The Journal of Open Source Software

einprot: flexible, easy-to-use, reproducible workflows for statistical analysis of quantitative proteomics data

Charlotte Soneson [1,2¶], **Vytautas Iesmantavicius** [1], **Daniel Hess** [1], **Michael B Stadler** [1,2,3], **and Jan Seebacher** [1]

1 Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland **2** SIB Swiss Institute of Bioinformatics, Basel, Switzerland **3** University of Basel, Switzerland ¶ Corresponding author

DOI: 10.21105/joss.05750

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @AnthonyOfSeattle
- @ByrumLab

Submitted: 05 August 2023
Published: 11 September 2023

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

Quantitative proteomics has come a long way - what used to be specialized analyses performed in proteomics research groups is nowadays a routine service in many proteomics core facilities, and a large collection of sophisticated quantification and analysis tools are available. Yet, the necessary reporting tasks, including statistical analyses of the resulting data, as well as describing all data processing steps, providing quality control, exploration opportunities and result visualizations for publication in a user-friendly way, are generally not routine or automated, and many different analysis workflows are conceivable (Peng et al., 2023). Moreover, additional downstream analyses and integration with other types of data are often necessary, and these are more likely to succeed when all steps of the routine data analysis workflow are transparent and well documented.

The einprot R package provides accessible workflows accommodating quality control, filtering, exploratory analysis and statistical analysis of proteomics data quantified with several commonly used tools, including label-free quantification (LFQ) with MaxQuant (Cox & Mann, 2008) or FragPipe (Kong et al., 2017), and tandem mass tag (TMT)-multiplexed data quantified with Proteome Discoverer (Orsburn, 2021). Each workflow is provided in the form of a template R Markdown (Rmd) file (Xie et al., 2018), containing the code to be executed as well as text descriptions and explanations of the individual steps. This can significantly reduce the amount of time spent on routine processing tasks, e.g., for a core facility, and enables the entire analysis process to be shared with collaborators or data generators in a way that is comprehensive and easy to follow. In addition, the report contains several interactive tables and plots that allow users to immediately explore their data.

To run a workflow, the user calls a single function in their R session, to which they provide the path to the quantification file(s) as well as a number of additional arguments specifying details about the experiment and the requested analyses. einprot then copies the template Rmd script from the package location to the designated output directory, injects the arguments specified by the user to create a stand-alone file, and compiles this into an html report describing the complete analysis. The stand-alone Rmd file is retained and can, if necessary, be manually modified and fine-tuned by the user and recompiled. Alternatively, the analyst can provide their own template Rmd file if custom analyses are desired. A collection of example reports generated by einprot are provided at https://csoneson.github.io/einprot_examples/. The einprot functions used in the workflows can also be called directly in the R session for an interactive analysis or recreation of specific plots (see the online vignette for examples). While compiling the report, einprot exports a set of text files and publication-ready plots for further inspection and dissemination of the results (see the vignette for a full list of output files and Figure 1 and Figure 2 for examples of figures generated by the workflows).

The Journal of Open Source Software

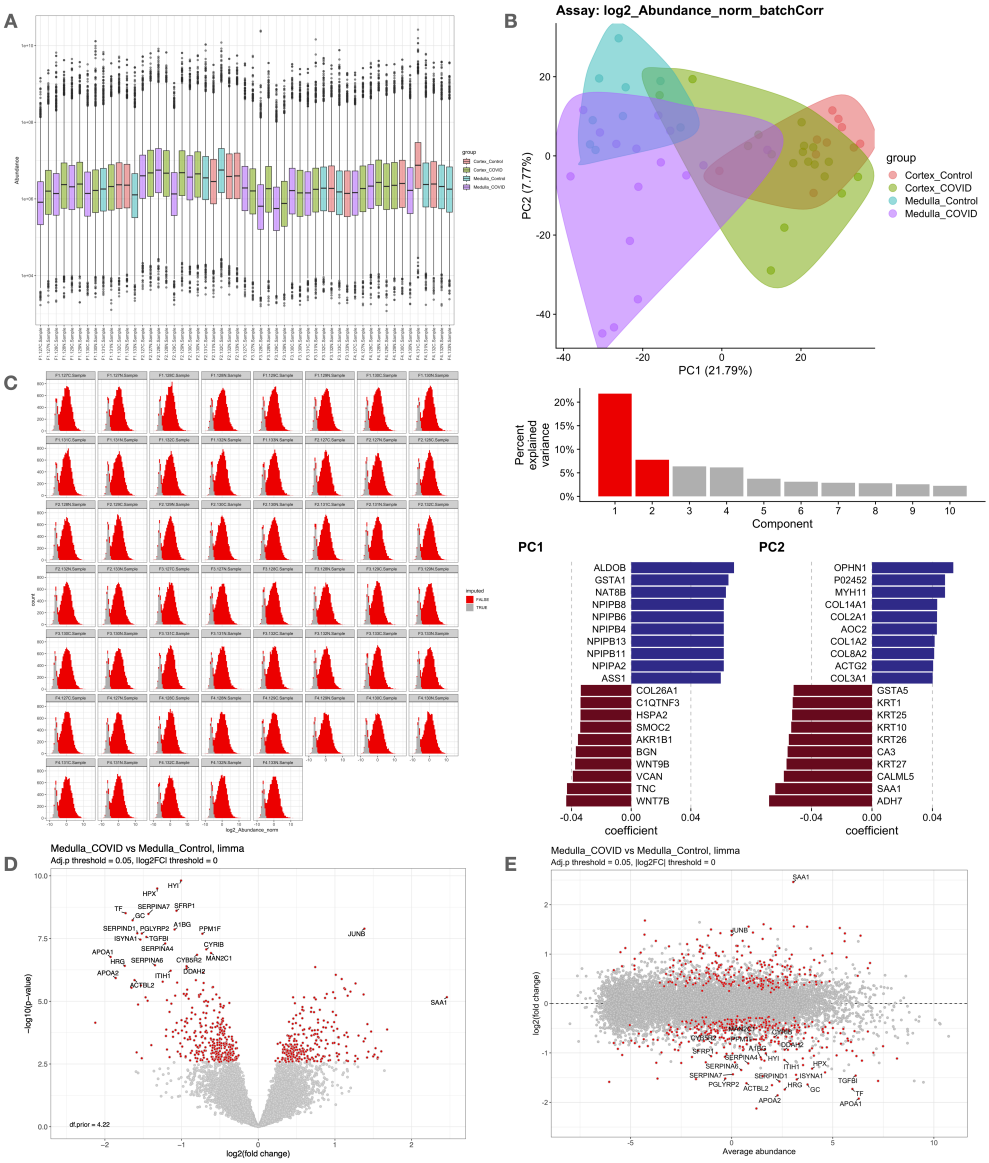

Figure 1: Example figures generated by the einprot workflow, based on a data set from Nie et al. (2021), requantified by He et al. (2022). The figures can also be generated programmatically from the SingleCellExperiment object generated by the einprot workflow. A. Distribution of abundances in each sample. B. Principal component analysis representation of samples, percent variance explained by each principal component, and the proteins with the highest loadings in the first two components. C. Histograms showing the distribution of observed and imputed abundance values in each sample. D. Volcano plot for one of the tested contrasts. E. MA plot for the contrast shown in D.

The Journal of Open Source Software

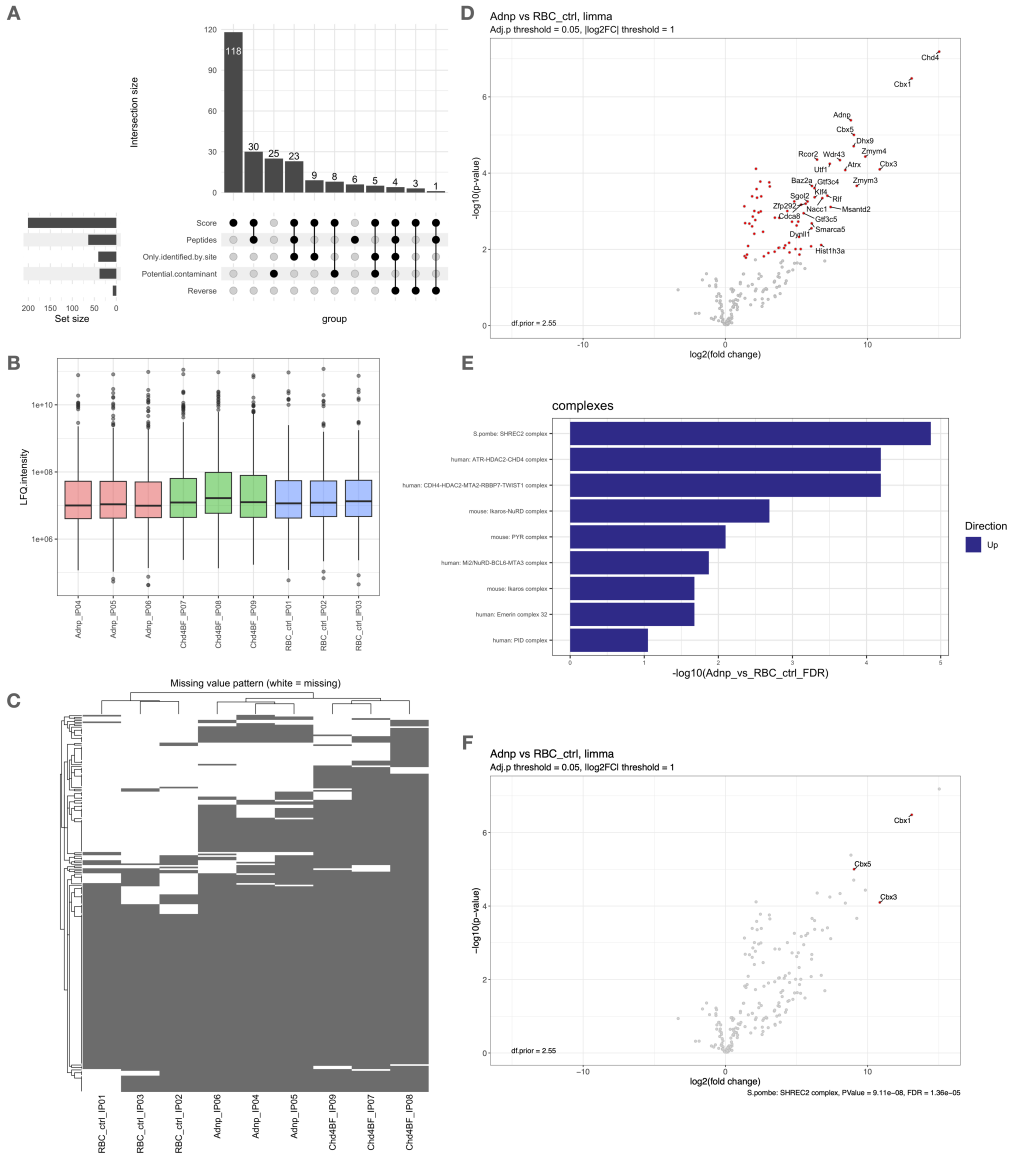

Figure 2: Example figures generated by the einprot workflow, based on an IP-MS data set from Ostapcuk et al. (2018), quantified with MaxQuant. A. Overview of the proteins filtered out by the workflow. B. Distribution of LFQ intensities across samples. C. Global missing value pattern. D. Volcano plot for one of the tested contrasts. E. List of known protein complexes most strongly associated with the contrast in D. F. The same volcano plot as in D, but with the members of the top-ranked complex from E highlighted.

The current version of einprot fully supports five common organisms: *Mus musculus*, *Homo sapiens*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. For these organisms, the user can elect to perform automatic enrichment testing of Gene Ontology terms (Ashburner et al., 2000; Gene Ontology Consortium et al., 2023) and known protein complexes, obtained from CORUM (Giurgiu et al., 2019), PomBase (Harris et al., 2022) and CYC2008 (Pu et al., 2009) and mapped to the organism of interest using the ortholog mapping from the babelgene R package (Dolgalev, 2022). Other species can be analyzed by skipping (parts of) the automatic enrichment testing. Finally, for improved interoperability with other tools, especially from the Bioconductor ecosystem (Huber et al., 2015), einprot stores all raw and processed values (including, e.g., results from differential abundance analysis and

dimensionality reduction via principal component analysis) in a SingleCellExperiment object (Amezquita et al., 2020). In addition, einprot automatically generates an R script that can be sourced to launch a customized interactive exploration session using the iSEE R package (Rue-Albrecht et al., 2018) (Figure 3).

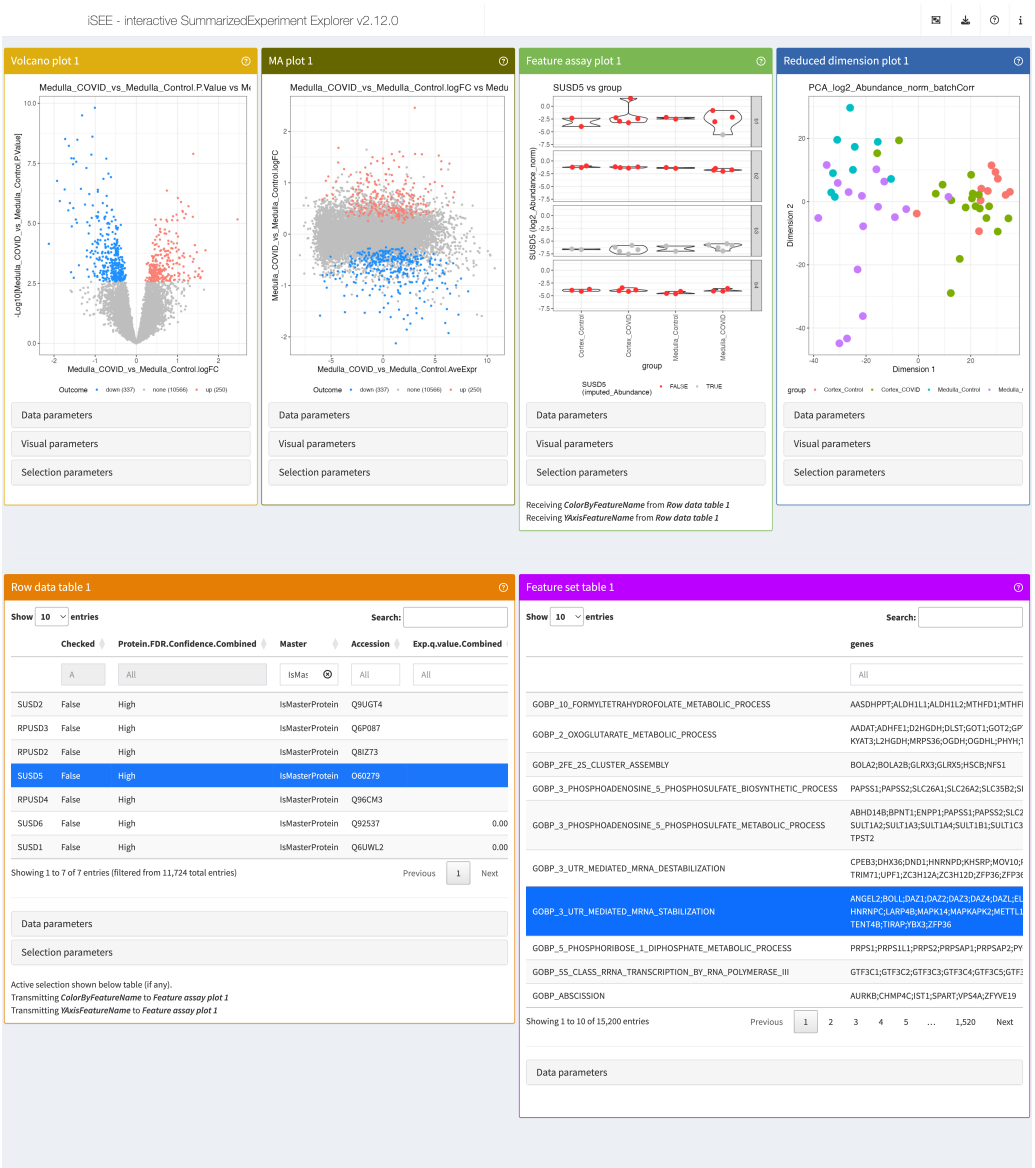

Figure 3: A part of the interactive interface configured for each einprot run, based on the iSEE package. A variety of interactive panels allows the user to explore the statistical results, abundance values and imputation status for individual proteins, a low-dimensional representation of the samples, as well as to browse all results in table form. In addition, all the flexibility provided by the iSEE package is retained, and the user can shape the interface and panel composition according to their own needs.

Statement of need

Several other toolkits are available for analyzing proteomics data (Bai et al., 2023). Arguably, routine statistical analyses are most commonly performed using vendor- or community-developed software suites accessible via graphical user interfaces, such as Perseus (Tyanova et al., 2016) or Proteome Discoverer (Orsburn, 2021). While they provide comprehensive analysis capabilities,

they are typically not fully open source, and sometimes only available for selected operating systems. In addition, the manual "point-and-click" interface means that they are difficult to incorporate into an automated or programmable data analysis pipeline. Among the broad range of community-developed open-source tools, many (e.g., POMAShiny (Castellano-Escuder et al., 2021), LFQ-Analyst (Shah et al., 2020), amica (Didusch et al., 2022), Eatomics (Kraus et al., 2021), ProVision (Gallant et al., 2020), and the proteomics workflows provided via the Galaxy platform (Galaxy Community, 2022; Hiltemann et al., 2023)) provide a graphical user interface where the user steps through the analysis manually; some additionally allow a summary report to be exported. Other software packages provide extensive collections of analysis functions, but require familiarity with programming to use (e.g. protti (Quast et al., 2021), POMA (Castellano-Escuder et al., 2021), SafeQuant (Glatter et al., 2012), Proteus (Gierlinski et al., 2018), DEP (Zhang et al., 2018), prolfqua (Wolski et al., 2023), MSstats (Choi et al., 2014)). With einprot, we attempt to hit a middle ground by providing a collection of fully reproducible workflows that do not require extensive coding skills to run, yet return comprehensive, self-contained reports that contain all the executed code and are fully customizable if needed. In addition, by returning the data and results in a standardized format, the interoperability with other packages, including some of the ones mentioned above, is simplified and allows the user to take advantage of a vast ecosystem of tools. One such example is the direct interface to iSEE, which lets the user interactively explore all the results and processed values generated by and documented in the reproducible workflows. In this way, einprot is used productively by the proteomics and protein analysis facility at FMI and as the basis for published and submitted articles (Welte et al., 2023).

Availability and installation

einprot is available from GitHub (https://github.com/fmicompbio/einprot) and can be installed using the standard installation processes for R packages (e.g., using the remotes package (Csárdi et al., 2021)). Documentation and example usage are available from https://fmicompbio.github.io/einprot/. A collection of example reports can be browsed at https://csoneson.github.io/einprot_examples/.

Acknowledgements

The authors would like to thank Merle Skribbe, Seraina Steiger, Thomas Welte, Patrick Matthias, Helge Grosshans and Marc Bühler for testing and feedback on the software, and Laurent Gatto for valuable discussions. This work was supported by the Novartis Research Foundation. The funding body did not have any role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

References

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, *17*(2), 137–145. https://doi.org/10.1038/s41592-019-0654-x

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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Bai, M., Deng, J., Dai, C., Pfeuffer, J., Sachsenberg, T., & Perez-Riverol, Y. (2023). LFQ-Based peptide and protein intensity differential expression analysis. *Journal of Proteome Research*, *22*(6), 2114–2123. https://doi.org/10.1021/acs.jproteome.2c00812

Castellano-Escuder, P., González-Domínguez, R., Carmona-Pontaque, F., Andrés-Lacueva, C., & Sánchez-Pla, A. (2021). POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis. *PLoS Computational Biology*, *17*(7), e1009148. https://doi.org/10.1371/journal.pcbi.1009148

Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., & Vitek, O. (2014). MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, *30*(17), 2524–2526. https://doi.org/10.1093/bioinformatics/btu305

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M., & Tenenbaum, D. (2021). *Remotes: R package installation from remote repositories, including 'GitHub'*. https://CRAN.R-project.org/package=remotes

Didusch, S., Madern, M., Hartl, M., & Baccarini, M. (2022). Amica: An interactive and user-friendly web-platform for the analysis of proteomics data. *BMC Genomics*, *23*(1), 817. https://doi.org/10.1186/s12864-022-09058-7

Dolgalev, I. (2022). *Babelgene: Gene orthologs for model organisms in a tidy data format*. https://CRAN.R-project.org/package=babelgene

Galaxy Community. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, *50*(W1), W345–W351. https://doi.org/10.1093/nar/gkac610

Gallant, J. L., Heunis, T., Sampson, S. L., & Bitter, W. (2020). ProVision: A web-based platform for rapid analysis of proteomics data processed by MaxQuant. *Bioinformatics*, *36*(19), 4965–4967. https://doi.org/10.1093/bioinformatics/btaa620

Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., … Westerfield, M. (2023). The gene ontology knowledgebase in 2023. *Genetics*, *224*(1). https://doi.org/10.1093/genetics/iyad031

Gierlinski, M., Gastaldello, F., Cole, C., & Barton, G. J. (2018). Proteus: An R package for downstream analysis of MaxQuant output. *bioRxiv Doi:10.1101/416511*. https://doi.org/10.1101/416511

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., & Ruepp, A. (2019). CORUM: The comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Research*, *47*(D1), D559–D563. https://doi.org/10.1093/nar/gky973

Glatter, T., Ludwig, C., Ahrné, E., Aebersold, R., Heck, A. J. R., & Schmidt, A. (2012). Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *Journal of Proteome Research*, *11*(11), 5145–5156. https://doi.org/10.1021/pr300273g

Harris, M. A., Rutherford, K. M., Hayles, J., Lock, A., Bähler, J., Oliver, S. G., Mata, J., & Wood, V. (2022). Fission stories: Using PomBase to understand Schizosaccharomyces pombe biology. *Genetics*, *220*(4). https://doi.org/10.1093/genetics/iyab222

He, T., Liu, Y., Zhou, Y., Li, L., Wang, H., Chen, S., Gao, J., Jiang, W., Yu, Y., Ge, W., Chang, H.-Y., Fan, Z., Nesvizhskii, A. I., Guo, T., & Sun, Y. (2022). Comparative evaluation of Proteome Discoverer and FragPipe for the TMT-Based proteome quantification. *Journal of Proteome Research*, *21*(12), 3007–3015. https://doi.org/10.1021/acs.jproteome.2c00390

Hiltemann, S., Rasche, H., Gladman, S., Hotz, H.-R., Larivière, D., Blankenberg, D., Jagtap, P. D., Wollmann, T., Bretaudeau, A., Goué, N., Griffin, T. J., Royaux, C., Le Bras, Y., Mehta, S., Syme, A., Coppens, F., Droesbeke, B., Soranzo, N., Bacon, W., … Batut, B. (2023). Galaxy training: A powerful framework for teaching! *PLoS Computational Biology*, *19*(1), e1010752. https://doi.org/10.1371/journal.pcbi.1010752

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, *14*(5), 513–520. https://doi.org/10.1038/nmeth.4256

Kraus, M., Mathew Stephen, M., & Schapranow, M.-P. (2021). Eatomics: Shiny exploration of quantitative proteomics data. *Journal of Proteome Research*, *20*(1), 1070–1078. https://doi.org/10.1021/acs.jproteome.0c00398

Nie, X., Qian, L., Sun, R., Huang, B., Dong, X., Xiao, Q., Zhang, Q., Lu, T., Yue, L., Chen, S., Li, X., Sun, Y., Li, L., Xu, L., Li, Y., Yang, M., Xue, Z., Liang, S., Ding, X., … Guo, T. (2021). Multi-organ proteomic landscape of COVID-19 autopsies. *Cell*, *184*(3), 775–791.e14. https://doi.org/10.1016/j.cell.2021.01.004

Orsburn, B. C. (2021). Proteome Discoverer-A community enhanced data processing suite for protein informatics. *Proteomes*, *9*(1). https://doi.org/10.3390/proteomes9010015

Ostapcuk, V., Mohn, F., Carl, S. H., Basters, A., Hess, D., Iesmantavicius, V., Lampersberger, L., Flemr, M., Pandey, A., Thomä, N. H., Betschinger, J., & Bühler, M. (2018). Activity-dependent neuroprotective protein recruits HP1 and CHD4 to control lineage-specifying genes. *Nature*, *557*(7707), 739–743. https://doi.org/10.1038/s41586-018-0153-8

Peng, H., Wang, H., Kong, W., Li, J., & Goh, W. W. B. (2023). Optimizing proteomics data differential expression analysis via high-performing rules and ensemble inference. *bioRxiv Doi:10.1101/2023.06.26.546625*. https://doi.org/10.1101/2023.06.26.546625

Pu, S., Wong, J., Turner, B., Cho, E., & Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, *37*(3), 825–831. https://doi.org/10.1093/nar/gkn1005

Quast, J.-P., Schuster, D., & Picotti, P. (2021). Protti: An R package for comprehensive data analysis of peptide- and protein-centric bottom-up proteomics data. *Bioinformatics Advances*, *2*(1), vbab041. https://doi.org/10.1093/bioadv/vbab041

Rue-Albrecht, K., Marini, F., Soneson, C., & Lun, A. T. L. (2018). iSEE: Interactive SummarizedExperiment Explorer. *F1000 Research*, *7*, 741. https://doi.org/10.12688/f1000research.14966.1

Shah, A. D., Goode, R. J. A., Huang, C., Powell, D. R., & Schittenhelm, R. B. (2020). LFQ-Analyst: An Easy-To-Use interactive web platform to analyze and visualize Label-Free proteomics data preprocessed with MaxQuant. *Journal of Proteome Research*, *19*(1), 204–211. https://doi.org/10.1021/acs.jproteome.9b00496

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics

data. *Nature Methods*, *13*(9), 731–740. https://doi.org/10.1038/nmeth.3901

Welte, T., Goulois, A., Stadler, M. B., Hess, D., Soneson, C., Neagu, A., Azzi, C., Wisser, M. J., Seebacher, J., Schmidt, I., Estoppey, D., Nigsch, F., Reece-Hoyes, J., Hoepfner, D., & Großhans, H. (2023). Convergence of multiple RNA-silencing pathways on GW182/TNRC6. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2023.06.001

Wolski, W. E., Nanni, P., Grossmann, J., d'Errico, M., Schlapbach, R., & Panse, C. (2023). Prolfqua: A comprehensive R-Package for proteomics differential expression analysis. *Journal of Proteome Research*, *22*(4), 1092–1104. https://doi.org/10.1021/acs.jproteome.2c00441

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The definitive guide*. Chapman; Hall/CRC. ISBN: 9781138359338

Zhang, X., Smits, A. H., Tilburg, G. B. A. van, Ovaa, H., Huber, W., & Vermeulen, M. (2018). Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nature Protocols*, *13*(3), 530–550. https://doi.org/10.1038/nprot.2017.147