Introduction to Galaxy

Stéphanie Le Gras
(slegras@igbmc.fr)

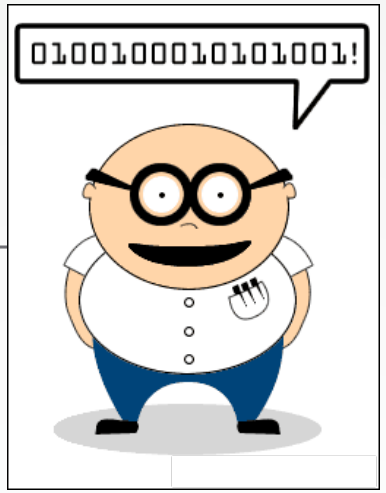
Guidelines

- 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

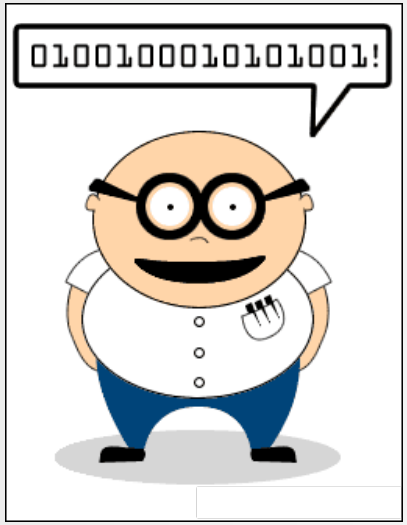
Bioinformatics analyses

Informatics data



nature

Bioinformatics analyses



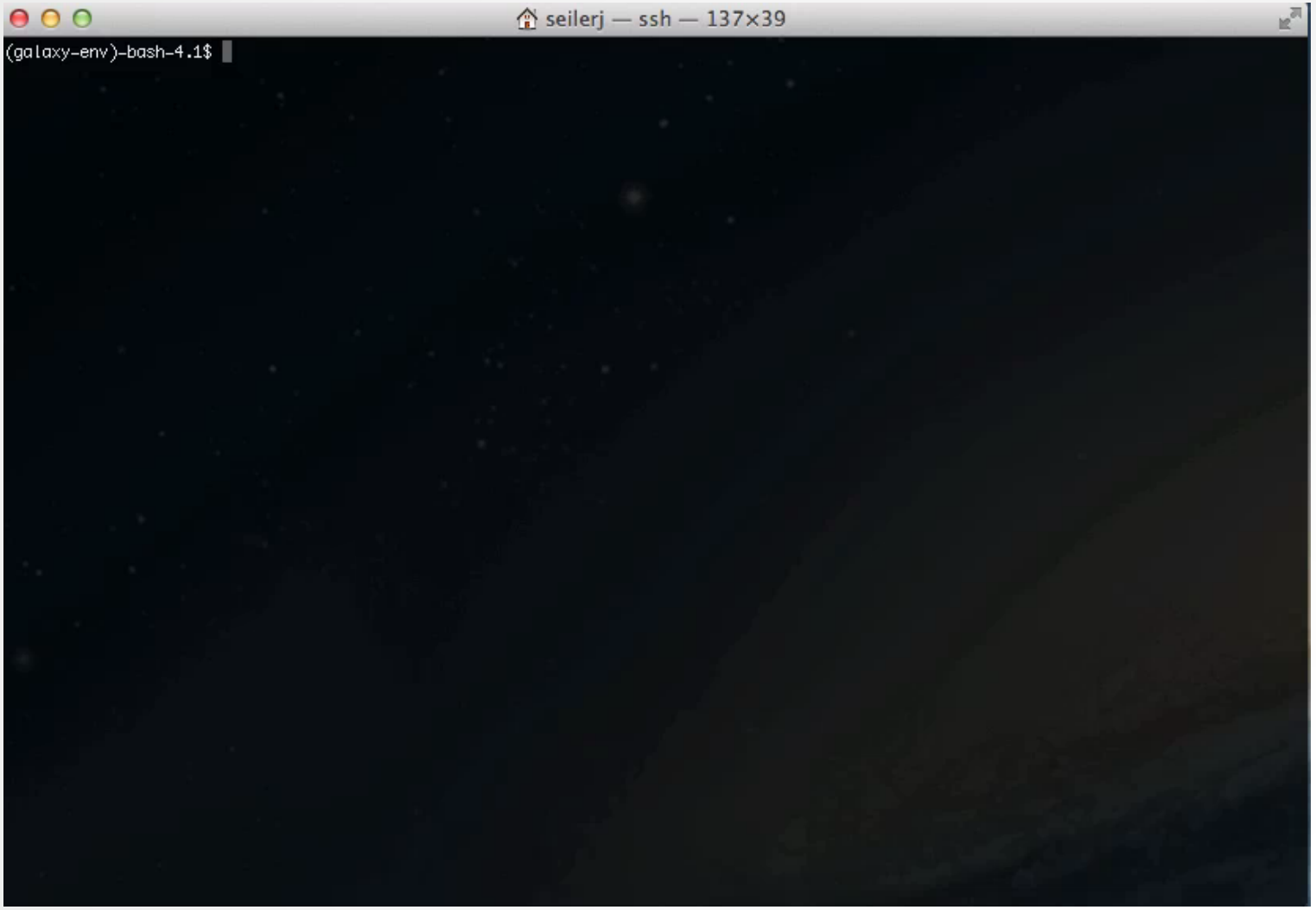
Scripts, softwares

```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.
```

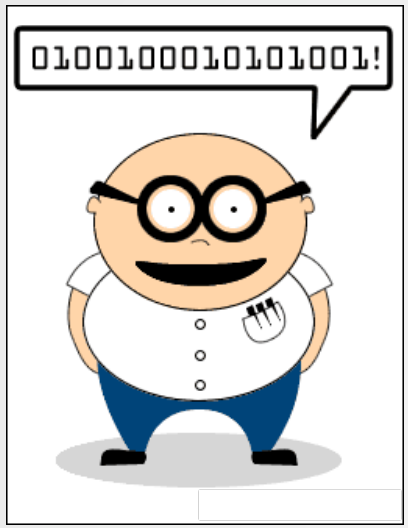
Command line

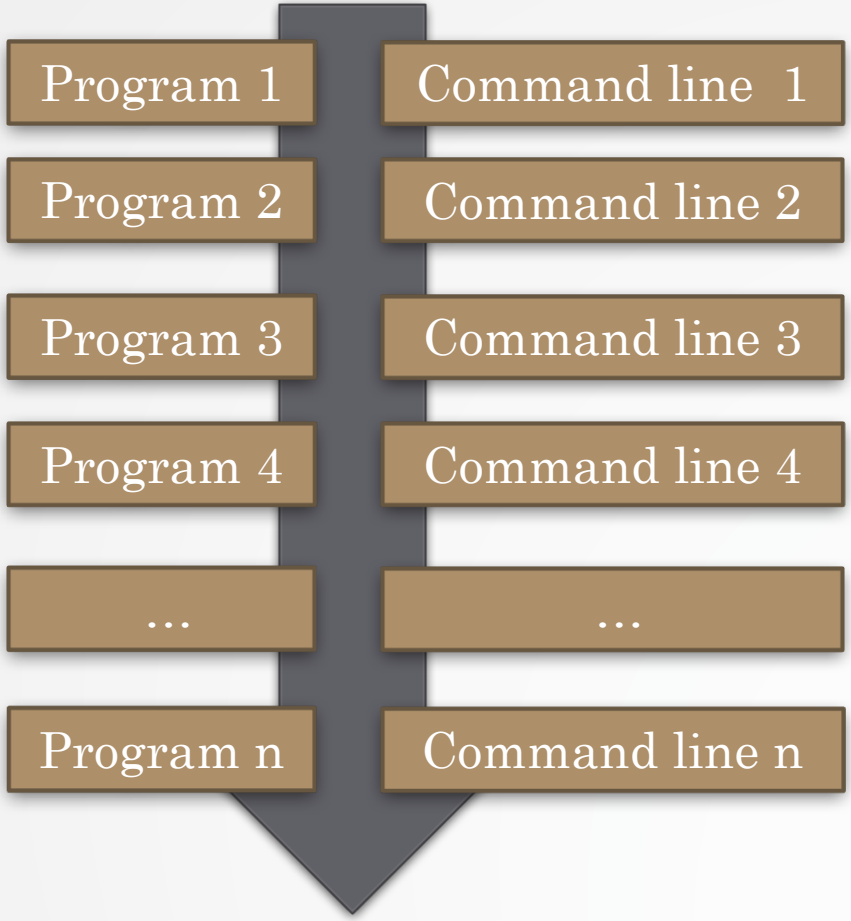
```
macs14 -t treatment.sort.bed -c control.bed -f BED -g mm --name=name1 --llocal=50000
 --slocal=5000 > macs1.nohup 2>&1 &
```

Bioinformatics analyses



Bioinformatics analyses



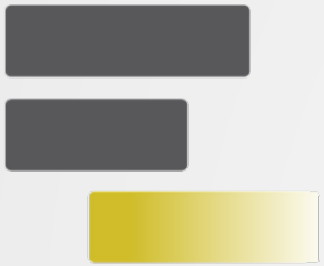


PIPELINE/ WORKFLOW

Workflow tools

- Cipres phylo.org
- eHive ensembl.org/info/docs/eHive/index.html
- Galaxy main.g2.bx.psu.edu
- Knime knime.org
- Mobyle mobyle.pasteur.fr
- Taverna taverna.org.uk

Galaxy ?



Galaxy

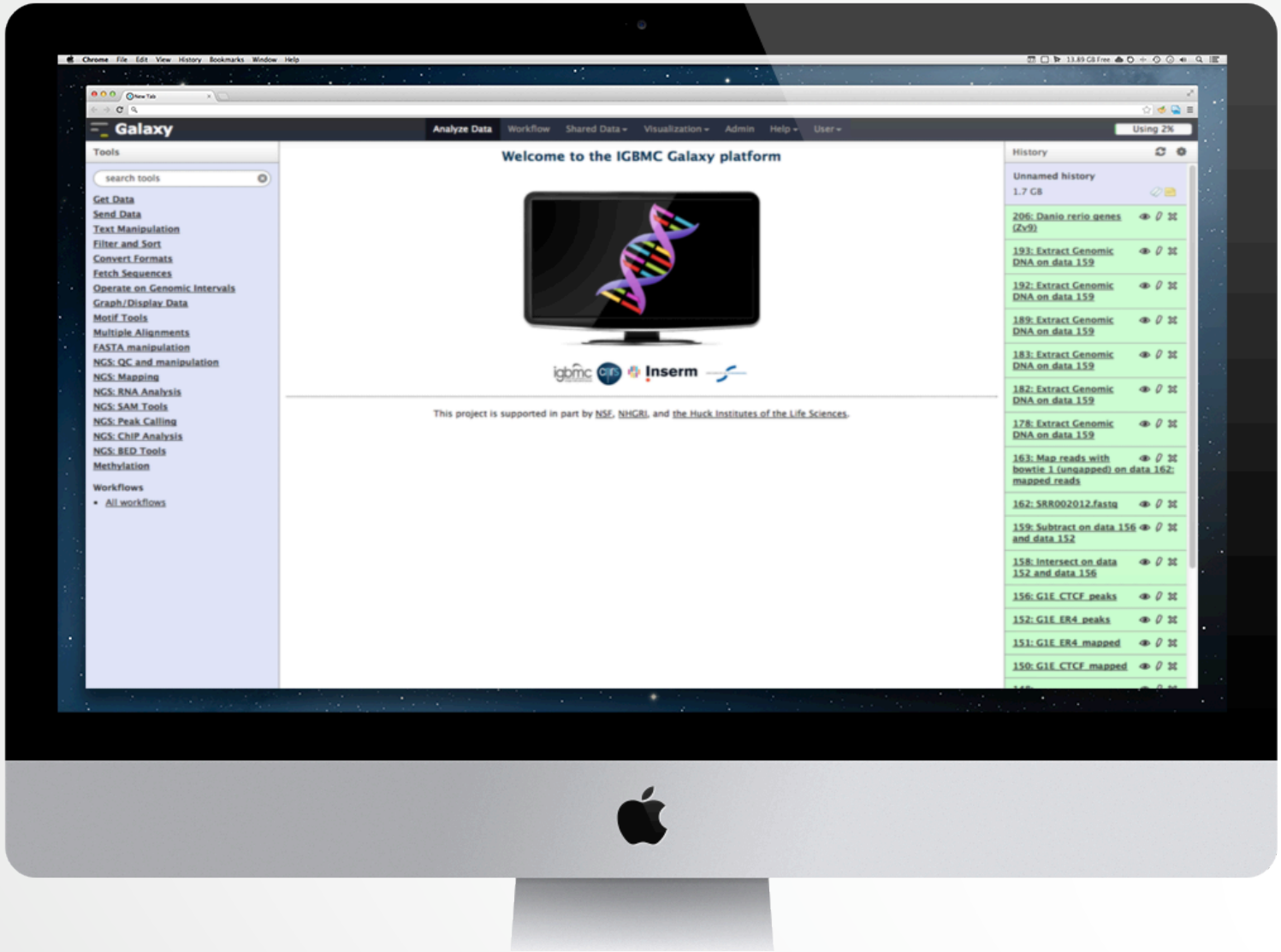
PROJECT

Galaxy project

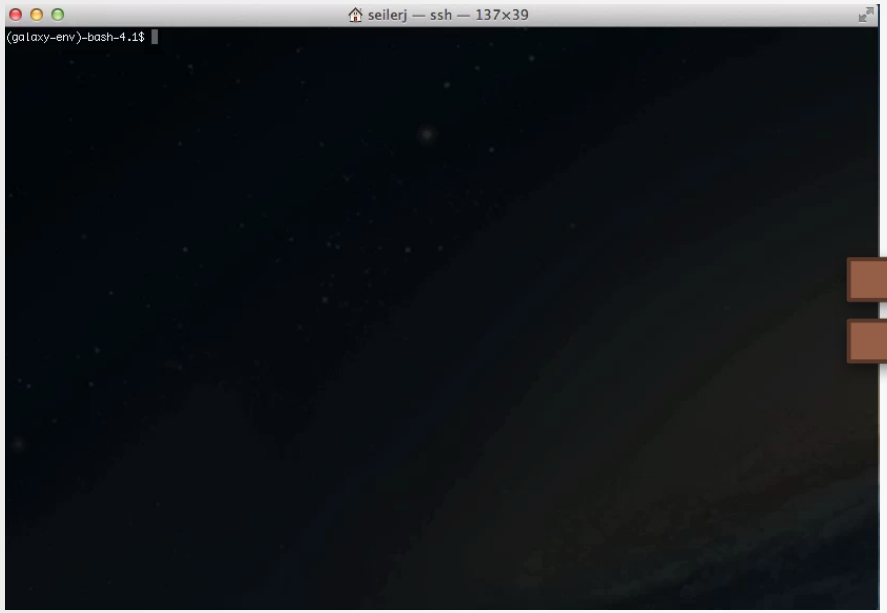
What is Galaxy ?

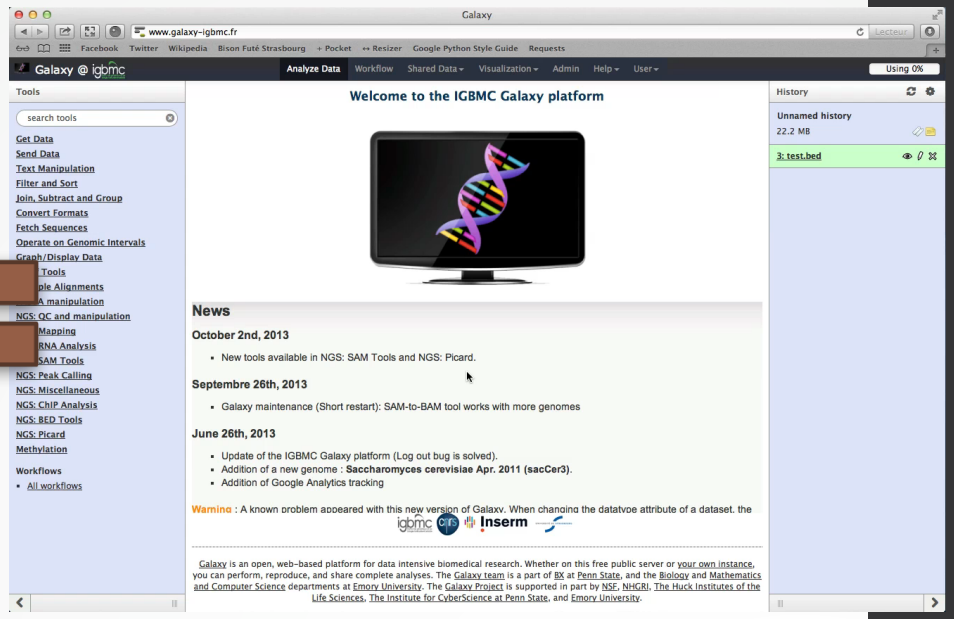
Galaxy is a **computing platform** that enables people to **run complex bioinformatics tools** on a **compute cluster** through a **simple web interface**.

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**
- Cloud
- Docker

Galaxy **public servers**

- Galaxy Project's public server (UseGalaxy.org)
- There are several public remote Galaxy instances worldwide
  - General Purpose servers (15)
  - Domain Servers (35)
  - Tool Publishing Servers (41)

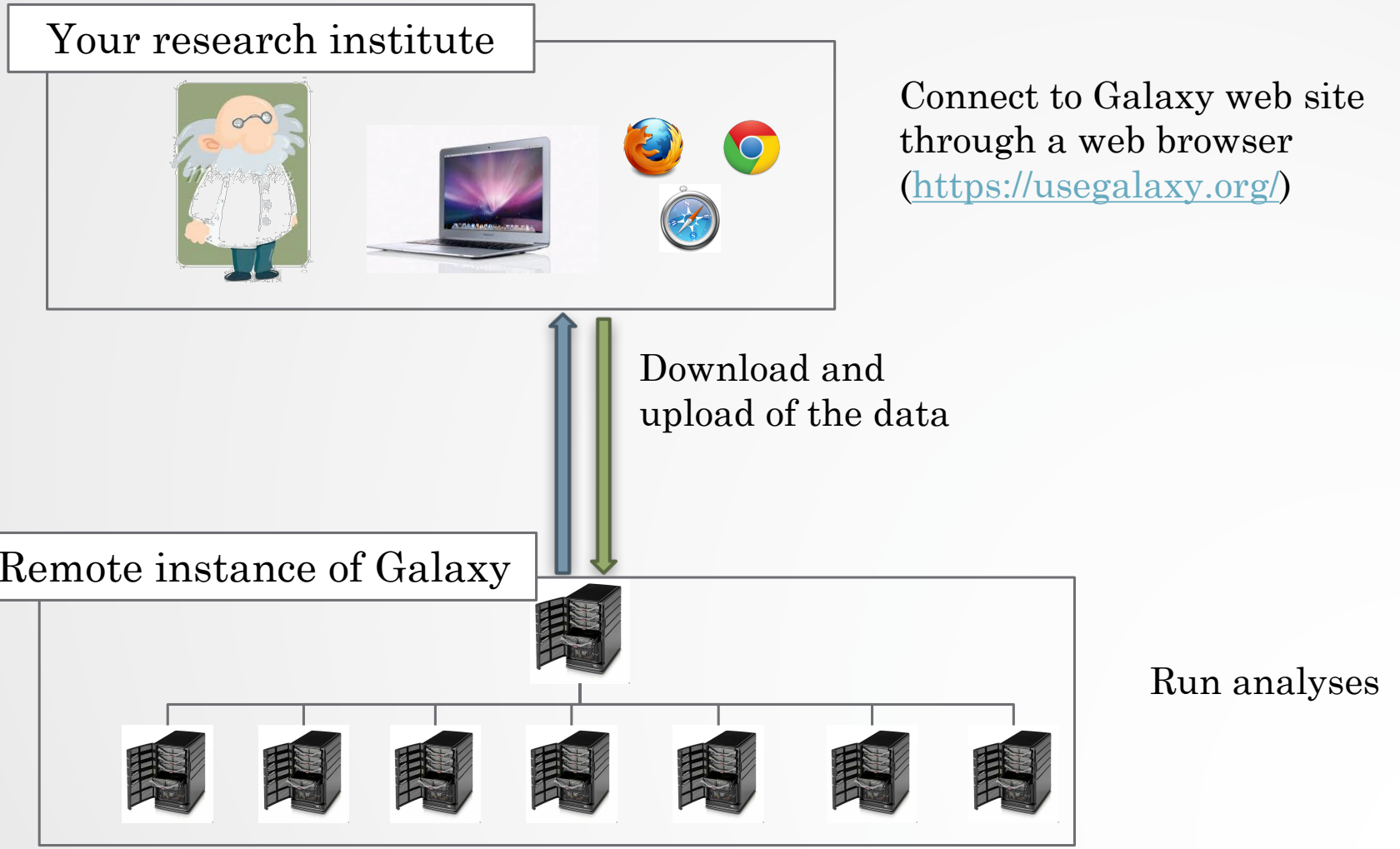
Public Galaxy Servers list :
https://wiki.galaxyproject.org/PublicGalaxyServers
Last Update on: 2017, January 12nd

Galaxy **public servers**

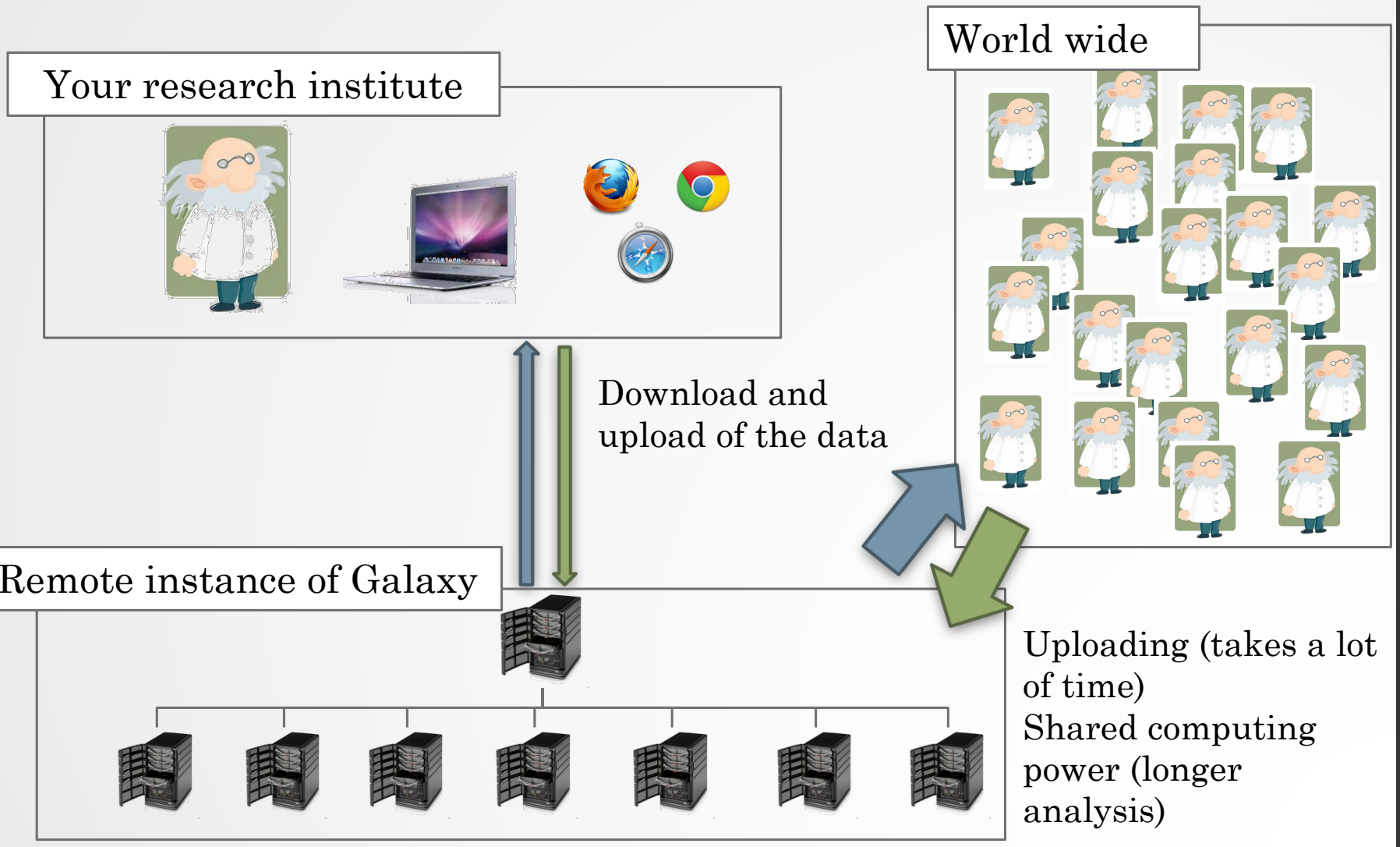
- All analyses are run on remote computing infrastructures
- No need to have a Supercomputer to use Galaxy
- Web browser




Galaxy **public servers**



Galaxy **public servers**

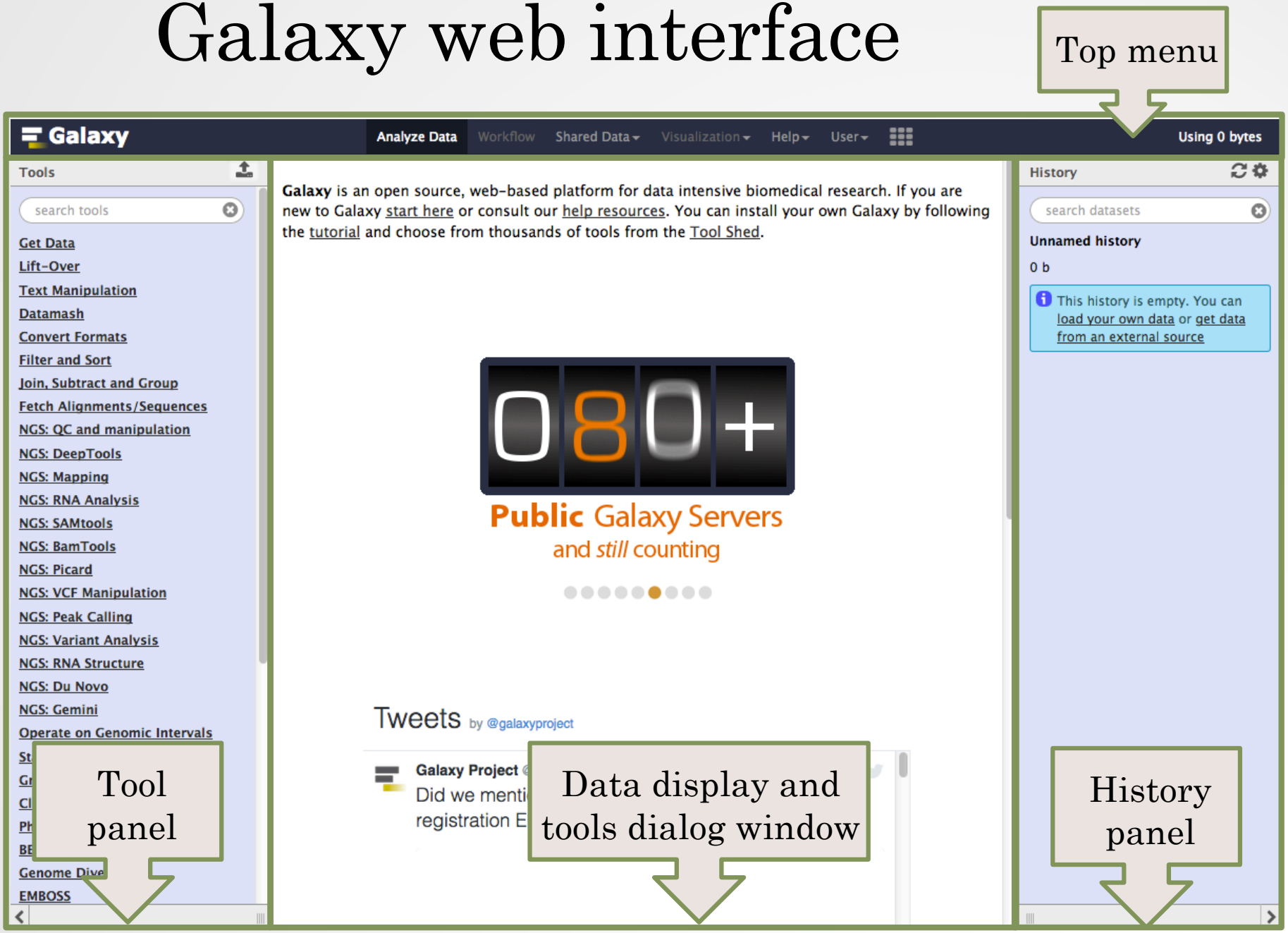


Galaxy **local server**

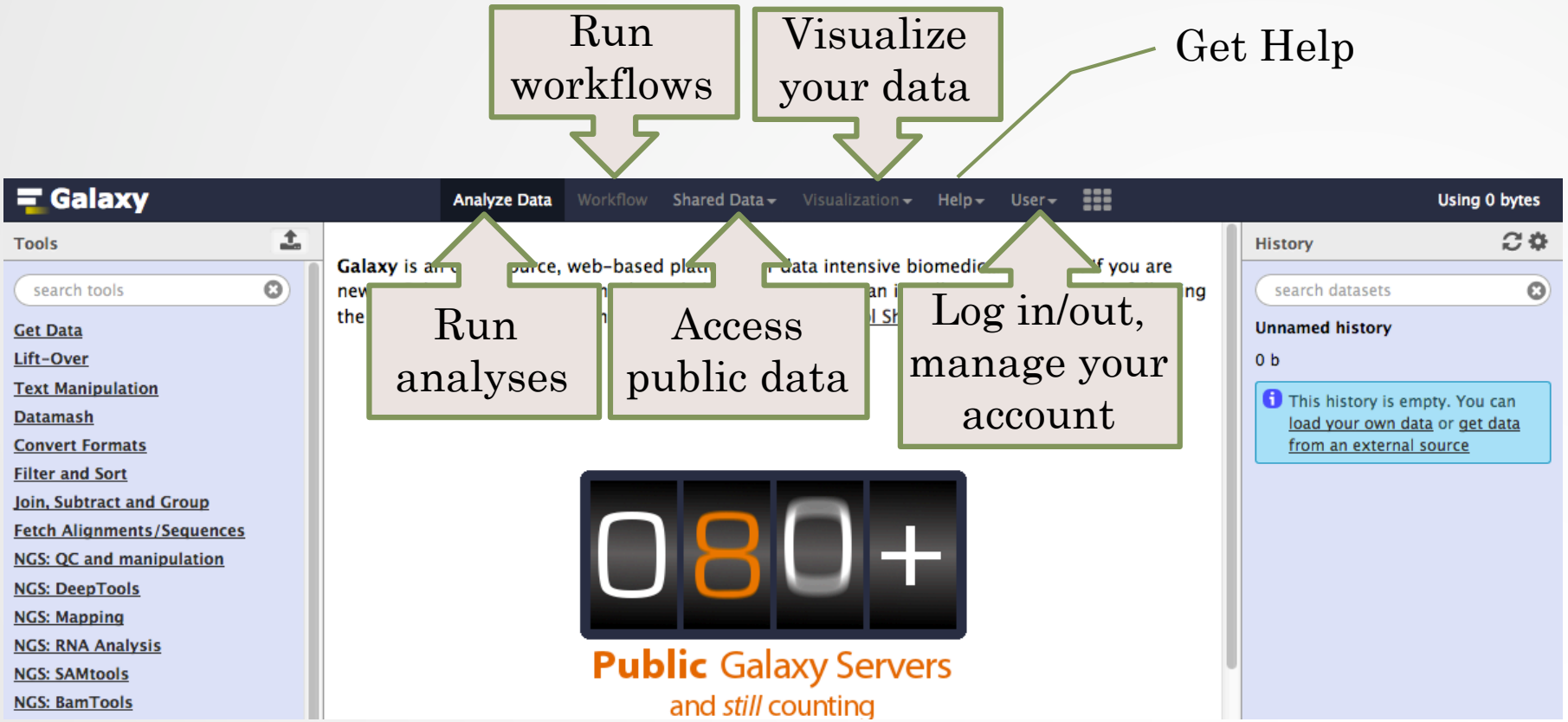
- Run a local production Galaxy because you want to
  - install and use tools unavailable on public Galaxies
  - use sensitive data (e.g. clinical)
  - process large datasets that are too big for public Galaxies
  - plug-in new datasources
  - Develop Galaxy tools
  - Develop Galaxy itself

Description of the main features of Galaxy

Galaxy web interface



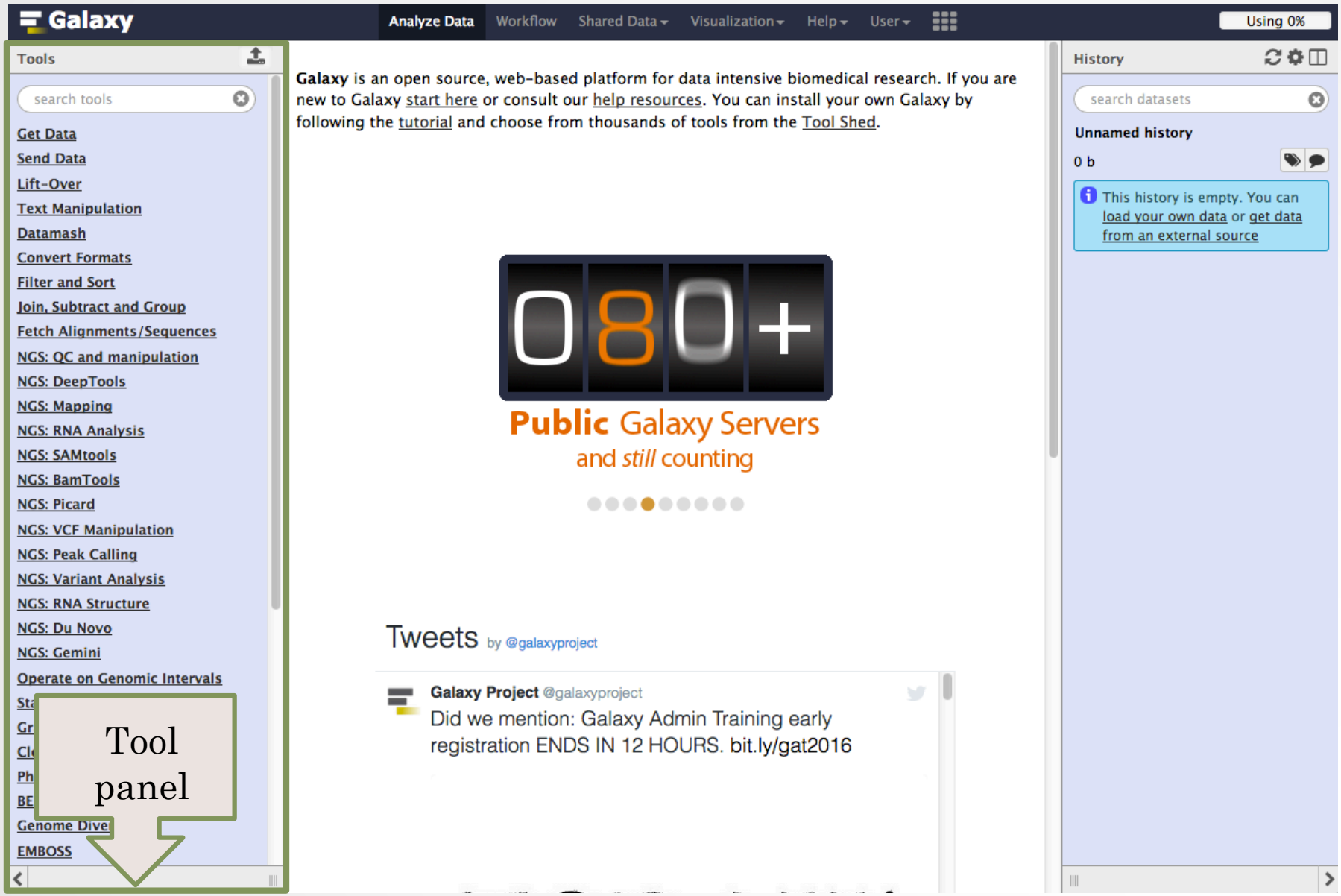
Top menu


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
0 8 0 +
Public Galaxy Servers
and still counting

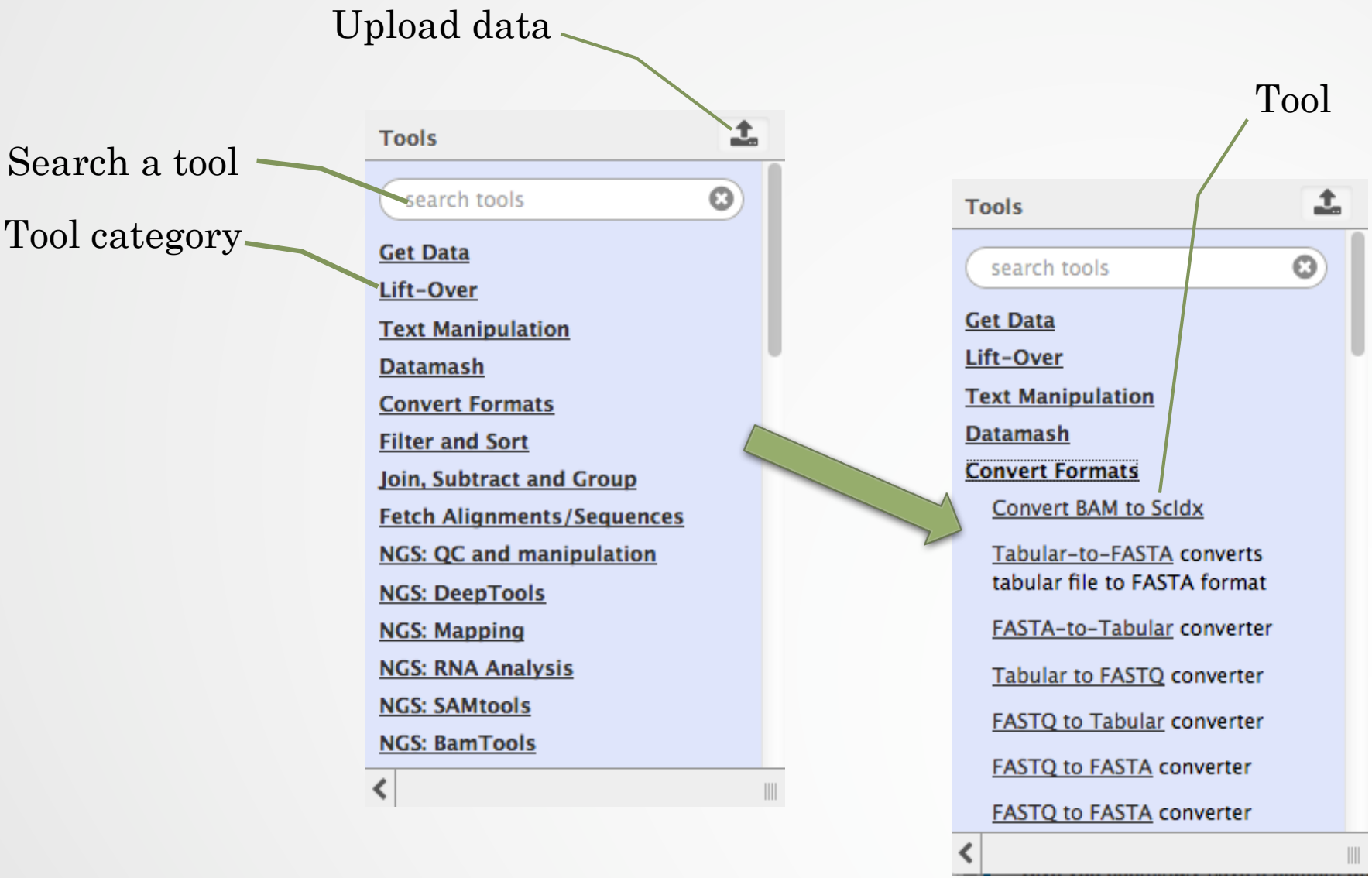
Exercise 1 : Log in

- Go to http://use.galaxeast.fr
- Log in with your training account (look to the post-it pasted into your booklet).

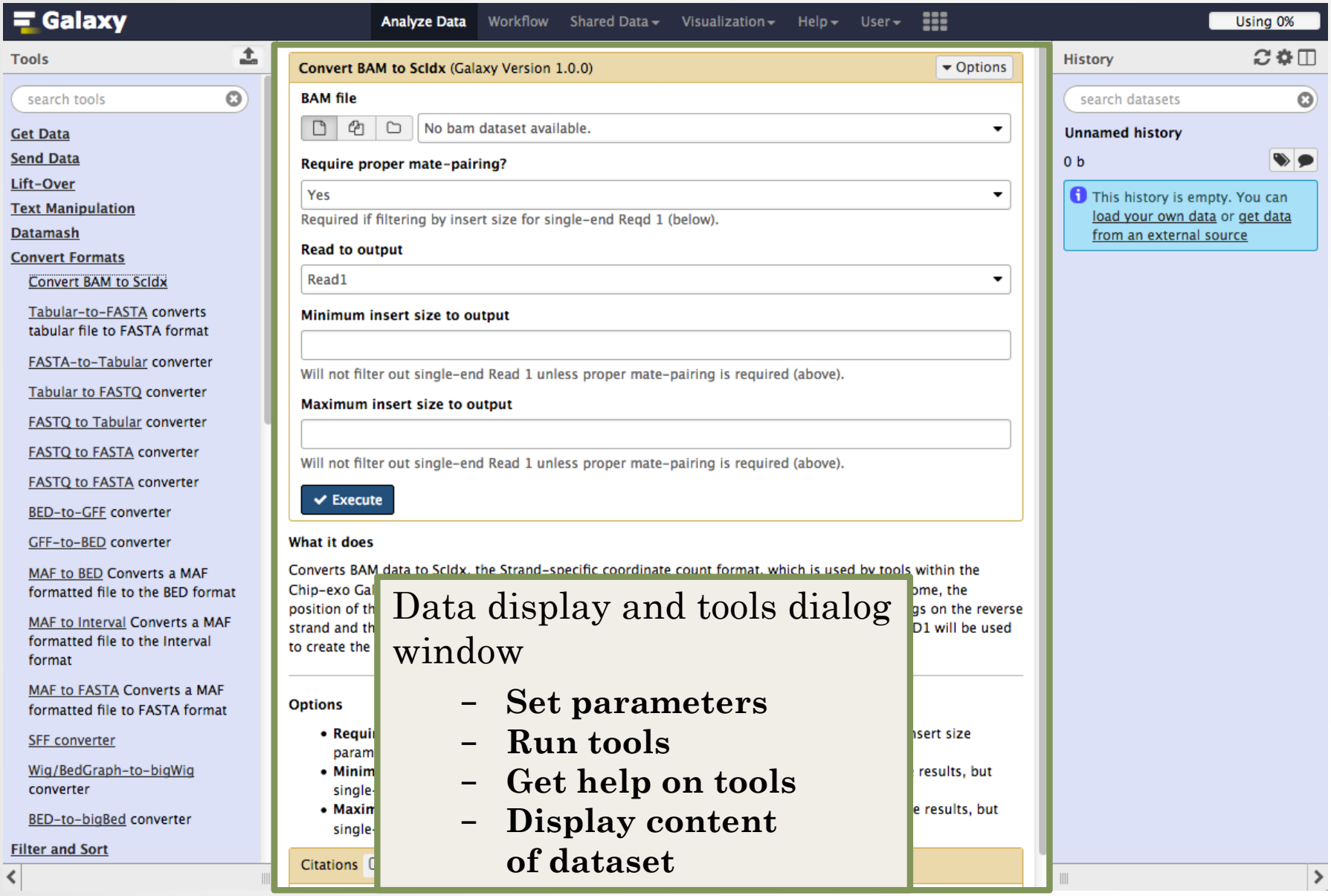
Tool Panel / Run analyses



Tool Panel / Run analyses



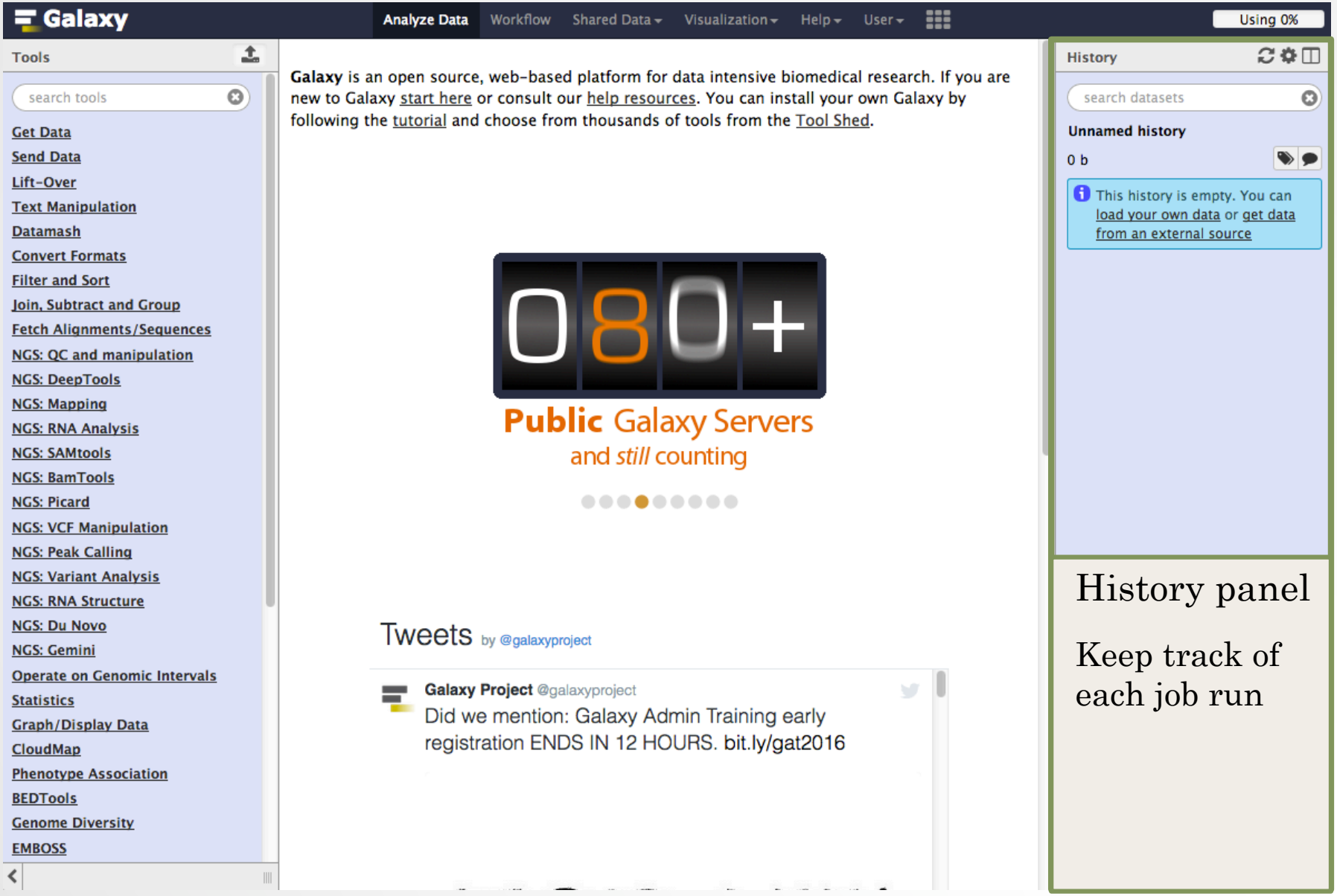
Tools dialog window



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

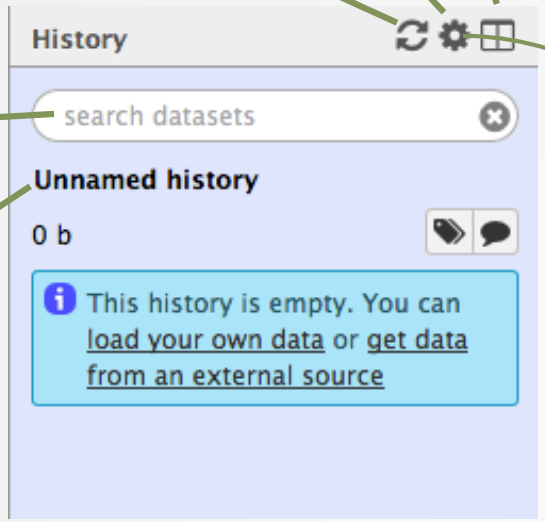
View all histories

History options

Refresh History

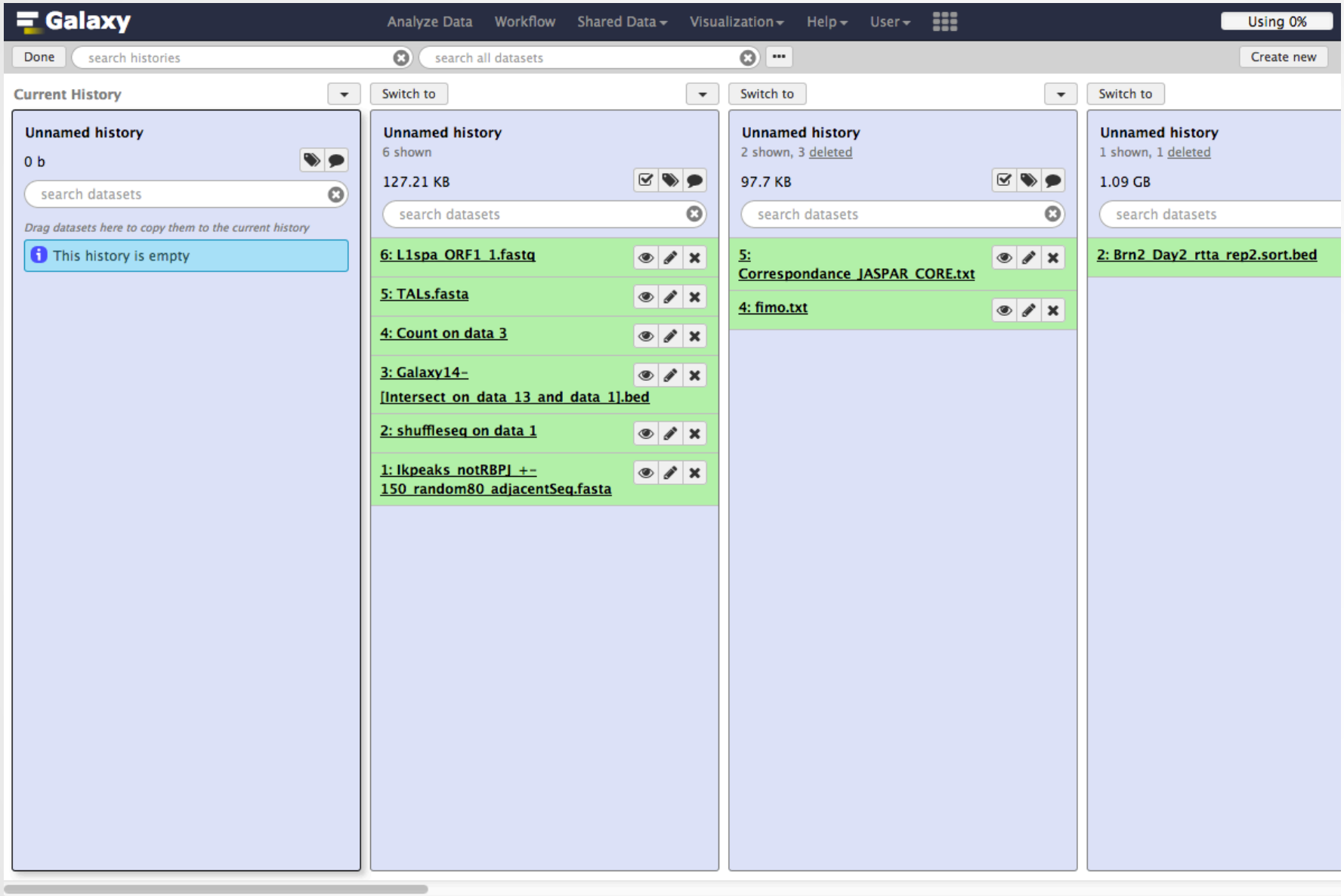
Search datasets

History name



HISTORY 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



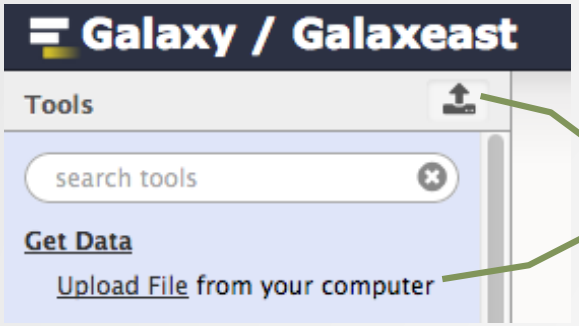
Exercise 2 : History

- Create a new history
- Change the name of the history to “RNA-seq data analysis”

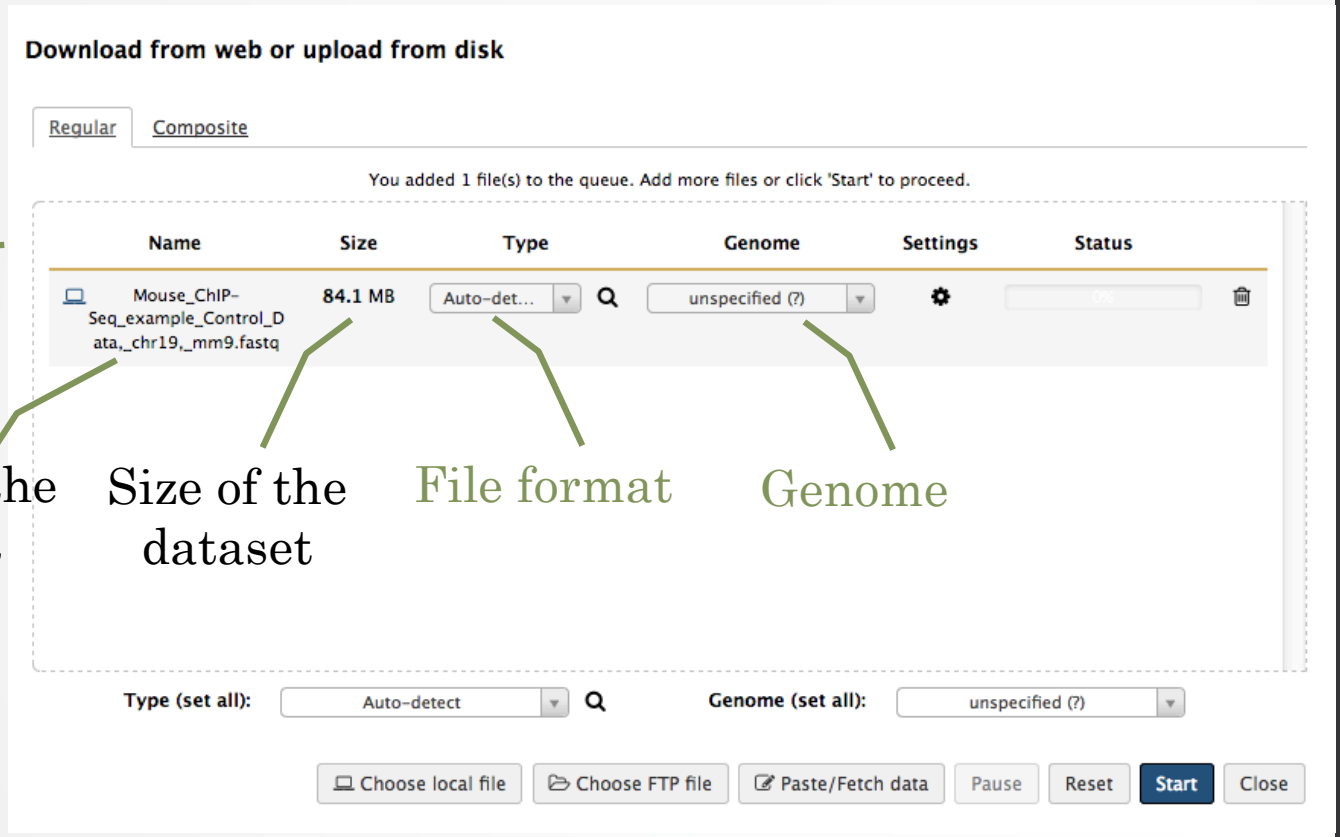
Import data into Galaxy

- 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



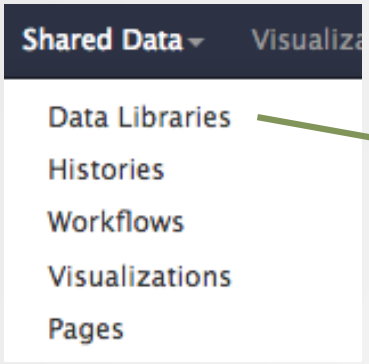
Name of the dataset

Size of the dataset

File format

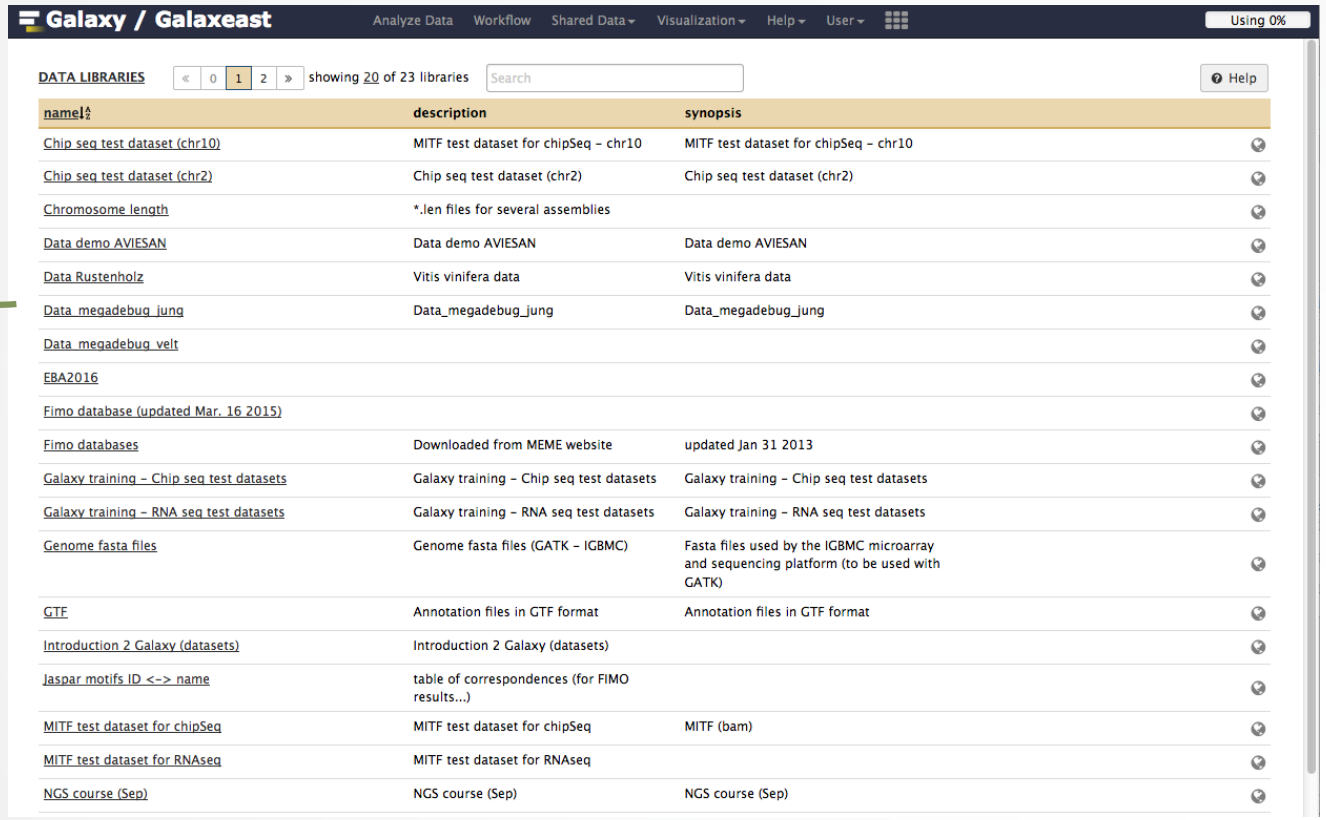
Genome

Import shared data (data libraries)



(Top menu)

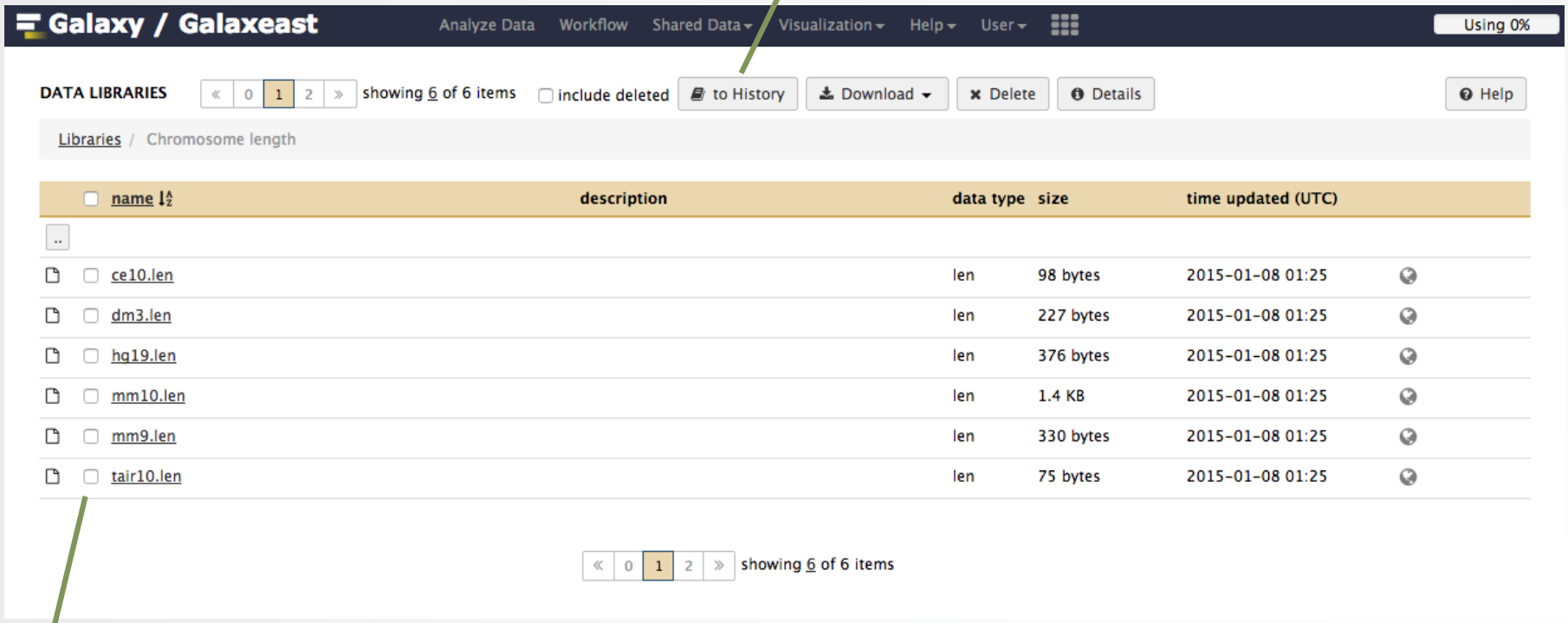
Data Libraries



name	description	synopsis
Chip seq test dataset (chr10)	MITF test dataset for chipSeq – chr10	MITF test dataset for chipSeq – chr10
Chip seq test dataset (chr2)	Chip seq test dataset (chr2)	Chip seq test dataset (chr2)
Chromosome length	*.len files for several assemblies	
Data demo AVIESAN	Data demo AVIESAN	Data demo AVIESAN
Data Rustenholz	Vitis vinifera data	Vitis vinifera data
Data_megadebug_jung	Data_megadebug_jung	Data_megadebug_jung
Data_megadebug_velt		
EBA2016		
Fimo database (updated Mar. 16 2015)		
Fimo databases	Downloaded from MEME website	updated Jan 31 2013
Galaxy training – Chip seq test datasets	Galaxy training – Chip seq test datasets	Galaxy training – Chip seq test datasets
Galaxy training – RNA seq test datasets	Galaxy training – RNA seq test datasets	Galaxy training – RNA seq test datasets
Genome fasta files	Genome fasta files (GATK – IGBMC)	Fasta files used by the IGBMC microarray and sequencing platform (to be used with GATK)
GTF	Annotation files in GTF format	Annotation files in GTF format
Introduction 2 Galaxy (datasets)	Introduction 2 Galaxy (datasets)	
Jaspar motifs ID <-> name	table of correspondences (for FIMO results...)	
MITF test dataset for chipSeq	MITF test dataset for chipSeq	MITF (bam)
MITF test dataset for RNAseq	MITF test dataset for RNAseq	
NGS course (Sep)	NGS course (Sep)	NGS course (Sep)

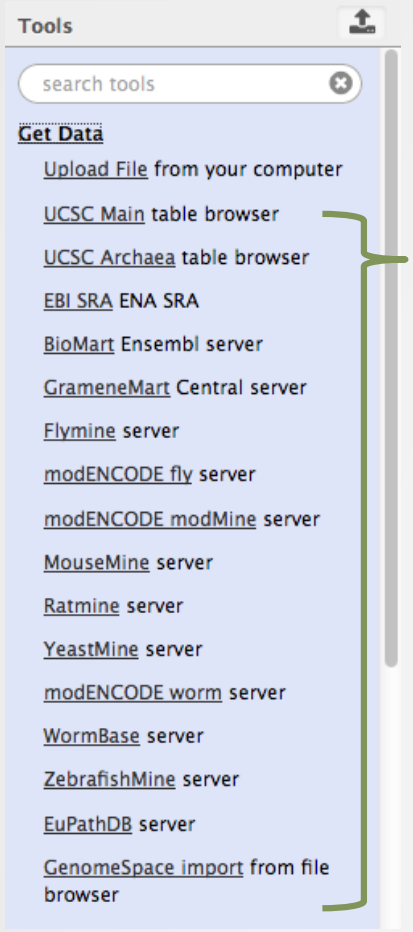
Import shared data (data libraries)

2. Import selected dataset to history

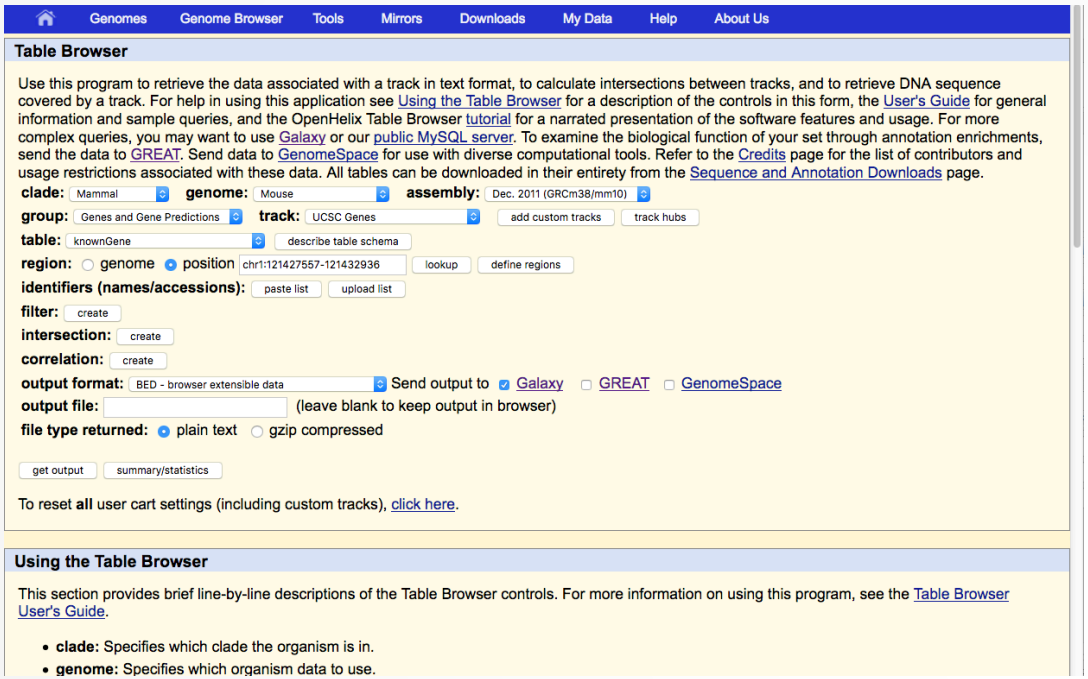


1. Select dataset

Import public data



Browse and import external data from public databases



Exercise 3 : Import data to Galaxy

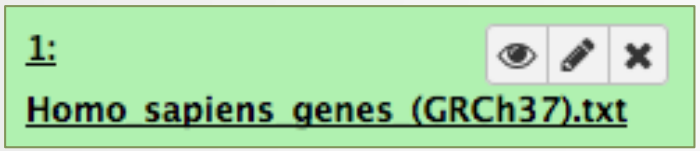
- 1. Import to Galaxy the file siLuc3_S12040.fastq from the “Shared data > Data Libraries > CNRS training > RNAseq > rawdata” to the history called “RNA-seq data analysis”
- 2. You should be in the history “RNA-seq data analysis” (Switch to it if needed)
- 3. Import to Galaxy the file sample.bed.gz located in the directory galaxy.
  - The genome is : Mouse (mm9)
  - The format is : bed

Datasets/Jobs in the History

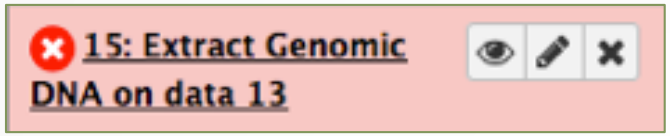
Grey: the job is waiting to run

Yellow: the job is running

Green: the job is successfully done



Red: the job encountered a problem

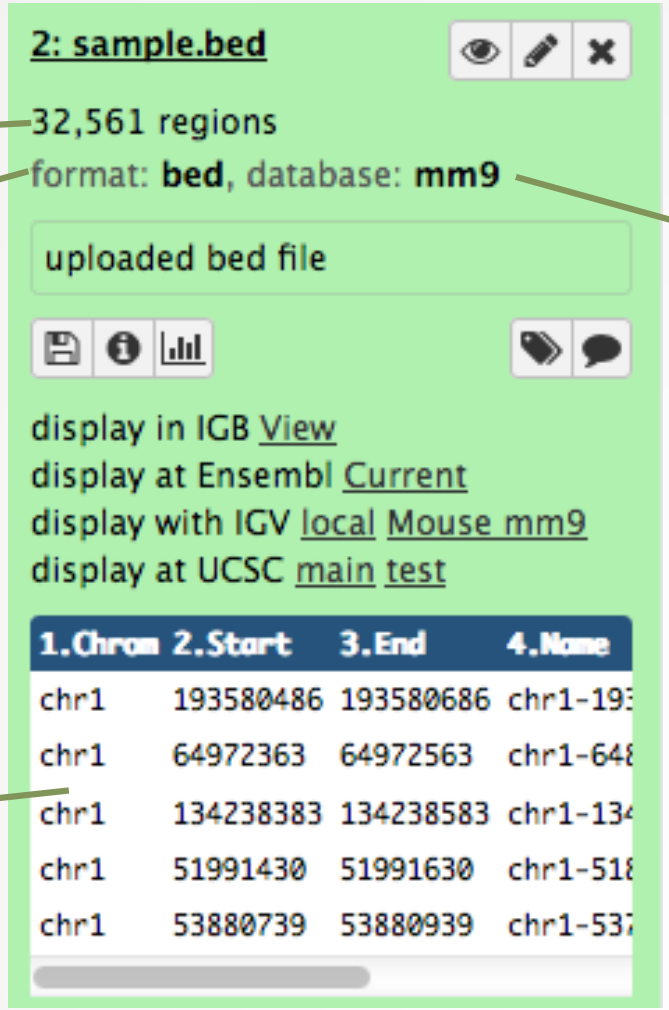


Datasets/Jobs in the History

Number of lines in the file or size of the file

Format

Genome

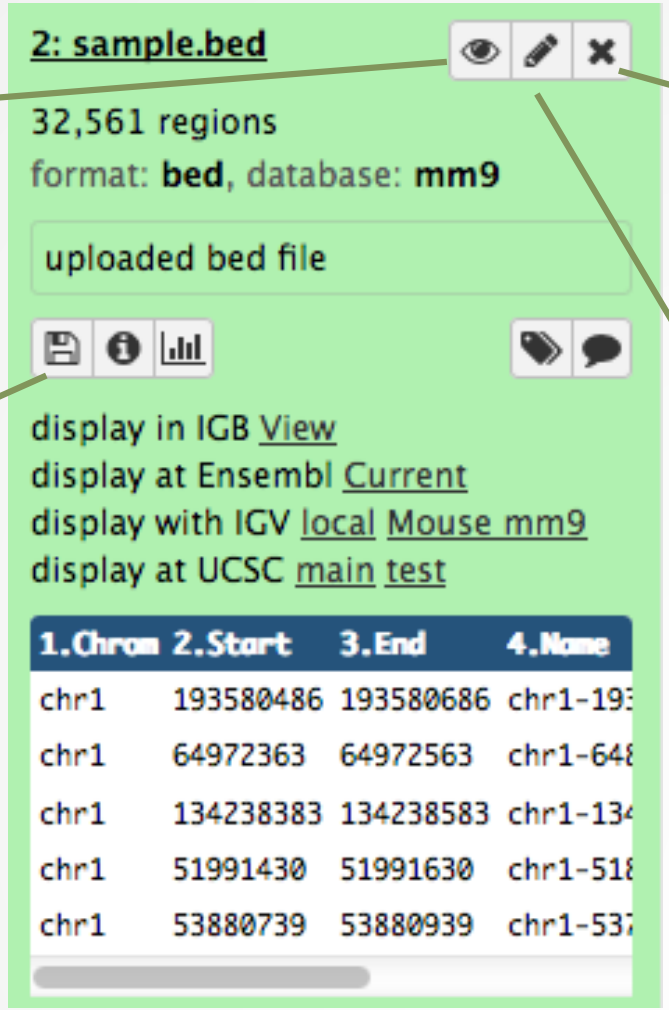


If the dataset is a text file, the first lines of the file are displayed

Datasets/Jobs in the History

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

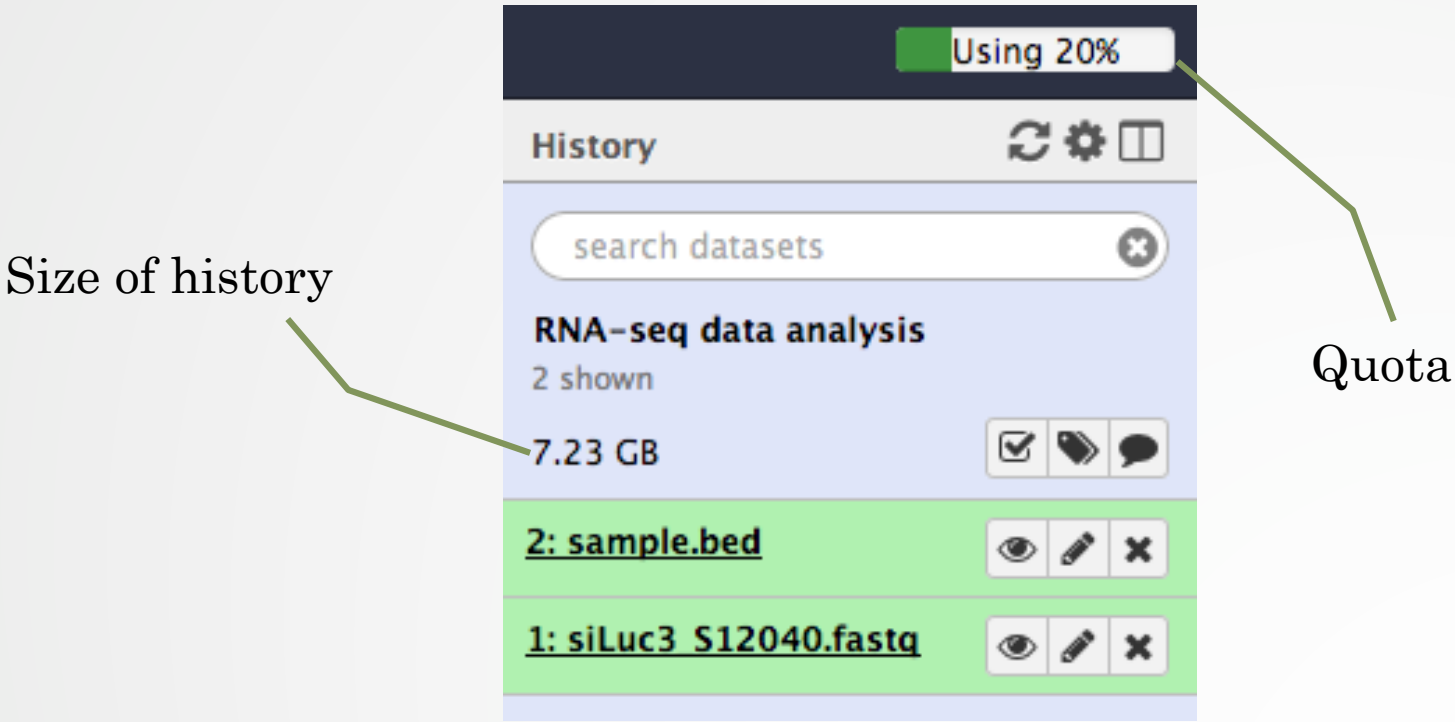
Download dataset



Delete dataset

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

Size of histories and quota



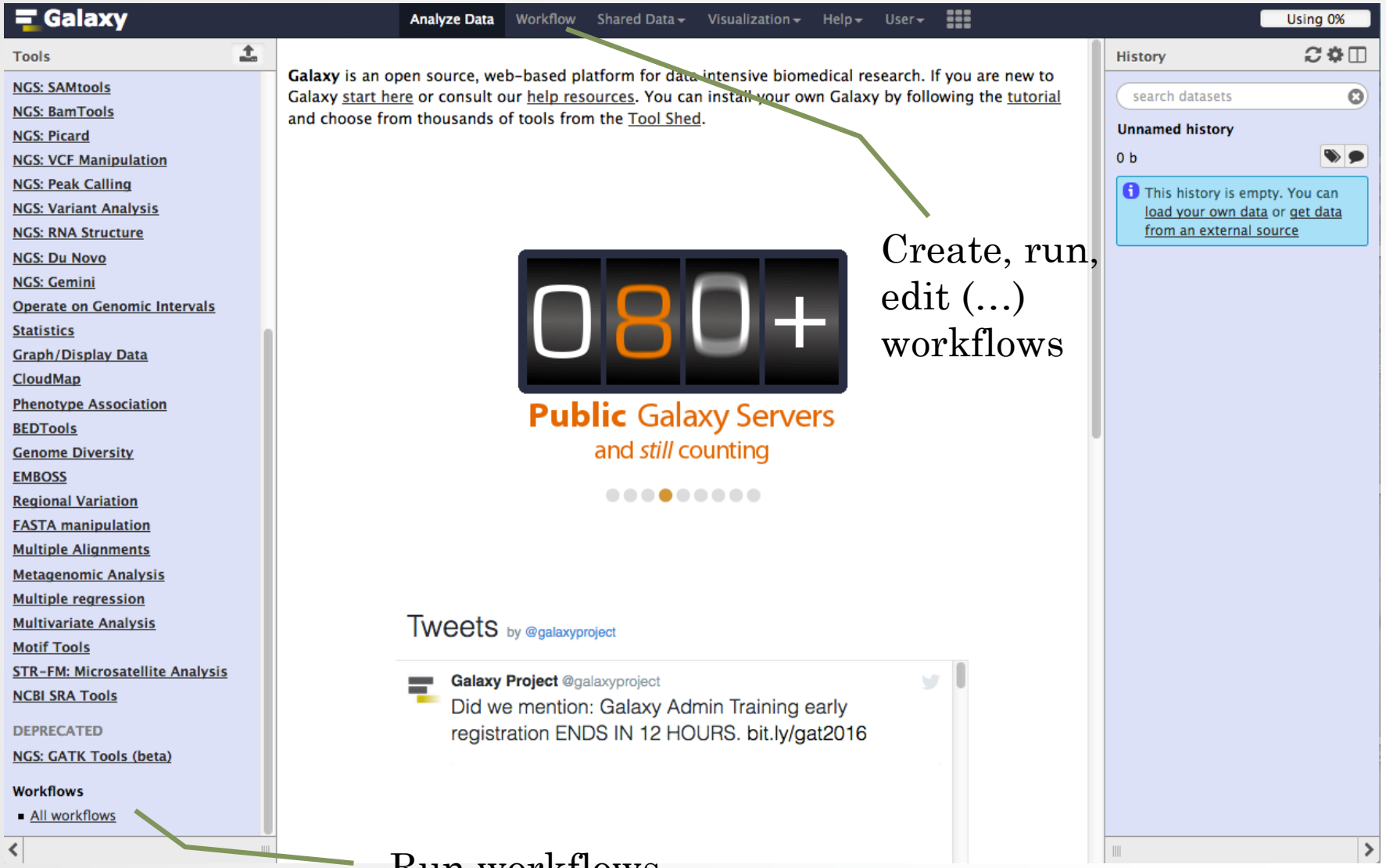
Exercise 4 : remove dataset

- 1. Remove the dataset sample.bed from your history by clicking on the button 
- 2.
  - A. Click on “deleted” in the top of the history panel (below the history name). Remove definitely the file from the disk by clicking on ” Permanently remove it from disk”.
  - B. Click on “hide deleted”

Exercise 5 : Running a tool

- 1. Use the tool "FastQC Read Quality reports" to compute quality analysis on the dataset "siLuc3_S12040.fastq"
  - Use default parameters.

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

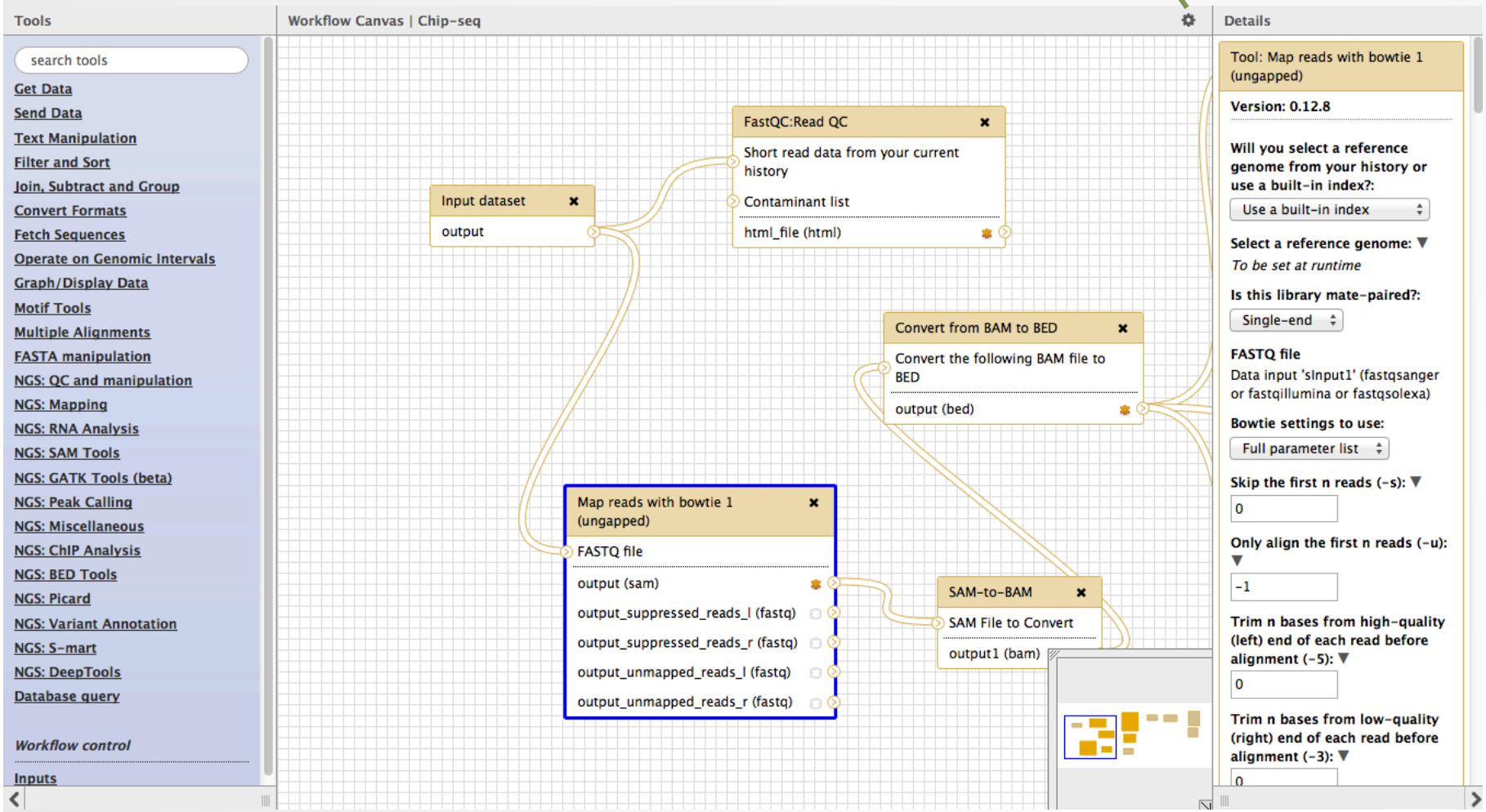
Workflows

Create new workflow

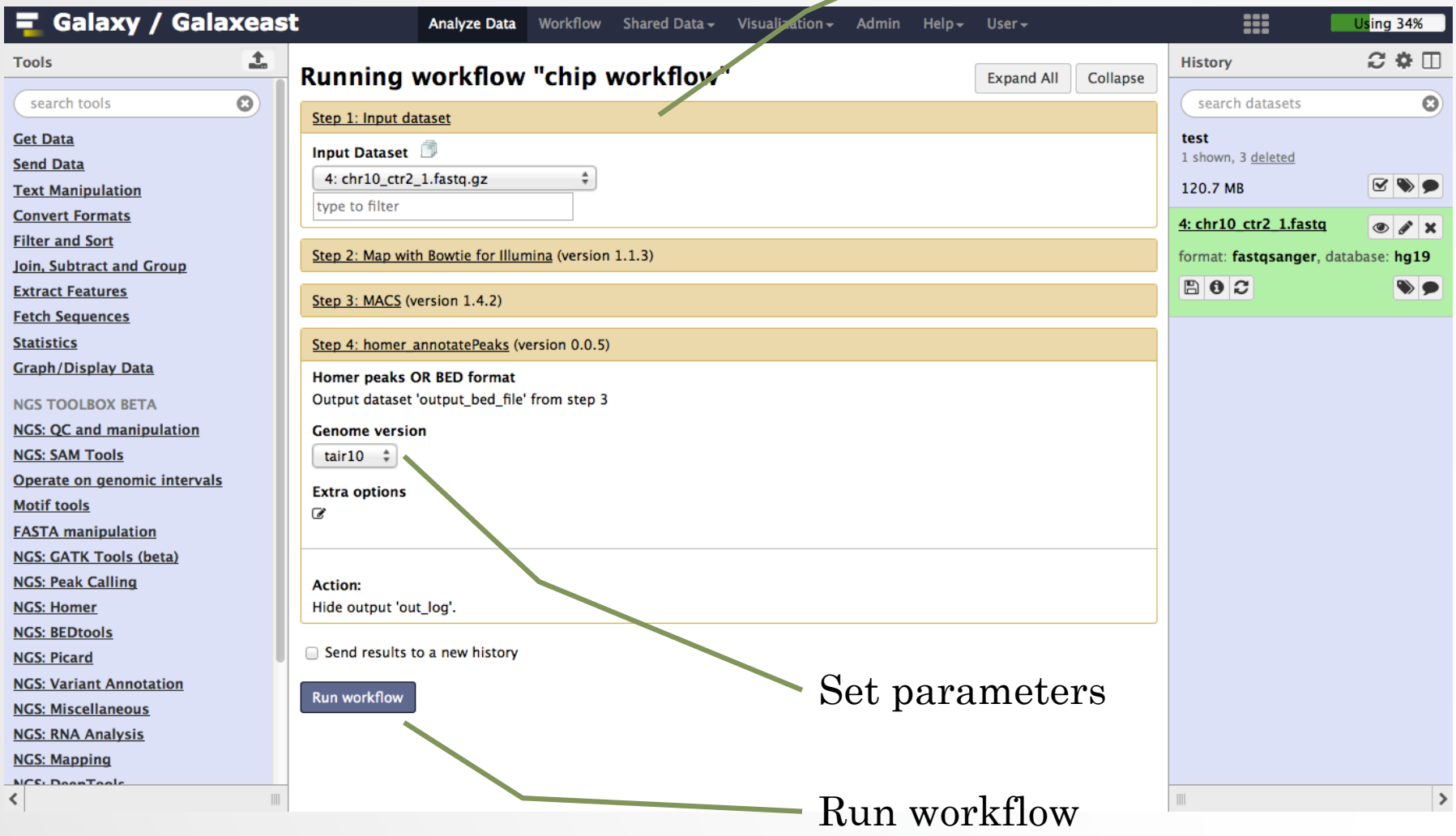
Edit, run, share
(…) workflows

Workflows

Save, run (…) workflows



Workflows


Set input file(s)
Galaxy / Galaxeast
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 34%
Tools
search tools
Get Data
Send Data
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Statistics
Graph/Display Data
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: SAM Tools
Operate on genomic intervals
Motif tools
FASTA manipulation
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: Homer
NGS: BEDtools
NGS: Picard
NGS: Variant Annotation
NGS: Miscellaneous
NGS: RNA Analysis
NGS: Mapping
Running workflow "chip workflow"
Expand All
Collapse
Step 1: Input dataset
Input Dataset
4: chr10_ctr2_1.fastq.gz
type to filter
Step 2: Map with Bowtie for Illumina (version 1.1.3)
Step 3: MACS (version 1.4.2)
Step 4: homer_annotatePeaks (version 0.0.5)
Homer peaks OR BED format
Output dataset 'output_bed_file' from step 3
Genome version
tair10
Extra options
Action:
Hide output 'out_log'.
Send results to a new history
Run workflow
History
search datasets
test
1 shown, 3 deleted
120.7 MB
4: chr10_ctr2_1.fastq
format: fastqsanger, database: hg19
Set parameters
Run workflow

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