












proteoDA: a package for quantitative proteomics

Timothy J. Thurman ^{1*}, Charity L. Washam ^{1*}, Duah Alkam ¹, Jordan T. Bird ¹, Allen Gies ¹, Kalyani Dhusia ¹, Michael S. Robeson II ², and Stephanie D. Byrum ^{1,3¶}

¹ Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR, USA ² Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA ³ Arkansas Children's Research Institute, Little Rock, AR, USA ¶ Corresponding author * These authors contributed equally.

DOI: [10.21105/joss.05184](https://doi.org/10.21105/joss.05184)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Jacob Schreiber](#) 

Reviewers:

- [@shahmoradi](#)
- [@MohmedSoudy](#)

Submitted: 02 February 2023

Published: 30 May 2023

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

proteoDA is an R package designed to analyze high resolution intensity-based mass spectrometry data. proteoDA was designed to be streamlined and user-friendly. proteoDA is built around a DAList, a custom R class, which is used to hold the data, statistical design, and results for a differential abundance analysis. Users import protein abundance data, protein annotation data, and sample metadata into a DAList, and all further functions operate on that list. Once the data are in DAList, proteoDA provides functions for further steps of the analysis [Figure 1](#).

proteoDA includes functions for 1) evaluation of multiple normalization methods using a graphical report based on the proteiNorm normalization tool ([Chawade et al., 2014](#); [Graw et al., 2020](#)), 2) generation of graphical quality control reports to assess data quality and sample clustering, 3) flexible specification and fitting of differential abundance models (including mixed models), using the R package limma to perform model fitting ([Law et al., 2020](#); [Ritchie et al., 2015](#)), and 4) generation of tabular results files, as well as interactive and portable HTML result files, using the R package Glimma ([Su et al., 2017](#)). Please see the [ByrumLab/proteoDA/vignettes/tutorial.html](#) for more details.

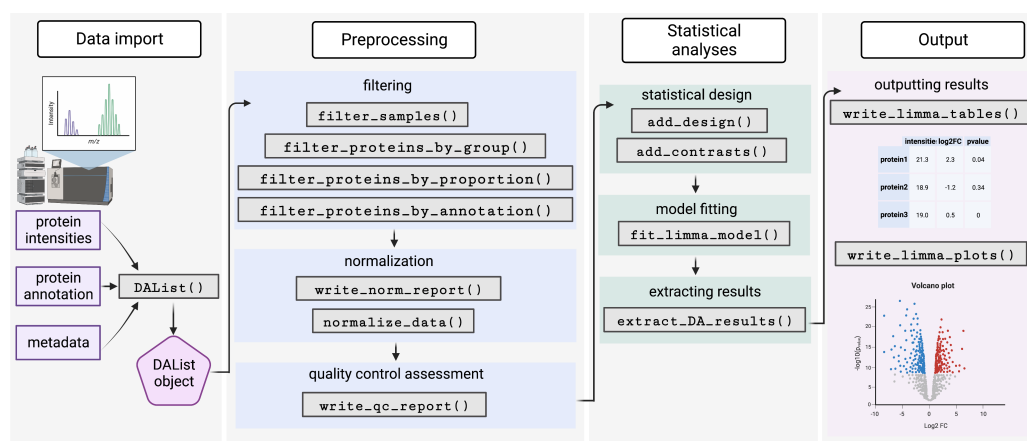


Figure 1: A flowchart of the proteoDA workflow.

proteoDA provides all the functions necessary to evaluate the quality, remove unwanted samples, filter proteins with missing values, normalize the data, perform statistical analysis, and export the results into an interactive HTML report.

State of the field

Quantitative proteomics analysis is growing in popularity with the advancement of mass spectrometer instrumentation and software that allows for the identification of more and more proteins from complex sample mixtures (Arora & Somasundaram, 2019; Deutsch et al., 2022; Perez-Riverol et al., 2021). Sequencing depths have increased from a few thousand to 10,000 proteins commonly identified from discovery experiments. Most proteomics mass spectrometry workflows follow the shotgun or bottom-up approach (Tariq et al., 2021) in which a mixture of proteins is digested by an enzyme into peptides and then analyzed by LC MS/MS. Expertise in sample preparation, mass spectrometer parameters, experimental study designs, biostatistics, and programming languages are all needed in order to properly analyze and interpret proteomics data. ProteoDA was designed to make the process of analyzing bottom-up MS intensity-based data streamlined with flexible model designs. The functions in proteoDA allow a practical approach (Schessner et al., 2022) for the evaluation of normalization methods, easy selection of limma model designs to account for batch effects or other factors, and interactive visualization of the results. In contrast to other comprehensive proteomics analysis packages, proteoDA delivers an interactive, sharable data product that allows domain researchers to visualize their results in a web browser without the need for dedicated server-side infrastructure (Ranathunge et al., 2023; Wolski et al., 2022).

Statement of need

proteoDA was designed to help researchers with minimal knowledge of R extract insights from proteomics data. proteoDA allows users to quickly assess the quality of a mass spectrometry experiment, normalize the data, control for batch effects, and define flexible linear model designs for a wide variety of proteomic experiments. In addition to providing quantitative analysis for experiments, proteoDA delivers interactive and sharable data visualizations that empower data analysis and domain researchers alike to explore data finding, save high quality figures, and generate new insights. proteoDA is widely used in the classroom as well as the IDeA National Resource for Quantitative Proteomics workshops for faculty and students. The package is robust but flexible to account for a wide variety of proteomics experiments and provides a training tutorial explaining normalization and linear model designs.

Acknowledgements

The development of this R package was supported by the National Institutes of Health National Institute of General Medical Sciences (NIH/NIGMS) grants P20GM121293, R24GM137786, the National Science Foundation Award No. OIA-1946391, and the UAMS Winthrop P. Rockefeller Cancer Institute.

References

- Arora, A., & Somasundaram, K. (2019). Targeted proteomics comes to the benchside and the bedside: Is it ready for us? *BioEssays*, 41(2), 1800042. <https://doi.org/10.1002/bies.201800042>
- Chawade, A., Alexandersson, E., & Levander, F. (2014). Normalyzer: A tool for rapid evaluation of normalization methods for omics data sets. *Journal of Proteome Research*, 13(6), 3114–3120. <https://doi.org/10.1021/pr401264n>
- Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J., Mendoza, L., Kundu, D. J., Wang, S., Bandla, C., Kamatchinathan, S., Hewapathirana, S., Pullman, B. S., Wertz, J., Sun, Z., Kawano, S., Okuda, S., Watanabe, Y., MacLean, B., MacCoss, M.

- J., ... Vizcaíno, J. A. (2022). The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Research*, 51(D1), D1539–D1548. <https://doi.org/10.1093/nar/gkac1040>
- Graw, S., Tang, J., Zafar, M. K., Byrd, A. K., Bolden, C., Peterson, E. C., & Byrum, S. D. (2020). proteiNorm – a user-friendly tool for normalization and analysis of TMT and label-free protein quantification. *ACS Omega*, 5(40), 25625–25633. <https://doi.org/10.1021/acsomega.0c02564>
- Law, C., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G., & Ritchie, M. (2020). A guide to creating design matrices for gene expression experiments [version 1; peer review: 2 approved]. *F1000Research*, 9(1444). <https://doi.org/10.12688/f1000research.27893.1>
- Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D., Prakash, A., Frericks-Zipper, A., Eisenacher, M., Walzer, M., Wang, S., Brazma, A., & Vizcaíno, J. A. (2021). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1), D543–D552. <https://doi.org/10.1093/nar/gkab1038>
- Ranathunge, C., Patel, S. S., Pinky, L., Correll, V. L., Chen, S., Semmes, O. J., Armstrong, R. K., Combs, C. D., & Nyalwidhe, J. O. (2023). Promor: A comprehensive r package for label-free proteomics data analysis and predictive modeling. *bioRxiv*. <https://doi.org/10.1101/2022.08.17.503867>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. <https://doi.org/10.1093/nar/gkv007>
- Schessner, J. P., Voytik, E., & Bludau, I. (2022). A practical guide to interpreting and generating bottom-up proteomics data visualizations. *PROTEOMICS*, 22(8), 2100103. <https://doi.org/10.1002/pmic.202100103>
- Su, S., Law, C. W., Ah-Cann, C., Asselin-Labat, M.-L., Blewitt, M. E., & Ritchie, M. E. (2017). Glimma: interactive graphics for gene expression analysis. *Bioinformatics*, 33(13), 2050–2052. <https://doi.org/10.1093/bioinformatics/btx094>
- Tariq, M. U., Haseeb, M., Aledhari, M., Razzak, R., Parizi, R. M., & Saeed, F. (2021). Methods for proteogenomics data analysis, challenges, and scalability bottlenecks: A survey. *IEEE Access*, 9, 5497–5516. <https://doi.org/10.1109/ACCESS.2020.3047588>
- Wolski, W. E., Nanni, P., Grossmann, J., d'Errico, M., Schlapbach, R., & Panse, C. (2022). Prolfqua: A comprehensive r-package for proteomics differential expression analysis. *bioRxiv*. <https://doi.org/10.1101/2022.06.07.494524>