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

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

\textsuperscript{1}University of Minnesota, Minneapolis
\textsuperscript{2}Max Planck Institute for Developmental Biology, Tübingen, Germany
\textsuperscript{3}University of Freiburg, Freiburg, Baden-Württemberg
\textsuperscript{4}Ghent University, Ghent, Belgium
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

Gene (*potential*) → Transcript (*intention*) → Protein (*action*)
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


*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

Quantified peptides (MS1 Intensities) → Aggregate by protein → Identify up-/down-regulated functions → Analyze and interpret functions of differentially expressed proteins

Aggregate by functional assignments → Identify and interpret up-/down-regulated functions

Identify up-/down-regulated proteins
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**: open source, freely available web platform for accessible bioinformatic analysis

**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

- Collect MS/MS spectra
- Identify peptides
- Quantify peptides
- Normalize peptide intensities
- Align peptides to proteins
- Associate peptide intensities with protein assignments
- Identify differentially expressed proteins

- Visualize differential expression statistics (volcano plot)
- Cluster protein intensities and samples (heatmap)
- View separation of samples by condition (PCA plot)

黄色 = Galaxy tool wrapper developed for this project
Case study: sucrose and the oral microbiome
Oral microbiome in a sugar-heavy diet

  - “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
1. Peptide identification

2. Peptide quantitation, normalization

3. Peptide mapping to protein, functional annotation

4. Differential expression analysis
Tools incorporated into Galaxy

- **moFF (Argentini, et al., Nat. Methods 2016)**
  - Obtains precursor intensities from Thermo raw files (or mzMLs)

  - 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

- 65,690 peptides identified and quantified
- 56,704 peptides mapped to proteins
- 47,240 unique proteins with 2+ distinct peptide hits
- 1,741 proteins present in all 6 samples
- 101 DE proteins at FDR < 5%

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:

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

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

Original experimenter
Joel Rudney, PhD

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

UMN Center for Mass Spectrometry and Proteomics