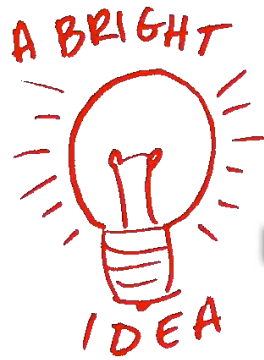# Introduction to Galaxy

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

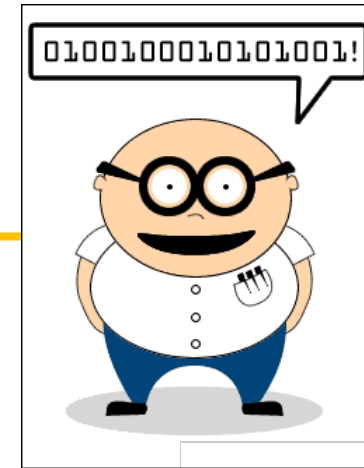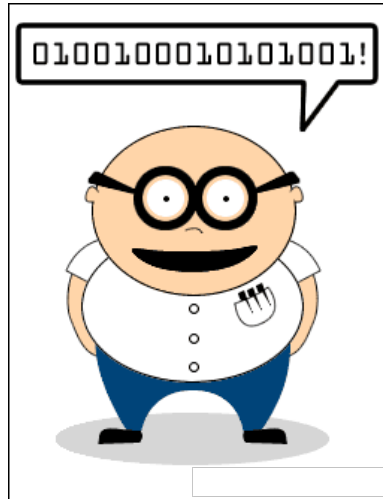# Analyzing biological data with informatics tools

Bioinformatics analyses

Informatics data

Enter

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.
```
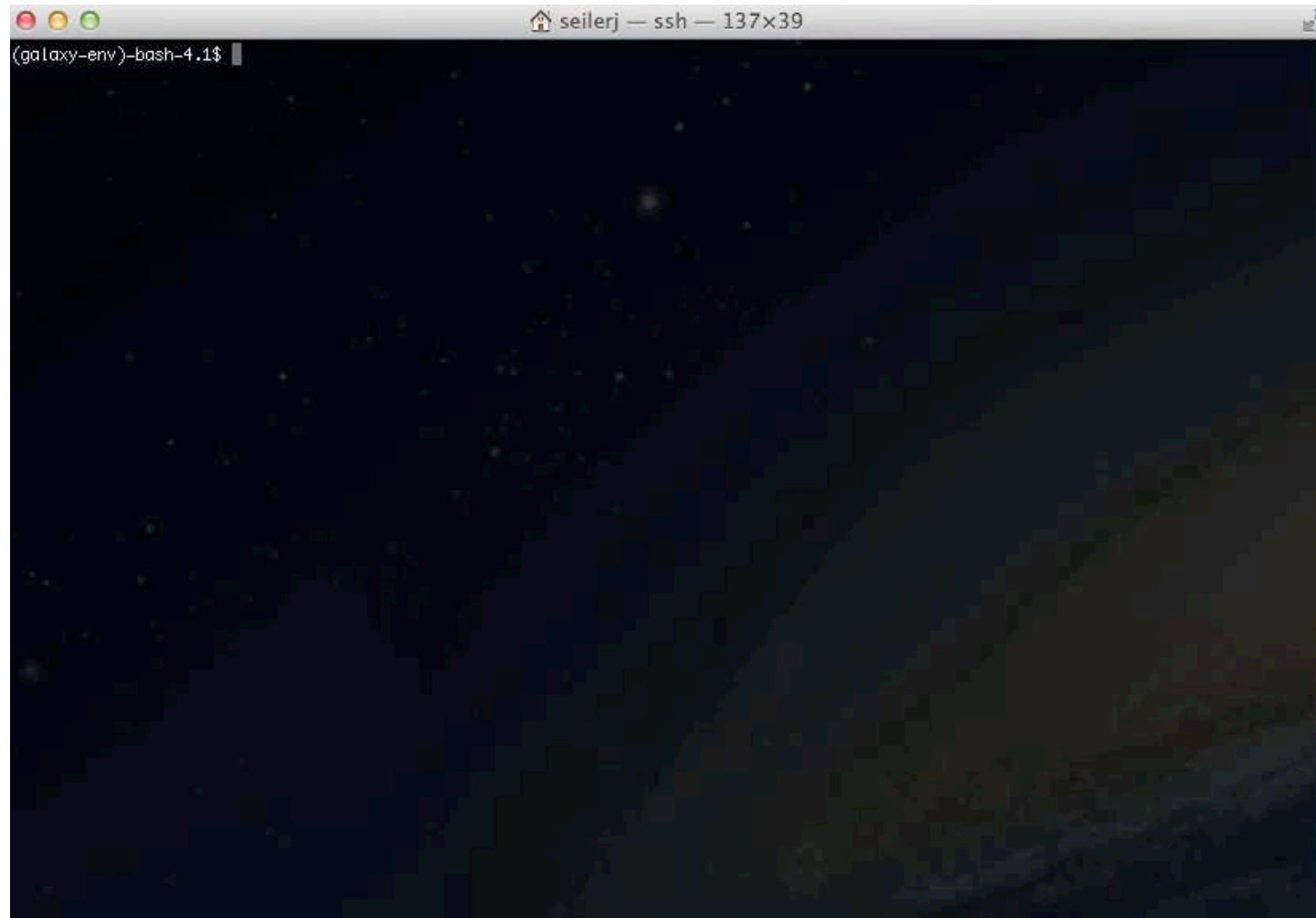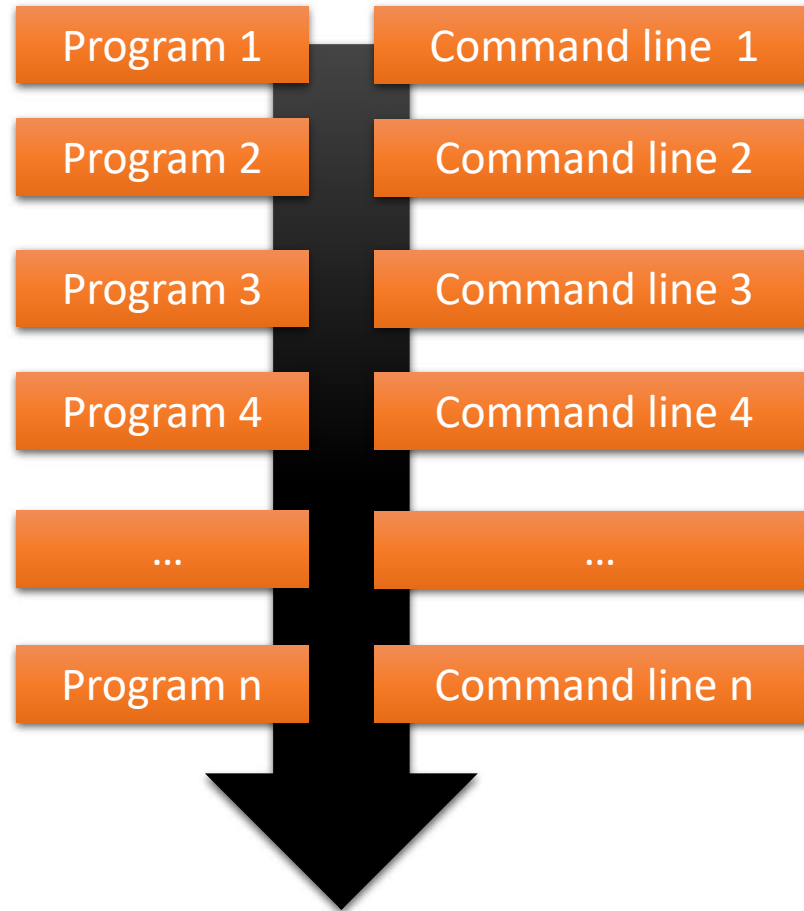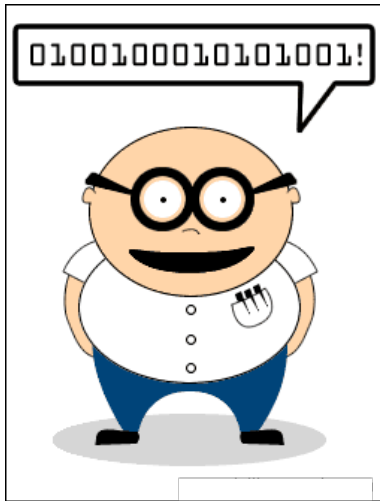
Command line

```
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

Galaxy ?

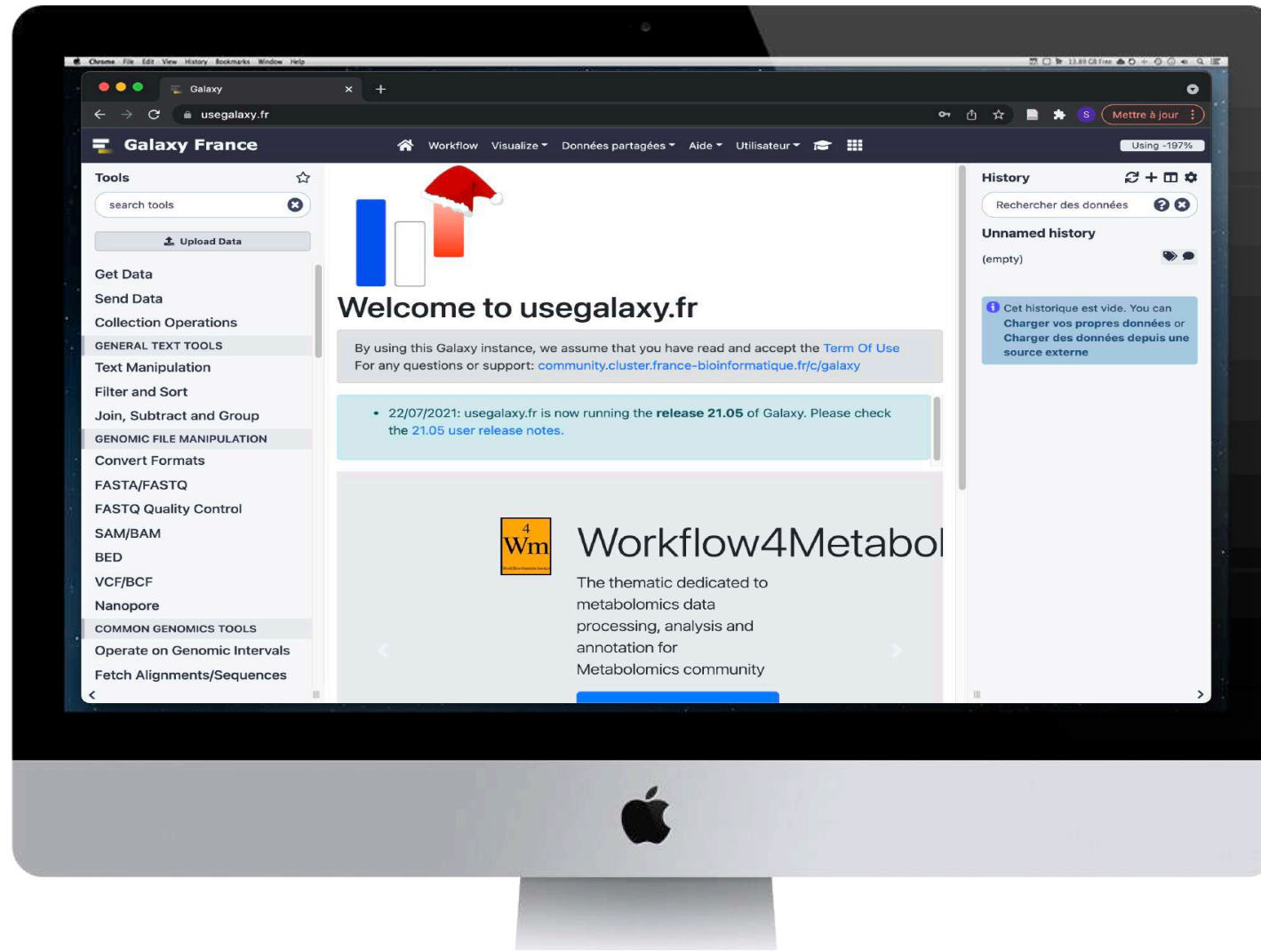# Galaxy project

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**.
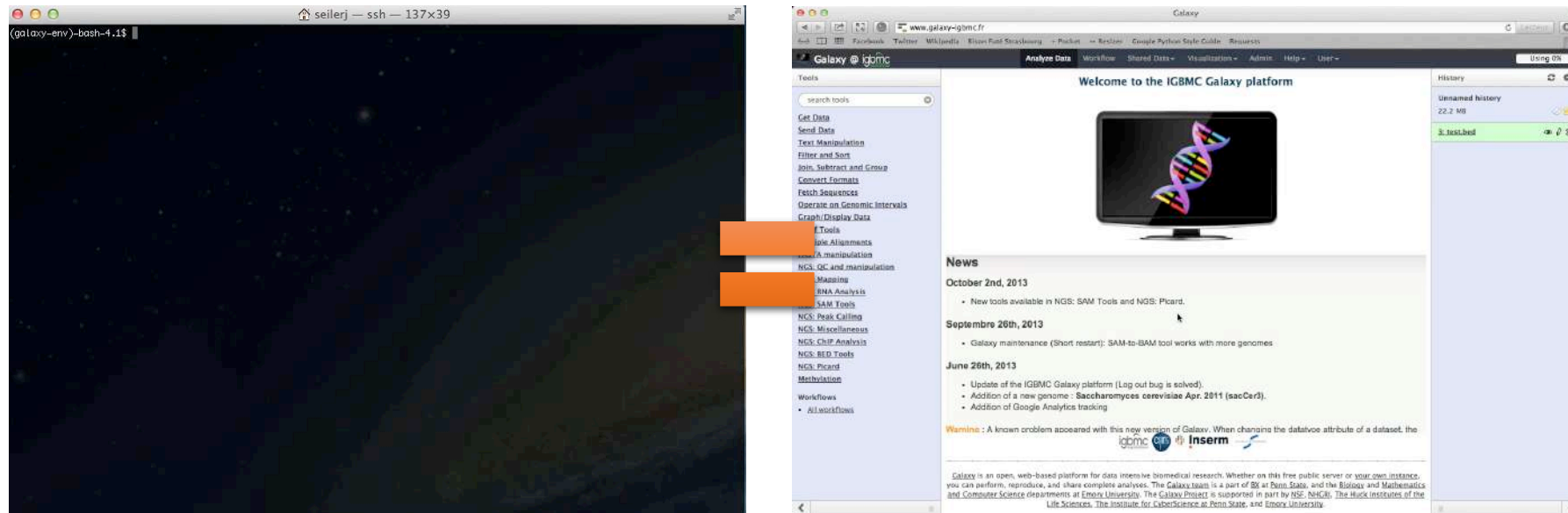
This is Galaxy

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

Use Galaxy

	UseGalaxy Servers	Public Servers	TIaaS	Academic Clouds	Commercial Clouds	Containers	VMs	Local
Free to use	Yes	Yes	Yes	Yes[1]	No	Yes	Yes	Yes
Uses your local compute infrastructure	No	No	No	No	No	Yes[2]	Yes[2]	Yes
Datasets (including intermediate) total > 250GB	No	?[5]	Yes	Yes	Yes	Yes[3]	Yes[3]	Yes
Computational requirements are similarly large	No	?[5]	Yes	Yes	Yes	Yes[3]	Yes[3]	Yes
Share Galaxy objects outside your organization	Yes	Yes	Yes	Yes	Yes	Yes[4]	Yes[4]	Yes[5]
Install custom tools and reference genomces	No	No	No	Yes[5]	Yes	Yes	Yes	Yes
Have absolute data security requirements	No	No	No	?[5]	?[5]	?[5]	?[5]	Yes

* TIaaS: Training Infrastructure as a Service

- https://galaxyproject.org/use/
  - 171 resources for using Galaxy (Last Update on: 2021, June 7th)

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

Your research institute

World wide

Download and
upload of the data

Remote instance of Galaxy

Uploading (takes a lot of
time)
Shared computing power
(longer analysis)

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

Tool panel

Data display and tools dialog window

History panel

Top menu

Exercise 1

History

History panel

Keep track of each job run

History

History/Dataset options

View all histories side by side

Create new history

Refresh History

Search datasets

History ⟳ + ⊡ ⚙

Rechercher des données ❓ ❌

Unnamed history

(empty) 🏷 💬

History name

ⓘ Cet historique est vide. You can
Charger vos propres données or
**Charger des données depuis une
source externe**

History Actions

Copy

Partager et publier

Montrer la structure

Extraire un Workflow

Set Permissions

Make Private

Reprendre les processus en pause

Actions sur les jeux de données

Copier des jeux de données

Réduire les données étendues

Afficher les données cachées

Supprimer les données cachées

Purger les données supprimées

Télécharger

Exporter les citations des outils

Exporter l'Historique dans un fichier

Beta Features

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

Galaxy France

Tools ☆

search tools ⊗

⬆ **Upload Data**

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Display the drag and drop utility used to upload local files

Download from web or upload from disk

| Regular | Composite | Collection | Rule-based |

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
💻 CRN-107_11-R1.fast⟨	**18.5** MB	Auto-de... ⊽ 🔍	----- Additional ... ⊽	⚙	0% 🗑

Name of the dataset

Size of the dataset

File format

Genome

Type (set all): Auto-detect ⊽ 🔍 **Genome (set all):** ----- Additional ... ⊽

⬜ Choose local files   📂 Choose remote files   ☑ Paste/Fetch data   **Start**   Pause   Reset   Close

Import shared data (data libraries)

1.   2.

List of data
libraries

Import shared data (data libraries)
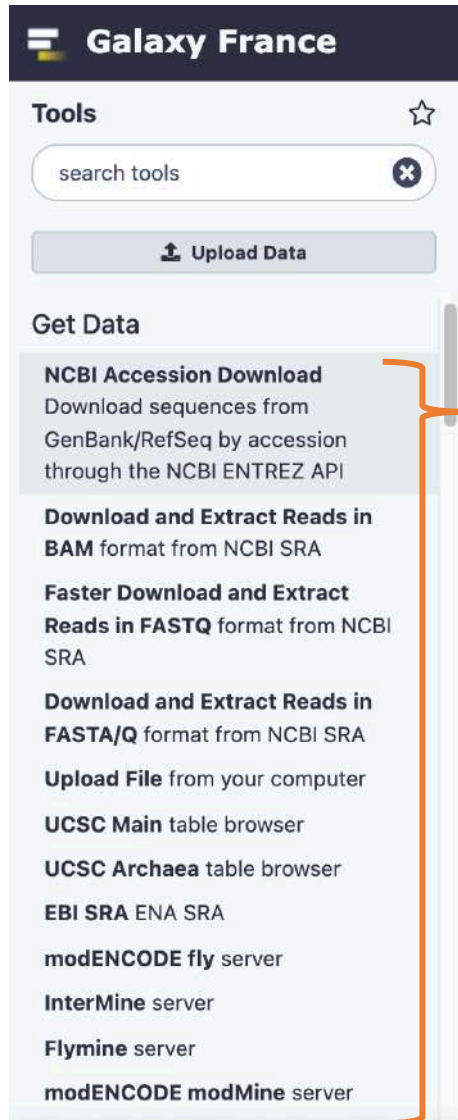
2. Import selected dataset
to history

1. Select dataset

Import public data

Browse and import external data from public databases

Exercise 3

Datasets/Jobs in the History

Grey: the job is waiting to run

🕐 4: sample.bed.gz   👁 ✏ ✖

Orange: the job is running

⋰ 4: sample.bed.gz   👁 ✏ ✖
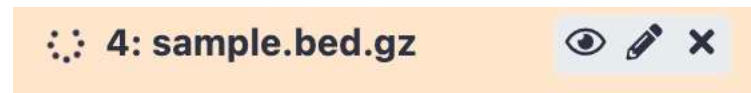
Green: the job is successfully done

4: sample.bed.gz   👁 ✏ ✖
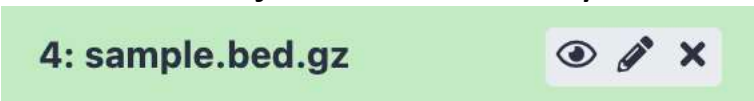
Red: the job encountered a problem

❌ 15: Extract Genomic DNA on data 13   👁 ✏ ✖
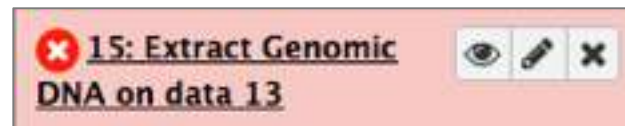
Datasets/Jobs in the History

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

4: sample.bed.gz

32,561 regions
format: **bed**, génome de référence: **mm10**

uploaded bed file

display in IGB View
display with IGV local
display at UCSC main test

1.Chrom	2.Start	3.End	4.Name
chr1	193580486	193580686	chr1-1934573
chr1	64972363	64972563	chr1-6486016
chr1	134238383	134238583	chr1-1341694
chr1	51991430	51991630	chr1-5187923
chr1	53880739	53880939	chr1-5376854

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

Size of history

Quota

Exercise 4

Tool Panel / Run analyses

Tool Panel / Run analyses

Upload data

Search a tool

Tool category

Tool

Tools ☆

search tools ✖

⬆ Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

Convert Formats
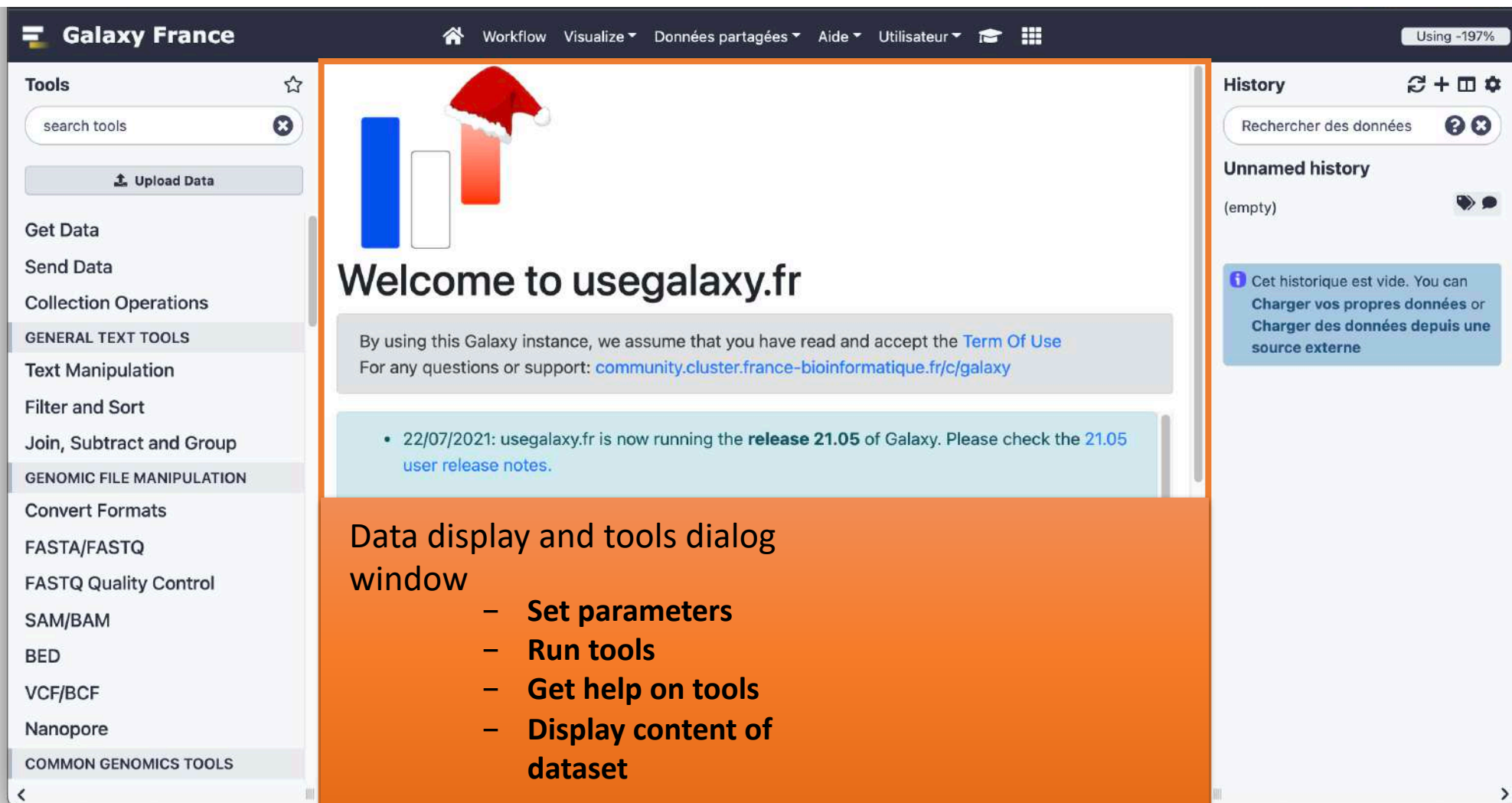
FASTA/FASTQ

Text Manipulation

Query Tabular using sqlite sql

annotateMyIDs annotate a generic set of identifiers

Compute an expression on every row

Replace column by values which are defined in a convert file

Sort a row according to their columns

Column Regex Find And Replace

Regex Find And Replace

43

Tools dialog window

Data display and tools dialog window
- – **Set parameters**
- – **Run tools**
- – **Get help on tools**
- – **Display content of dataset**

Exercise 5

Exercise 6

Workflow

What if we'd mix all together

Galaxy workflow
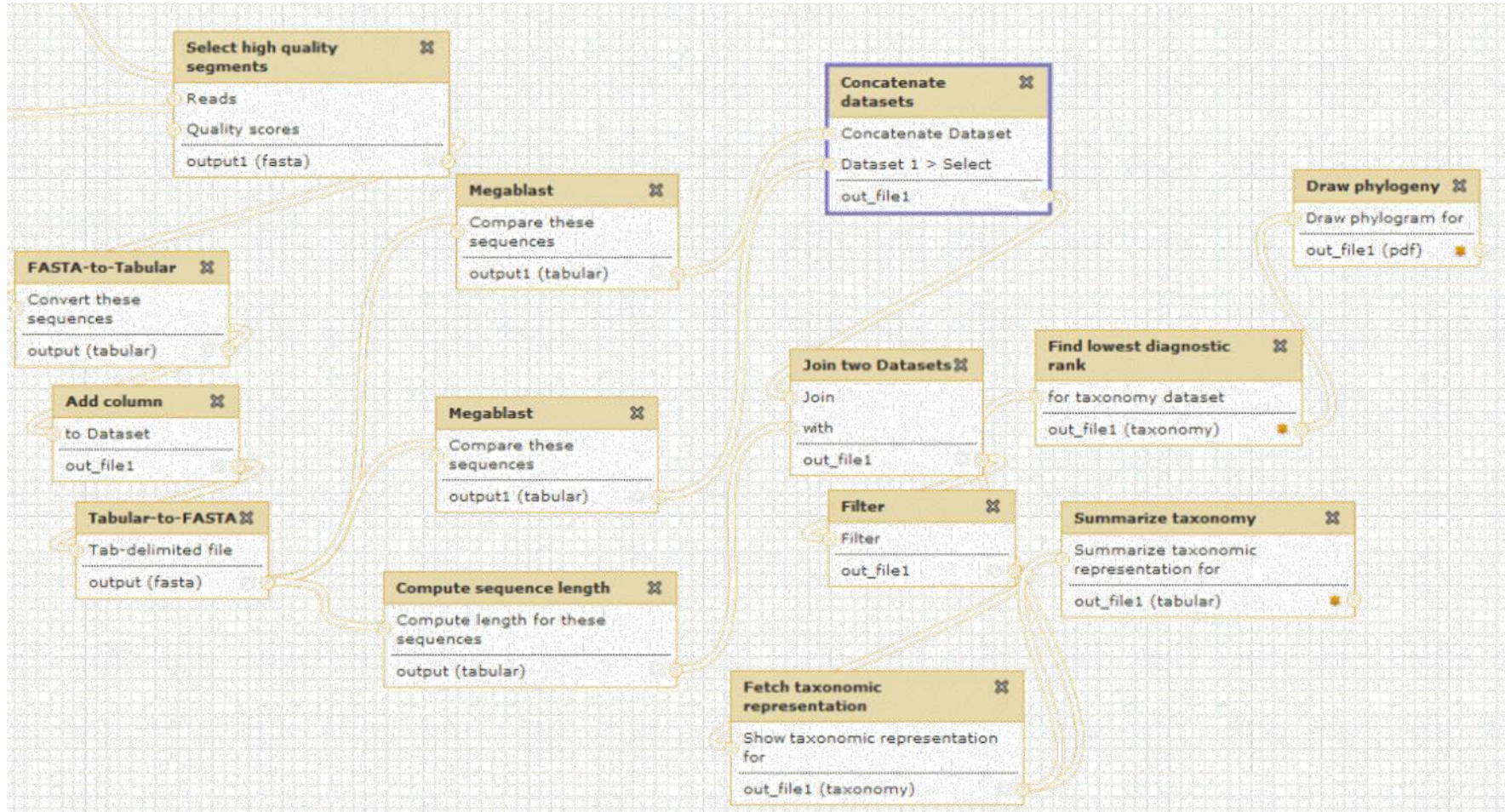
Galaxy workflows

- 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

51

Workflows

Create workflows

List of Workflows

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

56

Workflow creation

Save workflow

Run

Run workflows

Set input file(s). It has to be a dataset from your current history

Run workflow

Set parameters

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