Using Galaxy-P Documentation

Release 0.1

John Chilton, Pratik Jagtap

October 26, 2015

Contents

CHAPTER 1

Introduction

This guide covers intermediate and advanced usage of the Galaxy-P platform for proteomic and mass spec data analysis. For an gentle introduction to Galaxy, please check out the Galaxy 101 maintained at Penn State.

A proteomics version of this Galaxy 101 targeting Galaxy-P has been assembled by Pratik Jagtap and can be found here

CHAPTER 2

Galaxy-P 101 - Building Up and Using a Proteomics Workflow

This section is heavily inspired by the Galaxy 101 maintained at Penn State by the *core Galaxy team <http://wiki.galaxyproject.org/GalaxyTeam>*.

In this walkthrough we will introduce you to basics of Galaxy-P:

- Uploading database using Protein Database Downloader.
- Getting raw data from an external source.
- Peaklist processing including options and parameters.
- X! tandem search.
- PeptideProphet processing of X!tandem search.
- Converting PepXML to a table.
- Using ProteinProphet to process PeptideProphet results.
- FDR analysis.
- Performing simple data manipulation.
- Understanding Galaxy's history system.
- Creating and editing workflows.
- Applying workflows to your data.

2.1 What Are We Trying to Do?

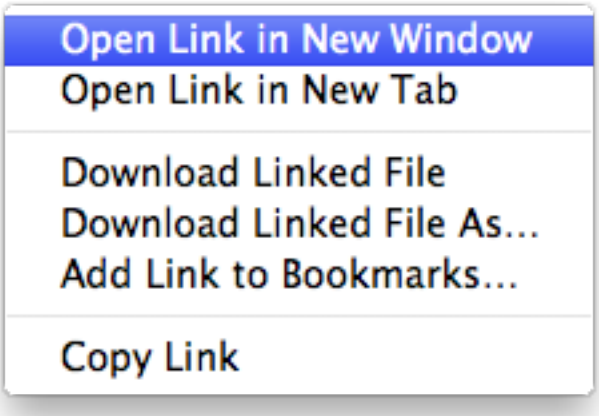
We are trying to set up a search of one RAW file (your experimental data acquired on an LTQ/Orbitrap instrument) against a database using a search algorithm named X!tandem. The idea is to estimate number of valid matches by using PeptideProphet or ProteinProphet and FDR analysis. Once we have created this workflow we can apply this workflow to another dataset and get results. So let's try it...

2.2 Setting Up Your Galaxy-P Account

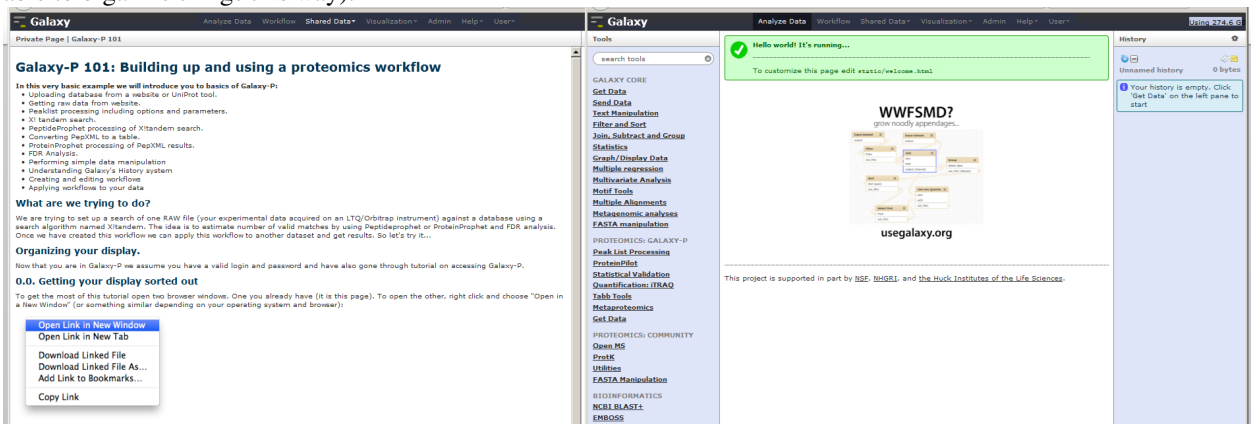
Now that you are in Galaxy-P, we assume you have a valid login and password and have knowledge about accessing Galaxy-P.

2.3 Organizing Your Windows

To get the most of this tutorial open two browser windows. One you already have (it is this page). To open the other, right click and choose "Open in a New Window" (or something similar depending on your operating system and browser):



Then organize your windows as something like this (depending on the size of your monitor you may or may not be able to organize things this way):



2.4 Loading a Search Database

There are a many options for how you can upload your search database (FASTA file with protein sequences). Three among these are:

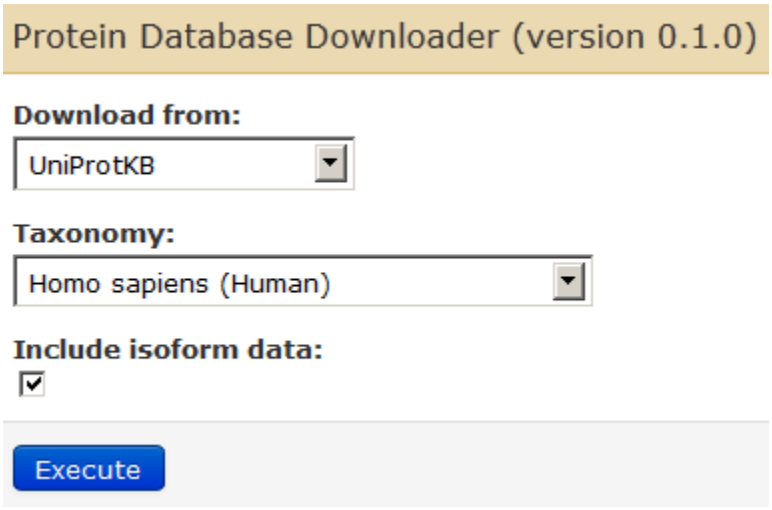
- Protein Database Downloader.
- Use website link for the database (see this short tutorial).
- Upload database from the data library.

In this tutorial, we will explore using Protein Database Downloader for database search.

2.5 Using Protein Database Downloader

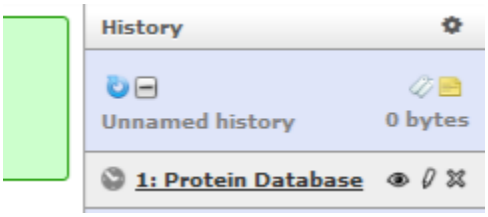
Now we get our search database. First thing we will do is to go to Tools –> Proteomics: Galaxy-P –> Get Data. Then click on Protein Database Downloader.

Since we are going to analyze a human dataset, select human database along with isoforms.

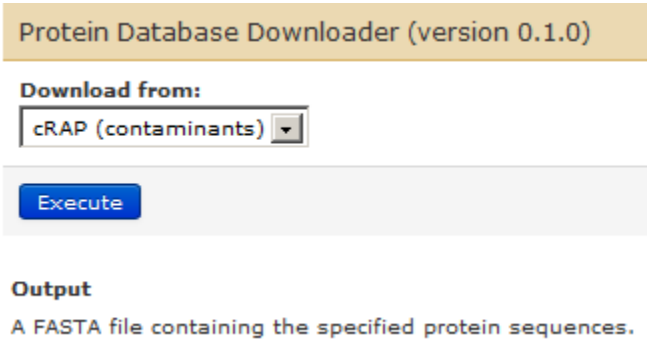


Clicking on "Execute" will download the human canonical isoform database.

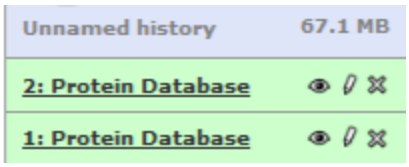
After this you will see your first History Item in Galaxy's right pane. It will go through gray (preparing) and yellow (running) states to become green:



For contaminants database, we will use the Protein database Downloader, select cRAP (contaminants) database and "Execute" much the same way you did for the human UniProt database.

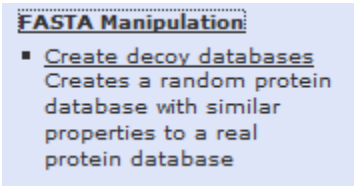


After this you will see your second History Item in Galaxy's right pane.

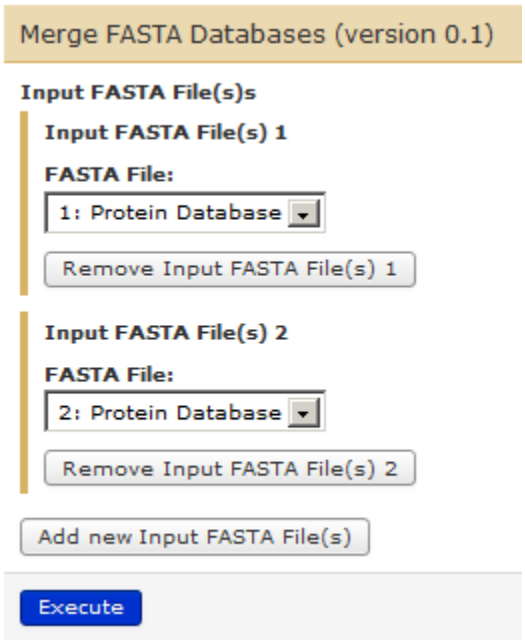


2.6 Using the Database Merge Tool

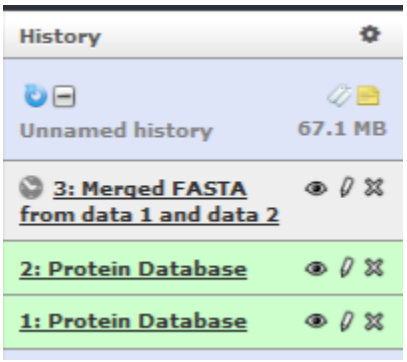
In order to merge these two FASTA files in your history, go to Tools –> FASTA Manipulation –> Merge FASTA Databases.



Click on "Add new Input FASTA File(s)" – twice – so that your set up for merging files looks like this:



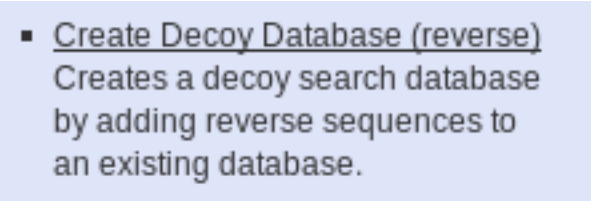
Click on Execute and you will be able to see your third History Item in Galaxy's right pane.



2.7 Creating a Decoy Database

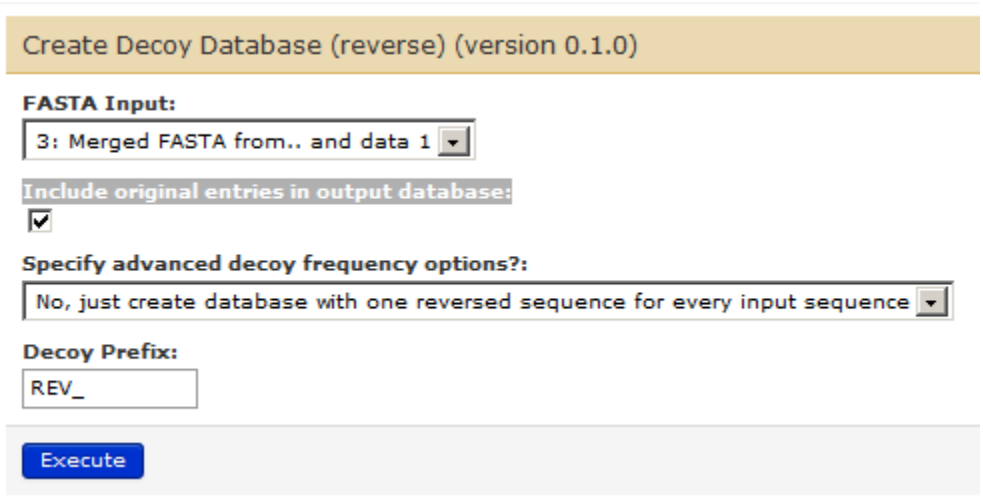
Our next step is to create a decoy database out of the merged file (3rd history item on the list).

For this, go to Tools -> FASTA Manipulation –> 'Create Decoy Database'.



Ensure that history item 3 shows up in the "FASTA Input:" box and that the box for "Include original entries in output database:" is checked. Change Decoy prefix to **REV_**.

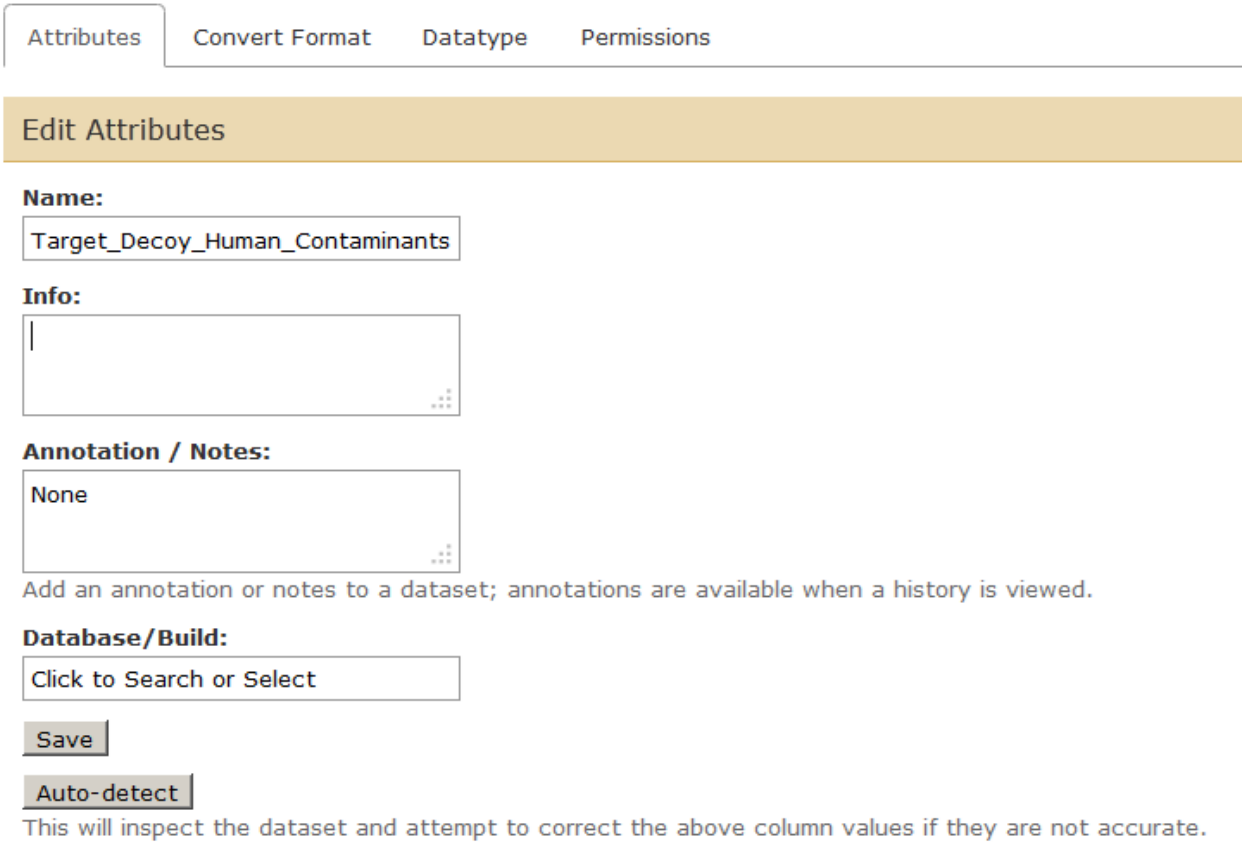
Your parameters for creating decoy database (reverse) tool should look like this.



Click on "Execute" to generate the fourth item in your history list. This is the FASTA database that we will be using for our search.

Now we will rename the history items to "Human UniProt", "cRAP", "Merged Human UniProt cRAP" and "Target_Decoy_Human_Contaminants" by clicking on the Pencil icon adjacent to each item. Also we will rename history to "Galaxy-P 101" (or whatever you want) by clicking on "Unnamed history" so everything looks like this:

Please feel free to explore tabs – Convert Format, Datatype, Permissions while you are editing the attributes. This is especially important while troubleshooting for steps that fail wherein a datatype has not been set properly or needs to be changed for subsequent steps.

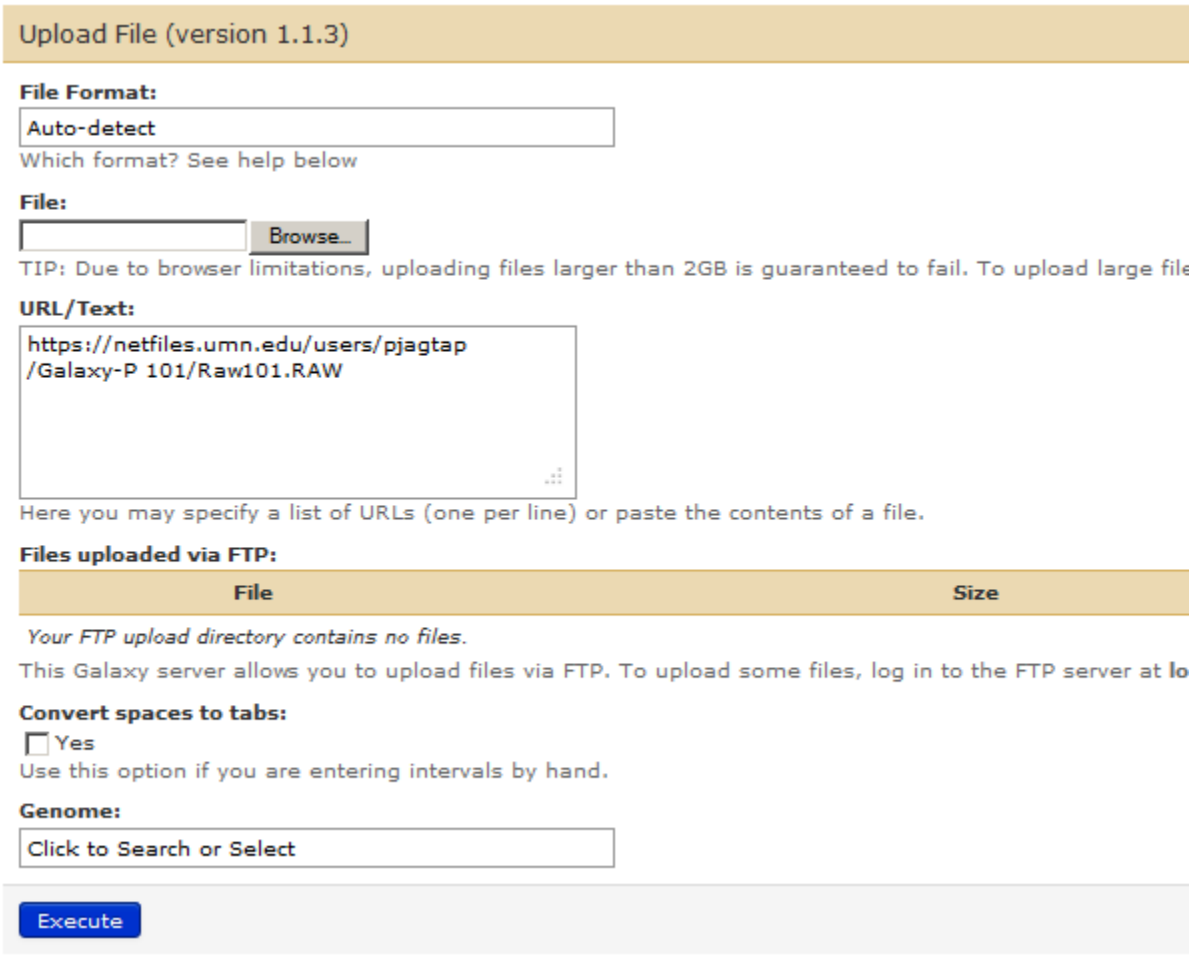


The next step is to upload RAW file.

There are a few options on how you can upload yours spectral data (RAW file acquired on a LTQ/Orbitrap) * Upload from your computer using your web browser. * Upload from your computer using the Galaxy-P FTP server. * Use website link for the RAW file (something we will use for this tutorial). * Import data from the data library.

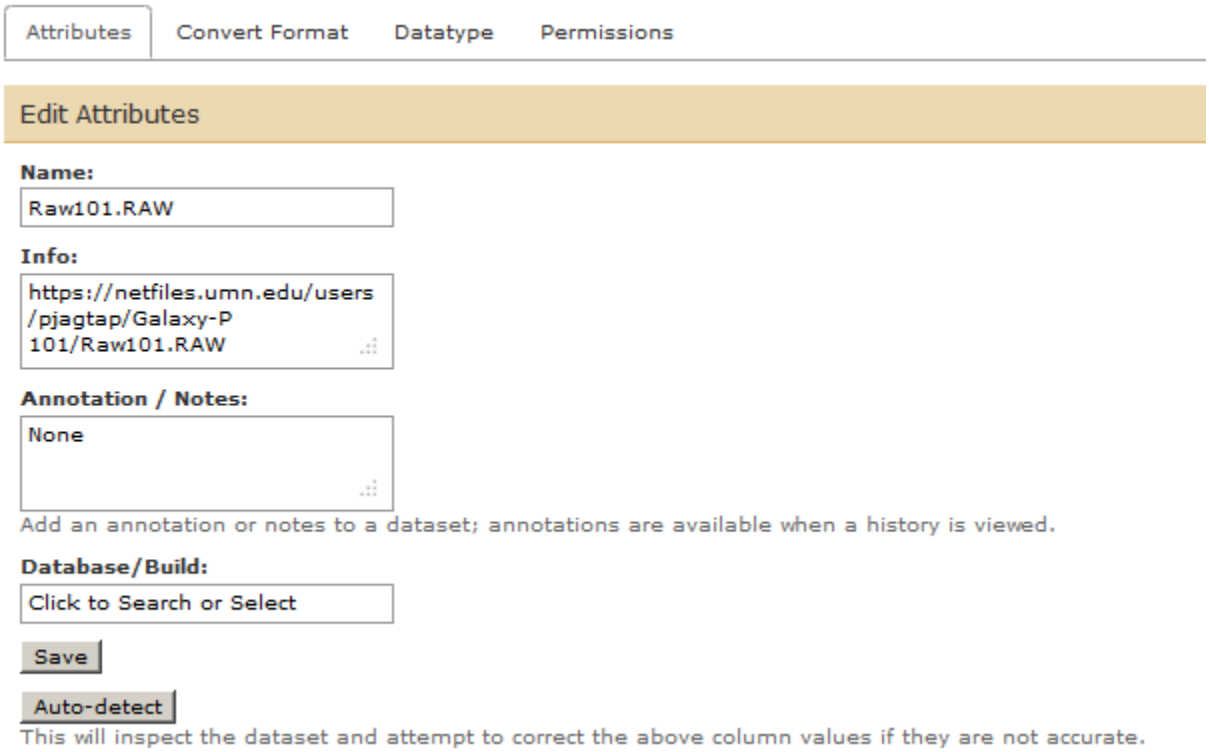
While there are many ways of uploading a RAW file, we will use a weblink to upload a fractionated human salivary RAW file.

In order to upload a RAW file, you can go to "Upload File" in Tools section and type in https://netfiles.umn.edu/users/pjagtap/Galaxy-P 101/Raw101.RAW and "Execute".



This will give us our 5th history item :

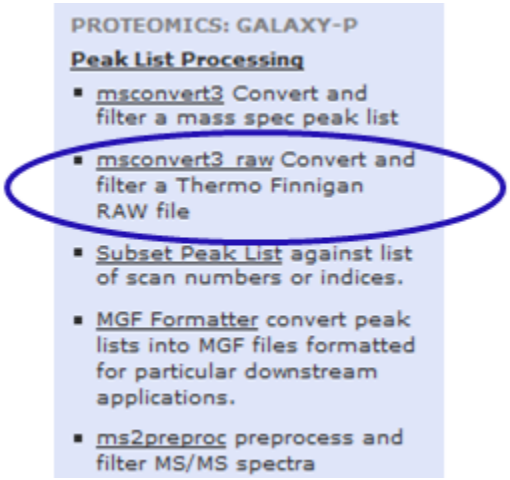
We can change the name of the RAW file to Raw101.RAW by using the pencil icon button.



2.8 Peaklist Processing

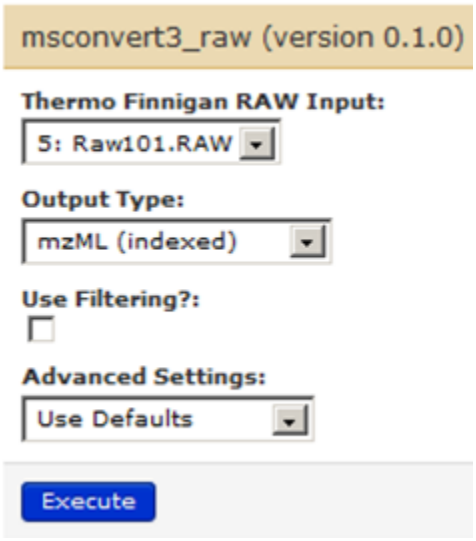
In the next step we will process the Raw file into a peaklist in mzml format so that it can be searched using X!tandem.

For this go to Tools –> Proteomics: Galaxy-P –> Peak List Processing –> msconvert3 raw



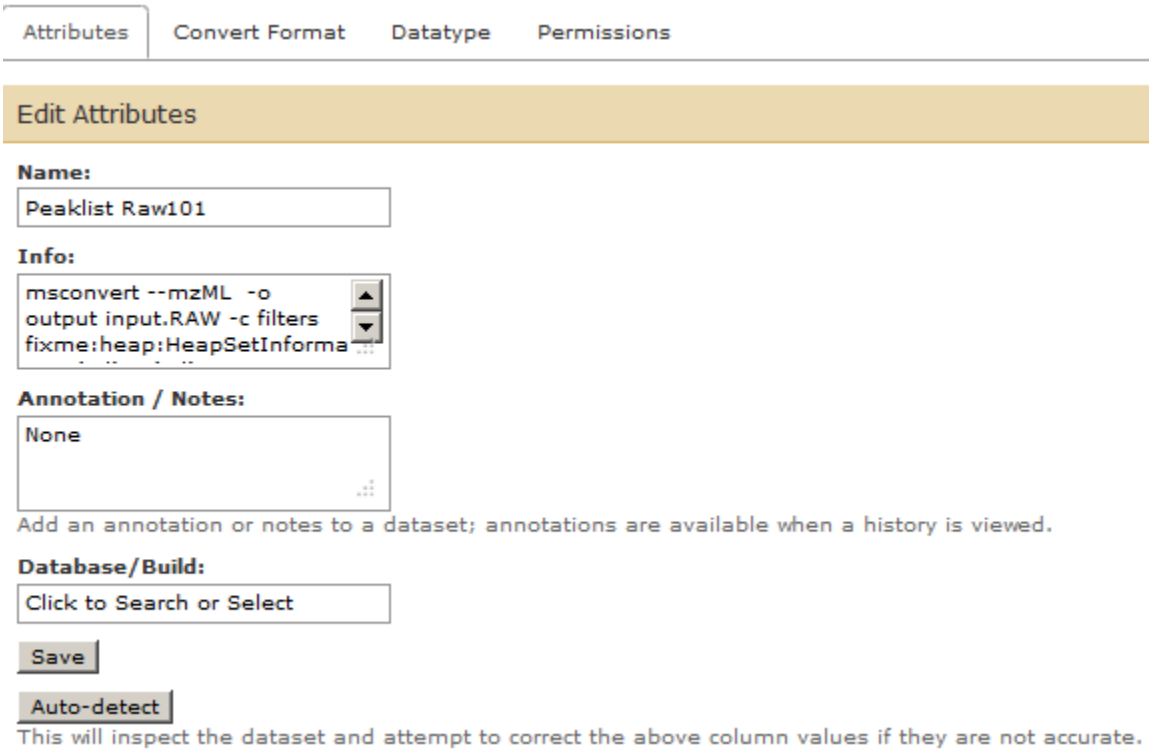
Explore the features by clicking on Use Filtering. However, for this tutorial we are going to use default settings for msconvert.

Click on “Execute” by ensuing that the settings appear as shown below:



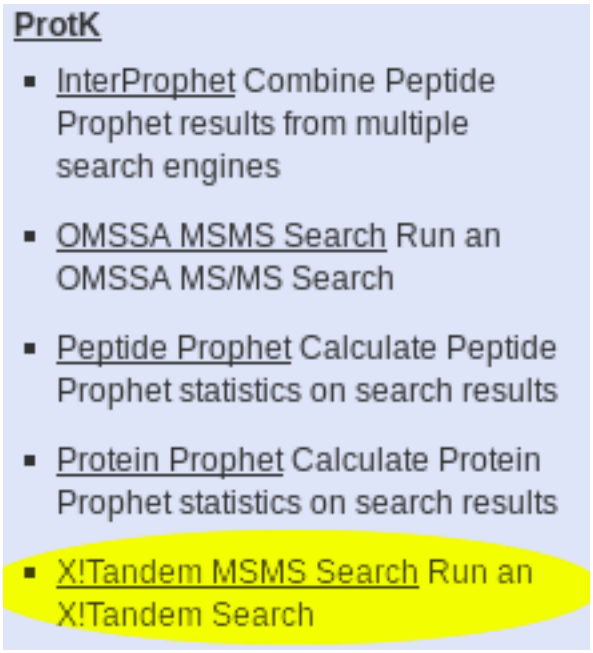
We have our sixth history item in the history list now.

Change the name of the mzml file to “Peaklist Raw101”. Note that this peaklist is in mzml format.



2.9 X! Tandem Search

For setting up a search of peaklist (Peaklist Raw101) against FASTA database (Target_Decoy_Human_Contaminants) using X!tandem, go to Tools–> ProtK (under PROTEOMICS: COMMUNITY) –> X!Tandem MSMS Search.



Set up search parameters for X!tandem as shown below. Ensure that you have the right FASTA file (4: Tar-

get_Decoy_Human_Contaminants), MSMS File (6: Peaklist Raw101) and click on Oxidation M for variable modifications. Ensure that you have 2 Missed Cleavages allowed, Trypsin as an enzyme, Fragment ion tolerance at 0.5 Da and Precursor ion tolerance at 10.0 ppm. The parameters file should look like below:

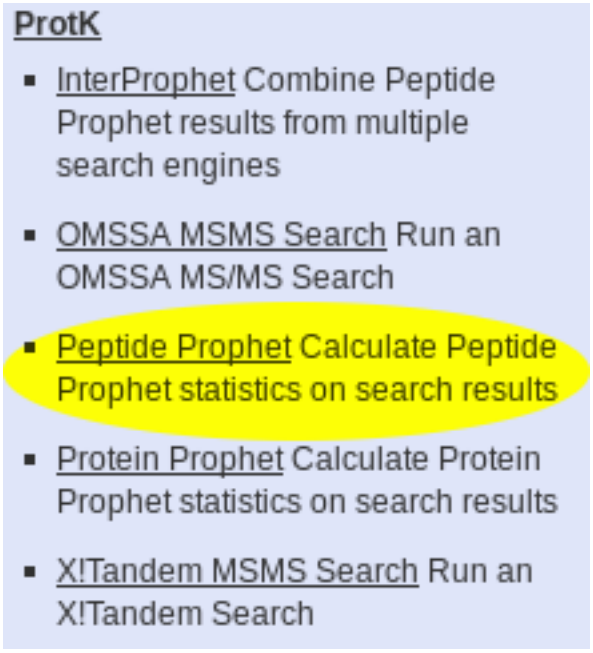

X!Tandem MSMS Search (version 1.0.0)
Database source:
Your Upload File
Uploaded FASTA file:
4: Target_Decoy_Huma..s on data 3
MSMS File:
6: Peaklist Raw101
An mzML file with MS/MS data
Variable Modifications:
Carbamidomethyl C
Glycocapture-N
Oxidation M
Hold the appropriate key while clicking to select multiple items
Custom Variable Modifications
You can specify a modification when present in a motif. For instance, 0.998@N!{P}[ST] is a deamida
Add new Custom Variable Modifications
Fixed Modifications:
Carbamidomethyl C
Glycocapture-N
Oxidation M
Hold the appropriate key while clicking to select multiple items
Custom Fixed Modifications
You can specify a modification when present in a motif. For instance, 0.998@N!{P}[ST] is a deamida
Add new Custom Fixed Modifications
Missed Cleavages Allowed:
2
Allow peptides to contain up to this many missed enzyme cleavage sites
Enzyme:
Trypsin
Fragment ion tolerance:
0.5
Fragment Ion Tolerance in Daltons
Precursor ion tolerance:
10.0
Precursor Ion Tolerance (Da or ppm)
Precursor Ion Tolerance Units:
ppm
Enable semi-cleavage rules:
Match peptides where one end or the other does not conform to specified enzyme rule.
Allow multi-isotope search:
This allows peptide candidates in windows around -1 Da and -2 Da from the acquired mass to be cor
is not corrected to the monoisotopic mass.
Use K-Score Alogirthm:
The k-score alogirthm is a Tandem score plug-in implementing the COMET score function as describ
Thresholds:
ISB k-Score defaults
Set various X! Tandem threshold values.

Your history list should look like this now.

2.10 Peptide Prophet Processing of X! Tandem Search

In this step we will process PepXML results from X! Tandem search to provide peptide probability scores for further analysis.

For this, go to Tools–> ProtK (under PROTEOMICS: COMMUNITY) –> 'Peptide Prophet'.

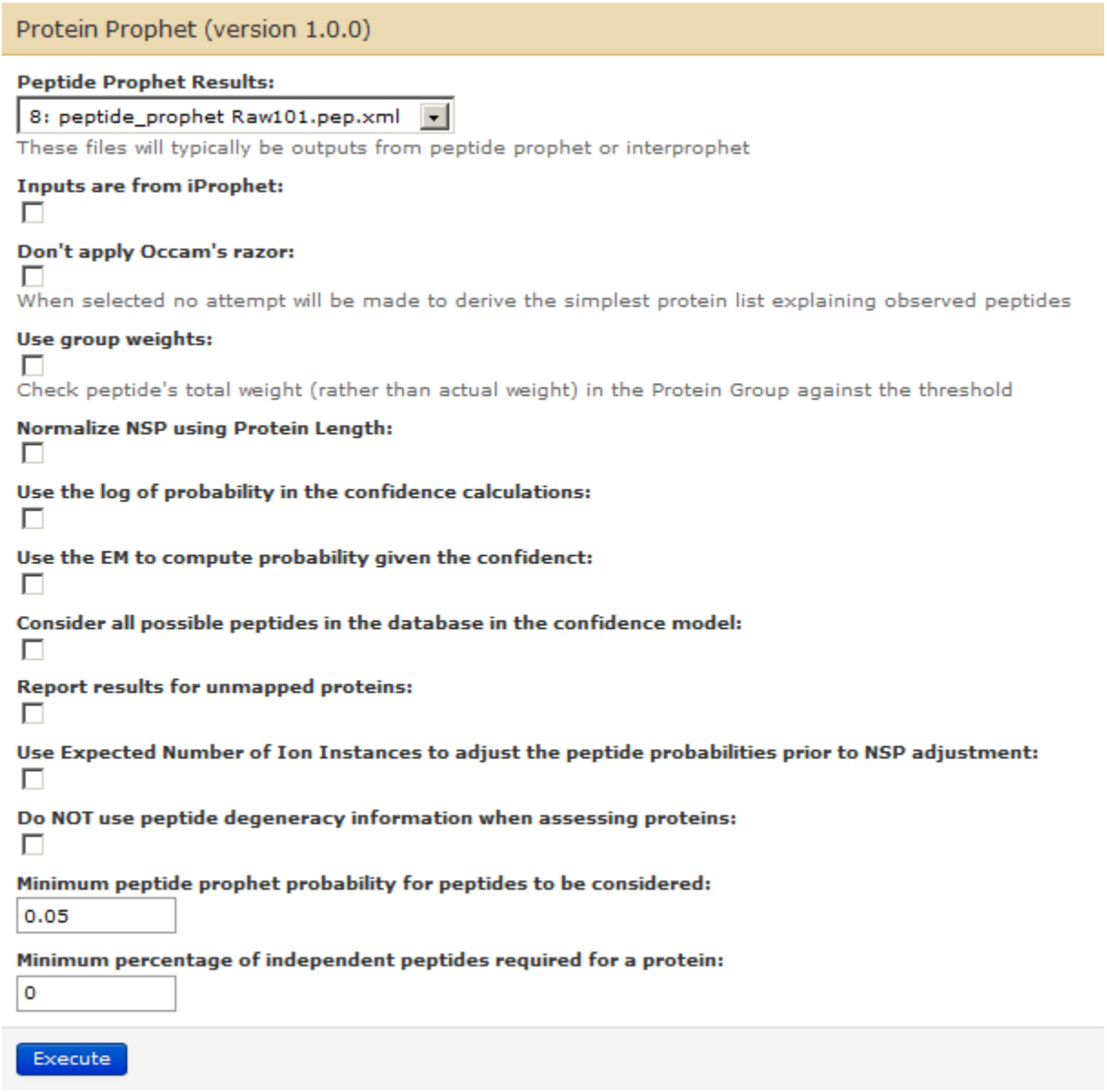


The Peptide Prophet parameters should be specified as follows:

Change the name of the Peptide Prophet file to “peptide_prophet Raw101.pep.xml” by using the pencil icon.

In this step, we will use ProteinProphet to process PeptideProphet results from X!tandem search to provide protein probability scores for further analysis.

For this, go to Tools –> ProtK (under PROTEOMICS: COMMUNITY)–> Protein Prophet.

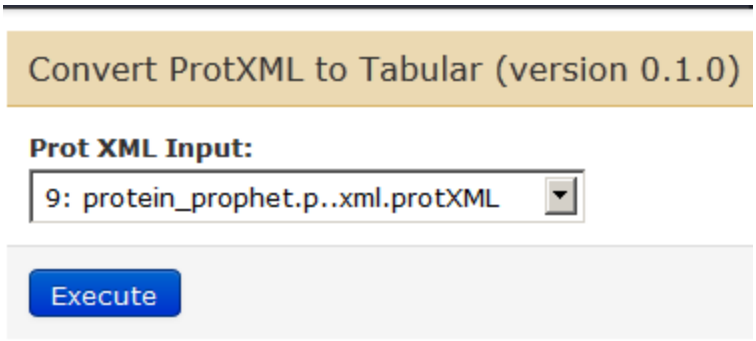


This will generate 9th history item in the list.

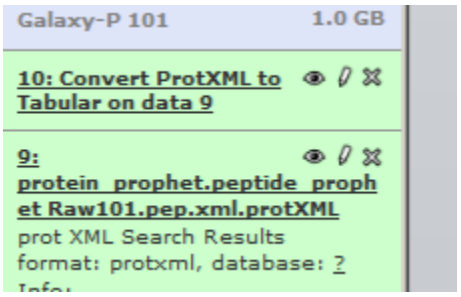
2.11 Converting ProtXML to a Table

In this step we will convert Protein Prophet results to a tabular format so that they can be viewed or processed for further analysis.

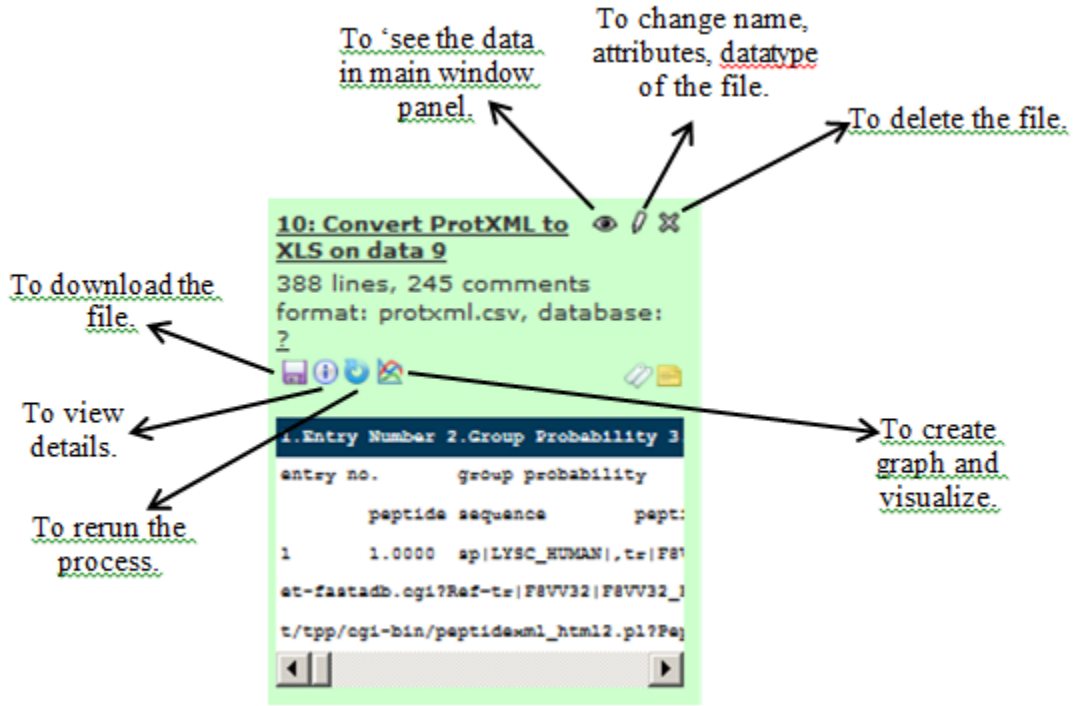
For this, go to Tools –> Utilities (under PROTEOMICS: COMMUNITY) –> Convert ProtXML to Tabular. Ensure that the latest file in the history (protXML file – 9th in the list).



Clicking on Execute will give us our 10th file in our history.



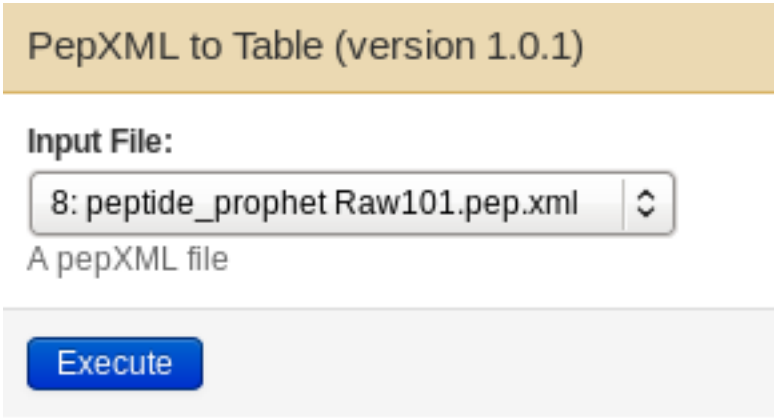
Explore the icons in the generate the file to see ('eye') the data OR to download ('floppy disk'), view details ('i" sign) OR rerun the analysis with changed parameters (blue 'rerun' icon). You can also create graphs and visualize using the 'graph' icon.



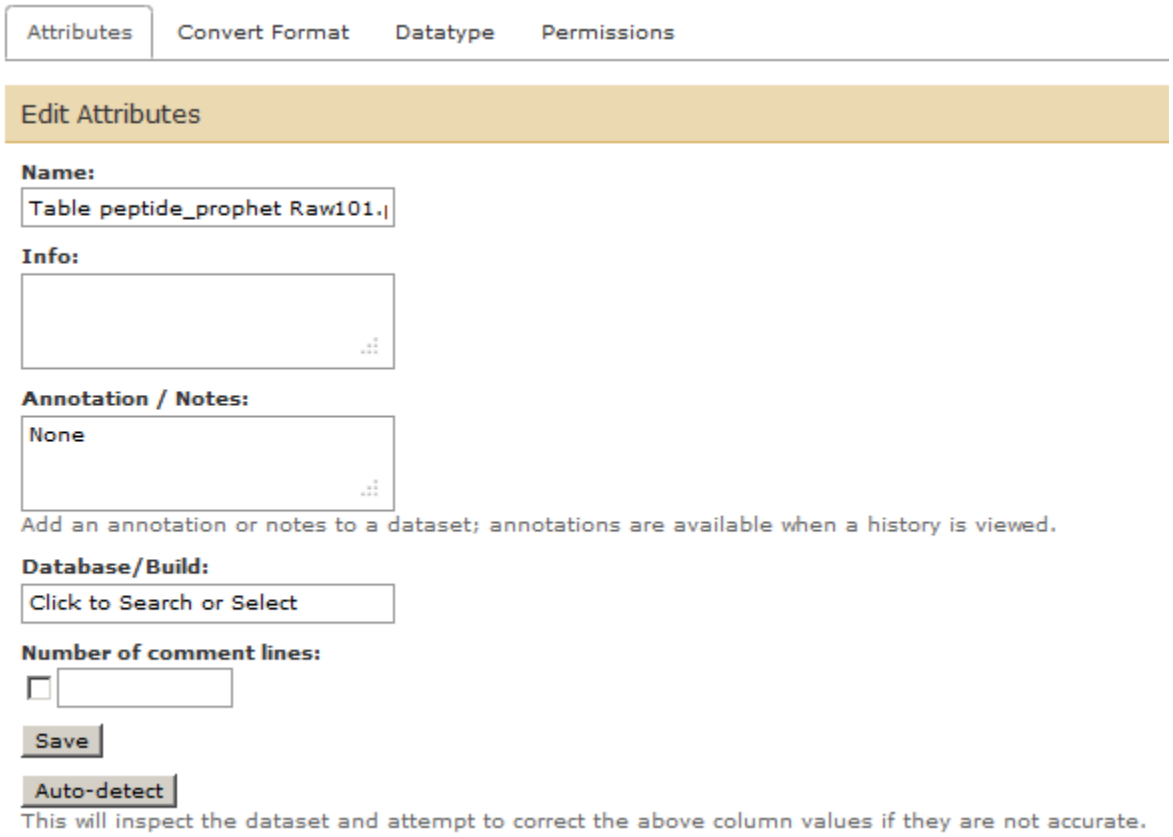
2.12 False Discovery Rate Analysis

In order to calculate FDR at the peptide level, we will first convert PeptideProphet file to a tabular format.

For this, go to Tools –> ProtK (under PROTEOMICS: COMMUNITY) –> Convert PepXML to Table. Ensure that the peptide prophet file in the history (peptide prophet file – 8th in the list) is highlighted.

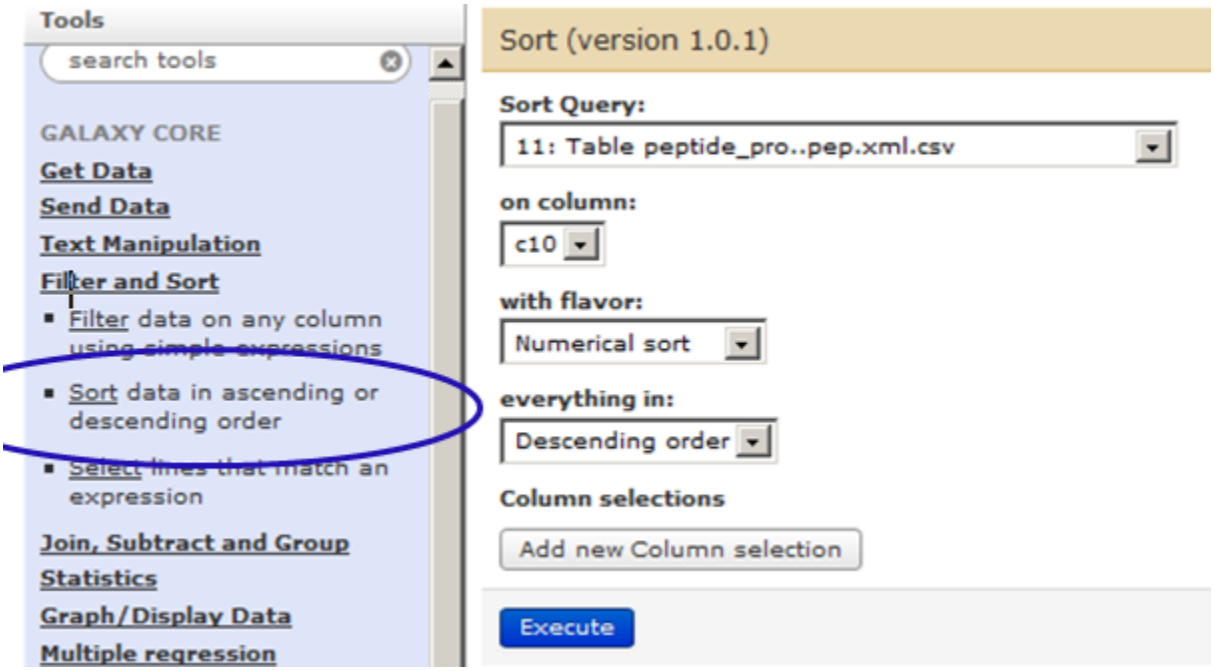


Change the name of this 11th file in history list to add "Table. . . " as shown in the image below.



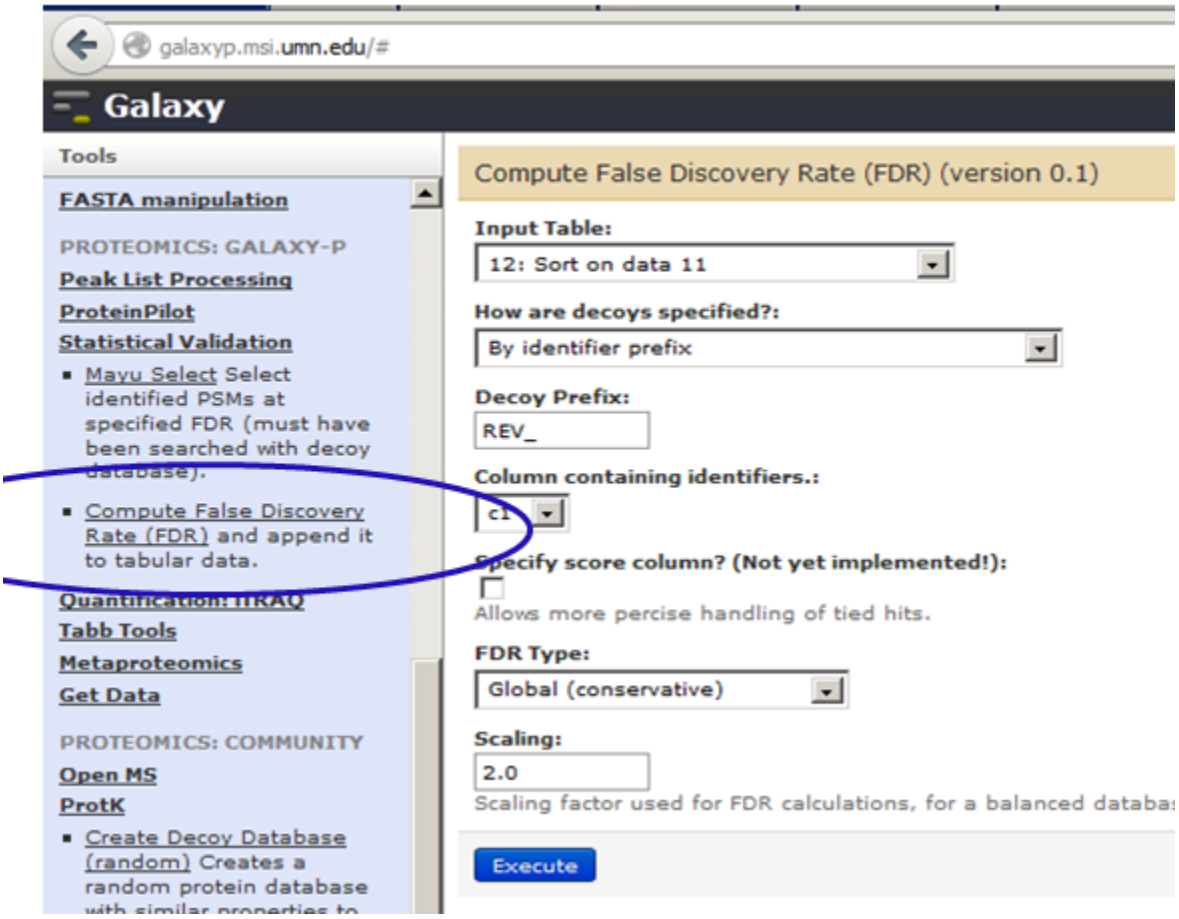
Next step is to sort the table with descending peptide probability scores. For this go to Tools –> Filter and Sort–> Sort. Ensure that table (csv file – 11th in the list) is highlighted.

Also select c10 (10th column which has peptide probability values) for sorting in descending order.



Click on 'Execute'. This will create you 12th list in history.

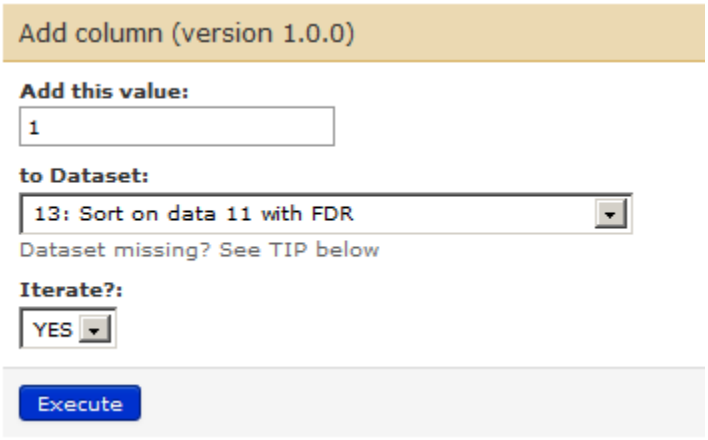
To compute FDR on this file, go to Tools –> Statistical validation (under PROTEOMICS: GALAXY-P) –> Compute False discovery Rate (FDR). Ensure that the sorted table (tsv file – 12th in the list) is highlighted. Also select c1 (1st column which has identifiers) for sorting in descending order. The parameters for this processing should look as below:



This will give you an output with a column with false discovery rate. You can use the 'eye' icon to visualize your data.

You can use the text manipulation tools such as 'Add column" and "cut" in order to visualize your data better.

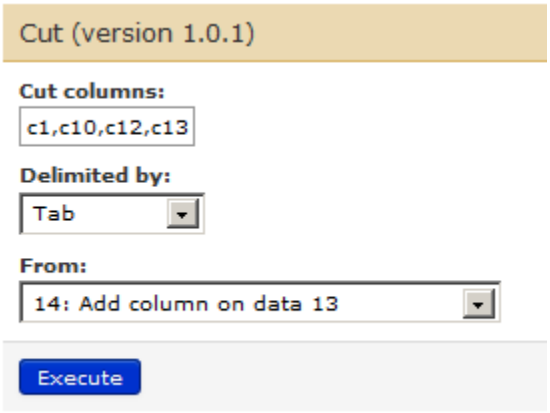
Go to Tools –> Text manipulation –> Add column. Ensure that the correct file (number 13 in the list) is chosen as a dataset. Change Iterate column to yes and click on execute.



This will give you your 14th history item. Use the eye icon to confirm that a column has been added at the end of the file.

Next, we will cut columns 1, 10 12 and 13 from history dataset number 14. For this go to Tools –> Text manipulation –> Cut. Ensure that the correct file (number 14 in the list) is chosen as a dataset. Type in c1,c10,c12,c13 in 'cut

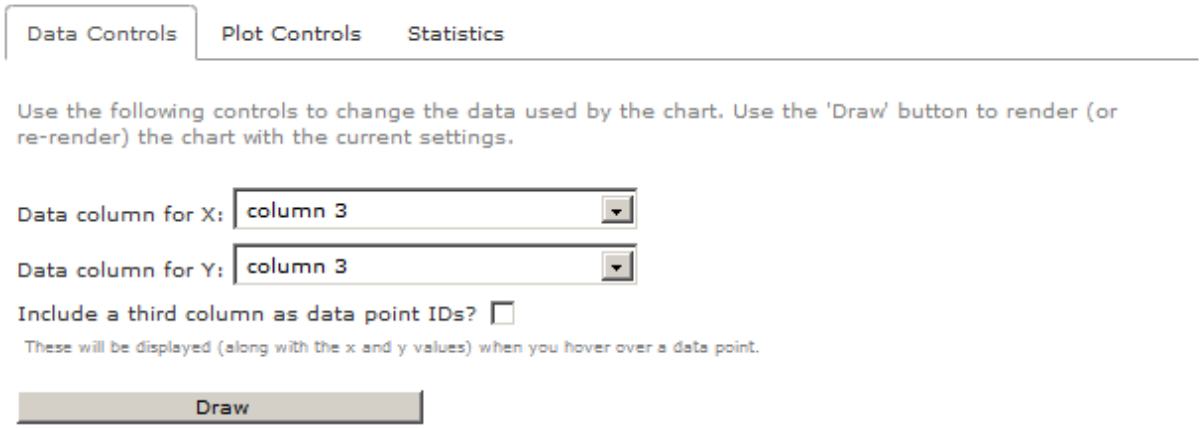
columns' box and press execute.



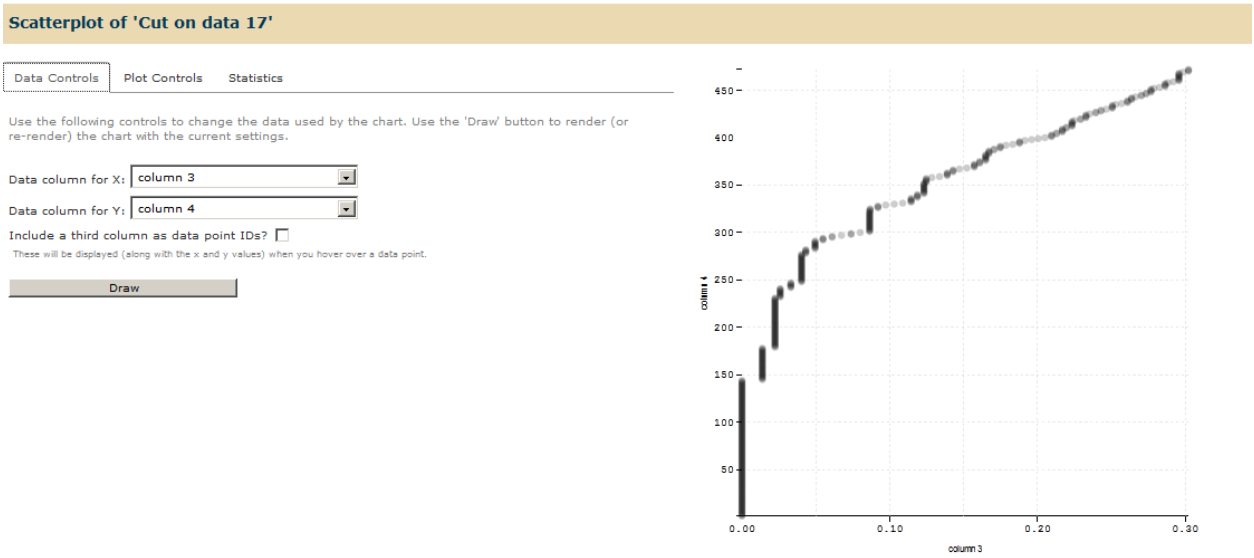
This will give you 15th history dataset.


History
Galaxy-P 101
1.0 GB
15: Cut on data 14
14: Add column on data 13
13: Sort on data 11 with FDR
12: Sort on data 11
11: Table peptide_prophet Raw101.pep.xml.csv
10: Convert ProtXML to Tabular on data 9
9: protein_prophet.peptide_prophet. X!tandem vs Target Decoy Human Contaminants.Peaklist Raw010.Peaklist Raw101.pepXML.pep.xml.protXML
8: peptide_prophet Raw101.pep.xml
7: X!tandem vs Target Decoy Human Contaminants.Peaklist Raw010.Peaklist Raw101.pepXML
6: Peaklist Raw101
5: Raw101.RAW
4: Target_Decoy_Human_Contaminants on data 3
3: Merged Human UniProt cRAP
2: CRAP
1: Human UniProt

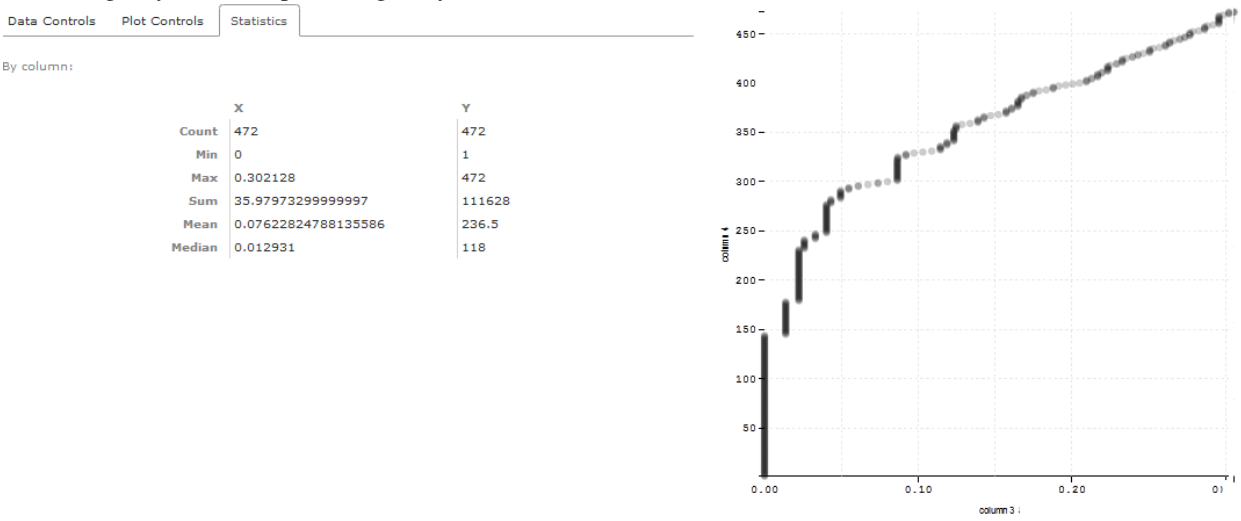
If you click on the Visualize / scatterplot icon you will see a scatterplot controls in the central window.



Keep the Data column X: column 3 and change Data column Y: column 4. Click on Draw.

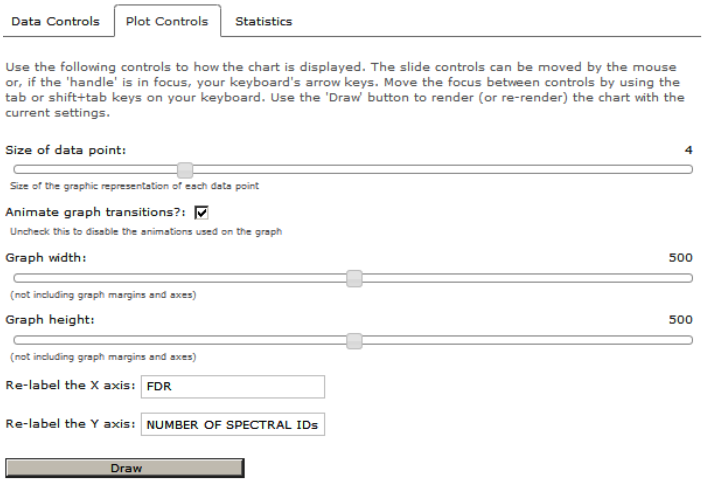


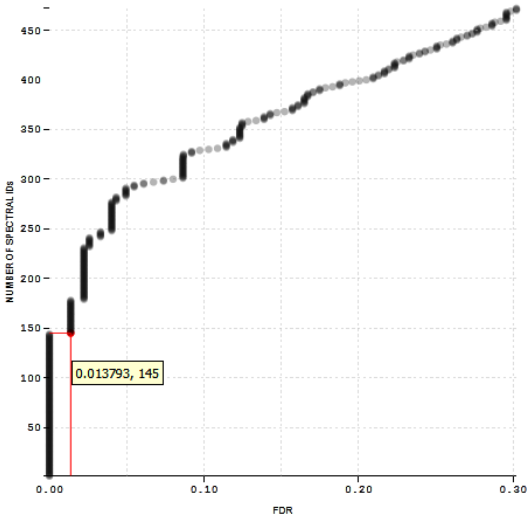
This will give you a ROC plot that gives you statistics and number of identifications at various levels of FDR.



Go to “Plot Controls” Tab and change the controls on your chart. Relabel X axis as “FDR” and Y-axis as ‘NUMBER OF SPECTRAL IDs”. Click on Draw to render the chart on current settings.

If you scroll on the graph, you will be able to see that 144 spectra were identified in this dataset before the first reverse match was encountered (FDR of 1.3%). Thus, 144 spectra were identified below 1% global FDR.





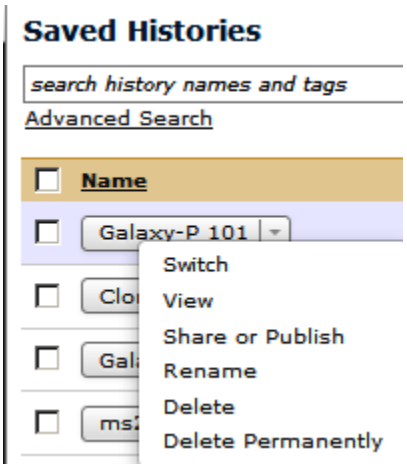
2.13 Understanding Galaxy Histories

In Galaxy, your analyses live in histories such as this one:


History
Galaxy-P 101
1.0 GB
15: Cut on data 14
14: Add column on data 13
13: Sort on data 11 with FDR
12: Sort on data 11
11: Table peptide_prophet Raw101.pep.xml.csv
10: Convert ProtXML to Tabular on data 9
9: protein_prophet.peptide_prophet. X!tandem vs Target Decoy Human Contaminants.Peaklist Raw010.Peaklist Raw101.pepXML.pep.xml.protXML
8: peptide_prophet Raw101.pep.xml
7: X!tandem vs Target Decoy Human Contaminants.Peaklist Raw010.Peaklist Raw101.pepXML
6: Peaklist Raw101
5: Raw101.RAW
4: Target_Decoy_Human_Contaminants on data 3
3: Merged Human UniProt cRAP
2: CRAP

Histories can be very large, you can have as many histories as you want, and all history behavior is controlled by the Options button on the top of the History pane:

Many of the options here are self explanatory. If you create a new history, your current history does not disappear. If you would like to list all of your histories just choose Saved Histories and you will see a list of all your histories in the central window pane:



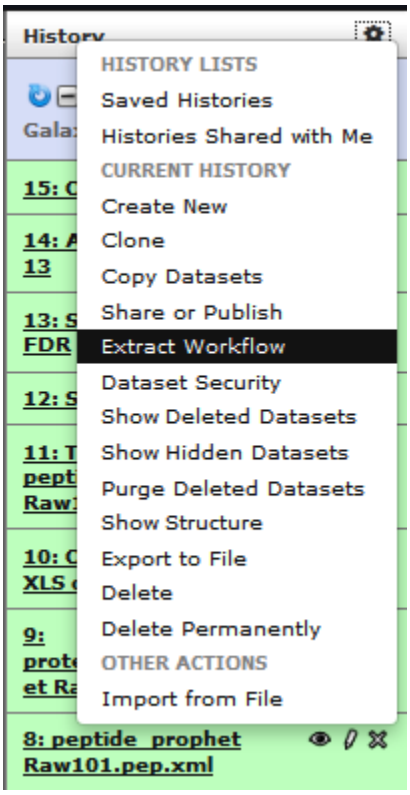
2.14 Converting Histories to Workflows

One of the history options listed is very special. It allows you to easily convert existing histories into analysis workflows. Why would you want to create a workflows out of a history? To redo the analysis again with minimal clicking.

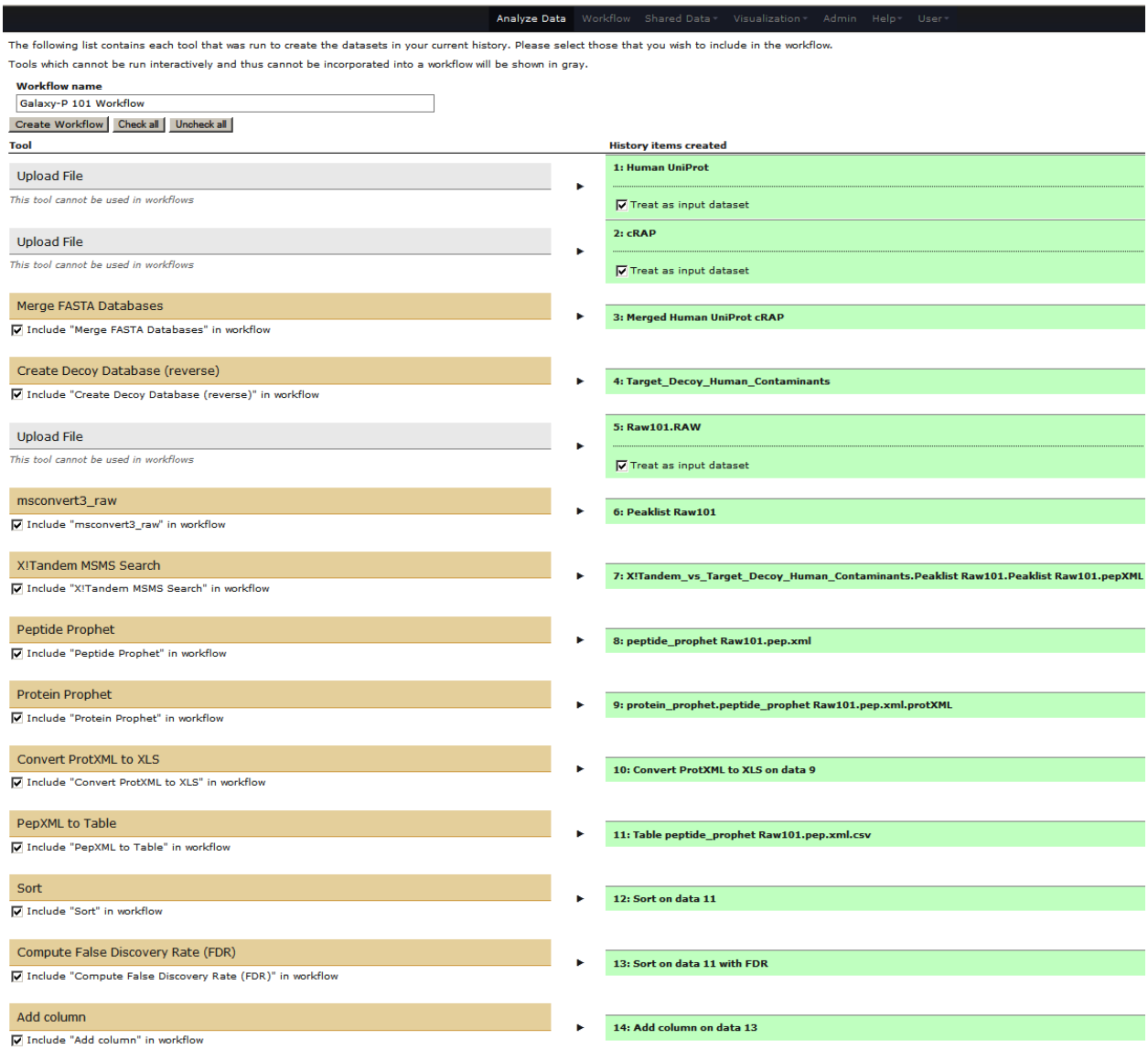
Lets take a look at the history again:


History
Galaxy-P 101
1.0 GB
15: Cut on data 14
14: Add column on data 13
13: Sort on data 11 with FDR
12: Sort on data 11
11: Table peptide_prophet Raw101.pep.xml.csv
10: Convert ProtXML to Tabular on data 9
9: protein_prophet.peptide_prophet. X!tandem vs Target Decoy Human Contaminants.Peaklist Raw010.Peaklist Raw101.pepXML.pep.xml.protXML
8: peptide_prophet Raw101.pep.xml
7: X!tandem vs Target Decoy Human Contaminants.Peaklist Raw010.Peaklist Raw101.pepXML
6: Peaklist Raw101
5: Raw101.RAW
4: Target_Decoy_Human_Contaminants on data 3
3: Merged Human UniProt cRAP
2: CRAP
1: Human UniProt

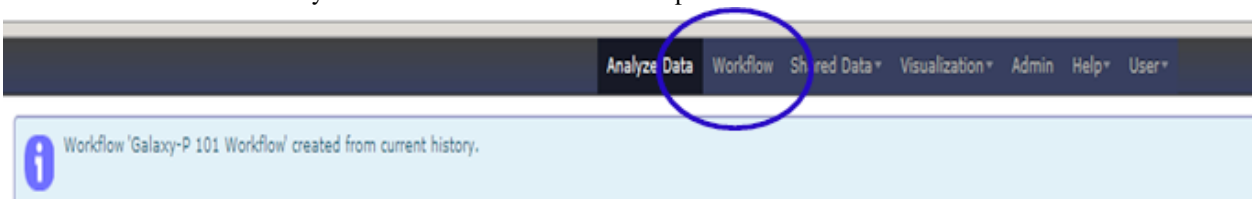
You can see that this history contains all steps of our analysis. So by building this history we have actually built a complete record of our analysis with Galaxy preserving all parameter settings applied at every step. Wouldn't it be nice to just convert this history into a workflow that we'll be able to execute again and again? This can be done by clicking on Options button and selecting Extract Workflow option:



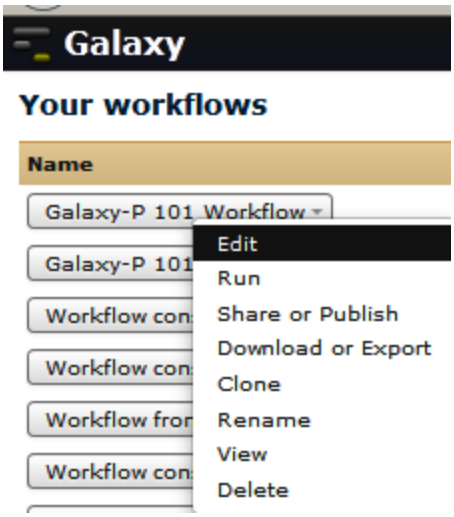
The center pane will change as shown below and you will be able to choose which steps to include/exclude and how to name the newly created workflow. In this case I named it "Galaxy-P 101 Workflow":



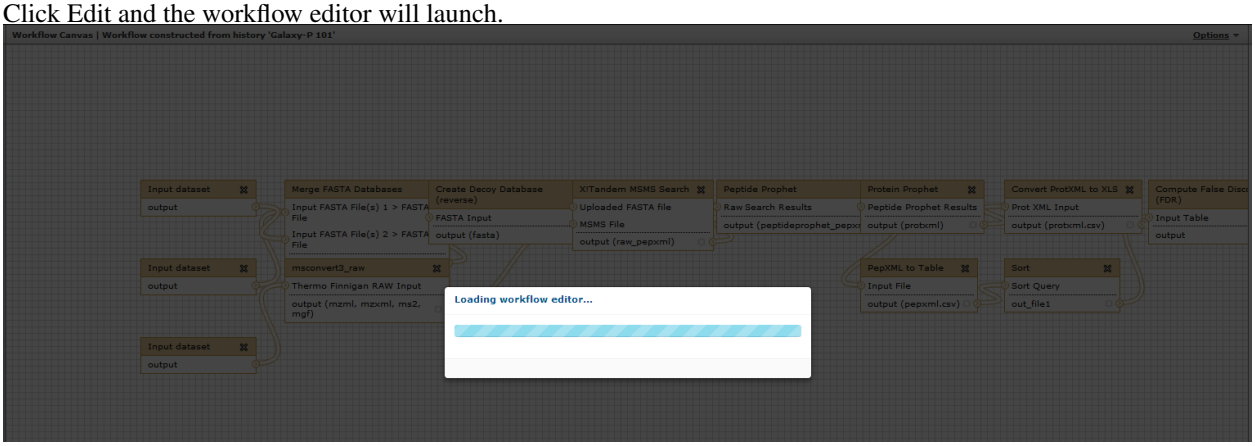
Once you click Create Workflow you will get the following message: "Workflow 'Galaxy-P 101 Workflow' created from current history". But where did it go? Click on Workflow link at the top of Galaxy interface and you will a list of all workflows with "Galaxy-P 101 Workflow" listed at the top:



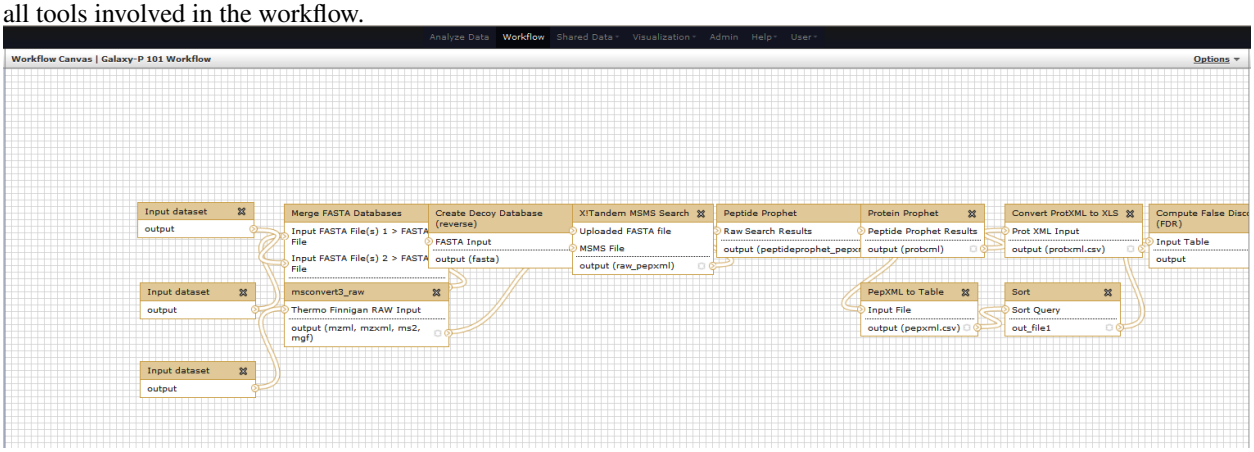
If you click on a triangle adjacent to the workflow's name you will see the following dialogue:



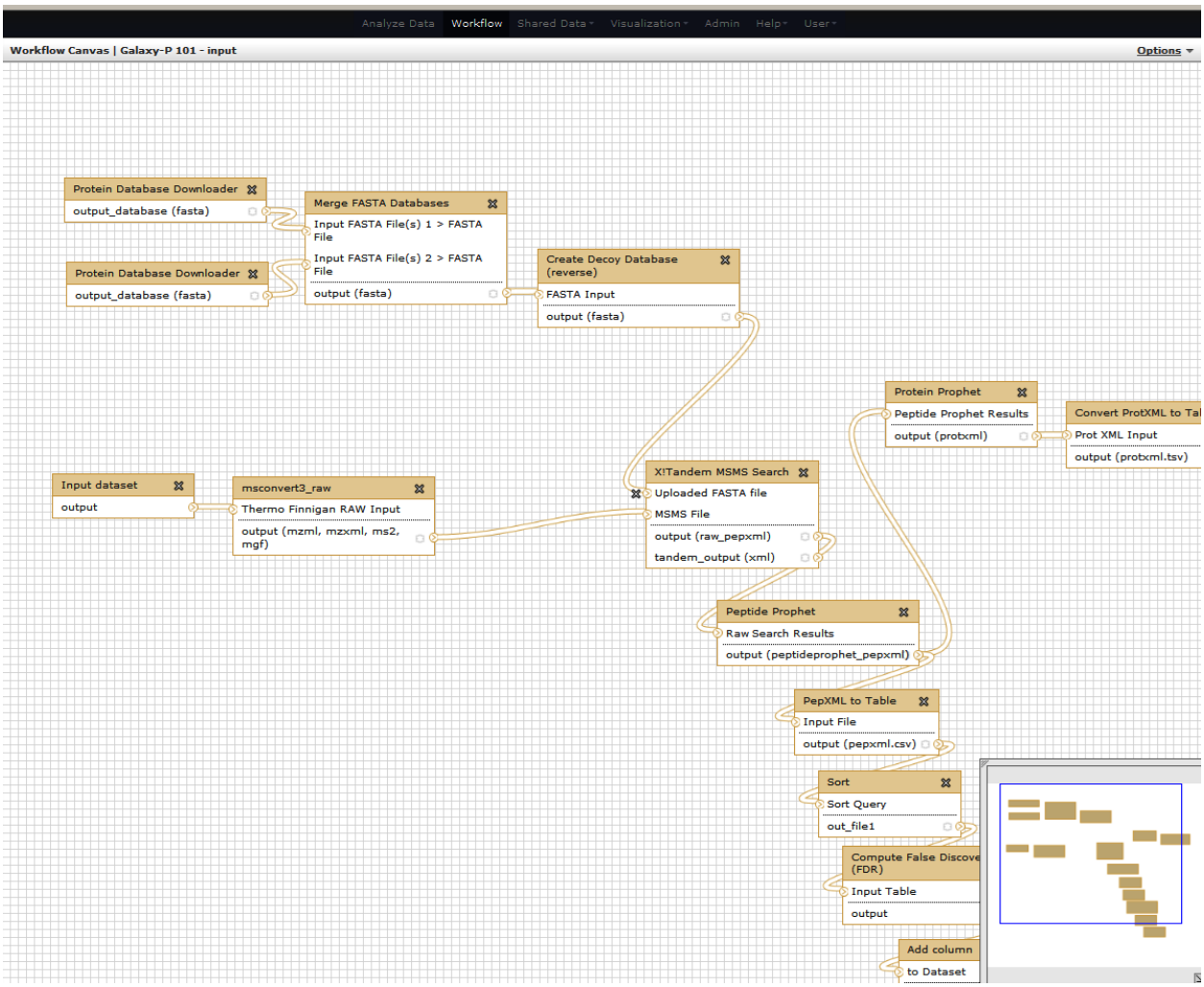
Click Edit and the workflow editor will launch.



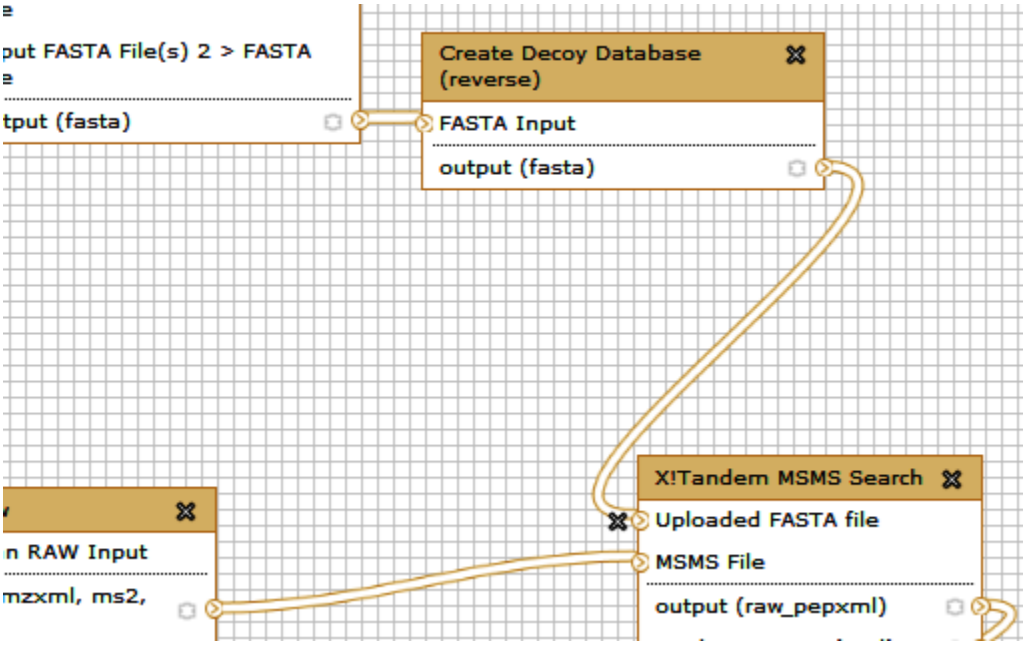
It will allow you to examine and change settings of this workflow as shown below. Note that you can click on each box so that you can see parameters of this tool on the right pane. This is how you can view and change parameters of all tools involved in the workflow.



You can also reorganize your workflow so that it makes intuitive sense. For example, I have compartmentalized this workflow as follows. However, as in any form of art – this is not the only way of representing your workflow. Re-architecture as the best way you wish to.

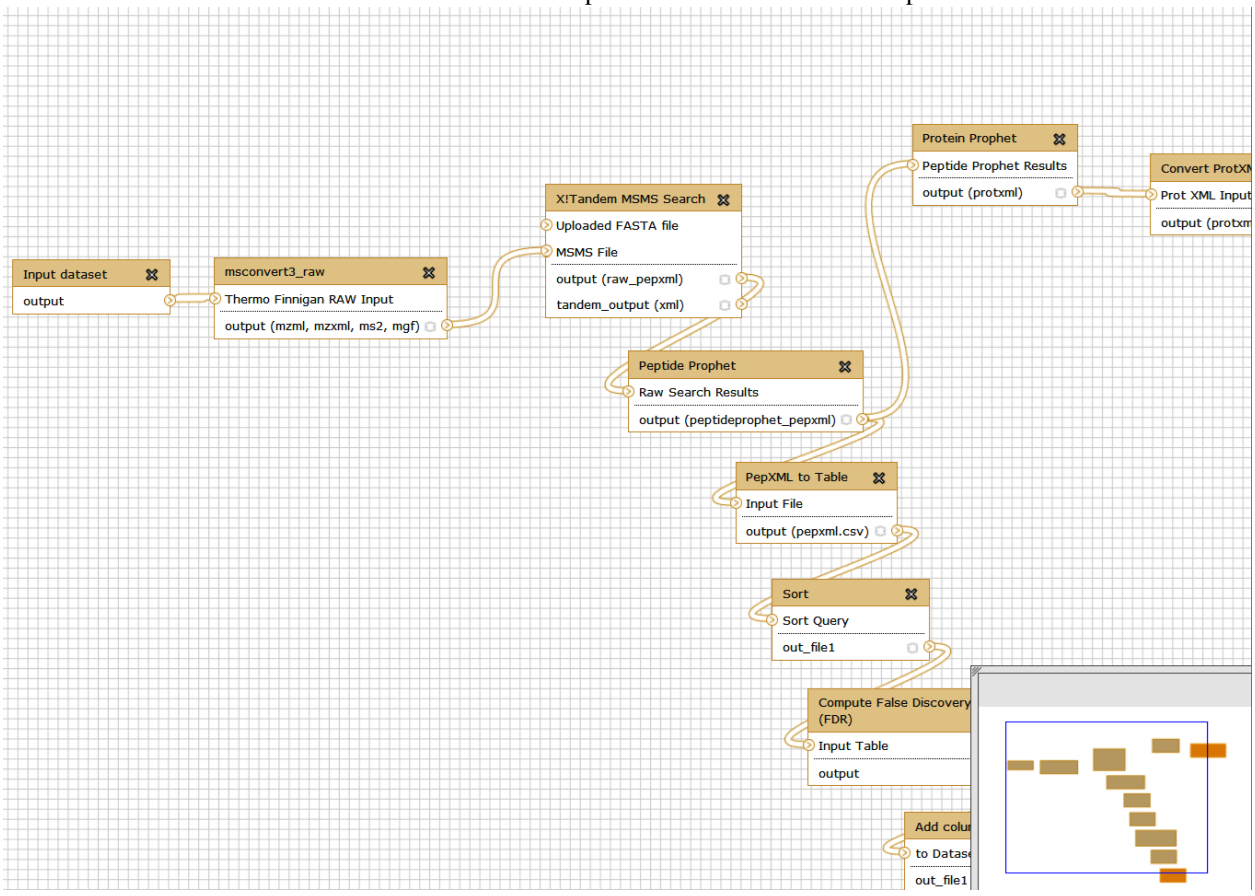


For this tutorial, we are going to reformat the workflow by replacing the earlier part of database creation with an input database. For this we excise the part where uploaded FASTA file merges with the X! Tandem search step.



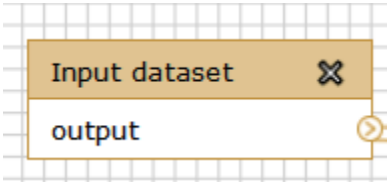
You can delete all the steps involving FASTA database creation. Delete one step at a time...

Your workflow now will look like this. Note that the Uploaded FASTA file has not input.

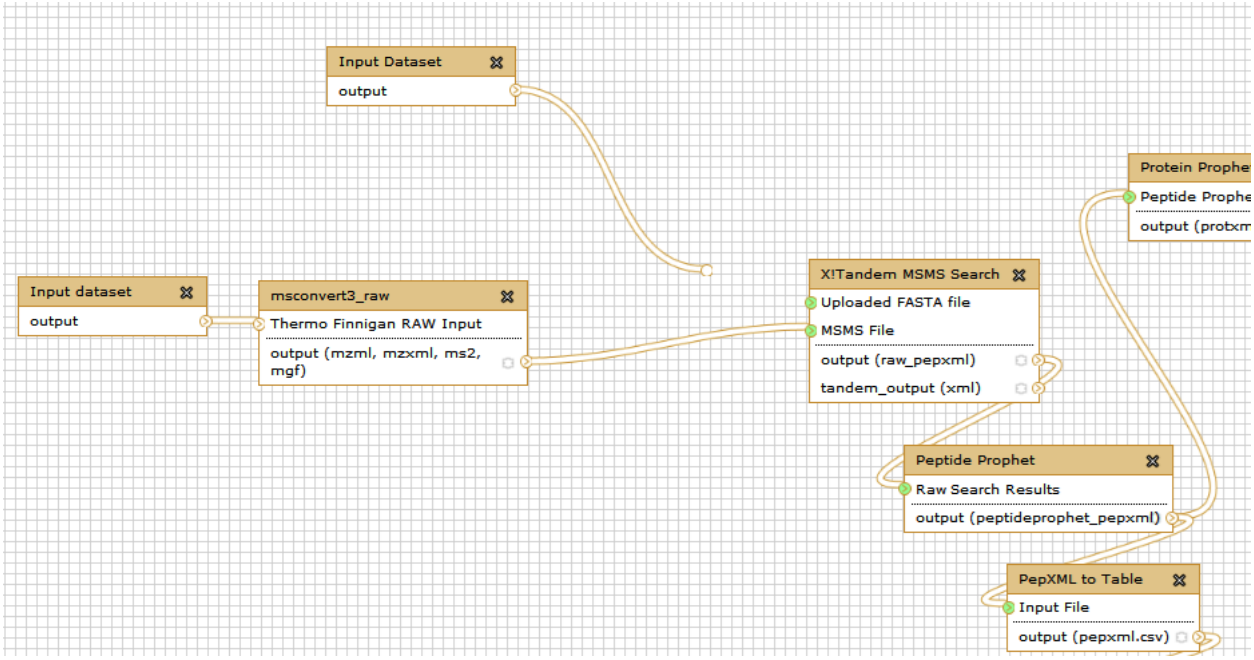


You can now scroll down to the left bottom corner of the screen and click on inputs. 'Input dataset' option will show up. Click on it.

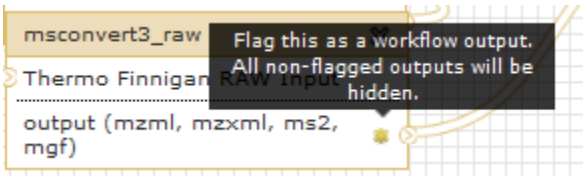
Once this is done, 'input dataset' box will appear in the central Window pane.



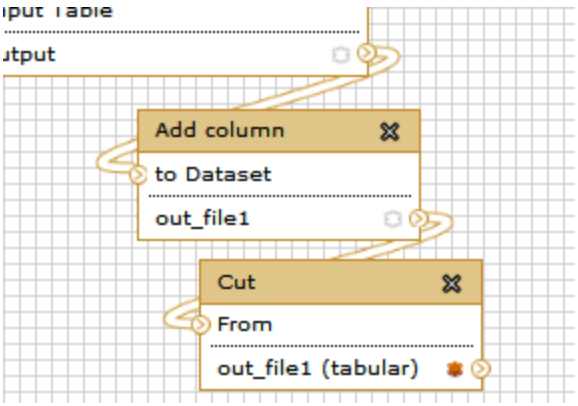
Position the Input dataset pane in such a way that you can join the output arrow on this box with the input arrow for the X!tandem search. In other words, we are providing a FASTA database input for the X!tandem search. The significance of this step will become clearer when we use this workflow.



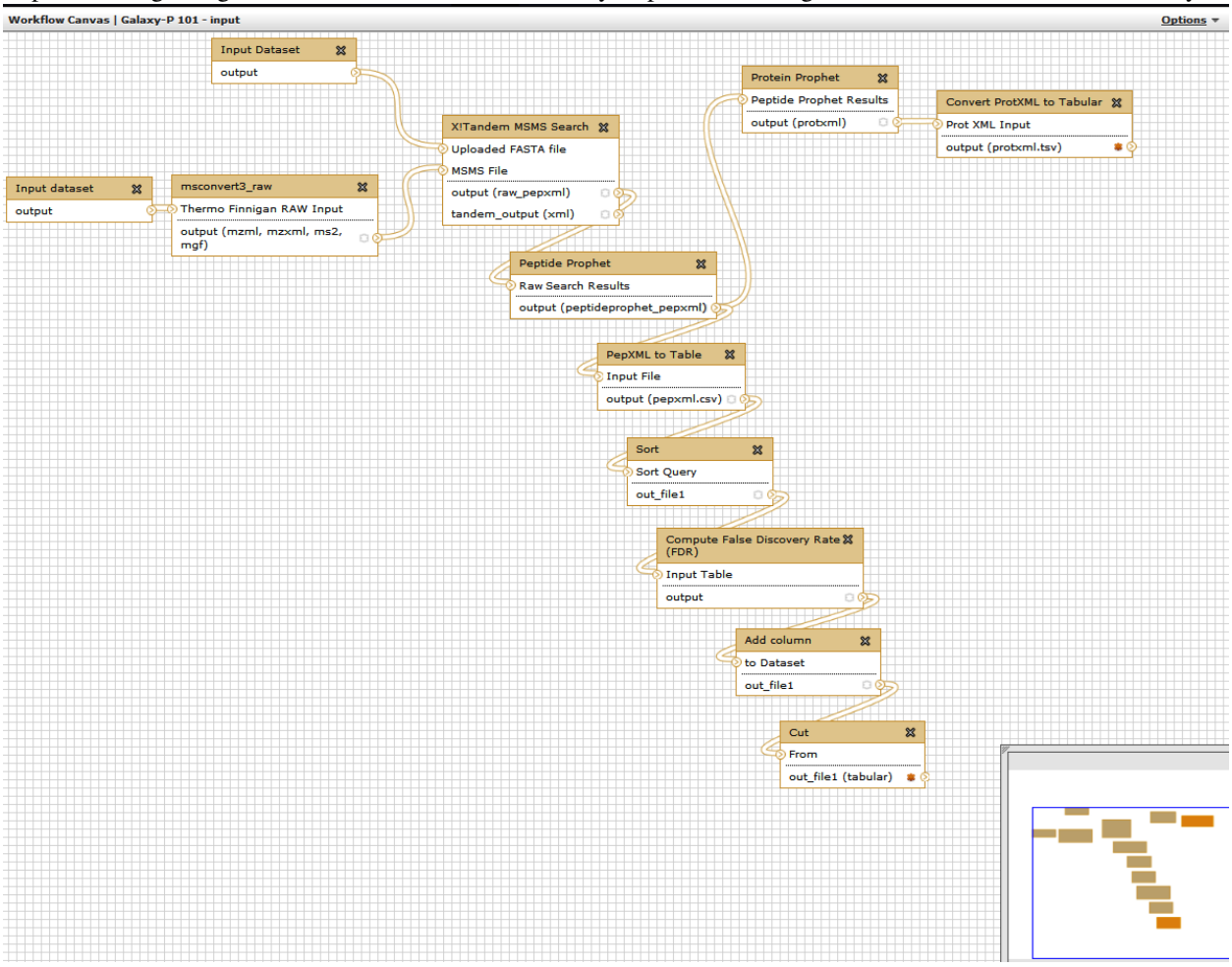
Below is mentioned one of the many things that you can do with workflows. When workflow is executed one is usually interested in the final product and not in the intermediate steps. These steps can be hidden by mousing over a small asterisk in the lower right corner of every tool box:



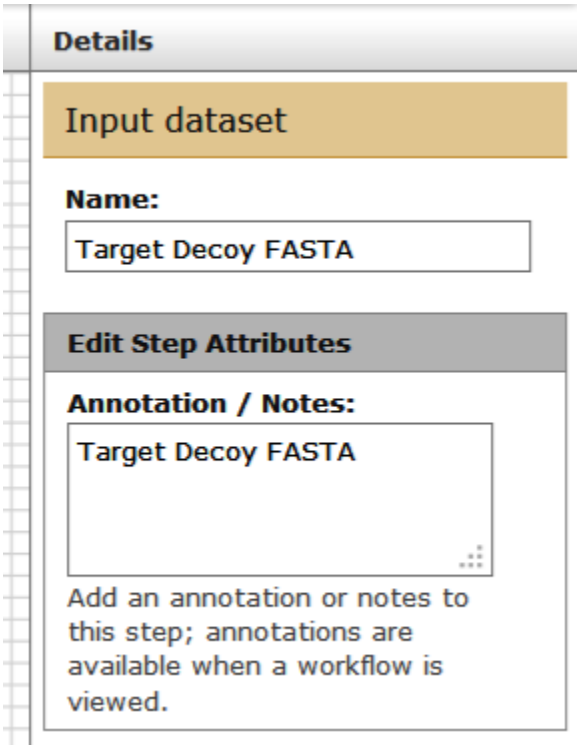
Yet there is a catch. In a newly created workflow all steps are hidden by default and default behavior of Galaxy is that if all steps of a given workflow are hidden, then nothing gets hidden in the history. This may be counterintuitive, but this is done to decrease the amount of clicking if you do want to hide some steps. So in our case if we want to hide all intermediate steps with the exception of the last one we will click that asterisk in last step of the workflow:



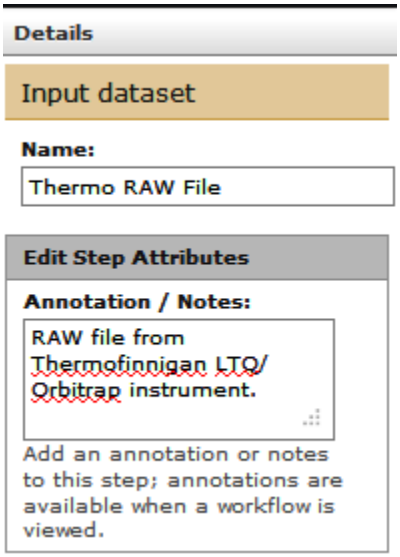
Once you do this the representation of the workflow in the bottom right corner of the editor will change with these steps becoming orange. This means that these are the only steps, which will generate datasets visible in the history:



Right now both inputs to the workflow look exactly the same. This is a problem as will be very confusing which input should be FASTA files and which should be RAW file. In your workflow you will see that the top input dataset connects to the X!tandem search , so it must correspond to the Target Decoy database. If you click on this box you will be able to rename the dataset in the right pane:



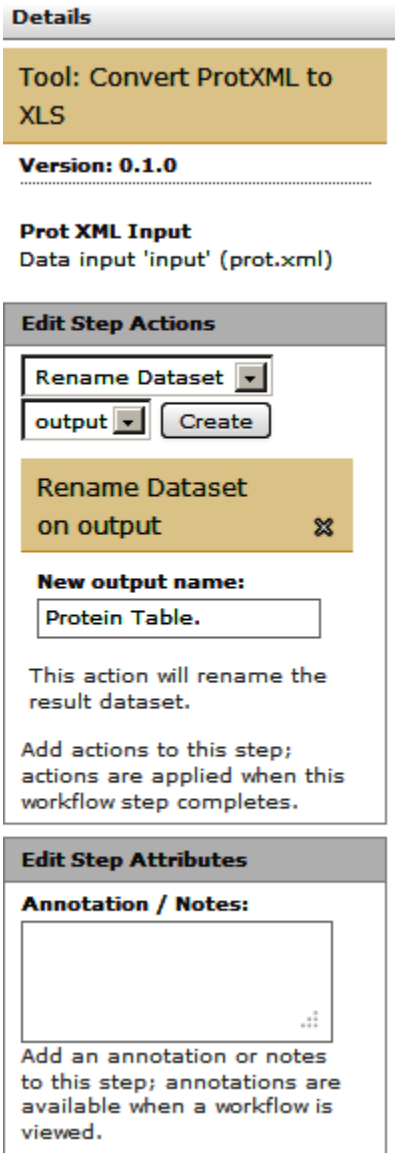
... and Thermo RAW file (second input).



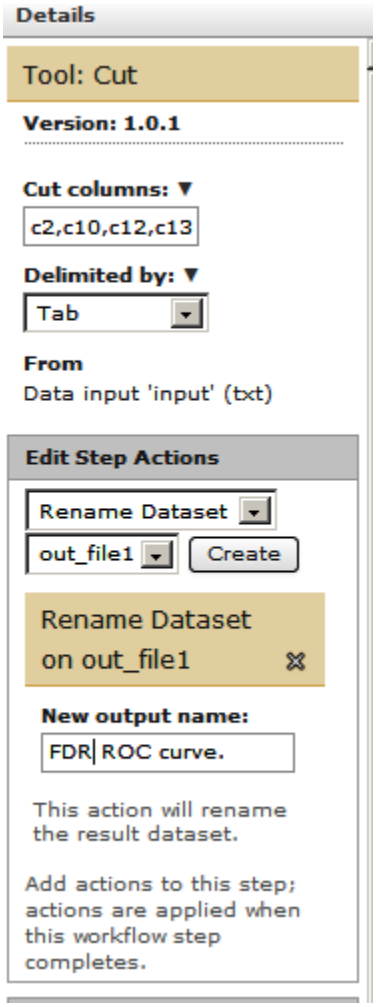
Feel free to annotate as many steps as you can so that it can be easier for you to revisit and understand the workflow or easier to share it with others.

Renaming outputs

Finally let's rename the workflow's output. For this click on one of the last datasets ("Convert ProtXML to Tabular") and in the Edit Step Actions dialogue box select "Rename Dataset".



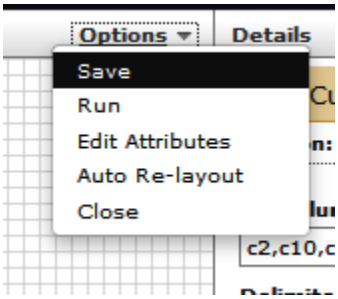
Similarly, for the other output dataset ("Cut") , click on the box and in the Edit Step Actions dialogue box select "Rename Dataset".



Remember you can highlight as many "outputs" as you want to and rename them for sake of a more complete and shareable annotation.

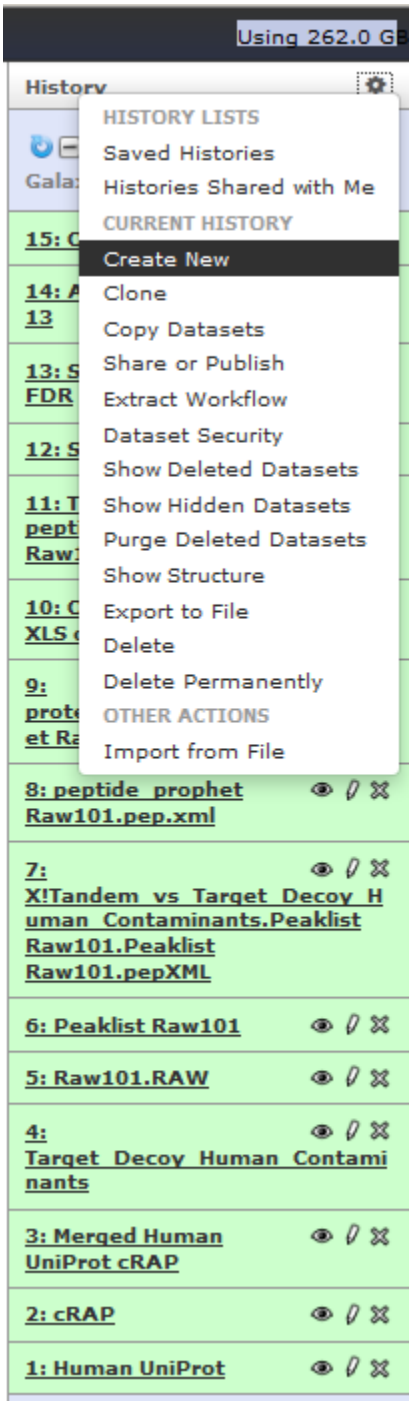
Save! It is important...

Now let's save the changes we've made by clicking Options (top of the center pane) and selecting Save:

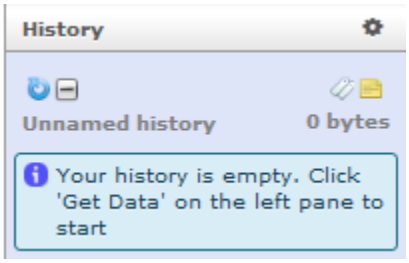


2.15 Applying Workflows to Your Data

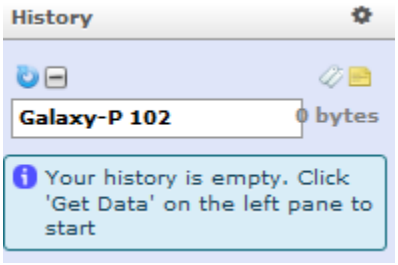
Let us use this workflow on another Raw file. For this go to Analyze Data –> History options icon and select create new history.



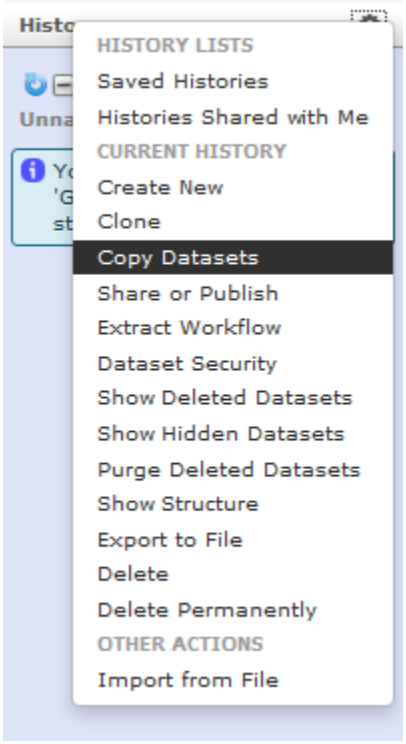
This will create a new history with no datafiles.



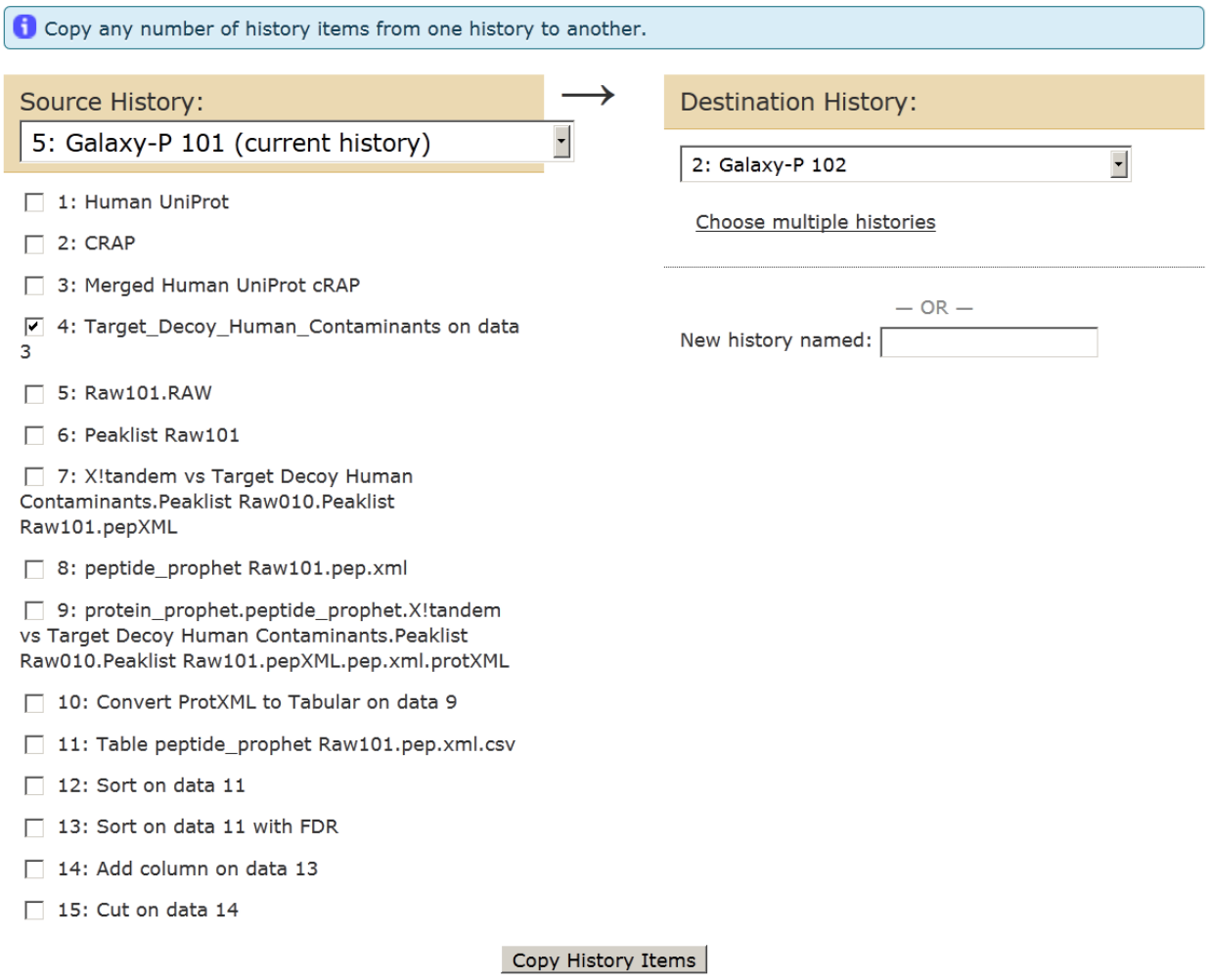
Let us name this history Galaxy-P 102.



We will transfer some of the files from our Galaxy-P 101 history to Galaxy-P 102 history. For this go to History options icon and click on copy datasets.



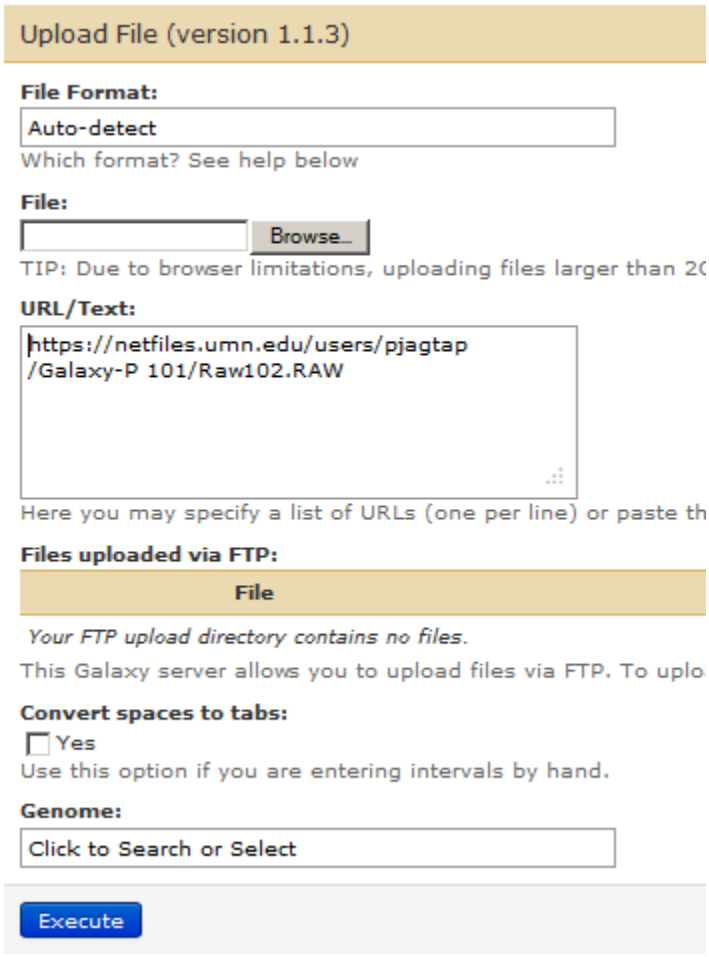
This will open up a new central window pane. Transfer the target-decoy FASTA file from Galaxy P-101 (Source History) to Galaxy-P 102 (Destination history).



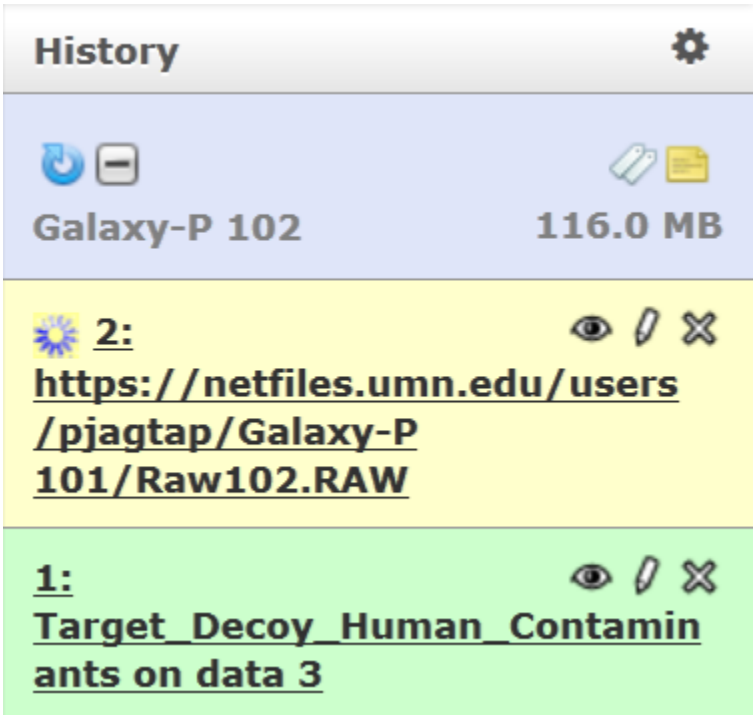
This will transfer the target-decoy FASTA file in your current history.

Click on Galaxy-P 102 history and let us download another raw file from the following link:

https://netfiles.umn.edu/users/pjagtap/Galaxy-P 101/Raw102.RAW

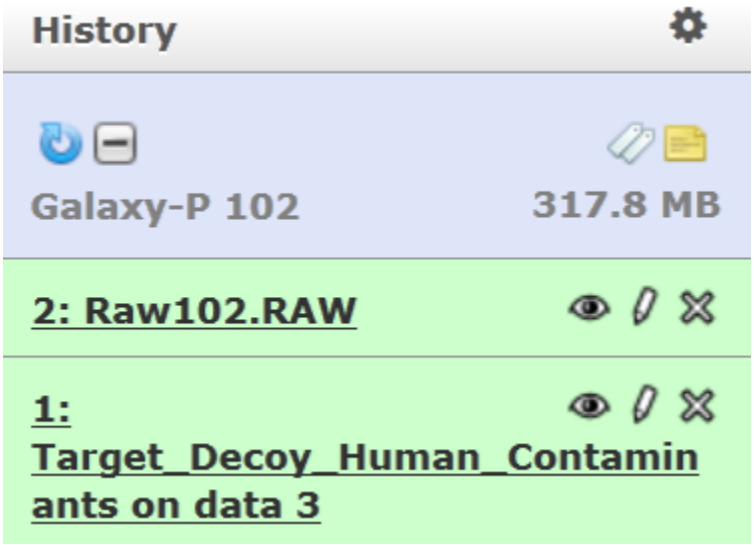


Clicking on Execute will add second file to this history.



Rename the RAW file by using the pencil icon. Change the name to “Raw102”.

Now, for your rerun, your history template with 2 files (Galaxy-P 102) is ready.



Go to Workflows tab and open the Galaxy-P 101 workflow and select “Run”.

Select appropriate files from your current history as inputs.

Step 1: Input dataset

RAW

RAW Input

2: Raw102.RAW

type to filter

Step 2: Input dataset

Target Decoy FASTA

Target Decoy FASTA

1: Target_Decoy_Huma..s on data 3

type to filter

Step 3: msconvert3_raw (version 0.1.0)

Step 4: X!Tandem MSMS Search (version 1.0.0)

Step 5: Peptide Prophet (version 1.1.0)

Step 6: PepXML to Table (version 1.0.0)

Step 7: Protein Prophet (version 1.0.0)

Step 8: Sort (version 1.0.1)

Step 9: Convert ProtXML to Tabular (version 0.1.0)

Step 10: Compute False Discovery Rate (FDR) (version 0.1)

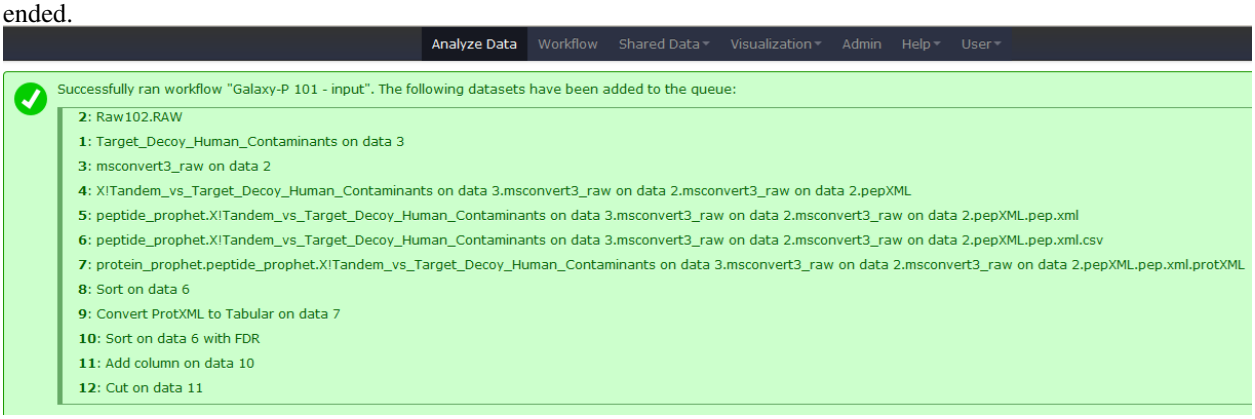
Step 11: Add column (version 1.0.0)

Step 12: Cut (version 1.0.1)

Send results to a new history

You can choose to run this in the same history or create a history of "only outputs" in a new history from this analysis. For this tutorial, we will run it in the same history.

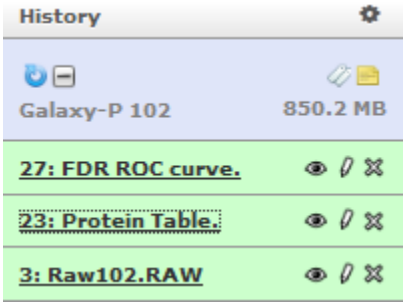
Click on Run Workflow. Luckily you do not have to wait as Galaxy will automatically start jobs once uploads have ended.




History
Galaxy-P 102
317.8 MB
12: Cut on data 11
11: Add column on data 10
10: Sort on data 6 with FDR
9: Convert ProtXML to Tabular on data 7
8: Sort on data 6
7: protein_prophet.peptide_prophet.X!Tandem_vs_Target_Decoy_Human_Contaminants on data 3.msconvert3_raw on data 2.msconvert3_raw on data 2.pepXML.pep.xml.protXML
6: peptide_prophet.X!Tandem_vs_Target_Decoy_Human_Contaminants on data 3.msconvert3_raw on data 2.msconvert3_raw on data 2.pepXML.pep.xml.csv
5: peptide_prophet.X!Tandem_vs_Target_Decoy_Human_Contaminants on data 3.msconvert3_raw on data 2.msconvert3_raw on data 2.pepXML.pep.xml
4: X!Tandem_vs_Target_Decoy_Human_Contaminants on data 3.msconvert3_raw on data 2.msconvert3_raw on data 2.pepXML
3: msconvert3_raw on data 2
2: Raw102.RAW
1: Target_Decoy_Human_Contaminants on data 3

Get coffee

As we mentioned above this will take some time, so go get coffee and then you will see this. Note that because all intermediate steps of the workflow were hidden, once it is finished you will only see the final dataset.



CHAPTER 3

FTP

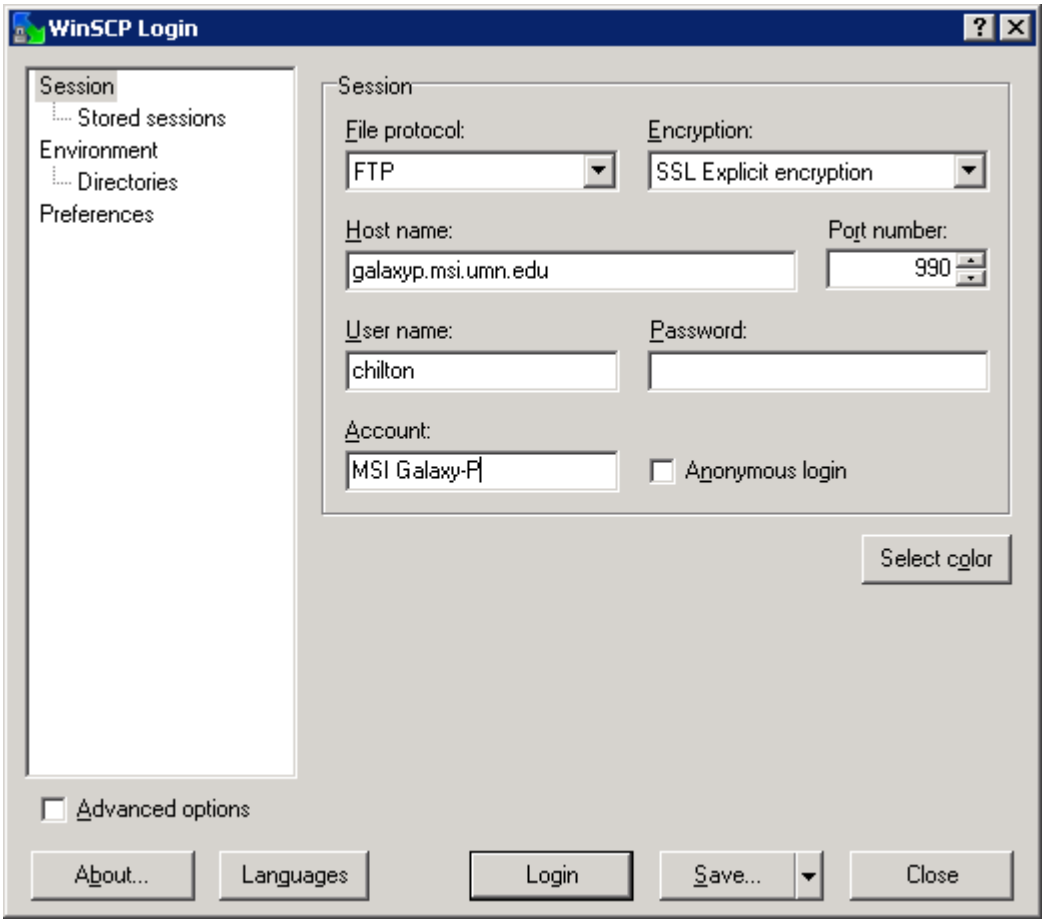
Galaxy provides the ability to upload files via the web interface (under the tool menu "Data Source" -> "Upload"), however this only allows one to upload one file at a time and web browsers generally limit uploads to 2 GB. Galaxy-P provides an FTP server that can be used to upload more and larger files.

3.1 Uploading Files

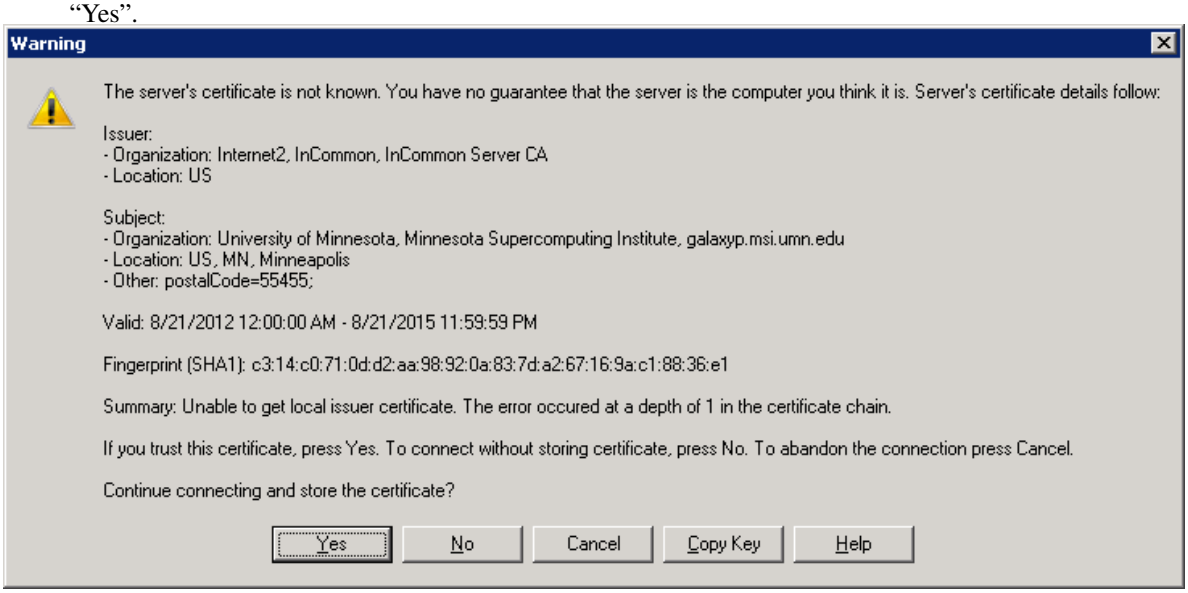
Property	Value
Hostname	galaxyp.msi.umn.edu
Connection	FTPS / FTP + Encyrption
Username	Your MSI username
Port	990

The following walkthrough demonstrates how to connect to Galaxy-P FTP server using WinSCP and upload RAW data files. WinSCP is demonstrated because it is a popular piece of freely available software, but many other tools could be used as long as they support FTP with encryption.

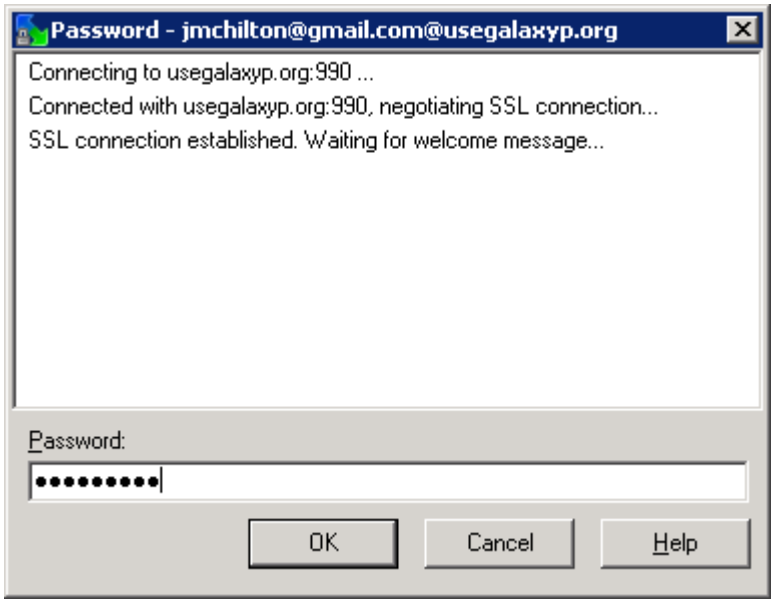
- Open WinSCP, specify connection information, and then click login. You may also want to give this connection a name and save it for later reuse.



- The first time you connect, you will likely be prompted to store the hosts SSL certificate, do this by clicking "Yes".

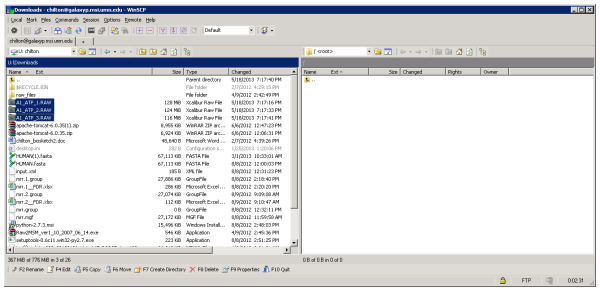


- When prompted for your password, please enter it and click "Okay".

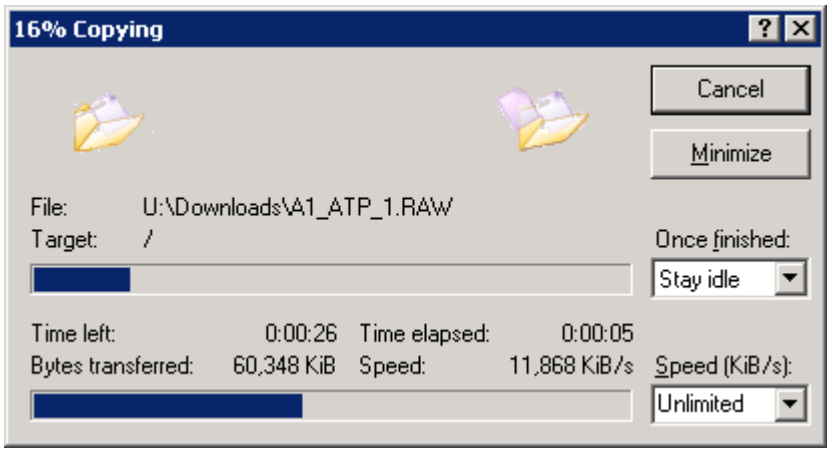


- If everything has gone well, you should now see two file browsers. The one on your left is your computer's files and the one on the right is your Galaxy-P staging area (which should be initially empty).

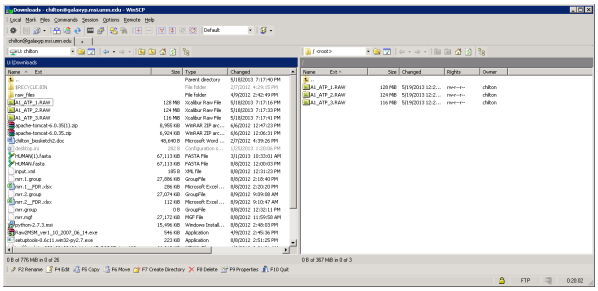
  Using the left file browser, navigate to the files you wish to upload and select them.



- Drag and drop these files to the right panel to begin the transfer and wait as the files are transfered.

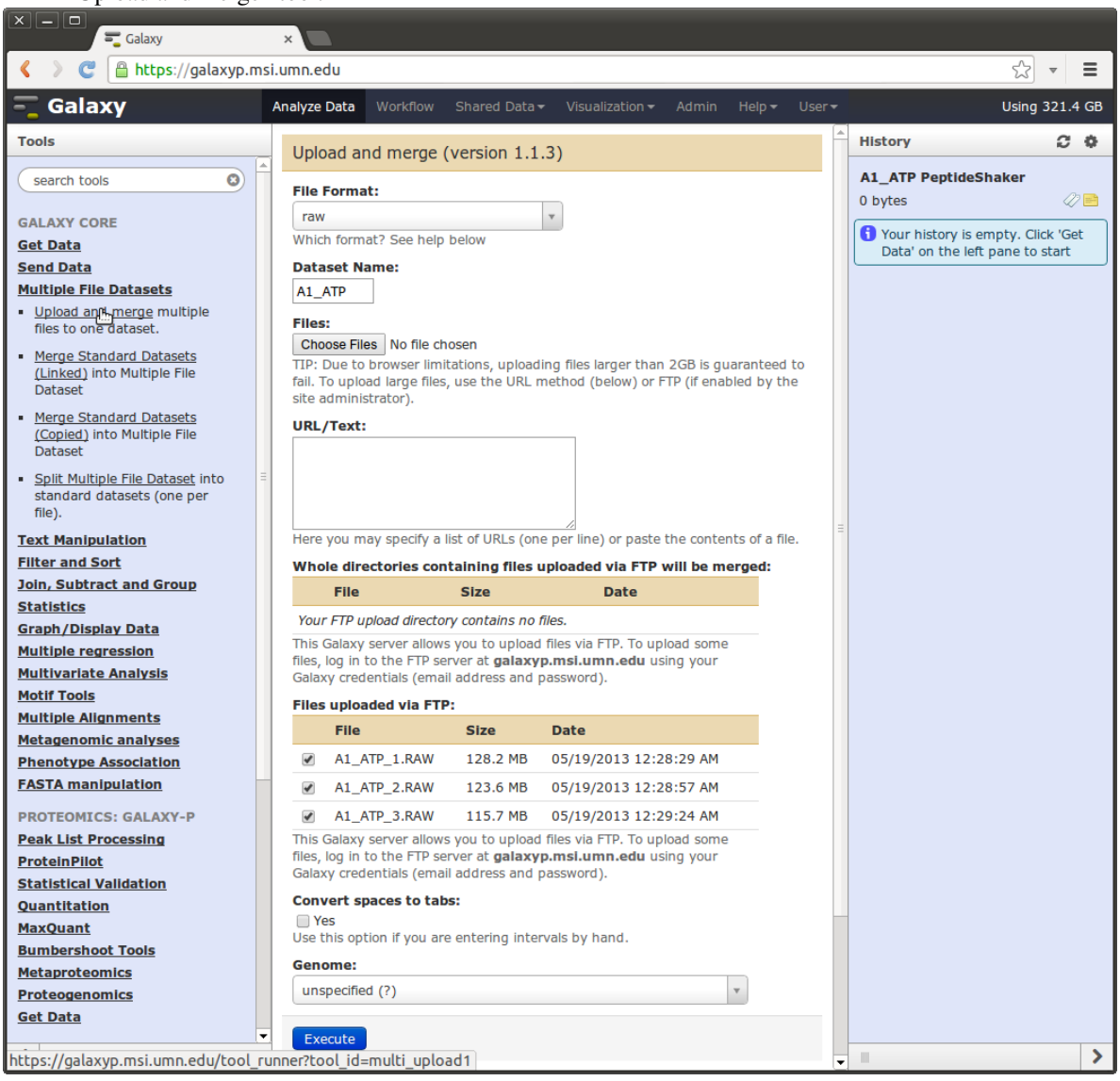


- Verify your files have been copied.



- These files may now be imported into a Galaxy history using the “Data Source” -> “Upload” tool. When filling out the upload information, instead of using the browser upload option simply check the uploaded files in the “Files uploaded by FTP” section.

  Likewise, these a multiple file dataset can be created using these files and the “Multiple File Datasets” -> “Upload and merge” tool.



3.2 See Also

- Galaxy Project Documentation on FTP Uploads

CHAPTER 4

Multiple File Datasets

4.1 Introduction

Traditional Galaxy workflows require a fixed number of input files and may produce a large number of intermediate files per input. Additionally, Galaxy renames datasets with each step, making tracking samples or fractions across long workflows very difficult. These issues together make Galaxy a problematic platform for workflows and application areas that require dealing with a large number of files. The prevalence of fractionated samples in mass spectrometry makes proteomics such an application area.

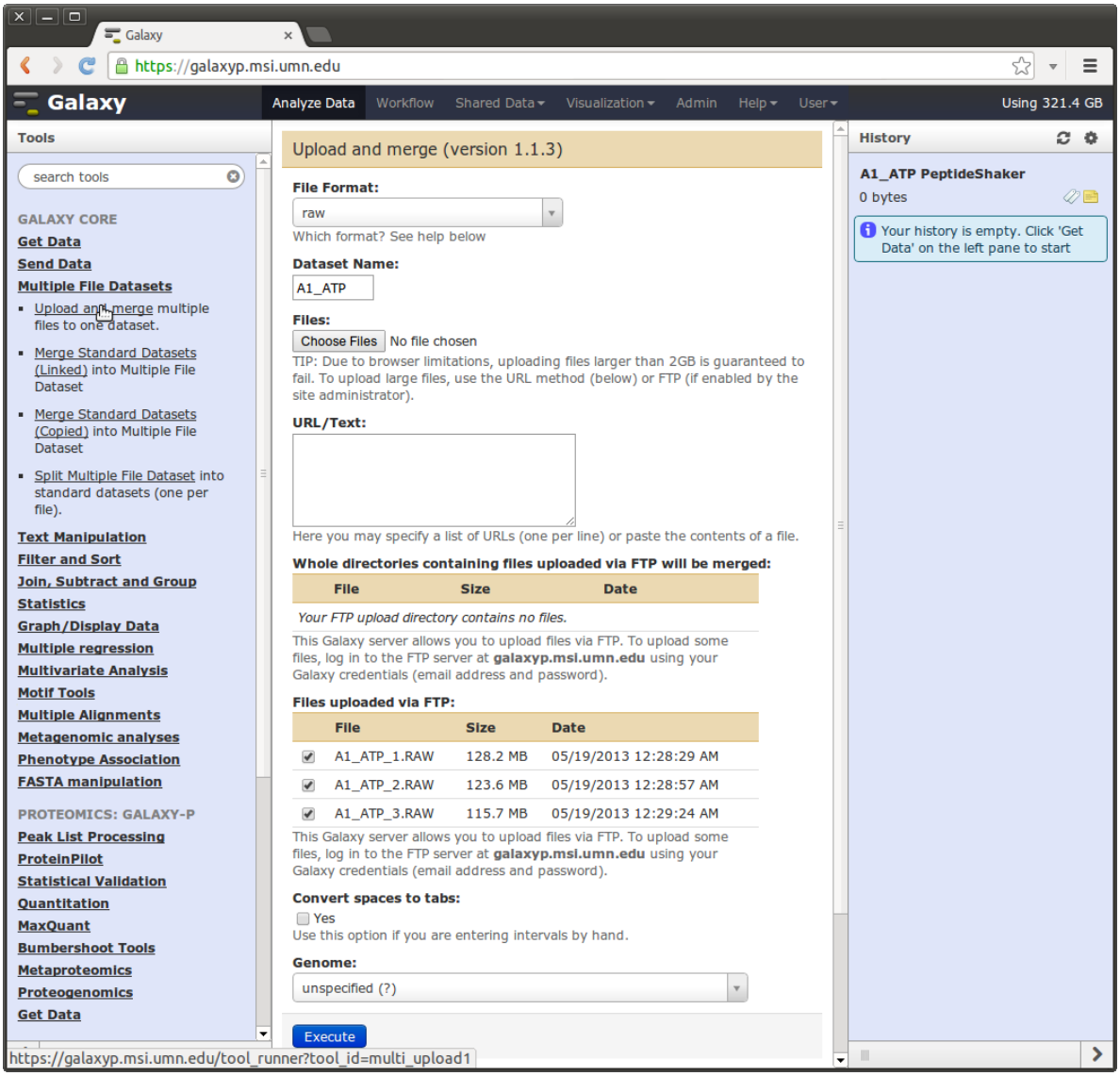
Galaxy-P however utilizes an extension to the core Galaxy framework called "Multiple File Datasets" designed to address these shortcomings in Galaxy. In simple terms, Galaxy-P can group many similar files into a single dataset (called a multiple file dataset). Normal Galaxy tools can then use these datasets and produce multiple file datasets of their own, operating on each item in parallel. In addition to keeping the number of datasets in the Galaxy managable, this allows for the creation of workflows that can operate over any number of files and the individual files in the multiple file dataset are given consistent, trackable names across such complex workflows making sample tarcking trivial.

4.2 Creating a Multiple File Dataset

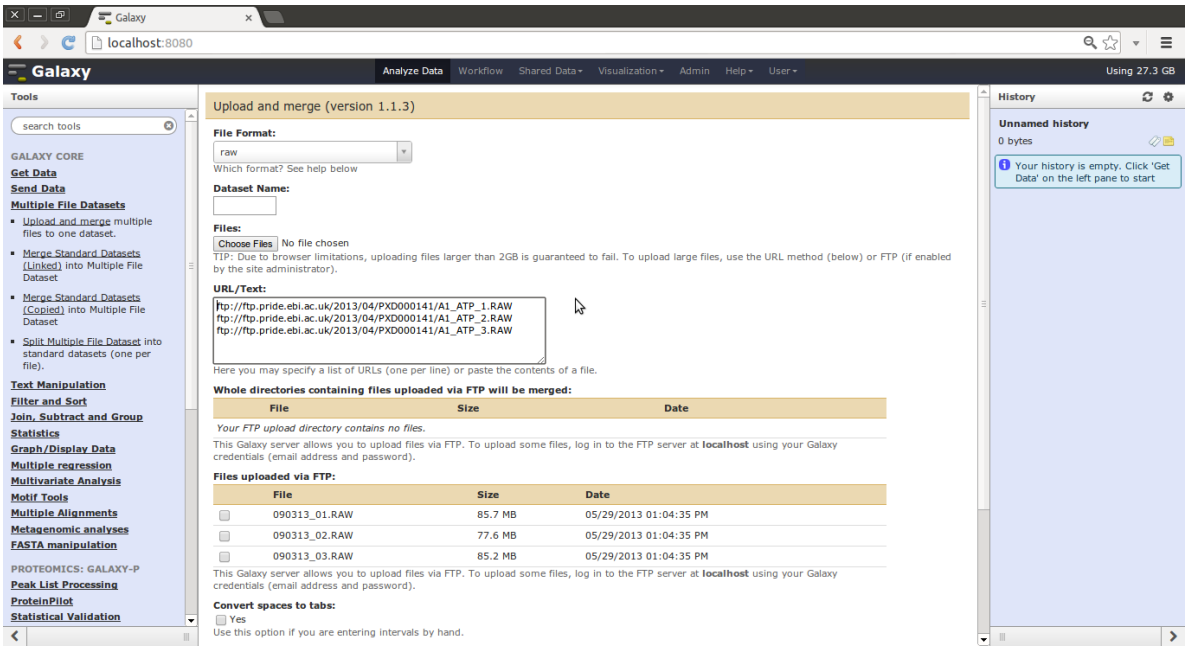
Once you have an initial multiple file dataset, most tools when operating on these datasets will in return produce multiple file dataset outputs. So for most proteomic workflows, the first step is simply to create a multiple file RAW dataset or a multiple file peak list (e.g. mzML or mgf) dataset.

There are several ways to do this.

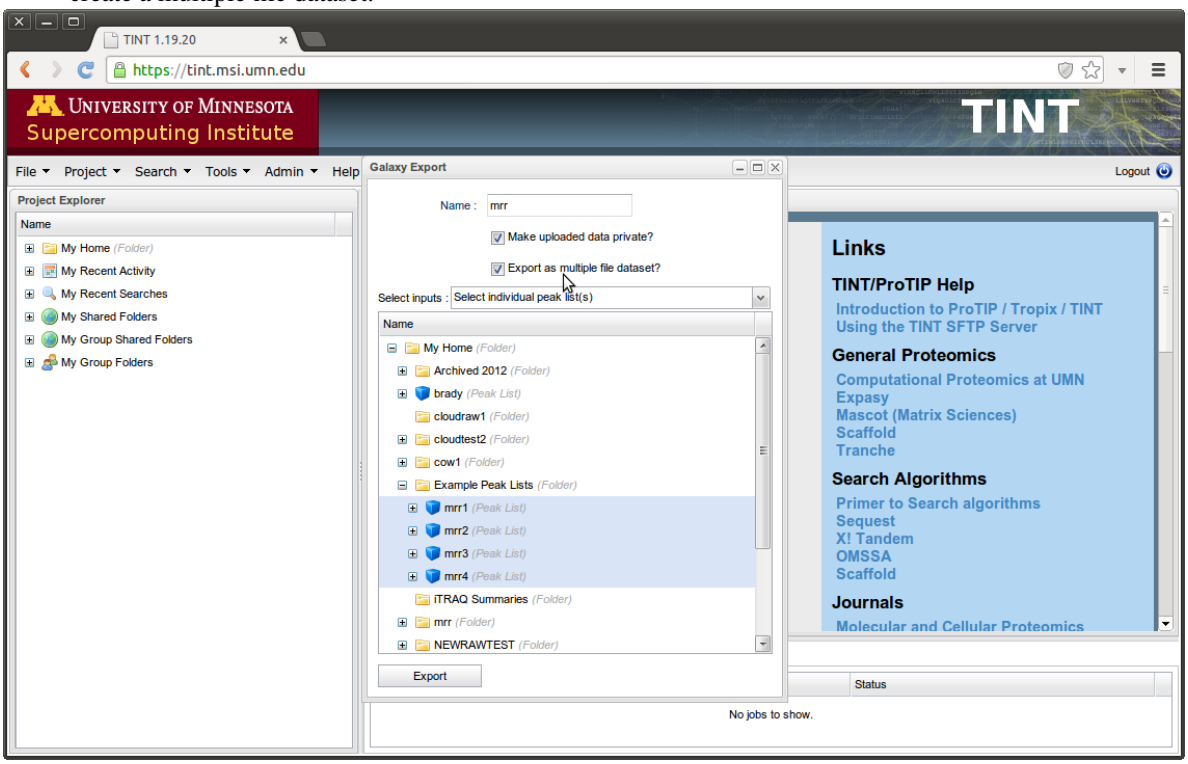
- One can use the "Multiple File Datasets" -> "Upload and Merge" tool along to create a multiple file datasets from files or a directory of files uploaded via FTP.



- One can use the “Multiple File Datasets” -> “Upload and Merge” tool and simply click the “Choose Files” button and select multiple file for upload.
- One can use the “Multiple File Datasets” -> “Upload and Merge” tool and paste in multiple URLs to have Galaxy download these files and create a multiple file dataset.



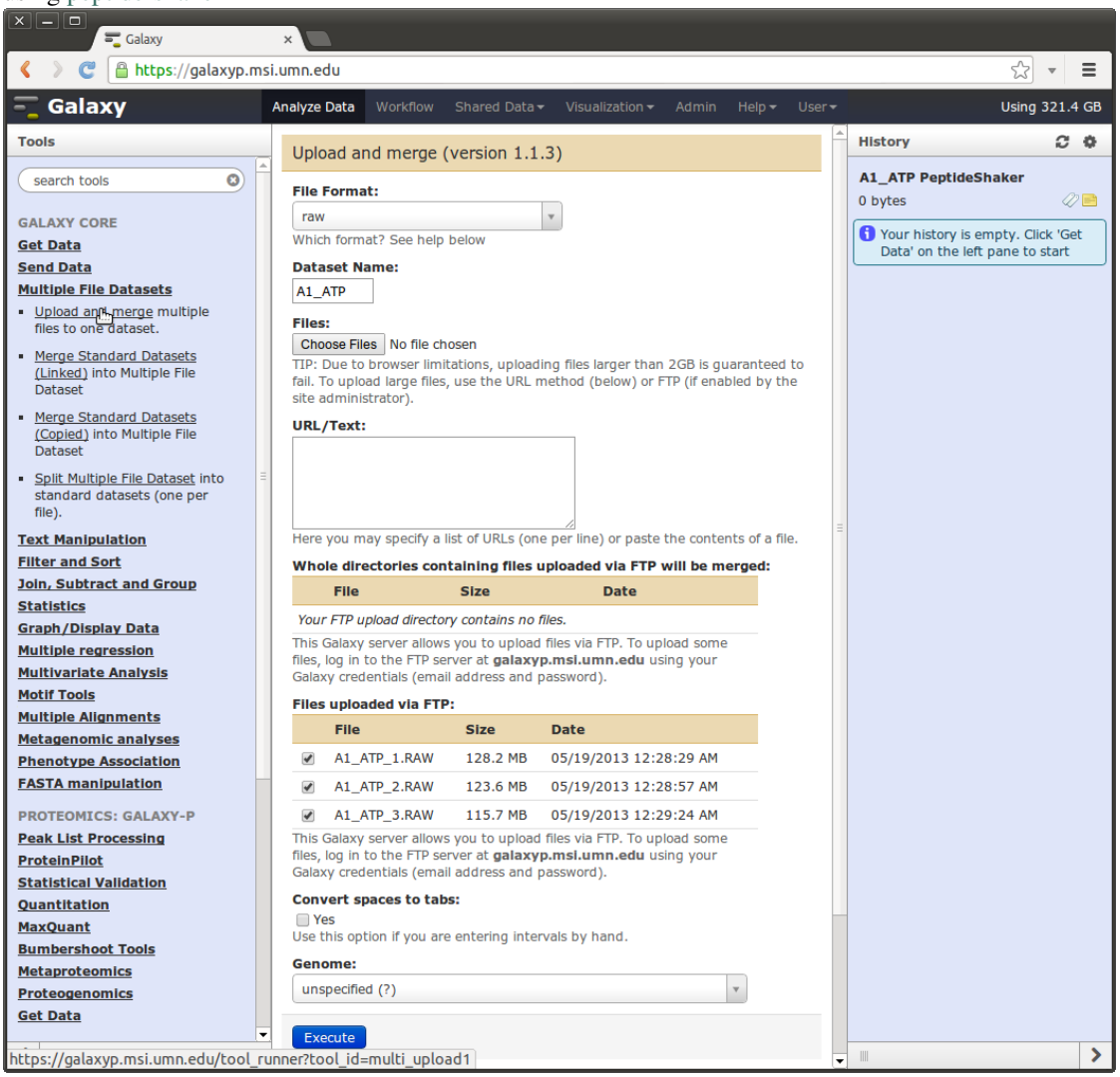
- When exporting data from TINT to Galaxy-P, simply check the "Export as multiple file dataset" checkbox to create a multiple file dataset.

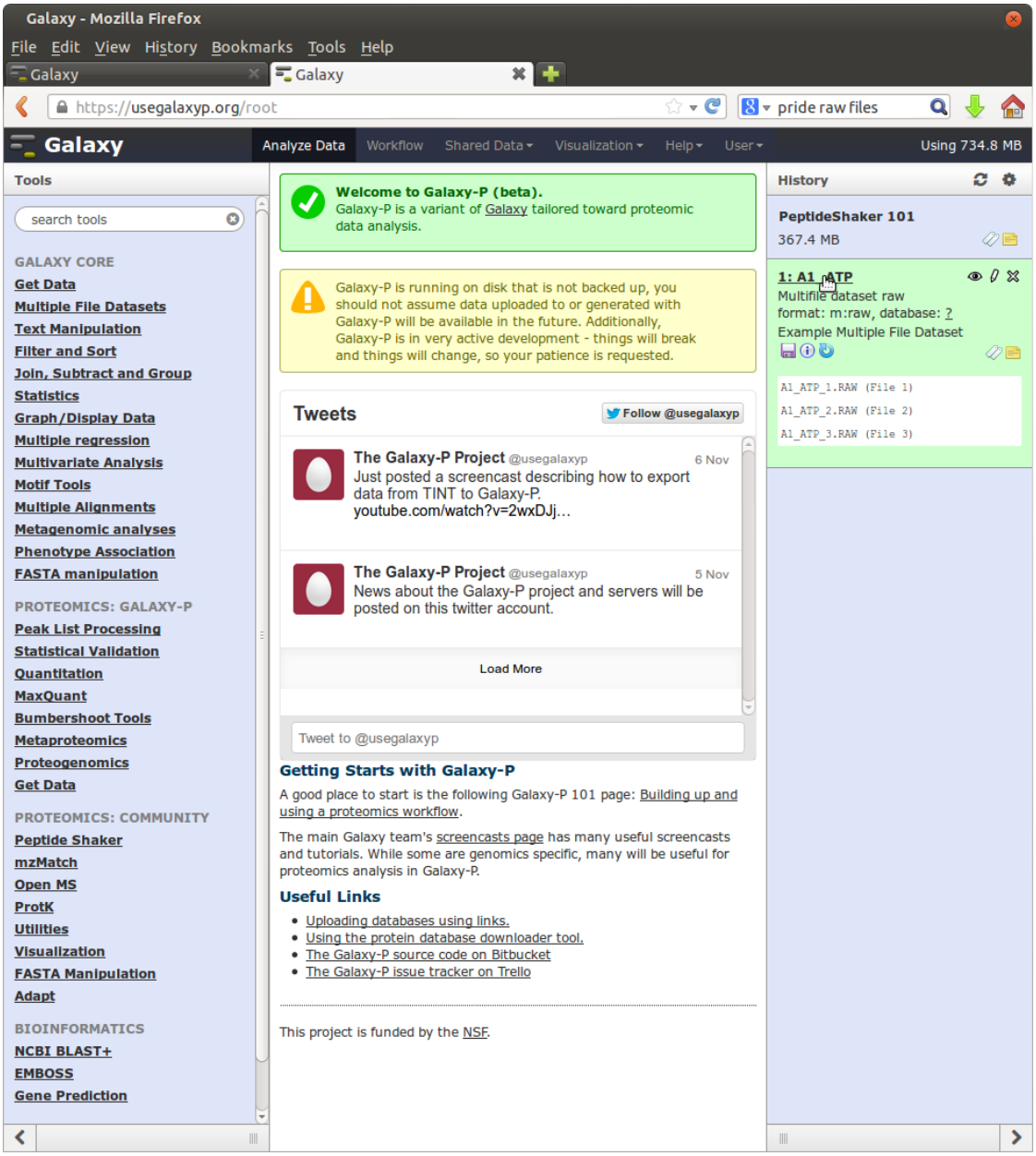


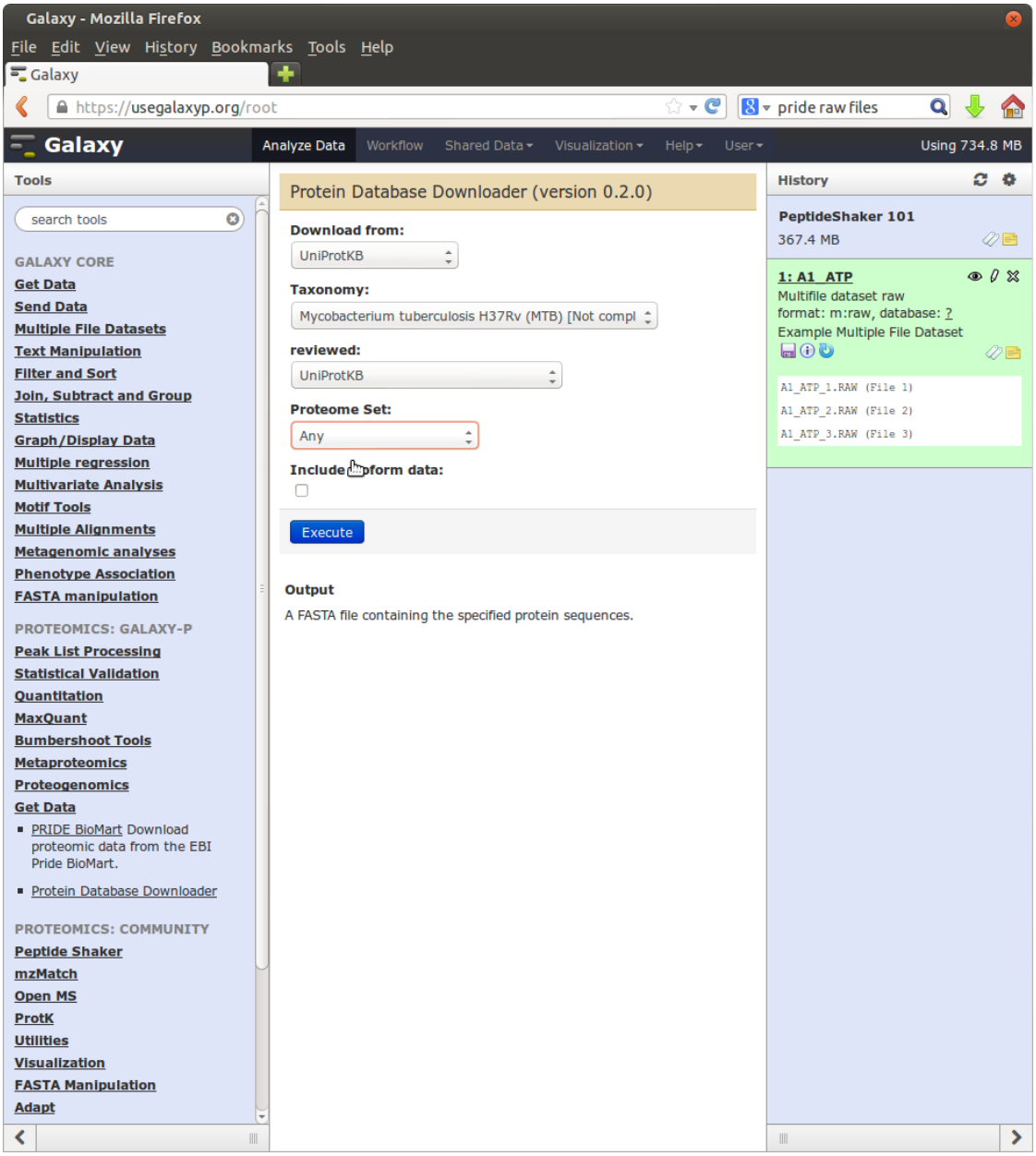
CHAPTER 5

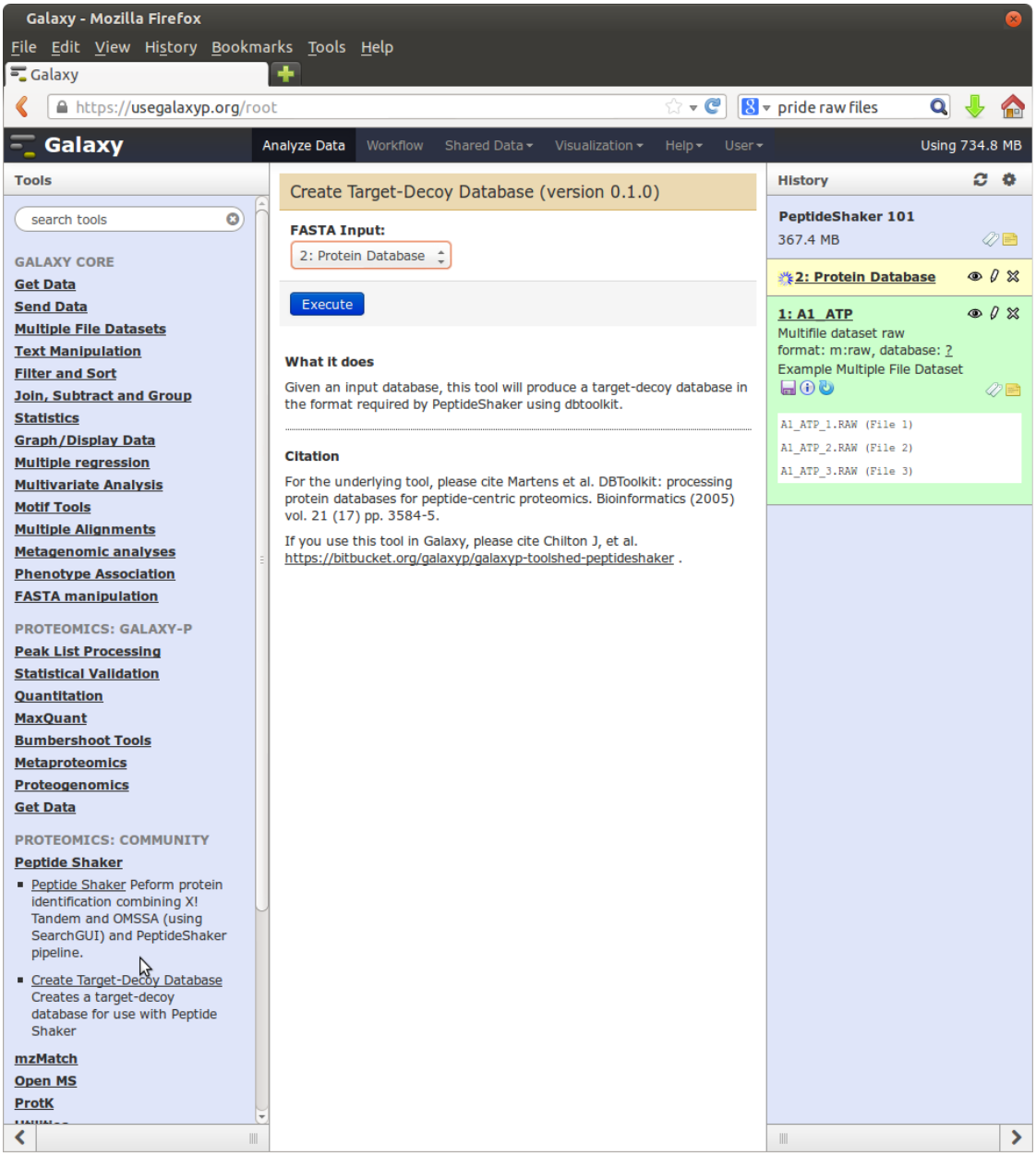
A Simple Worklfow using peptide-shaker

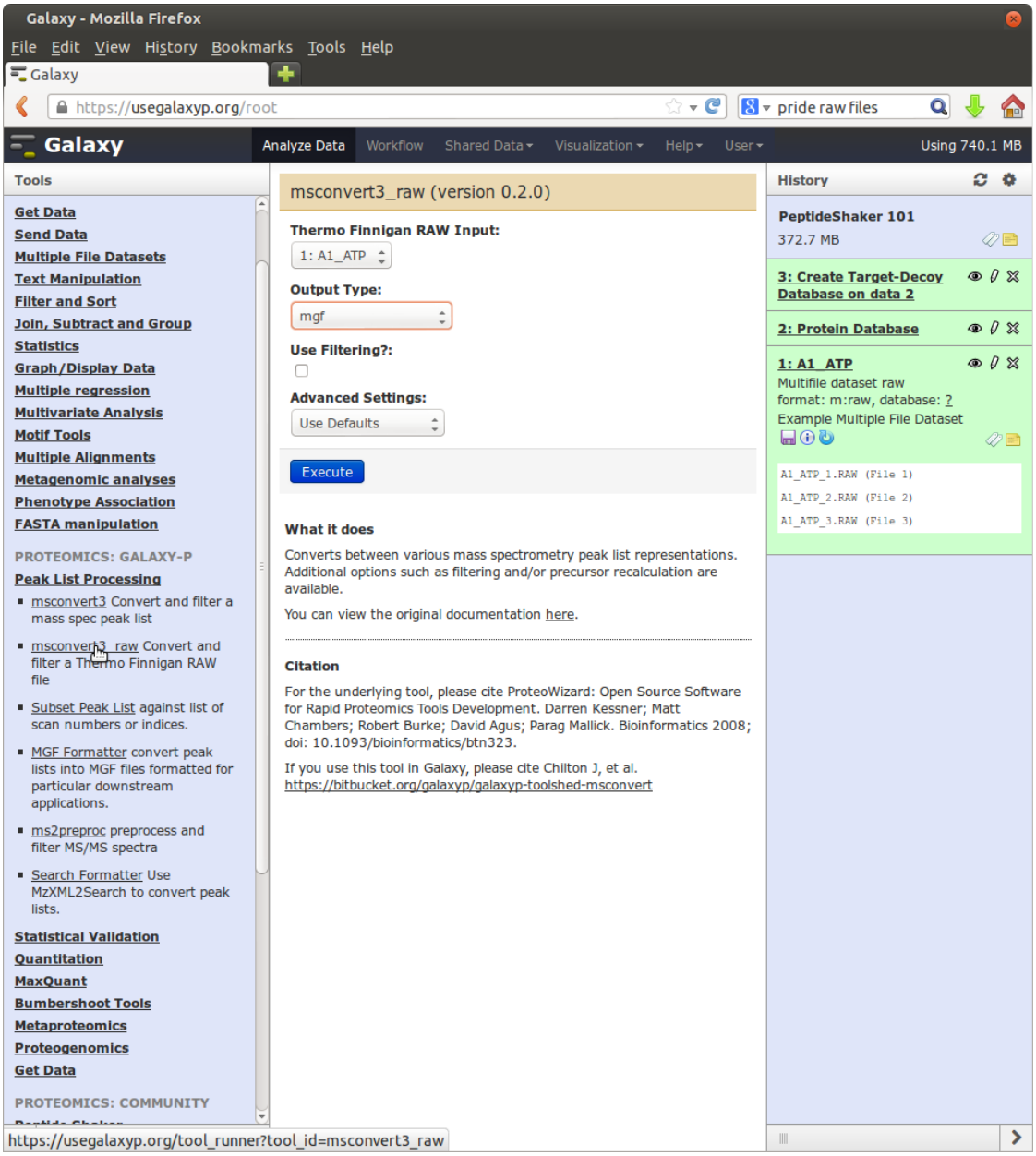
This section is a simple walkthrough of using Galaxy-P and multiple file datasets to analysis a collection of RAW files using peptide-shaker

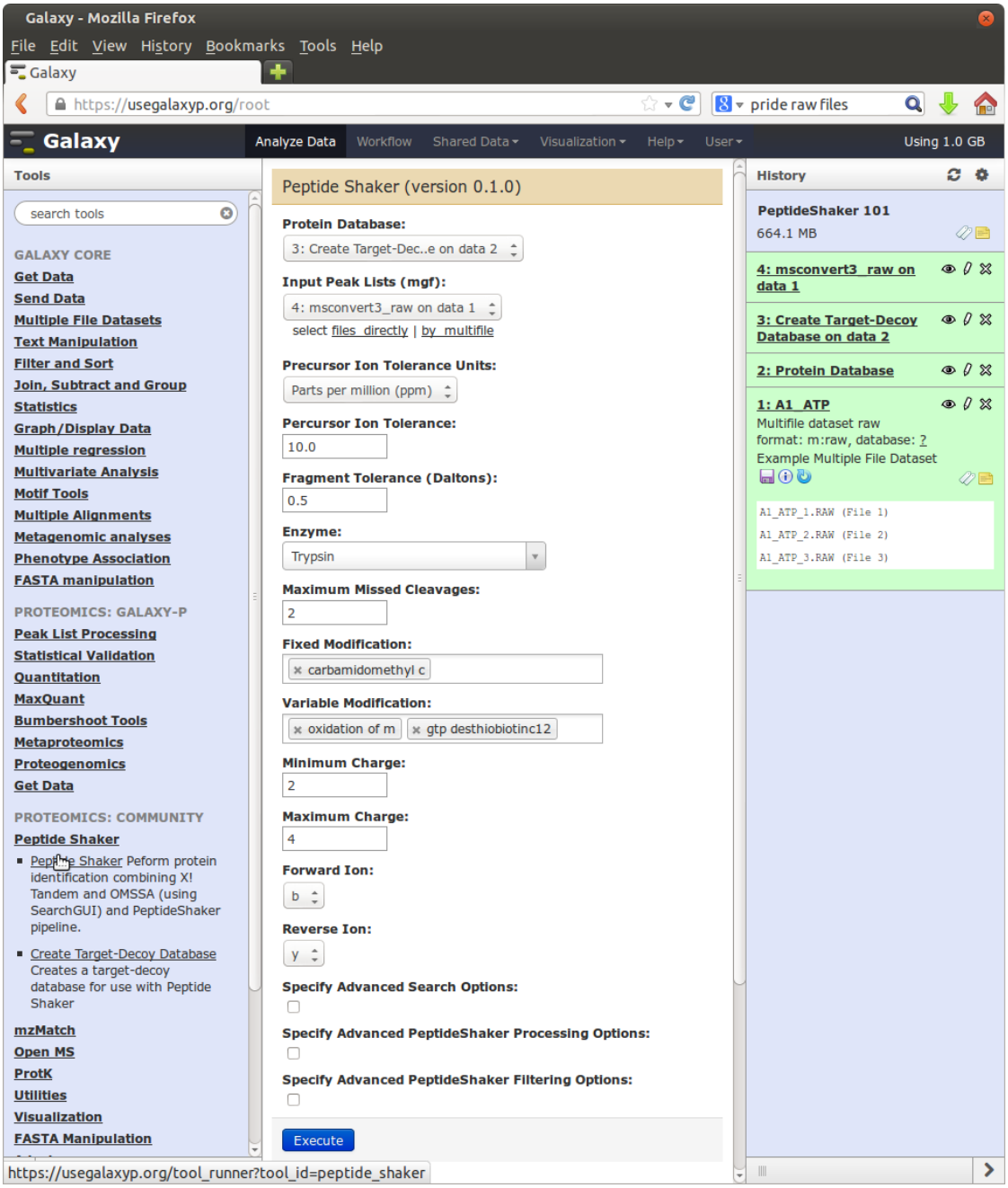


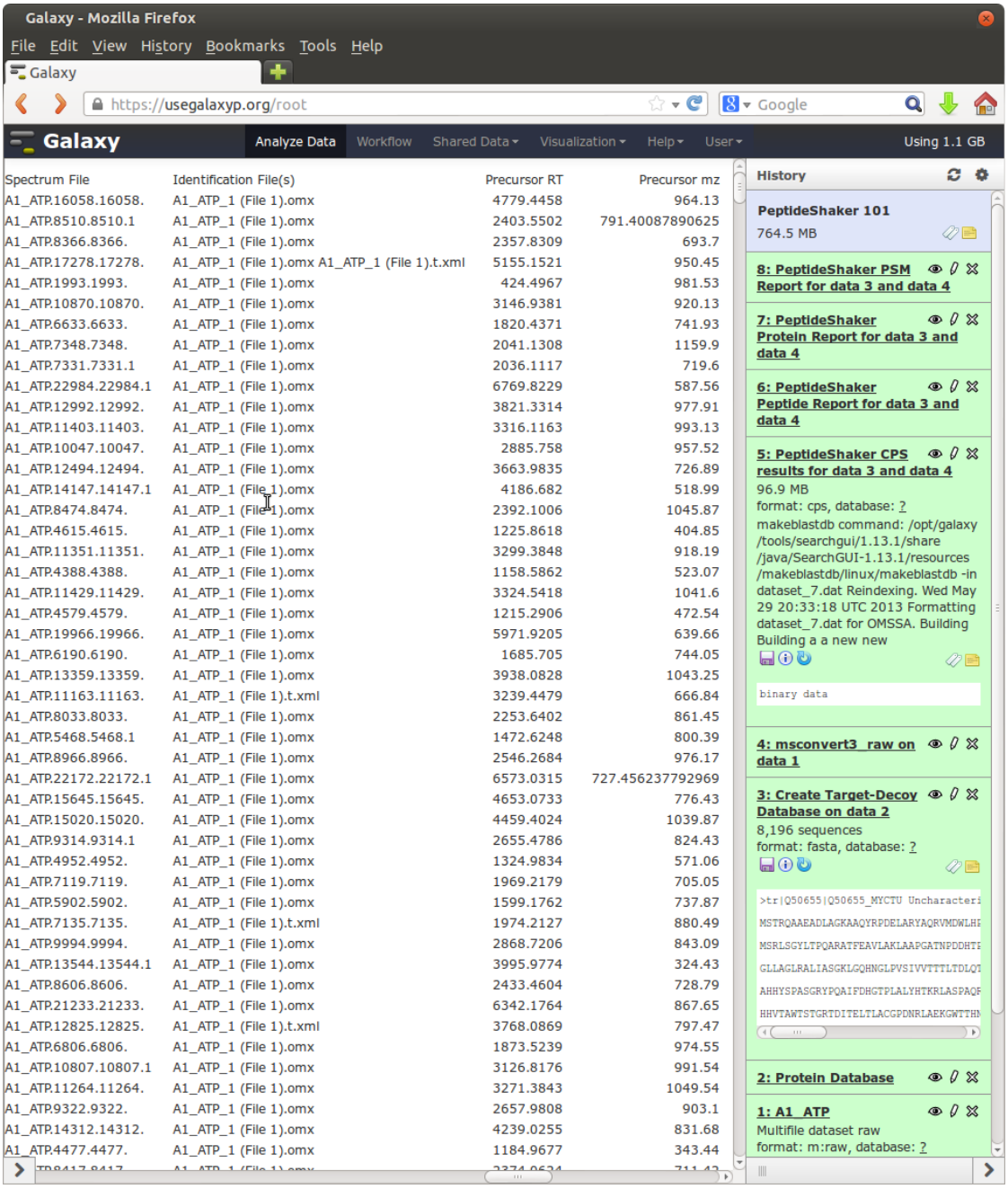

Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Galaxy
Galaxy
https://usegalaxyp.org/root
pride raw files
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 734.8 MB
Tools
search tools
GALAXY CORE
Get Data
Multiple File Datasets
Text Manipulation
Filter and Sort
Join, Subtract and Group
Statistics
Graph/Display Data
Multiple regression
Multivariate Analysis
Motif Tools
Multiple Alignments
Metagenomic analyses
Phenotype Association
FASTA manipulation
PROTEOMICS: GALAXY-P
Peak List Processing
Statistical Validation
Quantitation
MaxQuant
Bumbershoot Tools
Metaproteomics
Proteogenomics
Get Data
PROTEOMICS: COMMUNITY
Peptide Shaker
mzMatch
Open MS
ProtK
Utilities
Visualization
FASTA Manipulation
Adapt
BIOINFORMATICS
NCBI BLAST+
EMBOSS
Gene Prediction
Welcome to Galaxy-P (beta).
Galaxy-P is a variant of Galaxy tailored toward proteomic data analysis.
Galaxy-P is running on disk that is not backed up, you should not assume data uploaded to or generated with Galaxy-P will be available in the future. Additionally, Galaxy-P is in very active development - things will break and things will change, so your patience is requested.
Tweets
Follow @usegalaxyp
The Galaxy-P Project @usegalaxyp 6 Nov
Just posted a screencast describing how to export data from TINT to Galaxy-P. youtube.com/watch?v=2wxDJj…
The Galaxy-P Project @usegalaxyp 5 Nov
News about the Galaxy-P project and servers will be posted on this twitter account.
Load More
Tweet to @usegalaxyp
Getting Starts with Galaxy-P
A good place to start is the following Galaxy-P 101 page: Building up and using a proteomics workflow.
The main Galaxy team's screencasts page has many useful screencasts and tutorials. While some are genomics specific, many will be useful for proteomics analysis in Galaxy-P.
Useful Links
Uploading databases using links.
Using the protein database downloader tool.
The Galaxy-P source code on Bitbucket
The Galaxy-P issue tracker on Trello
This project is funded by the NSF.
History
PeptideShaker 101
367.4 MB
1: A1_ATP
Multifile dataset raw
format: m:raw, database: ?
Example Multiple File Dataset
A1_ATP_1.RAW (File 1)
A1_ATP_2.RAW (File 2)
A1_ATP_3.RAW (File 3)


Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Galaxy
https://usegalaxyp.org/root
pride raw files
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 734.8 MB
Tools
search tools
GALAXY CORE
Get Data
Send Data
Multiple File Datasets
Text Manipulation
Filter and Sort
Join, Subtract and Group
Statistics
Graph/Display Data
Multiple regression
Multivariate Analysis
Motif Tools
Multiple Alignments
Metagenomic analyses
Phenotype Association
FASTA manipulation
PROTEOMICS: GALAXY-P
Peak List Processing
Statistical Validation
Quantitation
MaxQuant
Bumbershoot Tools
Metaproteomics
Proteogenomics
Get Data
PRIDE BioMart Download proteomic data from the EBI Pride BioMart.
Protein Database Downloader
PROTEOMICS: COMMUNITY
Peptide Shaker
mzMatch
Open MS
ProtK
Utilities
Visualization
FASTA Manipulation
Adapt
Protein Database Downloader (version 0.2.0)
Download from:
UniProtKB
Taxonomy:
Mycobacterium tuberculosis H37Rv (MTB) [Not compl
reviewed:
UniProtKB
Proteome Set:
Any
Include isoform data:
Execute
Output
A FASTA file containing the specified protein sequences.
History
PeptideShaker 101
367.4 MB
1: A1_ATP
Multifile dataset raw
format: m:raw, database: ?
Example Multiple File Dataset
A1_ATP_1.RAW (File 1)
A1_ATP_2.RAW (File 2)
A1_ATP_3.RAW (File 3)


Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Galaxy
https://usegalaxyp.org/root
pride raw files
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 734.8 MB
Tools
search tools
GALAXY CORE
Get Data
Send Data
Multiple File Datasets
Text Manipulation
Filter and Sort
Join, Subtract and Group
Statistics
Graph/Display Data
Multiple regression
Multivariate Analysis
Motif Tools
Multiple Alignments
Metagenomic analyses
Phenotype Association
FASTA manipulation
PROTEOMICS: GALAXY-P
Peak List Processing
Statistical Validation
Quantitation
MaxQuant
Bumbershoot Tools
Metaproteomics
Proteogenomics
Get Data
PROTEOMICS: COMMUNITY
Peptide Shaker
Peptide Shaker Peform protein identification combining X! Tandem and OMSSA (using SearchGUI) and PeptideShaker pipeline.
Create Target-Decoy Database Creates a target-decoy database for use with Peptide Shaker
mzMatch
Open MS
ProtK
Create Target-Decoy Database (version 0.1.0)
FASTA Input:
2: Protein Database
Execute
What it does
Given an input database, this tool will produce a target-decoy database in the format required by PeptideShaker using dbtoolkit.
Citation
For the underlying tool, please cite Martens et al. DBToolkit: processing protein databases for peptide-centric proteomics. Bioinformatics (2005) vol. 21 (17) pp. 3584-5.
If you use this tool in Galaxy, please cite Chilton J, et al. https://bitbucket.org/galaxyp/galaxyp-toolshed-peptideshaker .
History
PeptideShaker 101
367.4 MB
2: Protein Database
1: A1_ATP
Multifile dataset raw format: m:raw, database: ?
Example Multiple File Dataset
A1_ATP_1.RAW (File 1)
A1_ATP_2.RAW (File 2)
A1_ATP_3.RAW (File 3)


Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Galaxy
https://usegalaxyp.org/root
pride raw files
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 740.1 MB
Tools
Get Data
Send Data
Multiple File Datasets
Text Manipulation
Filter and Sort
Join, Subtract and Group
Statistics
Graph/Display Data
Multiple regression
Multivariate Analysis
Motif Tools
Multiple Alignments
Metagenomic analyses
Phenotype Association
FASTA manipulation
PROTEOMICS: GALAXY-P
Peak List Processing
msconvert3 Convert and filter a mass spec peak list
msconvert3_raw Convert and filter a Thermo Finnigan RAW file
Subset Peak List against list of scan numbers or indices.
MGF Formatter convert peak lists into MGF files formatted for particular downstream applications.
ms2preproc preprocess and filter MS/MS spectra
Search Formatter Use MzXML2Search to convert peak lists.
Statistical Validation
Quantitation
MaxQuant
Bumbershoot Tools
Metaproteomics
Proteogenomics
Get Data
PROTEOMICS: COMMUNITY
msconvert3_raw (version 0.2.0)
Thermo Finnigan RAW Input:
1: A1_ATP
Output Type:
mgf
Use Filtering?:
Advanced Settings:
Use Defaults
Execute
What it does
Converts between various mass spectrometry peak list representations. Additional options such as filtering and/or precursor recalculation are available.
You can view the original documentation here.
Citation
For the underlying tool, please cite ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. Darren Kessner; Matt Chambers; Robert Burke; David Agus; Parag Mallick. Bioinformatics 2008; doi: 10.1093/bioinformatics/btn323.
If you use this tool in Galaxy, please cite Chilton J, et al. https://bitbucket.org/galaxyp/galaxyp-toolshed-msconvert
History
PeptideShaker 101
372.7 MB
3: Create Target-Decoy Database on data 2
2: Protein Database
1: A1_ATP
Multifile dataset raw format: m:raw, database: ?
Example Multiple File Dataset
A1_ATP_1.RAW (File 1)
A1_ATP_2.RAW (File 2)
A1_ATP_3.RAW (File 3)
https://usegalaxyp.org/tool_runner?tool_id=msconvert3_raw


Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Galaxy
https://usegalaxyp.org/root
pride raw files
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 1.0 GB
Tools
search tools
GALAXY CORE
Get Data
Send Data
Multiple File Datasets
Text Manipulation
Filter and Sort
Join, Subtract and Group
Statistics
Graph/Display Data
Multiple regression
Multivariate Analysis
Motif Tools
Multiple Alignments
Metagenomic analyses
Phenotype Association
FASTA manipulation
PROTEOMICS: GALAXY-P
Peak List Processing
Statistical Validation
Quantitation
MaxQuant
Bumbershoot Tools
Metaproteomics
Proteogenomics
Get Data
PROTEOMICS: COMMUNITY
Peptide Shaker
Peptide Shaker Peform protein identification combining X! Tandem and OMSSA (using SearchGUI) and PeptideShaker pipeline.
Create Target-Decoy Database Creates a target-decoy database for use with Peptide Shaker
mzMatch
Open MS
ProtK
Utilities
Visualization
FASTA Manipulation
Peptide Shaker (version 0.1.0)
Protein Database:
3: Create Target-Dec..e on data 2
Input Peak Lists (mgf):
4: msconvert3_raw on data 1
select files directly | by multifile
Precursor Ion Tolerance Units:
Parts per million (ppm)
Percursor Ion Tolerance:
10.0
Fragment Tolerance (Daltons):
0.5
Enzyme:
Trypsin
Maximum Missed Cleavages:
2
Fixed Modification:
carbamidomethyl c
Variable Modification:
oxidation of m
gtp desthiobiotinc12
Minimum Charge:
2
Maximum Charge:
4
Forward Ion:
b
Reverse Ion:
y
Specify Advanced Search Options:
Specify Advanced PeptideShaker Processing Options:
Specify Advanced PeptideShaker Filtering Options:
Execute
History
PeptideShaker 101
664.1 MB
4: msconvert3_raw on data 1
3: Create Target-Decoy Database on data 2
2: Protein Database
1: A1_ATP
Multifile dataset raw format: m:raw, database: ?
Example Multiple File Dataset
A1_ATP_1.RAW (File 1)
A1_ATP_2.RAW (File 2)
A1_ATP_3.RAW (File 3)
https://usegalaxyp.org/tool_runner?tool_id=peptide_shaker


Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Galaxy
https://usegalaxyp.org/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 1.1 GB
Spectrum File
Identification File(s)
Precursor RT
Precursor mz
A1_ATP.16058.16058. A1_ATP_1 (File 1).omx 4779.4458 964.13
A1_ATP.8510.8510.1 A1_ATP_1 (File 1).omx 2403.5502 791.40087890625
A1_ATP.8366.8366. A1_ATP_1 (File 1).omx 2357.8309 693.7
A1_ATP.17278.17278. A1_ATP_1 (File 1).omx A1_ATP_1 (File 1).t.xml 5155.1521 950.45
A1_ATP.1993.1993. A1_ATP_1 (File 1).omx 424.4967 981.53
A1_ATP.10870.10870. A1_ATP_1 (File 1).omx 3146.9381 920.13
A1_ATP.6633.6633. A1_ATP_1 (File 1).omx 1820.4371 741.93
A1_ATP.7348.7348. A1_ATP_1 (File 1).omx 2041.1308 1159.9
A1_ATP.7331.7331.1 A1_ATP_1 (File 1).omx 2036.1117 719.6
A1_ATP.22984.22984.1 A1_ATP_1 (File 1).omx 6769.8229 587.56
A1_ATP.12992.12992. A1_ATP_1 (File 1).omx 3821.3314 977.91
A1_ATP.11403.11403. A1_ATP_1 (File 1).omx 3316.1163 993.13
A1_ATP.10047.10047. A1_ATP_1 (File 1).omx 2885.758 957.52
A1_ATP.12494.12494. A1_ATP_1 (File 1).omx 3663.9835 726.89
A1_ATP.14147.14147.1 A1_ATP_1 (File 1).omx 4186.682 518.99
A1_ATP.8474.8474. A1_ATP_1 (File 1).omx 2392.1006 1045.87
A1_ATP.4615.4615. A1_ATP_1 (File 1).omx 1225.8618 404.85
A1_ATP.11351.11351. A1_ATP_1 (File 1).omx 3299.3848 918.19
A1_ATP.4388.4388. A1_ATP_1 (File 1).omx 1158.5862 523.07
A1_ATP.11429.11429. A1_ATP_1 (File 1).omx 3324.5418 1041.6
A1_ATP.4579.4579. A1_ATP_1 (File 1).omx 1215.2906 472.54
A1_ATP.19966.19966. A1_ATP_1 (File 1).omx 5971.9205 639.66
A1_ATP.6190.6190. A1_ATP_1 (File 1).omx 1685.705 744.05
A1_ATP.13359.13359. A1_ATP_1 (File 1).omx 3938.0828 1043.25
A1_ATP.11163.11163. A1_ATP_1 (File 1).t.xml 3239.4479 666.84
A1_ATP.8033.8033. A1_ATP_1 (File 1).omx 2253.6402 861.45
A1_ATP.5468.5468.1 A1_ATP_1 (File 1).omx 1472.6248 800.39
A1_ATP.8966.8966. A1_ATP_1 (File 1).omx 2546.2684 976.17
A1_ATP.22172.22172.1 A1_ATP_1 (File 1).omx 6573.0315 727.456237792969
A1_ATP.15645.15645. A1_ATP_1 (File 1).omx 4653.0733 776.43
A1_ATP.15020.15020. A1_ATP_1 (File 1).omx 4459.4024 1039.87
A1_ATP.9314.9314.1 A1_ATP_1 (File 1).omx 2655.4786 824.43
A1_ATP.4952.4952. A1_ATP_1 (File 1).omx 1324.9834 571.06
A1_ATP.7119.7119. A1_ATP_1 (File 1).omx 1969.2179 705.05
A1_ATP.5902.5902. A1_ATP_1 (File 1).omx 1599.1762 737.87
A1_ATP.7135.7135. A1_ATP_1 (File 1).t.xml 1974.2127 880.49
A1_ATP.9994.9994. A1_ATP_1 (File 1).omx 2868.7206 843.09
A1_ATP.13544.13544.1 A1_ATP_1 (File 1).omx 3995.9774 324.43
A1_ATP.8606.8606. A1_ATP_1 (File 1).omx 2433.4604 728.79
A1_ATP.21233.21233. A1_ATP_1 (File 1).omx 6342.1764 867.65
A1_ATP.12825.12825. A1_ATP_1 (File 1).t.xml 3768.0869 797.47
A1_ATP.6806.6806. A1_ATP_1 (File 1).omx 1873.5239 974.55
A1_ATP.10807.10807.1 A1_ATP_1 (File 1).omx 3126.8176 991.54
A1_ATP.11264.11264. A1_ATP_1 (File 1).omx 3271.3843 1049.54
A1_ATP.9322.9322. A1_ATP_1 (File 1).omx 2657.9808 903.1
A1_ATP.14312.14312. A1_ATP_1 (File 1).omx 4239.0255 831.68
A1_ATP.4477.4477. A1_ATP_1 (File 1).omx 1184.9677 343.44
History
PeptideShaker 101
764.5 MB
8: PeptideShaker PSM Report for data 3 and data 4
7: PeptideShaker Protein Report for data 3 and data 4
6: PeptideShaker Peptide Report for data 3 and data 4
5: PeptideShaker CPS results for data 3 and data 4
96.9 MB
format: cps, database: ?
makeblastdb command: /opt/galaxy/tools/searchgui/1.13.1/share/java/SearchGUI-1.13.1/resources/makeblastdb/linux/makeblastdb -in dataset_7.dat Reindexing. Wed May 29 20:33:18 UTC 2013 Formatting dataset_7.dat for OMSSA. Building Building a a new new
binary data
4: msconvert3_raw on data 1
3: Create Target-Decoy Database on data 2
8,196 sequences
format: fasta, database: ?
>tr|Q50655|Q50655_MYCTU Uncharacteri
MSTRQAAEADLAGKAAQYRPDELARYAQRVMDWLHE
MSRLSGYLTPQARATFEAVLAKLAAPGATNPDDHTE
GLLAGLRALIASGKLGQHNGLPVSIVVTTTLTDLQT
AHHYSPASGRYPQAIFDHGTPLALYHTKRLASPAQF
HHVTAWTSTGRTDITELTLACGPDNRLAEKGWTTHN
2: Protein Database
1: A1_ATP
Multifile dataset raw
format: m:raw, database: ?

CHAPTER 6

Indices and tables

- genindex
- modindex
- search