Statistical Analysis and Visualization of Mass Spectrometry Data using R

Joseph Hennessey, Linda Berg Luecke, Rebekah L. Gundry,
Department of Cellular and Integrative Physiology, University of Nebraska Medical Center, Omaha, NE 68198

Background

Mass spectrometry (MS) identifies the mass-to-charge ratio of ions. In proteomic research, MS can be used to identify proteins in a sample by digesting proteins into peptides, ionizing the peptides, then detecting the peptides and fragments of peptides via a mass spectrometer. However, most proteomic experiments identify large numbers of proteins and at times it can be difficult to efficiency communicate results of large datasets. Therefore, we sought to apply various visualization techniques in R to allow for fast and effective processing of large MS datasets.

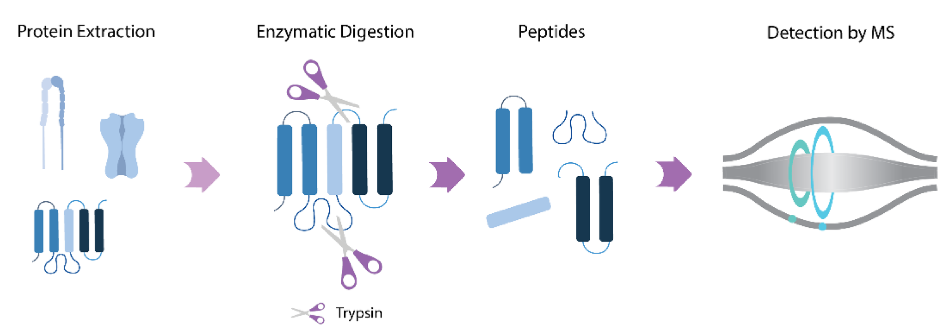

Figure 1. Overview of sample preparation and detection of peptides by mass spectrometry. Proteins are digested into peptides. Peptides are ionized and detected via a mass spectrometer.

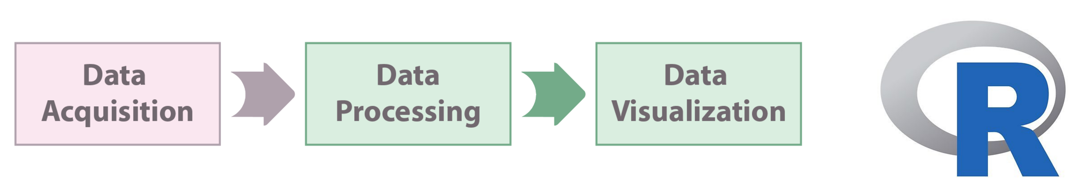

Figure 2. Flowchart overview showing the pipeline from data acquisition to data visualization. After data has been acquired, data is processed and visualized using R.

Goals

1. Develop efficient methods for visualizing mass spectrometry data.
2. Promote accessibility by using colorblind-safe color palettes in graphs.

What is R?

R is a language developed for statistical computing and graphics. R allows user to effectively handle and manipulate data and allows for facilitation of data analysis and display. It is a well-developed and simple programming language that can be easily applied to large MS datasets.
Packages used in this project include:

- ggplot2: creation of compelling visualizations
- dplyr: analysis, manipulation, and grouping of large datasets
- viridis: sequential color palettes designed to maintain perceptual uniformity and remain colorblind-safe
- Rcolorbrewer: qualitative colorblind-safe palettes

Colorblind-Safe Color Maps

Many color maps are not "colorblind-safe", leading to issues with perception for people with colorblindness. The viridis package in R designed colorblind-safe color maps, and primarily utilize the blue-yellow axis to convey changes in color.

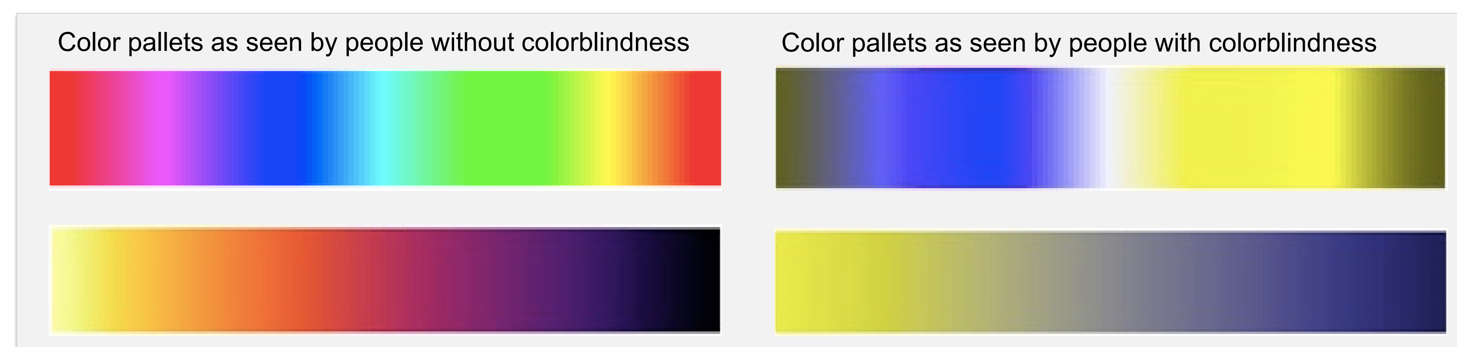

Figure 3. Color pallets as seen by people with and without colorblindness.

Visualization of MS Data: Heatmap

Dataset consisted of:

- 3 groups: control, virus, virus gene deletion
- 3 biological replications for each group
- 2 technical replicates for each biological replicate

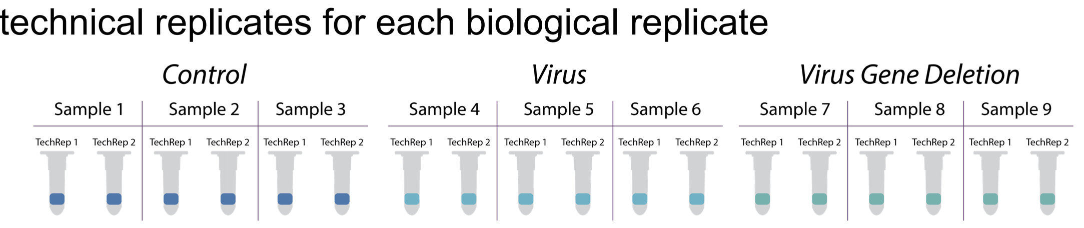

Figure 4. Overview of proteomics dataset. The dataset consisted of the control, virus, and virus gene deletion groups. Each groups contains 3 biological replicates.

Heatmaps are used to cluster proteins by similarity and show changes in abundances across experimental groups

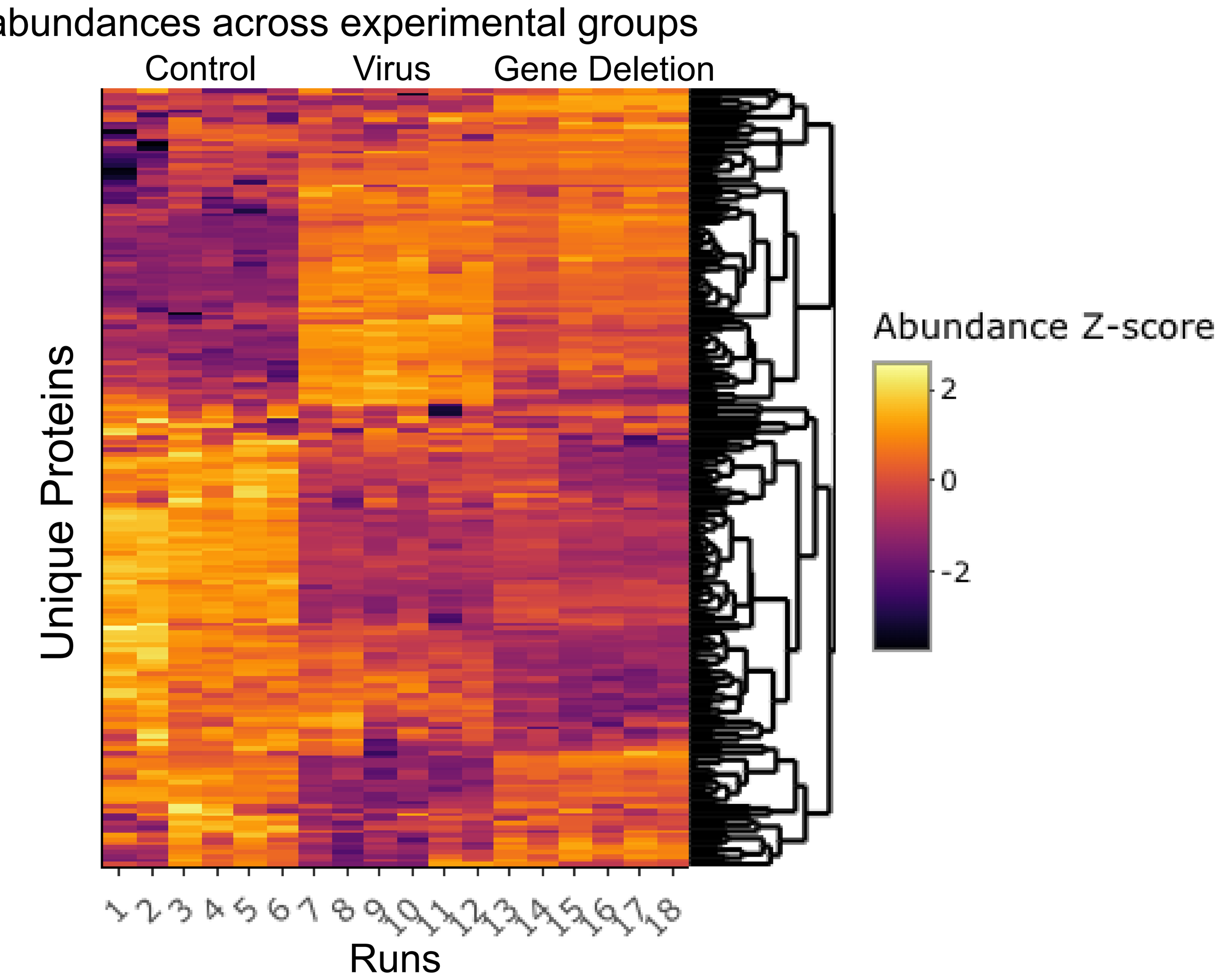

Figure 5. Heatmap of protein abundances of 18 technical replicates. The viridis "inferno" color map is used. Proteins are clustered using Optimal Leaf Ordering. An interactive graphic is available to the user locally, and gives individual protein names, run number, and abundance.

Visualization of MS Data: Volcano Plot

A volcano plot is a type of scatter-plot that shows magnitude of change (fold change) versus statistical significance (p value). It allows for visual identification of proteins with large fold changes in abundances.

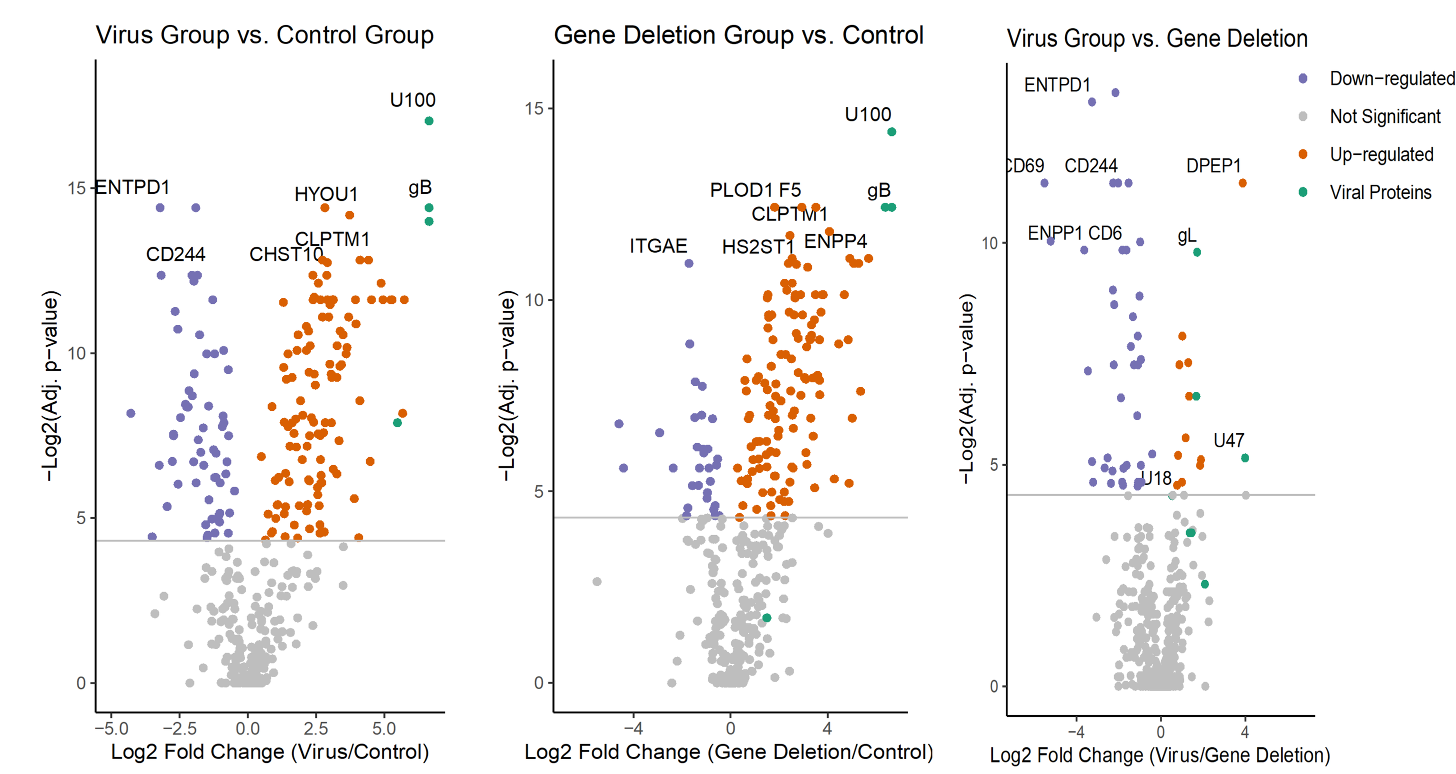

Figure 6. Pairwise volcano plots of log2 Fold Change vs. adjusted p-value across 3 experimental groups. P-values were adjusted to control for an FDR of 0.05 using the Benjamini-Hochberg correction, and a horizontal grey line reflects the significance threshold. Selected human and viral protein gene names are depicted in the figure. Viral proteins were used as an additional control in the experiment. The plots show that there were few viral proteins in the control samples, which is consistent with our expectations.

Conclusions

- Heatmaps are useful for evaluating changes in "clusters" of proteins across experimental groups and replicates
- Volcano plots allow for quick identification of proteins with high or low abundances in different experimental conditions
- Use of colorblind-safe color palettes is practical and easy to implement

Future Directions

- Optimize the clustering functions used in the heatmap to cluster proteins by biological similarity
- Explore other visualization methods, such as Principal Component Analysis (PCA)