Proteomics data analysis

Updated 10/02/2020

Please email us at EICC@emory.edu with any questions or comments regarding data analysis. For proteomics data acquisition please contact Emory Integrated Proteomics Core (EIPC) EIPC@emory.edu

1. Description of Steps

All Proteomics projects start with the following steps:

Raw file processing

All raw files will be processed by EIPC in MaxQuant [1]. MaxQuant reports summed intensity for each protein, as well as its iBAQ value. Proteins which share all identified peptides will be combined into a single protein group. Peptides which match multiple protein groups ("razor" peptides) are assigned to the protein group with the most unique peptides. MaxQuant employs the proprietary MaxLFQ algorithm for label-free quantitation (LFQ). Quantification will be performed using razor and unique peptides, including those modified by acetylation (protein N-terminal), oxidation (Met) and deamidation (NQ).

Descriptive and hypothesis-testing statistical analysis

All analysis will be performed by the EICC using the R package Proteus[2]. Proteus implements the algorithm for differential expression used in the R package limma (Linear Models for Microarray)[3]. While originally developed for microarray data, limma has been successfully used in numerous transcriptomic and proteomic study publications. Raw intensity data provided by Max Quant (protein group files) and information about the samples (provided by the client) are read into Proteus. If applicable, LFQ intensities are normalized with a log2 transformation. Next, intensities are fit into a linear model for each protein. This test allows for simple pairwise comparisons and more complex comparisons which account for covariates like sex, age, ethnicity, or other variables of interest. Then, the standard errors of each model are smoothed using an empirical Bayes estimation to determine differential expression between groups of interest. Finally, the user is supplied with the results of the differential expression analysis and provided output with which to visualize the findings. These include distance matrices, volcano plots, or customized graphics available upon request. Thresholds for statistical significance of proteins will vary by the project and can be set upon request but typically range from nominal p-values between 0.05-0.1 or adjusted p-values/FDR 0.01-0.2.

For detailed walkthroughs of the code used for differential expression analysis of proteomics data, please see our GitHub page.

2. Example Methods section

Below are examples of Methods sections for manuscripts using these methodologies. Please feel free to adapt them in your own words using your project-specific parameters.

Raw files are processed by MaxQuant with parameters best fitting the data [1]. MaxQuant employs the proprietary MaxLFQ algorithm for label-free quantitation (LFQ). Quantification will be performed using razor and unique peptides, including those modified by acetylation (protein N-terminal), oxidation (Met) and deamidation (NQ). The resulting protein groups information is read in R and analyzed using Proteus to determine differentially expressed proteins between groups. Next, LFQ intensities of each sample are log-2 transformed and compared using a linear model. Nominal p-values are transformed using the Benjamini-Hochberg correction to account for multiple hypothesis testing. Proteins considered to be significantly differentially expressed are those with adjusted p-values, also called False Discovery Rates (or FDR) controlled at a specific threshold. Graphs are generated using package-specific recommended data transformations and vary by project. Please see our Data Visualization Menu for some examples of graphs we have made in the past.

3. Required Files (provided to EIPC)

1. Raw data files to be put into MaxQuant

2. Metadata spreadsheet with information about samples, replicates, groupings of interest etc..

4. Deliverables from EICC

1. A power point presentation including all agreed upon analyses and results with interpretations

2. Excel spreadsheets of raw counts and complete differential expression analysis

3. Image files of selection of agreed upon number of graphs/plots with interpretations
    - volcano plots, PCA plots, MA plots, heatmaps etc.

5. References

1. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

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

3. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, (2015).