental design. More apps are being continuously developed and updated for access to the community.

Case studies and results

Case I: Database search and automated report of data set overviews

One individual microbiome was cultured with or without the presence of diclofenac (an NSAID drug) in triplicates, the data set was taken from our previously published work [6]. The protein digests were analyzed using a 1.5-h gradient with Orbitrap Q-Exactive. MetaLab Desktop (V2.2) was used to search the six samples against the IGC database using the default settings of closed search. By using four threads on a Windows server (Two Intel Xeon E5649 processors, 96 GB RAM), this search took 14 h to complete. After the database search, a series of result files, including summary.txt, peptides.txt, proteinGroups.txt, and BuiltIn.taxa.all.csv, and functions.tsv, were generated. An iMetaReport was also automatically created. The report was presented as an html webpage consisting of five summary tabs for visualizing identification (ID), peptides, proteinGroups, taxonomy, and function. The ID summary results took the summary.txt as input. In Case I, results showed that there were 21,600 peptide sequences identified and 6601 protein groups quantified in total, with an average MS/MS identification rate of 44.9% (Figure 2A,B). Taking peptides.txt and proteinGroups.txt as inputs, respectively, both Peptide and ProteinGroup reports provided important parameters, such as peptide charge states, score distribution (Figure 2C), intensity distribution, and so on, for users to examine the overall quality of the data. Both reports also provided a heatmap and principal component analysis (PCA) score plots to visualize the experimental outcome. The visualizations are based on $\log_{10}$-transformed peptide intensities and proteinGroup label-free quantification (LFQ) intensities, respectively. For the proteinGroup PCA visualization, $\log_{10}$-transformed LFQ-intensities were imputed using a robust sequential algorithm to resolve possible data sparsity. In this example, the two groups showed a clear separation on PC1 (Figure 2D). In the ProteinGroup report, if users set up the meta-information in the database search, analysis of variance (ANOVA) will be performed between the user-input experimental groups based on LFQ-intensities, and FDR-adjusted p values are given for both matrix and pairwise comparisons (Figure 2E). In the Taxonomy report, the number of taxa identification, alpha and beta diversity, as well as stacked bar plots of microbial composition were provided. As an example, differences in genus-level protein biomass contribution in response to diclofenac treatment can be clearly observed (Figure 2F). In the functional report, functional compositions at different levels using various

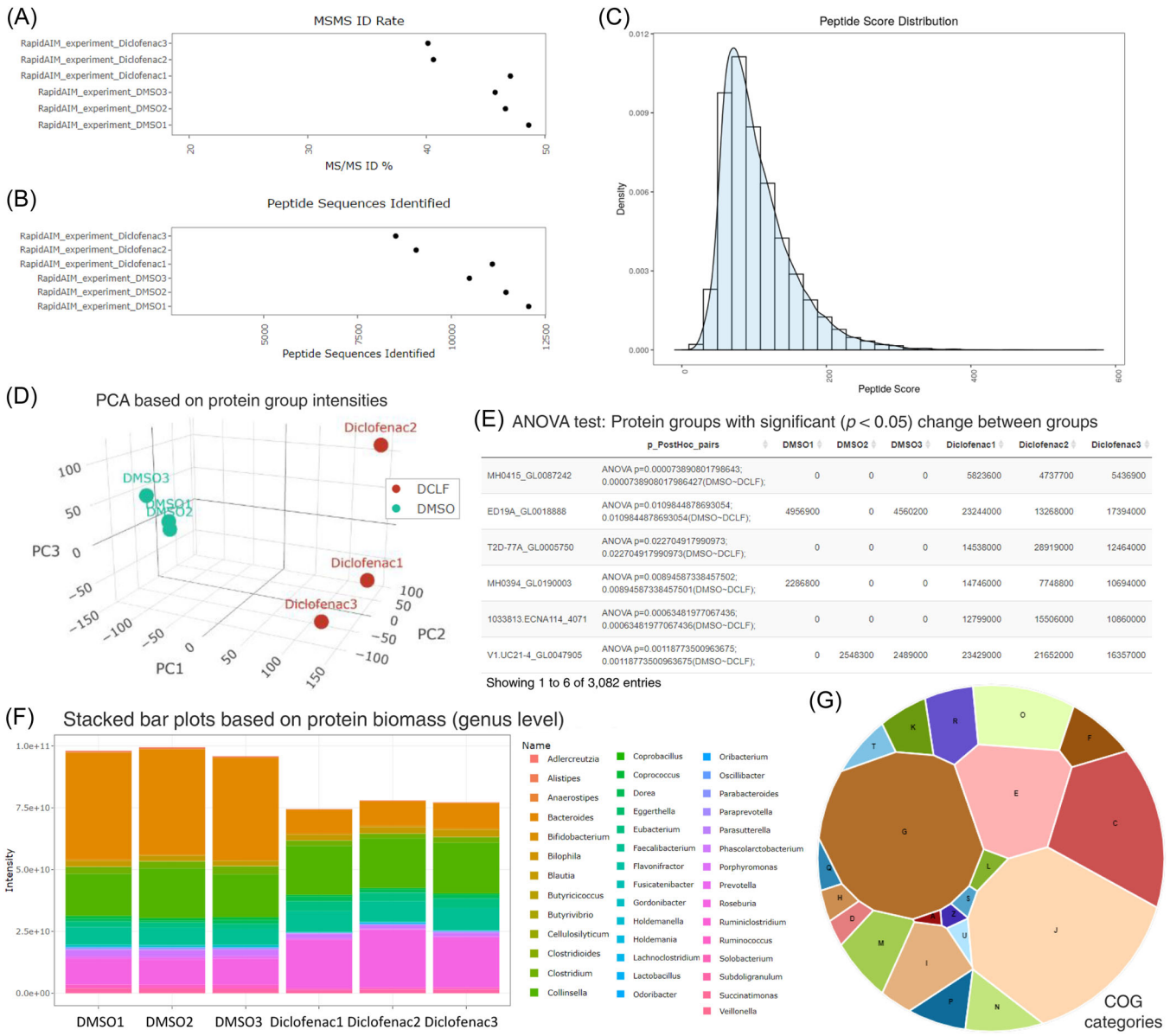

FIGURE 2 Examples from iMetaReport. (A) ID summary report: MS-MS identification rate of each sample. (B) ID summary report: Number of peptide sequences identified in each sample. (C) Peptide report: Peptide score distribution in the data set. (D) ProteinGroup report: principal component analysis based on protein group intensities. (E) ProteinGroup report: analysis of variance test based on protein group intensities. (F) Taxonomy report: stacked bar plots based on protein biomass (genus level). (G) Function report: composition of clusters of orthologues categories in the data set.

functional databases, including clusters of orthologues (COGs), were visualized (Figure 2G), and heatmap and PCA visualizations were also provided. In case users did not set up the meta-information during the database search, after the search, the user can remove the original report file, set up meta information, and click "run" again. MetaLab will check through all existing search files and skip the steps that have been performed, directly leading to a regeneration of the iMetaReport with updated meta information. A complete example of iMetaReport is available at https://report.imetalab.ca. Note that iMetaReport is aimed at quick sample overviews; it is recommended that the users perform further data analysis using iMetaShiny applications.

Case II: Differentially expressed protein groups and their taxonomy and functions

Protein LFQ intensities from the search results of Case I were uploaded to the Differential Protein Analyzer (https://shiny.imetalab.ca/Volcano_plot/). The data pre-processing option was turned on and navigated us to the

"Process Data" page. Here, we filtered out rows with 75% missing values and normalized them by columns. Users can also directly input their preprocessed protein expression table with the data preprocessing option kept off. We used default statistical parameters and a smooth-curve threshold for determining the significantly changed protein groups. The resulting volcano plot is shown in Figure 3A. We obtained 95 significantly increased and 117 significantly decreased protein groups in response to diclofenac treatment in this metaproteomics data set. The table can be downloaded under the "Result table download" panel. Next, we examined the enrichment profile of the differentially expressed proteins. IDs of these proteins were uploaded to the Enrichment Analysis tool

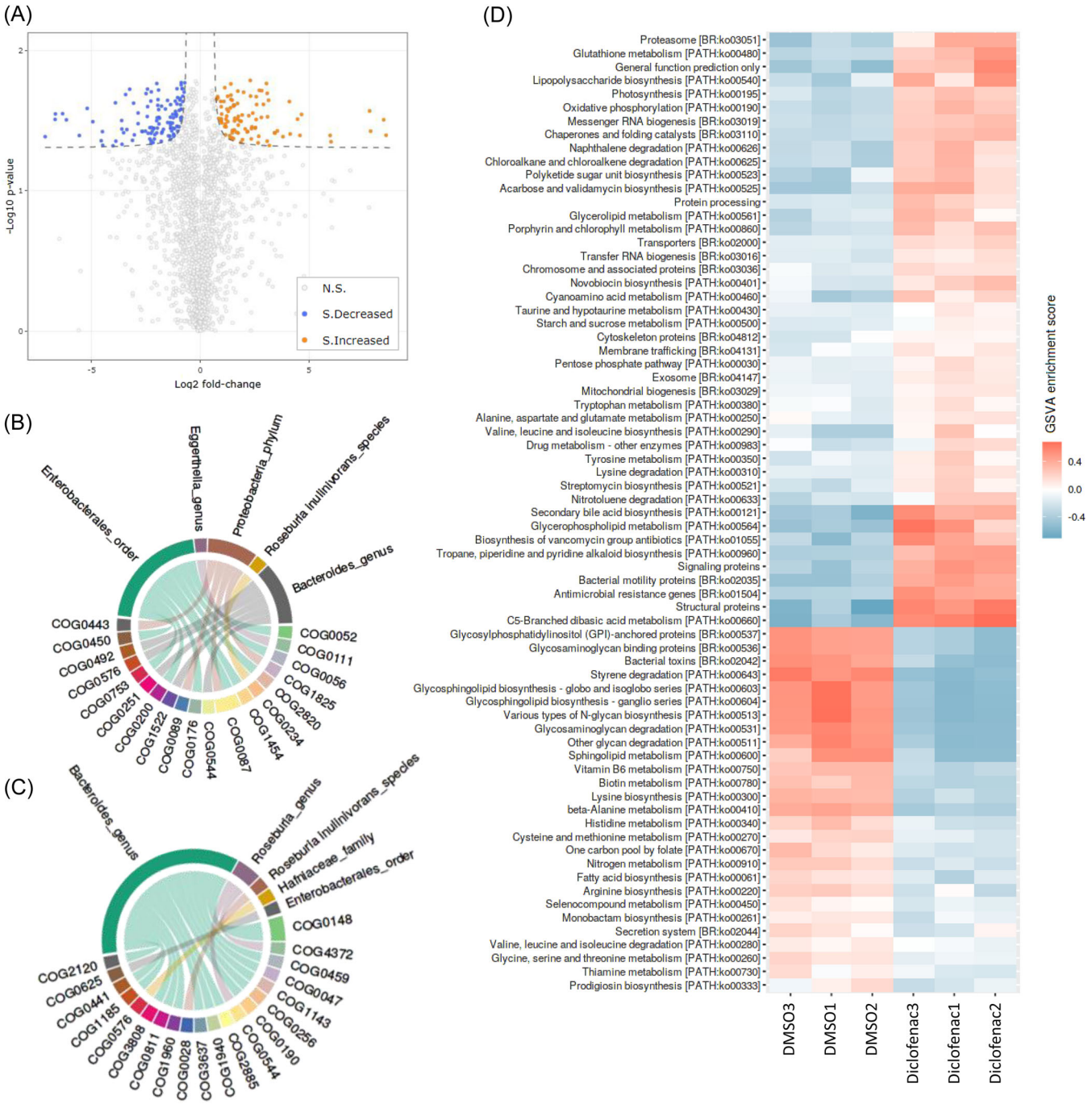

FIGURE 3 Examples of iMetaShiny applications. (A) Result of differential protein analysis from the diclofenac data set. Orange dots represent significantly increased protein groups, while blue dots represent significantly decreased protein groups. (B,C) Taxon-function enrichment analysis of the significantly changed protein groups (using top-1 protein in each protein group, $p < 0.05$). (D) Heatmap visualizing Gene Set Variation Analysis scores of the diclofenac data set.

(https://shiny.imetalab.ca/metaproteomics_enrichment/), Function and Taxon correlation was selected as the analysis type, and COG was selected as the functional group type. Protein IDs were assigned with taxonomic and functional information, and we were navigated to the visualization page. Here, we chose to visualize the data using the Circos plot. As shown in Figure 3B,C, significantly increased protein groups are mainly from Enterobacterales, and genus *Bacteroides* had the most significantly decreased COG functions.

Case III: Peptide-centric functional enrichment analysis

Besides using the LFQ protein group intensities, we demonstrate the peptide-centric workflow through our pepFunk [5] (https://shiny.imetalab.ca/pepFunk/). The peptides.txt table was uploaded to the application, DMSO was set as the control, and diclofenac was set as the treatment. Using Gene Set Variation Analysis adapted for peptide data, significantly enriched KEGG pathways showed clear differentiation between the treatment and the control (Figure 3D).

DISCUSSIONS

With iMetaLab Suite, we aim to maximize the accessibility of metaproteomic bioinformatics workflow to scientists with all levels of bioinformatics expertise in the field of microbiome research, as well as those in conventional proteomics/systems biology. We are actively developing novel database search workflows and strategies, as well as more statistical approaches for downstream functional, taxonomic, and ecological analysis of the metaproteomics data. These will be actively updated into the iMetaLab Suite and we welcome feedback and suggestions from users to improve the user experience and performance of the tools.

METHODS

MetaLab Desktop is developed in Java and integrates open-source third-party libraries/tools MzJava [7], PRIDE Cluster [8], X!Tandem [9], MaxQuant [10], and Msconvert. iMetaReport is developed using R Markdown [11] with packages, including ggplot2 [12], plotly [13], tidyverse [14], vegan [15], ggdendro, d3heatmap, pheatmap, and so on. The server is hosted via openCPU [16] and therefore can be accessed publicly. User database search result is submitted by the MetaLab software to the openCPU server to generate the report. iMetaShiny apps are developed using R and the Shiny package [13], other frequently used packages are DT, data.table, shinyBS, htmlwidgets, and so on. It is hosted via shiny server. All these servers are hosted on Amazon cloud AWS.

AUTHOR CONTRIBUTIONS

Daniel Figeys and Zhibin Ning conceptualized the framework of the iMetaLab Suite. Zhibin Ning established the webserver and bioinformatics frameworks for iMetaLab.ca, iMetaReport, and iMetaShiny. Zhibin Ning and Leyuan Li developed iMetaReports and iMetaShiny tools. Leyuan Li wrote the manuscript. Kai Cheng developed and maintains the MetaLab Desktop software. Xu Zhang developed the MetaPro-IQ pipeline. Caitlin M. A. Simopoulos developed the pepFunk tool in iMetaShiny. All authors have tested the toolsets, revised the manuscript, read the final manuscript, and approved it for publication.

ACKNOWLEDGMENTS

This study was supported by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-114), CIHR grant (ECD-144627), the Natural Sciences and Engineering Research Council of Canada (NSERC, Grant no. 210034), the Ontario Ministry of Economic Development and Innovation (REG1-4450). Daniel Figeys acknowledges a Distinguished Research Chair from the University of Ottawa. Caitlin M. A. Simopoulos was funded by a stipend from the NSERC CREATE in Technologies for Microbiome Science and Engineering (TECHNOMISE) Program.

CONFLICTS OF INTEREST

Daniel Figeys cofounded MedBiome, a clinical microbiomics company. Other authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All LC-MS/MS sequencing data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under submission number PXD033624. The database search results and reports are saved in GitHub (https://github.com/northomics/iMetaLab_paper). Supporting Information materials (graphical abstract, slides, videos, Chinese translated version, and update materials) may be found in the online DOI or iMeta Science http://www.imeta.science/.

ORCID

Daniel Figeys https://orcid.org/0000-0002-5373-7546

REFERENCES

1. Li, Junhua, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam,

et al. 2014. "An integrated catalog of reference genes in the human gut microbiome." *Nature Biotechnology* 32: 834–841. https://doi.org/10.1038/nbt.2942

2. Zhang, Xu, Zhibin Ning, Janice Mayne, Jasmine I. Moore, Jennifer Li, James Butcher, Shelley Ann Deeke, et al. 2016. "MetaPro-IQ: A universal metaproteomic approach to studying human and mouse gut microbiota." *Microbiome* 4: 31. https://doi.org/10.1186/s40168-016-0176-z
3. Cheng, Kai, Zhibin Ning, Xu Zhang, Leyuan Li, Bo Liao, Janice Mayne, Alain Stintzi, and Daniel Figeys. 2017. "MetaLab: An automated pipeline for metaproteomic data analysis." *Microbiome* 5: 157. https://doi.org/10.1186/s40168-017-0375-2
4. Cheng, Kai, Zhibin Ning, Xu Zhang, Leyuan Li, Bo Liao, Janice Mayne, Daniel Figeys. 2020. "MetaLab 2.0 enables accurate post-translational modifications profiling in metaproteomics." *Journal of the American Society for Mass Spectrometry* 31: 1473–1482. https://doi.org/10.1021/jasms.0c00083
5. Simopoulos, Caitlin M. A., Zhibin Ning, Xu Zhang, Leyuan Li, Krystal Walker, Mathieu Lavallée-Adam, and Daniel Figeys. 2020. "pepFunk: A tool for peptide-centric functional analysis of metaproteomic human gut microbiome studies." *Bioinformatics* 36: 4171–4179. https://doi.org/10.1093/bioinformatics/btaa289
6. Li, Leyuan, Zhibin Ning, Xu Zhang, Janice Mayne, Kai Cheng, Alain Stintzi, and Daniel Figeys. 2020. "RapidAIM: A culture- and metaproteomics-based rapid assay of individual microbiome responses to drugs." *Microbiome* 8: 33. https://doi.org/10.1186/s40168-020-00806-z
7. Horlacher, Oliver, Frederic Nikitin, Davide Alocci, Julien Mariethoz, Markus Müller, and Frederique Lisacek. 2015. "MzJava: An open source library for mass spectrometry data processing." *Journal of Proteomics* 129: 63–70. https://doi.org/10.1016/j.jprot.2015.06.013
8. Griss, Johannes, Joseph M. Foster, Henning Hermjakob, and Juan Antonio Vizcaíno. 2013. "PRIDE cluster: Building a consensus of proteomics data." *Nature methods* 10: 95–96. https://doi.org/10.1038/nmeth.2343
9. Craig, Robertson, and C. Beavis Ronald. 2004. "TANDEM: Matching proteins with tandem mass spectra." *Bioinformatics* 20: 1466–1467. https://doi.org/10.1093/bioinformatics/bth092
10. Cox, Jürgen, Matthias Mann. 2008. "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." *Nature Biotechnology* 26: 1367–1372. https://doi.org/10.1038/nbt.1511
11. Baumer, Benjamin, and Dana Udwin. 2015. "R markdown." *Wiley Interdisciplinary Reviews: Computational Statistics* 7: 167–177. https://doi.org/10.1002/wics.1348
12. Wickham, Hadley. 2011. "ggplot2." *Wiley Interdisciplinary Reviews: Computational Statistics* 3: 180–185. https://doi.org/10.1002/wics.147
13. Sievert, Carson. 2020. *Interactive web-based data visualization with R, plotly, and shiny*. CRC Press. https://doi.org/10.1201/9780429447273
14. Wickham, Hadley. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4: 1686. https://doi.org/10.21105/joss.01686
15. Dixon, Philip. 2003. "VEGAN, a package of R functions for community ecology." *Journal of Vegetation Science* 14: 927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x
16. Ooms, Jeroen. 2014. "The OpenCPU system: Towards a universal interface for scientific computing through separation of concerns." *arXiv preprint arXiv* 1406: 4806. https://doi.org/10.48550/arXiv.1406.4806

How to cite this article: Li, Leyuan, Zhibin Ning, Kai Cheng, Xu Zhang, Caitlin M. A. Simopoulos, and Daniel Figeys. 2022. "iMetaLab Suite: A one-stop toolset for metaproteomics." *iMeta* 1, e25. https://doi.org/10.1002/imt2.25