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

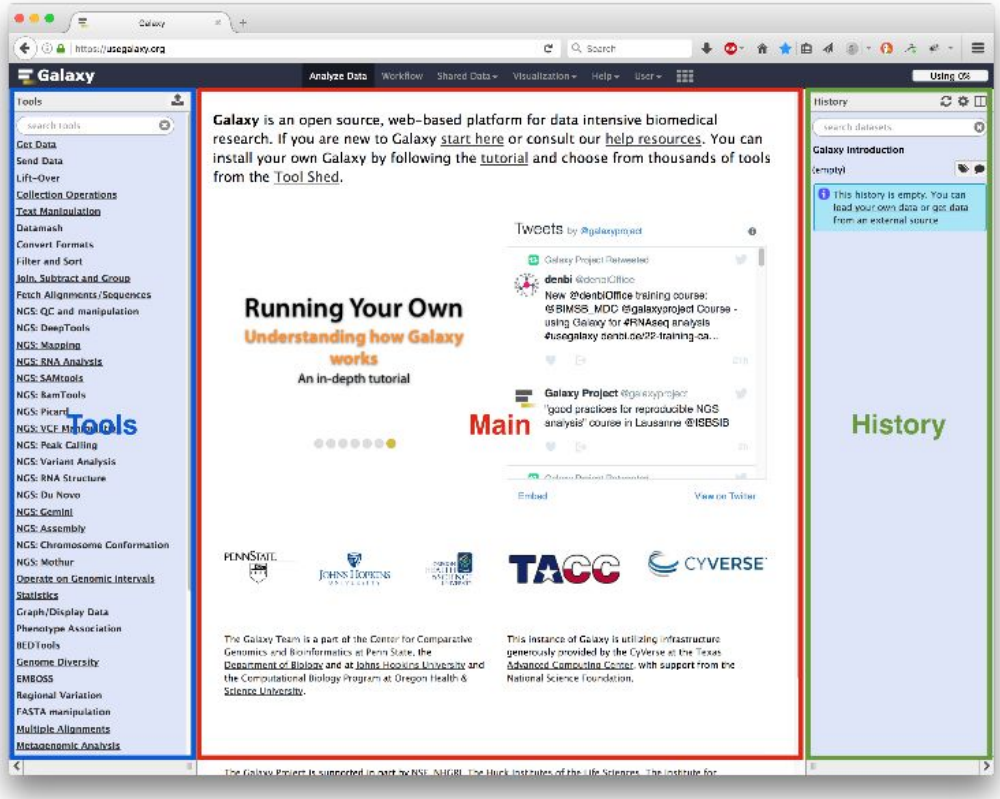
Core values

- **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

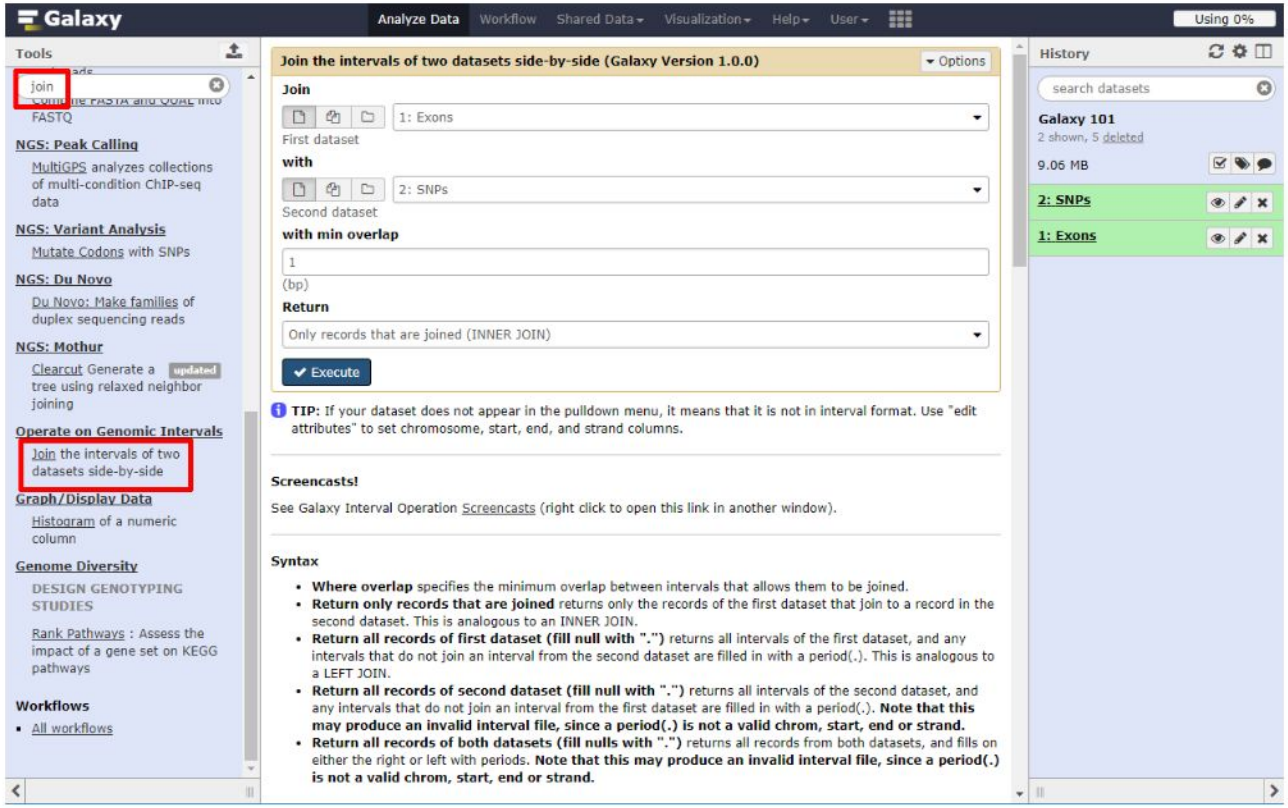


Home page divided into 3 panels

Top menu

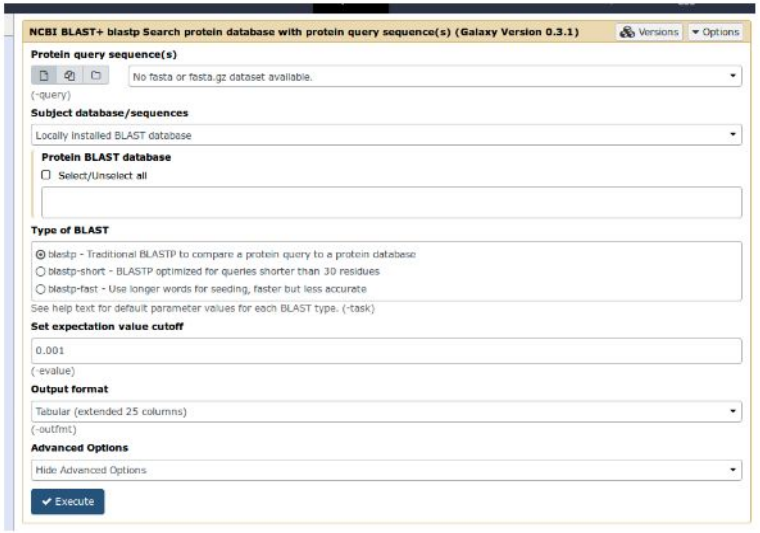
Link	Usage
Analyze Data	go back to the homepage
Workflow	access existing workflows or create new one using the editable diagrammatic pipeline
Visualize	create new visualisations and launch Interactive Environments
Shared data	access data libraries, histories, workflows, visualizations and pages shared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

Tools



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

Tool interface



- 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

- 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

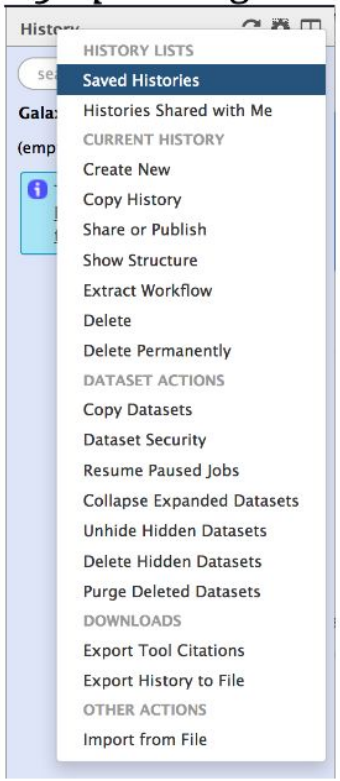
- Location of all analyses
  - collects all datasets produced by tools
  - collects all operations performed on the data
- For each dataset (the heart of Galaxy's reproducibility), the history tracks
  - name, format, size, creation time, datatype-specific metadata
  - tool id, version, inputs, parameters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

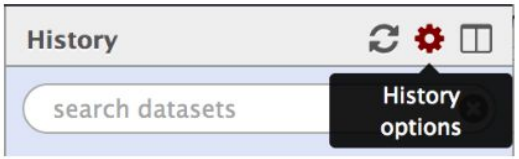
Multiple histories

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

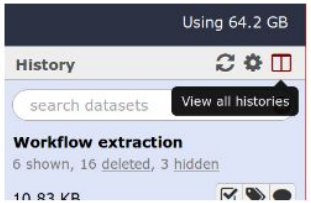
History options menu

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)

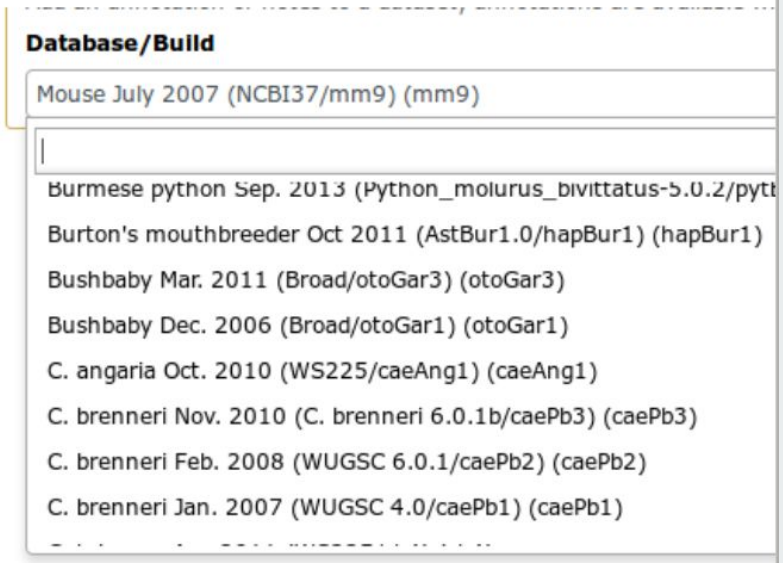
Datatypes

- 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*

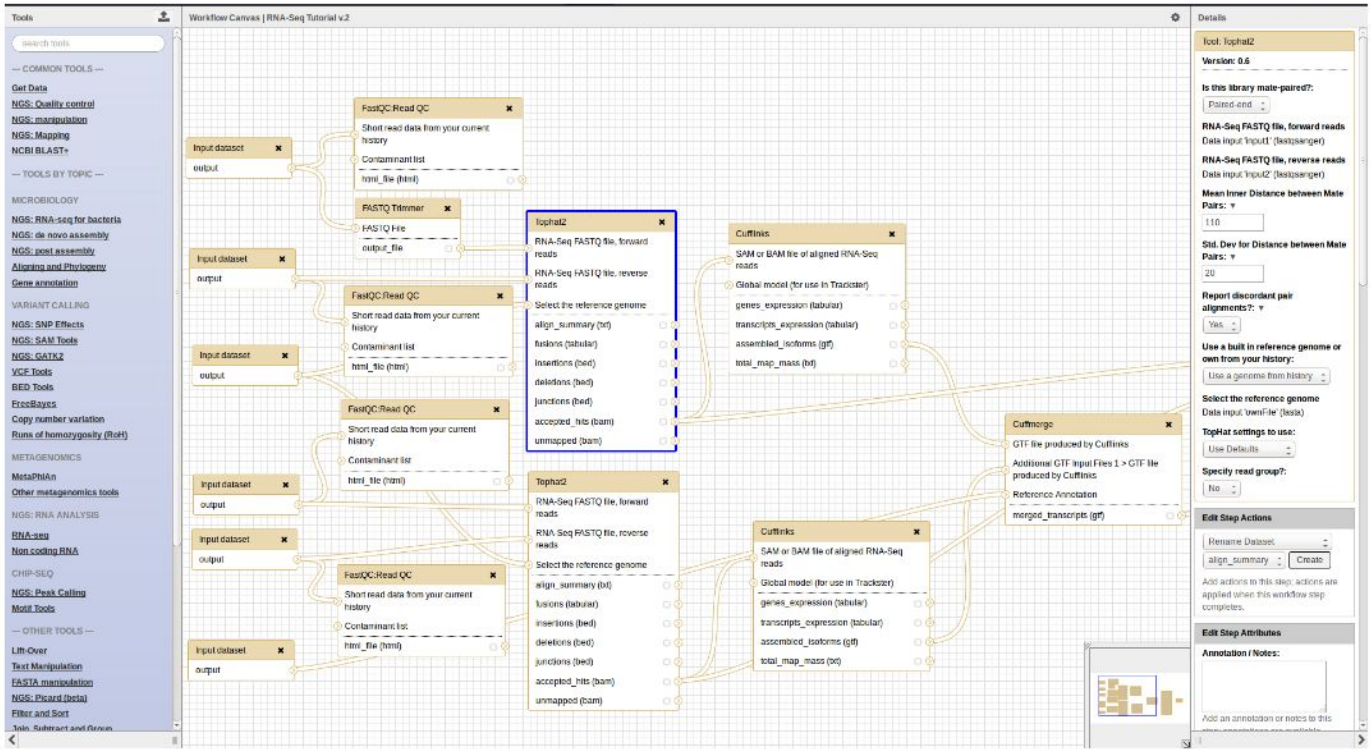
Reference datasets

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



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