

**The Galaxy Platform for Multi-Omic Data Analysis and Informatics**  
**ABRF 2016 Workshop**  
Saturday, February 20, 2016

Organizers:  
Pratik Jagtap  
Tim Griffin  
University of Minnesota



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Workshop objectives

- 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 multi-omic analysis (RNA-seq and proteomic integrative analysis)
- Lay the foundation for attendees to implement Galaxy at their own facility or institution



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Workshop instructors and acknowledgements

- **Instructors**
  - Dave Clements – core Galaxy team (Johns Hopkins University) 
  - Candace Guerrero
  - Getiria Onsongo
  - Pratik Jagtap
- **Support**
  - Tom McGowan, James Johnson (JJ), Ben Lynch (University of Minnesota)
  - Karen Reddy, Mo Heydarian (Johns Hopkins University)
- **Sponsors**



## Topic of workshop: multi-omics

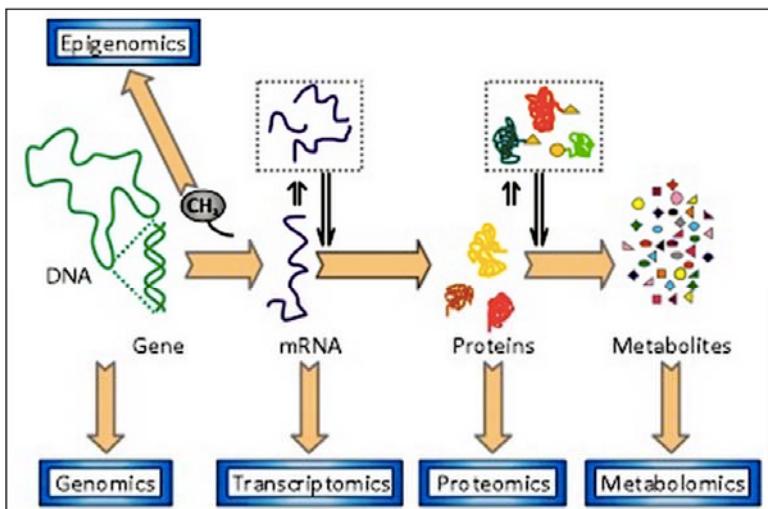


Image Source: Goodacre, J. Exp. Bot 2005.

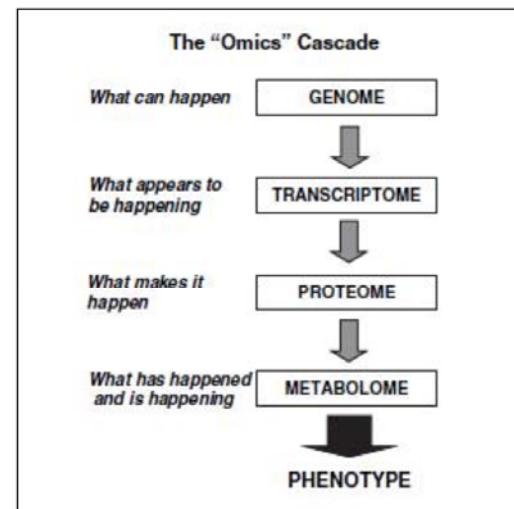


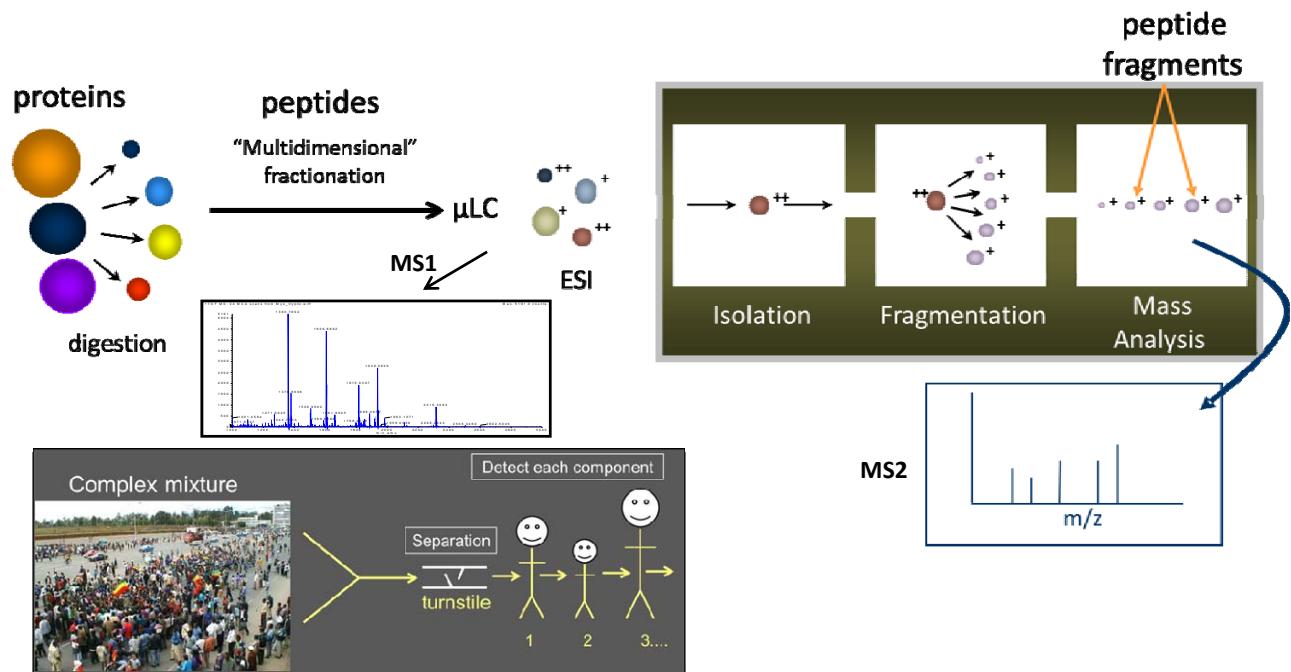
Image Source:  
<http://fluorous.com/images/omics.JPG>



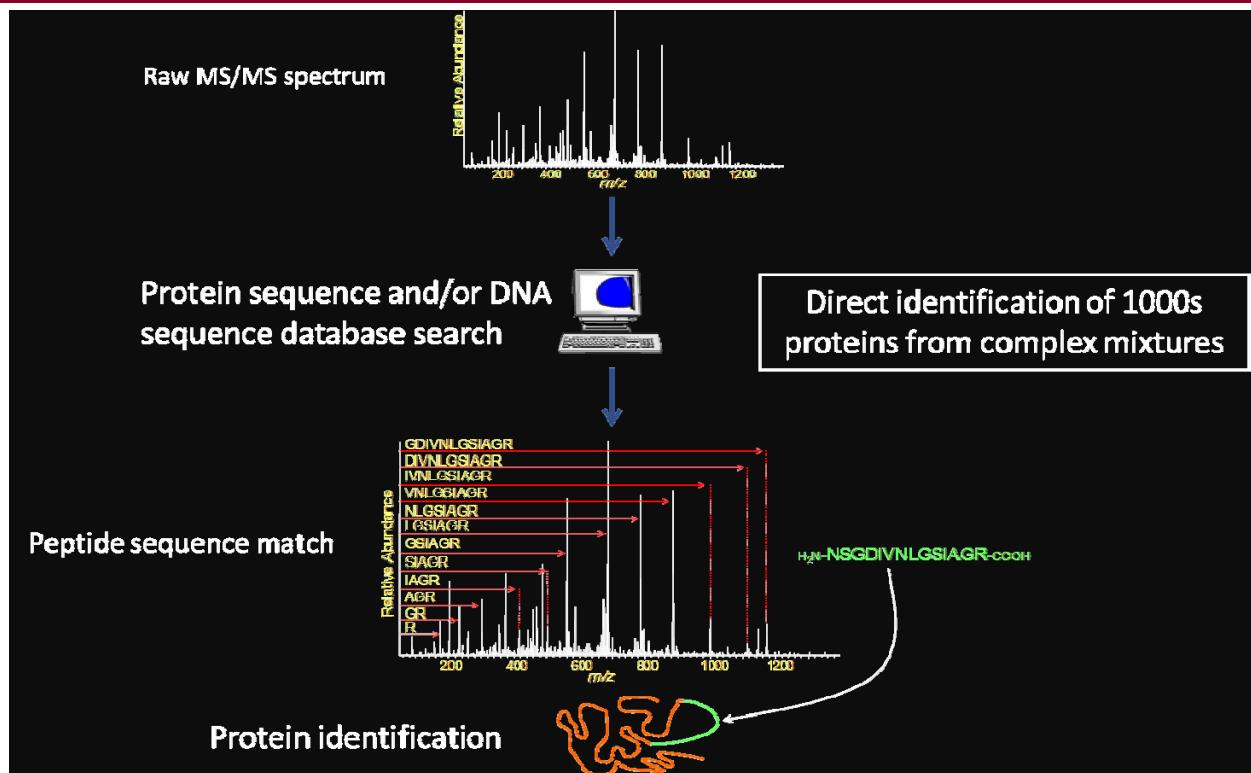
UNIVERSITY OF MINNESOTA  
Driven to Discover™

## Technologies: Proteomics

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

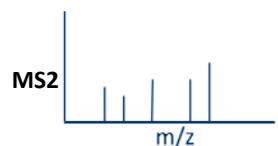


## From sticks on a graph to protein identities



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Inferring protein identity from peptide sequence matches



CAQCHTVEK

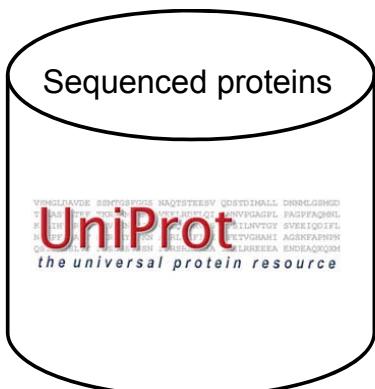
Cytochrome C

$\text{NH}_2$ GDVEKGKKIFVQK**CAQCHTVEK**GGKHK**TGPNLHGL**  
**FGRKTGQAPGFTYTDANKNGITWKEETLMEYLENPK**  
KYIPGTMIFAGIKKKTER**EDLIAYLK**KATNE<sub>COOH</sub>



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

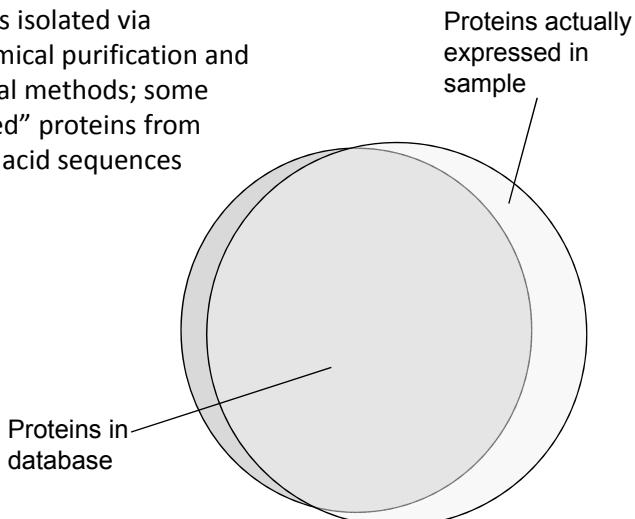
## MS-based proteomics only as good as the database used...



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

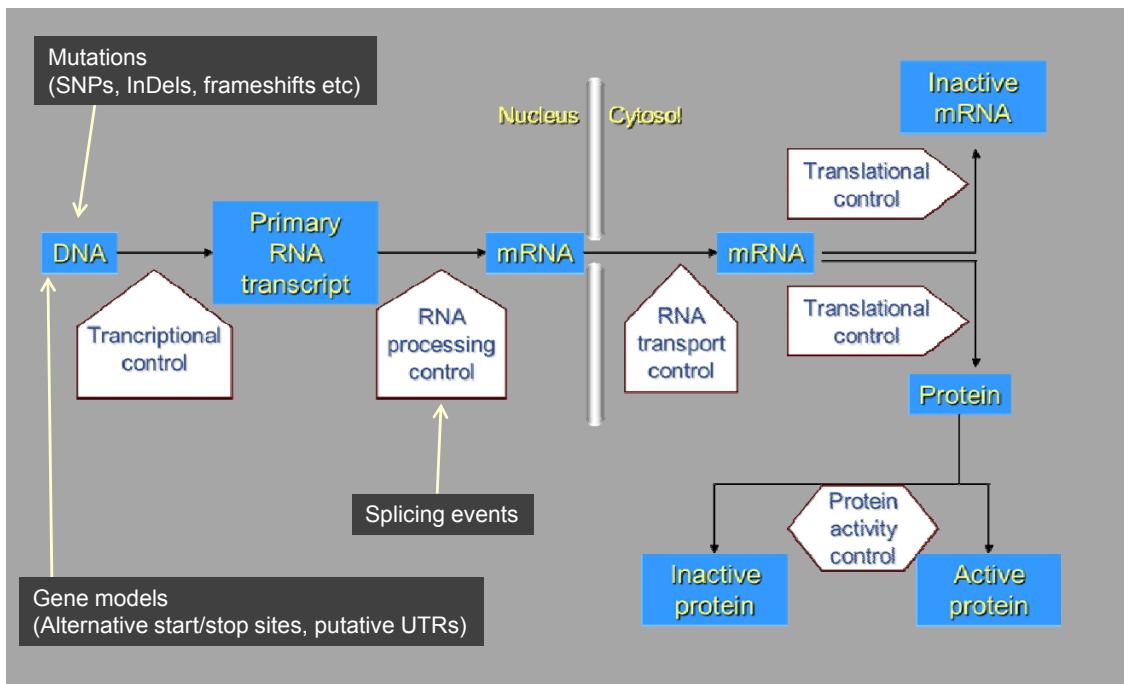
-- Sir Isaac Newton

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



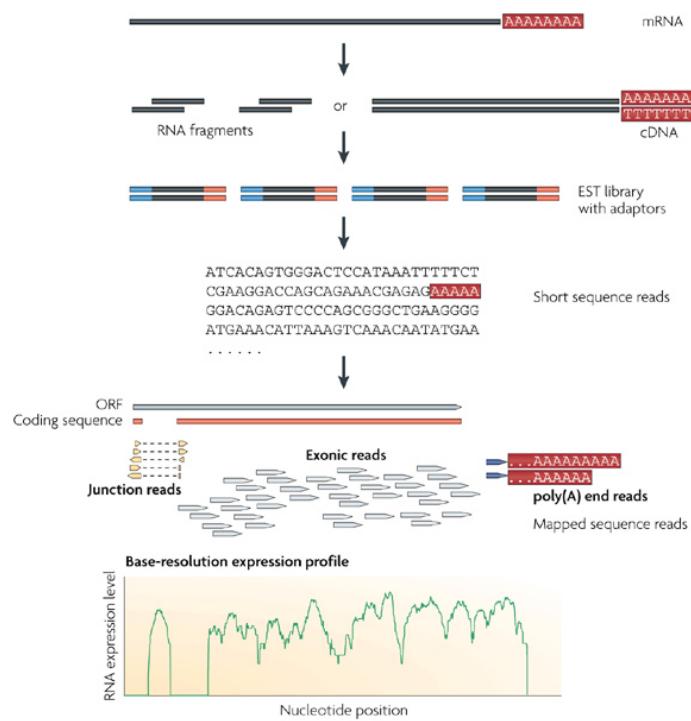
UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Sources of un-expected protein sequences?



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Technologies: RNA-seq

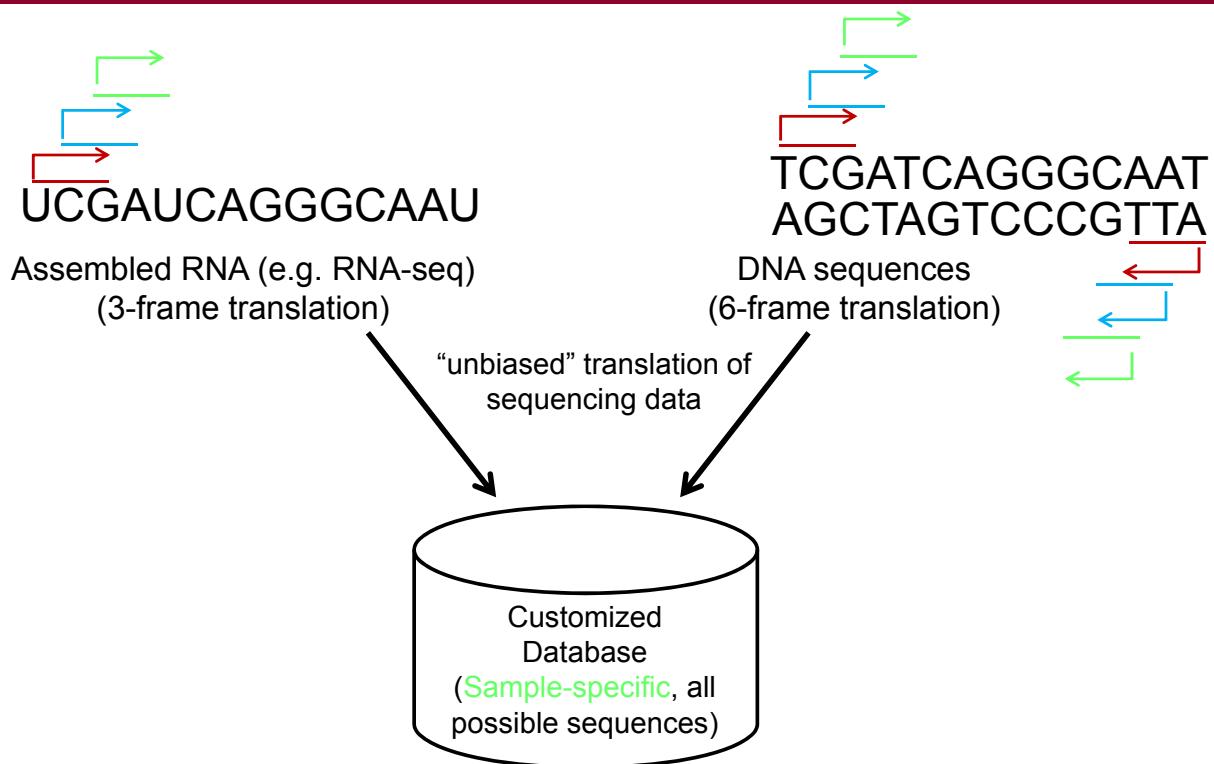


Nature Reviews | Genetics 10, 57-63



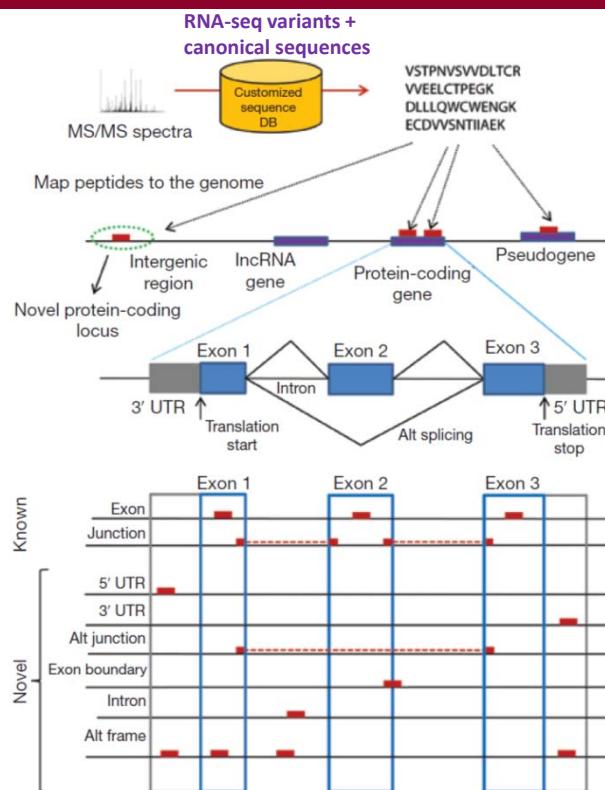
UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## A multi-omic example: Integrating Next-Gen sequencing and proteomics



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## A multi-omic example: Integrating Next-Gen sequencing and proteomics



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

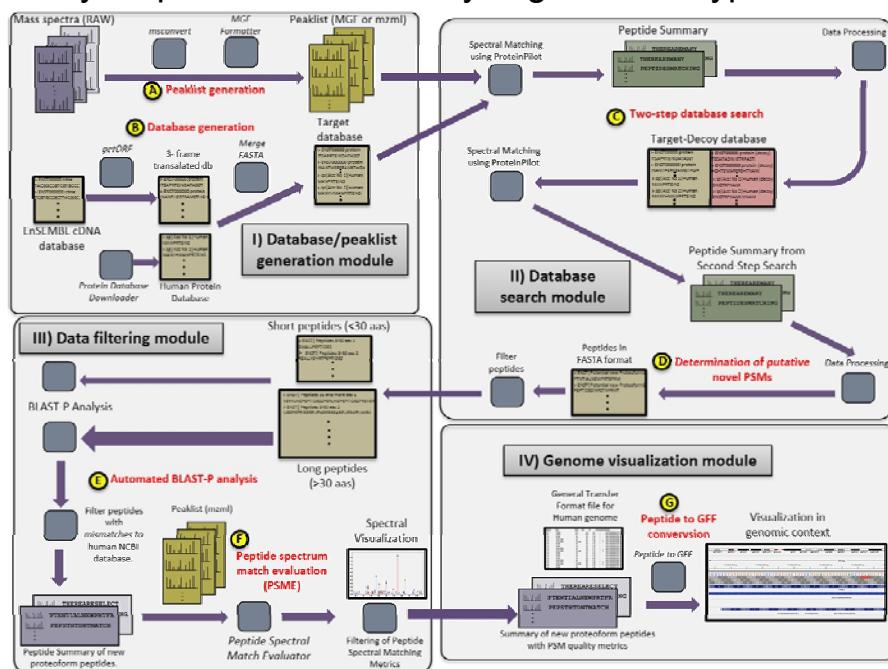
Nesvizhskii NATURE METHODS | VOL.11 NO.11 | NOVEMBER 2014 |



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Driven by bioinformatics

- Many step workflow for analyzing different types of data



J. Proteome Res. 2014, 13, 5898–5908



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Solving the informatics bottleneck in multi-omics: Galaxy



- A web-based, community developed bioinformatics framework/platform/workbench
- Originally designed to address issues in genomic informatics including:
  - Software accessibility and usability
  - Analytical transparency
  - Reproducibility
  - Scalability
  - Share-ability
- **In a nutshell:** Galaxy provides an open framework into which disparate software programs can be deployed, integrated into novel workflows for typical to advanced applications, which can be shared in their entirety with other users

Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010, **11**: R86.



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# The Galaxy interface

Tools

Main viewing window  
(workflow development, results visualization etc)

History

The screenshot shows the GalaxyP interface. On the left is the 'Tools' panel, which includes a search bar and sections for CORE TOOLS, PROTEOMICS, and NGS: QC and manipulation. The central area is the 'Main viewing window' showing a 'Welcome to GalaxyP' page with text about the platform and a 'Updates' section listing recent activity. The right side is the 'History' panel, displaying a list of workflow runs with details like name, date, and size. Red arrows point from the labels to their respective parts of the interface.

Galaxy / GalaxyP

Analyze Data Workflow Shared Data Visualization Help User

Using 35.5 GB

Tools

search tools

CORE TOOLS

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Statistics
- Graph/Display Data
- FASTA manipulation

PROTEOMICS

- MS Data Conversion
- Sequence Database Tools
- NGS: QC and manipulation
- Protein/Peptide Search
- Algorithms
- Data Conversion Tools
- Visualizers
- Quantification

Welcome to GalaxyP

GalaxyP

Updates

February 13, 2015  
All Galaxy Tools running nominally

February 12, 2015  
All Galaxy Tools running nominally

History

search datasets

imported: Test1 Workflow for human proteogenomics : Peptides to PSME - PeptideShaker Output

18 shown, 15 hidden

4.0 MB

33: BLAST-P Filtered Peptide Report

32: BLAST-P Filtered Peptides

16: Regex Find And Replace on data 14

15: Regex Find And Replace on data 13

14: Filter sequences by length on data 12

13: Filter sequences by length on data 12

12: Tabular-to-FAST

UNIVERSITY OF MINNESOTA

Driven to Discover<sup>SM</sup>

## Enabling integration of tools across the 'omic domains

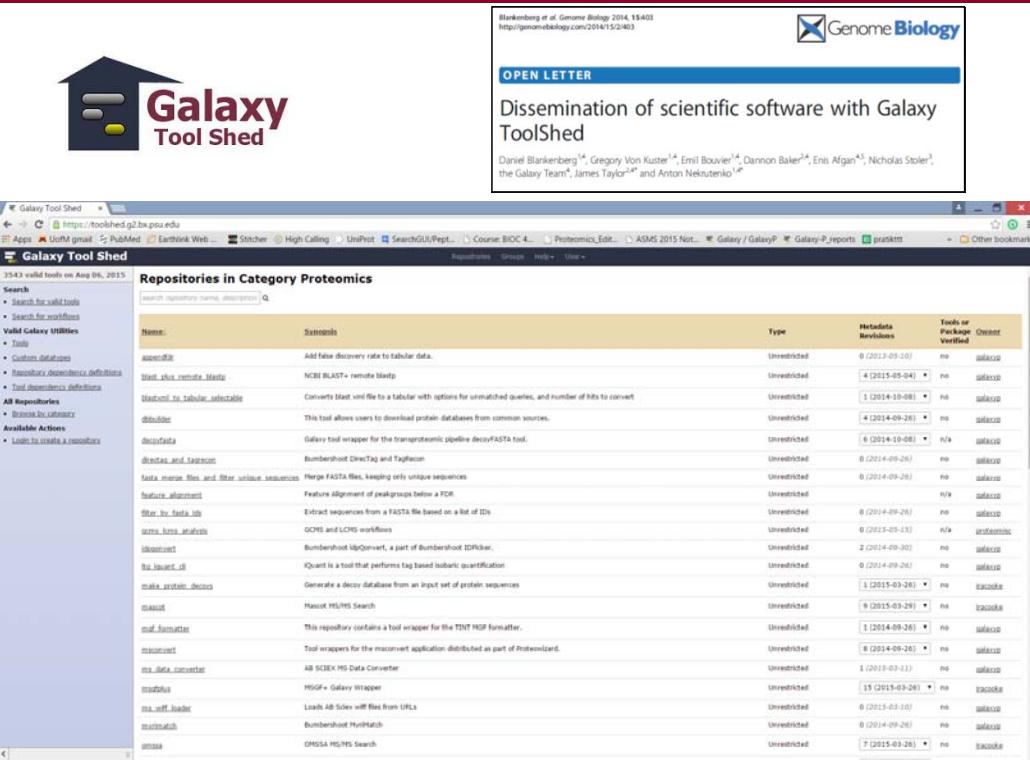
The diagram illustrates the integration of tools across 'omic domains through three interconnected publications:

- Top Publication:** "Colib'read on galaxy: a tools suite dedicated to biological information extraction from raw NGS reads" (GigaScience, TECHNICAL NOTE, Open Access). This publication is cited in the second publication.
- Middle Publication:** "Galaxy Integrated Omics: Web-based Standards-Compliant Workflows for Proteomics Informed by Transcriptomics" (Molecular & Cellular Proteomics 14.11, Jun Fan et al., Open Access). This publication is cited in the third publication.
- Bottom Publication:** "Web-based visual analysis for high-throughput genomics" (BMC Genomics 2013, Jeremy Goeckl et al., SOFTWARE, Open Access). This publication features a screenshot of the GalaxyP interface.

**GalaxyP Logo:** A red rounded square icon containing the text "GalaxyP" above two vertical bars.

**University of Minnesota Logo:** The iconic gold "M" logo next to the text "UNIVERSITY OF MINNESOTA" and "Driven to Discover™".

## Built for the community: sharing software tools



The screenshot shows the Galaxy ToolShed interface. At the top left is the Galaxy ToolShed logo. To its right is a window titled "OPEN LETTER" from "Genome Biology" with the title "Dissemination of scientific software with Galaxy ToolShed". Below this is a browser window showing a list of repositories under the heading "Repositories in Category Proteomics". The list includes various tools like "Samtools", "NCBI BLAST+", "decovFasta", "Bumbleshot DirectTag and TagScanner", and "msa2t2". Each entry has columns for Name, Type, Metadata Revisions, Tools or Package Verified, and Owner.

Name	Type	Metadata Revisions	Tools or Package Verified	Owner
Samtools	Unrestricted	0 (2013-09-10)	n/a	galaxy
aspcBLAST	Unrestricted	4 (2015-05-04)	• n/a	galaxy
blast_plus_remote_Hanta	Unrestricted	4 (2014-10-08)	• n/a	galaxy
blastxml_to_tablear_selectable	Unrestricted	1 (2014-09-28)	• n/a	galaxy
dbidbaser	Unrestricted	4 (2014-09-28)	• n/a	galaxy
decovFasta	Unrestricted	6 (2014-10-08)	• n/a	galaxy
directtag_and_tagscanner	Unrestricted	0 (2014-09-28)	n/a	galaxy
fasta_merger_fix_and_filter_unique_sequences	Unrestricted	0 (2014-09-28)	n/a	galaxy
feature_alignment	Unrestricted	n/a	n/a	galaxy
filter_by_fasta_ids	Unrestricted	0 (2014-09-28)	n/a	galaxy
gcnv_kmer_analyst	Unrestricted	0 (2013-05-13)	n/a	galaxy
isobamquant	Unrestricted	2 (2014-09-30)	n/a	galaxy
is_quant_if	Unrestricted	0 (2014-09-28)	n/a	galaxy
make_protein_database	Unrestricted	1 (2015-03-26)	• n/a	galaxy
msa2t2	Unrestricted	9 (2015-03-29)	• n/a	galaxy
msaf_formatter	Unrestricted	1 (2014-09-28)	• n/a	galaxy
msaconvert	Unrestricted	8 (2014-09-28)	• n/a	galaxy
msi_data_converter	Unrestricted	1 (2013-03-11)	n/a	galaxy
msdab2da	Unrestricted	15 (2015-03-26)	• n/a	galaxy
msa2ff_loader	Unrestricted	0 (2015-03-30)	n/a	galaxy
mscblast	Unrestricted	0 (2014-09-28)	n/a	galaxy
omssa	Unrestricted	7 (2015-03-26)	• n/a	galaxy

**GalaxyP**  UNIVERSITY OF MINNESOTA  Driven to Discover<sup>SM</sup>

## Built for the community: sharing complete workflows

**Galaxy-P 101: Building up and using a proteomics workflow**

In this very basic example we will introduce you to basics of Galaxy-P:

What are we trying to do?

We are trying to set up a workflow to analyse the raw experimental data acquired on an iTRAQ instrument against a database using a search algorithm named trypsin. This is a relatively simple task involving a few steps:

Organizing your display.

Now that you are in Galaxy-P, we assume you have a solid basis of knowledge and have knowledge about accessing Galaxy-P.

Q-D. Getting your display sorted out:

How often you sort your browser windows, (or if you already have 10 in this page), to open the other, right click and choose "Open in a New Window" (or something similar depending on your browser).

Open Link in New Tab  
Download Link to File...  
Download Link to File As...  
Add Link to Bookmarks...  
Copy Link

How organize your windows or something like this (depending on the size of your monitor you may or may not be able to implement these links):

Galaxy-P URLs: Building up and using a proteomics workflow

**HISTORY:** <https://galaxyp.msi.umn.edu/u/pjagtap/h/itraq-search-yang-2-xtandem-scaffold>

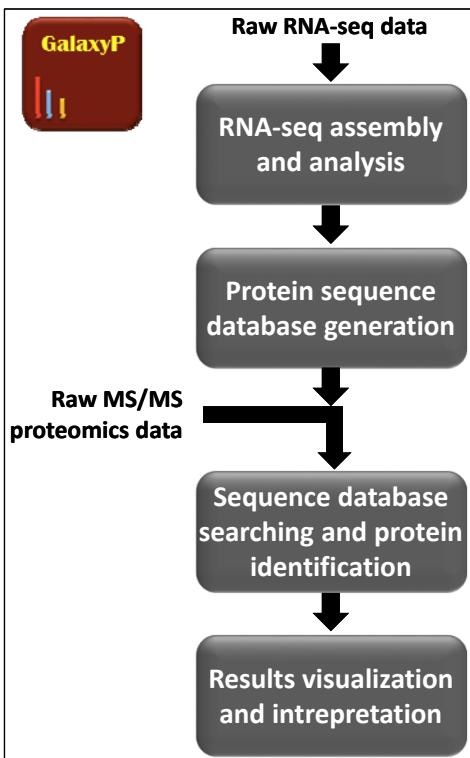
**WORKFLOW:** <https://galaxyp.msi.umn.edu/u/pjagtap/w/workflow-for-4-plex-itraq-xtandem-search-scaffold-processing>

Galaxy-Workflow-Workflow\_for\_4-plex\_iTRAQ\_X\_tandem\_Search\_Scaffold\_Processing.ga



UNIVERSITY OF MINNESOTA  
Driven to Discover™

# Structure of workshop



## Session 1

- RNA-seq and quantitative MS-proteomics data:

**Proteomics & Bioinformatics** Open Access  
Heydarian et al., J Proteomics Bioinform 2014, 7:2  
<http://dx.doi.org/10.4173/jpb.1000302>

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

Mohammad Heydarian<sup>1</sup>, Teresa Romeo Luperchio<sup>1</sup>, Jevon Cutler<sup>1,2</sup>, Christopher J. Mitchell<sup>1,2</sup>, Min-Sik Kim<sup>1</sup>, Akhilesh Pandey<sup>1</sup>, Barbara Sollner-Webb<sup>1</sup> and Karen Reddy<sup>1,2\*</sup>

<sup>1</sup>Johns Hopkins University, Department of Biological Chemistry, 725 North Wolfe Street, Baltimore, USA  
<sup>2</sup>Johns Hopkins University, Center for Epigenetics, 855 North Wolfe Street, Baltimore, USA  
Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, 733 North, Broadway Avenue, Baltimore, USA

## Session 3

- Documentation to complement hands-on instruction

## Session 2

## Session 4



UNIVERSITY OF MINNESOTA  
Driven to Discover™

## Agenda and schedule

### **8:00am-8:30am**

Introduction to Galaxy Platform and multi-omic studies. (Tim Griffin)

### **8:30am-10:00am**

Galaxy Tutorial – Basics of Data Analysis using Galaxy platform. (Dave Clements)

### **10:30am – 12:00pm**

Hands-on session for proteomics data analysis using Galaxy. (Candace Guerrero)

### **1:00pm – 1:30pm**

Introduction to multi-omics: proteogenomics and metaproteomics (Pratik Jagtap)

### **1:30pm-3:00pm**

Proteogenomic analysis using Galaxy framework I: RNASeq data and sequence database generation. (Getiria Onsongo)

### **3:30pm – 4:30pm**

Proteogenomic analysis using Galaxy framework II: Database search, filtering and visualization. (Pratik Jagtap)



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

## Instructions for accessing cloud instance

- Wifi
  - Connect to “Staff”
  - Password: simple@1950
- Cloud instance
  - You will be given an IP address to access your own instance
  - Username: [user@abrf.org](mailto:user@abrf.org)
  - Password: password
- Ensure that Java 7 or later version is installed on your computer



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# Galaxy Tutorial - Basics of Data Analysis using Galaxy platform, Condensed

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

Dave Clements, Galaxy Project, Johns Hopkins University  
8:30-10:00am, Saturday February 20, 2016

## ABRF 2016

[Explore Galaxy](#)

[Left/Tools panel](#)

[Right/History panel](#)

[Center panel](#)

[Create an account](#)

[Our goal: Identify transcripts using RNA-Seq in pre-pro-B and pro-B mouse cells](#)

[The data](#)

[Get the data / Data libraries](#)

[FASTQ datasets](#)

[FASTQ Format](#)

[Check FASTQ dataset quality](#)

[Run our first tool: FastQC](#)

[Per base sequence quality \(boxplot\)](#)

[Per tile sequence quality](#)

[Per sequence quality scores](#)

[Per base sequence content](#)

[Adapter Content](#)

[And the rest](#)

[Rerun FastQC on pro-B dataset](#)

[Improve the quality: Trimmomatic](#)

[Processing Multiple datasets with one submission](#)

[Select Trimmomatic Filters](#)

[Trimmomatic Results](#)

[Using Galaxy Scratchpad to view multiple datasets](#)

[Map the reads](#)

[HISAT2](#)

[HISAT2 Results](#)

[Updating dataset and history metadata: Renaming](#)

[But what do the HISAT2 mappings look like?](#)

[Transcript Prediction with StringTie](#)

[And there is more!](#)

[Workflows: repeating an analysis](#)

[Working with multiple histories](#)

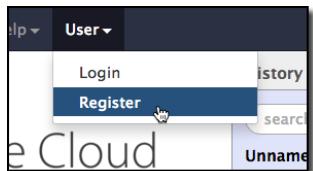
[Sharing and publishing](#)

[Galaxy pages](#)

A longer version of the same material is available online at <http://bit.ly/gxyabrf16intro-long>

This handout is available online at <http://bit.ly/gxyabrf16intro>

## Create an account



To create an account, click on the **User** pulldown and select **Register**.

Use an email address you can remember, **and a low security password**. Note that these servers use HTTP, not HTTPS, and therefore your password is going out on the net unencrypted. (Galaxy can use HTTPS and most servers do, but the default cloud installation uses HTTP.)

A screenshot of a registration form titled 'Create account'. It includes fields for 'Email address' (clements@galaxyproject.org), 'Password' (a masked password), 'Confirm password' (a masked password), and 'Public name' (clements). There is also a note about public names being identifiers for sharing. A 'Submit' button is at the bottom.

Reload the page by *clicking Analyze Data* or **Return to the home page**.

## Our goal: Identify transcripts using RNA-Seq in pre-pro-B and pro-B mouse cells

We are going to use RNA-seq data to detect transcripts, possibly novel transcripts, in pre-pro-B and pro-B mouse cells.

This data comes from

Heydarian *et al.*, (2014) Prediction of Gene Activity in Early B Cell Development Based on an Integrative Multi-Omics Analysis. *J Proteomics Bioinform* 7: 050-063.  
doi:10.4172/jpb.1000302

The paper uses RNA-Seq, ChIP-Seq, GRO-Seq, and iTRAQ techniques to “demonstrate that active chromatin modifications at promoters are good indicators of transcription and steady state mRNA levels.”

A screenshot of a journal article abstract. The title is 'Prediction of Gene Activity in Early B Cell Development Based on an Integrative Multi-Omics Analysis'. It's a Research Article published in 2014, Volume 7, page 050-063. The authors are Mohammad Heydarian, Teresa Romeo Luperchio, Jevon Cutler, Christopher J. Mitchell, Min-Sik Kim, Akhilesh Pandey, Barbara Sollner-Webb, and Karen Reddy. The journal is 'Proteomics &amp; Bioinformatics'. The abstract discusses the use of ChIP-seq, RNA-seq, and proteome abundance measurements to predict gene activity. It highlights that active chromatin modifications at promoters are good indicators of transcription and steady state mRNA levels. The paper also notes that many genes whose promoters have non-differential but active chromatin modifications also displayed changes in abundance of their cognate proteins, which was uncoupled from chromatin status. A specific protein, 2410004B18Rik, was regulated by a post-transcriptional mechanism mediated by a micro-RNA.

## The data

We have two RNA-Seq datasets, one from mouse pre-pro-B cells, and one from mouse pro-B cells. Each dataset consists of ~90 million single-end reads. Sequencing was done on an Illumina HiSeq 2000 and all reads are 97bp long.

We are going to use striped down datasets consisting of reads that tend to map to the particular regions of the genome we'll focus on today, plus some additional reads, just to make things interesting.

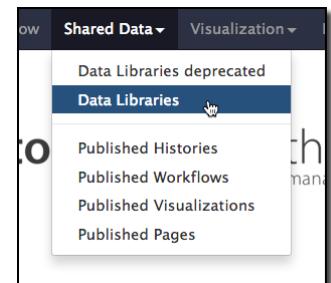
*Note that while reducing the datasets will make tool execution very fast, it can be a dangerous cheat. Many NGS tools are built on statistical models that assume they are being given data for the entire genome. With our data that is decidedly not the case. Downsampling your data is still a useful technique for early experimentation when building your analysis, but results and assumptions should be checked when scaling up your analysis to complete datasets.*

## Get the data / Data libraries

Our reduced datasets are available in a *data library* on your server. Click on the **Shared Data** pull-down, and then select **Data Libraries**.

Then click on the **RNA-Seq Transcript Prediction** library and select all 3 datasets in this library and then click on **to History** to import them into your current history.

This will ask which history to import these datasets into. So far, you only have one ("Unnamed history"). Click **Import** to add them to that history.



A screenshot of the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Import selected datasets into history' (which has a red arrow pointing to it), 'Help', 'User', and 'Using 0 bytes'. Below this is a 'DATA LIBRARIES' section with a table of three items. The table columns are 'name', 'data type', 'size', and 'time updated (UTC)'. The items listed are: 'GTF Annotation chrX/12 regions' (gtf, 55.8 KB, 2016-02-12 12:59 AM), 'pre-pro-B chrX/12 regions FASTQ raw' (fastqsanger, 4.4 MB, 2016-02-12 12:59 AM), and 'pro-B chrX/12 regions FASTQ raw' (fastqsanger, 7.9 MB, 2016-02-12 12:59 AM). A 'to History' button is visible in the toolbar above the table.

Now click **Analyze Data** to see what we have.

A screenshot of the Galaxy History panel. It shows an 'Unnamed history' containing three datasets: '3: pro-B chrX/12 regions FASTQ raw', '2: pre-pro-B chrX/12 regions FASTQ raw', and '1: GTF Annotation chrX/12 regions'. Each dataset entry has a preview icon, edit icon, and delete icon.

## FASTQ datasets

And what we have is a history consisting of three datasets. Two of them are the raw read datasets for the pre-pro-B and pro-B cells. These datasets are in FASTQ format.

A screenshot of a dataset preview window. It shows the name '3: pro-B chrX/12 regions FASTQ raw', size '7.9 MB', format 'fastqsanger', and database 'mm10'. Below this, it says 'pro-B reads that map to select regions of chrX and chr12, plus random sampling of other reads'. At the bottom, there is a sequence viewer showing a portion of the FASTQ file.

To see a preview of any dataset, click on the dataset name. This displays some metadata about the dataset, and if it's in a text-based format, it will also show the first few lines of the dataset.

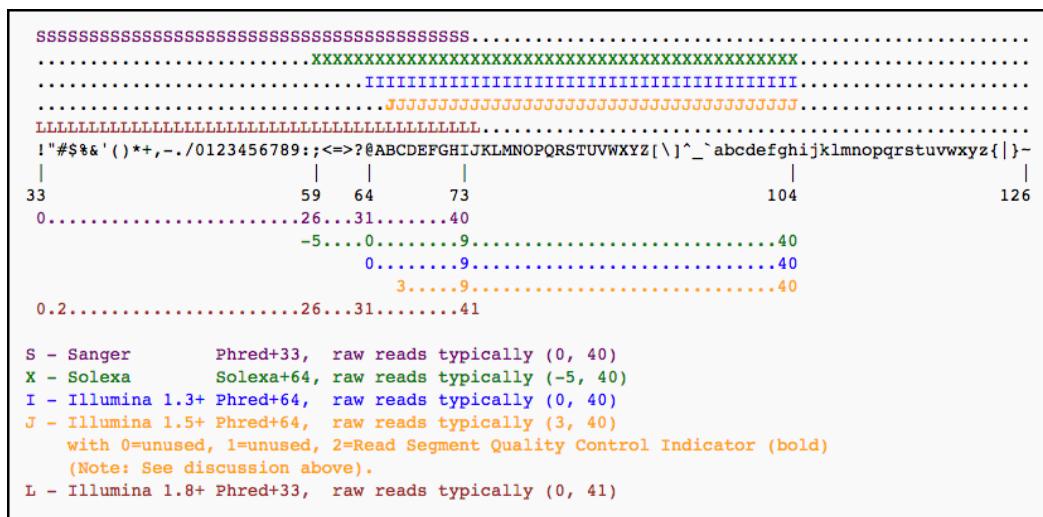
To view the full dataset, *poke it in the eye*. This will display the content of this FASTQ dataset in the middle panel.

A screenshot of the Galaxy dataset viewer. It shows the same dataset '3: pro-B chrX/12 regions FASTQ raw' with size '7.9 MB', format 'fastqsanger', and database 'mm10'. Below this, it says 'View data'. The main area displays the full sequence content of the FASTQ file.

## FASTQ Format

Each FASTQ entry is 4 lines long and incorporates three types of information:

- Line 1 starts with an @ and is the read's identifier
  - Line 2 shows the called bases in this read
  - Line 3 is a separator (the + character).
  - Line 4 shows the instrument's confidence in each of the base calls.



[https://en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

*Phred quality scores* are assigned to each base call. Phred scores are logarithmic and represent the likelihood that given base call is incorrect. Most datasets use Sanger scoring which typically ranges from 0 to 40. Some reference points in that spectrum:

<b>Score</b>	<b>Instrument thinks it will be wrong</b>	
10	10% of the time	(1 out of 10 times)
20	1% of the time	(1 out of 100 times)
30	0.1% of the time	(1 out of 1000 times)
40	0.01% of the time	(1 out of 10000 times)

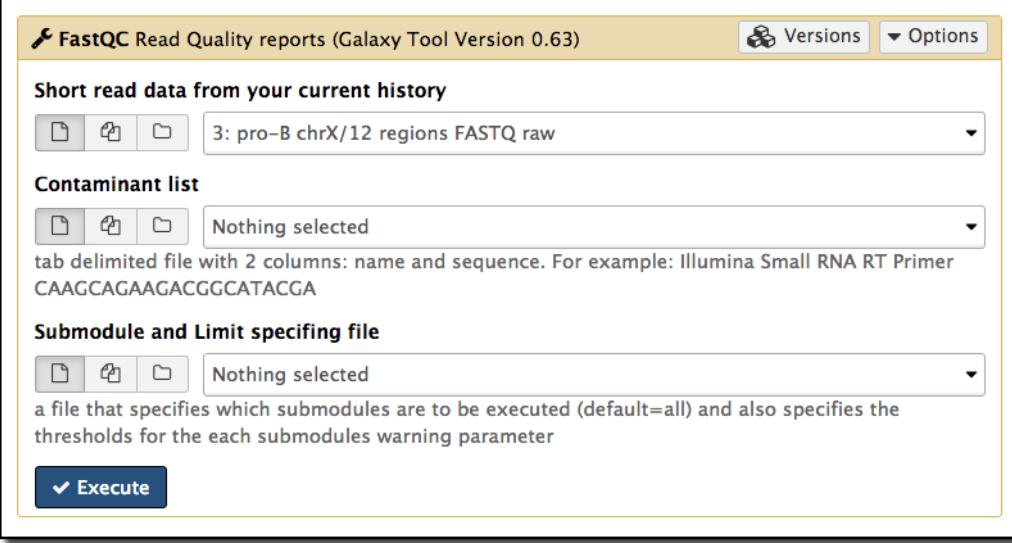
We don't see these scores in the FASTQ files because each score is encoded to a single character according the FASTQSanger convention. FASTQSanger maps scores to a range of adjacent ASCII characters, starting with ! for 0, and ending with Z for 40.

## Check FASTQ dataset quality

There are two ways to find tools on a Galaxy server. If you know the name of the tool you can enter it in the tool panel search box. You can also try searching for more general terms like “trim” or “quality.” Searching with more general terms will be hit or miss.

You can also just browse each Toolbox. In this case we are investigating the quality of an NGS dataset, so the **NGS: QC and Manipulation** toolbox looks the most promising. Click on the toolbox name to see the tools in it. In this case **FastQC Read Quality reports** is right at the top, and that sounds exactly like what we are looking for. Click on it.

FastQC is a widely used tool for summarizing the quality of FASTQ datasets. It's good at presenting the big picture, and at identifying specific



Metagenomic analyses  
Motif Tools  
**NGS: QC and manipulation**  
FastQC Read Quality reports  
Tabular to FASTQ converter  
FASTQ Quality Trimmer by sliding window  
FASTQ Trimmer by column

areas that may need to be addressed.

This is a typical, if simple, Galaxy tool form. Tool pages generally have these sections

- *Versions and options* - If multiple versions are supported, there will be link to run the others. And the options menu enables you see more about the tool and how it is wrapped in Galaxy.
- *The form* - where you configure and run the tool
- *Documentation* - Describes the tool and provides help. This section often include links to a tool's external documentation.
- *Citation* - A paper, if there is one, for this tool.

## Run our first tool: FastQC

FastQC has only three parameters that can be set. Let's set only one, **Short read data from your current history**. Click this option's dataset pulldown menu and select **2: pre-pro-B chrX/12 regions FASTQ raw**. We aren't going to set the contaminant list or limit the submodules/tests that FastQC runs. We'll request that it run all of its tests. Click **Execute**.

Now a quick succession of things (hopefully) happens

1. A **big green box** appears in the middle panel and two **small gray boxes** appear at the top of the history. These mean that the task has been successfully queued: It's ready to run, but isn't running yet.
2. The gray boxes in the history are replaced with **yellow boxes**. This means the job has started and is actively running
3. The yellow box are replaced with **green boxes** indicating that the tool finished successfully.

## Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Once a tool has finished running you can preview the output datasets (by clicking on the dataset name) or view the data itself (by poking it in the eye). Poke the **FastQC Web page** dataset *in the eye*.

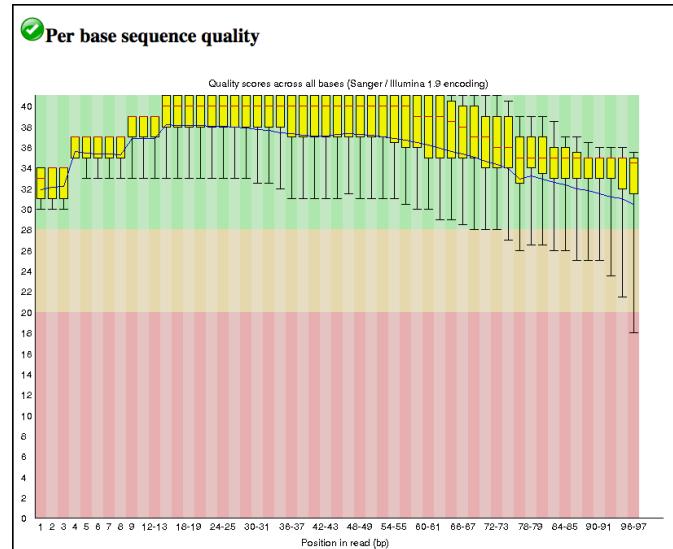
This brings up 12 different reports about the pre-pro-B FASTQ dataset.

For each test it runs, FastQC has predefined thresholds for what constitutes good, not-so-good, and bad. This dataset has passed 7 of the tests and got warning on the rest. Let's take a look at some of these tests.

## Per base sequence quality (boxplot)

Boxplots are the most common way to summarize the quality of a FASTQ dataset.

- **X axis is position in the read.** That is, the first column summarizes all the base calls in the first position of all the reads; the second column summarizes the second call in all the reads and so on. Starting at position 10, the information is aggregated for every two bases.
- **Y axis is Phred quality score**, from 0 to 40.
- Each column summarizes the quality scores across all reads at that position
  - **Yellow boxes** bound the middle 50% of quality scores at each position in the read.
  - **Whiskers** bound the middle 80% of quality scores at that position.
  - **Blue line** is the average score at each position
  - **Red lines** are the median score at each position.

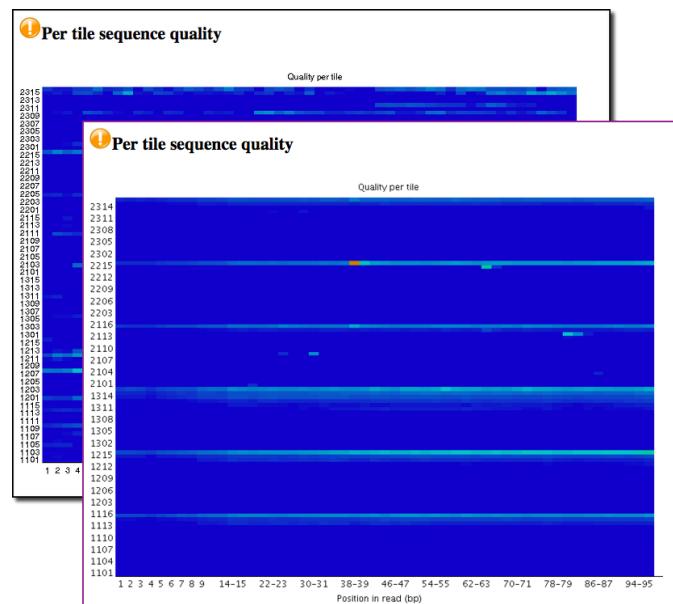


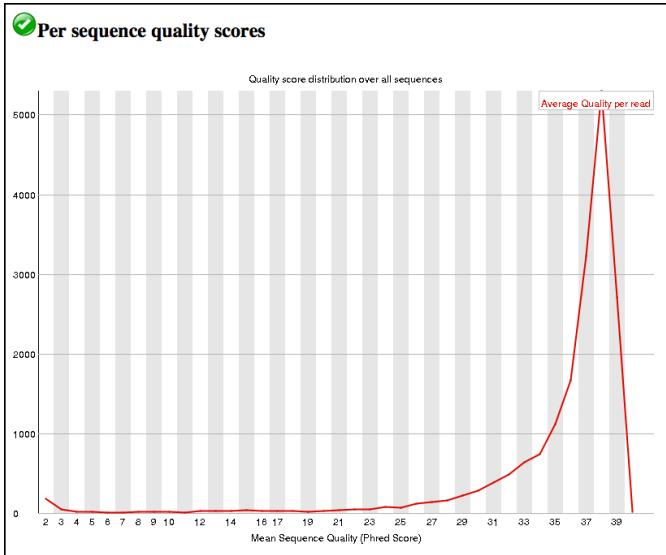
## Per tile sequence quality

This graph is particularly significant for this crowd. It highlights where on the slide the low quality reads tended to be. As core staff you can use this information to identify trouble spots with the sequencing run.

For Illumina data, tile information is embedded in the ID line of FASTQ entries.

*Note: We're showing the graphic for the complete ~90 million read dataset.*





## Per sequence quality scores

This graph gives us an idea just how bad (or good) our bottom 10% is. Here, the X axis is the quality score, and the Y axis is the number of times that score occurs across the entire dataset. The scores in this dataset are pretty good and there are only a small number of calls with confidence below 20. However, there is a small spike at 2. We'll want to drop those.

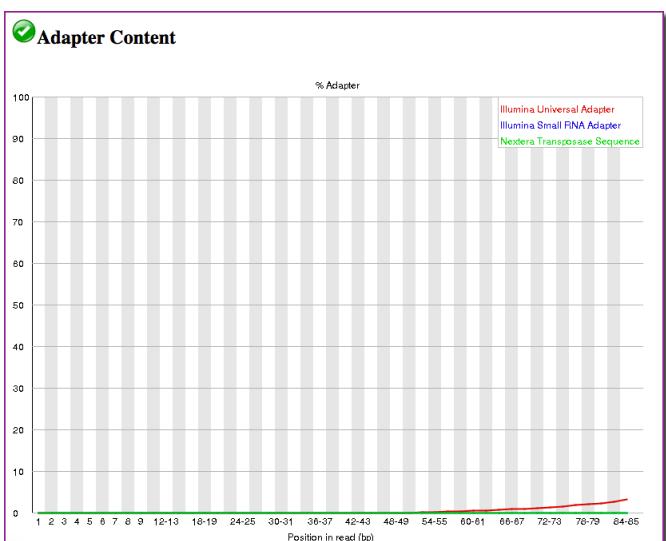
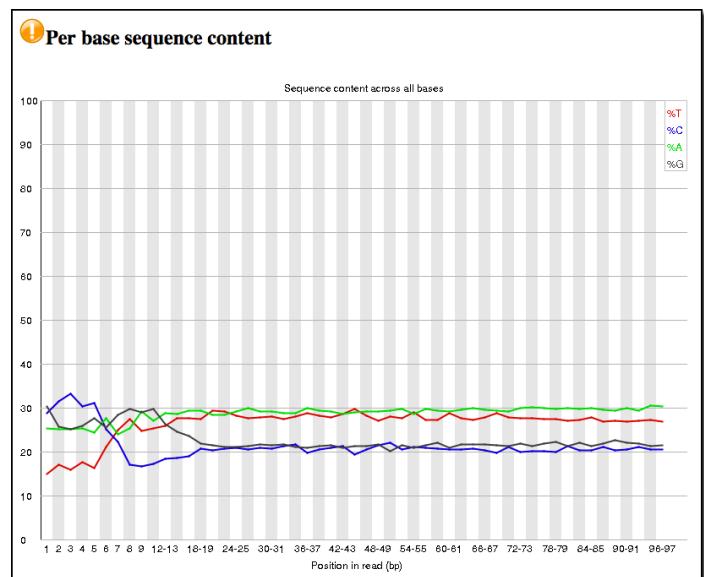
## Per base sequence content

This graph is good at highlighting untrimmed adapter or barcode problems.

Flat lines mean there is no bias for particular bases at particular positions. The turbulence at the front is caused by the use of random hexamers in Illumina library preparation. See

Hansen, et al., “[Biases in Illumina transcriptome sequencing caused by random hexamer priming](#)” *Nucleic Acids Research*, Volume 38, Issue 12 (2010)

The summary is: *It's not our fault.*



## Adapter Content

At the end of the reads the dataset is 3 to 4 percent untrimmed adapter. This falls within FastQC's tolerance, but not ours.

## And the rest

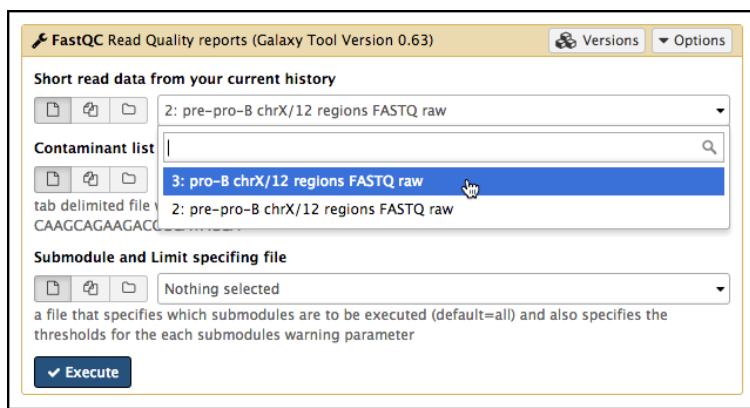
The remaining tests are either not informative (look at Sequence Length Distribution), flat out good (Per base N content), or look significantly different when the whole dataset is scanned (Per sequence GC content, Sequence Duplication Levels, Kmer Content). We'll ignore all these today.

## Rerun FastQC on pro-B dataset

Now let's run FastQC on the other dataset, the pro-B reads. We could rerun FastQC by finding it in the Tool panel again, or, we could rerun FastQC from the History panel. Let's do that.

Preview one of the **FastQC** datasets in your history by *clicking* on its name. In the preview *click* on the **Run this job again** looping arrow icon.

This launches the tool form in the middle panel and pre-populates it with the same settings it was run with before.



S: FastQC on data 2: Raw Data  
4: FastQC on data 2: Webpage  
320.2 KB  
format: html, database: mm10  
Run this job again  
3: pro-B chrX/12 regions FASTQ raw  
2: pre-pro-B chrX/12 regions FASTQ raw  
1: GTF Annotation chrX/12 regions

Click the dataset pulldown menu and select **3: pro-B chrX/12 regions FASTQ raw**. Then click **Execute**. Poke the new **FastQC Webpage** in the eye to review the results. The quality is similar.

## Improve the quality: Trimmomatic

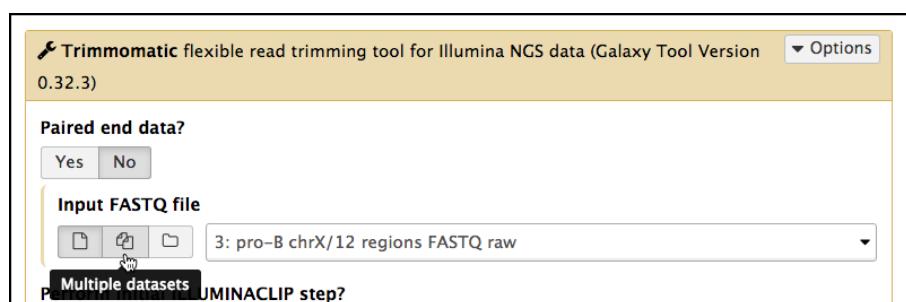
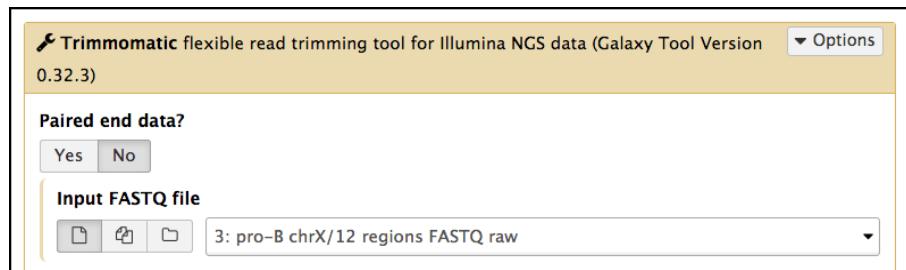
Use Trimmomatic to separate the good from the bad.

FastQC was the first tool in the **NGS: QC and manipulation** toolbox. **Trimmomatic** is the last. Click on it. First, **change Paired end data?** to **No**.

## Processing Multiple datasets with one submission

We could run Trimmomatic the same way we ran FastQC - by running it on one dataset, and then re-running it on the 2nd.

We can achieve the same effect more efficiently by using the **Multiple datasets** feature, another available shortcut.



To enable this *click* on the **Multiple datasets** icon under Input FASTQ file.

This changes the parameter selection from a pull-down to a multi-select field. *Select* both datasets.

## Select Trimmomatic Filters

We want to use three different filters in Trimmomatic:

1. **ILLUMINACLIP** - This will trim known Illumina adapters from reads in the dataset.
2. **SLIDINGWINDOW** - Cutting once the average quality within a sliding window falls below a threshold. Trims from both ends
3. **MINLEN** - Drop any reads that are shorter than a given length after trimming.

First, *select Yes* under **Perform initial ILLUMINACLIP step?**. Then *change the Adapter sequences to use pulldown to TruSeq3 (single-ended, for MiSeq and HiSeq)*. Leave other parameters at the default.

Update the default

SLIDINGWINDOW step to use the same window definition as the paper. **Change Number of bases to average across to 3**.

To deal with reads that are too short add another operation.

Trimmomatic Operation  
1: Trimmomatic Operation  
Select Trimmomatic operation to perform  
Sliding window trimming (SLIDINGWINDOW)  
Number of bases to average across  
3  
Average quality required  
20  
+ Insert Trimmomatic Operation

**Click on + Insert Trimmomatic Operation. Change Select Trimmomatic operation to perform to Drop reads below a specified length (MINLEN) and set Minimum length of reads to be kept to 32.** All 3 steps are set up. *Click Execute*.

2: Trimmomatic Operation  
Select Trimmomatic operation to perform  
Drop reads below a specified length (MINLEN)  
Minimum length of reads to be kept  
32  
+ Insert Trimmomatic Operation  
Execute  
Execute: Trimmomatic (0.32.3)  
What it does

## Trimmomatic Results

To see the net decrease in dataset size, preview the before and after datasets, and compare their size.

To see the full Trimmomatic summary, *click* the ⓘ icon (**View details**) in the dataset preview. This displays the metadata for this dataset.

Two links for **stdout** and **stderr** are included with the metadata. These are standard files in Unix to record *standard output* and *standard error* from the tool.

*Click stdout* to see the full summary. This is informative, but let's run FastQC again on both and compare those reports with the original reports.

/mnt/galaxy/files/000/dataset\_606  
\* ILLUMINACLIP:/mnt/galaxy  
View details  
6:D0PLGACXX:8:1101:10093:  
CCTCAGATGAATATTAAATATTGCCACTTGGGGAGAC  
+  
What it does

Tool version:	
Tool Standard Output:	<a href="#">stdout</a>
Tool Standard Error:	<a href="#">stderr</a>
Tool Exit Code:	0

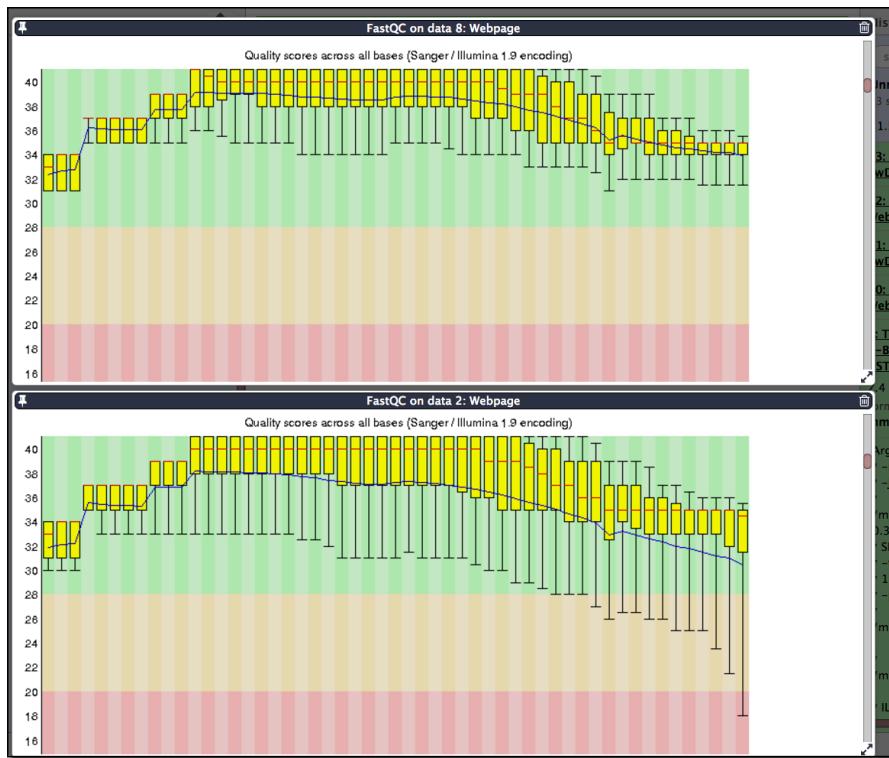
**Click FastQC** in the tool panel, **select the Multiple datasets icon** under **Short read data** from your current history, and then **select both Trimmomatic on ... datasets**. Click **Execute**.

*Note: We aren't using paired-end data today, but if we were Trimmomatic would automatically maintain read pairings throughout all steps.*

The screenshot shows the Galaxy FastQC tool interface. At the top, it says "FastQC Read Quality reports (Galaxy Tool Version 0.63)". Below that is a toolbar with icons for Versions and Options. The main area is titled "Short read data from your current history". A list of datasets is shown, with the last three highlighted in purple: "9: Trimmomatic on pro-B chrX/12 regions FASTQ raw", "8: Trimmomatic on pre-pro-B chrX/12 regions FASTQ raw", and "3: pro-B chrX/12 regions FASTQ raw". A note below the list states: "This is a batch mode input field. A separate job will be triggered for each dataset." There are also icons for saving and deleting datasets.

## Using Galaxy Scratchpad to view multiple datasets

We want to compare pre and post-Trimmomatic FastQC reports. Scratchpad will enable us to do this without going insane. To enable Scratchpad, click on the grid icon ( ) in the top menu bar. This turns it yellow ( )! Now, scroll down in your history to first FastQC report on the pre-pro-B data (most likely named **4: FastQC on data 2: Webpage**) and *poke it in the eye*.



*Click somewhere outside the displayed dataset. Scroll up the history to the second FastQC report on the pre-pro-B data (most likely the fourth dataset from the top) and *poke it in the eye*.*

After some scrolling and resizing you can now compare the two reports at once.. Additional datasets can be added as desired.

We see that the quality has improved.

To switch back to the standard view either click outside the displayed datasets or poke the yellow eye ( ). To disable the Scratchbook display, click on the yellow grid icon.

## Map the reads

The data has now been cleaned and we are ready to map them against the mouse reference genome.

## HISAT2

Bring up the HISAT2 tool form and **change Single end or paired reads?** to **Individual unpaired reads** and then **select the Multiple datasets icon** under **Reads**. Select both **Trimmomatic** datasets.

HISAT is an “option-rich” tool. The HISAT tool wrapper takes a common approach of placing them in sections that can be expanded as needed. HISAT has 5 such sections:

If you expand all sections, HISAT options go from 1 screen to almost 6. Understanding all these options is the right thing to do, but it’s also daunting. One of Galaxy’s strengths is that it allows you to *experiment with tools* and *learn them incrementally*. We are going to start with setting just one parameter.

**Change Spliced alignment parameters to Specify spliced alignment parameters.** This causes almost 15 additional options to be displayed, most of them giving you fine-grained control over how reads that are split across multiple exons are scored. Ignore all those and scroll down to **GTF file with known splice sites** and **select 1: GTF Annotation chrX/12 regions** and **click Execute**.

## HISAT2 Results

HISAT2 creates a mapped reads dataset for each of the two inputs. These files are in BAM format, an efficient binary format for representing mapped data.

To see the full HISAT2 summary **click the ⓘ icon (View details)** in the dataset preview. Then **click the stderr link**. Our overall alignment rates are over 98% for both datasets.

9: Trimmomatic on pro-B chrX/12 regions FASTQ raw  
8: Trimmomatic on pre-pro-B chrX/12 regions FASTQ raw  
3: pro-B chrX/12 regions FASTQ raw  
2: pre-pro-B chrX/12 regions FASTQ raw

This is a batch mode input field. A separate job will be triggered for each dataset.

Source for the reference genome to align against

Use a built-in genome

Alignment options

Use default values

Input options

Use default values

Scoring options

Use default values

Spliced alignment parameters

Use default values

Paired alignment parameters

Use default values

Disable spliced alignment

Yes No

GTF file with known splice sites

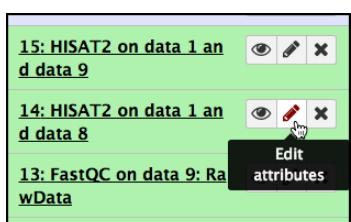
1: GTF Annotation chrX/12 regions

Transcriptome assembly reporting

Use default reporting.  
 Report only those alignments within known transcripts.  
 Report alignments tailored for transcript assemblers including StringTie.  
 Report alignments tailored specifically for Cufflinks.

```
[bam_header_read] EOF marker is absent. The input is probably truncated.  
[samopen] SAM header is present: 66 sequences.  
31843 reads; of these:  
 31843 (100.00%) were unpaired; of these:  
   482 (1.51%) aligned 0 times  
   26747 (84.00%) aligned exactly 1 time  
   4614 (14.49%) aligned >1 times  
 98.49% overall alignment rate
```

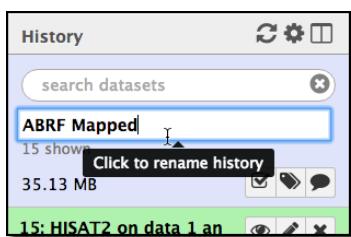
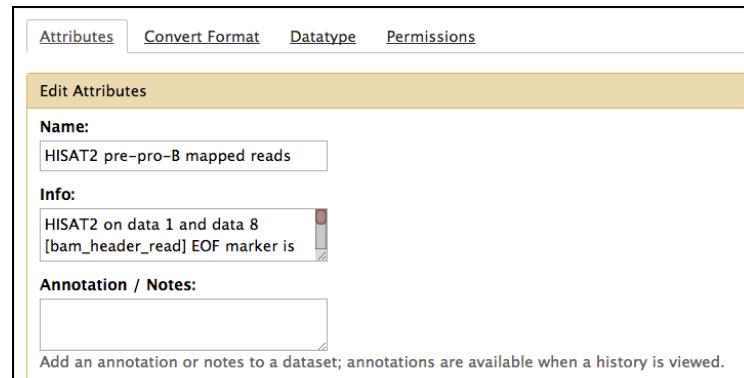
## Updating dataset and history metadata: Renaming



The HISAT2 output datasets have names like **HISAT2 on data 1 and data 8**. Rename them to clearly indicate what's in them.

**Click on the pencil (Edit attributes) icon** for the pre-pro-B

**HISAT2 dataset**. Name it something like **HISAT2 pre-pro-B mapped reads**. **Click Save** and then *repeat* with the pro-B dataset.



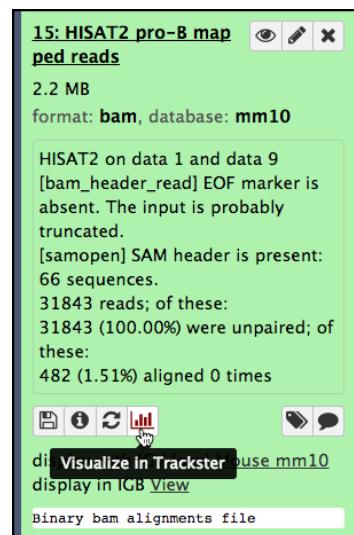
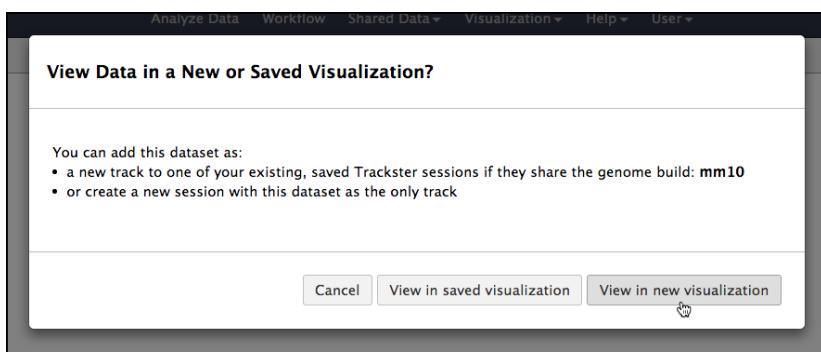
We should also give our history a name. A best practice is to always name your histories. To do this, *click on Unnamed history*, *enter* a new and informative history name, and then *press return* or *enter*.

## But what do the HISAT2 mappings look like?

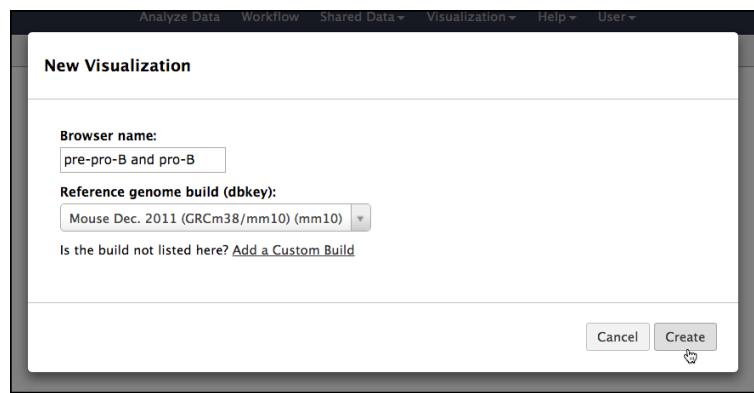
There are summary tools that provide more statistics on things like mapping state (**Flagstat**) and how many reads mapped to each chromosome (**IdxStats**).

We can visually inspect the mappings in the context of the genomic regions they mapped to. Later today you'll use the IGV desktop program to visualize genomic data. Right now, let's learn how to do this inside Galaxy, using the Trackster visualization tool.

To launch Trackster, open a preview of one of the HISAT2 datasets and *click* on the **Visualize in Trackster** icon.



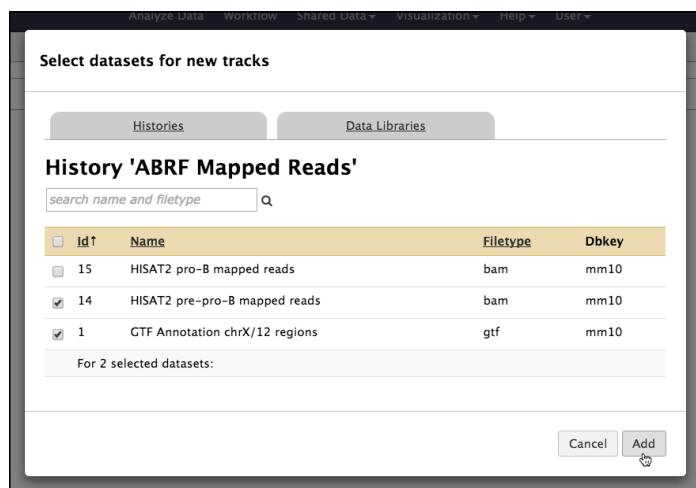
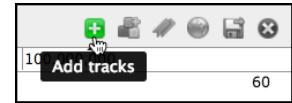
Select **View in new visualization** and then give the visualization a meaningful name and *click Create*.



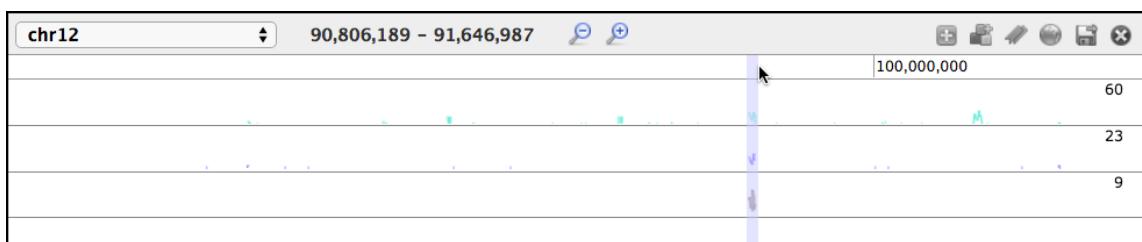
This will bring up a new Trackster browser with a status message saying that it is indexing the dataset for display. *Switch from chr1 to chr12 (or chrX), where we have mapped data.*



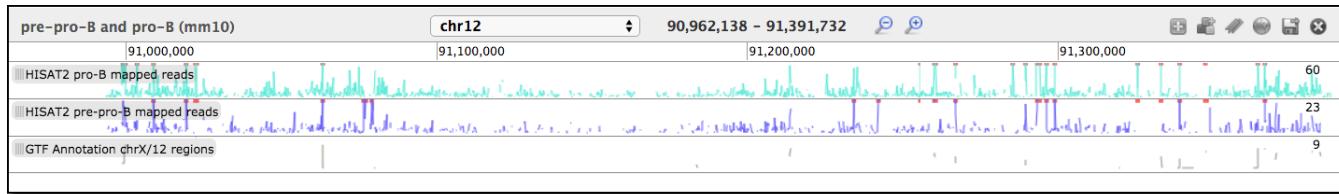
While the track is loading, add other datasets to the display. *Click on the + (Add tracks) icon in the upper right, and then select the other mapped dataset and the reference annotation.*



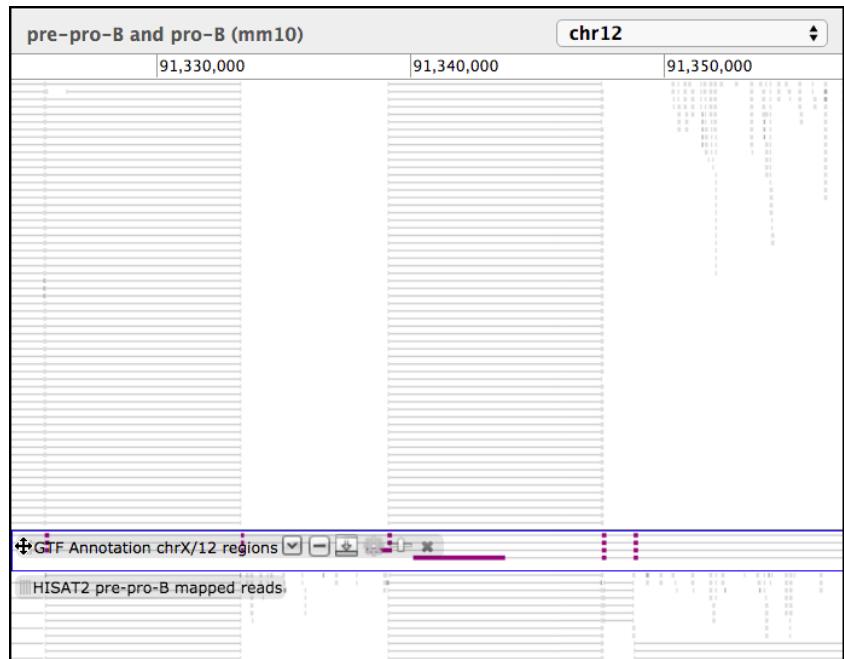
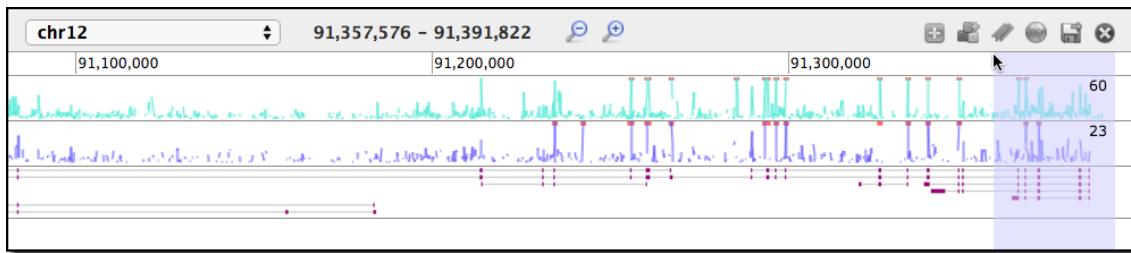
Once those new tracks have loaded, zoom in on the region of interest - the area with the annotations: *Click and drag on the coordinates track (near the top) and include the entire region with annotation.*



You may have to repeat that a few times to get the region of interest:

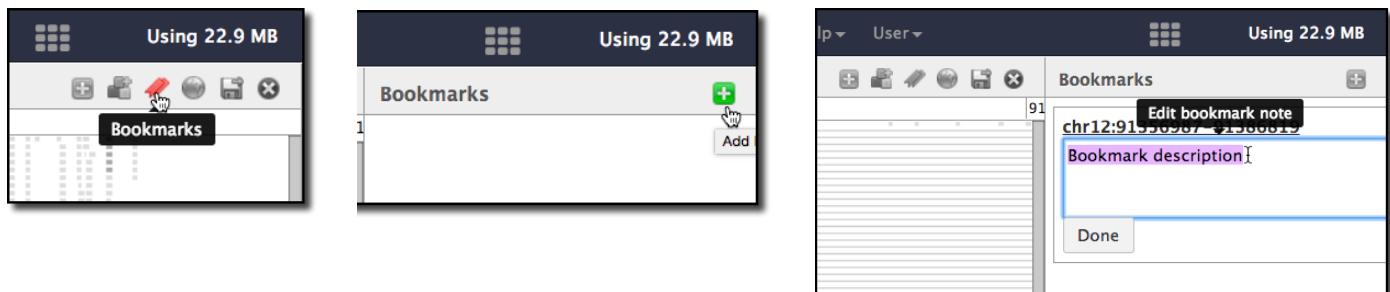


The top 2 tracks show the read depth for any position in this region, and the bottom track shows the annotation (but not very well). **Hover over the GTF Annotation chrX/12 regions track and click the Set display mode icon and select Squish.** To get a better idea of what the aligned reads look like, zoom in the last 5 or so introns on the far right.



This changes the display mode for all tracks and now we can see individual reads, including where splice junctions were introduced to map the reads.  
Drag the annotation track to the middle by **clicking and dragging the vertical lines icon** at the left of the track title upwards to between the two BAM tracks.

We'll want to come back to this location once we have predicted transcripts for these datasets. Bookmark this location by *clicking* on the **bookmark ribbon icon** in the upper right corner. Then *click* the **+ (Add bookmark) icon** and give the bookmark a meaningful description. Finally, *click* the **save icon** to save your visualization.



## Transcript Prediction with StringTie

HISAT2 makes no prediction about how those reads might assemble into transcripts. That job is handled by StringTie.

Search for **StringTie** in the tool panel and launch the tool form. Select **Multiple datasets** under **Mapped reads to assemble transcripts from**, and then select both **HISAT2** datasets. Change **Use GFF file to guide assembly** to **Use GFF**. Click **Execute**.

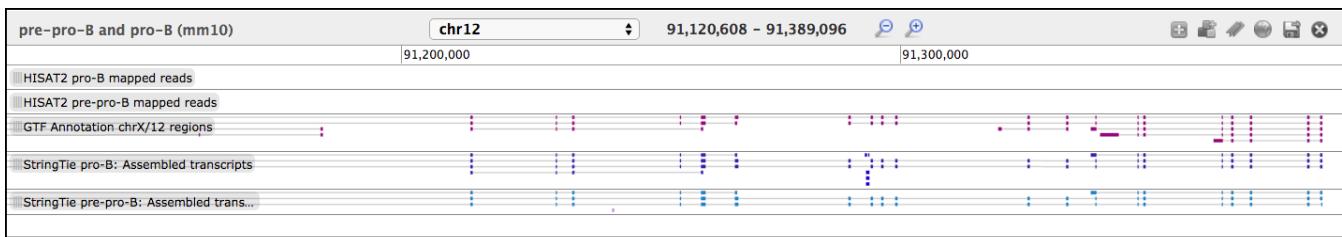
StringTie generates 3 output datasets for each input dataset:

- Coverage - All the transcripts in the reference GTF that are fully covered by reads.
- Gene abundance estimates - This is created but it's empty because we did not ask for it in the advanced options section.
- Assembled transcripts - This is the one we care most about.

### Rename the two **assembled transcripts**

datasets to include **pre-pro-B** or **pro-B** in the dataset name. Then add the assembled transcripts datasets to the visualization you created earlier.

Once in Trackster, add the other assembled transcripts dataset and hide the content of the two mapped reads datasets, and zoom out a few times. Set the display mode for both new datasets to Squish.



A couple of things to note about this region:

- the pro-B assembled transcripts have some transcripts that are not in the reference, and even some new exons
- the pre-pro-B assembled transcripts are largely a subset of the pro-B predicted transcripts.
- several annotated transcripts and exons are not detected at all in these datasets.

Save the visualization and go back to **Analyze Data**.

## And there is more!

### Workflows: repeating an analysis

We'll want to apply the same analysis to our full datasets, and maybe with other similar experiments as well. To do this create a workflow - a repeatable recipe for doing an analysis.

Galaxy supports *de novo* workflow creation, and creating them from histories. To create a workflow from our history, click the **cog** at the top of the history panel, and then select **Extract Workflow**.

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

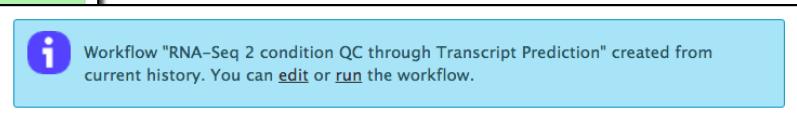
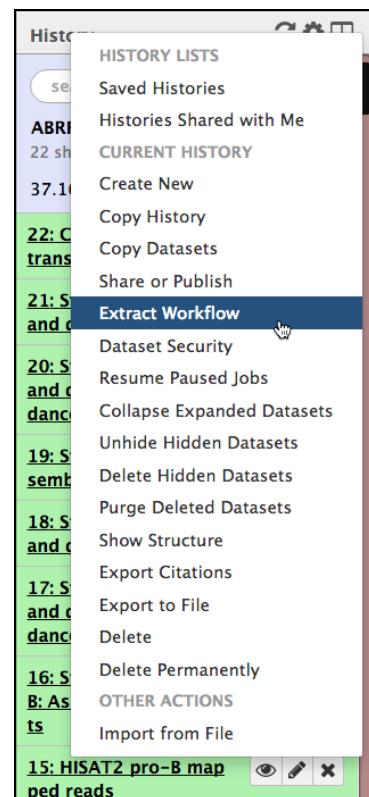
**Workflow name**: RNA-Seq 2 condition QC through Transcript Prediction

**Create Workflow** | **Check all** | **Uncheck all**

Tool	History items created
Unknown <i>This tool cannot be used in workflows</i>	1: GTF Annotation chrX/12 regions ✓ Treat as input dataset
Unknown <i>This tool cannot be used in workflows</i>	2: pre-pro-B chrX/12 regions FASTQ raw ✓ Treat as input dataset
Unknown <i>This tool cannot be used in workflows</i>	3: pro-B chrX/12 regions FASTQ raw ✓ Treat as input dataset
FastQC <input checked="" type="checkbox"/> Include "FastQC" in workflow	4: FastQC on data 2: Webpage
FastQC <input checked="" type="checkbox"/> Include "FastQC" in workflow	5: FastQC on data 2: RawData
Trimmomatic <input checked="" type="checkbox"/> Include "Trimmomatic" in workflow	6: FastQC on data 3: Webpage
Trimmomatic <input checked="" type="checkbox"/> Include "Trimmomatic" in workflow	7: FastQC on data 3: RawData
Trimmomatic <input checked="" type="checkbox"/> Include "Trimmomatic" in workflow	8: Trimmomatic on pre-pro-B chrX/12 regions FASTQ raw

This takes you to the create workflow page, where you can specify which of the steps and input in your current history you wish to include in the new workflow.

Give the new workflow an informative name and click **Create Workflow**. This message appears:



Test the new workflow on the exact same inputs that we just analyzed manually

Click the **run** link in the message (or you can also get there by clicking Workflow in the top bar). This brings up a form asking you to define inputs to the workflow. Set the first three input datasets for this run to the first three datasets from your current history:

Scroll down to the bottom of the form and check **Send results to a new history**.

Name the workflow and click **Run workflow**.

This displays an enormous green box in the middle panel. Click the link near the top of the box to get to the new history.

**Running workflow "RNA-Seq 2 condition QC through Transcript Prediction"**

**Step 1: Input dataset**

**Input Dataset** 1: GTF Annotation chrX/12 regions  
type to filter

**Step 2: Input dataset**

**Input Dataset** 2: pre-pro-B chrX/12 regions FASTQ raw  
type to filter

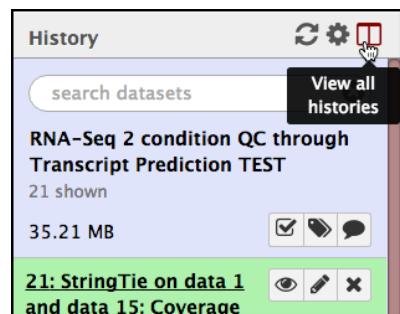
**Step 3: Input dataset**

**Input Dataset** 3: pro-B chrX/12 regions FASTQ raw  
type to filter

Expand All | Collapse

## Working with multiple histories

Running the workflow and sending results to a new history means that we now have two histories in this Galaxy instance. There are a couple of ways to see all your existing histories and to switch between them. The most recent and easiest to use is the all histories view. This can be accessed by clicking on the **table icon (View all histories)** at the top of the history panel.



The all histories view presents all your saved histories (we only have two), with the current one pinned at the left, and your other histories listed in reverse chronological order, from left to right.

## Sharing and publishing

Galaxy histories, workflows, and visualization can all be shared and published with Galaxy. Sharing in Galaxy means sharing something with someone else either directly with their Galaxy account, or by creating a URL that can be shared. In addition Galaxy objects can be published, making it easy for anyone to discover the object. Anything that is published will be listed in the appropriate Shared Data section.

## Galaxy pages

The datasets and analysis used in the paper (and in this tutorial) are available here:

<https://usegalaxy.org/u/thereddylab/p/prediction-of-gene-activity-in-early-b-cell-development-based-on-an-integrative-multi-omics-analysis>

This is a *Galaxy Page*, a document integrated into a Galaxy server that embeds and provides direct links to Galaxy objects such as histories, workflows, and datasets. Galaxy Pages are a way to bundle together all related analysis and to describe the semantics of your analysis.

# Mass Spectrometry-Based Proteomics Data Analysis Using Galaxy

---

*ABRF 2016 ANNUAL MEETING  
February 20-23*

# Introduction

1	Introduction .....	3
1.1	<i>Scope of this tutorial</i> .....	3
1.2	<i>Why proteomics?</i> .....	3
1.3	<i>Basics of Mass Spectrometry for Proteomics</i> .....	3
1.4	<i>Outline of tutorial</i> .....	4
2	Getting Started .....	5
2.1	<i>Getting Started</i> .....	6
3	Inputs: Peaklists and Database Collection .....	8
3.1	<i>Dataset Collection</i> .....	9
3.2	<i>MGF Formatter</i> .....	10
4	Database Generation .....	11
4.1	<i>Creating a UniProt Reference Proteome</i> .....	12
4.2	<i>Creating a Contaminates Reference Proteome</i> .....	13
4.3	<i>Merging UniProt, cRAP, and RNA-seq Derived Databases</i> .....	14
5	MS database search with search algorithm, SearchGUI .....	15
5.1	<i>SearchGUI Parameters for Protein Algorithm Searches</i> .....	16
6	Protein and Peptide Evaluation with PeptideShaker .....	17
6.1	<i>PeptideShaker</i> .....	18
6.2	<i>PeptideShaker Outputs</i> .....	19
7	mz to SQLite Database Generation .....	20
7.1	<i>Generating the sqlite Database</i> .....	21
7.2	<i>PSM Evaluator in Galaxy-P</i> .....	22
8	Generating and Running Workflows within Galaxy .....	23
8.1	<i>Extract Workflow from Current History</i> .....	24
8.2	<i>Edit the Workflow</i> .....	25
8.3	<i>Running Workflows</i> .....	26

## 1 Introduction

### 1.1 Scope of this tutorial

This is a practical, hands-on tutorial designed to give participants experience with mass spectrometry-based proteomics data analysis in Galaxy. During this tutorial, users will learn about available tools in Galaxy for proteomics (referred to herein as “Galaxy-P”). Furthermore, we will discuss how to build and implement workflows utilizing proteomic tools within Galaxy. Participants are expected to be familiar with mass spectrometry, database searches and Galaxy.

### 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, a proteomic analysis 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, Galaxy-P 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 samples need to be processed before introduction 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 proteomics**). 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.

## Introduction

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.

### Reference materials

Galaxy 101: Building Up and Using a Proteomics Workflow

[http://usegalaxyp.readthedocs.org/en/latest/sections/galaxyp\\_101.html](http://usegalaxyp.readthedocs.org/en/latest/sections/galaxyp_101.html)

### 1.4 Outline of tutorial

- 1 Introduction
- 2 Getting Started
- 3 Inputs: Peaklists and Database Collection
- 4 Database Generation
- 5 MS database search with search algorithm, SearchGUI
- 6 Protein and Peptide Evaluation with PeptideShaker
- 7 mz to SQLite Database Generation
- 8 Generating and Running Workflows

## 2 Getting Started

### ★ Tutorial Dataset ([Sect 2.1 page 6](#))

This tutorial will identify peptide and protein identifications for mass spectrometric data. The sample dataset used in this tutorial was created from a combination of published dataset ([J Proteomics Bioinform. 2014 Feb 17;7](#)) and unpublished data from Dr. Reddy's lab courtesy of [Mohammad Heydarian](#). The FASTA files consist of protein sequences derived RNA-Seq data from *Ebf1* -/- pre-pro B cells and *Rag2* -/- pro-B cells.

### ★ Accessing Galaxy for Proteomics

For the purposes of this tutorial at ABRF 2016, a special cloud instance of Galaxy-P 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 Galaxy-P server](#) to associate themselves with the Galaxy framework and current open-source proteomic tools.

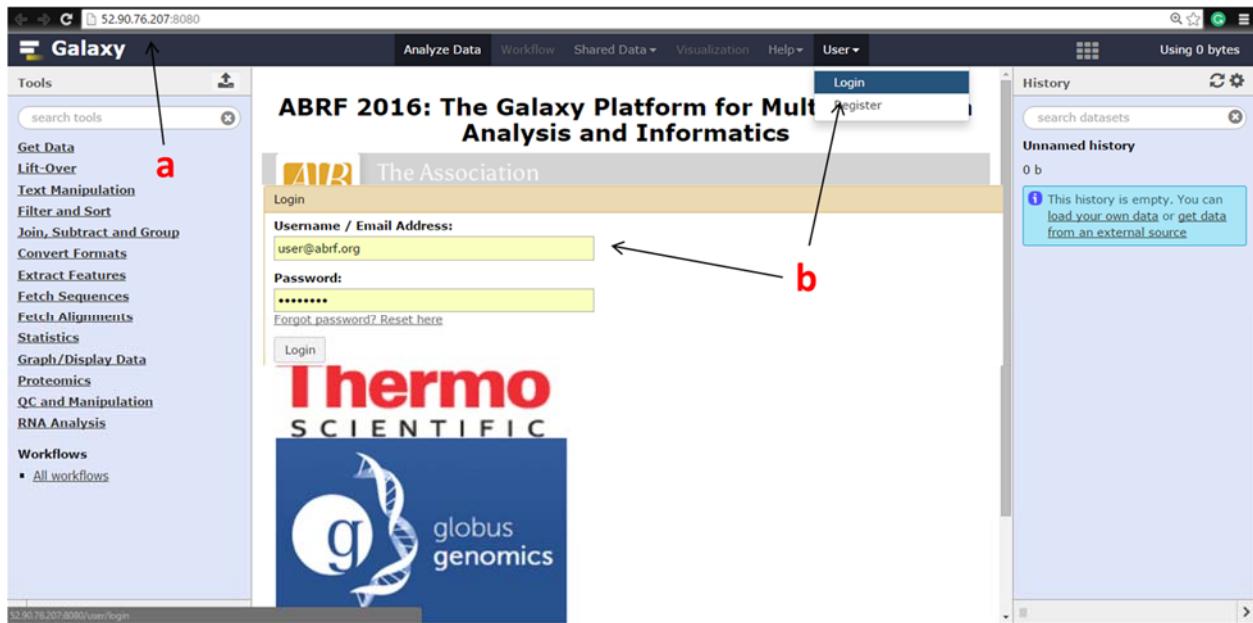
*NOTE: This dataset was chosen to allow fast processing and response times in a workshop setting where dozens of people will be submitting jobs at once to the server.*

# Getting Started

## 2.1 Getting Started

### ★Tutorial Dataset

- a) Open a web browser and navigate to the Galaxy-P cloud instance (instructions will be provided)
- b) Go to → User → Login and enter provided credentials and password
- c) At the top of the screen select “User” then migrate to “Saved Histories.”
- d) Select “Session 2 Start History” from the list of the published histories.
- e) “Session 2 Start History” should appear in your right panel History



# Getting Started

The screenshot shows the Galaxy Platform interface with the title "ABRF 2016: The Galaxy Platform for Multi-Omic Data Analysis and Informatics". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The user is logged in as "user@abrf.org". A context menu is open at the top right, labeled "Saved Histories", which includes options like "Saved Datasets", "Saved Pages", and "API Keys". The main area displays a "Saved Histories" list with the following entries:

Name	Datasets
Session 2 Start History	4
Unnamed history	0
BED Files from Large Searches	2
OUTPUT From Entire ProteoGenomics Workflow	19
INPUT For Entire ProteoGenomics Workflow	6
END History for Session 4	11
ABRF SW4 Session 4 START History	4
Unnamed history	2
Session 2 End History	12

Annotations with red letters and arrows point to specific elements:

- "e" points to the "Session 2 Start History" entry in the history list.
- "d" points to the "Session 2 Start History" entry in the history list.
- "C" points to the "Saved Histories" menu item in the top right.

### 3 Inputs: Peaklists and Database Collection

#### ★ Raw files and Peaklists ([Sect 3.2 page 10](#))

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](#).

#### ★ Database Collection ([Sect 3.1 page 9](#))

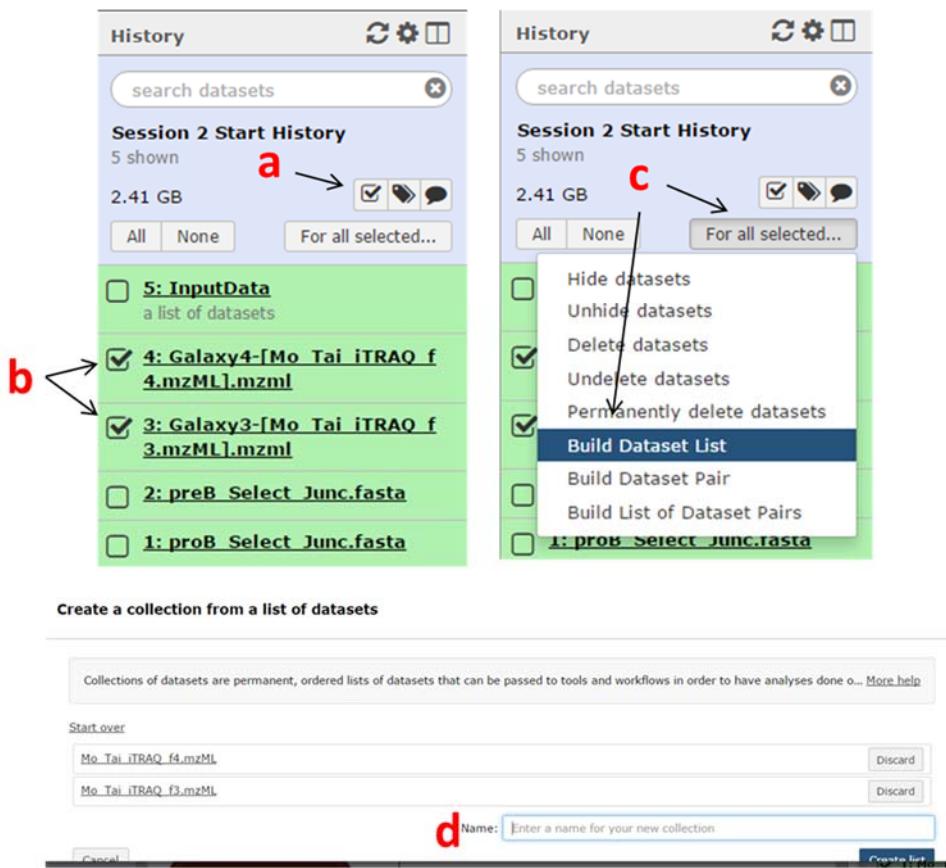
Typically, one biological sample can result in multiple RAW files. These RAW files need to be processed collectively on the backend to obtain an overall result of identified proteins for the starting sample. Galaxy was originally designed for analysis of a single data file generated for a single sample. However, Galaxy-P now utilizes a function called “Dataset collections” wherein multiple files of the same type can be defined as a collection, and processed together throughout the proteomic data analysis workflow. This capability simplifies the analysis process as the search engine processes the individual peaklist files during the sequence database search and then groups the results together as one output. Typically, a Dataset Collection is defined on processed results (mzML or mgf files) after raw file conversion using Msconvert and/or MGF Formatter.

## Inputs: Peaklists and Database Collection

### 3.1 Dataset Collection

#### ★Database Collection

- a) In the history panel click the checked box.
- b) Once you have clicked on the check box, notice that all inputs have an option to be checked. Please check the mzml files which you would like to combine for dataset collection.
- c) Proceed by clicking on “For All Selected...” and from the drop-down menu choose Build Dataset List.
- d) Give the desire dataset collection a name and click “Create List”
- e) Once a new input for dataset collection has been created in the history, unclick the checked box that you checked in part a.

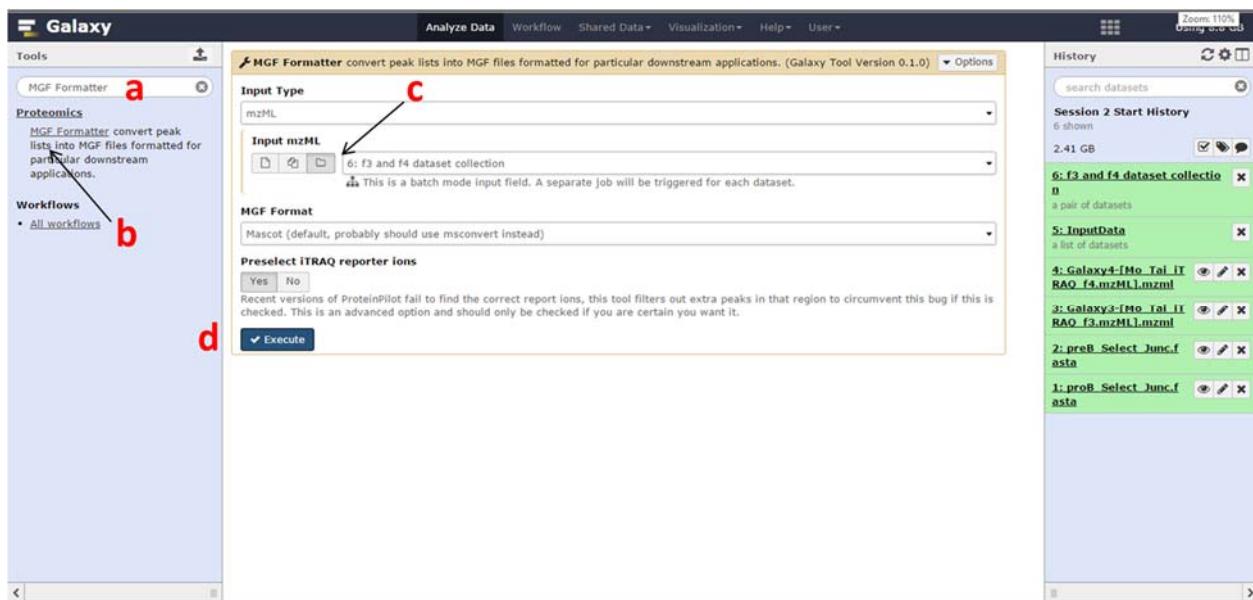


## Inputs: Peaklists and Database Collection

### 3.2 MGF Formatter

#### ★ Raw files and Peaklists

- In the search box under “Tools” type MGF Formatter.
- Click on the MGF Formatter tool and the tool will be prompted in the center panel.
- The following parameters should be selected.
  - Input Type → mzML
  - Input File → Dataset collection is represented by a folder. Choose the file for dataset collection.
  - MGF Formatter → Mascot
  - Preselect iTRAQ reporter ions → Yes
- Click Execute



## 4 Database Generation

### ★ Database Generation ([Sect 4.1 page 12](#))

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. Galaxy-P 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 Galaxy-P.

### ★ RNA-seq Derived Databases

Real time assessments of expressed protein variants can be done using RNA-seq data to derive a protein database. Generation of these RNA-seq derived protein databases will be discussed in Session 3.

### ★ Merging Multiple Databases ([Sect 4.3 page 14](#))

It is not uncommon to use a combination several databases as one general protein database for a protein database algorithm search. For Session 2, a merged protein database will include a UniProtKB reference, contaminates reference, and the two RNA-seq derived protein databases. The order in which the databases are merged will identify common peptides to one distinct protein database.

Example: Two databases merged

1. UniProtKB
2. Prob Selected Juncs

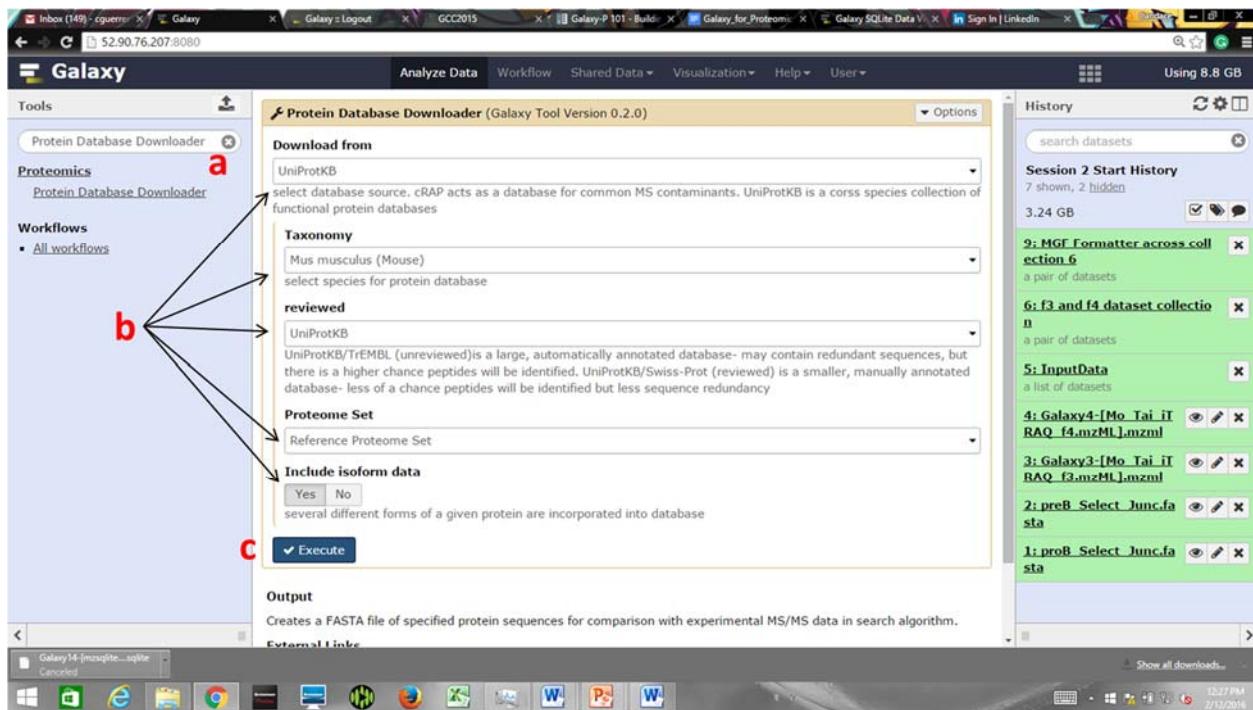
All peptides identified by both, will be further associated with UniProtKB

## Database Generation

### 4.1 Creating a UniProt Reference Proteome

#### ★Database Generation

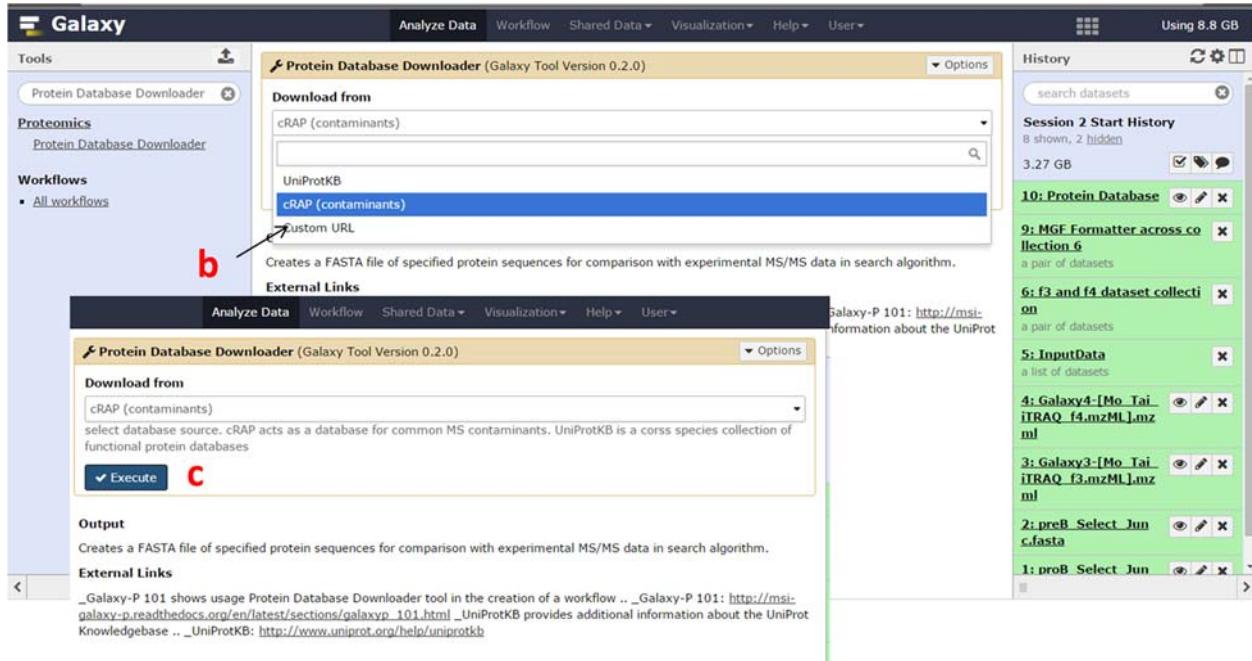
- In the search box under “Tools” type “Protein Database Downloader” and double-click on the tool
- From the flowing drop-down menu set the following parameters
  - Download From → UniProtKB
  - Taxonomy → Mus musculus (Mouse)
  - Reviewed → UniProtKB
  - Proteome Set → Reference Proteome Set
  - Include Isoform Data → Yes
- Click Execute



## Database Generation

### 4.2 Creating a Contaminates Reference Proteome

- Click again on the “Protein Database Downloader” tool
- From the “Download from” drop-down menu select “cRAP (contaminants)”
- Click Execute

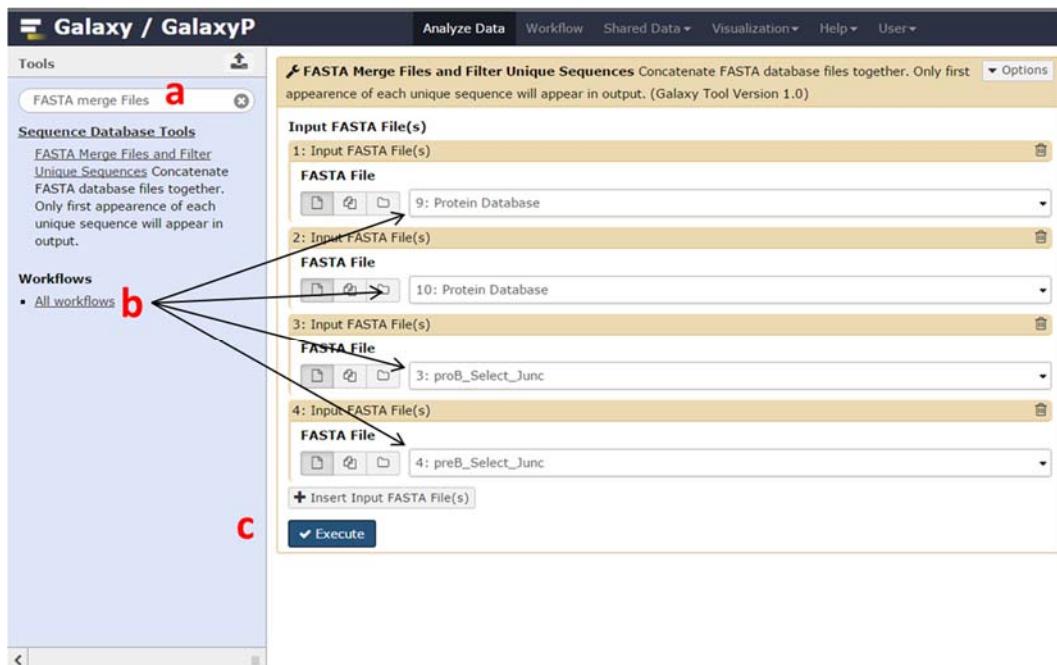


## Database Generation

### 4.3 Merging UniProt, cRAP, and RNA-seq Derived Databases

#### ★ Merging Multiple Databases

- a) In the search box under “Tools” type “FASTA Merge File and Filter Unique Sequences” and double-click on the tool and click “Add FASTA file”
- b) From the drop-down menus select the following parameters
  - 1: Input FASTA Files → UniProtKB Database
  - 2: Input FASTA Files → cRAP (contaminants) Database
  - 3: Input FASTA Files → proB Select Junc
  - 4: Input FASTA Files → preB Select Junc
- c) Click Execute



## 5 MS database search with search algorithm, SearchGUI

### ★ 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 \(2011\)](#).

### ★ SearchGUI in Galaxy-P

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 ([Sect 5.1 page 16](#))

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 within SearchGUI (see [SearchGUI source code](#) for more details). We will set-up our workflows to avoid any of these errors.

### ★ 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-P. 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-P for more detail).

# MS database search with search algorithm, SearchGUI

## 5.1 SearchGUI Parameters for Protein Algorithm Searches

### ★Setting Parameters

- In the search box under “Tools” type “SearchGUI” and double-click on the tool
- Select the parameter settings displayed below
- Click Execute

**b**

The screenshot shows the 'Search GUI' interface for protein identification. The 'Protein Database' dropdown is set to '11: Merged and Filtered FASTA from data-4, data-3, and others'. Under 'DB-Search Engines', several engines are selected: X!Tandem, MyriHatch, MS-Amanda, MS-GF+, OMSSA, and Comet. The 'Create a concatenated target/decoy database before running PeptideShaker' option is set to 'Yes'. In the 'Input Peak Lists (mgf)' section, 'B: MGF Formatter across collection 5' is selected. The 'Precursor Ion Tolerance' is set to 10 ppm. The 'Fragment Tolerance (Daltons)' is set to 0.1. The 'Enzyme' is set to 'Trypsin'. 'Maximum Missed Cleavages' is set to 2. Under 'Fixed Modifications', 'Carbamidomethylation of C' is selected. Under 'Variable Modifications', 'Dissociation of M' is selected. The 'Minimum Charge' is set to 2, and the 'Maximum Charge' is set to 4. The 'Forward Ion' is set to 'b' and the 'Reverse Ion' is set to 'y'. Under 'SearchGUI Options', all engines have 'Default' selected. The 'Execute' button at the bottom is highlighted.

## 6 Protein and Peptide Evaluation with PeptideShaker

### ★ 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>

### ★ 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.

### ★ PeptideShaker in Galaxy-P ([Sect 6.1 page 18](#))

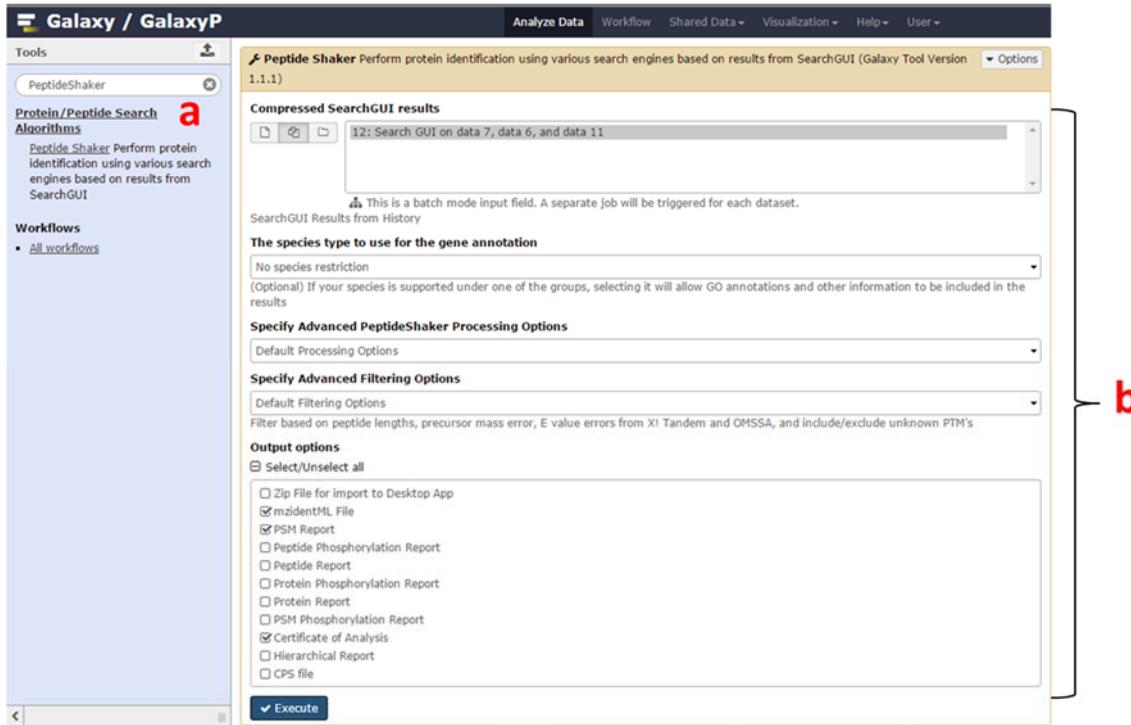
To interpret the protein/peptide identifications by SearchGUI, Galaxy-P uses a platform called PeptideShaker. It reports information about spectra and PSMs, proteins, peptides, and also provides an mzIdentML file that can be used for PSM visualization. Refer to the next section for more information about 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 \(2015\)](#).

## Protein and Peptide Evaluation with PeptideShaker

### 6.1 PeptideShaker

#### ★PeptideShaker in Galaxy-P

- In the search box under “Tools” type “PeptideShaker” and double-click on the tool
- Following the parameter settings
- Click Execute



## Protein and Peptide Evaluation with PeptideShaker

### 6.2 PeptideShaker Outputs

To the left in bold font are the possible output files from PeptideShaker. Under these outputs are different components available for that particular output option. Mzid files can be generated and exported for application in other viewing and post analysis software. On the right hand side Galaxy-P outputs are shown.

<b>Protein Report</b> <ul style="list-style-type: none"><li>- valid proteins</li><li>- coverage</li><li>- molecular weight</li></ul>	<b>7: Peptide Shaker on data 8: Protein Report</b>
<b>Peptide Report</b> <ul style="list-style-type: none"><li>- valid peptides</li><li>- potential novel proteoforms based on accession numbers</li><li>- sequences</li><li>- modifications and localization score</li><li>- confidence</li></ul>	<b>6: Peptide Shaker on data 8: Peptide Report</b>
<b>Spectrum (PSM) Report</b> <ul style="list-style-type: none"><li>- valid spectra</li><li>- potential novel proteoforms based on accession numbers</li><li>- sequences</li><li>- modifications and localization score</li><li>- confidence</li><li>- m/z, charge state, Δm/z</li></ul>	<b>5: Peptide Shaker on data 8: PSM Report</b>
<b>Summary (Parameters)</b> <ul style="list-style-type: none"><li>- valid peptides</li><li>- valid proteins</li><li>- valid spectra</li></ul>	<b>4: Peptide Shaker on data 8: Parameters</b>
<b>Archive (zipped file)</b> <ul style="list-style-type: none"><li>- CPS file to visualize data</li></ul>	<b>3: Peptide Shaker on data 8: Archive</b>
<b>Mzid</b> <ul style="list-style-type: none"><li>- PSM Visualization</li><li>- SWATH Analysis</li><li>- Skyline</li><li>- Scaffold</li></ul>	<b>2: Peptide Shaker on data 8: mzidentML_file</b>

## 7 mz to SQLite Database Generation

### ★ 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 PeptideShaker Results: sqlite Database ([Sect 7.1 page 21](#))

The **mz to sqlite** Galaxy-P 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 an mz.sqlite dataset. This is a special SQLite database schema that provides a **PSM Viewer** Galaxy-P visualization plugin for interactively analyzing the data.

### ★ Visualization with Peptide-Spectrum-Match Evaluator ([Sect 7.2 page 22](#))

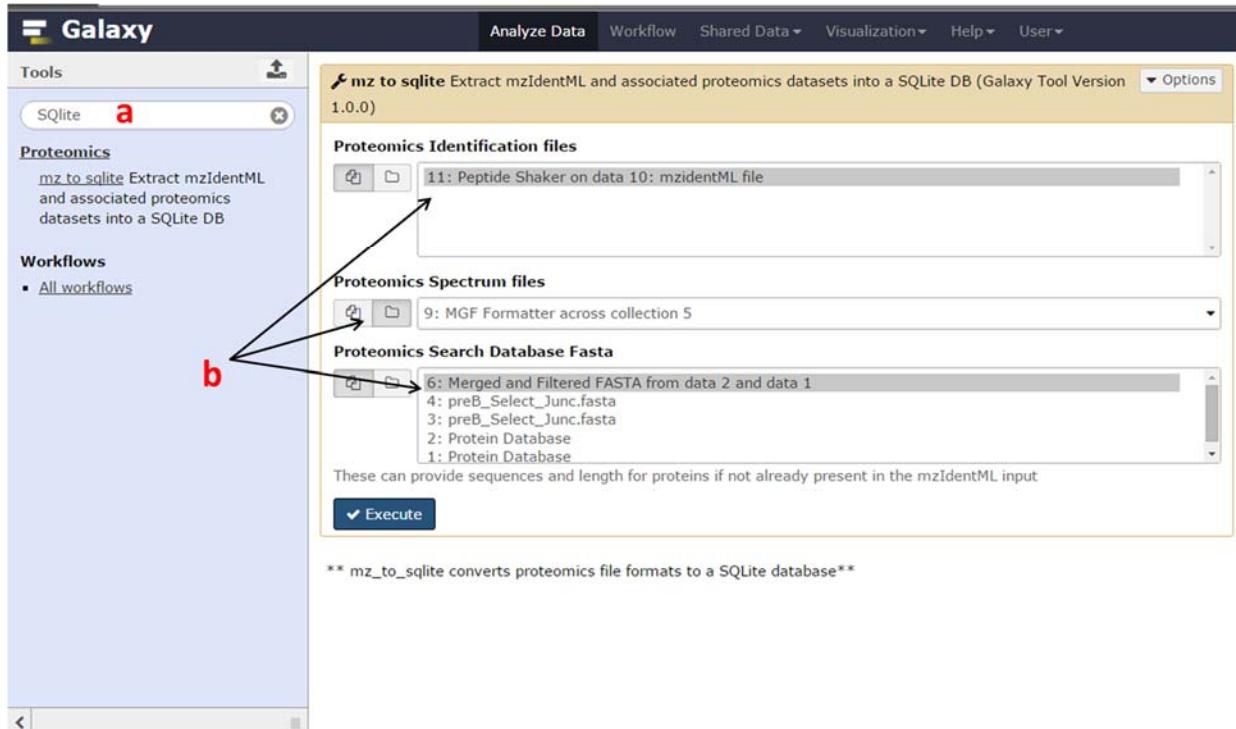
The Peptide-Spectrum-Match (PSM) evaluator tool is a unique visualization tool to Galaxy-P. Using the experimental peak lists and corresponding analyzed database search report (peptide report or mzid) PSM Evaluator will render each peptide for visual validation. PSM Evaluator can visualize fragmentation 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.

## mz to SQLite Database Generation

### 7.1 Generating the sqlite Database

#### ★Visualizing PeptideShaker Results: sqlite Database

- In the search box under “Tools” type “mz to sqlite” and double-click on the tool
- Follow the display settings below
- Click Execute



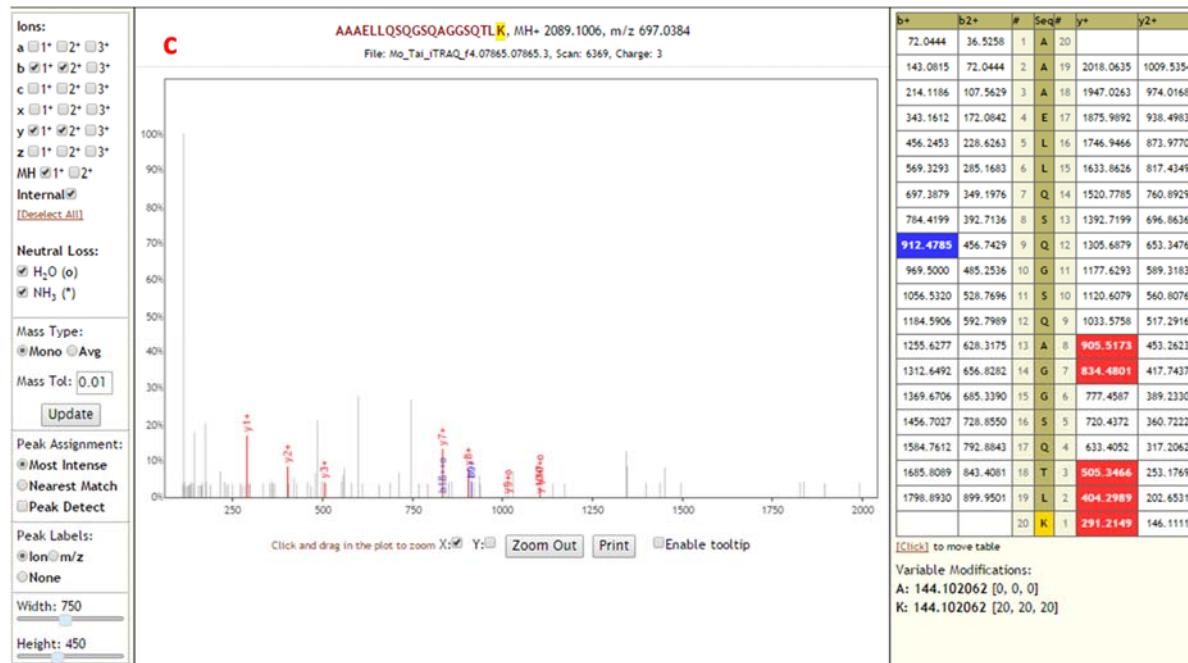
## mz to SQLite Database Generation

### 7.2 PSM Evaluator in Galaxy-P

#### ★ Visualizing Peptide with Peptide-Spectrum-Match Evaluator

“mztosqlite” tool allows for an interactive evaluation of spectrum through an PSM plug-in icon

- Click on the “mztosqlite” output in the history. Clicking on a previous ran job will create a drop-down menu for each tool with added features.
- Select the graphical bar figure icon. This will bring you to an outside browser for PSM visualization
- Peptides of interest can be selected for visualization.



## 8 Generating and Running Workflows within Galaxy

### ★ Galaxy Proteomic Workflows ([Sect 8.1 page 24](#))

For the sake of time, this workflow was performed on two reduced mzML files generated from two fractions from a biological sample from mouse. In total there are 9 fractions from this sample that could be processed collectively. The same workflow can be utilized for the proteomic database search of all nine mzML files. Galaxy workflows provided an easy method to automate this mass spectrometry-base proteomic analysis.

### ★ Workflow Parameters ([Sect 8.2 page 25](#))

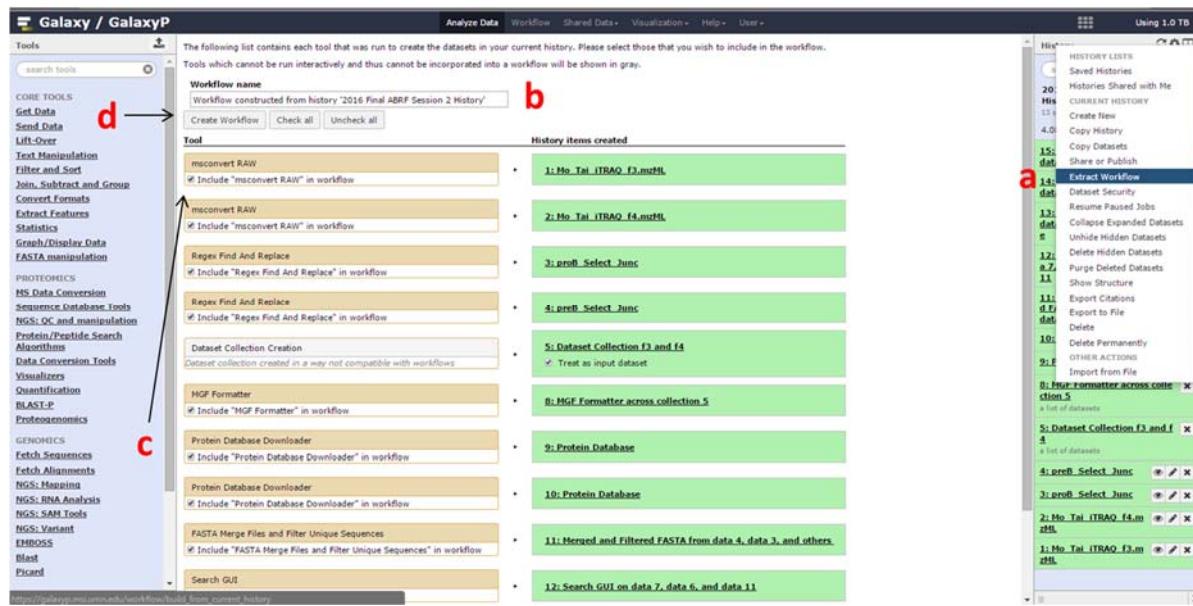
The workflow flow to be set up will run MGF Formatter, Protein Database Downloader, FASTA Merge Files and Filter Unique Sequences, SearchGUI, PeptideShaker, and mz to sqlite.

# Generating and Running Workflows within Galaxy

## 8.1 Extract Workflow from Current History

### ★ Galaxy Proteomic Workflows

- a) At the top of the history pan click on the small gear icon and select “Extract Workflow” from the pop-up menu
- b) In the “Workflow” name box enter “Session 2 Proteomics”
- c) Confirmed that all the tools are checked
- d) Click “Create Workflow” under the workflow name

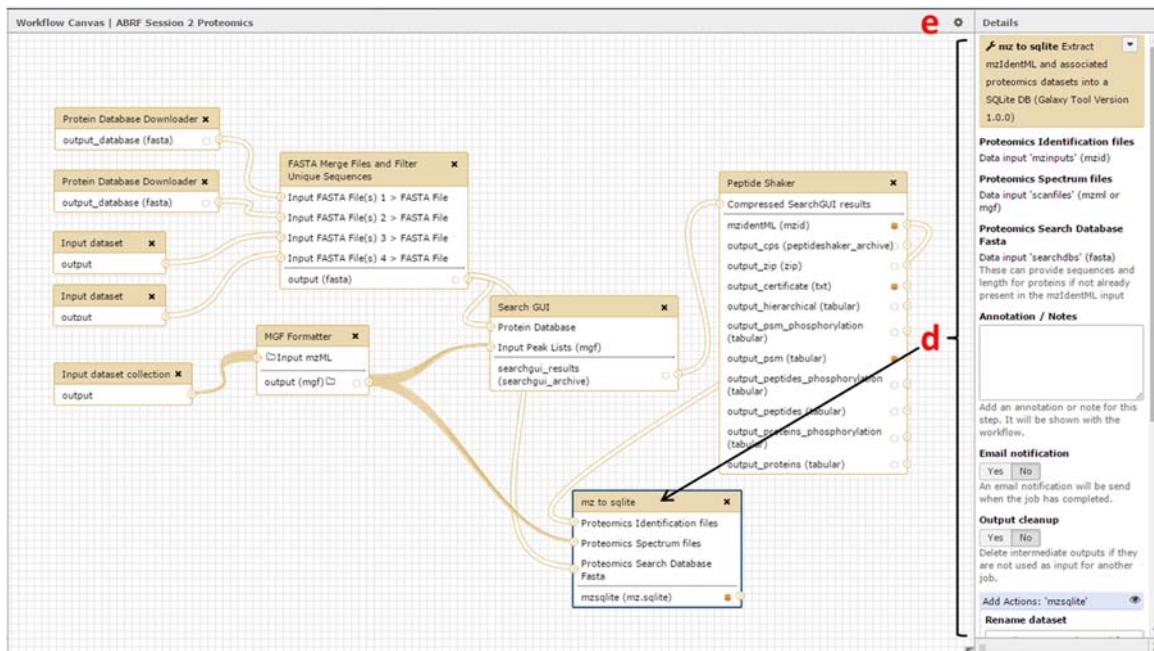


# Generating and Running Workflows within Galaxy

## 8.2 Edit the Workflow

### ★Workflow Parameters

- a) Click on “Workflow” at the top of the Galaxy window
- b) Click on the drop-down menu for the recently created workflow and select “Edit”
- c) Move the elements of the workflow around to make it easier to see how they are connected
- d) By clicking on individual tools, the right side panel will allow you to change inputs for the tool
- e) Once finished go to the mechanical wheel at the top right of the workflow canvas and click “save” from the drop-down menu.



## Generating and Running Workflows within Galaxy

### 8.3 Running Workflows

- a) First, load a “History” with the appropriate input files for the desired workflow.
- b) Click on “Workflow” at the top of the Galaxy window
- c) Click on the drop-down menu for the workflow and select “Run”
- d) The “Running workflow” screen will be prompted in the center panel. You will then need to choose the correct input files for the steps via the drop down menus.
- e) There will be the option to send the workflow to a “new history” or the workflow can be run in the current history.
- f) When ready, click “Execute Workflow.”

Analyze Data Workflow Shared Data Visualization Help User

Running workflow "ABRF Session 2 Proteomics" d

Step 1: Protein Database Downloader (version 0.2.0)

Step 2: Protein Database Downloader (version 0.2.0)

Step 3: Input dataset

proB - JUNC FASTA 3: proB\_Select\_Junc type to filter

Step 4: Input dataset

pre-proB - JUNC FASTA 4: preB\_Select\_Junc type to filter

Step 5: Input dataset collection

mzml Input Dataset Collection 8: MGF Formatter across collection 5 type to filter

Step 6: FASTA Merge Files and Filter Unique Sequences (version 1.0)

Step 7: MGF Formatter (version 0.1.0)

Step 8: Search GUI (version 2.1.1)

Step 9: Peptide Shaker (version 1.1.1)

Step 10: mz to sqlite (version 1.0.0)

Send results to a new history e

Run workflow

# Proteogenomics analysis using Galaxy framework I: *RNA-Seq data and sequence database generation*

---

*ABRF 2016 ANNUAL MEETING*  
*February 20-23*

## Introduction

1	Introduction .....	3
1.1	<i>Scope of this tutorial</i> .....	3
1.2	<i>Outline of tutorial</i> .....	3
2	Starting Galaxy .....	4
2.1	<i>Accessing Galaxy</i> .....	5
2.2	<i>Import tutorial datasets into current history</i> .....	6
3	Mapping with Tophat.....	8
3.1	<i>First Tophat run</i> .....	9
3.2	<i>Second Tophat run</i> .....	11
4	Generating protein database of novel proteoforms .....	12
4.1	<i>Identify novel splice-junctions</i> .....	13
4.2	<i>Fetch genomic sequence for novel splice junction</i> .....	14
4.3	<i>Generate peptide fasta using 3-frame translation</i> .....	15
5	Visualizing alignments with IGV.....	17
5.1	<i>Downloading IGV from the broad</i> .....	18
5.2	<i>Ensure correct genome is loaded in IGV</i> .....	20
5.3	<i>Load files into IGV</i> .....	21

## 1 Introduction

### 1.1 Scope of this tutorial

This is a practical, hands-on tutorial designed to give participants experience with RNA-Seq based proteogenomics data analysis in Galaxy. The analysis in this tutorial is typical of experiments in eukaryotic species with high-quality genomes and genome annotation available. Participants are expected to be familiar with next-generation sequence data, basic theory of RNA-Seq, and Galaxy.

### Reference materials:

RNA-Seq Lecture PDFs on MSI website: [www.msi.umn.edu/content/bioinformatics-analysis](http://www.msi.umn.edu/content/bioinformatics-analysis)

Galaxy 101: NGS data analysis hands-on tutorial:

[www.msi.umn.edu/content/bioinformatics-analysis](http://www.msi.umn.edu/content/bioinformatics-analysis)

Tophat manual: [ccb.jhu.edu/software/tophat/manual.shtml](http://ccb.jhu.edu/software/tophat/manual.shtml)

### 1.2 Outline of tutorial

- 1 Introduction
- 2 Starting Galaxy
- 3 Mapping with Tophat
- 4 Generating protein database of novel proteoforms
- 5 Visualizing alignments with IGV

## 2 Starting Galaxy

### ★Tutorial Dataset

This tutorial will identify novel splice junctions. The sample dataset used in this tutorial was created from a combination of published datasets ([\[J Proteomics Bioinform. 2014 Feb 17;7\]](#)) and unpublished data from Dr. Reddy's lab courtesy of Mohammad Heydarian).

*NOTE: This dataset was chosen to allow fast processing and response times in a workshop setting. A typically analysis would usually take a lot longer and with significantly more data points to look at.*

### ★GTF Files

A GTF file identifies the genomic locations of genes and their exons. GTF file for most organisms can be found online at sites such as [ccb.jhu.edu/software/tophat/igenomes.shtml](#), [www.ensembl.org/info/data/ftp/index.html](#), [genome.ucsc.edu/cgi-bin/hgTables?command=start](#), or NCBI.

### ★Quality Control

It is important to always verify the integrity of your dataset before you begin analysis. Quantifying dataset quality may uncover problems that might otherwise go undetected. Data quality problems such as sequencing adaptor contamination or low read quality require trimming and filtering not covered in this tutorial. For more information on dealing with data quality problems see the main Galaxy tutorial site ([https://wiki.galaxyproject.org/Learn](#) )

## Starting Galaxy

### 2.1 Accessing Galaxy

- Open a web browser and navigate to GalaxyP website [galaxy.msi.umn.edu](http://galaxy.msi.umn.edu)
- Log in with your username and password
- On the right pane, Click on “Unnamed history” and change the history name to something more descriptive e.g., “proteogenomics-tutorial” and press enter

The screenshot shows the GalaxyP web interface with three main panes:

- Tools pane (left):** A sidebar containing a search bar and lists of tools categorized under CORE TOOLS, PROTEOMICS, and other sections like Sequence Database Tools and Protein/Peptide Search Algorithms.
- Center pane (middle):** A green header with a checkmark icon and the text "Welcome to GalaxyP". Below it is a detailed description of GalaxyP's purpose and a note about its public nature. To the right is a "Updates" section with a GalaxyP logo and two entries: "Sept. 11, 2015 ProteinPilot is available." and "February 13, 2015 All Galaxy Tools running normally".
- History pane (right):** A panel titled "History" showing a single entry named "Unnamed history" with a red circle around it. A tooltip message says: "This history is empty. You can load your own data or get data from an external source".

Tools pane

Center pane

History pane

# Starting Galaxy

## 2.2 Import tutorial datasets into current history

### ★Tutorial Dataset

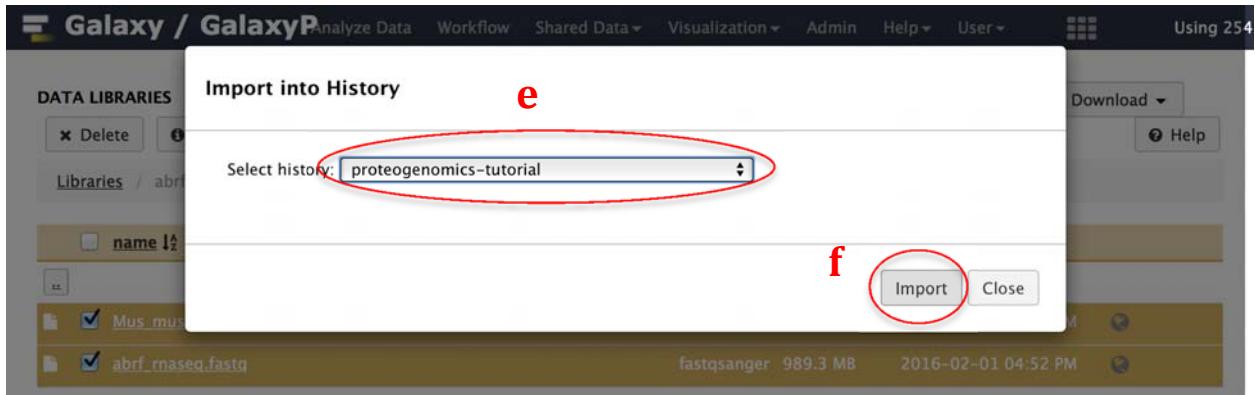
- At the top of the screen select “Shared Data -> Data Libraries”
- Select “abrf\_tutorial\_2016” from the list of data libraries
- Select the files “Mus\_musculus.GRCm38.74.gtf” and “abrf\_rnaseq.fastq”
- Near the top of the page click “to History”

The figure consists of three vertically stacked screenshots of the Galaxy web interface.

- Screenshot a:** Shows the main Galaxy dashboard. A red circle highlights the "Shared Data" dropdown menu in the top navigation bar. A red arrow points to the "Data Libraries" option in the dropdown menu, which is currently selected. The "Data Libraries" menu also includes options for "Published Histories", "Published Workflows", "Published Visualizations", and "Published Pages".
- Screenshot b:** Shows the "Data Libraries" page for the "abrf\_tutorial\_2016" library. A red circle highlights the library name "abrf\_tutorial\_2016" in the top navigation bar. A red arrow points to the "Import selected datasets into history" button in the top right corner of the page.
- Screenshot c:** Shows the list of datasets in the "abrf\_tutorial\_2016" library. A red circle highlights the two checked datasets: "Mus\_musculus.GRCm38.74.gtf" and "abrf\_rnaseq.fastq". A red arrow points to the "to History" button in the top right corner of the list table.

## Starting Galaxy

- e) Select your working history (name you changed your history to) from the drop down menu
- f) Click Import
- g) Click “Analyze Data”



The screenshot shows the Galaxy web interface with the title bar "Galaxy / GalaxyP". The "Analyze Data" tab is highlighted with a red circle and labeled 'g'. Below the tabs, there is a library list titled "Libraries / abrf\_tutorial\_2016". The list contains two items: "Mus\_musculus.GRCm38.74.gtf" and "abrf\_rnaseq.fastq", both of which have checkboxes next to them. The "Mus\_musculus.GRCm38.74.gtf" item has its checkbox checked. The table columns are "name", "data type", "size", and "time updated (UTC)". The "data type" column shows "gtf" for the first item and "fastqsanger" for the second. The "size" column shows "261.0 MB" for the first item and "989.3 MB" for the second. The "time updated (UTC)" column shows "2016-02-03 02:09 PM" for the first item and "2016-02-01 04:52 PM" for the second. The background shows the main Galaxy interface with a toolbar.

### 3 Mapping with Tophat

#### ★ Reference Genomes

It is important the reference genome you align against is generated from the same reference genome as the GTF you are using. Chromosome names and coordinates used in the GTF file must be the same as those used in the database. If your genome build is mm10, make sure your GTF file is from the same build.

#### ★ Mapping Statistics

It is important to determine how well the RNA-Seq reads align to the reference genome. Low mapping rates require further investigation to determine the cause.

### 3.1 First Tophat run

- Load the Tophat tool from the tool pane: "Tophat for Illumina"
- RNA-Seq FASTQ file -> abrf\_rnaseq.fastq
- Select a reference genome -> Mouse Ensembl GRCm38 (GRCm38\_canon GATK mm10)
- Is this library mate-paired -> Single-end
- Default settings -> Full parameter list

**a** TopHat for Illumina Find splice junctions using RNA-seq data

**b** 2: abrf\_rnaseq.fastq

**c** Mouse Ensembl GRCm38 (GRCm38\_canon GATK mm10)

**d** Single-end

**e** Full parameter list

## Mapping with Tophat

- f) Scroll down, "Use Own Junctions -> Yes"
- g) Use Gene Annotation Model -> Yes
- h) Gene Model Annotations -> Mus\_musculus.GRC.m38.74.gtf
- i) Only look for supplied junctions -> Yes
- j) Use Coverage Search -> No
- k) Click "Execute" to submit the job

The screenshot shows the configuration interface for the Tophat tool. Various options are set and annotated:

- Use Own Junctions:** Set to **Yes** (circled in red, labeled **f**).
- Use Gene Annotation Model:** Set to **Yes** (circled in red, labeled **g**).
- Gene Model Annotations:** Set to **1: Mus\_musculus.GRCm38.74.gtf** (circled in red, labeled **h**). A tooltip below explains: "TopHat will use the exon records in this file to build a set of known splice junctions for each gene, and will attempt to align reads to these junctions even if they would not normally be covered by the initial mapping."
- Use Raw Junctions:** Set to **No**.
- Only look for supplied junctions:** Set to **Yes** (circled in red, labeled **i**).
- Use Closure Search:** Set to **No**.
- Use Coverage Search:** Set to **No** (circled in red, labeled **j**).
- Use Microexon Search:** Set to **No**. A note below states: "With this option, the pipeline will attempt to find alignments incident to microexons. Works only for reads 50bp or longer."
- Execute:** A blue button with a checkmark icon, circled in red and labeled **k**.

## 3.2 Second Tophat run

- Click on the name of any one of the Tophat output files in the history pane to expand it, and click on the circular arrow icon to display the Tophat tool in the central pane with the parameters preset from the last Tophat run
- Scroll down until you get to “Only look for supplied junctions” and select “No”
- Click “Execute” to submit the second Tophat run

The screenshot shows the Galaxy web interface. The top navigation bar includes "Galaxy / GalaxyP", "Analyze Data", "Workflow", "Shared Data", "Visualization", "Admin", "Help", "User", and "Using 254.6 GB". The left sidebar under "Tools" lists various bioinformatics tools categorized by domain (Send Data, Text Manipulation, Filter and Sort, etc.). The main area displays the history pane with a green message box indicating a successful job addition:

**History**

1 job has been successfully added to the queue - resulting in the following datasets:

- 3: TopHat for Illumina on data 1 and data 2: insertions
- 4: TopHat for Illumina on data 1 and data 2: deletions
- 5: TopHat for Illumina on data 1 and data 2: splice Junctions
- 6: TopHat for Illumina on data 1 and data 2: accepted\_hits

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

The history pane also shows several other jobs listed with their status (e.g., accepted, splice junctions, deletions, insertions) and a note that job 2 is currently running. A red circle labeled 'a' highlights the status of job 2.

The screenshot shows the configuration form for the Tophat tool. The form consists of several dropdown menus and descriptive text fields:

- Use Raw Junctions:** Set to "No".
- Only look for supplied junctions:** Set to "No" (circled in red). A red circle labeled 'b' is placed over this field.
- Use Closure Search:** Set to "No".
- Use Coverage Search:** Set to "No".
- Use Microexon Search:** Set to "No".
- Description:** Text area stating "With this option, the pipeline will attempt to find alignments incident to microexons. Works only for reads 50bp or longer."
- Execute:** A blue button with a checkmark icon labeled "Execute" (circled in red). A red circle labeled 'c' is placed next to the button.

## 4 Generating protein database of novel proteoforms

The first Tophat run will output known splice junctions (based on the GTF file we supplied) for which there is supporting RNA-Seq evidence in our dataset. The second Tophat run will search for novel splice junctions in addition to known splice junctions. In this section, we will compare results from these two Tophat runs to identify novel splice junctions.

## 4.1 Identify novel splice-junctions

- Find the tool “Filter BED on splice junctions” and click on the name
- Bed file -> “9 TopHat for Illumina on data 1 and data 2: splice junctions”
- reference bed file -> “5 TopHat for Illumina on data 1 and data 2: splice junctions”
- Extend the start position -> 66
- Extend the end position -> 66
- Click “Execute” to submit the job

**a** Galaxy / GalaxyP Analyze Data Workflow Shared Data Visualization Admin Help User Using 254.6 GB

Tools JUNCTIONS USING RNA-SEQ DATA Filter BED on splice junctions that are not in a reference bed file (Galaxy Tool Version 0.0.1)

**b** Filter BED on splice junctions that are not in a reference bed file (Galaxy Tool Version 0.0.1)

**c** BED file 9: TopHat for Illumina on data 1 and data 2: splice junctions e.g. tophat junctions.bed run without GTF option or no -novel-junctions

**c** reference bed file 5: TopHat for Illumina on data 1 and data 2: splice junctions e.g. tophat junctions.bed run with GTF option and no -novel-junctions

**d** Extend the start position 66

The number of base pairs to extend the start of the exon before the junction position

**e** Extend the end position 66

The number of base pairs to extend the end of the exon after the junction position

**f** Execute

## 4.2 Fetch genomic sequence for novel splice junction

- Find the tool “Extract Genomic DNA” and click on the name
- Fetch sequences for interval in -> “11 Filter BED on splice junctions on data 5 ....”
- Interpret features when possible -> No
- Output data type -> Interval
- Click “Execute” to submit the job

**a** Galaxy / GalaxyP

Analyze Data Workflow Shared Data Visualization Admin Help User Using 254.6 GB

Tools

**Fetch Sequences**

Extract Genomic DNA using coordinates from assembled/unassembled genomes

1 job has been successfully added to the queue – resulting in the following datasets:

11: Filter BED on splice junctions on data 5 and data 9

You can check the status of queued jobs and view the resulting data.

History

search datasets

proteogenomics-tutorial  
11 shown  
115.72 KB

**b**

**c**

**d**

**e**

**Extract Genomic DNA using coordinates from assembled/unassembled genomes (Galaxy Tool Version 2.2.3)**

**Fetch sequences for intervals in**

11: Filter BED on splice junctions on data 5 an...

**Interpret features when possible**

No

Only meaningful for GFF, GTF datasets.

**Source for Genomic Data**

Locally cached

If 'Locally cached' is selected, it will use a genomic reference file that matches the input file's dbkey. First it looks whether there are corresponding \*.nib files in alignseq.loc. If that is not available, it searches for a corresponding \*.2bit in twobit.loc.

**Output data type**

Interval

Execute

### 4.3 Generate peptide fasta using 3-frame translation

NOTE: When Tophat identifies splice junctions, it indicates frame of translation. A 3-frame translation is thus sufficient to identify all potential novel-proteoforms supported by our RNA-Seq data.

- Find the tool “Translate BED sequences” and click on the name
- BED file with added sequence column -> “12: Extract Genomic DNA....”
- Fasta ID source, e.g. generic -> generic
- ID prefix for generated source -> abrf\_

- e) Genome reference name -> GRCh38
- f) The SEQTYPE:STATUS to include in the fasta ID lines -> pep:splice
- g) Add the bed score field fasta ID line with this tag name -> depth
- h) Stop codon filtering start position base pairs-> 66
- i) Stop codon filtering end position base pairs -> 66
- j) Minimum length of a translation to be reported -> 10
- k) Click "Execute" to submit the job

**Genome reference name**

GRCh38 **e**

By default, the database metadata will be used.

**The SEQTYPE:STATUS to include in the fasta ID lines**

pep:splice **f**

For example: pep:splice

**Add the bed score field fasta ID line with this tag name**

depth **g**

For example: with the tag name 'depth' and bed score 12: depth:12

**Filter out translations with stop codons before the splice site**

Yes

**Stop codon filtering start position base pairs**

66 **h**

Do not reject translation if stop\_codons are within base pairs of the BED start position for positive strand

**Stop codon filtering end position base pairs**

66 **i**

Do not reject translation if stop\_codons are within base pairs of the BED end position for negative strand

**Trim translations to stop codons**

Yes

**Minimum length of a translation to be reported**

10 **j**

**Execute** **k**

## 5 Visualizing alignments with IGV

### ★Visualization

Visualizing alignments is a quick and easy way to check for major problems with data. You may wish to verify that RNA-Seq reads do indeed overlap potential novel proteoforms identified by your workflow. Depth of read coverage across splice junction is a good indicator of high quality novel proteoforms. A splice junction with over 30 RNA-Seq reads supporting the junction is better relative to one with a few RNA-Seq reads supporting the splice junction.

### ★Galaxy Visualization Options

Galaxy supports three genome browsers for visualizing data:

- a) The Integrative Genomics Viewer (IGV) is a popular visualization tool for genomic datasets. It is fast, powerful, and easy to use. We will be using IGV in this tutorial
- b) Trackster is a genome browser built into Galaxy. Any data file that can be viewed in Trackster will have a Trackster icon  next to it in the history pane.
- c) The Integrated Genome Browser (IGB) is similar to IGV, but most users prefer to use IGV.

### ★IGV

Developed at the Broad Institute, IGV is a high-performance visualization tool for interactive exploration of large, integrated genomics datasets. We will download IGV and use it to visually inspect quality of reads supporting our novel splice junctions.

## Visualizing alignments with IGV

### 5.1 Downloading IGV from the broad

- a) Navigate to <https://www.broadinstitute.org/software/igv/home>
- b) Under **Downloads** click *Register*
- c) Fill in the registration form and click *Agree*

**a**

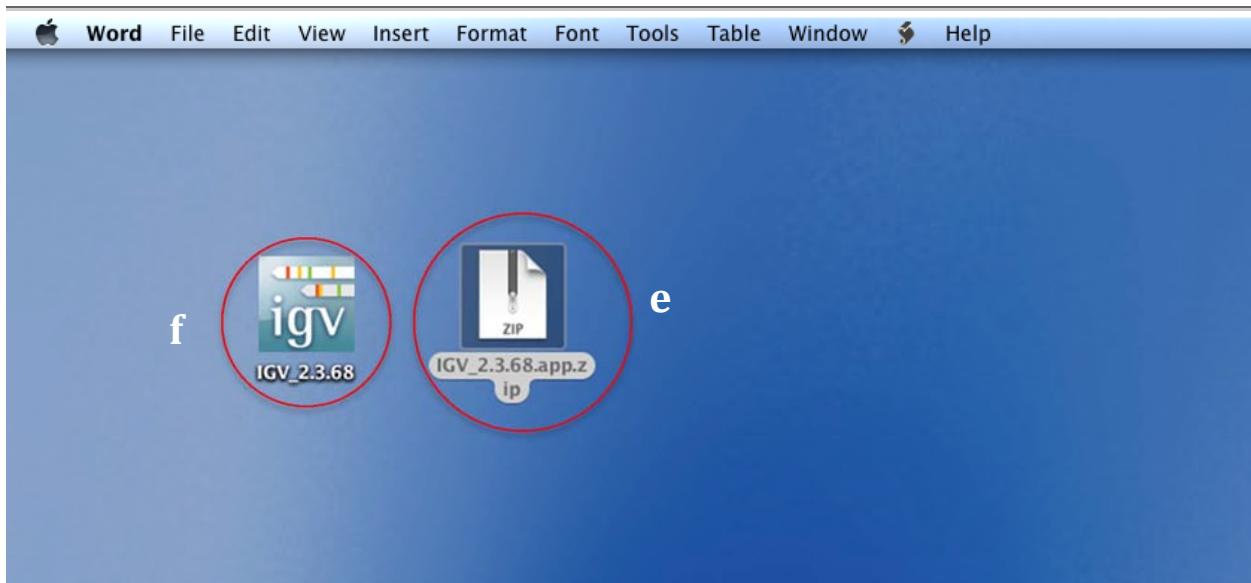
**b**

**c**

## Visualizing alignments with IGV

- d) Find the appropriate download link for your system and click download. These tutorial covers instructions for a Mac.
- e) Navigate to where the application was downloaded and double click the zipped archived
- f) Double click the IGV application to run it.

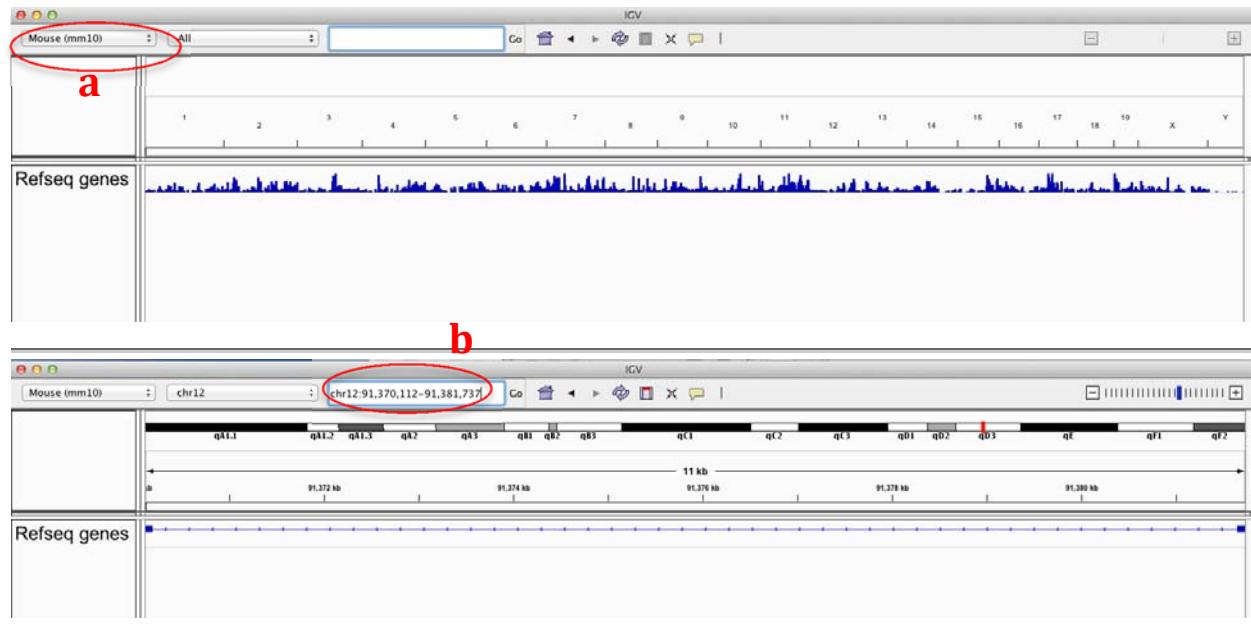
The screenshot shows the IGV website's 'Downloads' page. On the left, there's a sidebar with links like Home, Downloads (which is selected), Documents, and Contact. The main content area is titled 'Install IGV' and contains instructions for Mac users. It says: 'Download and unzip the Mac App archive, then double-click the IGV application to run it. The application can be moved to the "Applications" folder, or anywhere else.' A red circle labeled 'd' highlights the 'Download Mac App' button. Below this, there's a section for Windows users with a 'Download Windows Package' button.



## Visualizing alignments with IGV

### 5.2 Ensure correct genome is loaded in IGV

- a) Once the application loads, ensure the correct genome build is loaded (mm10)
- b) Enter “chr12:91,370,112-91,381,737” into the region field and press “Enter”. This will zoom IGV into this region. This is one of the regions we will be examining. It is a good illustration of reads supporting a splice junction. Zooming into the region makes loading BAM files faster. IGV only loads reads for the region in view

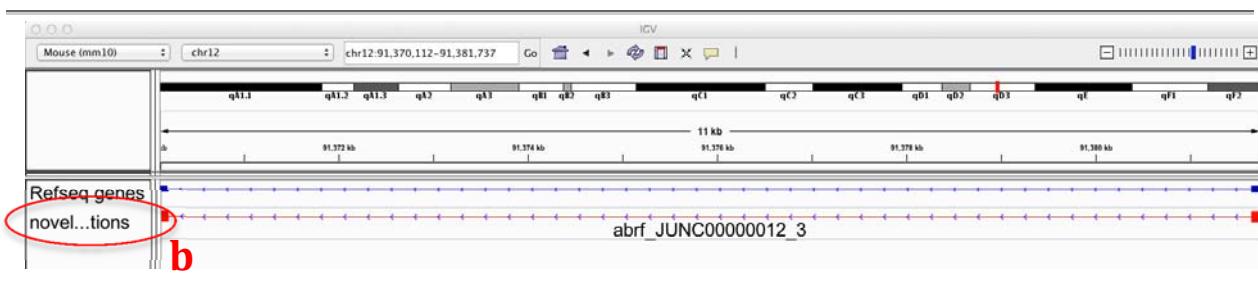


# Visualizing alignments with IGV

## 5.3 Load files into IGV

- a) Click on the “**...Translate BED Sequences on data...**” file in the history pane to expand it and click the link “local” next to “display with IGV web current”
  - b) You should now see a track in IGV named “novel\_junction\_translations”

Analyze Data		Workflow		Shared Data		Visualization		Admin		Help		User	
1	2	3	4	5	6	7	8	9	10	11	12	13	14
		track name="novel_junctioni_translations" description="test"											
X	13041959	13043605	abrf_JUNC00000001_1	7	+	13041959	1						
X	48694976	48695328	abrf_JUNC00000002_2	14	+	48694976	4						
X	48695281	48697453	abrf_JUNC00000003_3	68	+	48695281	4						
X	48695354	48697418	abrf_JUNC00000004_1	5	+	48695354	4						
X	48695358	48697413	abrf_JUNC00000004_2	5	+	48695358	4						
X	48695317	48697453	abrf_JUNC00000004_3	5	+	48695317	4						
X	48695628	48697453	abrf_JUNC00000005_1	2	+	48695628	4						
X	48695578	48697418	abrf_JUNC00000005_2	2	+	48695578	4						
X	48695606	48697413	abrf_JUNC00000005_3	2	+	48695606	4						
X	48695628	48697418	abrf_JUNC00000006_1	8	+	48695628	4						
X	48695606	48697453	abrf_JUNC00000006_3	8	+	48695606	4						
X	48697360	48706437	abrf_JUNC00000007_3	28	+	48697360	4						
X	157591529	157593797	abrf_JUNC00000008_1	11	-	157591529	15						
X	157591459	157593775	abrf_JUNC00000008_2	11	-	157591459	15						
X	157591428	157593837	abrf_JUNC00000008_3	11	-	157591428	15						
X	157591529	157598616	abrf_JUNC00000009_1	15	-	157591529	15						
X	157591459	157598615	abrf_JUNC00000009_2	15	-	157591459	15						
X	157593720	157598463	abrf_JUNC00000010_1	11	-	157593720	15						
X	157593859	157598461	abrf_JUNC00000010_3	11	-	157593859	15						
12	91370112	91381737	abrf_JUNC00000012_3	66	-	91370112	9						



## Visualizing alignments with IGV

- c) Back in GalaxyP, click on the output from your second Tophat run with “accepted\_hits” as part of the name (**10 TopHat for Illumina on data 1 and data 2: accepted hits**) to expand it.
- d) Click the link “local” next to “display with IGV web current”

Analyze Data Workflow Shared Data Visualization Admin Help User Using 255.3 GB

1	2	3	4	5	6	7	8
X	13041959	13043605	abrf_JUNC00000001_1	7	+	13041959	
X	48694976	48695328	abrf_JUNC00000002_2	14	+	48694976	
X	48695281	48697453	abrf_JUNC00000003_3	68	+	48695281	
X	48695354	48697418	abrf_JUNC00000004_1	5	+	48695354	
X	48695358	48697413	abrf_JUNC00000004_2	5	+	48695358	
X	48695317	48697453	abrf_JUNC00000004_3	5	+	48695317	
X	48695628	48697453	abrf_JUNC00000005_1	2	+	48695628	
X	48695578	48697418	abrf_JUNC00000005_2	2	+	48695578	
X	48695606	48697413	abrf_JUNC00000005_3	2	+	48695606	
X	48695628	48697418	abrf_JUNC00000006_1	8	+	48695628	
X	48695606	48697453	abrf_JUNC00000006_3	8	+	48695606	
X	48697360	48706437	abrf_JUNC00000007_3	28	+	48697360	
X	157591529	157593797	abrf_JUNC00000008_1	11	-	157591529	
X	157591459	157593775	abrf_JUNC00000008_2	11	-	157591459	
X	157591428	157593837	abrf_JUNC00000008_3	11	-	157591428	
X	157591529	157598616	abrf_JUNC00000009_1	15	-	157591529	
X	157591459	157598615	abrf_JUNC00000009_2	15	-	157591459	
X	157593720	157598463	abrf_JUNC00000010_1	11	-	157593720	
X	157593859	157598461	abrf_JUNC00000010_3	11	-	157593859	
12	91370112	91381737	abrf_JUNC00000012_3	66	-	91370112	

History

```
track name="novel_junctioni_translations" description="test"
X 13041959 13043605 abrf_JUNC00000001_1 7 + 13041959
X 48694976 48695328 abrf_JUNC00000002_2 14 + 48694976
X 48695281 48697453 abrf_JUNC00000003_3 68 + 48695281
X 48695354 48697418 abrf_JUNC00000004_1 5 + 48695354
X 48695358 48697413 abrf_JUNC00000004_2 5 + 48695358
X 48695317 48697453 abrf_JUNC00000004_3 5 + 48695317
X 48695628 48697453 abrf_JUNC00000005_1 2 + 48695628
X 48695578 48697418 abrf_JUNC00000005_2 2 + 48695578
X 48695606 48697413 abrf_JUNC00000005_3 2 + 48695606
X 48695628 48697418 abrf_JUNC00000006_1 8 + 48695628
X 48695606 48697453 abrf_JUNC00000006_3 8 + 48695606
X 48697360 48706437 abrf_JUNC00000007_3 28 + 48697360
X 157591529 157593797 abrf_JUNC00000008_1 11 - 157591529
X 157591459 157593775 abrf_JUNC00000008_2 11 - 157591459
X 157591428 157593837 abrf_JUNC00000008_3 11 - 157591428
X 157591529 157598616 abrf_JUNC00000009_1 15 - 157591529
X 157591459 157598615 abrf_JUNC00000009_2 15 - 157591459
X 157593720 157598463 abrf_JUNC00000010_1 11 - 157593720
X 157593859 157598461 abrf_JUNC00000010_3 11 - 157593859
X 91370112 91381737 abrf_JUNC00000012_3 66 - 91370112
```

12: Extract Genomic DNA on data 11  
11: Filter BED on splice junctions on data 9 and data 5 C  
10: TopHat for Illumina on data 1 and data 2: accepted\_hits  
9: TopHat for Illumina on data 1 and data 2: splice junctions  
8: TopHat for Illumina on data 1 and data 2: deletions  
7: TopHat for Illumina on data 1 and data 2: insertions  
6: TopHat for Illumina on data 1 and data 2: accepted\_hits  
5: TopHat for Illumina on data 1 and data 2: splice junctions  
4: TopHat for Illumina on data 1 and data 2: deletions  
3: TopHat for Illumina on data 1 and data 2: insertions  
2: tutorial\_rnaseq.fasta  
1: Mus\_musculus.GRCm38.74.abrf.gtf

Analyze Data Workflow Shared Data Visualization Admin Help User Using 70.7 KB format: bam, database: GRCm38\_canon

1	2	3	4	5	6	7	8
X	13041959	13043605	abrf_JUNC00000001_1	7	+	13041959	
X	48694976	48695328	abrf_JUNC00000002_2	14	+	48694976	
X	48695281	48697453	abrf_JUNC00000003_3	68	+	48695281	
X	48695354	48697418	abrf_JUNC00000004_1	5	+	48695354	
X	48695358	48697413	abrf_JUNC00000004_2	5	+	48695358	
X	48695317	48697453	abrf_JUNC00000004_3	5	+	48695317	
X	48695628	48697453	abrf_JUNC00000005_1	2	+	48695628	
X	48695578	48697418	abrf_JUNC00000005_2	2	+	48695578	
X	48695606	48697413	abrf_JUNC00000005_3	2	+	48695606	
X	48695628	48697453	abrf_JUNC00000006_1	8	+	48695628	
X	48695606	48697453	abrf_JUNC00000006_3	8	+	48695606	
X	48697360	48706437	abrf_JUNC00000007_3	28	+	48697360	
X	157591529	157593797	abrf_JUNC00000008_1	11	-	157591529	
X	157591459	157593775	abrf_JUNC00000008_2	11	-	157591459	
X	157591428	157593837	abrf_JUNC00000008_3	11	-	157591428	
X	157591529	157598616	abrf_JUNC00000009_1	15	-	157591529	
X	157591459	157598615	abrf_JUNC00000009_2	15	-	157591459	
X	157593720	157598463	abrf_JUNC00000010_1	11	-	157593720	
X	157593859	157598461	abrf_JUNC00000010_3	11	-	157593859	
12	91370112	91381737	abrf_JUNC00000012_3	66	-	91370112	

History

```
TopHat v1.4.0
tophat -p 4 -a 8 -m 0 -i 70 -l 500000 -g 20 -G
/home/mscls/apps/galaxy2014/files/000/121/dataset_121215.dat --no-novel-juncs --library-type fr-unstranded --max-insertion-length 3 --max-deletion-length 3 --no-coverage-search --no-closure-s
```

12: Extract Genomic DNA on data 11  
11: Filter BED on splice junctions on data 9 and data 5  
10: TopHat for Illumina on data 1 and data 2: accepted\_hits  
70.7 KB format: bam, database: GRCm38\_canon  
d  
display with IGV web current local  
display in IGB View  
Binary bam alignments file  
9: TopHat for Illumina on data 1 and data 2: splice junctions  
8: TopHat for Illumina on data 1 and data 2: deletions

## Visualizing alignments with IGV

- e) The BAM file with rna-seq reads should now be loaded in IGV
- f) Go back to GalaxyP again, click on the output from your first Tophat run with “accepted\_hits” as part of the name to expand it and click the link “local”

The screenshot shows the Galaxy Platform interface with two main panels:

**Top Panel (IGV View):**

- Panel title: IGV
- Genomic track: chr12 (chr12:91,368,970-91,385,274)
- Scale: 16 kb
- Genomic tracks: q41.3, q41.2, q41.3, q42, q43, q41, q42, q41, q42, q43, q41, q42, q41, q42, q43, q41, q42.
- Coordinates: 91,370 Kb, 91,372 Kb, 91,374 Kb, 91,376 Kb, 91,378 Kb, 91,380 Kb, 91,382 Kb, 91,384 Kb.
- Annotations: TopHat for Illumina on data 1 and data 2: accepted\_hits.
- Bottom panel: Refseq genes novel...tions abrf\_JUNC00000012\_3.
- A red arrow labeled 'e' points to the 'TopHat for Illumina on data 1 and data 2: accepted\_hits' annotation.

**Bottom Panel (Workflow History):**

- Panel title: Using 255.3 GB
- Table header: Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Table rows (selected row highlighted):
 

1	2	3	4	5	6	7	8
track name="novel_junctions_translations" description="test"							History
X 13041959	13043605	abrf_JUNC00000001_1	7	+	13041959		9: TopHat for Illumina on data 1 and data 2: splice junctions
X 48694976	48695328	abrf_JUNC00000002_2	14	+	48694976		8: TopHat for Illumina on data 1 and data 2: deletions
X 48695281	48697453	abrf_JUNC00000003_3	68	+	48695281		7: TopHat for Illumina on data 1 and data 2: insertions
X 48695354	48697418	abrf_JUNC00000004_1	5	+	48695354		6: TopHat for Illumina on data 1 and data 2: accepted_hits
X 48695358	48697413	abrf_JUNC00000004_2	5	+	48695358		85.9 KB
X 48695317	48697453	abrf_JUNC00000004_3	5	+	48695317		format: bam, database: GRCm38_canon
X 48695628	48697453	abrf_JUNC00000005_1	2	+	48695628		TopHat v1.4.0
X 48695578	48697418	abrf_JUNC00000005_2	2	+	48695578		tophat -p 4 -a 8 -m 0 -i 70 -l 500000 -g 20 -G
X 48695606	48697413	abrf_JUNC00000005_3	2	+	48695606		/home/mscls/apps/galaxyP2014/files/000/121/dataset_121215.dat --library-type
X 48695628	48697418	abrf_JUNC00000006_1	8	+	48695628		fr-unstranded --max-insertion-length 3 --max-deletion-length 3 --no-coverage-
X 48695606	48697453	abrf_JUNC00000006_3	8	+	48695606		search --no-closure-search --initial-r
X 48697360	48706437	abrf_JUNC00000007_3	28	+	48697360		
X 157591529	157593797	abrf_JUNC00000008_1	11	-	157591529		
X 157591459	157593775	abrf_JUNC00000008_2	11	-	157591459		
X 157591428	157593837	abrf_JUNC00000008_3	11	-	157591428		
X 157591529	157598616	abrf_JUNC00000009_1	15	-	157591529		
X 157591459	157598615	abrf_JUNC00000009_2	15	-	157591459		
X 157593720	157598463	abrf_JUNC00000010_1	11	-	157593720		
X 157593859	157598461	abrf_JUNC00000010_3	11	-	157593859		
- Annotation: display with IGV web current local display in IGB View (circled with a red circle).
- Annotation: Binary bam alignments file
- Annotation: 5: TopHat for Illumina on data 1 and data 2: splice junctions

## Visualizing alignments with IGV

- g) Click on the GTF file to expand it and then click on the link “local” to also load this GTF file into IGV
- h) Zoom out a little bit to ensure your novel splice junction is in view. Does it make sense why this splice-junction was marked as novel? Hint: Look at the difference between our GTF file and RefSeq genes. Here are a few other regions to explore (X:157593659-157598661, X:13041888-13045518, X:48694458-48698245)

The screenshot shows the IGV (Integrating Genomics Viewer) interface. At the top, there is a menu bar with options like Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User. A status bar indicates "Using 255.3 GB".

**Left Panel (Data View):**

- Header: Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Panel 1: GTF file content (novel\_junction\_translations.gtf). It lists various transcript entries with columns 1-12. An example entry is: X 13041959 13043605 abrf\_JUNC00000001\_1 7 + 13041959.
- Panel 8: History section. It shows several items:
  - display with IGV web current local
  - display in IGB View
  - binary bam alignments file
  - 5: TopHat for Illumina on data 1 and data 2: splice junctions
  - 4: TopHat for Illumina on data 1 and data 2: deletions
  - 3: TopHat for Illumina on data 1 and data 2: insertions
  - 2: tutorial\_rnaseq.fastq
  - 1: Mus\_musculus.GRCm38.74.abrf.gtf
- Bottom of the panel: 27 lines, format: gtf, database: GRCm38\_canon, uploaded gtf file.

**Right Panel (Genomic View):**

- Header: IGV, Mouse (mm10), chr12, chr12-91,368,005-91,385,745, Go, zoom controls.
- Panel 1: TopHat for Illumina on data 1 and data 2: accepted\_hits. It shows a genomic track with a red arrow pointing to a specific position labeled 'h'.
- Panel 2: TopHat for Illumina on data 1 and data 2: accepted\_hits. It shows a genomic track with a red arrow pointing to a specific position labeled 'g'.
- Panel 3: Refseq genes, novel..tions, Mus\_m...f.gtf. It shows a genomic track with a red arrow pointing to a specific position labeled 'g'.
- Bottom right: A small box displays sequencing quality scores: Total count: 0, A: 0, C: 0, G: 0, T: 0, N: 0.
- Bottom right: Gene information: ENSMUST00000129111, ENSMUST00000141429.

# Proteogenomics analysis using Galaxy framework II: *Search results filtering and visualization*

---

*ABRF 2016 ANNUAL MEETING*  
*February 20-23*

## Introduction

1	Introduction .....	3
1.1	<i>Scope of this tutorial</i> .....	3
1.2	<i>Outline of tutorial</i> .....	3
2	PeptideShaker Outputs.....	4
2.1	<i>PSM Report</i> .....	5
2.2	<i>Current history</i> .....	6
2.3	<i>Import tutorial datasets into current history</i> .....	6
3	<i>Running workflow for session 4</i> .....	7
3.1	<i>Inputs for the session 4 workflow</i> .....	7
3.2	<i>Workflow for session 4</i> .....	7
3.3	<i>Generating a PSM summary of peptides derived from RNA-Seq derived db</i> .....	11
3.4	<i>Converting peptide list into a FASTA format</i> .....	12
3.5	<i>BLAST-P searches and filtering</i> .....	14
3.6	<i>List of novel peptides and peptide-spectral match visualization</i> .....	15
3.7	<i>Peptide spectral summary and genome visualization</i> .....	22
3.8	<i>Genome visualization</i> .....	23
4	Running Entire Proteogenomic workflow .....	33
5	Presenters and acknowledgements.....	36

# 1 Introduction

## 1.1 Scope of this tutorial

This session of proteogenomics workshop will take you through the processing of search results (PSM Report) generated via SearchGUI / PeptideShaker analysis. This will include a blueprint workflow for a) Generating a PSM summary of peptides derived from RNA-Seq derived db; b) Converting peptide list into a FASTA format (as an input for BLAST-P analysis); c) BLAST-P searches and filtering ; d) PSM visualization and e) Visualization of peptides on the genome.

Reference materials

Salivary proteogenomics workflow manuscript:

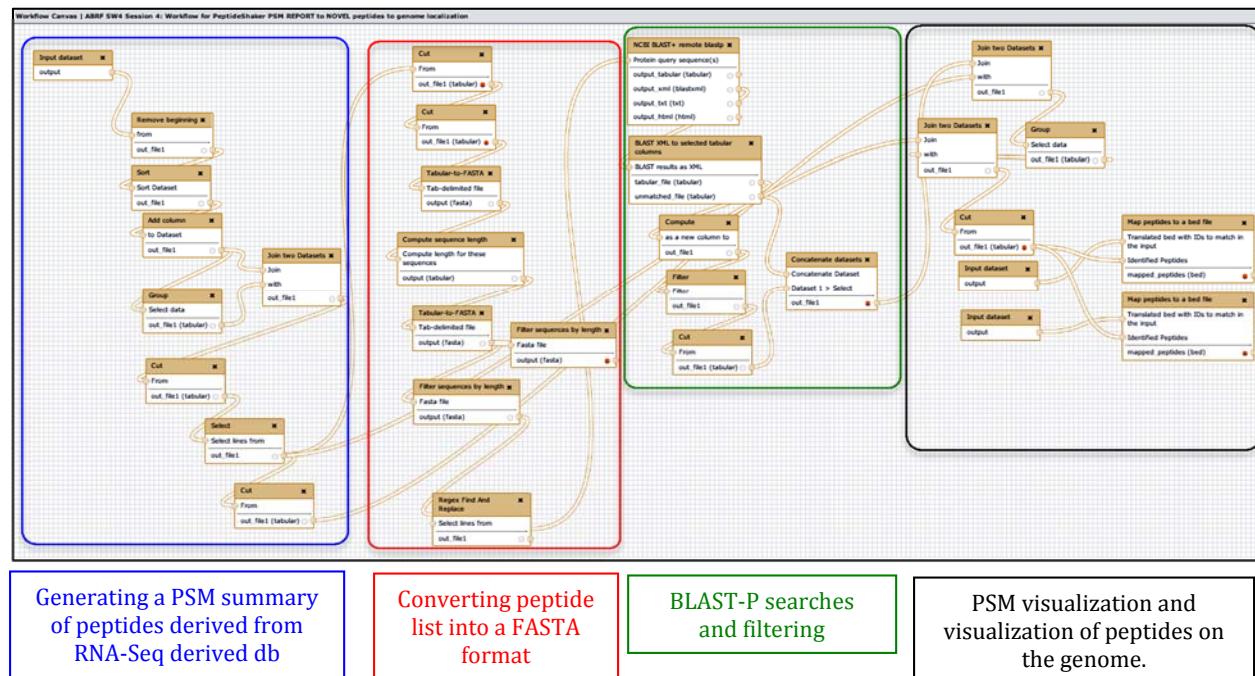
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4261978/>

Hibernation proteogenomics manuscript:

<http://www.ncbi.nlm.nih.gov/pubmed/26435507>

Multi-omics overview: <http://www.nature.com/nbt/journal/v33/n2/full/nbt.3134.html>

## 1.2 Outline of tutorial



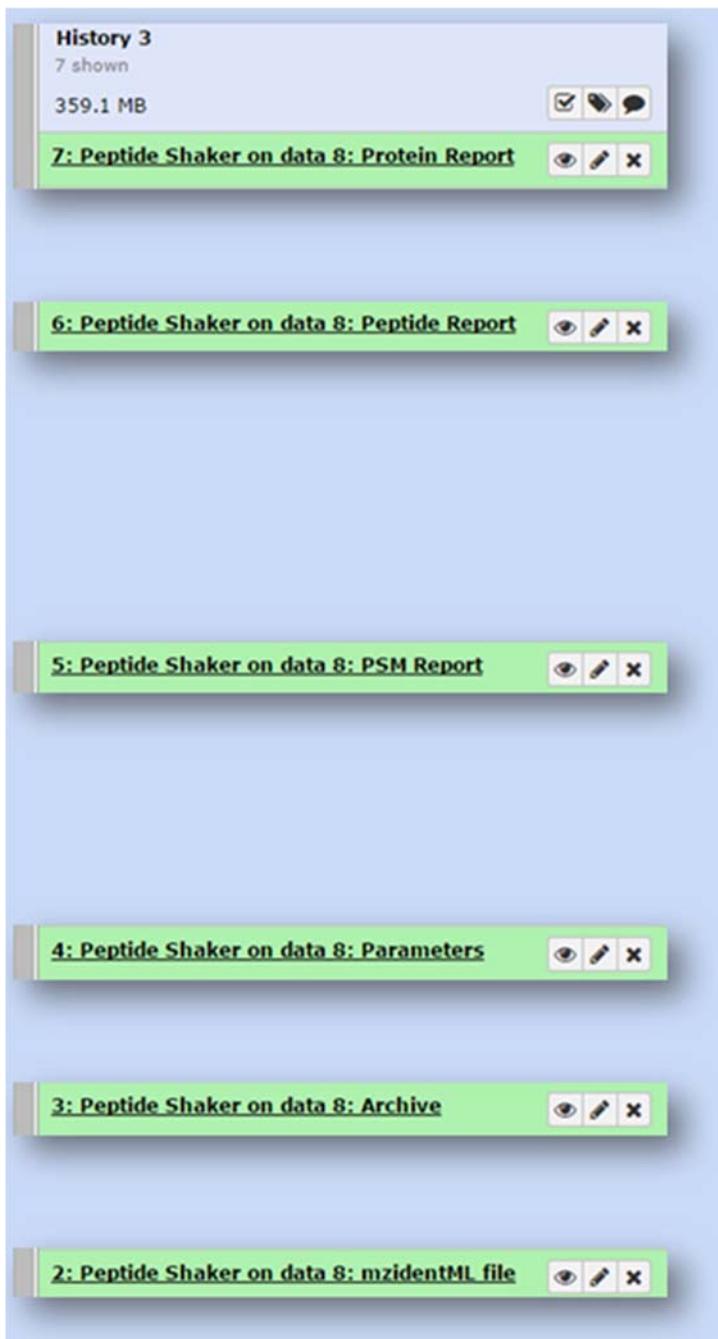
Generating a PSM summary  
of peptides derived from  
RNA-Seq derived db

Converting peptide  
list into a FASTA  
format

BLAST-P searches  
and filtering

PSM visualization and  
visualization of peptides on  
the genome.

## 2 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

### mzIdentML

- PSM Visualization
- SWATH Analysis
- Skyline
- Scaffold

## PeptideShaker Outputs

### 2.1 PSM Report (PeptideShaker Output)

The PSM Report contains information about the peptide-spectral matching of all spectra within the dataset. The report contains Sequence of the peptide (c3), the Spectrum scan information (c7) and its associated Confidence score (c19).

The screenshot shows the PeptideShaker software interface. At the top, it says "13: Peptide Shaker" and "on data 10: PSM Report". There are icons for eye, edit, and close. Below that, it says "5,415 lines" and "format: tabular, database: GRCm38\_canon". A message box displays: "Path configuration completed. Fri Feb 05 17:50:08 CST 2016 Unzipping searchgui\_input.zip. 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%". Below this, another message box says: "Fri Feb 05 17:52:30 CST 2016 Import process for Galaxy\_Experiment\_20160205174 (Sample: Sample\_20160205174914547)". At the bottom, there is a table titled "Protein(s) Sequence" with columns "TM score", "D-score", and "Confidence". The data in the table is:

	Protein(s)	Sequence	TM score	D-score	Confidence
1	A0A0R4J1E2, A0A0R4J1L2, D3YUQ9		.8	867.40094	2+ 2+
2	E9Q133, P80318, Q3U0I3				
3	Q8BTI8, Q8BTI8-2, Q8BTI8-3				I

#### 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**
- c20: Validation: *Confidence > 85 and delta ppm within 6 ppm are CONFIDENT PSMs*

## PeptideShaker Outputs

### 2.2 Current history

- ★ Your current status of history would be the output from Session 2. These would have a) mzIdentML file, b) Parameters, c) PSM Report and d) mzsQLite output to be used for PSM Visualization – as outputs from PeptideShaker analysis. You will need to import tutorial datasets into your current history.



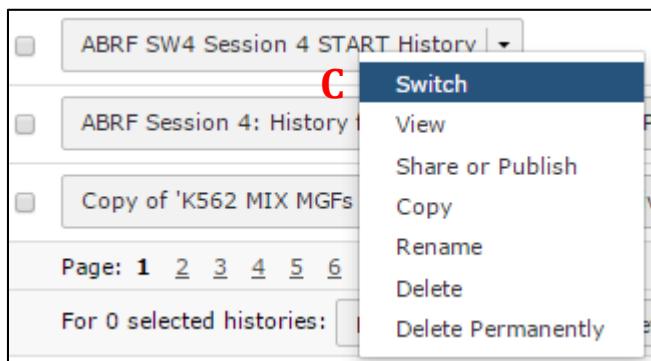
```
Project Details: PeptideShaker Version: 1.1.3
1: Date: Fri Feb 15 15:50:08 CST 2016
2: Project: ABRF_Galaxy_Flowcell_2016020517491454716158
3: Sample: Sample_2016020517491454716158
4: Replicate Number: 1
5: Identifying Algorithm: ORSSA, MS-DF+, Comet and MyrIMatch
Database Search Parameters
1: Precursor Accuracy Unit: ppm
2: Precursor Ion m/z Tolerance: 10.0
3: Fragment Ion m/z Tolerance: 0.1
4: Enzyme: Trypsin
5: Number of Missed Cleavages: Not Implemented
6: Database: input_database.fasta
7: Forward Ion: b
8: Reverse Ion: y
9: Fixed Modifications: Carbamidomethylation of C, iTRAQ 4-plex of K, iTRAQ 4-plex of peptide N-term
10: Variable Modifications: Oxidation of M, iTRAQ 4-plex of Y
11: Refinement Variable Modifications:
12: Refinement Fixed Modifications:
Input Filters
1: Minimal Peptide Length: 4
2: Maximal Peptide Length: 30
3: Maximal Tolerance: none
4: Precursor m/z Tolerance Unit: Yes
5: Unrecognized Modifications Disallowed: Yes
Validation Summary
1: #Validated Proteins: 1663.0
2: Protein Total: 1815.29
3:Protein FOR Limit: 0.94 %
4: Protein FDR Limit: 9.08 %
5: Protein Confidence Limit: 92.45 %
6: Protein PEP Limit: 7.55 %
7: Protein Confidence Accuracy: 2.13 %
8: #Validated Peptides: 2979.0
9: Peptide Total: 3057.0
10: Peptide FOR Limit: 0.97 %
11: Peptide FDR Limit: 14.76 %
12: Peptide Confidence Limit: 89.13 %
13: Peptide PEP Limit: 10.87 %
14: Peptide Confidence Accuracy: 1.22 %
15: #Validated PSM: 3857.0
16: #Validated PSM: 2357.0
17: PSM Total: 3777.78
18: PSM FOR Limit: 0.98 %
19: PSM FDR Limit: 0.98 %
20: PSM Confidence Limit: 98.98 %
21: PSM PEP Limit: 19.28 %
22: PSM Confidence Accuracy: 0.98 %
23: PSM Confidence Limit: 83.89 %
24: PSM Confidence Limit: 83.64 %
25: PSM PEP Limit: 16.11 %
26: PSM PEP Limit: 16.39 %
27: PSM Confidence Accuracy: 0.67 %
28: PSM Confidence Accuracy: 0.61 %
Posttranslational Modification Scoring Settings
1: A score: Yes
2: Accounting for Neutral Losses: No
3: False Location Rate: 1.0
Spectrum Counting Parameters
```

We will be processing the PSM Report and using its outputs to visualize PSM quality and localization of novel peptide identification on the mouse genome.

### 2.3 Import tutorial datasets into current history

Tutorial Dataset:

- At the top select “Saved Histories”
- Select “ABRF SW4 Session 4 START History” from the list of histories
- Select Switch.



## 3 Running workflow for session 4

### 3.1 Inputs for the session 4 workflow

For Session 4, the inputs that would be needed are PSM Report, Translate BED Sequences from pre-proB and ProB and mssqlite output for PSM Visualization. Read 2.3 to get the right inputs for this workflow.

**ABRF SW4 Session 4 START**

**History**

4 shown

159.16 MB

**4: mssqlite output to be used for PSM Visualization**

**3: Translate BED Sequences on data 12 bed**

**2: Translate BED Sequences on data 23 bed**

**1: Peptide Shaker on data 10: PSM Report**

### 3.2 Workflow for session 4

Select ABRF SW4 Session 4 START History as your active history.

Once you have selected your active history, select workflow for session 4 from a list of workflows.

Analyze Data **Workflow** Shared Data Visualization

Chain tools into workflows

## Running workflow for session 4

Run the workflow on the active history.

**Your workflows**

Name
ABRF SW4 Session 4: Workflow for PeptideShaker PSM REPORT to NOVEL peptides to genome localization

Localization ▾

- Edit
- Run**
- Share or Publish
- Download or Export
- Copy
- Rename
- View
- Delete

**Running workflow "ABRF SW4 Session 4: Workflow for PeptideShaker PSM REPORT to NOVEL peptides to genome localization"**

Step 1: Input dataset

PSM REPORT from PeptideShaker Output

- 1: Peptide Shaker on data 10: PSM Report
- 1: Peptide Shaker on data 10: PSM Report**
- 2: Translate BED Sequences on data 23 bed
- 3: Translate BED Sequences on data 12 bed

Step 2: Input dataset

bed file with genome coordinates of the protein search database. Produced by Translation of the PSM report.

proB-BED File: Translation bed file

- 3: Translate BED Sequences on data 12 bed
- 3: Translate BED Sequences on data 12 bed**

Step 3: Input dataset

Pre-proB BED File: Input Dataset

- 3: Translate BED Sequences on data 12 bed
- 2: Translate BED Sequences on data 23 bed**
- 3: Translate BED Sequences on data 12 bed

Step 4: Remove beginning (version 1.0.0)

Step 5: Sort (version 1.0.2)

Select ‘PSM Report’ for Step 1, ‘Translate BED Sequences on data 12 bed’ for Step 2 and ‘Translate BED Sequences on data 23 bed’ for Step 3.

## Running workflow for session 4

The screenshot shows a workflow step titled "Step 30: Map peptides to a bed file (version 0.1.0)" with the sub-instruction "Produces a bed file that can be displayed in IGV (probable)". Below it is another step titled "Step 31: Map peptides to a bed file (version 0.1.0)". At the bottom of the panel, there is a checkbox labeled "Send results to a new history" and a prominent blue "Run workflow" button, which is circled in red.

Run Workflow.

The screenshot displays a success message: "Successfully ran workflow 'ABRF SW4 Session 4: Workflow for PeptideShaker PSM REPORT to NOVEL peptides to genome localization'." Below the message is a list of 14 numbered steps:

- 1: Peptide Shaker on data 10: PSM Report
- 2: Translate BED Sequences on data 12 bed
- 3: Translate BED Sequences on data 12 bed
- 4: Remove beginning on data 1
- 5: Sort on data 4
- 6: Add column on data 5
- 7: Group on data 6
- 8: Join two Datasets on data 7 and data 6
- 9: Cut on data 8
- 10: Select on data 9
- 11: PSM REPORT of potential novel PEPTIDES
- 12: Cut on data 10
- 13: Potential Novel PEPTIDES
- 14: Tabular-to-FASTA on data 13

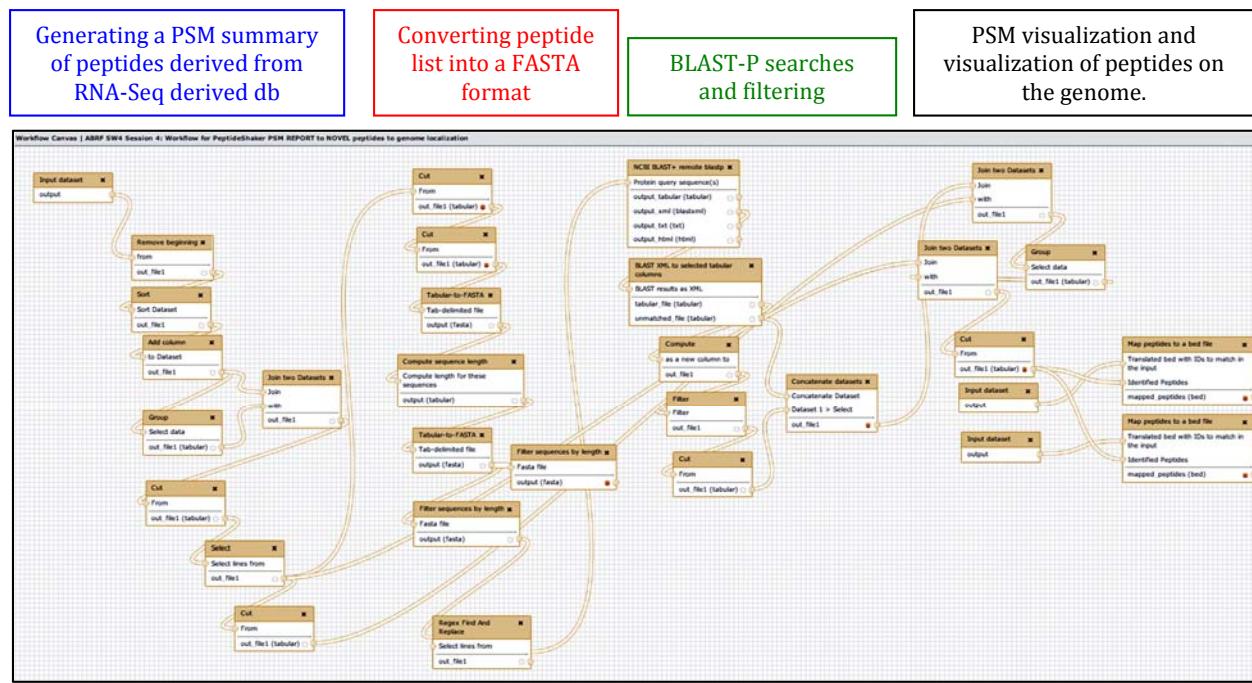
If your workflow ran successfully, we will use the history to go through the steps.

If not, then download the 'END History for Session 4' from Saved History.

The screenshot shows a list of saved histories. The "END History for Session 4" entry has a context menu open, with the "Switch" option highlighted in blue. The menu also includes "View", "Share or Publish", "Copy", "Rename", "Delete", and "Delete Permanently".

## Running workflow for session 4

*Workflow for Session 4* is a slightly complex workflow than the preceding one, which involved only SearchGUI and PeptideShaker. In this workflow, 29 processing steps are used to take the PSM Report from PeptideShaker and manipulate it to put through BLAST-P analysis to verify novel proteoforms.

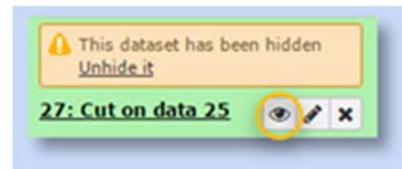


### Overview of Session 4 Workflow

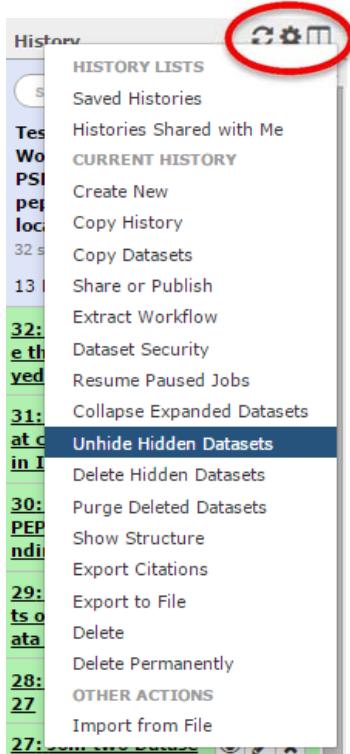
#### Step 1: Input dataset (PSM Report)

- Steps 5-11: Selects peptides with accession number from RNASeq-derived protein FASTA file. (See Section 3.3 below for details)
- Step 12: PSM Report of peptides identified from RNASeq-derived proteins.
- Steps 13-20: Conversion of peptide list into a FASTA format
- Step 21: Short BLAST-P on NCBI remote nr mouse database
- Steps 22-26: Brings up mismatched peptides.
- Step 27: Peptides corresponding to novel proteoforms.
- Steps 28-30: Conversion to PSM Report of peptides corresponding to novel proteoforms.
- Step 31: PSM Report of peptides corresponding to novel proteoforms.
- Steps 32 and 33: Produces BED files for IGV viewer.

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



### 3.3 Generating a PSM summary of peptides derived from RNA-Seq derived db



Unhide Hidden Datasets from your completed workflow OR from the “END History for Session 4” that you have as your current history.

**Once “unhidden” you should see 33 datasets within your history.**

Let us focus on steps 5 to 12

Step 5: Removes the beginning line of the PSM Report (Now we are without headers and will need to use columns as our headers!)

*Note: To view details of a step, click on the step number and then click on the ‘rerun’ icon. DO NOT hit rerun though!*

Step 6: Sorts PSM Report with increasing Spectrum Title (column 7) ascending order and Confidence (column 19) in descending order. This ensures that the highest ranking PSM for that spectrum title is at the top.

Step 7: Ranks columns based on the new sorting performed in Step 6.

Step 8: Group - helps in selecting only one PSM per Spectrum Title.

Step 9 and 10: Join and Cut - generates PSM Report of one PSM per spectrum title.

Step 11: Selects peptides with accession number starting with preB or proB. (Details in figure below)

Step 12: PSM Report of potential novel PEPTIDES – click on eye icon to view details of the PSM Report in the main panel.



## Running workflow for session 4

11: Select on data ✖

10  
5 lines  
format: tabular, database: GRCm38\_canon

Run this job again

1746 prcB\_JUNC000923898\_2 00SPAE  
176 prcB\_JUNC000923898\_2 00SPAE  
207 prcB\_JUNC000923898\_1 LNGEVS  
R (7: Very Confident), iTRAQ 4-plex of  
511 prcB\_JUNC000923898\_2 00SPAE  
2016 prcB\_JUNC000923898\_2 00SPAE

Select lines that match an expression (Galaxy Tool Version 1.0.1)

Select lines from 10: Cut on data 9

that Matching

the pattern `^d+\|pr.B[^t]*\|, pr.B[^t]*\|t.\$`

here you can enter text or regular expression (for syntax check lower part of this frame)

Step 11: Selects peptides with accession number starting with preB or proB. (Details in figure below)

### 3.4 Converting peptide list into a FASTA format (as an input for BLAST-P analysis)

Let us focus on steps 13 to 20

Step 13: Generates a peptide list along with ranking number by cutting column 3 (c3) and column 21 (c21) from the step 12 tabular format file.

Step 14: Generates a peptide list by cutting column 1 (c1) step 13 tabular format file.

Step 15: Generates a FASTA format from Step 14 in the following format:

>PEPTIDE  
PEPTIDE

Step 16: Computes sequence length on data 15.

Step 17: Generates a FASTA Format from Data 16.

>PEPTIDE\_length  
PEPTIDE

Step 18: Filters sequences from length 8 to 30 aas from the list of sequences.

Step 19: Filters sequences from length 31 to 50 aas from the list of sequences. In our test case for the tutorial we do not have any example.

Step 20: Converts Step 18 output to a format so that it can be searched by short BLAST-P search.

>PEPTIDE\_sequence length=length aa  
PEPTIDE

22: BLAST results a s tabular for data 2 ✖

1

21: blastp-short on remote nr ✖

20: Regex Find And Replace on data 18 ✖

19: FASTA file with peptides with length greater than 30 aas ✖

18: Filter sequence s by length on data 17 ✖

17: Tabular-to-FAST A on data 16 ✖

16: Compute sequence length on data 15 ✖

15: Tabular-to-FAST A on data 14 ✖

14: Potential Novel PEPTIDES ✖

13: Cut on data 11 ✖

12: PSM REPORT of potential novel PEP TIDES ✖

**Regex Find And Replace** (Galaxy Tool Version 0.1.0)

Select lines from  
18: Filter sequences by length on data 17

**Check**  
1: Check Delete

**Find Regex**  
`^(>.*)([0-9]+)$`  
here you can enter text or regular expression (for syntax check lower part of this frame)

**Replacement**  
`\1 sequence length=\2 aa`

**Buttons**  
+ Insert Check  Execute

Step 20: Converts Step 18 output to a format so that it can be searched by using short BLAST-P search.

### 3.5 BLAST-P searches and filtering

BLAST-P SEARCH

**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 href="#">31: PSM REPORT of PEPTIDES corresponding to novel proteoforms</a>			
<a href="#">30: Join two Databases on data 29 and data 11</a>			
<a href="#">29: Group on data 28</a>			
<a href="#">28: Join two Databases on data 13 and data 27</a>			
<a href="#">27: PEPTIDES corresponding to novel proteoforms</a>			
<a href="#">26: Cut on data 25</a>			
<a href="#">25: Filter on data 24</a>			
<a href="#">24: Compute on data 22</a>			
<a href="#">23: Query sequences with no hits for data 21</a>			
<a href="#">22: BLAST results as tabular for data 21</a>			
<a href="#">21: blastp-short on remote nr</a>			
<a href="#">20: Regex Find And Replace on data 18</a>			
<a href="#">19: FASTA file with peptides with length</a>			

BLAST-P Search and output processing is carried out from steps 21 to 27.

Step 21: This step performs short BLAST-P search on peptide FASTA sequences and generates a XML output. The short BLAST-P uses parameters for short peptide sequences (8-30 aas). Please use the rerun option to look at the parameters used.

Step 22: Converts BLAST XML output into a tabular output with various metrics such as a) ID of your sequence (c1); b) Percentage of identical matches (c3); c) Total number of gaps (c17) d) Alignment length (c4) and Query length (c23).

Step 23: Query sequences with no hits for data 21.

Step 24: Calculates percentage of alignment length versus actual query length and adds it as column 25.

Steps 25: Selects peptides with - Percentage of identical matches (c3) less than 100 OR Total number of gaps (c17) is at least one OR percentage of alignment length versus actual query length is less than 100.

Steps 26 and 27: Generates a list of **peptides corresponding to novel 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.

## 3.6 List of peptides corresponding to novel proteoforms and PSM visualization.

### Peptide Spectral Match (PSM) Visualization

#### Importance of PSM Visualization

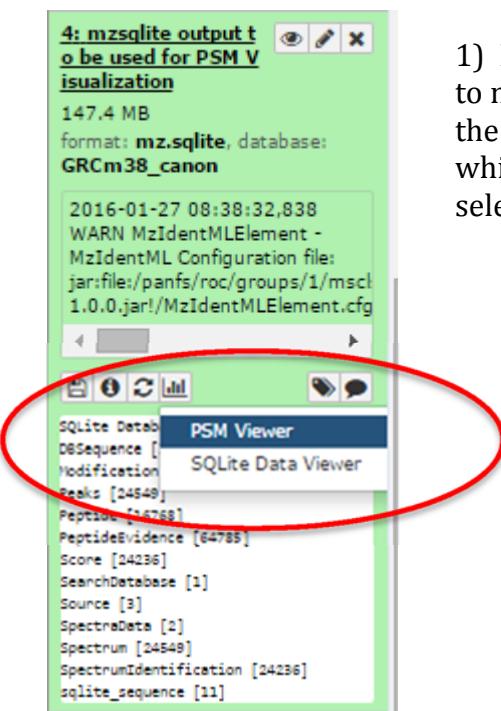
Scoring matrices and FDR thresholding of data acquired from search algorithms serves to reduce the number of many poor peptide-spectral matches (PSMs). However, some low quality identifications elude filtering and must be manually evaluated. PSM Visualization may reveal that a reported high-scoring spectrum 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 mwid) 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.



1) In order to visualize PSMs of the peptides corresponding to novel proteoforms, go to step 6 and click on the title of the step. Once the box has expanded, click on the icon which representing a bar diagram (for visualization) and select PSM Viewer and click.

2) This would result in opening of another Tab.



3) Select on 'Sequences from History' and Select Potential Novel PEPTIDES:

The screenshot shows the Galaxy SQLite Data Viewer interface. At the top, there is a tab bar with 'Galaxy SQLite Data Viewer' and 'New Tab'. Below the tab bar, the URL 'psmeviz?dataset\_id=f72ec72df8287025' is displayed. The main menu bar includes 'PSM Viewer', 'Scans to History', 'Save Visualization to Galaxy', 'Sequences from History' (which is highlighted with a red oval), and 'IGV Viewer'. A dropdown menu is open under 'Sequences from History', listing various data processing options. The option 'Potential Novel PEPTIDES' is also highlighted with a red oval.

mzsqlite\_output

Sequences from History

Potential Novel PEPTIDES

- Peptide\_Shaker\_on\_data\_10\_PSM\_Report.tabular
- Remove beginning on data 4
- Sort on data 5
- Add column on data 5
- Group on data 6
- Join two Datasets on data 7 and data 6
- Cut on data 8
- Select on data 9
- PSM REPORT of potential novel PEPTIDES
- Cut on data 10
- Potential Novel PEPTIDES**
- Compute sequence length on data 14
- BLAST results as tabular for data 20
- Query sequences with no hits for data 20
- Compute on data 21
- Filter on data 23
- Cut on data 24
- PEPTIDES corresponding to novel proteoforms
- Join two Datasets on data 12 and data 26
- Group on data 27
- Join two Datasets on data 28 and data 10
- PSM REPORT of PEPTIDES corresponding to novel proteoforms

4) A new tab shows up named "Choose a sequence(s) for filtering":

The screenshot shows the Galaxy SQLite Data Viewer interface with a new tab open. The tab title is 'mzsqlite\_output\_to\_be\_used\_for\_PSM\_Visualization.mz.sqlite'. The top menu bar is identical to the previous screenshot. A dropdown menu is open under the 'Choose a sequence(s) for filtering' button, which is highlighted with a red oval. The dropdown menu contains several peptide sequences.

Choose a sequence(s) for filtering

mzsqlite\_output\_to\_be\_used\_for\_PSM\_Visualization.mz.sqlite

- GGSPAEGAIR
- GGSPAEGAIR
- LNSEYSMAETPSR
- GGSPAEGAIR
- GGSPAEGAIR

5) Select on peptide 'LNSEYSMAETPSR'

The screenshot shows the Galaxy SQLite Data Viewer interface with the same tab as the previous screenshot. The dropdown menu under 'Choose a sequence(s) for filtering' is still open, and the peptide 'LNSEYSMAETPSR' is highlighted with a red oval. The other peptides listed are GGSPAEGAIR, GGSPAEGAIR, GGSPAEGAIR, and GGSPAEGAIR.

Choose a sequence(s) for filtering

mzsqlite\_output\_to\_be\_used\_for\_PSM\_Visualization.mz.sqlite

- GGSPAEGAIR
- GGSPAEGAIR
- LNSEYSMAETPSR**
- GGSPAEGAIR
- GGSPAEGAIR

## 6) Click on Filter Peptides.

PSM Viewer Scans to History Save Visualization to Galaxy Sequences from History IGV Viewer Choose a sequence(s) for filtering

mzssqlite\_output\_to\_be\_used\_for\_PSM\_Visualization.mz.sqlite

Available Tables: Peptides

Showing 1 to 10 of 16,235 entries

peptide.sequence	count(si.pkid)
AAAAAAPAGGPAAAAPSGENAEESR	2
AAAAADHR	1
AAAELSSSSGSER	1
AAAAGALAP/GPL_POLAAR	1
AAAAPAGGNPGEGR	3
AAAASAEAQIATPGTEDSDALLK	2
AAADTLQGPMQAYR	5
AAAEPQEAAAASDGTAESGVPAK	2
AAEELLQSQQSAGGSQTLK	6
AAAEDQNETVVVK	1

Show 10 entries Previous 1 2 3 4 5

Filter Proteins

Find peptides by sequence(s)

## 7) Click on Filter Peptide Tab

PSM Viewer Scans to History Save Visualization to Galaxy Sequences from History IGV Viewer Choose a sequence(s) for filtering

mzssqlite\_output\_to\_be\_used\_for\_PSM\_Visualization.mz.sqlite

Available Tables: Peptides

Showing 1 to 1 of 1 entries

peptide.sequence	count(si.pkid)
LNSEYSMAETPSR	1

Show 10 entries Previous 1 Next

Filter Proteins

Find peptides by sequence(s)

Show 1 to 1 of 1 entries

Score."PeptideShaker PSM score"	Score."OMSSA:evalue"	Score."MS-GF:SpecValue"	Score."Comet:expectation value"	Score."theoretical mass"	Score."PeptideShaker PSM confidence"	Score."PeptideShaker PSM confidence type"	Score."MyriMatch:MVH"	Spectrum.precursorCharge	Spectrum.precursorIntensity
40.57	0.0472989373184082	6.054829e-8	0.0214	1787.8603475229902	95.86	0	8.10765714265142e-26	2	

Show 10 entries Previous 1 Next

- 8) Click on bottom Tab with Score information and it opens a Peptide Sequence Tab for visualization.

The screenshot shows the PSM Viewer interface with the following details:

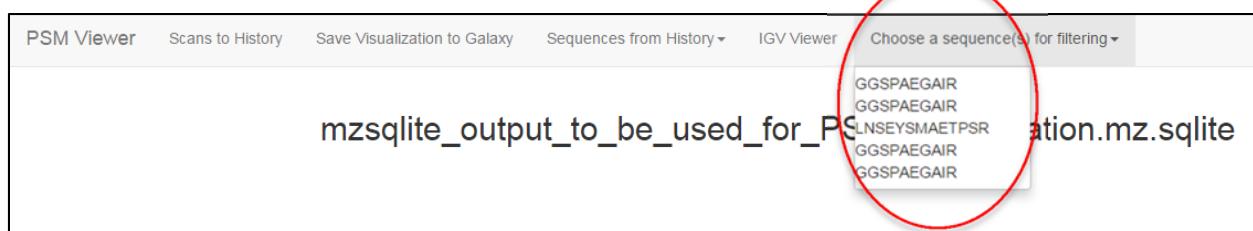
- Header:** PSM Viewer, Scans to History, Save Visualization to Galaxy, Sequences from History, IGV Viewer, Choose a sequence(s) for filtering.
- Title:** mzssqlite\_output\_to\_be\_used\_for\_PSM\_Visualization.mz.sqlite
- Available Tables:** Peptides
- Table Headers:** Showing 1 to 1 of 1 entries, peptide.sequence, count(si.pkid).
- Table Data:** LNSEYSMAETPSR, count(si.pkid) 1.
- Filtering:** Show 10 entries, Filter Proteins, Find peptides by sequence(s) LNSEYSMAETPSR, Filter Peptides, Clear Filtering.
- Score Headers:** Score."PeptideShaker PSM score", Score."OMSSA:evalue", Score."MS-GF:SpecEValue", Score."Comet:expectation value", Score."theoretical mass", Score."PeptideShaker PSM confidence", Score."PeptideShaker PSM confidence type", Score."MyriMatch:MH".
- Score Data:** 0.0472989373184082, 6.084829e-8, 0.0214, 1787.8603475229902, 95.86, 0, 0.07657142651428e-26.
- Buttons:** Show 10 entries, Toggle View.
- Sequence Selection:** LNSEYSMAETPSR 3827, with a red circle around it.

- 9) Clicking on the Tab and appropriate parameters reveals the PSM for ‘LNSEYSMAETPSR’ – our first candidate:



- To visualize fragment ion series select the desired series under the ions interface. An accepted approach to validate spectra is to visualize all b<sup>1+</sup>, y<sup>1+</sup>, MH<sup>1+</sup>, Internal Ions, all Neutral Losses. Visualize at a Mass Tolerance of 0.05 Daltons.
- Users can validate the table visually from the spectral graph or from fragment ion table (on the right).

- 10) Click on “Choose a sequence(s) for filtering” and Select on peptide ‘GGSPAEGAIR’:



11) Click on Filter Peptides.

Available Tables: Peptides

Showing 1 to 2 of 2 entries

**peptide.sequence**

GGSPAEGAIR
LNSEYSMAETPSR

Show 10 entries

**Filter Proteins**

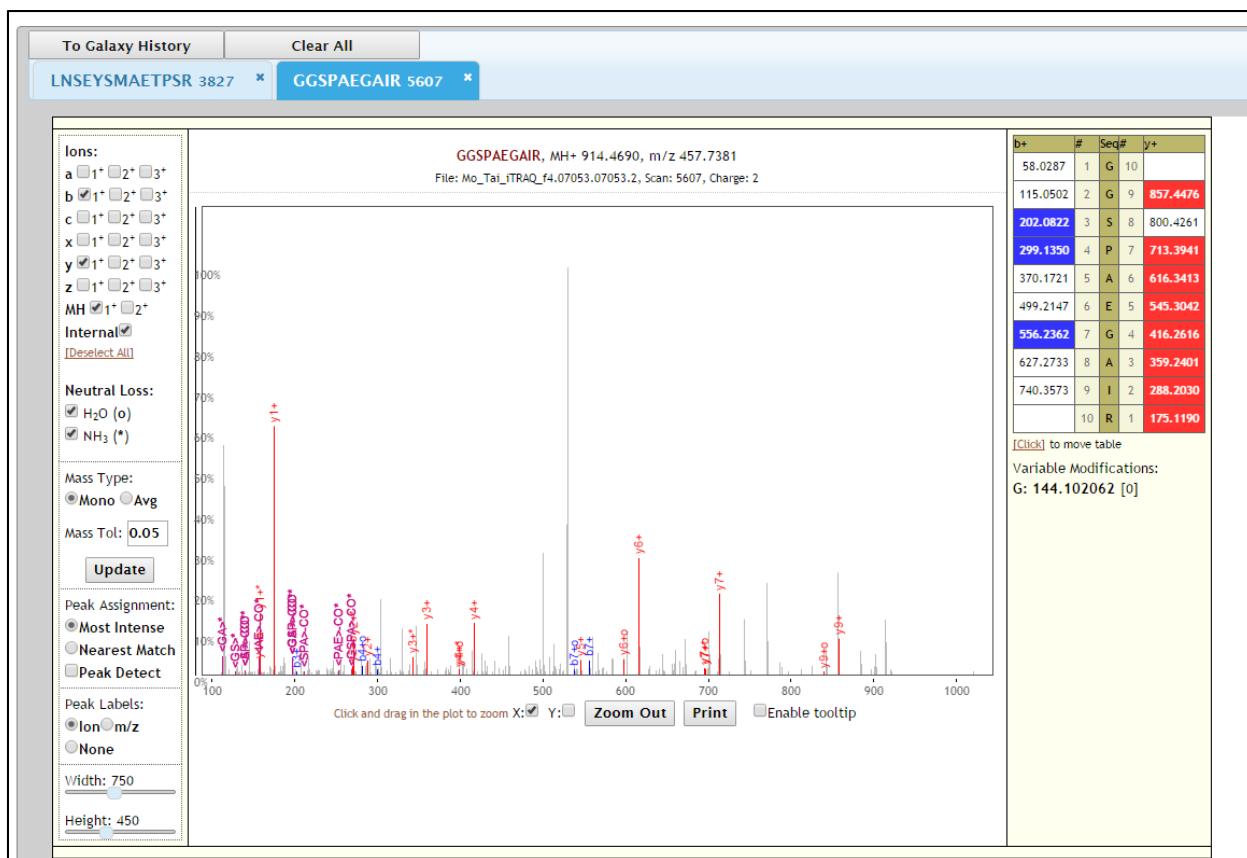
Find peptides by sequence(s)

**Filter Peptides** (button circled in red)

Clear Filtering

Showing 1 to 1 of 1 entries

12) Click on 'GGSPAEGAIR' under peptide.sequence and Click on bottom Tab with Score information and it opens a Peptide Sequence Tab for visualization.



Users can switch between spectra by selecting the sequence at the top of the visualizer.  
Compare matched ions and fragment ion table on the right to assess the quality of spectrum.  
Determine on why some ions are matched and others are not. Do you think this is a good spectrum?

### 3.7 Peptide spectral summary and genome visualization.

32: proB BED file that can be displayed in IGV

31: PSM REPORT of PEPTIDES corresponding to novel proteoforms

2 lines  
format: tabular, database: GRCh38\_canon

1	2	3
3839	proB_2JUC00092898_2	GSSPA
297	proB_2JUC00092617_1	LISEYI

30: Join two Datasets on data 29 and data 11

29: Group on data 2 8

28: Join two Datasets on data 13 and data 27

27: PEPTIDES corresponding to novel proteoforms

2 lines  
format: tabular, database: GRCh38\_canon

1
GSSPAEGAIR
LISEYIHAETPSR

26: Cut on data 25

2 lines  
format: tabular, database:

In order to visualize peptides onto a genome, we will first need to generate a PSM Report for the novel peptides. This is done in steps 28 to 31.

Step 28: Joins two tabular files Step 27 and Step 13 with Column 1 (Peptide sequence). This generates three columns – first two are peptide sequences and third one is ranking of the PSM.

Step 29: Generates a list of distinct peptides by grouping column 1 (Peptide Sequence).

Step 30: The join datasets function joins distinct peptides from Step 29 with Step 11 PSm Report using the ranking column 21.

Step 31: PSM Report of PEPTIDES corresponding to novel proteoforms. This PSM Report is used to generate subset BED files for mapping peptides onto the genome.

## 3.8 Genome Visualization

### Mapping peptides corresponding to novel proteoforms to the genome.

#### Peptides to Genome

Considerable information can be obtained from the identity of 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 read [review by Nesvizhskii et al \(2014\)](#).

#### Map Peptides to a BED File

A **BED** 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 **Map Peptides to a BED file** Galaxy tool generates a BED 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 BED 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.

## Generating a subset BED file

The screenshot shows a Galaxy workflow interface with two main panels. On the left, a list of completed steps is shown:

- 33: pre-proB BED file that can be displayed in IGV
- 32: proB BED file that can be displayed in IGV
- 1: region, 2 comments  
format: bed, database: GRCm38\_canon
- Run this job again nb\_current [local]  
display in IGV view  
display with IGV [local]
- 1.Grm
- track name="novel\_junction\_peptides" 1  
gffTags
- 31: PSM REPORT of PEPTIDES corresponding to novel proteoforms
- 2 lines  
format: tabular, database: GRCm38\_canon
- 1 2 3
- 3339 proB\_2UIC000921898\_2 003PA
- 297 preB\_2UIC00092617\_1 LIGEV
- N (7: Very Confident), iTRAQ 4-plex or
- 4
- 30: Join two Datasets on data 29 and d

On the right, the "Map peptides to a bed file" tool configuration is displayed:

- Translated bed with IDs to match in the input: 31: PSM REPORT of PEPTIDES corresponding to novel proteoforms
- Identified Peptides
- peptide column: Column: 3 (highlighted with a red circle)
- protein name column: Column: 2 (highlighted with a red circle)
- peptide offset column: Nothing selected
- Use #gffTags in output: Yes
- Execute button

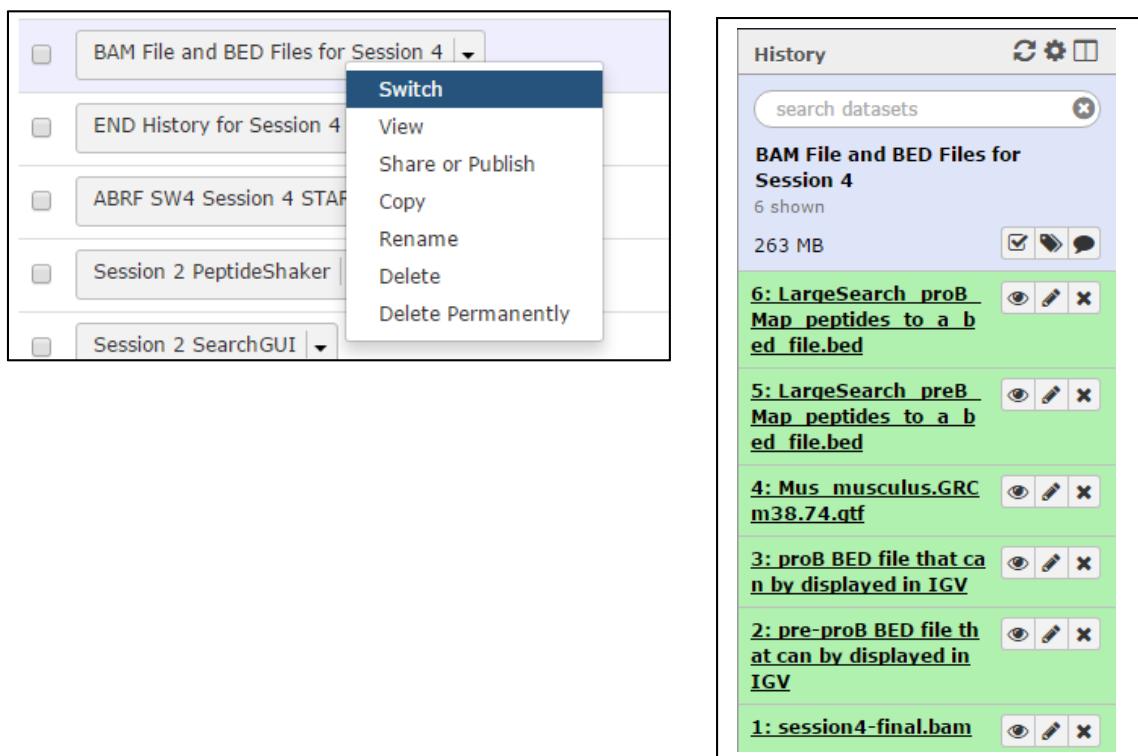
- The Map peptides to a bed file tool uses BED Sequences generated from Session 2 and PSM Report from Step 31. The tool also needs coordinates for peptide sequence (column 3) and protein (column 2).
- The tool generates BED files in steps 32 and 33.

## Localizing a subset BED file onto a IGV Browser.

- a) Open your IGV Browser from Session 3 (Section 5).
- b) Enter “chr12: 85331772 -85332611” into the region field and press “Enter”. This will zoom IGV into this region. This is one of the regions we will be examining. Zooming into the region makes loading BAM files faster.
- c) Remove all existing tracks.

### Load files into IGV

- d) Switch to the history with the name “**BAM Files and BED Files for Session 4**” and switch to the history.



- e) Click on the “...session 4 final.bam” file in the history pane to expand it and click the link “local” next to “display with IGV web current”.

The screenshot shows the Galaxy History pane. At the top, there is a search bar labeled "search datasets". Below the search bar, the title "BAM File and BED Files for Session 4" is displayed, followed by "6 shown" and a total size of "263 MB". There are three icons in a row: a checkmark, a clipboard, and a speech bubble. The list of datasets is as follows:

- 6: LargeSearch proB**  
Map peptides to a b  
ed file.bed
- 5: LargeSearch preB**  
Map peptides to a b  
ed file.bed
- 4: Mus musculus.GRC**  
m38.74.gtf
- 3: proB BED file that ca**  
n be displayed in IGV
- 2: pre-proB BED file th**  
at can be displayed in  
IGV
- 1: session4-final.bam**

For the "1: session4-final.bam" entry, the size is listed as "2.0 MB" and the format is "bam". A question mark icon indicates a database. Below this, there is a text input field containing "uploaded bam file". At the bottom of the list, there are two links: "display with IGV local" and "display in IGB View".

- f) You should now see a track in IGV named “session4 final”.

- g) Click on the “pre-proB BED file that can by displayed in IGV” file in the history pane to expand it and click the link “local” next to “display with IGV web current.”

**BAM File and BED Files for Session 4**  
6 shown  
263 MB

- 6: LargeSearch\_proB**  
Map peptides to a bed file.bed
- 5: LargeSearch\_preB**  
Map peptides to a bed file.bed
- 4: Mus\_musculus.GRCm38.74.gtf**
- 3: pre-proB BED file that can by displayed in IGV**
- 2: pre-proB BED file that can by displayed in IGV**

1 region, 2 comments  
format: **bed**, database: **GRCm38\_canon**

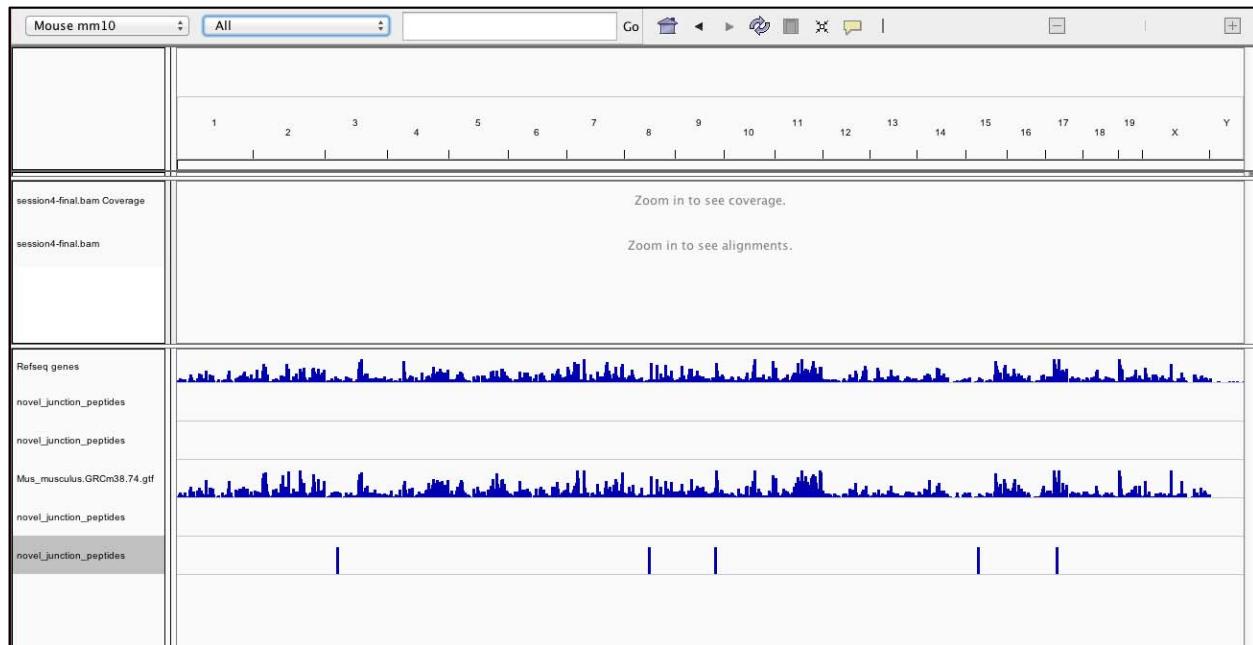
display in IGB [View](#)  
display with IGV [local](#)  
display with IGV [local](#)

```
1.Chrom 2.Start 3.End 4
track name="novel_junction_peptides" type=
#gffTags
12      85331772 85332611 ID=preB_JUNC000
```

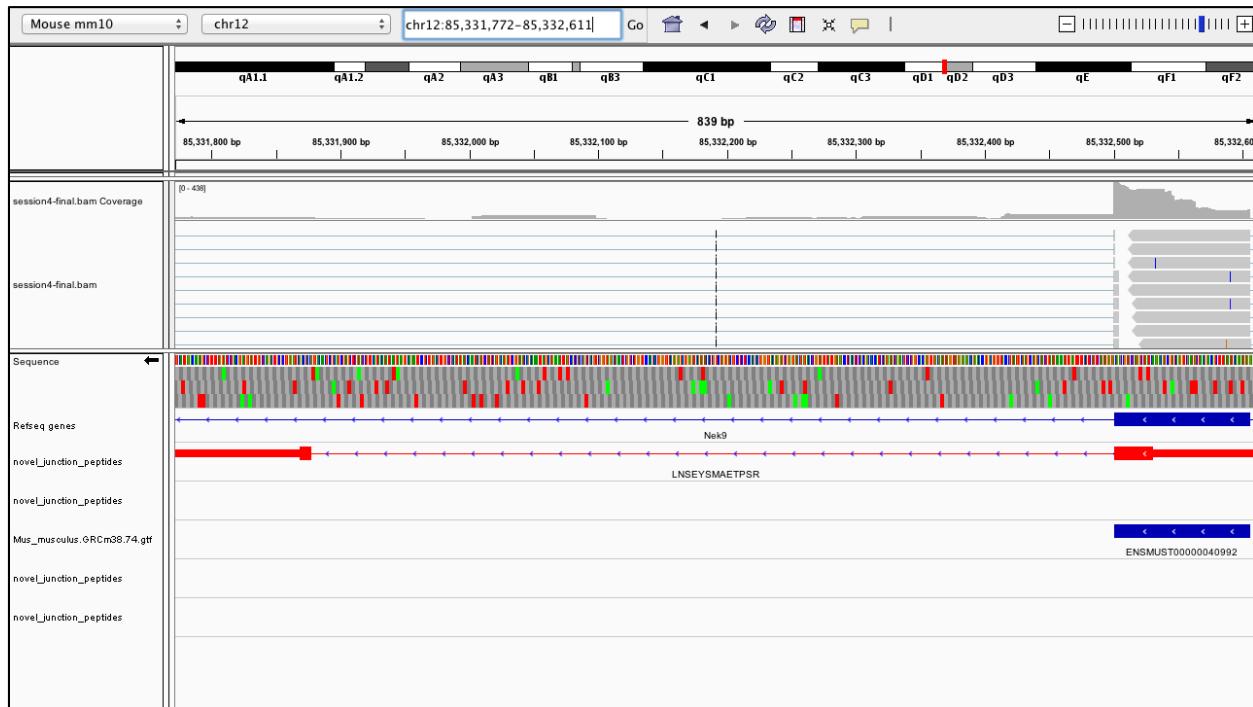
**1: session4-final.bam**

- h) You should now see a track in IGV named “novel junction peptides”. Change the name of the track to preB\_small if you wish.

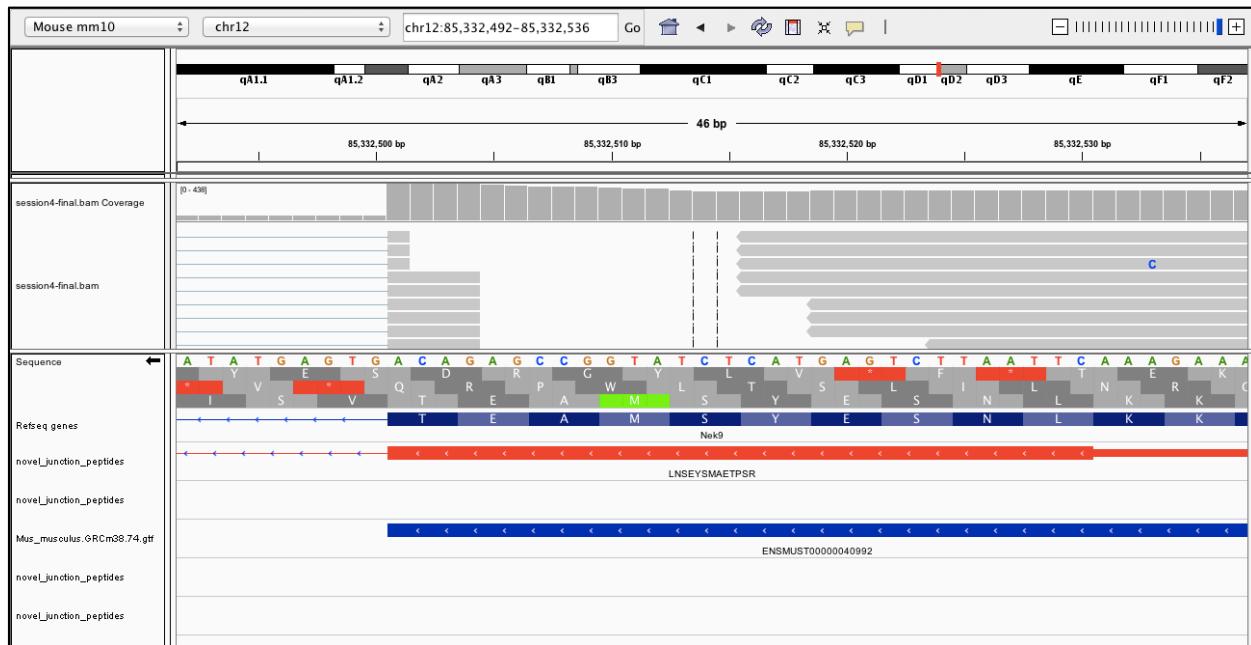
i) Repeat steps g and h for datasets 3, 4, 5 and 6.



j) Enter “chr12: 85331772 -85332611” into the region field and press “Enter”.



k) Zoom out a little bit to ensure your novel splice junction is in view. Does it make sense why this splice-junction was marked as novel? Explore both the junctions to find out if there is anything worth noting. Also look at the BAM files RNASeq evidence.



This was the novel peptide wherein the PSM was not of good quality.

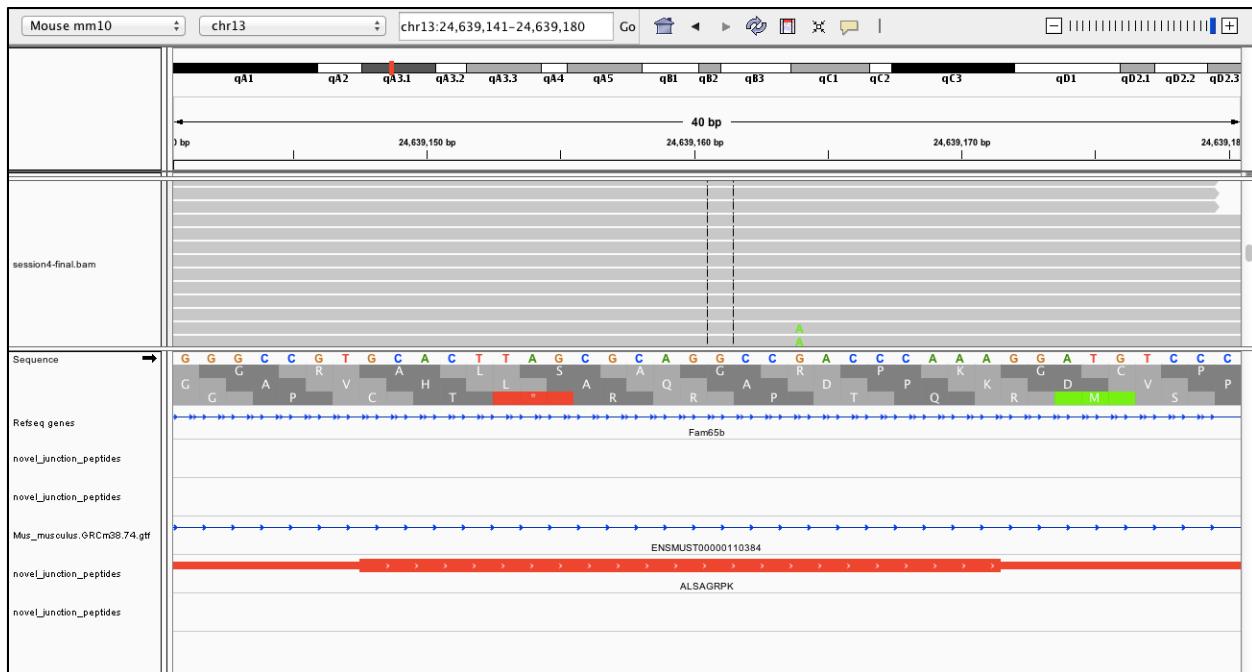
- l) Enter “chr10: 80777493-80778343” into the region field and press “Enter”. This will zoom into the peptide which had slightly better PSM quality.



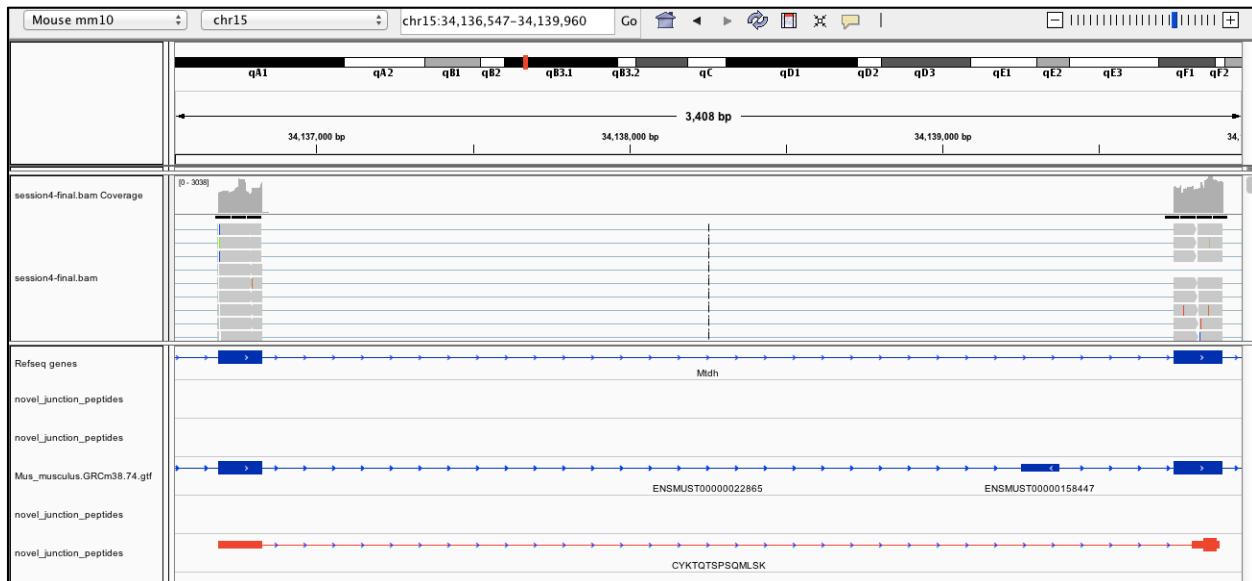
Explore both the junctions to find out if there is anything worth noting. Also look at the BAM files RNASeq evidence.



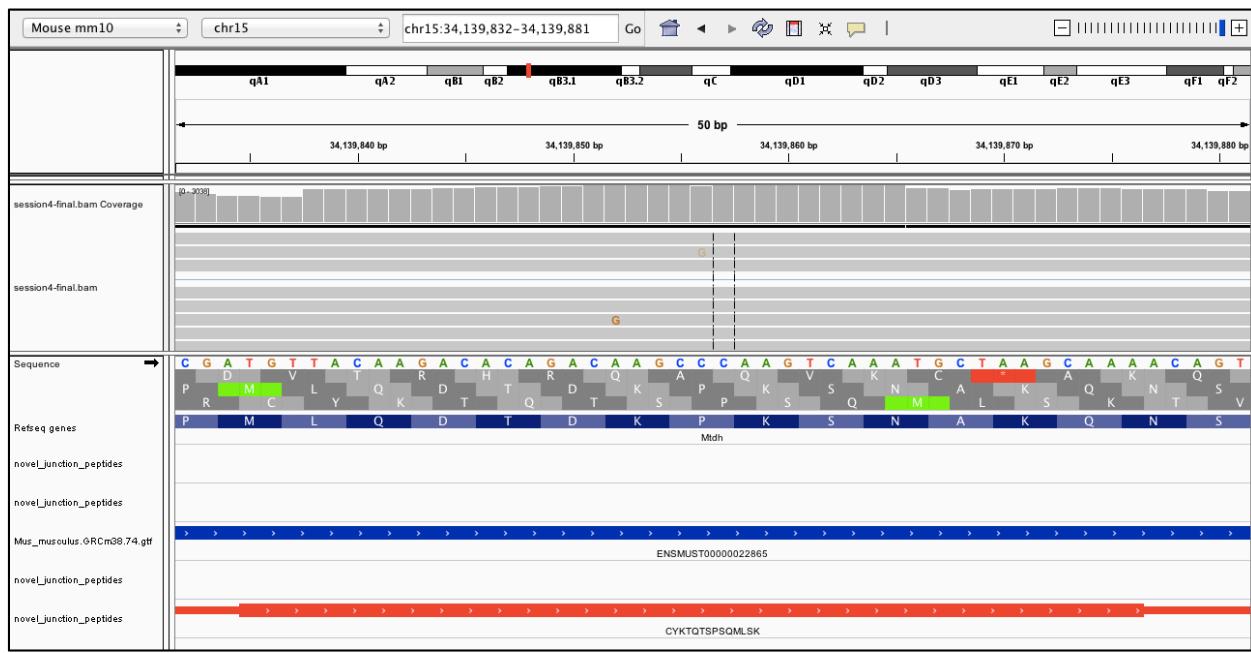
m) Here is another region to explore - “chr13: 24639141-24639180”



n) Here is another region to explore - “chr15: 34136547-34139960”



Explore the junction with identified peptide.



## 4 Running Entire Proteogenomics workflow

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

- A. Select the INPUT For Entire ProteoGenomics Workflow from the list of saved histories.

Saved Histories

search history names and tags

Advanced Search

Name

INPUT For Entire ProteoGenomics Workflow

END History for Session 4

ABRF SW4 Session 4 START Hist

Unnamed history

Session 2 End SearchGUI

Switch

- View
- Share or Publish
- Copy
- Rename
- Delete
- Delete Permanently

History

search datasets

**INPUT For Entire ProteoGenomics Workflow**

5 shown

10.45 MB

7: Translate BED Sequences on data 23 bed.bed

6: Translate BED Sequences on data 12 bed.bed

5: InputData

a list of datasets

2: preB Select Junc.fasta

1: proB Select Junc.fasta

## Running Entire Proteogenomics workflow

At the top of the screen select Workflows and click and Run for SW4\_ABRF2016: Entire ProteoGenomics Workflow

The screenshot shows the Galaxy interface with the title 'Your workflows'. A context menu is open over the workflow named 'SW4\_ABRF2016: Entire ProteoGenomics Workflow'. The menu options are: Edit, Run, Share or Publish, Download or Export, Copy, Rename, View, and Delete. The 'Run' option is highlighted with a blue background.

Appropriately assign each input database from the input history to the corresponding input or the workflow. Run the workflow. Run the workflow.

The screenshot shows the 'Running workflow "SW4\_ABRF2016: Entire ProteoGenomics Workflow"' interface. It consists of ten steps, each with a dropdown menu for selecting an input dataset:

- Step 1: Input dataset collection**: Input Dataset Collection mzml dropdown with item 5: InputData selected.
- Step 2: Protein Database Downloader (version 0.2.0)**: Step header.
- Step 3: Protein Database Downloader (version 0.2.0)**: Step header.
- Step 4: Input dataset**: proB database dropdown with item 1: proB\_Select\_Junc.fasta selected.
- Step 5: Input dataset**: pre-proB database dropdown with item 2: preB\_Select\_Junc.fasta selected.
- Step 6: Input dataset**: Translated BED File preB dropdown with item 6: Translate\_BED\_Sequences\_on\_data\_12\_bed.bed selected.
- Step 7: Input dataset**: Translated BED File proB dropdown with item 7: Translate\_BED\_Sequences\_on\_data\_23\_bed.bed selected.
- Step 8: MGF Formatter (version 0.1.0)**: Step header.
- Step 9: Regex Find And Replace (version 0.1.0)**: Step header.
- Step 10: Renex Find And Replace (version 0.1.0)**: Step header.

## Running Entire Proteogenomics workflow

The workflow output would be like 'OUTPUT From Entire ProteoGenomics Workflow' which is stored in saved histories.

OUTPUT From Entire ProteoGenomics Workflow			
19 shown, 28 hidden			
4.82 GB	<input checked="" type="checkbox"/>		
<a href="#"><u>49: pre-proB BED file that can be displayed in IGV</u></a>			
<a href="#"><u>48: proB BED file that can be displayed in IGV</u></a>			
<a href="#"><u>47: PSM REPORT of PEPTIDES corresponding to novel proteoforms</u></a>			
<a href="#"><u>43: PEPTIDES corresponding to novel proteoforms</u></a>			
<a href="#"><u>35: FASTA file with peptides with length greater than 30 aas</u></a>			
<a href="#"><u>30: Potential Novel PEPTIDES</u></a>			
<a href="#"><u>28: PSM REPORT of potential novel PEP TIDES</u></a>			
<a href="#"><u>20: mzsqlite output to be used for PSM Visualization</u></a>			
<a href="#"><u>19: Peptide Shaker on data 16: PSM Report</u></a>			
<a href="#"><u>18: Peptide Shaker on data 16: Parameters</u></a>			
<a href="#"><u>17: Peptide Shaker on data 16: mzidentML file</u></a>			



## 5 Presenters and acknowledgements

---

### Presenters:

The main presenters were members of research team at the University of Minnesota, working an ongoing project developing Galaxy for multi-omic applications ([National Science Foundation Grant 1147079](#)). We have in-depth experience in Galaxy and its use for multi-omics data analysis.  
(z.umn.edu/galaxypreferences).

Speakers in our session include:

- [Tim Griffin](#), Associate Professor, and Faculty Director, CMSP, University of Minnesota. Dr. Griffin is the Principal Investigator on the project developing Galaxy for multi-omics. (ABRF member).
  - [Dave Clements](#), from Johns Hopkins University, coordinates training and outreach efforts for the Galaxy Project. As part of the Galaxy Project, he organizes meetings and courses, prepares training materials, and is involved in Galaxy's documentation, wiki, and web presence.
  - [Pratik Jagtap](#), Managing Director, CMSP, University of Minnesota. ABRF Member. Member of Protein Research Group (PRG).
  - [Getiria Onsongo](#), Analyst, Research Informatics Support Systems, University of Minnesota Supercomputing Institute. Dr. Onsongo is a Galaxy expert, working with users across the domains of transcriptomics and proteomics.
  - [Candace Guerrero](#), Post-Doctoral Researcher, University of Minnesota. Dr. Guerrero is a post-doctoral researcher in Tim Griffin's Lab and works on using Galaxy platform for mass spectrometry data analysis.
-

## Presenters and acknowledgements

### Thanks to our KILO Sponsors

Flexible, scalable, affordable genomics analysis for all biologists.

[globus.org/genomics](http://globus.org/genomics)

“At ICBi, we are working very closely with leading researchers to advance the frontiers of genomic science. By adopting Globus Genomics, we are much better positioned to deliver on our mission to enhance clinical and translational research at the medical center.”

Dr. Subha Madhavan  
Director of the Innovation Center for Biomedical Informatics  
Georgetown University Medical Center

**Thermo SCIENTIFIC**  
A Thermo Fisher Scientific Brand

**More Proteins. More accurately. Faster than ever.**

Biology is complex and understanding it is a big challenge. Identify and quantify more proteins and complexities such as PTMs faster and more accurately with our new portfolio of LC-MS instruments, sample prep solutions and software. HRAM solutions using Thermo Scientific™ Orbitrap™ MS quantifies all detectable proteins and peptides with high specificity and fewer false positives, while triple quadrupole MS delivers SRM sensitivity and speed to detect targeted proteins more quickly. Join us in meeting today's challenges. Together we'll transform proteomics.

**Quantitation transformed.**

- Attend our sponsored workshop: (SW4) The Galaxy Platform for Multi-Omic Data Analysis and Informatics

Also thanks to Amazon Web Services Education Research Grant.