



# Introduction to Galaxy

Last updated: Oct 2020

David Morais

# Galaxy

**Data Intensive *analysis* for everyone**

- Versatile and reproducible workflows
- **Web** platform
- **Open source** under [Academic Free License](#)
- Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with substantial outside contributions



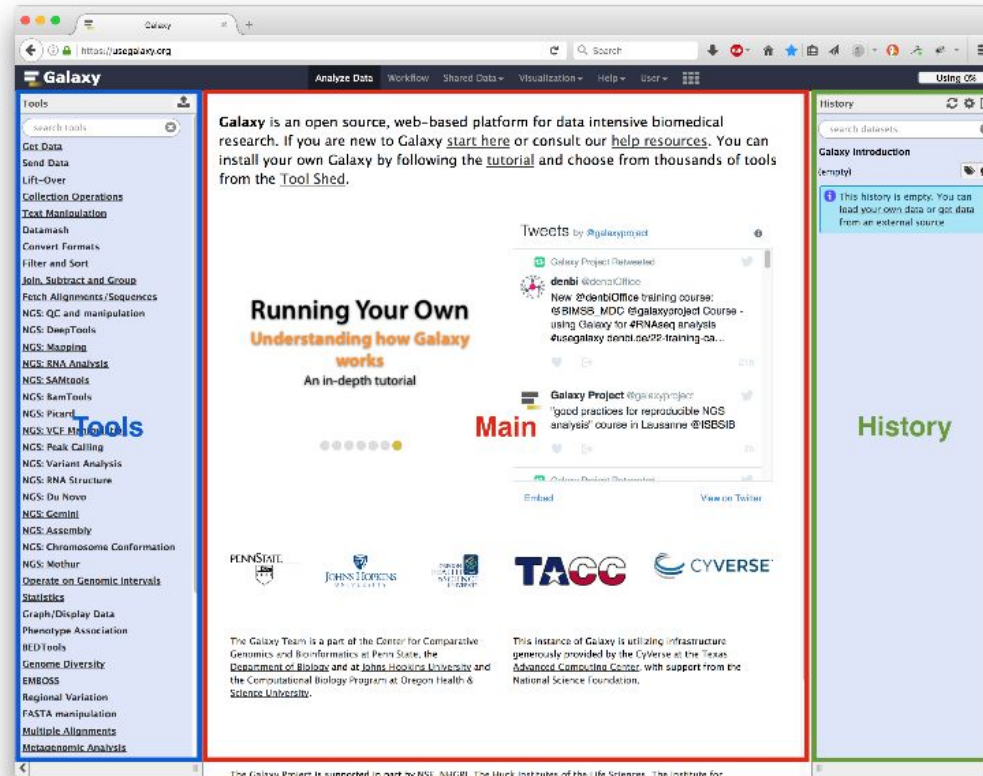
## Core values

- **Accessibility**
  - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
  - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
  - Users can share or publish their analyses (histories, workflows, visualizations)
  - Pages: online Methods for your paper

## Galaxy growth

- More than 7,000 ready to use tools for users
- More than 9,500 [citations](#)
- More than 350 [public Galaxy resources](#)
  - 120+ public servers, many more non-public
  - Both general-purpose and domain-specific

## Main Galaxy interface



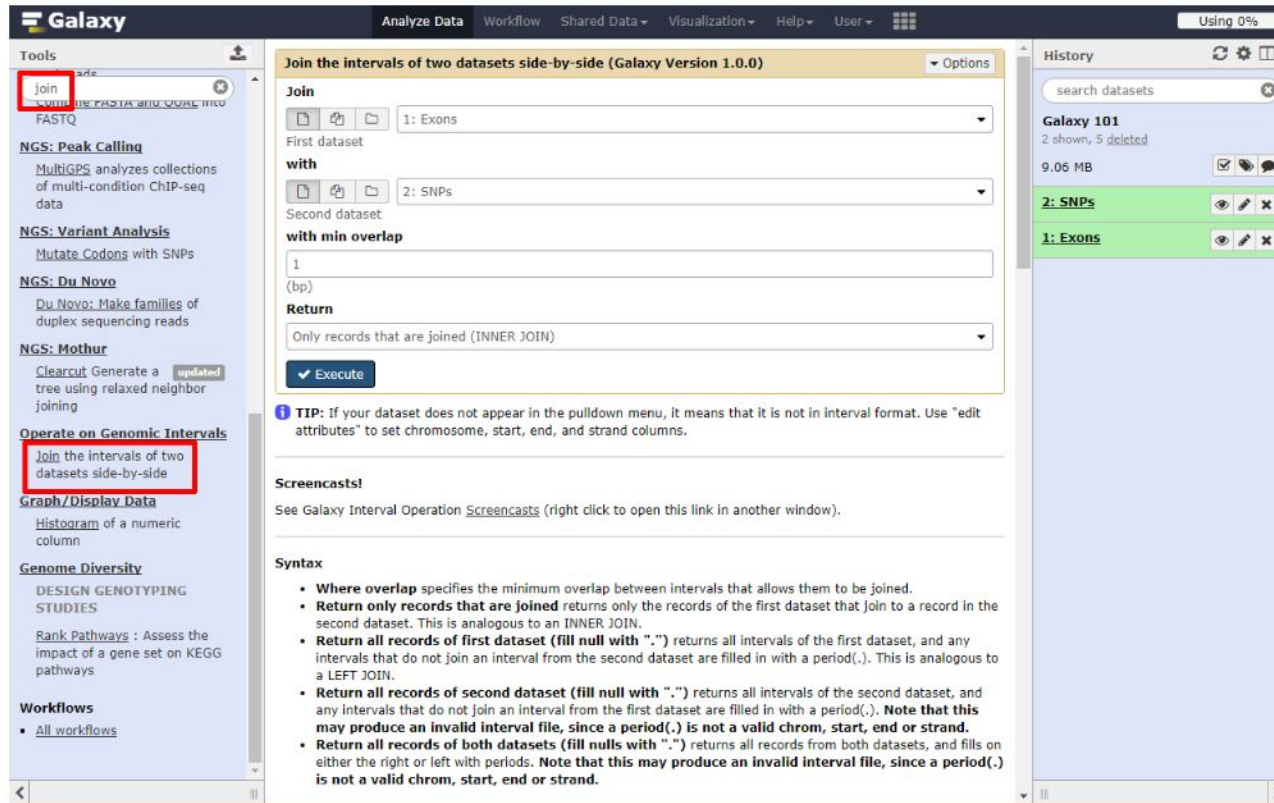
Home page divided into 3 panels

## Top menu



Link	Usage
<i>Analyze Data</i>	go back to the homepage
<i>Workflow</i>	access existing workflows or create new one using the editable diagrammatic pipeline
<i>Visualize</i>	create new visualisations and launch Interactive Environments
<i>Shared data</i>	access data libraries, histories, workflows, visualizations and pages shared with you
<i>Help</i>	links to Galaxy Help Forum (Q&A), Galaxy Community Hub (Wiki), and Interactive Tours
<i>User</i>	your preferences and saved histories, datasets, pages and visualizations

## Tools



The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar is visible, with the tool 'Join the intervals of two datasets side-by-side' highlighted in a red box. The main panel displays the tool's configuration page, which includes a title 'Join the intervals of two datasets side-by-side (Galaxy Version 1.0.0)', a 'Join' dropdown menu, two dataset selection fields (1: Exons, 2: SNPs), a 'with min overlap' field set to '1', and a 'Return' dropdown set to 'Only records that are joined (INNER JOIN)'. Below the configuration is a 'Execute' button and a 'TIP' section. The right sidebar shows the 'History' panel with a search bar and a list of datasets, including '2: SNPs' and '1: Exons'.

**Join the intervals of two datasets side-by-side (Galaxy Version 1.0.0)**

**Join**

First dataset: 1: Exons

with

Second dataset: 2: SNPs

with min overlap: 1 (bp)

Return: Only records that are joined (INNER JOIN)

**Execute**

**TIP:** If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

**Screencasts!**

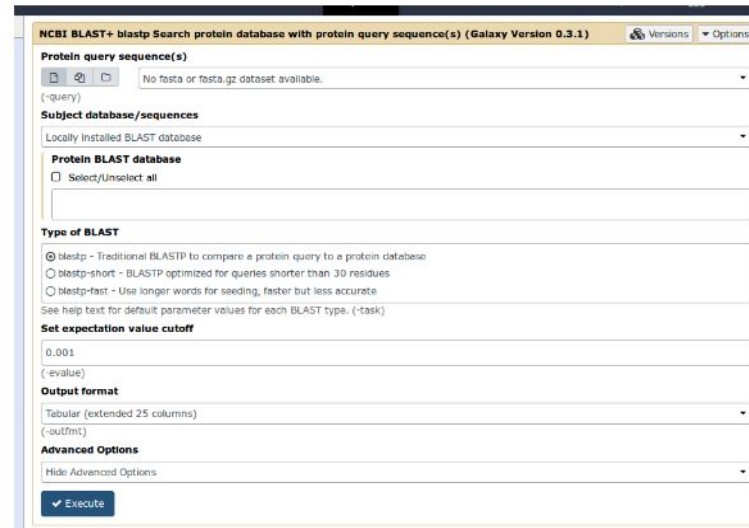
See Galaxy Interval Operation [Screencasts](#) (right click to open this link in another window).

**Syntax**

- **Where overlap** specifies the minimum overlap between intervals that allows them to be joined.
- **Return only records that are joined** returns only the records of the first dataset that join to a record in the second dataset. This is analogous to an INNER JOIN.
- **Return all records of first dataset (fill null with ".")** returns all intervals of the first dataset, and any intervals that do not join an interval from the second dataset are filled in with a period(.). This is analogous to a LEFT JOIN.
- **Return all records of second dataset (fill null with ".")** returns all intervals of the second dataset, and any intervals that do not join an interval from the first dataset are filled in with a period(.). **Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.**
- **Return all records of both datasets (fill nulls with ".")** returns all records from both datasets, and fills on either the right or left with periods. **Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.**

- The tool search helps in finding a tool in a crowded toolbox

## Tool interface



The screenshot shows the NCBI BLAST+ web interface. At the top, it says "NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 0.3.1)". Below this, there are several sections:
 

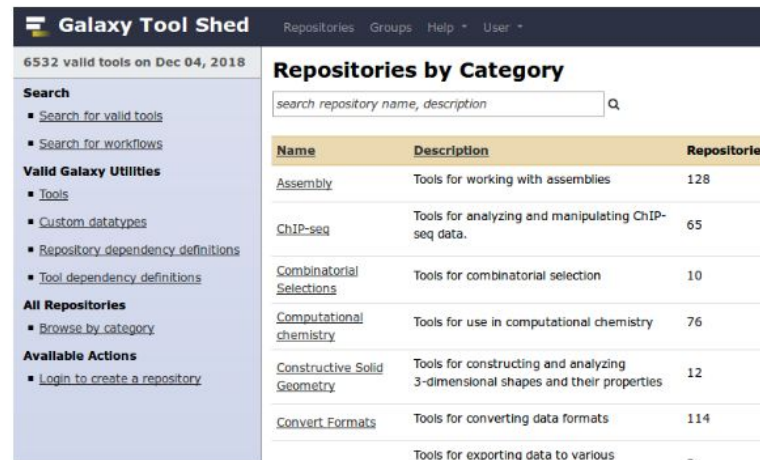
- Protein query sequence(s)**: A text input field with a placeholder "No fasta or fasta.gz dataset available." and a small icon to the left.
- Subject database/sequences**: A dropdown menu currently set to "Locally installed BLAST database".
- Protein BLAST database**: A checkbox labeled "Select/Unselect all" and an empty text input field below it.
- Type of BLAST**: Three radio button options:
  - blastp - Traditional BLASTP to compare a protein query to a protein database
  - blastp-short - BLASTP optimized for queries shorter than 30 residues
  - blastp-fast - Use longer words for seeding, faster but less accurate
 Below these is a note: "See help text for default parameter values for each BLAST type. (-task)"
- Set expectation value cutoff**: A text input field containing "0.001" and a label "(-evalue)" below it.
- Output format**: A dropdown menu set to "Tabular (extended 25 columns)" with a label "(-outfmt)" below it.
- Advanced Options**: A dropdown menu set to "Hide Advanced Options".

 At the bottom left of the form is a blue button labeled "Execute".

- A tool form contains:
  - input datasets and parameters
  - help, citations, metadata
  - an Execute button to start a job, which will add some output datasets to the history
- New tool versions can be installed without removing old ones to ensure reproducibility



# Tool Shed



Galaxy Tool Shed Repositories Groups Help User

6532 valid tools on Dec 04, 2018

**Search**

- Search for valid tools
- Search for workflows

**Valid Galaxy UTILITIES**

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

**All Repositories**

- Browse by category

**Available Actions**

- Login to create a repository

**Repositories by Category**

search repository name, description

Name	Description	Repositories
<a href="#">Assembly</a>	Tools for working with assemblies	128
<a href="#">ChIP-seq</a>	Tools for analyzing and manipulating ChIP-seq data.	65
<a href="#">Combinatorial Selections</a>	Tools for combinatorial selection	10
<a href="#">Computational chemistry</a>	Tools for use in computational chemistry	76
<a href="#">Constructive Solid Geometry</a>	Tools for constructing and analyzing 3-dimensional shapes and their properties	12
<a href="#">Convert Formats</a>	Tools for converting data formats	114
	Tools for exporting data to various	

- Free "app" store: [Galaxy Tool Shed](#)
  - Thousands of tools already available
  - Most software can be integrated
    - If a tool is not available, ask the Galaxy community for help!
  - Only a Galaxy admin can install tools

# History

- Location of all analyses
  - collects all datasets produced by tools
  - collects all operations performed on the data
- For each dataset (the heart of Galaxy's reproducibility), the history tracks
  - name, format, size, creation time, datatype-specific metadata
  - tool id, version, inputs, parameters
  - standard output (stdout) and error (stderr)
  - state (waiting, running, success, failed)
  - hidden, deleted, purged



## Multiple histories

- You can have as many histories as you want
  - each history should correspond to a **different analysis**
  - and should have a meaningful **name**

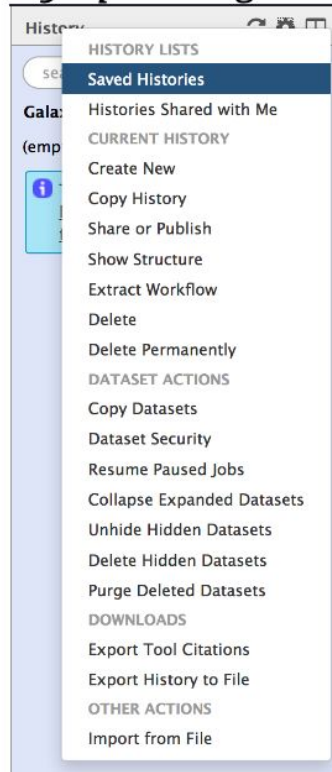


The screenshot displays the Galaxy / Europe web interface with four history panels open, illustrating the concept of multiple histories for different analyses.

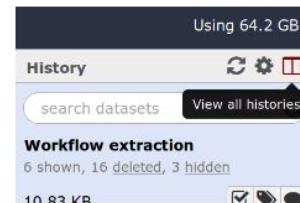
- Current History:** Workflow extract error (10.83 KB). Contains items like '24: data 7 (flattened)', '23: Venn on collection 1: svq', '22: Venn on collection 1: sharedotus', '5: Venn on collection 1: svq', '4: Venn on collection 1: sharedotus', and '1: Sub.sample on data 76: subsample.shared'.
- Unnamed history (910.45 MB):** Contains items like '127: Heatmap.sim on collection 86: heatmap.sim.svg', '119: Plotting tool on collection 83', '113: Classify.seqs on data 48, data 9, and others: tree.sum', '117: Classify.seqs on data 48, data 9, and others: tax.summary', '87: Rarefaction.single on data 79: rarefaction.curves', '86: Dist.shared on data 76: dist.files', '85: Summary.single on data 76: summary', '84: Summary.single on data 76: ave.std.summary', '83: Rarefaction.single on data 76: rarefaction.curves', and '82: Sub.sample on data 76: subsample.shared'.
- Training: 16S rRNA sequencing with mother (1.05 GB):** Contains items like '236: Krona pie chart on data', '235: HTML', '234: Taxonomy-to-Krona on collection 184: krona-formatted taxonomy file', '232: Make.biom on collection 189 and collection 184: biom.files', '231: Newick Display on data', '218: Tree Graph', '217: Tree.shared on collection 199: tre', '214: Venn on collection 189: svq', '213: Venn on collection 189: sharedotus', and '206: Heatmap.sim on collection 199: heatmap.sim.svg'.
- Unnamed history (163.07 MB):** Contains items like '41: samples', '40: https://zenodo.org/record/800651/files/Mock\_R2.fastq', '39: https://zenodo.org/record/800651/files/Mock\_R1.fastq', '38: https://zenodo.org/record/800651/files/F3D9\_R2.fastq', '37: https://zenodo.org/record/800651/files/F3D9\_R1.fastq', '36: https://zenodo.org/record/800651/files/F3D8\_R2.fastq', '35: https://zenodo.org/record/800651/files/F3D8\_R1.fastq', '34: https://zenodo.org/record/800651/files/F3D7\_R2.fastq', '33: https://zenodo.org/record/800651/files/F3D7\_R1.fastq', '32: https://zenodo.org/record/800651/files/F3D6\_R2.fastq', and '31: https://zenodo.org/record/800651/files/F3D6\_R1.fastq'.

## History options menu

History behavior is controlled by the *History options* (gear icon)



- *Create New* history will **not** make your current history disappear
- To see all of your histories, use the history switcher





- *Copy Datasets* from one history to another and save disk space for your quota

## Importing data

- Copy/paste from a file
- Upload data from a local computer
- Upload data from internet using URL
- Upload data from online databases: UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Import from Shared Data (libraries, histories, pages)

## Datatypes

- Tools only accept input datasets with the appropriate datatypes
- When uploading a dataset, its datatype can be either:
  - automatically detected
  - assigned by user
- Dataset produced by a tool: datatype assigned by the tool
- To change the datatype of a dataset:
  -  *Edit Attributes and Datatype*
  -  *Edit Attributes and Convert Formats*

## Reference datasets

### Example: reference Genome

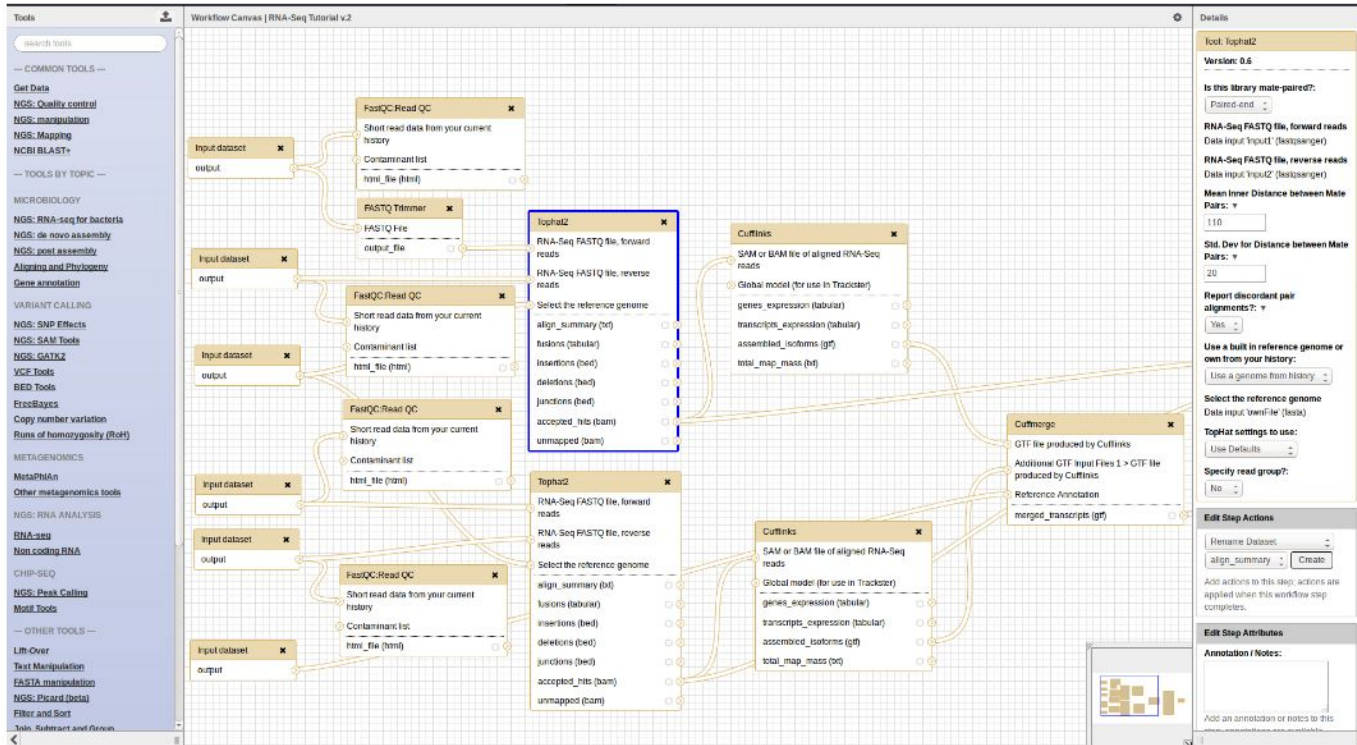
- Genome build specifies which genome assembly a dataset is associated with
  - e.g. mm10, hg38...
- Can be automatically detected or assigned by user
- Users can create custom genome builds
- New builds can be added by the admin

#### Database/Build

Mouse July 2007 (NCBI37/mm9) (mm9)

Burmese python Sep. 2013 (Python\_molurus\_bivittatus-5.0.2/pytt  
 Burton's mouthbreeder Oct 2011 (AstBur1.0/hapBur1) (hapBur1)  
 Bushbaby Mar. 2011 (Broad/otoGar3) (otoGar3)  
 Bushbaby Dec. 2006 (Broad/otoGar1) (otoGar1)  
 C. angaria Oct. 2010 (WS225/caeAng1) (caeAng1)  
 C. brenneri Nov. 2010 (C. brenneri 6.0.1b/caePb3) (caePb3)  
 C. brenneri Feb. 2008 (WUGSC 6.0.1/caePb2) (caePb2)  
 C. brenneri Jan. 2007 (WUGSC 4.0/caePb1) (caePb1)

## Workflow Editor



- **Extracted** from a history
- **Built manually** by adding and configuring tools using the canvas
- **Imported** using an existing shared workflow



## Why would you want to create workflows?

- **Re-run** the same analysis on different input data sets
- **Change parameters** before re-running a similar analysis
- Make use of the workflow job **scheduling**
  - jobs are submitted as soon as their inputs are ready
- Create **sub-workflows**: a workflow inside another workflow
- **Share** workflows for publication and with the community

## Sharing data

- Share everything you do in Galaxy - histories, workflows, and visualizations
  - Directly using a Galaxy account's email addresses on the same instance
  - Using a web link, with anyone who knows the link
  - Using a web link and publishing it to make it accessible to everyone from the *Shared Data* menu

THANK YOU

# Link for Demo

<https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html>