Introduction to Galaxy

Stéphanie Le Gras
(slegras@igbmc.fr)

Training material

- Wiki:

http://genomeast.igbmc.fr/wiki/

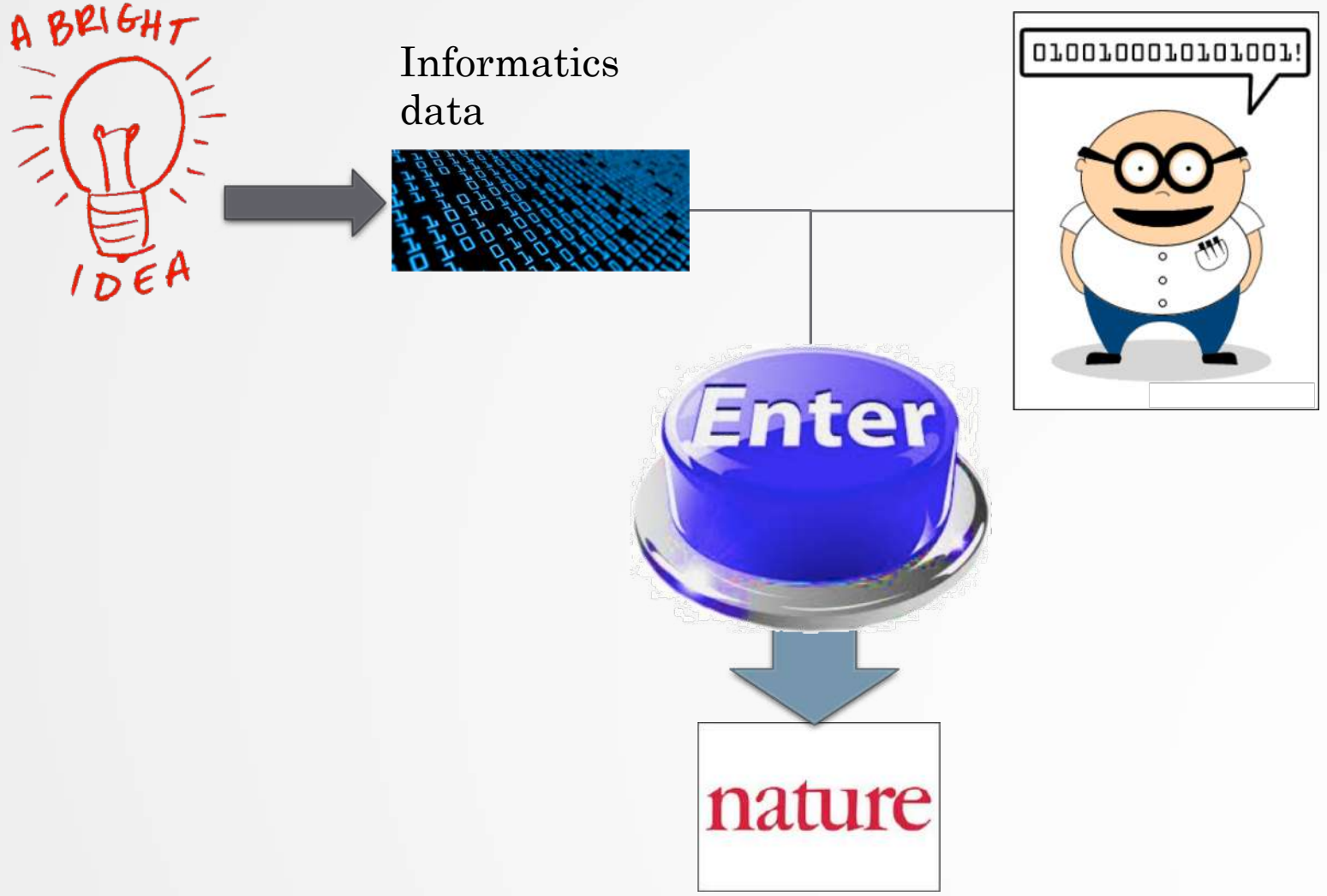
- Go to: Training > DU Dijon > Introduction to Galaxy > Hands-on
- http://use.galaxeast.fr
- Login: user[1..22]
- Password: training

Guidelines

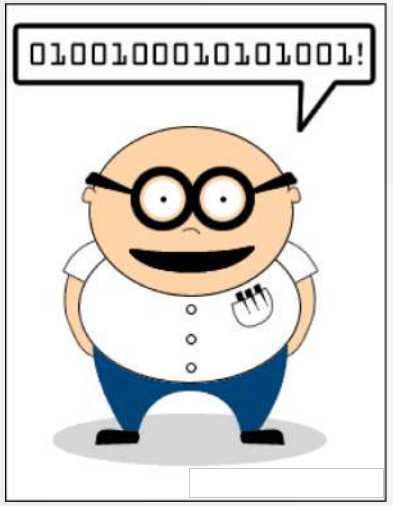
- Analyzing biological data with informatics tools
- Presentation of the Galaxy project
- Description of the main features of the Galaxy platform
- Workflow

Analyzing biological data with informatics tools

Bioinformatics analyses



Bioinformatics analyses



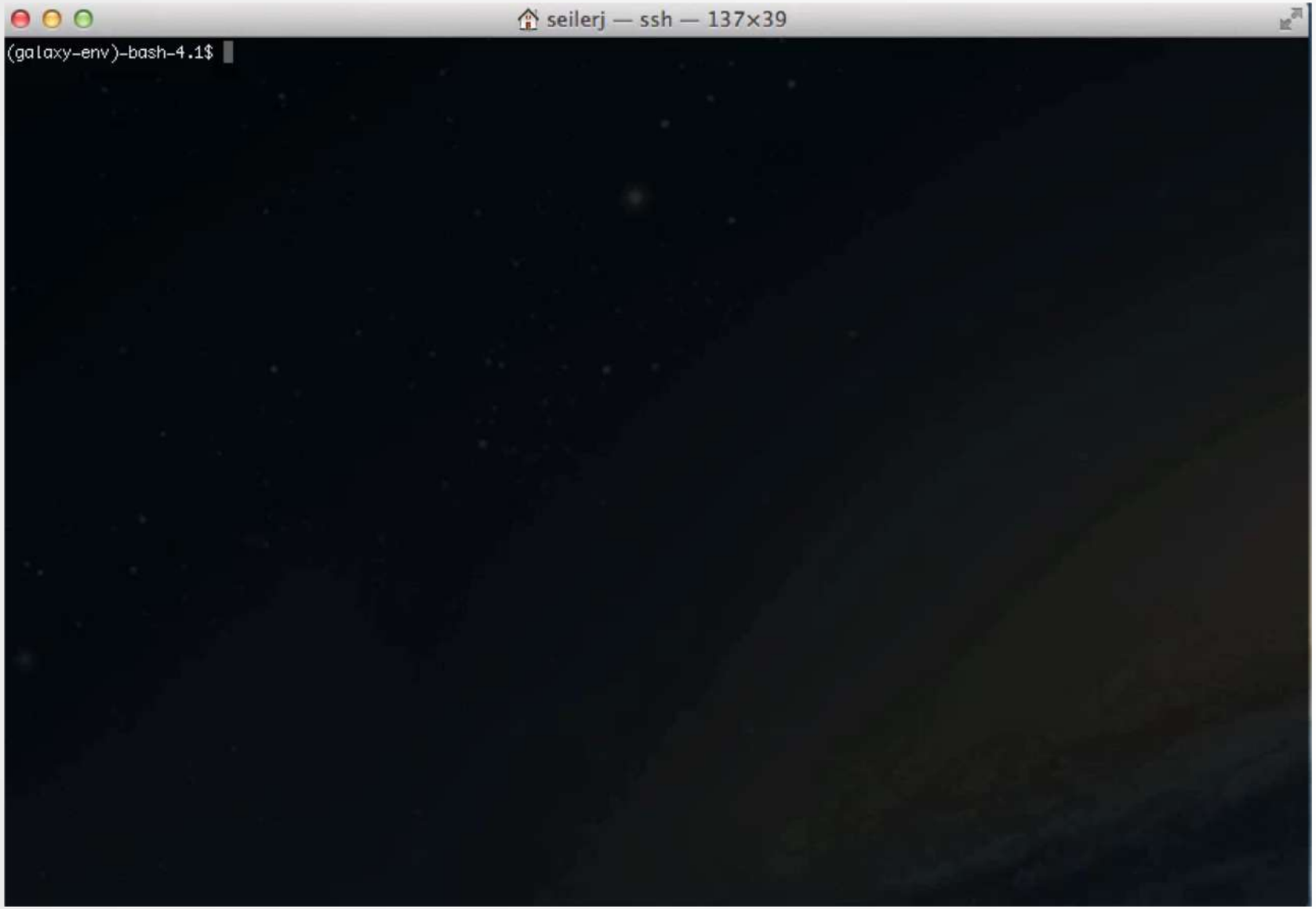
Scripts, software

```perl
#! /usr/bin/perl

use strict;
use warnings;
use Getopt::Long;

## Date : 22 fev 2011
## Author : Stephanie Le Gras

## Objectives :

my $num_arg  = scalar @ARGV;
my $progname = "ExtractID.pl";
my $input;
my $out;
my $id;

my $result = GetOptions(
    "id=s" => \$id,
    "out=s"      => \$out,
    "input=s"       => \$input,
);

my $usage = <<END;

Usage: $progname --id=FILENAME --out=FILENAME --input=FILENAME

END

die $usage unless ($result);
my @files = @ARGV;
die "Enter at least two files\n$usage" if ( $num_arg < 2 );
die $usage if ( $num_arg == 0 );

my %ids;
$out = ( defined $out ) ? $out : "results.txt";

## first, every lines of each files are put in the hash table ids. Variant ids are used as keys of the
## hash table and it contains a table.
```

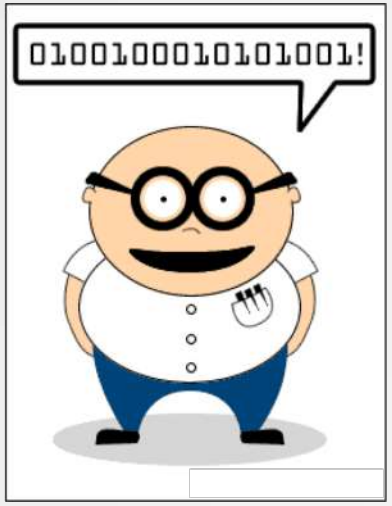
Command line

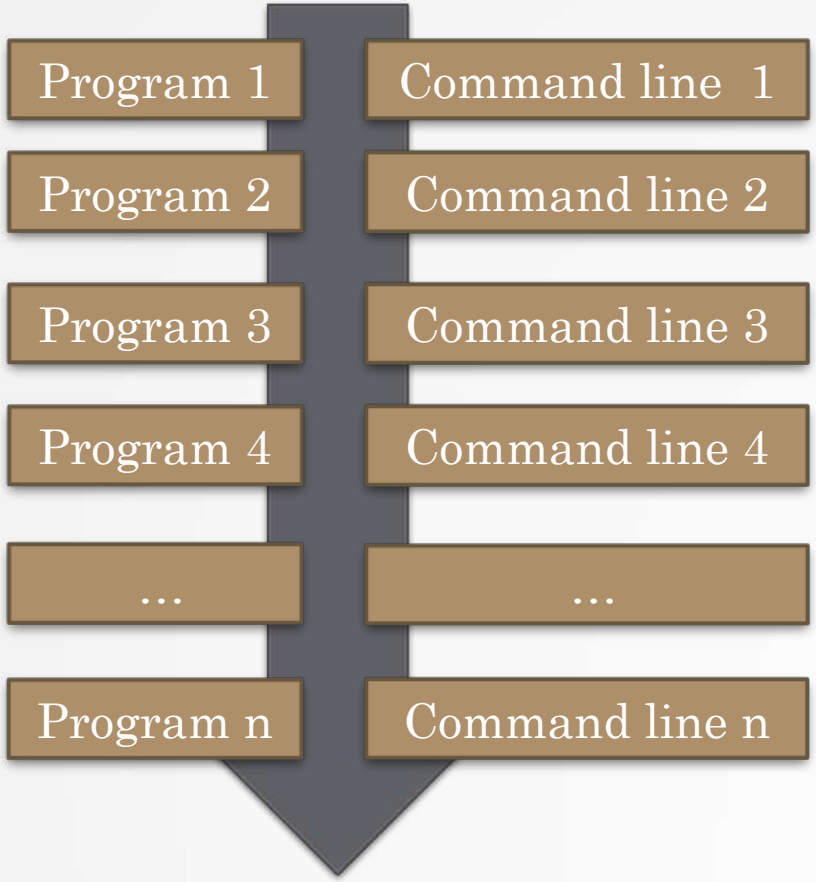
```
macs14 -t treatment.sort.bed -c control.bed -f BED -g mm --name=name1 --llocal=50000
 --slocal=5000 > macs1.nohup 2>&1 &
```

Bioinformatics analyses



Bioinformatics analyses





PIPELINE/ WORKFLOW

Galaxy ?

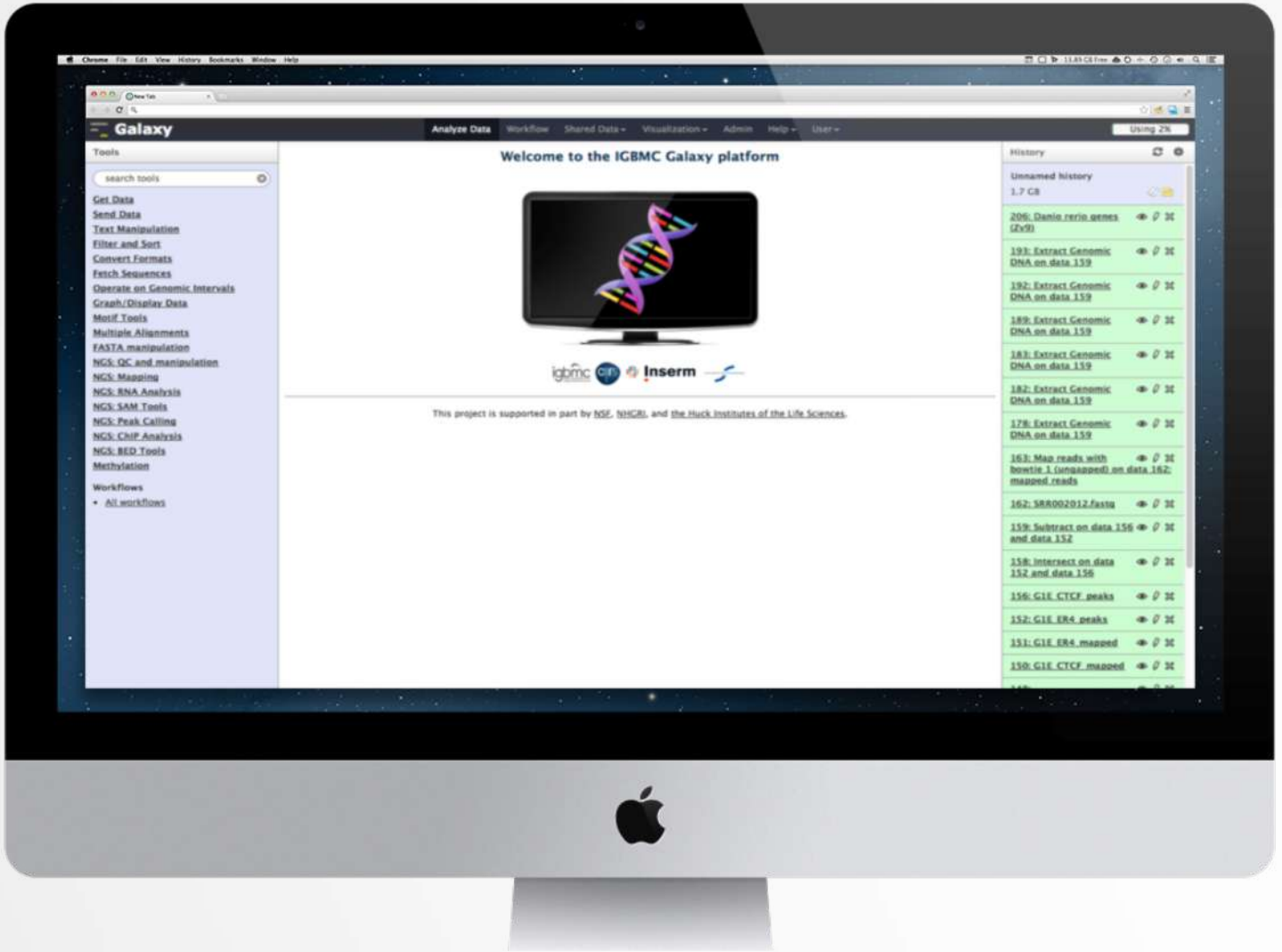

Galaxy
PROJECT

Galaxy project

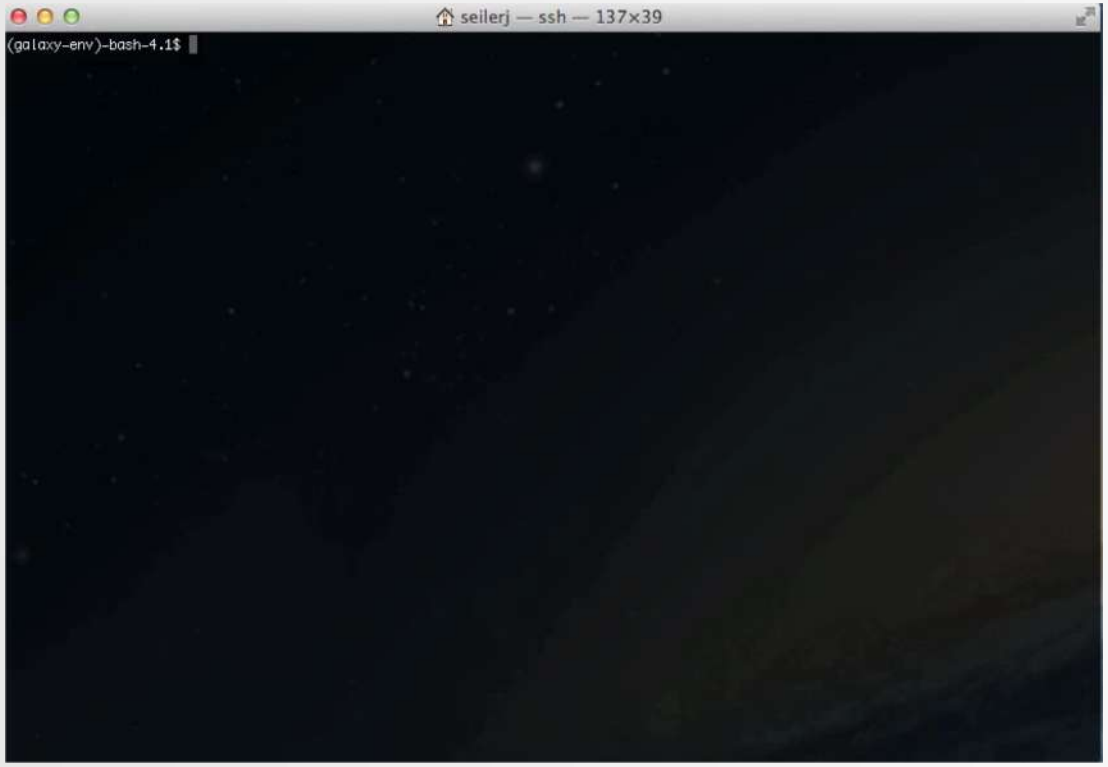
What is Galaxy ?

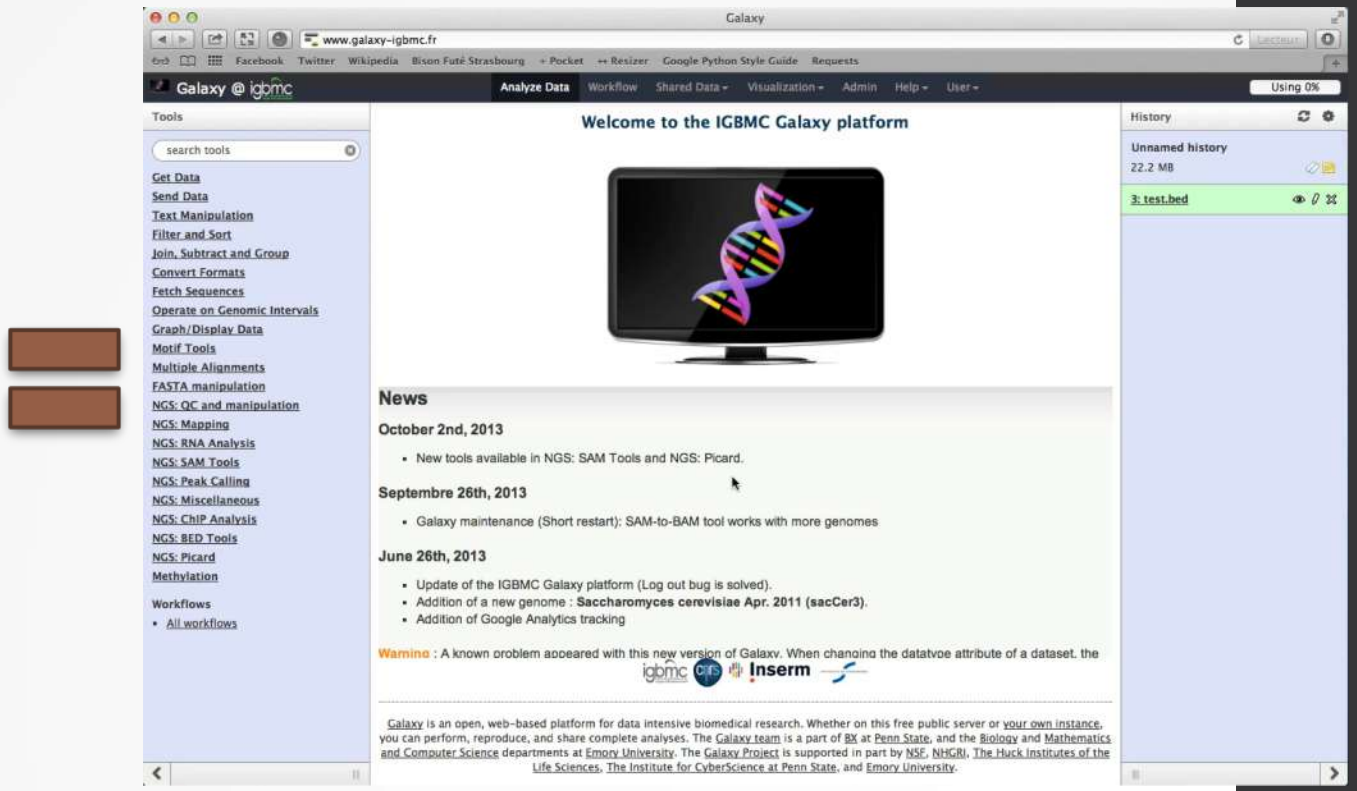
Galaxy is a **computing platform** that enables people to **run complex bioinformatics tools** on a **compute cluster** through a **simple web interface**.

This is Galaxy



Running analyses with tools





Galaxy philosophy

- **Perform**, and **share** complete analysis
- **No programming skills** required
- **Open source** and **free** solution
- **Very large** and **active** community
- **Reproducibility/Usability/Transparency**

How to use Galaxy

Use Galaxy

- **Public servers**
- **Local servers**
- Clouds (Public, Commercial or Academic)
- Docker
- Virtual Machines

Galaxy **public servers**

- Galaxy Project's public server (https://usegalaxy.org/)
- There are several public remote Galaxy instances worldwide (156)
  - Genomics Servers
  - Domain Servers
  - Tool Publishing Servers

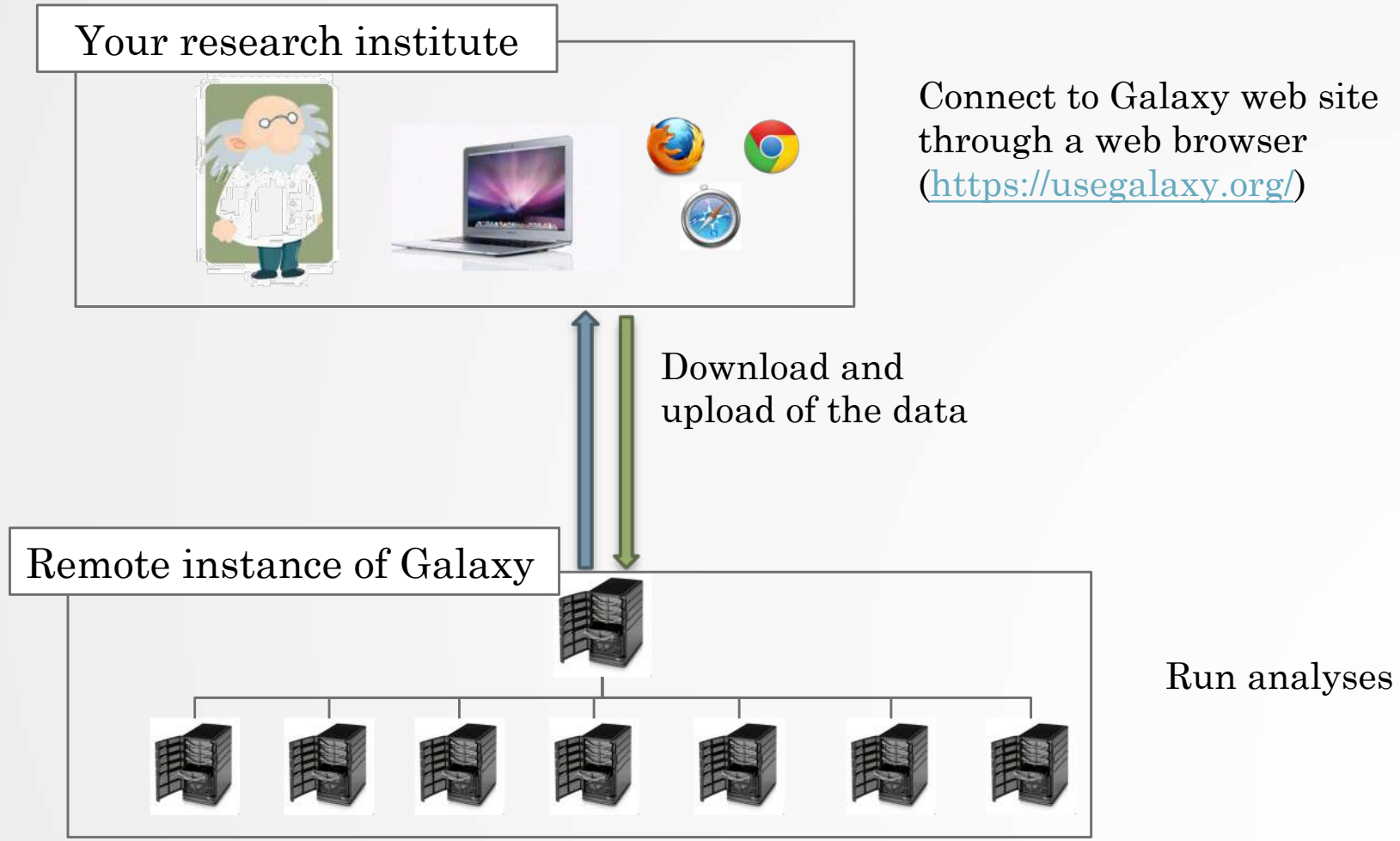
Public Galaxy Servers list :
https://galaxyproject.org/use/
Last Update on: 2019, November 14th

Galaxy **public servers**

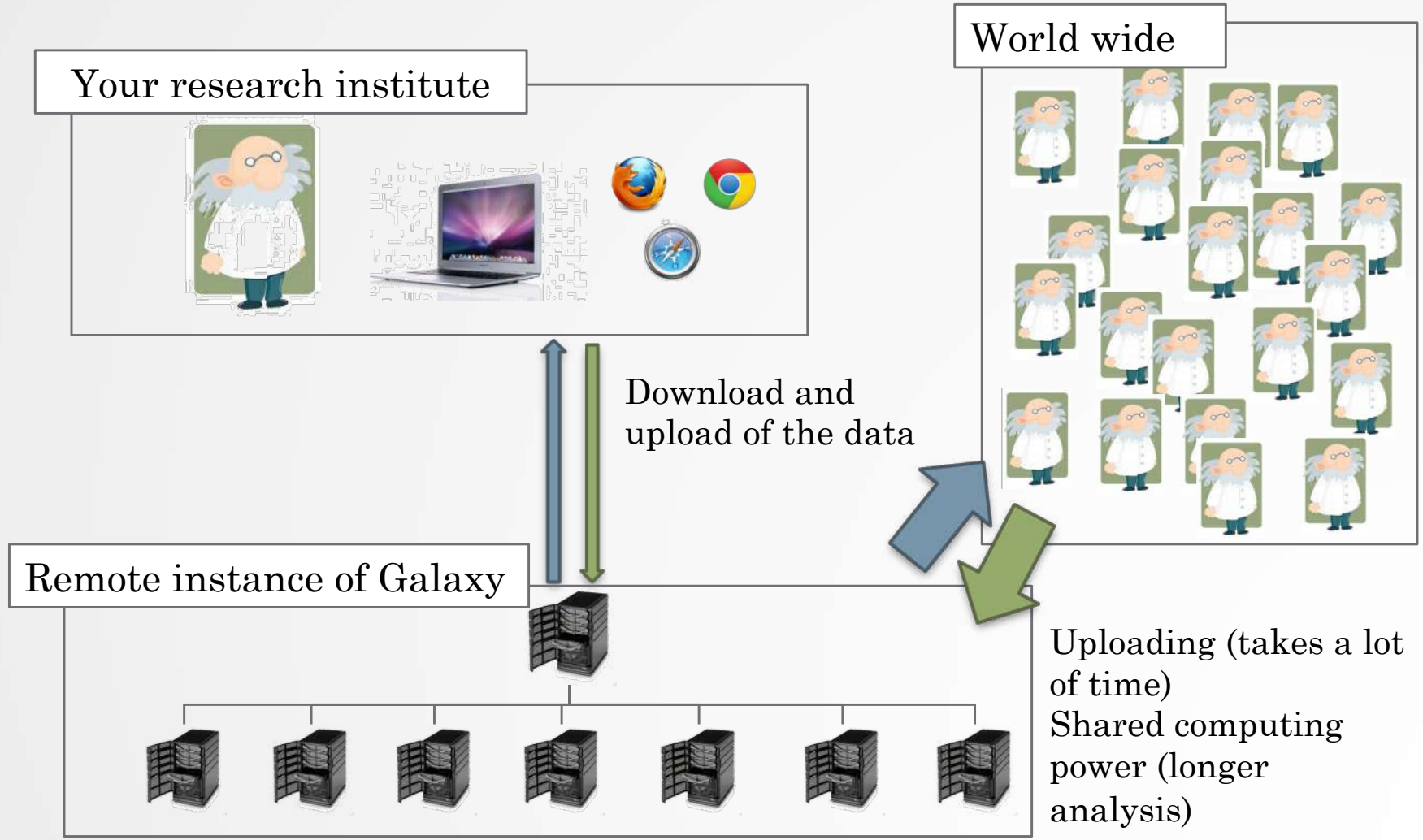
- All analyses are run on remote computing infrastructures
- No need to have a Supercomputer to use Galaxy
- Web browser




Galaxy **public servers**



Galaxy **public servers**

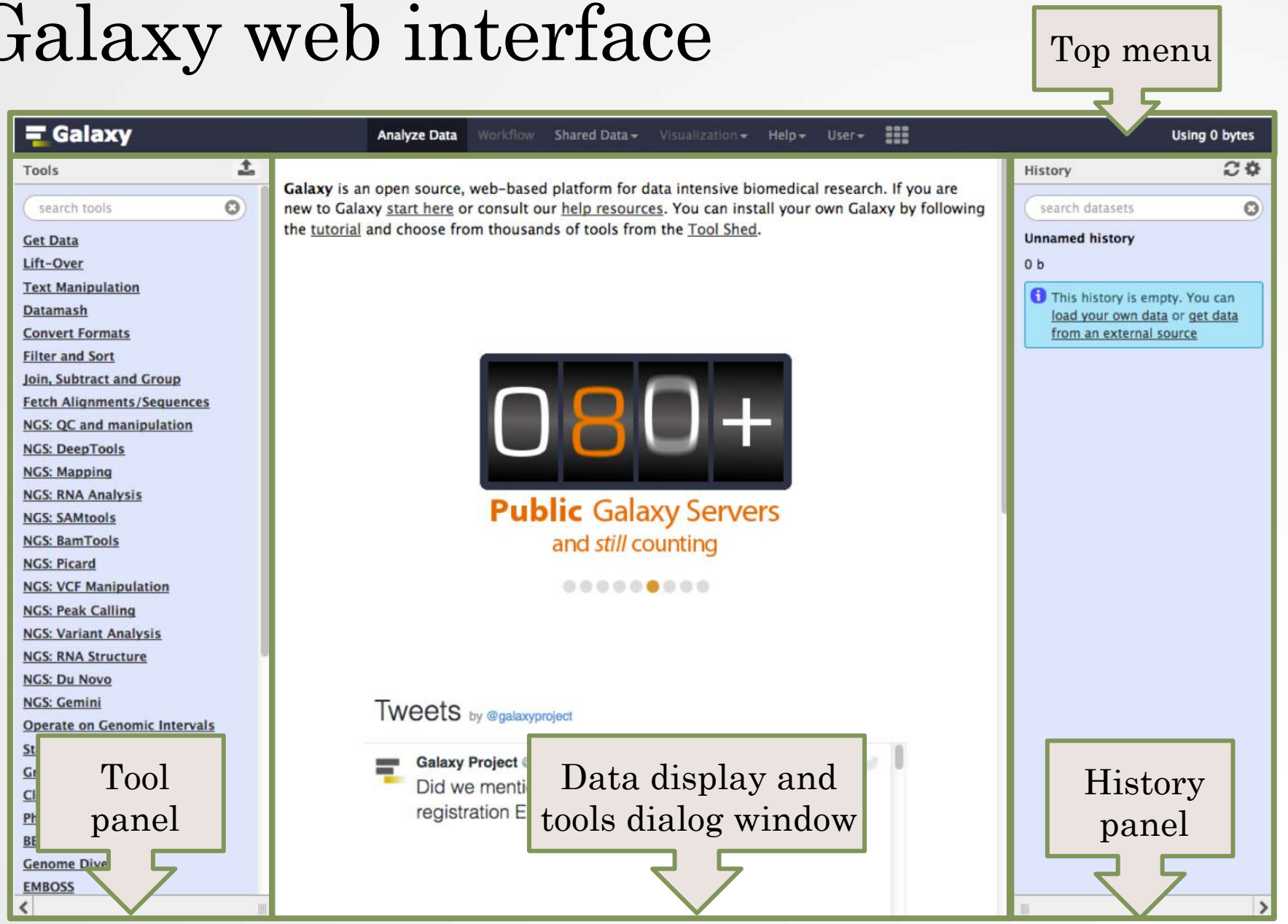


Galaxy **local server**

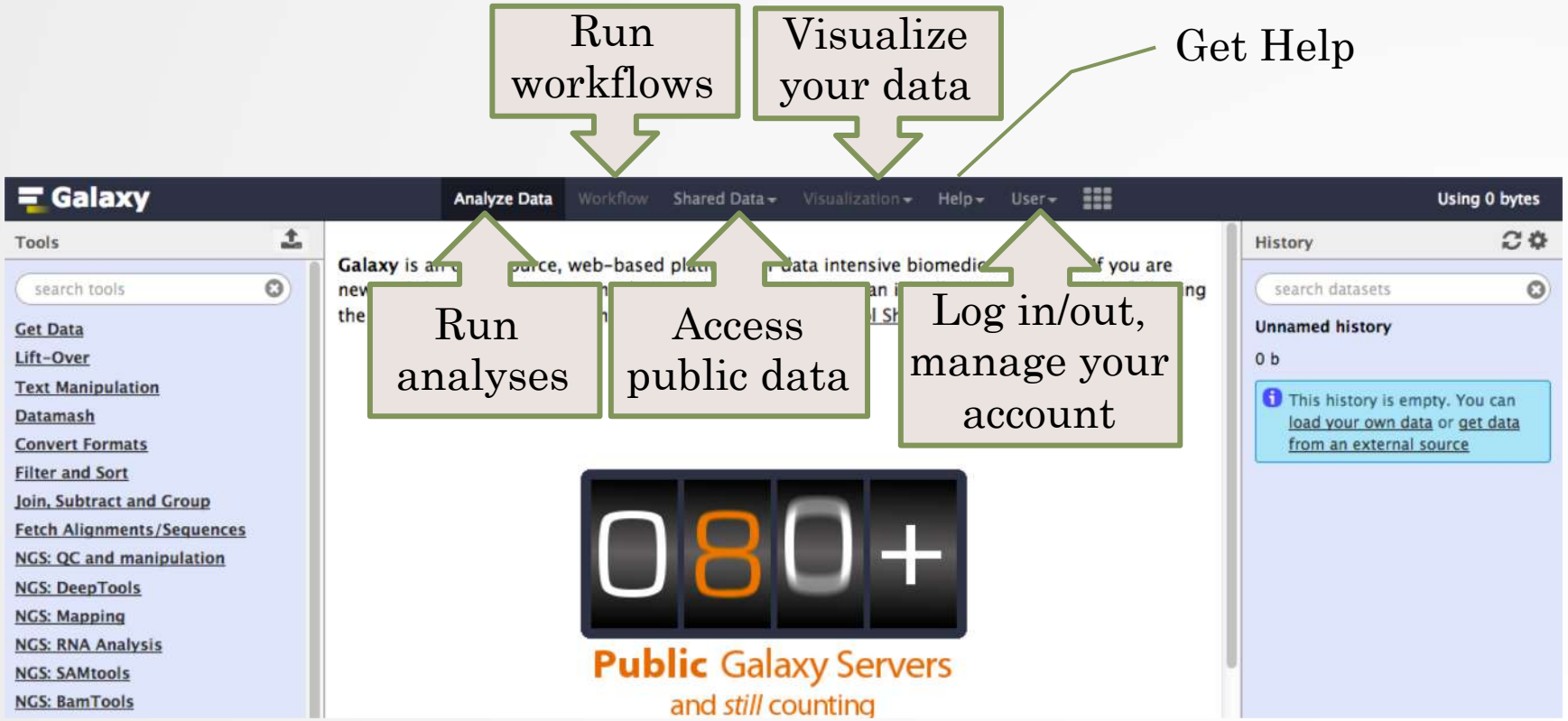
- Run a local production Galaxy because you want to
  - install and use tools unavailable on public Galaxies
  - use sensitive data (e.g. clinical)
  - process large datasets that are too big for public Galaxies
  - Develop Galaxy tools
  - Develop Galaxy itself

Description of the main features of Galaxy

Galaxy web interface

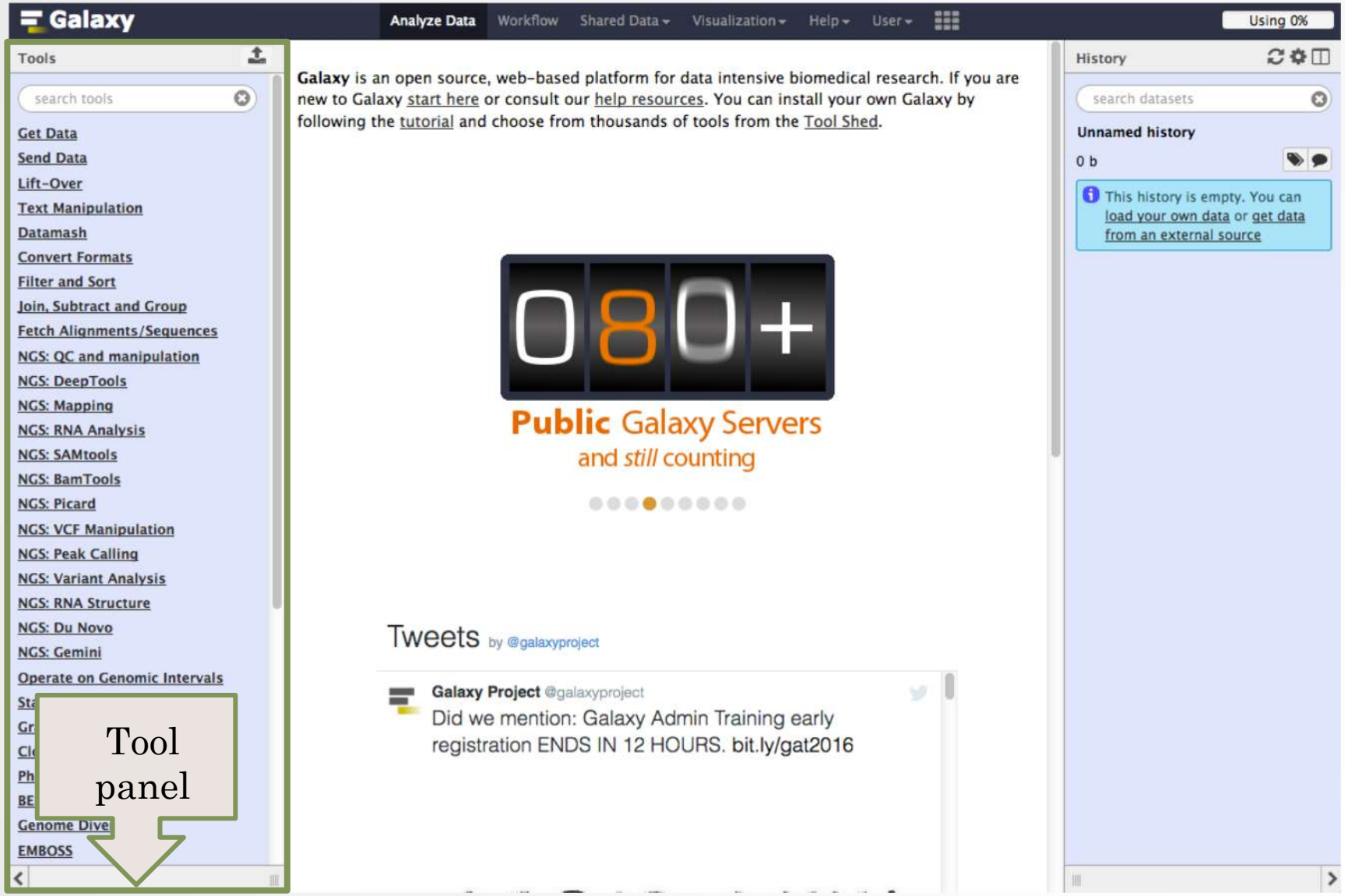


Top menu

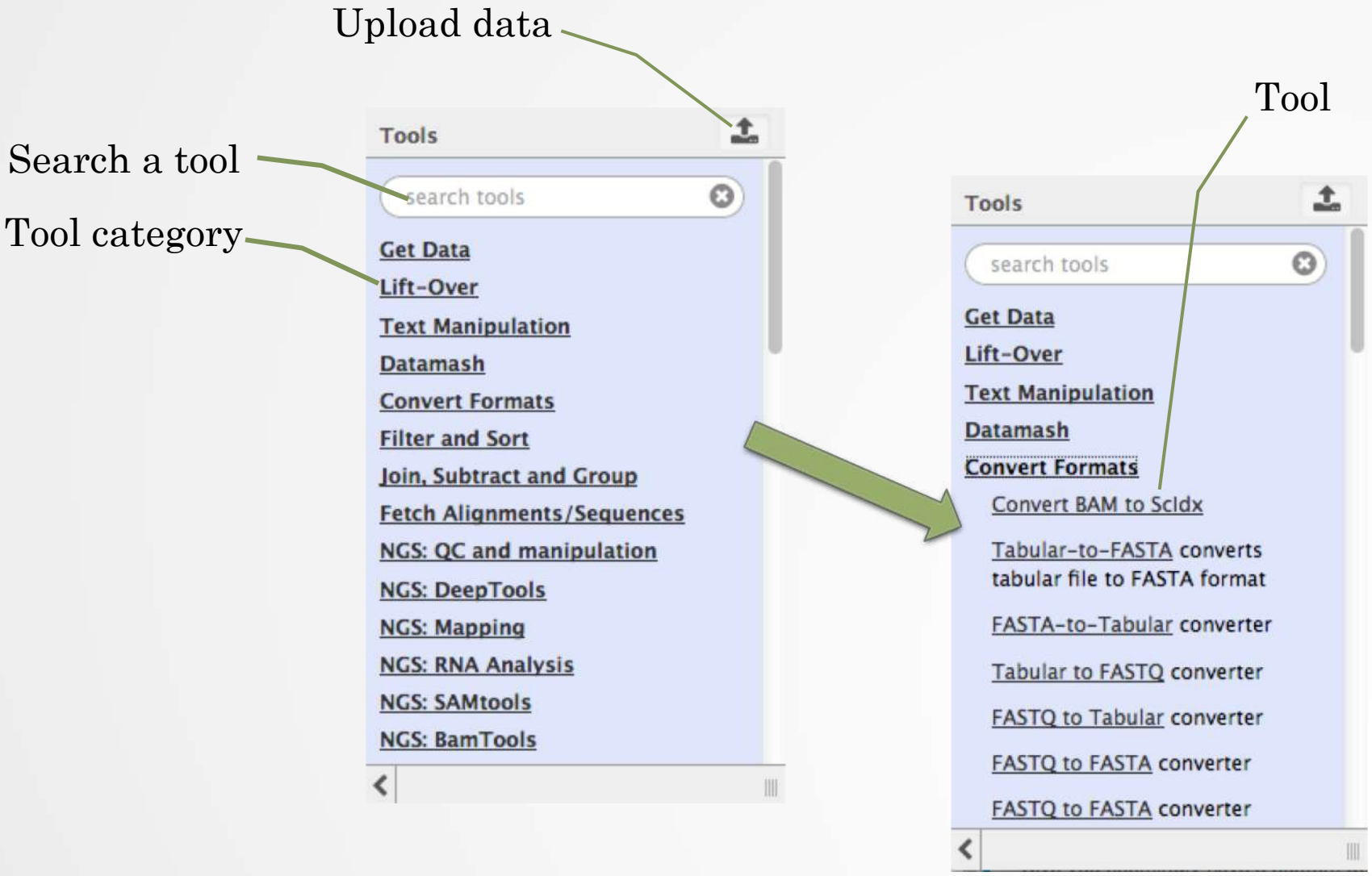

Run workflows
Visualize your data
Get Help
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
Get Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
Run analyses
Access public data
Log in/out, manage your account
History
search datasets
Unnamed history
0 b
This history is empty. You can load your own data or get data from an external source
Public Galaxy Servers
and still counting

Exercise 1

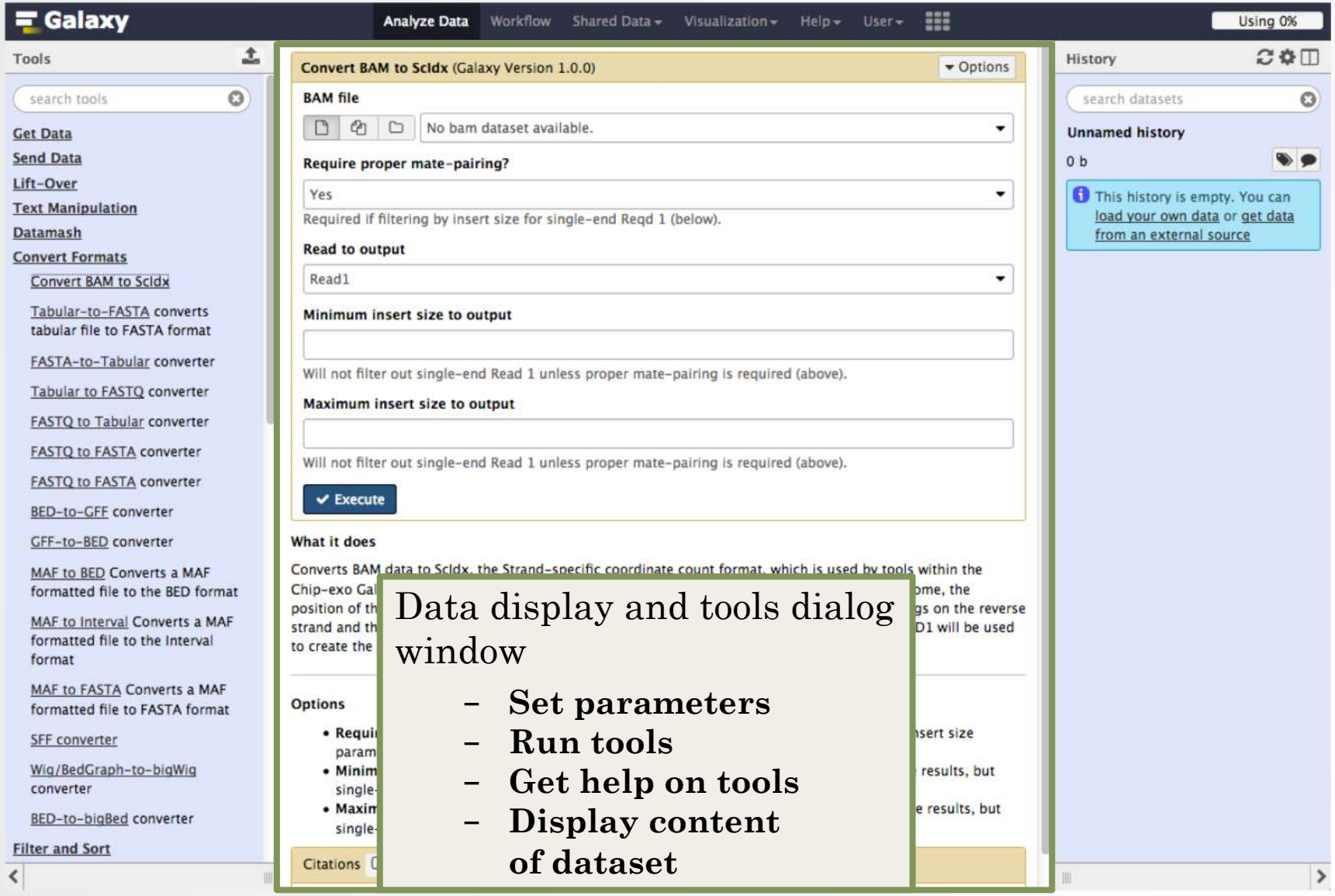
Tool Panel / Run analyses



Tool Panel / Run analyses



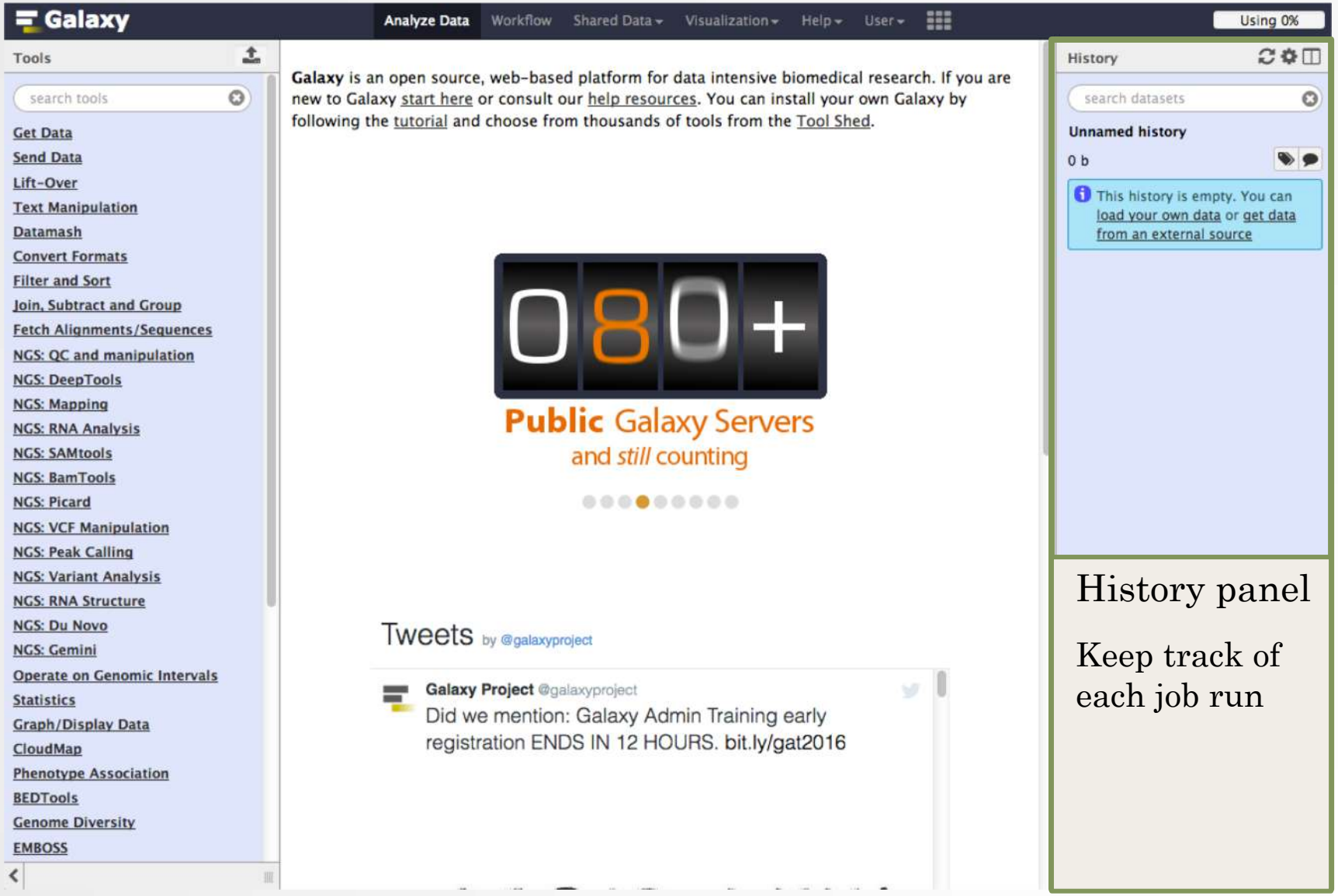
Tools dialog window



Data display and tools dialog window

- **Set parameters**
- **Run tools**
- **Get help on tools**
- **Display content of dataset**

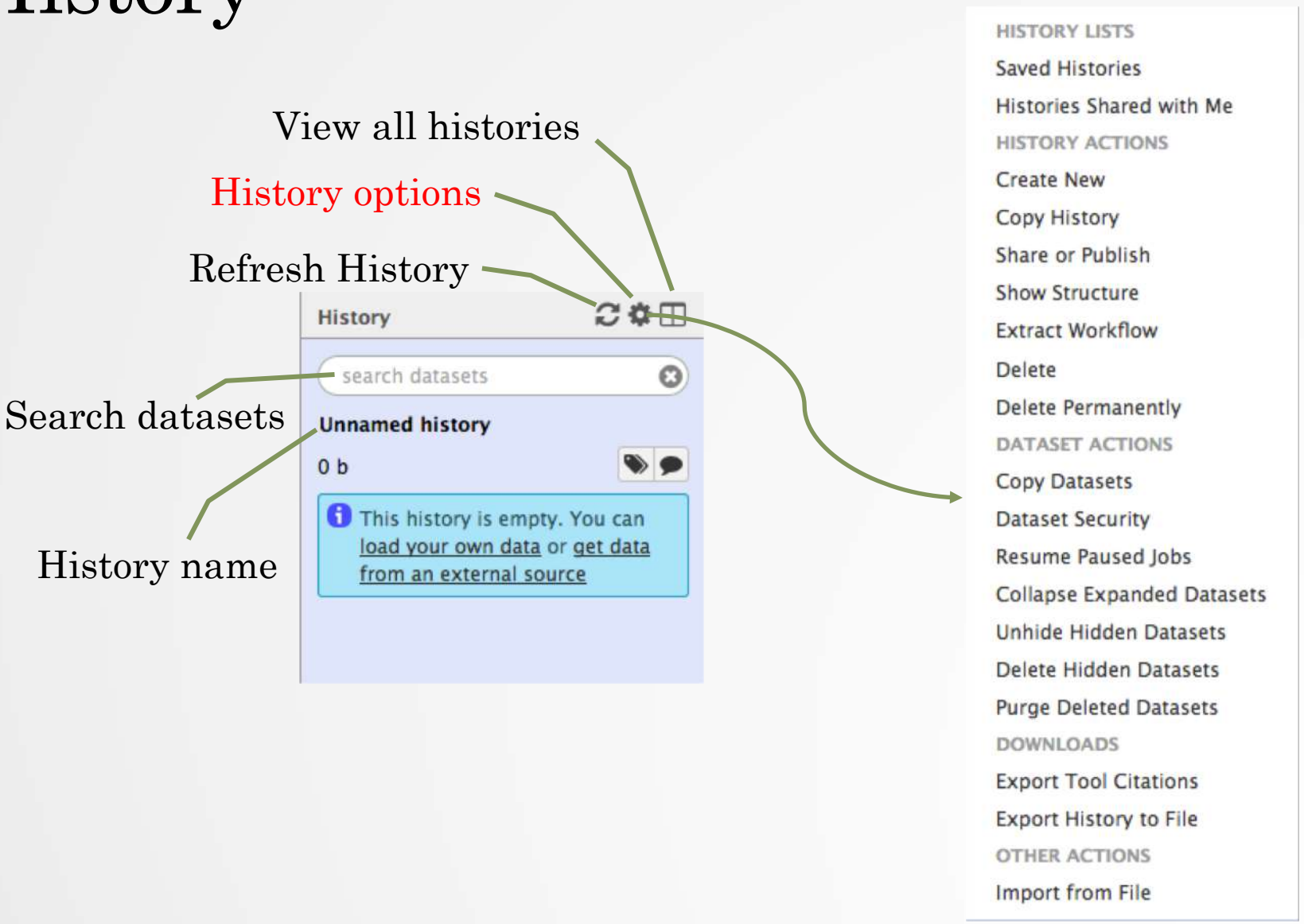
History



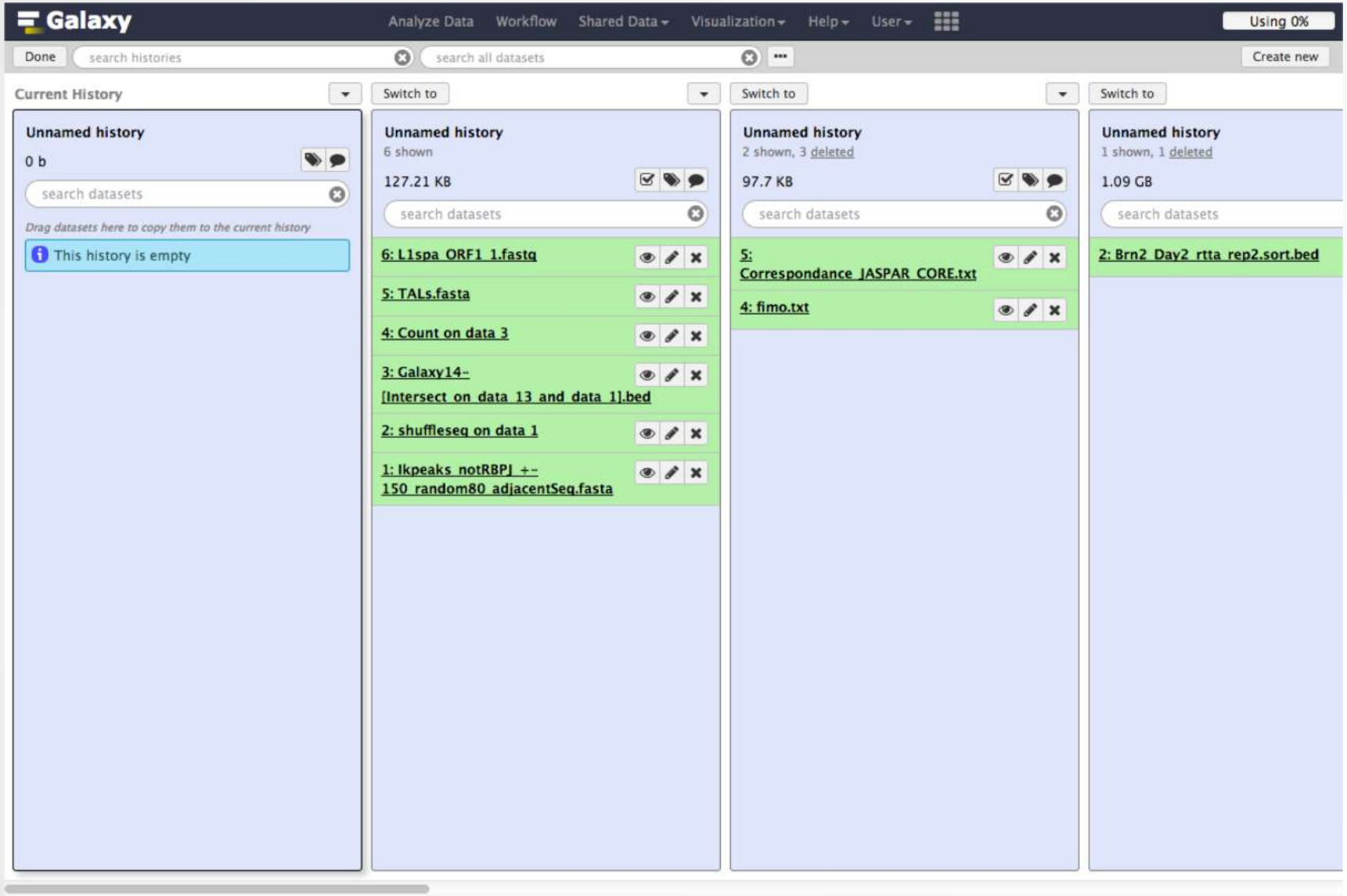
History panel

Keep track of each job run

History


View all histories
History options
Refresh History
Search datasets
History name
History
search datasets
Unnamed history
0 b
This history is empty. You can load your own data or get data from an external source
HISTORY LISTS
Saved Histories
Histories Shared with Me
HISTORY ACTIONS
Create New
Copy History
Share or Publish
Show Structure
Extract Workflow
Delete
Delete Permanently
DATASET ACTIONS
Copy Datasets
Dataset Security
Resume Paused Jobs
Collapse Expanded Datasets
Unhide Hidden Datasets
Delete Hidden Datasets
Purge Deleted Datasets
DOWNLOADS
Export Tool Citations
Export History to File
OTHER ACTIONS
Import from File

View all histories

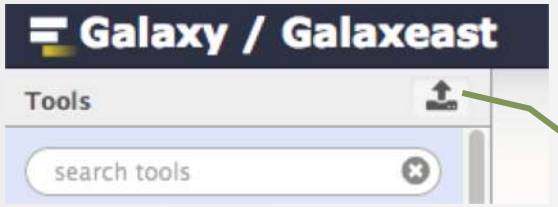


Exercise 2

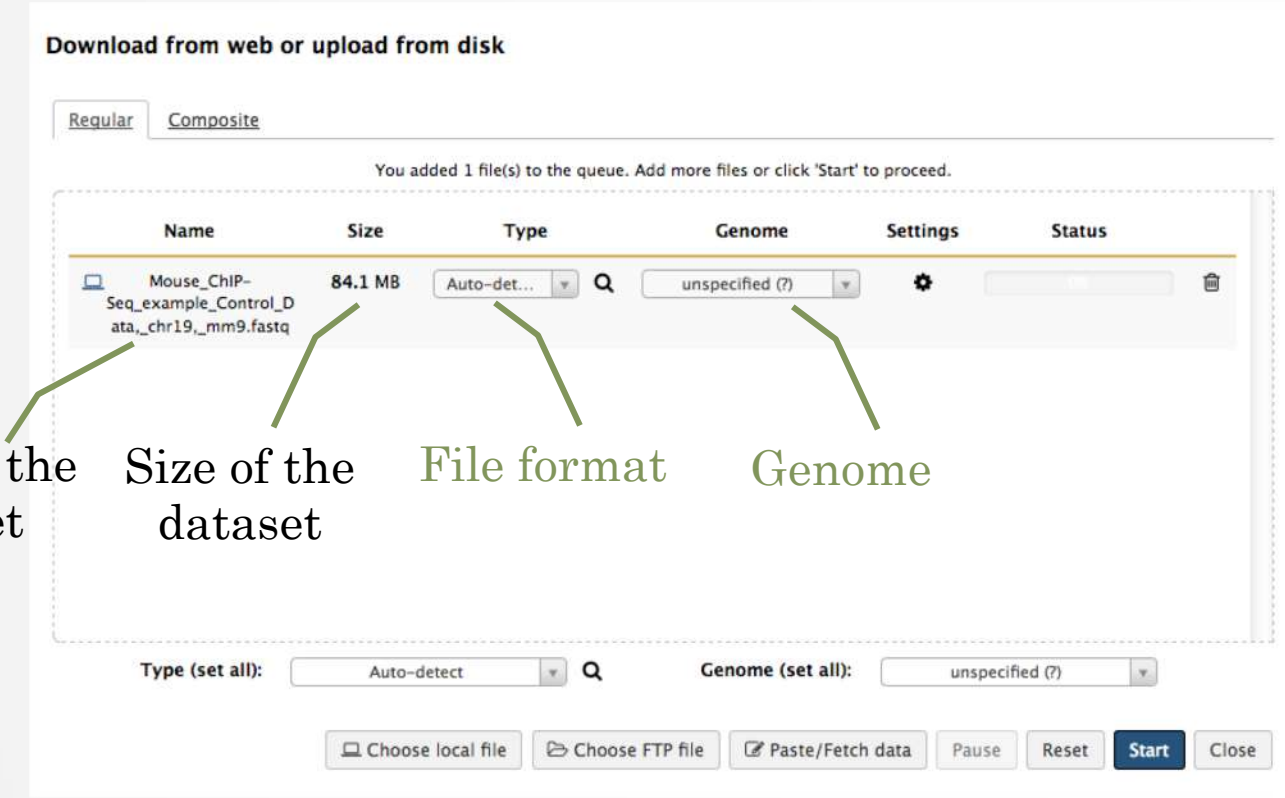
How to import data into Galaxy

1. Your own data (from your computer)
2. Shared data
3. Data from external sources

1. Import your own data to Galaxy



Display the drag and drop utility used to upload local files



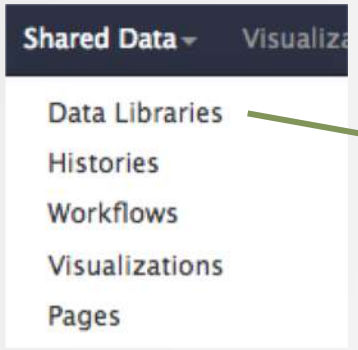
Name of the dataset

Size of the dataset

File format

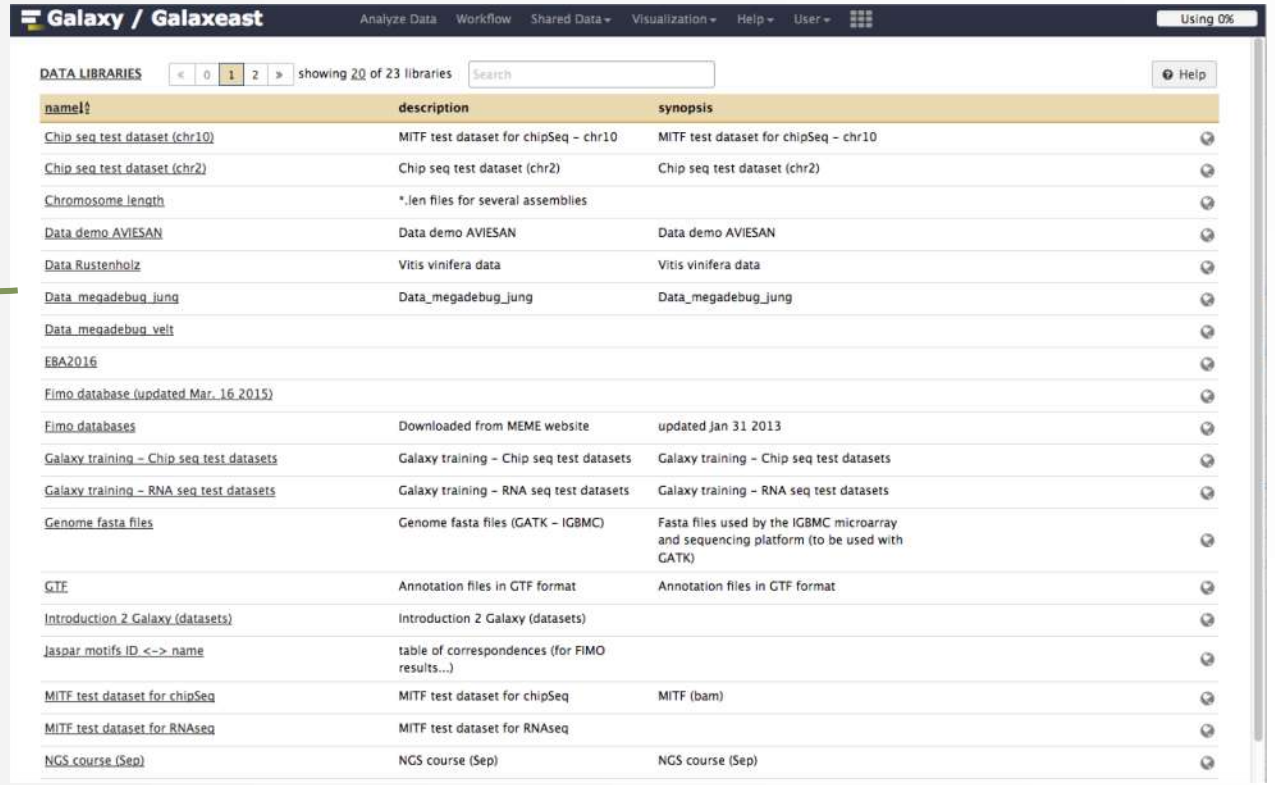
Genome

2. Import shared data (data libraries)



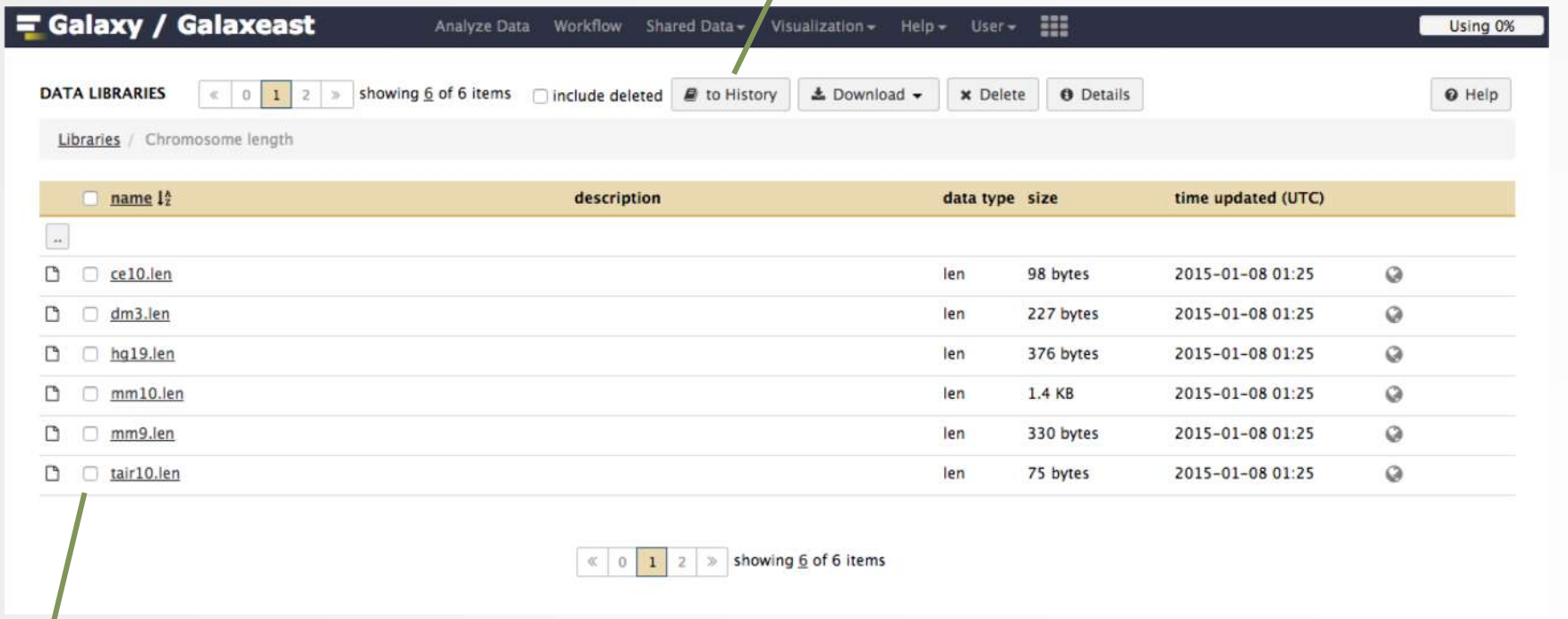
(Top menu)

Data Libraries



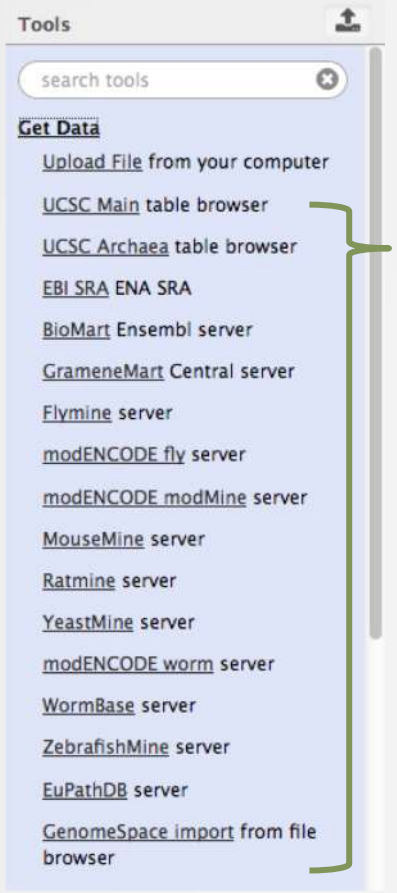
2. Import shared data (data libraries)

2. Import selected dataset to history

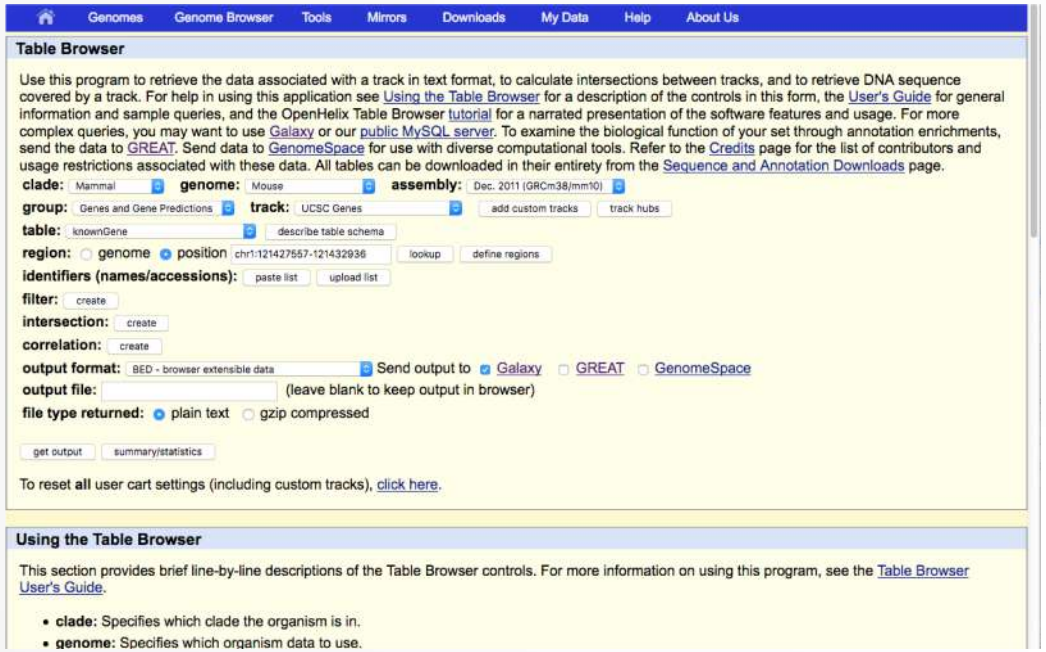


1. Select dataset

3. Import public data



Browse and import external data from public databases



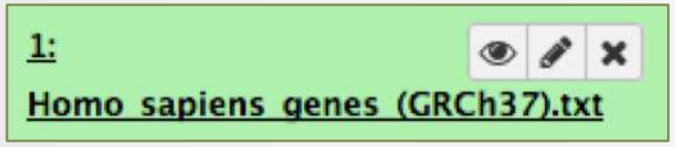
Exercise 3.1

Datasets/Jobs in the History

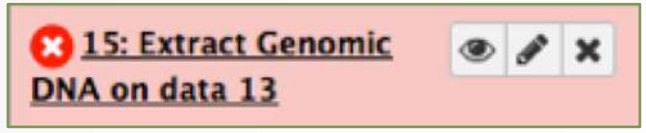
Grey: the job is waiting to run

Yellow: the job is running

Green: the job is successfully done



Red: the job encountered a problem

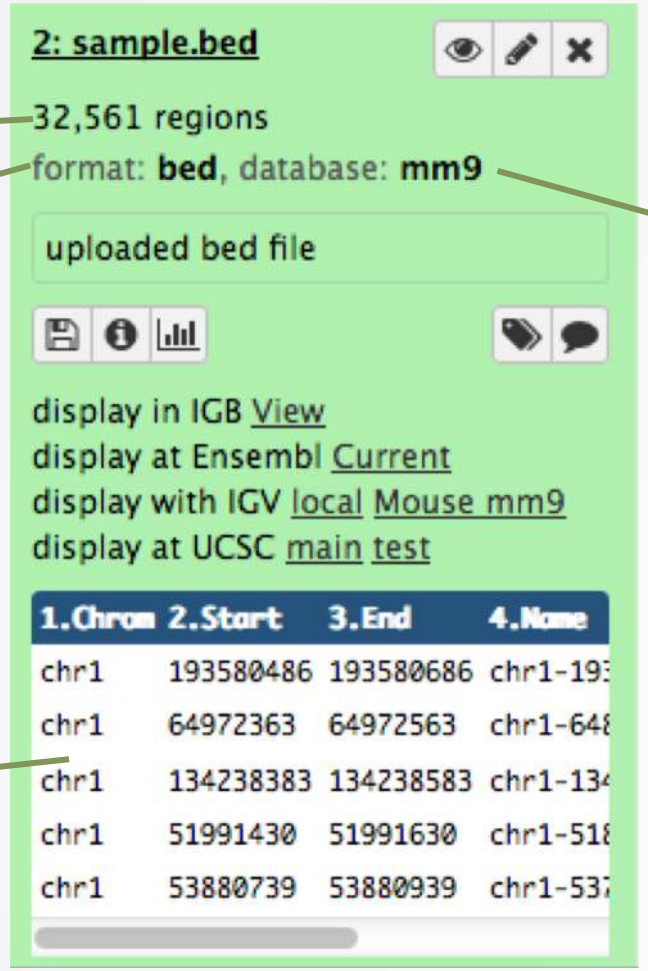


Datasets/Jobs in the History

Number of lines in the file or size of the file

Format

Genome

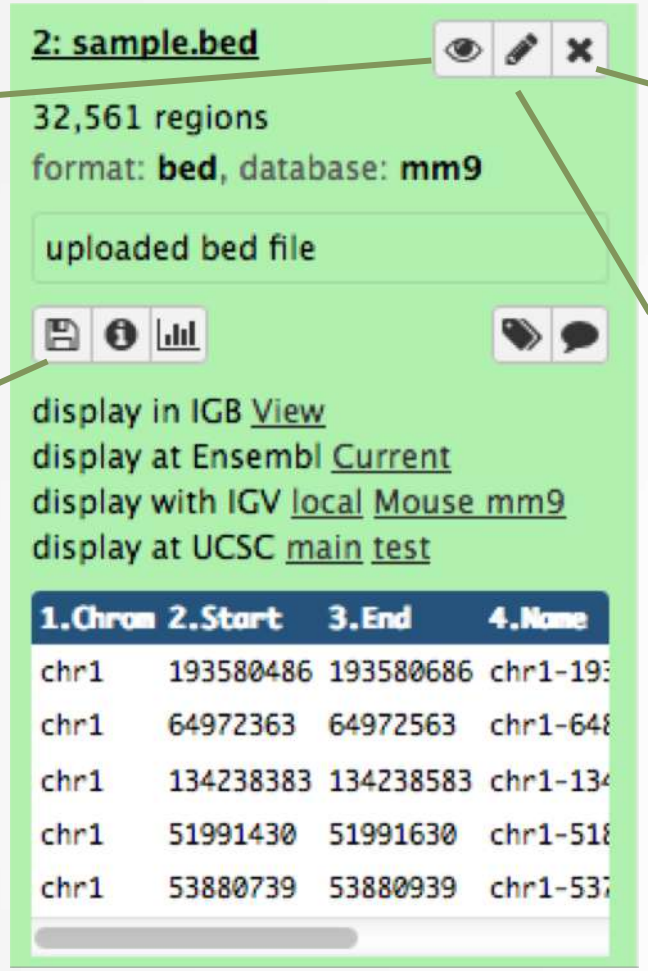


If the dataset is a text file, the first lines of the file are displayed

Datasets/Jobs in the History

View dataset (if possible) in the middle panel of Galaxy

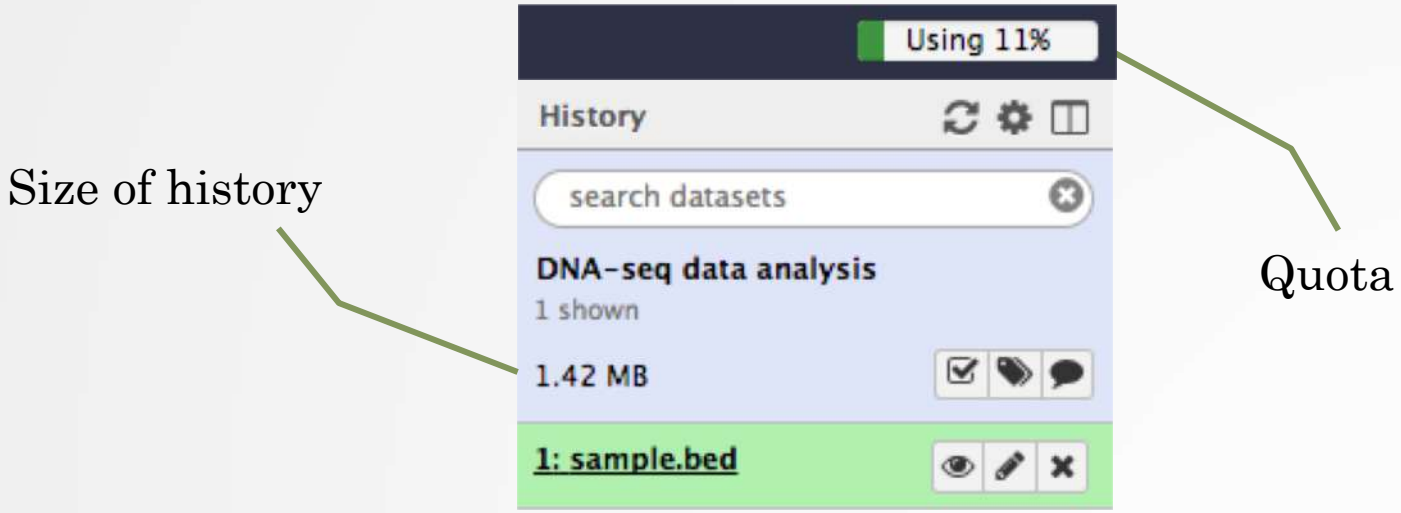
Download dataset



Delete dataset

Edit attributes of the dataset (change name, format, genome, permission)

Size of histories and quota



Exercise 3.2

Exercise 4

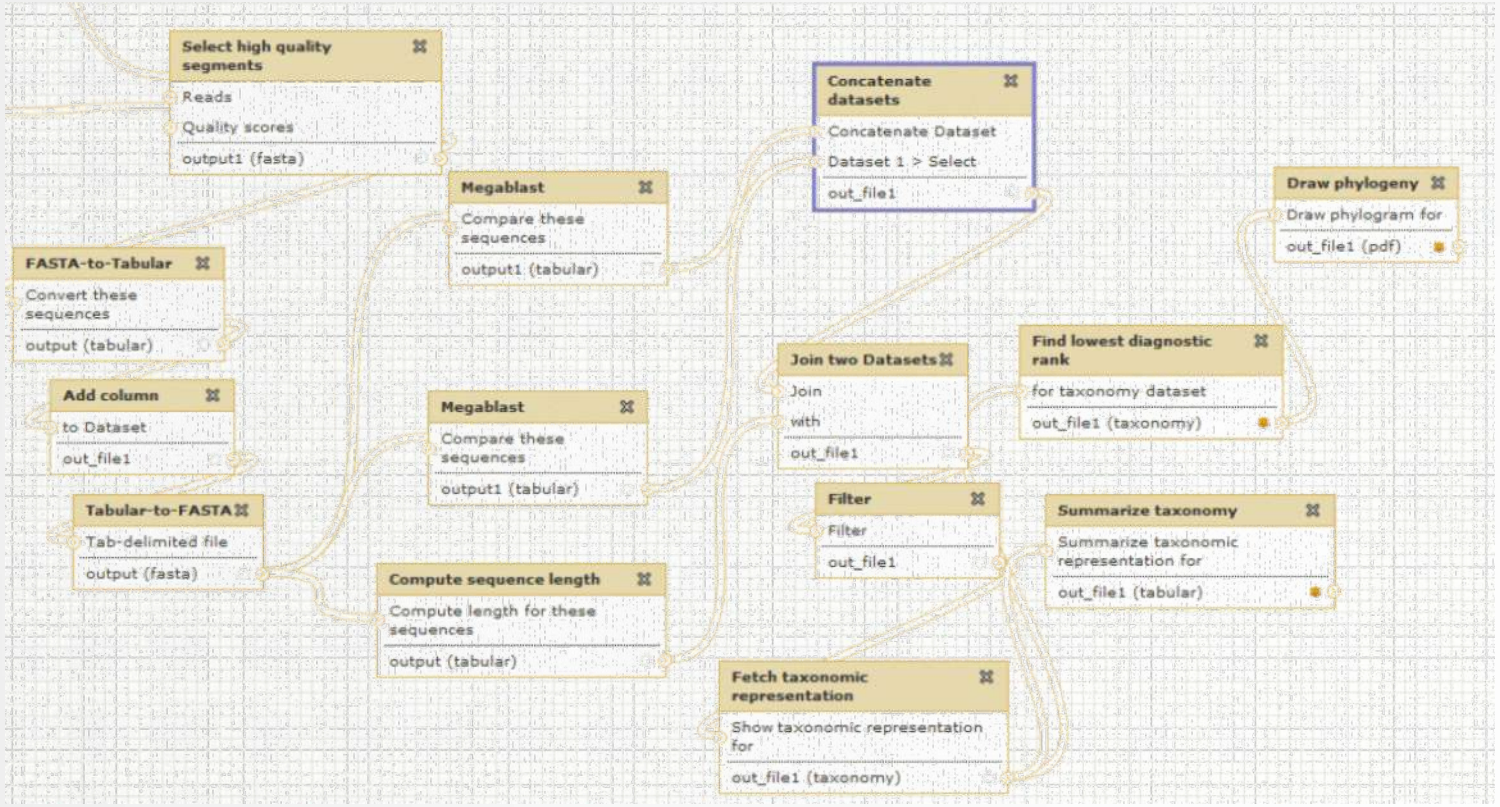
Exercise 5

Exercise 6

Workflow

What if we'd mix all together

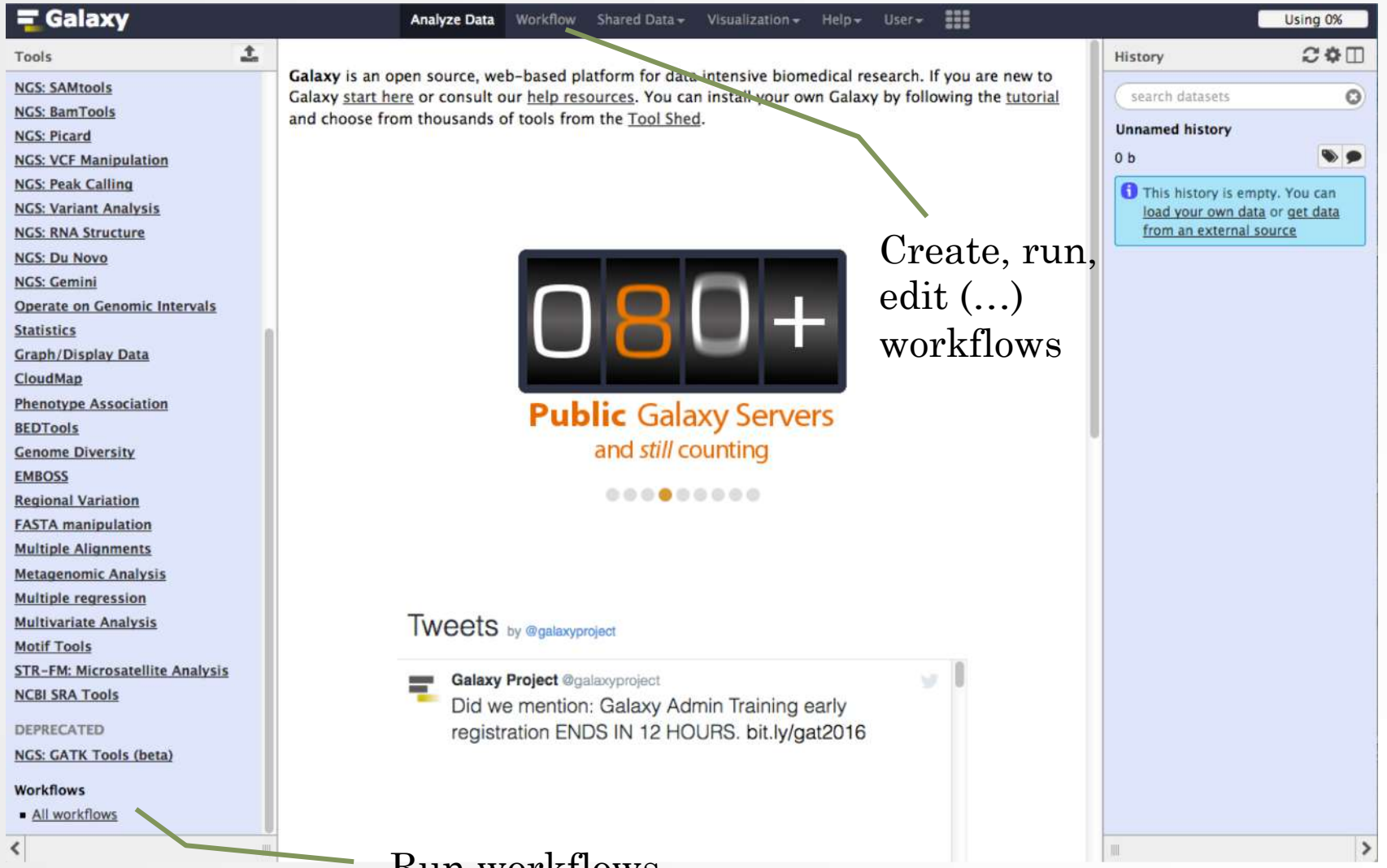
Galaxy workflow



Galaxy workflows

- Workflow:
  - Analysis protocol with several steps (tools)
  - The output of a step is used as the input of the next next so file formats between two steps should be compatible!
- Workflows are often made general so that they can be run on various datasets
- Some of the parameters are pre-defined while others are set at runtime

Workflows

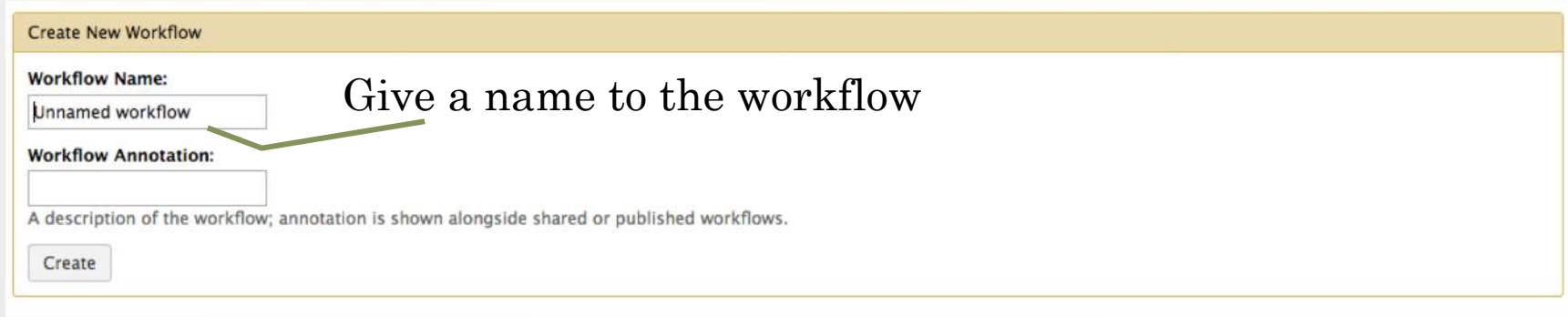


Create, run, edit (…) workflows

Run workflows

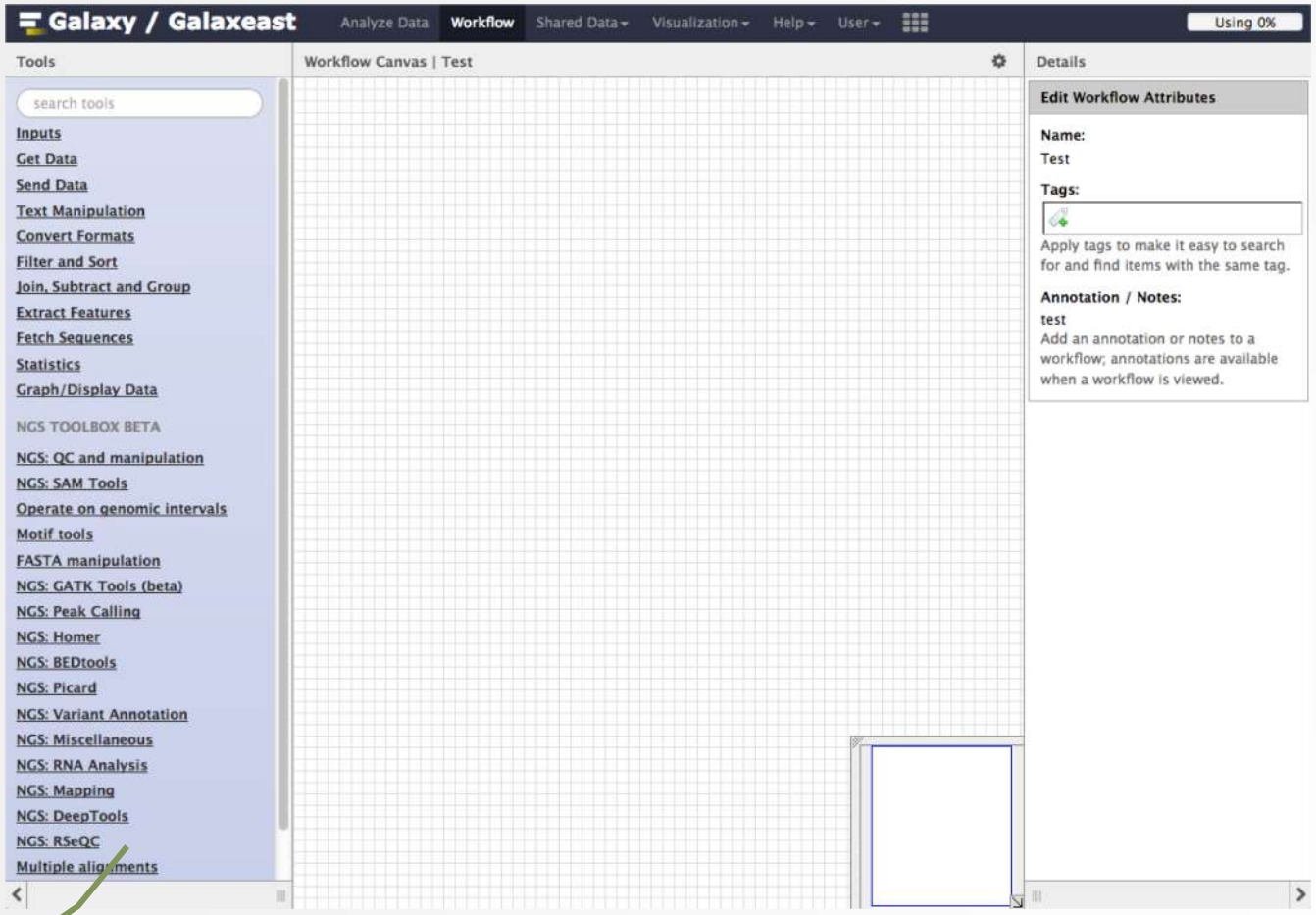
Workflows

Create workflows



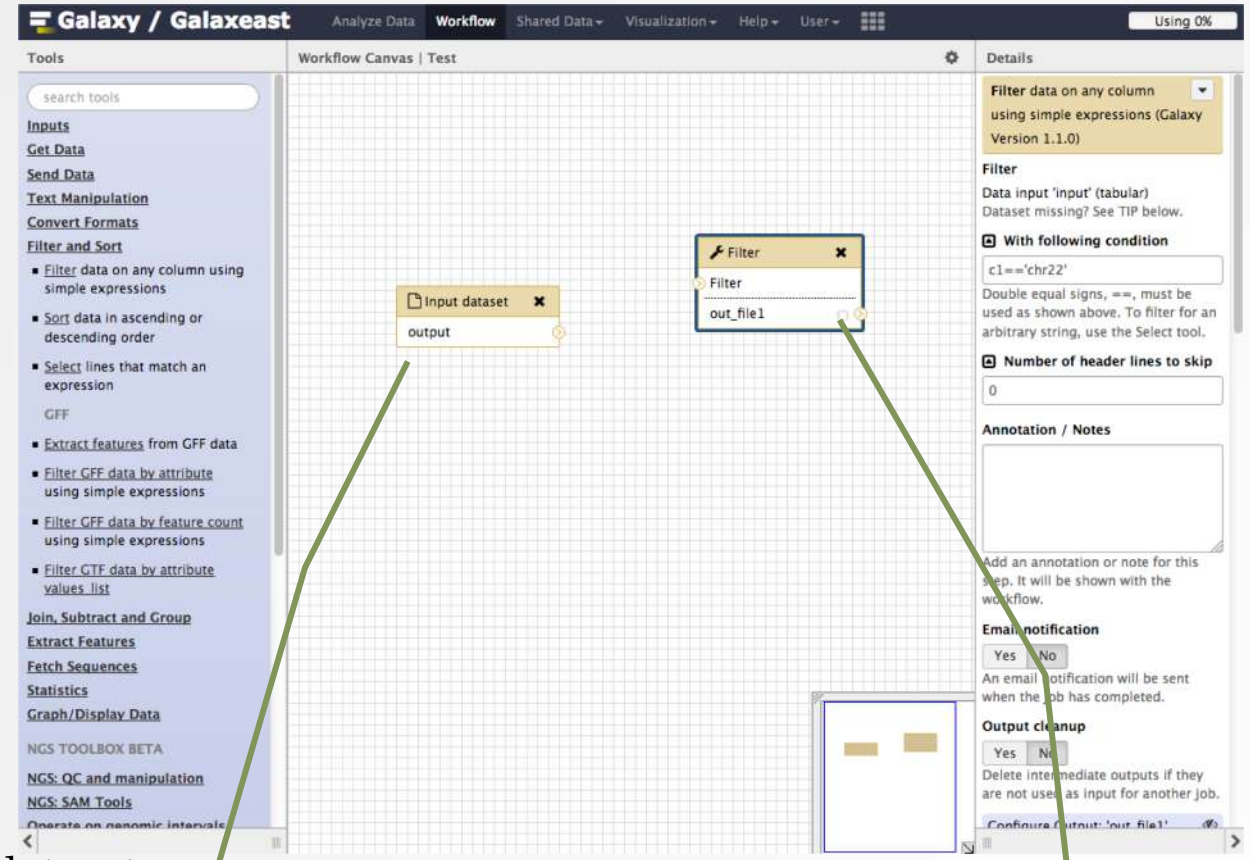
Give a name to the workflow

Workflow creation



Add tools or input datasets to the workflow

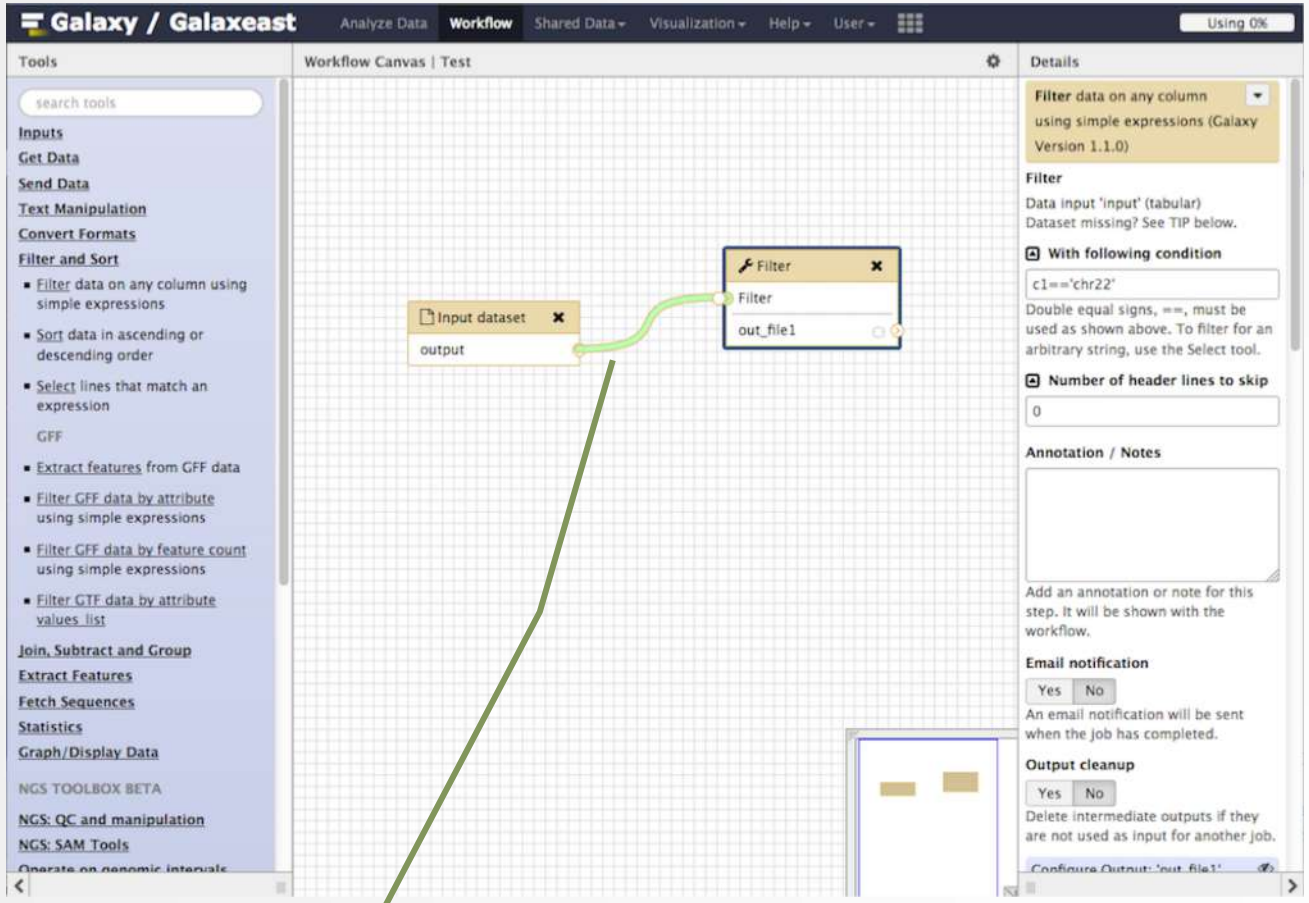
Workflow creation



Input dataset.
Most of the time, a workflow starts with an input dataset to which analyses are applied.
In Galaxy, the file format of the input dataset will be limited to the input file format of the subsequent step

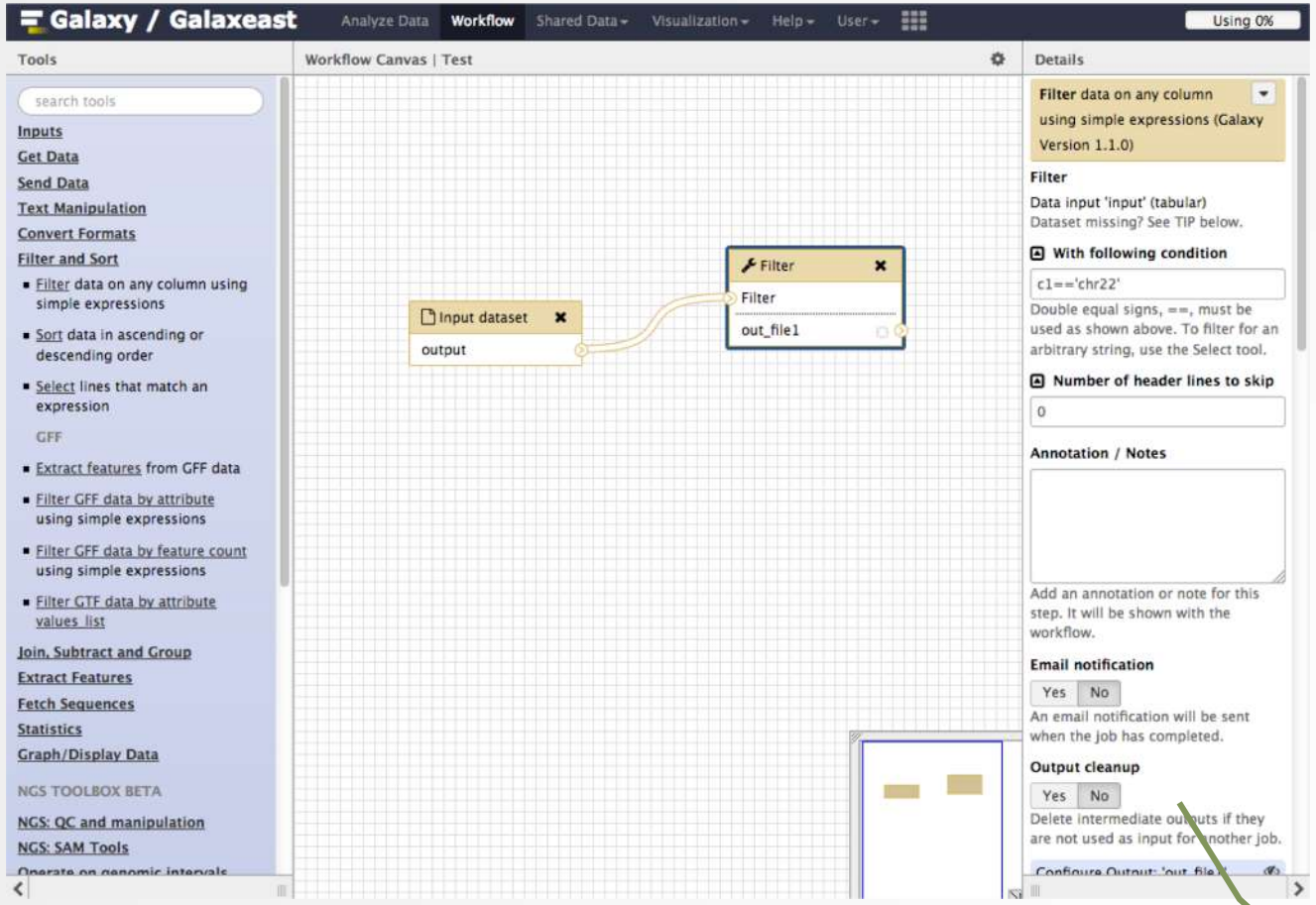
Tool to be run

Workflow creation



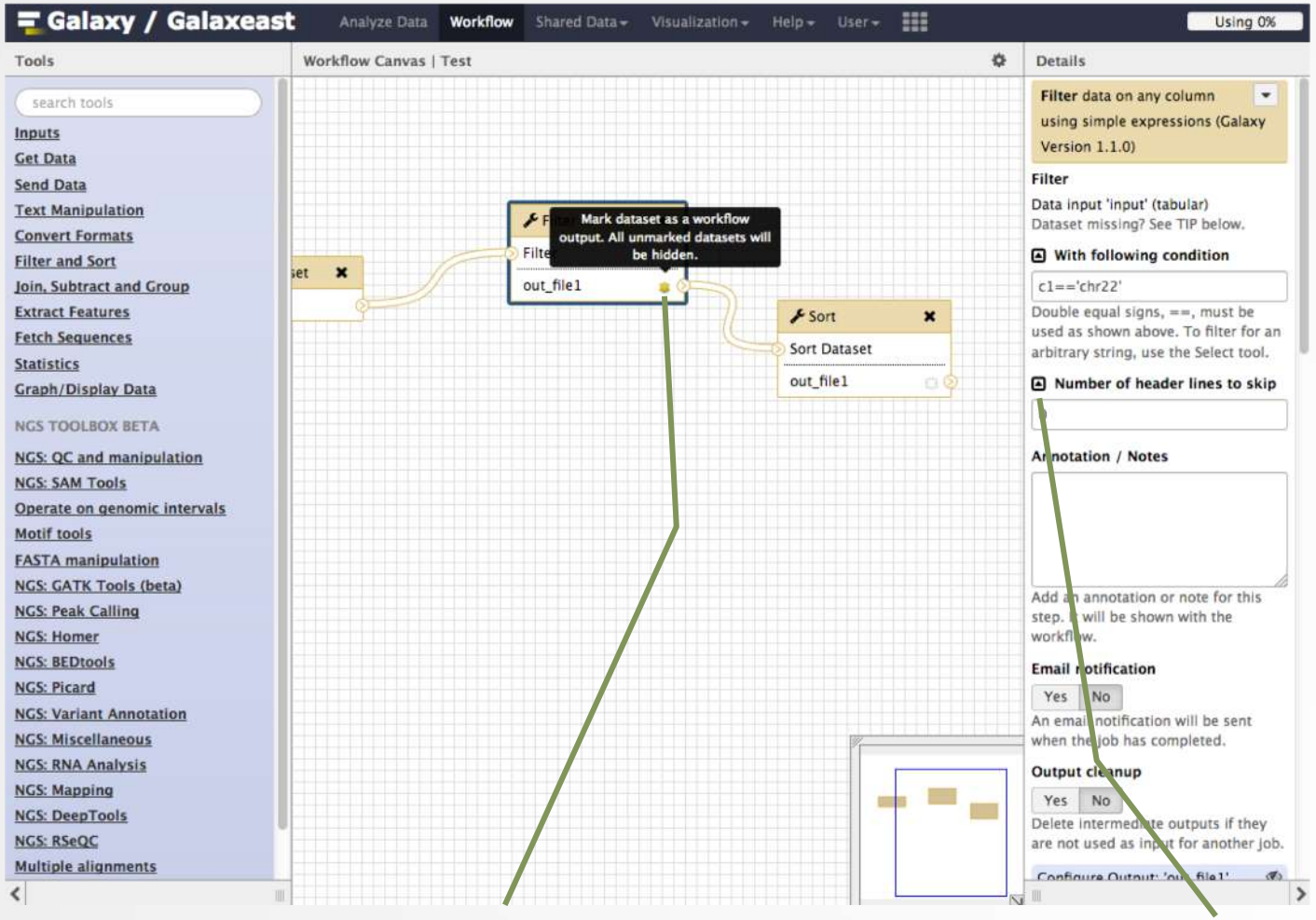
If two steps can be linked together, the link between the two boxes is green

Workflow creation



Pre-configure tool parameters and configure parameters to be set at run time

Workflow creation

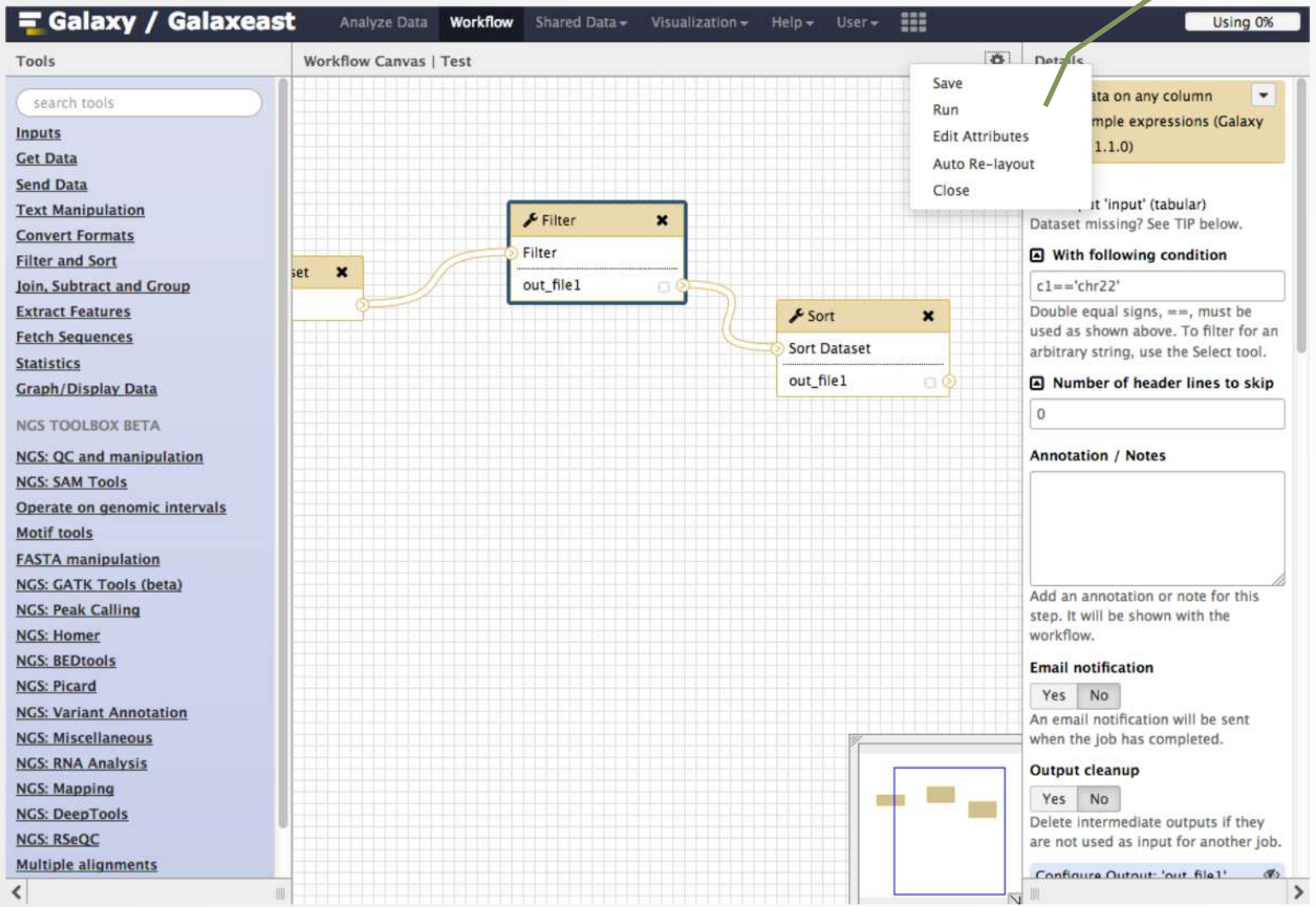


Click on star to select which datasets will be displayed in the history generated when running of the workflow

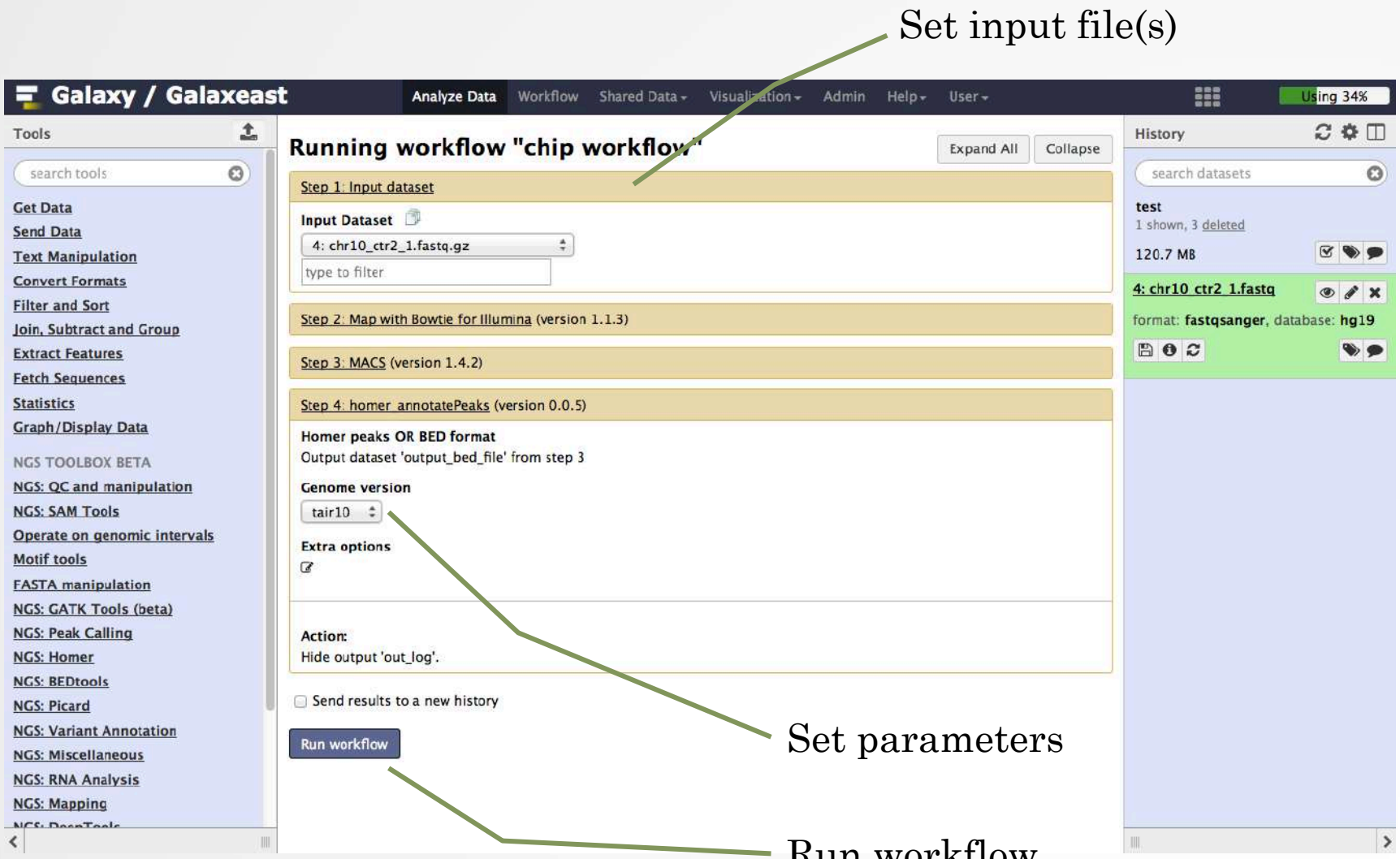
Click to get the parameter to be set at runtime

Workflow creation

Save, run workflows



Run workflows



Exercise 7

Exercise 8

Privacy

- By default datasets, workflows, histories are private to the user that generated/uploaded them.
- They can be shared across Galaxy users (of the same Galaxy instance) or via links