SW2: THE GALAXY PLATFORM AS AN ACCESSIBLE, CORE LABORATORY SOLUTION FOR PROTEOGENOMIC ANALYSIS AND INFORMATICS

ABRF 2018 Satellite Workshop
Sunday, April 22, 2018

Pratik Jagtap
Timothy Griffin
Praveen Kumar
James Johnson

University of Minnesota
WORKSHOP INSTRUCTORS AND ACKNOWLEDGEMENTS

• Instructors
  • Praveen Kumar
  • Prof. Timothy Griffin
  • Pratik Jagtap
  • James Johnson

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

• Sponsors

• Funding
8:00 am – 8:15 am: Introduction to Galaxy Platform and multi-omic studies

8:15 am – 9:15am: RNASeq Data Processing: Data Analysis using Galaxy platform

9:15 am – 10:00 am: Hands-on session for proteomics data analysis using Galaxy

10:00 am – 10:30 am: Morning Refreshment Break

10:30 am – 12:00 pm: Identification of novel proteoforms and visualization

12:00 pm – 1:00 pm: Lunch

SW2: THE GALAXY PLATFORM AS AN ACCESSIBLE, CORE LABORATORY SOLUTION FOR PROTEOGENOMIC ANALYSIS AND INFORMATICS
MULTI-OMICS


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

6-frame DNA sequences.
3-frame cDNA sequences.

RNASEq data (Customized OR Combined)
Benefits of Galaxy

- A **WEB-BASED** bioinformatics data analysis platform.
- **ACCESSIBILITY**, **USABILITY**, **SHARE-ABILITY** of software tools, workflows & histories.
- **REPRODUCIBILITY** and ability to compare results after using variable parameters.

- **Software tools** can be used in a sequential manner to generate **ANALYTICAL WORKFLOWS** that can be reused, shared and creatively modified for multiple studies.

TOOLS & WORKFLOWS

• **Software tools** can be used in a sequential manner to generate **analytical workflows** that can be reused, shared and creatively modified for multiple studies.

For example, Protein Database Downloader downloads UniProt protein FASTA databases of various organisms.
WORKSHOP SCHEDULE

Session 1
- 8:15 am – 9:15am: RNASeq Data Processing: Data Analysis using Galaxy platform

Session 2
- 9:15 am – 10:00 am: Hands-on session for proteomics data analysis using Galaxy
  - 10:00 am – 10:30 am: Morning Refreshment Break

Session 3
- 10:30 am – 12:00 pm: Identification of novel proteoforms and visualization
  - 12:00 pm – 1:00 pm: Lunch
• 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.

z.umn.edu/abrf2018

Galaxy Instance on JetStream with documentation, tools & workflows for the ABRF 2018 workshop
ABRF 2018 Workshop

(SW2) The Galaxy Platform as an Accessible, Core Laboratory Solution for Proteogenomic Analysis and informatics

Organizer: Pratik Jagtap, University of Minnesota
Instructors: Tim Griffin, Pratik Jagtap and Praveen Kumar

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: Sunday April 22, 2018

8:00AM – 8:15AM  Introduction to Galaxy Platform and multi-omic studies. (Pratik Jagtap)
8:15AM – 9:15AM  RNASeq Data Processing: Data Analysis using Galaxy platform. (Praveen Kumar)
9:15AM – 10:00AM  Hands-on session for proteomics data analysis using Galaxy. (Tim Griffin)
10:00AM – 10:30AM  Morning Refreshment Break
10:30AM – 12:00PM  Identification of novel proteoforms and visualization. (Pratik Jagtap)
12:00PM – 1:00PM  Lunch

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 NIH, NSF, The Hack Institute of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.
SW2: THE GALAXY PLATFORM AS AN ACCESSIBLE, CORE LABORATORY SOLUTION FOR PROTEOGENOMIC ANALYSIS AND INFORMATICS

- 8:00 am – 8:15 am: Introduction to Galaxy Platform and multi-omic studies
- 8:15 am – 9:15am: RNASeq Data Processing: Data Analysis using Galaxy platform
- 9:15 am – 10:00 am: Hands-on session for proteomics data analysis using Galaxy
- 10:00 am – 10:30 am: Morning Refreshment Break
- 10:30 am – 12:00 pm: Identification of novel proteoforms and visualization
- 12:00 pm – 1:00 pm: Lunch
• 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

GALAXY
- INPUT

WORKFLOW
- OUTPUT
GALAXY INTERFACE

Left Pane
- Tools

Center Pane
- Tool details
- Preview data
- Workflow details
- Workflow canvas
- List of Histories

Right Pane
- History (list of input and output data)
- Tool details (when editing workflow)

ABRF 2018 Workshop
(SW2) The Galaxy Platform as an Accessible, Core Laboratory Solution for Proteogenomic Analysis and Informatics
Organizer: Pratik Jagtap, University of Minnesota
Instructors: Tim Griffin, Pratik Jagtap and Praveen Kumar

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: Sunday April 22, 2018
8:00AM – 8:15AM Introduction to Galaxy Platform and multi-omic studies. (Pratik Jagtap)
8:15AM – 9:15AM RNASeq Data Processing: Data Analysis using Galaxy platform. (Praveen Kumar)
9:15AM – 10:00AM Hands-on session for proteomics data analysis using Galaxy. (Tim Griffin)
10:00AM – 10:30AM Morning Refreshment Break
10:30AM – 12:00PM Identification of novel proteoforms and visualization. (Pratik Jagtap)
12:00PM – 1:00PM Lunch

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 partners.
The Galaxy Project is supported by DOE, NIGMS, The Joint Institute of Life Sciences, The Institute for Cell Engineering at Johns Hopkins, and Johns Hopkins University.
IMPORT HISTORY
ABRF 2018 Workshop

(SW2) The Galaxy Platform as an Accessible, Core Laboratory Solution for Proteogenomic Analysis and Informatics

Organizer: Pratik Jagtap, University of Minnesota
Instructors: Tim Griffin, Pratik Jagtap and Praveen Kumar

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: Sunday April 22, 2018

8:00AM – 8:15AM  Introduction to Galaxy Platform and multi-omic studies. (Pratik Jagtap)
8:15AM – 9:15AM  RNASeq Data Processing: Data Analysis using Galaxy platform. (Praveen Kumar)
9:15AM – 10:00AM  Hands-on session for proteomics data analysis using Galaxy. (Tim Griffin)
10:00AM – 10:30AM  Morning Refreshment Break
10:30AM – 12:00PM  Identification of novel proteoforms and visualization. (Pratik Jagtap)
12:00PM – 1:00PM  Lunch

Accessing Proteogenomics Galaxy instance on JetStream
Workshop documentation
Workshop slides

Please provide us with your feedback
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
IMPORT WORKFLOW
### RUNNING A WORKFLOW

#### Galaxy Interface

- **Your workflows**
  - **Name**: imported: ABBF Workflow1
  - **Tags**: You
  - **Owner**: You
  - **# of Steps**: 22
  - **Published**: No

- **Workflow Details**
  - **Name**: 2_Database Search_HistP_ready
    - **Name**: 1_Novel_peptide_analysis
      - **Name**: WERD_RegEx_Rep_CompRes_DE (imported from uploaded file)
        - **Name**: 4_MuTal MGPs
          - **Name**: 3_Ref_5000_uniprot_sRA
            - **Name**: 2_Hu_maximus_GRCm38
              - **Name**: 1_PoA_fasta
SELECTING INPUT FILES TO RUN A WORKFLOW

- FASTQ file (# 1)
- GTF file (# 2)
- FASTA file (# 3)
## JOB STATUS (HISTORY PANES)

<table>
<thead>
<tr>
<th>Job Status</th>
<th>Job ID</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job in queue</td>
<td>89: Peptide Shaker on data 38: Peptide Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>88: Peptide Shaker on data 38: PSM Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>87: Peptide Shaker on data 38: Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>86: Peptide Shaker on data 38: Archive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>85: Peptide Shaker on data 38: mzidentML file</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job running</td>
<td>71: Peptide Shaker on data 38: Peptide Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>70: Peptide Shaker on data 38: PSM Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>69: Peptide Shaker on data 38: Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>68: Peptide Shaker on data 38: Archive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>67: Peptide Shaker on data 38: mzidentML file</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job successful</td>
<td>150: Peptide Shaker on data 145: PSM Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>149: Peptide Shaker on data 145: Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>148: Peptide Shaker on data 145: Archive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>147: Peptide Shaker on data 145: CPS file</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>146: Peptide Shaker on data 145: mzidentML file</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job failed</td>
<td>75: Peptide Shaker on data 70: PSM Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>74: Peptide Shaker on data 70: Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>73: Peptide Shaker on data 70: Archive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72: Peptide Shaker on data 70: CPS file</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>71: Peptide Shaker on data 70: mzidentML file</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Emojis:**
- 🟥 Job in queue
- 🌶 Job running
- 🟢 Job successful
- 🟥 Job failed

**Status Icons:**
- 🟥 Job failed
- 🟢 Job successful
Workflow #1
RNA-Seq to Variant
FASTA database

Workflow #2
Database Searching
Using MS/MS data

Workflow #3
Identifying Novel Variants
And Visualization
WORKFLOW #1: RNA-SEQ TO VARIANT PROTEIN

SAV / In-Del Variants

Assembly Workflow
POTENTIAL NOVEL PEPTIDE IDENTIFICATIONS

5' Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7 Exon 8 3'
Expressed 5' UTR Known Peptides
Alternate start
Alternate frame
Novel Exon
Novel Spliceform
Exon extension
Expressed 3' UTR
/Alternate stop
Intergenic
/Novel gene
Single amino acid variant
InDels

+2 +2 +2 +3

+1

+3

Start Stop

A+1
RNA-SEQ TO FASTA DATABASE CREATION

- Genome
- RNA-Seq FASTQ
- GTF

**Alignment tool**
- HISAT

**Variant Calling**
- FREEBAYES

**RNA-Seq to transcripts**
- STRINGTIE

**Variant annotation**
- CustomPro DB

**Genome mapping**
- Mapping Files

**Sequence FASTA**
- Translate transcripts

**SAV / In-Del Variants**
- Evaluates the assembly with annotated transcripts
- GFF COMPARE

**Files**
- SAV / In-Del Variants
ALIGNMENT

Reference gene

Mapping to gene/genome

HISAT2: Outputs BAM file (Dataset #9)
VARIANT CALLING

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

STRINGTIE
- RNA-Seq to transcripts

HISAT
- Alignment tool

FREEBAYES
- Variant Calling

Evaluates the assembly with annotated transcripts

GFF COMPARE
- Translate transcripts

CustomPro DB
- Variant annotation
- Genome mapping

Mapping Files
- Sequence FASTA
ALIGNMENT

Mapping to gene/genome

Reference gene
TRANSCRIPT ASSEMBLY

Reference gene

Mapping to gene/genome

Splicing

Assembled Transcript

3-Frames Translation

FASTA Sequence
### FASTA Sequence File

>generic|ENSMUSP00000107433|Erp29|ER protein 29
MAAAAGVSGAASLSPLLSVLLGLLLLFAPHGGSLHTKGALPLDVTFTYKSRLLLGP

>generic|ENSMUSP00000120715|Rps2|ribosomal protein S2
MADDAGAAGGPGPPLGGGRRGFRGGFSGLRGRGRRGRGRGRGRRGRARGGKAEDKEWIPVTKLGLRKIMSKLEEEIXLFSLPIKESEIIDFFLGASLKLDEVLKIMPVQKQTRAGQR

### Genomic Mapping File

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Start Position</th>
<th>End Position</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>ENSMUSP00000107433</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>709</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24721622</td>
<td>24721727</td>
<td>+</td>
<td>814</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24721802</td>
<td>24721897</td>
<td>+</td>
<td>909</td>
</tr>
</tbody>
</table>
SW2: THE GALAXY PLATFORM AS AN ACCESSIBLE, CORE LABORATORY SOLUTION FOR PROTEOGENOMIC ANALYSIS AND INFORMATICS

- 8:00 am – 8:15 am: Introduction to Galaxy Platform and multi-omic studies
- 8:15 am – 9:15 am: RNASeq Data Processing: Data Analysis using Galaxy platform
- 9:15 am – 10:00 am: Hands-on session for proteomics data analysis using Galaxy
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
9:15 AM – 10:00 AM: HANDS-ON SESSION FOR PROTEOMICS DATA ANALYSIS USING GALAXY

Protein FASTA: reference proteins + potential variants

- RNA-Seq database
- MGF

Multiple algorithms for matching MS/MS to peptides

SearchGUI
Proteomics. 11:996-9

PEPTIDE SHAKER
Histone ChIP-Seq
Nat Biotechnol. 33:22-4

Organization and scoring of peptide spectral matches (PSMs)

mzIdentMl

Mz to sqlite

Generation of an SQLite database for downstream data visualization and filtering

Putative variant peptide sequences for further verification and analysis

Peptides for BLASTP

GalaxyP
SW2: THE GALAXY PLATFORM AS AN ACCESSIBLE, CORE LABORATORY SOLUTION FOR PROTEOGENOMIC ANALYSIS AND INFORMATICS

- 8:00 am – 8:15 am: Introduction to Galaxy Platform and multi-omic studies
- 8:15 am – 9:15 am: RNASeq Data Processing: Data Analysis using Galaxy platform
- 9:15 am – 10:00 am: Hands-on session for proteomics data analysis using Galaxy
- 10:00 am – 10:30 am: Morning Refreshment Break
- 10:30 am – 12:00 pm: Identification of novel proteoforms and visualization
- 12:00 pm – 1:00 pm: Lunch
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
RECAP: DATABASE GENERATION

**Assembly Workflow**

- **STRINGTIE** (RNA-Seq to transcripts)
- **GFF COMPARE** (Evaluates the assembly with annotated transcripts)
- **FREEBAYES** (Creates variant file)
- **CustomPro DB**
  - Variant annotation
  - Genome mapping
- **Mapping Files**
- **Sequence FASTA**
- **Translate transcripts**

**SAV / In-Del Variants**

- **HISAT** (Alignment tool)
- **Genome**
- **RNA-Seq FASTQ**
- **GTF**
- **FREEBAYES** (Creates variant file)
- **SAV / In-Del Variants**

**Genome**

**RNA-Seq FASTQ**

**GTF**

**FREEBAYES** (Creates variant file)

**CustomPro DB**

**Mapping Files**

**Sequence FASTA**

**Translate transcripts**

**GFF COMPARE** (Evaluates the assembly with annotated transcripts)
RECAP: PROTEOMICS SEARCH

RNA-Seq database → SearchGUI → PEPTIDE SHAKER

MGF

mzidentml → Mz to sqlite → Peptides for BLASTP
YOUR CURRENT HISTORY

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>#ID</td>
<td>Sequence</td>
</tr>
<tr>
<td>GAILIJK PSM=3081 length=7</td>
<td>GAILIJK</td>
</tr>
<tr>
<td>LNMELSEK PSM=623 length=8</td>
<td>LNMELSEK</td>
</tr>
<tr>
<td>LSVVIPVR PSM=2071,2328 length=8</td>
<td>LSVVIPVR</td>
</tr>
<tr>
<td>ETHQQVRSR PSM=1264,1585 length=9</td>
<td>ETHQQVRSR</td>
</tr>
<tr>
<td>LLATGQRDR PSM=3149,4918 length=9</td>
<td>LLATGQRDR</td>
</tr>
<tr>
<td>SLNLTAFR PSM=4948 length=9</td>
<td>SLNLTAFR</td>
</tr>
<tr>
<td>ELGSSOLTAR PSM=4554 length=10</td>
<td>ELGSSOLTAR</td>
</tr>
<tr>
<td>FDTQPYGKEK PSM=1325,3728 length=10</td>
<td>FDTQPYGKEK</td>
</tr>
<tr>
<td>LVEPGSAEK PSM=3462 length=10</td>
<td>LVEPGSAEK</td>
</tr>
<tr>
<td>QGLDGLLSVK PSM=2783,3970 length=10</td>
<td>QGLDGLLSVK</td>
</tr>
<tr>
<td>CTIVASAPVK PSM=3271,4263,7777,4799 length=11</td>
<td>CTIVASAPVK</td>
</tr>
<tr>
<td>NVTLLSSCFK PSM=2493,3195,4155 length=11</td>
<td>NVTLLSSCFK</td>
</tr>
<tr>
<td>TYSYLTPDLWK PSM=2634,2814 length=11</td>
<td>TYSYLTPDLWK</td>
</tr>
<tr>
<td>CTIVASAPVK PSM=2179,2332 length=12</td>
<td>CTIVASAPVK</td>
</tr>
<tr>
<td>SPYQEFHLVQK PSM=399,529,2097 length=12</td>
<td>SPYQEFHLVQK</td>
</tr>
<tr>
<td>SPYQEFHLVQK PSM=767 length=12</td>
<td>SPYQEFHLVQK</td>
</tr>
<tr>
<td>LLVDKVETDEFFK PSM=290,872 length=13</td>
<td>LLVDKVETDEFFK</td>
</tr>
<tr>
<td>SLEELFSLIPK PSM=5147,5715 length=13</td>
<td>SLEELFSLIPK</td>
</tr>
<tr>
<td>AVPDSSAESQLR PSM=5730 length=14</td>
<td>AVPDSSAESQLR</td>
</tr>
<tr>
<td>DGCLENPLYSLVCAV PSM=4725,5467 length=15</td>
<td>DGCLENPLYSLVCAV</td>
</tr>
<tr>
<td>DDGASGSELSAAAR PSM=5463 length=15</td>
<td>DDGASGSELSAAAR</td>
</tr>
<tr>
<td>FVAEEDYNQVACLK PSM=296 length=16</td>
<td>FVAEEDYNQVACLK</td>
</tr>
<tr>
<td>AVPDSSAESQLRQIDR PSM=434 length=18</td>
<td>AVPDSSAESQLRQIDR</td>
</tr>
<tr>
<td>SGLLADRLVEVNGENVEK PSM=315,337,502 length=19</td>
<td>SGLLADRLVEVNGENVEK</td>
</tr>
<tr>
<td>ESSREALEEPETSSEPAILAR PSM=2168 length=21</td>
<td>ESSREALEEPETSSEPAILAR</td>
</tr>
<tr>
<td>SEQEPFPPAAADTHEAGDQNEAEK PSM=889 length=23</td>
<td>SEQEPFPPAAADTHEAGDQNEAEK</td>
</tr>
<tr>
<td>VIPSEQHLDIPLEFPESNGEQK PSM=3337 length=23</td>
<td>VIPSEQHLDIPLEFPESNGEQK</td>
</tr>
</tbody>
</table>
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

ACTIVE HISTORY FROM EARLIER WORKFLOW

Published Workflows

<table>
<thead>
<tr>
<th>Name</th>
<th>Style</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABRF_Workflow1_RNAseq_DBcreation</td>
<td>Dropdown</td>
</tr>
<tr>
<td>ABRF_Workflow2_Database_Search_BlastP_ready</td>
<td>Dropdown</td>
</tr>
<tr>
<td>ABRF_Workflow3_Novel_peptide_analysis</td>
<td>Dropdown</td>
</tr>
</tbody>
</table>

Advanced Search
WORKFLOW FOR THIS SECTION

- **Peptides for Blast**
  - Add tables to this Database
  - Output

- **Tabular-to-Fasta**
  - Tab-delimited file
  - Output (Fasta)

- **NCBI BLAST+**
  - Protein query sequence(s)
  - Output (Tabular, Ext. HTML, BlastXML)

- **PeptideNeaker_PSM**
  - Output

- **mz_to_sqlite**
  - Output

- **Genomic_mapping_sqlite**
  - Output

- **PSM_Novel_Peptides**
  - Add tables to this Database
  - Database Table 1 > Tabular Dataset for Table
  - Database Table 2 > Tabular Dataset for Table
  - SQLite (SQLite)
  - Output (Tabular)

- **Peptide Genomic Coordinate**
  - Peptide list (without any header line)
  - mz to sqlite (mzsqlite) file
  - Genomic mapping sqlite file
  - Genome_bed (Bed)

- **Novel_Peptides**
  - Add tables to this Database
  - Database Table 1 > Tabular Dataset for Table
  - SQLite (SQLite)
  - Output (Tabular)

- **PepPointer**
  - BED file with chromosomal coordinates of peptide
  - Classified (Tabular)

- **Final_Summary_Novel_Peptides_Output**
  - Add tables to this Database
  - Database Table 1 > Tabular Dataset for Table
  - Database Table 2 > Tabular Dataset for Table
  - SQLite (SQLite)
  - Output (Tabular)
WORKFLOW FOR THIS SECTION

Workshop Documentation: [z.umn.edu/abrf18doc](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**

### Peptides Overview

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Spectra Count</th>
<th>Protein Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVPDSSAEAGGLR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AVPDSSAEAGGLRAQDR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DGQILEPNVLYSGAV</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DSGWAGSILEEASAAR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ELSGLDTR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ESSREALVEPTSESRRALAR</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NVYITLSC</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>SPYREFTIDLVX</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**
MULTI-OMICS VISUALIZATION PLATFORM FOR VISUALIZING NOVEL PROTEOFORMS

SPECTRAL QUALITY VISUALIZATION (Lorikeet Viewer)

GENOMIC LOCALIZATION (Integrated Genomics Viewer)
GENOMIC LOCALIZATION (INTEGRATED GENOMICS VIEWER)
# NOVEL PROTEOFORM ANALYSIS

<table>
<thead>
<tr>
<th>ProteinSequence</th>
<th>SpectralCount</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>AVDPDSSAESGLR</td>
<td>1</td>
<td>chr11</td>
<td>115176499</td>
<td>115176491</td>
<td>+</td>
<td>CDS</td>
<td>chr11:115176499-115176491</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr11:115176499-115176491">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr11:115176499-115176491</a></td>
</tr>
<tr>
<td>AVDPDSSAESGLRQAQRD</td>
<td>1</td>
<td>chr11</td>
<td>115176499</td>
<td>115176503</td>
<td>+</td>
<td>CDS</td>
<td>chr11:115176499-115176503</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr11:115176499-115176503">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr11:115176499-115176503</a></td>
</tr>
<tr>
<td>DGDLENPVLVSAGVK</td>
<td>2</td>
<td>chr5</td>
<td>121445444</td>
<td>121445489</td>
<td>-</td>
<td>CDS</td>
<td>chr5:121445444-121445489</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr5:121445444-121445489">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr5:121445444-121445489</a></td>
</tr>
<tr>
<td>DSCGASGILEASAAR</td>
<td>1</td>
<td>chr17</td>
<td>228609597</td>
<td>22867042</td>
<td>-</td>
<td>five_prime_utr</td>
<td>chr17:228609597-22867042</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:228609597-22867042">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:228609597-22867042</a></td>
</tr>
<tr>
<td>ELGSSDLTAR</td>
<td>3</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>ESSREALVEPTSESPRLAR</td>
<td>1</td>
<td>chr11</td>
<td>115180006</td>
<td>115180069</td>
<td>+</td>
<td>CDS</td>
<td>chr11:115180006-115180069</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr11:115180006-115180069">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr11:115180006-115180069</a></td>
</tr>
<tr>
<td>NIYITLSCFK</td>
<td>3</td>
<td>chr17</td>
<td>58482350</td>
<td>58482383</td>
<td>+</td>
<td>intergene</td>
<td>chr17:58482350-58482383</td>
<td><a href="https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:58482350-58482383">https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&amp;position=chr17:58482350-58482383</a></td>
</tr>
<tr>
<td>SPYREFTDHLVK</td>
<td>1</td>
<td>chr17</td>
<td>24721702</td>
<td>24721826</td>
<td>+</td>
<td>Splice Junction</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 1: Protein Details

<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>115176499</td>
<td>115176491</td>
<td>AVDPDSSAESGLR</td>
<td>255</td>
<td>+</td>
<td>CDS</td>
</tr>
<tr>
<td>chr11</td>
<td>115176499</td>
<td>115176503</td>
<td>AVDPDSSAESGLRQAQRD</td>
<td>255</td>
<td>+</td>
<td>CDS</td>
</tr>
<tr>
<td>chr5</td>
<td>121445444</td>
<td>121445489</td>
<td>DGDLENPVLVSAGVK</td>
<td>255</td>
<td>-</td>
<td>CDS</td>
</tr>
<tr>
<td>chr17</td>
<td>228609597</td>
<td>22867042</td>
<td>DSCGASGILEASAAR</td>
<td>255</td>
<td>-</td>
<td>five_prime_utr</td>
</tr>
<tr>
<td>chr2</td>
<td>91155262</td>
<td>91155292</td>
<td>ELGSSDLTAR</td>
<td>255</td>
<td>-</td>
<td>intron</td>
</tr>
<tr>
<td>chr11</td>
<td>115180006</td>
<td>115180069</td>
<td>ESSREALVEPTSESPRLAR</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>Splice Junction</td>
</tr>
</tbody>
</table>
CDART BLAST SEARCH

CONSERVED Domain Architecture Retrieval Tool

([Query] &nbsp; (Local query sequence) NP_036160.1 Na(+)/H(+) exchange regulatory cofactor
Total architectures: 3261

- 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

- PDZ and LIM domain protein 5
  taxonomy span: Boreoeutheria
  Similarity score: 2
  Total nr sequences: 5

EBP50_C superfamily

c07569, 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.
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
Collaborators
- 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
Collaborators
- Smith (UW-Madison)
- Martens (Ghent University/VIAB)
- 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

---

GalaxyP

---
ACKNOWLEDGMENTS

University of Minnesota
Tim Griffin
Pratik Jagtap
Praveen Kumar
Candace Guerrero
Subina Mehta
Adrian Hegeman
Art Eschenlauer
Shane Hubler

Biologists / collaborators
Joel Rudney
Maneesh Bhargava
Kevin Viken
Amy Skubitz
Kristin Boylan
Marnie Peterson
Somiah Afiuni
Chris Wendt
Brian Sandri
Alexa Pragman

GalaxyP

Brook Nunn
U Washington
Josh Elias
Stanford University
Lloyd Smith
Michael Shortreed
UW-Madison

Minneapolis Supercomputing Institute
James Johnson
Tom McGowan
Getiria Onsongo
Michael Milligan

Harald Barsnes & Marc Vaudel
University of Bergen, Norway

Lennart Martens
Bart Mesuere
Felix Van der Jeugt
VIB, UGhent, Belgium

Judson Hervey
Naval Research Institute
Washington, D.C.

Alessandro Tanca
Porto Conte
Ricerche, Italy

Karen Reddy
Mo Heydarian
Johns Hopkins University

Bjoern Gruening
Clemens Blank
Bérénice Batut
University of Freiburg,
Freiburg, Germany

Anamika Krishnapal
Persistent Systems Limited

Carolin Kolmeder
University of Helsinki,
Finland

Thilo Muth
Bernhard Renard
Robert Koch Institut

Ira Cooke
Melbourne, Australia

Funding

NSF
NIH
NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

Thomas Doak
Jeremy Fisher
Indiana University

Lennart Martens
Bart Mesuere
Felix Van der Jeugt
VIB, UGhent, Belgium

Judson Hervey
Naval Research Institute
Washington, D.C.

Alessandro Tanca
Porto Conte
Ricerche, Italy

Karen Reddy
Mo Heydarian
Johns Hopkins University

Bjoern Gruening
Clemens Blank
Bérénice Batut
University of Freiburg,
Freiburg, Germany

Anamika Krishnapal
Persistent Systems Limited

Carolin Kolmeder
University of Helsinki,
Finland

Thilo Muth
Bernhard Renard
Robert Koch Institut

Ira Cooke
Melbourne, Australia

Funding

NSF
NIH
NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research
QUESTIONS?

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

Feedback: https://z.umn.edu/abrf18fb