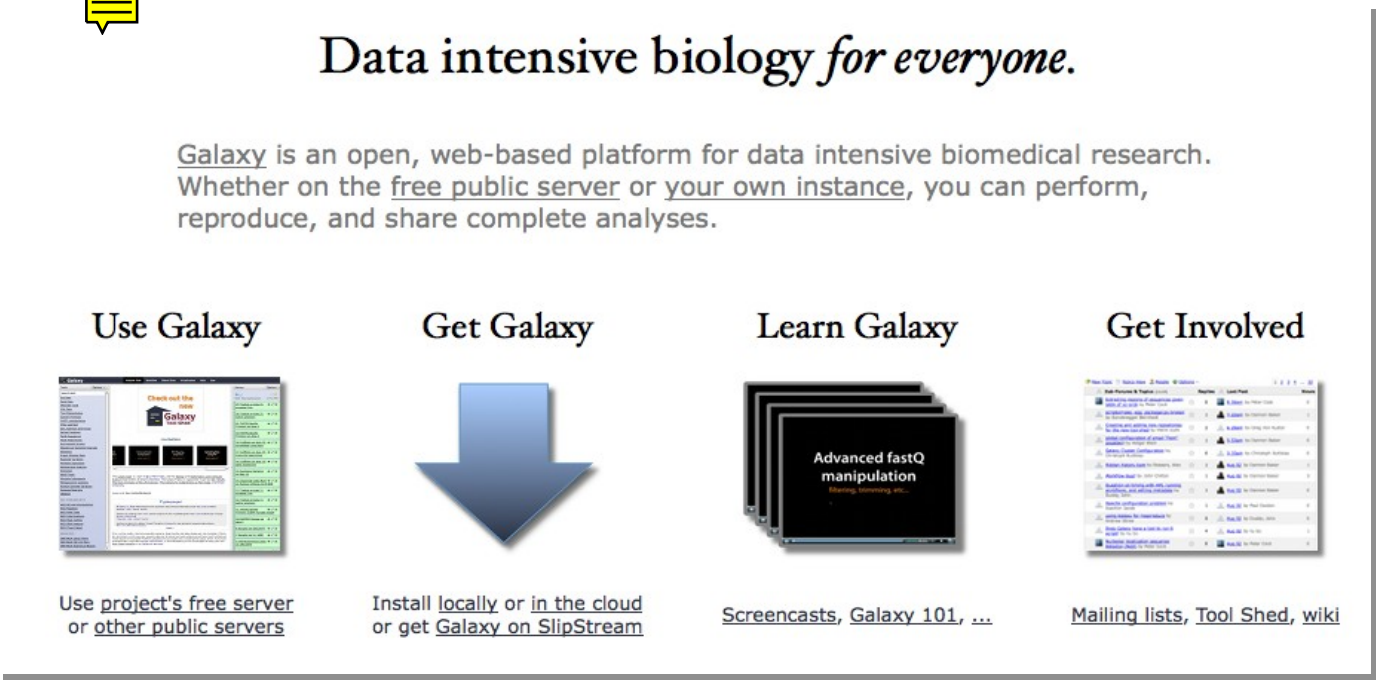
Galaxy

http://galaxyproject.org/



The Galaxy Project we page contains several sections:

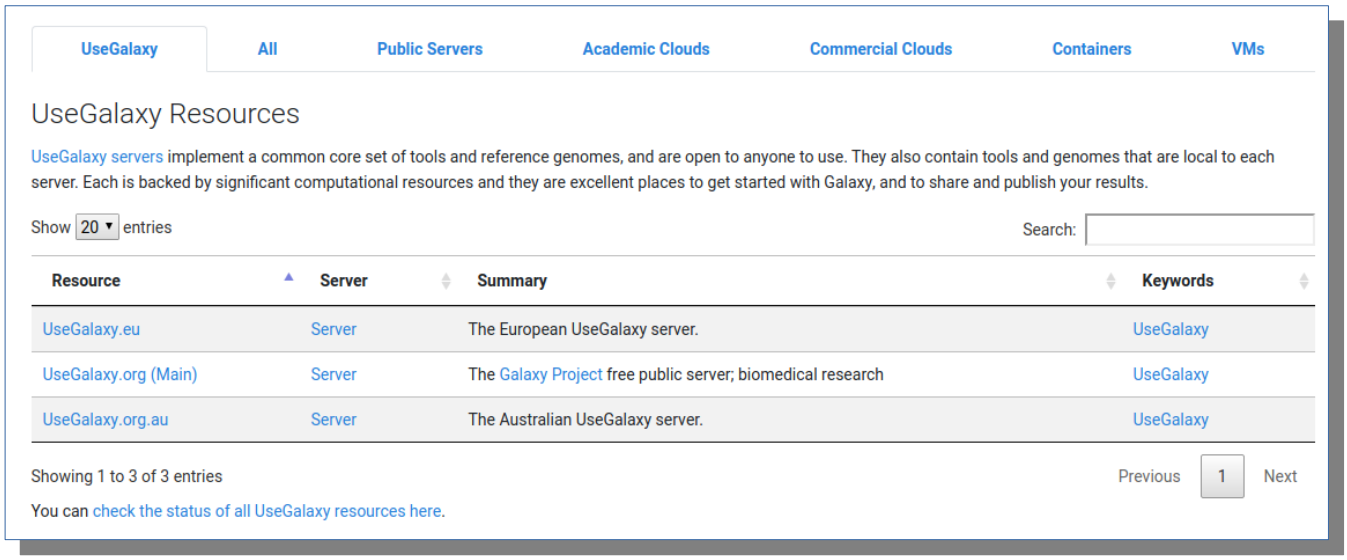
Use Galaxy: contains a list of Galaxy platform in the world in which you can run different analysis pipelines. In this tutorial we will use usegalaxy.org, hosted by the galaxy project itself.

Get Galaxy: Intructions to install a Galaxy platform on a server, on your own computer or in the cloud.

Learn Galaxy: Tutorials of the different aspects of Galaxy, from normal usage to administration.

Get Involved: How to contribute to this Open Source project

Use Galaxy Section



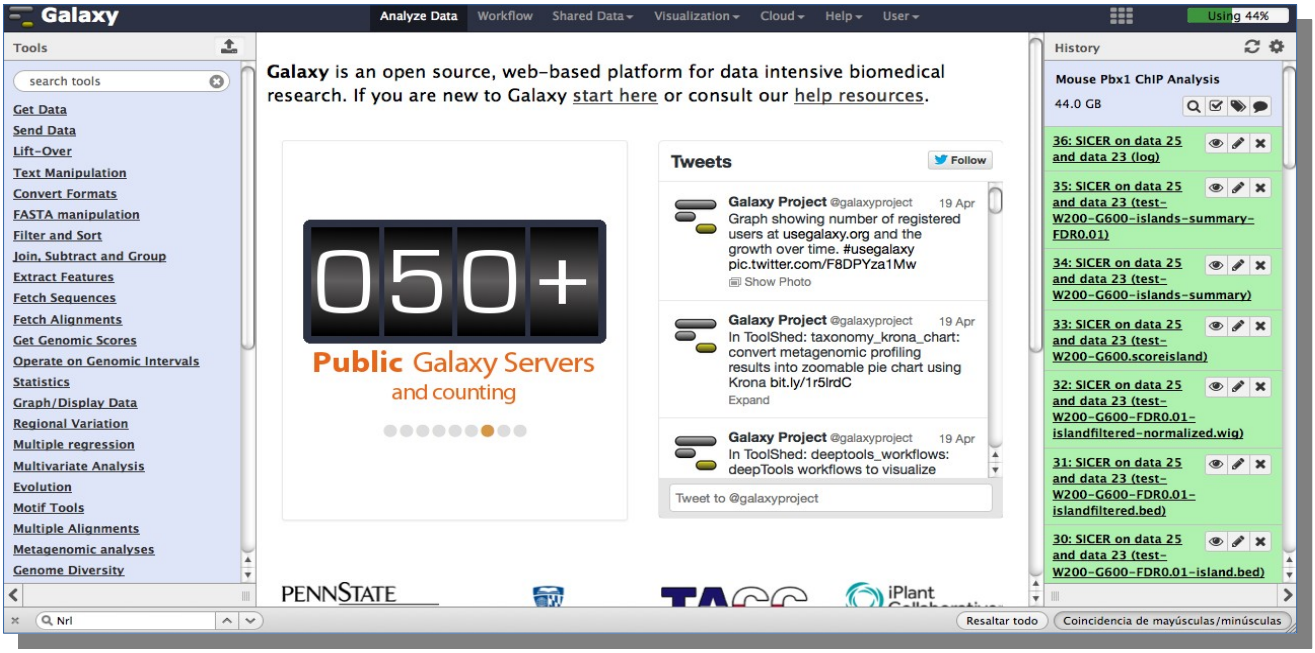
Here we can find many dedicated Galaxy platforms to perform many data analysis.

Use Galaxy: Presents 3 servers dedicated to learn and perform small simple analysis.

All: In this tab you can find servers provided by different private and public institutions where you can register and run specific analysis.

Try to find here servers dedicated to proteomics analysis

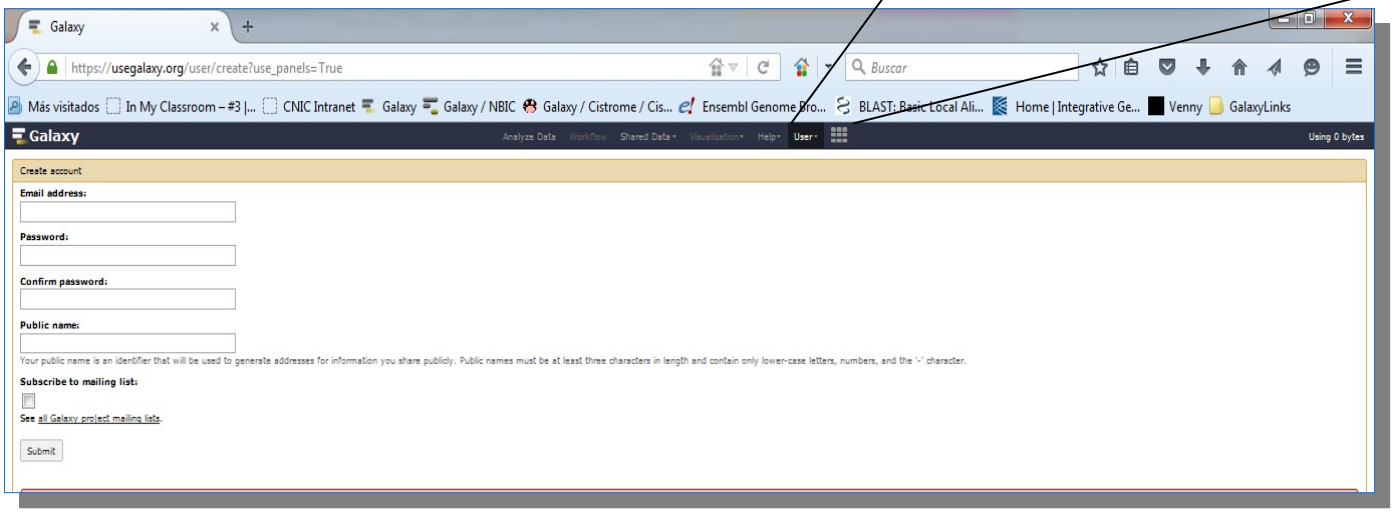
http://usegalaxy.org Or http://usegalaxy.eu



In any of these servers you can setup an account and perform small analysis

Create A Galaxy Account

Create an Account using your gmail or any other email. Remember to confirm registration via your email

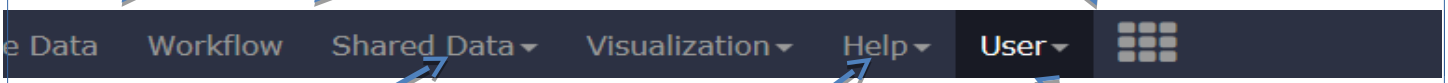


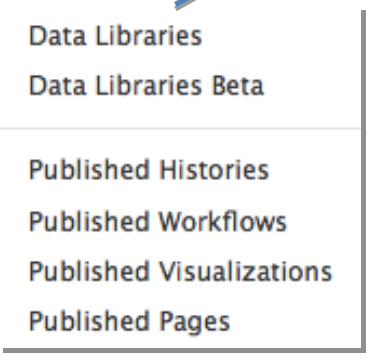
Main Menu:

Change to Analysis Window

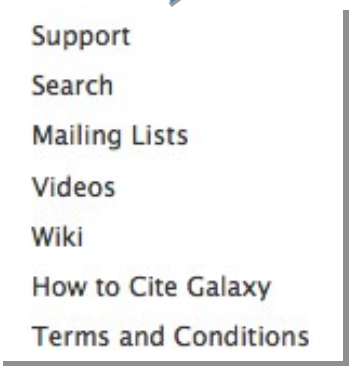
Change to Workflow Design

Quota Usage

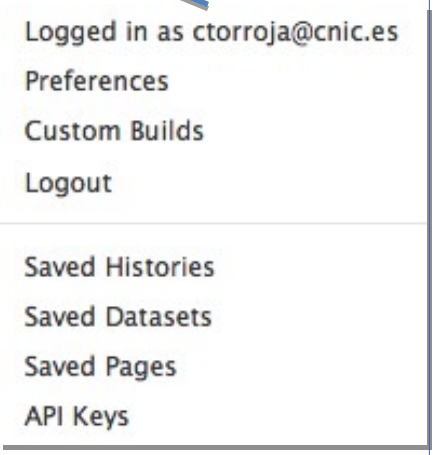




Explore shared Datasets, Histories and Workflows

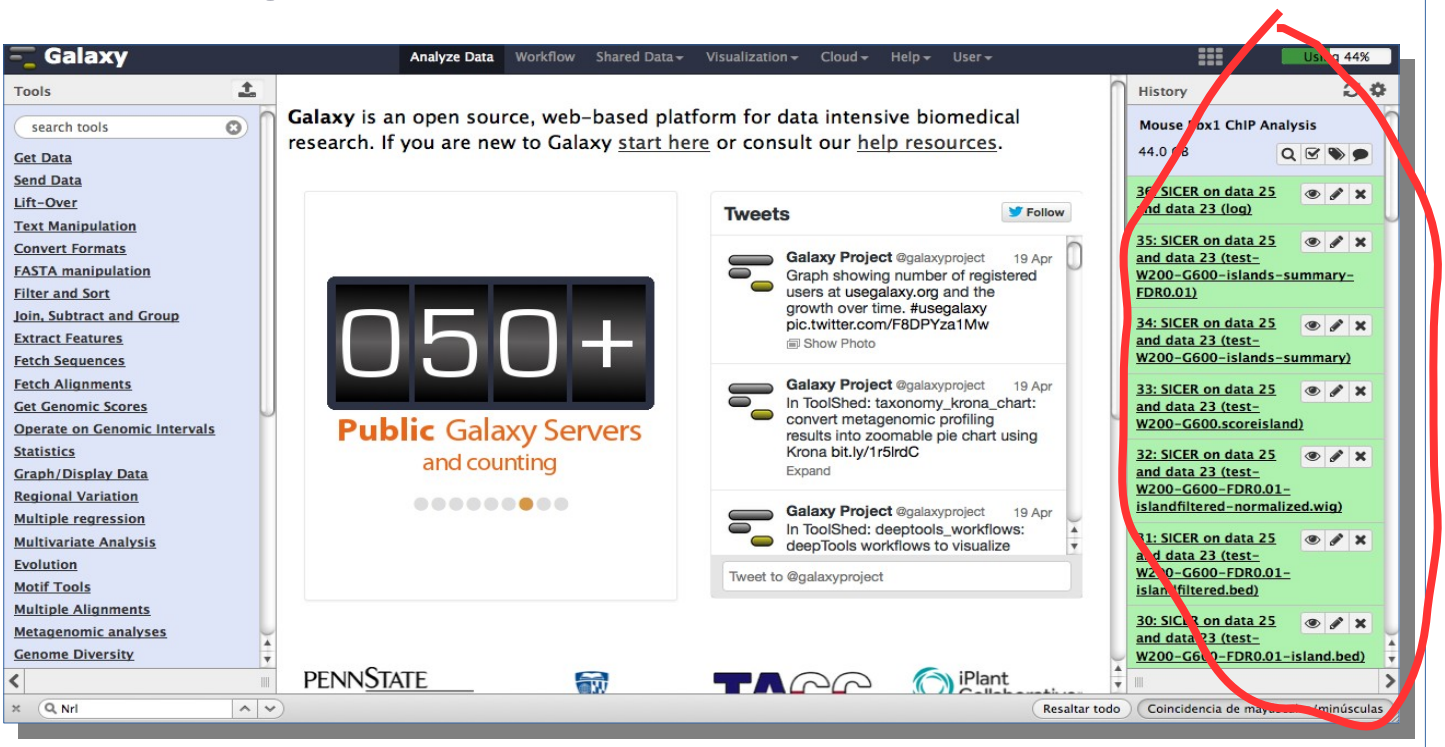


Get Help, Tutorials,...

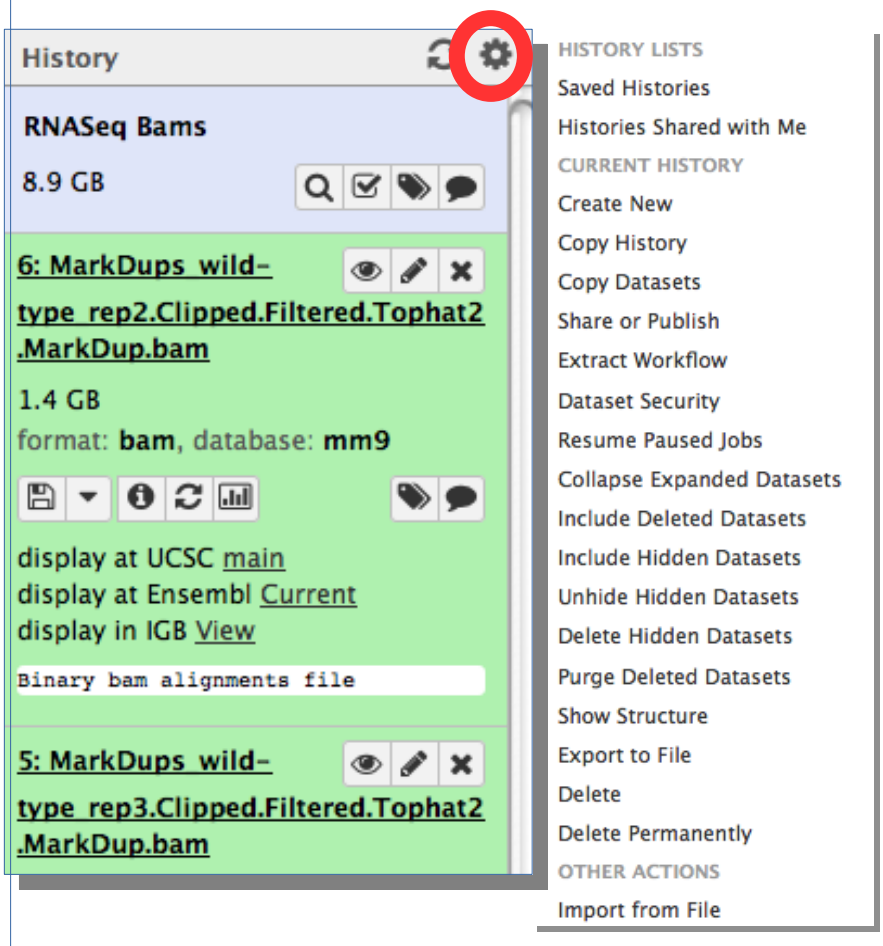


Your profile, Preferences

History Panel



History Settings



Display the list of Datasets generated dunring the analysis.

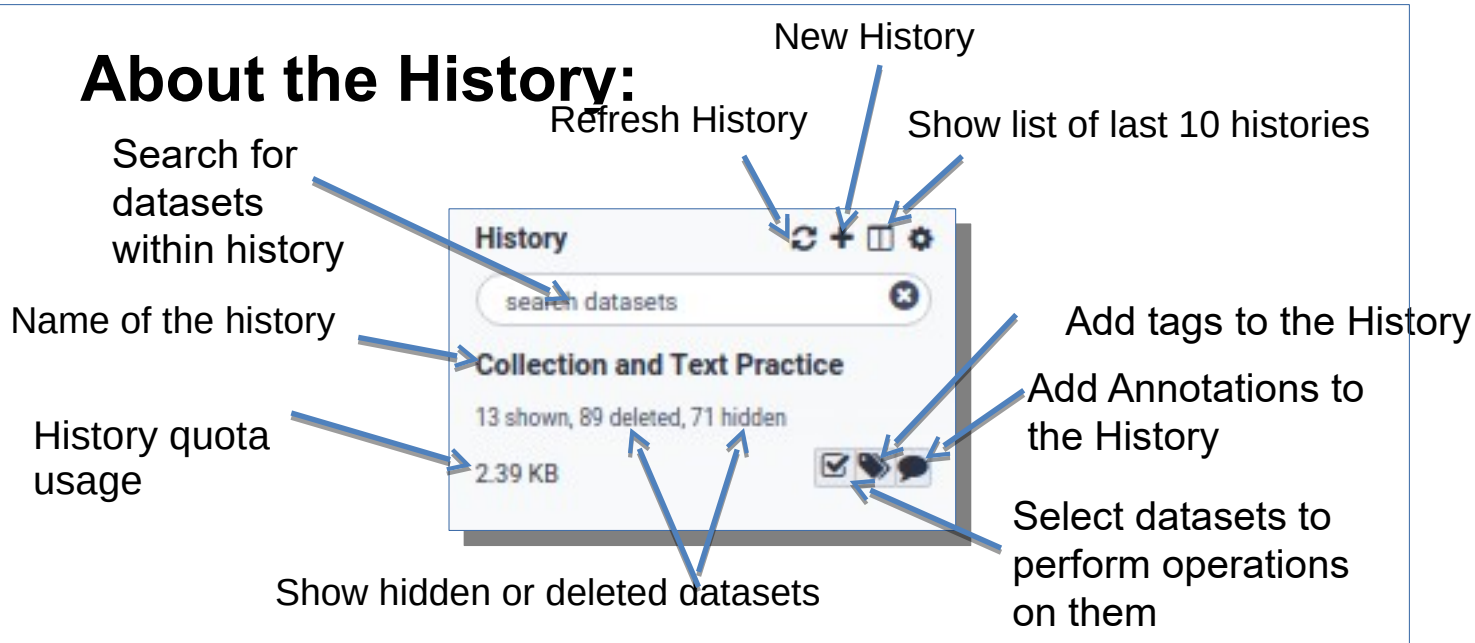
Datasets can be renamed, annotated, deleted, hidden, undeleted, downloaded, visualized, etc.

Progress of the analysis can be also followed.

Full history can be shared with other users.

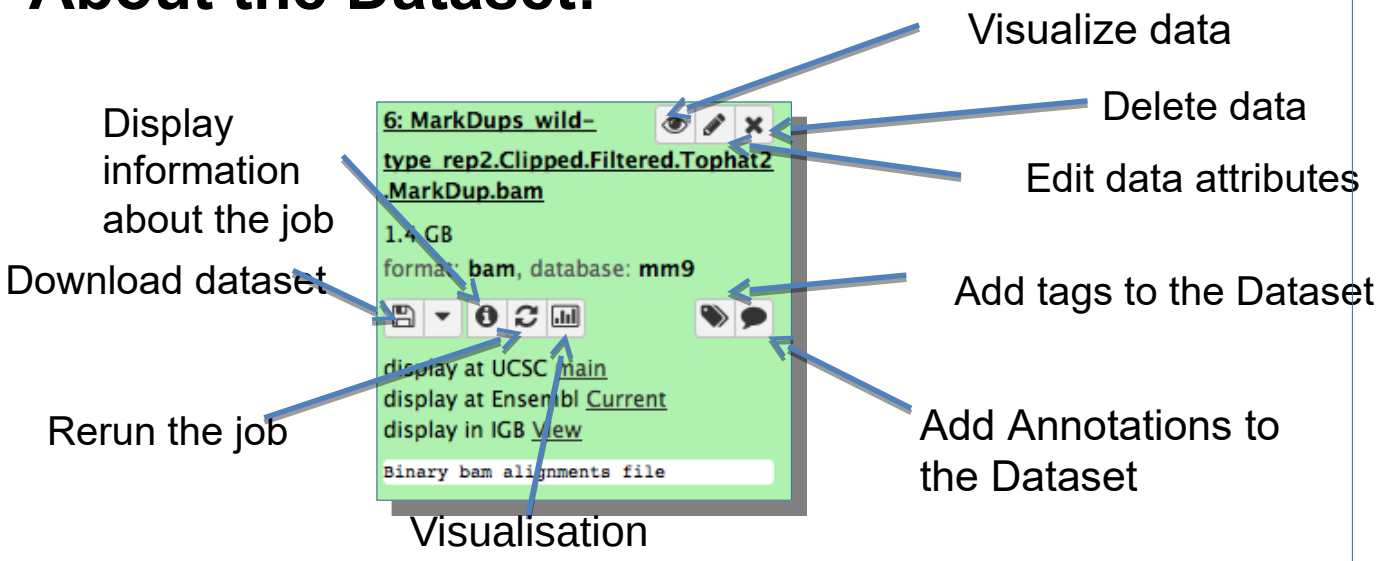
Extract a workflow from the history

About the History:



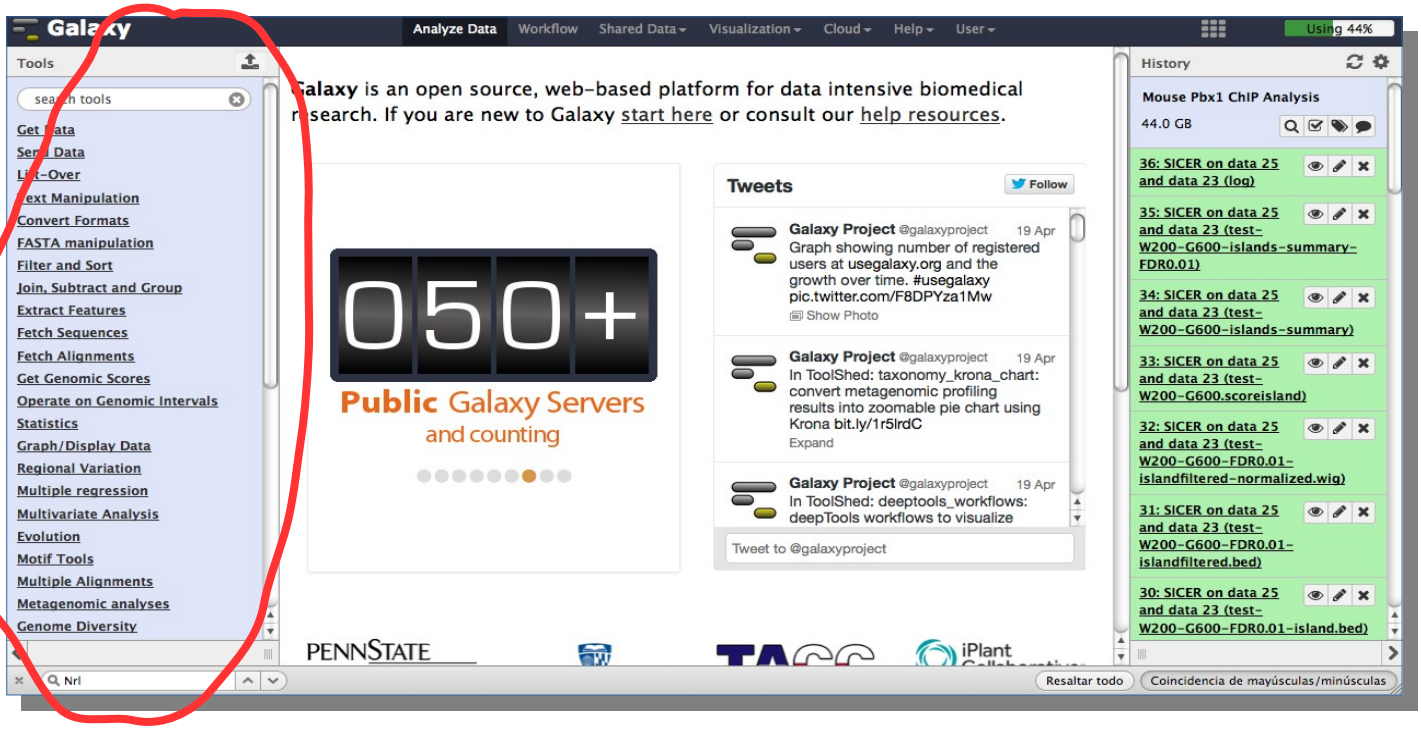
Name: Can be change by clicking directly on it
Tags: One word annotations to classify my analysis histories to help me later find easily histories.
Annotations: A small description of the Analysis
Select: Allows to select several datasets of my history to perform an operation on them.
Search: Allows to search for datasets in my history

About the Dataset:



Visualize Data: Shows the content of the dataset in the central working panel whenever is a text file. If binary, there are specific tools too explore the data.
Delete data: The dataset is hidden and labeled to be deleted. It could be recovered unless the "Purge Deleted Datasets" option is selected in the "Options Menu" of the History.
Edit Attributes: Allows to change the name of the dataset, change it's format type and describe it. (Annotations also accesible through "Tags" and "Annotation" icons).
Download Dataset: Allows to download directly the dataset to your computer.
Information: Will display information of the status of the job sent to the cluster, it's parametrization and command.
Rerun job: In some cases there may be the need to rerun the job.
Visualisation: It presents some ways to visualise the data. Also present as a list below.

Tools Panel



Contains all the tools available on the Galaxy instance grouped by categories.

The tools can be search for using the search box. The search will look for the name, but also in the description of the tool.

When a category is selected it expands and shows all the tools in it.

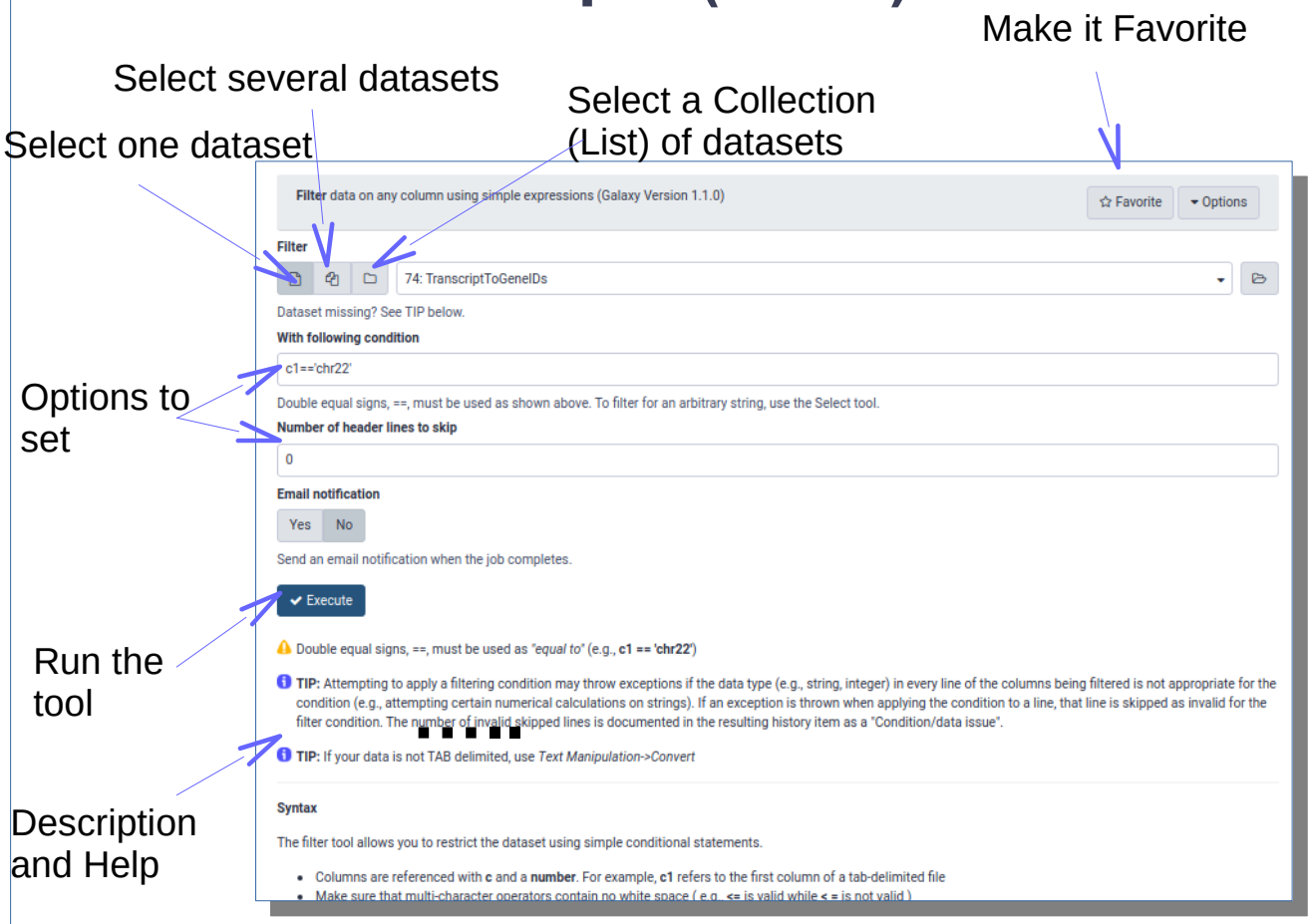
If a tool is selected, a form will be displayed in the central working panel.

The form will ask for the inputs and settings this tool requires.

At the bottom of the form there will be a description of what it does and some examples of how to use it. If not, there might be a link to tha publication where this tool was created.

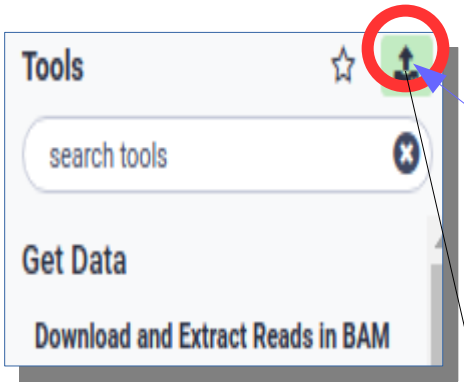
See an example in the following page...

Tool Form Example (Filter)

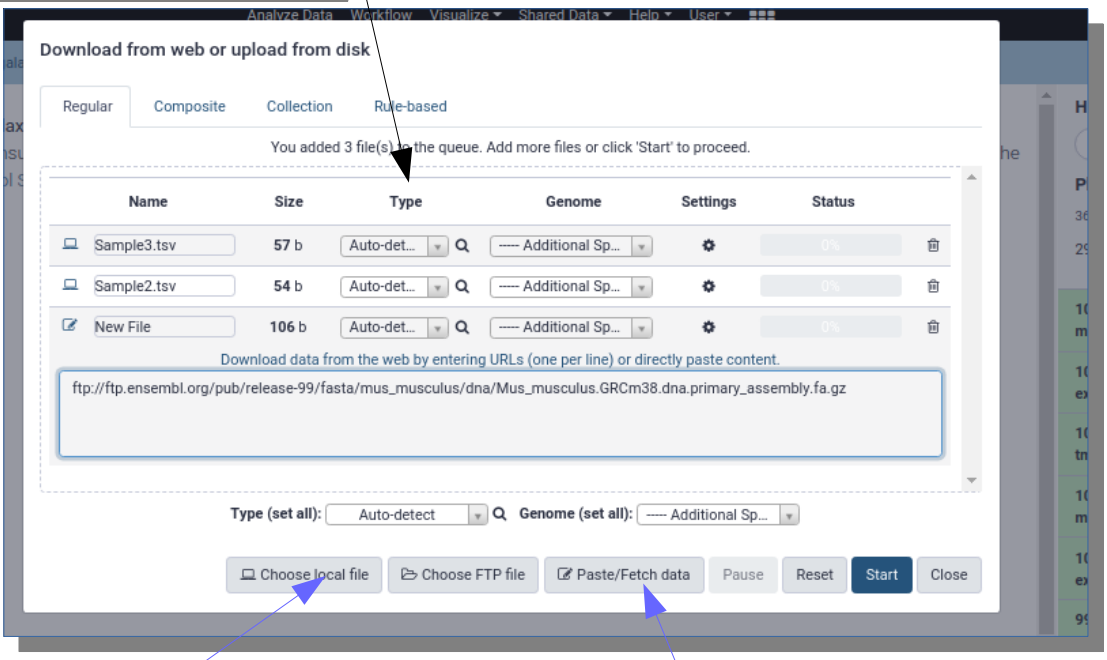


Many tools have some of the settings with a default which in general is the best guess of the person who created it. Leave it as it is unless you understand what that option implies and need to change it.

Upload Data to Galaxy:



This tool allows you to upload files from your computer or transfer files from other servers vía a URL address.



Use this option to select files from your computer to upload.

Use this option to transfer data from another server using an URL link.

You can select several files and add several URL links at once and then click Start. When the datasets are uploaded they will appear as green in your History.

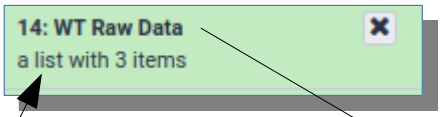
There are specific tools under “Get Data” category to download datasets from specific online databases like:

- **Faster Download and Extract Reads in FASTQ** format from NCBI SRA. To download fastq files from experiments stored in Sequence Read Archive repository.
- **UCSC:** Will let you get genome related data from UCSC Genome browser.
- **....**

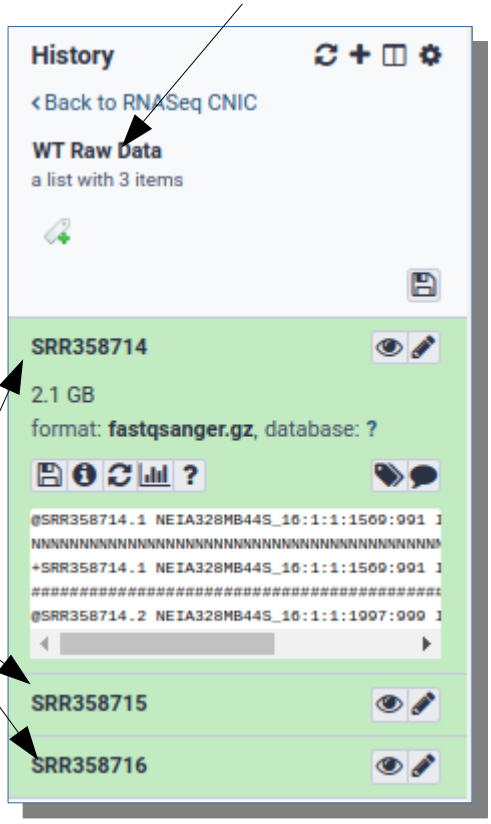
Collection of Datasets

The name of the collection can be changed here

Click on the name to expand it's content



The term "list" appear in the entry information to indicate this is a collection.

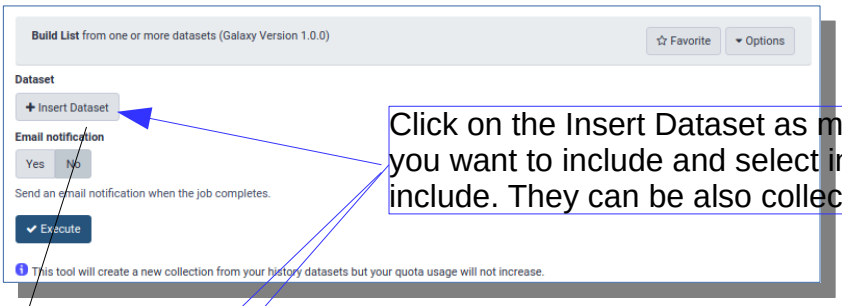


This collection has 3 datasets

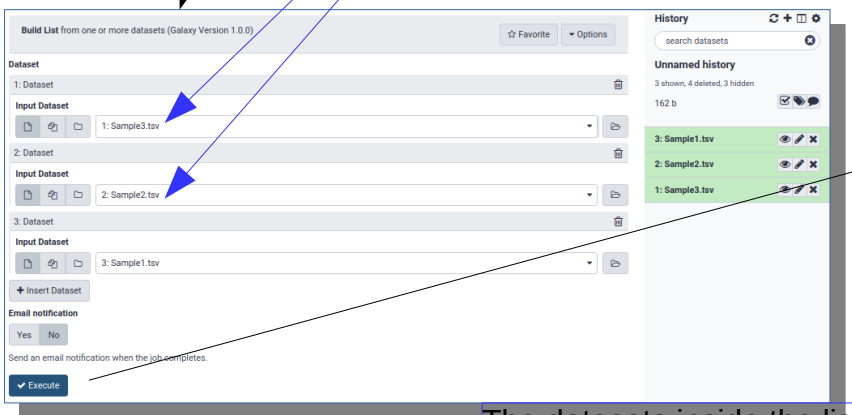
- A collection is a list of datasets that will be all of the same format (all fastq files or all tabular files …).
- A collection will have datasets that we want to process in the same way.
- The collection allows us to maintain a tidy history.
- The collection allows us to reduce the number of processes we want to perform on them. We will need to setup a tool with all it's parameters once per all datasets in it. Otherwise we will need to do the same many times, one for each dataset.
- The tools applied to collections also output collections.
- **Collection Operations:** There is a special set of tools in the Tool Panel to manage collections.
- **Nested Collections:** There may be several collections inside a collection.
- **Pair Collection:** It is a special type of collection to store a pair of files belonging to the same sample. Like the two files of a pair-end sequencing, one with the 1st read pair and the other for the 2nd read pair.
    - There are tools that need this Pair Collection to manage properly the pair-end sequencing.

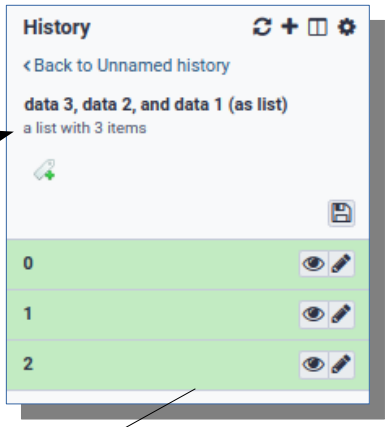
Collection Operations:

Build List: will create a collection from the datasets selected



Click on the Insert Dataset as many times as datasets you want to include and select in each box the ones to include. They can be also collections.

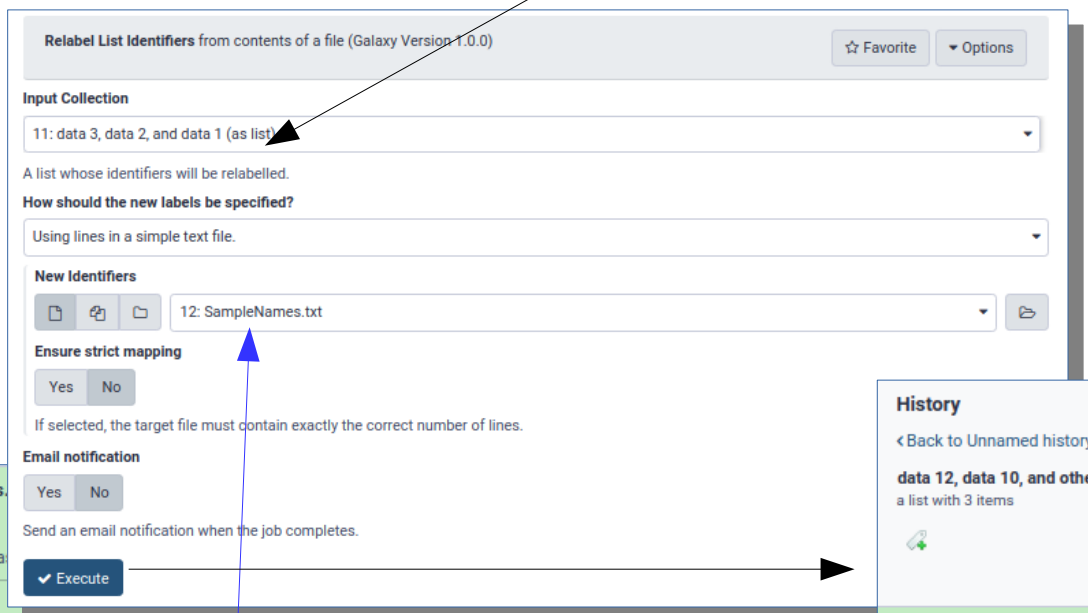


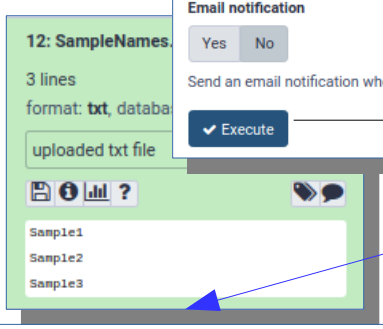


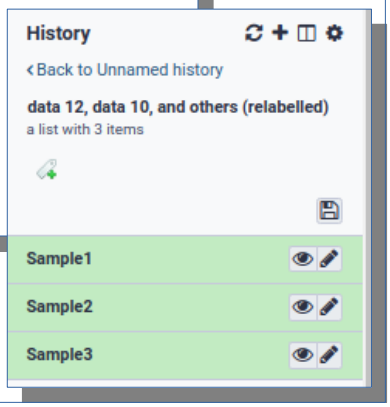
The datasets inside the list are labelled with consecutive numbers

We need to relable the datasets to keep good track

Relabel List Identifiers: Will change the name of the datasets inside the collection



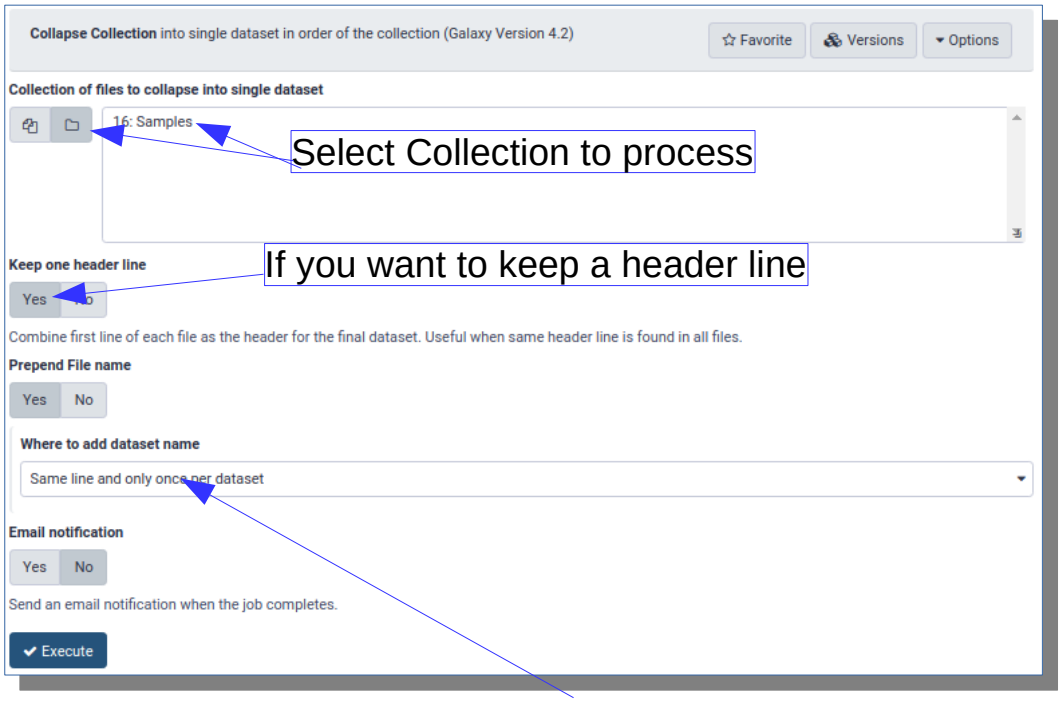




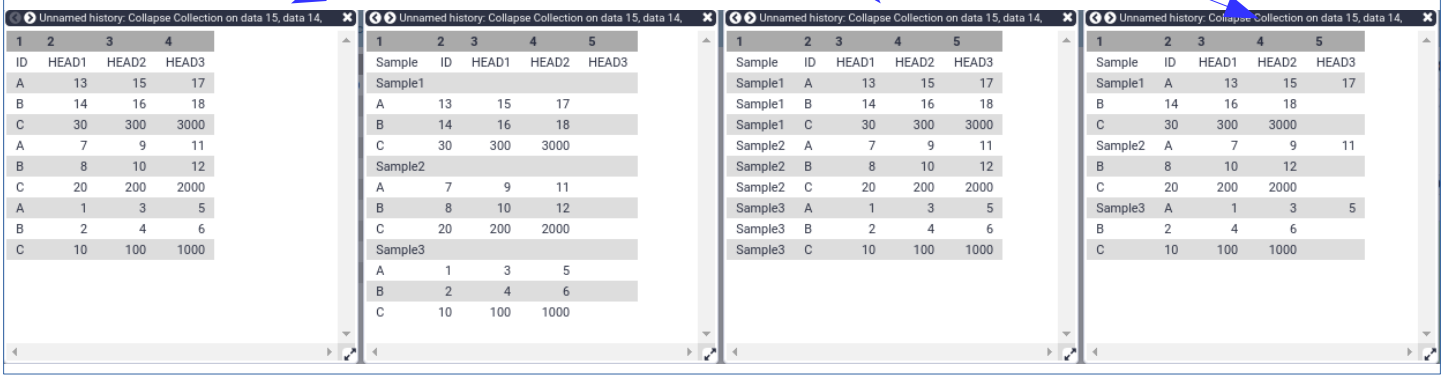
Upload a text file with one name per line in the same order as they appear in the collection

Collection Operations:

Collapse Collection into single dataset in order of the collection



Some options on how to identify each line afterwards. See the different results you may get bellow



Unnamed history: Collapse Collection on data 15, data 14,

1	2	3	4
ID	HEAD1	HEAD2	HEAD3
A	13	15	17
B	14	16	18
C	30	300	3000
A	7	9	11
B	8	10	12
C	20	200	2000
A	1	3	5
B	2	4	6
C	10	100	1000

Unnamed history: Collapse Collection on data 15, data 14,

1	2	3	4	5
Sample	ID	HEAD1	HEAD2	HEAD3
Sample1				
A	13	15	17	
B	14	16	18	
C	30	300	3000	
Sample2				
A	7	9	11	
B	8	10	12	
C	20	200	2000	
Sample3				
A	1	3	5	
B	2	4	6	
C	10	100	1000	

Unnamed history: Collapse Collection on data 15, data 14,

1	2	3	4	5
Sample	ID	HEAD1	HEAD2	HEAD3
Sample1	A	13	15	17
Sample1	B	14	16	18
Sample1	C	30	300	3000
Sample2	A	7	9	11
Sample2	B	8	10	12
Sample2	C	20	200	2000
Sample3	A	1	3	5
Sample3	B	2	4	6
Sample3	C	10	100	1000

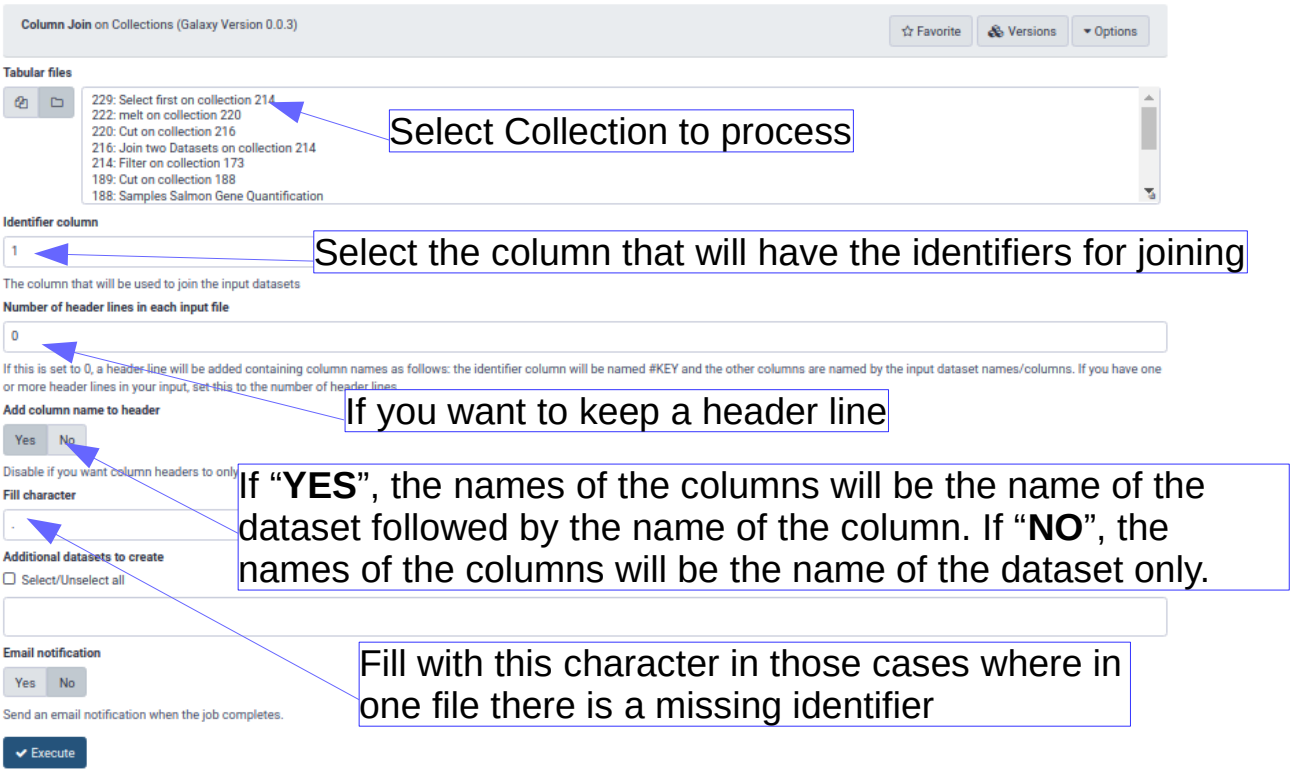
Unnamed history: Collapse Collection on data 15, data 14,

1	2	3	4	5
Sample	ID	HEAD1	HEAD2	HEAD3
Sample1	A	13	15	17
B	14	16	18	
C	30	300	3000	
Sample2	A	7	9	11
B	8	10	12	
C	20	200	2000	
Sample3	A	1	3	5
B	2	4	6	
C	10	100	1000	

Use Sample files 1 to 5 to play with the different Collection tools.

Collection Operations:

Column Join on Collections



This tool will generate a dataset that will be a tabular file (matrix) joining side by side all the columns present in the datasets of the collection matched through the column of identifiers selected.

If some files miss some identifiers, their columns will be filled with dots or whatever character you chose in the Fill character option.

Examples joining a Collection of 3 Samples and asking to leave the column name. In this case is important because several columns per sample are going to be joined.

1	2	3	4	5	6	7	8	9	10	11	12	13
ID	Sample1_HEAD1	Sample1_HEAD2	Sample1_HEAD3	Sample2_HEAD1	Sample2_HEAD2	Sample2_HEAD3	Sample3_HEAD1	Sample3_HEAD2	Sample3_HEAD3	Sample4_HEAD1	Sample4_HEAD2	Sample4_HEAD3
A	1	3	5	7	9	11	13	15	17	19	21	23
B	2	4	6	8	10	12	14	16	18	20	22	24
C	10	100	1000	20	200	2000	30	300	3000	40	400	4000

In this example some identifiers are unique for one sample and the rest of samples the tool fills the columns with NAs

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ID	Sample1_HEAD1	Sample1_HEAD2	Sample1_HEAD3	Sample2_HEAD1	Sample2_HEAD2	Sample2_HEAD3	Sample3_HEAD1	Sample3_HEAD2	Sample3_HEAD3	Sample4_HEAD1	Sample4_HEAD2	Sample4_HEAD3	Sample5_HEAD1	Sample5_HEAD2	Sample5_HEAD3
A	13	15	17	7	9	11	1	3	5	19	21	23	NA	NA	NA
B	14	16	18	8	10	12	2	4	6	20	22	24	NA	NA	NA
C	30	300	3000	20	200	2000	1	100	1000	40	400	4000	NA	NA	NA
D	NA	NA	NA	NA	NA	NA	N	NA	NA	NA	NA	NA	25	27	29
E	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	26	28	30
F	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	50	500	5000

Text Manipulation Tools:

Help to manipulate text files and transform them to prepare them for the following operation.

- **Cut**: Extracts a column from the file
- **Sort**: Sort the rows by specified column. Can remove duplicates
- **Unique**: Remove duplicated rows if they have been previously sorted
- **Concatenate**: Paste the rows of one file after the other
- **Join**: Join the columns of two files. Will sort them if they are not in the same order. Has several options when some identifiers do not match between columns.
- **Multi-Join:** Same for several values at the same time. It has some unexpected behaviour if files are not sorted first. Avoid.
- **Select First:** Extract several rows from the beginning of the file
- **Select Last**: Extract several rows form the end or starting from the specified row. Very useful
- **Add Column**: To add a column with a constant value to the file.
- Merge Column: Merge the content of two column into one.
- **Replace**: Replace some text from each row. Can be done on specific columns. It uses Regular expressions if desired.
- **Melt**: Convert a 2D matrix with several rows a columns (all of the same type of value) into a 3 column notation (x,y,value). Useful for several kinds of visualisation plots.
- **...**

For some of these tools the are more of one version with different options of configuration. Use the one fits better to your needs.

- Practice all this tools with the Sample files provided and see how they behave.
- Also, use the gtf file with the gene definitions to practice **Replace** with Regular Expressions

A small tutorial on building Regular Expressions:
https://www.geeksforgeeks.org/write-regular-expressions/

Filter and Sort Tools:

Essentially also tools to manipulate text or tabular files.

- **Filter**: Will help to select the rows based on an operation over one or several columns. ==, >=, >, <=, <, != (not equal). I.e: c1+c2 >= 25.
- **Sort**: Help you sort based on one or more columns
- **Select**: Filter rows based on a Regular Expression over a column.

Practice with the gtf file to filter records

Join, Subtract and Group Tools:

Essentially also tools to manipulate text or tabular files.

- **Substract**: Will remove the rows of one file that match with the rows of the other. It is desgined thinking on files with annotations for genomic regions (like gtf or bed formats), but any text file can be processed.
- **Sort**: Help you sort based on one or more columns.
- **Select**: Filter rows based on a Regular Expression over a column.

Practice with the gtf file to filter records using Select. And then substract it from the original

Remember to read the instructions and examples at the end of each tool form to understand its function.