Mass spectrometry-based multi-omics for studying health and disease

Timothy J. Griffin
University of Minnesota

tgriffin@umn.edu
Acknowledgements

**Biochemistry, Molecular Biology & Biophysics**

Dr. Pratik Jagtap (Co-leader, Galaxy-P)
Praveen Kumar
Subina Mehta
Caleb Easterly
Ray Sajulga
Andrew Rajczewski
Dr. Shane Hubler
Mark Esler
Dr. Art Eschenlauer
Dr. Candace Guerrero
Matt Chambers
Adrienne Bunn

**Collaborators**

Adrian Hegeman
Mo Heydarian/Karen Reddy
Brian Crooker/Wanda Weber
Bart Mesuere
Brook Nunn
Thilo Muth
Magnus Øverlie Arntzen

James Johnson
Tom McGowan
Dr. Getiria Onsongo
Dr. Michael Milligan

**NIH award U24CA199347**

**COMMUNITY-BASED SOFTWARE DEVELOPMENT**

Harald Barsnes and Marc Vaudel
University of Bergen, Bergen, Norway
Bjoern Gruening (Galaxy community...)
University of Freiburg, Freiburg, Germany
Lennart Martens
VIB Department of Medical Protein Research, UGent, Belgium
Lloyd Smith/Michael Shortreed
University of Wisconsin-Madison

ITCR groups
Rachel Karchin/Michael Ryan
Johns Hopkins University/In-Silico Solutions

Tom Doake/Jeremy Fischer
Indiana University

**Center for Mass Spectrometry and Proteomics**
Outline

• **Multi-omics: An introduction**
  • A focus on mass spectrometry-based multi-omics

• **Proteogenomics**
  • Background and informatics challenges
  • Overview of software and ongoing development

• **Metaproteomics**
  • Background and informatics challenges
  • Overview of software and ongoing development

• **Accessibility**
Multi-omics: one view


Image Source:
http://fluorous.com/images/omics.JPG
Vision: A multi-omics workflow for precision health care

- Integrative multi-omics
  - Pathology associations
  - Functional/pathway prioritization
  - A.I. prediction

Informed treatment and health management

Advances in Genetics, 93:147-190

Medical records
Overview of our multi-omics work

Proteogenomics → Multi-omic informatics hub for cancer researchers → Metabolomics

Metabolomics
(Adrian Hegeman)

→ Metaproteomics

MS-based data

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo
Proteogenomics: A primer

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)
Matching amino acid sequences to MS/MS data

Raw MS/MS spectrum

Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification
Inferring protein identity from peptides

Cytochrome C

\[ \text{NH}_2 \text{GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGL} \]

\[ \text{FGRKGTGQAPGFTYTDANKNKG} \text{ITWKEETLMEYLENPK} \]

\[ \text{KYIPGTKMIFAGIKKKTEREDLIAYLKKATNE}_{\text{COOH}} \]
Proteogenomics: building a more specific and comprehensive database

RNA sequences (e.g. RNA-seq) (3-frame translation)

DNA sequences (6-frame translation)

In-silico translation

Comprehensive Database (Sample-specific, all possible sequences)
Proteogenomic outcomes

- Confirms translation of variants
- Direct evidence of potential functional variants
- Applications in neoantigen discovery (immuno-oncology)
Bringing proteogenomics to the masses: informatics challenges

- Many software tools, integration, automation....

- RNA-Seq assembly and analysis
- Customized protein dB generation
- Matching sequences to MS/MS data
- Filtering and QC!
- Interpretation! Beyond a list....

J. Proteome Res., 2014, 13, pp 5898–5908
One possible solution....

http://notapipe.biz/quality-quantity-and-infinite-monkeys/
Galaxy-P: A new community-based informatics paradigm for MS-based proteomics
Solution...Galaxy?


- A web-based, community developed bioinformatics workbench for integrating disparate software -- flexible
- Geared towards use by bench scientists; many training resources available
- Already home to genomic/transcriptomic tools
- Provenance tracking, sharing and reproducibility
- Amenable to other ‘omic tools (e.g. Galaxy for proteomics project, Galaxy-P)

*Working philosophy:*
Galaxy: an integrative workbench well-suited for multi-omics

Integrate datasets, analysis tools, visualizations, and computing resources for large-scale biomedical data science

Data Sets

Web UI
Programmatic API

Interfaces

Galaxy

Analysis Tools and Visualizations

Filter by columns
Differential expression
Drug response predictor

Computing Resources

Courtesy Jeremy Goecks, OHSU
Integrative data processing: RNA-Seq + proteomics
What’s next? Beyond a big list....
I. Multi-Omics Visualization Platform: Characterizing the nature of detected variants

- HTML-based Galaxy plugin
- Interactive reading of mzsqlite dB
Multi-Omics Visualization Platform: Characterizing the nature of detected variants
II. Assessing potential impact of variants: CRAVAT-P

Cancer-Related Analysis of Variants Toolkit (cravat.us) (Rachel Karchin/Michael Ryan)
Viewing results in CRAVAT-P

- HTML-based Galaxy plugin
- Interactive viewer

https://jraysajulga.github.io/cravatp-galaxy-docker/
Unleashing the power of CRAVAT on proteogenomic results

III. Quantitative proteo-transcriptomics
Quantitative proteotranscriptomics

Praveen Kumar
(Krishanpal Anamika/Priyabrata Panigrahi)

QuanTP: interactive visualization of RNA-protein response

QuanTP: Association between abundance ratios of transcript and protein

Input data summary

- Abbreviations used: PE (Proteome data) and TE (Transcriptome data)
- Input Proteome data dimension (Row Columns): 2417 x 5
- Input Transcriptome data dimension (Row Columns): 2417 x 5

Table of Contents:

- Sample distribution
- Correlation
- Regression analysis
- Influential observations
- Cluster analysis

SAMPLE DISTRIBUTION

- Boxplot: Transcriptome data
- Boxplot: Proteome data
Multiple studies have correlated microbial composition with physiological conditions.
Potential to unravel the mechanistic details of microbial interactions with host/environment by analyzing the functional dynamics of the microbiome.

Image from https://thedoctorweighsin.com/what-everyone-should-know-about-the-infant-microbiome/
Functional state as a better measure than composition?

Unpublished Dataset

**TAXONOMY**
Cluster Separation: 0.19056

**FUNCTION**
Cluster Separation: 0.93352
Metaproteomic Challenges

**SINGLE-ORGANISM PROTEOMICS**
- **Size**: SMALL TO MEDIUM SIZE (10 K TO 100K SEQUENCES)
- **Complexity**: SINGLE

**METAPROTEOMICS**
- **Size**: LARGE (1 MILLION AND ABOVE)
- **Complexity**: MULTI-ORGANISM DATABASE WITH HOMOLOGOUS PROTEINS

Disparate tools and multiple processing steps.

- **Search Algorithms** being developed to address large and complex database searches
- **Protein Grouping** at multi-organism level
- **Identification Statistics** affected by large databases
- **Taxonomy** based on unique peptide identifications
- **Functional analysis** based on proteins identified
Metaproteomics Workflow

**DATABASE GENERATION**
- FASTQ

**DATABASE SEARCH & STRATEGIES**
- Protein / Peptide FASTA
- Search Algorithm
- Spectra

**QUANTITATIVE ANALYSIS**
- Peptides
- Spectral counts OR Intensity data

**FUNCTIONAL ANALYSIS**
- Known Function
- Hypothetical Function
- Unknown Function
- Shared Taxonomy
- Unassigned Taxonomy
- Unique Peptides
- Taxonomy Analysis
Metaproteomics Quantitation

### Peptide Quantification

- Peptide quantitation and normalization
- Peptides with normalized intensities

### Peptide Identification

- Peptide quantitation
- Normalization
- Peptide Shaker

### Function and Taxonomy Annotation

- Unipept 4.0
  - peptinfo
  - peptide

### Data Sources

- metaQuantome databases
- Gene Ontology (GO) term database
- ENZYME database
- NCBI taxonomy database

### Statistical Analysis

- metaQuantome filter

### Visualization

- metaQuantome viz
  - Bar chart
  - Heatmap
  - PCA plot
  - Volcano plot

### Experimental Design

- metaQuantome expand
metaQuantome

- Enables quantitative analysis of the taxonomic and functional state of a microbiome.
- Unravels the complex and hierarchical data structure of taxonomic and functional ontologies.
- Enables data exploration, tests hypotheses, and generates high-quality visualizations.
- Deciphers the contribution of taxa to a functional process and vice versa.

Case Study: Sucrose-induced oral dysbiosis

- Mass spectral data was acquired from plaque samples from **twelve subjects at high risk for dental caries grown in biofilm reactor** in the presence (With Sucrose, or **WS**) and absence of sucrose (No Sucrose, or **NS**) (12 in each group, 24 total samples).

- Mass spectra were searched against the Human Oral Microbiome database (HOMD) to identify microbial peptides.

- Quantitation, functional annotation, and taxonomic assignment was performed in Galaxy; **metaQuantome** was used to analyze the results.

*Rudney et al., BMC Microbiome DOI: 10.1186/s40168-015-0136-z*
metaQuantome enables data exploration, tests hypotheses, and generates visualizations

**Oral dysbiosis results: taxonomy**

**Most abundant Genera in NS**

- Fusobacterium: $1 \times 10^{10}$
- Veillonella: $8 \times 10^{9}$
- Streptococcus: $6 \times 10^{9}$
- Haemophilus: $4 \times 10^{9}$
- Granulicatella: $2 \times 10^{9}$

**Most abundant Genera in WS**

- Streptococcus: $6 \times 10^{9}$
- Veillonella: $5 \times 10^{9}$
- Fusobacterium: $4 \times 10^{9}$
- Granulicatella: $2 \times 10^{9}$
- Haemophilus: $1 \times 10^{9}$
metaQuantome enables data exploration, tests hypotheses, and generates visualizations.

**Oral dysbiosis results:**
**Volcano plots**
metaQuantome deciphers the contribution of taxa to a functional process

Oral dysbiosis results:
Taxonomy units contribution to carbohydrate metabolism
Oral dysbiosis results:
Functional distribution of Streptococcaceae peptides

metaQuantome deciphers the functional processes expressed by a taxon

Proportion of peptide intensity

Functional Terms

nitrogen compound metabolic process
metabolic process
transport
carbohydrate metabolic process
small molecule metabolic process
biosynthetic process
response to stress
RNA metabolic process
protein folding

Functional Terms

nitrogen compound metabolic process
metabolic process
carbohydrate metabolic process
translation
protein folding
FNA metabolic process
biosynthetic process
regulation of metabolic process
response to stress
Workflow engine enables testing of new pipelines

- Sectioning of large database into smaller sub-databases to increase sensitivity
New algorithms enable improved and deeper data

i
Assigned Enzyme Commission Number (EC)

Traditional

346

205

Sectioning

10

ii
Top 15 Protein Enzyme Groups (EC)

Number of PSMs

EC Terms
Bringing it all together

• Develop a tool based on methods for correlation/comparison of outputs from workflows for host-proteogenomics; host-transcriptomics; metatranscriptomics & metaproteomics.
Accessing the Multi-omic Workflows

PUBLIC INSTANCES

Proteogenomics Gateway: z.umn.edu/proteogenomicsgateway
Step-by-step instructions: z.umn.edu/pginnov18

Metaproteomics Gateway: z.umn.edu/metaproteomicsgateway
Step-by-step instructions: z.umn.edu/suppS1

Tools and Workflows also available on: https://proteomics.usegalaxy.eu/
Accessing the Multi-omic Workflows

**ALSO AVAILABLE ON:**

GitHub: [https://github.com/galaxyproteomics](https://github.com/galaxyproteomics)

Galaxy Toolshed: [https://toolshed.g2.bx.psu.edu/](https://toolshed.g2.bx.psu.edu/)

Docker: [https://jraysajulga.github.io/cravatp-galaxy-docker/](https://jraysajulga.github.io/cravatp-galaxy-docker/)

Training Workflows also available on: [https://training.galaxyproject.org](https://training.galaxyproject.org)
We can be Reached at:

Published Manuscripts: z.umn.edu/galaxypreferences

Galaxy-P Presentations: http://galaxyp.org/conference-presentations

Contact: http://galaxyp.org/contact/

Twitter: twitter.com/usegalaxyp