amica: an interactive and user-friendly web-platform for the analysis of proteomics data

Sebastian Didusch [1,2,*], Moritz Madern [1], Markus Hartl [1] and Manuela Baccarini [1]

[1]Center for Molecular Biology, Max Perutz Labs, University of Vienna, A-1030, Vienna, Austria.
[2]Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, A-1030, Vienna, Austria.

Abstract

Summary: Quantitative proteomics has become an increasingly prominent tool in the study of life sciences. A substantial hurdle for many biologists are, however, the intricacies involved in the associated high troughput data analysis. In order to facilitate this task for users with little background knowledge in proteomics, we have developed amica, a freely available open-source web-based software that accepts proteomic input files from different sources and provides quality control, differential expression, biological network and over-representation analysis on the basis of minimal user input. Scientists can use amica interactively to compare proteins across multiple groups, create customized output graphics, and ultimately export the results in a tab-separated format that can be shared with collaborators.
Availability and Implementation: The code for the application, input data and documentation can be accessed online at https://github.com/tbaccata/amica and is also incorporated in the web application. A freely available version of amica is available at https://bioapps.maxperutzlabs.ac.at/app/amica.
Contact: sebastian.didusch@univie.ac.at

1 Introduction

Mass spectrometry (MS)-based proteomics enables deep qualitative and quantitative characterization of any organism's proteome which is crucial for the understanding of the underlying cell biology, physiology and biochemistry. The constant technological advancement of instruments and data-acquisition techniques, as well as the parallel development of a broad variety of methods such as proximity-dependent labeling for the study of protein-protein interactions (PPIs), has expanded the scope and relevance of MS-based approaches in tackling specific biological questions. As a result, proteomic approaches have become increasingly popular, but the complex analysis of MS-based proteomics data is an obstacle for many novices in the field, complicating and delaying the interpretation of experimental outcomes.

Moreover, the many software platforms available for processing proteomics raw data and their different output formats require advanced knowledge for obtaining final interpretable results. Several attempts to solve this problem have been made in the past. The software Perseus (Tyanova *et al.*, 2016b) for example is widely used because it provides a graphical user interface and has extensive options for both label-based and label-free methods. Other software tools revolve around the R programming language, for example MSstats (Choi *et al.*, 2014) or MSnbase (Gatto and Lilley, 2012). These tools have the advantage of automating many of the data processing steps that would have to be performed manually in the Perseus interface, albeit knowledge of the programming language R is required.

Recently developed applications integrate their software into R-Shiny apps, allowing for interactive user interactions and visualizations for output from the MaxQuant package, one of the most widely used software platforms in MS-based proteomics (Tyanova *et al.*, 2016a). LFQ-Analyst (Shah *et al.*, 2019) allows for automatic analysis of label-free data, Eatomics (Kraus *et al.*, 2020) allows for the input of enhanced experimental designs and ProVision (Gallant *et al.*, 2020) can process label-free and TMT labeled data and integrates PPI networks. These tools provide appealing solutions for first-time users but are limited to output from MaxQuant. In addition, none of these tools allow the systematic comparison of proteins across multiple groups or the integration and comparison of multiple proteomics experiments.

As a solution to these issues, we have developed amica, a user-friendly web-based platform for comprehensive quantitative proteomics data analysis that can automatically handle multiple database search tool outputs such as for instance data from FragPipe - a recent but increasingly popular open-source software package (Teo *et al.*, 2020; da Veiga Leprevost *et al.*, 2020) - as well as any generic tab-separated dataset such as RNA-seq data and datasets that were previously analyzed with amica. By allowing the upload of a second amica file to be compared with the current data input, the tool enables multi-omics integration.

2 Results

amica was developed as a user-friendly R-Shiny app with interactive and customizable visualizations that can be exported in a publication-ready vector graphic format.

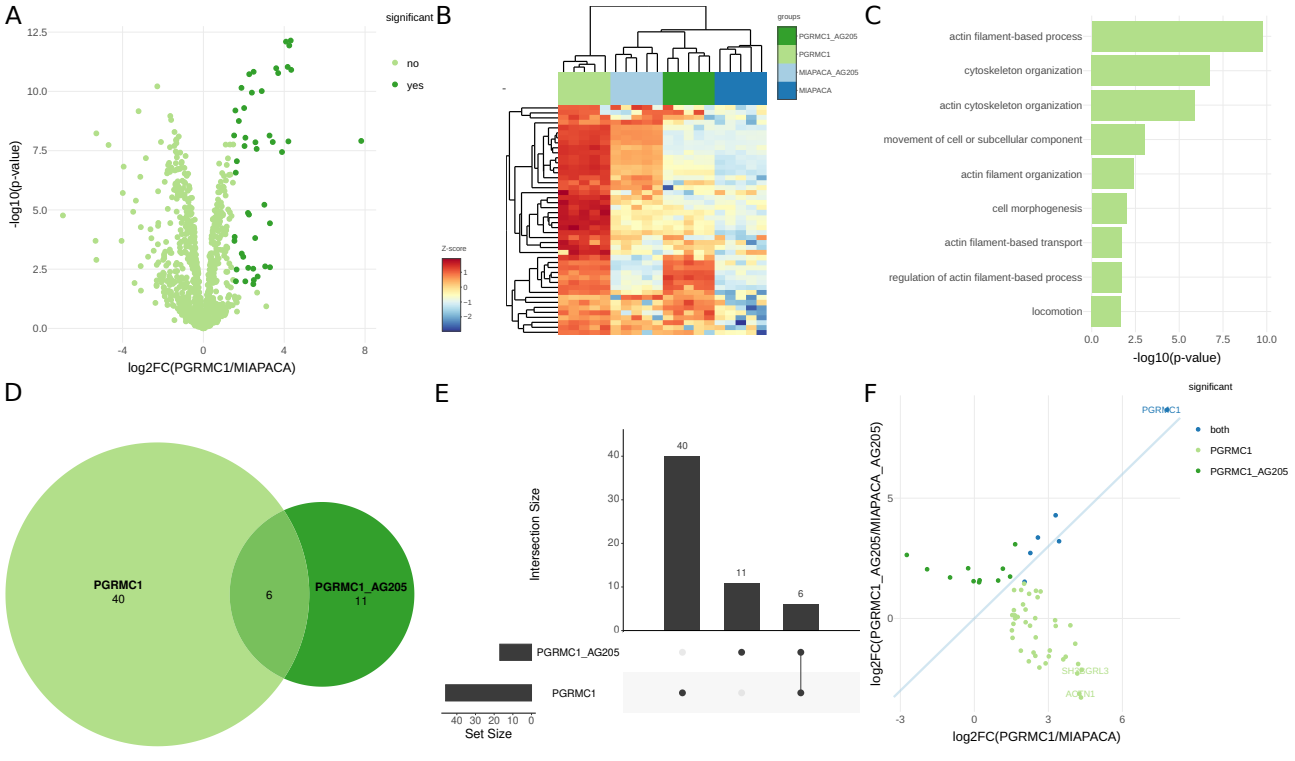

Fig. 1. Examples of amica graphics, produced with the provided example dataset (Teakel et al., 2020). Differentially abundant proteins can be represented with volcano plots (A) or interactive heatmaps (Galili et al., 2018) in (B). Results from functional enrichment in (C). Proteins specific to group comparisons are visualized in Euler diagrams (D), Upset plots (E) or fold-change plots (F).

amica requires three input files: a) a database search result such as MaxQuant's proteinGroups.txt file, FragPipe's combined_proteins.txt file or a custom tab-separated file b) a tab-separated file denoting the experimental design in the experiment, i.e. the mapping of samples to distinct biological groups and c) a tab-separated file containing a contrast matrix that specifies the desired pairwise group comparisons to be made. The parameters for the analysis address filtering by valid values, normalization, imputation of missing values, differential abundance testing with limma (Ritchie *et al.*, 2015) or DEqMS (Zhu *et al.*, 2020) and can be changed in the user interface. The results of the analysis can be downloaded in a custom amica format, which can later be used as the input file for subsequent re-inspection in amica.

Quality control can be visualized for raw intensities, LFQ intensities, normalized and imputed intensities. This includes barplots for coefficient of variations, intensities, contaminants, identified proteins, missing values, most abundant proteins per sample; as well as density plots, boxplots, scatter plots, correlation plots and a PCA plot.

Differentially abundant proteins can be visualized as volcano (Fig. 1A) - and MA - plots for single group comparisons. Unlike other R-Shiny apps for proteomics data analysis amica can generate Euler diagrams (Fig. 1D) and UpSet plots (Fig. 1E) for visualizing the overlap of significant proteins from multiple selected group comparisons. Statistically significant proteins are displayed in a data table and can be further filtered, allowing for the immediate inspection of relevant proteins specific to a comparison in subsequent visualizations. These include fold-change plots (Fig. 1F), profile plots and interactive heatmaps (Fig. 1B).

For human proteomics studies, amica allows users to map significant proteins to a PPI network from the IntAct database (Orchard *et al.*, 2014), and enables the retrieval of subcellular localization predictions from the humancellmap database (Go *et al.*, 2021). As an example, selecting a particular cellular compartment in the interactive web interface will highlight all proteins that map to this localization. This information can be downloaded in gml format for visualization in a network visualization tool such as Cytoscape (Shannon *et al.*, 2003). A functional enrichment analysis of differentially abundant proteins using gprofiler2 (Raudvere *et al.*, 2019) assists in unveiling changes in the underlying biology (Fig. 1C).

Finally, amica accepts and analyzes any generic tab-separated dataset provided a specification file is uploaded that maps the relevant search engine specific columns to a standard format. This feature, combined with the possibility to upload previously analyzed datasets to be compared with the current data input, facilitates multi-omics integration.

3 Conclusions and Outlook

We have developed amica as a general software tool for analyzing and visualizing MS-based proteomics data. amica's user-friendly interface provides a customizable data analysis workflow and its analysis results can conveniently be exported, shared and reloaded into the amica environment for re-inspection at a later time. The data analysis workflow in amica includes quality control and standard differential expression testing, as well as the integration of PPI networks and pathway and gene ontology enrichment analysis for differentially abundant proteins - all of which can help researchers to better interpret the biology underlying the results of their proteomics experiment. The code for the application and online documentation can be found at https://github.com/tbaccata/amica and the software is available at https://bioapps.maxperutzlabs.ac.at/app/amica.

Acknowledgements

We would like to thank all members of the Baccarini lab, the Max Perutz Labs MS facility and IT team and the Menche lab with a special thanks to Joerg Menche and Matthias Baernthaler for their supportive feedback and helpful comments.

Funding

This work was supported by FWF grant W1261, Doctoral Program Signaling mechanism in Cellular Homeostasis, FWF Special Research Program SFB F70 and subproject F7007.

Conflict of interest: None declared.

References

Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., and Vitek, O. (2014). Msstats: an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, **30**(17), 2524–2526.

da Veiga Leprevost, F., Haynes, S. E., Avtonomov, D. M., Chang, H.-Y., Shanmugam, A. K., Mellacheruvu, D., Kong, A. T., and Nesvizhskii, A. I. (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nature methods*, **17**(9), 869–870.

Galili, T., O'Callaghan, A., Sidi, J., and Sievert, C. (2018). heatmaply: an r package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*, **34**(9), 1600–1602.

Gallant, J. L., Heunis, T., Sampson, S. L., and Bitter, W. (2020). Provision: a web-based platform for rapid analysis of proteomics data processed by maxquant. *Bioinformatics*, **36**(19), 4965–4967.

Gatto, L. and Lilley, K. S. (2012). Msnbase-an r/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, **28**(2), 288–289.

Go, C. D., Knight, J. D., Rajasekharan, A., Rathod, B., Hesketh, G. G., Abe, K. T., Youn, J.-Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L. Y., *et al.* (2021). A proximity-dependent biotinylation map of a human cell. *Nature*, pages 1–5.

Kraus, M., Mathew Stephen, M., and Schapranow, M.-P. (2020). Eatomics: Shiny exploration of quantitative proteomics data. *Journal of Proteome Research*, **20**(1), 1070–1078.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., *et al.* (2014). The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, **42**(D1), D358–D363.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, **47**(W1), W191–W198.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47–e47.

Shah, A. D., Goode, R. J., Huang, C., Powell, D. R., and Schittenhelm, R. B. (2019). 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.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**(11), 2498–2504.

Teakel, S. L., Ludescher, M., Thejer, B. M., Poschmann, G., Forwood, J. K., Neubauer, H., and Cahill, M. A. (2020). Protein complexes including pgrmc1 and actin-associated proteins are disrupted by ag-205. *Biochemical and biophysical research communications*, **524**(1), 64–69.

Teo, G. C., Polasky, D. A., Yu, F., and Nesvizhskii, A. I. (2020). Fast deisotoping algorithm and its implementation in the msfragger search engine. *Journal of Proteome Research*, **20**(1), 498–505.

Tyanova, S., Temu, T., and Cox, J. (2016a). The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, **11**(12), 2301–2319.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016b). The perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods*, **13**(9), 731–740.

Zhu, Y., Orre, L. M., Tran, Y. Z., Mermelekas, G., Johansson, H. J., Malyutina, A., Anders, S., and Lehtiö, J. (2020). Deqms: a method for accurate variance estimation in differential protein expression analysis. *Molecular & Cellular Proteomics*, **19**(6), 1047–1057.