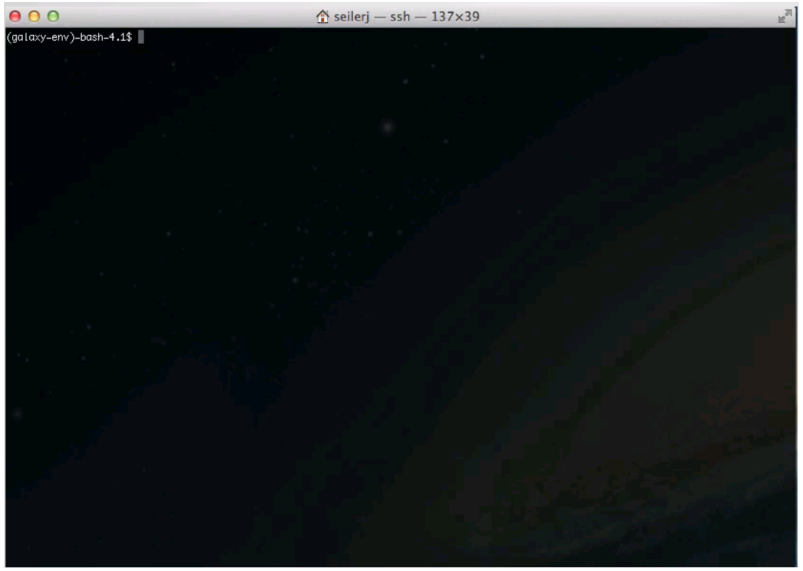
Your Bioinformatic Analyses with Galaxy

Sarah Maman – Maria Bernard

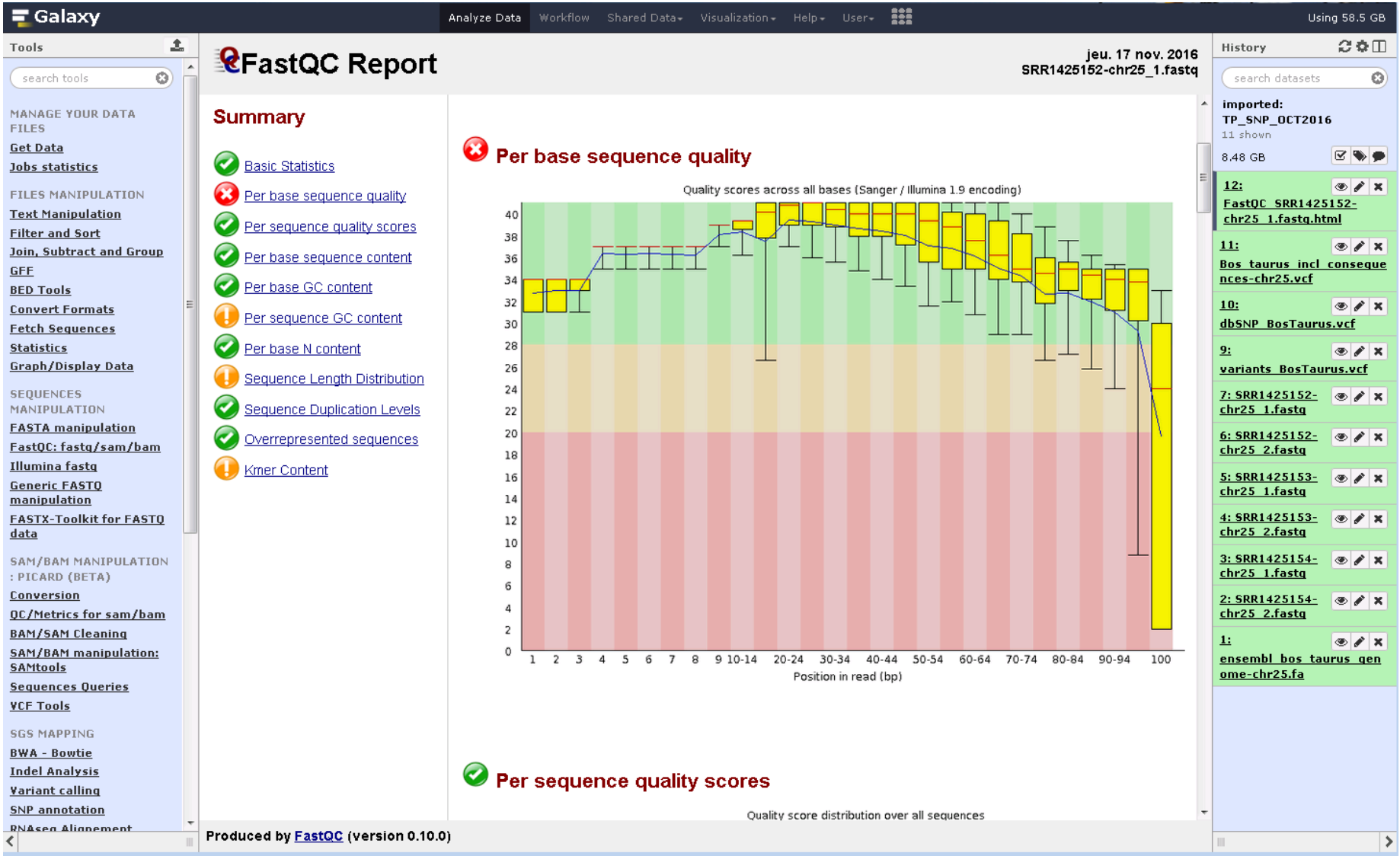
Montpellier 2016

Classical bioinformatic




```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
```

Bioinformatic with Galaxy



Galaxy

Initial Galaxy Project team (https://www.galaxyproject.org/) :

- The Center for Comparative Genomics and Bioinformatics of the University of Penn State,
- Departments of Biology and Mathematics and Computer Science of the University of Emory.

Anton Nekrutenko
Penn State

Nate Coraor
Penn State

James Taylor
Emory

Galaxy

- Galaxy Phylosophy

Useless to know:

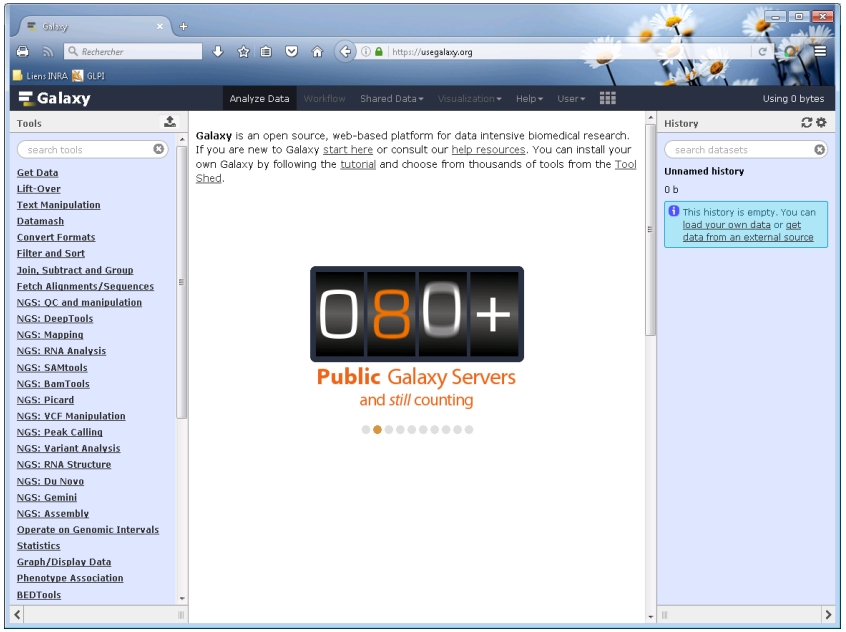
Command lines
Programming in perl, python, shell …
To launch a script

To Allow:

Automatisation of analyses,
Reproduce them,
Share them,
Publish then in a **transparency** way

Galaxy, that's:

An **intuitive way** to do bioinformatic,
Open source,
In **constant evolution** thanks to a very **large community of (bio)informatic scientists**



A Galaxy among others

- Galaxy is a web server whith more and more instances across the world (+ 91 public servers and + 5 public tool repositories)

Public server (https://main.g2.bx.psu.edu/):

Free
Limited quota
Unprotected data
➔ usefull to familiarise yourself on common tool on small dataset

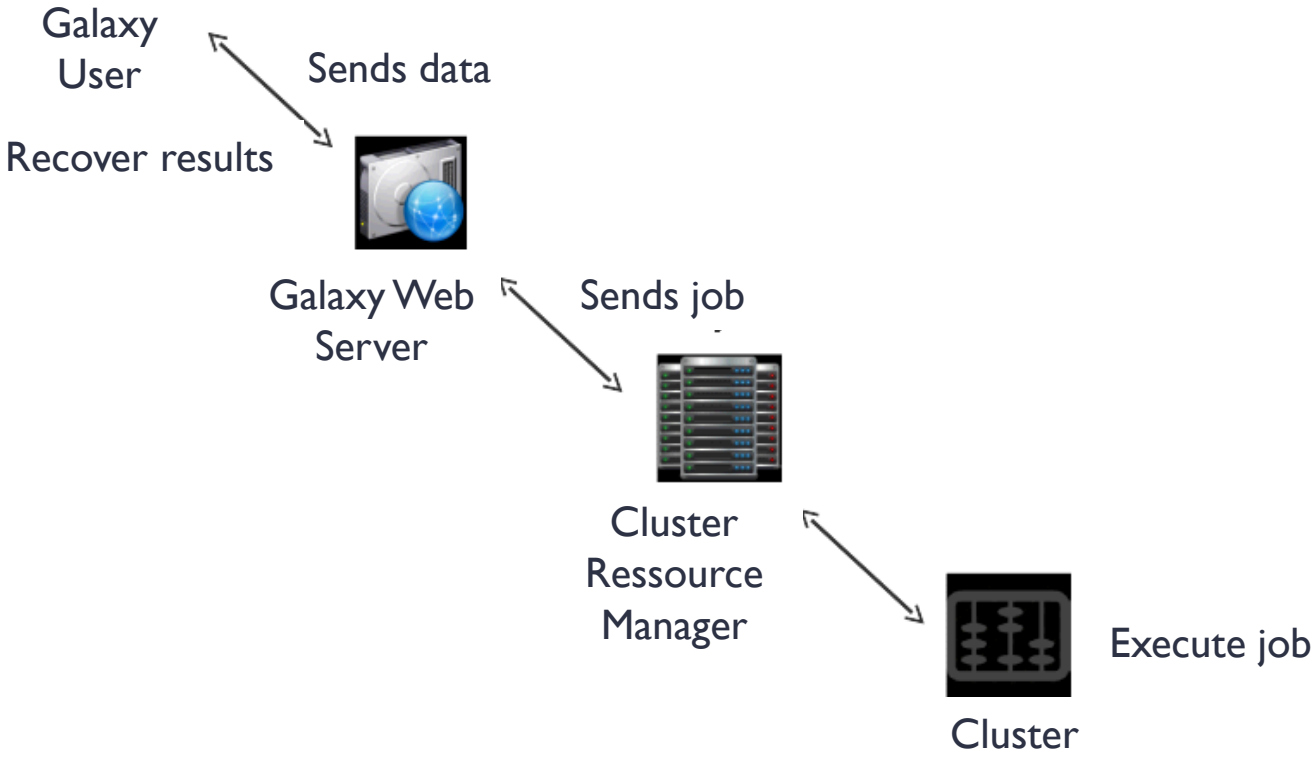
In France : in particular http://www.france-bioinformatique.fr/fr/groupes-de-travail/galaxy


…

How Galaxy works ?

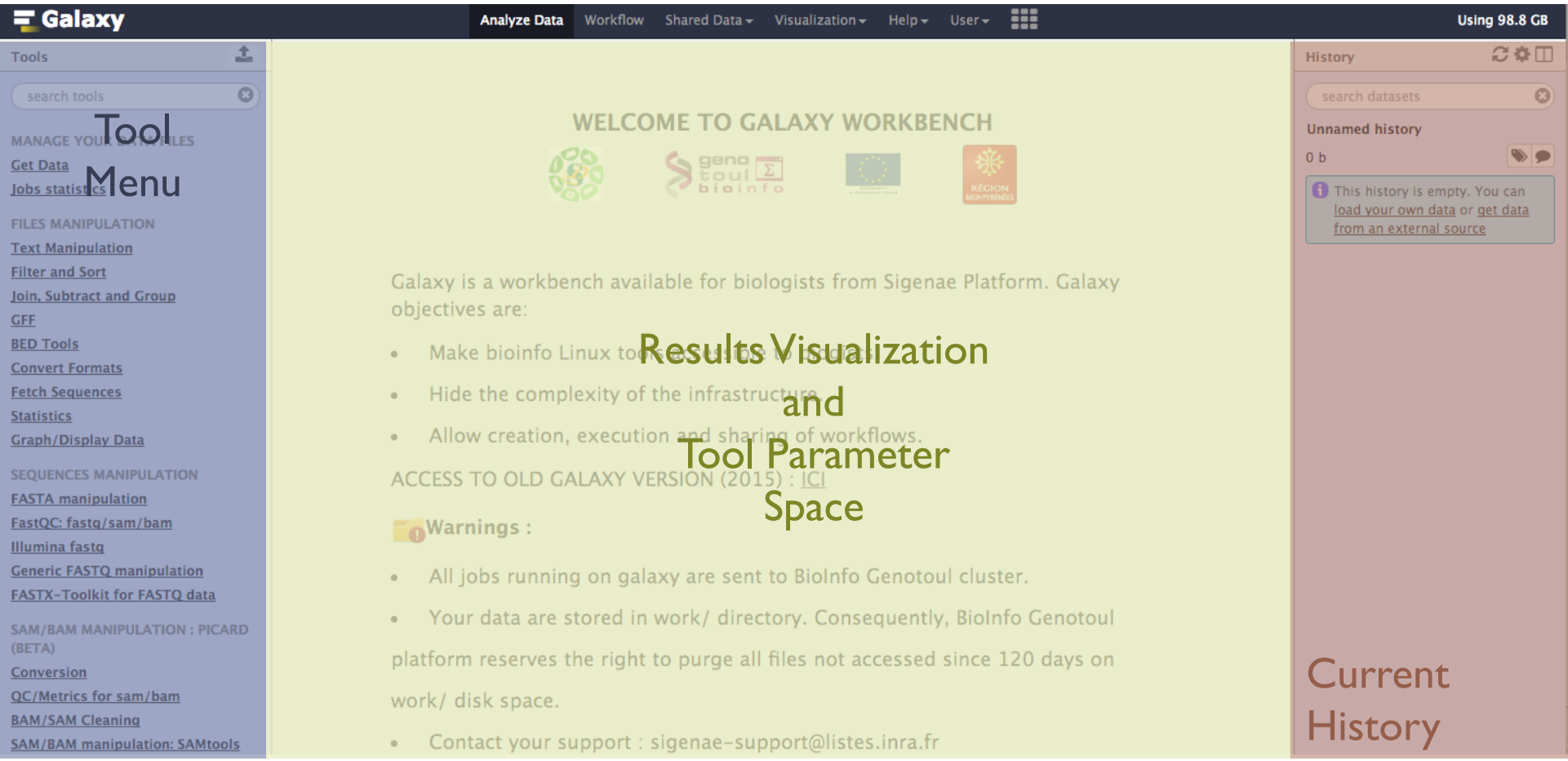
Useless to wait the end of a treatment:

- Possible to **launch multiple** jobs in parallel
- Go away, ..**close the Internet** !
- See the results later

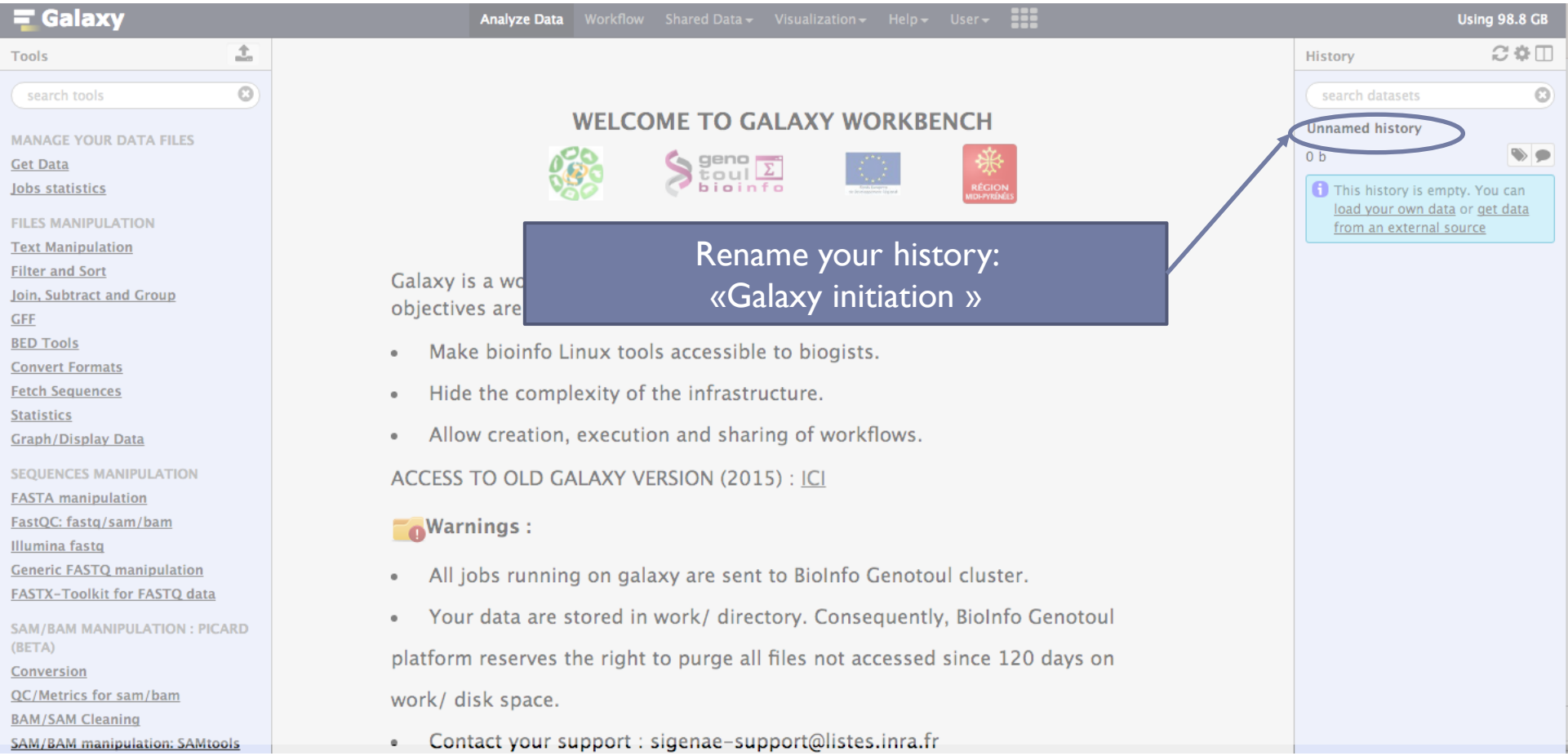


http://sigenae-workbench.toulouse.inra.fr

Galaxy

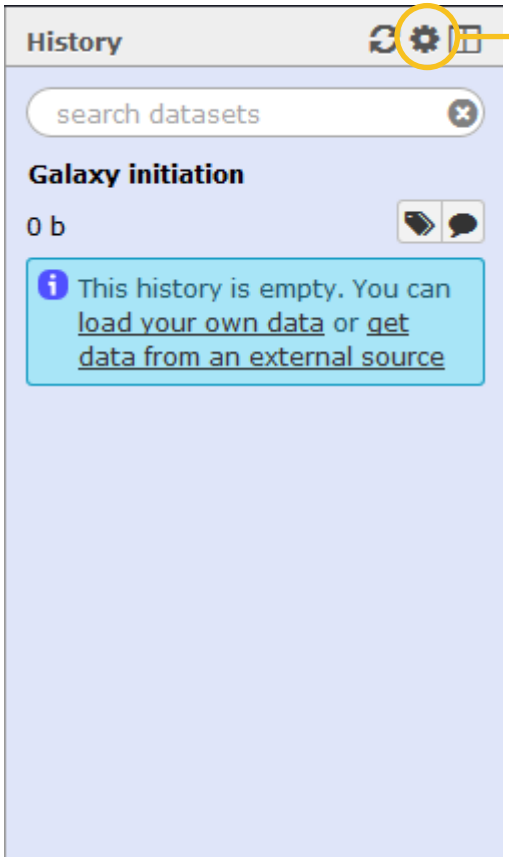

Administration Menu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 98.8 GB
Tools
search tools
Tool Menu
Get Data
Jobs statistics
FILES MANIPULATION
Text Manipulation
Filter and Sort
Join, Subtract and Group
GFF
BED Tools
Convert Formats
Fetch Sequences
Statistics
Graph/Display Data
SEQUENCES MANIPULATION
FASTA manipulation
FastQC: fastq/sam/bam
Illumina fastq
Generic FASTQ manipulation
FASTX-Toolkit for FASTQ data
SAM/BAM MANIPULATION : PICARD (BETA)
Conversion
QC/Metrics for sam/bam
BAM/SAM Cleaning
SAM/BAM manipulation: SAMtools
WELCOME TO GALAXY WORKBENCH
Galaxy is a workbench available for biologists from Sigenae Platform. Galaxy objectives are:
Hide the complexity of the infrastructure
Allow creation, execution and sharing of workflows.
ACCESS TO OLD GALAXY VERSION (2015) : ICI
Warnings :
All jobs running on galaxy are sent to BioInfo Genotoul cluster.
Your data are stored in work/ directory. Consequently, BioInfo Genotoul platform reserves the right to purge all files not accessed since 120 days on work/ disk space.
Contact your support : sigenae-support@listes.inra.fr
Results Visualization and Tool Parameter Space
History
search datasets
Unnamed history
0 b
This history is empty. You can load your own data or get data from an external source
Current History

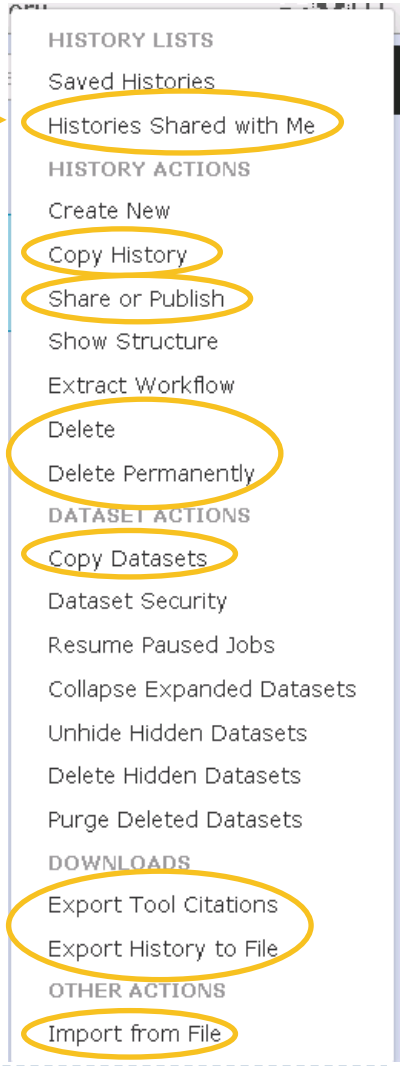
History


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 98.8 GB
Tools
search tools
MANAGE YOUR DATA FILES
Get Data
Jobs statistics
FILES MANIPULATION
Text Manipulation
Filter and Sort
Join, Subtract and Group
GFF
BED Tools
Convert Formats
Fetch Sequences
Statistics
Graph/Display Data
SEQUENCES MANIPULATION
FASTA manipulation
FastQC: fastq/sam/bam
Illumina fastq
Generic FASTQ manipulation
FASTX-Toolkit for FASTQ data
SAM/BAM MANIPULATION : PICARD (BETA)
Conversion
QC/Metrics for sam/bam
BAM/SAM Cleaning
SAM/BAM manipulation: SAMtools
WELCOME TO GALAXY WORKBENCH
Rename your history:
«Galaxy initiation »
Galaxy is a wo
objectives are
Make bioinfo Linux tools accessible to biogists.
Hide the complexity of the infrastructure.
Allow creation, execution and sharing of workflows.
ACCESS TO OLD GALAXY VERSION (2015) : ICI
Warnings :
All jobs running on galaxy are sent to BioInfo Genotoul cluster.
Your data are stored in work/ directory. Consequently, BioInfo Genotoul platform reserves the right to purge all files not accessed since 120 days on work/ disk space.
Contact your support : sigenae-support@listes.inra.fr
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

- Explore the history Menu

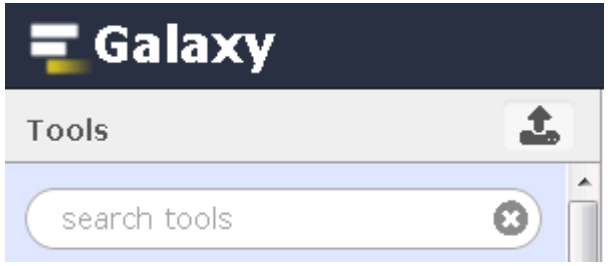




- You may:
  - **Publish** your History
    ➔ Share or Publish
  - **Access** to shared histories
    ➔ Histories Shared With Me
  - **Copy** ou **Export/Import**, or **delete**

Tools and Datasets

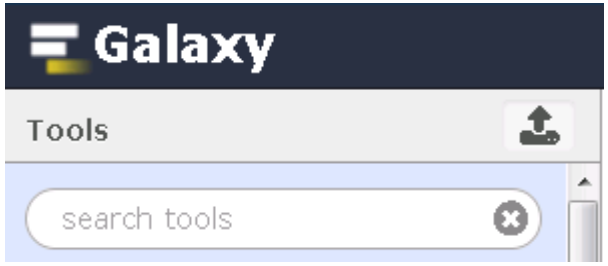
- How to find a tool in Galaxy ?



How much tools do you find for « FROGS » ?

Tools and Datasets

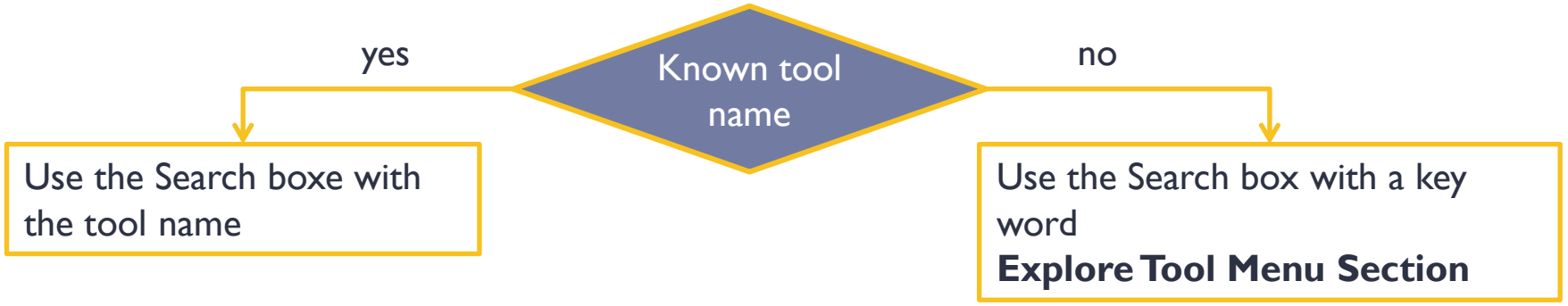
- How to find a tool in Galaxy ?



How much tools are linked with the fastq format ?

Tools and Datasets

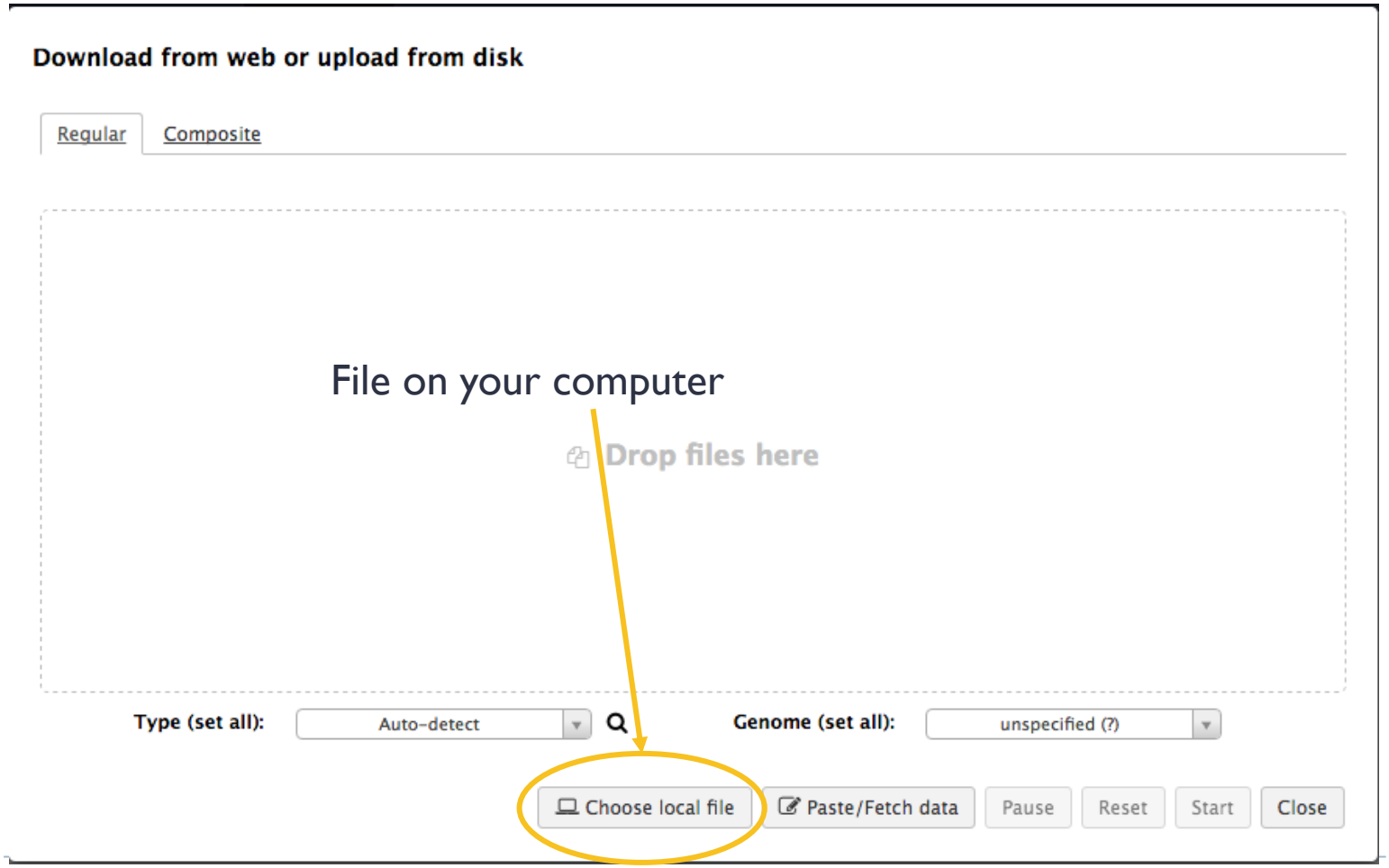
- How to find a tool in Galaxy ?



Which section may you use for uploading data ?

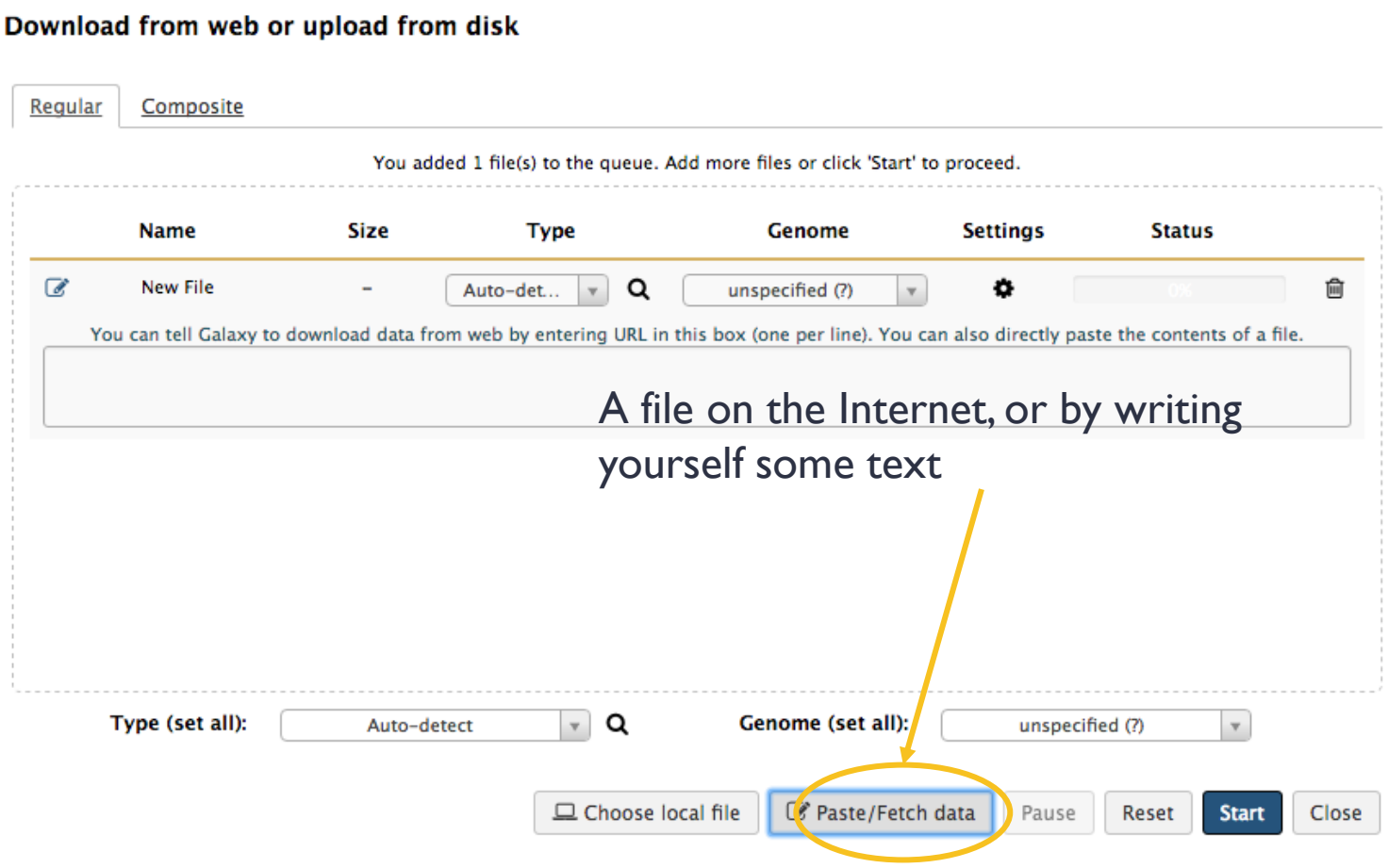
Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- For files **smaller than 2Go** use the tool: **Upload File from your computer**



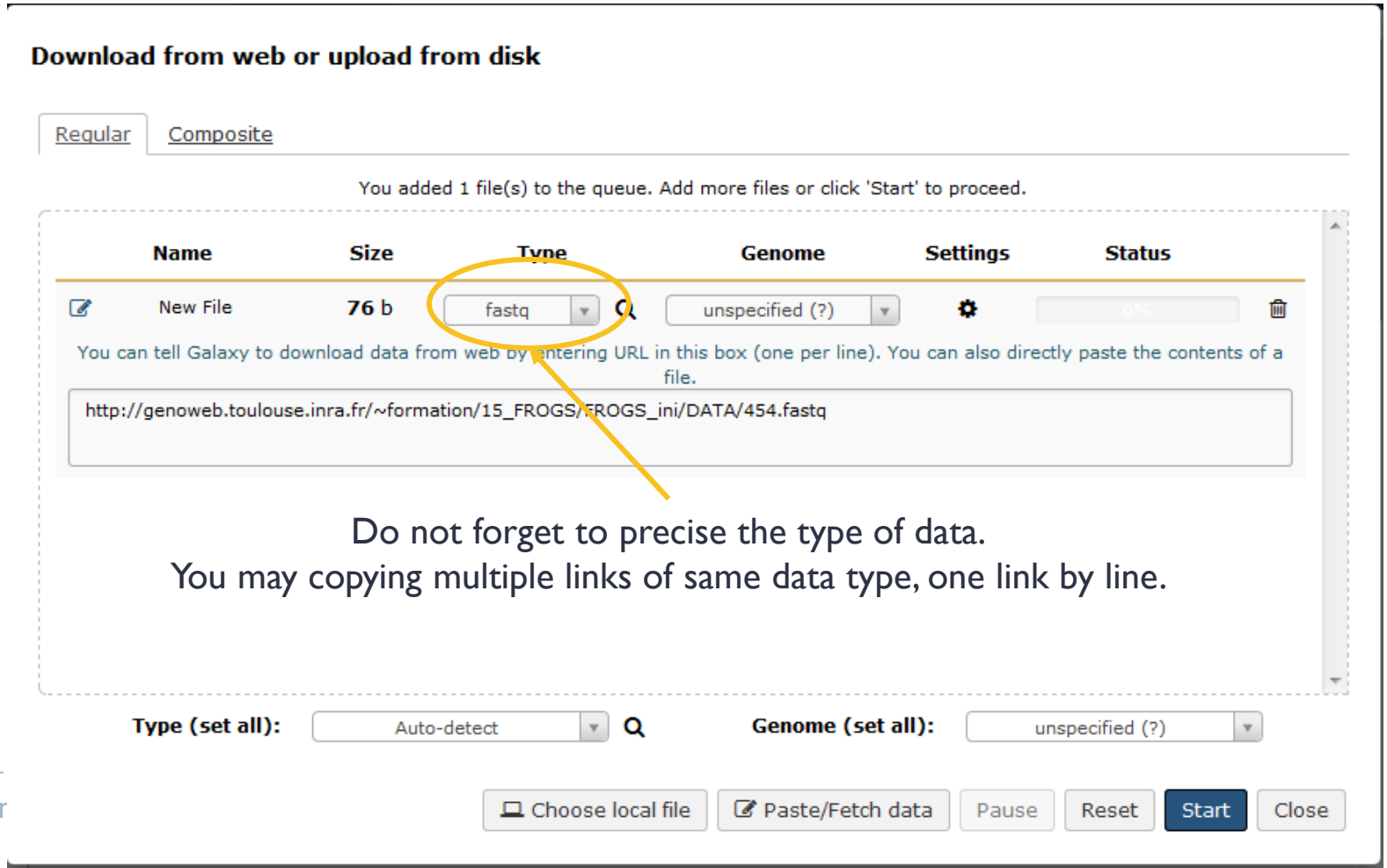
Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- For files **smaller than 2Go** use the tool: **Upload File from your computer**



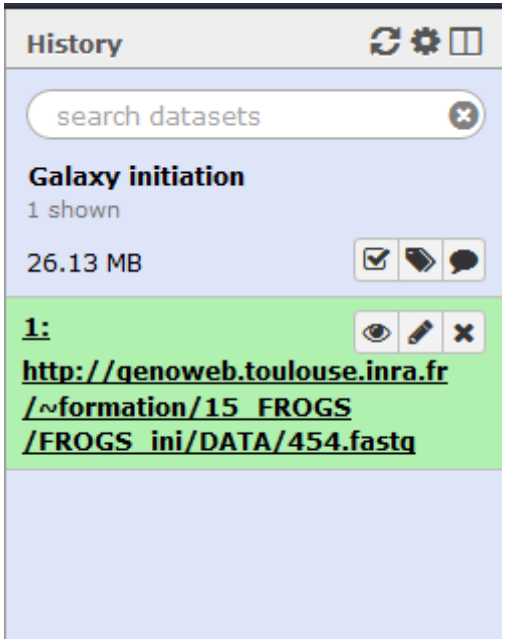
Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- **Go to :** http://genoweb.toulouse.inra.fr/~formation/15_FROGS/FROGS_ini/DATA/
- **Upload file 454.fastq by « copying link adress »**



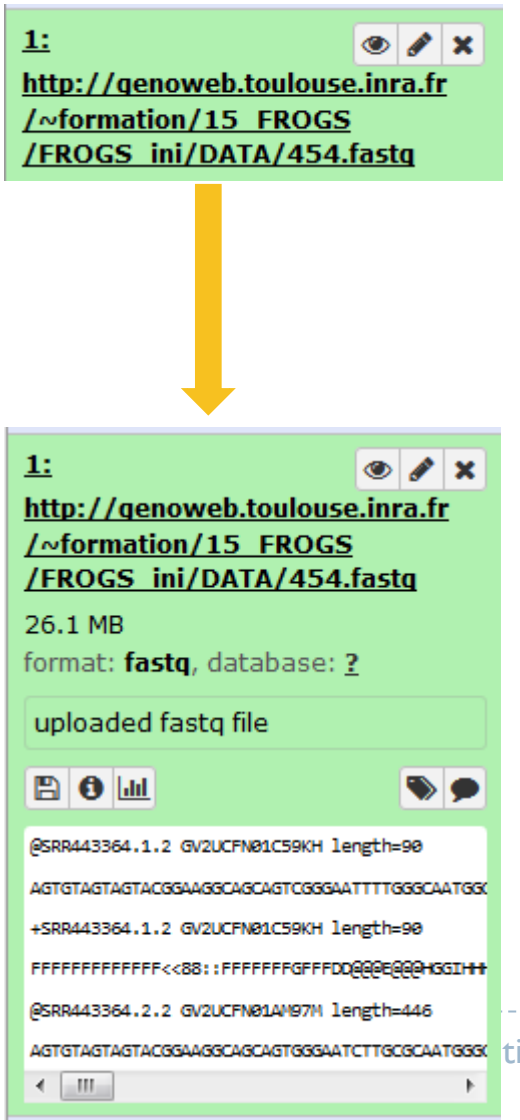
Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- Go to : http://genoweb.toulouse.inra.fr/~formation/15_FROGS/FROGS_ini/DATA/
- Upload file 454.fastq by « copying link adress »



Tools and Datasets

- **To what correponds each icon : « eye », « pen », « cross », « diskette »**



Visualize content in the main space

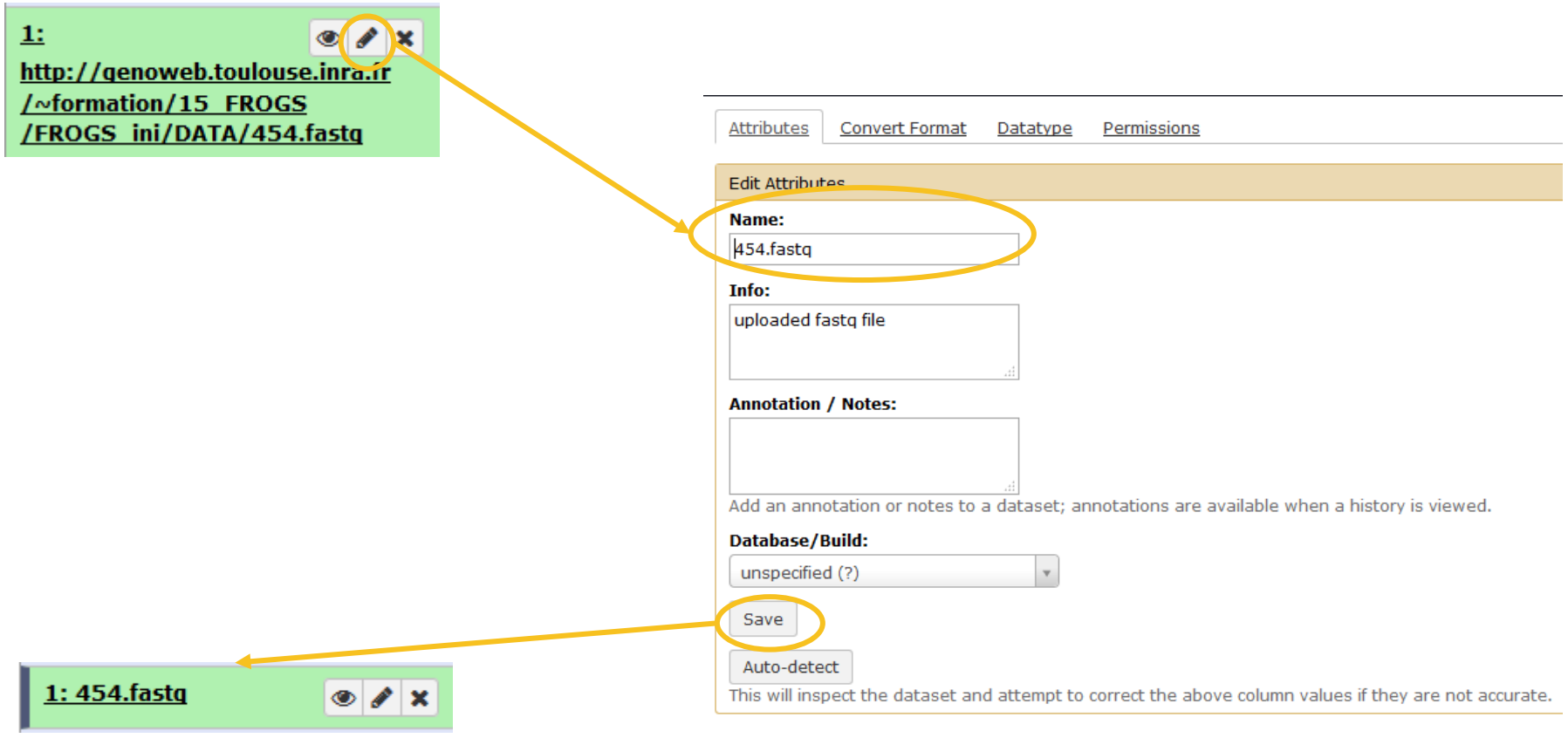
Delete dataset

Edit attributs : name, genome reference, type...

Download the dataset

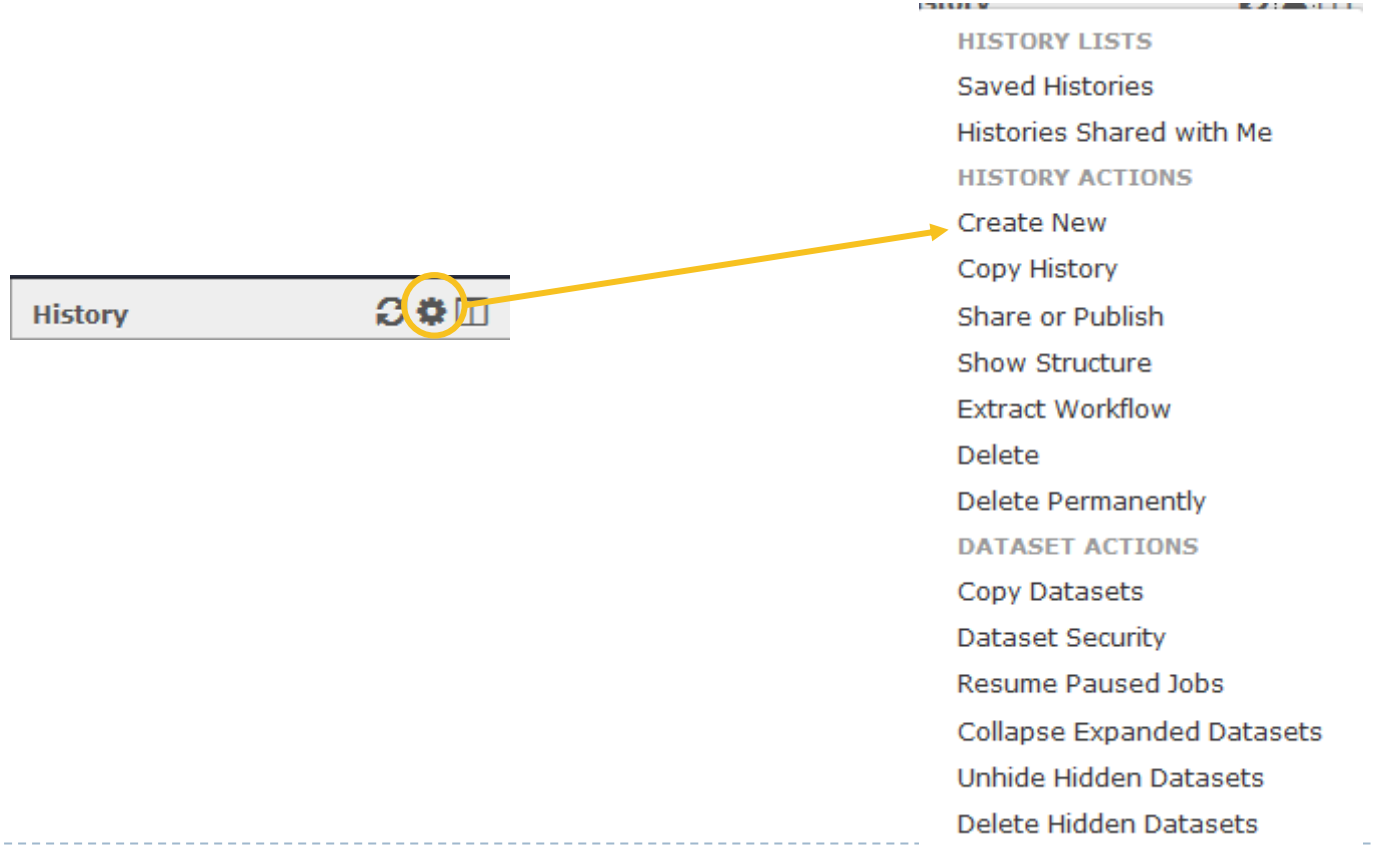
Tools and Datasets

- **Rename your dataset.**



Tools and Datasets

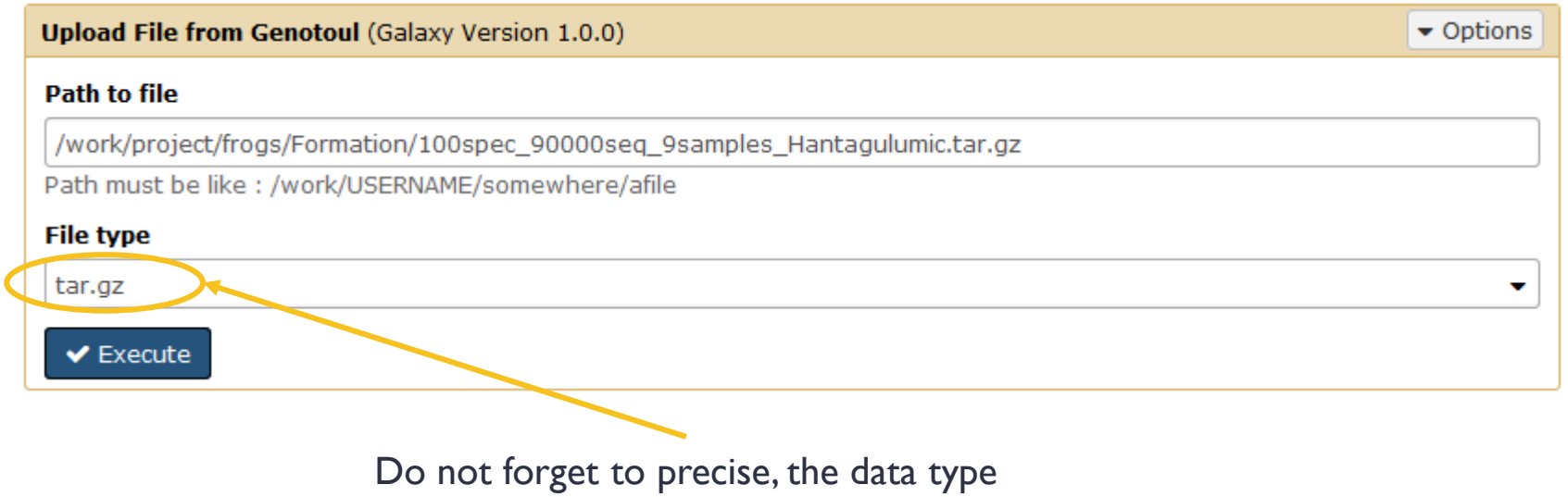
- How to upload Data in Galaxy ? ⇒ section **Get Data**
- For files **bigger than 2Go** use the tool: **Upload File from Genotoul**
- **Create a new history, named « Hantagulumic training »**



Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- For Files bigger than 2Go use the tool: **Upload File from Genotoul**
- Create a new history, name « Hantagulumic training »
- **Upload the tar.gz archive available on Genotoul thanks to « Upload File from Genotoul »:**

/work/project/frogs/Formation/100spec_90000seq_9samples_Hantagulumic.tar.gz



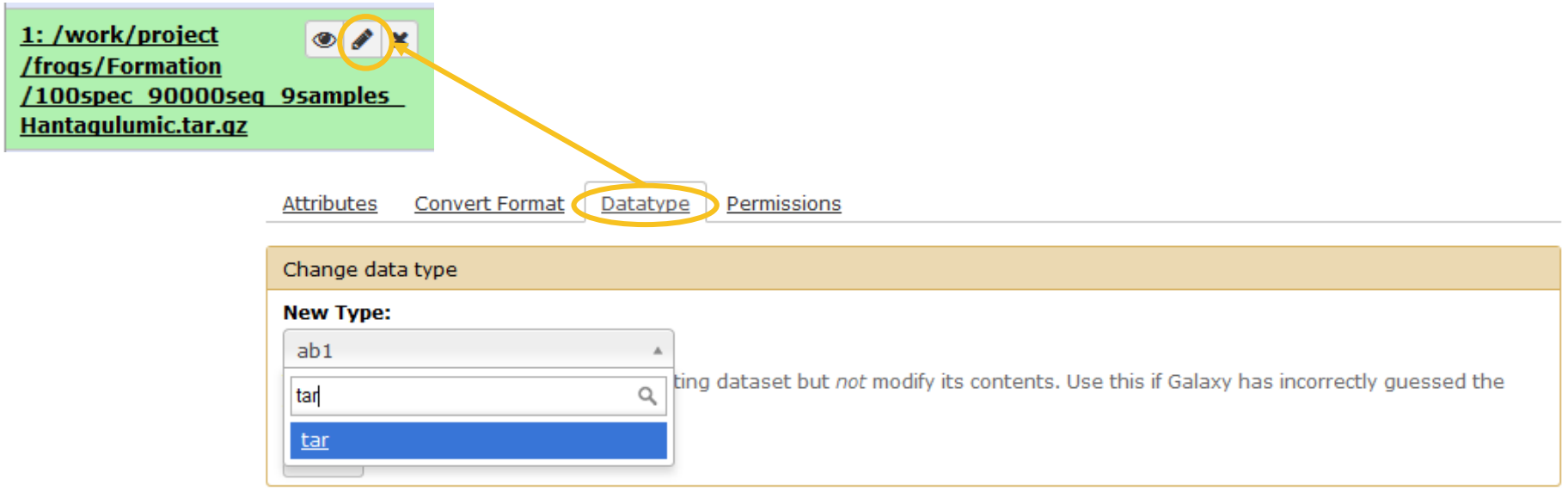
Do not forget to precise, the data type

Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- For Files bigger than 2Go use the tool: **Upload File from Genotoul**
- Create a new history, name « Hantagulumic training »
- Upload the tar.gz archive available on Genotoul:

/work/project/frogs/Formation/100spec_90000seq_9samples_Hantagulumic.tar.gz

- **May you visualize the content of this dataset?**
- **How to change data type, if you forget to change it during upload?**



Tools and Datasets

- How to upload Data in Galaxy ? ⇒ section **Get Data**
- For Files bigger than 2Go use the tool: **Upload File from Genotoul**
- Create a new history, name « Hantagulumic training »
- Upload the tar.gz archive available on Genotoul**:**

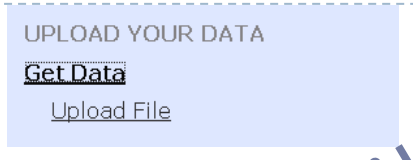
/work/project/frogs/Formation/100spec_90000seq_9samples_Hantagulumic.tar.gz

- **May you visualize the content of this dataset?**
- **How to change data type, if you forget to change it during upload?**

Datatype are important. Each tool will take as input a precise type of data.

- **Do not forget to rename dataset**

History, Tools and Dataset

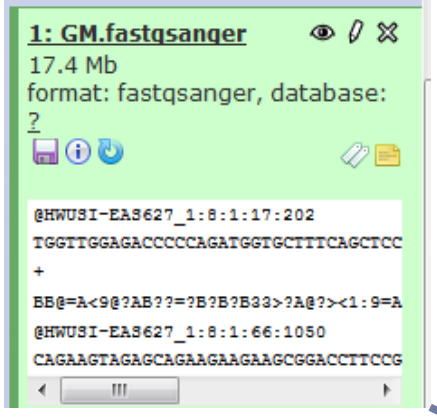


Left menu lists the **TOOLS**.

A **TOOL**

generates

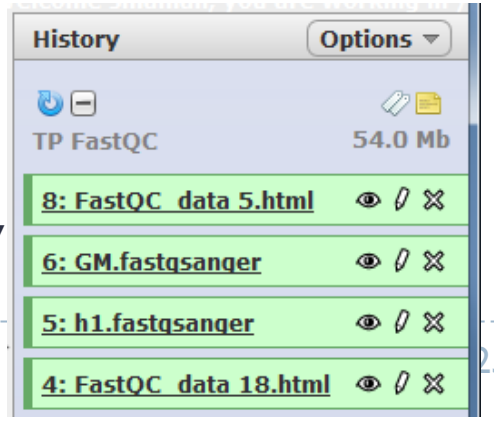
A/some **DATASET(S)**



Each green bloc, a **DATASET**, represents an input or output data file of a tool

Which together forms

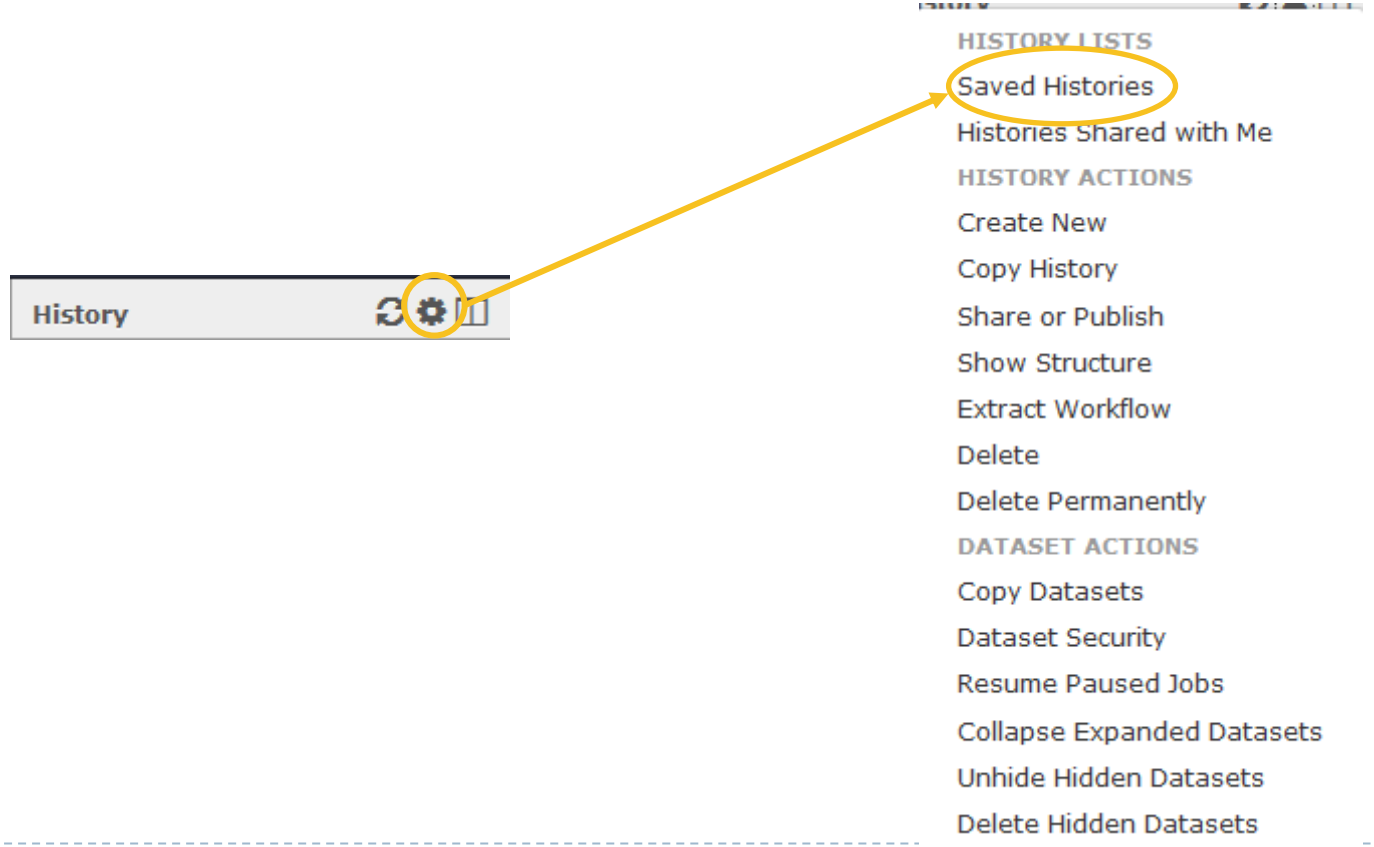
A **HISTORY**



All together forms a **HISTORY**, that finally represents a complete analysis

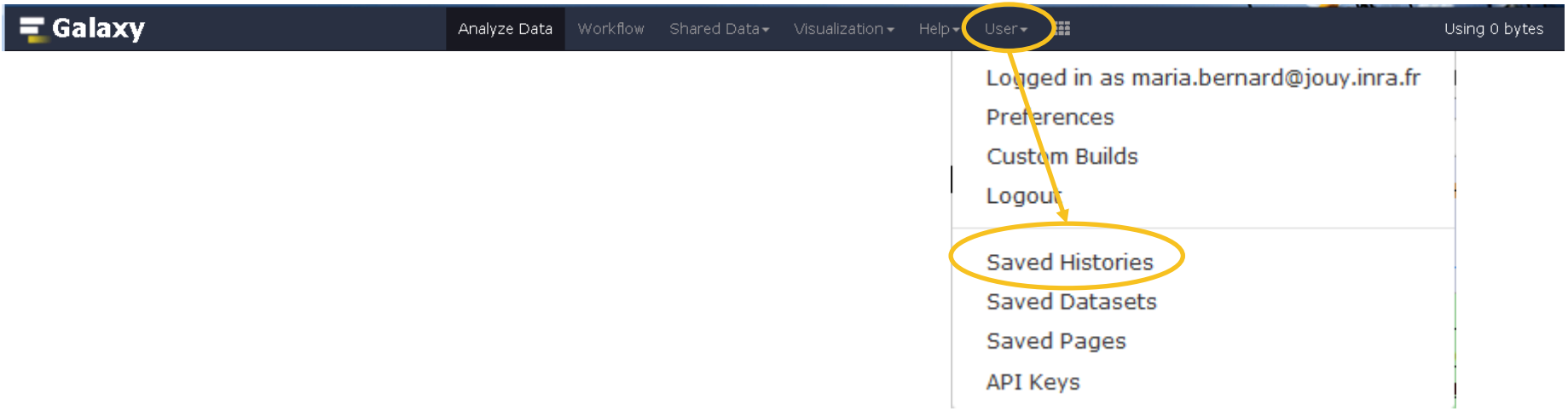
Back to History

- How to list all histories ?



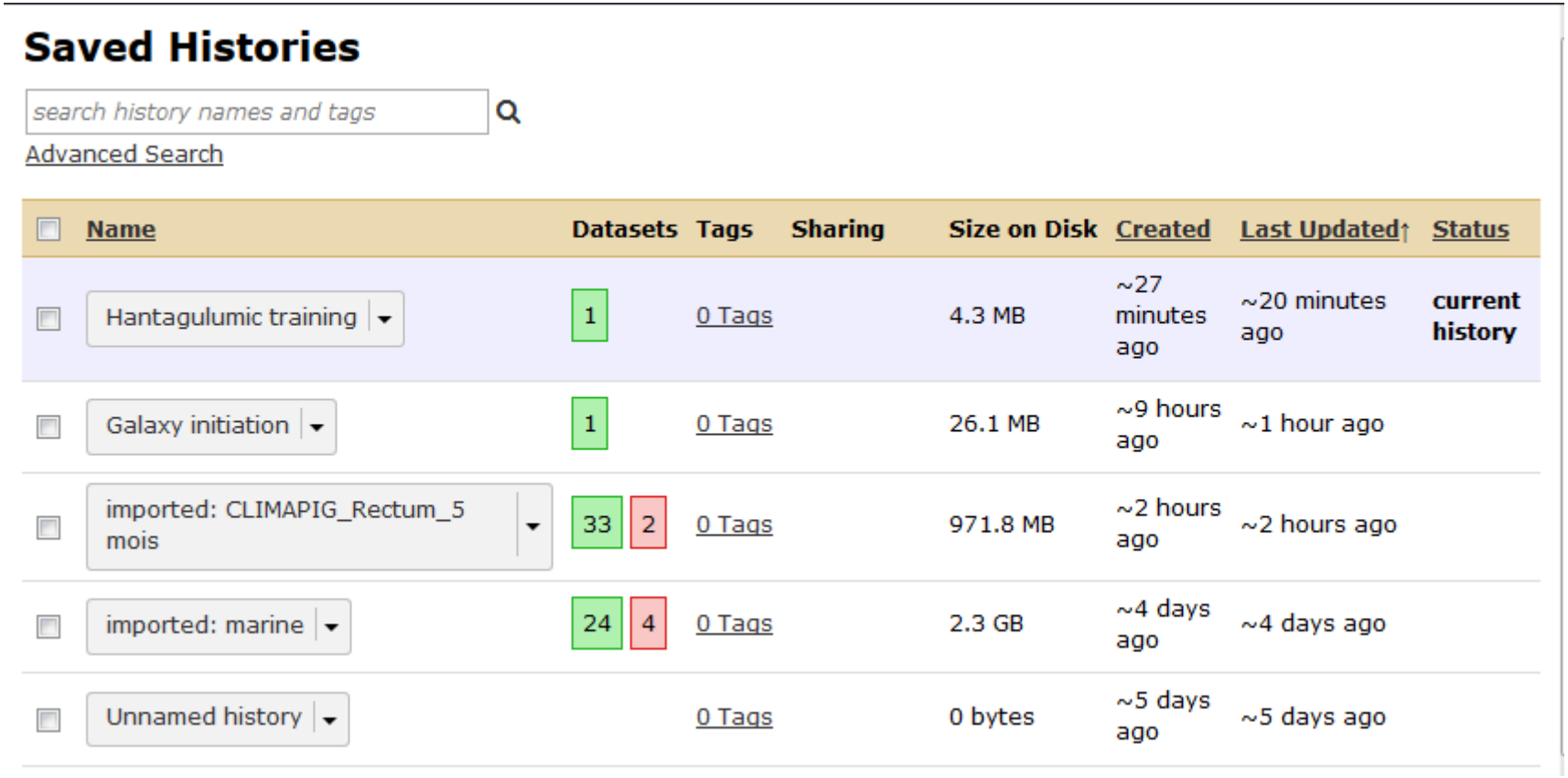
Back to History

- How to list all histories ?



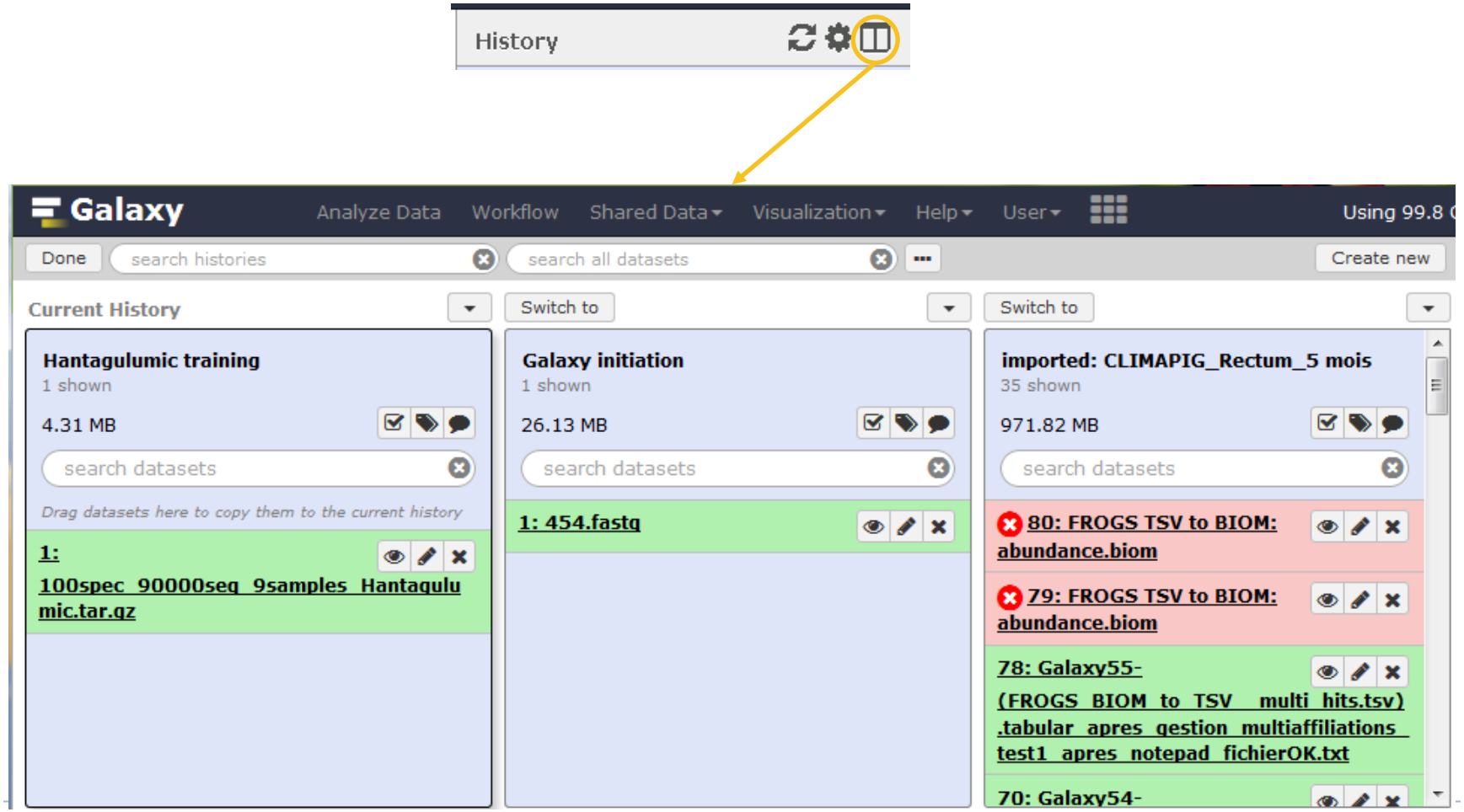
Back to History

- How to list all histories ?



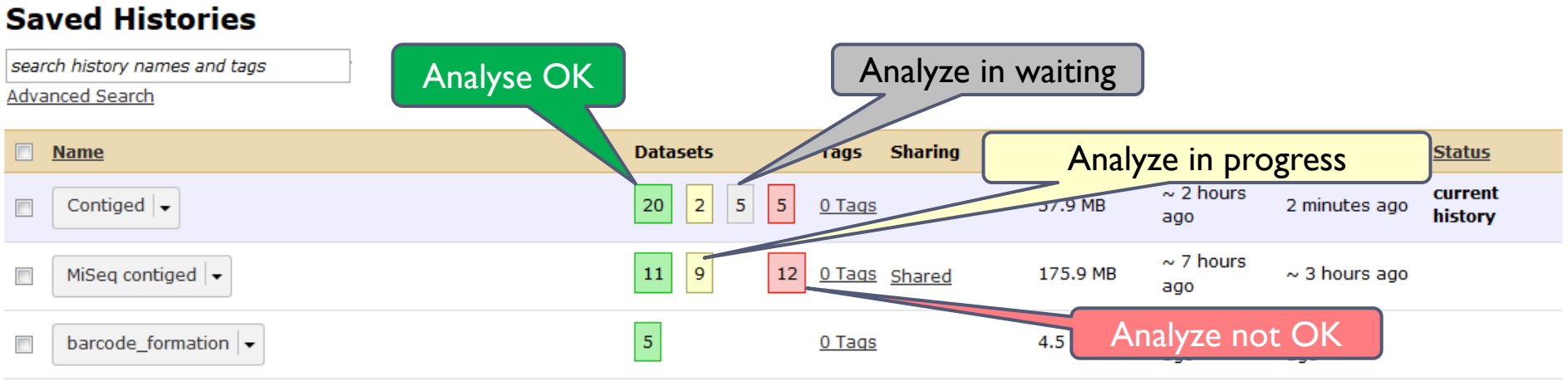
Back to History

- How to list all histories ?



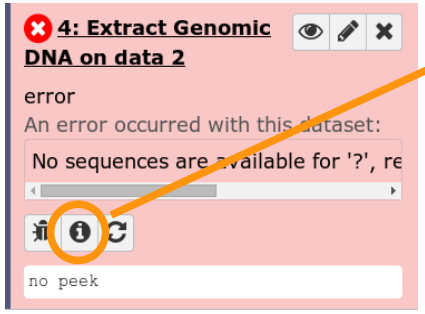
Back to History

- How to list all histories ?
- Meaning of color code:



How to debug?

- What to do when something goes wrong ?



Tool: Extract Genomic DNA

Name:	Extract Genomic DNA on data 2
Created:	Thu Sep 17 15:20:08 2015 (UTC)
Filesize:	0 bytes
Dbkey:	?
Format:	fasta
Galaxy Tool ID:	Extract genomic DNA 1
Galaxy Tool Version:	2.2.3
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	f09437b8822035f7
History ID:	5a1cff6882ddb5b2
UUID:	4bb33045-5f78-41f7-96e5-854c09711c06

Input Parameter	Value	Note for rerun
Fetch sequences for intervals in	2: mes_positions.txt	
Interpret features when possible	Yes	
Source for Genomic Data	cached	
Output data type	FASTA	

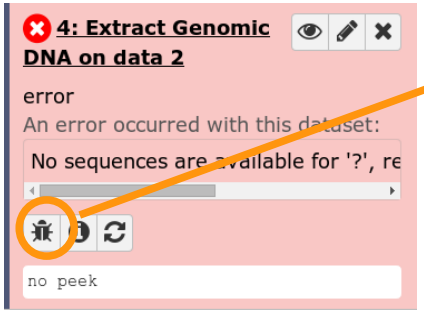
Inheritance Chain

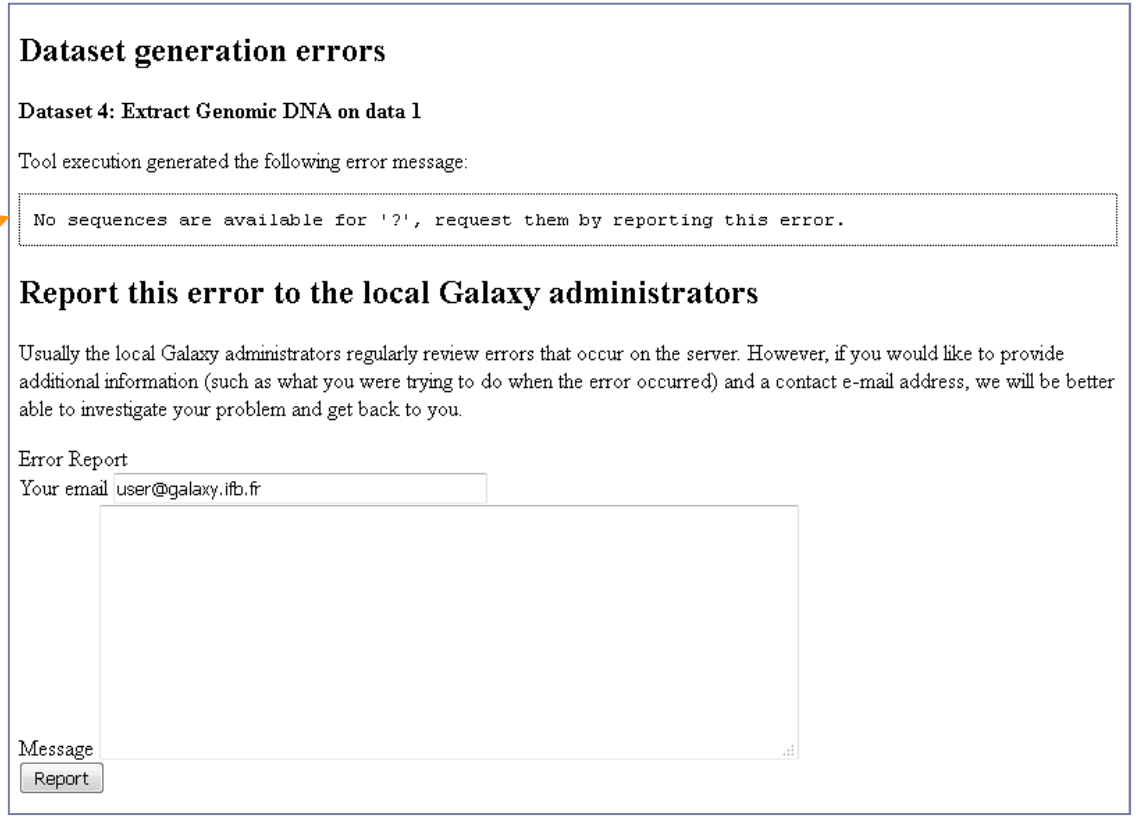
Extract Genomic DNA on data 2

```
No sequences are available for '?', request them by reporting this error.
```

How to debug?

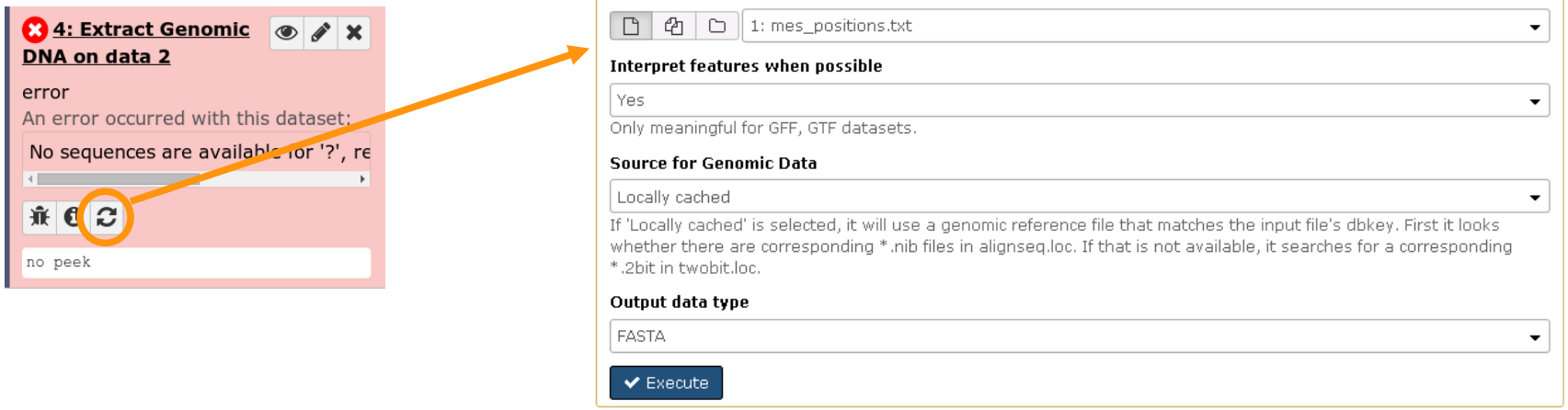
- What to do when something goes wrong ?





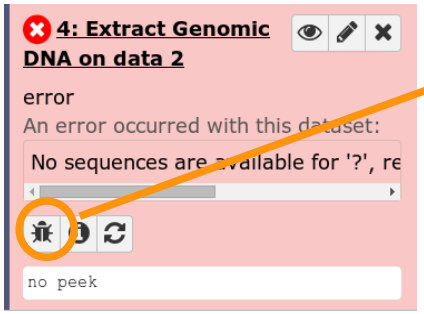
How to debug?

- What to do when something goes wrong ?
- Use « double arrow » to reload the tool with the previous parameters. Correct them and Execute



How to debug?

- What to do when something goes wrong ?
- Use « double arrow » to reload the tool with the previous parameters. Correct them and Execute
- What to do when the job still crash ?



Dataset generation errors

Dataset 4: Extract Genomic DNA on data 1

Tool execution generated the following error message:

```
No sequences are available for '?', request them by reporting this error.
```

Report this error to the local Galaxy administrators

Usually the local Galaxy administrators regularly review errors that occur on the server. However, if you would like to provide additional information (such as what you were trying to do when the error occurred) and a contact e-mail address, we will be better able to investigate your problem and get back to you.

Error Report

Your email user@galaxy.ifb.fr

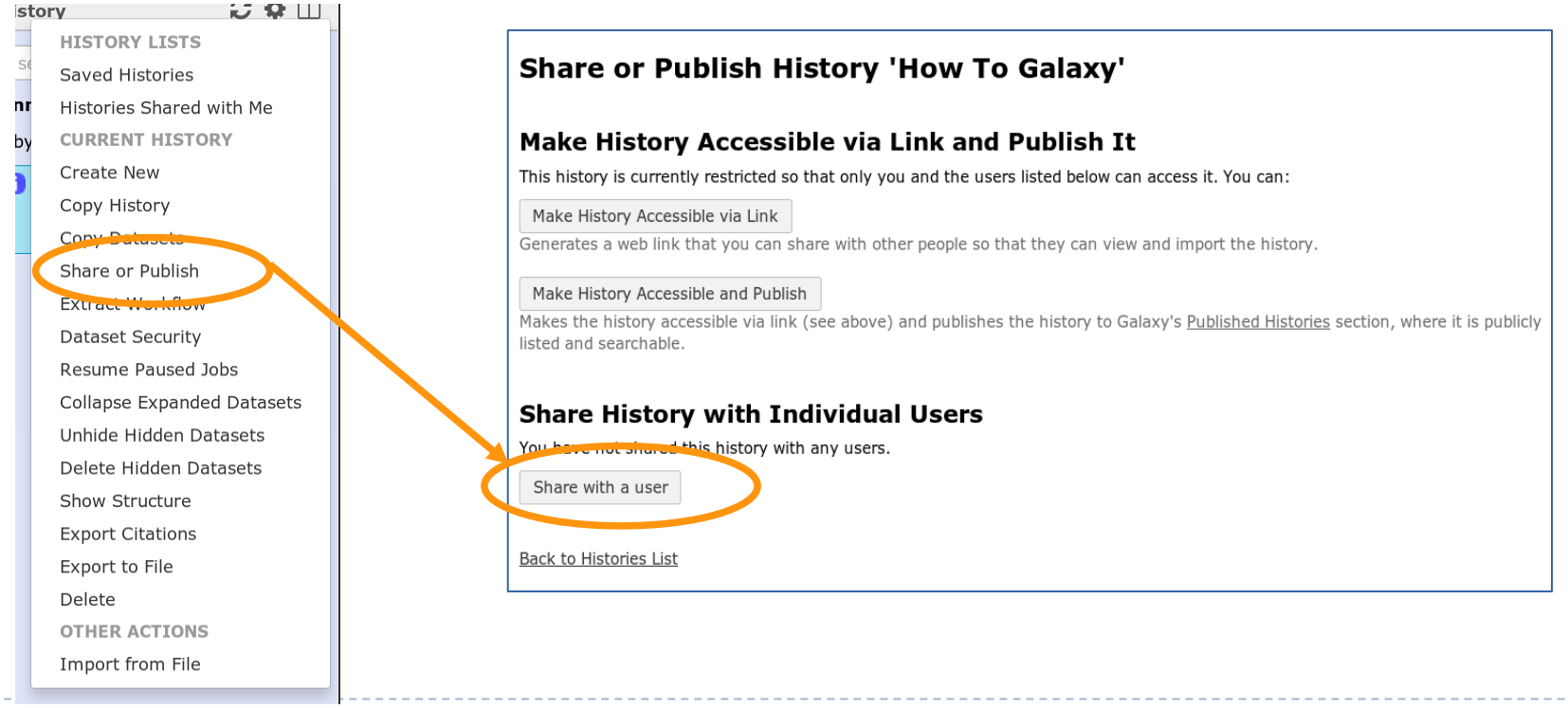
Send a message to the support. He will probably ask you to share your history!

Message

Report

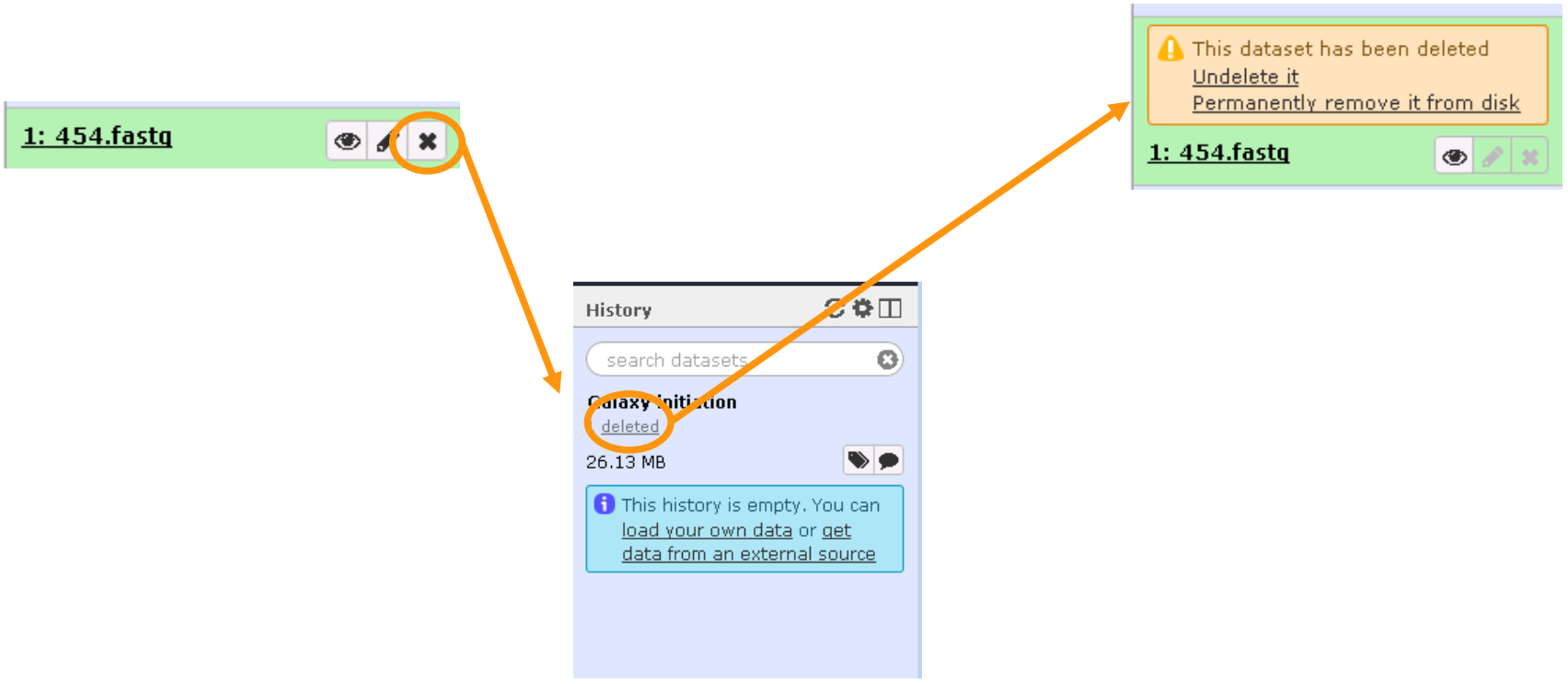
How to debug?

- What to do when something goes wrong ?
- Use « double arrow » to reload the tool with the previous parameters. Correct them and Execute
- What to do when the job still crashed ?
- How to share your history ?



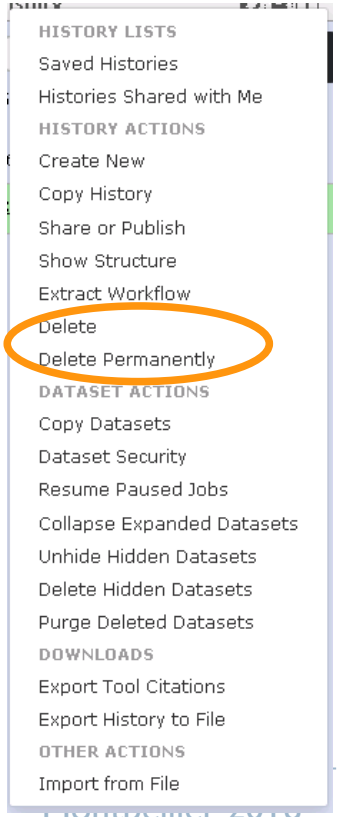
How to clean?

- Very often, there is a space disk quota on your Galaxy account. So you need to clean your histories.
- How to delete dataset ?
  - **Go back to your « Galaxy initiation » history.**
  - **Delete your dataset.**

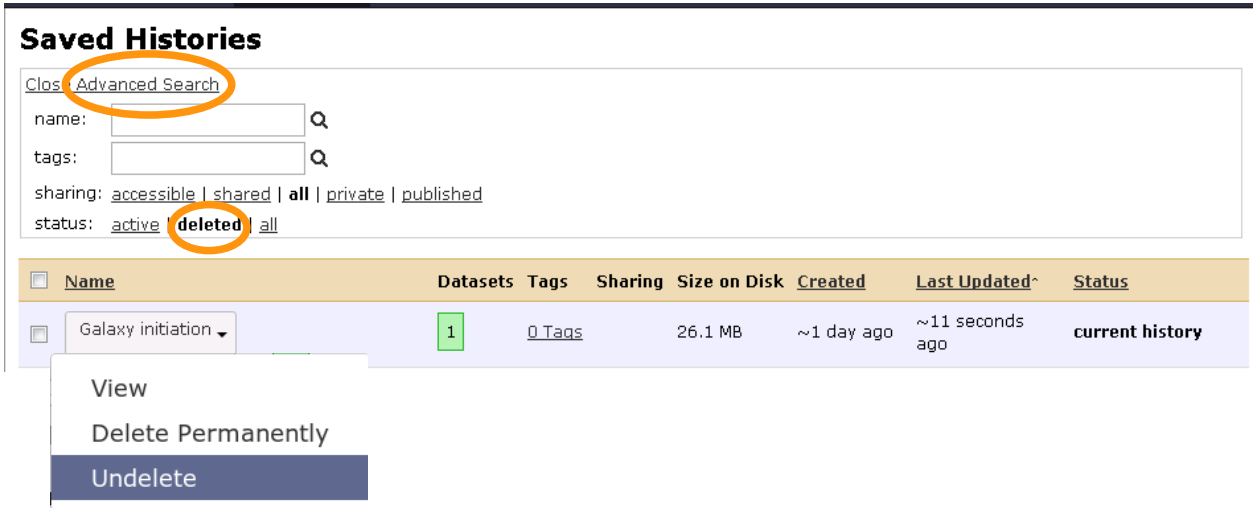


How to clean?

- Very often, there is a space disk quota on your Galaxy account. So you need to clean your histories.
- How to delete history ?
  - **Delete your history.**



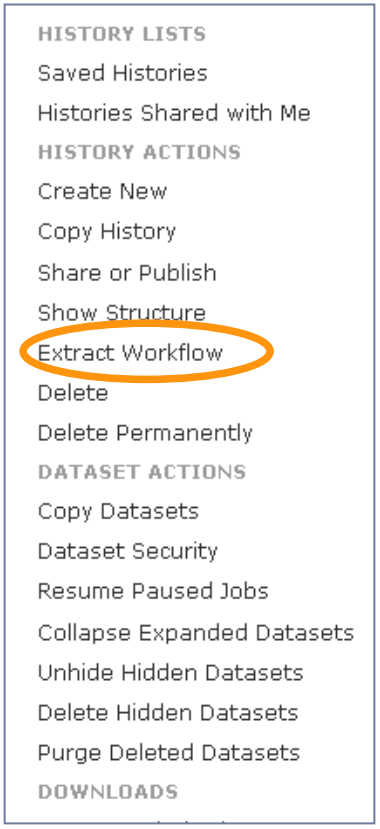
How to undelete history ?



Use « delete permenently » to save space

Workflow

- A workflow chains automatically all steps you want.
- How to create a workflow ?
- 1) Based on a **history** created manually



Workflow

- A workflow chains automatically all steps you want.
- How to create a workflow ?
- 2) Thanks to the **Workflow tab**:

Citations

- Cite used tools and their version (see the « info » icône and stdout output)
- Cite Galaxy publication:

Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Greg Von Kuster, Eric Rasche, Nicola Soranzo, Nitesh Turaga, James Taylor, Anton Nekrutenko, and Jeremy Goecks. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research (2016) doi: 10.1093/nar/gkw343

- Cite which instance you used.