Using Galaxy for proteomics and integrative multi-omic analysis

Part 1: Introduction and Proteomics Search

Tim Griffin
Pratik Jagtap
James Johnson (JJ)
The Galaxy-P Team

http://sched.co/5Yg3
Objectives for this workshop:

- Gain some hands-on experience using Galaxy (user interface, Histories, workflows, managing data, sharing Histories and workflows)
- Learn some about integrated multi-omic tools in Galaxy
- Learn about some ways to use Galaxy for multi-omics yourself
- *Forum for questions, comments and suggestions*

Workflow we will follow: Proteogenomics analysis

**Part 1**
- Build FASTA sequence database (including RNA-seq variants)
- Sequence database search using SEARCHGUI
- Identify proteins and peptides using PeptideShaker

**Part 2**
- Determine novel peptide sequences
- Visualize and interpret (MVP and IGV)

**Part 3**

*Cooking show model*
Acknowledgements

**Galaxy-P team**
Pratik Jagtap
Tom McGowan
James Johnson

**Galaxy community collaborators**
Ira Cooke (La Trobe University, Australia)
Bjoern Gruening (University of Freiburg, Germany)
Galaxy core team (Penn State/Johns Hopkins)

Mo Heydarian/Karen Reddy (Johns Hopkins University)

**Globus Genomics**
Ravi Madduri and team

**Funding**

 aws educate

 NSF

 NIH

 University of Minnesota

 Driven to Discover℠
Topic of workshop: proteomics and multi-omics


Image Source:
http://fluorous.com/images/omics.JPG
Technologies: Proteomics

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

- Proteomics
  - TOF MS: 24 MCA scans from Myo_ trypptic.wiff
  - Max. 5191.0 counts.

- Proteins → Peptides
  - "Multidimensional" fractionation
  - μLC

- Isolation
- Fragmentation
- Mass Analysis

- ESI

- Complex mixture
  - Separation
  - turnstile

- MS1
- MS2

- peptide fragments

- m/z

- Detect each component
- 1, 2, 3, ...

- University of Minnesota
  - Driven to Discover™
From sticks on a graph to protein identities

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 peptide sequence matches

Cytochrome C

\[ \text{NH}_2\text{GDVEKGKKIFVQKCAQCHTVEKKGGKHKTGPNLHGLFGRK} \]
\[ \text{QAPGFTYTDANKNKGITWKEETLMEYLENPKKYIPG} \]
\[ \text{TKMIFAGIKKKTTEREDLIAYLKKATNE}_\text{COOH} \]
"What we know is a drop, what we don't know is an ocean."

-- Sir Isaac Newton

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

MS-based proteomics only as good as the database used...

Proteins in database

Proteins actually expressed in sample
Multi-omic example: integrating RNA-seq and proteomic data

- Genome annotation
- Gene expression regulation
- Protein variants in disease
- Functional outcomes of genome mutation
Multi-omic example: integrating RNA-seq and proteomic data
Let’s get started…..

We will go through this process together:

You need a laptop connected to WiFi

1. Go to https://www.globusid.org/create
2. Sign up for Globus account (username, password, email address)
3. Verify email and initiate account
4. Go to msi.globusgenomics.org
Concluding thoughts:
- Online form: http://z.umn.edu/gcc2016fb
  - Provide email and feedback via the form
  - We will follow-up with more information (documentation etc.)

How can I access Galaxy for proteomics?
- Docker container + documentation (information will follow by email)
  - Install locally
  - Cloud deploy
- Globus Genomics
  - Subscription for supported service and scalable deployment in AWS
Create a Globus ID

Username: griffithst2
griffithst2@globusid.org is available
Username may contain both letters and numbers, but must begin with a letter and be between 3 and 31 characters long.
NOTE: this is an ID you are creating — not a working e-mail address

Password: **************
Better passwords are longer, use mixed-case letters with punctuation and numbers.
This password is strong

Full Name: Tim Griffin

E-mail: tgriff1@comcast.net

This account will be used for
- non-profit research or educational purposes
- commercial purposes

Organization: University of Minnesota

I have read and agree to the Globus Terms of Service and Privacy Policy

Please e-mail me updates about Globus

Create ID
Verify E-mail Address

An email was sent to tgriffin@comcast.net.

Please check your e-mail and click on the verification link in that e-mail or enter the verification code that appears in that e-mail into the text box below.

Verification Code

This is a text string like the following:
12345678-90ab-cdef-1234-567895abcdef

Can't find the verification e-mail? Check your spam folder or search for an e-mail from support@globus.org. You may also re-send the verification e-mail.
After verifying email:

- Open pilot2.globusgenomics.org
- Click “Authenticate using Globus”
- If prompted, do not link to any existing accounts
- If prompted, click “Allow” to allow Globus web app
Using Galaxy for proteomics and integrative multi-omic analysis

Part 2: Determining peptides that correspond to novel proteoforms

Tim Griffin
Pratik Jagtap
James Johnson (JJ)
The Galaxy-P Team

http://sched.co/5Yg3
“Our complex workflow (approximately 140 steps) can be easily shared using built-in Galaxy functions, enabling their use and customization by others. Our results provide a blueprint for the establishment of the Galaxy framework as an ideal solution for the emerging field of proteogenomics.”

FILTERING TOOLS FOR Quality Control

Putative novel peptides

Automated, BLAST-P analysis via Galaxy

Novel peptide sequences

“Our customized Galaxy-based software includes automated, batch-mode BLASTP searching and a Peptide Sequence Match Evaluator tool, both useful for evaluating the veracity of putative novel peptide identifications.”
doi: 10.1021/pr500812t
PSM Report

c1: Column 1: Rank of protein group

c2: Protein(s): Accession numbers of protein groups

c3: Sequence: Amino acid sequence of the identified peptide

c4: Variable modifications

c5: Fixed Modifications

c6: Spectrum File: Input MGF file of the identified PSM

c7: Spectrum Title: Fraction number, scan number and charge state

c8: Spectrum Scan Number

c9: Retention Time

c10: m/z: Mass to charge ratio

c11: Measured Charge

c12: Identification Charge

c13: Theoretical Mass: Calculated from identified peptide sequence

c14: Isotope Number

c15: Precursor m/z Error [ppm]

c16: Localization Confidence

c17: probabilistic PTM score

c18: D-score

c19: Confidence

Validation: Confidence > 85 and delta ppm within 6 ppm are CONFIDENT PSMs
Generating a PSM summary of peptides derived from RNA-Seq derived db

Converting peptide list into a FASTA format

BLAST-P searches and filtering
Overview of Workflow

Step 1: Input dataset (PSM Report)

Steps 2-8: Selects peptides with accession number from RNASEq-derived protein FASTA file.
Step 9: PSM Report of peptides identified from RNASEq-derived proteins.

Steps 10-18: Conversion of peptide list into a FASTA format
Step 19: Short BLAST-P on NCBI remote nr mouse database
Step 20: BLAST-P on NCBI remote nr mouse database

Steps 21-28: Identifies mismatched peptides.
Step 29: Peptides corresponding to novel proteoforms.

Steps 30-32: Conversion to PSM Report of peptides corresponding to novel proteoforms.
Step 33: PSM Report of peptides corresponding to novel proteoforms.
View MS/MSMS

- Multiple Spectra can be compared.
- Full functionality of Lorikeet is available
- Backing data can be exported to your Galaxy history
- Backing data can be downloaded as CSV.
MultiOmics Visualization Platform

Filter on Galaxy Data Sets
- Access tabular data sets from history
- Select single or multiple sequences for data search

Select Peptide Sequence(s) for Filtering Data
- Filter by peptide sequence(s)

Filter on Score Values
- Score fields are dynamic
- Score distributions are graphed
- User can choose which scores are viewed/filtered

University of Minnesota
Driven to Discover™
MultiOmics Visualization Platform

Data Exploration and Validation
Starting from your Galaxy History

MVP Viewer is invoked from the existing Galaxy visualization framework

The MVP Viewer has API access back to the originating Galaxy history.

- User can send data back to their history
- User can access other datasets from their Galaxy history
- User will be able to save the entire MVP, including state, back to Galaxy
- User will be able to share a saved MVP with other Galaxy users
Using Galaxy for proteomics and integrative multi-omic analysis

Part 3: Visualizing proteoform peptides using Integrative Genomics Viewer

Tim Griffin
Pratik Jagtap
James Johnson (JJ)
The Galaxy-P Team

http://sched.co/5Yg3
RNAseq to Search Database

• Novel Splice Junctions
  – Tophat GTF vs Tophat allow novel
• Small Variants
  – Tophat/Hisat GATK Mpileup bcftools SnpEff
• Structural Variation
  – deFuse
• DeNovo Transcript Assembly and ORF calling
  – Trinity transcriptsToOrfs (map with blat)
Novel Splice Junctions Workflow
Novel Splice Junctions Workflow

**Protein Fasta**
- >generic|proBJUNC00112203_3|pep:splice chromosome:GRCm38_canon:17:31564812:31588506:1 depth:3
- RLDAGAALRVPIAAACQCDEDWHPDAGGHGVKNRARSQLL

**Translated BED**
- 17 31564812 31588506 proBJUNC00112203_3 3 + 31564812 31588506 255,0,0 2 76,47 0,23647 RLDAGAALRVPIAAACQCDEDWHPDAGGHGVKNRARSQLL

**Mapped Peptide LDAGAALR**
- 17 31564812 31588506 ID=proB_JUNC00112120_3;Name=LDAGAALR 3 + 31564815 31564839 255,0,0 2 76,47 0,23647 LDAGAALR RLDAGAALRVPIAAACQCDEDWHPDAGGHGVKNRARSQLL
## Novel Peptides with Quality PSM

<table>
<thead>
<tr>
<th>Peptide ID</th>
<th>Peptide Sequence</th>
<th>iTRAQ 4-plex of peptide N-term(1)</th>
<th>iTRAQ 4-plex of peptide N-term(2)</th>
<th>iTRAQ 4-plex of peptide N-term(3)</th>
<th>iTRAQ 4-plex of peptide N-term(4)</th>
<th>iTRAQ 4-plex of peptide N-term(5)</th>
<th>iTRAQ 4-plex of peptide N-term(6)</th>
<th>Oxidation of M (3: Very Confident, 6: Very Confident)</th>
<th>Oxidation of M (3: 100.0, 6: 100.0)</th>
<th>Con(99.8333)</th>
</tr>
</thead>
<tbody>
<tr>
<td>proB_JUNC000463249_3</td>
<td>DPSQIGEDR</td>
<td>Mo_Tai_iTRAQ_f6.03569.2</td>
<td>Mo_Tai_iTRAQ_f6.03569.2</td>
<td>35</td>
<td>580.78502</td>
<td>1159.5591</td>
<td>-2.99391554</td>
<td>80.90</td>
<td>Con(99.8333)</td>
<td>286nt</td>
</tr>
<tr>
<td>proB_JUNC0011218920_3</td>
<td>LDAGAALR</td>
<td>Mo_Tai_iTRAQ_f5.05285.2</td>
<td>Mo_Tai_iTRAQ_f5.05285.2</td>
<td>52</td>
<td>465.77852</td>
<td>929.54161</td>
<td>1.031484219</td>
<td>97.00</td>
<td>Con(99.8333)</td>
<td>498nt</td>
</tr>
<tr>
<td>proB_JUNC0009282598_2</td>
<td>GGSPAEGAIR</td>
<td>Mo_Tai_iTRAQ_f5.08716.2</td>
<td>Mo_Tai_iTRAQ_f5.08716.2</td>
<td>87</td>
<td>529.79002</td>
<td>1057.5638</td>
<td>-5.785289619</td>
<td>98.55</td>
<td>Con(100.0)</td>
<td>749nt</td>
</tr>
<tr>
<td>proB_JUNC0008542689_2</td>
<td>QAMGIMNSFVNDIFER</td>
<td>Mo_Tai_iTRAQ_f6.17790.2</td>
<td>Mo_Tai_iTRAQ_f6.17790.2</td>
<td>17</td>
<td>1024.48422</td>
<td>2046.9624</td>
<td>-4.043104841</td>
<td>94.00</td>
<td>Con(100.0)</td>
<td>804nt</td>
</tr>
<tr>
<td>proB_JUNC0008542689_3</td>
<td></td>
<td>Mo_Tai_iTRAQ_f5.00930.2</td>
<td>Mo_Tai_iTRAQ_f5.00930.2</td>
<td>93</td>
<td>424.74002</td>
<td>847.46336</td>
<td>2.501779203</td>
<td>89.10</td>
<td>Con(100.0)</td>
<td>6256nt</td>
</tr>
</tbody>
</table>
Bedtools Intersect for BAM & GTF
SSVAAGGR chr8
SSVAAGGR chr8
NSQTLFQNSLSR chr15
NSQTLFQNSLSR chr15
• Workshops

• Publications
z.umn.edu/galaxypreferences

• Oral Presentations
Proteogenomics GCC 2016 presentation: z.umn.edu/galaxypf1000
http://sched.co/73yn
Metaproteomics ASMS 2016 presentation: z.umn.edu/asmsmp2016vimeo

• Email : tgriffin@umn.edu

• Follow us on: twitter.com/usegalaxyp