A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers
Pratik Jagtap and Timothy J. Griffin
University of Minnesota

tgriffin@umn.edu
pjagtap@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

GalaxyP

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

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

NIH award U24CA199347

GalaxyP

NSF

NIH
NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

Jetstream
Outline

- **ITCR project:** A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers
  - 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 and future directions**
Overview of ITCR project

Proteogenomics

Multi-omic informatics hub for cancer researchers

Metabolomics
(Adrian Hegeman)

Metaproteomics

MS-based data

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo
**METABOLOMICS**

- **GOAL**: user-friendly workflow for MS-based metabolite quant and ID within Galaxy framework
- Leveraging and contributing to Galaxy-based community of metabolomic informatics developers (Workflow4metabolomics, W4M)
- User-driven resource: meeting requirements for reproducibility, throughput, statistical analysis
- “Dockerized” appliance for portability is under development
- **Availability of tools**: metabolomics.usegalaxy.eu

**Recent and ongoing developments:**
- **New tool development for W4M:**
  - *Sample Subset* – prepare data for statistical analysis of metabolite features
  - *OPLS-DA Contrasts* – selection of OPLS-DA pairs of sample classes
  - *w4mkmeans* - an alternative to multivariate analysis for discovering relationships among samples
  - *VKMZ* – a tool for predicting metabolite composition from high res metabolite data and visualizing via a van Krevelen diagram

Galaxy-based workflow for high-throughput metabolite quant and annotation (ID); xCMS based
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
Detecting protein variants via proteogenomics

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

In-silico translation

Comprehensive Database
(Sample-specific, all possible sequences)

TCGATCAGGGCAATAGCTAGTCCCCGTTA
DNA sequences
(6-frame translation)
Proteogenomic outcomes

- Confirms translation of variants
- Direct evidence of potential functional variants
- Applications in neoantigen discovery (immunoncology)
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....
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

Datasets

Interfaces
- Web UI
- Programmatic API

Analysis Tools and Visualizations
- Filter by columns
- Differential expression
- Drug response predictor

Computing Resources

Courtesy Jeremy Goecks, OHSU
Integrative data processing: RNA-Seq + proteomics

- RNA-Seq
  - Reference genome
  - GTF
    - HISAT
    - STRINGTIE
    - GFF COMPARE
      - Translate variant transcripts
        - Transcript variant detection
    - Peptide sequence variants
  - SAV/InDel variants detection
    - Cancer Res. 77:e43-e46
  - VCF file
    - CustomPro DB
      - Genomic mapping file
        - Protein sequence database
          - MS/MS proteomics data
            - mz-to-sqlite
            - mzssqlite database
  - Variant peptide detection and verification
    - SEARCHGUI/PeptideShaker
      - BLASTP
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)
Assessing potential impact of protein-level variants: CRAVAT-P

- Intersecting of transcript variants and confirmed protein variants
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 proteotranscriptomics

mRNA versus protein abundance ratios, Gal/Eth

mRNA abundance ratio (log₁₀) vs. protein abundance ratio (log₁₀)
Quantitative proteotranscriptomics

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 Column): 2817 x 5
- Input Transcriptome data dimension (Row Column): 2817 x 5

Table of Contents:
- Sample distribution
- Correlation
- Regression analysis
- Influential observations
- Cluster analysis

**SAMPLE DISTRIBUTION**

<table>
<thead>
<tr>
<th>Boxplot: Transcriptome data</th>
<th>Boxplot: Proteome data</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Transcriptome data Boxplot" /></td>
<td><img src="image2" alt="Proteome data Boxplot" /></td>
</tr>
</tbody>
</table>

---

The diagram shows the distribution of transcriptome and proteome data, with boxplots illustrating the spread and central tendency of the data across different groups.
Outline

• ITCR project: A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers
  • 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 and future directions
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/
Biogas Cellulose Degradation Dataset

Unpublished Dataset

**TAXONOMY**
Cluster Separation: 0.19056

**FUNCTION**
Cluster Separation: 0.93352

PCA1 (78.8%)

PCA1 (95.3%)
• Microorganisms have been implicated for their role in cancer progression.

• Challenges exist in understanding the mechanism of cancer host-microbiome interaction.

• We seek to use complementary microbiome -omics techniques to provide functional information by quantifying microbial RNA & protein expression levels.
Metaproteomic Challenges

**SINGLE-ORGANISM PROTEOMICS**
- **Size**: Small to medium size (10 K to 100K sequences)
- **Complexity**: Single + Contaminants

**METAPROTEOMICS**
- **Size**: Large (1 million and above)
- **Complexity**: Multi-organism database with homologous proteins

**Search Database**
- **Size**
  - SMALL TO MEDIUM SIZE (10 K TO 100K SEQUENCES)
  - LARGE (1 MILLION AND ABOVE)
- **Complexity**
  - SINGLE + CONTAMINANTS
  - 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
- Protein / Peptide FASTA

**DATABASE SEARCH & STRATEGIES**
- Search Algorithm
- Spectra
- Peptides

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

**FUNCTIONAL ANALYSIS**
- Known Function
- Proteins
- Hypothetical Function
- Unknown Function
- Shared Taxonomy
- Unassigned Taxonomy
- Unique Peptides

**TAXONOMY ANALYSIS**
Unipept 4.0

- Unipept is an application for metaproteome data analysis.
- Unipept generates outputs for taxonomy and functional analysis.
- Unipept functional analysis is based on GO terms and EC numbers.

Singh et al. J. Proteome Res. 18: 606–615
DOI: 10.1021/acs.jproteome.8b00716

https://unipept.ugent.be/
Metaproteomics Quantitation

**Peptide Identification**
- Spectrum files
- msconvert
- SearchGUI
- Peptide Shaker

**Peptide Quantification**
- FlashLFQ
- Limma
- Normalization
- Peptides with normalized intensities
- Peptides with functional assignments
- Peptides with taxonomic assignments

**Function and Taxonomy Annotation**
- Unipept 4.0
- pepinfo
- pep2lca

**Statistical Analysis Visualization**
- metaQuantome
  - Experimental Design
  - Gene Ontology (GO) term database
  - Enzyme database
  - NCBI taxonomy database
- metaQuantome stats
- metaQuantome filter
- metaQuantome expand
- metaQuantome create samples file
- metaQuantome viz
- Bar chart
- Heatmap
- PCA plot
- Volcano plot
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: 1e+10
- Veillonella: 8e+09
- Streptococcus: 6e+09
- Haemophilus: 4e+09
- Granulicatella: 2e+09

**Most abundant Genera in WS**
- Streptococcus: 6e+09
- Veillonella: 5e+09
- Fusobacterium: 4e+09
- Haemophilus: 3e+09
- Granulicatella: 1e+09

**Abundant taxa**
metaQuantome enables data exploration, tests hypotheses, and generates visualizations.

Oral dysbiosis results: volcano plots
metaQuantome enables data exploration, tests hypotheses, and generates visualizations.

**Oral dysbiosis results:**
**Principal component analysis**
metaQuantome enables data exploration, tests hypotheses, and generates visualizations.
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.
Metaproteomics Workflow

MICROBIOME

imetaQuantome workflow
@ASMS 2019

METAPROTEOMICS
Metatranscriptomics Workflow

MICROBIOME

ASaiM
Batut et al Gigascience. 2018 7(6)
doi: 10.1093/gigascience/giy057

imetaQuantome workflow
@ASMS 2019
For metatranscriptomics data, the AsaiM workflow assembles, extracts, explores, and enables functional and taxonomy visualization, and taxonomy-functional interaction analysis data.
For metatranscriptomics data, the AsaiM workflow assembles, extracts, explores, and enables functional and taxonomy visualization, and taxonomy-functional interaction analysis data.
A 100 µl aliquot of an enriched community from a biogas reactor was transferred to 27 anaerobic bottles containing a rich medium and 10g/L of cellulose as sole carbon source and incubated at 65 °C.

Samples were collected at 9 different time points (0, 8, 13, 18, 23, 28, 33, 38 and 43 h) and processed in triplicates. **Metatranscriptomics analysis was performed on 7 time points.**
ASaiM workflow enables functional and taxonomy visualization, and taxonomy-functional interaction analysis data.
Ongoing Work

• Develop a statistical analysis and visualization tool/workflow for metatranscriptomics analysis.
• Develop a tool based on methods for correlation/comparison of outputs from workflows for metatranscriptomics and metaproteomics analyses.
Future Plans

- Develop a tool based on methods for correlation/comparison of outputs from workflows for host-proteogenomics; host-transcriptomics; metatranscriptomics & metaproteomics.
Outline

• **ITCR project: A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers**
  • 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 and future directions**
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/supportS1

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