msVolcano: a flexible web application for visualizing quantitative proteomics data

Sukhdeep Singh [1*], Marco Y. Hein [2] and A.Francis Stewart[1*]

[1]Genomics, Biotechnology Center, Technische Universitaet Dresden, Tatzberg 47, 01307 Dresden, Germany
[2]Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, CA 94143, USA

Abstract

We introduce msVolcano, a web application, for the visualization of label-free mass spectrometric data. It is optimized for the output of the MaxQuant data analysis pipeline of interactomics experiments and generates volcano plots with lists of interacting proteins. The user can optimize the cutoff values to find meaningful significant interactors for the tagged protein of interest. Optionally, stoichiometries of interacting proteins can be calculated. Several customization options are provided to the user for flexibility and publication-quality outputs can also be downloaded (tabular and graphical).

Availability: msVolcano is implemented in R Statistical language using Shiny and is hosted at server in-house. It can be accessed freely from anywhere at http://projects.biotec.tu-dresden.de/msVolcano/

1 Introduction

The analysis of protein-protein interactions and complex networks using affinity purification or affinity enrichment and mass spectrometry (AP/MS,AE/MS) is one of the most commonly used applications in proteomics. The technology produces high quality protein interaction data [1] and is scalable to proteome-wide levels [2]. Even though isotope labeling methods have been developed to detect and quantify protein-protein interactions [3], label-free approaches are gaining momentum due to their simplicity and applicability [4]. While different quantification strategies exist for label-free data, such as those based on spectral counting, methods that make use of peptide intensities (also known as extracted ion currents) are regarded as the most accurate [5, 6]. Such methods generate quantitative profiles of peptides or proteins across samples, which can be analyzed by established statistical methods, e.g. by a modified t-test across replicate experiments [7].

MaxQuant is an integrated suite of algorithms for the analysis of high-resolution quantitative MS data [8]. Its MaxLFQ module normalizes the contribution of individual peptide fractions and extracts the maximum available quantitative information to calculate highly reliable relative label-free quantification (LFQ) intensity profiles [6], which are exported as tab-limited text files for the downstream analysis.

To identify interactors of a tagged protein of interest (termed the 'bait'), in the presence of a vast number of background binding proteins, replicates of affinity-enriched bait samples are compared to a set of negative control samples. A student's t-test or Welch's test can be used to determine those proteins that are significantly enriched along with the specific baits. A volcano plot is a good way to visualize this kind of analysis [9]. When the resulting differences between the logarithmized mean protein intensities between bait and the control samples are plotted against

*to whom correspondence should be addressed

the negative logarithmic p values derived from the statistical test, unspecific background binders center around zero. The enriched interactors appear on the right section of the plot, whereas ideally no proteins should appear on the left section when compared to an empty control (because these would represent proteins depleted by the bait). The higher the difference between the group means (i.e. the enrichment) and the p value (i.e. the reproducibility), the more the interactors shift towards the upper right section of the plot, which represents the area of highest confidence for a true interaction.

Though label-free methods are nowadays as accurate as the isotope-based methods, false positives may appear enriched alongside the true positive interaction partners [10]. Identifying background proteins and defining a threshold that separates these from true interactors is a critical step during data analysis, and often benefits from some manual optimization.

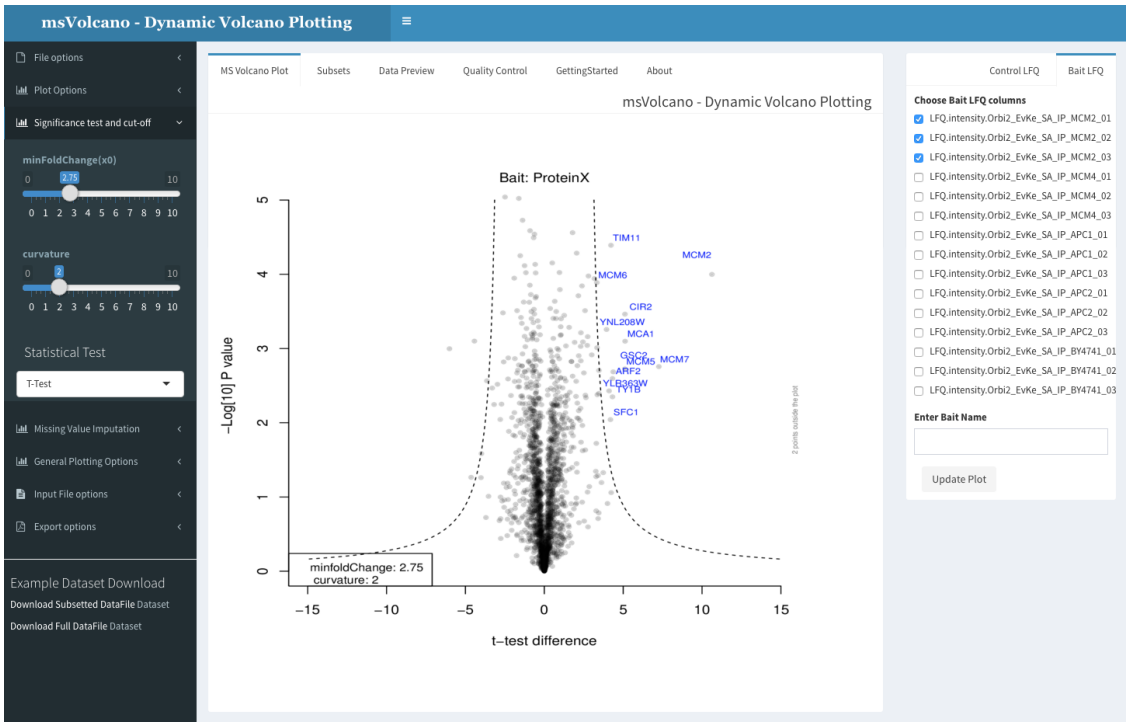

Figure 1: The interface is divided into three sections, sidebar, body panel and column selection panel (left to right). Sidebar provides an access to the file upload, plot aesthetics, cutoff parameters, missing data imputation, stoichiometry and the export options. The body panel has five different tabs, where the default panel labeled as "MS Volcano Plot" displays the volcano plot. Second tab, "Subsets" displays the filtered input data for the significant interactors. "Data Preview" tab displays the user inputted data for scrutiny, "GettingStarted" and "About" tab display the specific and general information about the web interface. When user uploads a file or enters a ftp link, all LFQ columns are scanned and displayed in the column selection tab on the right side. User now selects respective bait and control columns (minimum two) and optionally enters the name of bait in the provided text box. As the 'Update Plot' button is pressed, the plot is generated simultaneously.

2 Description

To facilitate the analysis and presentation of AE-MS data, we present msVolcano, which is a user modulated, freely accessible web application. It requires the MaxQuant output of an interaction dataset that was analysed using the MaxLFQ module. LFQ intensity profiles retain the absolute scale from the original summed- up peptide intensities [6], serving as a proxy for absolute protein abundance. The purpose of msVolcano is to implement all steps of downstream data analysis into a simple and intuitive user interface that requires no bioinformatics knowledge or specialized

software. To this end, msVolcano automatically extracts relevant data columns, filters out hits to the decoy database and potential contaminants and imputes missing values by simulated noise. A visual Quality Control (QC) output is generated allowing the user to monitor the correlation between replicates, fraction of missing values and behaviour of the population of imputed values.

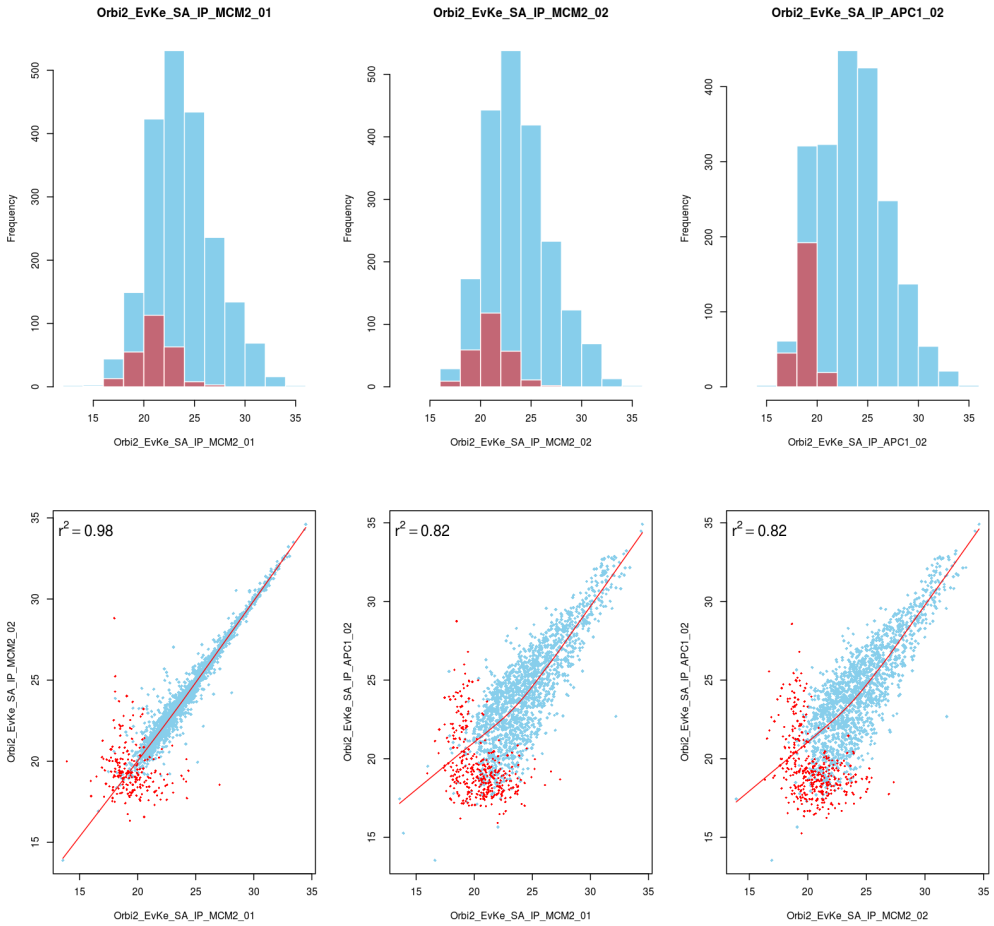

Figure 2: Default QC plot using a dataset from budding yeast study (sample data in msVolcano) [11] A) top row displaying the distribution of the raw values (LFQ intensites - in blue) overlayed with the distribution of imputed values (in red) per LFQ column selected B) 2x2 scatter plots between the chosen LFQ columns with local regression(lowess) displayed as a red line with pearson's correlations coefficient. For the visual aestheticity, the number of scatter plots are restricted to the number of histograms displayed above them.

A user-defined statistical test is then performed between selected bait and control samples and the tool generates a volcano plot. We implemented a recently introduced hyperbolic curve threshold [11], based on the given formula

$$y = \frac{c}{x - x0} \tag{1}$$

where c = curvature, x0 = minimum fold change, thus dividing enriched proteins into mildly and strongly enriched [11]. The cutoff parameters can be adjusted by the user and monitored by the graphical output. The user has access to the plot aesthetics and can view the original input file and its subset for significant interactors in the inbuilt browser. A publication-quality PDF plot can be generated and exported along with the subset of original data limited to the significant interactors. Next to the identities of interacting proteins, their stoichiometries relative to their bait is crucial for the understanding of the molecular function of protein complexes [2, 12]. Thus, optional stoichiometry calculations have been implemented in the code. We use a modified version of intensity-based absolute quantification (iBAQ) [13] for the estimation of protein abundance for

stoichiometry calculations, where LFQ intensities are normalized by the number of theoretical tryptic peptides between 7 and 30 amino acids, as described [2](Fig 1b). Theoretical peptides are pre-calculated for the most commonly used proteomes of model organisms and are matched based on the proteins' uniprot IDs. Stoichiometry calculations are based on the given formula

$$\mathrm{st(i)} = \frac{\mathrm{sIp(i)}}{\mathrm{sIp(bait} \text{ — } \mathrm{most\_abundant)}} \tag{2}$$

where

$$\mathrm{sIp(i)} = \frac{\mathrm{Ip(i)}}{\mathrm{trp(i)}} \tag{3}$$

where

$$\mathrm{Ip(i)} = \mathrm{mean(Ip(case))\text{-}mean(Ip(control))} \tag{4}$$

where st = stoichiometry, sIp = size normalised protein intensity,
Ip = protein intensity, trp = number of theoretical peptides of protein

3 Conclusion

msVolcano provides a web-platform for the quick visualization of label-free mass spectrometric data and can be freely accessed globally. With the underlying hyberbolic curve parameters and other statistics, user can intuitively isolate the true protein interaction partners from the false positives, without the need of writing a code. With its ftp file input support, the user can quickly analyse and re-analyse the results of the interactomics experiment present on their own cloud servers and along with the calculated optional stoichiometries, all the results can be exported in publication quality tabular or graphical format.

Funding

This work was supported by the EU 7th Framework integrated project, SyBoSS (www.syboss.eu).

References

[1] Loic Royer, Matthias Reimann, A Francis Stewart, and Michael Schroeder. Network compression as a quality measure for protein interaction networks. *PloS one*, 7(6):e35729, 2012.

[2] Marco Y Hein, Nina C Hubner, Ina Poser, Jürgen Cox, Nagarjuna Nagaraj, Yusuke Toyoda, Igor A Gak, Ina Weisswange, Jörg Mansfeld, Frank Buchholz, et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–723, 2015.

[3] Shao-En Ong and Matthias Mann. A practical recipe for stable isotope labeling by amino acids in cell culture (silac). *Nature protocols*, 1(6):2650–2660, 2006.

[4] Stephen Tate, Brett Larsen, Ron Bonner, and Anne-Claude Gingras. Label-free quantitative proteomics trends for protein–protein interactions. *Journal of proteomics*, 81:91–101, 2013.

[5] Hyungwon Choi, Timo Glatter, Mathias Gstaiger, and Alexey I Nesvizhskii. Saint-ms1: protein–protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *Journal of proteome research*, 11(4):2619–2624, 2012.

[6] Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular & Cellular Proteomics*, 13(9):2513–2526, 2014.

[7] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[8] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

[9] Nina C Hubner, Alexander W Bird, Jürgen Cox, Bianca Splettstoesser, Peter Bandilla, Ina Poser, Anthony Hyman, and Matthias Mann. Quantitative proteomics combined with bac transgeneomics reveals in vivo protein interactions. *The Journal of cell biology*, 189(4):739–754, 2010.

[10] H Christian Eberl, Cornelia G Spruijt, Christian D Kelstrup, Michiel Vermeulen, and Matthias Mann. A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Molecular cell*, 49(2):368–378, 2013.

[11] Eva C Keilhauer, Marco Y Hein, and Matthias Mann. Accurate protein complex retrieval by affinity enrichment mass spectrometry (ae-ms) rather than affinity purification mass spectrometry (ap-ms). *Molecular & Cellular Proteomics*, 14(1):120–135, 2015.

[12] Arne H Smits, Pascal WTC Jansen, Ina Poser, Anthony A Hyman, and Michiel Vermeulen. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic acids research*, page gks941, 2012.

[13] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.