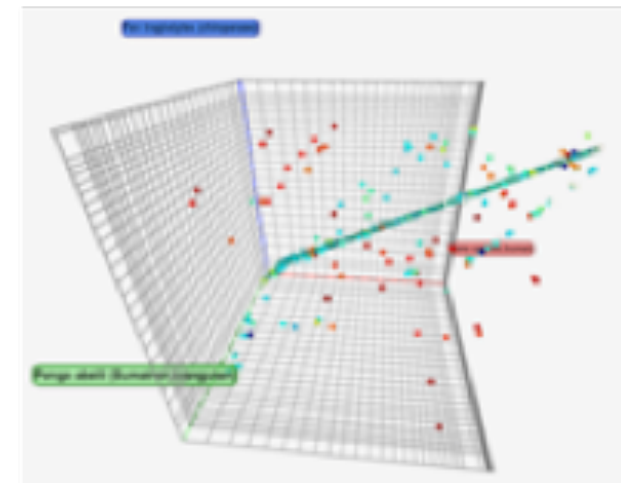
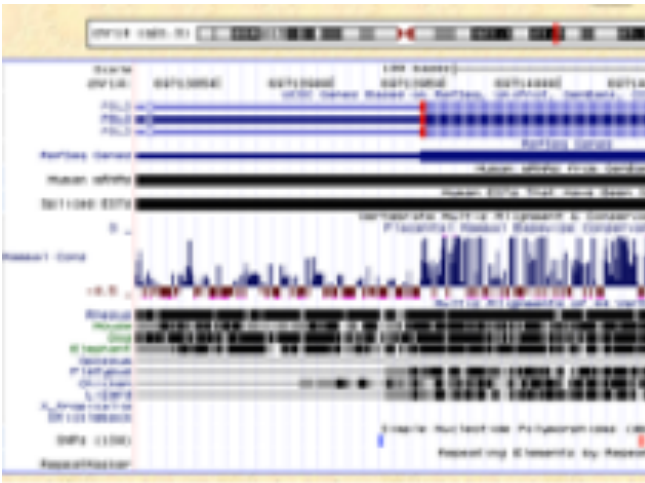


Computational Genomics

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

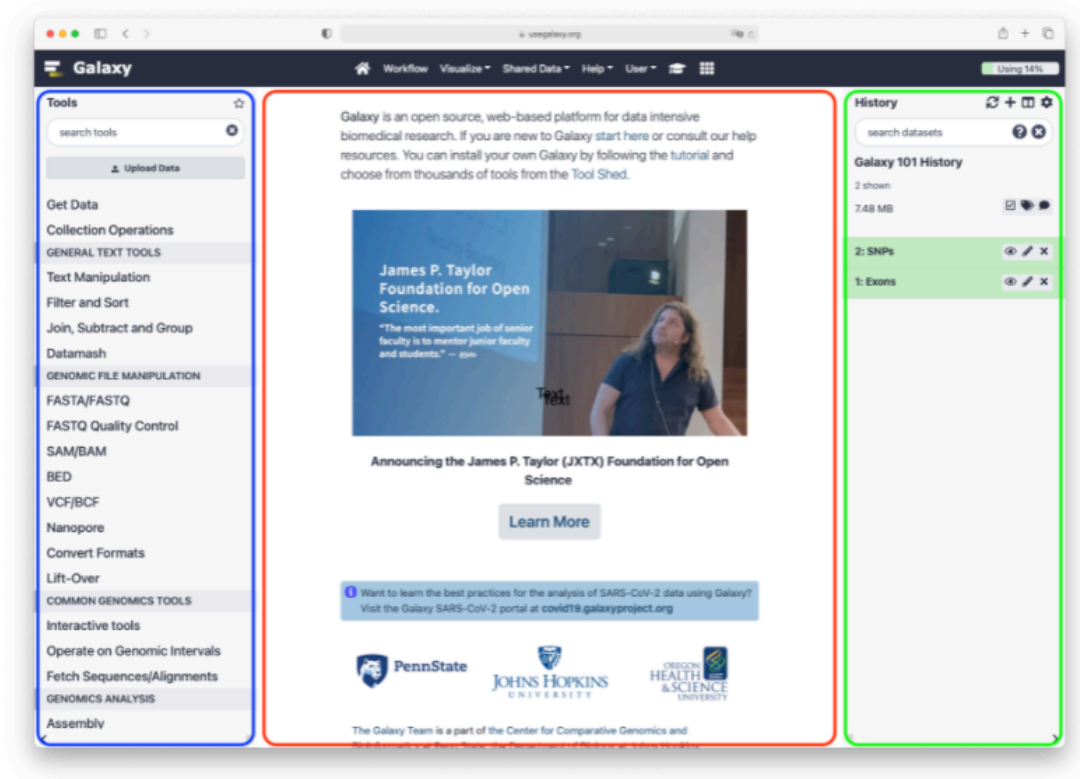
- **Accessibility**
 - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
 - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
 - Users 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 <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 displays the Galaxy web interface. On the left, the 'Tools' panel is visible, with the 'Join' tool highlighted in a red box. The main panel shows the 'Join the intervals of two datasets side-by-side (Galaxy Version 1.0.0)' tool configuration. The 'First dataset' is set to '1: Exons' and the 'Second dataset' is set to '2: SNPs'. The 'with min overlap' field is set to '1 (bp)'. The 'Return' dropdown is set to 'Only records that are joined (INNER JOIN)'. The 'Execute' button is visible. A tip states: '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.' The right panel shows the 'History' section with a search bar and a list of datasets: 'Galaxy 101' (2 shown, 5 deleted, 9.06 MB), '2: SNPs', and '1: Exons'. The 'Workflows' section at the bottom left shows 'All workflows'.

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

Tool interface

Sort data in ascending or descending order (Galaxy Version 1.1.0) ☆ Favorite ▾ Options

Sort Dataset

1: R901.fasta (as tabular) ▾

on column

with flavor

Numerical sort ▾

everything in

Descending order ▾

Column selection

+ Insert Column selection

Number of header lines to skip

0

characters are already considered as comments and kept

Email notification

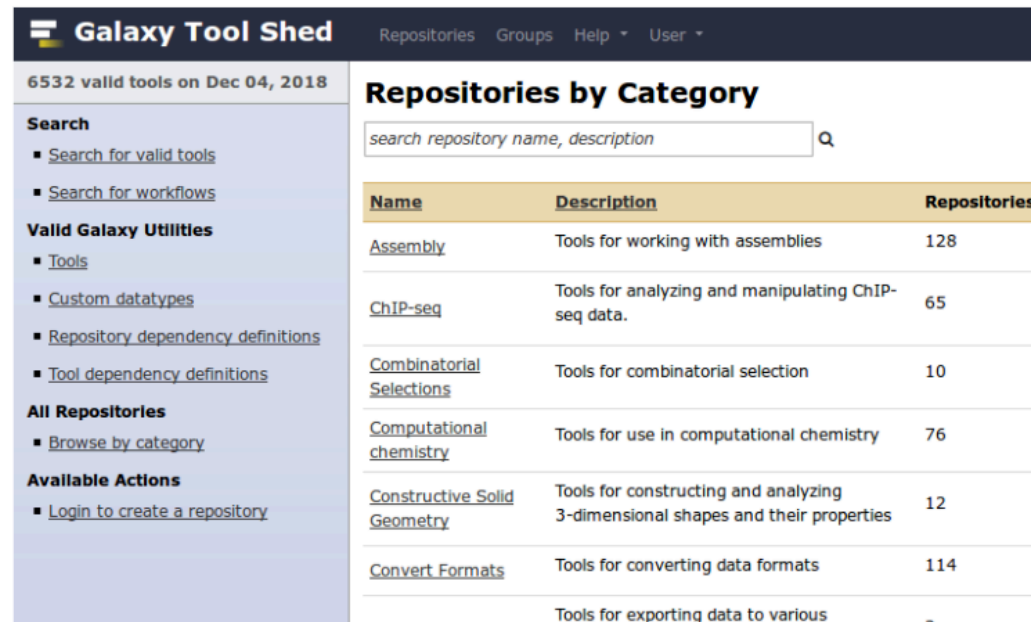
Yes No

Send an email notification when the job completes.

✓ 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 |
|---|--|--------------|
| Assembly | Tools for working with assemblies | 128 |
| ChIP-seq | Tools for analyzing and manipulating ChIP-seq data. | 65 |
| Combinatorial Selections | Tools for combinatorial selection | 10 |
| Computational chemistry | Tools for use in computational chemistry | 76 |
| Constructive Solid Geometry | Tools for constructing and analyzing 3-dimensional shapes and their properties | 12 |
| Convert Formats | 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**

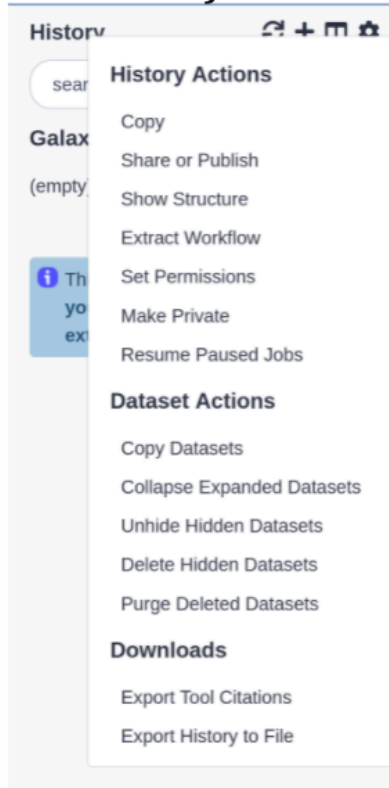
The screenshot displays the Galaxy / Europe web interface, which is designed for managing and analyzing biological data. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', and 'User', along with a 'Using 64.2 GB' status indicator. Below the navigation bar, there are search bars for 'search histories' and 'search all datasets'. The main content area is divided into four vertical panels, each representing a different history:

- Current History:** Titled 'Workflow extract error', it shows 6 shown, 16 deleted, and 3 hidden datasets. It includes a search bar and a list of datasets, some of which are marked with red 'X' icons indicating errors.
- Unnamed history:** Titled 'Unnamed history', it shows 86 shown, 3 deleted, and 44 hidden datasets. It includes a search bar and a list of datasets.
- Training: 16S rRNA sequencing with mothur:** Titled 'Training: 16S rRNA sequencing with mothur', it shows 134 shown, 54 deleted, and 56 hidden datasets. It includes a search bar and a list of datasets.
- Unnamed history:** Titled 'Unnamed history', it shows 41 shown datasets. It includes a search bar and a list of datasets.

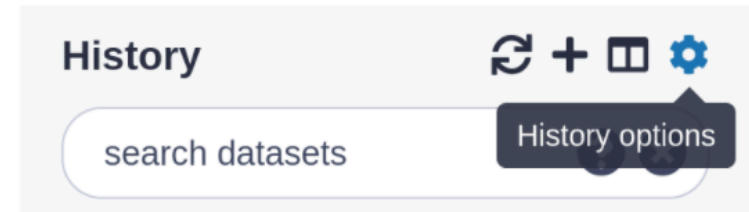
Each history panel has a 'Switch to' button and a 'search datasets' bar. The datasets are listed with their names, sizes, and various icons for viewing, deleting, and sharing. The interface is designed to allow users to manage multiple analyses simultaneously, each with its own history and dataset collection.

History options menu

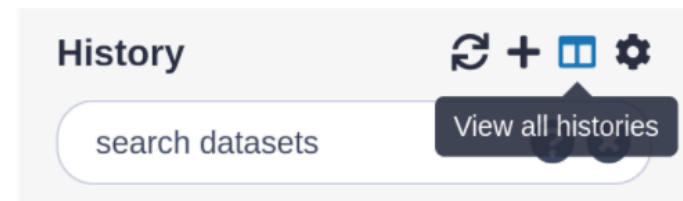
History behavior is controlled by the *History*



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

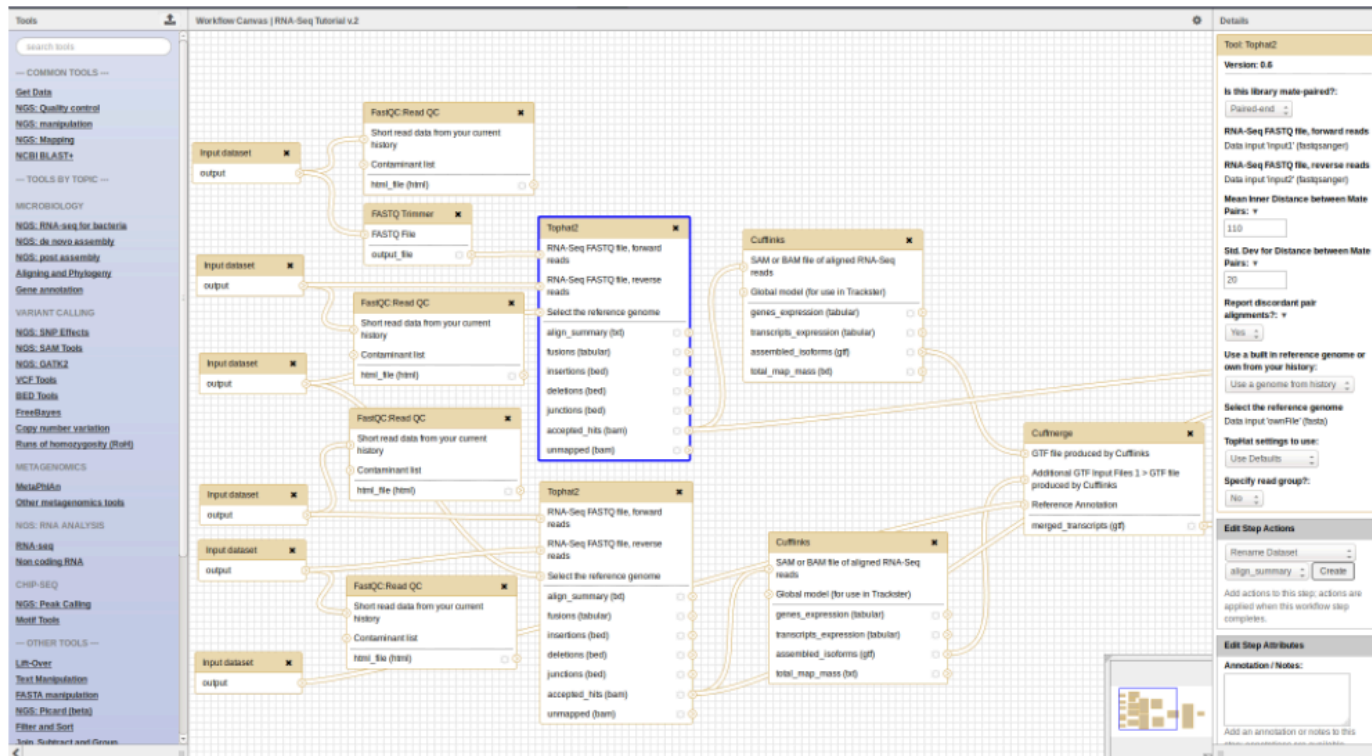
Database/Build

Mouse July 2007 (NCBI37/mm9) (mm9)

Burmese python Sep. 2013 (Python_molurus_bivittatus-5.0.2/pytBiv1) (pytBiv1)
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)

Workflows

Workflow Editor

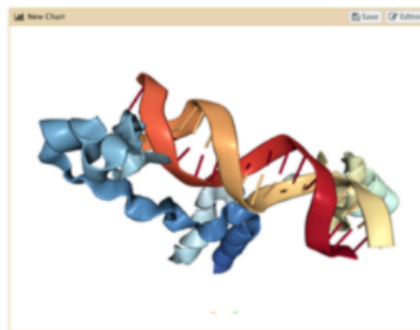
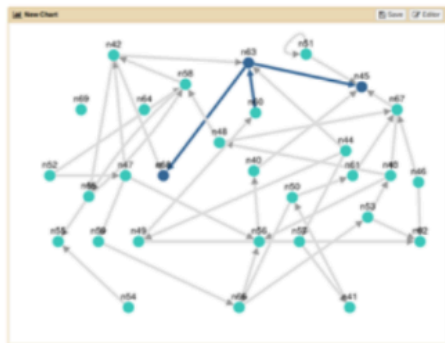
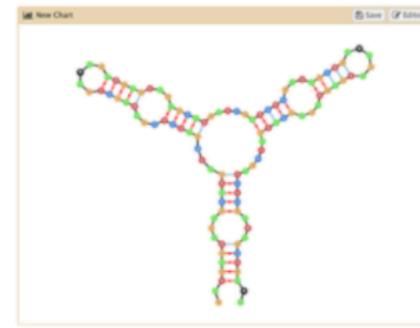
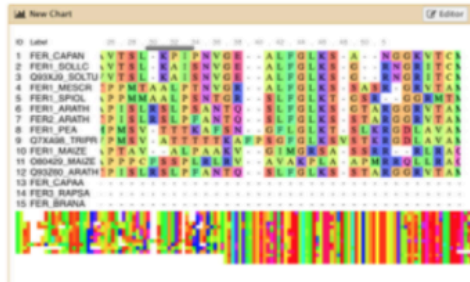


- **Extracted** from a history
- **Built manually** by adding and configuring tools using the canvas
- **Imported** using an existing shared workflow

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](#)

 **GalaxyHelp**

search topics, posts, users, or categories

Sign Up Log In

all categories all tags **Latest** Top Categories

| Topic | Category | Users | Replies | Views | Activity |
|--|--|--|---------|-------|----------|
|   Troubleshooting resources for errors or unexpected results Start by reviewing the troubleshooting FAQ. Common reasons and solutions for tool errors are explained. Most job errors can be resolved by correcting your input data's format/content. Others indicate a tool setting/param... read more |  usegalaxy.org support |  | 1 | 85 | 7d |
|   Welcome to Galaxy Community Help For assistance with a specific Galaxy server please post into appropriate category. | |   | 1 | 75 | 15d |

- 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](#)