Metabolomic data processing and feature selection pipeline

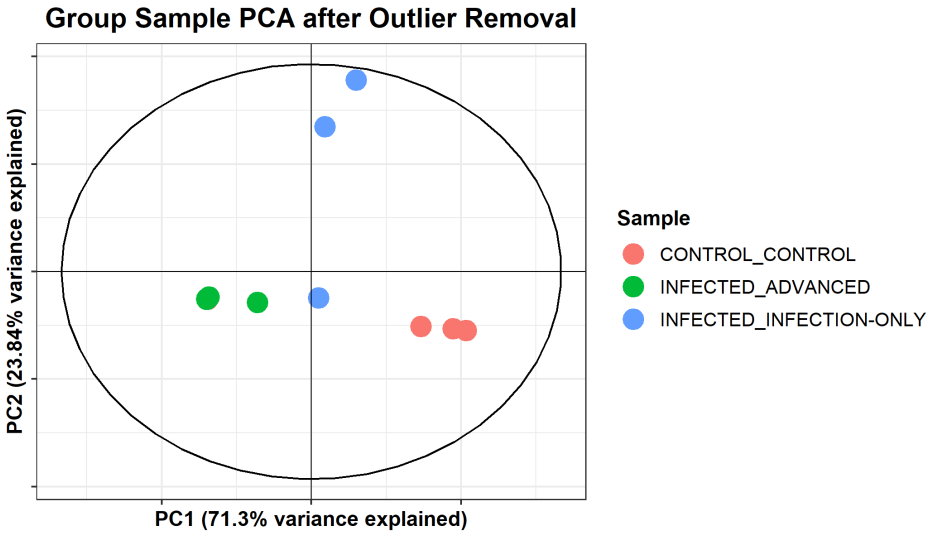
Seth Rhoades
Fulgens Consulting, LLC

Key technical elements in this work:

- Raw LC-MS metabolomic data processing, feature extraction and normalization
- Dimensionality reduction and machine learning-guided candidate feature selection

*** Code for this analysis is available at www.fulgensconsulting.com ***

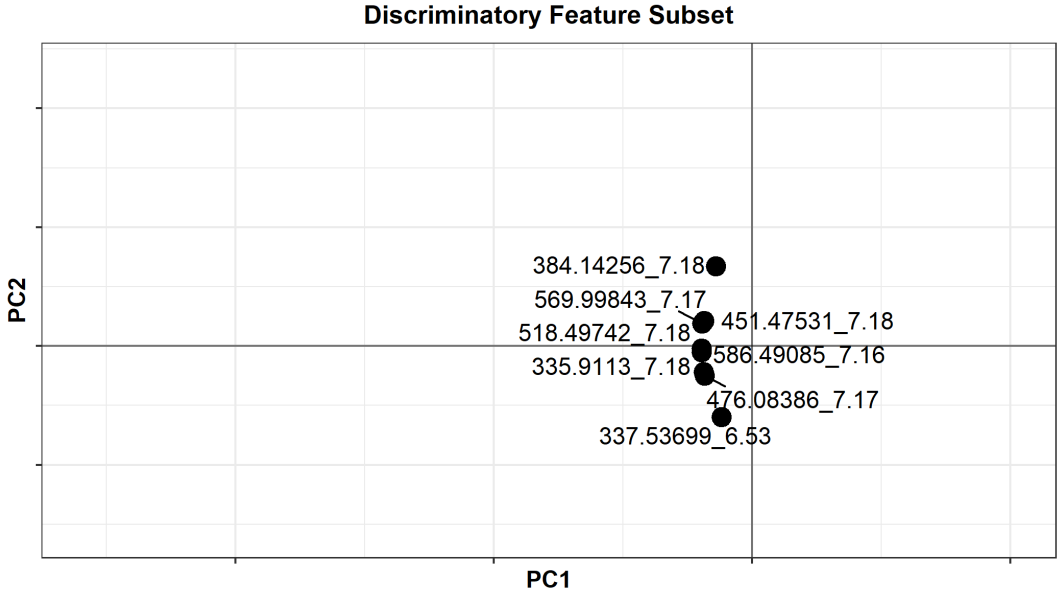
The hardware advancements in analytical profiling, including that used in both metabolomics and proteomics, allow an ever greater capture of chemical space. Such breadth of data inevitably requires computational means of processing and knowledge extraction. As with other 'omic technologies, bioinformatics for metabolomics minimally requires raw data processing and normalization prior to statistical analyses and biologically-relevant interpretations. This sample workflow covers a few of these bioinformatic processes for high-resolution, untargeted metabolomics data. In particular, metabolomics acquired through Orbitrap instrumentation is increasingly popular, but this data is large, complex, and may require bespoke processing methods from the underlying .raw data files.



Principal component scores on mz:rt features, after removal of one sample which fell outside the Hotellings T2 distribution. Samples are colored by group. The first Principal component contains variance across the control group through advanced infections.

Raw LC-MS data from a Thermo Q-Exactive Orbitrap mass spectrometer is distilled into distinct features based on mass and retention times ("mz:rt").. Ion counts for these features are summed over expected peak widths across samples, and scaled by total sample signal intensities. Note that this bespoke processing method adopts a relatively quick and simple approach to peak alignment, although more statistically rigorous approaches, including

regression and probabilistic feature assignment are possible. These mz:rt predictive variables are then fit to principal components analysis, to both delineate the variance amongst study groups, and to flag and remove outlier samples prior to downstream analysis. After outlier removal, Ridge regression is performed to identify a small subset of features which drive most of the study variance. These mz:rt features are then candidates for additional analysis and metabolite identification.



Principal component loadings on mz:rt features selected from Ridge regression variable selection. These mz:rt features are candidates for further exploration and identification.

References

This data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, https://www.metabolomicsworkbench.org where it has been assigned Project IDPR000646 . The data can be accessed directly via it's Project DOI: 10.21228/M85M3S This work is supported by NIH grant U2C-DK119886.

Kockmann, T and Panse, C. The rawrr R Package: Direct Access to Orbitrap Data and Beyond. J Prot Res. 2021 20 (4), 2028-2034. DOI: 10.1021/acs.jproteome.0c00866