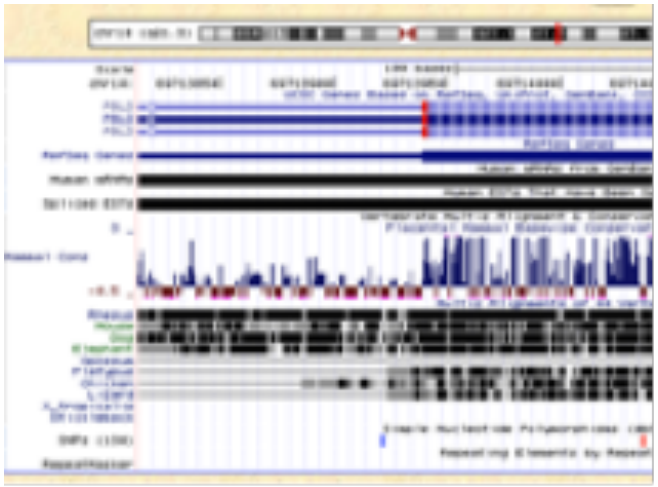
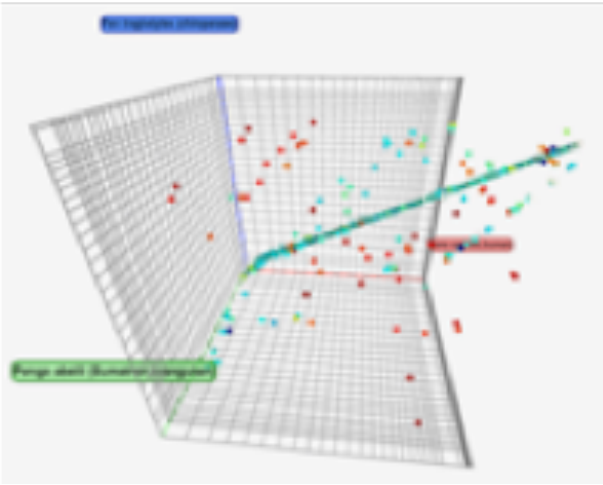
Computational Genomics

Introduction to Galaxy





Introduction to Galaxy

Andrea Bagnacani Bérénice Batut Saskia Hiltemann Anne Pajon
Nicola Soranzo Helena Rasche Christopher Barnett Michele Maroni
Anne Fouilloux Nadia Goué Olha Nahorna Dave Clements

What is Galaxy?

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

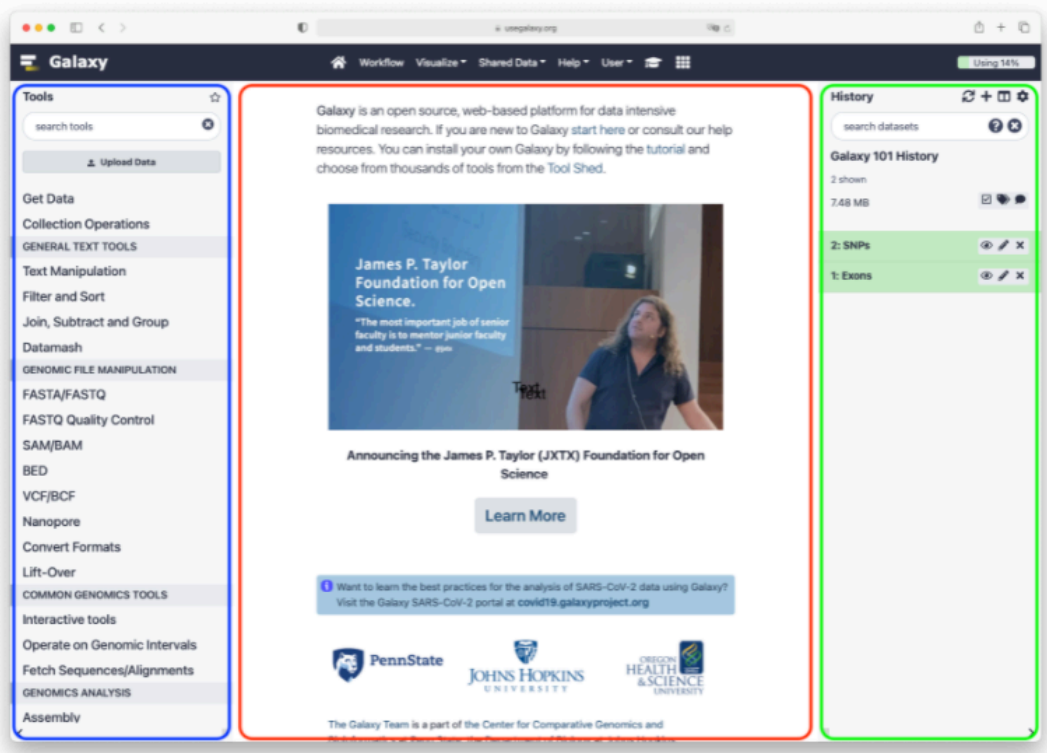
- **Accessibility**
  - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
  - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
  - Users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

Galaxy growth

- More than 8,400 ready to use tools for users
- More than 11,700 citations
- More than 170 public Galaxy resources
- - 130+ public servers, many more non-public
- - Both general-purpose and domain-specific

User Interface

Main Galaxy interface

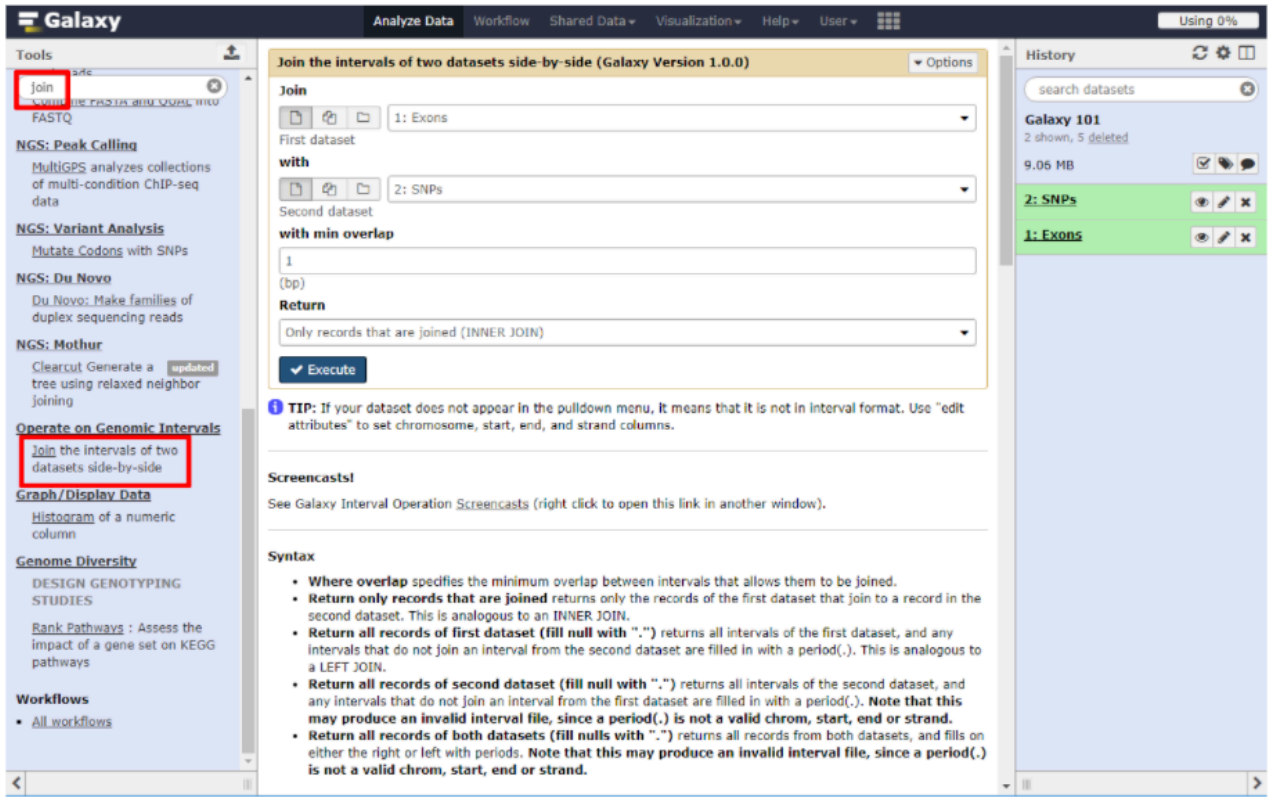


Home page divided into 3 panels

Top menu

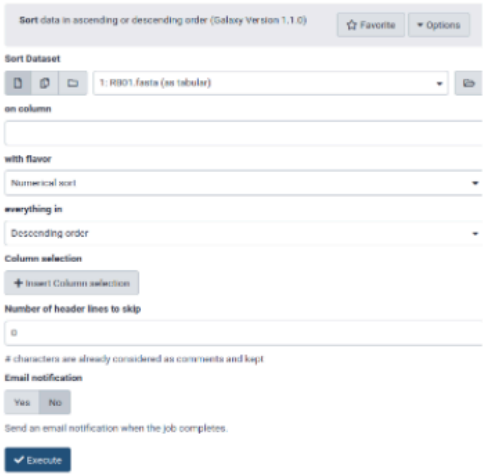
Link	Usage
(or *Analyze Data*)	go back to the homepage
Workflow	access existing workflows or create new one using the editable diagrammatic pipeline
Visualize	create new visualisations and launch Interactive Environments
Shared Data	access data libraries, histories, workflows, visualizations and pages 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



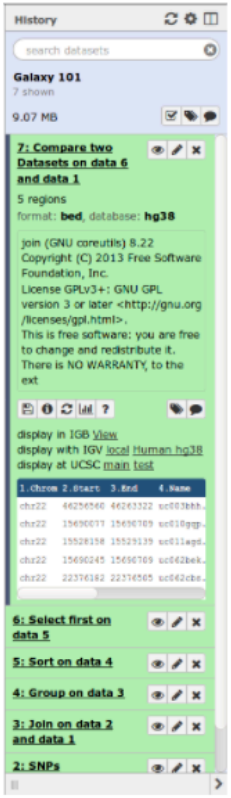
- A tool form contains:
  - input datasets and parameters
  - help, citations, metadata
  - an `Execute` button to start a job, which will add some output datasets to the history
- New tool versions can be installed without removing old ones to ensure reproducibility

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

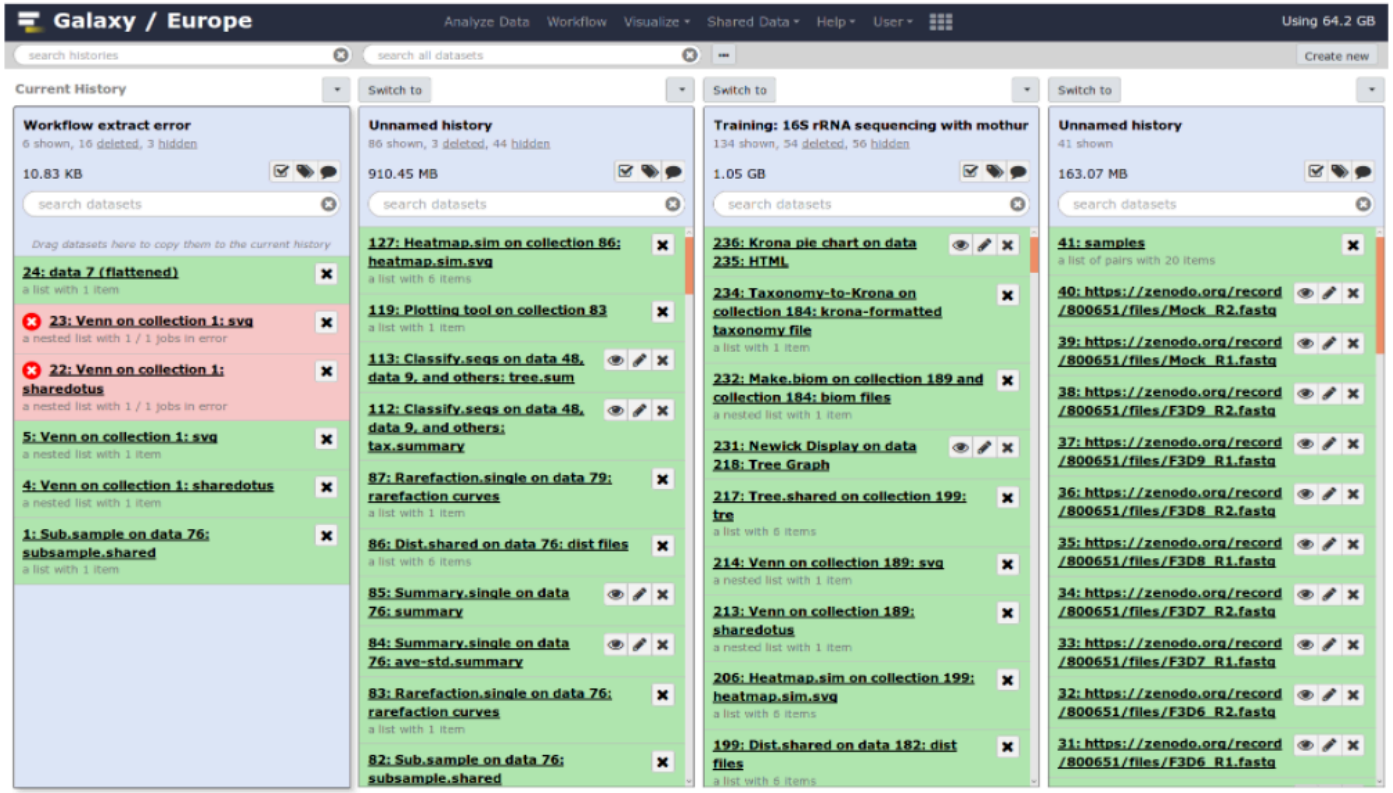
History

- 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



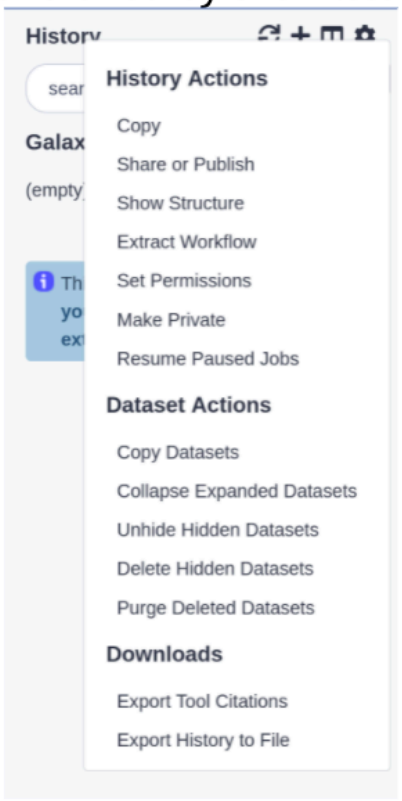
Multiple histories

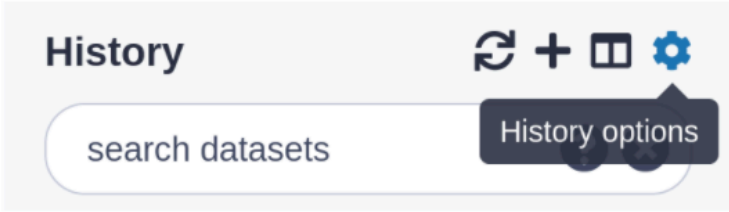
- You can have as many histories as you want
  - each history should correspond to a **different analysis**
  - and should have a meaningful **name**



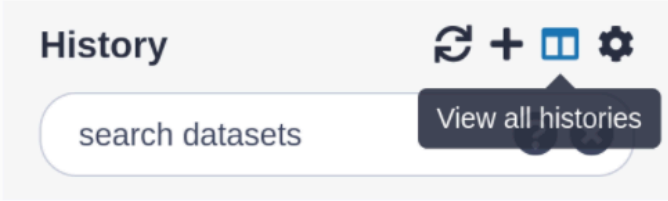
History options menu

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





- *Create new history* (+ icon) 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

Loading data

Importing data

- Copy/paste some text
- Upload files from your local computer
- Upload data from an internet URL
- Upload data from online databases: UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Import from Shared Data (libraries, histories, pages)
- Upload data from FTP

See Getting data into Galaxy

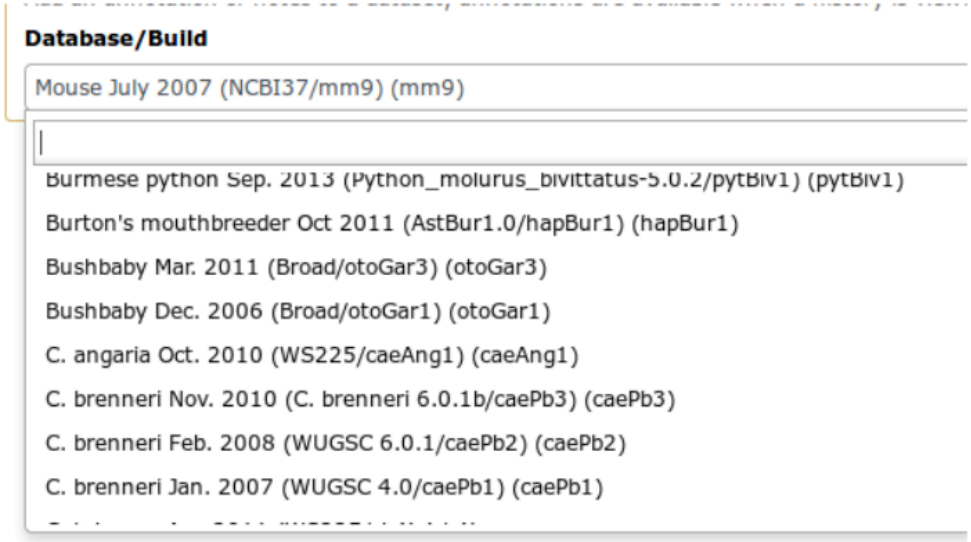
Datatypes

- Tools only accept input datasets with the appropriate datatypes
- When uploading a dataset, its datatype can be either:
  - automatically detected
  - assigned by the user
- Datasets produced by a tool have their datatype assigned by the tool
- To change the datatype of a dataset, either:
  - ✎ *Edit attributes* and *Datatypes* (if original wrong), or
  - ✎ *Edit attributes* and *Convert*

Reference datasets

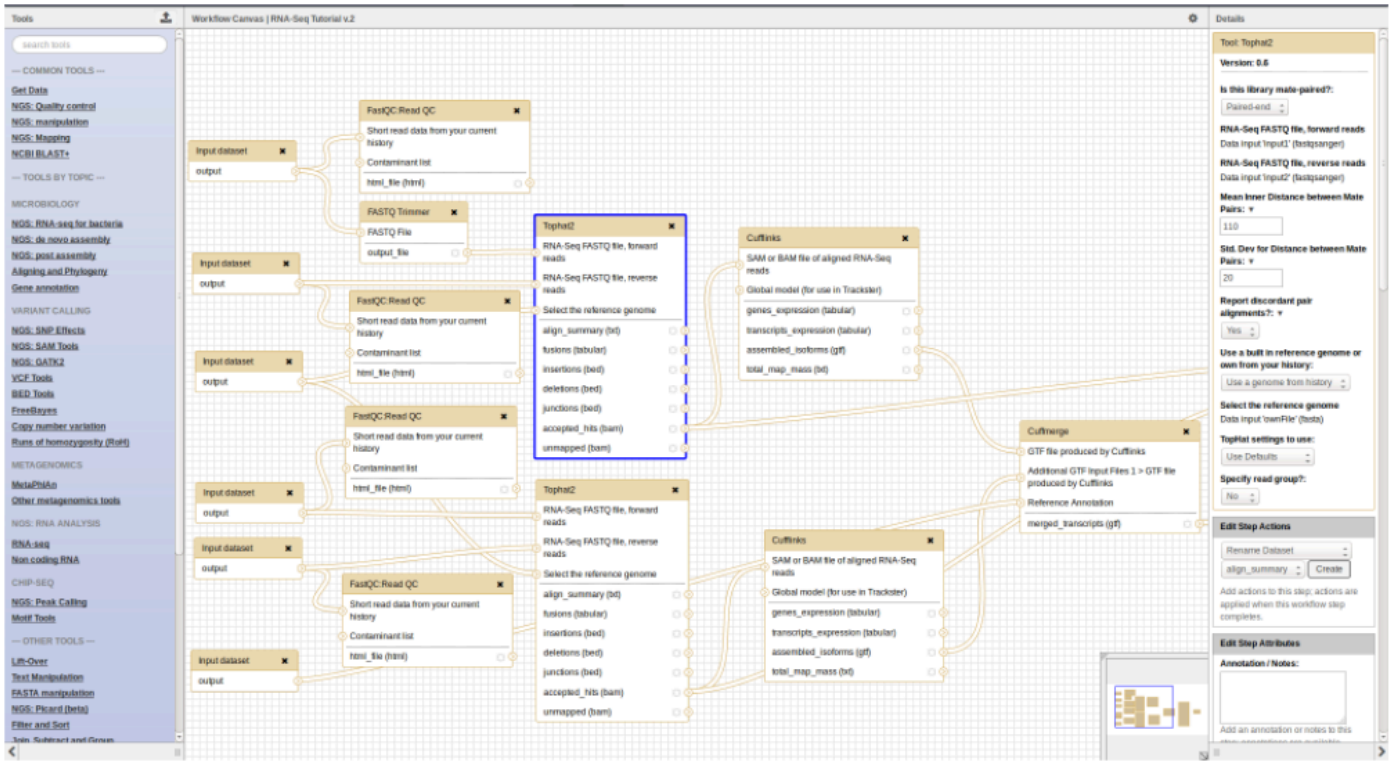
Example: reference Genome

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



Workflows

Workflow Editor

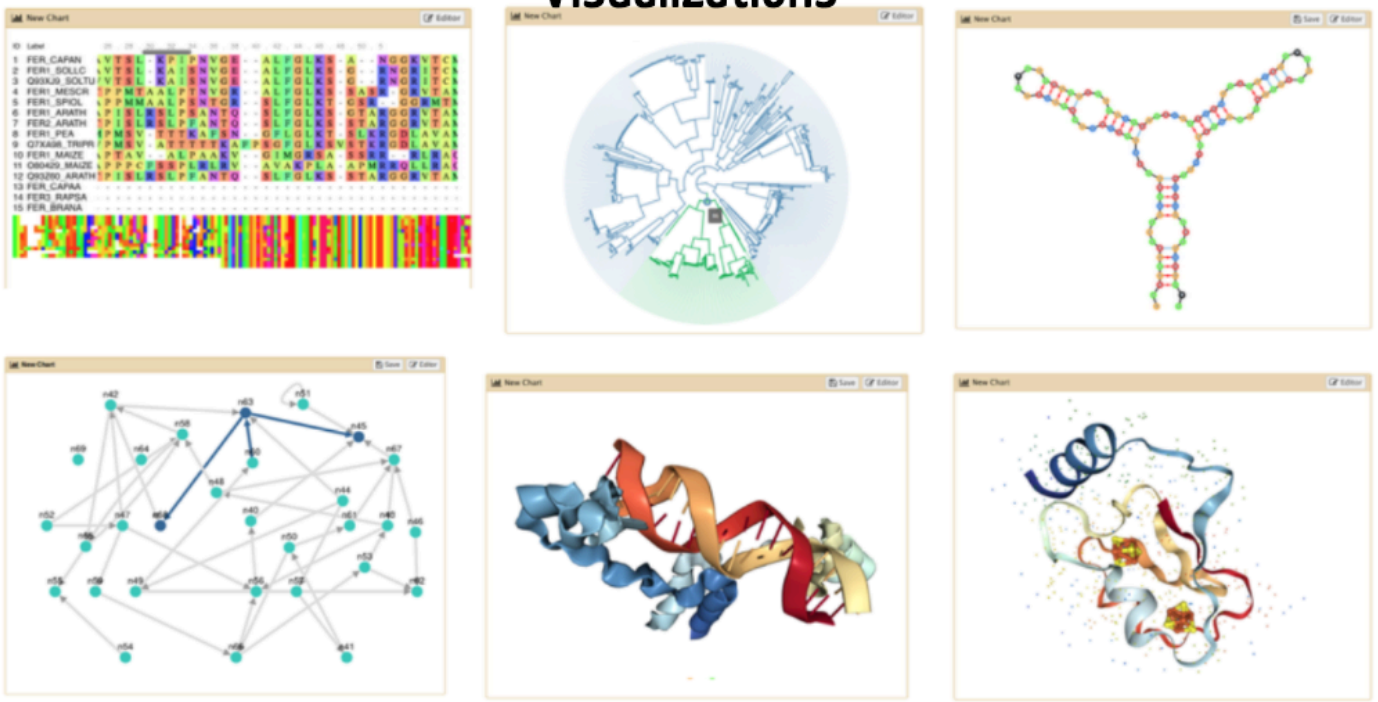


- **Extracted** from a history
- **Built manually** by adding and configuring tools using the canvas
- **Imported** using an existing shared workflow

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

Visualizations



- Datatypes know what tools can be used to visualize datasets:
  - Sequencing data has a button for visualizing in IGV
  - Tabular data will prompt you to build charts
  - Protein data can be seen in a 3D viewer
- Interactive environments: Jupyter, RStudio, etc

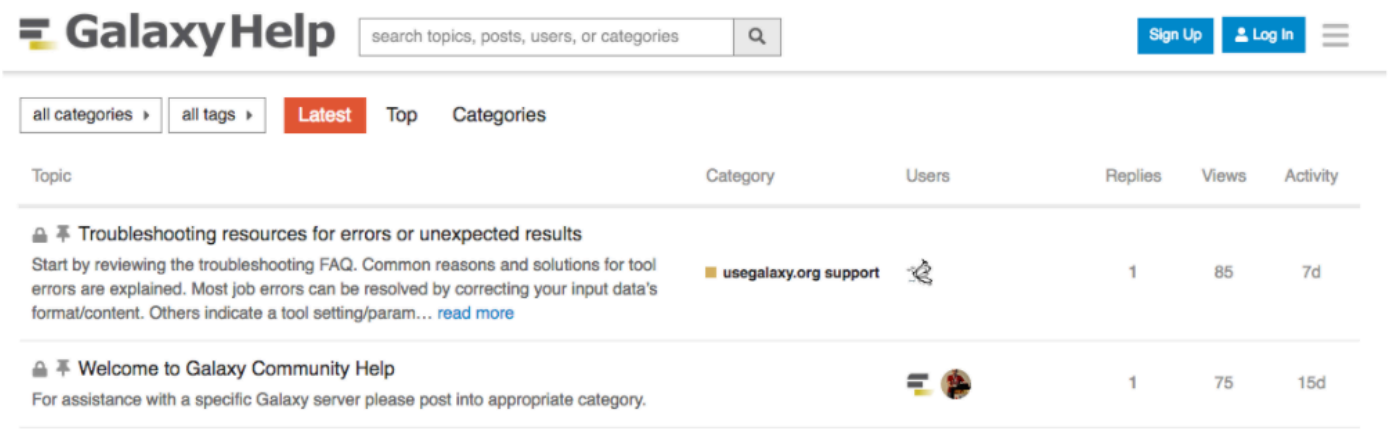
Sharing data

- 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

See Sharing your History in Galaxy

Community

- Support forum: Galaxy Help



- Community curated documentation: Galaxy Community Hub
- Events all around the world
- Galaxy Training for scientists, developers, admins, instructors: Galaxy Training Community
  - Training questions? Chat with us on Gitter