Mass Spectrometry-Based Proteomics Data Analysis Using GalaxyP

GCC 2015 GalaxyP Workshop

July 6th, 2015
Norwich, UK

Presenters:
Tim Griffin, Pratik Jagtap and James Johnson

Documentation:
Kevin Murray, Ray Sajulga and Pratik Jagtap

Infrastructure Support:
Thomas McGowan

Contents

1 Introduction
1.1 What is the GalaxyP Project? 3
1.2 Why Proteomics? 3
1.3 Basics of Mass Spectrometry for Proteomics 3
1.4 Scope of this Tutorial 4

2 Inputs: Peaklists and Database Generation
2.1 What Do You Need for Proteomic Analysis. 5
2.2 Getting Started - GCC2015 GalaxyP Tutorial. 6

3 Search Algorithms
3.1 Basics of Search Algorithms 8
3.2 SearchGUI in GalaxyP 8

4 Interpreting Protein/Peptide Identifications
4.1 Protein Inference. 11
4.2 Target-Decoy Search. 11
4.3 PeptideShaker in GalaxyP 11
4.4 PeptideShaker Outputs. 13

5 BLAST-P Search
5.1 Proteoforms 14
5.2 BLAST-P 14
5.3 Data Manipulation via Workflow. 15

6 Visualizing Proteomic Results
6.1 Importance of Visualization 15
6.2 PSM Evaluator in GalaxyP. 16

7 Proteogenomics
7.1 Peptides to Genome 18
7.2 Navigating to IGV Browser from GalaxyP. 19

8 Running entire Proteogenomics workflow. 22

1 Introduction

1.1 What is the GalaxyP Project?

GalaxyP is a web-based multiple 'omics' data analysis platform with particular emphasis on mass spectrometry based proteomics. GalaxyP is developed at the University of Minnesota, deployed at the Minnesota Supercomputing Institute, and is an extension of the popular Galaxy project. The GalaxyP project is supported by a grant from NSF.

1.2 Why Proteomics?

The original Galaxy project was created to provide a bioinformatics platform for accessibility, user-friendliness, and standardization in genomics. Although genomics can further the understanding of an organism in a mostly holistic sense, it cannot accurately predict the protein products that arise from gene expression at specific times and environments. On the other hand, proteomic analysis can track which proteins are expressed, how much of each type are expressed, and how they associate with one another under different conditions. Proteomic analysis can also identify and characterize post-translational modifications (PTMs) and protein complexes. That being said, proteomics is still an emerging field in need of optimization at multiple steps of analysis. Analyses from the same sample using different sample preparation methods, instruments and data analysis platforms can yield different results between laboratories - thus requiring standardization. Also, like genomics, proteomic analyses uses multiple techniques and steps that can be difficult to keep track of without a central hub. Much like how Galaxy was developed for genomics, GalaxyP was created to solve these problems.

For more about proteomics: http://proteomics.cancer.gov/whatisproteomics

1.3 Basics of Mass Spectrometry for Proteomics

As mentioned previously, there are many techniques used to analyze proteomes, but at the moment, mass spectrometry is the main technique, generating large amounts of data from complex biological samples.
Protein sample needs to be processed before introducing into a mass spectrometer. To begin, the proteins in a sample are separated from other cell parts using fractionation or affinity selection methods (e.g. SDS-PAGE, immunoprecipitation etc.). Next, the proteins can either be digested into peptides (**bottom-up/shotgun proteomics**) or kept intact as a protein (**top-down proteomic**s). The proteins/peptides are separated further through a liquid chromatography (LC) system as they are introduced into the mass spectrometer.

A mass spectrometer analyzes molecules of all types using three main steps:

1. The creation of fragment ions is carried out in the gas phase through volatilization and ionization in a vacuum. The most common techniques to achieve this are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI).
2. Ions are then separated based on their **mass-to-charge ratio (m/z)**. Different mass spectrometers achieve this separation using different physical mechanisms, such as quadrupole mass filters or time-of-flight.
3. Separated ions are detected when they strike a detector which converts their kinetic energy to a current, which is detected as a signal. The recorded information is converted to a mass spectrum consisting of m/z values and corresponding intensities.

Protein Identification

Knowing only the mass-to-charge ratios (m/z) of intact proteins or digested peptides and their corresponding intensities can prove difficult for identification since multiple proteins and peptides might share the same known m/z. By using tandem mass spectrometry (MS/MS), m/z values of ionized peptides can be measured (also called"MS1" spectrum or precursor ion). Each individual peptide can be isolated and fragmented, and the m/z values of the fragments detected ("MS2" or "MS/MS" spectrum or fragment ions). The fragments are made up of a small number of amino acids that make up the parent peptide. By using the m/z of the intact peptide (MS1 spectrum) and comparing the fragmentation pattern detected in the MS2 spectrum to a database of expected fragmentation patterns of known or predicted peptide sequences, the amino acid sequence of detected peptides can be determined. The peptide sequence correlated to the full sequences of known proteins, inferring the presence of that protein within the sample.

Evaluating the Results

Instruments and experiments are prone to differences in performance, leading to variability in data quality in proteomics experiments. Some common terms are used pertaining to data quality produced by mass spectrometers. One term, **mass accuracy**, evaluates how far an experimentally measured m/z value deviates from the actual value. Usually expressed in parts per million (ppm), a lower value is preferred, meaning the instrument used has high mass accuracy. Another term, **mass resolution**, represents the ratio between the mass of an analyte and the width of its observed peak. In other words, it evaluates how easy it is to differentiate between different peaks of very similar m/z values. Wider peaks can mask two different analytes with close values. Narrower peak widths provide the ability to resolve peaks close in m/z value, and provide a higher mass resolution value, which is preferred.

1.4 Scope of this Tutorial

- ➔ Learn the basics of data analysis for proteomics and proteogenomics.
- ➔ Learn about the tools available in Galaxy for proteomics.
- ➔ Learn how to build some workflows and use tools for proteomic/proteogenomic analysis.

The workflow modules in this tutorial can be run as a single, complete workflow (see section 8 below) - but for sake of this tutorial we will run the modules separately so as to understand the details of steps used.

2 Inputs: Peaklists and Database Generation

2.1 What do you need for Proteomic Analysis?

Getting GalaxyP
For the purposes of this tutorial at GCC 2015, a special cloud instance of GalaxyP has been generated. To obtain a local instance of Galaxy download the latest Galaxy source code and install the Proteomics tools from the Galaxy Tool Shed. Users can also take advantage of the public GalaxyP server to associate themselves with the Galaxy framework and current open-source proteomic tools.

What are RAW files and Peaklists?
RAW files are datasets generated experimentally using Thermofinnigan instruments and contain raw information pertaining to a mass spectrometry run. Peaklists (e.g. mzml and MGF files) are generated by processing multiple RAW data files. Processing includes multiple steps such as peak detection (including intensity), noise removal, baseline correction, monoisotope peak correction, charge state derivation, etc. The MGF (short for Mascot generic format) file is used as an input for multiple search algorithms (Mascot, ProteinPilot, OMSSA, etc.). The file encodes multiple experimental MS/MS spectra in a single file with m/z and its associated intensity pairs separated by headers. The header for each spectral scan has information about Peptide mass, charge state, scan number etc. For more information about commonly used file formats read manuscript by Deutsch.

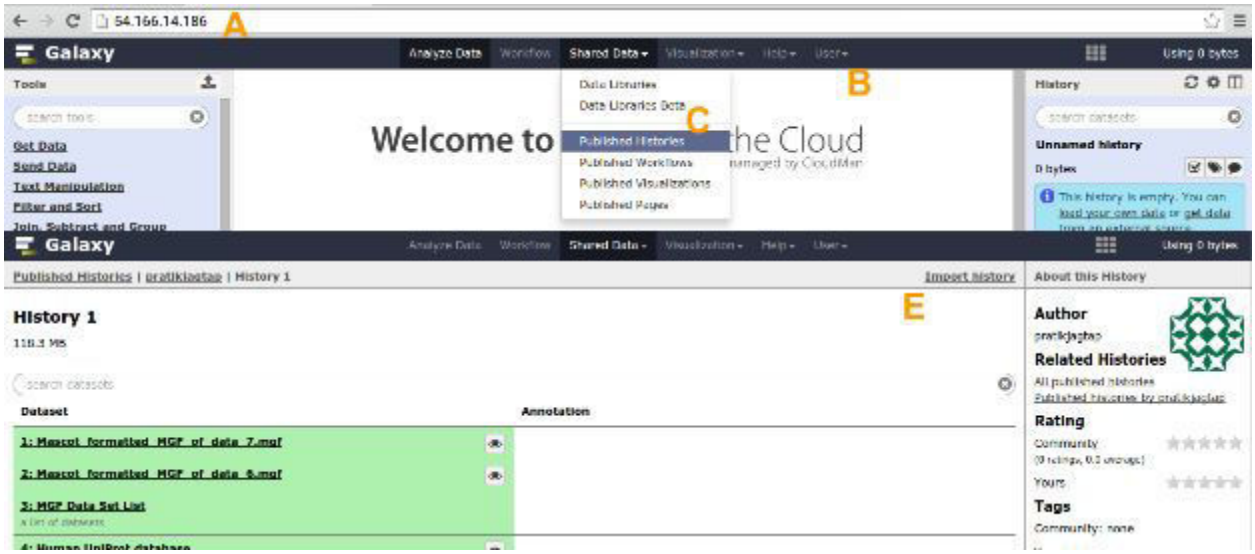
Generating a Database
Database searching is one of the two most common approaches for "bottom-up" proteomics, and involves correlating mass spectra with peptide sequences in a protein database. To positively identify proteins within your sample from a database search, a comprehensive yet precise database is required. Incomplete databases will overlook spectra and large databases have a high false discovery rate and low sensitivity. UniProt offers links to reference proteomes for multiple organisms whose whole genome sequence information and annotation is available. GalaxyP custom tool, Protein Database Downloader, possesses a wide array of commonly used databases for numerous species. Missing databases can easily be uploaded and formatted into GalaxyP.

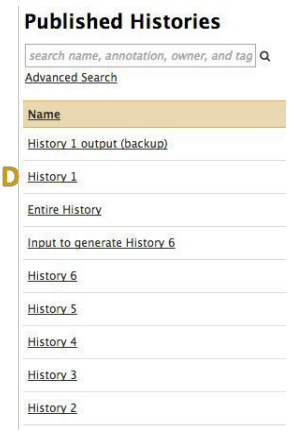
2.2 Getting Started - GCC2015 GalaxyP Tutorial

A. Open a web browser and navigate to the GalaxyP cloud instance - http://54.205.17.20

B. At the top of the screen select *User* and *Register* and *Log in* for your Cloud username (your email) and password.

C. At the top of the screen select *Shared Data* then migrate to *Published Histories*.

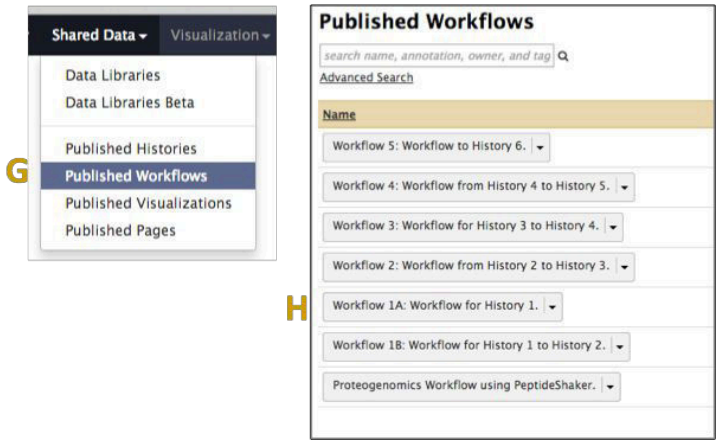




D. Select *History 1* from the list of published histories.

E. Select *Import history* to add the selected history to your user histories.

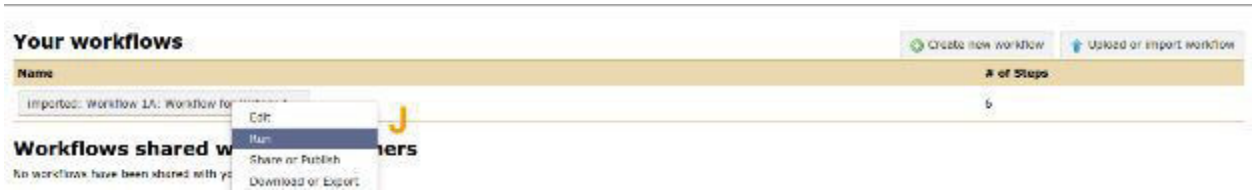
F. On the confirmation screen select *start using this history* to navigate to this history in the Galaxy view.



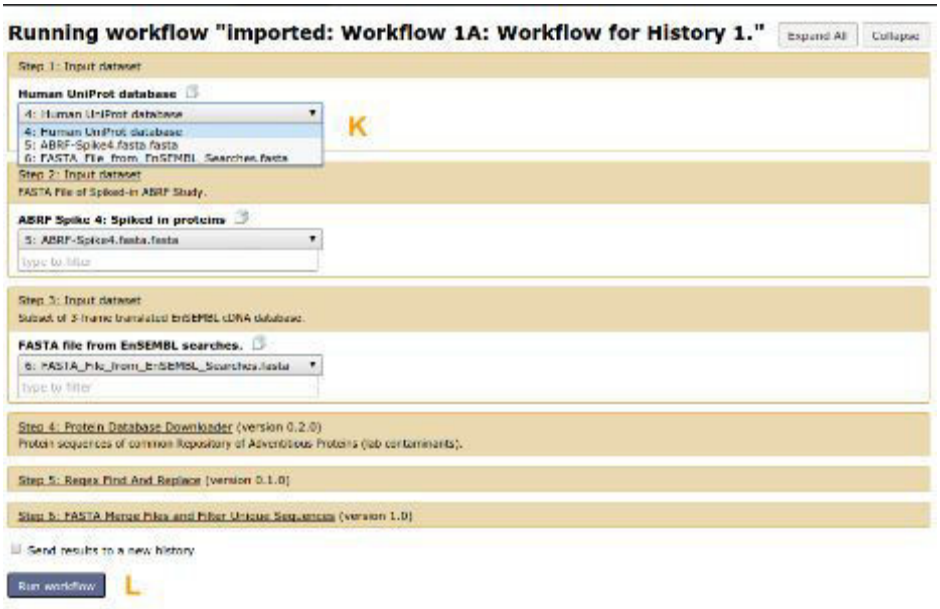
G. At the top of the screen select *Shared Data* then migrate to *Published Workflows*.

H. Select *Workflow 1A: Workflow for History 1* from the list of published workflows and choose *import* to copy the workflow into your user workflows.

I. On the confirmation screen, select *start using this workflow* to navigate to your user workflows.



J. In the workflows menu select *Run* Workflow 1A: Workflow for History 1 from the drop down menu.



K. Appropriately assign each input database from History 1 to the corresponding input or the workflow.

L. Run the workflow.

3 Search Algorithms

3.1 Basics of Search Algorithms

What Does a Search Algorithm Do?
To confirm the presence of a protein within a sample, the observed fragmentation of the peptides within a sample must be correlated to an amino acid sequence of a protein within a database. A peptide-spectrum match (PSM) represents the correlation of a peptide fragmentation to a sequence within a database. A search algorithm provides information of the database hit and assigns a score to each PSM based on the quality of the match based on features such as number of matched ions, modifications used for search, mass accuracy at MS and MS/MS level, etc. The results of a search algorithm may be visualized and analyzed by various software to filter for high-quality PSMs and validate identifications.

Why So Many Search Algorithms?
No two search algorithms are perfectly alike. Each algorithm identifies different PSMs from a database search. For a number of search algorithms, many PSM identifications overlap. However some algorithms may identify proteins that another does not. So why not use every search algorithm? Database searching is a long, intensive process requiring considerable system resources. Some search algorithms are more efficient than others, making the use of some algorithms redundant. However, using too few or inefficient algorithms results in missed identifications. Each experiment requires consideration to determine the quantity and quality of search algorithms that align with experimental aims and resources. For more information about search algorithms please read review by Eng et al (2011).

3.2 SearchGUI in GalaxyP

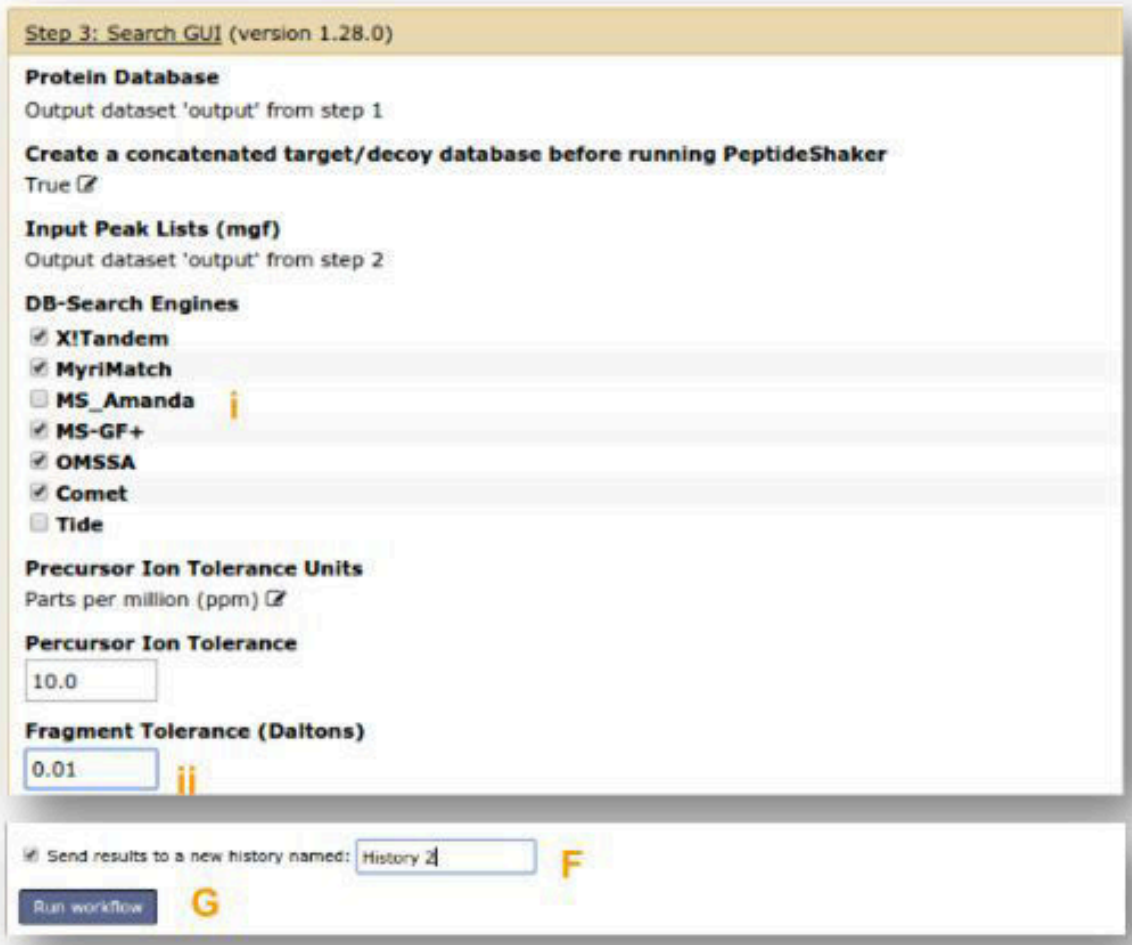
SearchGUI in GalaxyP
Optimizing SearchGUI parameters for your database search will result in improved identifications for further analysis. SearchGUI can utilize data generated from different instrumentations, utilizing generic MGF peaklist files as an input. As part of optimization of SearchGUI parameters, also see Basics of Database Generation for further advice on generating a comprehensive database that is right for your data.

Setting Parameters
Selecting the proper parameters to match your sample preparation and instruments specificity is crucial for obtaining accurate database matches using SearchGUI. Search parameters may be edited directly from the tool menu of SearchGUI or within the workflow interface. Not all modifications are handled correctly by some search engines (see SearchGUI source code for

more details).

Search Engines
SearchGUI implements multiple search engines to obtain proteomic identifications. SearchGUI currently supports X!Tandem, MS-GF+, MS Amanda, MyriMatch, Comet, Tide, and OMSSA search engines; however, full implementation of all search engines is pending within Galaxy. Users may choose to use any number of search engines when performing a database search with SearchGUI. PeptideShaker is the suggested tool for visualizing and analyzing results from SearchGUI (see PeptideShaker in Galaxy for more detail).

A. From *Published Workflows*, *import* *Workflow 1B: Workflow for History 1 to History 2*.
B. On the confirmation screen, select *start using this workflow* to navigate to your user workflows.
C. *Run* Workflow 1B on your current history (History 1).
D. Appropriately select each dataset from History 1 that corresponds to input within the workflow.



E. In Step 3: SearchGUI some parameters must be set at run time:
    a. Choose to run X!Tandem, MyriMatch, MS-GF+, OMSSA, and Comet
    b. Change the Fragment Tolerance to 0.01 Daltons
F. Select the box *Send results to a new history named:* and enter *History 2* into the field provided.
G. Select *Run* and navigate to History 2 in the normal Galaxy view. **For the purposes of this tutorial, a completed History 2 may be imported from Published Histories (refer to section 2.2).**

4 Interpreting Protein/Peptide Identifications

4.1 Protein Inference

In shotgun proteomics (see Section 1.3), proteins are analyzed by breaking them down experimentally into peptides; extracting the resulting peptide fragment information, and then using that information to reconstruct them into proteins computationally.

More information about protein inference can be found in the second half of the following SciVee Conferences Demo: http://www.scivee.tv/node/12671

4.2 False Discovery Rate based on Target-Decoy Search

In order to maintain accuracy and effectiveness in spectral / peptide / protein identification, a target-decoy search strategy can be used to discern how correct and incorrect a spectral or peptide or protein match is. The most popular approach for generating decoy databases is the 'reverse database'approach. Essentially, protein sequences are reversed to generate a 'decoy' database. Any matches and their associated scores against a target and decoy database are noted (with the premise that matches against decoy matches are incorrect). Later the matches are ranked according to descending scores and 'decoy matches' are used to calculate false discovery rate (FDR) to set a threshold for valid identifications. The FDR approach allows for a fairer comparison of datasets across labs, machines and proteomic workflows. Please read manuscript by Elias and Gygi (2010) for more information.

4.3 PeptideShaker in GalaxyP

To interpret the protein/peptide identifications by SearchGUI, GalaxyP uses a platform called PeptideShaker. It reports information about spectra and PSMs, proteins, peptides, and also provides an mzidentML file that can be used for PSM visualization. Refer to the next section for more information about the outputs. For more PeptideShaker-related information visit the Github webpage for Galaxy version of PeptideShaker, Download link for standalone version or manuscript by Vaudel et al (2015).

A. At the top of the screen, click *Shared Data* then select the option: *Published Workflows*.
B. Click *Workflow 2: Workflow for History 2 to History 3* and choose *import*.
C. Select *start using this workflow*.
D. Click *imported: Workflow 2: Workflow for History 2 to History 3* and choose *edit* to open the workflow editor.

E. To view the whole workflow (two tools*), navigate to the right and down by clicking and dragging the grid…

    i. ...or use the minimap in the lower right corner.

**note: Galaxy workflows are usually more complex with many more tools and connections; for the sake of this tutorial we will start simple.*

F. Click on the PeptideShaker tool to view its details in the right panel.

To run, click the cogwheel at the top right corner of the middle frame and select *run* from the resulting menu.

a. The input for this workflow under SearchGUI Results should be *8: Search GUI on data 7, data 2, and data 1* from *History 3*. *If this is not the case, then refer to section 2.2 on how to import a complete history. Alternatively, create the history yourself by following section 3.2.*

b. Check *Send results to a new history* then input the name *History 3*.

c. Select *Run workflow*.

4.4 PeptideShaker Outputs

Protein Report
- valid proteins
- coverage
- molecular weight

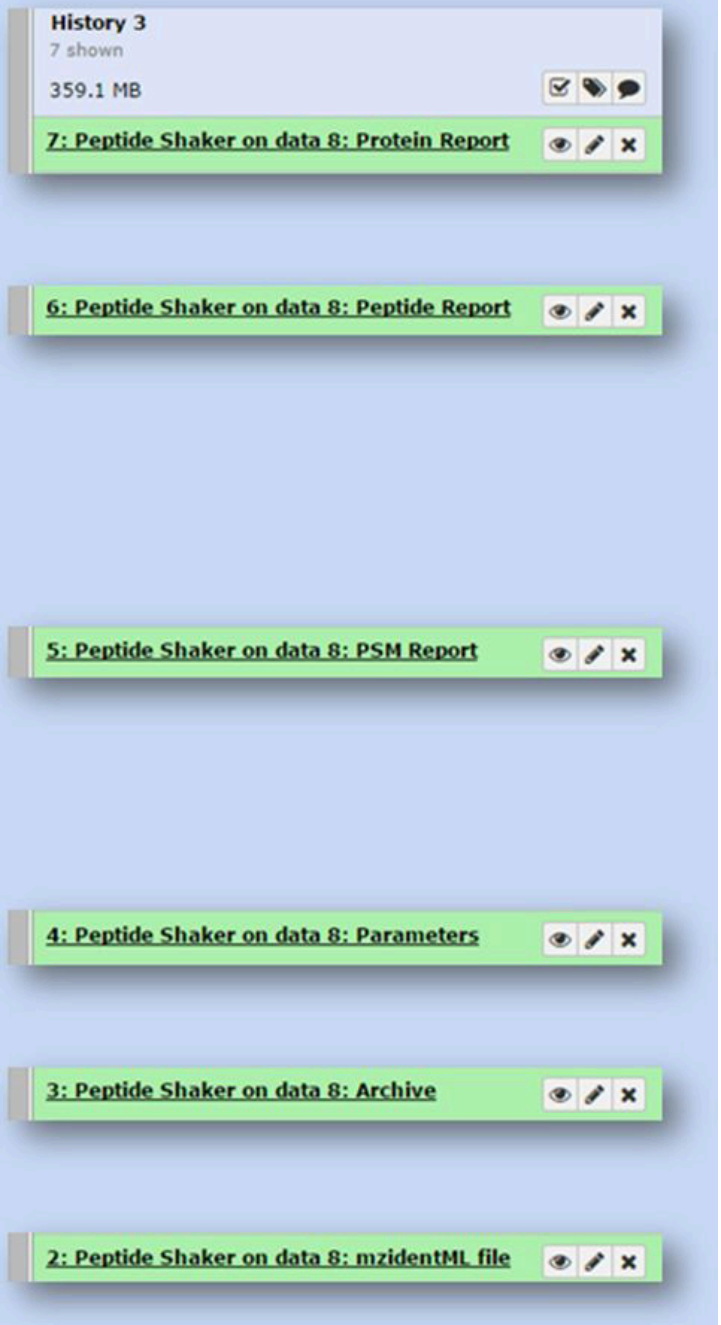
Peptide Report
- valid peptides
- potential novel proteoforms based on accession numbers
- sequences
- modifications and localization score
- confidence

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



5 BLAST-P Search

5.1 Proteoforms

What are Proteoforms?
Due to the genomic complexity and redundancy of proteins and the associated post-translational modifications that can occur during or after their expression, there can be a number of proteoforms associated with a protein. A proteoform is the product that results from a protein's specific genetic code and all the modifications molding it (e.g. post-translational modifications) or its transcription (e.g. alternatively spliced RNA and allelic variations). For more information about proteoforms please read manuscript by Smith and Kelleher (2013).

Why are they so important?
Proteoforms contribute to biological diversity. Because of chemical differences, proteoforms not only differ in structure, but in function as well. This leads to several different process modulations that affect cells differently, contributing to variation between and within individuals.

Identifying Peptides Corresponding to Novel Proteoforms
Proteoforms retain a lot of similarity with one another, which can make it hard to identify them from one another. Since the advent of proteomics, peptides corresponding to novel proteoforms are continually being identified after verification through BLAST analysis. Once validated, these proteoforms help in a more complete annotation of the genome and also identification of a role for such novel biomarkers in disease and physiological states such as cancer.

5.2 BLAST-P

BLAST (Basic Local Alignment Search Tool) is a web-based tool used to compare biological sequences. BLAST-P, matches protein sequences against a protein database. More specifically, it looks at the amino acid sequence of proteins and can detect and evaluate the amount of differences between say, an experimentally derived sequence and all known amino acid sequences from a database. It can then find the most similar sequences and allow for identification of known proteins or for identification of potential peptides associated with novel proteoforms..

a. At the top of the screen, click *Shared Data* then select the option: *Published Workflows*.
b. Click *Workflow 3: Workflow for History 3 to History 4* and choose *import*.
c. Select *start using this workflow*.
d. To run, click *Workflow 3: Workflow from History 3 to History 4* and choose *run*.
    i. **Change** your input file to *6: Peptide Shaker on data 8: Peptide Report*

ii. Notice the number of steps in the workflow (refer to Section 5.3 for more detail)

iii. Check *Send results to a new history* at the bottom and name it *History 4*

iv. Run workflow

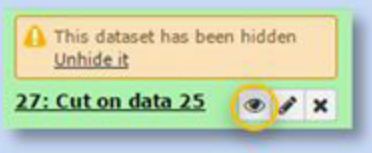
5.3 Data Manipulation via Workflow

Workflow 3: Workflow from History 3 to History 4 is a significantly more complex workflow than the preceding one, which involved only PeptideShaker. In this workflow, 31 steps are used to take the peptide report from PeptideShaker and manipulate it to put through BLAST-P analysis to verify novel proteoforms.

Workflow 3: Workflow from History 3 to History 4

- Step 1: Input dataset
- Step 5: Selects peptides with accession number from 3-frame translated cDNA library
- Step 9: Tabular-to-FASTA
- Steps 13/14: Splits FASTA file into large and short peptide sequences.
- Steps 15/16: Uses python regular expression to format derived sequence length information from step 10.
- Steps 17/18: BLAST-P analysis
- Steps 21-24: Brings up mismatched peptides.
- Steps 25-31: Produces peptide report

To view other steps in detail, search specific tools using the left panel. To view outputs and intermediate steps, use the right panel to load *History 4*, click *hidden* underneath the title to reveal all the steps, and clicking any of the eye icons.



6 Visualizing Proteomic Results

6.1 Why Visualize Your Results?

Importance of PSM Visualization

Scoring matrices and FDR thresholding of data acquired from search algorithms serves to reduce the number of many poor peptide-spectral matches (PSMs). However, some low quality identifications elude filtering and must be manually evaluated. PSM Visualization may reveal that a reported high-scoring spectrum is in fact a result of several unmatched ions. Validation of PSMs is often considered the final step before reporting protein identifications. Visualization may be forgone at the risk of misreporting identifications.

Visualizing Peptide Shaker Results: sqlite Database
The **mz to sqlite** Galaxy tool consolidates the information in the mzIdentML output dataset from a search algorithm like PeptideShaker along with the peaklist input datasets (e.g. mzml and MGF files) and the fasta SearchDB into a mz.sqlite dataset. This is a special SQlite database schema that provides a **PSM Viewer** Galaxy visualization plugin for interactively analyzing the data.

Visualization with Peptide-Spectrum-Match Evaluator
The Peptide-Spectrum-Match (PSM) evaluator tool is a unique visualization tool to GalaxyP. Using the experimental peak lists and corresponding analyzed database search report (peptide report or mzid) PSM Evaluator will render each peptide for visual validation. PSM Evaluator can visualize fragmentation ion series and precursor ions for use in validation. In addition to this PSM metrics can also be used to decide on which PSMs need to be visually validated before reporting as those corresponding from novel proteoforms.

6.2 Generating a sqlite Database

a. From *Shared Data* at the top of the screen, *import* *History 5* and *Workflow 4: History 4 to History 5*.
b. From the *workflows menu*, *run Workflow 4: History 4 to History 5* on History 4.
c. In the workflow run options, check that all inputs are correct and select the box *Send results to a new history named:* and enter *History 5* into the field provided.
d. Select *Run* and navigate to History 5 in the normal Galaxy view. **For the purposes of this tutorial, a completed History 5 may be imported from Published Histories.**

6.3 PSM Evaluator in GalaxyP

a. From *Shared Data* at the top of the screen, *import* History 5
b. *Click on mz_to_sqlite* dataset to expand, select *Visualize in PSM Viewer*.
c. Select Peptide View from PSM Table at the top of the screen.

d. To view the spectra of the novel peptides input the novel peptide sequences (LLSVGGLR, SPVLKPSR) into the filter peptides by sequence(s) field and select the sequence to generate the spectral data.

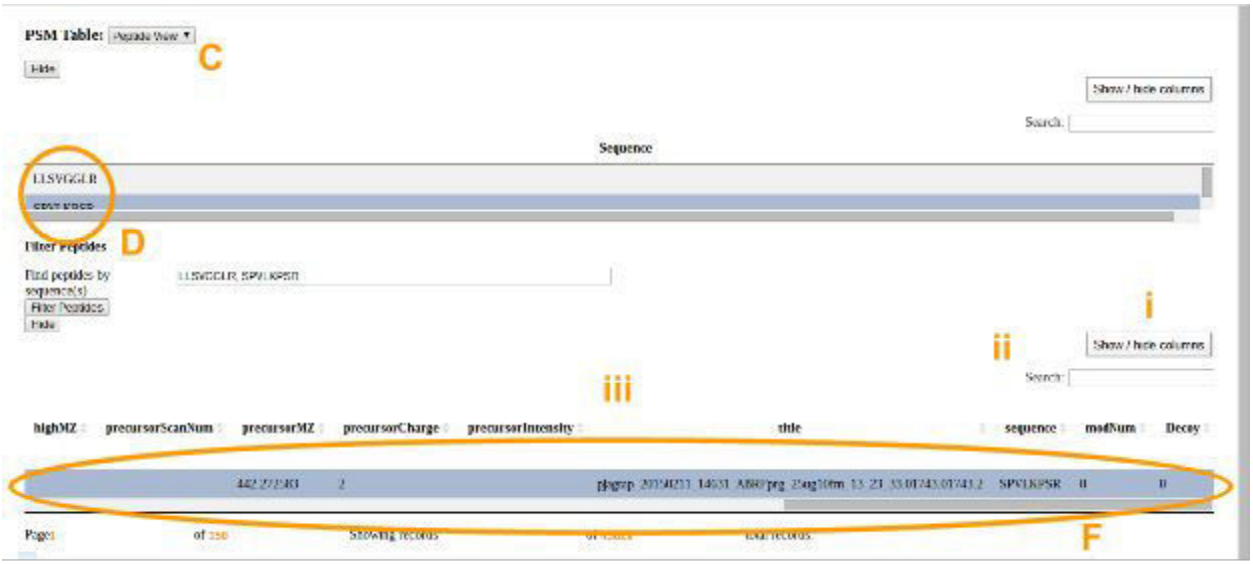
e. Within the viewer users can utilize numerous organization tools to sort through data of each peptide.

   i. Users may select which columns they prefer to see in their PSM Tables

   ii. Users may search the table for a specific piece of data.

   iii. Users may sort data in each column and organize the order in which columns appear.

f. To view the spectra of LLSVGGLR select anywhere with the peptide spectral data table.

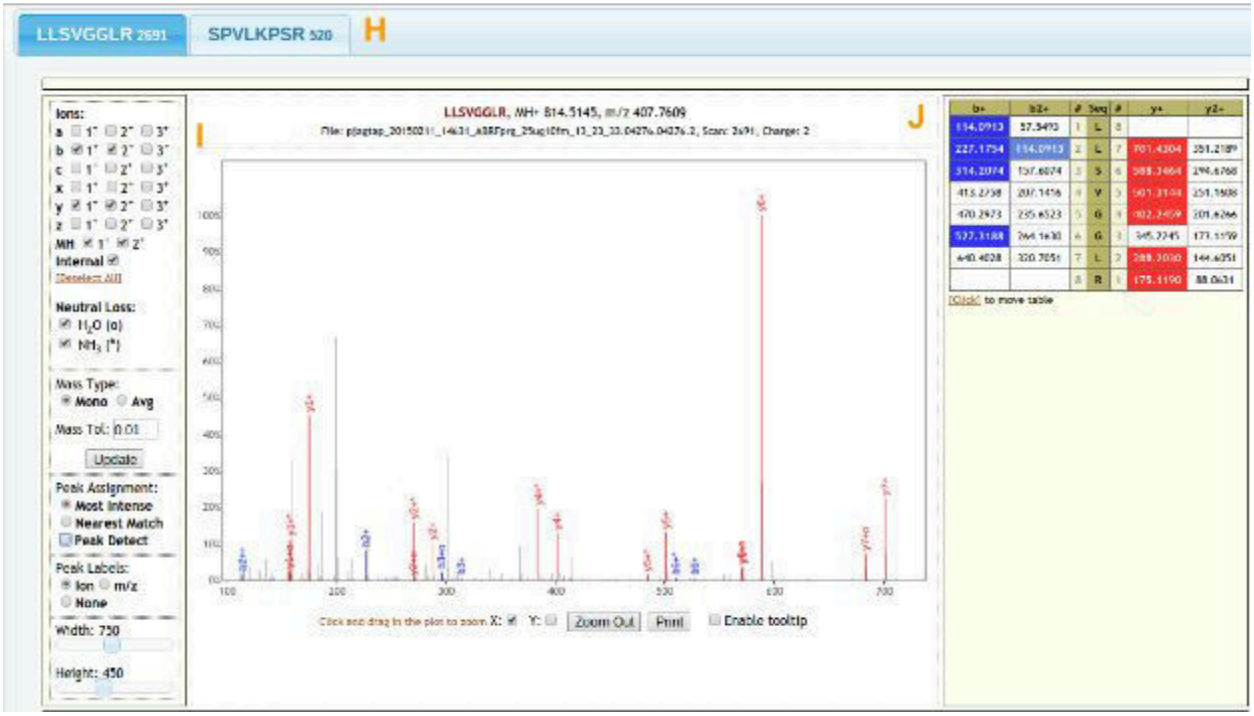


g. The spectra will appear near the bottom of the page, along with a matrix of all fragmentation ions.

h. Users can switch between spectra by selecting the sequence at the top of the visualizer.

i. To visualize fragment ion series select the desired series under the ions interface. An accepted approach to validate spectra is to visualize all b^{1+}, b^{2+}, y^{1+}, y^{2+}, MH^{1+}, MH^{2+}, Internal Ions, all Neutral Losses. Visualize at a Mass Tolerance of 0.01 Daltons.

j. Users can validate the table visually from the spectral graph or from fragment ion table.



7 Proteogenomics

7.1 Mapping Novel Proteoforms to the Genome

Peptides to Genome
Considerable information can be obtained from the identity of a peptides corresponding to novel proteoforms. Beyond the identity of the aberrant proteins, the localization of each peptide can reveal intriguing genomic architecture. In essence, proteogenomics involves the mapping of an experimental proteome to an established genome. Clustering of proteoforms in a particular genomic region may implicate a point of interest for further research. For an excellent review on proteogenomics please read review by Nesvizhskii et al (2014).

Peptides to GFF
A **GGF** file contains the mapping of features, e.g. genes and exons, to a reference genome. Most genome browsers support the display of GFF files.

The **Peptides to GFF** Galaxy tool generates a GFF3 file to map peptides to a reference genome. Most search algorithms have an output that associate peptides to proteins from the search database. The tool will first map each peptide to the associated protein sequence, then map the protein sequence to the reference genome. The final peptide mapping may have multiple lines in the GFF3 if the peptide sequence mapping is split across exon boundaries.

Ideally the construction of a search database will have associated files to aid the mapping of the protein sequence to the genomic sequence. For example, one might perform a 3-frame translation of EnsEMBL cDNA sequences to search for frameshift peptides. The associated EnsEMBL GTF file can then be used to map the cDNA sequence to genome accounting for the splice junctions.

Navigating to IGV Browser from GalaxyP

IGV Browser is a stand-alone software that may be conveniently accessed from the Galaxy user interface. In order to install a local IGV Browser on your machine, go to https://www.broadinstitute.org/igv/home and register to download the appropriate version from https://www.broadinstitute.org/software/igv/?q=download

Click on "Launch". Save the JNLP file and click on the downloaded igv_hm JNLP file and click "Run".

7.2 Generating a GFF file

a. From *Shared Data* at the top of the screen, *import Input for History 6* and *Workflow 5: Workflow to History 6* from Published Histories and Workflows, respectively.
b. From the workflow menu, *Run* Workflow 5 on Inputs for History 6.
c. Select the appropriate inputs for Workflow 5 from Inputs to History 6.
d. Check to *send the results to a new history named:* History 6 and *Run* the workflow. **For the purposes of this tutorial, a completed History 6 may be imported from Published Histories.**

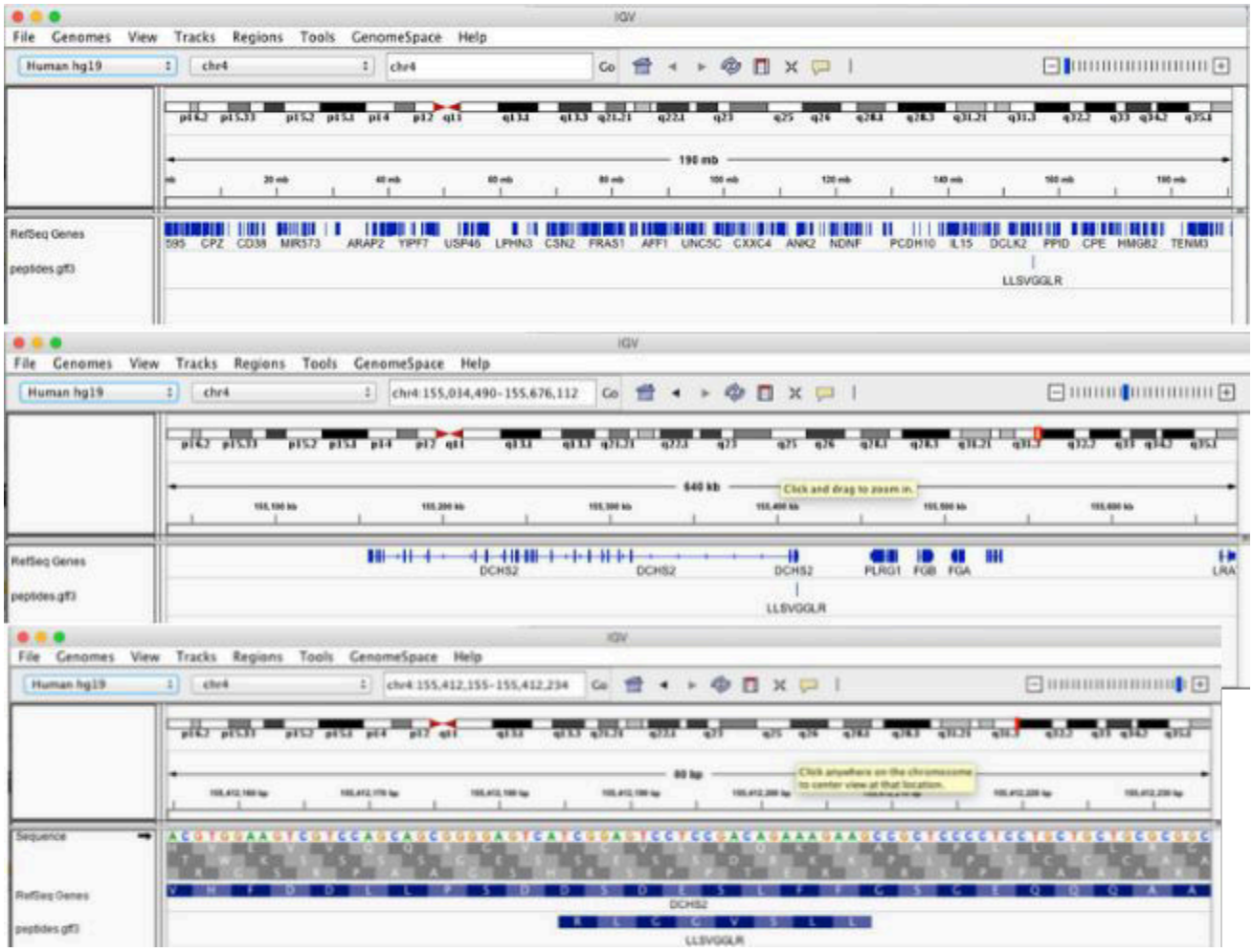
7.3 From GFF to IGV Browser

***** A local installation of IGV Browser is required for the next portion of the tutorial.**

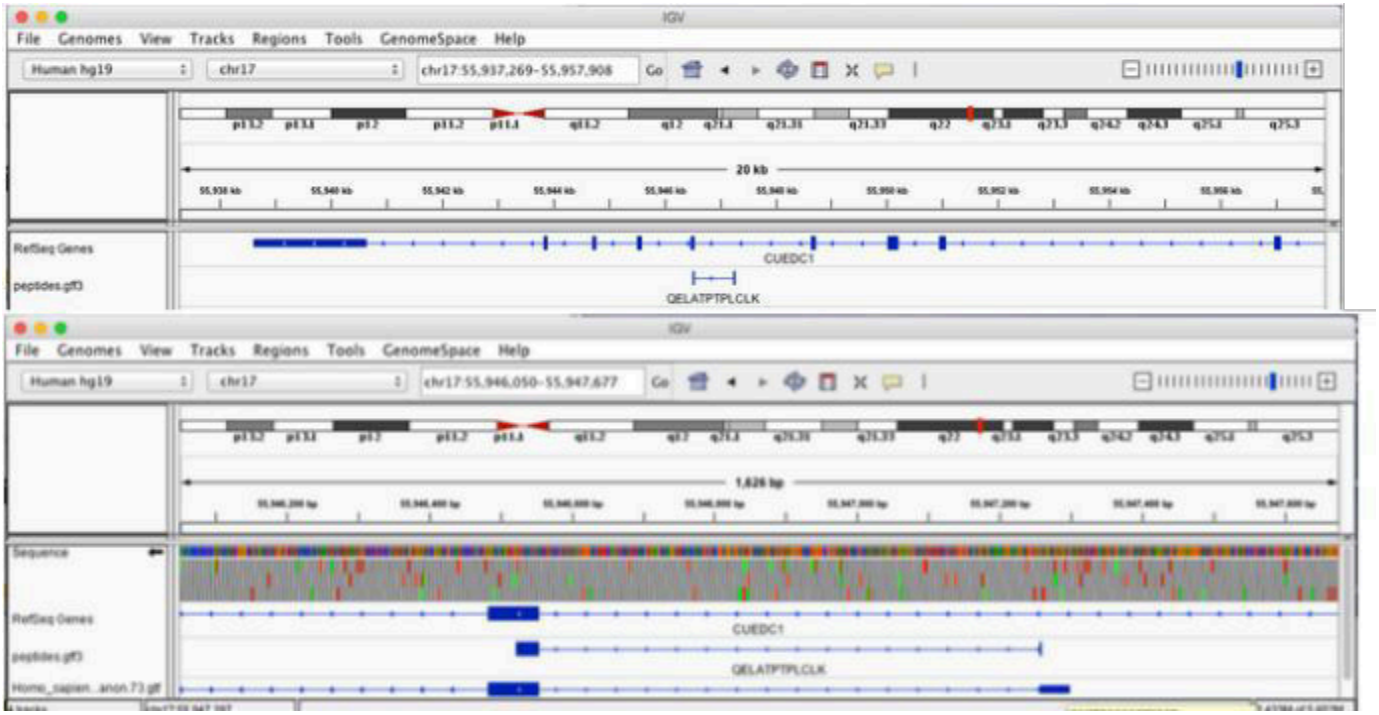
(http://www.broadinstitute.org/software/igv/download)

a. From *Shared Data* at the top of the screen, *import History 6* from Published Histories.

b. To navigate to a local installation of IGV Browser from GalaxyP, select *display with IGV: local* on the LLSVGGLR GFF file.
c. To visualize novel peptide LLSVGGLR on the genome change the view to focus on *Chromosome 4* (The precise genomic region is given within the GFF file).
d. *Drag out a region* in the IGV header to zoom in on the novel peptide.
e. When zoomed in far enough the 3-frame translation of the genomic region surrounding the peptide.
f. If the observed sequence does not match any of observed 3-frame translation, the reverse 3-frame translation may be visualized by selecting to reverse the sequence.
g. To visualize the other novel peptide, select *display with IGV: local* on the SPVLKPSR GFF file and change the view to Chromosome 17.
h. Drag out a region to zoom in on the novel peptide. Note that the peptide does not match the annotated RefSeq exons of this region.
i. Add the EnSEMBL GTF file from History 6 as a track on the IGV browser, using the *display with IGV: local*.
j. The novel peptide sequence can now be compared to the latest EnSEMBL track for the genomic region.



This image shows 3 successive views of IGV. 1) The entire chromosome 4 showing the location of the peptides on the chromosome. 2) Zoomed in to show the structure of a gene in the peptide region. 3) Zoomed in to display the peptide sequence and the 3-frame translation of the genomic strand. NOTE: You may need to click the IGV Sequence arrow to display the reverse strand.



This image shows a peptide that was mapped across a splice junction to 2 exons.

8 Running Entire Proteogenomics Workflow

The proteogenomics workflow can be run as a single, complete workflow.

A. Select *History 8A: Input for Entire Workflow* from the list of published histories.

B. Select *Import history* to add the selected history to your user histories.

C. On the confirmation screen select *start using this history* to navigate to this history in the Galaxy view.

D. At the top of the screen select *Shared Data* then migrate to *Published Workflows*.

E. Select *Workflow 8: Entire Proteogenomics Workflow* from the list of published workflows and choose *import* to copy the workflow into your user workflows.

F. On the confirmation screen, select *start using this workflow* to navigate to your user workflows.

G. In the workflows menu select *Run* from the drop down menu.

H. Appropriately assign each input database from History 8A to the corresponding input or the workflow.

I. Run the workflow.

J. The workflow output would be like *'Entire History'* which is stored in published histories.