Reproducible and robust quantitative functional analysis of metaproteomes using the Galaxy platform

Caleb Easterly[1], Joel Rudney[1], James Johnson[1], Carolin Kolmeder[2], Andrea Argentini[4], Thomas McGowan[1], Björn Grüning[3], Praveen Kumar[1], Subina Mehta[1], Lennart Martens[4], Tim Griffin[1] and Pratik Jagtap[1]

[1]University of Minnesota, Minneapolis
[2]Max Planck Institute for Developmental Biology, Tübingen, Germany
[3]University of Freiburg, Freiburg, Baden-Württemberg
[4]Ghent University, Ghent, Belgium

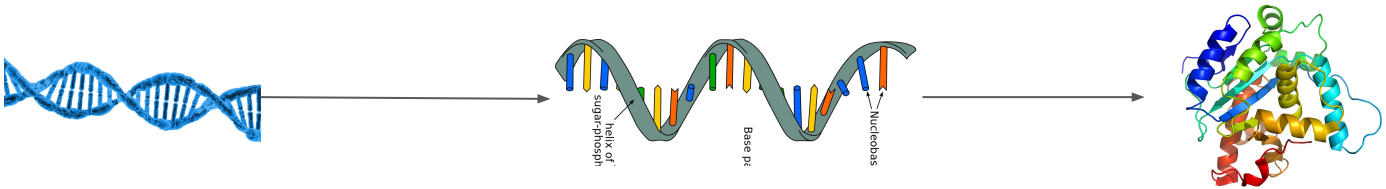
UNIVERSITY OF MINNESOTA
Driven to Discover℠

Outline

1) Introduction to metaproteomics
2) Why look at differentially expressed proteins in microbiota?
3) Steps in the analysis pipeline
4) Galaxy as a workflow engine
5) Case study: the oral microbiome and a model of a sugar-heavy diet
6) Future directions

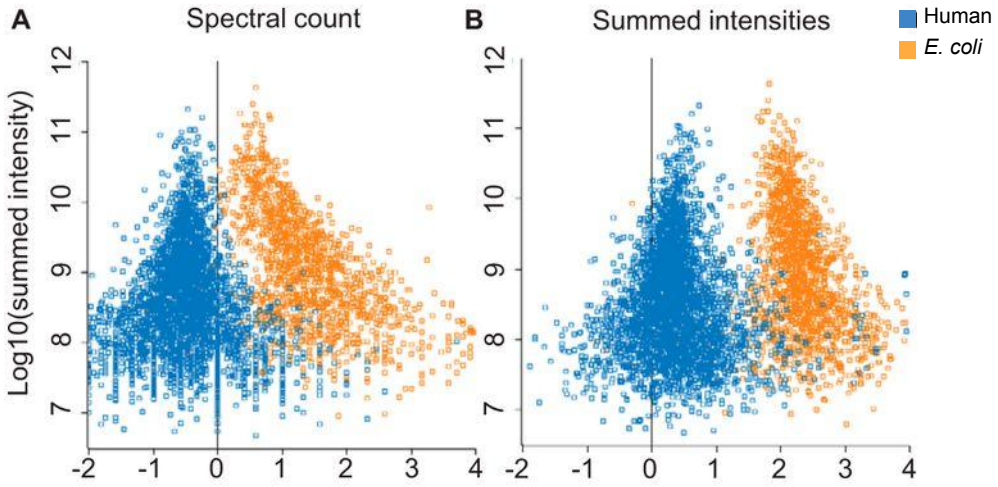
Metaproteomics

- Take a snapshot of activities that the microbiome is carrying out at moment of analysis



Quantitation in Metaproteomics

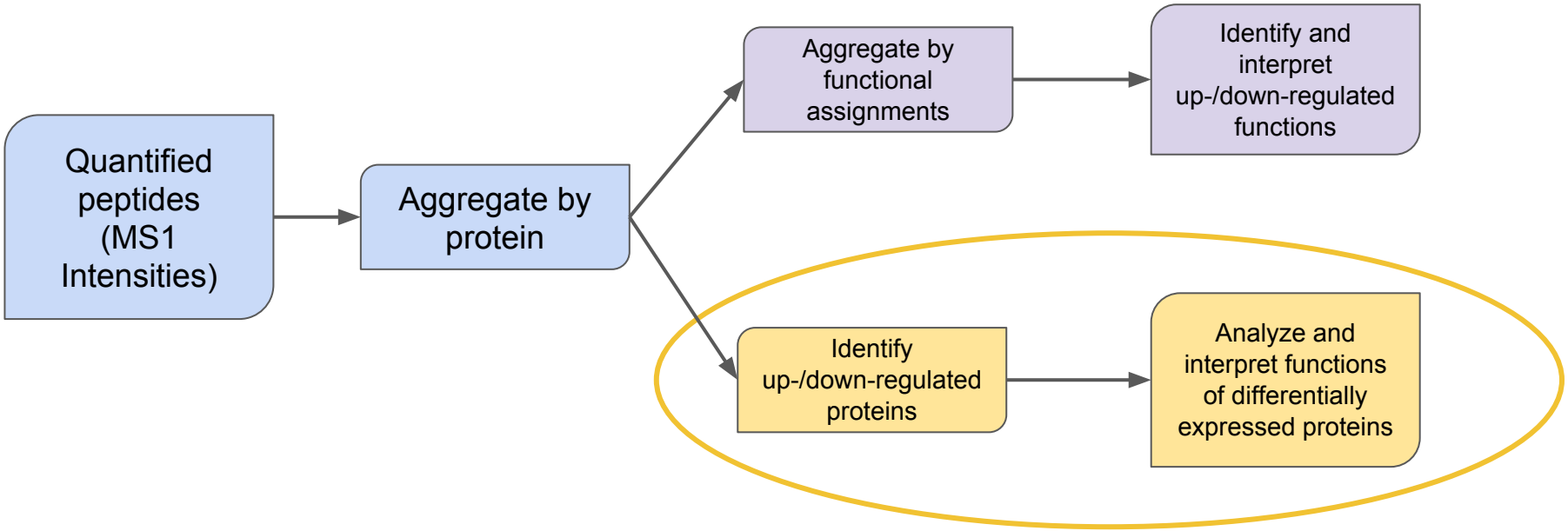
- Metaproteomic studies are often qualitative
- Some studies use spectral counts
- Precursor (MS1) intensity can produce more accurate fold changes estimates than spectral counts (Cox, et al. 2014, *Molecular & Cellular Proteomics*)
- However, spectral counts vs. precursor intensity is controversial



From Cox, et al. 2014 *Molecular & Cellular Proteomics*.

E. coli and human proteomes mixed at predetermined ratios; *E. coli* proteins were differentially expressed and human proteins were constant

Two methods for quantitative functional analysis



Differential expression protein analysis

- **Differentially expressed protein:** has systematically higher abundance in one condition versus another condition
- DE analysis is common in single-organism proteomics and transcriptomics
  - e.g. biomarker discovery
- In metaproteomics, can identify functioning of microbiome in different scenarios
  - Examples:
    - Oral microbiome: high-sugar diet versus low-sugar diet
    - Gut microbiome: before and after treatment with antibiotics

Galaxy and Galaxy-P

Galaxy

Galaxy: open source, freely available web platform for accessible bioinformatic analysis

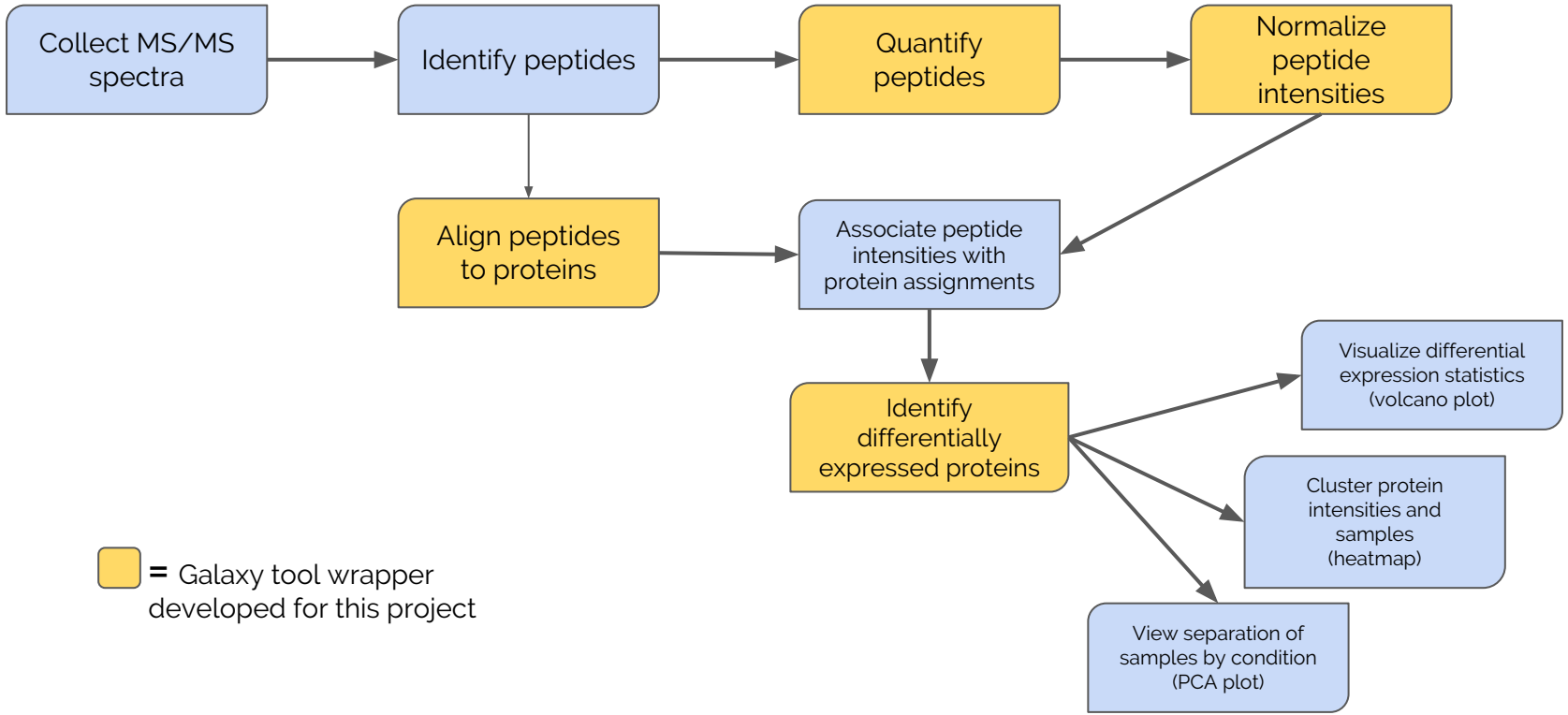
GalaxyP

Galaxy-P: based at University of Minnesota. Develops tools for proteomic data analysis within Galaxy.

Why Galaxy?

- Graphical interface
- Use existing software by "wrapping" it
- Develop custom Galaxy tools to accomplish specific tasks
- Software can be linked together in a **workflow**, a Galaxy object that can be reused and shared.

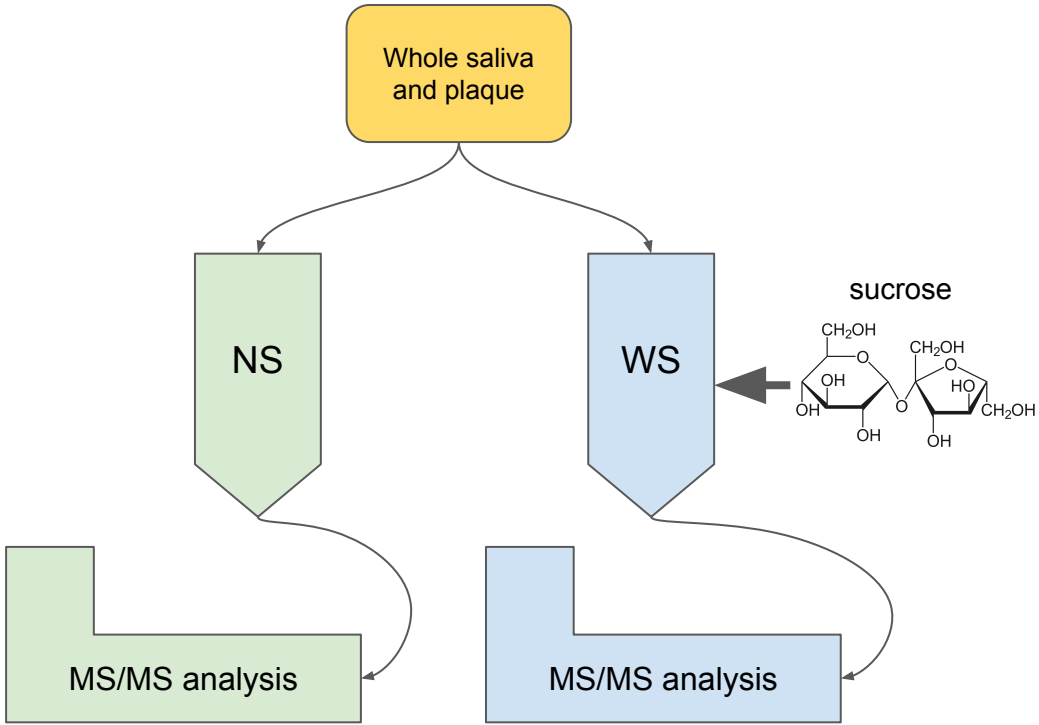
The Analysis Pipeline



Case study: sucrose and the oral microbiome

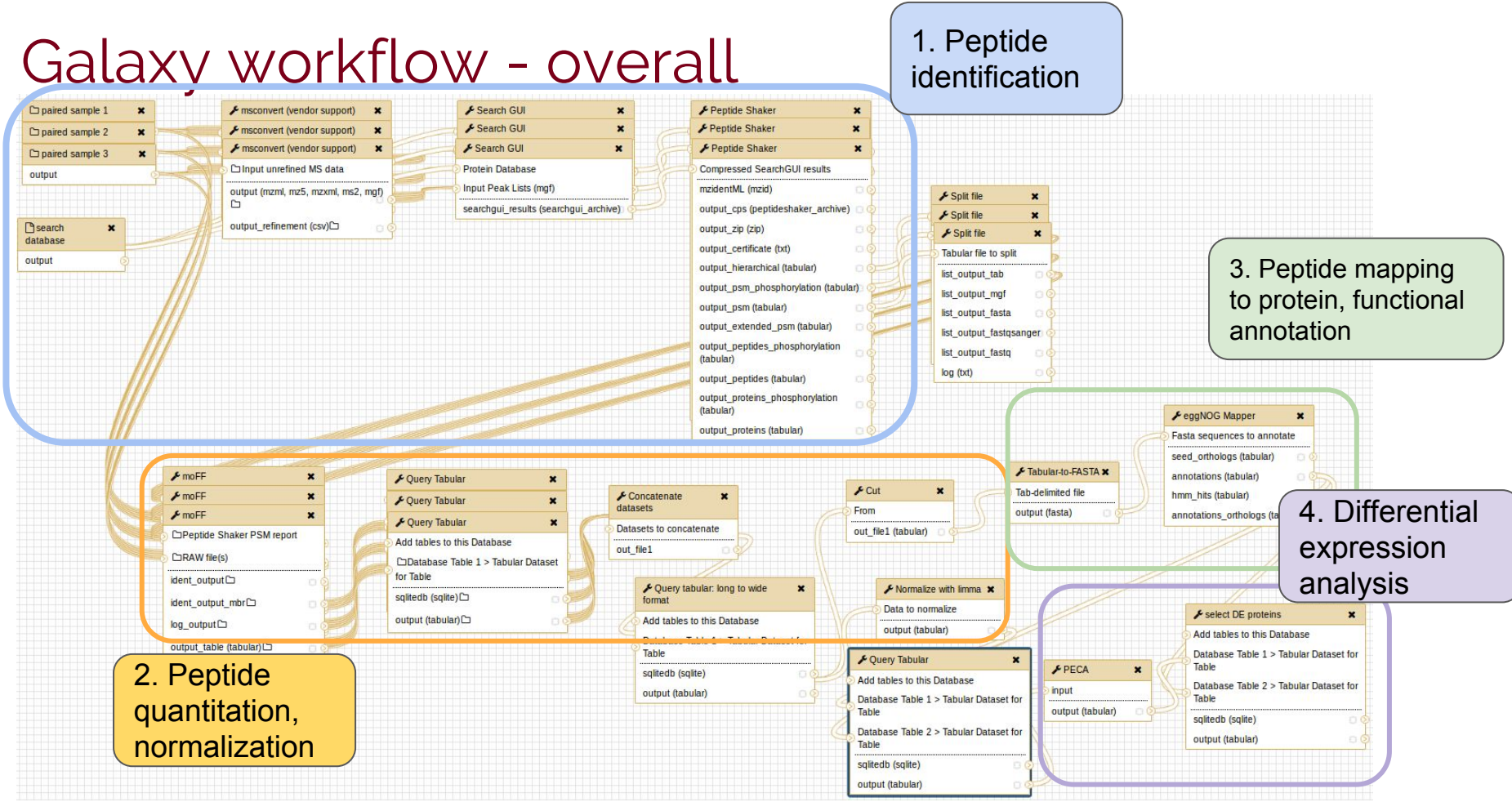
Oral microbiome in a sugar-heavy diet

- Expectorated whole saliva and plaque incubated in paired biofilm reactors (Rudney, et al. *Microbiome*, 2015)
    - "With sucrose" (WS) reactor was sucrose-pulsed 5x per day
- 12 pairs collected - we analyzed 3 pairs for illustration purposes
- Publicly available on PRIDE (PXD003151)



WS = with sucrose
NS = no sucrose

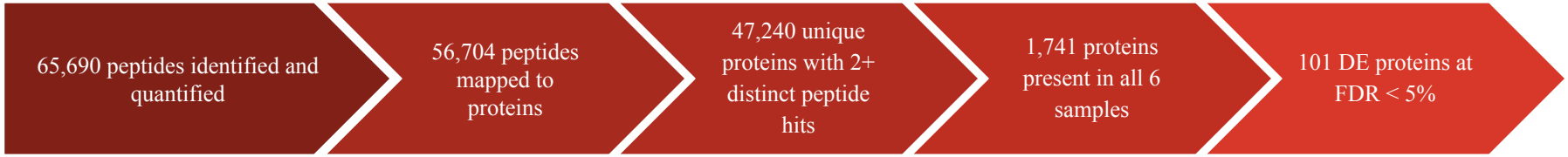
Galaxy workflow - overall



Tools incorporated into Galaxy

- moFF (Argentini, et al., Nat. Methods 2016)
  - Obtains precursor intensities from Thermo raw files (or mzMLs)
- eggNOG mapper (Huerta-Cepas, et al., Mol Biol Evol, 2017)
  - Analogous to BLAST, but searches against eggNOG database, which has detailed functional information
- limma (Ritchie, et al., Nucleic Acids Res. 2015)
  - Many functions - used here for normalization
- PECA (Suomi, et al., J. Proteome Res. 2015)
  - Aggregates peptides to proteins and calculates differential expression statistics
- Quality control filtering (Galaxy-P team, manuscript in preparation)
  - remove proteins with only 1 peptide hits
  - keep only proteins expressed in every sample

Results



- The eggNOG mapper results offer (when available):
  - Taxonomy ID of protein
  - Gene name
  - KEGG KO
  - GO terms
  - BiGG reactions
  - Free text functional annotation

Results

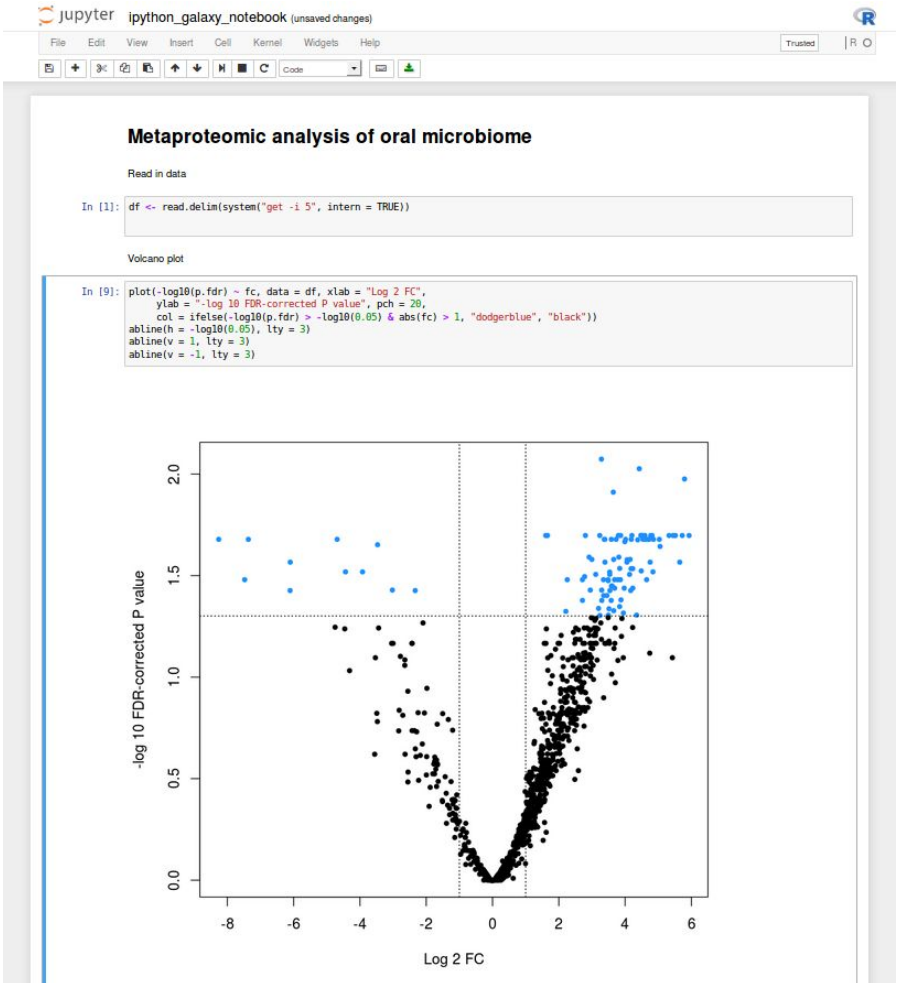
Several glycolytic enzymes are upregulated (FDR≤5%) in a sucrose-rich environment:

Enzyme	**Fold change (WS to NS)**	**Species**
Pyruvate kinase	32.9 20.0 18.2 15.0	*S. oralis* *S. cristatus* *S. mutans* *S. sp. M334*
Glyceraldehyde-3-phosphate dehydrogenase	23.8 6.3	*S. sp. M143* *S. uberis*
Enolase	20.6 16.5 13.2	*S. oralis* *S. cristatus* *S. sp. M334*
Phosphofructokinase-1	8.7	*S. infantis*
L-lactate dehydrogenase	22.6 13.5	*S. salivarius* *S. sp. M334*

Others upregulated as well - DnaK = chaperone; GroL & GroS = chaperonins, etc.

Results

- Data can be loaded into Jupyter notebooks (Gruening, et al., PLoS Comp. Bio. 2017)
- Allows using programming languages such as R and Python within Galaxy platform



Volcano plot: $-\log_{10}(p\text{-value})$ plotted against $\log_2(\text{fold change})$

Conclusions

- Differential expression analysis of microbial proteins can help identify changes in function across experimental conditions
- The full analysis can be carried out within Galaxy
- Workflows can be reused and shared in publications, can be accessed from any computer
    - Provides more transparent and reproducible data analysis

Future Directions

1. Optimize, test workflow and make it available on z.umn.edu/metaproteomicsgateway, a publicly available Galaxy server provided by Galaxy-P
2. Scale workflow to many samples
3. Develop interactive visualizations and data interpretation tools
4. Explore methods for direct differential expression analysis of function

Acknowledgements

Biochemistry, Molecular Biology & Biophysics

Tim Griffin, PhD
Pratik Jagtap, PhD
Praveen Kumar
Subina Mehta
Caleb Easterly
Ray Sajulga
Andrew Rajczewski
Shane Hubler, PhD
Mark Esler
Art Eschenlauer, PhD
Candace Guerrero, PhD
Matt Chambers

CMSP

UMN Center for Mass Spectrometry and Proteomics

Original experimenter
Joel Rudney, PhD

Minnesota Supercomputing Institute

James Johnson
Tom McGowan
Getiria Onsongo, PhD
Michael Milligan, PhD

NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

NIH award U24CA199347

NSF

NSF award 1458524