Protstatmd: A NextFlow Containerized Analysis Pipeline for Spectral Count Proteomic Analysis Doubles the Number of Pairwise Comparisons between Beer Samples

Jordan B. Burton [1,2], Nicholas J. Carruthers [3], Paul M. Stemmer [2]

1.Wayne State University, Department of Chemistry, Detroit, MI

2. Wayne State University, Institute of Environmental Health Sciences, Detroit, MI

3. University of Michigan, Bioinformatics Core Facility, Ann Arbor, MI

PROJECT GOAL

The default proteomicsLFQ Nextflow workflow uses area under the curve (AUC) as a measure of abundance and MSstats to evaluate pairwise comparisons. Unquantified proteins are treated as missing values. Spectral counting includes proteins with Peptide Spectral Matches (PSMs) that are missing AUC information thus allowing statistical assessment. protstatmd was appended to the proteomicsLFQ workflow to facilitate installation of common R packages and computing environments to produce interactive html documents using RMarkdown. Protstatmd performs statistical analysis of spectral count data enabling comparisons lacking AUC measurements. Beer metaproteomic studies are evaluated in the proteomicsLFQ workflow. Effects of yeast, hops, grains and brewing conditions on the beer proteome are shown.

Mass Spectrometry

- Orbitrap Fusion Tribrid & nLC-1000 (Thermo Fisher Scientific)
  - Data Dependent Analysis
  - 60 min acquisition
  - 3 beer samples x 3 replicates = 9 samples
    - PRG = Proteomics Research Group Lager (Control)
    - BBA = Bourbon Barrel Aged
    - IFHA = Imperial Farmhouse Ale

Proteomics Nextflow Pipeline

Nextflow is a command line tool that links containers containing the scripts and computing environments for:

- Raw File Conversion
- Database Searching
  - MS-GF+ Algorithm
  - Comet Algorithm
- False Discovery Rate Correction
  - Percolator Algorithm
- Statistical Analysis
  - MSstats (proteomicsLFQ default)
  - EdgeR (Protstatmd)

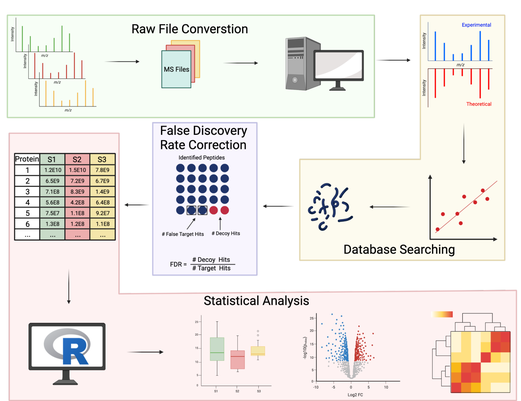

Figure 1: Diagram of a Nextflow workflow for proteomic analysis.

Database Search Results

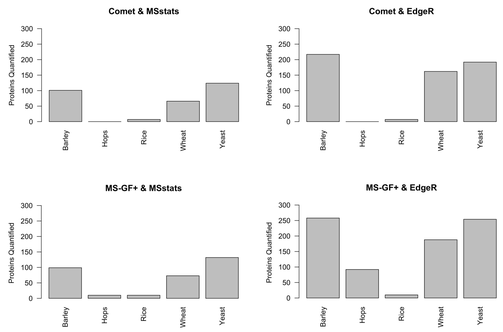

Figure 2: Total proteins quantified for each species in three beer samples.

- Using EdgeR with Comet or MS-GF+ doubles the number of proteins quantified in all beer samples for four of the five database searches compared to MSstats.
- The MS-GF+ database search increased the number of quantified proteins 6 - 30% more than Comet for yeast, barley, hops, and wheat database searches.
  - 258, 92, 10, 188, and 254 proteins were quantified after searching the data against individual barley, hops, rice, wheat, and yeast databases using MS-GF+ with EdgeR.

MSstats Statistical Analysis

- MSstats is incorporated into the default proteomicsLFQ Nextflow pipeline to identify differentially abundant proteins using protein AUC values.
  - The greatest number of differentially abundant proteins (q-value < 0.1) were from yeast for the BBA vs. IFHA (101 proteins), BBA vs. PRG (109 proteins), and IFHA vs. PRG (82) comparison.

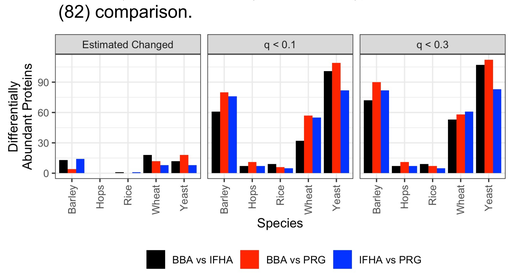

Figure 3: The number of differentially abundant barley, hops, rice, wheat, and yeast proteins are shown for each comparison. Differentially abundant proteins are categorized into: Estimate Changed, q-value < 0.1, or q-value < 0.3.

- Yeast, barley and wheat proteins are expected to have the largest number of differentially abundant proteins between beer types due to differences in brewing ingredients and the yeast used for fermentation.

EdgeR Statistical Analysis

- EdgeR was incorporated into the proteomicsLFQ Nextflow pipeline with the prostatmd package to identify differentially abundant proteins using spectral count data.
  - The most differentially abundant proteins at q-value < 0.1 were yeast proteins for the BBA vs. IFHA (92 proteins) and BBA vs, PRG (169 proteins) comparisons and barley proteins for the IFHA vs. PRG (81 proteins) comparison.

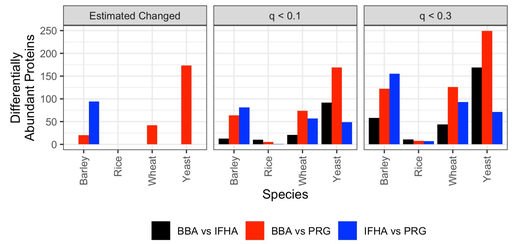

Figure 4: The number of differentially abundant barley, rice, wheat, and yeast proteins are shown for group comparisons. Differentially abundant proteins are categorized into: Estimate Changed, q-value < 0.1, and q-value < 0.3.

- p-value histograms are exported by the protstatmd Nextflow pipeline to validate the number of differentially abundant proteins.
  - The distribution of p-values is skewed to the right as is expected because three different styles of beer are compared in this study and each style is expected to have many differentially abundant proteins when compared to another style.

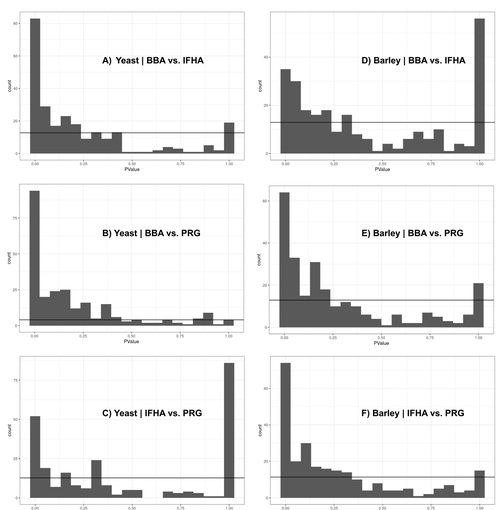

Figure 5: p-value histograms are shown.

Fold Changes

- The fold change for each comparison is exported by the protstatmd Nextflow pipeline in the form of a volcano plot.
  - Volcano plots are interactive in the exported document, allowing users to hover over each point and identify proteins of interest.
  - There are more differentially abundant yeast proteins in each comparison than non-differentially abundant yeast proteins.
  - There are more non-differentially abundant barley proteins in each comparison than differentially abundant barley proteins.

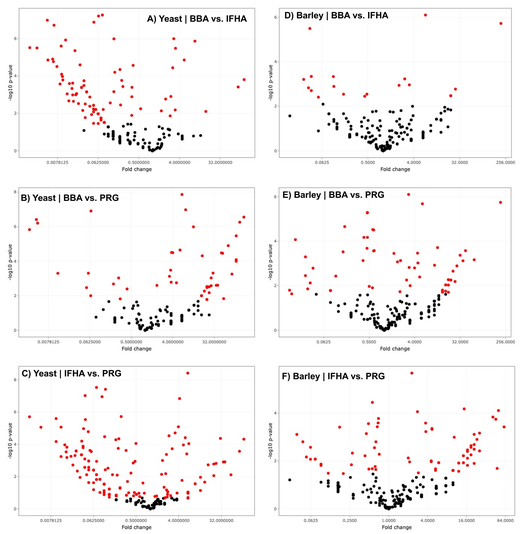

Figure 6: Volcano plots displaying the fold change.

Conclusions

- More proteins were quantified with MS-GF+ and EdgeR in the protstatmd Nextflow workflow
- More differentially abundant proteins were identified with EdgeR as compared to MSstats
  - Spectral count data allowed more comparisons to be made between beers as there were many proteins that were missing AUC data.

ACKNOWLEDGEMENTS

- National Institute of Health (P30 ES020957, P30 CA022453, and S10 OD010700)
- Wayne State University Funding
- Wayne State University High Performance Computing Grid