JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 10, 2018

Pp. 1123–1127
DOI: 10.1089/cmb.2018.0050

MixProTool: A Powerful and Comprehensive Web Tool for Analyzing and Visualizing Multigroup Proteomics Data

SHISHENG WANG, WEN ZHENG, LIQIANG HU, MENG GONG, and HAO YANG

ABSTRACT

Deciphering and visualizing proteomics data are a big challenge for high-throughput proteomics research. In this work, we develop a free interactive web software platform, MixProTool, for processing multigroup proteomics data sets. This tool provides integrated data analysis workflow, including quality control assessment, normalization, soft independent modeling of class analogy, statistics, gene ontology enrichment, and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis. This software is also highly compatible with the identification and quantification results of various frequently used search engines, such as MaxQuant, Proteome Discoverer, or Mascot. Moreover, all analyzed results can be visualized as vector graphs and tables for further analysis. MixProTool can be conveniently operated by users, even those without bioinformatics training, and it is extremely useful for mining the most relevant features among different samples. MixProTool is deployed at the public shinyapps io server.

Keywords: multigroups, proteomics data, web tool, data analysis workflow.

1. INTRODUCTION

Next-generation sequencing has been extensively applied to biomedical research and has largely fueled the development of precision medicine. However, genomic sequence is only the first step for decoding genome function. The comprehensive understanding of molecular mechanisms, processes, and pathways requires to translate genome into proteome and connect genome to phenotypic characteristics. In the past decade, with the tremendous improvements of instruments, mass spectrometry (MS) coupled with ether liquid chromatography (LC-MS) or gas chromatography (GC-MS) is increasingly becoming a prevalent and powerful approach for the identification and quantification of differentially expressed proteins (Graves and Haystead, 2002; Conrads and Petricoin, 2016). With the augmentation in the sensitivity and quantification accuracy of these systems and the development of new methods (Meier et al., 2018), thousands of proteins can be identified across each group sample. As a result, deciphering biologically correlative proteomics data is an arduous and complex process, even for professionals specialized in bioinformatics.

We herein present the development of a free, fast, and comprehensive web-based graphical user interface (GUI) tool, named MixProTool, to further analyze proteomics results acquired from software such as MaxQuant (Cox and Mann, 2008), Proteome Discoverer (Thermo Scientific), or Mascot (Koenig et al., 2008). Shiny (Gatto and Christoforou, 2014) provides an elegant and formidable web framework for building web

Lab of Proteomics and Metabolomics, West China-Washington Mitochondria and Metabolism Research Center, Key Lab of Transplant Engineering and Immunology, MOH, West China Hospital, SCU, Chengdu, China.

applications; thus, we developed MixProTool as a Shiny app that includes a variety of functions and automates analysis of multigroup proteomics data. Overall, we aim to provide convenience for users—even those not engaged in proteomics—to complete analysis and visualization of their proteomics data. MixProTool is available online at https://wsslearning.shinyapps.io/MixProTool.

2. SOFTWARE DESIGN AND IMPLEMENTATION

2.1. Compilation and workflow

The GUI of MixProTool compiled in R (www.r-project.org) was developed in Shiny and deployed on the free shinyapp.io sever. The GUI contains mainly three parts (Fig. 1A): module names, the parameter setting panel, and the results presentation panel. Ten kernel modules cover data input, quality control, soft independent modeling of class analogy (SIMCA) analysis, univariate and multivariate statistical analyses, visualization, and enrichment analysis (Fig. 1B). In the parameter setting panel, users can follow the workflow (Fig. 1B) and indicate appropriate parameters for their data analysis in each step. Within the results presentation panel, the results will be demonstrated immediately after suitable parameters are set up.

2.2. Example data set

By default, MixProTool provides example data (PXD008522), which were obtained from the triple negative parental human breast cancer cell lines (Das et al., 2018) to appraise its quality and accuracy.

2.3. Implementation

2.3.1. Data import. The identification and quantification results output from popular proteomics software packages (e.g., MaxQuant, Proteome Discoverer, or Mascot) can be saved as .txt, .csv, .xls, or .xlsx files, with very minor edition, and then uploaded into MixProTool without redundant submission in subsequent analysis.

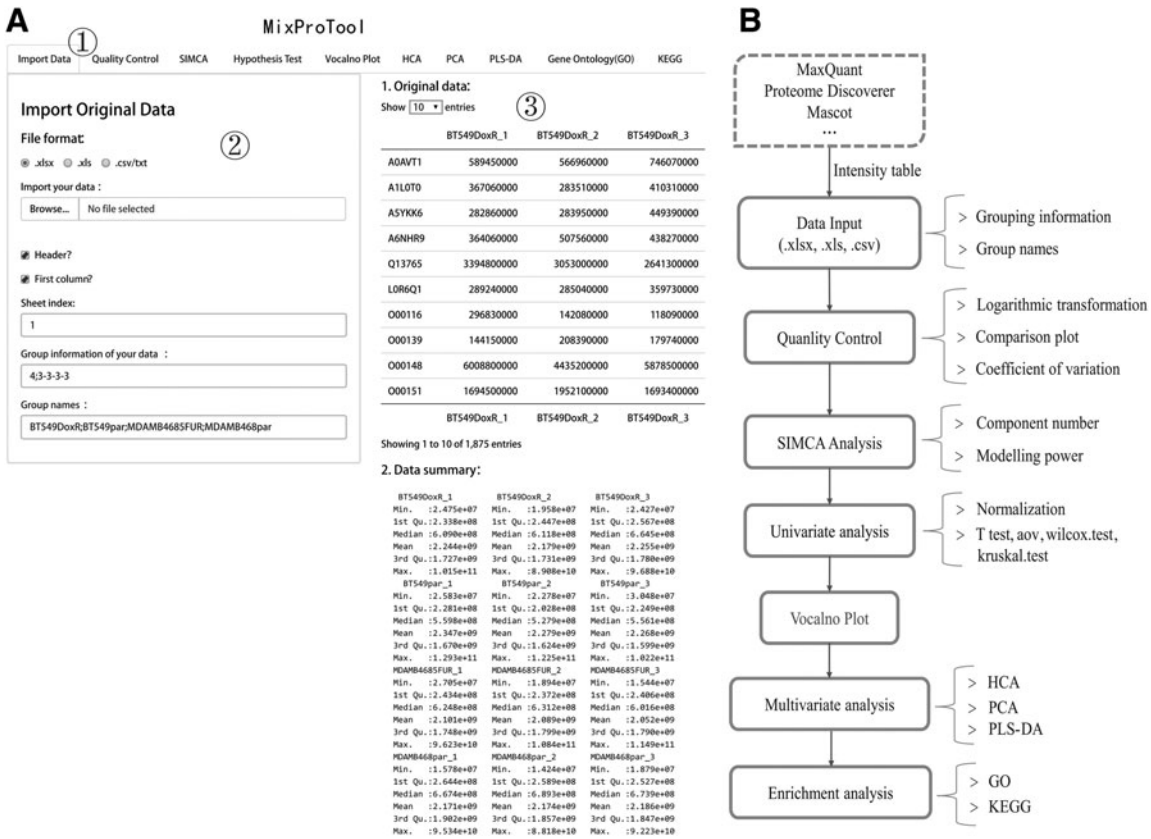

FIG. 1. GUI and workflow: **(A)** module names and operation panels of MixProTool, and **(B)** basic workflow for users to analyze their data and generate relevant graphs and tables in the output. GO, gene ontology; GUI, graphical user interface; HCA, hierarchical cluster analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; PCA, principal components analysis; PLS-DA, partial least squares–discriminant analysis; SIMCA, soft independent modeling of class analogy.

2.3.2. Quality control. This module can transform raw intensities into log values ($\log_2$ or $\log_{10}$) and calculate the coefficient of variation (CV) for each group to evaluate their precision and repeatability. Moreover, the quantification values of every two samples are compared with a linear regression model.

2.3.3. SIMCA analysis. Proteomics data sets usually contain unwanted variation introduced by signal drift or multiplicative noise across the dynamic range. These effects can detrimentally impact discovery of significant signals. SIMCA analysis can counteract this situation by calculating the features' modeling power, which is a measure of the contribution of each variable to the model, and removing those below the threshold.

2.3.4. Univariate and multivariate statistical analysis. For univariate statistical analysis, MixProTool offers parametric statistical tests, nonparametric statistical tests, and the false discovery rate (FDR)-corrected *p* value by using the Benjamini–Hochberg FDR algorithm (Benjamini and Hochberg, 1995). For multivariate statistical analysis, principal components analysis (PCA), partial least squares–discriminant analysis (PLS-DA), and hierarchical cluster analysis (HCA) are provided for cluster analysis with relevant graphical results and diagnostics (Alonso-Gutierrez et al., 2015; Song et al., 2018).

2.3.5. Functional enrichment analysis. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis can be processed in this tool to determine which biological processes, molecular functions, and cellular compartments are modulated and the kinds of biological pathways in which those proteins are involved. Currently, MixProTool can perform GO enrichment analysis for 14 common organisms (9606, 10090, etc.) and KEGG pathway enrichment analysis for 5299 organisms (listed at www.genome.jp/kegg/catalog/org_list.html). The GO enrichment analysis was implemented using the topGO package (Alexa and Rahnenführer, 2009), and the KEGG pathway enrichment analysis was performed using the clusterProfiler package (Yu et al., 2012).

3. RESULTS

We prepared demo data containing 1875 proteins on MixProTool in an attempt to make every function and result more evident. For the demo data, in the quality assessment procedure, the CV of 1826 proteins, accounting for 97.39% of the demo data, was <5% (Table 1), implying the decent repeatability of this assay.

In addition, SIMCA analysis calculates the modeling power for every protein, and those above zero were retained for subsequent hypothesis test analysis (Fig. 2A). According to the group index set by users, a volcano plot can be displayed, as shown in Figure 2B. For multivariate statistical analysis, two-dimensional graphs of PCA and PLS-DA are shown in Figure 2C and D, respectively, from which we can discover that replicates of the same group cluster close to each other, whereas different samples are apart. An HCA plot is shown in Figure 2E, which illustrates the variation trend of similar expression of each protein. For enrichment analysis, 310 GO IDs and 35 KEGG IDs are enriched, and Figure 2F displays the top 10 terms for each category.

4. CONCLUSIONS

MixProTool offers powerful and comprehensive data processing while being easy to operate and capable of dealing with large-scale proteomics data sets. In contrast to existing software, such as GiaPronto (Weiner et al., 2017), MixProTool deftly handles multigroup samples and provides more useful modules, that is, SIMCA analysis, HCA, PLS-DA, and KEGG enrichment analysis, which can offer deeper understanding of users' data sets. In addition, MixProTool is built in Shiny, with a menu-driven interactive interface, so there are no requirements of programming techniques from the users. Overall, this tool provides a convenient platform for rapid and consistent analysis of proteomics data sets, which could notably promote such research in related fields.

Table 1. Statistics of Coefficient of Variation of All Proteins in Demo Data

Groups	*Number of proteins.*	*Ratio*
[0, 5%]	1826	97.39%
(5%, 10%]	47	2.51%
(10%, 15%]	2	0.1%

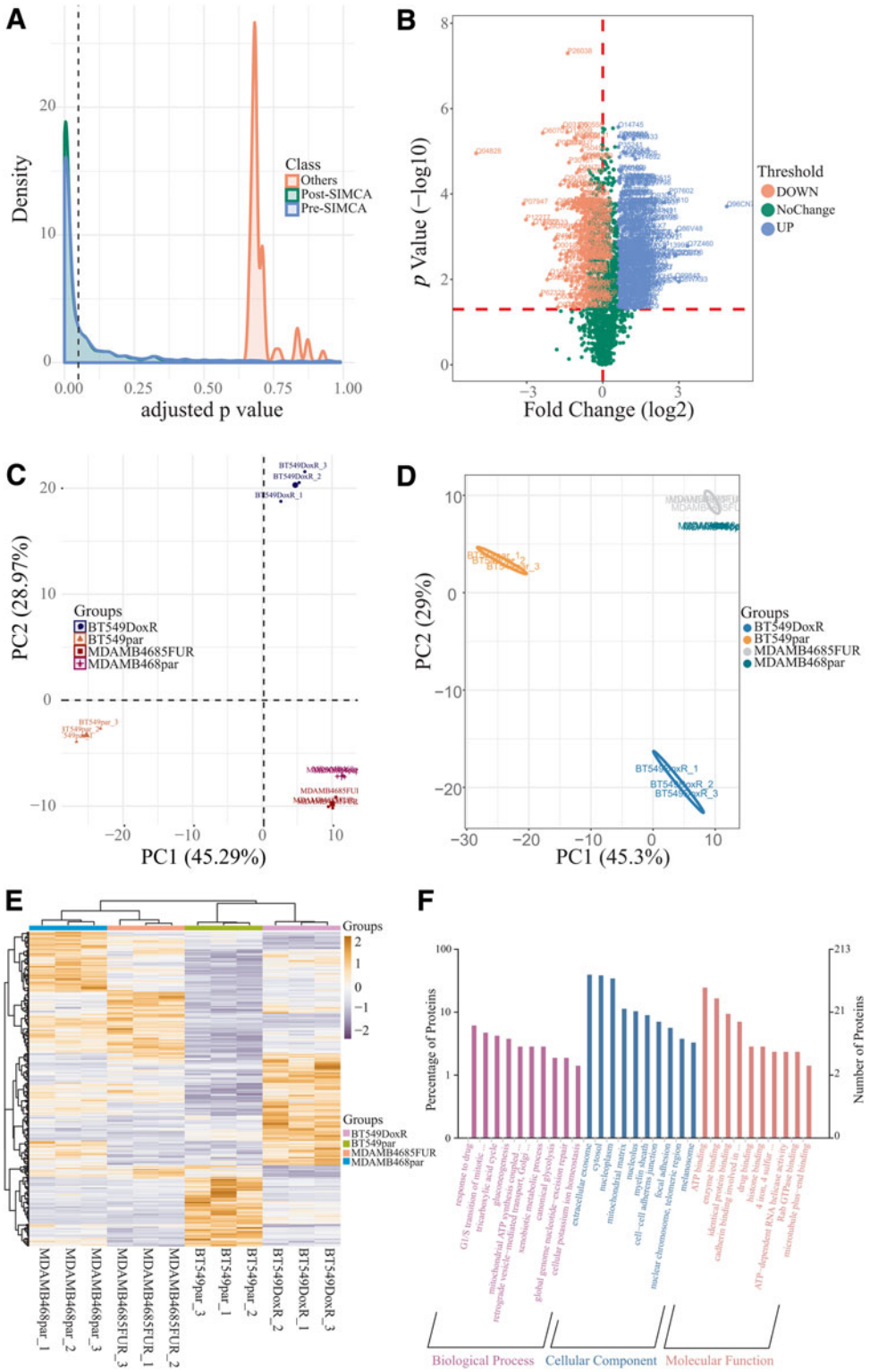

FIG. 2. Statistical analysis and relevant visualization of MixProTool: **(A)** distribution of the adjusted *p* value of each protein with SIMCA analysis (labeled post-SIMCA), without SIMCA analysis (labeled pre-SIMCA), and those with modeling powers below zero (labeled others); **(B)** volcano plot; **(C)** two-dimensional plot of PCA; **(D)** two-dimensional plot of PLS-DA; **(E)** cluster of expression of proteins in HCA plot; and **(F)** bar plot of top 10 terms for each category (i.e., molecular function, cellular component, and biological process) in GO enrichment analysis.

ACKNOWLEDGMENTS

This work was supported by the Science and Technology Department of Sichuan Province (Grant No. 2017HH0036) and the National Natural Science Foundation of China (Grant No. 81102366).

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Alexa, A., and Rahnenführer, J. 2009. Gene set enrichment analysis with topGO. R package version 2.30.1. Available at: www.bioconductor.org. Accessed March 9, 2018.

Alonso-Gutierrez, J., Kim, E.M., Batth, T.S., et al. 2015. Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng*. 28, 123–133.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*. 1, 289–300.

Conrads, T.P., Petricoin, E.F., 3rd. 2016. The Obama Administration's Cancer Moonshot: A call for proteomics. *Clin. Cancer Res.* 22, 4556–4558.

Cox, J., and Mann, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol*. 26, 1367–1372.

Das, C.K., Linder, B., Bonn, F., et al. 2018. BAG3 overexpression and cytoprotective autophagy mediate apoptosis resistance in chemoresistant breast cancer cells. *Neoplasia*. 20, 263–279.

Gatto, L., and Christoforou, A. 2014. Using R and Bioconductor for proteomics data analysis. *Biochim. Biophys. Acta*. 1844, 42–51.

Graves, P.R., and Haystead, T.A. 2002. Molecular biologist's guide to proteomics. *Microbiol. Mol. Biol. Rev.* 66, 39–63.

Koenig, T., Menze, B.H., Kirchner, M., et al. 2008. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *J. Proteome. Res.* 7, 3708–3717.

Meier, F., Geyer, P.E., Virreira Winter, S., et al. 2018. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods*. [Epub ahead of print]; DOI:10.1038/s41592-018-0003-5. Available at: www.nature.com. Accessed May 6, 2018.

Song, W., Wang, H., Maguire, P., et al. 2018. Nearest clusters based partial least squares discriminant analysis for the classification of spectral data. *Anal. Chim. Acta.* 1009, 27–38.

Weiner, A.K., Sidoli, S., Diskin, S.J., et al. 2017. GiaPronto: A one-click graph visualization software for proteomics datasets. *Mol. Cell Proteomics*. [Epub ahead of print]; DOI:10.1074/mcp.TIR117.000438. Available at: www.mcponline.org. Accessed March 13, 2018.

Yu, G., Wang, L.G., Han, Y., et al. 2012. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS.* 16, 284–287.

Address correspondence to:
Prof. Meng Gong
Lab of Proteomics and Metabolomics
West China-Washington Mitochondria and Metabolism Research Center
Key Lab of Transplant Engineering and Immunology, MOH
West China Hospital, SCU
No. 88, Keyuan South Road
Hi-tech Zone, Chengdu 610041, China

E-mail: gongmeng@scu.edu.cn

Prof. Hao Yang
Lab of Proteomics and Metabolomics
West China-Washington Mitochondria and Metabolism Research Center
Key Lab of Transplant Engineering and Immunology, MOH
West China Hospital, SCU
No. 88, Keyuan South Road
Hi-tech Zone, Chengdu 610041, China

E-mail: yanghao@scu.edu.cn