Introduction to Galaxy

Stéphanie Le Gras
(slegras@igbmc.fr)

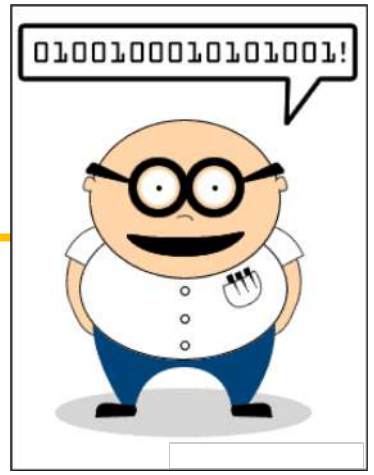
Guidelines

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

Analyzing biological data with informatics tools

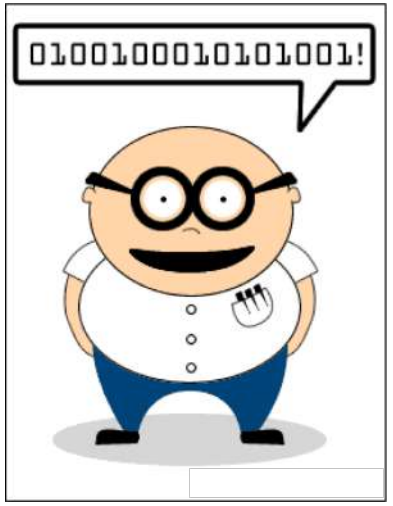
Bioinformatics analyses

Informatics data



nature

Bioinformatics analyses



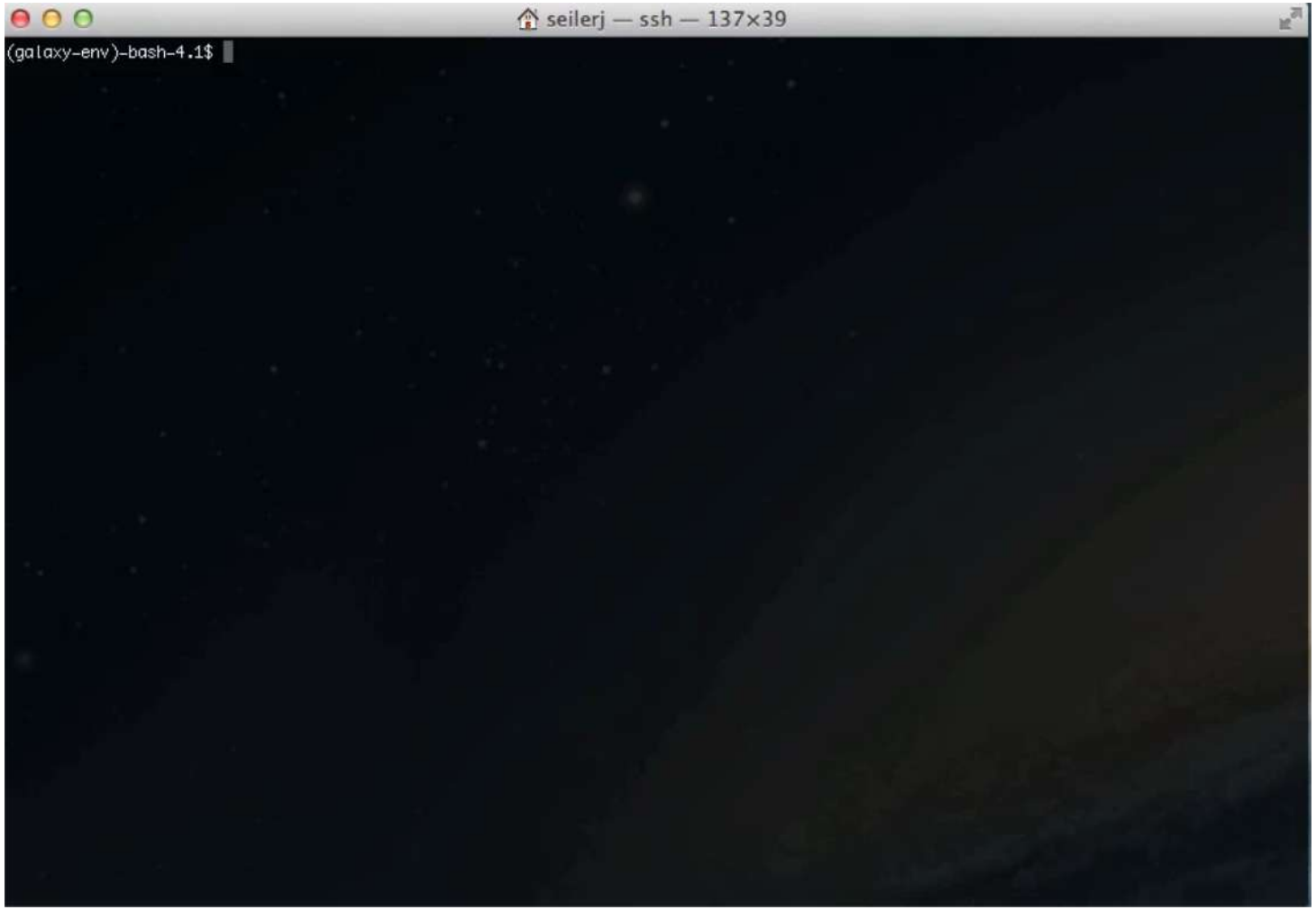
Scripts, softwares

```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.
```

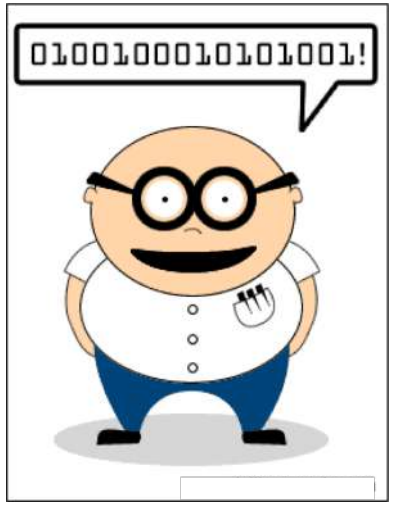
Command line

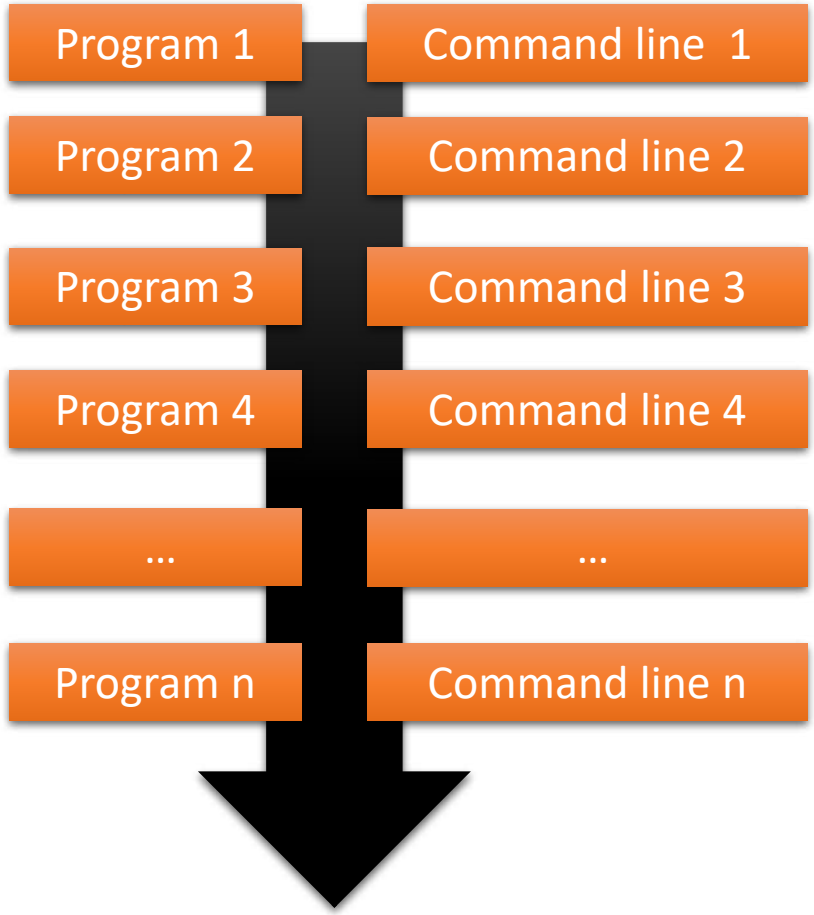
```
macs14 -t treatment.sort.bed -c control.bed -f BED -g mm --name=name1 --llocal=50000
 --slocal=5000 > macs1.nohup 2>&1 &
```

Bioinformatics analyses



Bioinformatics analyses





PIPELINE / WORKFLOW

Galaxy ?

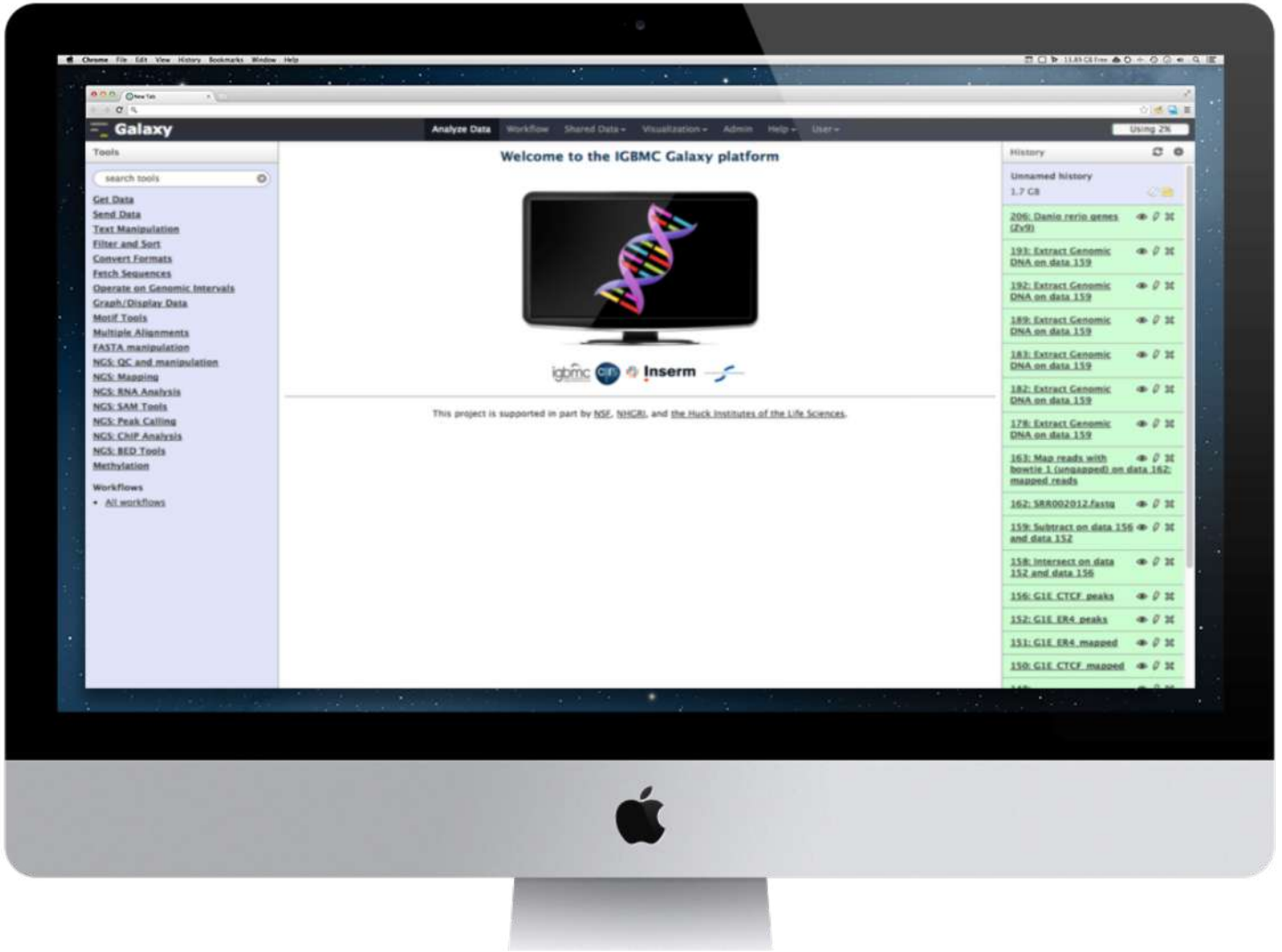

Galaxy
PROJECT

Galaxy project

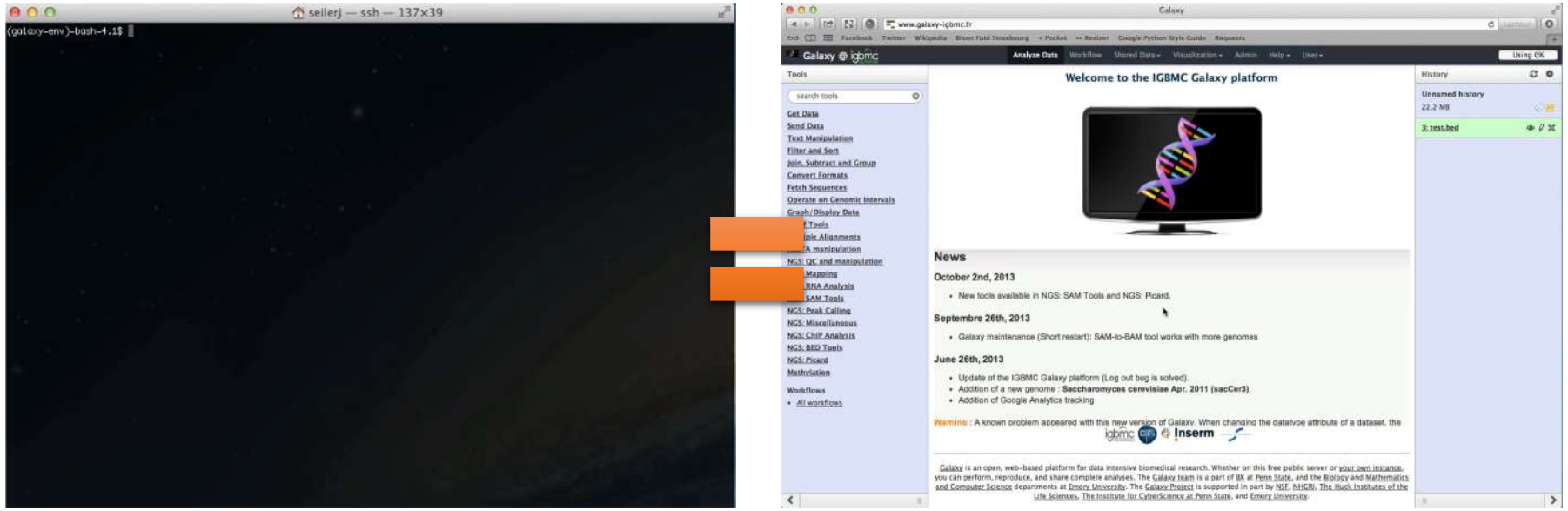
What is Galaxy ?

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/) (3)
- There are several public remote Galaxy instances worldwide (160)
  - Genomics Servers
  - Domain Servers
  - Tool Publishing Servers

Public Galaxy Servers list :
https://galaxyproject.org/use/
Last Update on: 2020, December 16th

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

Your research institute

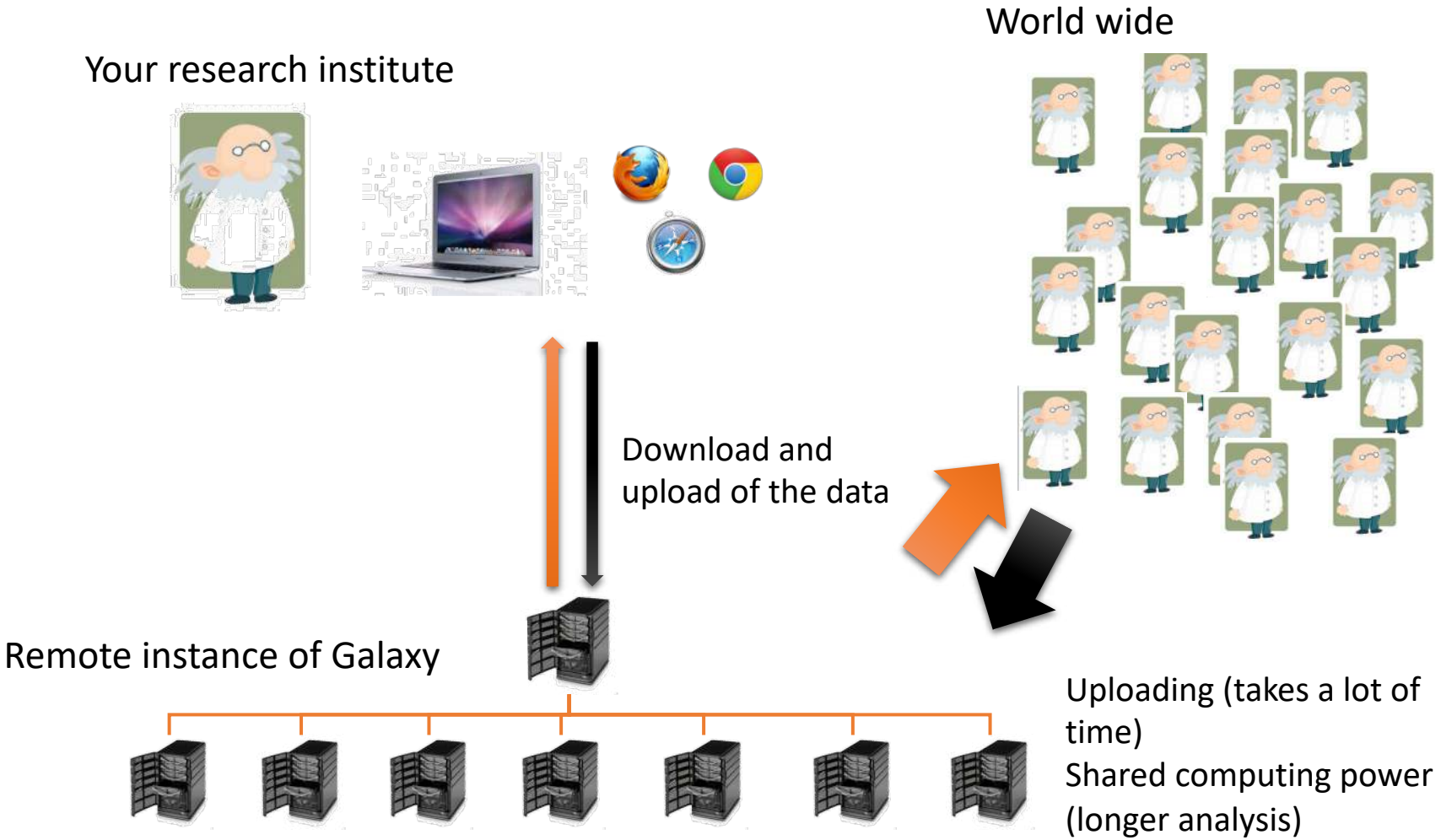
Connect to Galaxy web site through a web browser (https://usegalaxy.org/)

Download and upload of the data

Remote instance of Galaxy

Run analyses

Galaxy **public servers**



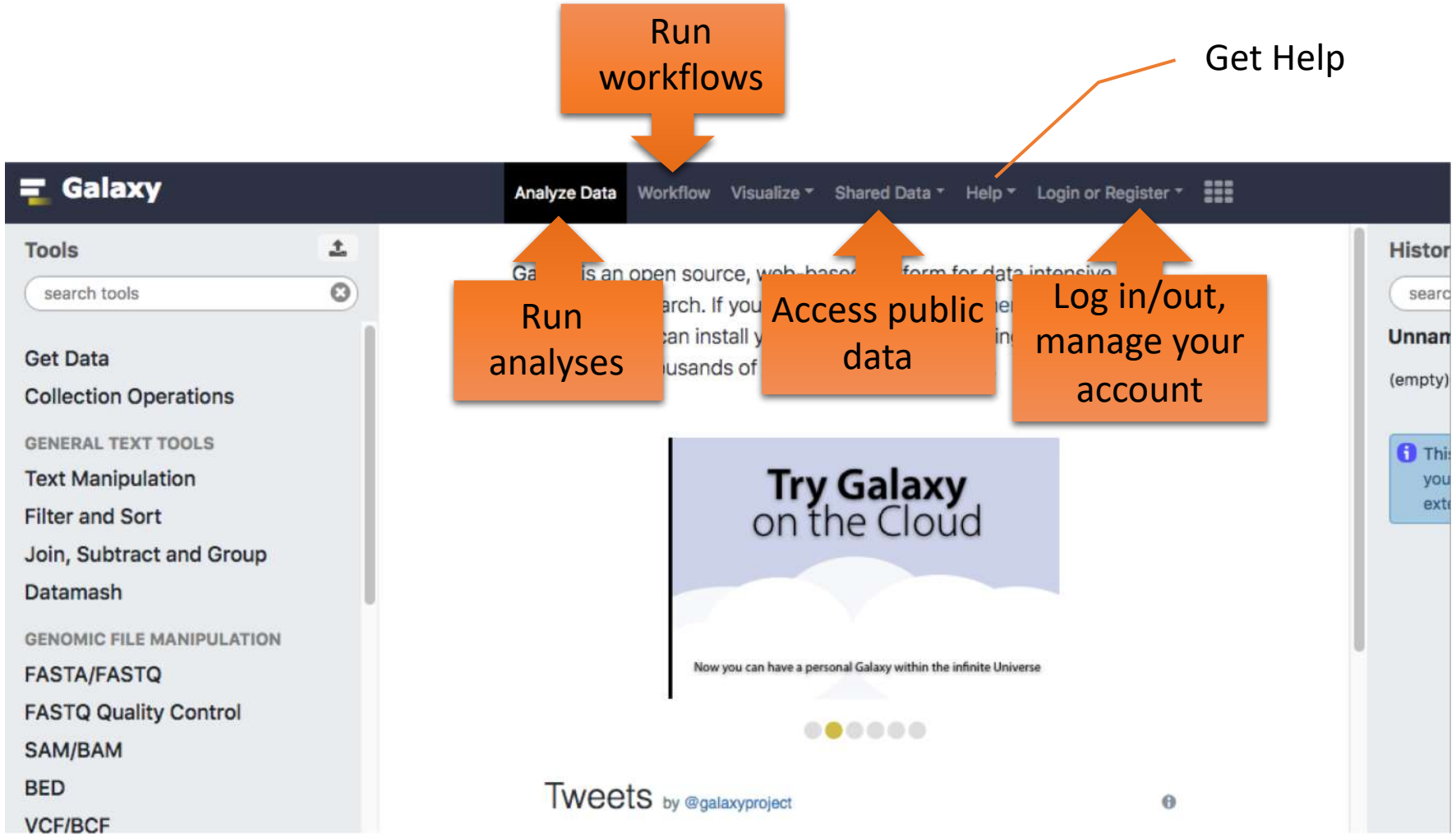
Galaxy local server

- 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

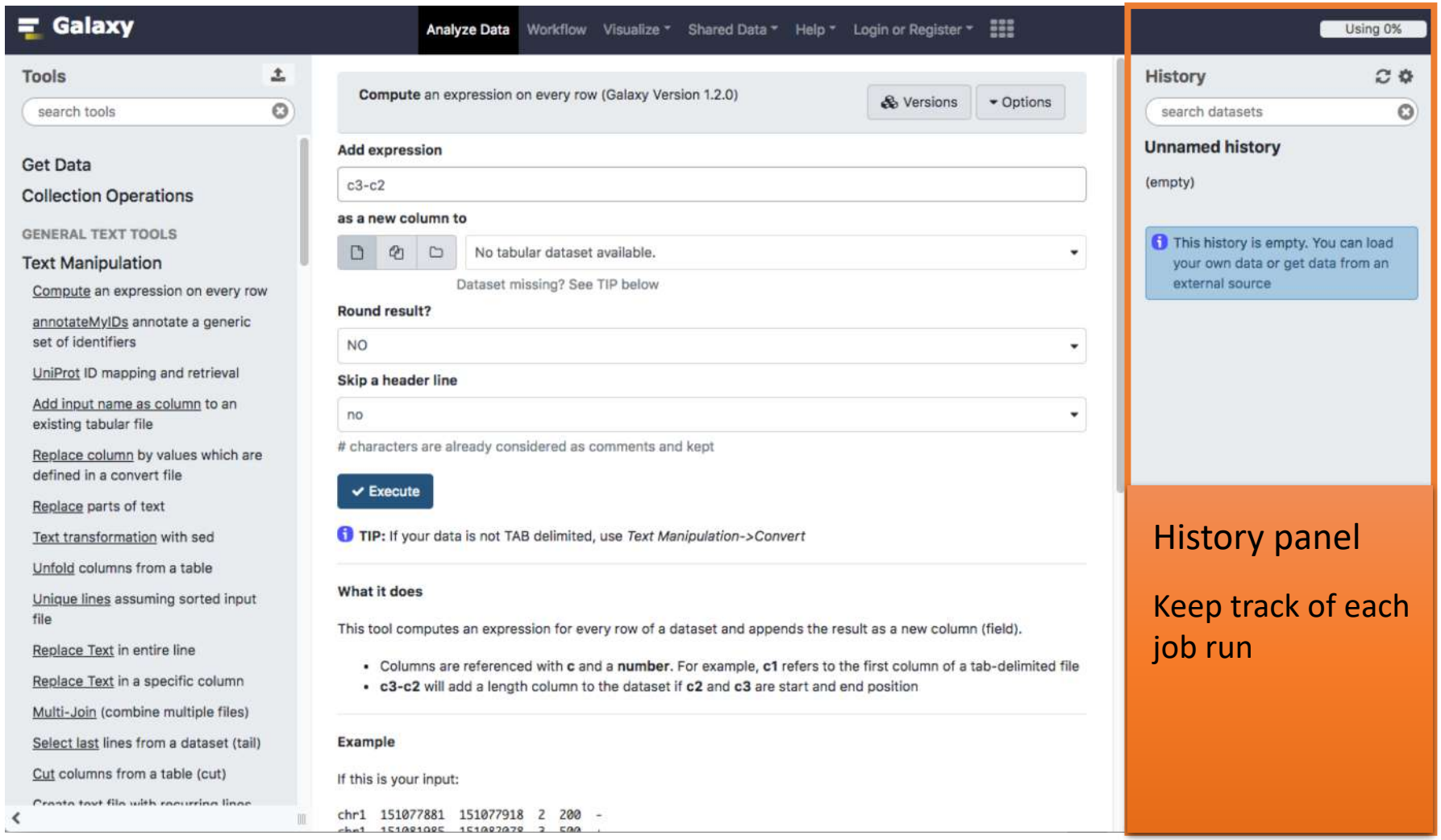
Galaxy web interface

Top menu

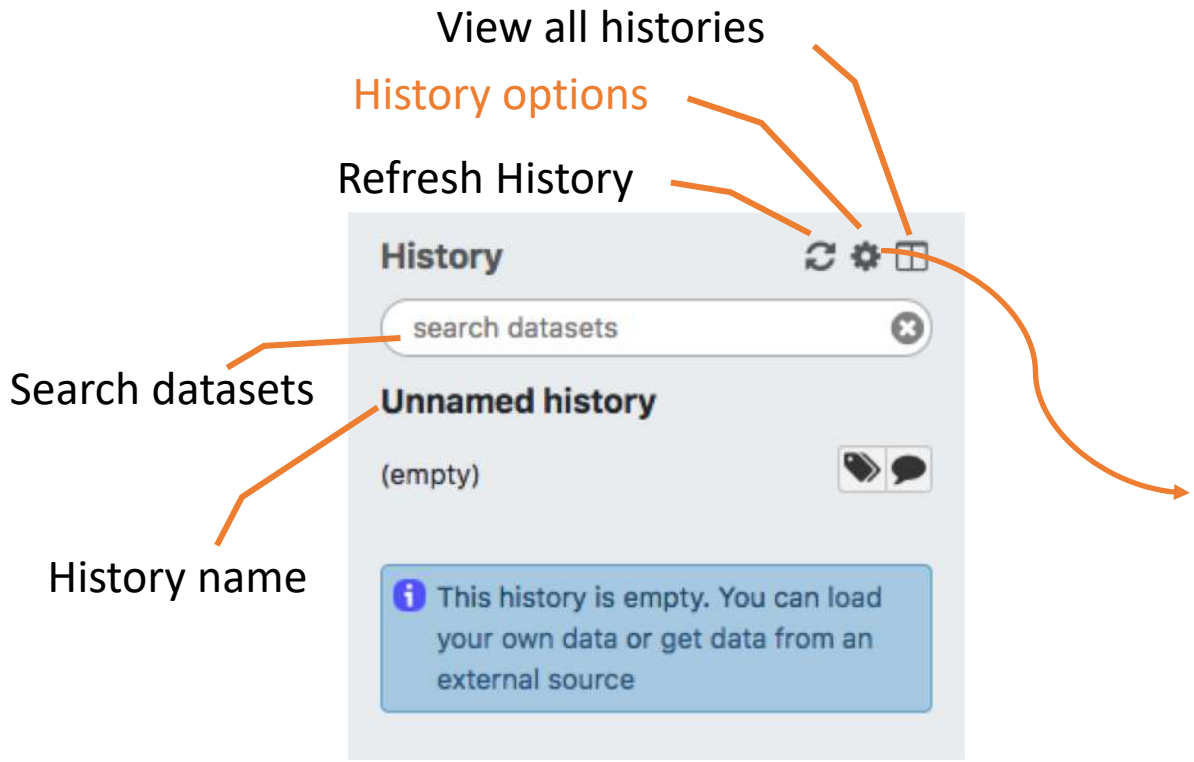

Run workflows
Get Help
Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
Login or Register
Run analyses
Access public data
Log in/out, manage your account
Tools
search tools
Get Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Datamash
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Try Galaxy on the Cloud
Now you can have a personal Galaxy within the infinite Universe
Tweets by @galaxyproject

Exercise 1

History

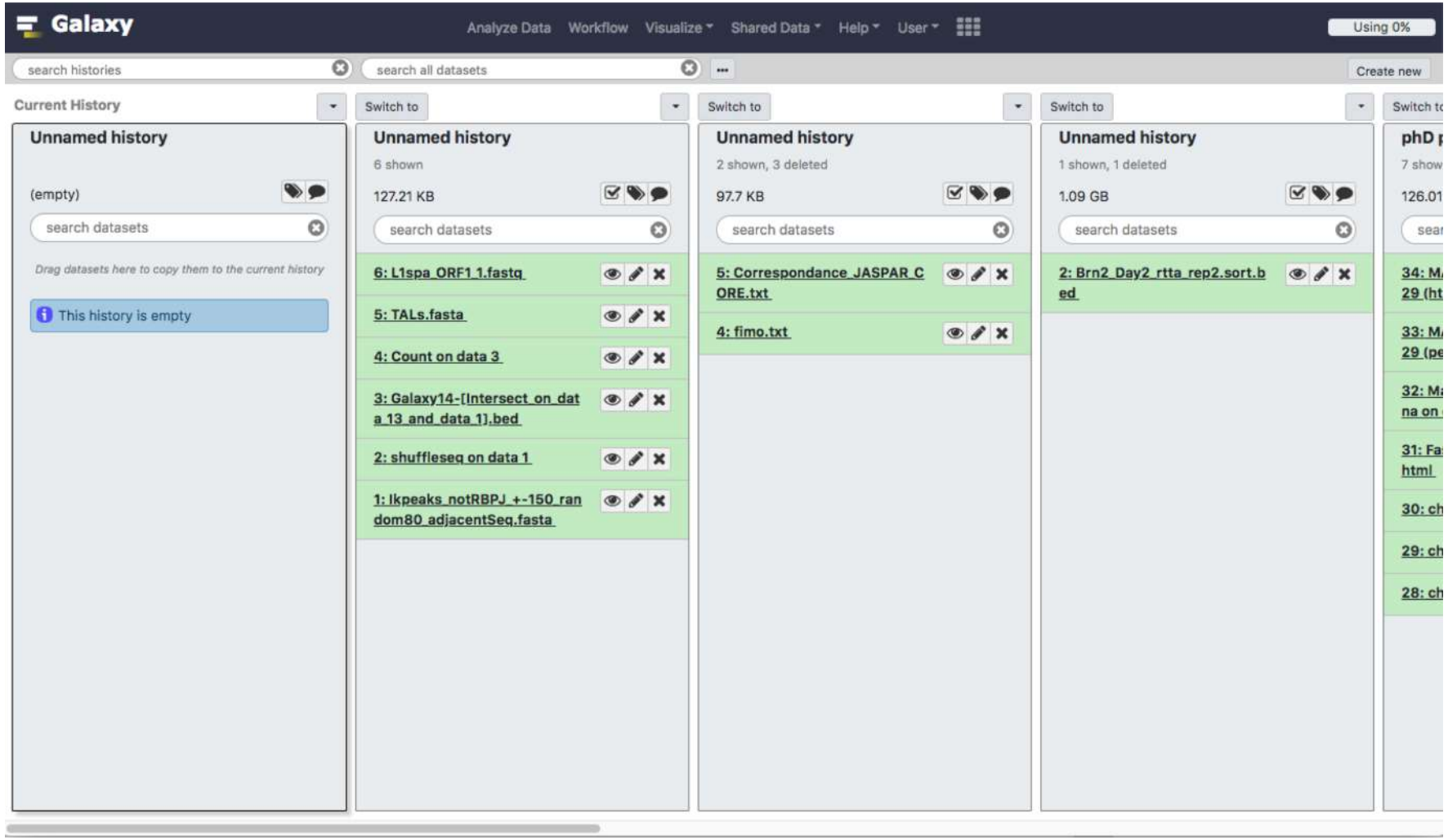

Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
Login or Register
Using 0%
Tools
search tools
Get Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Compute an expression on every row
annotateMyIDs annotate a generic set of identifiers
UniProt ID mapping and retrieval
Add input name as column to an existing tabular file
Replace column by values which are defined in a convert file
Replace parts of text
Text transformation with sed
Unfold columns from a table
Unique lines assuming sorted input file
Replace Text in entire line
Replace Text in a specific column
Multi-Join (combine multiple files)
Select last lines from a dataset (tail)
Cut columns from a table (cut)
Compute an expression on every row (Galaxy Version 1.2.0)
Versions
Options
Add expression
c3-c2
as a new column to
No tabular dataset available.
Dataset missing? See TIP below
Round result?
NO
Skip a header line
no
characters are already considered as comments and kept
Execute
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
What it does
This tool computes an expression for every row of a dataset and appends the result as a new column (field).
Columns are referenced with c and a number. For example, c1 refers to the first column of a tab-delimited file
c3-c2 will add a length column to the dataset if c2 and c3 are start and end position
Example
If this is your input:
chr1 151077881 151077918 2 200 -
History
search datasets
Unnamed history
(empty)
This history is empty. You can load your own data or get data from an external source
History panel
Keep track of each job run

History



HISTORY LISTS
Saved Histories
Histories Shared with Me
CURRENT HISTORY
Create New
Copy History
Share or Publish
Show Structure
Extract Workflow
Delete
Delete Permanently
Make Data Private
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

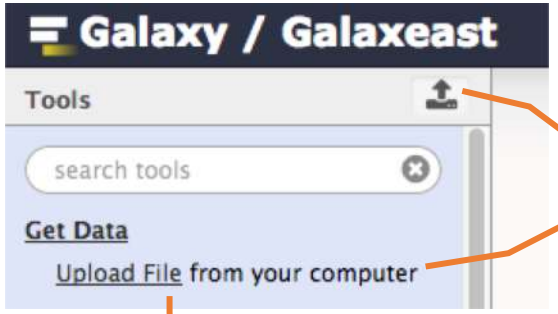


Exercise 2

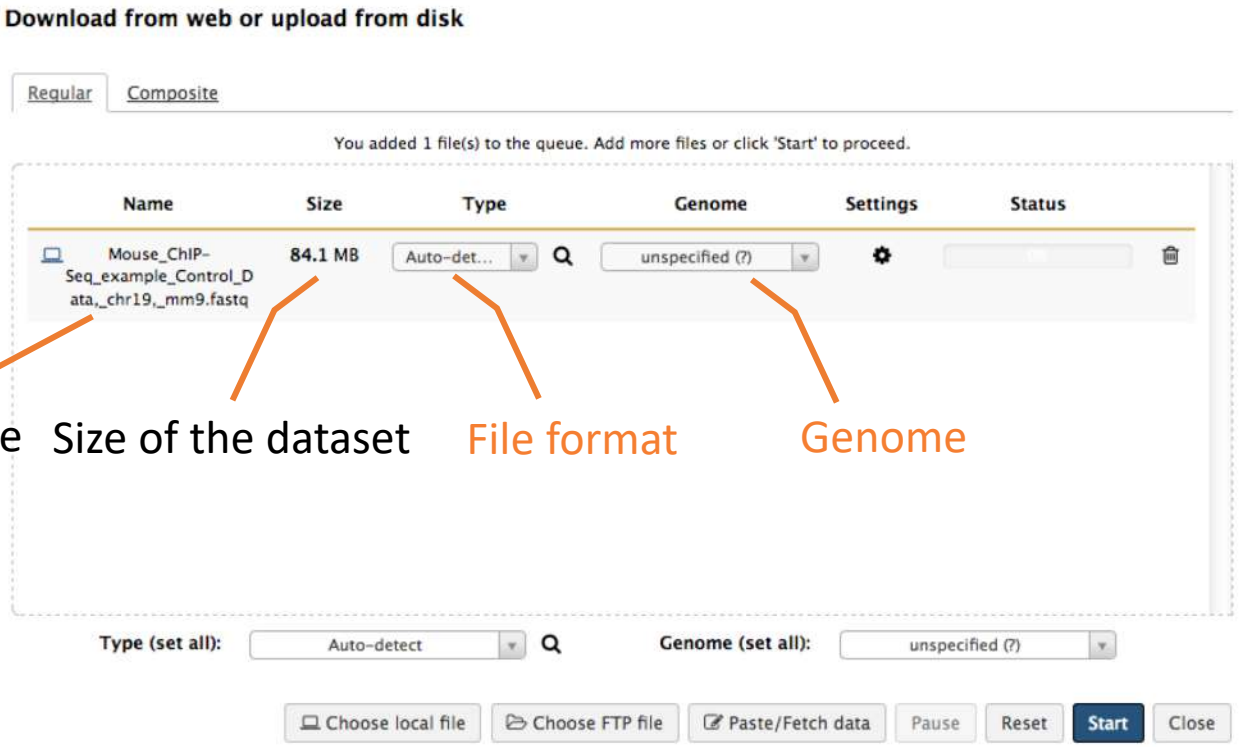
Import data into Galaxy

- Your own data (from your computer)
- Shared data
- Data from external sources

Import your own data to Galaxy



Display the drag and drop utility used to upload local files



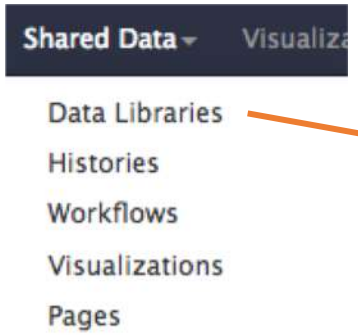
Name of the dataset

Size of the dataset

File format

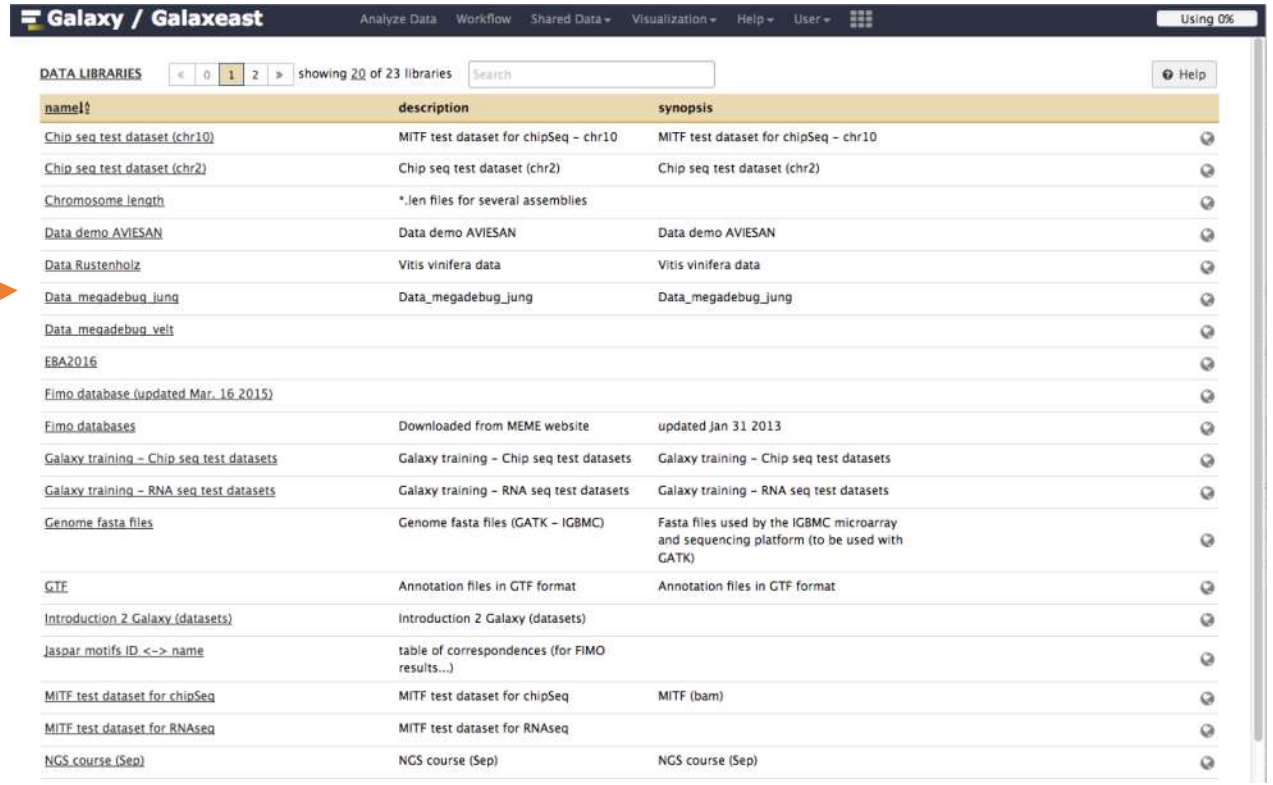
Genome

Import shared data (data libraries)



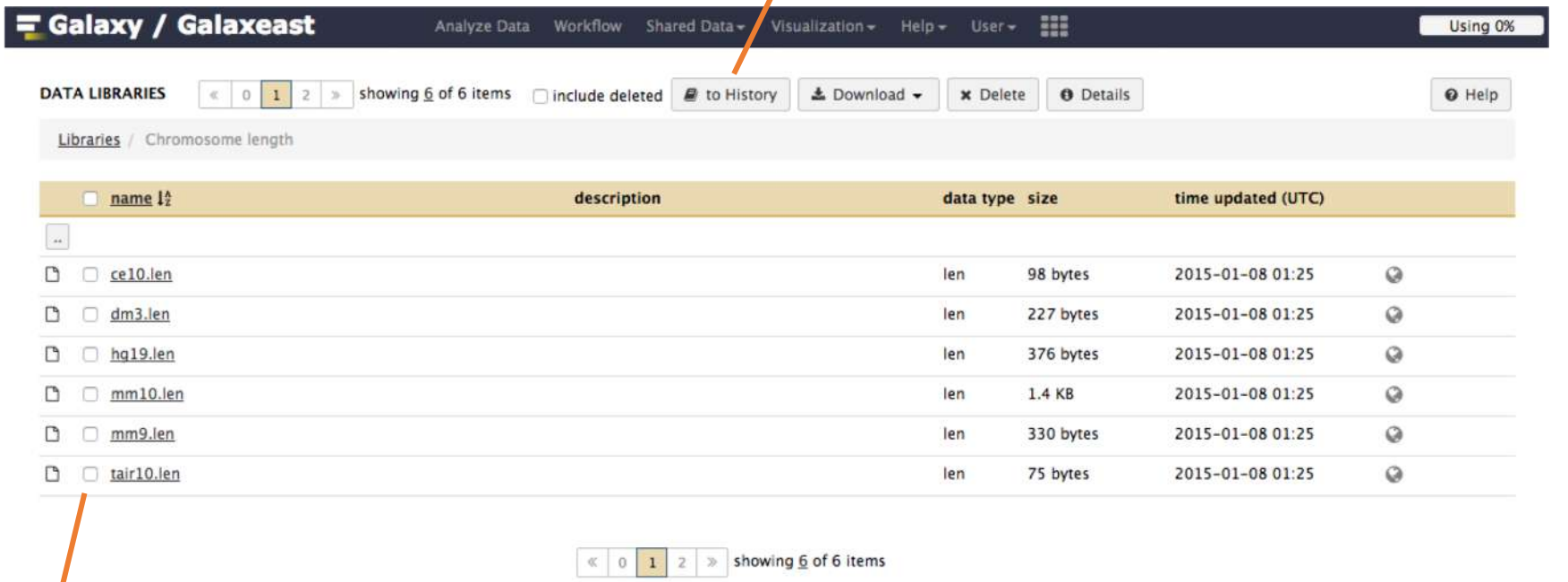
(Top menu)

Data Libraries



Import shared data (data libraries)

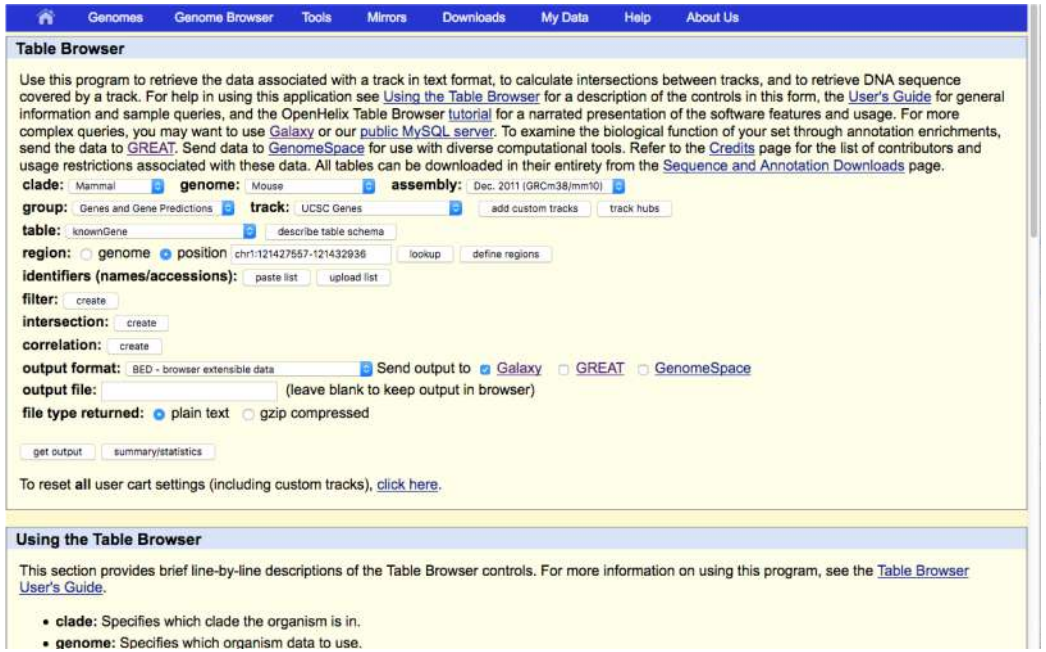
2. Import selected dataset to history



1. Select dataset

Import public data

Browse and import external data from public databases



Exercise 3.1

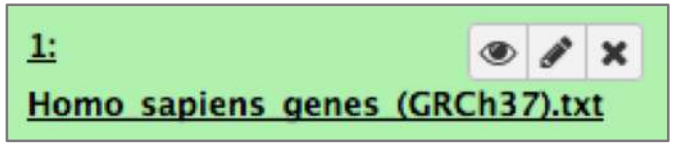
Exercise 3.2

Datasets/Jobs in the History

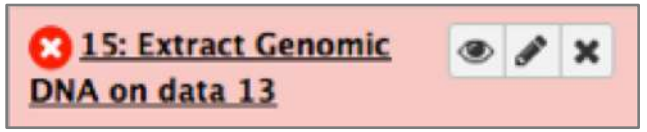
Grey: the job is waiting to run

Yellow: the job is running

Green: the job is successfully done



Red: the job encountered a problem

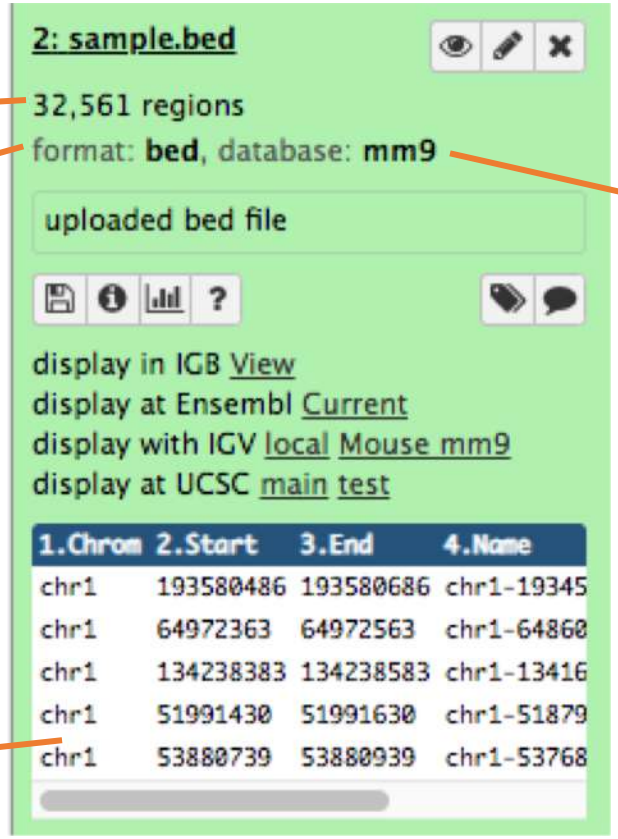


Datasets/Jobs in the History

Number of lines in the file or size of the file

Format

Genome

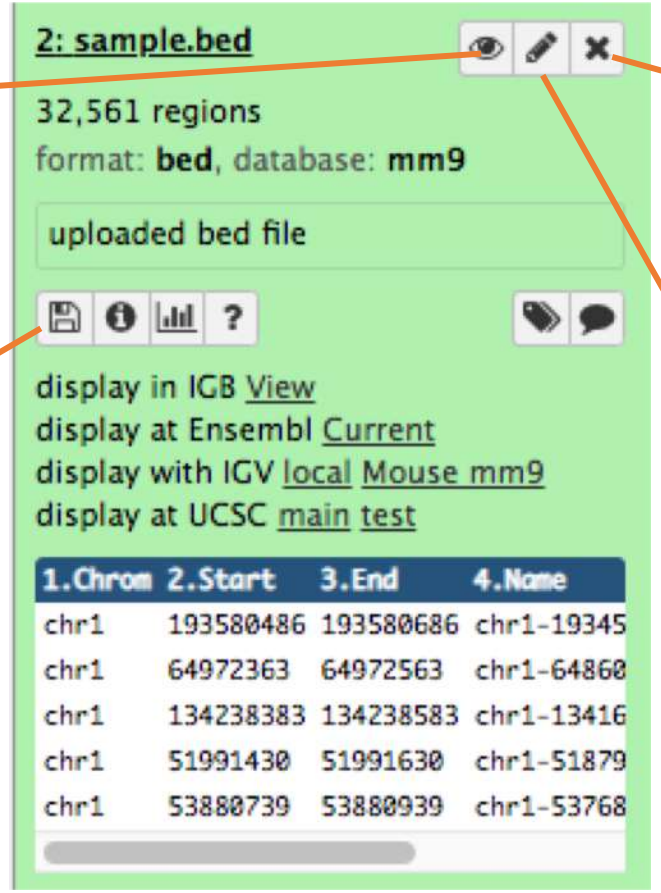


If the dataset is a text file, the first lines of the file are displayed

Datasets/Jobs in the History

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

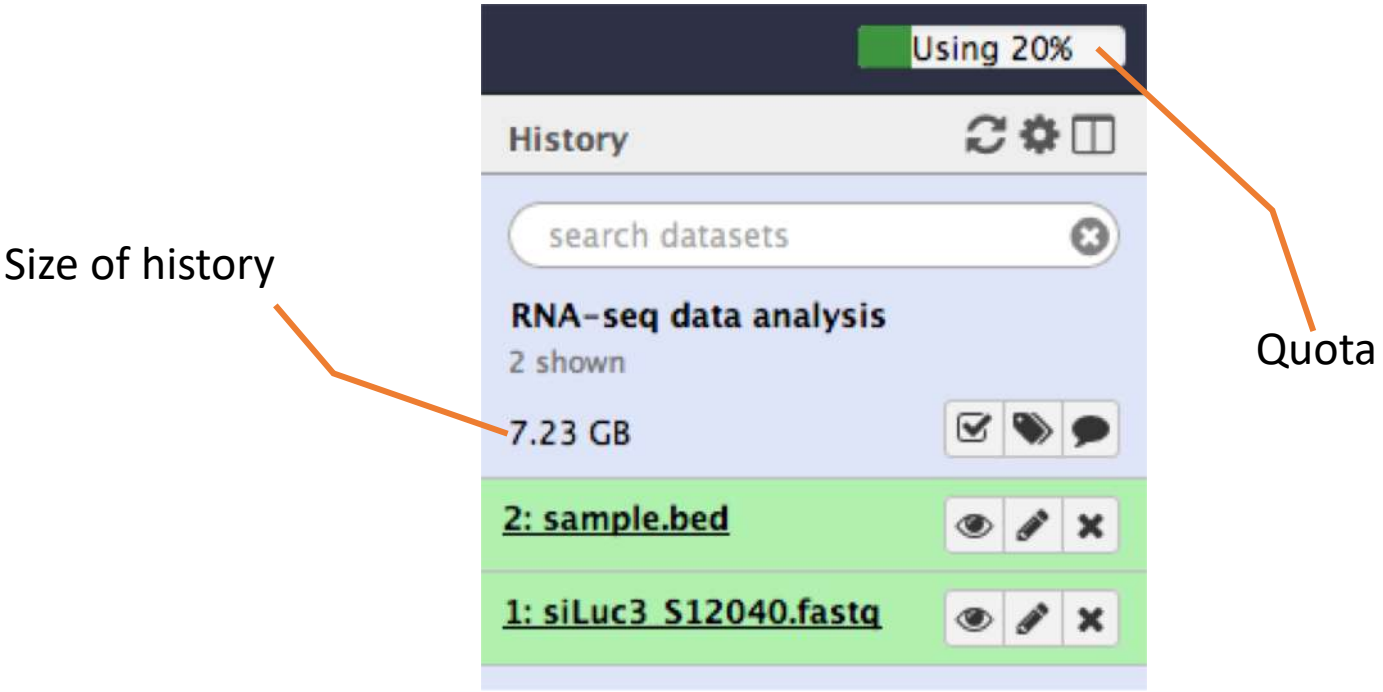
Download dataset



Delete dataset

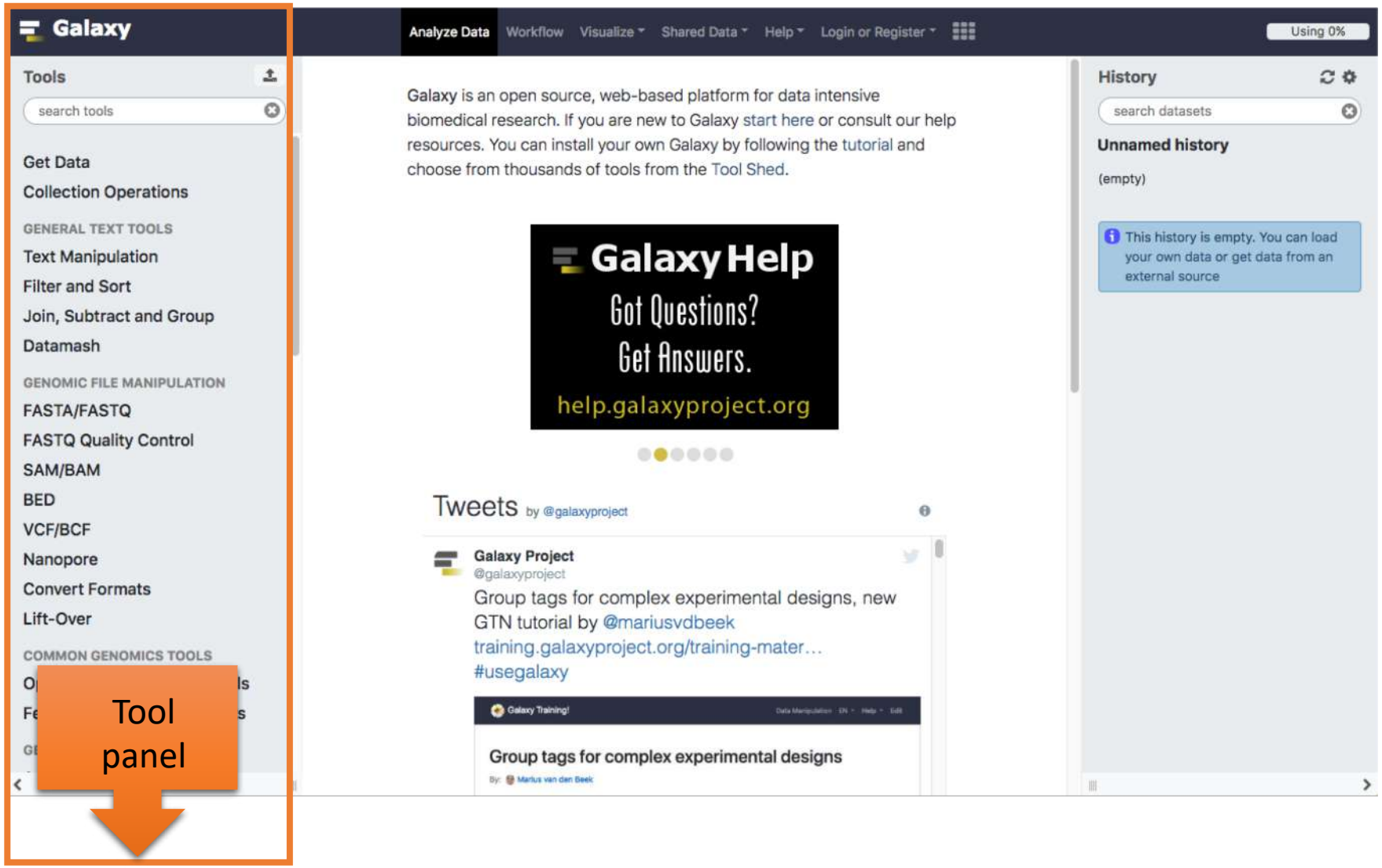
Edit attributes of the dataset (change name, format, genome, permission)

Size of histories and quota

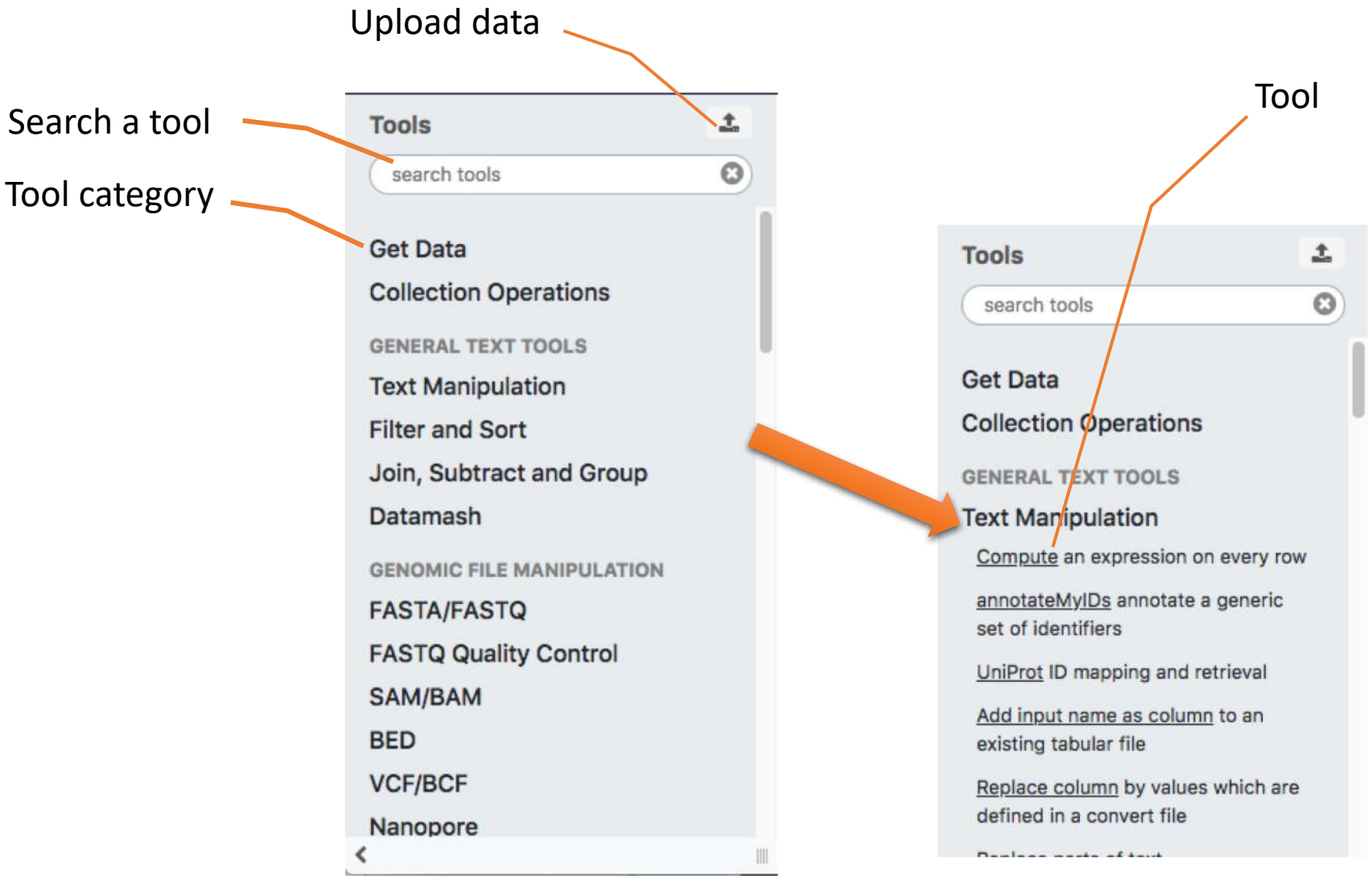


Exercise 4

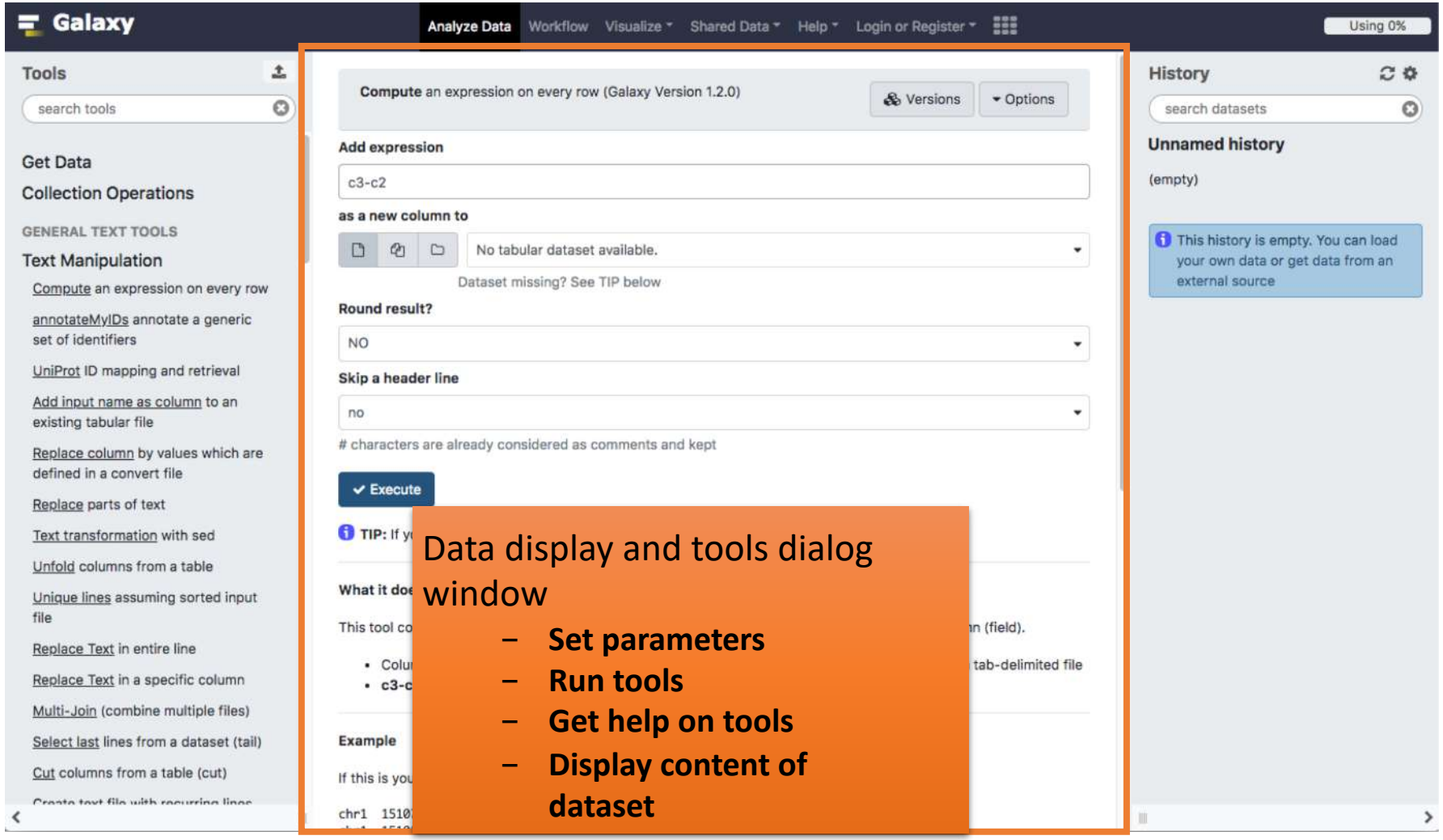
Tool Panel / Run analyses



Tool Panel / Run analyses



Tools dialog window



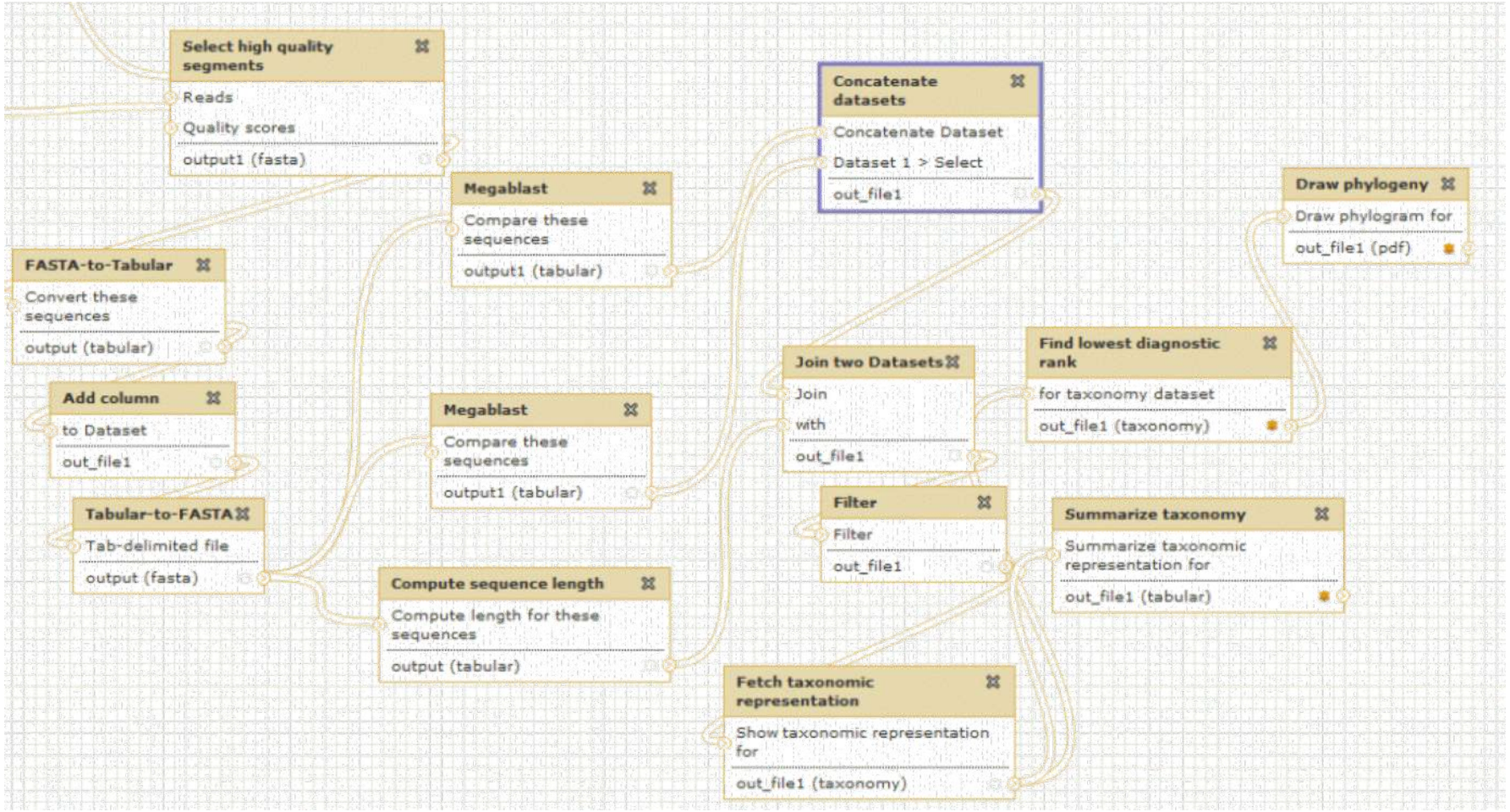
Exercise 5

Exercise 6

Workflow

What if we'd mix all together

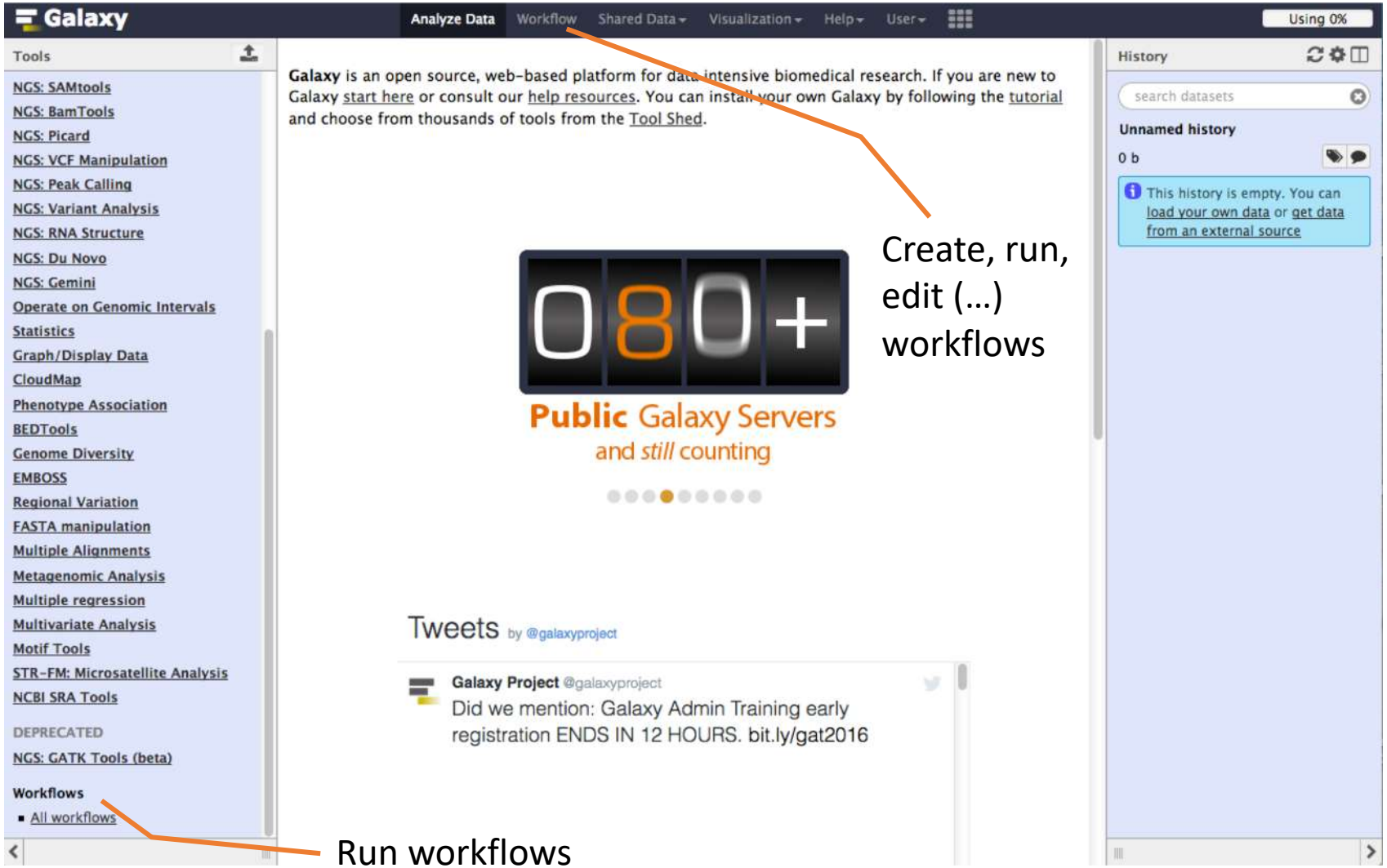
Galaxy workflow



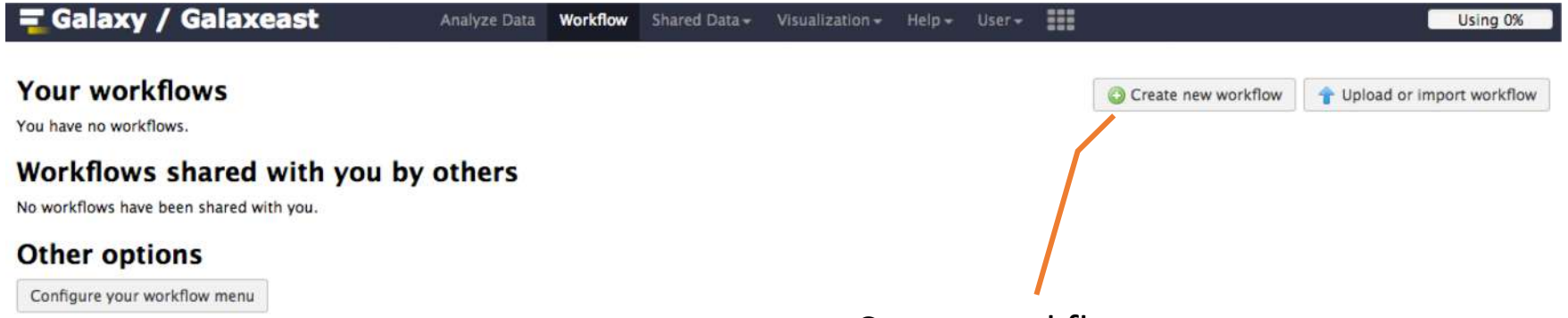
Galaxy workflows

- 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

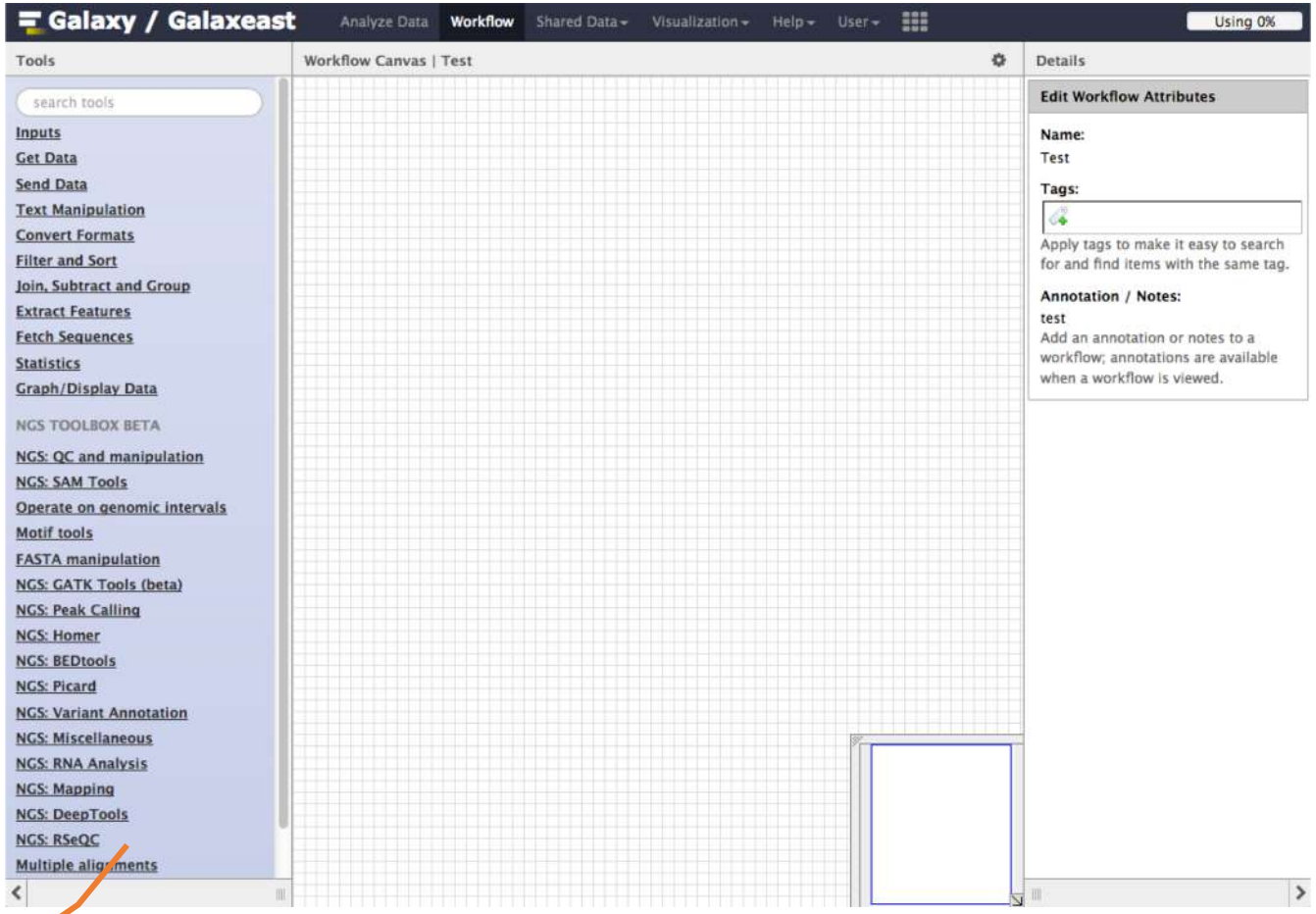

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0%
Tools
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
Operate on Genomic Intervals
Statistics
Graph/Display Data
CloudMap
Phenotype Association
BEDTools
Genome Diversity
EMBOSS
Regional Variation
FASTA manipulation
Multiple Alignments
Metagenomic Analysis
Multiple regression
Multivariate Analysis
Motif Tools
STR-FM: Microsatellite Analysis
NCBI SRA Tools
DEPRECATED
NGS: GATK Tools (beta)
Workflows
All workflows
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.
080+
Public Galaxy Servers
and still counting
Tweets by @galaxyproject
Galaxy Project @galaxyproject
Did we mention: Galaxy Admin Training early registration ENDS IN 12 HOURS. bit.ly/gat2016
History
search datasets
Unnamed history
0 b
This history is empty. You can load your own data or get data from an external source
Create, run, edit (…) workflows
Run workflows

Workflows


Galaxy / Galaxeast
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0%
Your workflows
You have no workflows.
Create new workflow
Upload or import workflow
Workflows shared with you by others
No workflows have been shared with you.
Other options
Configure your workflow menu
Create workflows

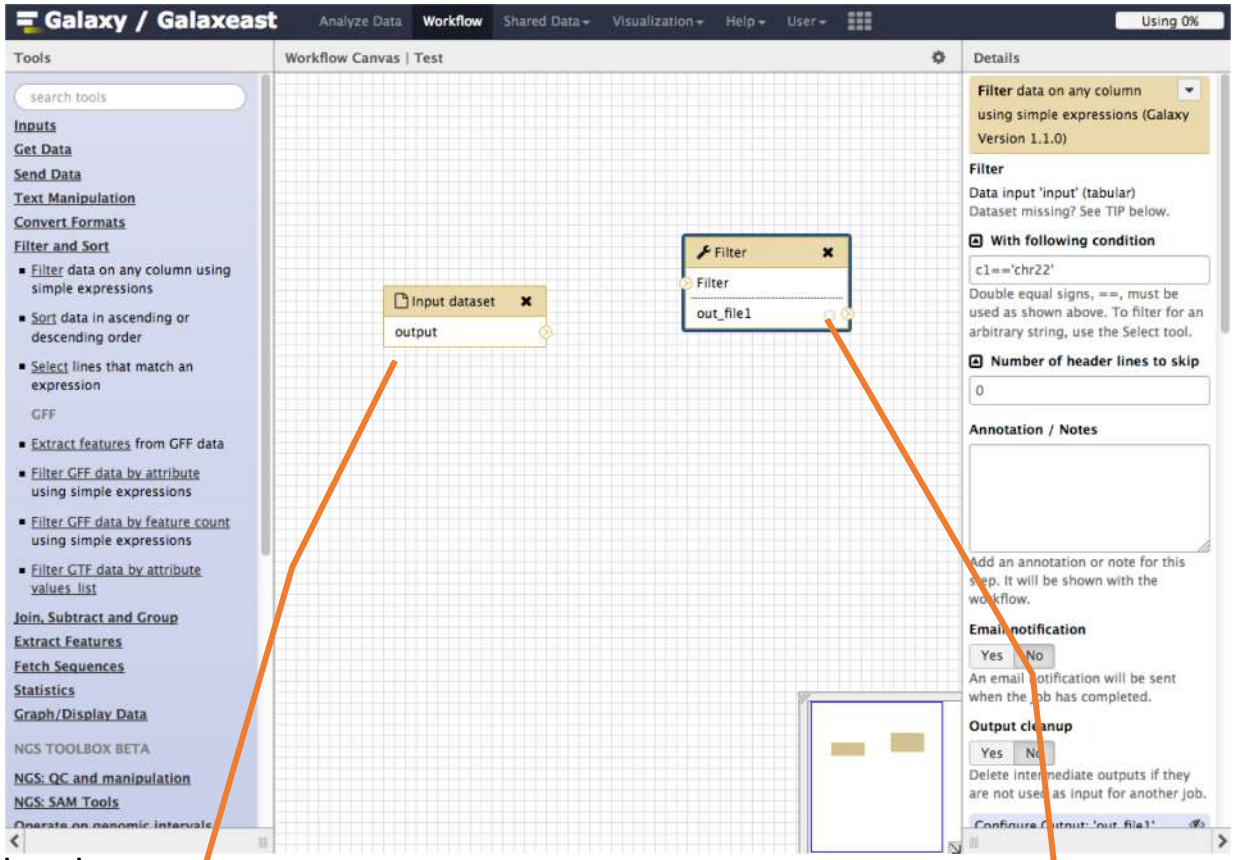

Create New Workflow
Workflow Name:
Unnamed workflow
Give a name to the workflow
Workflow Annotation:
A description of the workflow; annotation is shown alongside shared or published workflows.
Create

Workflow creation



Add tools or input datasets to the workflow

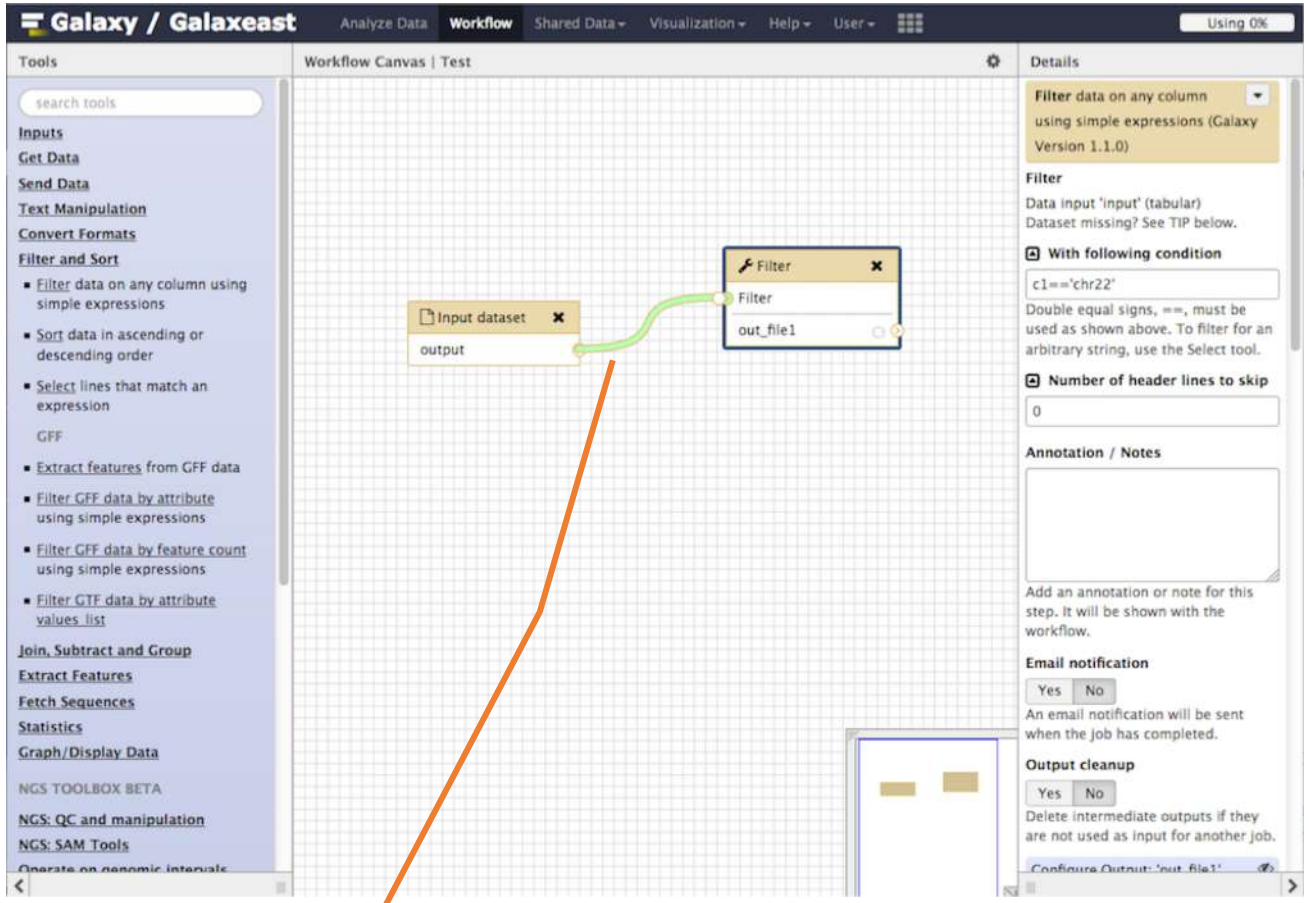
Workflow creation



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

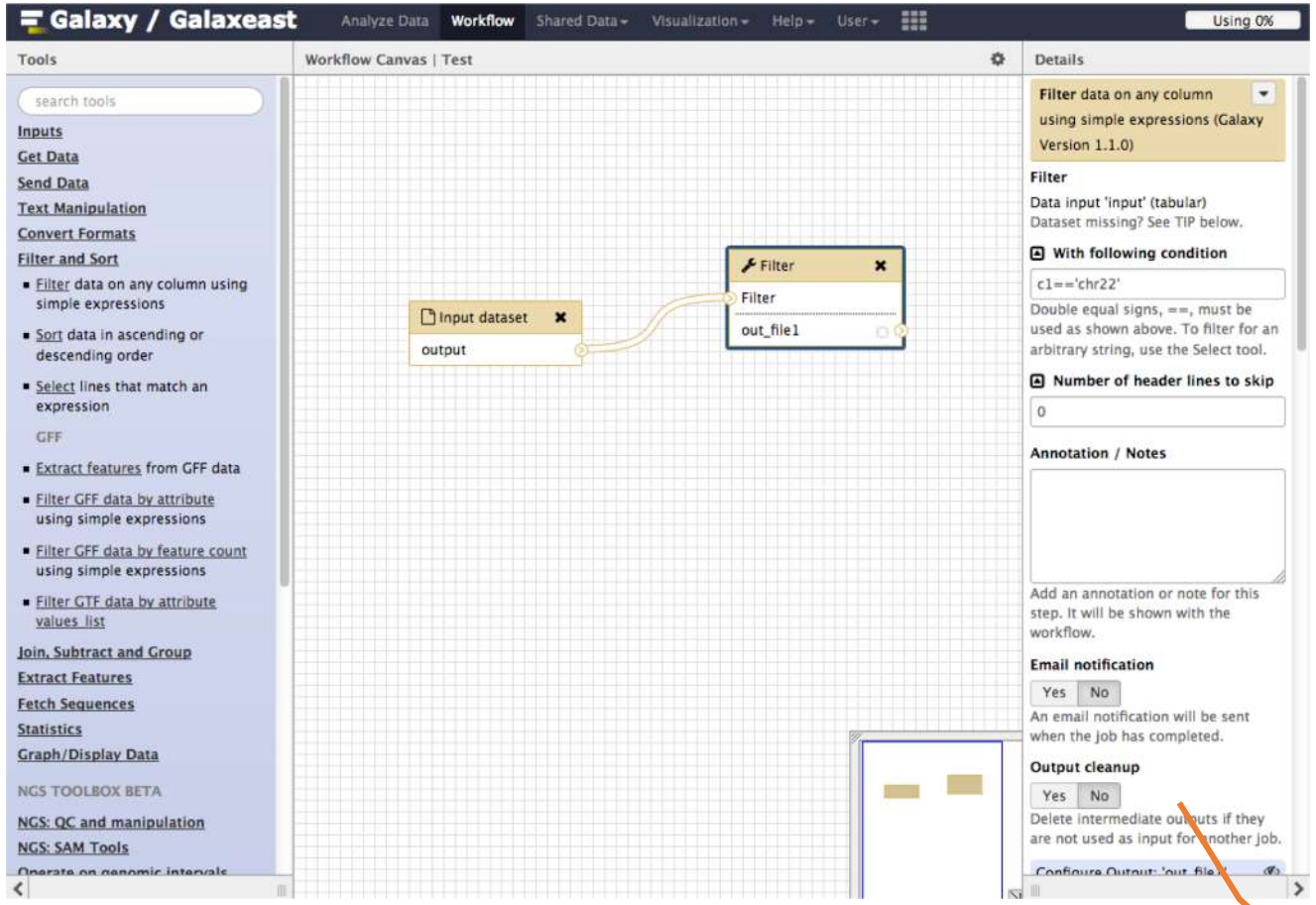
Tool to be run

Workflow creation



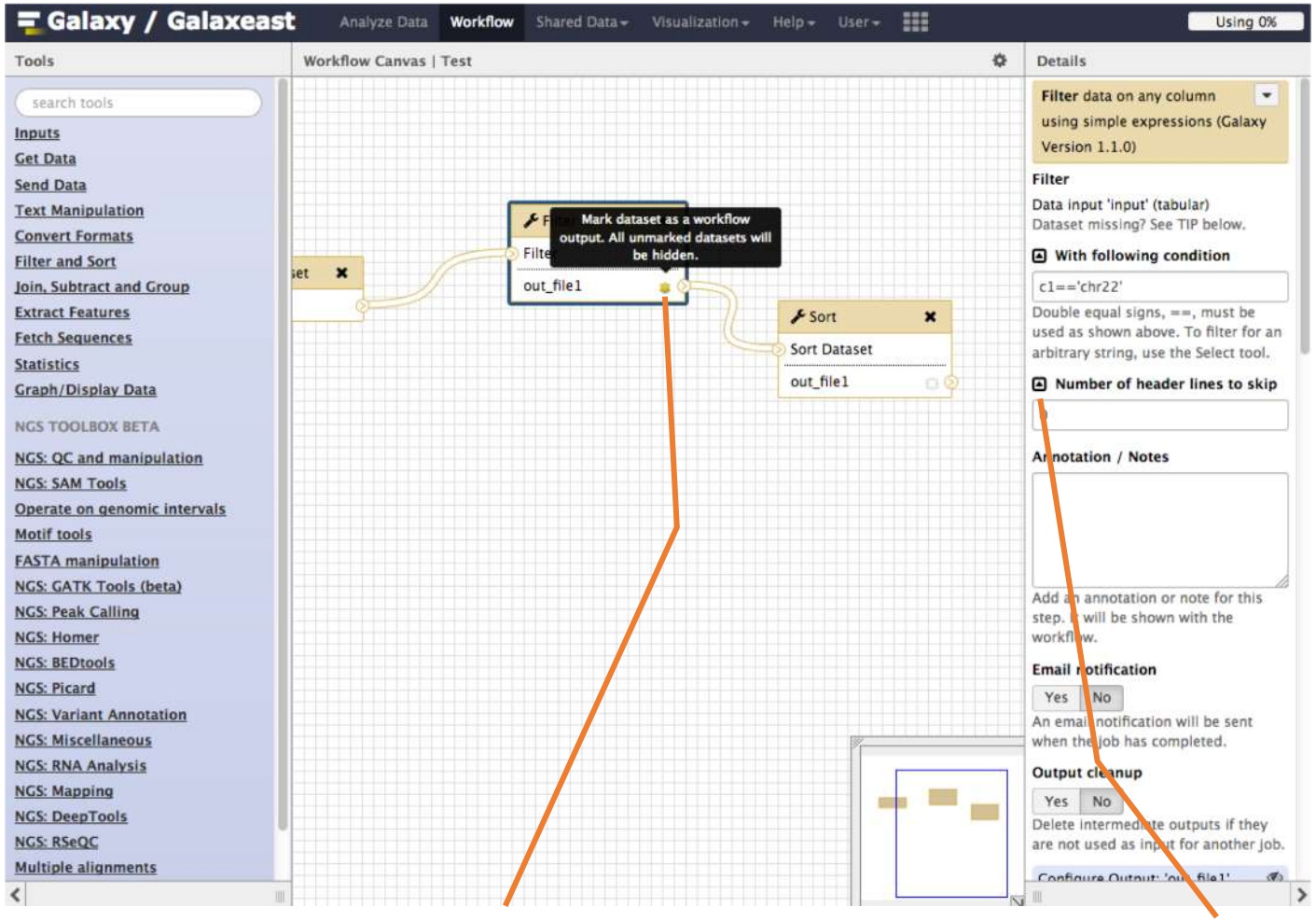
If two steps can be linked together, the link between the two boxes is green

Workflow creation



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

Workflow creation

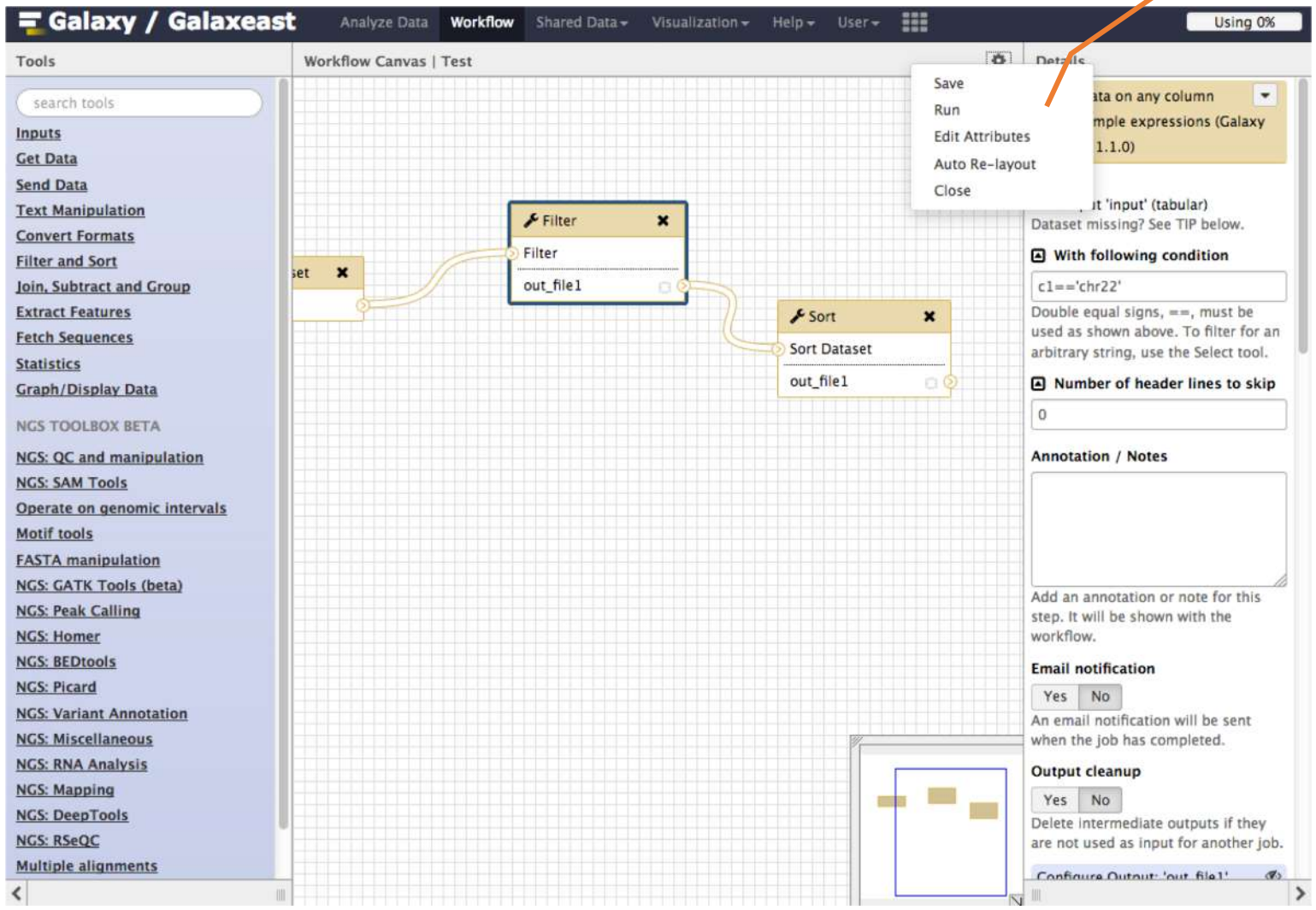


Click on star to select which datasets will be displayed in the history generated when running of the workflow

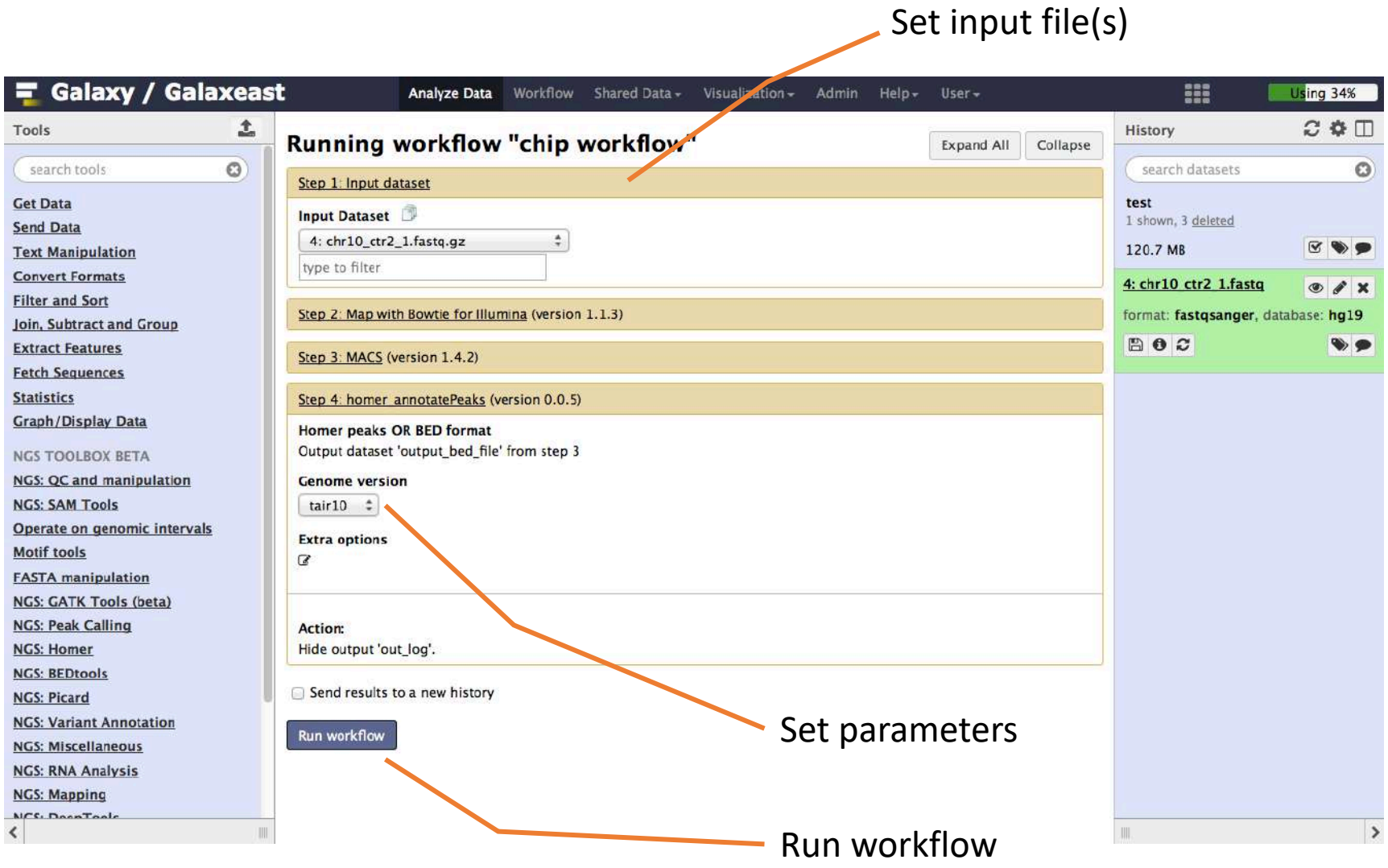
Click to get the parameter to be set at runtime

Workflow creation

Save, run workflows



Run workflows



Exercise 7

Exercise 8

Exercise 9

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