



<http://galaxyproject.org/>



Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy



Use project's free server or other public servers

Get Galaxy



Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy



Screencasts, Galaxy 101, ...

Get Involved



Mailing lists, Tool Shed, wiki

The Galaxy Project web page contains several sections:

**Use Galaxy:** contains a list of Galaxy platforms in the world in which you can run different analysis pipelines. In this tutorial we will use [usegalaxy.org](http://usegalaxy.org), hosted by the galaxy project itself.

**Get Galaxy:** Instructions 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

[UseGalaxy](#) [All](#) [Public Servers](#) [Academic Clouds](#) [Commercial Clouds](#) [Containers](#) [VMs](#)

## UseGalaxy Resources

[UseGalaxy servers](#) implement a common core set of tools and reference genomes, and are open to anyone to use. They also contain tools and genomes that are local to each server. Each is backed by significant computational resources and they are excellent places to get started with Galaxy, and to share and publish your results.

Show  entries Search:

Resource	Server	Summary	Keywords
<a href="#">UseGalaxy.eu</a>	Server	The European UseGalaxy server.	<a href="#">UseGalaxy</a>
<a href="#">UseGalaxy.org (Main)</a>	Server	The <a href="#">Galaxy Project</a> free public server; biomedical research	<a href="#">UseGalaxy</a>
<a href="#">UseGalaxy.org.au</a>	Server	The Australian UseGalaxy server.	<a href="#">UseGalaxy</a>

Showing 1 to 3 of 3 entries Previous **1** Next

You can [check the status of all UseGalaxy resources here](#).

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

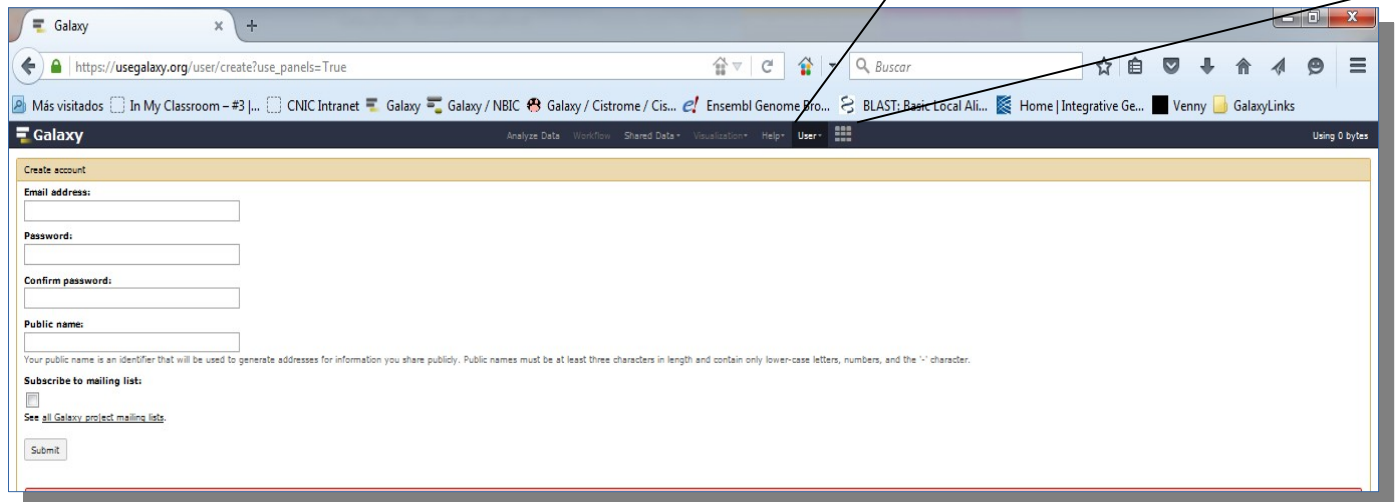
<http://usegalaxy.org> Or <http://usegalaxy.eu>

The screenshot displays the Galaxy web interface. On the left, a 'Tools' sidebar lists various bioinformatics tools like 'Get Data', 'Send Data', 'FASTA manipulation', etc. The main content area features a central banner for 'Public Galaxy Servers and counting' with a '050+' counter. To the right of the banner is a 'Tweets' section showing recent updates from the Galaxy Project. On the far right, a 'History' panel lists recent analyses, such as 'Mouse Pbx1 ChIP Analysis' and 'SICER on data 25 and data 23'. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'.

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

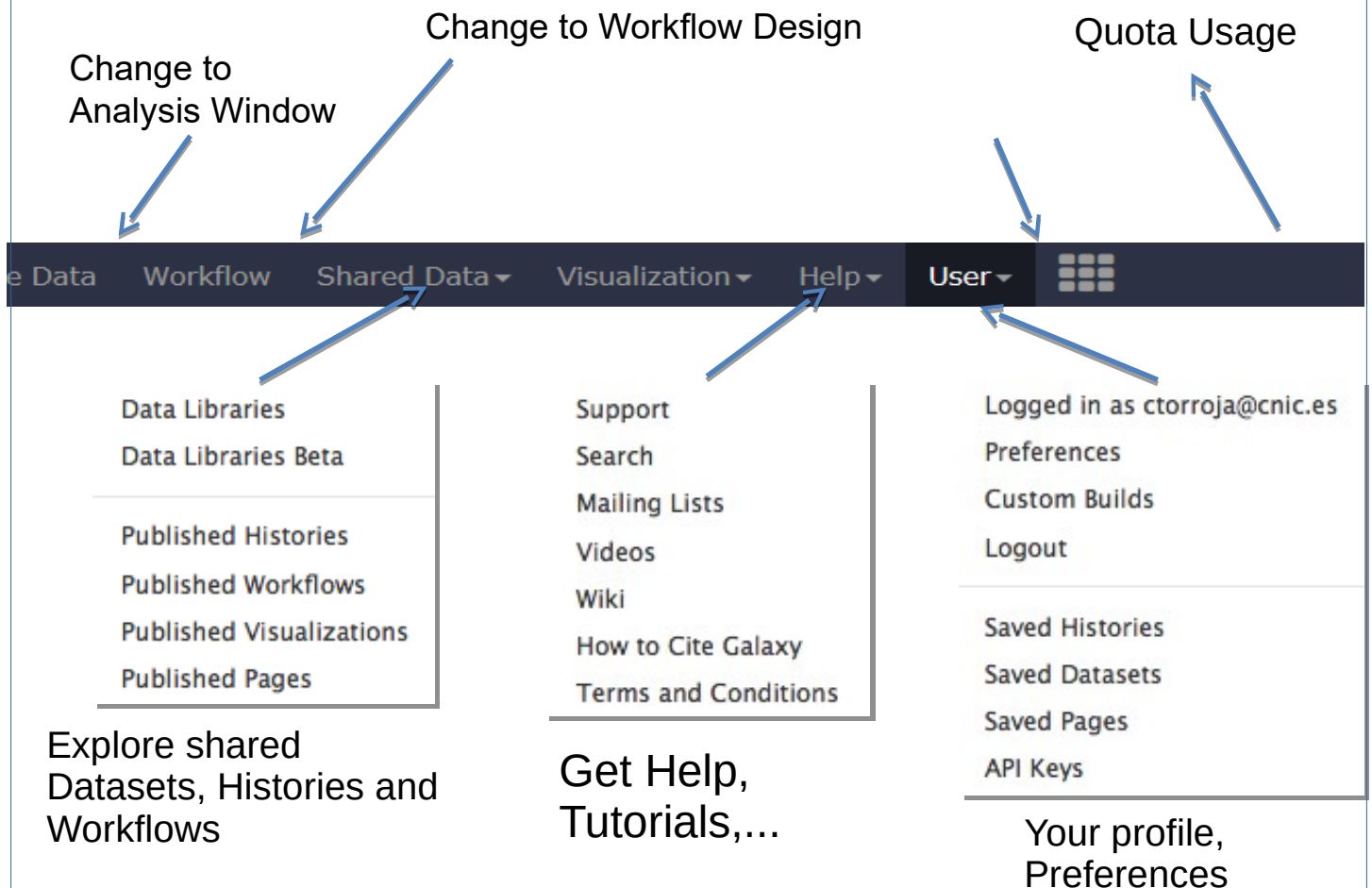
# Create A Galaxy Account

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



The screenshot shows the Galaxy web interface at the URL [https://usegalaxy.org/user/create?use\\_panels=True](https://usegalaxy.org/user/create?use_panels=True). The page has a header with the Galaxy logo and navigation links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The main content area is titled "Create account" and contains the following fields: "Email address:", "Password:", "Confirm password:", and "Public name:". Below these fields is a note: "Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least three characters in length and contain only lower-case letters, numbers, and the '-' character." There is also a checkbox for "Subscribe to mailing list:" with a link "See all Galaxy project mailing lists." and a "Submit" button at the bottom.

## Main Menu:



The diagram illustrates the Galaxy Main Menu with annotations for each section:

- Change to Analysis Window**: Points to the "Workflow" menu item.
- Change to Workflow Design**: Points to the "Shared Data" menu item.
- Quota Usage**: Points to the "User" menu item.

The menu items and their corresponding sub-items are:

- Workflow**:
  - Data Libraries
  - Data Libraries Beta
  - Published Histories
  - Published Workflows
  - Published Visualizations
  - Published Pages

Explore shared Datasets, Histories and Workflows
- Shared Data**:
  - Support
  - Search
  - Mailing Lists
  - Videos
  - Wiki
  - How to Cite Galaxy
  - Terms and Conditions

Get Help, Tutorials,...
- User**:
  - Logged in as [ctorroja@cnic.es](#)
  - Preferences
  - Custom Builds
  - Logout
  - Saved Histories
  - Saved Datasets
  - Saved Pages
  - API Keys

Your profile, Preferences

# History Panel

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. The left sidebar lists various tools like 'Get Data', 'Send Data', 'Lift-Over', etc. The main content area features a 'Public Galaxy Servers and counting' banner with a '050+' counter, a 'Tweets' section, and a description of Galaxy as an open-source platform. The right sidebar shows the 'History' panel, which is circled in red. It lists several datasets, including 'Mouse Fox1 ChIP Analysis' and multiple 'SICER on data 25 and data 23' entries. Each entry has icons for viewing, editing, and deleting.

# History Settings

The screenshot shows the 'History' panel in Galaxy. The top header has a gear icon circled in red. Below the header, the panel is divided into two main sections. The left section, titled 'RNaseq Bams', shows a dataset named '6: MarkDups wild-type rep2.Clipped.Filtered.Tophat2 .MarkDup.bam' with a size of 1.4 GB. It includes options to 'display at UCSC main', 'display at Ensembl Current', and 'display in IGB View'. The right section, titled 'HISTORY LISTS', contains a list of actions: 'Saved Histories', 'Histories Shared with Me', 'CURRENT HISTORY', 'Create New', 'Copy History', 'Copy Datasets', 'Share or Publish', 'Extract Workflow', 'Dataset Security', 'Resume Paused Jobs', 'Collapse Expanded Datasets', 'Include Deleted Datasets', 'Include Hidden Datasets', 'Unhide Hidden Datasets', 'Delete Hidden Datasets', 'Purge Deleted Datasets', 'Show Structure', 'Export to File', 'Delete', 'Delete Permanently', and 'OTHER ACTIONS' with 'Import from File'.

Display the list of Datasets generated during the analysis.

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

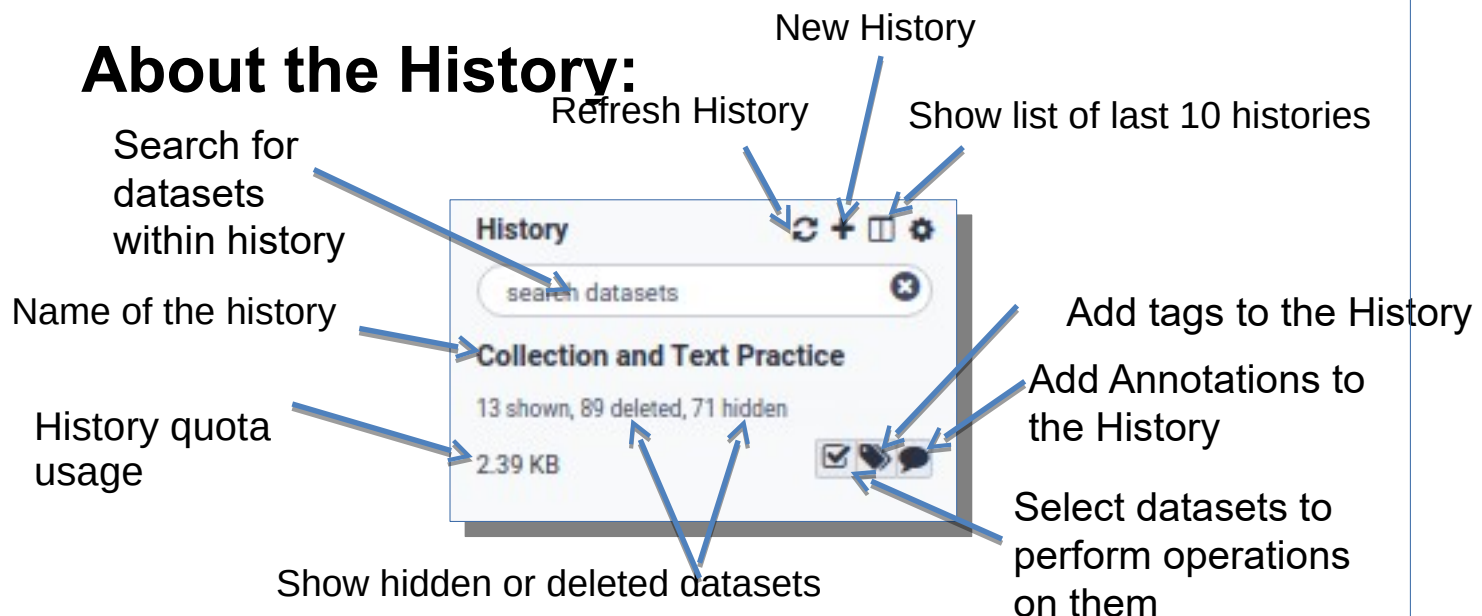
Progress of the analysis can be also followed.

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

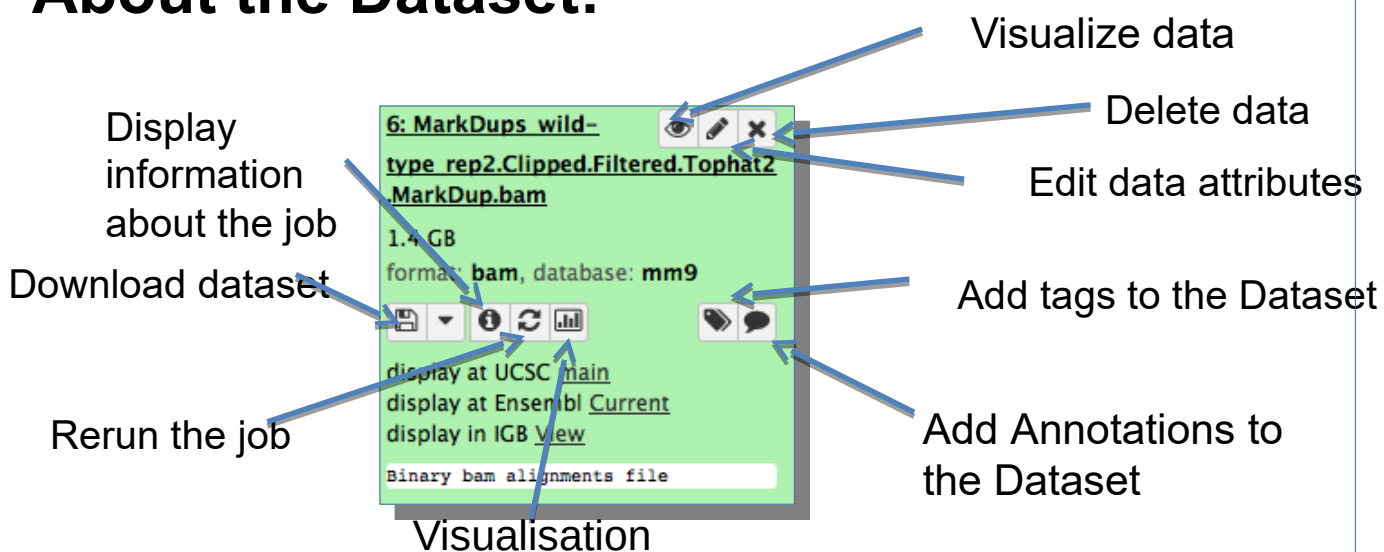
**Tags:** One word annotations to classify my analysis histories to help me later find easily histories.

**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).

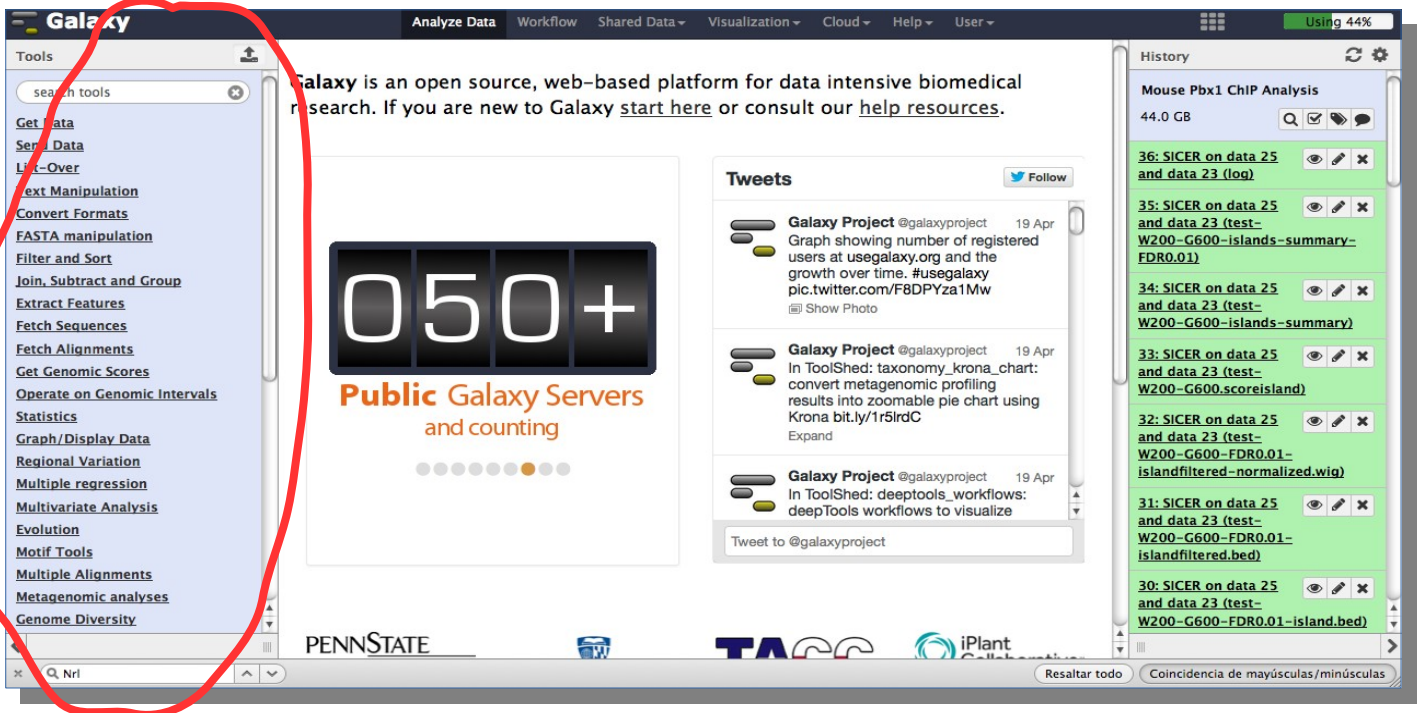
**Download Dataset:** Allows to download directly the dataset to your computer.

**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 the publication where this tool was created.

See an example in the following page...

# Tool Form Example (Filter)

Make it Favorite

Select several datasets

Select a Collection  
(List) of datasets

Select one dataset

Options to  
set

Run the  
tool

Description  
and Help

The screenshot shows the Galaxy 'Filter' tool interface. At the top, a header bar contains the text 'Filter data on any column using simple expressions (Galaxy Version 1.1.0)' and two buttons: '☆ Favorite' and '▼ Options'. Below this, the 'Filter' section has a dropdown menu showing '74: TranscriptToGeneIDs' and a 'Dataset missing? See TIP below.' message. The 'With following condition' section contains a text input field with 'c1=='chr22''. Below this is a note: 'Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.' The 'Number of header lines to skip' section has a text input field with '0'. The 'Email notification' section has 'Yes' and 'No' buttons, with 'Yes' selected. Below this is a 'Send an email notification when the job completes.' checkbox. The 'Execute' button is a blue button with a checkmark and the text 'Execute'. Below the button are two tips: a warning tip about double equal signs and a blue tip about filtering conditions. The 'Syntax' section at the bottom explains the filter tool's purpose and provides examples of column references and operators.

Filter data on any column using simple expressions (Galaxy Version 1.1.0) ☆ Favorite ▼ Options

Filter

74: TranscriptToGeneIDs

Dataset missing? See TIP below.

With following condition

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Email notification

Yes No

Send an email notification when the job completes.

✓ Execute

⚠ Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')

💡 TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

💡 TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

- Columns are referenced with c and a number. For example, c1 refers to the first column of a tab-delimited file
- Make sure that multi-character operators contain no white space (e.g., <= is valid while < = is not valid)

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:

The image shows a screenshot of the Galaxy web interface. In the top left, the 'Tools' menu is open, and a red circle highlights the 'Upload Data' icon. A callout box points to this icon with the text: 'This tool allows you to upload files from your computer or transfer files from other servers via a URL address.'

Below the 'Tools' menu, the 'Get Data' section is visible, with the option 'Download and Extract Reads in BAM' highlighted.

The main part of the image shows the 'Download from web or upload from disk' dialog. It has tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based'. The 'Regular' tab is selected. The dialog shows a table with 3 files added to the queue:

Name	Size	Type	Genome	Settings	Status
Sample3.tsv	57 b	Auto-det...	Additional Sp...	Settings	0%
Sample2.tsv	54 b	Auto-det...	Additional Sp...	Settings	0%
New File	106 b	Auto-det...	Additional Sp...	Settings	0%

Below the table, there is a text box for entering URLs, with the example: `ftp://ftp.ensembl.org/pub/release-99/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz`.

At the bottom, there are buttons for 'Choose local file', 'Choose FTP file', 'Paste/Fetch data', 'Pause', 'Reset', 'Start', and 'Close'.

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:

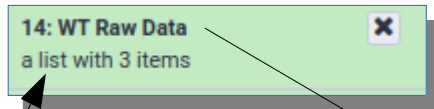
- **Faster Download and Extract Reads in FASTQ** format from NCBI SRA. To download fastq files from experiments stored in Sequence Read Archive repository.
- **UCSC:** Will let you get genome related data from UCSC Genome browser.
- ....



# Collection of Datasets

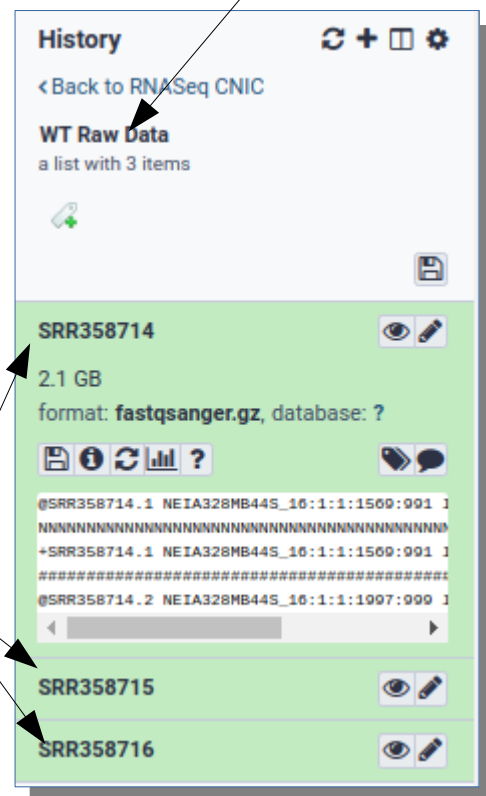
The name of the collection can be changed here

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 1<sup>st</sup> read pair and the other for the 2<sup>nd</sup> 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 relabel 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

**Collapse Collection into single dataset in order of the collection (Galaxy Version 4.2)** ☆ Favorite 🔄 Versions ⌵ Options

Collection of files to collapse into single dataset

16: Samples

Select Collection to process

Keep one header line

Yes No

Combine first line of each file as the header for the final dataset. Useful when same header line is found in all files.

Prepend File name

Yes No

Where to add dataset name

Same line and only once per dataset

Email notification

Yes No

Send an email notification when the job completes.

✓ Execute

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

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

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

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

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



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

# Collection Operations:

## Column Join on Collections

Column Join on Collections (Galaxy Version 0.0.3) ☆ Favorite 🔄 Versions ▼ Options

**Tabular files**

  229: Select first on collection 214  
222: melt on collection 220  
220: Cut on collection 216  
216: Join two Datasets on collection 214  
214: Filter on collection 173  
189: Cut on collection 188  
188: Samples Salmon Gene Quantification

**Identifier column**

1 Select the column that will have the identifiers for joining

The column that will be used to join the input datasets

**Number of header lines in each input file**

0 If you want to keep a header line

If this is set to 0, a header line will be added containing column names as follows: the identifier column will be named #KEY and the other columns are named by the input dataset names/columns. If you have one or more header lines in your input, set this to the number of header lines

**Add column name to header**

☒ Yes ☐ No If "YES", the names of the columns will be the name of the dataset followed by the name of the column. If "NO", the names of the columns will be the name of the dataset only.

Disable if you want column headers to only

**Fill character**

.

**Additional datasets to create**

☐ Select/Unselect all

**Email notification**

☒ Yes ☐ No Send an email notification when the job completes.

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	NA	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 from 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 there are more of one version with different options of configuration. Use the one that fits better to your needs.

- Practice all these 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.  $=$ ,  $>=$ ,  $>$ ,  $<=$ ,  $<$ ,  $\neq$  (not equal). I.e:  $c1+c2 \geq 25$ .
- **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.

- **Subtract:** Will remove the rows of one file that match with the rows of the other. It is designed 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 subtract it from the original

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