Galaxy for Proteogenomics!

GCCBOSC 2018 Workshop
Monday, June 25, 2018

Pratik Jagtap
Timothy Griffin
Praveen Kumar
Subina Mehta

University of Minnesota
WORKSHOP INSTRUCTORS AND ACKNOWLEDGEMENTS

• Instructors
  • Praveen Kumar
  • Prof. Timothy Griffin
  • Subina Mehta

  Galaxy-P team (University of Minnesota)

• Support
  • James Johnson and Thomas McGowan (University of Minnesota)
  • Matthew Chambers
  • Jetstream Cloud at Indiana University

• Funding

[NSF logo]
[National Institutes of Health logo]
GALAXY FOR PROTEOGENOMICS!

- 3:30 pm – 3:40 pm: Introduction to proteogenomics and multi-omic studies!
- 3:40 pm – 4:15 pm: RNASeq Data Processing: Data Analysis using Galaxy platform!
- 4:15 pm – Break
- 4:20 – 5:00 pm Hands-on session for proteomics data analysis using Galaxy!
- 5:00 – Break/questions
- 5:10 – 5:50 pm: Identification of novel proteoforms and visualization!
- 5:50 pm: Wrap up
MULTI-OMICS TECHNOLOGIES

- Next-Gen Sequencing
- RNASeq
- Mass Spectrometry
- Proteogenomics
- Proteo-transcriptomics
- Metaproteomics
- Meta-transcriptomics
- Metabolomics
LOOKING BEYOND THE KNOWN PROTEOME

Mass spectrum

Reference Protein Database from genomic annotation

Identification of peptides corresponding to novel proteoforms.

Cancer / Disease related Databases such as COSMIC, IARC p53, OMIM...

Deep genome sequencing data from ICGC, TCGA and CPTAC

RNASeq data (Customized OR Combined)

6-frame DNA sequences.

3-frame cDNA sequences.
WORKSHOP SCHEDULE

• 3:30 pm – 3:40 pm: Introduction to proteogenomics and multi-omic studies!

• 3:40 pm – 4:15 pm: RNASeq Data Processing: Data Analysis using Galaxy platform!

• 4:15 pm – Break

• 4:20 – 5:00 pm Hands-on session for proteomics data analysis using Galaxy!

• 5:00 – Break/questions

• 5:10 – 5:50 pm: Identification of novel proteoforms and visualization!

• 5:50 pm: Wrap up
DATASET FOR MULTI-OMICS ANALYSIS

- **Mouse cell culture.**

- **RNA-seq analysis**
  RNA-seq libraries were sequenced on a HiSeq 2000 (Illumina SY-401–1001) to a read depth of ~90,000,000 single end 97 bp reads per sample.

- **iTRAQ-labeling and Mass Spectrometry**
  Reversed phase liquid chromatography using Easy-nLC system (Thermo Scientific) and analyzed on a LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific).

---

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. 

Galaxy Instance on JetStream with documentation, tools & workflows for the ABRF 2018 workshop
REGISTER
Workflow #1
RNA-Seq to Variant
FASTA database

Workflow #2
Database Searching
Using MS/MS data

Workflow #3
Identifying Novel Variants
And Visualization
GALAXY FOR PROTEOGENOMICS!

- 3:30 pm – 3:40 pm: Introduction to proteogenomics and multi-omic studies!
- 3:40 pm – 4:15 pm: RNASEq Data Processing: Data Analysis using Galaxy platform!
- 4:15 pm – Break
- 4:20 – 5:00 pm Hands-on session for proteomics data analysis using Galaxy!
- 5:00 – Break/questions
- 5:10 – 5:50 pm: Identification of novel proteoforms and visualization!
- 5:50 pm: Wrap up
OBJECTIVE OF WORKFLOW 1

- Create custom variant database
- Retain genomic coordinate information
Workflow #1
RNA-Seq to Variant
FASTA database

Workflow #2
Database Searching
Using MS/MS data

Workflow #3
Identifying Novel Variants
And Visualization

FASTA Sequences
Genome Mapping Information
Select History 1
Import history
Start using this history

Select Workflow 1
Import workflow
Using the workflow
Run Workflow 1

INPUT
GALAXY
WORKFLOW
OUTPUT
IMPORT HISTORY
IMPORT HISTORY
INPUT DATA

GCC 2018 Workshop
Galaxy For Proteogenomics

Instructors: Timothy Griffin, Prajit Jagtap, Praveen Kumar and Subina Mehta

Workshop Goals
- Introduce the Galaxy framework as a solution for data analysis across ‘omics’ domains
- Provide hands-on experience to attendees in using Galaxy
- Demonstrate use of Galaxy for a proteogenomic analysis (RNA-seq and proteomic integrative analysis)
- Lay the foundation for attendees to implement Galaxy at their own facility or institution to meet ‘omics’ data analysis needs (either specific to one domain or for multi-omics)

Workshop Schedule: Monday June 24, 2018
3:30PM – 3:40PM Introduction to Galaxy Platform and Multi-omic Studies
3:40PM – 4:15PM RNASeq Data Processing: Data Analysis using Galaxy Platform
4:15PM – 4:20PM Break
4:20PM – 5:00PM Hands-on Session for Proteomics Data Analysis Using Galaxy
5:00PM – 5:10PM Break/Questions
5:10PM – 5:50PM Identification of Novel Proteoforms and Visualization
5:50PM – 6:00PM Wrap-up

Accessing Proteogenomics Galaxy instance on JetStream
Workshop documentation
Workshop slides

Please provide us with your feedback

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors. The Galaxy Project is supported in part by NIGMS, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.
INPUT DATA

• RNA-Seq FASTQ file: Reads in FASTQ format

• GTF file: Gene Transfer Format
  • Tabular file to describe genes and related features

• Known protein and contaminant protein sequence FASTA file

• Mass-spectrometry (MGF) file
IMPORT WORKFLOW

Published Workflows

- ABRF_Workflow1_RNAseq_DBcreation
- ABRF_Workflow2_Database_Search_BlastP_ready
- ABRF_Workflow3_Novel_peptide_analysis
### Running a Workflow

#### Your Workflows

<table>
<thead>
<tr>
<th>Name</th>
<th>Tags</th>
<th>Owner</th>
<th># of Steps</th>
<th>Published</th>
<th>Show in tools panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>imported: ABRF_Workflow1_CREATED_RNAseq_DBcreation</td>
<td></td>
<td>You</td>
<td>22</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2_Database_Search_BlastIP_ready</td>
<td></td>
<td>You</td>
<td>14</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>1_Novel_peptide_analysis</td>
<td></td>
<td>You</td>
<td>11</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>MED_Regex_HL_Complete_RNAseq_Sample_DB (imported from uploaded file)</td>
<td></td>
<td>You</td>
<td>47</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2_Database_Search_BlastIP_ready</td>
<td></td>
<td>You</td>
<td>14</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

#### History

- output: ABRF_workshop 4 steps, 4 tools
- 839.16 MB
- 57: 44.bed
- 47: identified peptides in .bam
- 45: msไท MGFs with 4 steps
- 43: Ref 5000 uniprot.fasta
- 40: 1 fasta
- 39: ABRF_RNAseq_Cal3
- 34: 8.86 off
- 1: Prok_fasta
SELECTING INPUT FILES TO RUN A WORKFLOW

- FASTQ file (# 1)
- GTF file (# 2)
- FASTA file (# 3)
<table>
<thead>
<tr>
<th>Job Status</th>
<th>Job Running</th>
<th>Job Successful</th>
<th>Job Failed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job in queue</td>
<td><img src="https://via.placeholder.com/150x150" alt="Job in queue" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job running" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job failed" /></td>
</tr>
<tr>
<td>Job running</td>
<td><img src="https://via.placeholder.com/150x150" alt="Job running" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job run successful" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job failed" /></td>
</tr>
<tr>
<td>Job successful</td>
<td><img src="https://via.placeholder.com/150x150" alt="Job run successful" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job run successful" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job failed" /></td>
</tr>
<tr>
<td>Job failed</td>
<td><img src="https://via.placeholder.com/150x150" alt="Job run successful" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job run successful" /></td>
<td><img src="https://via.placeholder.com/150x150" alt="Job failed" /></td>
</tr>
</tbody>
</table>

**JOB STATUS (HISTORY PANE)**

- **Job in queue**: 89: Peptide Shaker on data 38: Peptide Report
- **Job running**: 71: Peptide Shaker on data 38: Peptide Report
- **Job successful**: 150: Peptide Shaker on data 145: PSM Report
- **Job failed**: 75: Peptide Shaker on data 70: PSM Report
- **Job in queue**: 88: Peptide Shaker on data 38: PSM Report
- **Job running**: 70: Peptide Shaker on data 38: PSM Report
- **Job successful**: 149: Peptide Shaker on data 145: Parameters
- **Job failed**: 74: Peptide Shaker on data 70: Parameters
- **Job in queue**: 87: Peptide Shaker on data 38: Parameters
- **Job running**: 69: Peptide Shaker on data 38: Parameters
- **Job successful**: 148: Peptide Shaker on data 145: Archive
- **Job failed**: 73: Peptide Shaker on data 70: Archive
- **Job in queue**: 86: Peptide Shaker on data 38: Archive
- **Job running**: 68: Peptide Shaker on data 38: Archive
- **Job successful**: 147: Peptide Shaker on data 145: CPS file
- **Job failed**: 72: Peptide Shaker on data 70: CPS file
- **Job in queue**: 85: Peptide Shaker on data 38: mzidentML file
- **Job running**: 67: Peptide Shaker on data 38: mzidentML file
- **Job successful**: 146: Peptide Shaker on data 145: mzidentML file
- **Job failed**: 71: Peptide Shaker on data 70: mzidentML file
**Workflow #1**
RNA-Seq to Variant
FASTA database

**Workflow #2**
Database Searching
Using MS/MS data

**Workflow #3**
Identifying Novel Variants
And Visualization

WORKSHOP WORKFLOWS
WORKFLOW #1: RNA-SEQ TO VARIANT PROTEIN

SAV / In-Del Variants

Assembly Workflow
POTENTIAL NOVEL PEPTIDE IDENTIFICATIONS

Expressed 5' UTR
Alternate start
Alternate frame
Novel Exon
Novel Spliceform
Exon extension
Expressed 3' UTR
/Alternate stop
Intergenic
/Novel gene
Single amino acid variant
InDels

5' Exon 1 | 5' Exon 2 | 5' Exon 3 | 5' Exon 4 | 5' Exon 5 | 5' Exon 6 | 5' Exon 7 | 5' Exon 8

3' Exon 1 | 3' Exon 2 | 3' Exon 3 | 3' Exon 4 | 3' Exon 5 | 3' Exon 6 | 3' Exon 7 | 3' Exon 8
RNA-SEQ TO FASTA DATABASE CREATION

**Assembly Workflow**
- **HISAT Alignment tool**
- **STRINGTIE** RNA-Seq to transcripts
- **GFF COMPARE** Evaluates the assembly with annotated transcripts
- **FREEBAYES** Variant Calling
- **CustomPro DB**
- **Variant annotation**
- **Genome mapping**

**SAV / In-Del Variants**
- **RNA-Seq FASTQ**
- **GTF**
- **Mapping Files**
- **Sequence FASTA**

**Protein annotation**
- **Variant calling**
- **Genome mapping**

**Workflow components**
- **Genome**
- **RNA-Seq FASTQ**
- **GTF**
RNA-SEQ TO FASTA DATABASE CREATION

- **Genome**
  - **RNA-Seq FASTQ**
    - **HISAT Alignment tool**
      - **FREEBAYES Variant Calling**
        - **CustomPro DB**
          - Variant annotation
          - Genome mapping
          - SAV / In-Del Variants

- **GTF**
  - **STRINGTIE RNA-Seq to transcripts**
    - **GFF COMPARE**
      - Evaluates the assembly with annotated transcripts
      - **Translate transcripts**
        - **Mapping Files**
          - **Sequence FASTA**
Alignment

Reference gene/genome

Mapping to gene/genome

HISAT2: Outputs BAM file (Dataset #9)
Freebayes: Outputs VCF file (Dataset #14)

Garrison E., Marth G. Haplotype-based variant detection from short-read sequencing. (arXiv:1207.3907)
VIEWING SNP VARIANT IN IGV
RNA-SEQ TO FASTA DATABASE CREATION

Assembly Workflow

- **GENOME**
  - **RNA-SEQ FASTQ**
  - **GTF**

**Alignment tool**
- **HISAT**

**Variant Calling**
- **FREEBAYES**
  - Variant annotation
  - Genome mapping

**CustomPro DB**
- Mapping Files
  - Sequence FASTA

**RNA-Seq to transcripts**
- **STRINGTIE**

**Translate transcripts**

**Evaluates the assembly with annotated transcripts**
- **GFF COMPARE**

**Variant annotation**
- **Genome mapping**
ALIGNMENT

Mapping to gene/genome

Reference gene/genome
TRANSCRIPT ASSEMBLY

Reference gene/genome → Mapping to gene/genome → Splicing → Assembled Transcript → 3-Frames Translation → FASTA Sequence
FASTA Sequence File

>generic|ENSMUSP00000107433|Erp29|ER protein 29
MAAAAGVSGAASLSPLLSVLLGLLLLFAPHGGSGLHTKGALPLDTVTFYKSRLLLGP

>generic|ENSMUSP00000120715|Rps2|ribosomal protein S2
MADDAGAAGGPGGPGLGGRGFRGGFGSGSLRGRGRGRGRGRGRGRGRGGKAEDKEWIPVTKLNLVLDVLMKIKSLEEYLFSLPIKESEIIDDFFLGASLKDDEVKIMPVQKQTRAGQR

Genomic Mapping File

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSP00000107433</td>
<td>chr5</td>
<td>121452190</td>
<td>121452340</td>
<td>-</td>
<td>150</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr5</td>
<td>121449139</td>
<td>121449163</td>
<td>-</td>
<td>174</td>
</tr>
<tr>
<td>ENSMUSP00000107433</td>
<td>chr5</td>
<td>121452190</td>
<td>121452340</td>
<td>-</td>
<td>150</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr5</td>
<td>121449139</td>
<td>121449163</td>
<td>-</td>
<td>174</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24720275</td>
<td>24720452</td>
<td>+</td>
<td>177</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24720533</td>
<td>24720731</td>
<td>+</td>
<td>375</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24720968</td>
<td>24721302</td>
<td>+</td>
<td>375</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24721622</td>
<td>24721727</td>
<td>+</td>
<td>709</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24721802</td>
<td>24721897</td>
<td>+</td>
<td>909</td>
</tr>
</tbody>
</table>
GALAXY FOR PROTEOGENOMICS!

- 3:30 pm – 3:40 pm: Introduction to proteogenomics and multi-omic studies!

- 3:40 pm – 4:15 pm: RNASeq Data Processing: Data Analysis using Galaxy platform!

- 4:15 pm – Break

- 4:20 – 5:00 pm Hands-on session for proteomics data analysis using Galaxy!

- 5:00 – Break/questions

- 5:10 – 5:50 pm: Identification of novel proteoforms and visualization!

- 5:50 pm: Wrap up
WORKSHOP WORKFLOWS

Workflow #1
RNA-Seq to Variant
FASTA database

Workflow #2
Database Searching
Using MS/MS data

Workflow #3
Identifying Novel Variants
And Visualization
4:20 PM – 05:00 PM: HANDS-ON SESSION FOR PROTEOMICS DATA ANALYSIS USING GALAXY

**Protein FASTA:** reference proteins + potential variants

RNA-Seq database → SearchGUI

- Multiple algorithms for matching MS/MS to peptides

SearchGUI → PEPTIDE SHAKER

- Organization and scoring of peptide spectral matches (PSMs)

PEPTIDE SHAKER → Mz to sqlite

- Generation of an sqLite database for downstream data visualization and filtering

Mz to sqlite → Peptides for BLASTP

- Putative variant peptide sequences for further verification and analysis

MGF → Peaklist of MS/MS data
Peptide fractionation coupled to tandem mass spectrometry (MS/MS)
Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification

• Bundles a multiple freely-available algorithms for matching MS/MS to peptide sequences


• Infers proteins from peptide sequence matches
• Assigns confidence scores to peptide sequence matches and inferred proteins
• Provides outputs in standard formats (e.g. mzidentML) for further processing
<table>
<thead>
<tr>
<th>Process</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search GUI</td>
<td>Protein Database =&gt; searchgui_results (searchgui_archive)</td>
</tr>
<tr>
<td>Peptide Shaker</td>
<td>Compressed SearchGUI results =&gt; mzidentML (mzid) =&gt; output_cps (peptideshaker_archive) =&gt; output_2.zip (zip) =&gt; output_certificate (txt) =&gt; output hierarchical (tabular) =&gt; output psm (tabular) =&gt; output_psm_xml (tabular) =&gt; output_extended_psm (tabular) =&gt; output peptides (tabular) =&gt; output_proteins (tabular)</td>
</tr>
<tr>
<td>mz to sqlite</td>
<td>Proteomics Identification files =&gt; Proteomics Spectrum files =&gt; Proteomics Search Database =&gt; Fasta =&gt; sqlite_db (mz.sqlite)</td>
</tr>
<tr>
<td>Removing Reference Proteins</td>
<td>Add tables to this Database =&gt; Database Table 1 &gt; Tabular Dataset for Table =&gt; Database Table 2 &gt; Tabular Dataset for Table =&gt; Database Table 3 &gt; Tabular Dataset for Table</td>
</tr>
<tr>
<td>FASTA-to-Tabular</td>
<td>Convert these sequences =&gt; output (tabular)</td>
</tr>
<tr>
<td>Cut</td>
<td>From</td>
</tr>
<tr>
<td>Convert</td>
<td>In Dataset</td>
</tr>
<tr>
<td>Cut</td>
<td>From</td>
</tr>
<tr>
<td>Convert</td>
<td>In Dataset</td>
</tr>
</tbody>
</table>
GALAXY FOR PROTEOGENOMICS!

- 3:30 pm – 3:40 pm: Introduction to proteogenomics and multi-omic studies!
- 3:40 pm – 4:15 pm: RNASEq Data Processing: Data Analysis using Galaxy platform!
- 4:15 pm – Break
- 4:20 – 5:00 pm Hands-on session for proteomics data analysis using Galaxy!
- 5:00 – Break/questions
- 5:10 – 5:50 pm: Identification of novel proteoforms and visualization!
- 5:50 pm: Wrap up
WORKSHOP WORKFLOWS

Workflow #1
RNA-Seq to Variant
FASTA database

Workflow #2
Database Searching
Using MS/MS data

Workflow #3
Identifying Novel Variants
And Visualization
YOUR CURRENT HISTORY
In order to access the input for this part of the workshop, click on “Shared Data” → “Histories” → “Inputs_for_ABRF_workshop”. And click on ABRF_History 2.
Select ‘ABRF_Workflow3_Novel_peptide_analysis’ from Shared Directory

Import workflow

Start using this workflow

Run Workflow
WORKFLOW FOR THIS SECTION
WORKFLOW FOR THIS SECTION

Workshop Documentation: z.umn.edu/abrf18doc
5.2 BlastP analysis 32
5.3 Novel proteoform analysis 33
5.4 Using Multi-omics Visualization Platform for visualizing novel proteoforms 35
SELECT DISTINCT psm.*
FROM psm JOIN blast ON psm.sequence = blast.qseqid
WHERE blast.pident < 100 OR blast.gapopen >= 1 OR blast.length < blast.qlen
ORDER BY psm.sequence, psm.id
MULTI-OMICS VISUALIZATION PLATFORM FOR VISUALIZING NOVEL PROTEOFORMS

mz to sqlite on data 36, data 7, and others

Peptide Overview

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Spectra Count</th>
<th>Protein Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVPDSSAEASGLR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AVPDSSAEASGLRAQDR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DGDLENPVLYSGAVG</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DSGASGSILEASAAR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ELGSSDLTAR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ESSREALVEPTSESPPALAR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NIVTLLESR</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>SPYREFTDHLVK</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Showing 1 to 8 of 8 entries (filtered from 4,976 total entries)

Selected Peptide PSMs PSMs Filtered by Score

Filtering Sequences

Novel Peptides
AVPDSSAEASGLR to SPYREFTDHLVK
MULTI-OMICS VISUALIZATION PLATFORM FOR VISUALIZING NOVEL PROTEOFORMS

SPECTRAL QUALITY VISUALIZATION (Lorikeet Viewer)

GENOMIC LOCALIZATION (Integrated Genomics Viewer)
## NOVEL PROTEOFORM ANALYSIS

<table>
<thead>
<tr>
<th>#PeptideSequence</th>
<th>SpectralCount</th>
<th>Proteins</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
<th>Annotation</th>
<th>GenomeCoordinate</th>
<th>UCSC Genome Browser</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSGASGSILEAAR</td>
<td>1</td>
<td>STRC:13.1_92.325</td>
<td>chr17</td>
<td>22860957</td>
<td>22867642</td>
<td>-</td>
<td>five_prime_utr</td>
<td>chr17:22860957-22867642</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:22860957-22867642">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:22860957-22867642</a></td>
</tr>
<tr>
<td>ELGSSDLTA</td>
<td>1</td>
<td>STRG:17.1_64.379</td>
<td>chr2</td>
<td>91155262</td>
<td>91155292</td>
<td>-</td>
<td>intron</td>
<td>chr2:91155262-91155292</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr2:91155262-91155292">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr2:91155262-91155292</a></td>
</tr>
<tr>
<td>NIYITLSCFK</td>
<td>3</td>
<td>STRG:22.1_70.262</td>
<td>chr4</td>
<td>58482350</td>
<td>58482383</td>
<td>+</td>
<td>intergene</td>
<td>chr4:58482350-58482383</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr4:58482350-58482383">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr4:58482350-58482383</a></td>
</tr>
<tr>
<td>SPYREFTDHLVK</td>
<td>1</td>
<td>ENSMUSP00000013474.E590.Q267R</td>
<td>chr17</td>
<td>24721702</td>
<td>24721826</td>
<td>+</td>
<td>SpliceJunction</td>
<td>chr17:24721702-24721826</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:24721702-24721826">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:24721702-24721826</a></td>
</tr>
</tbody>
</table>

### Table:

<table>
<thead>
<tr>
<th>chrom</th>
<th>chromStart</th>
<th>chromStop</th>
<th>name</th>
<th>score</th>
<th>strand</th>
<th>annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr11</td>
<td>115176449</td>
<td>115176491</td>
<td>AVPDSSAESGLR</td>
<td>255</td>
<td>+</td>
<td>CDS</td>
</tr>
<tr>
<td>chr11</td>
<td>115176449</td>
<td>115176503</td>
<td>AVPDSSAESGLRAQRK</td>
<td>255</td>
<td>+</td>
<td>CDS</td>
</tr>
<tr>
<td>chr5</td>
<td>121445444</td>
<td>121445489</td>
<td>DGDLENPVLYSGAVK</td>
<td>255</td>
<td>-</td>
<td>CDS</td>
</tr>
<tr>
<td>chr17</td>
<td>22866997</td>
<td>22867042</td>
<td>DSGASGSILEAAR</td>
<td>255</td>
<td>-</td>
<td>five_prime_utr</td>
</tr>
<tr>
<td>chr2</td>
<td>91155262</td>
<td>91155292</td>
<td>ELGSSDLTA</td>
<td>255</td>
<td>-</td>
<td>intron</td>
</tr>
<tr>
<td>chr11</td>
<td>115180006</td>
<td>115180069</td>
<td>ESSREALVPTSESPRALAR</td>
<td>255</td>
<td>+</td>
<td>CDS</td>
</tr>
<tr>
<td>chr4</td>
<td>58482350</td>
<td>58482383</td>
<td>NIYITLSCFK</td>
<td>255</td>
<td>+</td>
<td>intergene</td>
</tr>
<tr>
<td>chr17</td>
<td>24721702</td>
<td>24721826</td>
<td>SPYREFTDHLVK</td>
<td>255</td>
<td>+</td>
<td>SpliceJunction</td>
</tr>
</tbody>
</table>
CDART BLAST SEARCH

CONSERVED Domain Architecture Retrieval Tool

[Diagram showing domain architectures and superfamily]

- **Query**: ezrin binding protein 50, partial
  - Taxonomy span: Bilateria
  - Similarity score: 2
  - Total nr sequences: 899
  - Lookup sequences in Entrez

- **Predicted**: PDZ and LIM domain protein 5
  - Taxonomy span: Eutheria
  - Similarity score: 2
  - Total nr sequences: 10
  - Lookup sequences in Entrez

- **EBP50_C superfamily**

  - cl07569, EBP50, C-terminal; This C terminal domain allows interaction of EBP50 with FERM (four-point one ERM) domains, resulting in the activation of Ezrin-radixin-moesin (ERM), with subsequent cytoskeletal modulation and cellular growth control. It includes a disordered section between two reasonably well conserved hydrophobic regions.
GALAXY FOR PROTEOGENOMICS!

- 3:30 pm – 3:40 pm: Introduction to proteogenomics and multi-omic studies!
- 3:40 pm – 4:15 pm: RNASeq Data Processing: Data Analysis using Galaxy platform!
- 4:15 pm – Break
- 4:20 – 5:00 pm Hands-on session for proteomics data analysis using Galaxy!
- 5:00 – Break/questions
- 5:10 – 5:50 pm: Identification of novel proteoforms and visualization!
- 5:50 pm: Wrap up
PROJECT OVERVIEW

**AIM 1. Extend our Galaxy-plugin MVP tool for visualization, interpretation and data exchange**
- UMN/MSi developers
- Feedback from DCP collaborators

**AIM 2. Extend Galaxy and the MVP tool for metabolite profiling in cancer research**
- Hegeman (UMN)
- Metabolomics tool developers
- ITCR groups
- Driving cancer projects (DCPs)

**AIM 3. Extend Galaxy and the MVP tool for integrative genomic-proteomic informatics and workflows**
- Smith (UW-Madison)
- Martens (Ghent University/VIB)
- Tool developer community
- ITCR groups
- Driving Cancer Projects

**AIM 4. Catalyze use by cancer researchers via dissemination, promotion and training activities**
- Deploy in accessible resources
- Promote easy Galaxy-instance installation and interoperability
- Promote usage (bench scientists and developers) publication, presentation, workshops
- Provide bench scientist training opportunities

Multi-omic informatics hub for cancer researchers
ACKNOWLEDGMENTS

Biologists / collaborators
Laurie Parker
Joel Rudney
Maneesh Bhargava
Amy Skubitz
Chris Wendt
Brian Crooker
Steven Friedenberg
Kevin Viken
Kristin Boylan
Marnie Peterson
Somiah Afiuni
Brian Sandri
Alexa Pragman
Wanda Weber
Amy Treeful

Minnesota Supercomputing Institute
James Johnson
Thomas McGowan
Lee Parsons
Michael Milligan

Harald Barsnes
Marc Vaudel
University of Bergen, Norway

Carolin Kolmeder
University of Helsinki,
Finland

Lennart Martens (Co-I)
Bart Mesuere
Robbert G Singh
VIB, UGhent, Belgium

Judson Hervey
Naval Research
Institute
Washington, D.C.

Alessandro Tanca
Porto Conte
Ricerche, Italy

Anamika Krishanpal
Priyabrata Panigrahi
Persistent Systems Limited

Karen Reddy
Mo Heydarian
UW-Madison

Thomas Doak
Jeremy Fisher
Indiana University

Matt Chambers
Nashville, TN

Bjoern Gruening
Bérénice Batut
University of Freiburg,
Freiburg, Germany

Ira Cooke
Melbourne, Australia

GalaxyP
www.galaxyp.org
QUESTIONS?

Workshop Documentation: z.umn.edu/abrf18doc
Visit: http://galaxyp.org
Follow us on twitter.com/usegalaxyp

Feedback: https://z.umn.edu/gcc18fb
Come discuss:

• **Current state of proteomics tools in Galaxy**
• **Plans for future**
• **Ideas for promoting and increasing user base**

Join us!

Tuesday @ 6:20 pm, Reed Commons Cafe