



Introduction to Galaxy

Authors: Andrea Bagnacani Bérénice Batut Saskia Hiltemann Anne Pajon

Nicola Soranzo Helena Rasche Christopher Barnett Michele Maroni Anne Fouilloux

Nadia Goué Olha Nahorna Dave Clements

Updated:

Plain-text slides

Tip: press P to view the presenter notes | ♦ Use arrow keys to move between slides

1 / 26



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 8,500 ready to use tools for users
- More than 11,800 [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

The screenshot shows the main Galaxy web interface with three distinct panels:

- Left Panel (Blue Border): Tools**
 - Upload Data
 - Get Data
 - Collection Operations
 - GENERAL TEXT TOOLS
 - Text Manipulation
 - Filter and Sort
 - Join, Subtract and Group
 - Datamash
 - GENOMIC FILE MANIPULATION
 - FASTA/FASTQ
 - FASTQ Quality Control
 - SAM/BAM
 - BED
 - VCF/BCF
 - Nanopore
 - Convert Formats
 - Lift-Over
 - COMMON GENOMICS TOOLS
 - Interactive tools
 - Operate on Genomic Intervals
 - Fetch Sequences/Alignments
 - GENOMICS ANALYSIS
 - Assembly
- Middle Panel (Red Border): Home**

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

James P. Taylor
Foundation for Open Science.
"The most important job of senior faculty is to mentor junior faculty and students." — jptn

Announcing the James P. Taylor (JXTX) Foundation for Open Science

Learn More

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at covid19.galaxyproject.org

PennState Johns Hopkins University Oregon Health & Science University

The Galaxy Team is a part of the Center for Comparative Genomics and
- Right Panel (Green Border): History**

Galaxy 101 History

2 shown

 - 7.48 MB
 - 2: SNPs
 - 1: Exons

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 shows the Galaxy web interface. On the left, a sidebar lists various tools under categories like NGS: Peak Calling, NGS: Variant Analysis, NGS: Du Novo, NGS: Mothur, and Operate on Genomic Intervals. The 'Join the intervals of two datasets side-by-side' tool is highlighted with a red box. On the right, a main panel displays the configuration for this tool. It has fields for 'First dataset' (set to 'Exons') and 'Second dataset' (set to 'SNPs'). Below these are options for 'with min overlap' (set to '1 (bp)'), 'Return' (set to 'Only records that are joined (INNER JOIN)'), and a 'Execute' button. A tip message at the bottom of the panel says: '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.' At the bottom of the main panel, there's a 'Syntax' section with detailed instructions about the 'join' command. To the right of the main panel is a 'History' sidebar showing a list of datasets: 'Galaxy 101' (2 shown, 5 deleted), '9.06 MB', '2: SNPs' (selected), and '1: Exons'. There are also icons for search, settings, and help.

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



Tool interface

Sort data in ascending or descending order (Galaxy Version 1.2.0)

Sort Dataset

on column: 1: 1.bed

with flavor: Numerical sort

everything in: Descending order

Column selection: + Insert Column selection

Number of header lines to skip: 0

Email notification: No

Execute

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

Syntax

This tool sorts the dataset on any number of columns in either ascending or descending order.

- Numerical sort orders numbers by their magnitude, ignores all characters besides numbers, and evaluates a string of numbers to the value they signify.

- 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

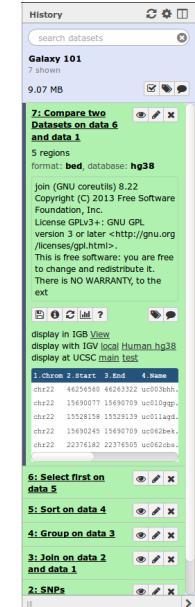
Name	Description	Repositories
Assembly	Tools for working with assemblies	128
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	65
Combinatorial Selections	Tools for combinatorial selection	10
Computational chemistry	Tools for use in computational chemistry	76
Constructive Solid Geometry	Tools 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 shows the Galaxy Europe web interface. At the top, there is a navigation bar with links for Workflow, Visualize, Shared Data, Help, User, and a sign-in button. Below the navigation bar, there are search fields for "search histories" and "search all datasets". The main area displays a list of histories:

- Current History
- Troubleshooting
- Imported-CSV Inferface
- Imported-CSV Filter, Plot and
- Imported-GSE13423Q
- Unnamed history

Each history item has a "Switch to" dropdown menu next to it.



13 / 26



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)

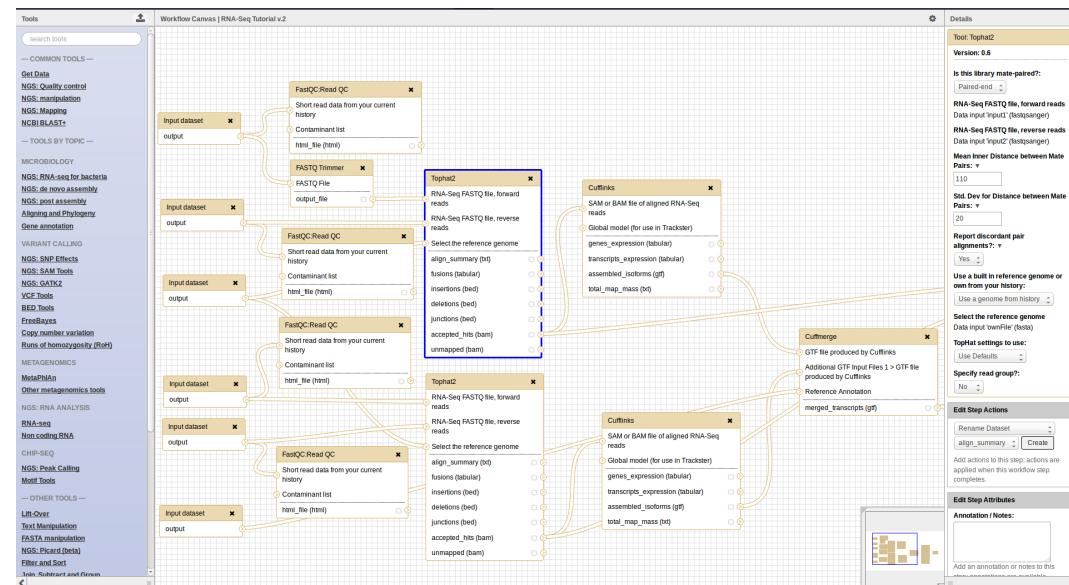
C. brenneri Jan. 2007 (WUGSC 4.0/caePb1) (caePb1)



Workflows



Workflow Editor

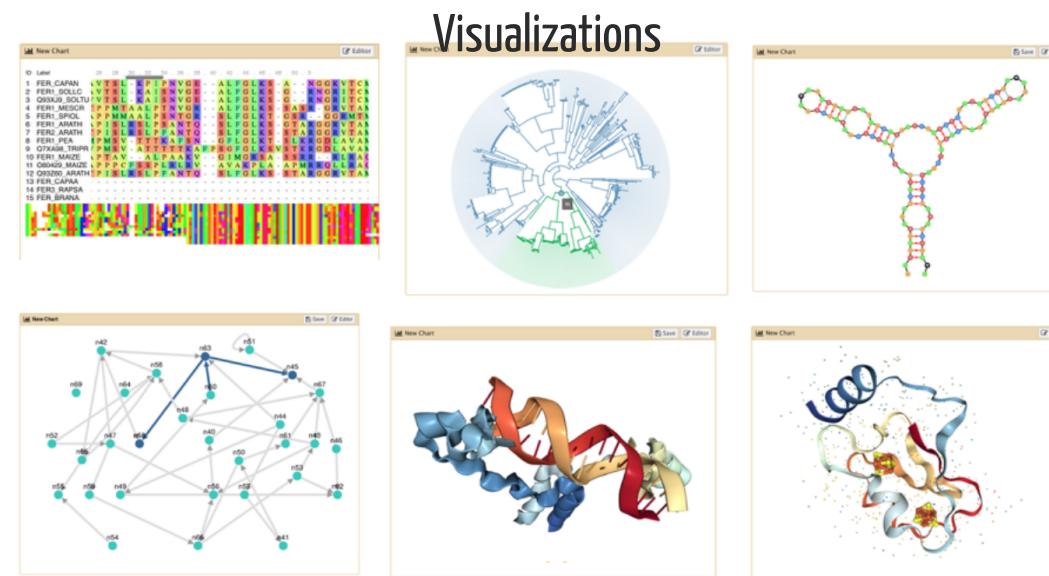


- Extracted from a history
- Built manually by adding and configuring tools using the canvas
- Imported using an existing shared workflow



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



- 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](#)

The screenshot shows the Galaxy Help forum homepage. At the top, there is a search bar and links for 'Sign Up' and 'Log In'. Below the header, there are navigation buttons for 'all categories', 'all tags', 'Latest' (which is highlighted in red), 'Top', and 'Categories'. The main content area displays two forum posts:

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](#)



Related tutorials



Thank You!

This material is the result of a collaborative work. Thanks to the [Galaxy Training Network](#) and all the contributors!

Authors: Andrea Bagnacani Bérénice Batut Saskia Hiltemann Anne Pajon

Nicola Soranzo Helena Rasche Christopher Barnett Michele Maroni Anne Fouilloux

Nadia Goué Olha Nahorna Dave Clements



This material is licensed under the [Creative Commons Attribution 4.0 International License](#).

