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℠

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

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Workshop instructors and acknowledgements

- **Instructors**
  - Dave Clements – core Galaxy team (Johns Hopkins University)
  - Candace Guerrero
  - Getiria Onsongo
  - Pratik Jagtap

  } Galaxy-P team (University of Minnesota)

- **Support**
  - Tom McGowan, James Johnson (JJ), Ben Lynch (University of Minnesota)
  - Karen Reddy, Mo Heydarian (Johns Hopkins University)

- **Sponsors**

Topic of workshop: multi-omics

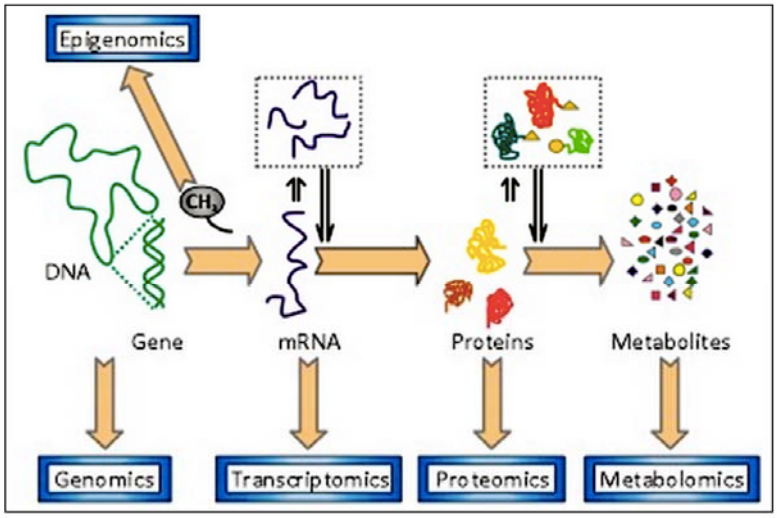


Image Source: Goodacre, J. Exp. Bot 2005.

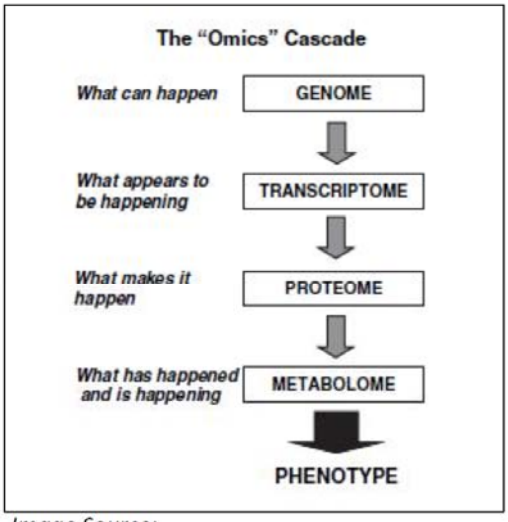


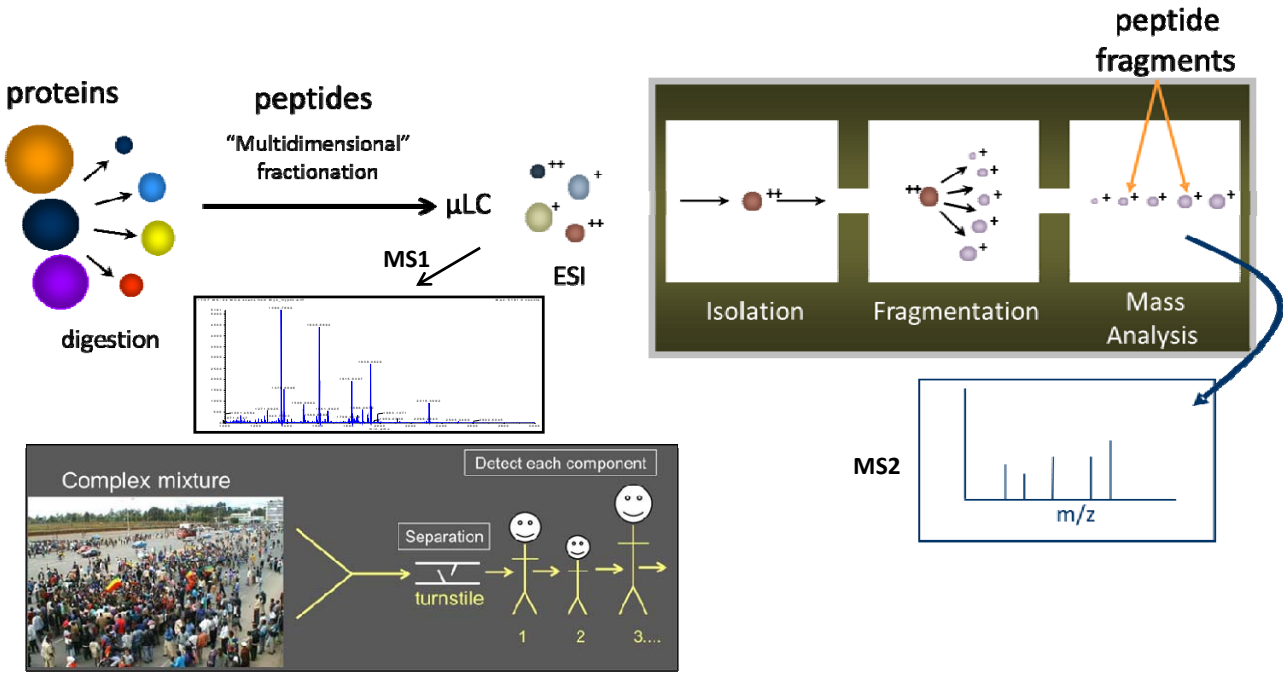
Image Source: http://fluorous.com/images/omics.JPG

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Technologies: Proteomics

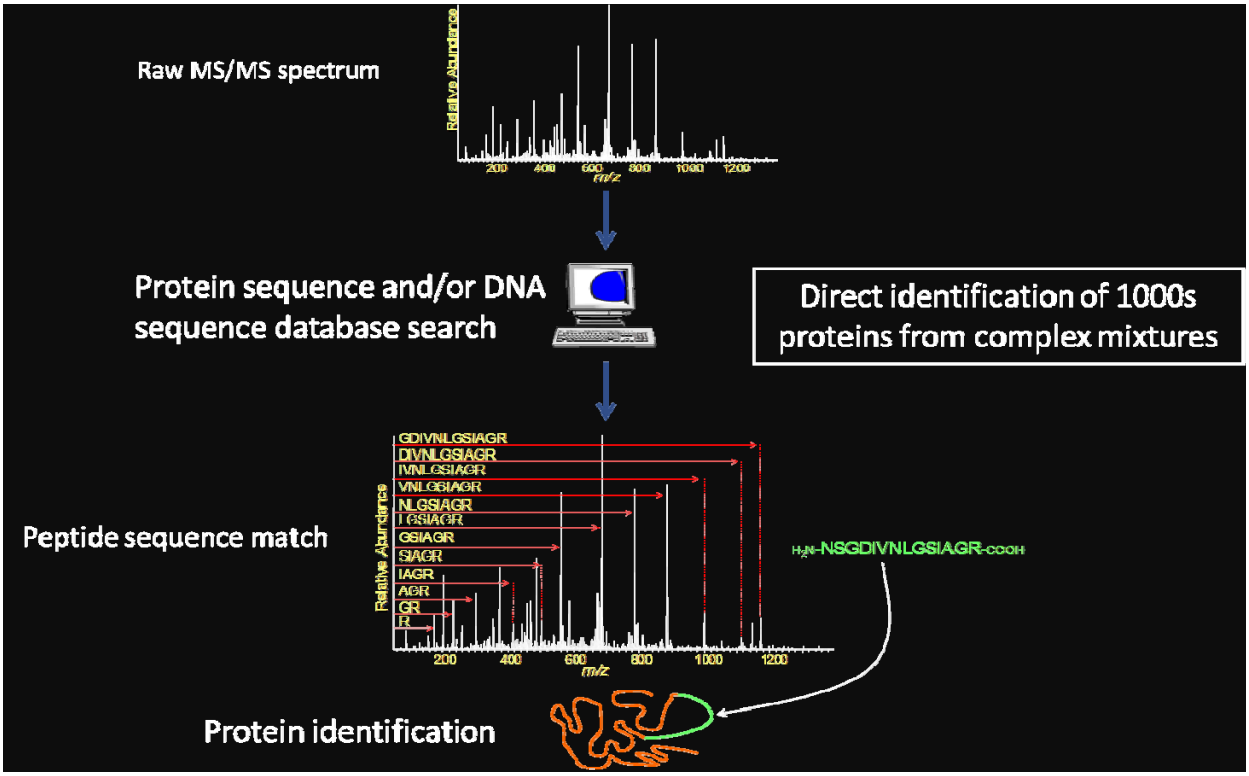
Peptide fractionation coupled to tandem mass spectrometry (MS/MS)



GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

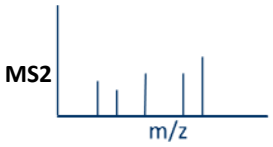
From sticks on a graph to protein identities



GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Inferring protein identity from peptide sequence matches



CAQCHTVEK

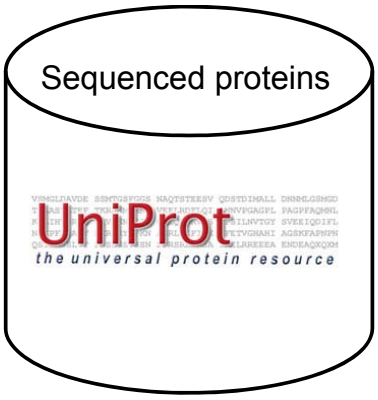
Cytochrome C

$_{NH_2}$GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGL
FGRKTGQAPGFTYTDANKNKGITWKEETLMEYLENPK
KYIPGTKMIFAGIKKKTEREDLIAYLKKATNE$_{COOH}$

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

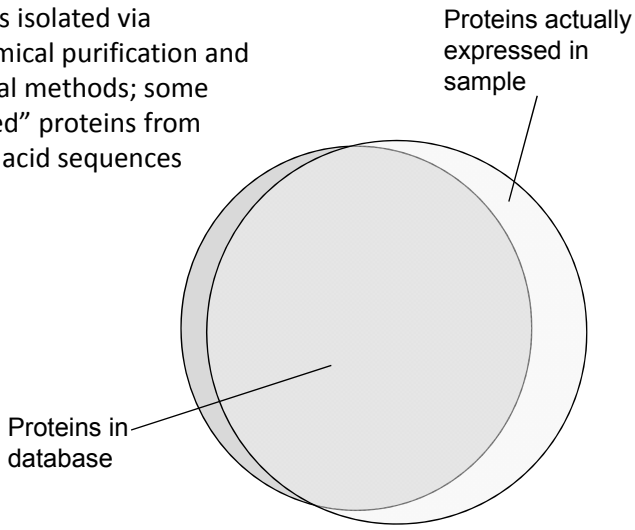
MS-based proteomics only as good as the database used…



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

-- Sir Isaac Newton

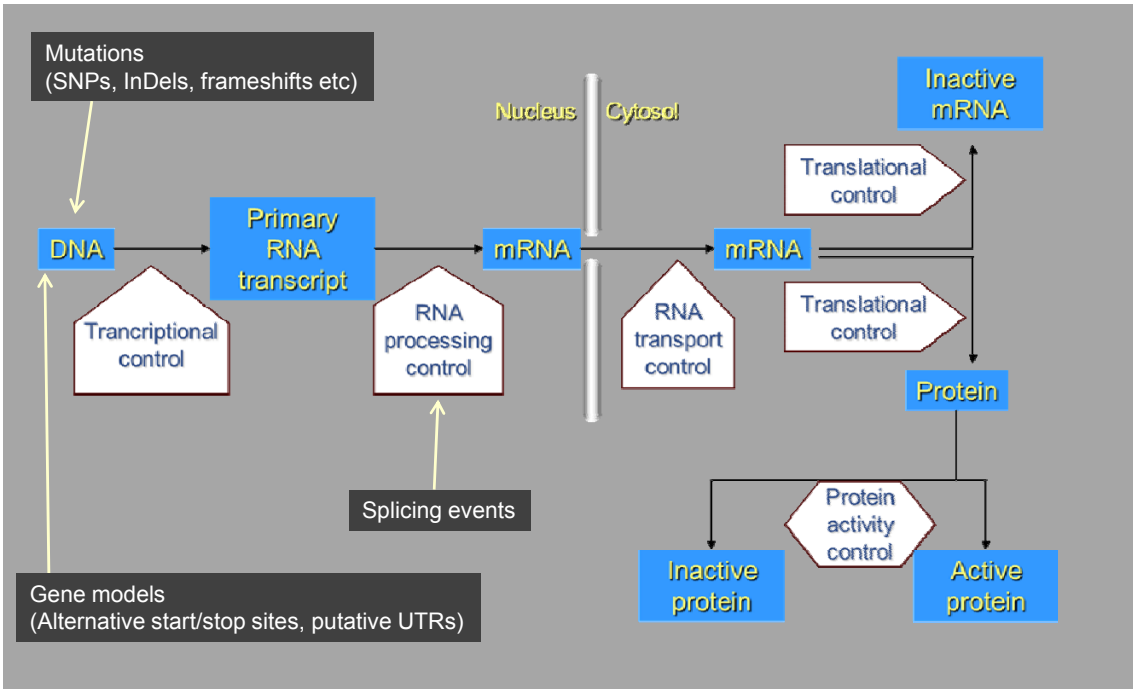
- Generally includes canonical protein sequences or single proteins isolated via biochemical purification and chemical methods; some "inferred" proteins from nucleic acid sequences



GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

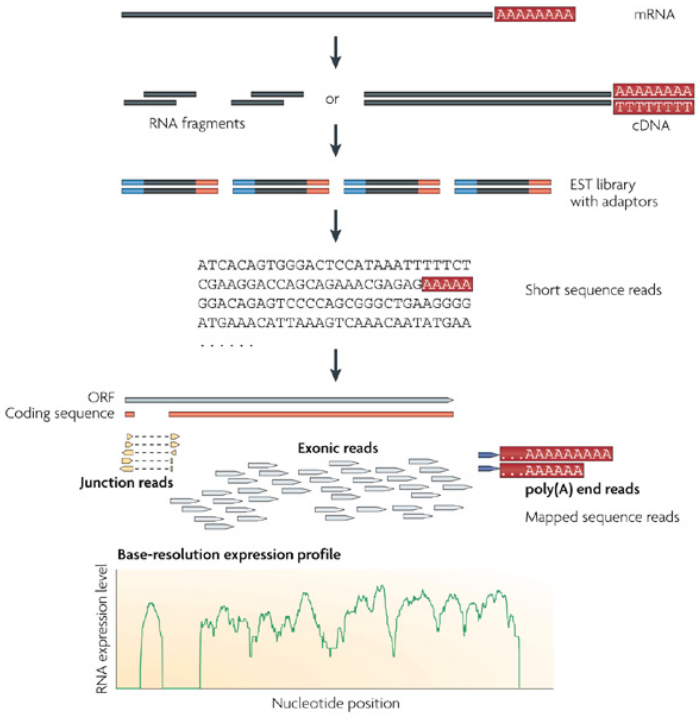
Sources of un-expected protein sequences?



GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Technologies: RNA-seq

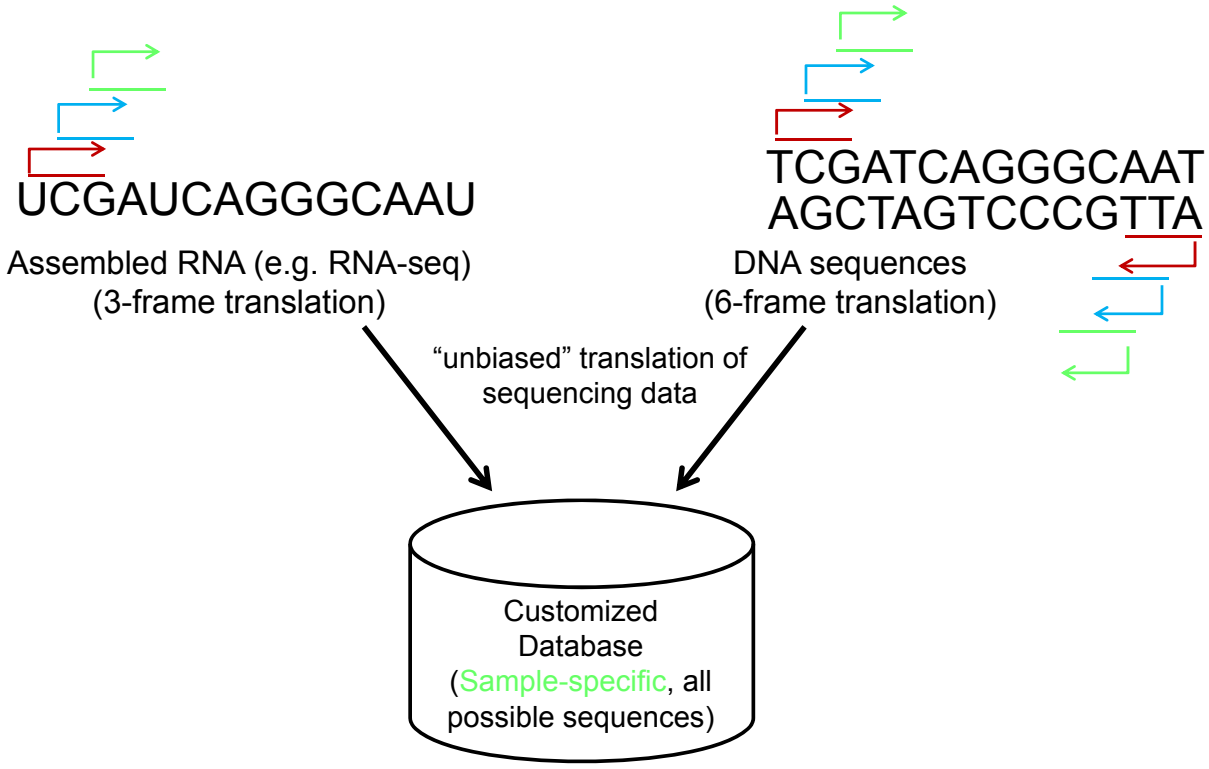


Nature Reviews | Genetics 10, 57-63

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

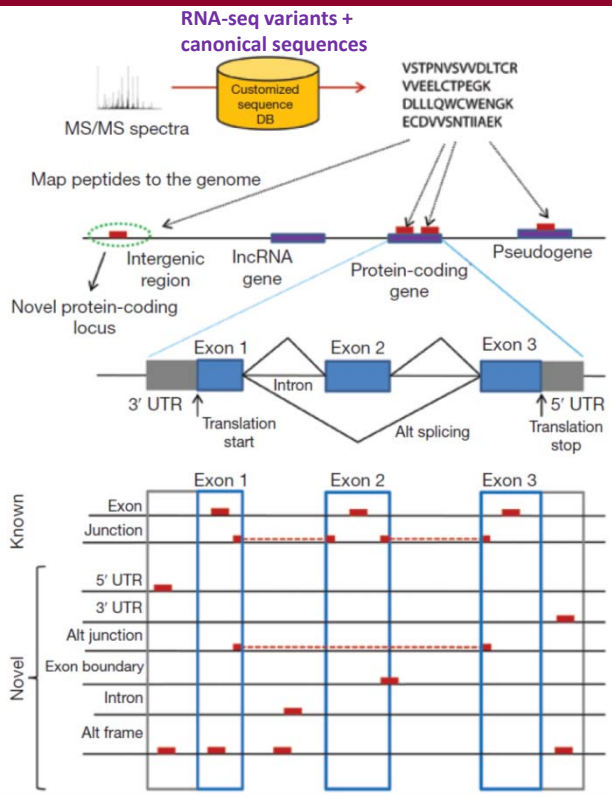
A multi-omic example: Integrating Next-Gen sequencing and proteomics



GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

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



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

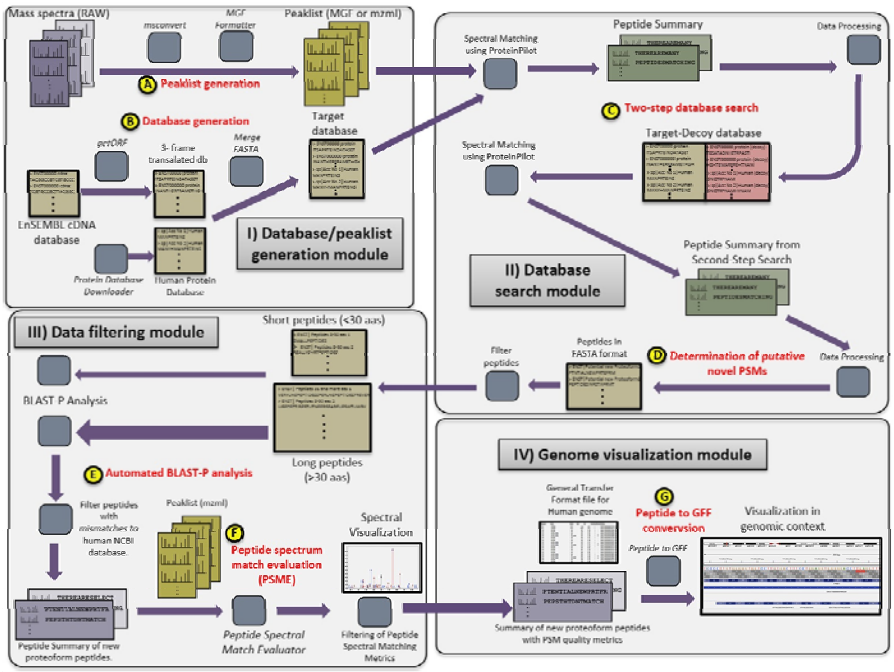
Nesvizhskii NATURE METHODS | VOL.11 NO.11 | NOVEMBER 2014 |

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Driven by bioinformatics

- Many step workflow for analyzing different types of data



J. Proteome Res. 2014, 13, 5898−5908

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

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℠

The Galaxy interface

Tools

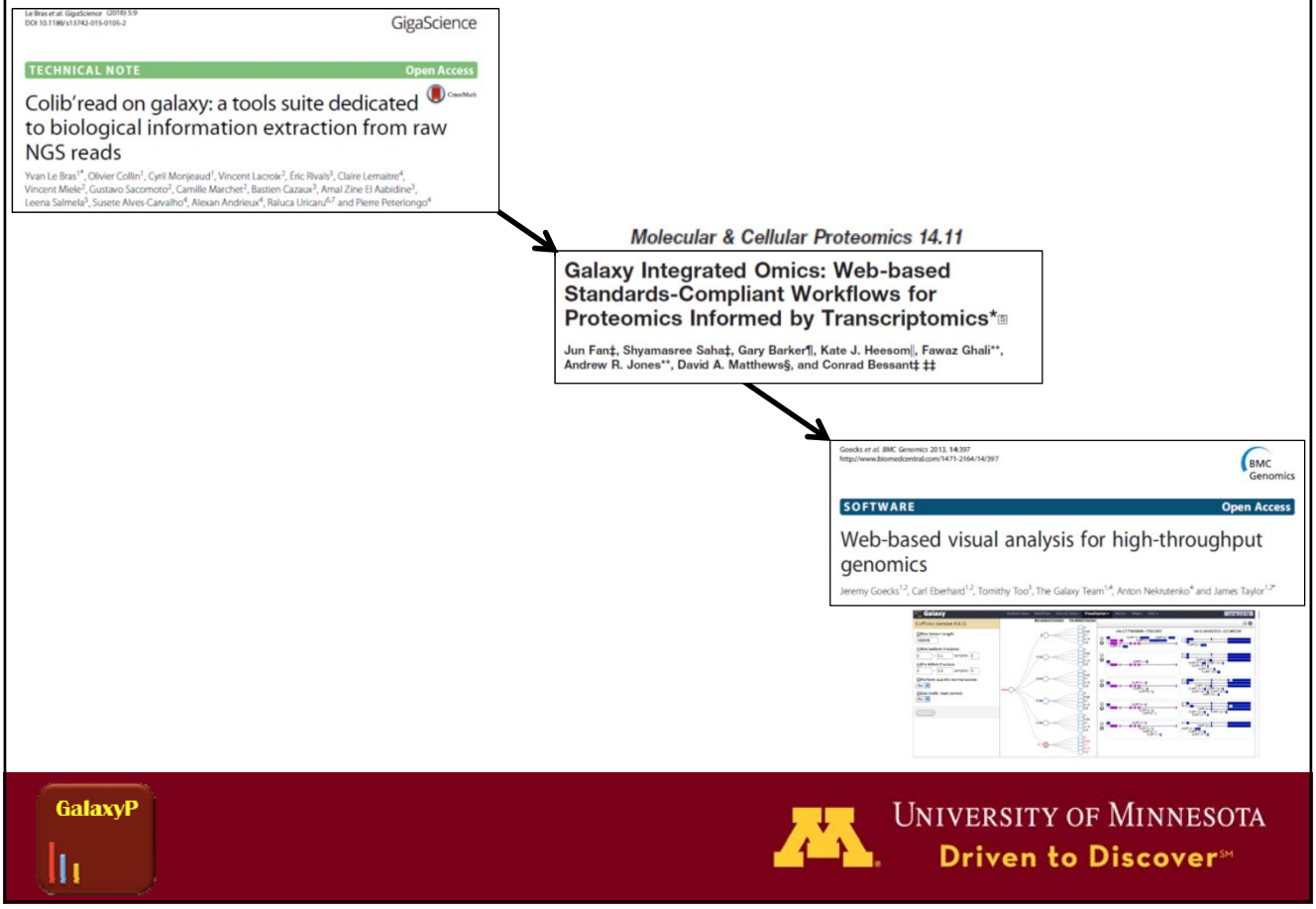
Main viewing window
(workflow development, results visualization etc)

History

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Enabling integration of tools across the ‘omic domains



Built for the community: sharing software tools

Galaxy Tool Shed

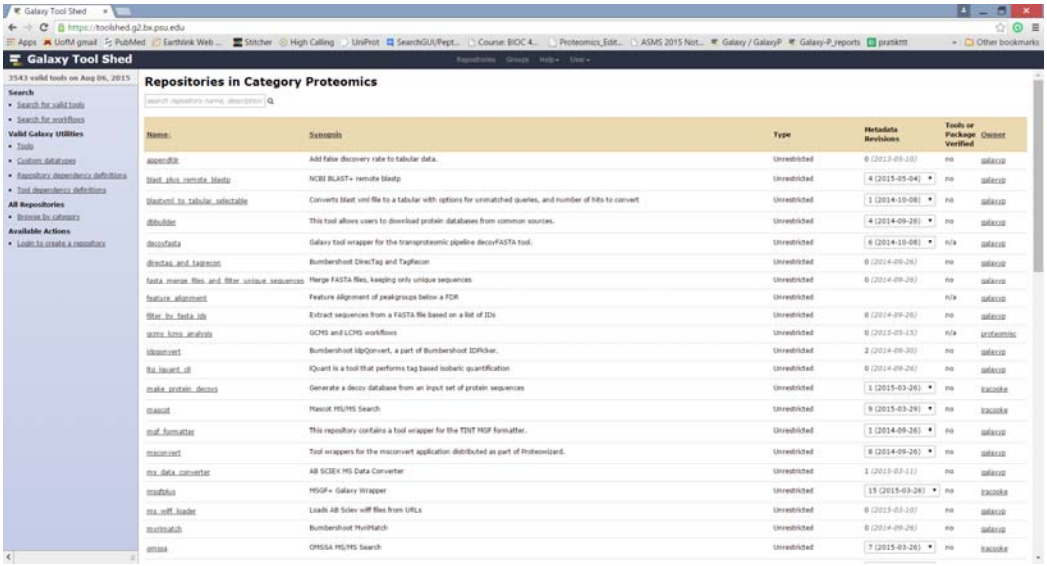

Blankenberg et al. Genome Biology 2014, 15:403
http://genomebiology.com/2014/15/2/403

Genome Biology

OPEN LETTER

Dissemination of scientific software with Galaxy ToolShed

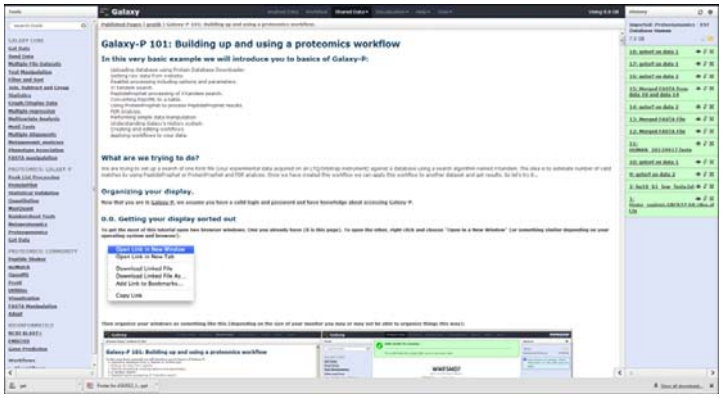
Daniel Blankenberg[1*], Gregory Von Kuster[1,4], Emil Bouvier[1,4], Dannon Baker[2,4], Enis Afgan[4,5], Nicholas Stoler[3], the Galaxy Team[4], James Taylor[2,4*] and Anton Nekrutenko[1,4*]



GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Built for the community: sharing complete workflows



URL export

File export (.ga)

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

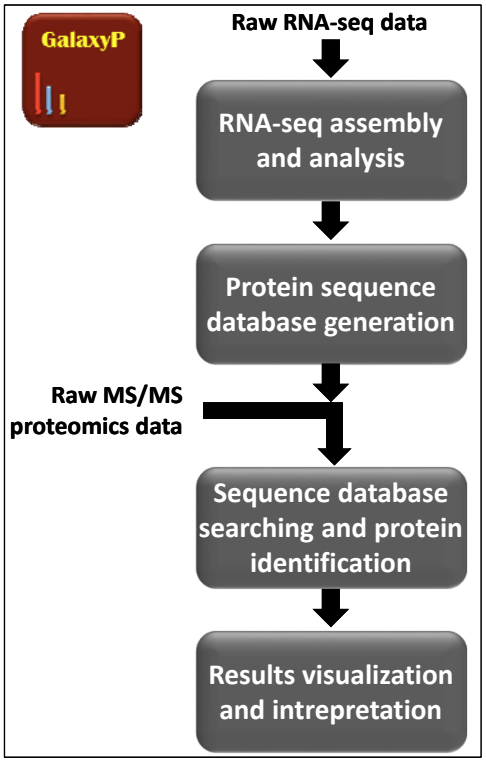
WORKFLOW: https://galaxyp.msi.umn.edu/u/pjagtap/w/workflow-for-4-plex-itraq-xtandem-search-scaffold-processing

Galaxy-Workflow-Workflow_for_4-plex_iTRAQ_X_tandem_Search_Scaffold_Processing.**ga**

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Structure of workshop



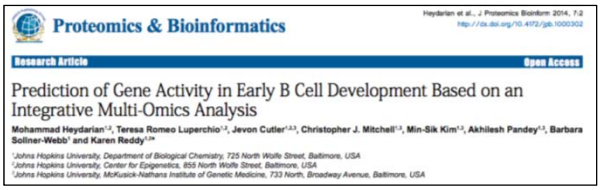
Session 1

Session 3

Session 2

Session 4

- RNA-seq and quantitative MS-proteomics data:


Proteomics & Bioinformatics

Research Article Open Access

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

- Documentation to complement hands-on instruction

GalaxyP

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)

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

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
  - Password: password
- Ensure that Java 7 or later version is installed on your computer

GalaxyP

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Galaxy Tutorial – Basics of Data Analysis using Galaxy platform, Condensed

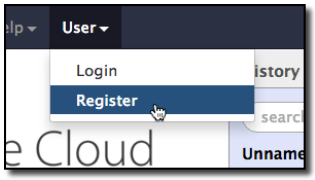
(SW4) The Galaxy Platform for Multi-Omic Data Analysis and Informatics

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

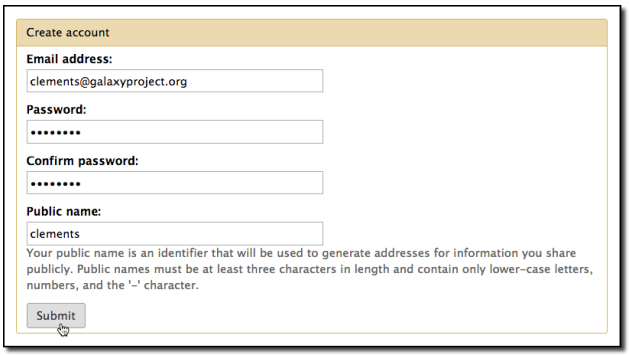
ABRF 2016

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 HTPPS, 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.)

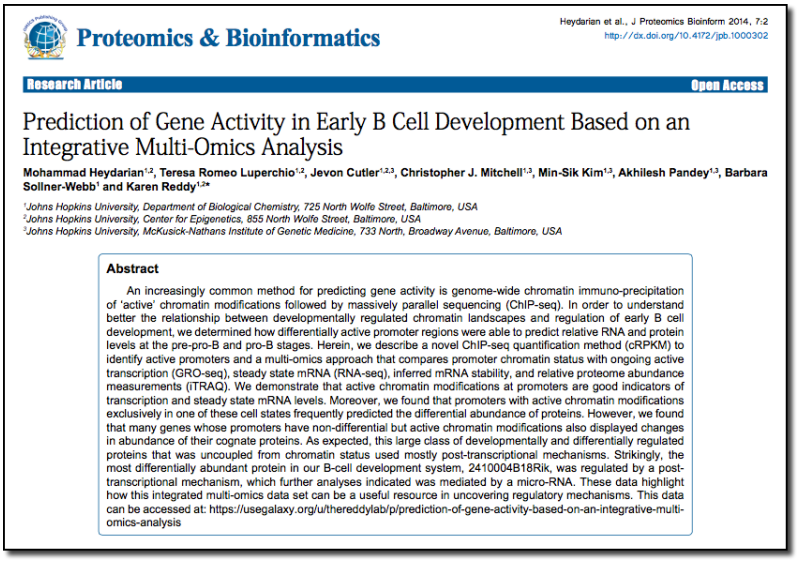
Reload the page by *clicking* **Analyze Data** or **Return to the home page**.

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


Heydarian et al., J Proteomics Bioinform 2014, 7:2
http://dx.doi.org/10.4172/jpb.1000302

Proteomics & Bioinformatics

Research Article Open Access

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

Mohammad Heydarian[1,2], Teresa Romeo Luperchio[1,2], Jevon Cutler[1,2,3], Christopher J. Mitchell[1,3], Min-Sik Kim[1,3], Akhilesh Pandey[1,3], Barbara Sollner-Webb[1] and Karen Reddy[1,2]*

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

Abstract

An increasingly common method for predicting gene activity is genome-wide chromatin immuno-precipitation of 'active' chromatin modifications followed by massively parallel sequencing (ChIP-seq). In order to understand better the relationship between developmentally regulated chromatin landscapes and regulation of early B cell development, we determined how differentially active promoter regions were able to predict relative RNA and protein levels at the pre-pro-B and pro-B stages. Herein, we describe a novel ChIP-seq quantification method (cRPKM) to identify active promoters and a multi-omics approach that compares promoter chromatin status with ongoing active transcription (GRO-seq), steady state mRNA (RNA-seq), inferred mRNA stability, and relative proteome abundance measurements (iTRAQ). We demonstrate that active chromatin modifications at promoters are good indicators of transcription and steady state mRNA levels. Moreover, we found that promoters with active chromatin modifications exclusively in one of these cell states frequently predicted the differential abundance of proteins. However, we found that many genes whose promoters have non-differential but active chromatin modifications also displayed changes in abundance of their cognate proteins. As expected, this large class of developmentally and differentially regulated proteins that was uncoupled from chromatin status used mostly post-transcriptional mechanisms. Strikingly, the most differentially abundant protein in our B-cell development system, 2410004B18Rik, was regulated by a post-transcriptional mechanism, which further analyses indicated was mediated by a micro-RNA. These data highlight how this integrated multi-omics data set can be a useful resource in uncovering regulatory mechanisms. This data can be accessed at: https://usegalaxy.org/u/thereddylab/p/prediction-of-gene-activity-based-on-an-integrative-multi-omics-analysis

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."

The data

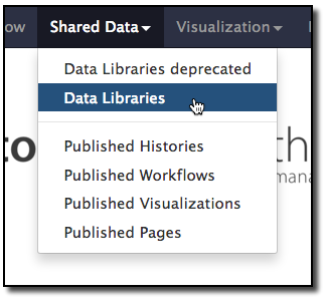
We have two RNA-Seq datasets, one from mouse pre-pro-B cells, and one from mouse pro-B cells. Each dataset consists of ~90 million single-end reads. Sequencing was done an Illumina HiSeq 2000 and all reads are 97bp long.

We are going to use striiped 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.*

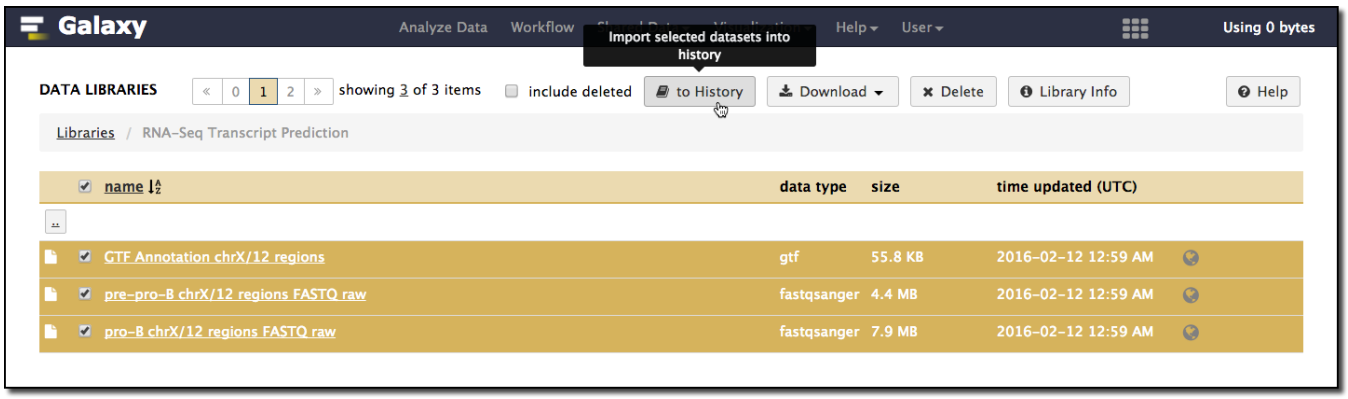
Get the data / Data libraries

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.

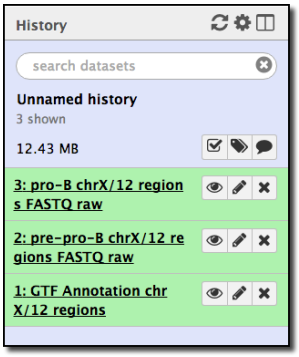
This will ask which history to import these datasets into. So far, you only have one ("Unnamed history"). *Click* **Import** to add them to that history.



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

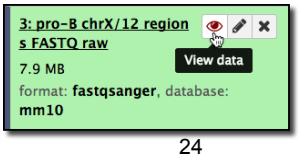
FASTQ datasets

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.



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.

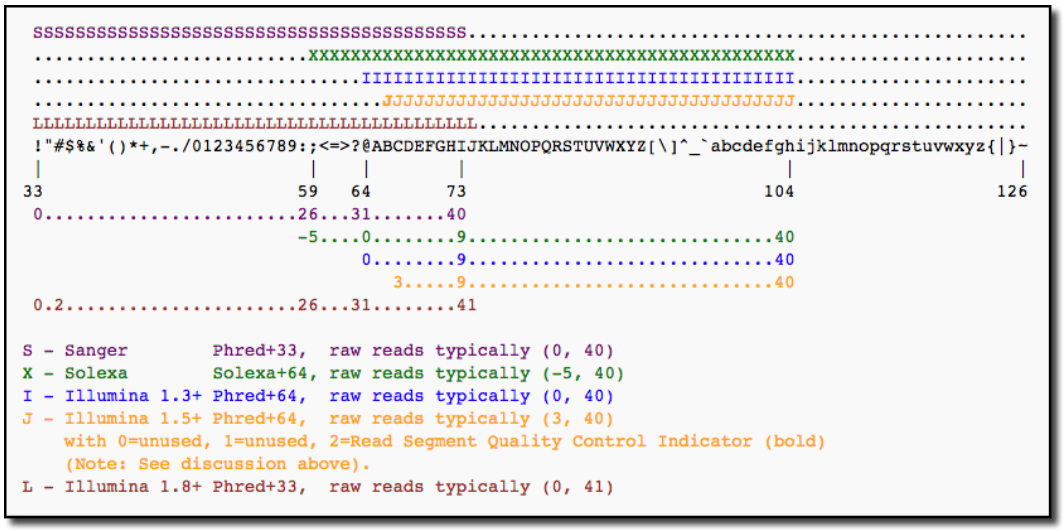
To view the full dataset, *poke it in the eye*. This will display the content of this FASTQ dataset in the middle panel.



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

Phred quality scores 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:

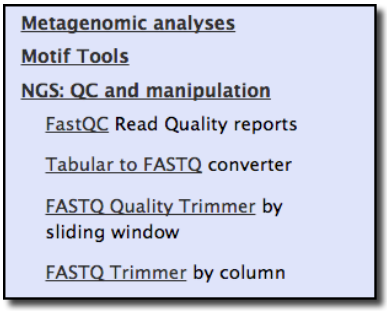
Score	Instrument thinks it will be wrong	
10	10% of the time	(1 out of 10 times)
20	1% of the time	(1 out of 100 times)
30	0.1% of the time	(1 out of 1000 times)
40	0.01% of the time	(1 out of 10000 times)

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 `I` for 40.

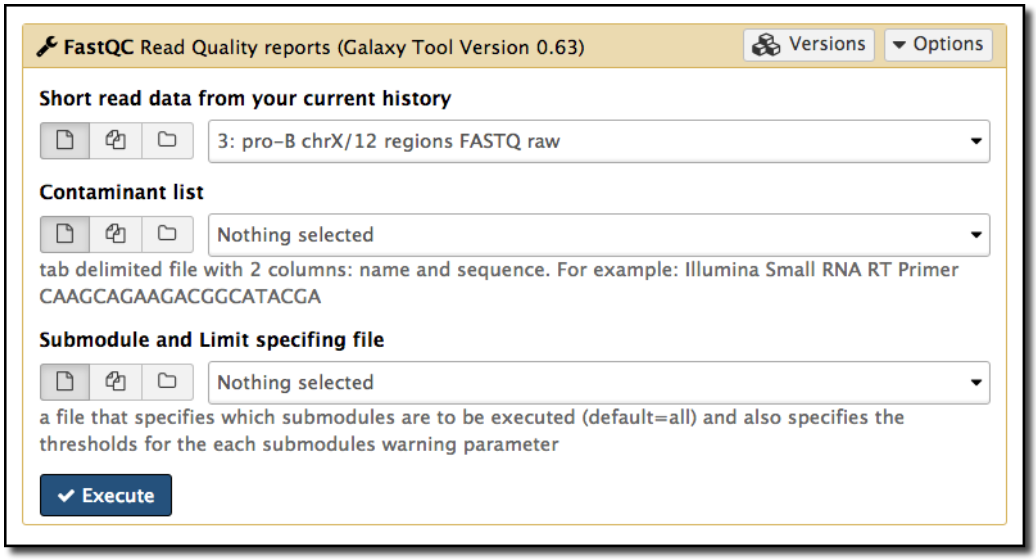
Check FASTQ dataset quality

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



areas that may need to be addressed.
This is a typical, if simple, Galaxy tool form. Tool pages generally have these sections

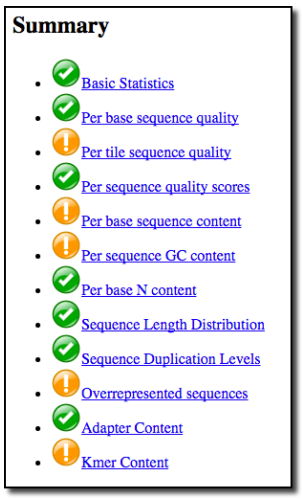
- *Versions and options* - If multiple versions are supported, there will be link to run the others. And the options menu enables you see more about the tool and how it is wrapped in Galaxy.
- *The form* - where you configure and run the tool
- *Documentation* - Describes the tool and provides help. 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.

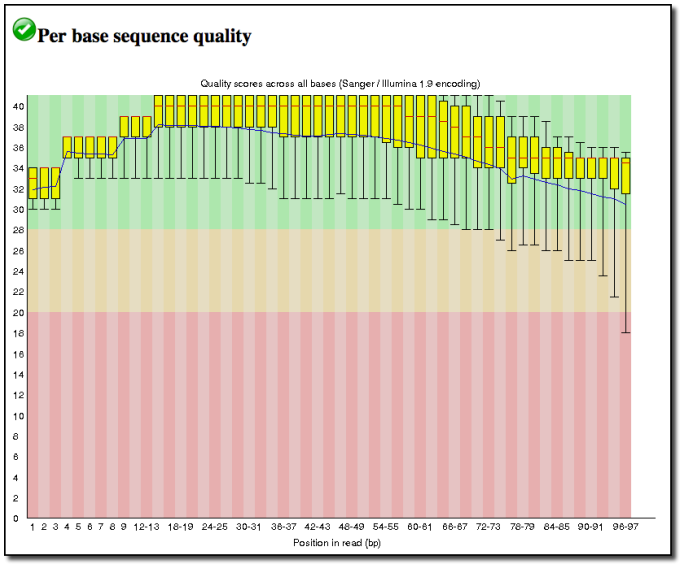


Once a tool has finished running you can preview the output datasets (by clicking on the dataset name) or view the data itself (by poking it in the eye). *Poke* the **FastQC Web page** dataset *in the eye*.

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.

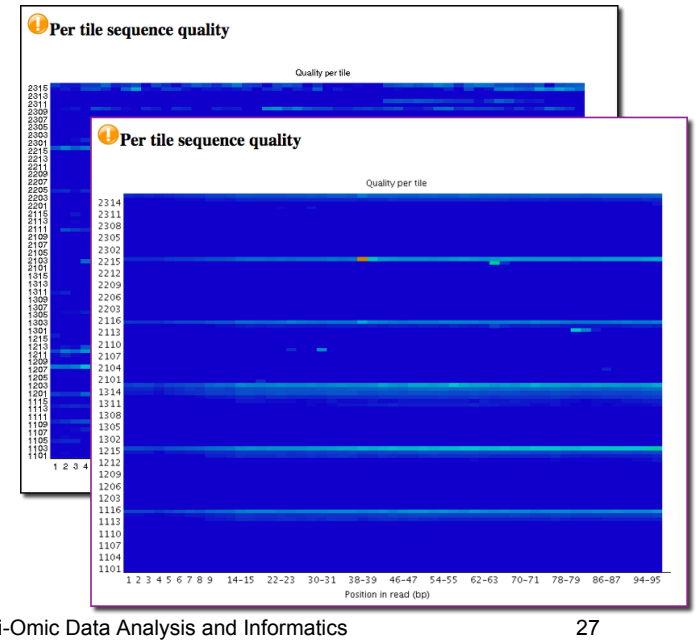
Per base sequence quality (boxplot)



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.

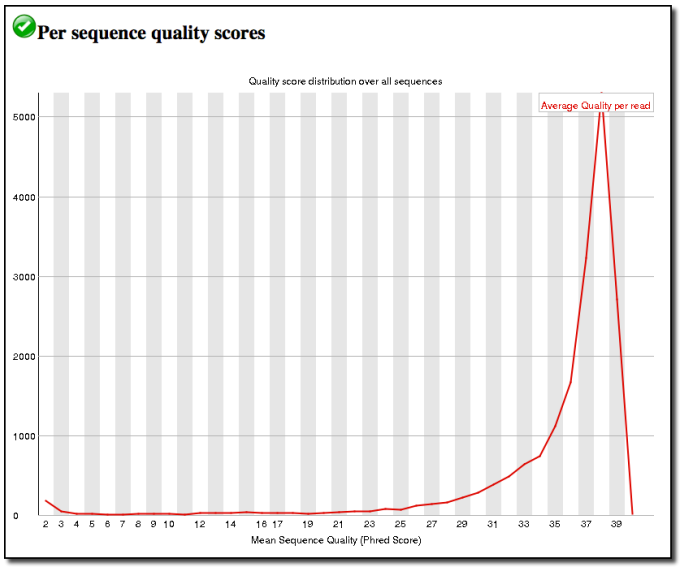
Per tile sequence quality



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.

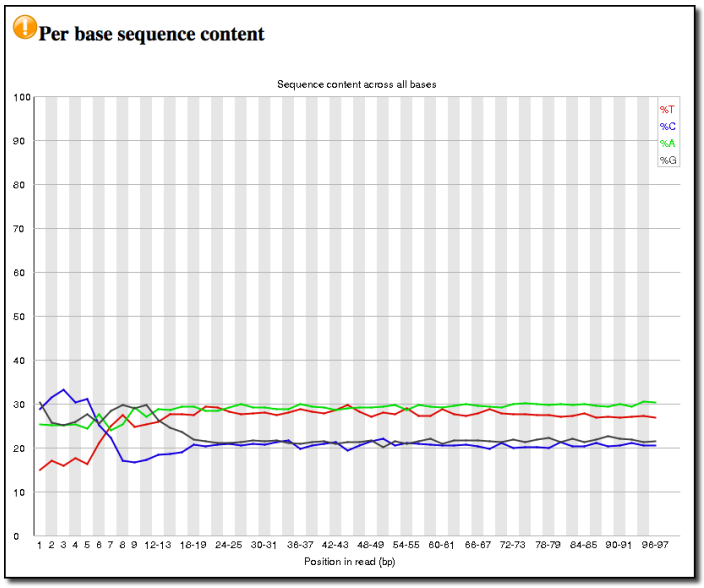
Per base sequence content

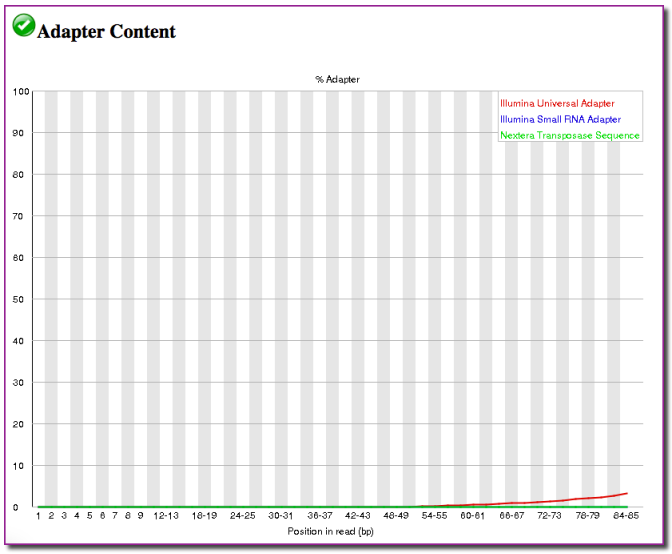
This graph is good at highlighting untrimmed adapter or barcode problems.

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.

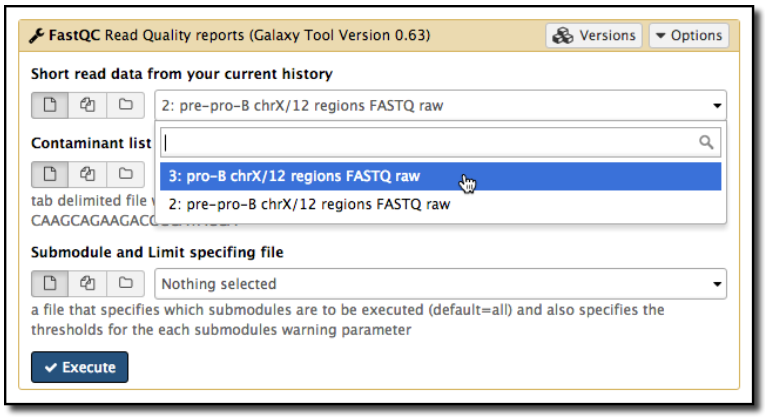
And the rest

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.

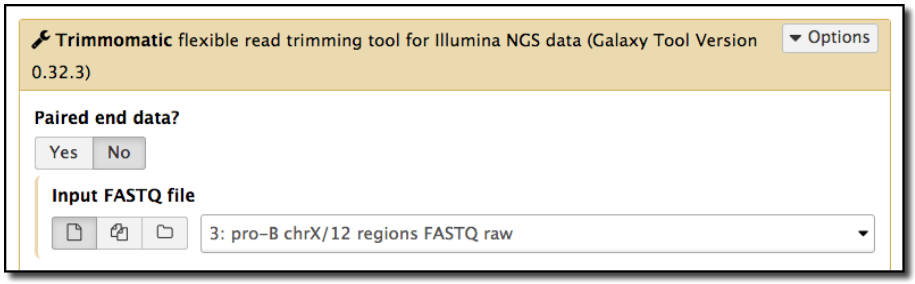
Preview one of the **FastQC** datasets in your history by *clicking* on its name. In the preview *click* on the **Run this job again** looping arrow icon. This launches the tool form in the middle panel and pre-populates it with the same settings it was run with before.



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

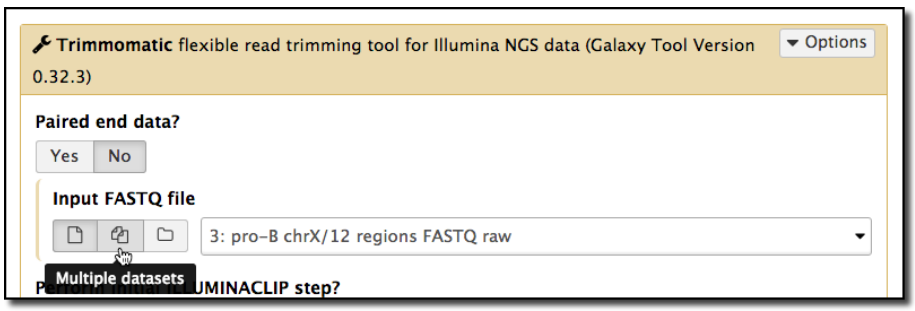
Use Trimmomatic to separate the good from the bad.



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.

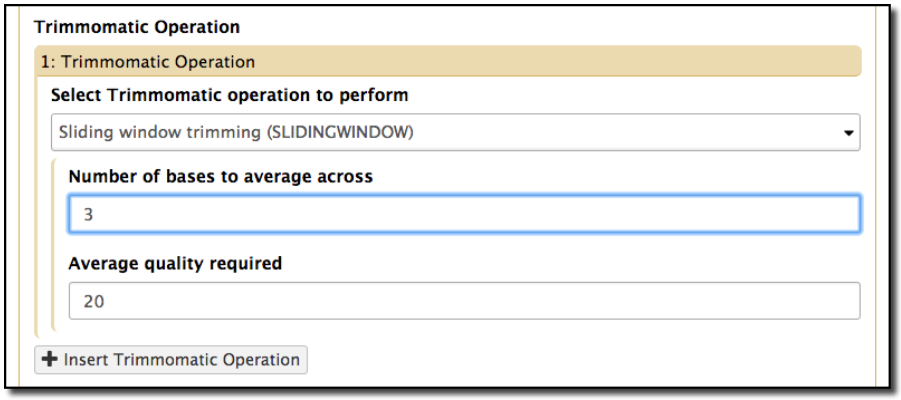
Select Trimmomatic Filters

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.

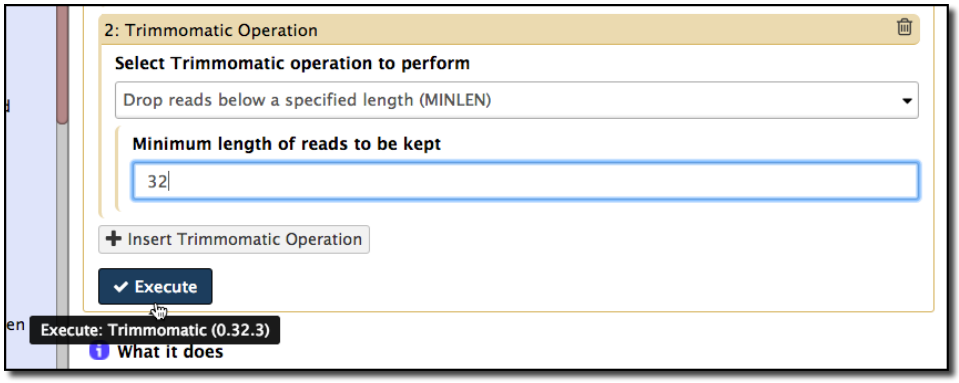
First, *select* **Yes** under **Perform initial ILLUMINACLIP step?**. Then *change* the **Adapter sequences to use** pulldown to **TruSeq3 (single-ended, for MiSeq and HiSeq)**. Leave other parameters at the default.

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



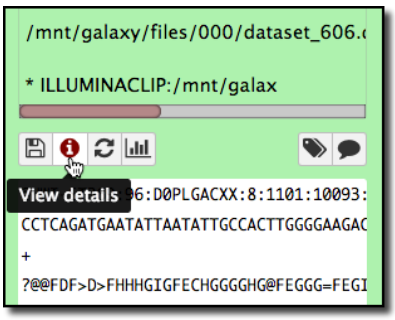
To deal with reads that are too short add another operation.

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. C*lick* **Execute**.



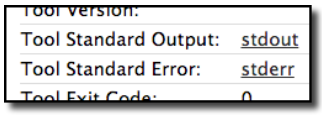
Trimmomatic Results

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



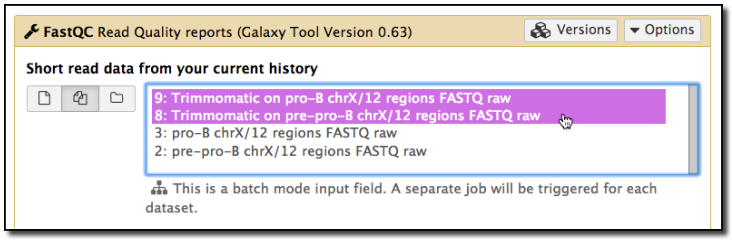
To see the full Trimmomatic summary, *click* the ⓘ **icon (View details)** in the dataset preview. This displays the metadata for this dataset.

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.

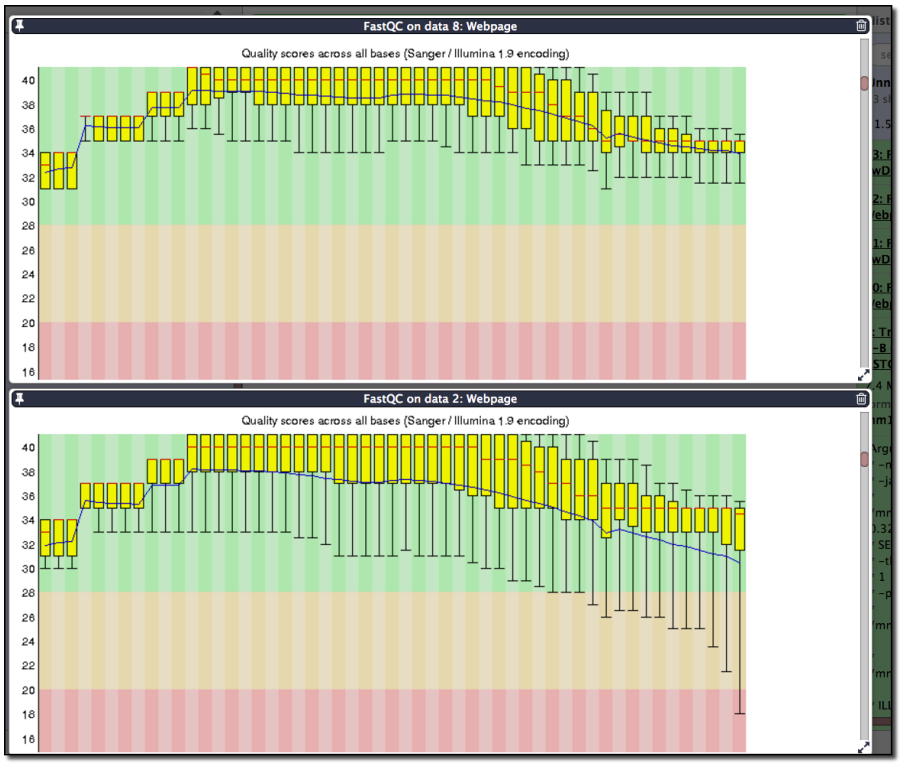
Click **FastQC** in the tool panel, *select* the **Multiple datasets icon** under **Short read data from your current history**, and then *select* both **Trimmomatic on …** datasets. *Click* **Execute.**



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

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.

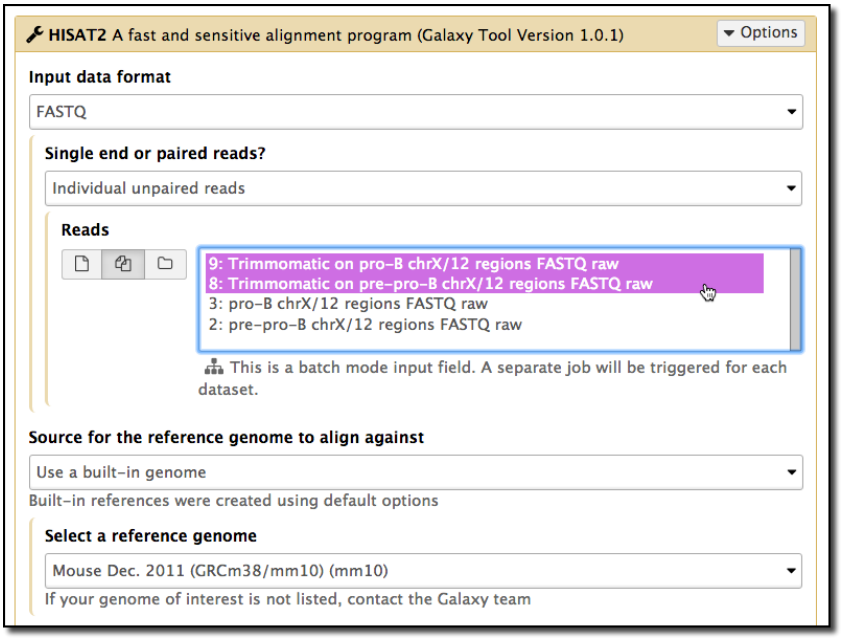
To switch back to the standard view either click outside the displayed datasets or poke the yellow eye ( ). To disable the Scratchbook display, click on the yellow grid icon.

Map the reads

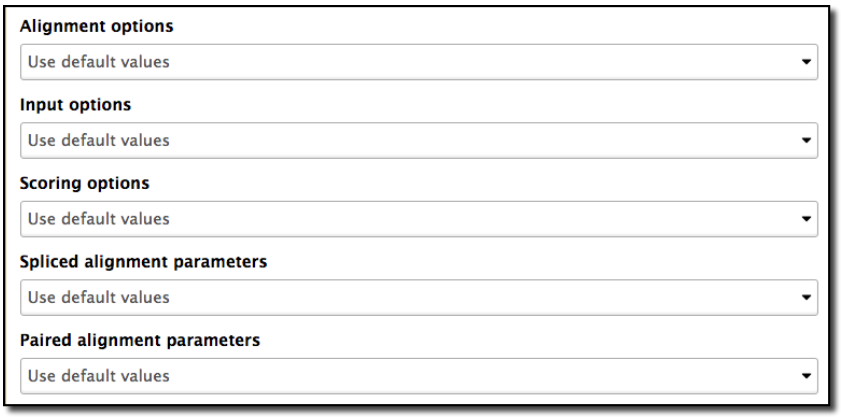
The data has now been cleaned and we are ready to map them against the mouse reference genome.

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.

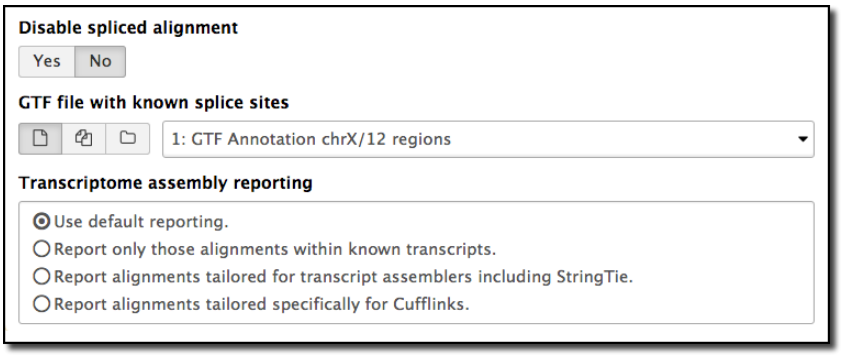


HISAT is an "option-rich" tool. The HISAT tool wrapper takes a common approach of placing them in sections that can be expanded as needed. HISAT has 5 such sections:



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**.



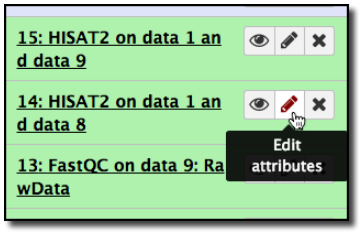
HISAT2 Results

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.

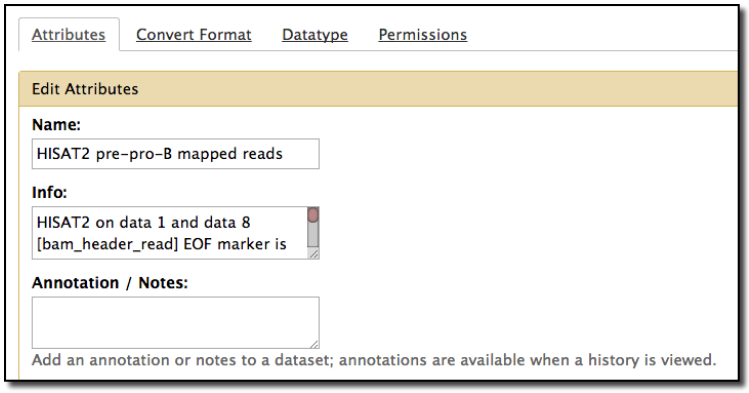
```
[bam_header_read] EOF marker is absent. The input is probably truncated.
[samopen] SAM header is present: 66 sequences.
31843 reads; of these:
  31843 (100.00%) were unpaired; of these:
    482 (1.51%) aligned 0 times
    26747 (84.00%) aligned exactly 1 time
    4614 (14.49%) aligned >1 times
98.49%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

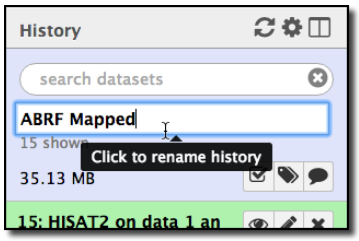
Updating dataset and history metadata: Renaming



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.**



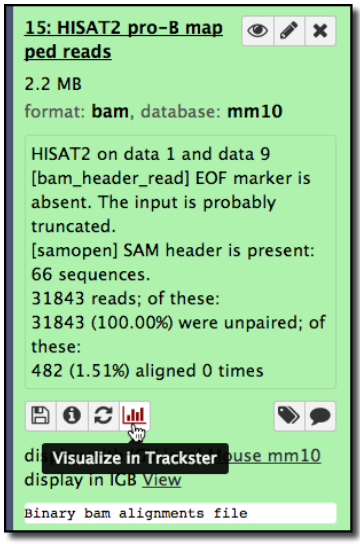


We should also give our history a name. A best practice is to always name your histories. To do this, *click* on **Unnamed history**, *enter* a new and informative history name, and then *press* **return** or **enter**.

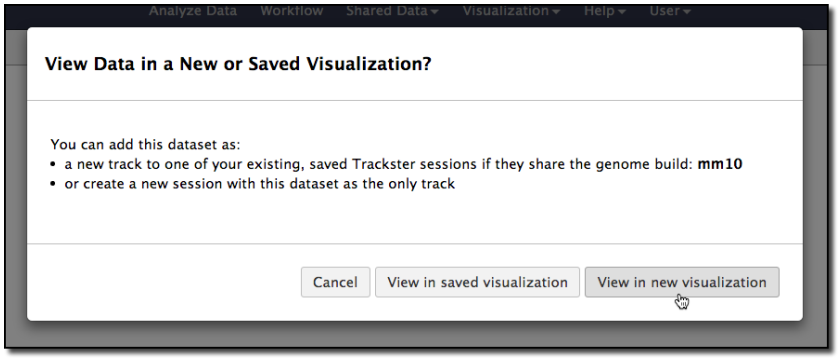
But what do the HISAT2 mappings look like?

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

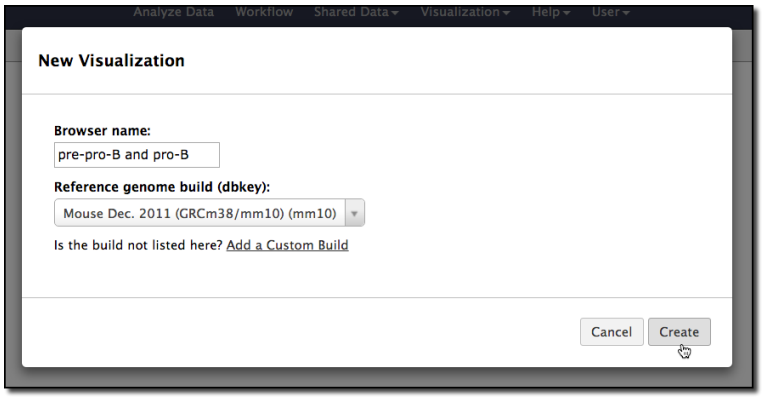
We can visually inspect the mappings in the context of the genomic regions they mapped to. Later today you'll use the IGV desktop program to visualize genomic data. Right now, let's learn how to do this inside Galaxy, using the Trackster visualization tool.



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

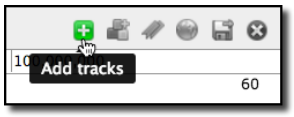


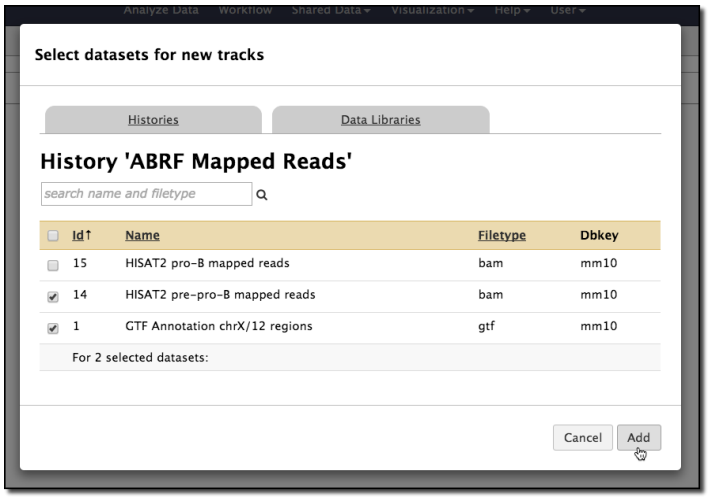
Select **View in new visualization** and then give the visualization a meaningful name and *click* **Create**.



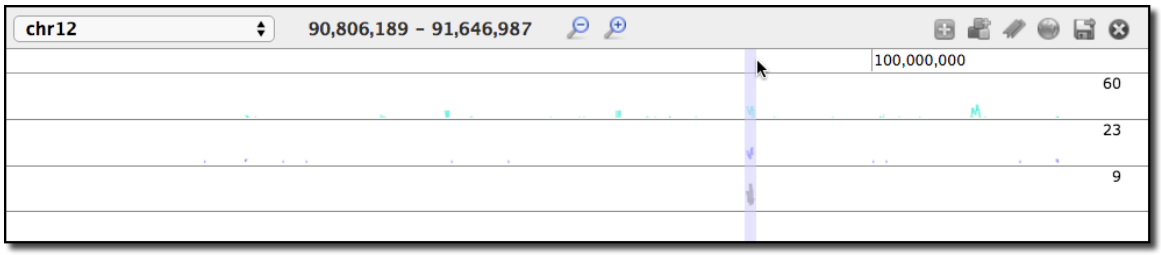
This will bring up a new Trackster browser with a status message saying that it is indexing the dataset for display. *Switch* from chr1 to **chr12** (or chrX), where we have mapped data.

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.

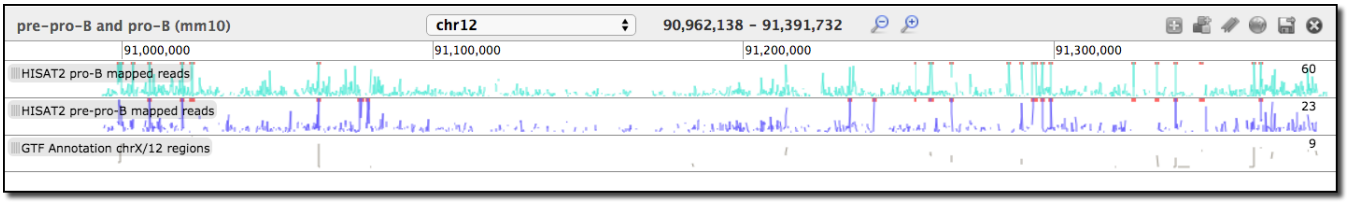




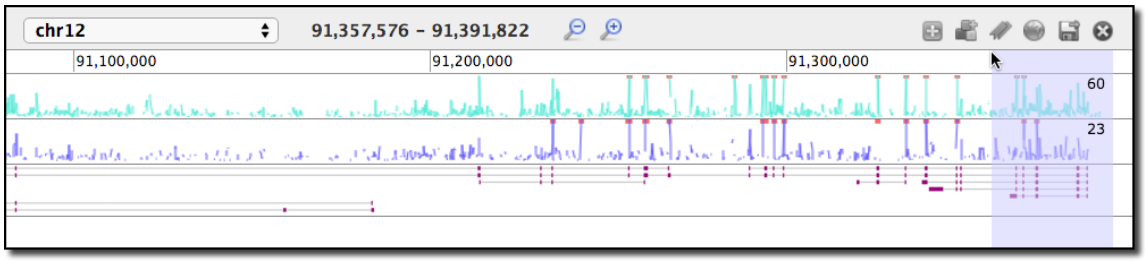
Once those new tracks have loaded, zoom in on the region of interest - the area with the annotations: *Click and drag* on the coordinates track (near the top) and include the entire region with annotation.

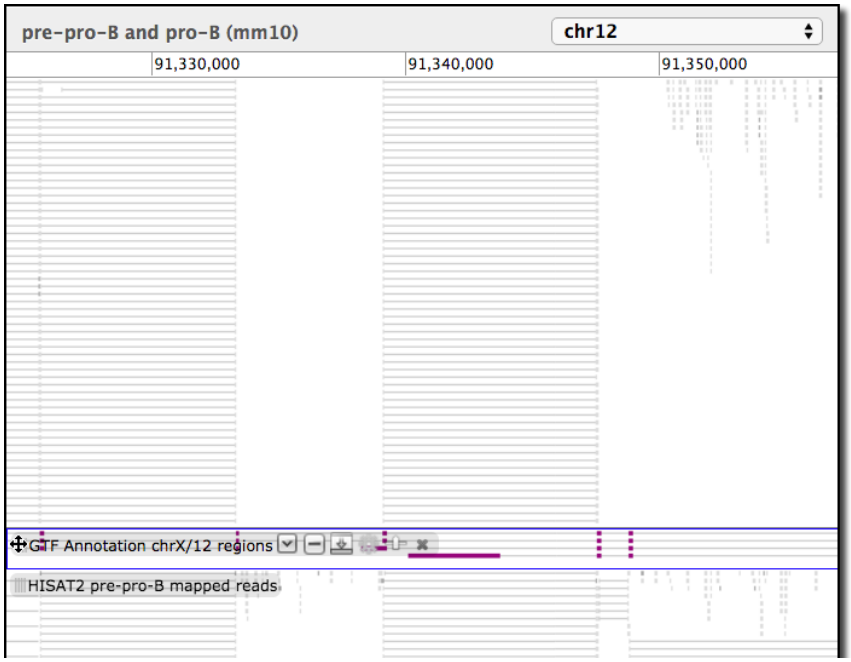


You may have to repeat that a few times to get the region of interest:



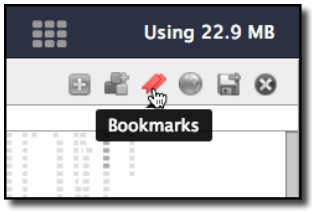
The top 2 tracks show the read depth for any position in this region, and the bottom track shows the annotation (but not very well). *Hover* over the **GTF Annotation chrX/12 regions** track and *click* the ⌄ **Set display mode icon** and *select* **Squish**. To get a better idea of what the aligned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

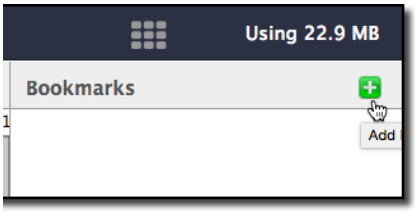


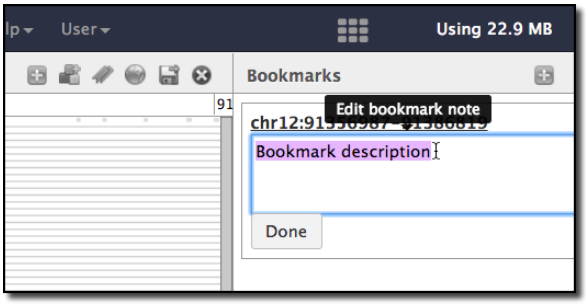


This changes the display mode for all tracks and now we can see individual reads, including where splice junction were introduced to map the reads. Drag the annotation track to the middle by *clicking and dragging* the **|||| vertical lines icon** at the left of the track title upwards to between the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



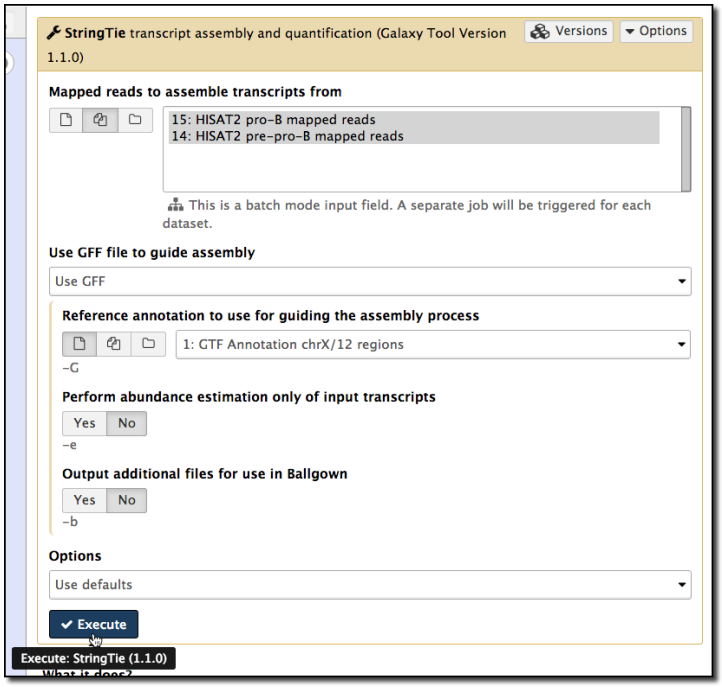




Transcript Prediction with StringTie

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**.

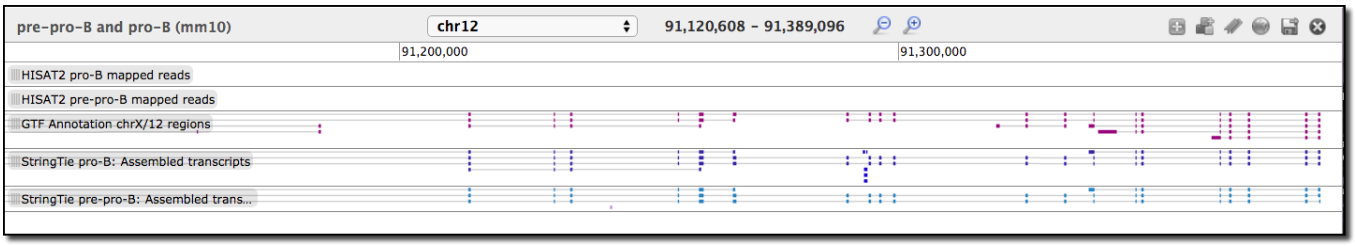


StringTie generates 3 output datasets for each input dataset:

- 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.

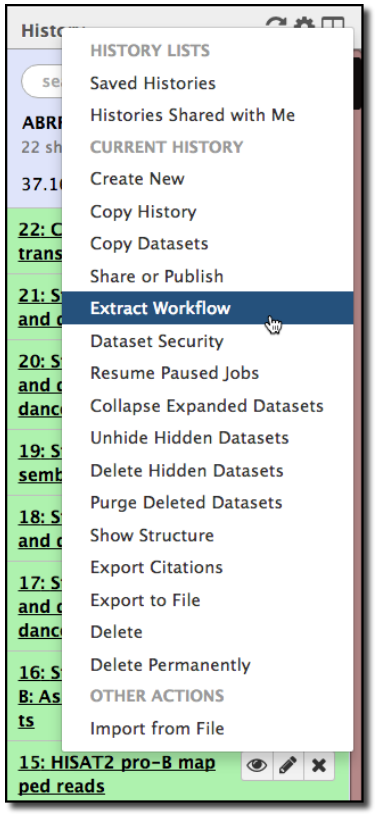
Save the visualization and go back to **Analyze Data**.

And there is more!

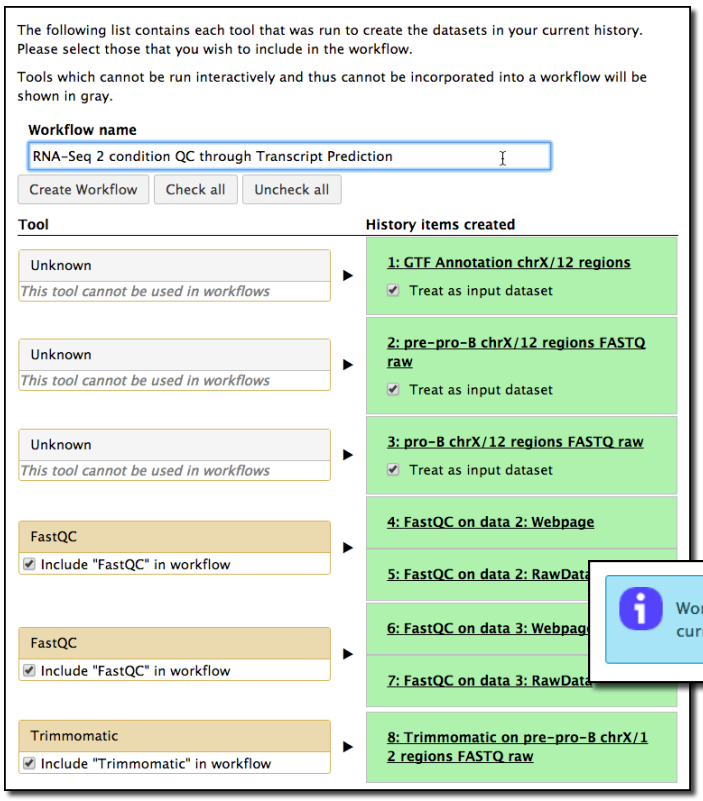
Workflows: repeating an analysis

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.

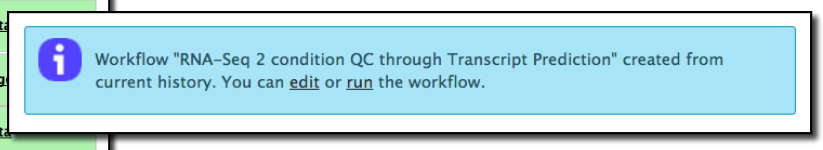
Galaxy supports *de novo* workflow creation, and creating them from histories. To create a workflow from our history, *click* the **cog** at the top of the history panel, and then *select* **Extract Workflow.**



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.

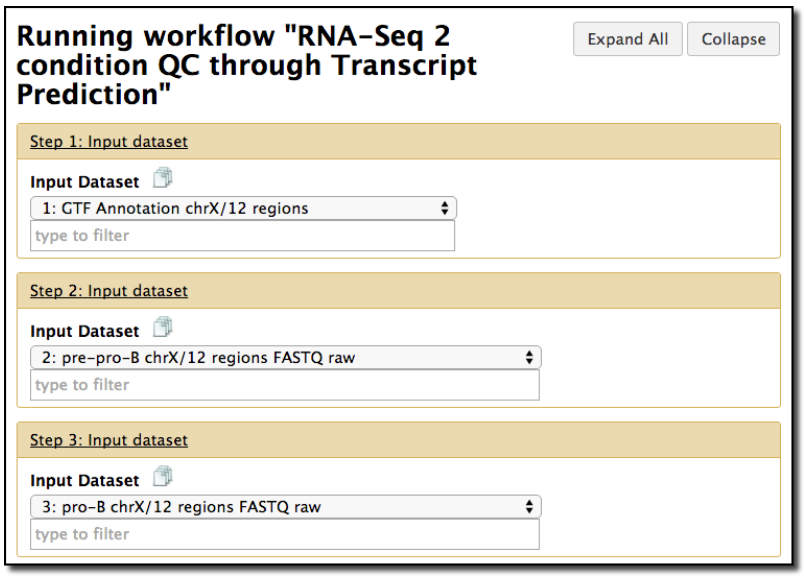


Give the new workflow an informative name and *click* **Create Workflow**. This message appears:



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

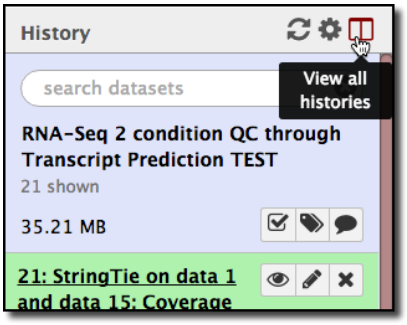
Click the **run** link in the message (or you can also get there by clicking Workflow in the top bar). This brings up a form asking you to define inputs to the workflow. *Set* the first three input datasets for this run to the first three datasets from your current history:
Scroll down to the bottom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

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

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 proteomic**s). The proteins/peptides are separated further through a liquid chromatography (LC) system as they are introduced into the mass spectrometer.

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

1. The creation of fragment ions is carried out in the gas phase through volatilization and ionization in a vacuum. The most common techniques to achieve this are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI).
2. Ions are then separated based on their **mass-to-charge ratio (m/z)**. Different mass spectrometers achieve this separation using different physical mechanisms, such as quadrupole mass filters or time-of-flight.

3. Separated ions are detected when they strike a detector which converts their kinetic energy to a current, which is detected as a signal. The recorded information is converted to a mass spectrum consisting of m/z values and corresponding intensities.

Protein Identification

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

Evaluating the Results

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

Reference materials

Galaxy 101: Building Up and Using a Proteomics Workflow
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 Pepide 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.

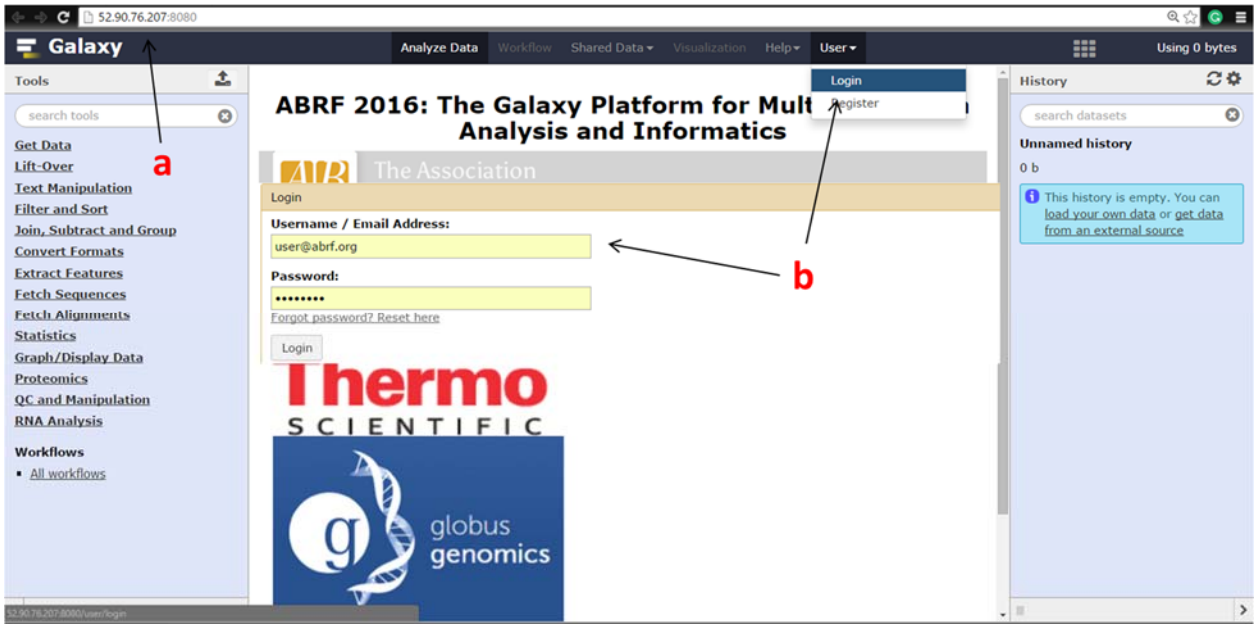
★ **Accessing Galaxy for Proteomics**
For the purposes of this tutorial at ABRF 2016, a special cloud instance of Galaxy-P has been generated. To obtain a local instance of Galaxy download the latest Galaxy source code and install the Proteomics tools from the Galaxy Tool Shed. Users can also take advantage of the public Galaxy-P server to associate themselves with the Galaxy framework and current open-source proteomic tools.

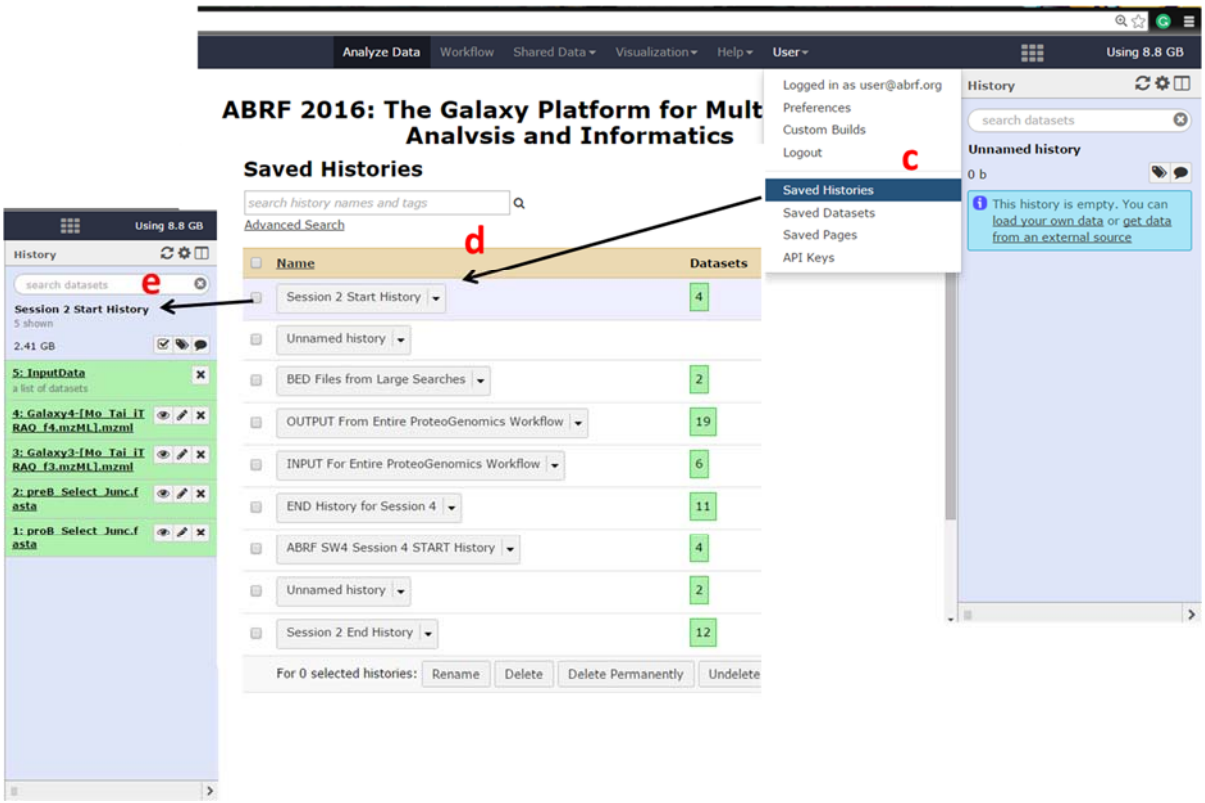
NOTE: This dataset was chosen to allow fast processing and response times in a workshop setting where dozens of people will be submitting jobs at once to the server.

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




Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 8.8 GB
ABRF 2016: The Galaxy Platform for Mult
Analysis and Informatics
Logged in as user@abrf.org
Preferences
Custom Builds
Logout
c
Saved Histories
Saved Datasets
Saved Pages
API Keys
History
search datasets
Unnamed history
0 b
This history is empty. You can load your own data or get data from an external source
Saved Histories
search history names and tags
Advanced Search
d
Name
Datasets
Session 2 Start History
4
Unnamed history
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
For 0 selected histories:
Rename
Delete
Delete Permanently
Undelete
Using 8.8 GB
History
search datasets
e
Session 2 Start History
5 shown
2.41 GB
5: InputData
a list of datasets
4: Galaxy4-[Mo_Tai_iTRAQ_f4.mzML].mzml
3: Galaxy3-[Mo_Tai_iTRAQ_f3.mzML].mzml
2: preB_Select_Junc.fasta
1: proB_Select_Junc.fasta

3 Inputs: Peaklists and Database Collection

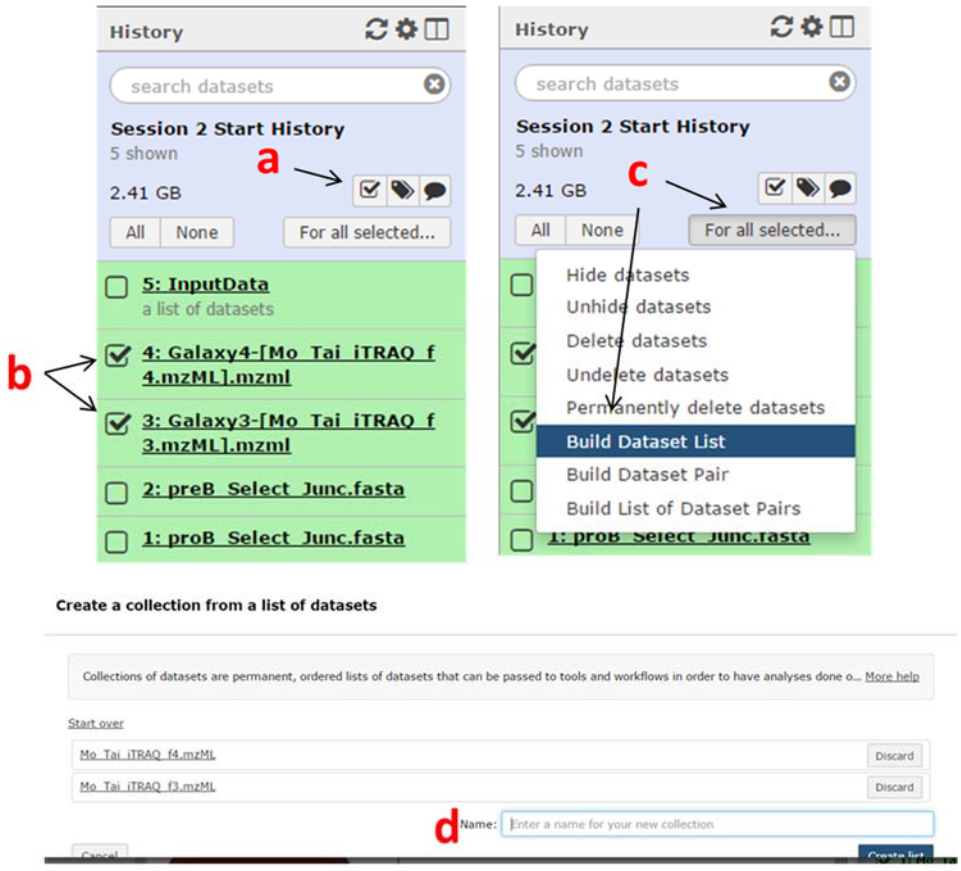
★ **Raw files and Peaklists** (Sect 3.2 page 10)
RAW files are datasets generated experimentally using Thermofinnigan instruments and contain raw information pertaining to a mass spectrometry run. Peaklists (e.g. mzml and MGF files) are generated by processing multiple RAW data files. Processing includes 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.

3.1 Dataset Collection

★**Database Collection**

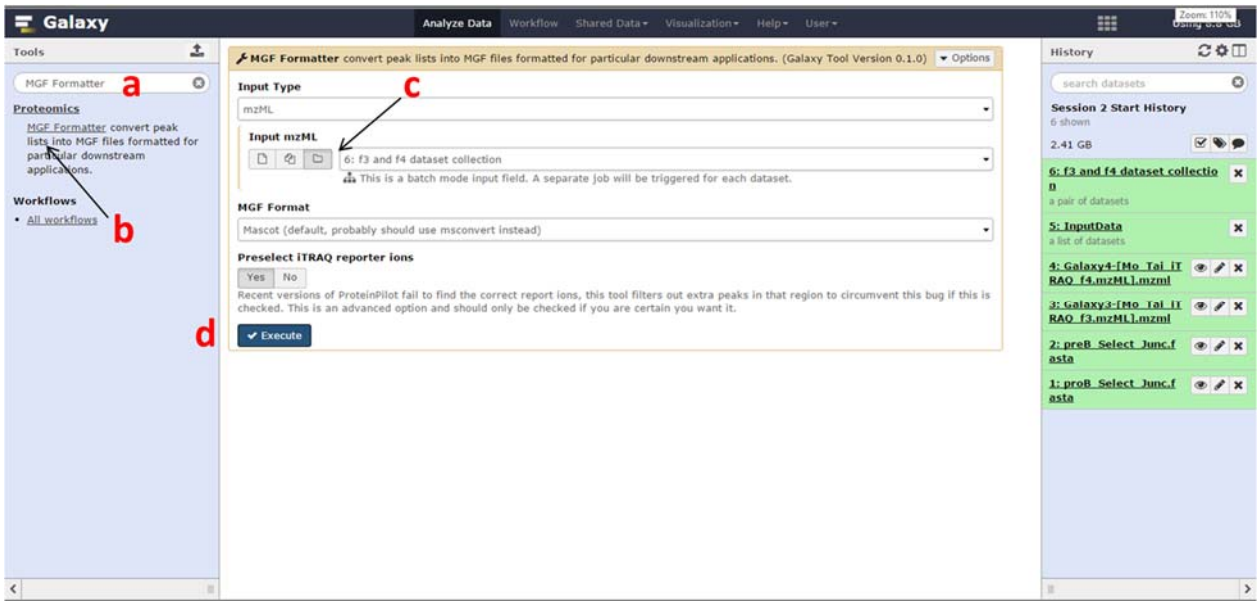
a) In the history panel click the checked box.
b) Once you have clicked on the check box, notice that all inputs have an option to be checked. Please check the mzml files which you would like to combine for dataset collection.
c) Proceed by clicking on "For All Selected…" and from the drop-down menu 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.



3.2 MGF Formatter

★**Raw files and Peaklists**

a) In the search box under "Tools" type MGF Formatter.
b) Click on the MGF Formatter tool and the tool will be prompted in the center panel.
c) 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
d) 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.

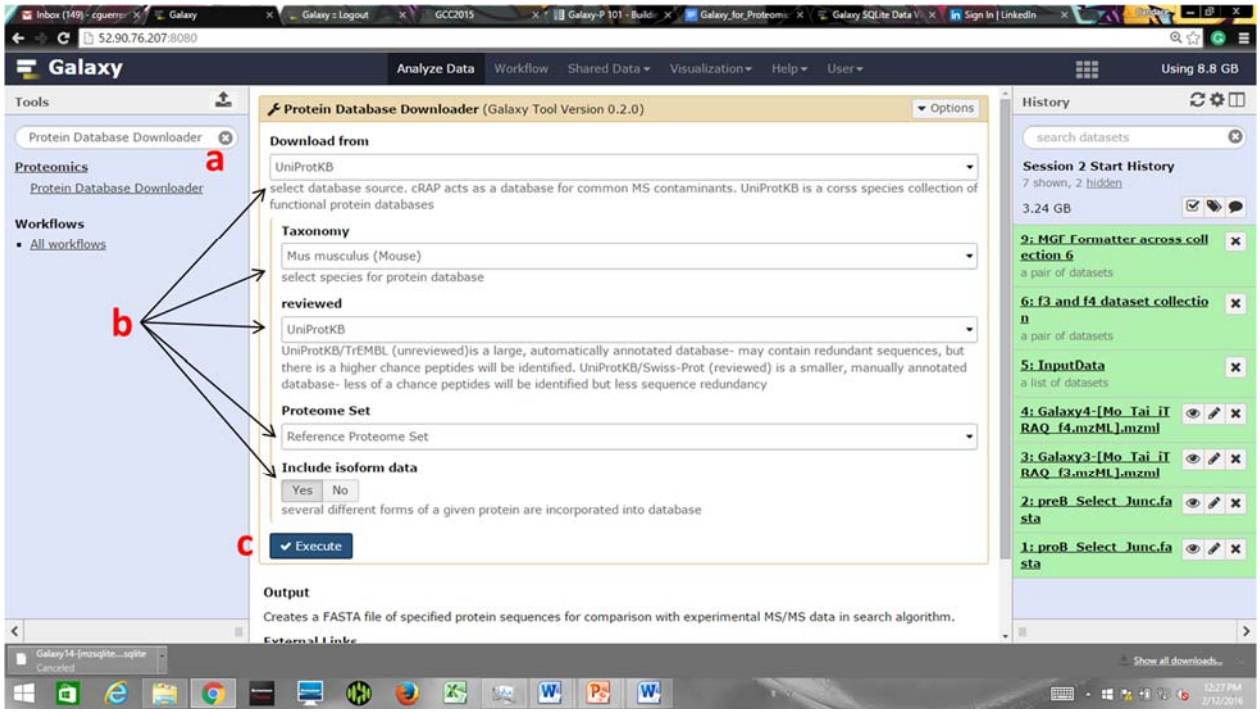
Example: Two databases merged
1. UniProtKB
2. Prob Selected Juncs

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

4.1 Creating a UniProt Reference Proteome

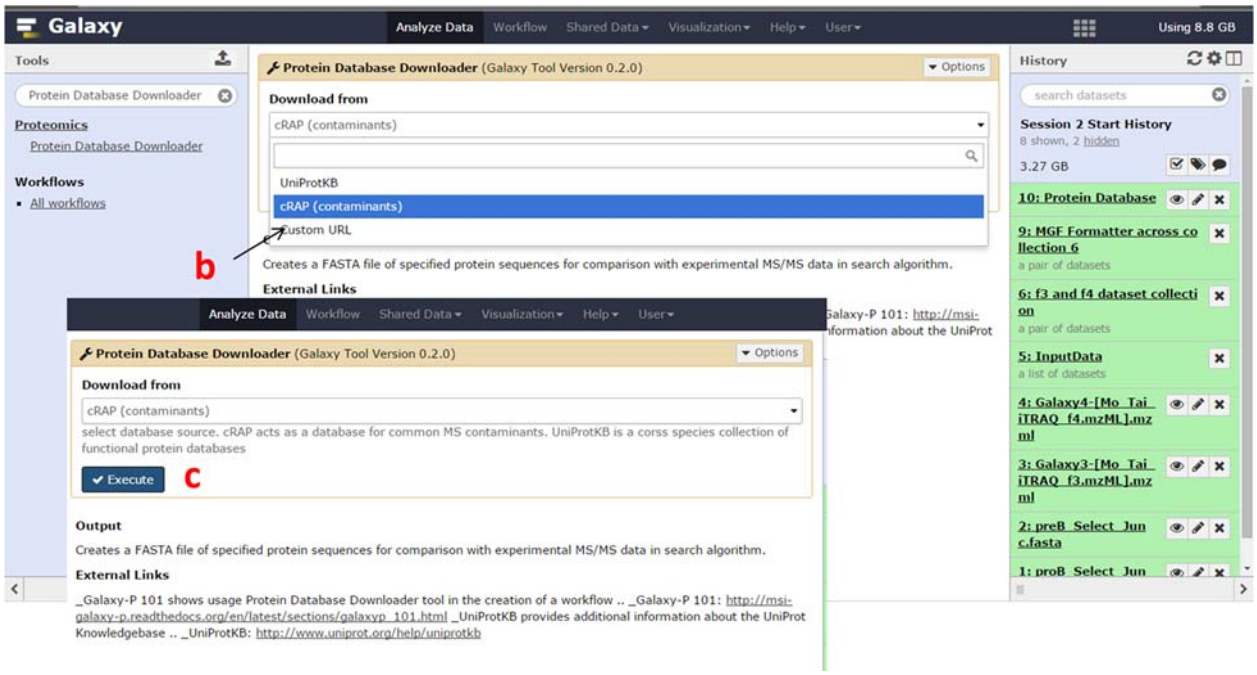
★**Database Generation**

a) In the search box under “Tools” type “Protein Database Downloader” and double-click on the tool
b) 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
c) Click Execute



4.2 Creating a Contaminates Reference Proteome

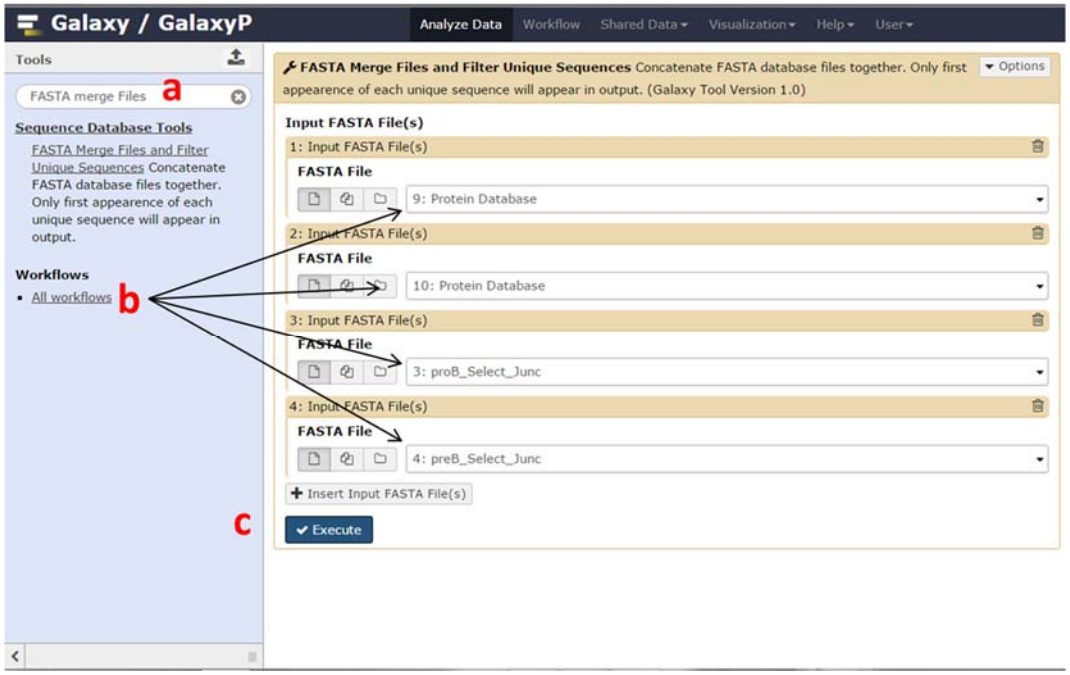
a) Click again on the “Protein Database Downloader” tool
b) From the “Download from” drop-down menu select “cRAP (contaminants)”
c) Click Execute



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.

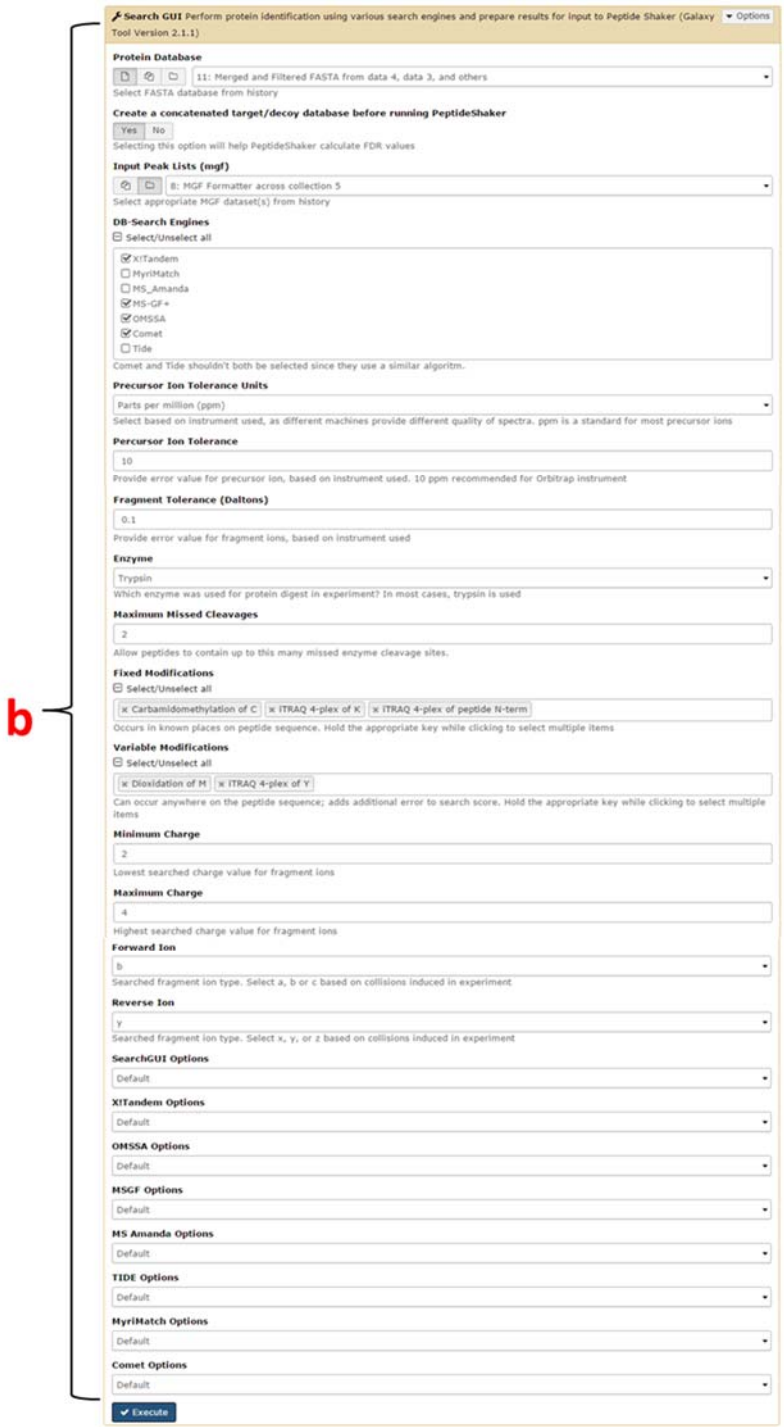
★**Setting Parameters** (Sect 5.1 page 16)
Selecting the proper parameters to match your sample preparation and instruments specificity is crucial for obtaining accurate database matches using SearchGUI. Search parameters may be edited directly from the tool menu of SearchGUI or within the workflow interface. 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).

5.1 SearchGUI Parameters for Protein Algorithm Searches

★**Setting Parameters**

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



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

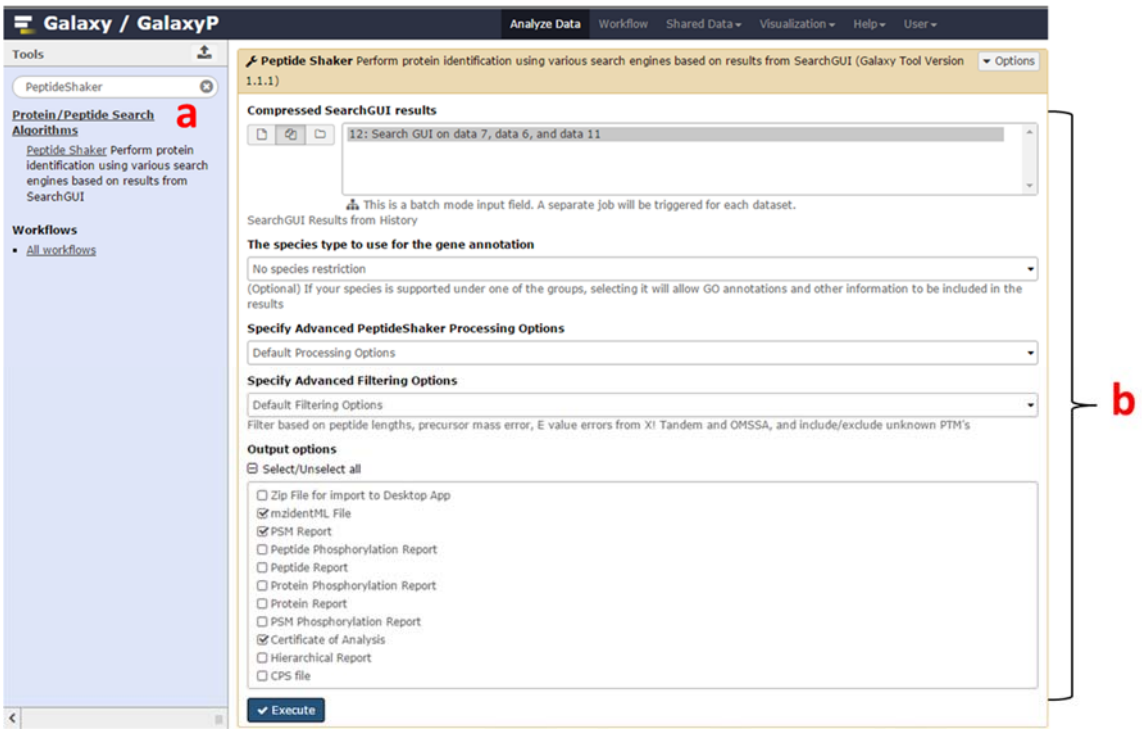
★**False Discovery Rate based on Target-Decoy Search**
In order to maintain accuracy and effectiveness in spectral / peptide / protein identification, a target-decoy search strategy can be used to discern how correct and incorrect a spectral or peptide or protein match is. The most popular approach for 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).

6.1 PeptideShaker

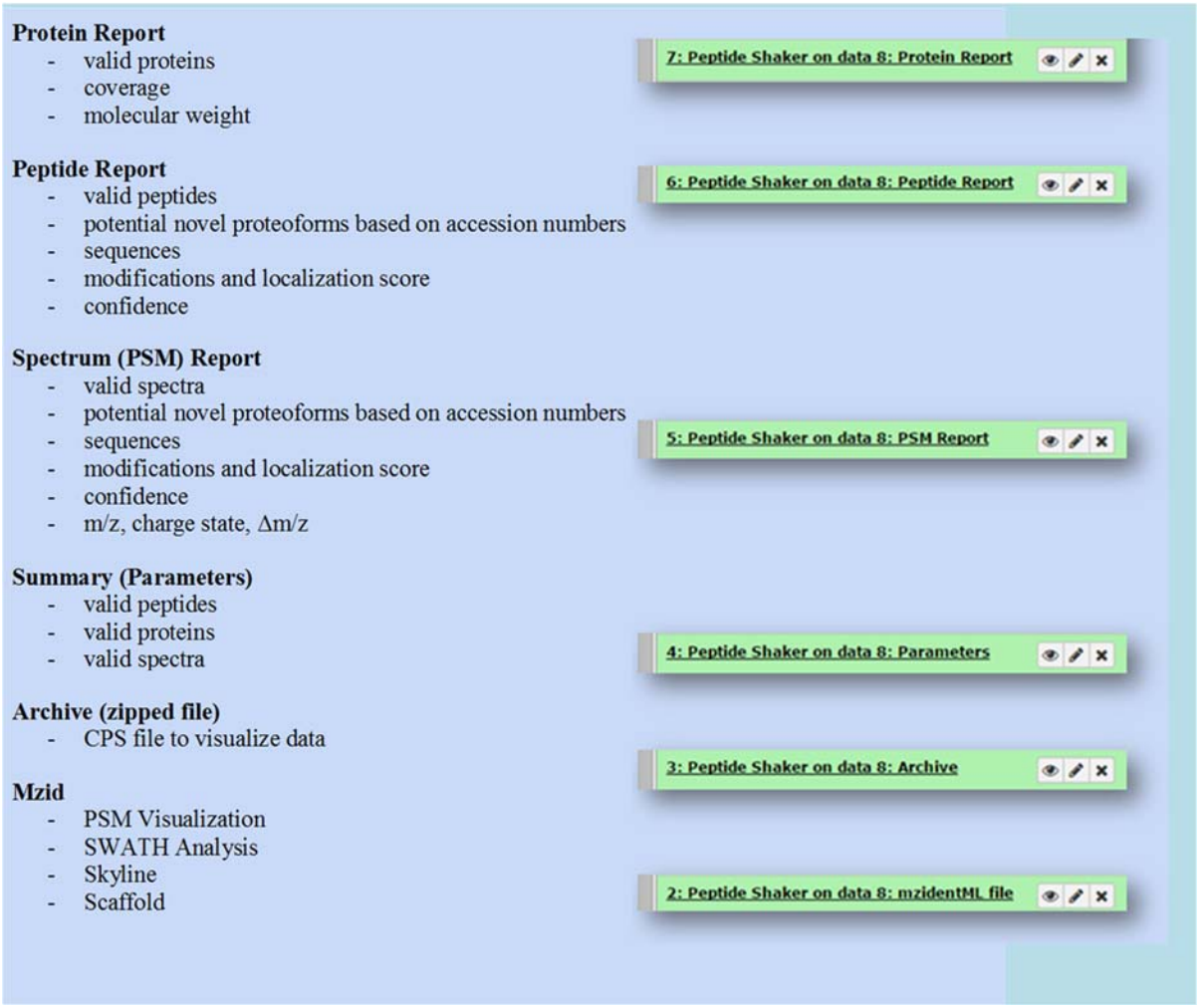
★PeptideShaker in Galaxy-P

a) In the search box under "Tools" type "PeptideShaker" and double-click on the tool
b) Following the parameter settings
c) Click Execute



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.



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.

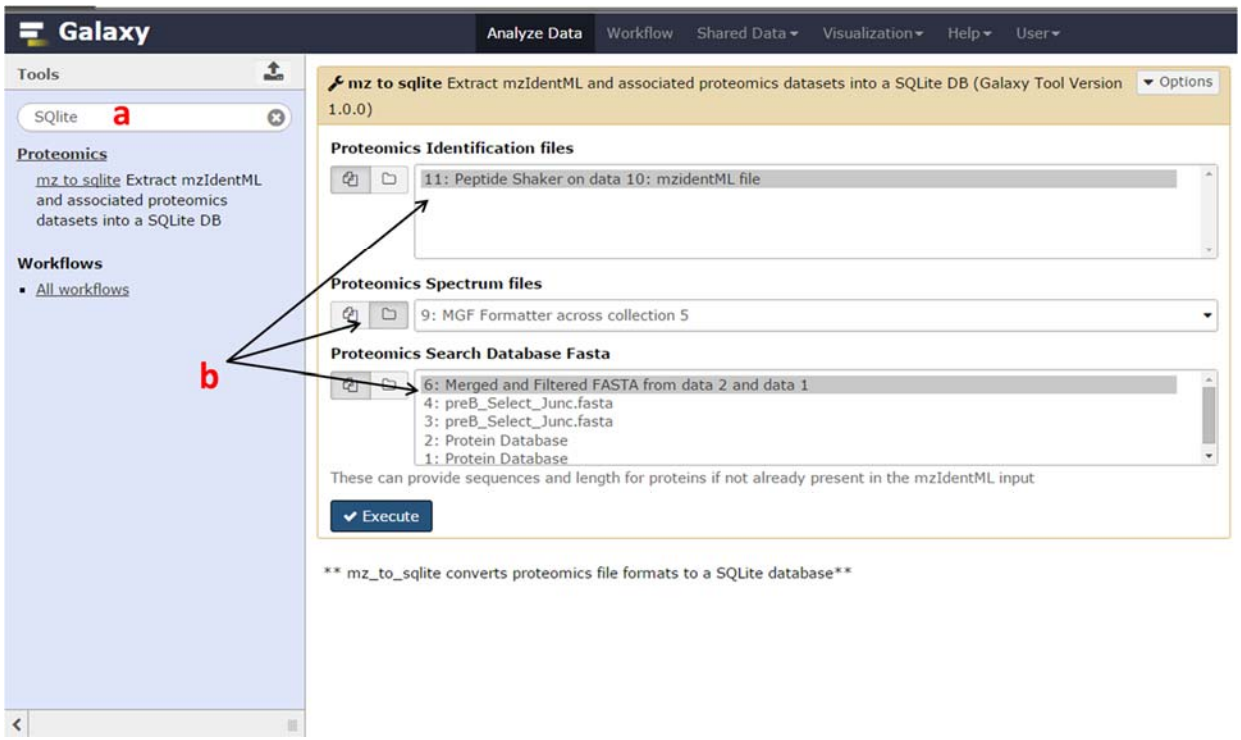
★**Visualizing PeptideShaker Results: sqlite Database** (Sect 7.1 page 21)
The **mz to sqlite** Galaxy-P tool consolidates the information in the mzIdentML output dataset from a search algorithm like PeptideShaker along with the peaklist input datasets (e.g. mzml and MGF files) and the fasta SearchDB 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.

7.1 Generating the sqlite Database

★**Visualizing PeptideShaker Results: sqlite Database**

a) In the search box under "Tools" type "mz to sqlite" and double-click on the tool
b) Follow the display settings below
c) Click Execute

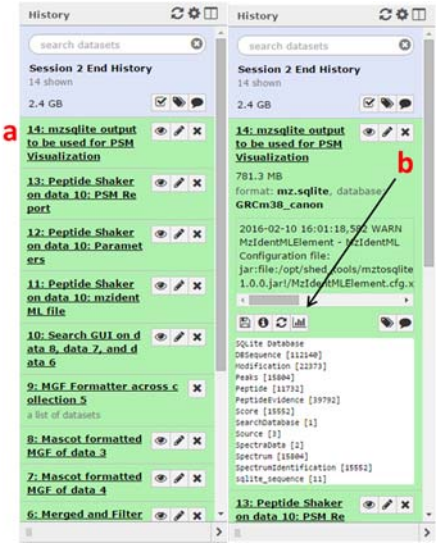


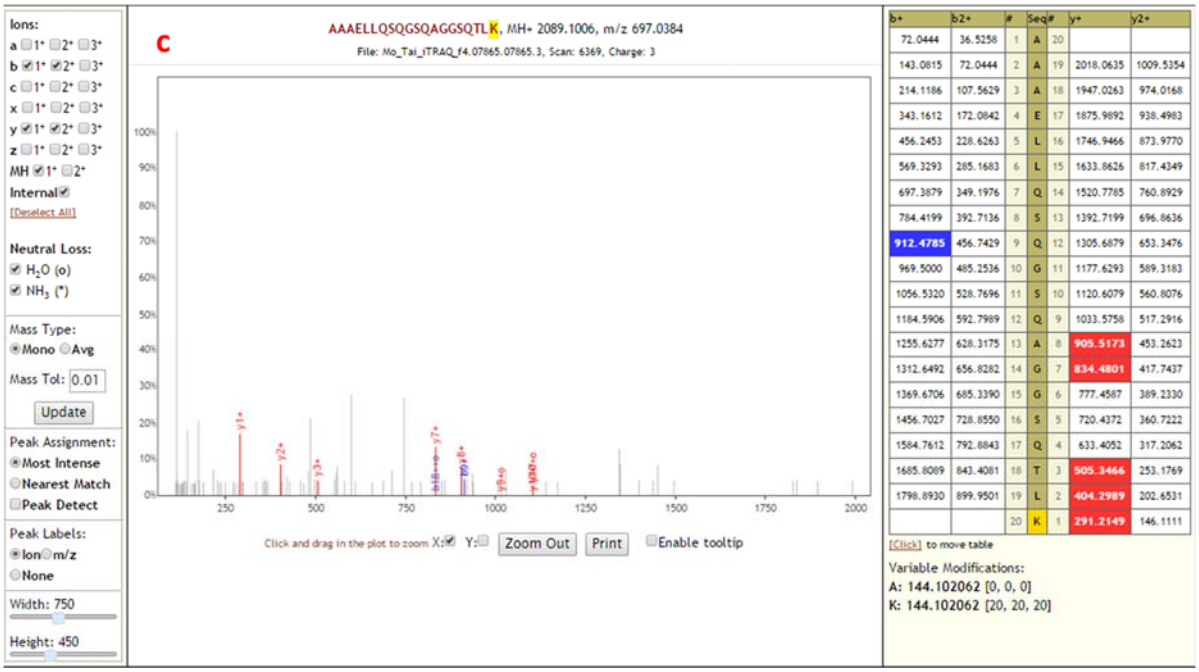
7.2 PSM Evaluator in Galaxy-P

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

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

a) 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.
b) Select the graphical bar figure icon. This will bring you to an outside browser for PSM visualization
c) Peptides of interest can be selected for visualization.





8 Generating and Running Workflows within Galaxy

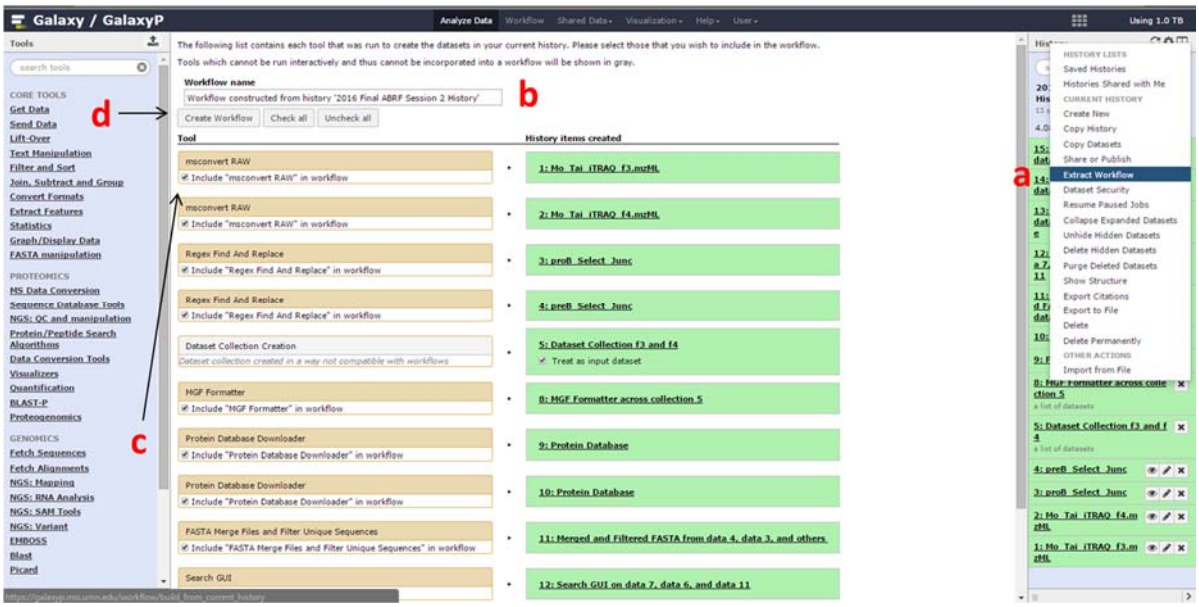
★ **Galaxy Proteomic Workflows** (Sect 8.1 page 24)
For the sake of time, this workflow was performed on two reduced mzML. files generated from two fractions from a biological sample from mouse. In total there are 9 fractions from this sample that could be processed collectively. The same workflow can 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.

8.1 Extract Workflow from Current History

★**Galaxy Proteomic Workflows**

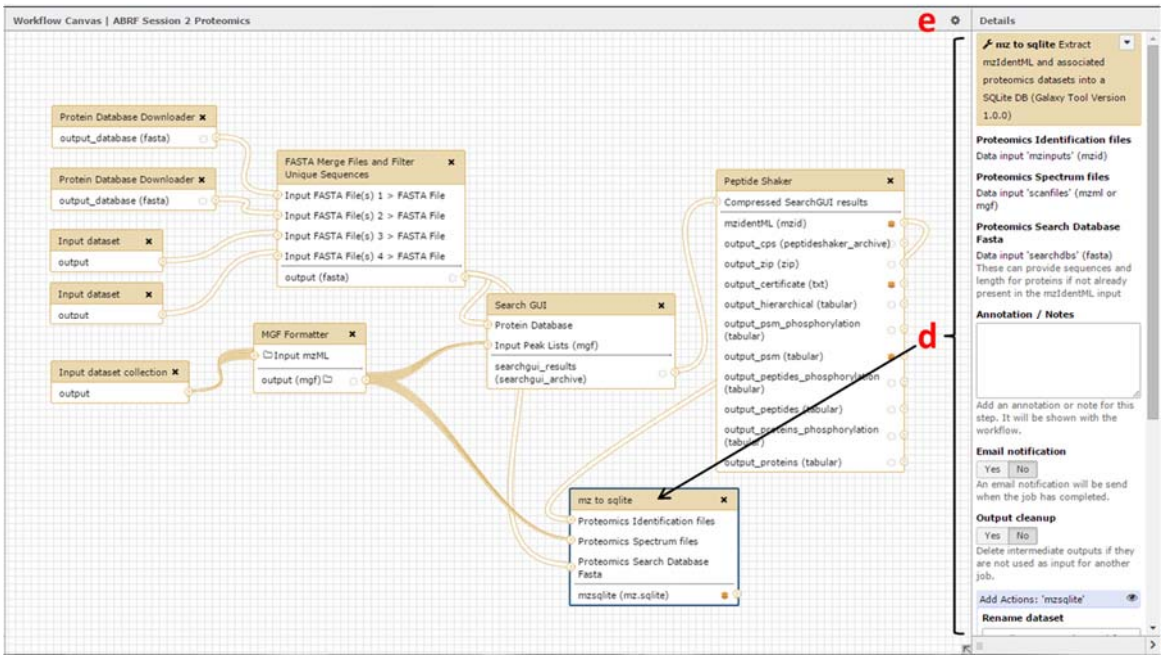
a) At the top of the history pan click on the small gear icon and select “Extract Workflow” from the pop-up menu
b) In the “Workflow” name box enter “Session 2 Proteomics”
c) Confirmed that all the tools are checked
d) Click “Create Workflow” under the workflow name



8.2 Edit the Workflow

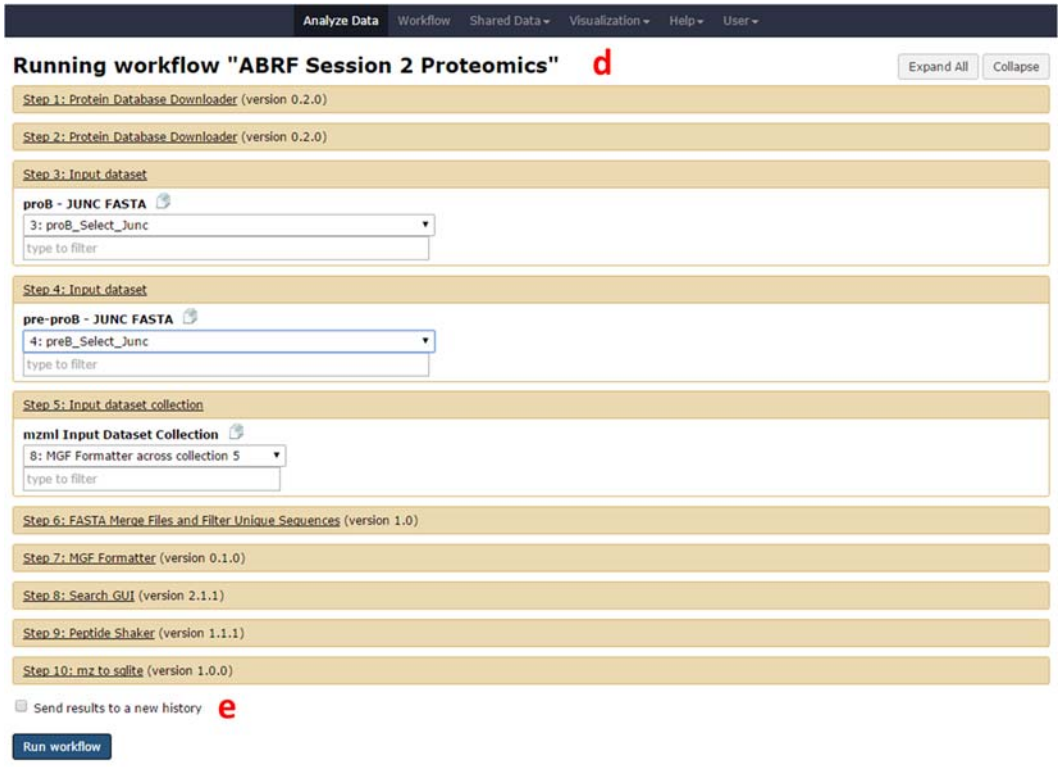
★**Workflow Parameters**

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.



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."



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

ABRF 2016 ANNUAL MEETING
February 20-23

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
Galaxy 101: NGS data analysis hands-on tutorial:
www.msi.umn.edu/content/bioinformatics-analysis
Tophat manual: 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)

2.1 Accessing Galaxy

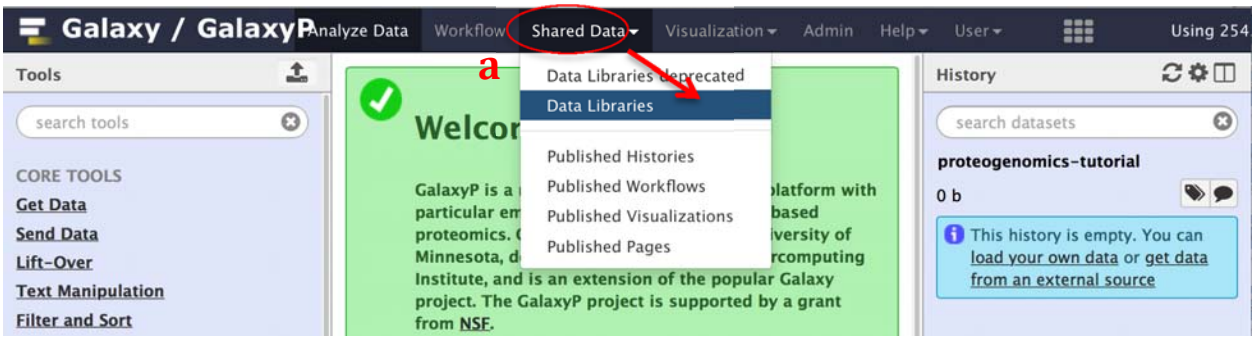
a) Open a web browser and navigate to GalaxyP website galaxy.msi.umn.edu
b) Log in with your username and password
c) On the right pane, Click on “Unnamed history” and change the history name to something more descriptive e.g., “proteogenomics-tutorial” and press enter

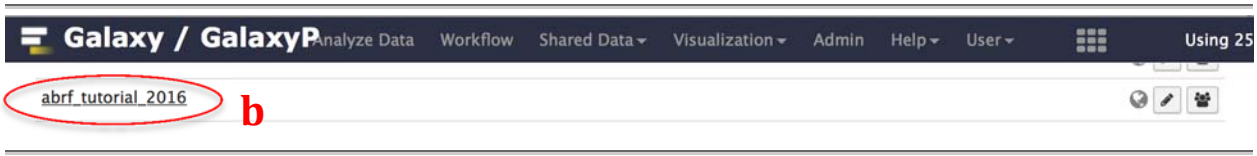
Tools pane Center pane History pane

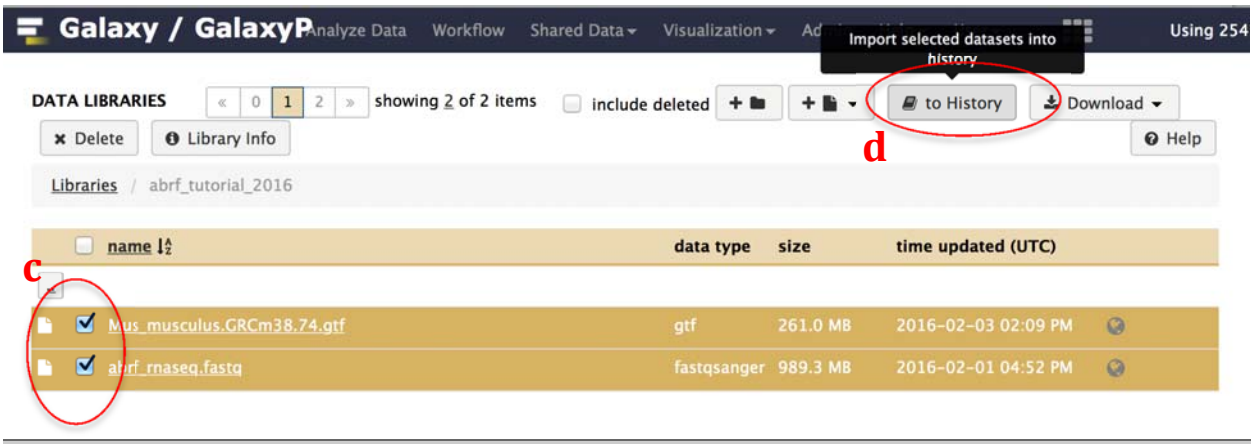
2.2 Import tutorial datasets into current history

★Tutorial Dataset

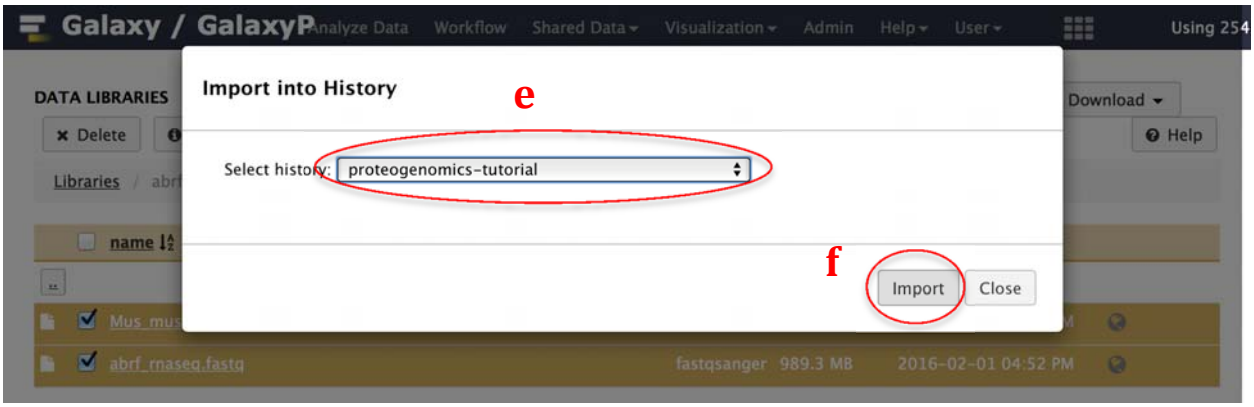
a) At the top of the screen select “Shared Data -> Data Libraries”
b) Select “abrf_tutorial_2016” from the list of data libraries
c) Select the files “Mus_musculus.GRCm38.74.gtf“ and “abrf_rnaseq.fastq”
d) Near the top of the page click “to History”

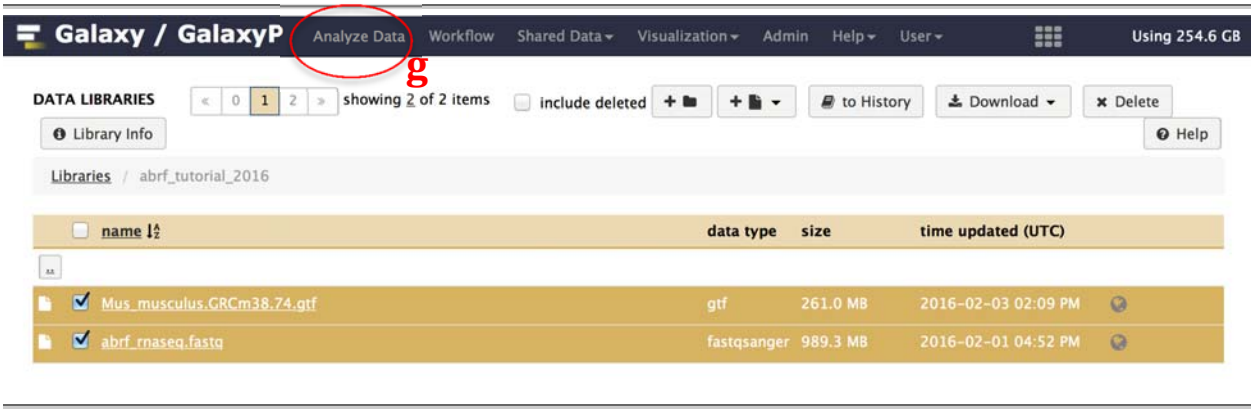






e) Select your working history (name you changed your history to) from the drop down menu
f) Click Import
g) Click “Analyze Data”





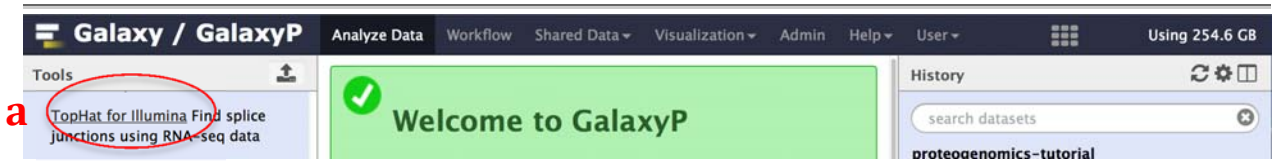
3 Mapping with Tophat

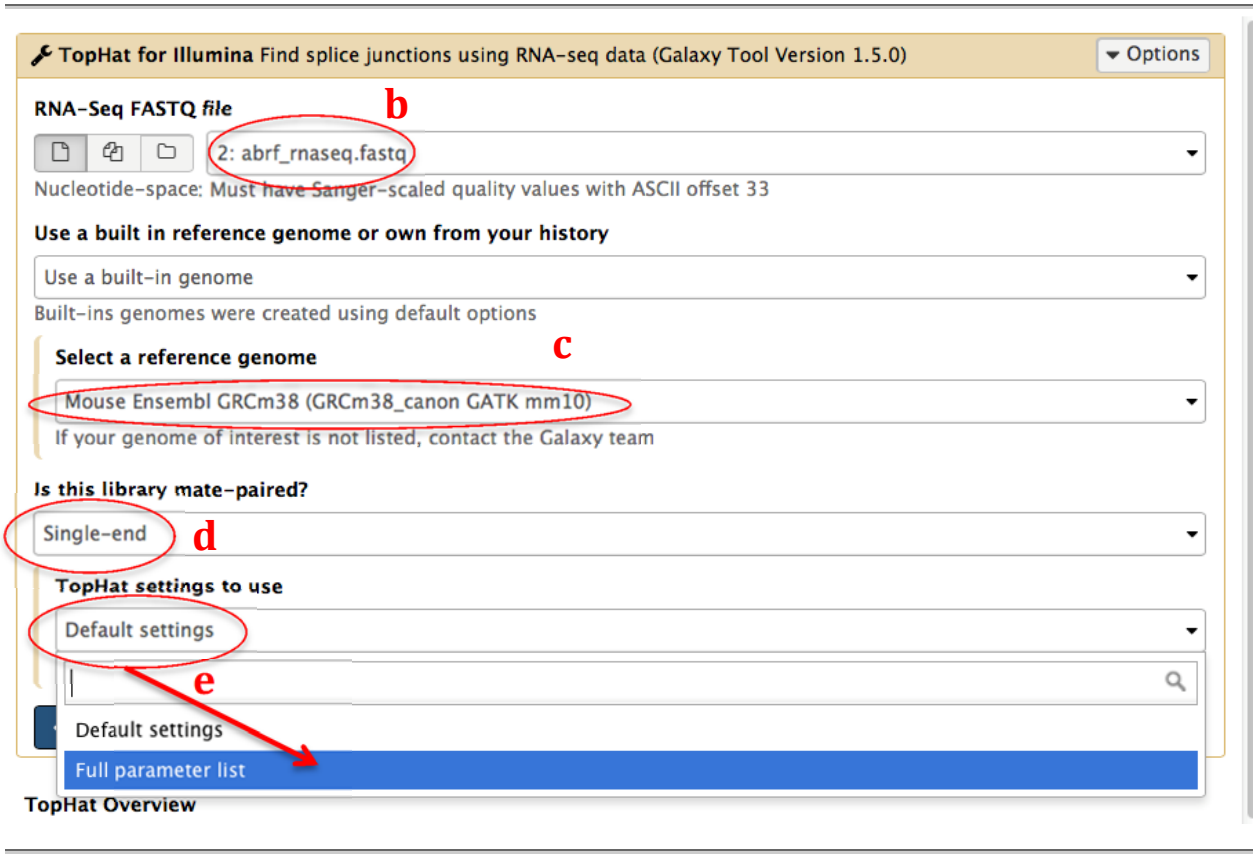
★ **Reference Genomes**
It is important the reference genome you align against is generated from the same reference genome as the GTF you are using. Chromosome names and coordinates used in the GTF file must be the same as those used in the database. If your genome build is mm10, make sure your GTF file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

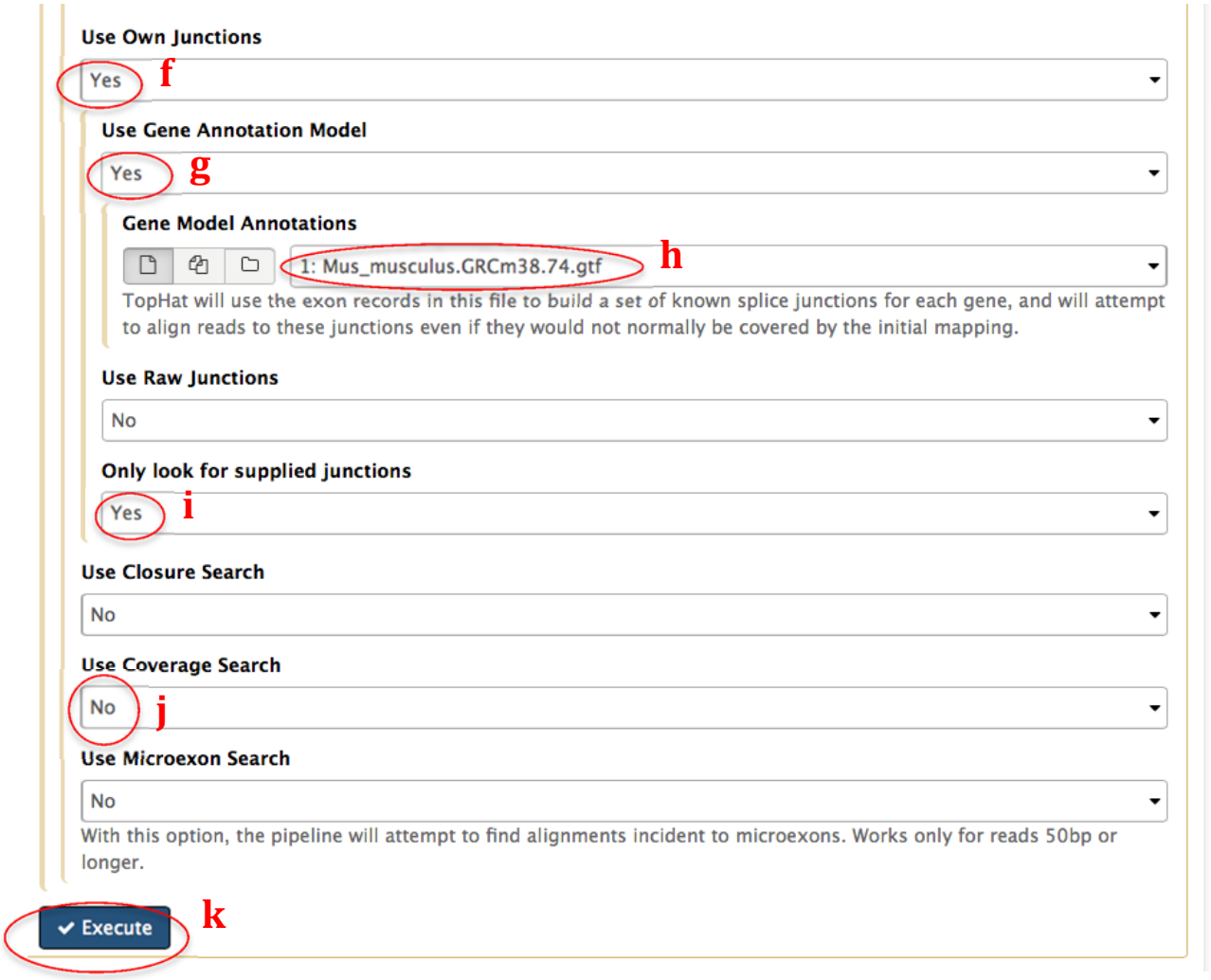
3.1 First Tophat run

a) Load the Tophat tool from the tool pane: “Tophat for Illumina”
b) RNA-Seq FASTQ file -> abrf_rnaseq.fastq
c) Select a reference genome -> Mouse Ensembl GRCm38 (GRCm38_canon GATK mm10)
d) Is this library mate-paired -> Single-end
e) Default settings -> Full parameter list



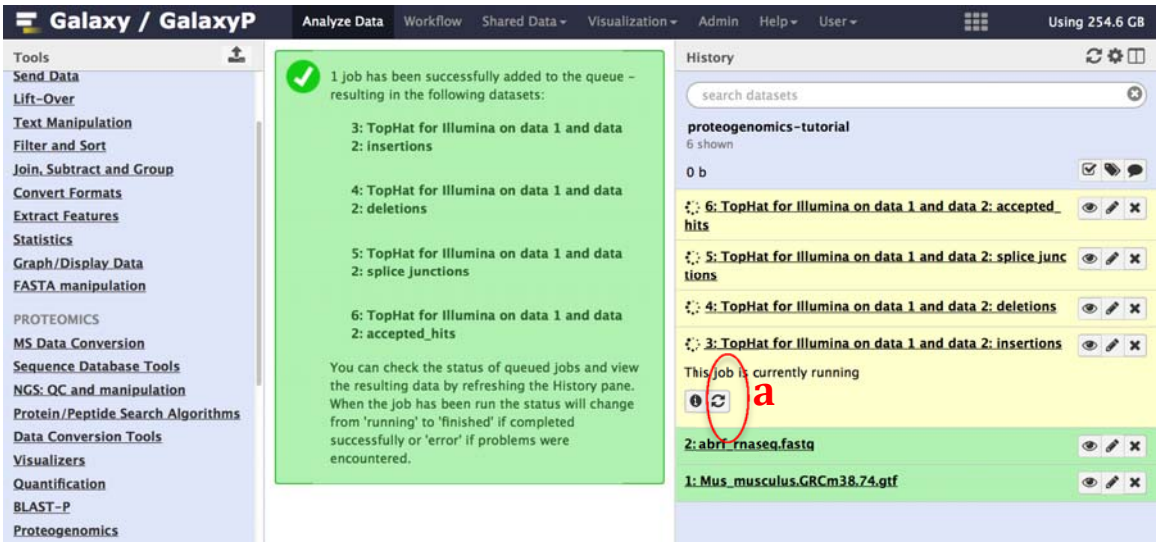


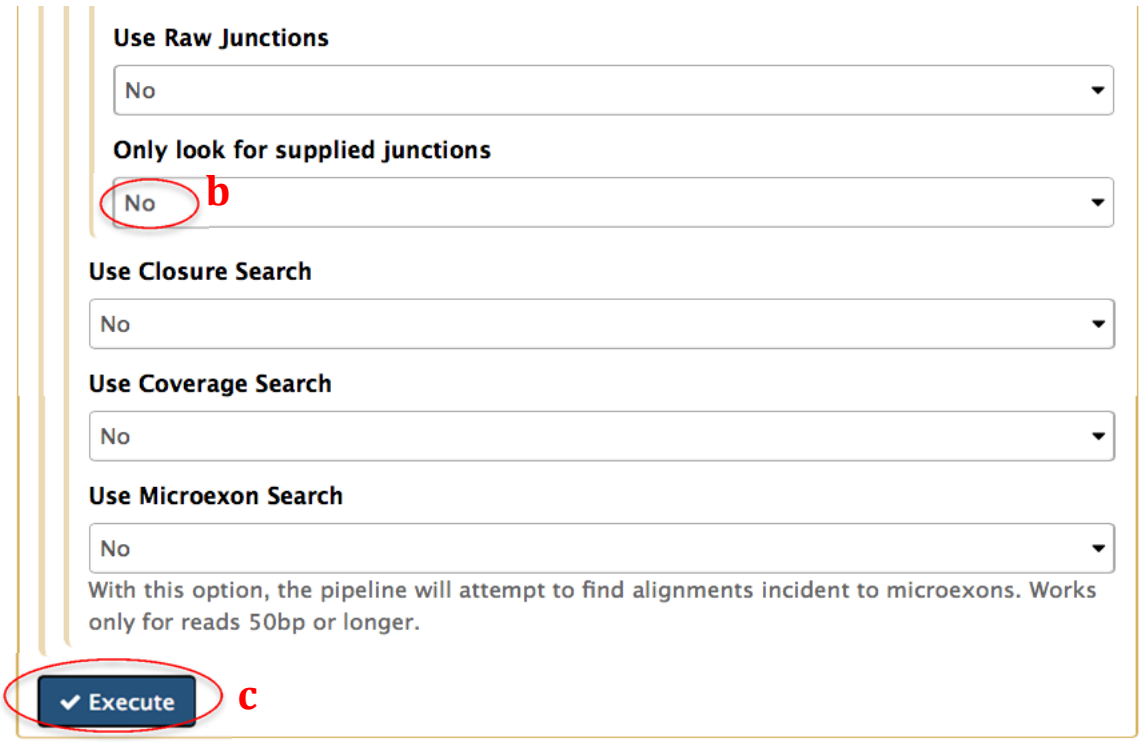
f) Scroll down, “Use Own Junctions -> Yes”
g) Use Gene Annotation Model -> Yes
h) Gene Model Annotations -> Mus_musculus.GRC.m38.74.gtf
i) Only look for supplied junctions -> Yes
j) Use Coverage Search -> No
k) Click “Execute” to submit the job



3.2 Second Tophat run

a) 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
b) Scroll down until you get to **"Only look for supplied junctions"** and select "**No**"
c) Click "Execute" to submit the second Tophat run



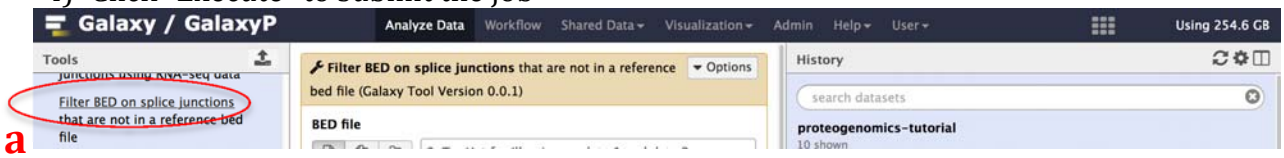


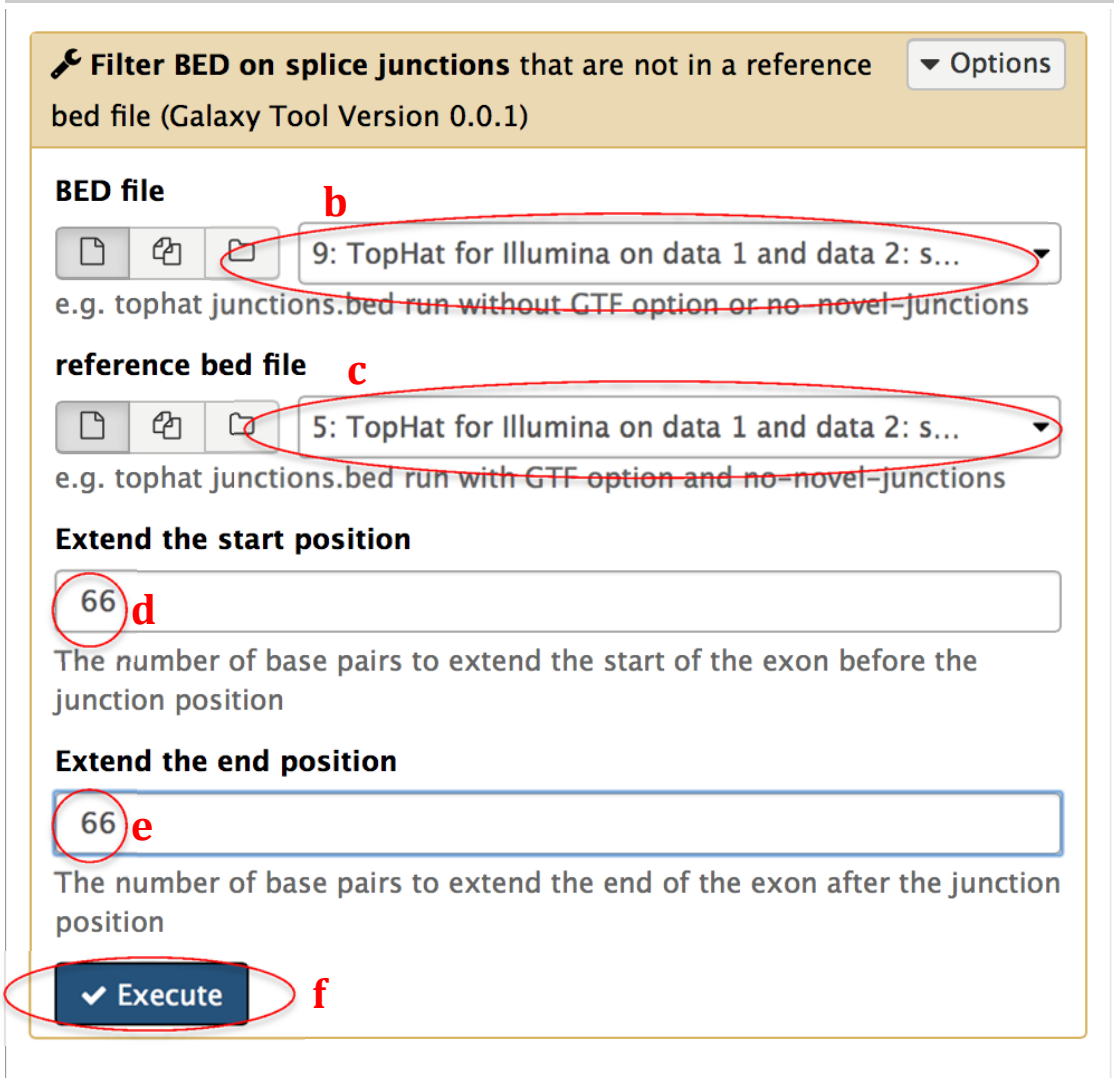
4 Generating protein database of novel proteoforms

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

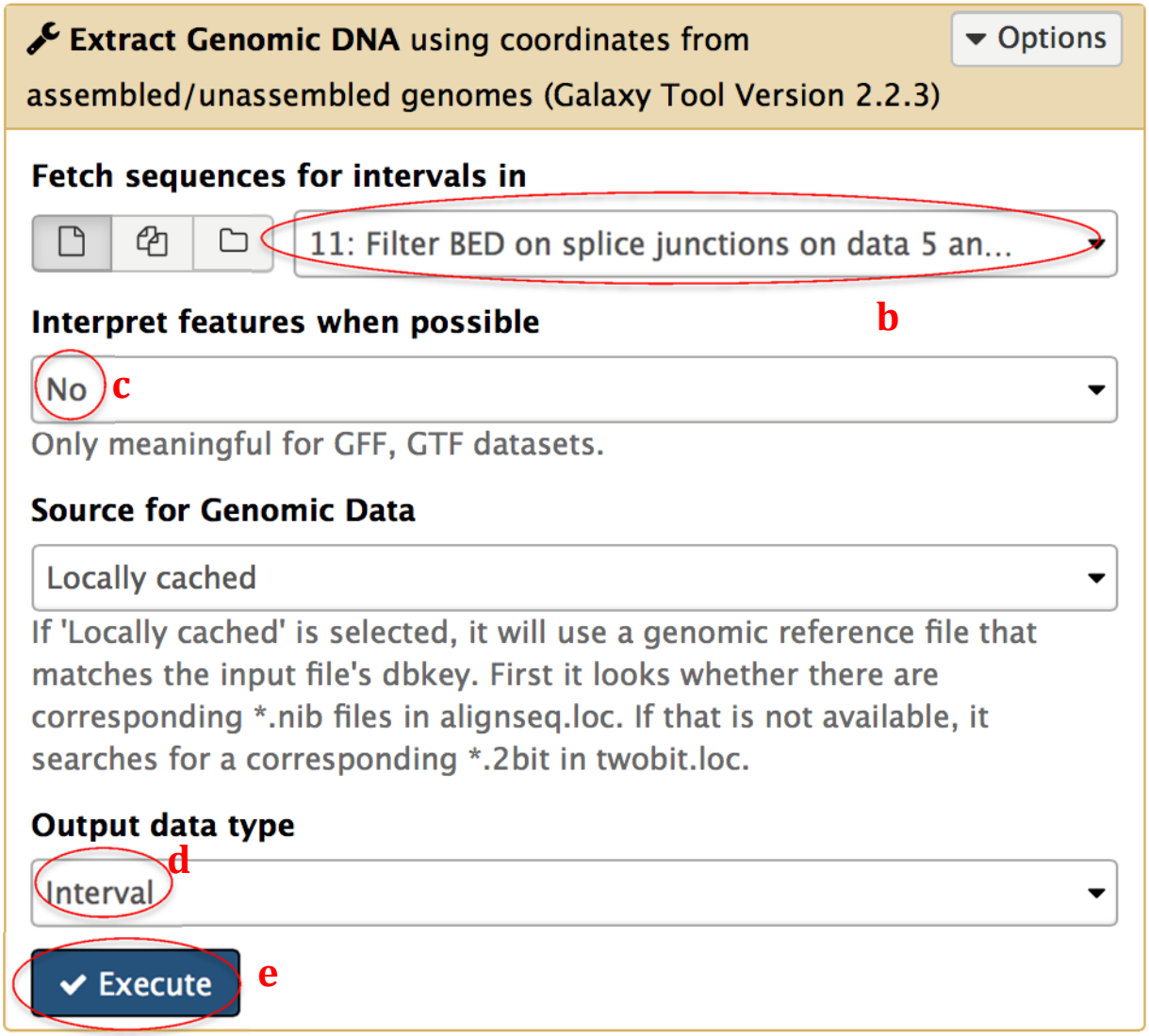
a) Find the tool “Filter BED on splice junctions” and click on the name
b) Bed file -> “9 TopHat for Illumina on data 1 and data 2: splice junctions”
c) reference bed file -> “5 TopHat for Illumina on data 1 and data 2: splice junctions”
d) Extend the start position -> 66
e) Extend the end position ->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





4.2 Fetch genomic sequence for novel splice junction

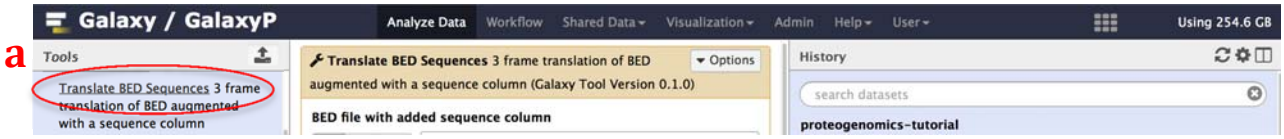
a) Find the tool "Extract Genomic DNA" and click on the name
b) Fetch sequences for interval in -> "11 Filter BED on splice junctions on data 5 ...."
c) Interpret features when possible -> No
d) Output data type -> Interval
e) Click "Execute" to submit the job

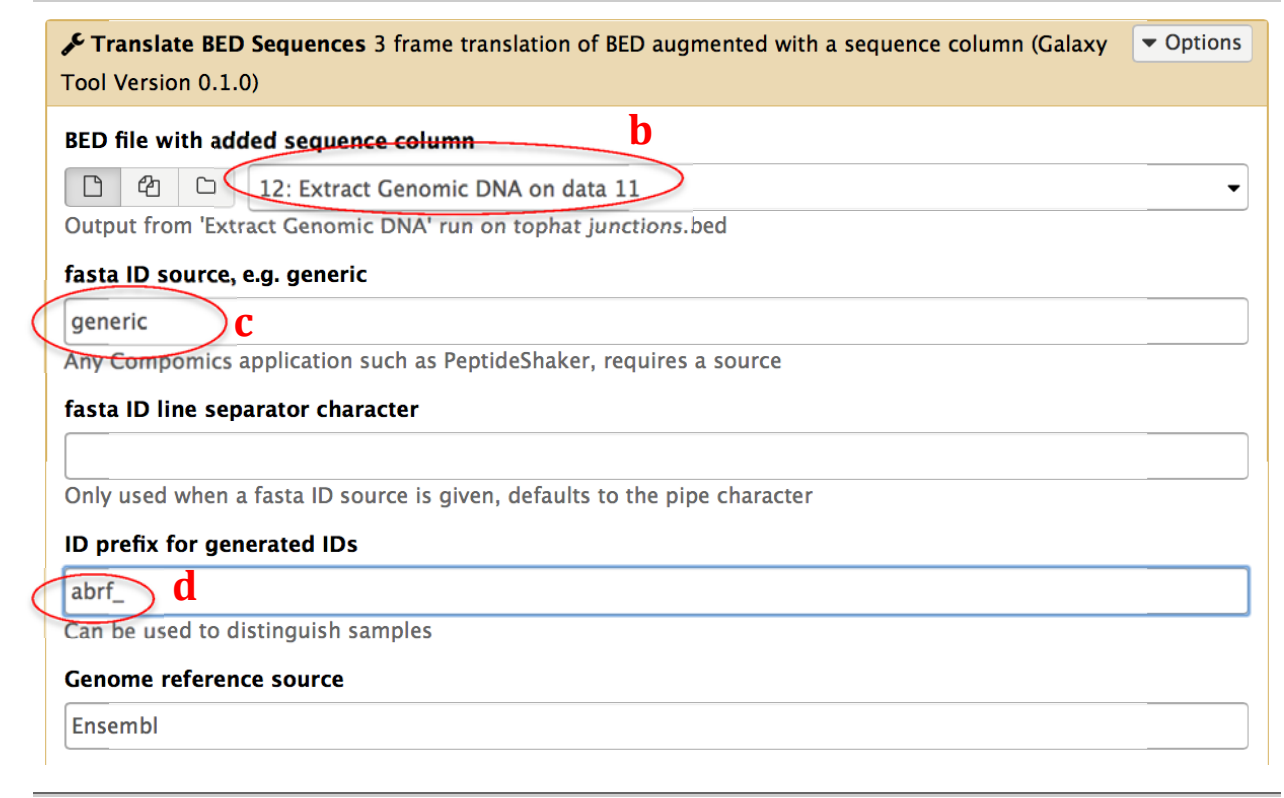


4.3 Generate peptide fasta using 3-frame translation

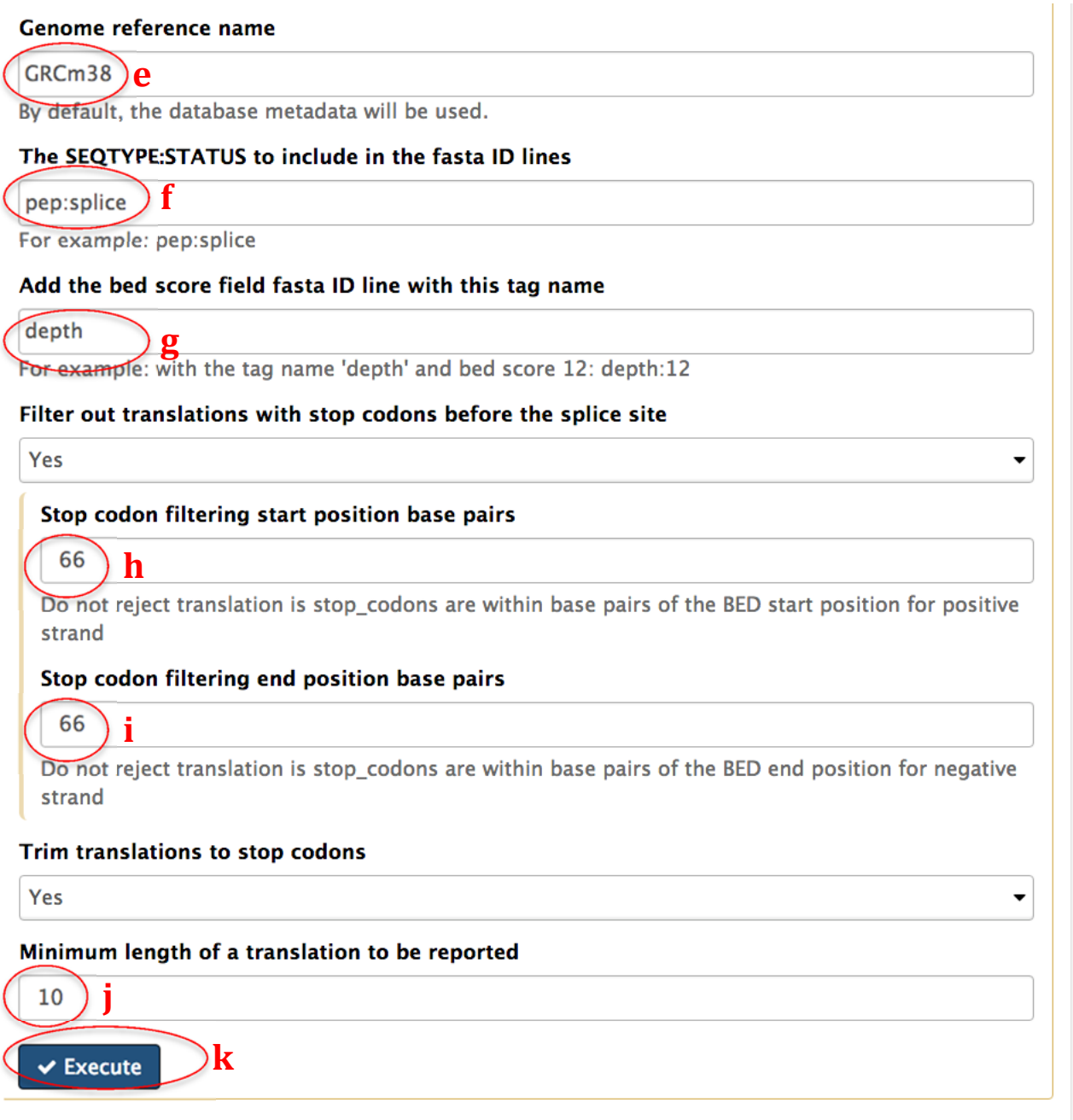
NOTE: When Tophat identifies splice junctions, it indicates frame of translation. A 3-frame translation is thus sufficient to identify all potential novel-proteoforms supported by our RNA-Seq data.

a) Find the tool "Translate BED sequences" and click on the name
b) BED file with added sequence column -> "12: Extract Genomic DNA…."
c) Fasta ID source, e.g. generic -> generic
d) ID prefix for generated source -> abrf_





e) Genome reference name -> GRCm38
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



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

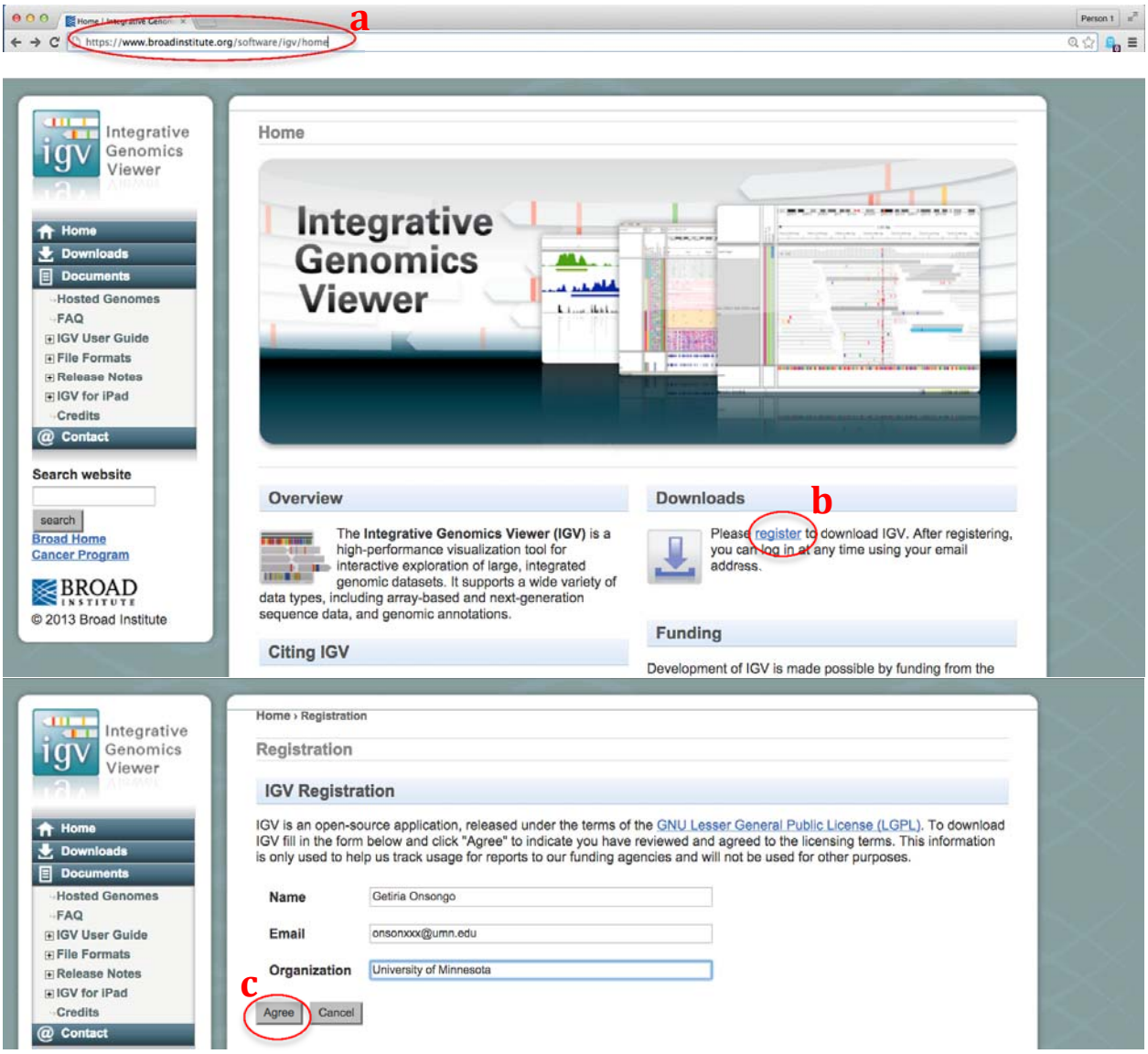
b) Trackster is a genome browser built into Galaxy. Any data file that can be viewed in Trackster will have a Trackster icon  next to it in the history pane.

c) The Integrated Genome Browser (IGB) is similar to IGV, but most users prefer to use IGV.

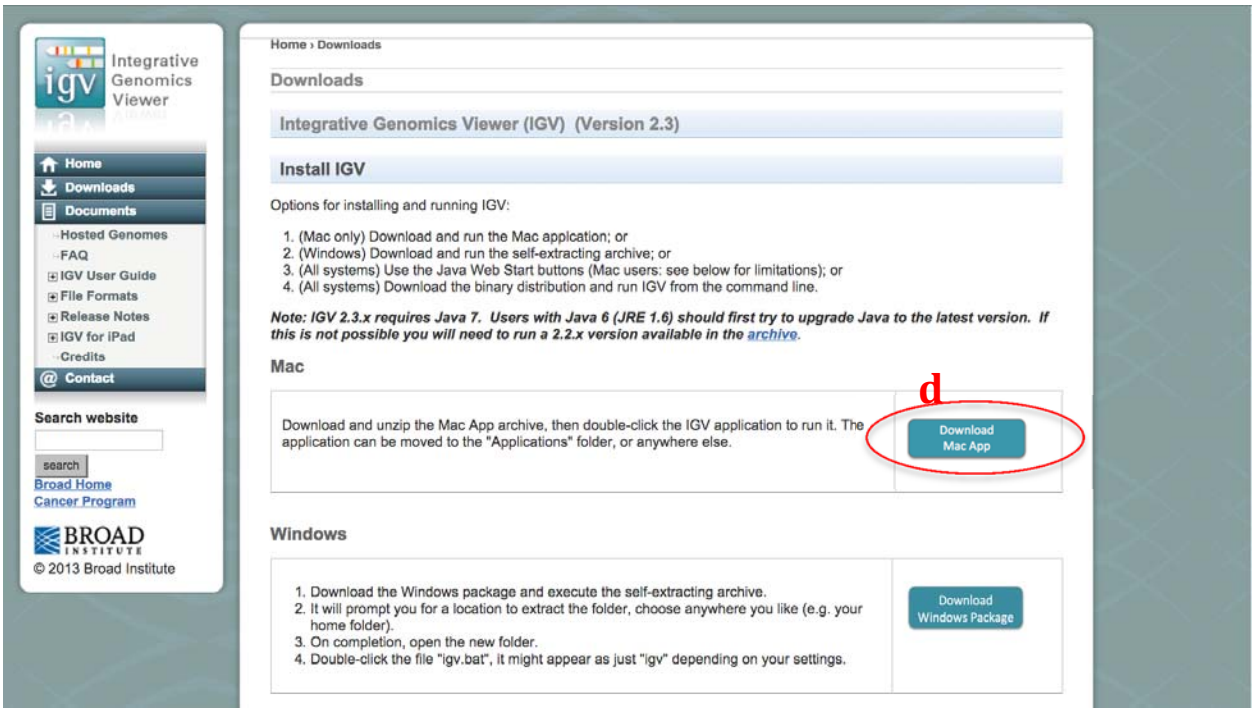
★**IGV**
Developed at the Broad Institute, IGV is a high-performance visualization tool for interactive exploration of large, integrated genomics datasets. We will download IGV and use it to visually inspect quality of reads supporting our novel splice junctions.

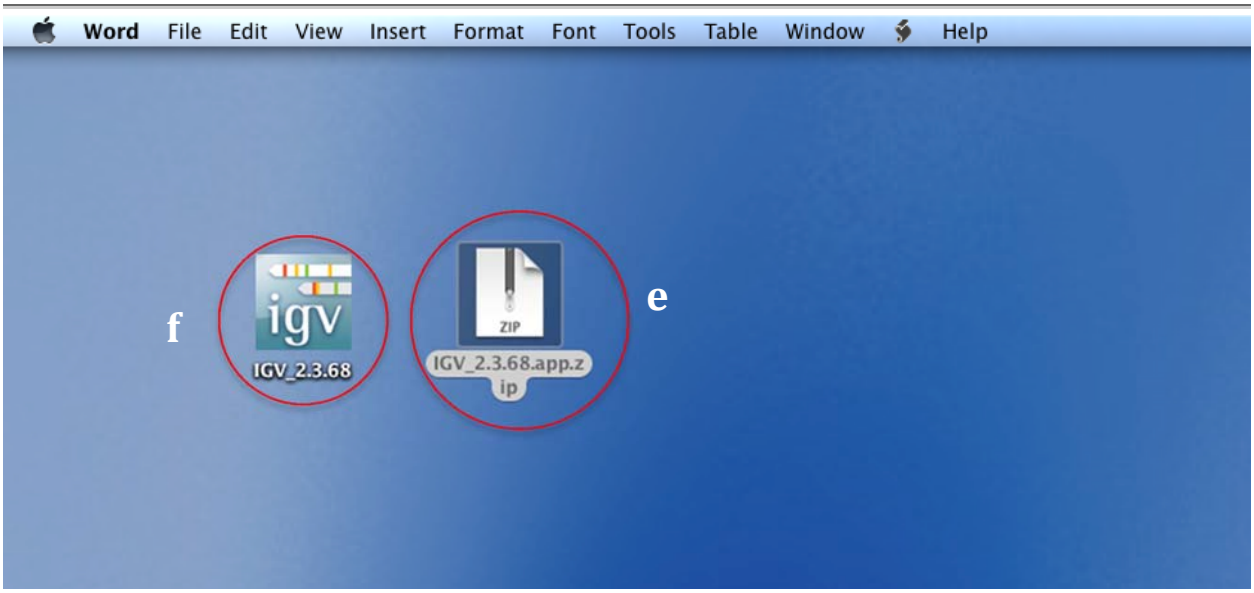
5.1 Downloading IGV from the broad

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



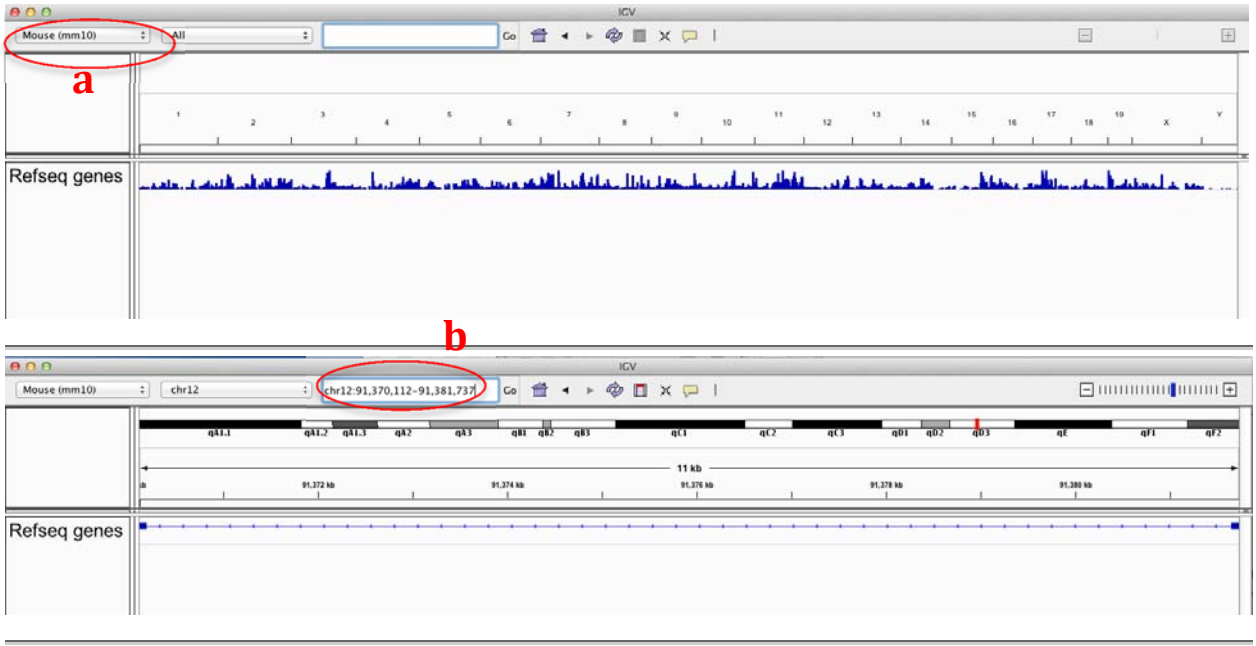
d) Find the appropriate download link for your system and click download. These tutorial covers instructions for a Mac.
e) Navigate to where the application was downloaded and double click the zipped archived
f) Double click the IGV application to run it.





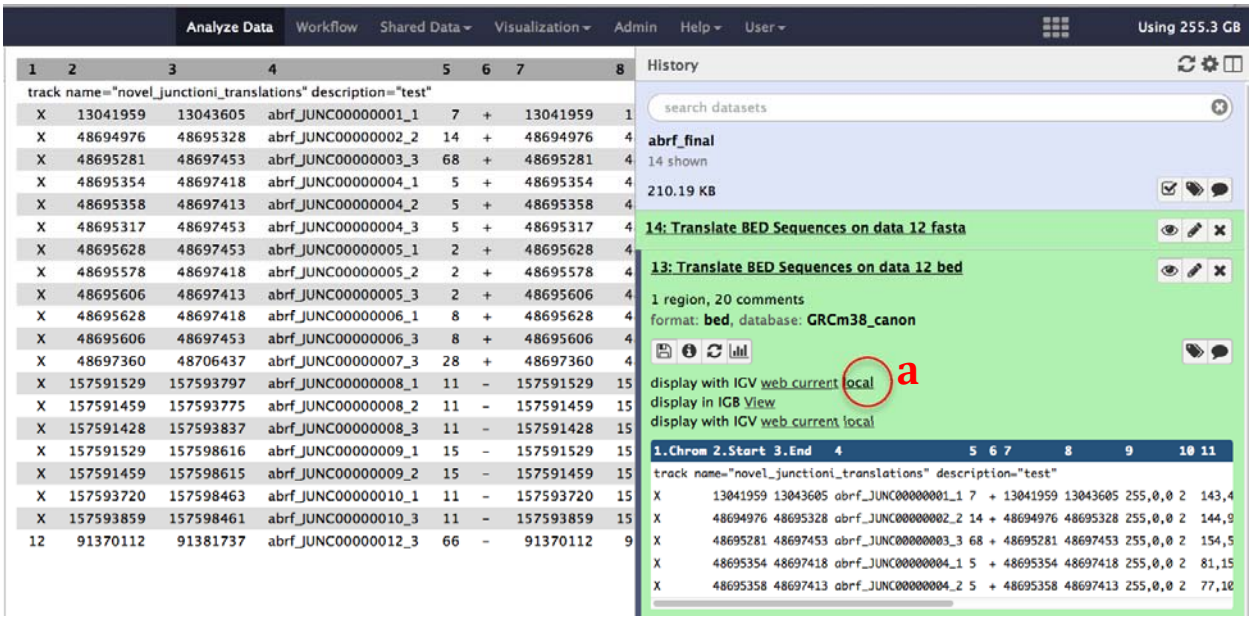
5.2 Ensure correct genome is loaded in IGV

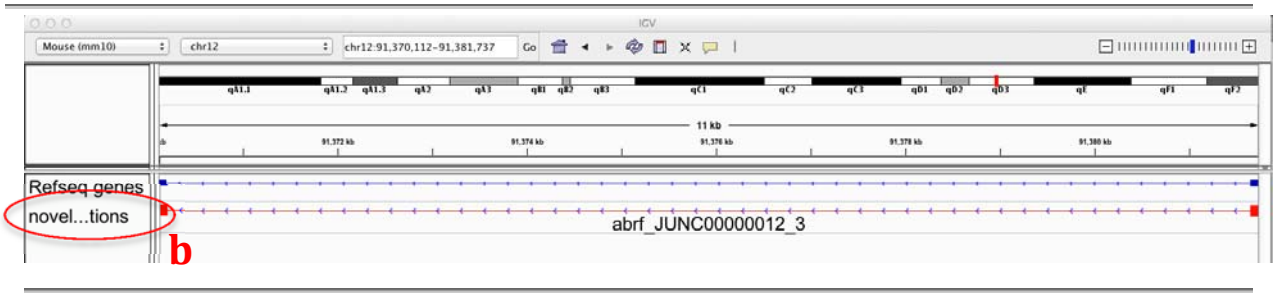
a) Once the application loads, ensure the correct genome build is loaded (mm10)
b) Enter "chr12:91,370,112-91,381,737" into the region field and press "Enter". This will zoom IGV into this region. This is one of the regions we will be examining. It is a good illustration of reads supporting a splice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



5.3 Load files into IGV

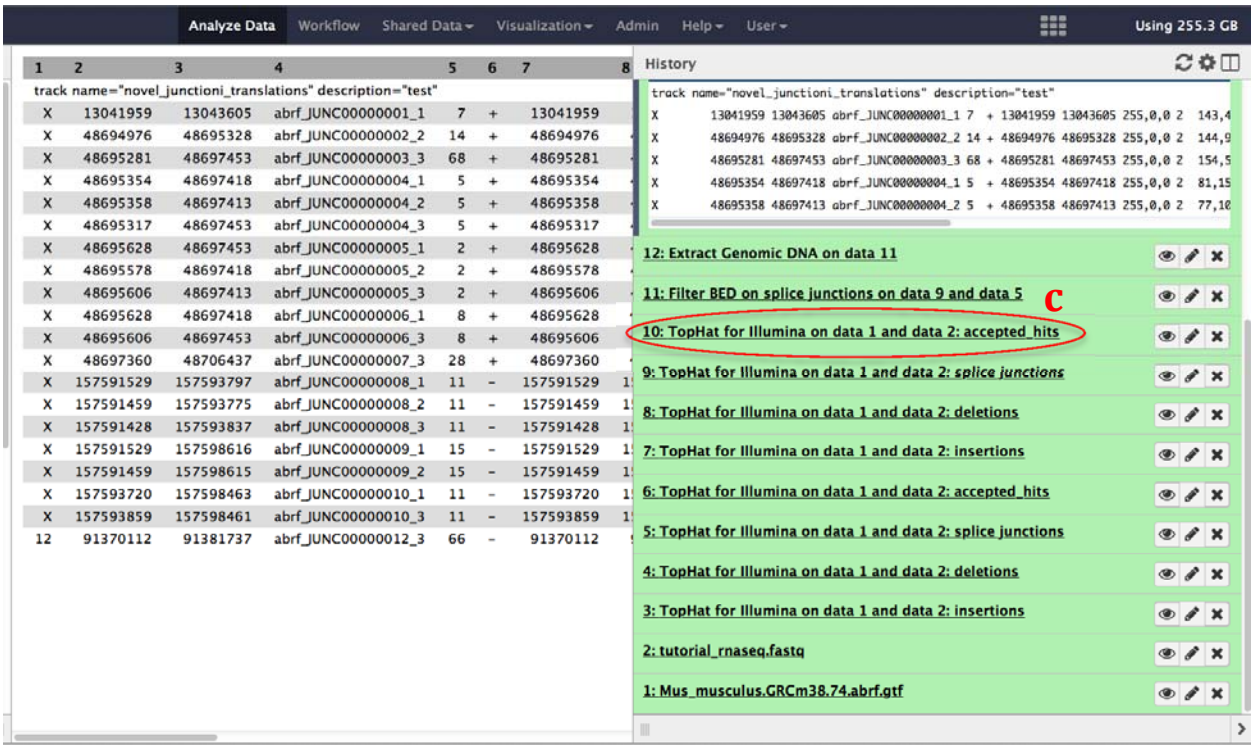
a) Click on the "**…Translate BED Sequences on data…**" file in the history pane to expand it and click the link "local" next to "display with IGV web current "
b) You should now see a track in IGV named "novel_junction_translations"

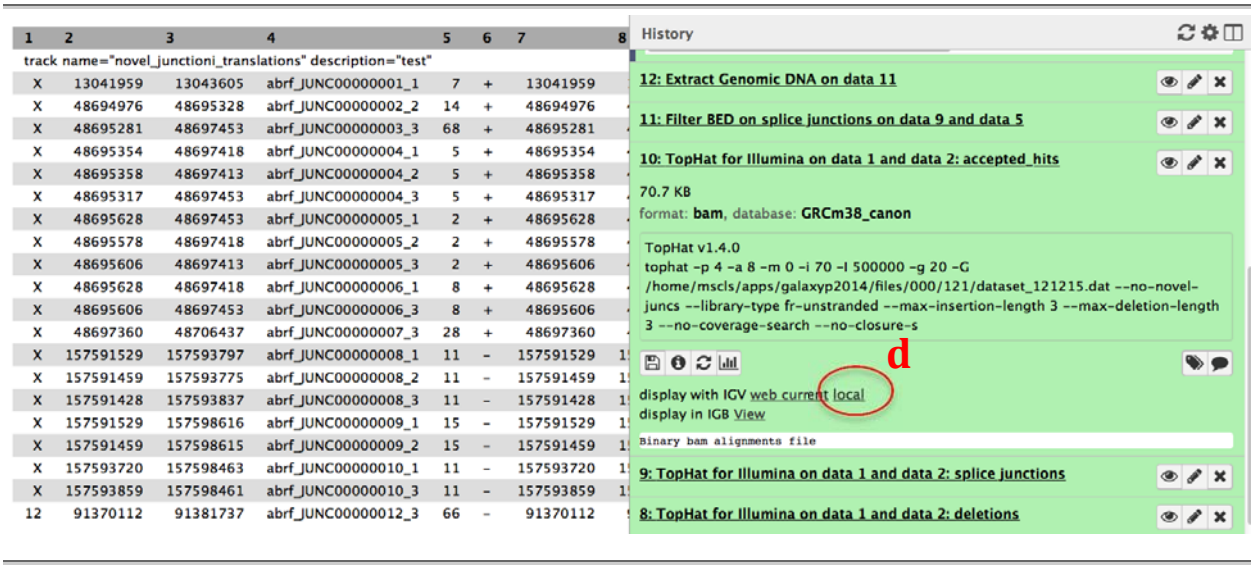




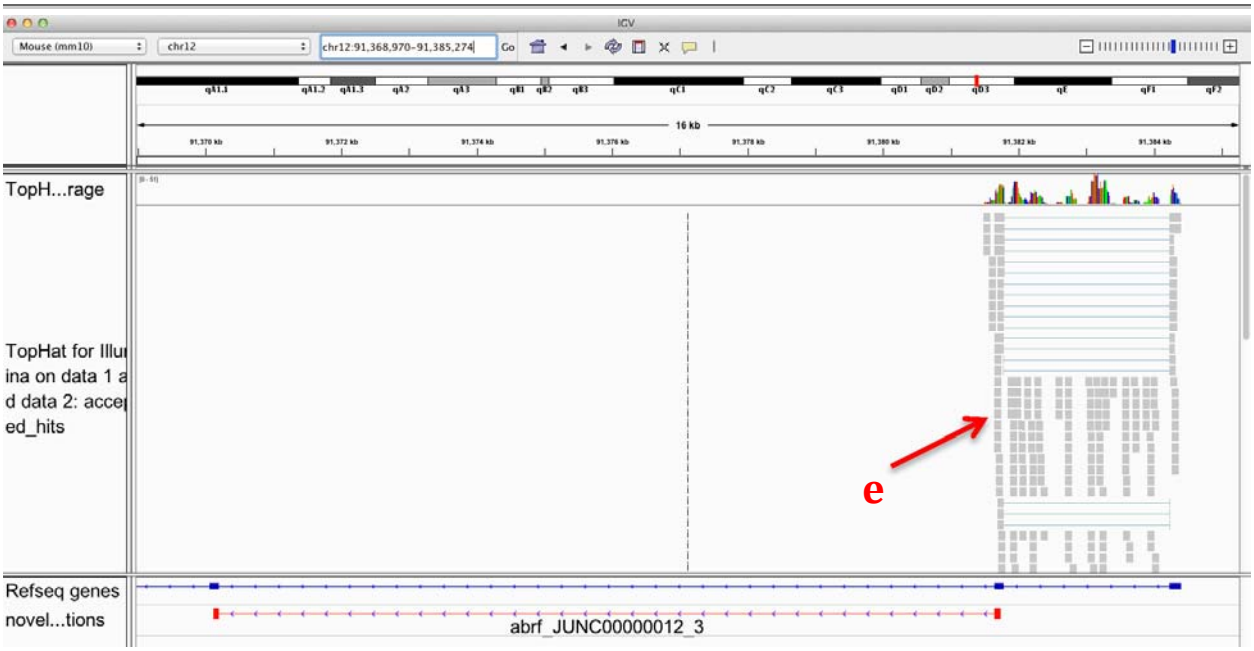
c) Back in GalaxyP, click on the output from your second Tophat run with "**accepted_hits"** as part of the name (**10 TopHat for Illumina on data 1 and data 2: accepted_hits)** to expand it.

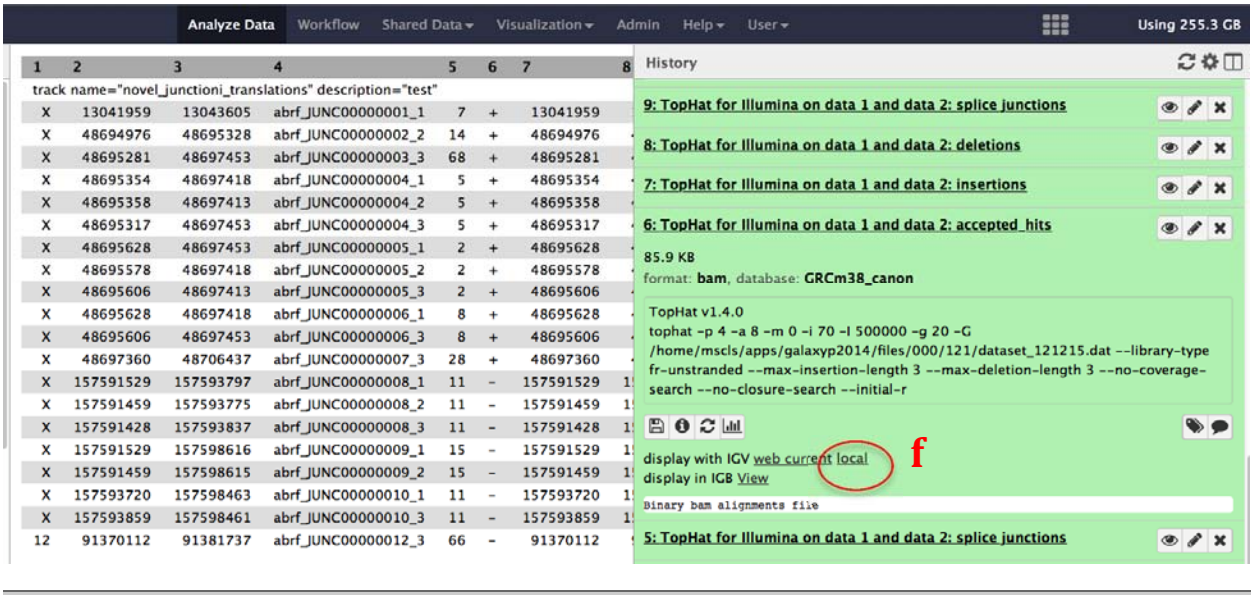
d) Click the link "local" next to "display with IGV web current "





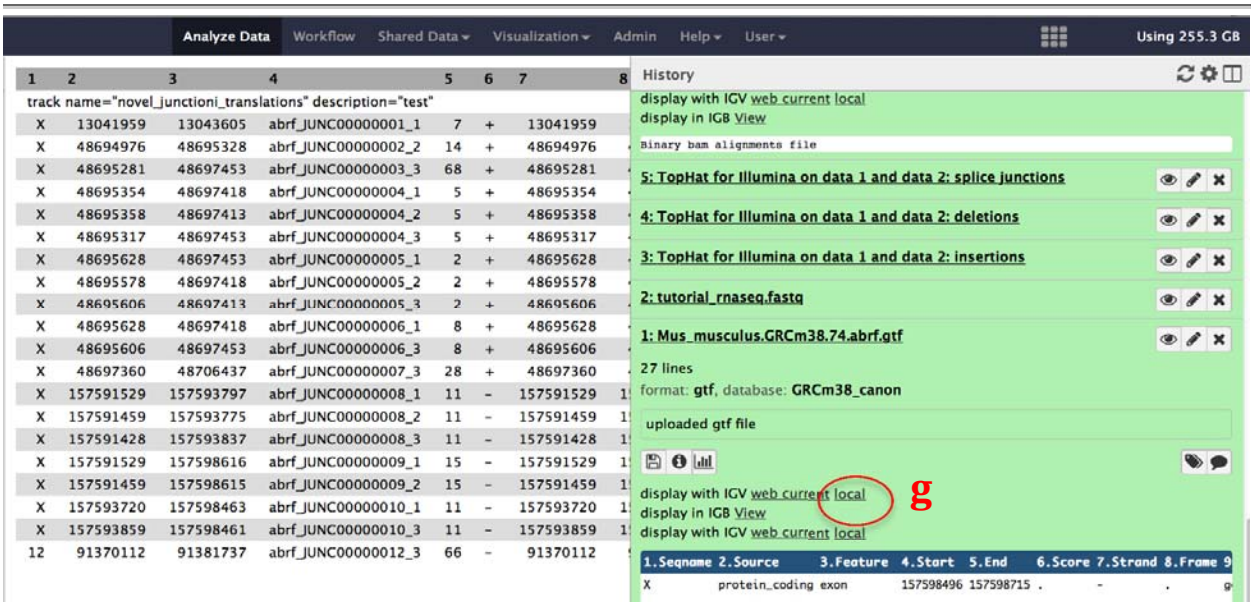
e) The BAM file with rna-seq reads should now be loaded in IGV
f) Go back to GalaxyP again, click on the output from your first Tophat run with "**accepted_hits"** as part of the name to expand it and click the link "local"

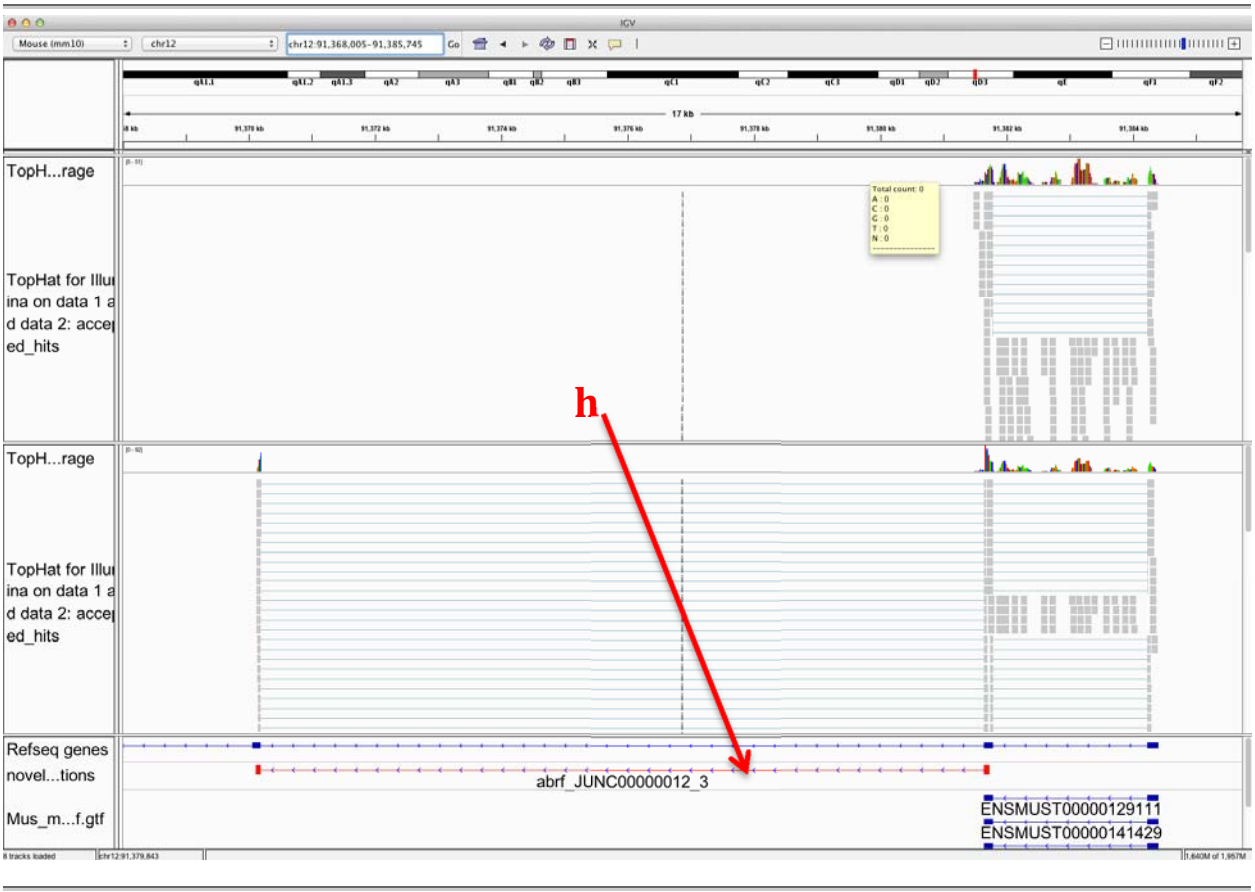




g) Click on the GTF file to expand it and then click on the link "local" to also load this GTF file into IGV

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)





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

ABRF 2016 ANNUAL MEETING
February 20-23

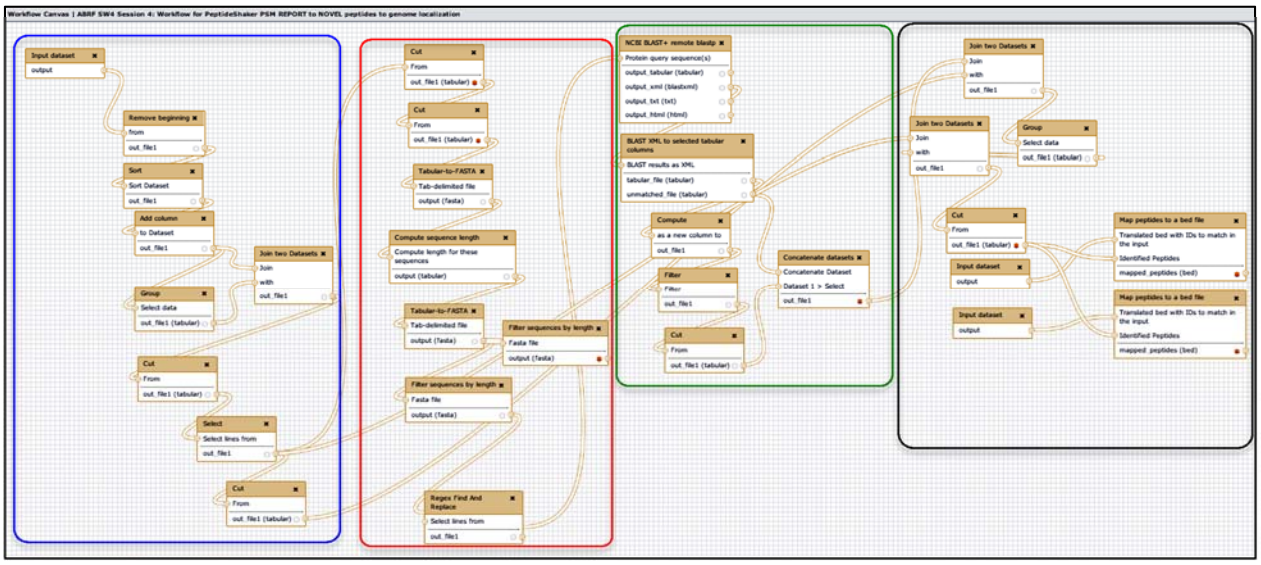
1 Introduction

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



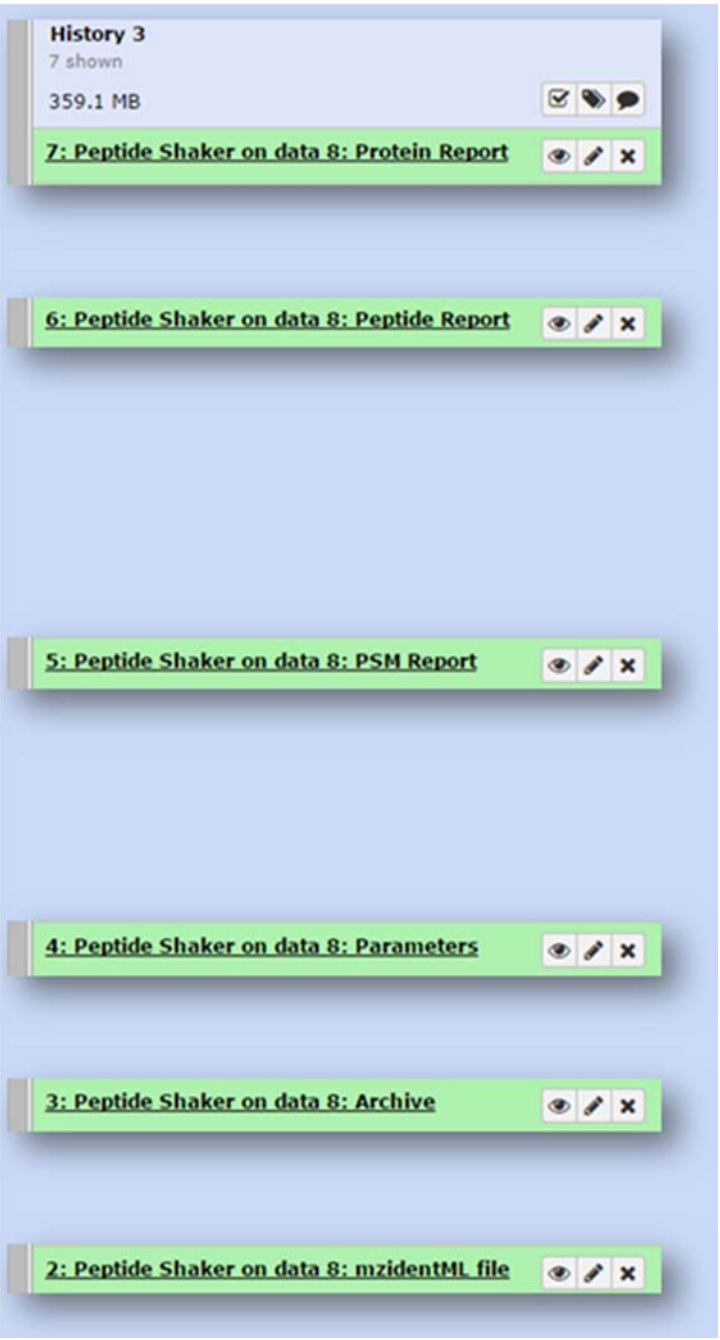
Generating a PSM summary of peptides derived from RNA-Seq derived db

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

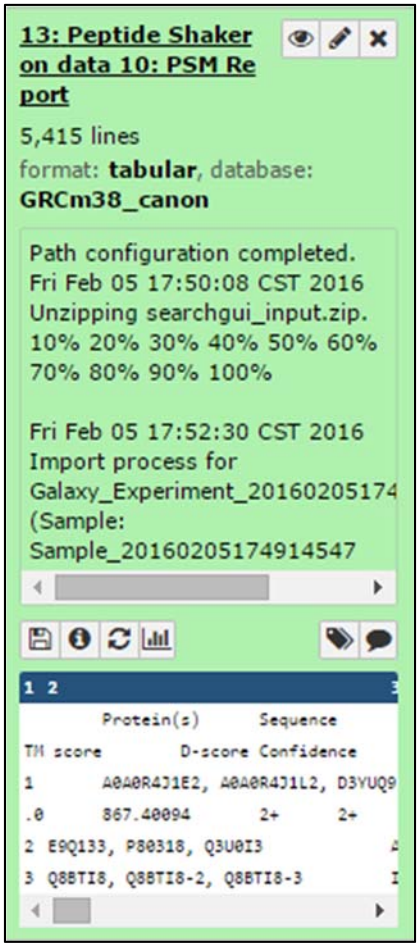
Summary (Parameters)
- valid peptides
- valid proteins
- valid spectra

Archive (zipped file)
- CPS file to visualize data

mzIdentML
- PSM Visualization
- SWATH Analysis
- Skyline
- Scaffold

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).

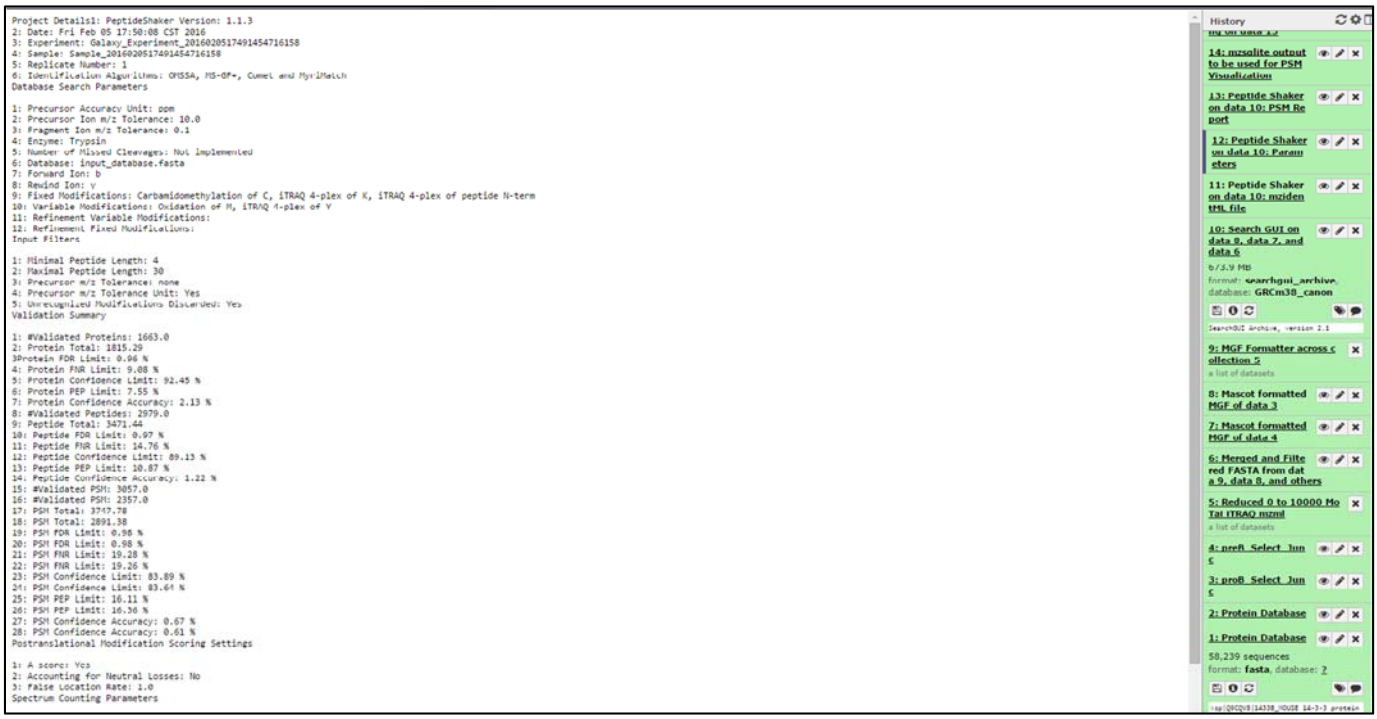


PSM Report

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*

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.

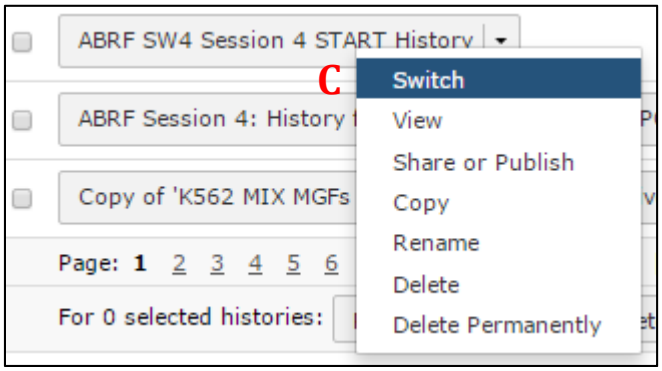


We will be processing the PSM Report and using its outputs to visualize PSM quality and localization of novel peptide identification on the mouse genome.

2.3 Import tutorial datasets into current history

Tutorial Dataset:

a) At the top select "Saved Histories"
b) Select "ABRF SW4 Session 4 START History" from the list of histories
c) Select Switch.



3 Running workflow for session 4

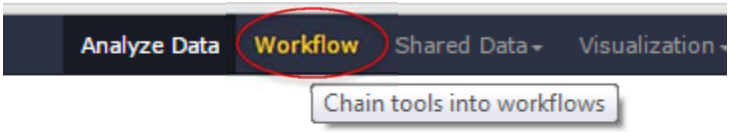
3.1 Inputs for the session 4 workflow

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

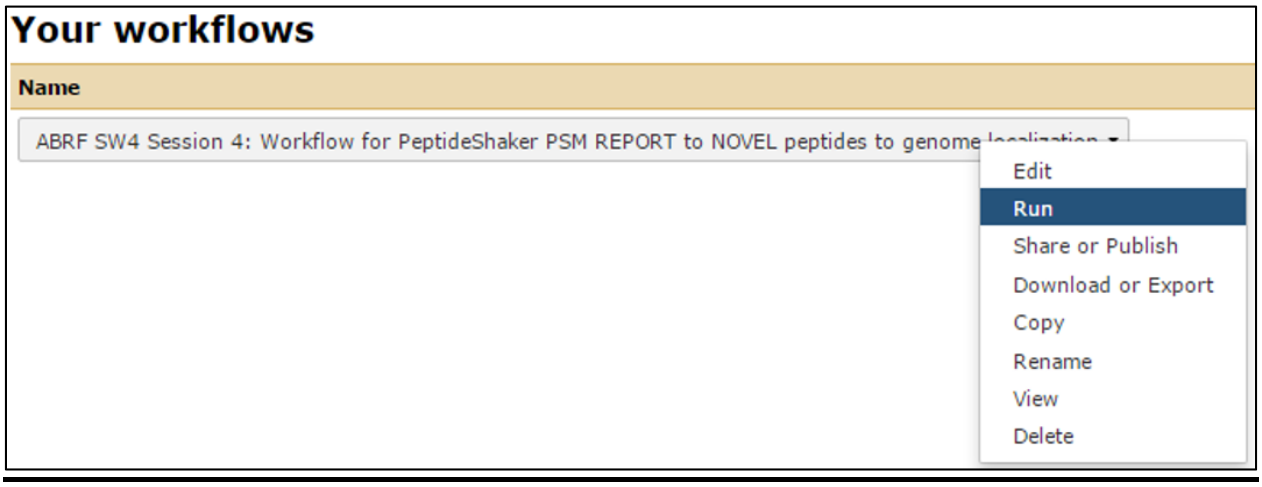
3.2 Workflow for session 4

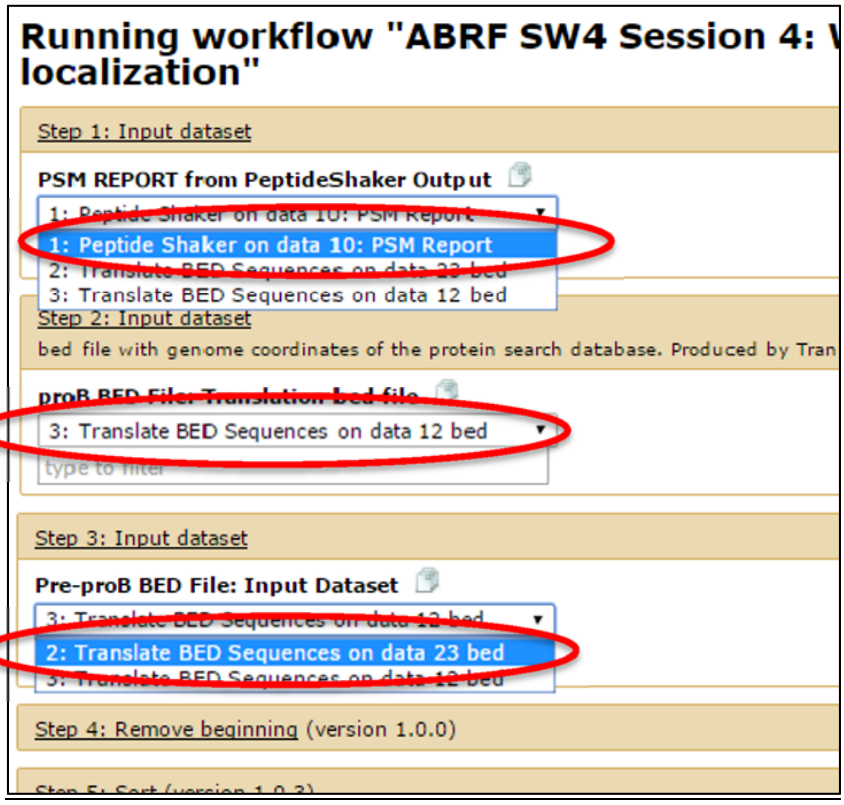
Select ABRF SW4 Session 4 START History as your active history.

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

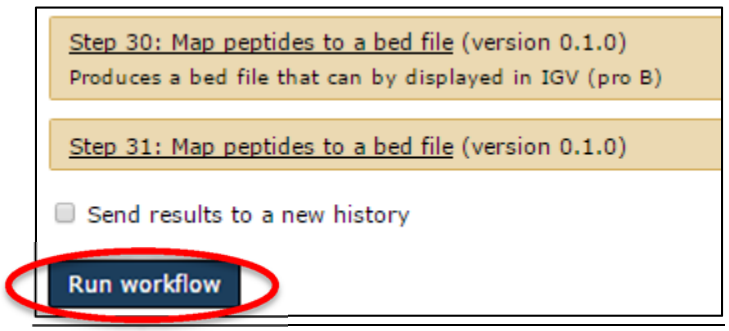


Run the workflow on the active history.

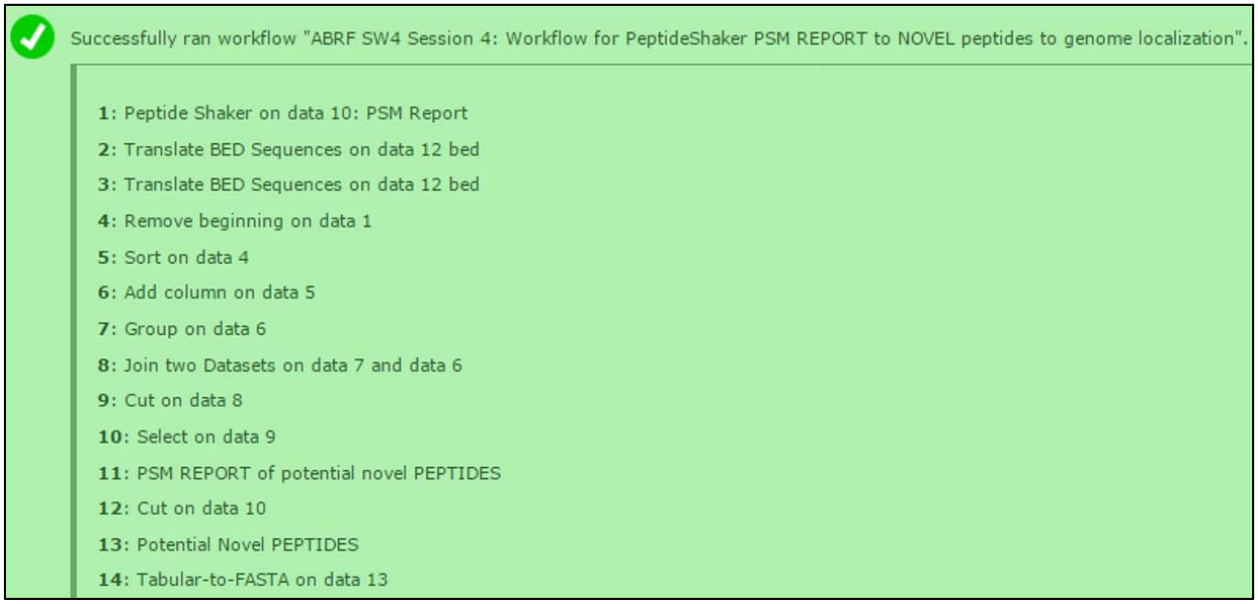




Select ‘PSM Report’, for Step1, ‘Translate BED Sequences on data 12 bed’ for Step 2 and ‘Translate BED Sequences on data 23 bed’ for Step 3.

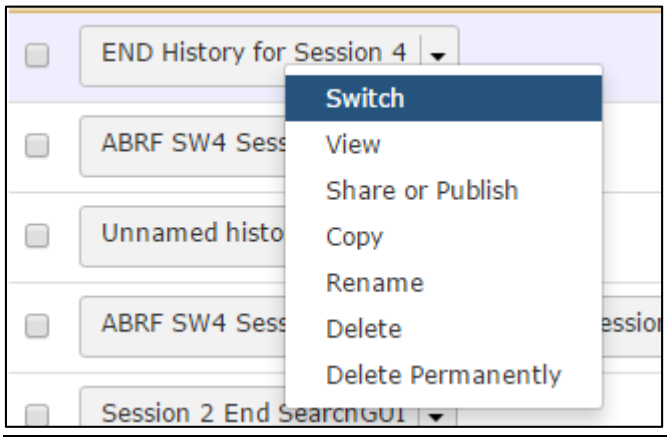


Run Workflow.



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.



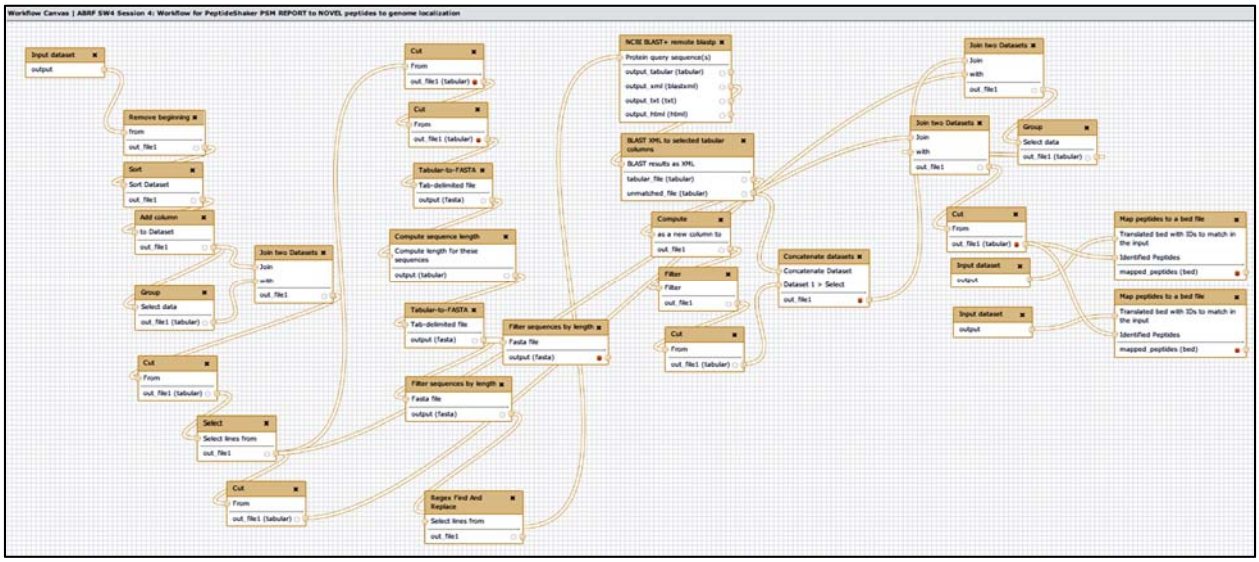
Workflow for Session 4 is a slightly complex workflow than the preceding one, which involved only SearchGUI and PeptideShaker. In this workflow, 29 processing steps are used to take the PSM Report from PeptideShaker and manipulate it to put through BLAST-P analysis to verify novel proteoforms.

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.

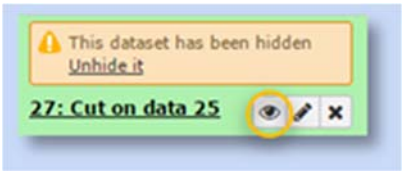


Overview of Session 4 Workflow

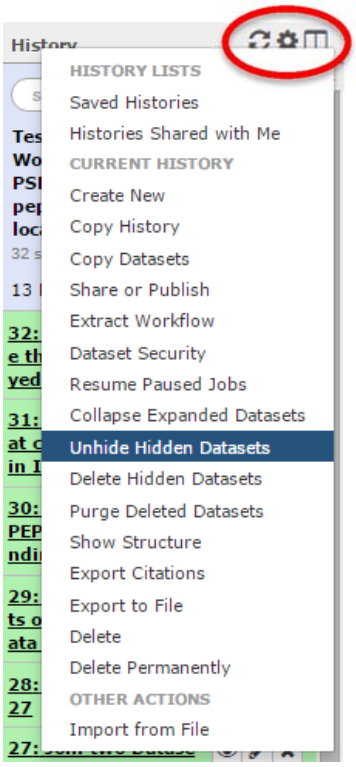
Step 1: Input dataset (PSM Report)

- 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.

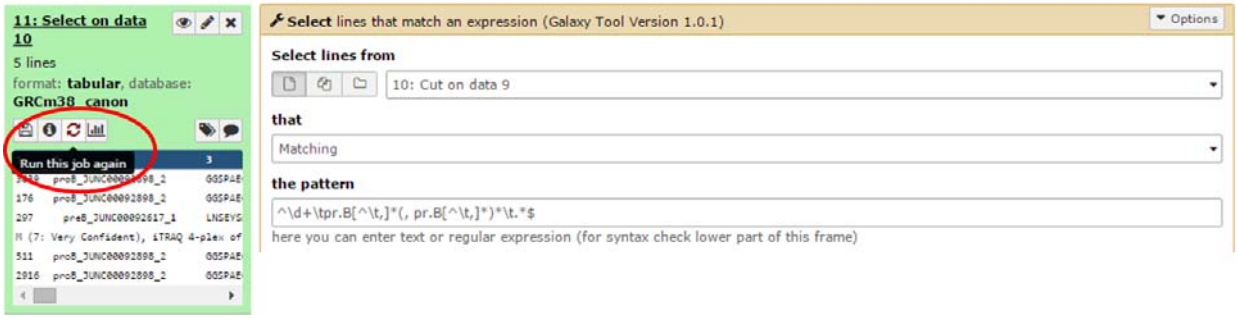
Step 7: Ranks columns based on the new sorting performed in Step 6.

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.

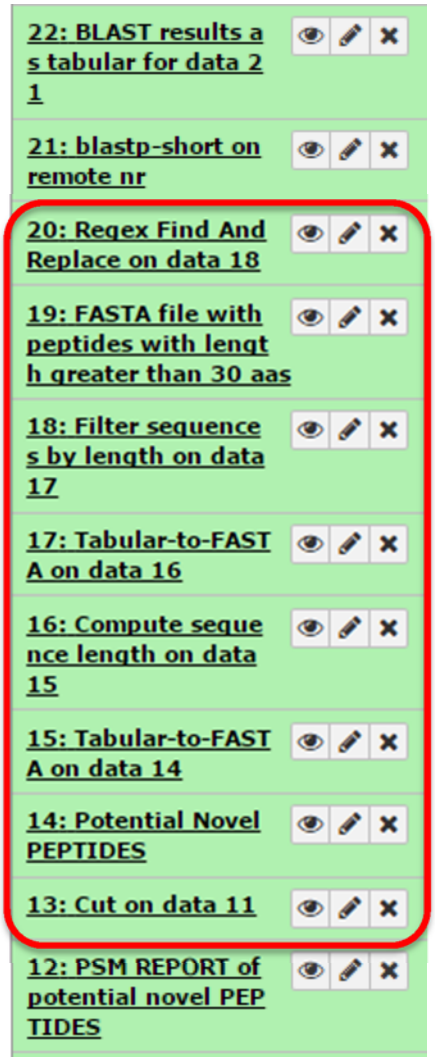
Step 11: Selects peptides with accession number starting with preB or proB. (Details in figure below)

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



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

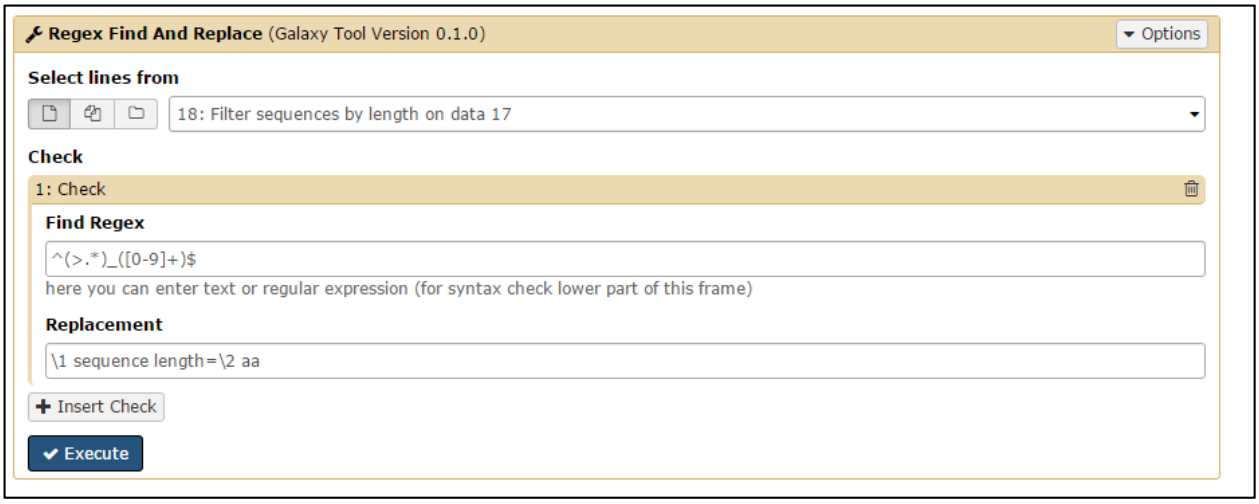
Step 16: Computes sequence length on data 15.

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 bu short BLAST-P search.
>PEPTIDE_sequence length=length aa
PEPTIDE

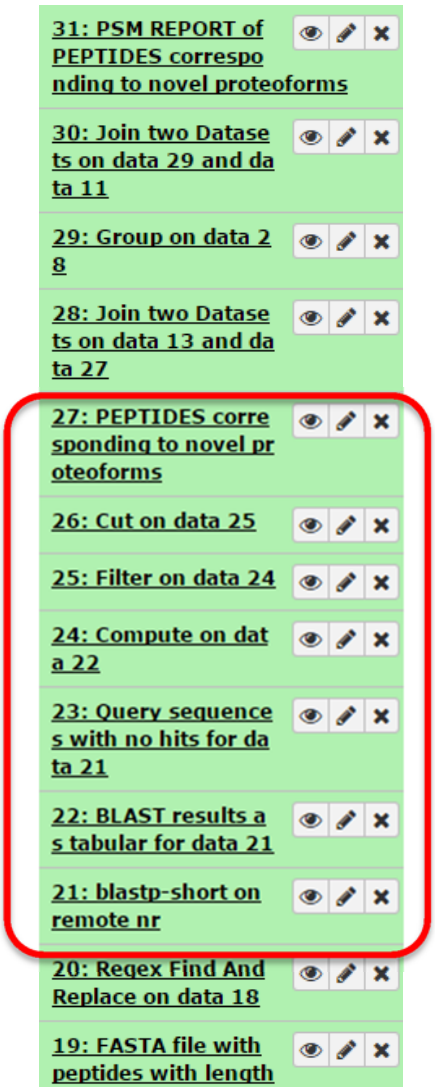


Step 20: Converts Step 18 output to a format so that it can be searched by using short BLAST-P search.

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.



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.

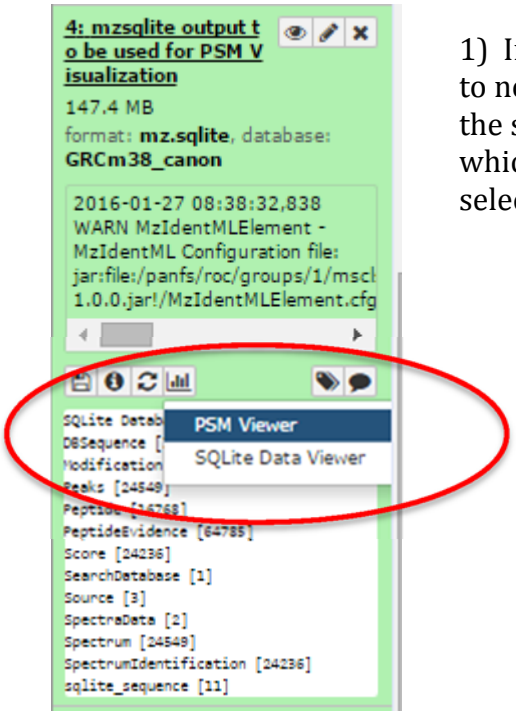
3.6 List of peptides corresponding to novel proteoforms and PSM visualization.

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.

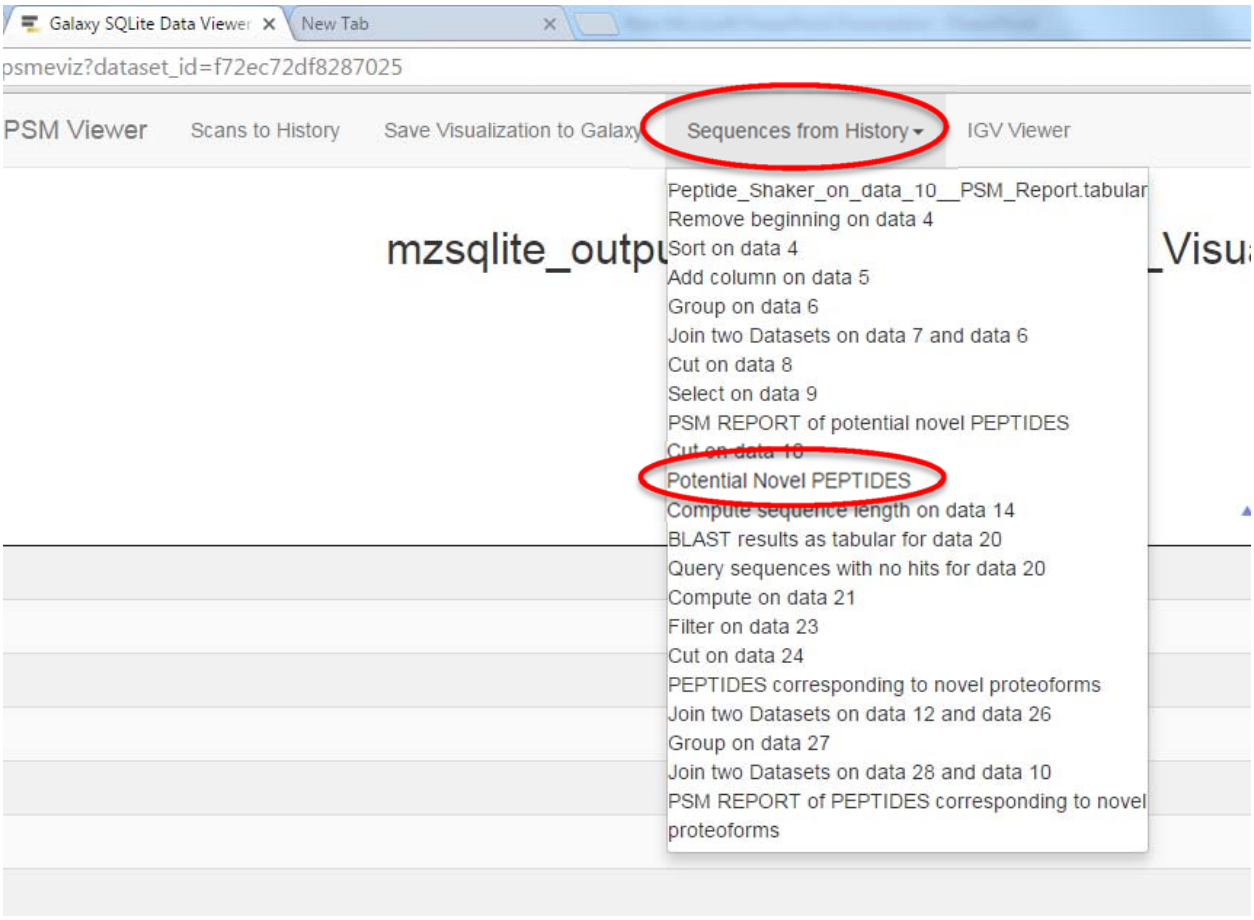
Visualization with Peptide-Spectrum-Match Evaluator
The Peptide-Spectrum-Match (PSM) evaluator tool is a unique visualization tool to GalaxyP. Using the experimental peak lists and corresponding analyzed database search report (peptide report or mzid) PSM Evaluator will render each peptide for visual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



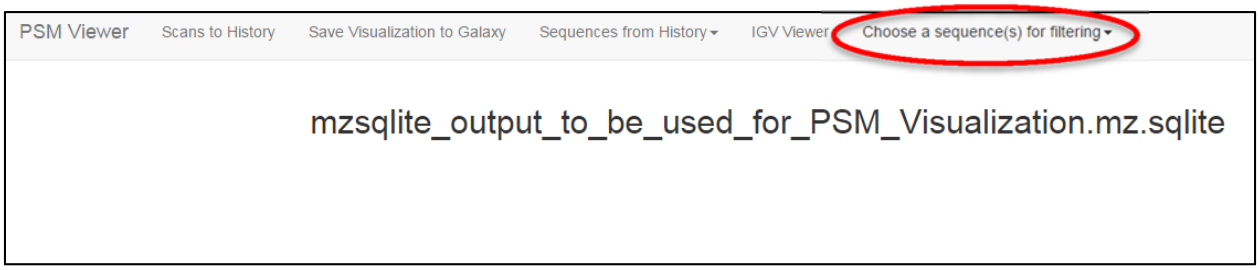
1) In order to visualize PSMs of the peptides corresponding to novel proteoforms, go to step 6 and click on the title of the step. Once the box has expanded, click on the icon which representing a bar diagram (for visualization) and select PSM Viewer and click.

2) This would result in opening of another Tab.

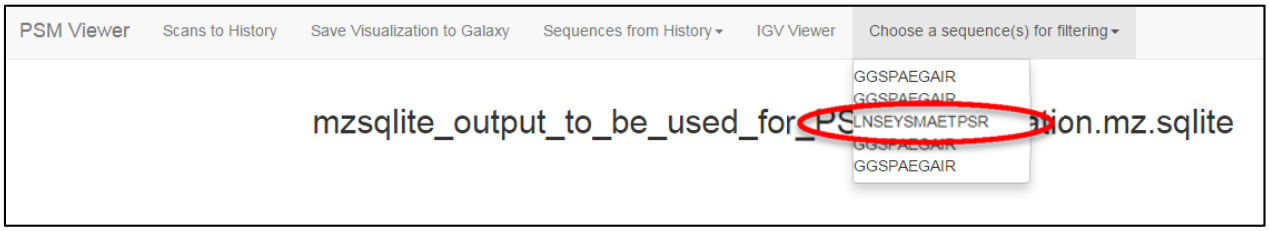
3) Select on 'Sequences from History' and Select Potential Novel PEPTIDES:



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

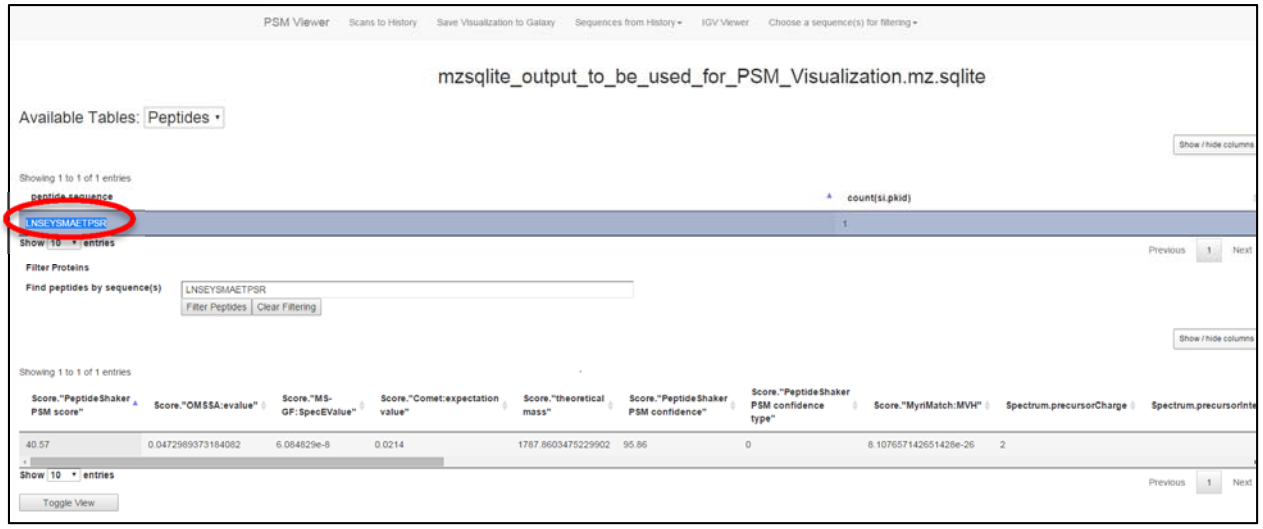


5) Select on peptide 'LNSEYSMAETPSR'

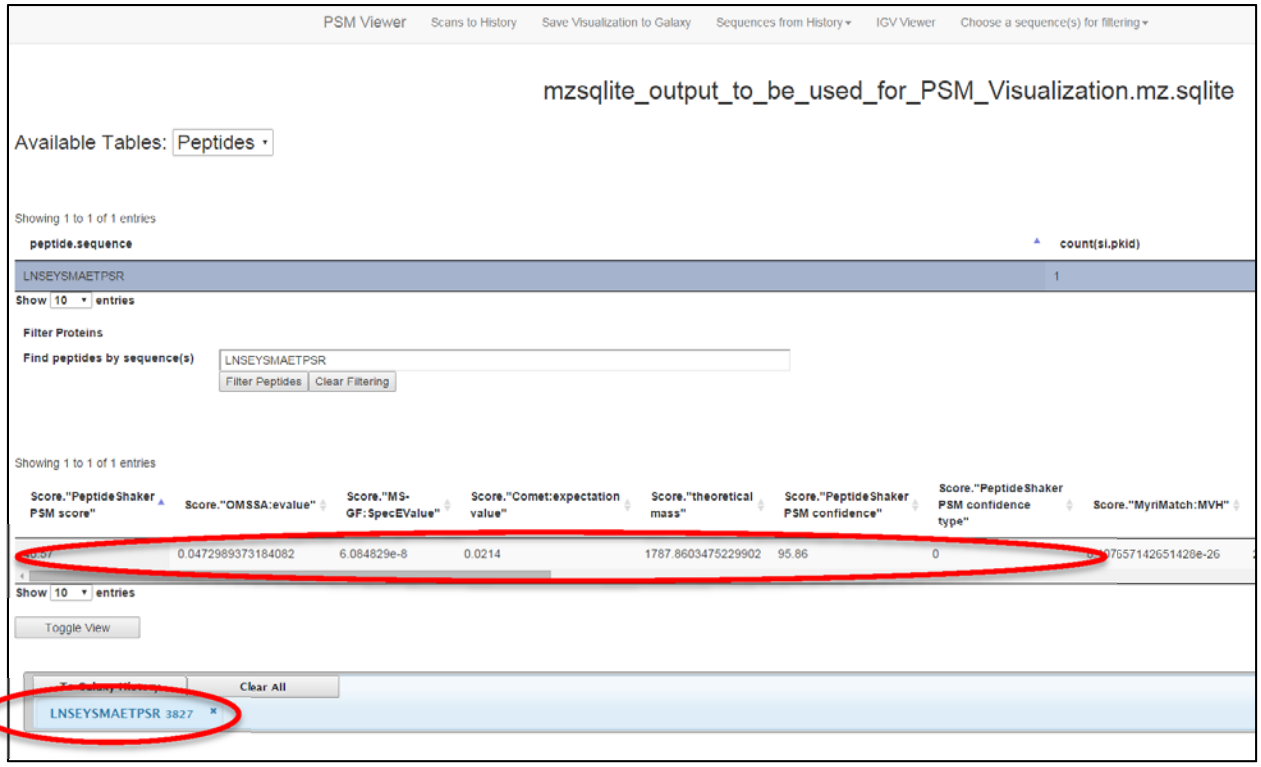


6) Click on Filter Peptides.

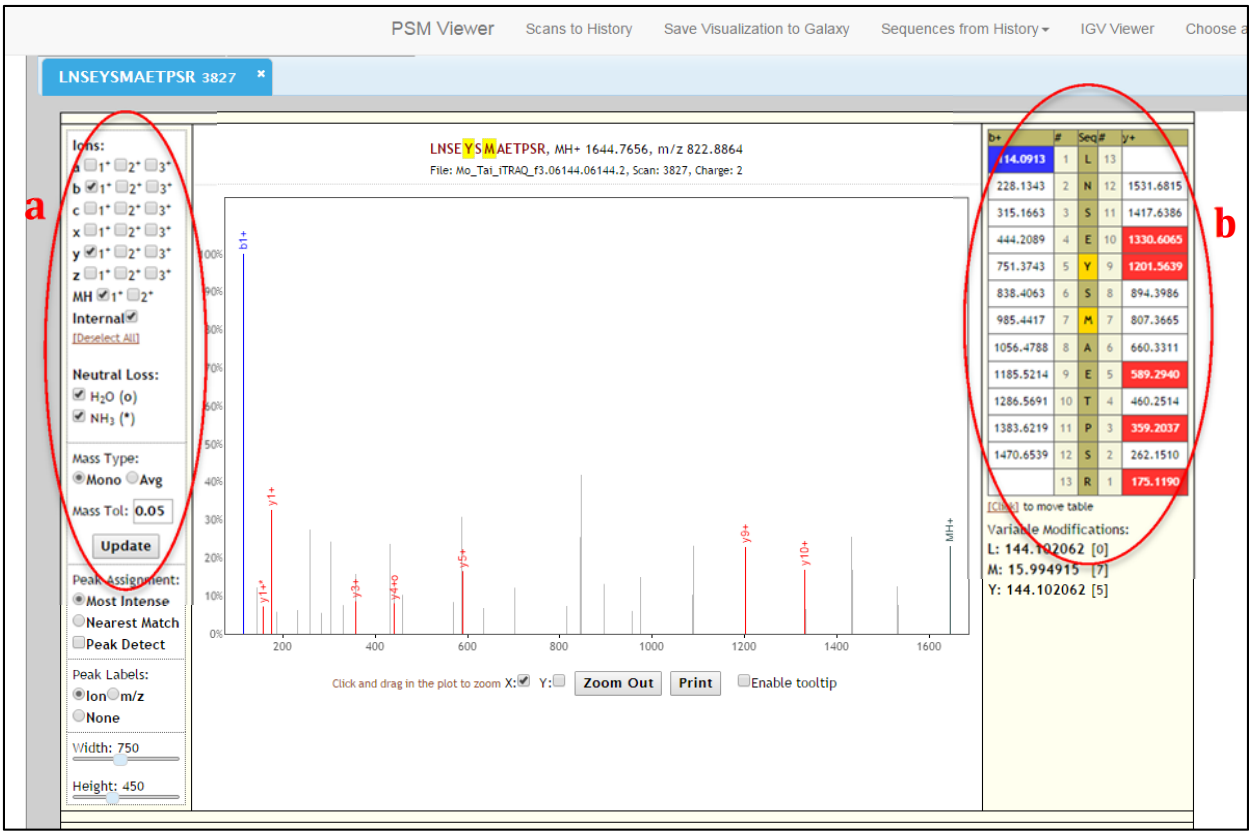
7) Click on Filter Peptide Tab



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



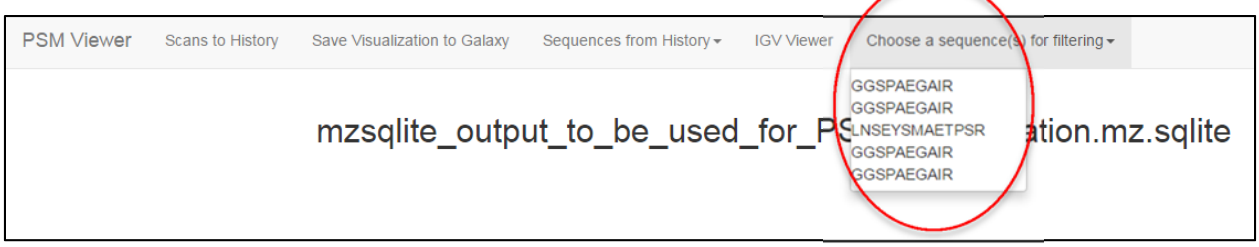
9) Clicking on the Tab and appropriate parameters reveals the PSM for 'LNSEYSMAETPSR' – our first candidate:



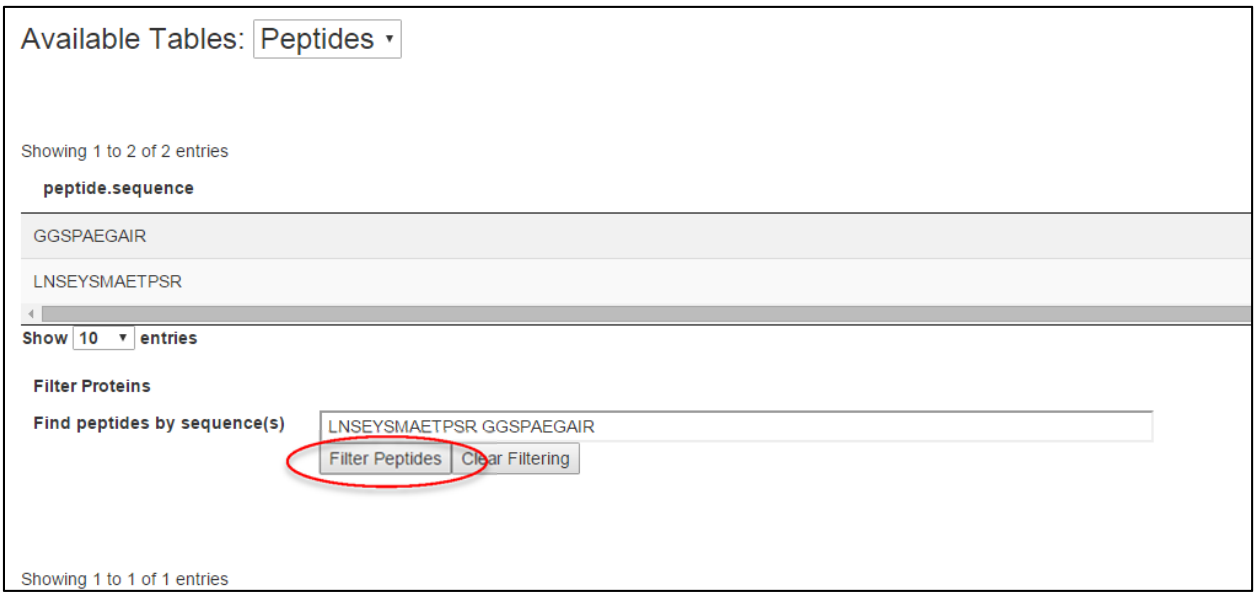
a. To visualize fragment ion series select the desired series under the ions interface. An accepted approach to validate spectra is to visualize all b^{1+}, y^{1+}, MH^{1+}, Internal Ions, all Neutral Losses. Visualize at a Mass Tolerance of 0.05 Daltons.

b. Users can validate the table visually from the spectral graph or from fragment ion table (on the right).

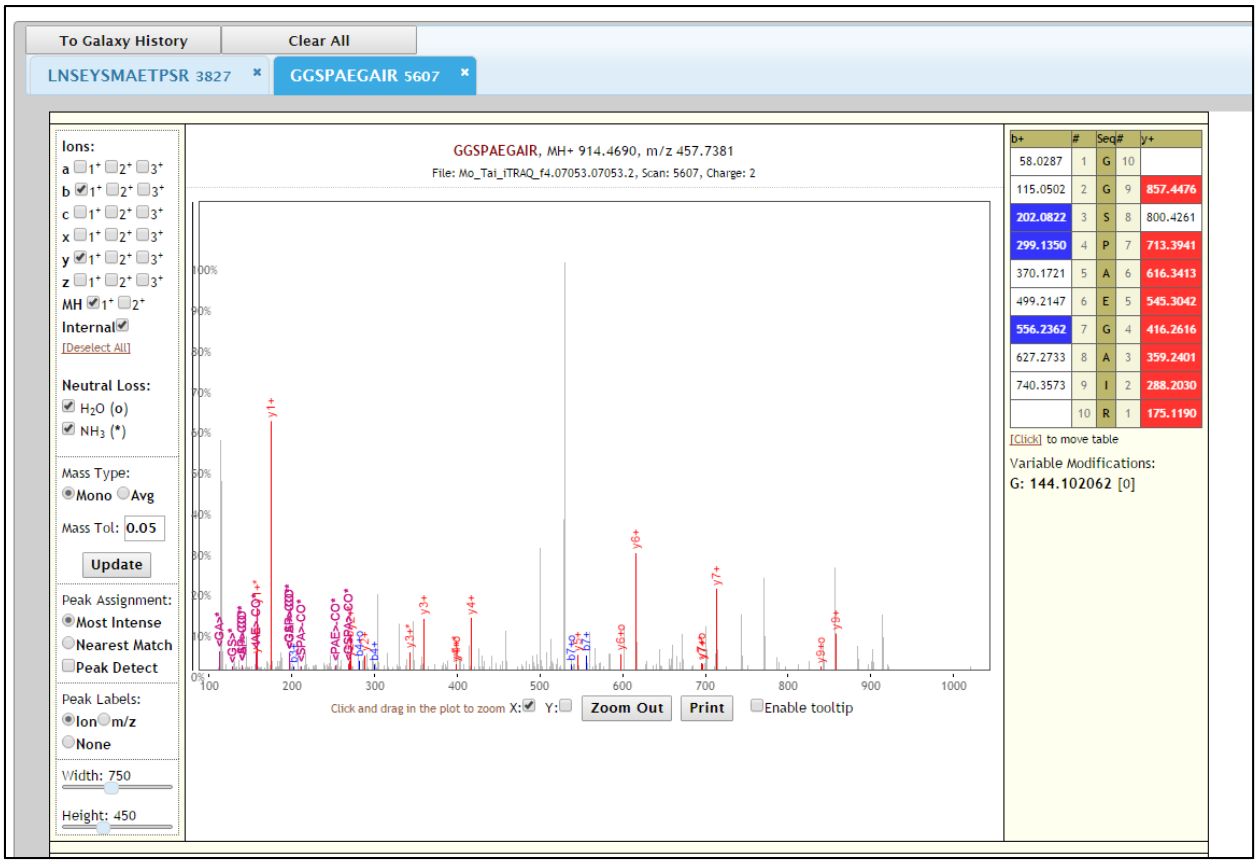
10) Click on "Choose a sequence(s) for filtering" and Select on peptide 'GGSPAEGAIR':



11) Click on Filter Peptides.



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.

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.

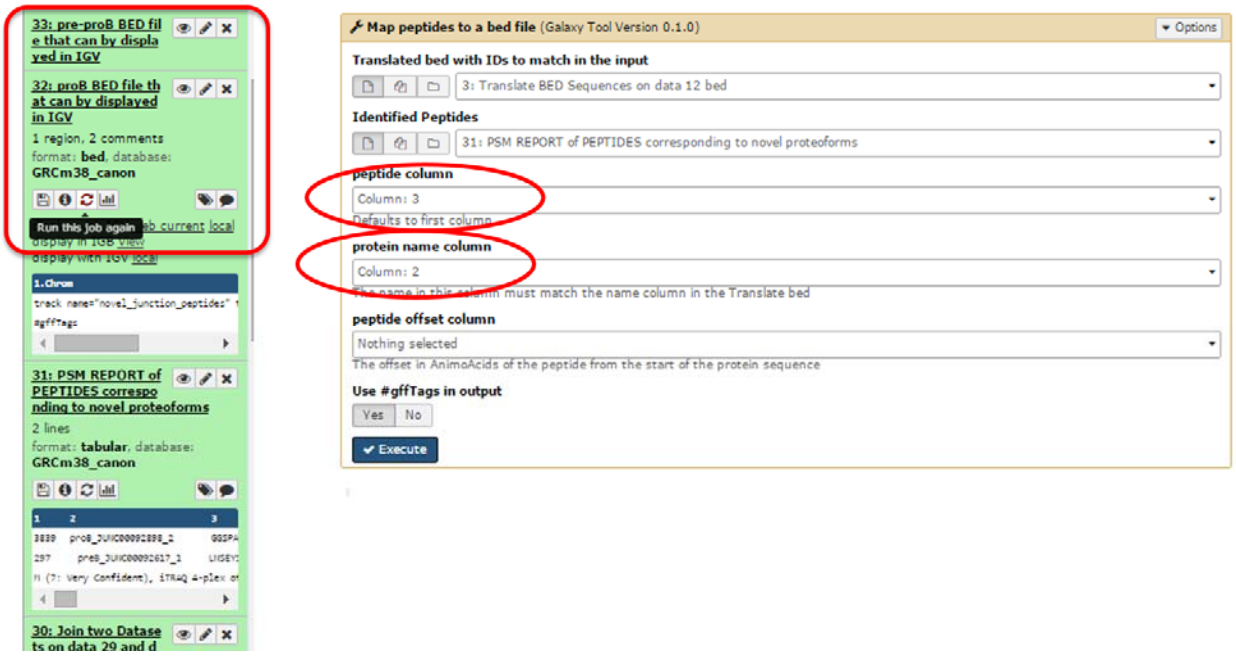
Peptides to Genome
Considerable information can be obtained from the identity of a peptides corresponding to novel proteoforms. Beyond the identity of the aberrant proteins, the localization of each peptide can reveal intriguing genomic architecture. In essence, proteogenomics involves the mapping of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



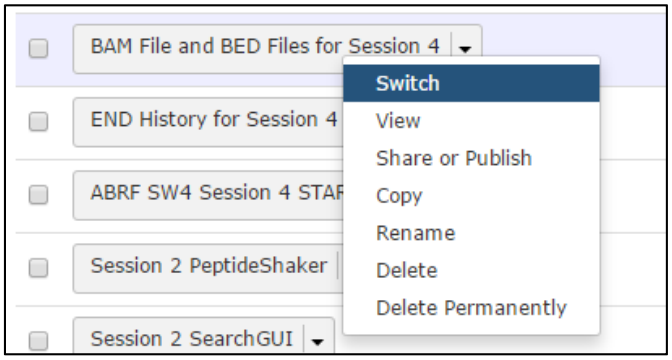
a. The Map peptides to a bed file tool uses BED Sequences generated from Session 2 and PSM Report from Step 31. The tool also needs coordinates for peptide sequence (column 3) and protein (column 2).
b. The tool generates BED files in steps 32 and 33.

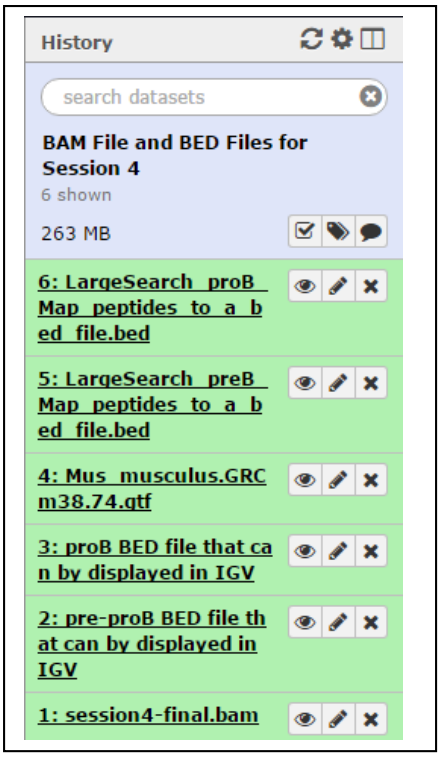
Localizing a subset BED file onto a IGV Browser.

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.

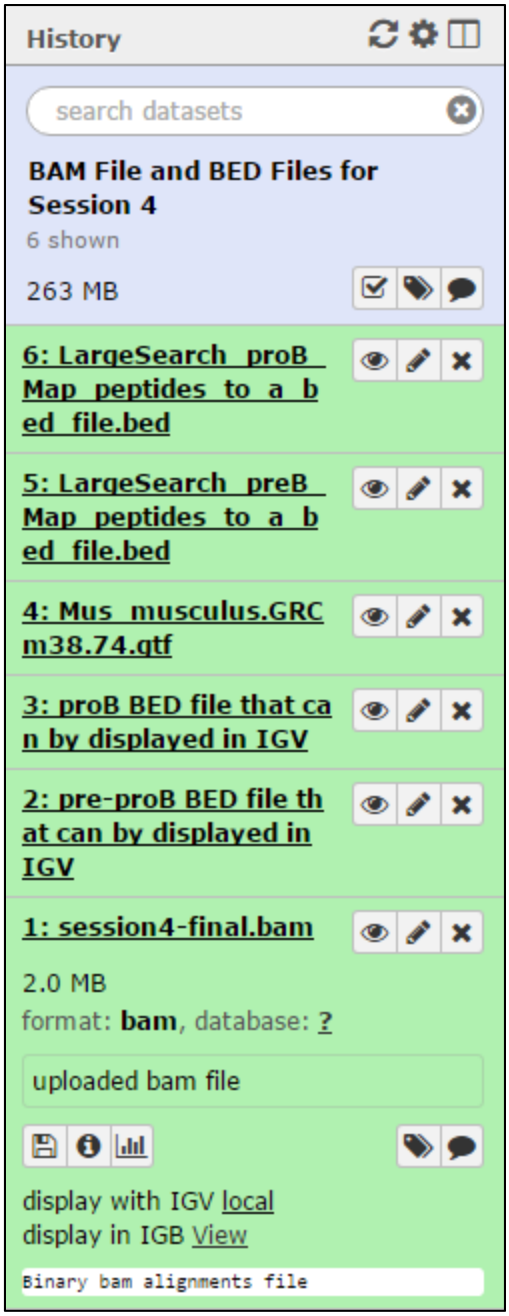
Load files into IGV

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 ".

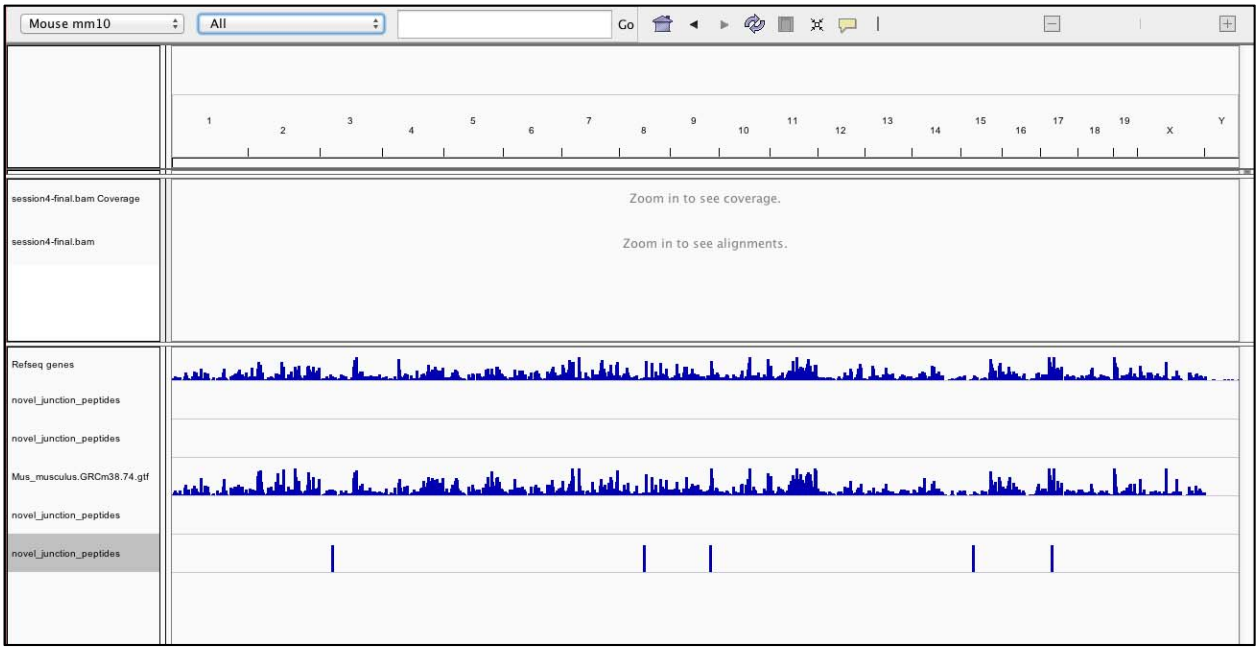


f) You should now see a track in IGV named "session4 final".

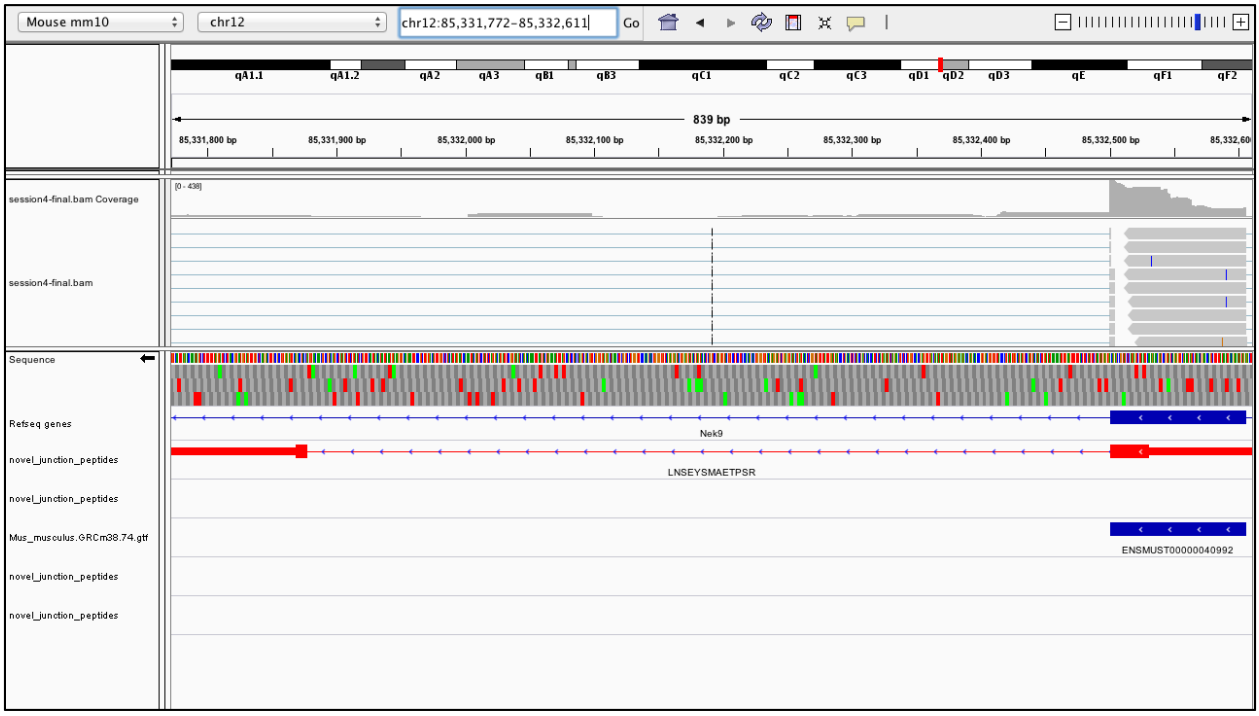
g) Click on the "**pre-proB BED file that can by displayed in IGV"** file in the history pane to expand it and click the link "**local**" next to "display with IGV web current.

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.

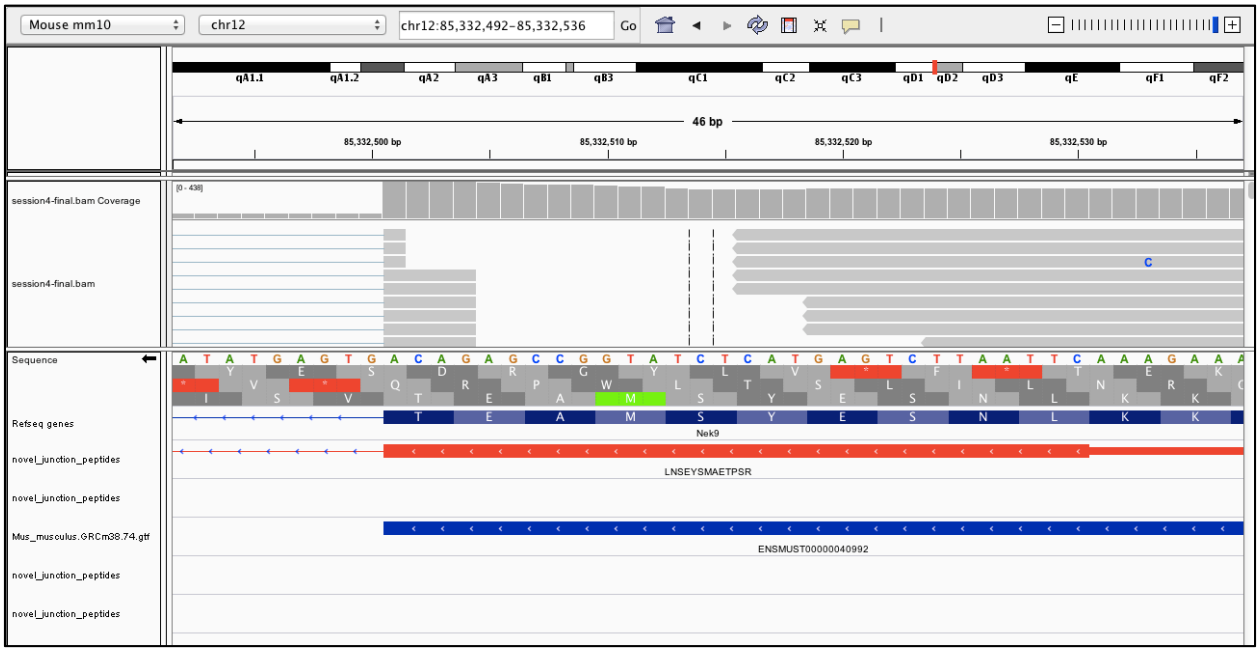
i) Repeat steps g and h for datasets 3, 4, 5 and 6.

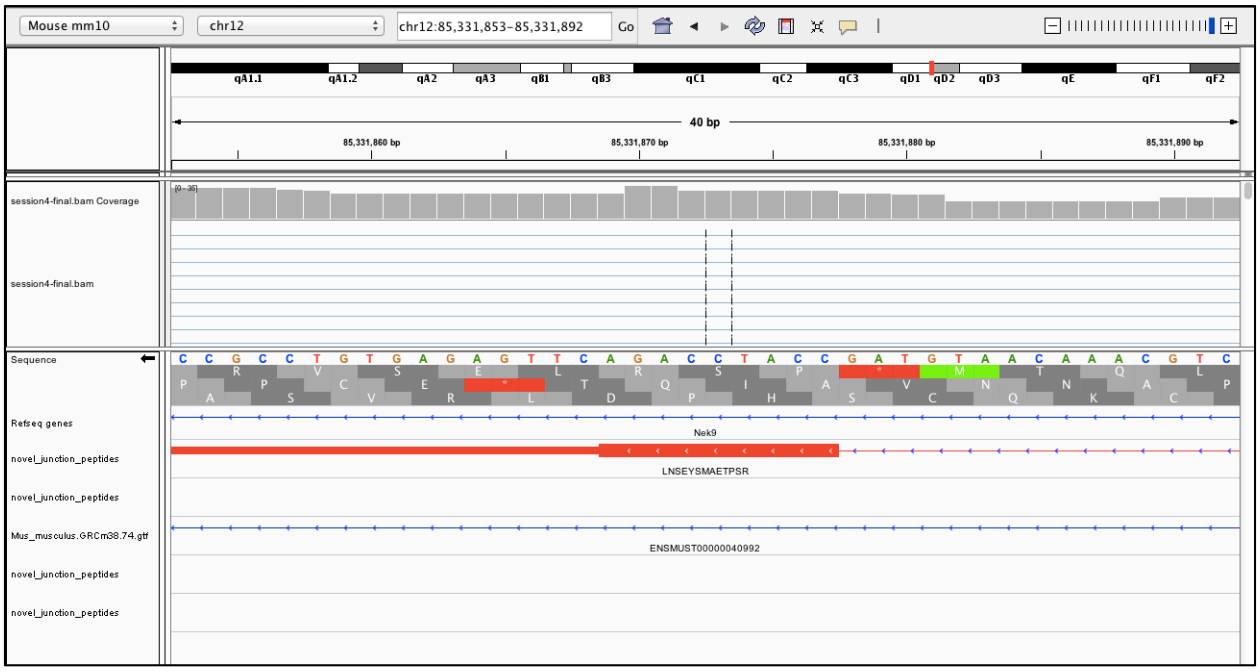


j) Enter "chr12: 85331772 -85332611" into the region field and press "Enter".



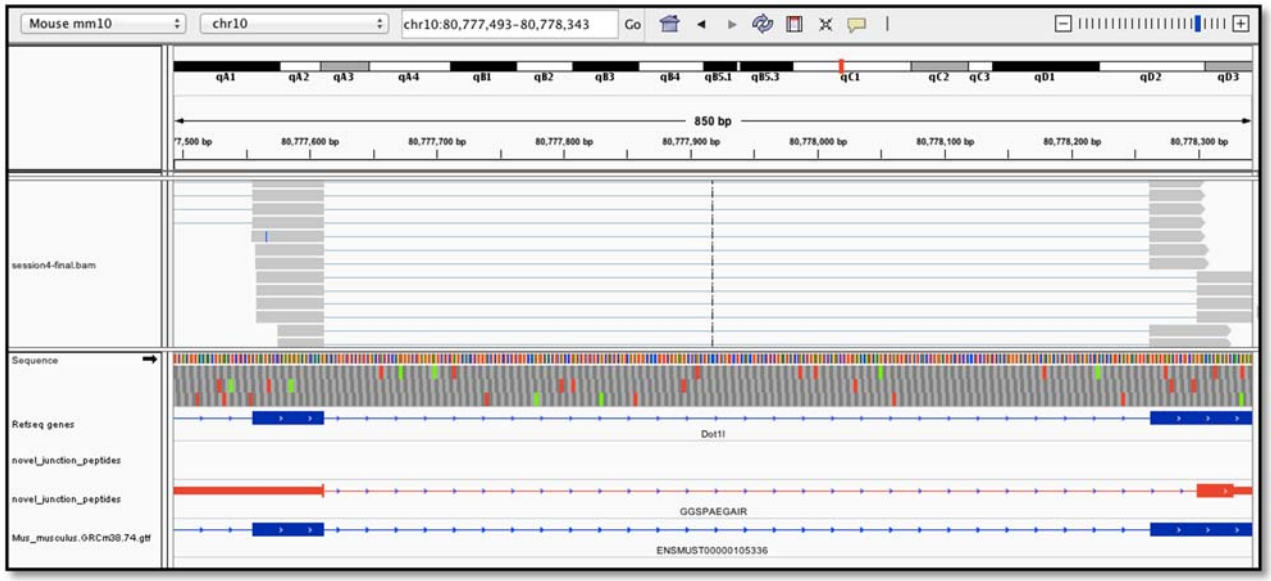
k) Zoom out a little bit to ensure your novel splice junction is in view. Does it make sense why this splice-junction was marked as novel? Explore both the junctions to find out if there is anything worth noting. Also look at the BAM files RNASeq evidence.



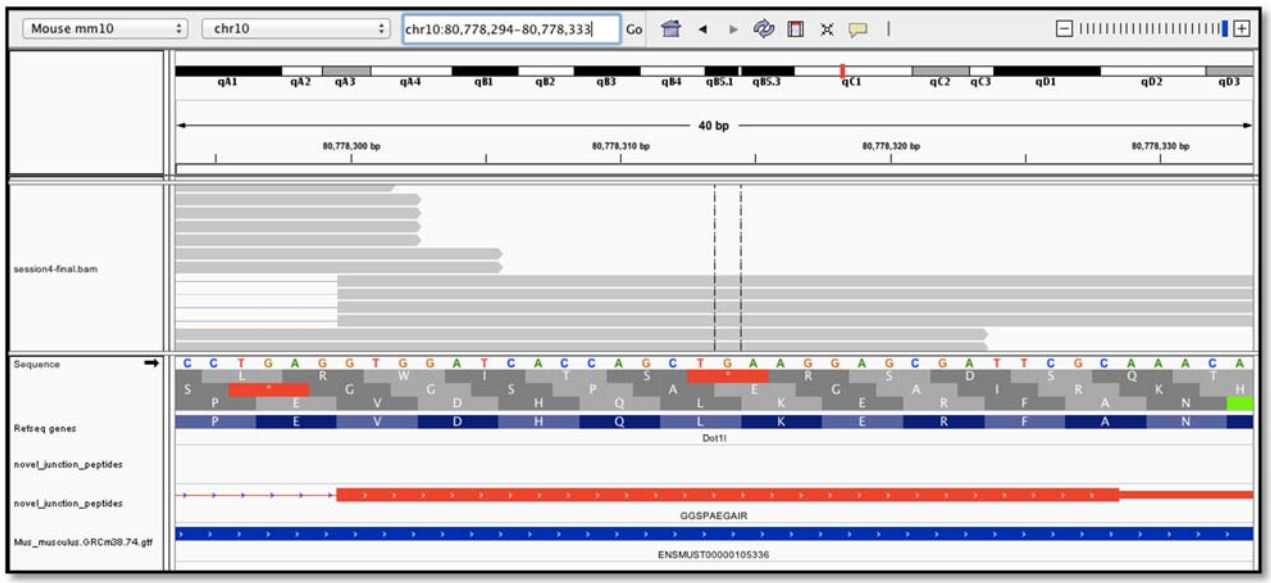


This was the novel peptide wherein the PSM was not of good quality.

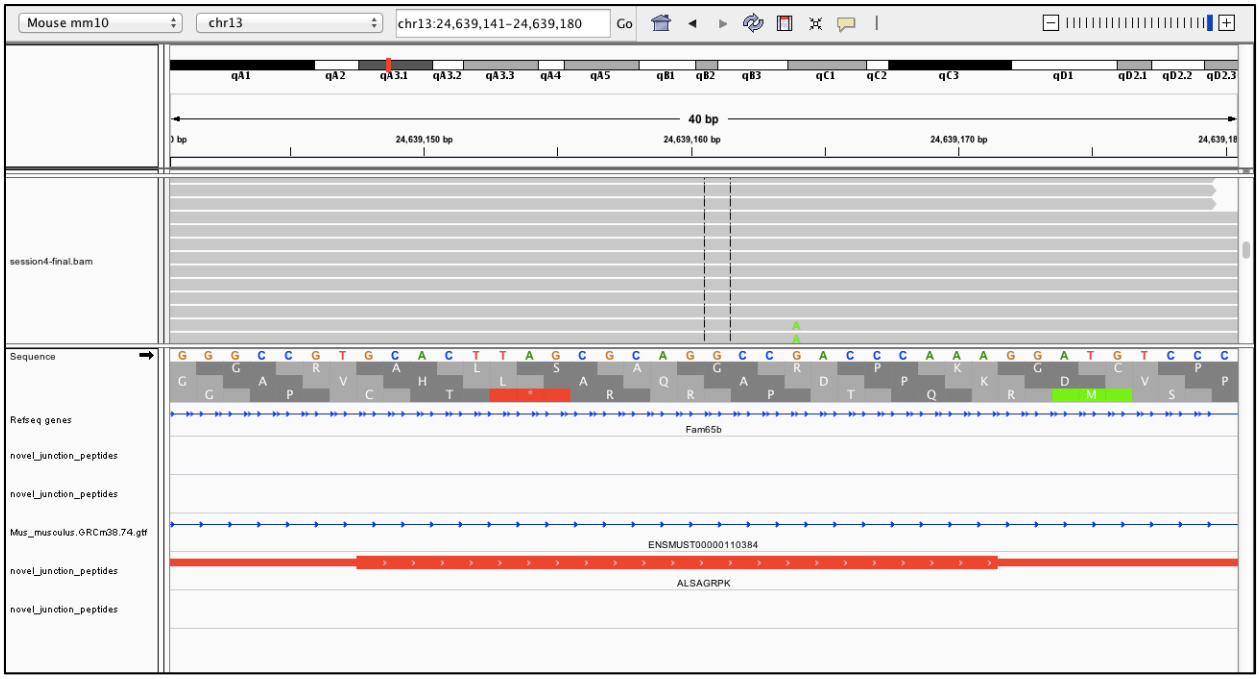
l) Enter "chr10: 80777493-80778343" into the region field and press "Enter". This will zoom into the peptide which had slightly better PSM quality.



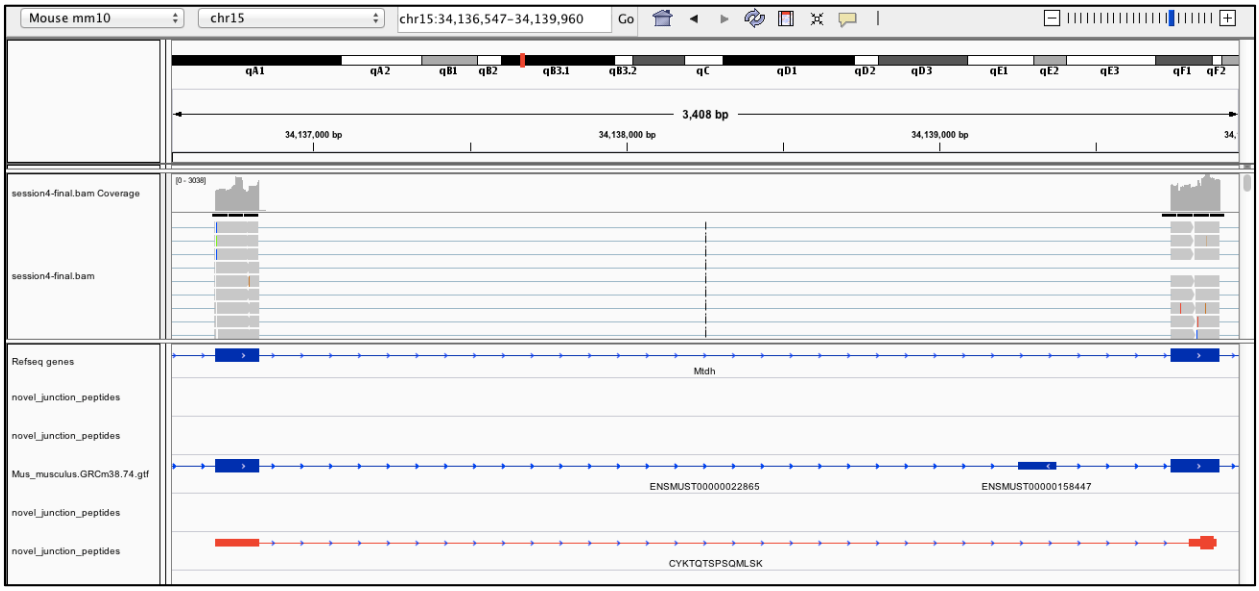
Explore both the junctions to find out if there is anything worth noting. Also look at the BAM files RNASeq evidence.



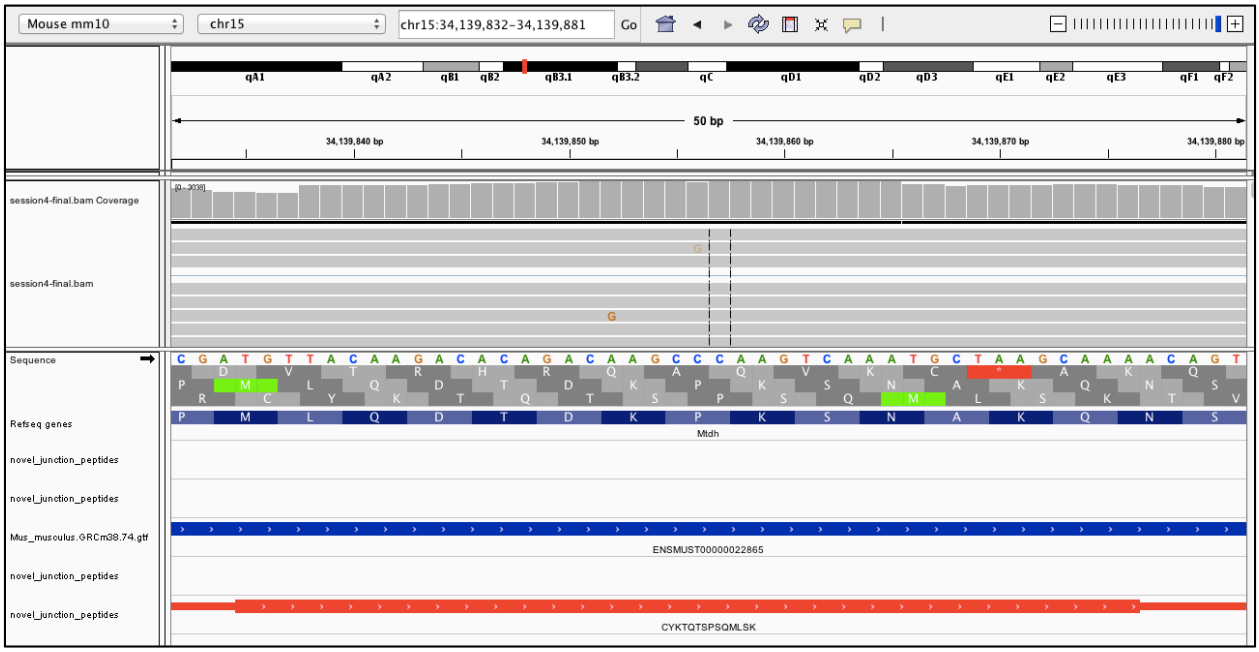
m) Here is another region to explore – “chr13: 24639141-24639180”



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



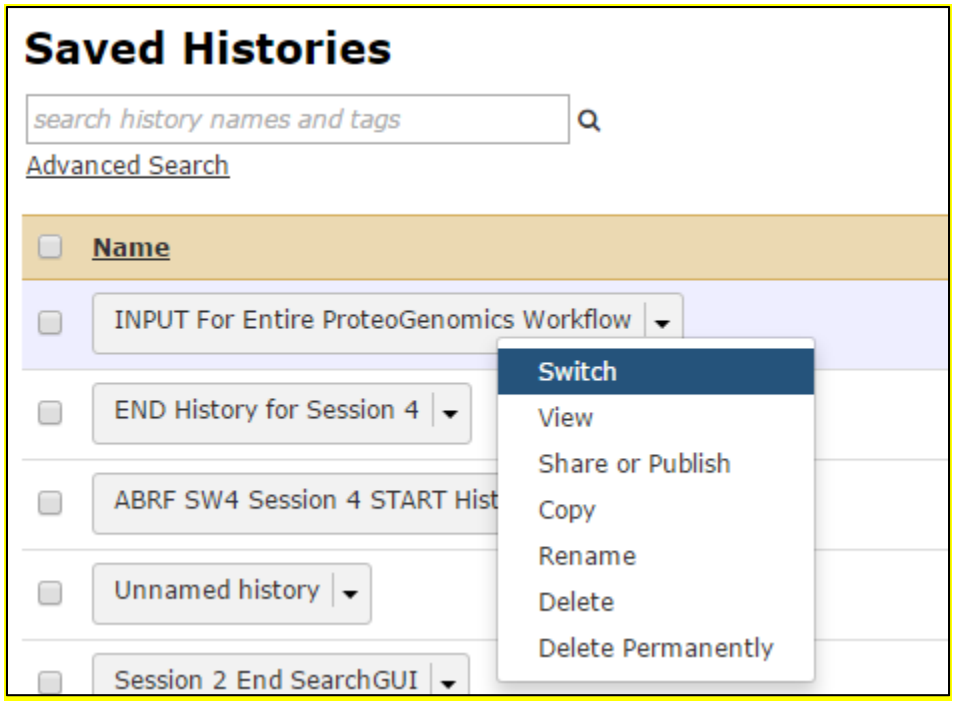
Explore the junction with identified peptide.

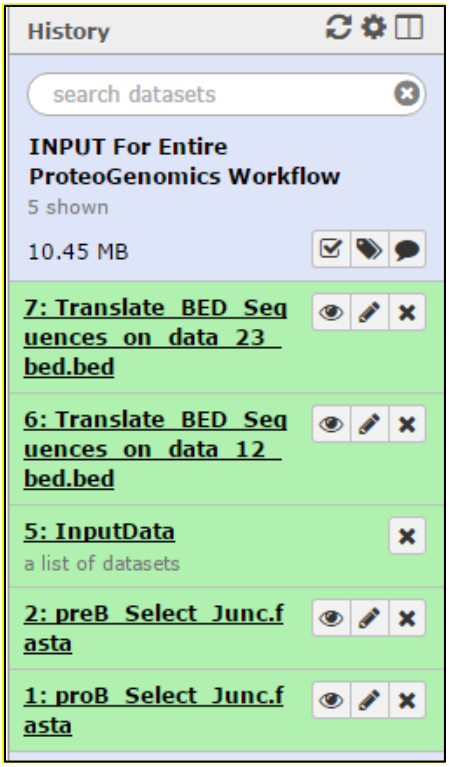

Mouse mm10
chr15
chr15:34,139,832-34,139,881
Go
qA1
qA2
qB1
qB2
qB3.1
qB3.2
qC
qD1
qD2
qD3
qE1
qE2
qE3
qF1
qF2
50 bp
34,139,840 bp
34,139,850 bp
34,139,860 bp
34,139,870 bp
34,139,880 bp
session4-final.bam Coverage
session4-final.bam
Sequence
Refseq genes
Mtdh
novel_junction_peptides
novel_junction_peptides
Mus_musculus.GRCm38.74.gtf
ENSMUST00000022865
novel_junction_peptides
novel_junction_peptides
CYKTQTSPSQMLSK

4 Running Entire Proteogenomics workflow

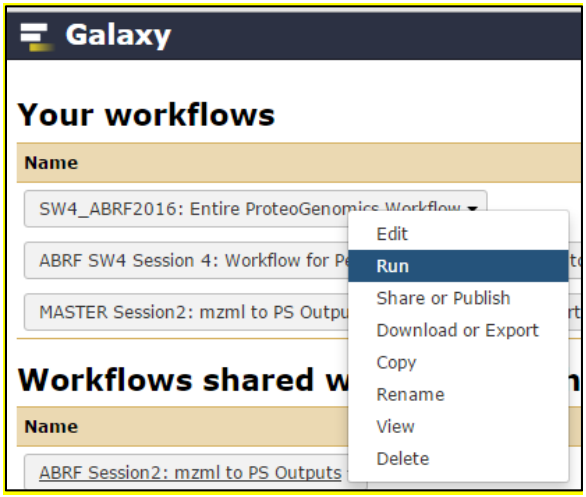
The proteogenomics workflow can be run as a single, complete workflow.

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

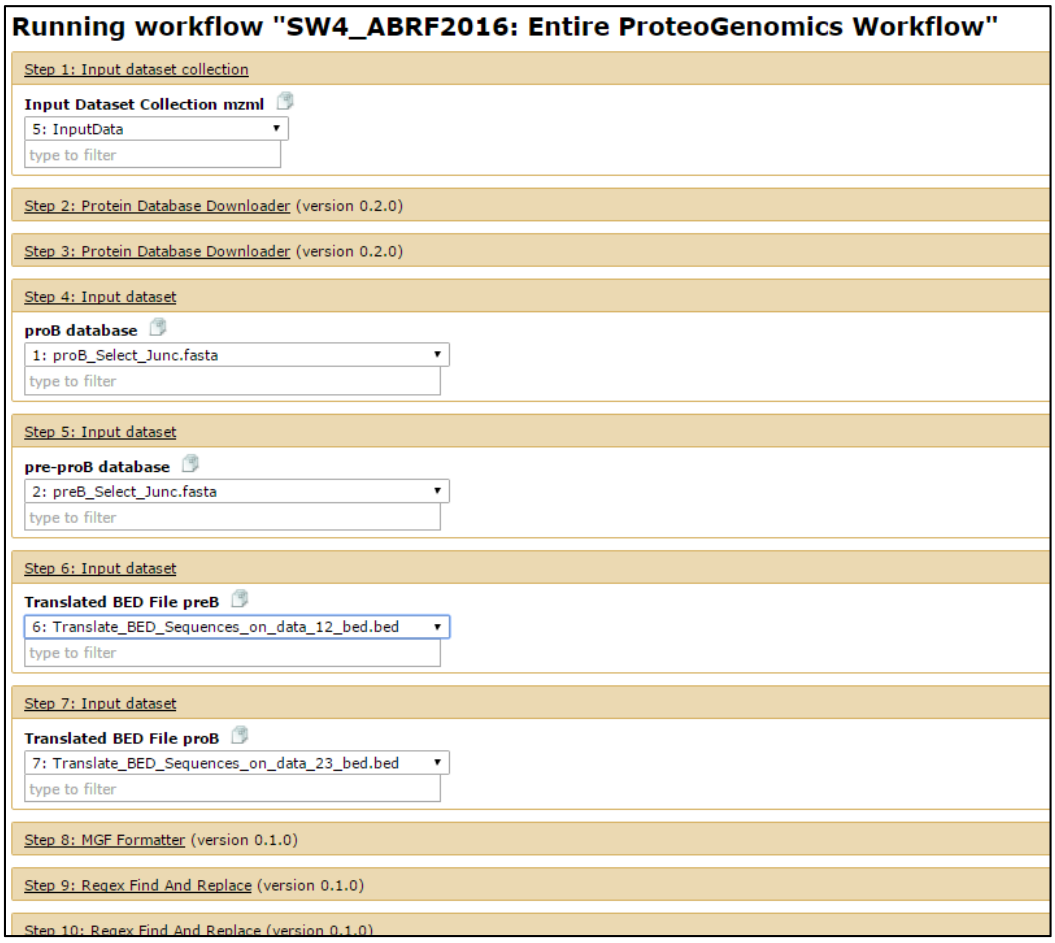




At the top of the screen select Workflows and click and Run for SW4_ABRF2016: Entire ProteoGenomics Workflow



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



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

5 Presenters and acknowledgements

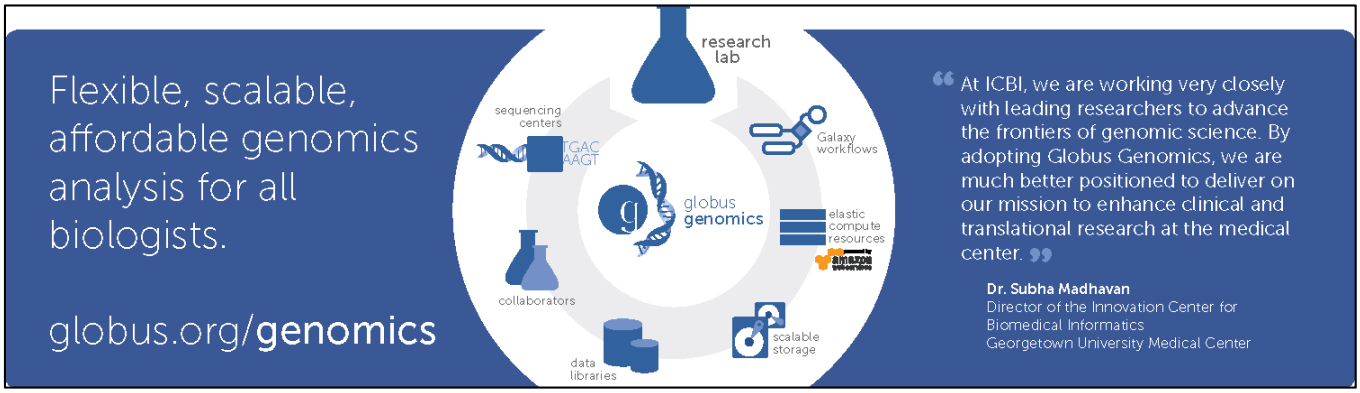
Presenters:

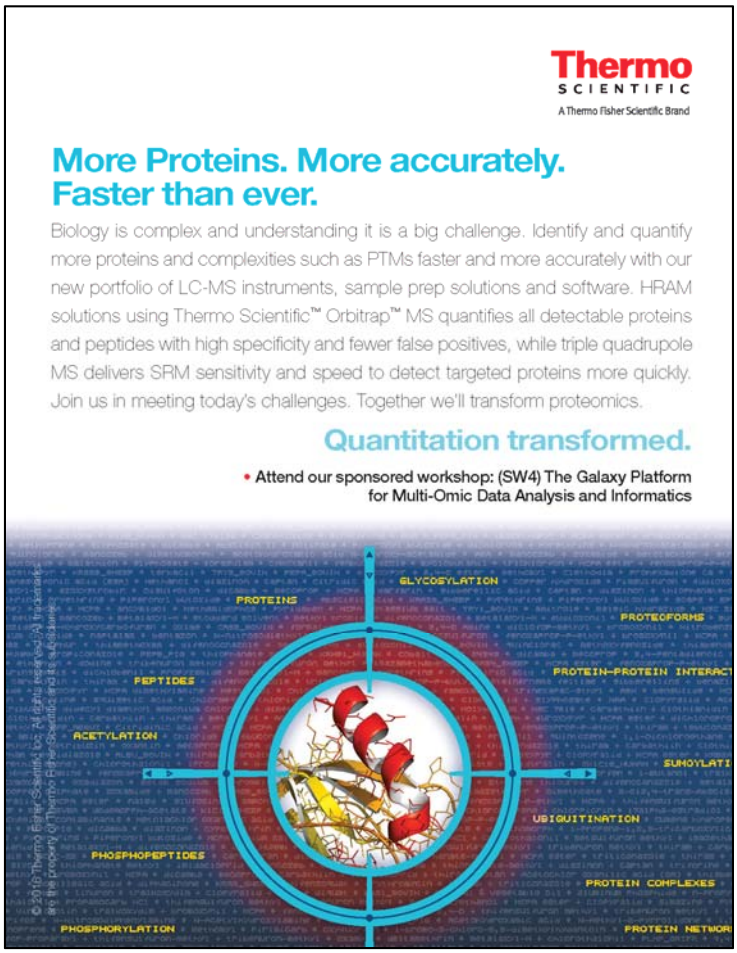
The main presenters were members of research team at the University of Minnesota, working an ongoing project developing Galaxy for multi-omic applications (National Science Foundation Grant 1147079). We have in-depth experience in Galaxy and its use for multi-omics data analysis. (z.umn.edu/galaxypreferences).

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.

Thanks to our KILO Sponsors





Also thanks to Amazon Web Services Education Research Grant.