SESSION 1
Basics of Data Analysis using Galaxy

(SW4) The Galaxy Platform for Multi-Omic Data Analysis and Informatics

ABRF 2016
February 20, 2016

Dave Clements
Galaxy Team
Johns Hopkins University
http://galaxyproject.org/

#usegalaxyp
#usegalaxy

@usegalaxyp
@galaxyproject
Acknowledgements

ABRF
AWS
Thermo Scientific
Globus
Genomics
Pratik Jagtap
Tim Griffin
Thomas McGowan
Getiria Onsongo
Candace Guerrero
Mohammad Heydarian
NIH
Johns Hopkins University
Penn State University
University of Minnesota
The Galaxy Team

http://wiki.galaxyproject.org/GalaxyTeam
The tutorial materials are available online:

Materials
Short (handout): bit.ly/gxyabrf16intro
Long: bit.ly/gxyabrf16intro-long

And, completely unrelated to today, but relevant to this audience, see this talk from WACD 2015:

Utilizing the Galaxy Analysis Framework at Core Facilities
Goals

Provide an introduction to using Galaxy for bioinformatic analysis.

Give you enough background for subsequent sessions today.

This session does cover RNA-Seq but you won't be an RNA-Seq expert at the end of the session.
2016 Galaxy Community Conference (GCC2016)

June 25-29, 2016
Bloomington, Indiana

galaxyproject.org/GCC2016
Galaxy Training Network

20+ training organizations around the world

bit.ly/gxygtn
Thanks
How to better understand bioinformatics & Galaxy

Experiment. (You are already used to the idea and) Galaxy makes it easy

Read tool documentation and tool and method review papers

Get Help!

http://biostars.org/
http://seqanswers.com/
https://biostar.usegalaxy.org/
http://galaxyproject.org/search
SESSION 2

Mass Spectrometry-Based Proteomics Data Analysis Using Galaxy

Candace Guerrero
University of Minnesota
February 20, 2016
Prediction of Gene Activity in Early B cell Development Based on an Integrative Multi-omics Analysis

![Diagram of early B cell development](image)

<table>
<thead>
<tr>
<th>Experiment type</th>
<th>pre-pro-B origin</th>
<th>pro-B origin</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3ac ChIP-seq</td>
<td>Ebf1 deficient</td>
<td>Rag1 deficient</td>
<td>Lin, Y.C. et al, Nature Immunology 2010 Geo accession: GSE21978</td>
</tr>
<tr>
<td>H3K4me3 ChIP-seq</td>
<td>Ebf1 deficient</td>
<td>Rag1 deficient</td>
<td>Lin, Y.C. et al, Nature Immunology 2010 Geo accession: GSE21978</td>
</tr>
<tr>
<td>GRO-seq</td>
<td>E2a deficient</td>
<td>Rag1 deficient</td>
<td>Lin, Y.C. et al, Nature Immunology 2012 Geo accession: GSE40173</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>Ebf1 deficient</td>
<td>Rag2 deficient</td>
<td>this study</td>
</tr>
<tr>
<td>iTRAQ</td>
<td>Ebf1 deficient</td>
<td>Rag2 deficient</td>
<td>this study</td>
</tr>
</tbody>
</table>

Figure 1: Experimental system and multi-omics data. (A) Schematic of early B cell development through three stages: MPP, pre-pro-B, and pro-B cells. Relevant receptors and protein expression are indicated. (B) Multi-omics' data used in this study and their respective sources.
Mass Spectrometry (MS)

• A technique that measures the mass-to charge ratio (m/z)* of ions

• Analyzes gas-phase ions

• Consists of three parts:
  – Ion source
  – Mass analyzer
  – Detector

* we can easily infer the charge “z” and then we can calculate the molecular weight
Ion source: Electrospray ionization (ESI)

High Voltage

Evaporating droplets containing peptide ions

Nobel-prize winning achievement!!
Technologies: Proteomics

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)

Proteins → Digestion → Peptides → "Multidimensional" Fractionation → μLC → MS1 → ESI → Isolation → Fragmentation → Mass Analysis → MS2 → Peptide Fragments

Complex mixture → Separation (turnstile) → 1, 2, 3, ... → Detect each component

Proteomics + TOF MS: 24 MCA scans from MyoTryptic.wiff
Max. 5191.0 counts.
Precursor Ion Fragmentation

Most common peptide fragments are b, y, and then a

We will on use b and y ions
Example of Peptide Fragments

Fragment with Collision Gas

**b ions**

- \( \text{NH}_2 \rightarrow \text{AE} \rightarrow \text{CO} \)
- \( \text{NH}_2 \rightarrow \text{AEP} \rightarrow \text{CO} \)
- \( \text{NH}_2 \rightarrow \text{AEPT} \rightarrow \text{CO} \)
- \( \text{NH}_2 \rightarrow \text{AEPTI} \rightarrow \text{CO} \)

**y ions**

- \( \text{NH}_2 \rightarrow \text{IR} \rightarrow \text{COOH} \)
- \( \text{NH}_2 \rightarrow \text{TIR} \rightarrow \text{COOH} \)
- \( \text{NH}_2 \rightarrow \text{PTIR} \rightarrow \text{COOH} \)
- \( \text{NH}_2 \rightarrow \text{EPTIR} \rightarrow \text{COOH} \)
Fragment ion MS

NH3+ A E P T I R COOH

Intensity

m/z

b1 72.0
y1 174.1
b2 201.1
y2 287.2
b3 298.1
y3 388.2
b4 399.2
y4 485.2
b5 512.2
y5 614.2

A E P T I R
Identifying Peptide Spectrum Matches

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
MS Approaches to Quantitation

• Mass spectrometry is not inherently quantitative, as different molecules have different MS responses

• Most accurate way is to label the sample with stable isotopes ($^{13}$C, $^{15}$N, $^{18}$O…)

• We can let the organism make isotope-labeled proteins (SILAC)

• Or we can label the proteins / peptides after collecting them (iTRAQ)

• (Label-free methods also exist)
iTRAQ

- Isobaric tag for relative and absolute quantitation
- Available in up to 8 different mass tags; apply each tag to a different sample
- Peptides are chemically labeled
- After labeling, combine all your samples and analyze in the same run
For testing purposes, these were mixed 1:1:1:1

In the MS² step, we learn which peptide this is (i.e. which protein is present) and how much is present in each sample. In this case, MS¹ is less informative but helps us identify the peptide.
Inferring protein identity from peptide sequence matches

Cytochrome C

\[ \text{NH}_2\text{GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMYLENPPKYIPGTMIFAGIKKKTEREDLIAYLKKATNECOOH} \]
Using Protein Database Searches to Infer Protein Identities

• There are a variety of algorithm searches to infer protein identity
  • Open source: Comet, X!Tandem, MS Amanda, etc.
  • Commercial: ProteinPilot

• Each search algorithm is varies in coding
  – Leads to different results
  – match MS/MS spectra to sequences

• SearchGUI
MS-based Proteomics only as good as the database use….

Sequenced proteins

- Generally includes canonical protein sequences or single proteins isolated via biochemical purification and chemical methods; some “inferred” proteins from nucleic acid sequences.

“What we know is a drop, what we don't know is an ocean.”

-- Sir Isaac Newton
General Proteomics Approach

Protein Database

MS Data

Protein Database Search

Analysis and Visualization Software
Galaxy Work Canvas
The Need for Dataset Collection

Whole Cell Lysate

Requires 3 LC-MS/MS runs for 1 biological sample

Trypsin Digested

Whole Cell Lysates (WCL)
SDS-PAGE 4-20% (Bio-Rad precast gel)
Blue Coomassie staining

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>MW (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>130</td>
</tr>
</tbody>
</table>
The Need for Dataset Collection

1 Collective Output for 1 Biological Sample

3 Independent Outputs for 1 Biological Sample
Creating Peaklists

RAW File
Instrument File Type

• Thermofisher files
• Info raw needs to be extracted out

mzml file

• Peak detection
• Noise removal
• Baseline correction
• Monoisotopic peak correction
• Charge state derivation

MGF file
Desired Database Input

• Input of SA
• Info. separated by headers
• Each spectral scan
  • Peptide mass
  • Charge state
  • Scan number
Galaxy Work Canvas
General Proteomics Approach

Protein Database

MS Data

Protein Database Search

Analysis and Visualization Software
SearchGUI

- Used for analysis MS/MS data to identify peptide spectrum matches
- Search Program that utilizes multiple search algorithms
  - Creates a streamlined user friendly interface to use open source search algorithms
  - Collectively processes results as one output
  - Allows simple parameters to be inserted for multiple searches
Identifying Peptide Spectrum Matches

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
PeptideShaker

- Statistical Validation
  - FDR (false discovery rate)
    - Decoy base
    - % Confidence
- Protein grouping
- Visualization
- Output files
Protein Report
- valid proteins
- coverage
- molecular weight

Peptide Report
- valid peptides
- potential novel proteoforms based on accession numbers
- sequences
- modifications and localization score
- confidence

Spectrum (PSM) Report
- valid spectra
- potential novel proteoforms based on accession numbers
- sequences
- modifications and localization score
- confidence
- m/z, charge state, Δm/z

Summary (Parameters)
- valid peptides
- valid proteins
- valid spectra

Archive (zipped file)
- CPS file to visualize data

Mzid
- PSM Visualization
- SWATH Analysis
- Skyline
- Scaffold
Galaxy Work Canvas

- Database Generation
- Generating Peaklists
- Database Search and Post Processing
- Visualization
SQLite Dataset

• **mz to sqlite tool**
  – Consolidates information in an indexed format
    • mzIdentML
    • peaklist input datasets (e.g. mzml and MGF files)
    • fasta SearchDB

• **SQLite database schema**
  – **PSM Viewer**
    • Galaxy-P visualization plugin
    • Interactively analyze data
POST-LUNCH SESSION

PROTEOGENOMICS & METAPROTEOMICS USING THE GALAXY PLATFORM

PRATIK JAGTAP
University of Minnesota
St. Paul / Minneapolis
Minnesota, USA
Session 1
- RNA-seq and quantitative MS-proteomics data:

Session 2
- Documentation to complement hands-on instruction

Session 3
- Prediction of Gene Activity in Early B Cell Development Based on an Integrative Multi-Omics Analysis

Session 4

http://z.umn.edu/abrf2016sw4
Schematic representation of the flow of biological information from DNA to RNA to protein and the proteomics workflow designed to study biological complexity at the functional level of proteins. The information from our genes is increased by many orders of magnitude following protein translation. To understand biological systems, protein function and dysfunction must be deciphered. Bottom-up proteomics and bioinformatics forms a cycle; proteomics to first recover, identify, and quantify the post-translation modifications, protein–protein interactions, signaling events, etcetera using advances in sample preparation and mass spectrometry while the software and bioinformatics packages reconstruct these bits of information to systems level understanding of complex biological processing.
Emerging Fields

- Next Generation Proteomics.
  - Proteogenomics (Uses data from RNASeq data)
  - Metaproteomics (uses metagenomics data)
  - SWATH (Data-independent acquisition)

Janice Mayne, Zhibin Ning, Xu Zhang, Amanda E. Starr, Rui Chen, Shelley Deeke, Cheng-Kang Chiang, Bo Xu, Ming Wen, Kai Cheng, Deeptee Seebun, Alexandra Star, Jasmine I. Moore, and Daniel Figeys

Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology and Immunology, University of Ottawa, 451 Smyth Rd., Ottawa, Ontario, Canada, K1H8M5

CONTENTS

Advances in Proteomic Related Techniques 96
   General Sample Preparation 97
      One-Pot Strategies 97
      High-Throughput Strategies 97
      Optimizing Digestion 97
   Sample Membrane Considerations for Membrane Proteins 98
   Mass Spectrometry-Based Approaches for Detecting Protein Interactions 98
      Affinity Purification Coupled to Mass Spectrometry 98
      Proximity Biotinylation Approach to AP-MS 99
      Data Independent Analysis of AP-MS Generated Spectra 99
   Enriching for Post-Translational Modifications 99
      Phosphorylation Enrichment 99
      Ubiquitination Enrichment 100
      Acetylation Enrichment 101
      Glycosylation Enrichment 101
      Methylation Enrichment 102
      Terminal Peptide Enrichment 102
   Liquid Chromatography Mass Spectrometry (LC-MS) Technological Advances 102
      Single Shot Long Gradient Versus Fractionation 103
      Two-Column Configuration To Boost Sample Throughput 103
   Protein Quantification in Bottom-up Proteomics 103
      NeuCode SILAC 103
      MS2 Level Quantification 103
      Label Free Quantification (LFQ) 103
      Absolute Quantification 104

Next Leap of Proteomics 107
   Proteogenomics 107
      Call for Proteogenomics 107
   Metaproteomics, Unveiling the Functions of Microbial Community 109
      Sample Preparation Considerations for Metaproteomic Studies 109
      Challenges in Microbial Peptide Identification and Protein Inference 110
      Taxonomic Assignment of Peptides and Proteins 110
      Functional Interpretation 110
      Quantitative Metaproteomics 111
      Proteomics for Understanding Host–Microbe Interactions 111
   Proteomics and Clinical Applications 113
      Biomarkers 113
      Drug Design 114
      Decision Making in Health Care 114
      The Future and Personalized Medicine 114

Perspectives 115

Author Information 115
   Corresponding Author 115
   Author Contributions 115
   Notes 115
   Biographies 115
   Acknowledgments 116
   References 116

Over 17,000 papers related to proteomics have been published since our 2011–2013 review. A subset are true gems reporting amazing applications of proteomics to better understand biology and medicine.
DEFINING PROTEOGENOMICS: LOOKING WITHIN AND WITHOUT

- Mass spectrum
- Reference Protein Database from genomic annotation
- RNASeq data
- Genome six-frame translation
- cDNA three-frame translation
PROTEOGENOMICS

- Genome annotation
- Gene expression regulation
- Protein variants in disease
- Functional outcomes of genome mutation

PROTEOGENOMICS : BIOINFORMATIC CHALLENGES

• Large database sizes (6-frame and 3-frame translation and metagenomic databases).

• False-positive sources and their elimination.

• Validation of the peptide identification. *(Search using BLAST-P)*

• PSM Evaluation / Targeted proteomics of identified peptides.

• Genomic localization.

• Disparate tools and numerous processing steps.
The dataset was searched against FASTA database with human proteins, contaminant proteins, 3-frame translated cDNA database from EnSEMBL and Human Oral Microbiome database (HOMD).

Table 1. Summary of Genomic Organization of Peptides Corresponding to Novel Proteoforms

<table>
<thead>
<tr>
<th>genomic rearrangements</th>
<th>peptides</th>
<th>chromosome location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternate frame</td>
<td>26</td>
<td>1, 3, 5, 7, 8, 9, 11, 12, 14, 16, and 19</td>
</tr>
<tr>
<td>untranslated region</td>
<td>15</td>
<td>2, 4, 6, 7, 8, 11, 12, 13, 14, and 19</td>
</tr>
<tr>
<td>pseudogenes</td>
<td>6</td>
<td>1, 3, 6, 14, 19, and X</td>
</tr>
<tr>
<td>intronic region</td>
<td>2</td>
<td>12 and 16</td>
</tr>
<tr>
<td>novel exon junctions</td>
<td>2</td>
<td>15 and 17</td>
</tr>
<tr>
<td>antisense</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

Figure 4. Representation of organization of identified peptides corresponding to a novel proteoform from PRB1 and PRB2 genes on chromosome 12. View is a zoomed-in screenshot of chromosome 12, which shows the orientation of expression, amino acid sequences within three frames of translation, reference files in the tracks, and amino acid sequence of the identified peptide corresponding to a novel proteoform. The red arrows indicate the direction and amino acid sequence (from amino-terminal to carboxy-terminal) of the identified peptides. A red asterisk indicates a stop codon in the normal coding frame. Block arrows in red indicate multiple distinct peptides identified during the proteogenomic analysis.

J Proteome Res. (2014) 13(12):5898-908. doi: 10.1021/pr500812t
Galaxy-P provides an integrated platform for every step of proteogenomic analysis.

- Build target database – download and translate EST databases.
- Numerous tools for identification and text manipulation.
- Workflow utilizing BLAST to identify novel peptides.
- Tool to assess peptide-spectrum matches and visualize spectra.
- Visualize identified peptides on the genome.
- 140 steps: Seamless, integrated proteogenomic workflow.

Flexible and accessible workflows for improved proteogenomic analysis using Galaxy framework.
J. Proteome Res. (2014) DOI: 10.1021/pr500812t
Link: z.umn.edu/pgfirstlook
Tracing of core body temperature ($T_b$, black line) from a single animal measured by a surgically implanted transmitter, along with the controlled ambient temperature (blue line) over the course of the hibernation season.

* **TOR** (Torpor), **J-IBA** (January IBA), **M-IBA** (March IBA)
The datasets were run in triplicates and were searched against proteomic dataset from RNASeq data.

Differentially expressed genes from RNASeq data and differentially expressed proteins from iTRAQ data were compared.

Functional analysis of differentially expressed proteins revealed that:

Protein expression in hibernation relative to AUGUST highlights fatty acid metabolism and altered calcium handling and contractile function in the heart.

162 novel peptide sequences were identified in all three replicates.
Proteogenomics is a rapidly evolving field at the intersection of genomics, transcriptomics, and proteomics. Whole genome, exome, and RNA sequencing are well-established techniques that can provide information at the DNA and RNA level with excellent sequencing coverage and depth. Although tens of thousands of clinical samples have been sequenced thus far, data integration and interpretation still remain largely incomplete. Recent advances in proteomic technologies have enabled the accurate and almost complete characterization of the proteomes of many tissues and biological fluids. Integration of multiomics data for the accurate annotation and reciprocal refinement of genomic and proteomic models is essentially the goal of proteogenomics.

This integrative approach has the potential to provide solid evidence for the translation of previously unknown transcripts. Those transcripts and the respective encoded proteins might be implicated in physiological or pathophysiological processes. Novel reported peptides can represent single amino acid variants, splice variants, gene fusions, RNA editing events, novel open reading frames, translated noncoding RNAs, and pseudogenes, among many others. Proteogenomic platforms can now be used lower false-positive and false-negative identification rates. However, like all areas of active research, proteogenomics in its current state is not free of drawbacks. Major limitations in the field are the sensitivity of the mass spectrometers, the increased false discovery rate for the novel peptide hits, and the inherent biophysical properties that render some peptides undetectable.

In this Q&A we discuss with 4 experts in the field the current status of proteogenomics and conditions that have to be met to deliver its promises.

**What are the key technologies that enabled the development of proteogenomics?**

**Alexey Nesvizhskii:** In most cases, especially when studying human or model organisms, proteogenomics is critically dependent on the knowledge (and often aims to refine that knowledge) assembled by large genome
“The large-scale characterization of the entire protein complement of environmental microbiota at a given point in time”


“Through the application of metaproteomics to different microbial consortia over the past decade, we have learnt much about key functional traits in the various environmental settings where they occur.”

DEFINING METAPROTEOMICS: LOOKING WITHIN AND WITHOUT

Mass spectrometry

Reference Protein Database
from genomic annotation

RNASEq data

Genome six-frame translation

cDNA three-frame translation

Metagenomic sequences
Figure 4. Comparison of genera identified and SEED functional groups within control (magenta) and OPML (tan) samples by using MEGAN5 analysis. (A) Radial plot showing the relative distribution of genus-level peptides. The width of the segment assigned to each genus is proportional to the relative abundance of peptides for that genus, and the heights of the magenta and tan spokes within each segment are proportional to the relative abundance of peptides for that genus assigned to the Control and OPML groups. (B) Radial plot showing the relative distribution of peptides into SEED functional roles. The width of the segment assigned to each functional role is proportional to the relative number of identified peptides for that category, and the heights of the magenta and tan spokes within each segment are proportional to the relative abundance of peptides assigned to the Control and OPML groups.
METAPROTEOMICS OF CHILDHOOD CARIES

- In vitro investigation of sucrose-induced changes in the metaproteomes of children with caries.
- Major shifts in taxonomy and function in paired microcosm oral biofilms grown without and with sucrose respectively.
  - Twelve replicates have been analyzed.

- SEED analysis of Oral microcosm biofilms showed characteristic NS and WS patterns of protein expression that were highly conserved across taxonomically diverse communities.

- Targeted proteomic approaches then can be used to determine whether those proteins are also expressed when plaque is exposed to sucrose in the mouth.
The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation

Thilo Muth, Alexander Behne, Robert Heyer, Fabian Kohrs, Dirk Benndorf, Marcus Hoffmann, Miro Lehteva, Udo Reichl, Lennart Martens, and Erdmann Rapp

Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany
Chair of Bioprocess Engineering, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany
Department of Medical Protein Research, VIB, B-9000 Gent, Belgium
Department of Biochemistry, Ghent University, B-9000 Gent, Belgium

Supporting Information

ABSTRACT: The enormous challenges of mass spectrometry-based metaproteomics are primarily related to the analysis and interpretation of the acquired data. This includes reliable identification of mass spectra and the meaningful integration of taxonomic and functional meta-information from samples containing hundreds of unknown species. To ease these difficulties, we developed a dedicated software suite, the MetaProteomeAnalyzer, an intuitive open-source tool for metaproteomics data analysis and interpretation, which includes multiple search engines and the feature to decrease data redundancy by grouping protein hits to so-called meta-proteins. We also designed a graph database back-end for the MetaProteomeAnalyzer to allow seamless analysis of results. The functionality of the MetaProteomeAnalyzer is demonstrated using a sample of a microbial community taken from a biogas plant.

KEYWORDS: bioinformatics, environmental proteomics, mass spectrometry, metaproteomics, microbial communities, software

INTRODUCTION

Mass spectrometry (MS)-based analysis of pure culture proteomes or simple mixed cultures has advanced rapidly in the past decade. There is now a growing interest in studying complex multispecies samples such as entire microbial communities. Microbial consortia are key players in geochemical cycles, biochemical networks, and biotechnological processes. Therefore, the challenges of metaproteomics data analysis and interpretation therefore must be addressed in a sophisticated manner as well. Another challenging issue is encountered in the functional annotation of proteins, as metaproteomics research is not only interested in (single-)protein identifications, but also focuses strongly on specific functions performed by microorganisms in an ecosystem. Unfortunately, no stand-alone software tool currently exists to aid metaproteomics research in...
COMMUNITY-BASED SOFTWARE DEVELOPMENT

Software Developers

SearchGUI / PeptideShaker

Galaxy Wrapper

Software tool deposited in Galaxy Toolshed

 Improvements to the software tool

Software tool installed in GalaxyP

USER FORUM / GITHUB

Users test the tools and provide feedback to developers.
Links to workflows, webcast, pages, documentation...

Workflows
Proteogenomic studies: http://z.umn.edu/pg140
Metaproteomic studies: http://z.umn.edu/metaproteomics1

Webcast
Using ProteinPilot within Galaxy-P: z.umn.edu/ppingp

Pages
Proteogenomics page: z.umn.edu/proteinpilotpage
Metaproteomics page: z.umn.edu/metaproteomicspage

Workshop / Tutorial on proteogenomics:
Mass Spectrometry-based Proteomics Data Analysis using Galaxy-P: z.umn.edu/gcc2015gp
Multi-omic data analysis using Galaxy

To the Editor: Controlling multi-omic data analysis has become a reality, largely due to the availability of high-throughput sequencing technologies and advancements in bioinformatics. The use of multiple data types in proteomics research offers unique opportunities for integrating genomic and proteomic data, enabling a holistic understanding of biological systems.

Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework

Pratik D. Jagtap, James E. Johnson, Getria Onsongo, Fredrik W. Sadler, Kevin Murray, Yuanbo Wang, Gloria M. Shemykman, Sricharan Bandhakavi, Lloyd M. Smith, and Timothy J. Griffin

1Center for Mass Spectrometry and Proteomics, University of Minnesota, 43 Gortner Laboratory, 1479 Gortner Avenue, St. Paul, Minnesota 55108, United States
2Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, 6155 Jackson Hall, 321 Church Street

Proteomic Profiles in Acute Respiratory Distress Syndrome Survivors from Non-Survivors

Maneesh Bharagav, Trisha L. Becker, Kevin J. Viken, Pratik D. Jagtap, Sanjoy Dey, Michael S. Steinbach, Baolin Wu, Virpal Kumar, Peter B. Bitterman, David H. Ingber, and Christine H. Weng

1Department of Medicine, University of Minnesota, Minneapolis, Minnesota, United States
2Memorial Sloan Kettering Cancer Center, New York, New York, USA
3University of Minnesota, Minneapolis, Minnesota, USA
4Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, Minnesota, USA
5University of Minnesota, Minneapolis, Minnesota, USA
6School of Medicine, University of Minnesota, Minneapolis, USA

Metaproteomic analysis using the Galaxy framework

F. D. Jagtap, A. Blakely, K. Murray, S. Steward, J. Johnson, M. Steinbach, T. Griffin

1Center for Mass Spectrometry and Proteomics, University of Minnesota, Minneapolis, MN, USA
2Department of Bioinformatics, University of Minnesota, Minneapolis, MN, USA
3Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, USA
4University of Minnesota, Minneapolis, MN, USA
5University of Minnesota, Minneapolis, MN, USA
6School of Medicine, University of Minnesota, Minneapolis, USA

Metaproteomic analysis is an emerging field that aims to understand the complex interactions between microorganisms and host health and disease. Galaxy-P is a powerful platform for metaproteomic analysis, enabling researchers to leverage RNA-Seq data for the discovery of novel protein variations. This integration of data types provides a comprehensive view of the biological system, facilitating the identification of disease biomarkers and therapeutic targets.

Open Access

Open Access

Open Access

Open Access

Open Access
The GalaxyP Project
@usegalxyp

Try our public server at usegalxyp.org.

usegalxyp.org
Joined November 2012

---

**Tweets**

1. **z.umn.edu/abrf2016sw4**: Documentation for the #abrf2016 conf.abrf.org/the-galaxy-pla... workshop - #multiomics #proteogenomics

2. **Happening soon in a #usegalaxy not so far, far away...conf.abrf.org/the-galaxy-pla... #multiomics abrf2016 proteogenomics**

3. **Attending the #ABRF2016? Register for the 'Galaxy platform for the #multiomic Data Analysis' satellite workshop! conf.abrf.org/the-galaxy-pla...**
ABRF 2016 Workshop

The Galaxy Platform for Multi-Omic Data Analysis and Informatics

Organizers: Pratik Jagtap and Tim Griffin, University of Minnesota

Instructors: Tim Griffin, Dave Clements, Candace Guerrero, Getiria Onsongo and Pratik Jagtap

The workshop material includes a basic introduction to Galaxy framework and its use for 'omics' data analysis.
Proteogenomics analysis using Galaxy framework I: RNA-Seq data and sequence database generation
Saturday, February 20, 2016

Getiria Onsongo
University of Minnesota
Overview of analysis workflow

- Identify all possible splice junctions
  - FASTQ
  - GTF

- Identify known splice junctions
  - TopHat for Illumina
  - RNA-Seq FASTQ file
  - Gene Model Annotations
  - insertions (bed)
  - deletions (bed)
  - junctions (bed)
  - accepted_hits (bam)

- Novel splice junctions
  - Filter BED on splice junctions
  - BED file
  - reference bed file
  - novel_junctions (bed)

- Genomic DNA
  - Extract Genomic DNA
  - Fetch sequences for intervals in
  - out_file1

- Novel peptides
  - Translate BED Sequences
  - BED file with added sequence column
  - translated_bed (bed)
  - output (fasta)
Data Formats

• fastq
  - stores sequence & quality of individual bases
  - single sequence spans 4 lines

```
@AMELIA:165:C03TUABXX:5:1101:1435:2073 1:N:0:GTGGCC
AGATCGGAAGAGCAGCAGTCTGAACCGCTACGCTATGCGCTGTCTTC
+
CCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCT
```

  - sequence ID
  - sequence bases
  - base quality

• GTF
  - Gene Transfer Format
  - 9 column tab delimited file (seqname, source, feature, start,...)

<table>
<thead>
<tr>
<th>Seqname</th>
<th>Source</th>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strand</th>
<th>Frame</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>protein_coding</td>
<td>exon</td>
<td>157598496</td>
<td>157598715</td>
<td>.</td>
<td>-</td>
<td>-</td>
<td>gene_id &quot;ENSMUSG00000046873&quot;; transcript_id &quot;ENSMUST&quot;</td>
</tr>
<tr>
<td>X</td>
<td>protein_coding</td>
<td>CDS</td>
<td>157598496</td>
<td>15759870</td>
<td>.</td>
<td>-</td>
<td>0</td>
<td>gene_id &quot;ENSMUSG00000046873&quot;; transcript_id &quot;ENSMUST&quot;</td>
</tr>
<tr>
<td>X</td>
<td>protein_coding</td>
<td>start_codon</td>
<td>157598568</td>
<td>157598570</td>
<td>.</td>
<td>-</td>
<td>0</td>
<td>gene_id &quot;ENSMUSG00000046873&quot;; transcript_id &quot;ENSMUST&quot;</td>
</tr>
</tbody>
</table>
Single-end vs Paired-end reads

cDNA fragments

SE

PE

exon
Main advantage of Paired-end reads

genomic DNA

exon 1

second reads anchors

reads can equally match
Single-end vs Paired-end reads

• Paired-end:
  - easier to do quality control (insert size, removing duplicates)
  - easier to map (areas with repetitive sequences)
  - Is cost is no issue, go with PE

• Single-end
  - cheaper
  - works fine with simpler genomes (e.g., bacteria)
Tutorial dataset

1. Small enough for jobs to take a few minutes to run

2. Decent quality RNA-Seq and Mass Spec data

3. Above two points explain use of single-end data
Do Section 2
Section 3: Mapping with Tophat
Caution!!

Reference DB matters
Section 3
Integrative Genomics Viewer (IGV)

Colored line = reference genome

forward read

reverse read

paired-reads

5’ UTR

bait

sequenced reads

SRPK3: exon 1

SRPK3: exon 2
Section 5