Training on Galaxy: Metagenomics
June 2018

Find, Rapidly, OTUs with Galaxy Solution

FRÉDÉRIC Escudié* and LUCAS Auer*, MARIA Bernard, LAURENT Cauquil, SARAH Maman, MAHENDRA Mariadassou, SYLVIE Combes, GUILLERMINA Hernandez-Raquet, GÉRALDINE Pascal & MALO Le Boulch

*THESE AUTHORS HAVE CONTRIBUTED EQUALLY TO THE PRESENT WORK.

Introduction to Galaxy platform and preparation of FROGS training

June 2018

MALO Le Boulch, SARAH Maman, GÉRALDINE Pascal

Chitchat time!

- What is your computer skills level?

- Have you ever heard of or used Galaxy?

Feedback on metagenomics:

- What are your needs in "metagenomics"?

- 454 / MiSeq ?

- Your background ?

Monday
Galaxy · Coffee break · Galaxy · Lunch · Frogs · Coffee break · Frogs

Tuesday
Frogs · Coffee break · Frogs · Lunch · Frogs · Coffee break · Frogs

Wednesday
Frogs · Coffee break · Frogs · Lunch · Statistics · Coffee break · Statistics

Thrusday
Statistics · Coffee break · Statistics · Lunch · Statistics · Coffee break · Statistics
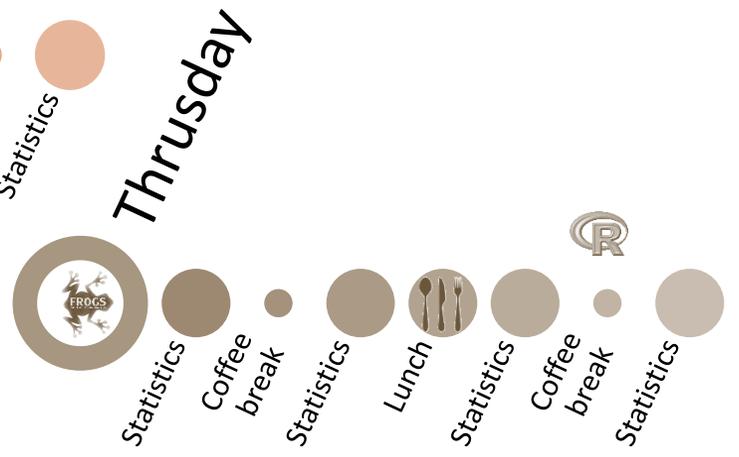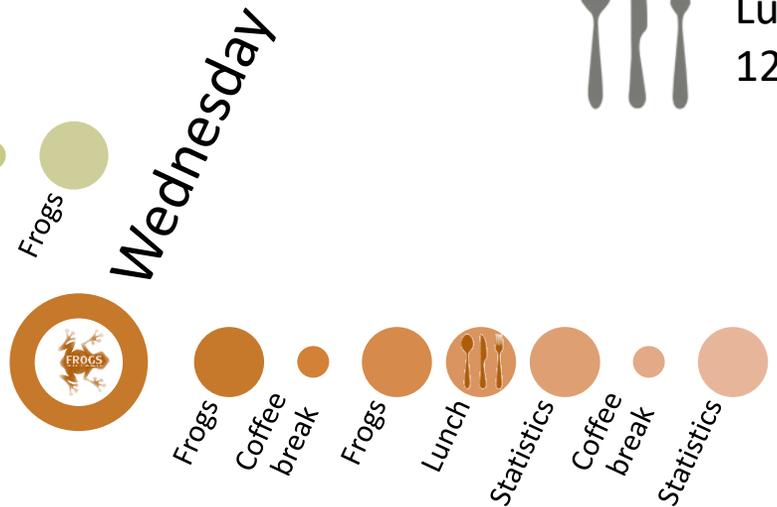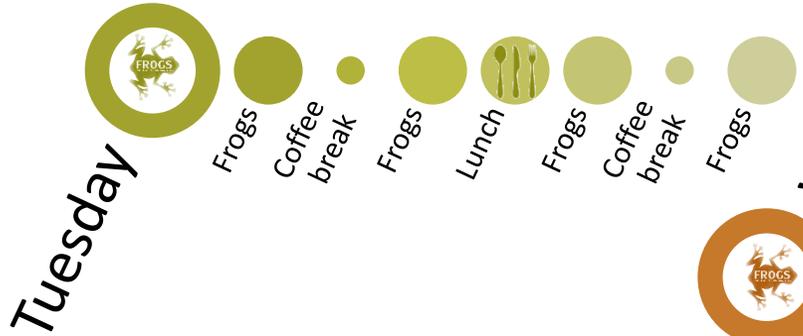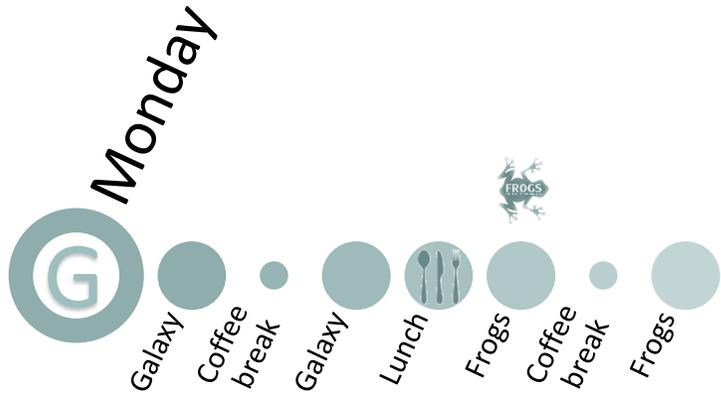
9 am to 5 pm

2 short coffee breaks morning and afternoon

Lunch
12.30 to 2.00 pm

Objectives

- Learn the basics of Galaxy

- Being independent when using it

- Prepare the datasets for FROGS formation

What is Galaxy?

Galaxy software framework

- Galaxy is an overlay that allows researchers without computer expertise to perform computational analyses online, automate them and share the results easily.

- Developed by the Galaxy Team Project.

- Relies on a computational infrastructure (Server).

- Open source

Galaxy software framework

User

Web browser

Servers

You

What you see

What you don't see

https://galaxyproject.org/tutorials/g101/
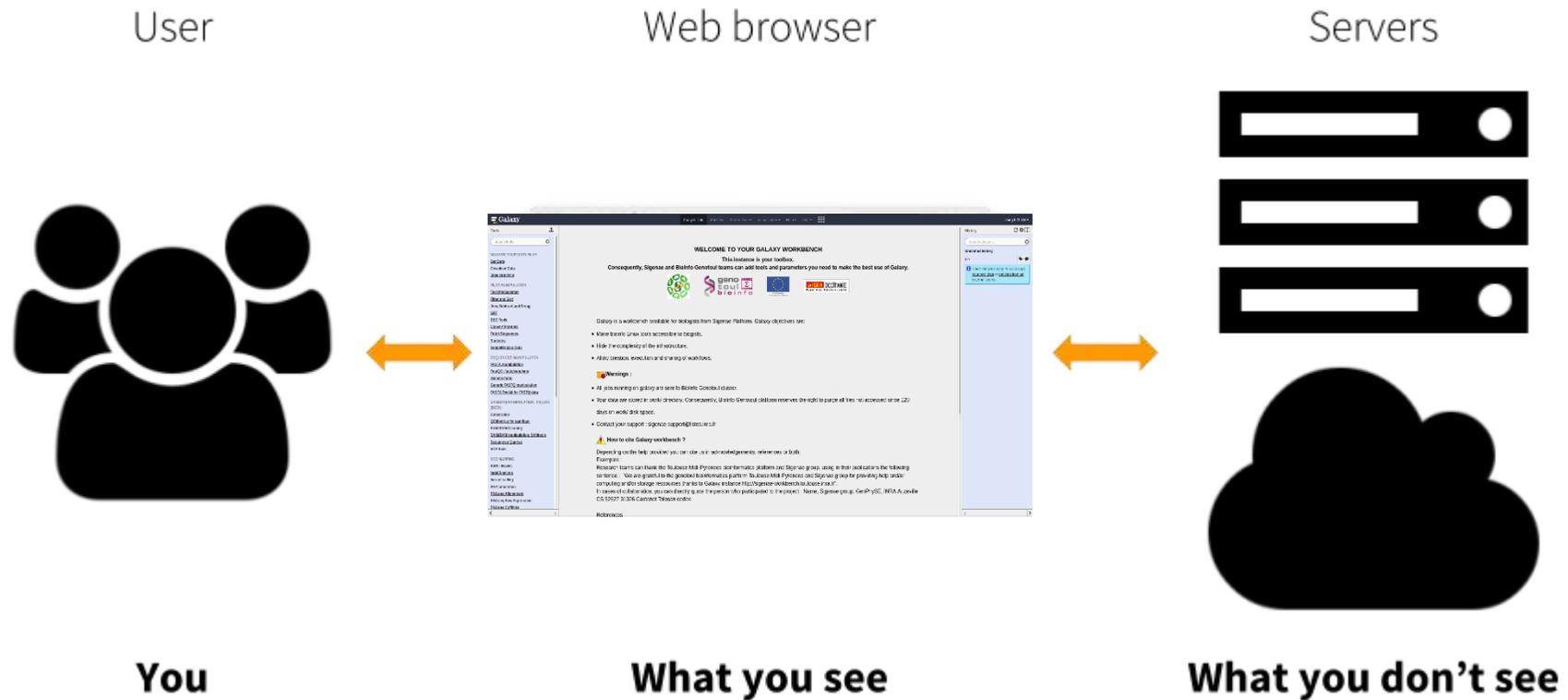
Where to use Galaxy?

- Galaxy software must be installed and run on powerful server farms (Cluster).

- Genotoul Bioinfo

  - Cluster: Many computers tightly connected that work together

  - High performance computer:

    - More than 5000 cores

    - 34 TB of RAM

    - More than 1 Peta Byte (1024 TB) of hard drive

Where to use Galaxy?

- Our Galaxy platform is http://sigenae-workbench.toulouse.inra.fr/galaxy/

⚠ The different platforms of Galaxy are not connected together

- Galaxy is installed on many clusters across the world.

- Some tools are in <u>our</u> Galaxy platform but not in <u>other</u> platforms.

- Your data is not shared with other Galaxy platforms than ours.

Exemple of 2 INRA Galaxy platforms

Exemple of 2 INRA Galaxy platforms

Your Turn!

CONNECT TO OUR GALAXY WORKBENCH

Exercise

During this formation, we will use a remote computer.

Connect to a distant computer by clicking on the remote desktop connection icon in the task bar.

Ask trainers for the address and password.

All your data will be erased at the end of the week of the remote computer

Exercise

During this formation if you have a personal account, use it.

This account will save all the work done during this week and your future work.

If you don't have a personal account ask us for a temporary account.

You can request a personal account later on : http://bioinfo.genotoul.fr/index.php/ask-for/create-an-account/
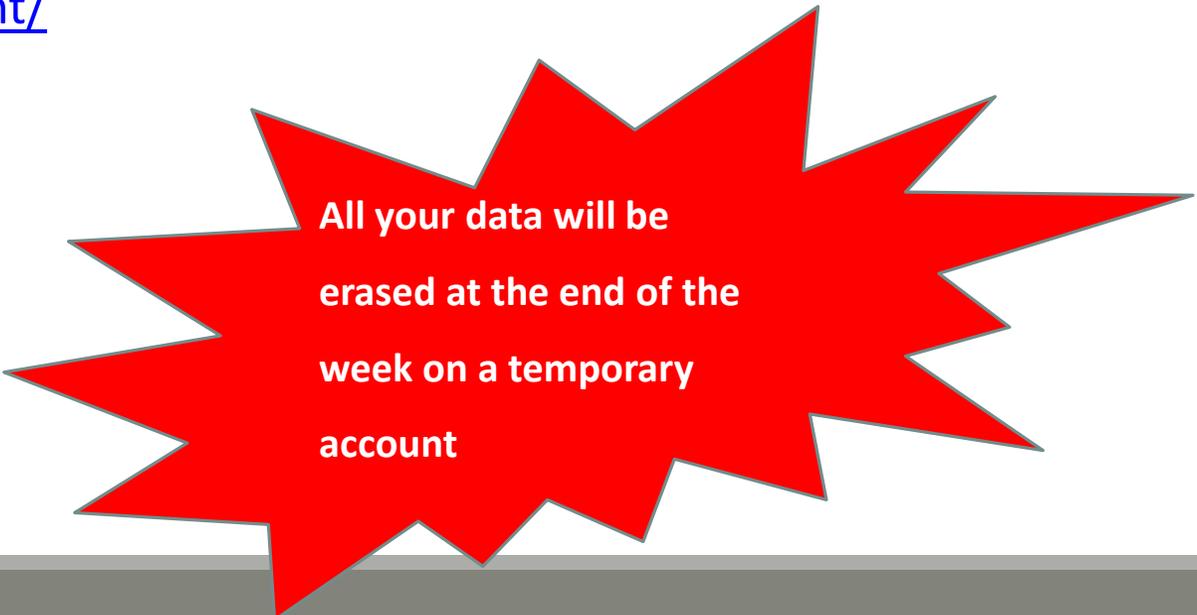
All your data will be erased at the end of the week on a temporary account

Exercise

Our Galaxy platform is: http://sigenae-workbench.toulouse.inra.fr/galaxy/

Be careful, to fully login you must enter your credentials twice:

- The first time in this pop-up window:
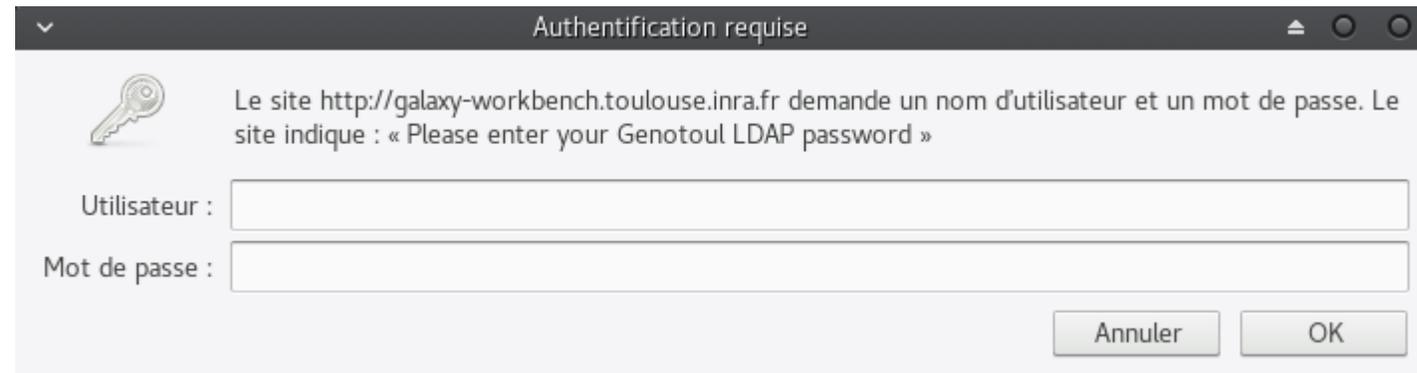
Exercise

- And a second time, in the browser:
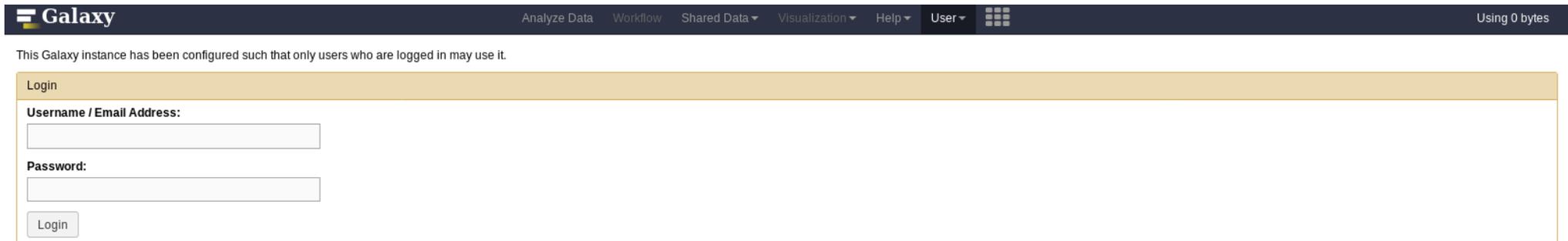
Change your password

- In order to change your password, you need to download PuTTY: http://www.putty.org/

- PuTTY is a terminal emulator, it allows to connect directly to the server in command line.

- You can not change your password via the Galaxy's interface for the moment.

Change your password

- Launch PuTTY.

- In the following window, you must enter your host name which is:

  YourGenotoulName@genotoul.toulouse.inra.fr

- The connection type must be SSH.

- Click on « Open » .
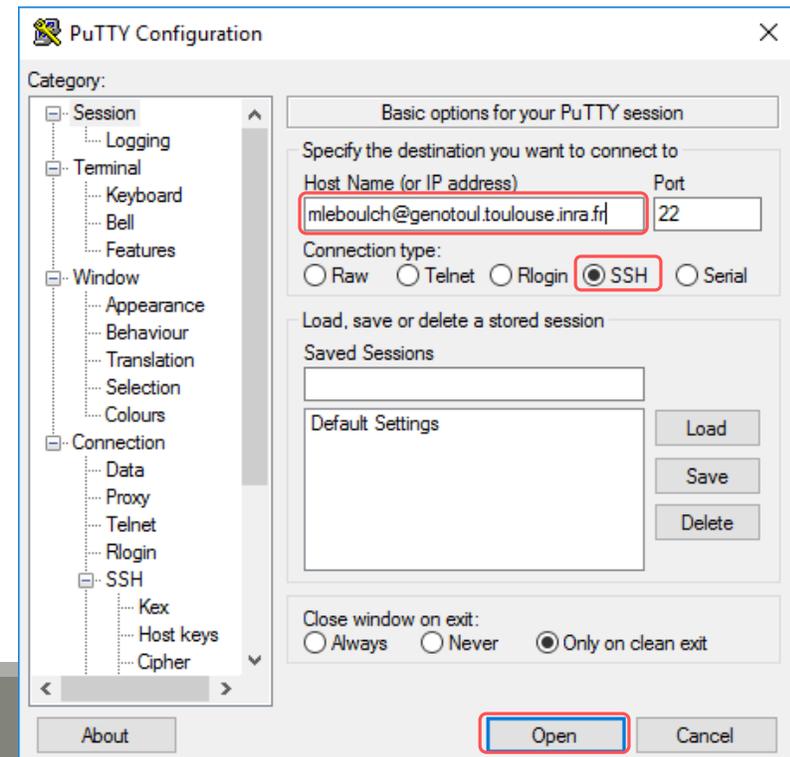
Change your password

- A new window appear, click on « Yes ».

- In the following window, enter your password and hit « Enter ».
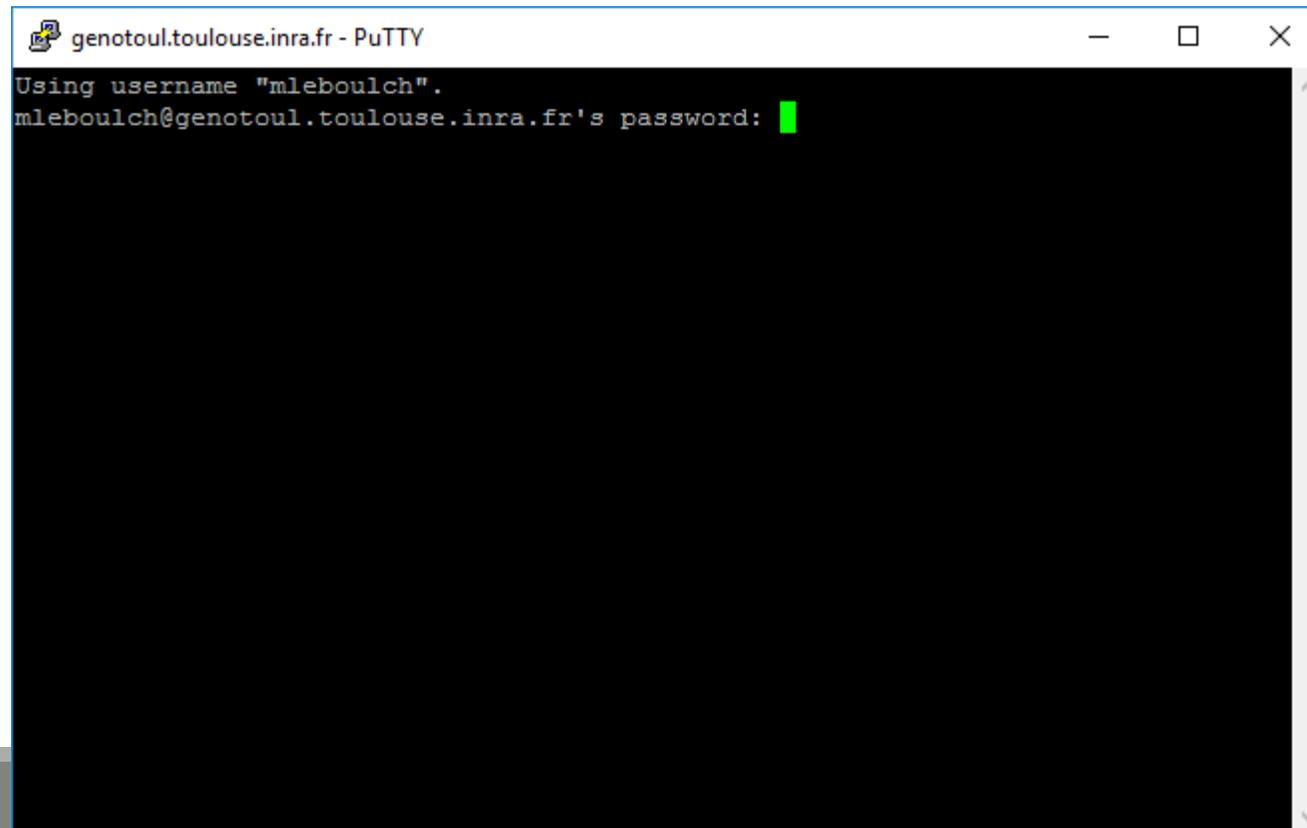
Change your password

- Type « passwd » and hit « Enter ».

Change your password

- Enter your current password and hit « Enter ».

Change your password

- Enter your new password (with an upper case, a number and a special character in it) and hit « Enter ».

Change your password

- Enter a second time your new password and hit « Enter ». You have changed your Genotoul password, you can close PuTTY.

Galaxy

Analyze Data | Workflow | Shared Data ▾ | Visualization ▾ | Help ▾ | User ▾

Using 0%

Tools

search tools

MANAGE YOUR DATA FILES
Get Data
Download Data
Jobs statistics

FILES MANIPULATION
Text Manipulation
Filter and Sort
Join, Subtract and Group
GFF
BED Tools
Convert Formats
Fetch Sequences
Statistics
Graph/Display

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
Sequences Queries
VCF Tools

SGS MAPPING
BWA - Bowtie
Indel Analysis
Variant calling
SNP annotation
RNAseq Alignement
RNAseq Raw Expression
RNAseq Cufflinks

AVAILABLE TOOLS

History

search datasets

Data Collection
3 shown

33 b

3: seq1.fasta

2: seq2.fasta

1: seq3.fasta

DATASETS HISTORY

**RESULTS VISUALISATION
AND
TOOL PARAMETER WINDOW**

Genotoul Bioinfo

LA RÉGION OCCITANIE
Pyrénées-Méditerranée

Fonds Europeen
de Développement Régional

How to cite Galaxy workbench ?

Depending on the help provided you can cite us in acknowledgements, references or both.

Examples :
Research teams can thank the Toulouse Midi-Pyrenees bioinformatics platform and Sigenae group, using in their publications the following sentence : "We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees and Sigenae group for providing help and/or computing and/or storage ressources thanks to Galaxy instance http://sigenae-workbench.toulouse.inra.fr".

In cases of collaboration, you can directly quote the person who participated to the project : Name, Sigenae group, GenPhySE, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex.

References
X. SIGENAE [http://www.sigenae.org/]

Ready to use Workflows

- 1/ FROGS - Find, Rapidly, Otus with Galaxy Solution
- 2/ Tax4Fun workflow
- 3/ Function Table for Tax4Fun matrix workflow

Galaxy News

Outlook Web App

⊕ Nouveau message

- 1/ How to use datasets collections ?
- 2/ How to upload several files in Galaxy ?
- 3/ New tools : Tax4Fun (manual), VIP (manual), Salmon, multiQC.
- 4/ Tools availables for virologie analyses
- 5/ Virology tools
- 6/ How to use Function Table tool ?

E-learning

formation permanente
Classe virtuelle

INRA

https://inra.classilio.com/Login
2 e-learning availables : Galaxy initiation and Quality sequences analysis

Vocabulary of Galaxy

- **Tools:**

  - A tool has a function which is explained when you click on it.

  - Each Galaxy platform has its own tools.

- **Dataset:**

  - A dataset is a file, uploaded to Galaxy by you or produced by a tool.

  - Be careful: a dataset has a datatype.

- **History:**

  - A tool generates datasets and these datasets are stored in the current history.

  - Everything is permanently saved.

  - If you log off your computer or browser, it's ok, everything will keep running and be saved!

Your Turn!

DISCOVER GALAXY

Exercise

1. Visit the Galaxy Platform.

2. Look at the tool list.

3. Display only FROGS tools.

4. Display all tools concerning fastq files.

Exercise

Search a tool by name.
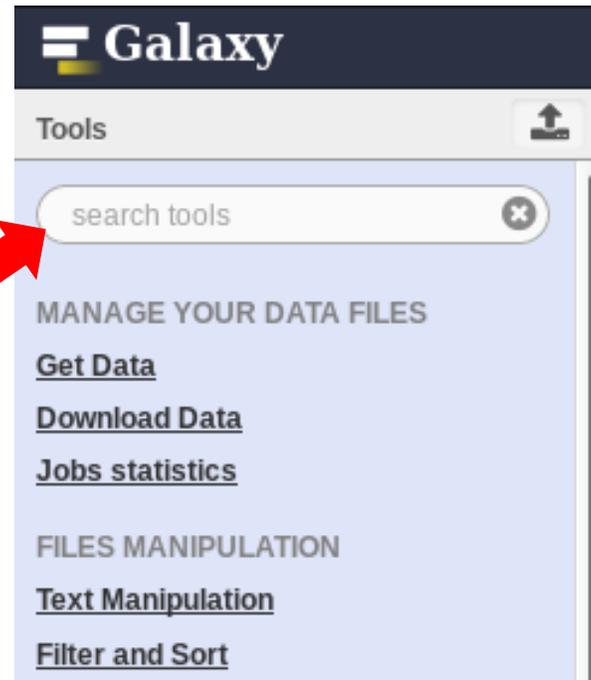
Manipulate Histories

Your Turn!

CREATE THE 4 HISTORIES NEEDED FOR THE FROGS FORMATION

To create a new history:

- Click on the wheel.

- Click on « create new ».
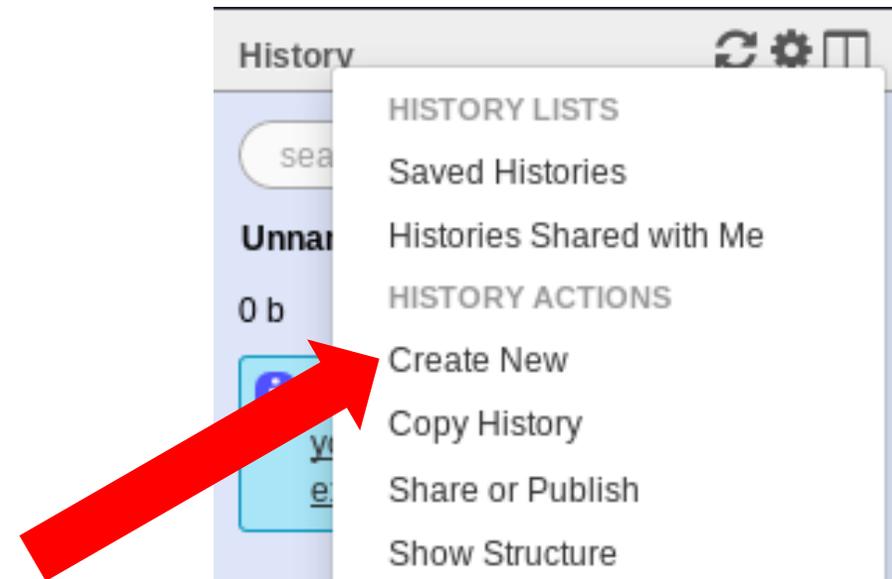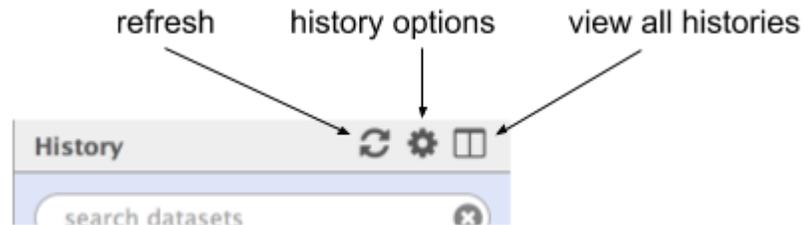
To rename a history:

- Click on the history name (at the top).

- Enter « multiplex ».

- Hit « Enter » to validate.
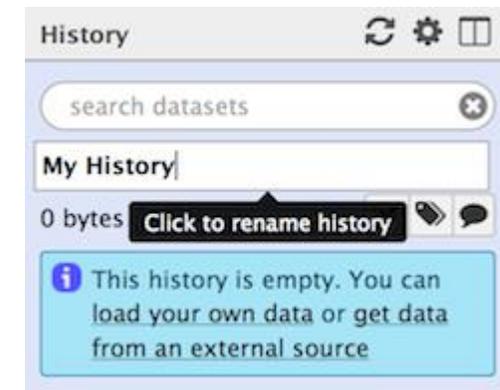
Don't use special characters or accents!

Exercise

- Create histories named:

  - 454

  - merged

  - temp

- Switch to the history named « multiplex » as current history.

- Go back to the main interface.

How to list all histories?

- To view all histories, click on this icon.
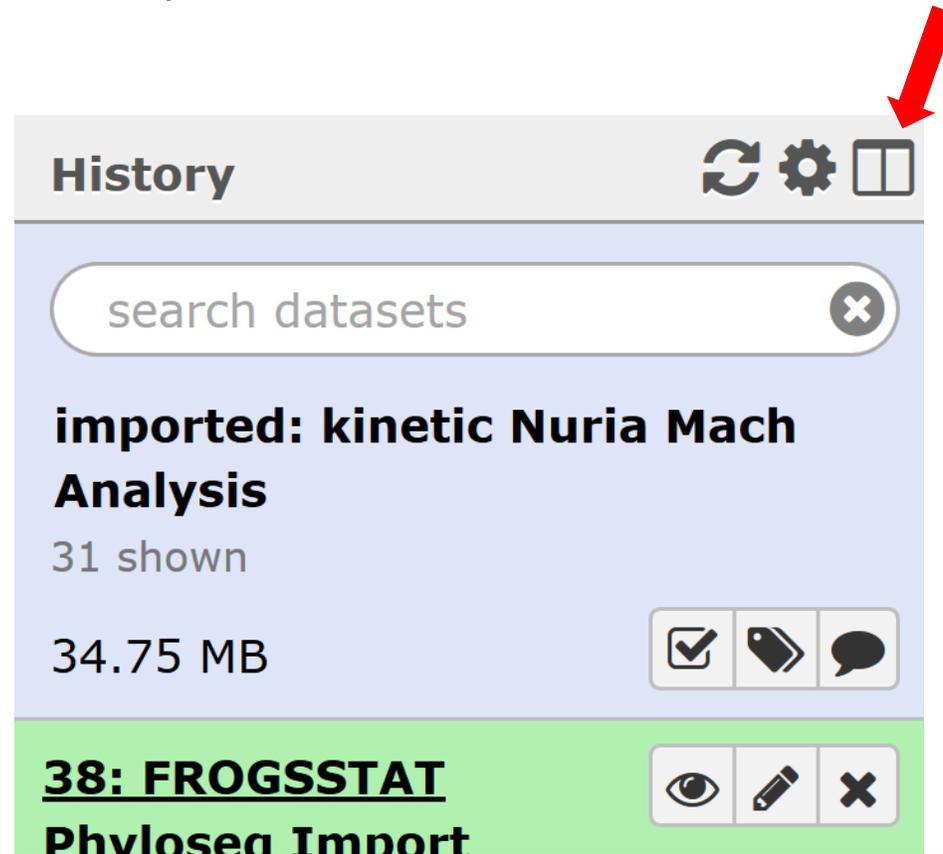
Explore the « View all histories » section

Switch current history

- Switch to the history named « multiplex » as current history.

- Click on "Done" to go back to the main interface.

Data import

How to import your data to Galaxy

- 5 ways to upload your data to Galaxy:

    - From your computer

    - By URL

    - From Genotoul Bioinfo clusters

    - Shared by other users of Galaxy

How to choose your upload method?
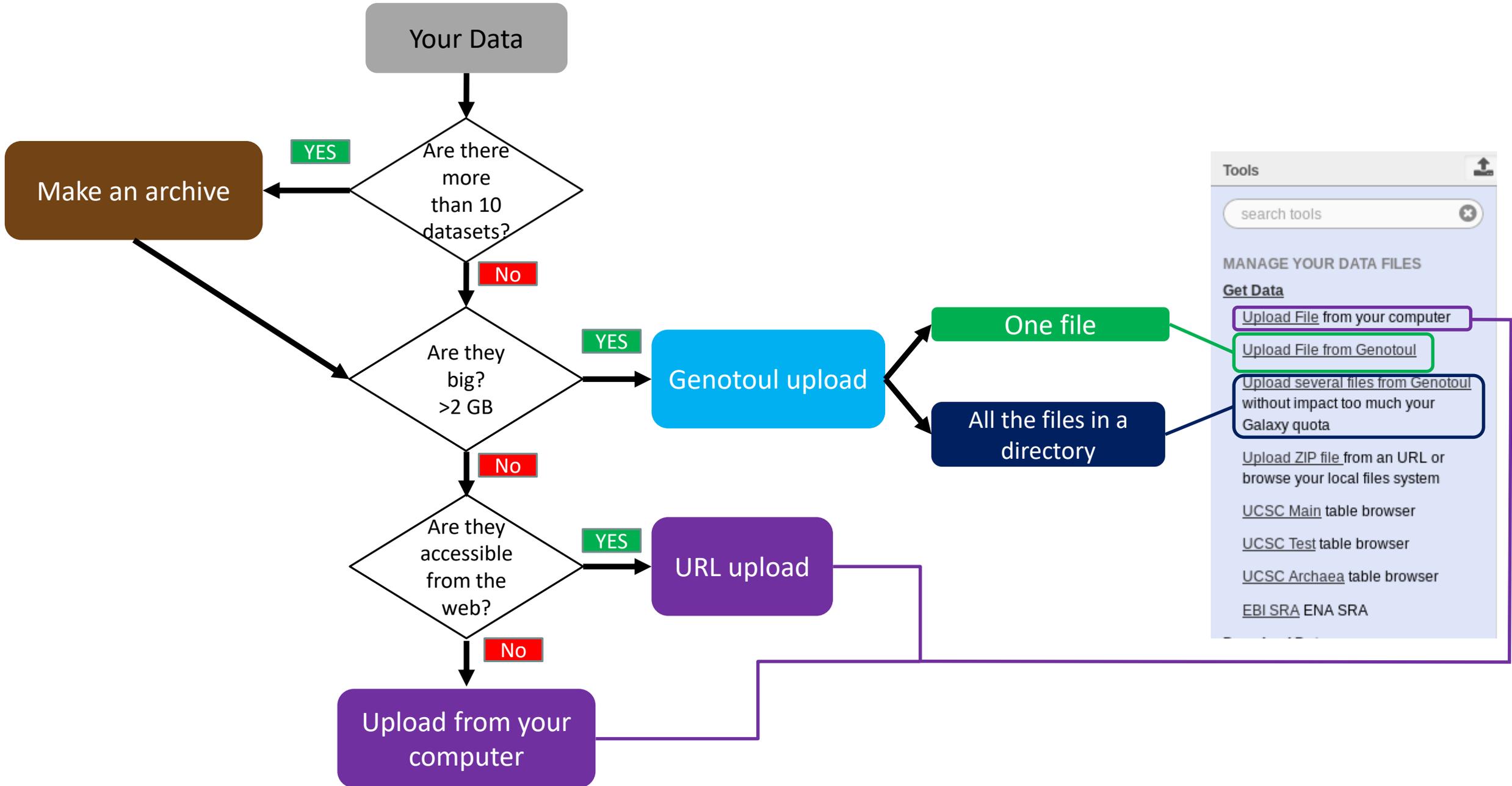
Your Turn!

PREPARE FILES

Formation data

- All the formation data are in the Data folder on your desktop.
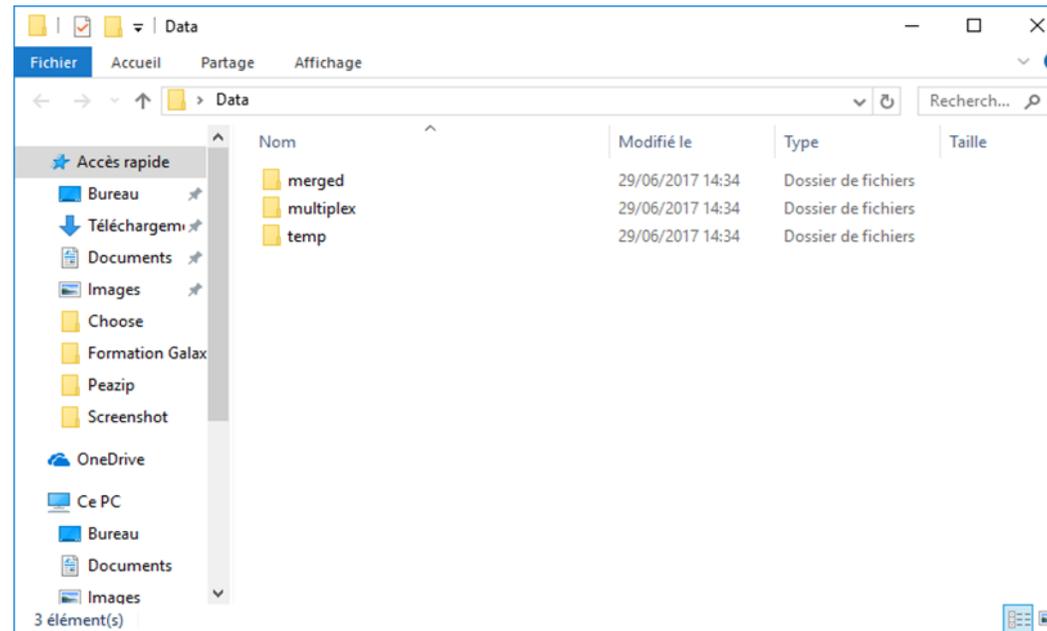
- If you don't have this directory on your desktop, ask us.

- You can find all the formation data online on:
http://genoweb.toulouse.inra.fr/~formation/15_FROGS/FROGS_ini/DATA/

Your Data

Are there more than 10 datasets?
- **YES** → Make an archive
- No ↓

Are they big? >2 GB
- **YES** → Genotoul upload
  - One file
  - All the files in a directory
- No ↓

Are they accessible from the web?
- **YES** → URL upload
- No ↓

Upload from your computer

44
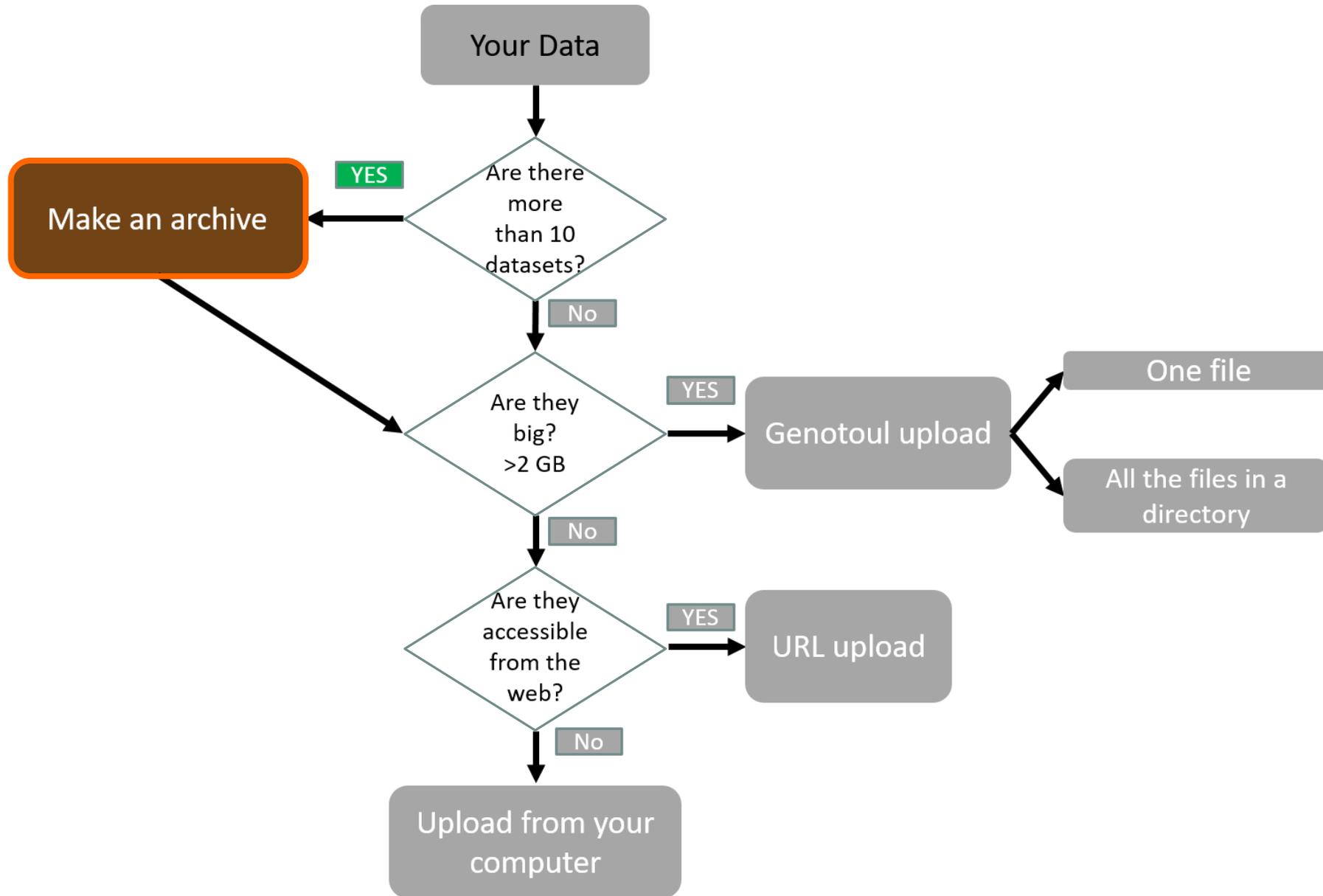
Create a Tar file

- Uploading multiple local files is time consuming.

- Solution: put all your files in an archive!

- What is a tar file?
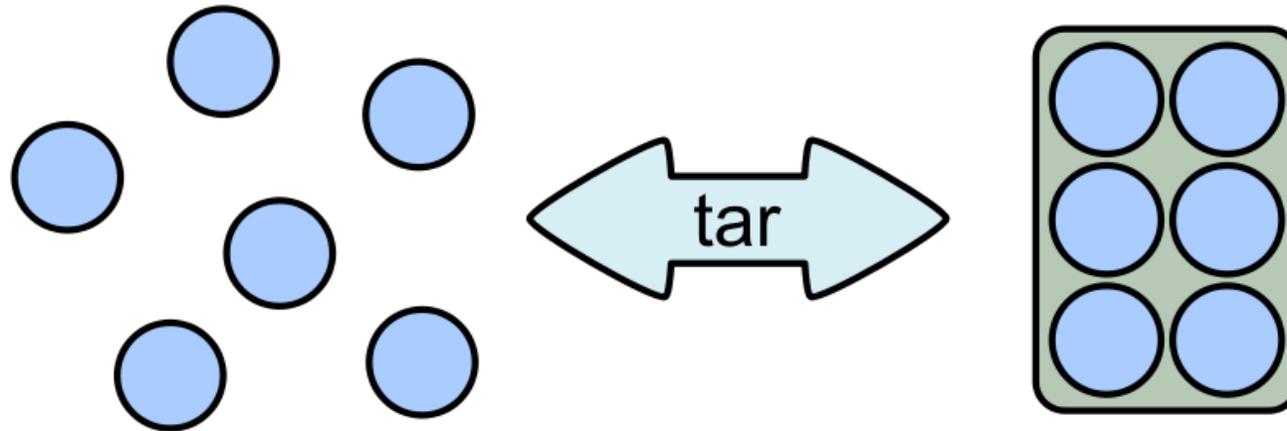
Create a Tar.gz

- Moreover, we can compress the archive to free up space.
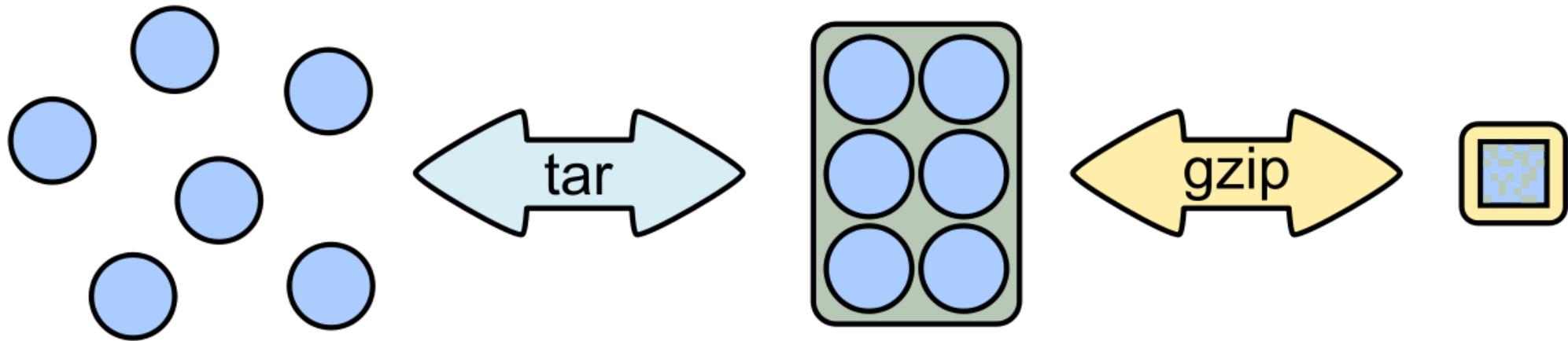
Your Turn!

CREATE AN ARCHIVE WITH THE FILES IN MERGED

7-Zip

- 7-Zip is a software called a file archiver.

- Can archive and compress files.

- Open source

- At the lab, you could download it at: http://www.7-zip.org/

Create a Tar file with 7-Zip

- Browse into the Data directory and next into the merged directory.
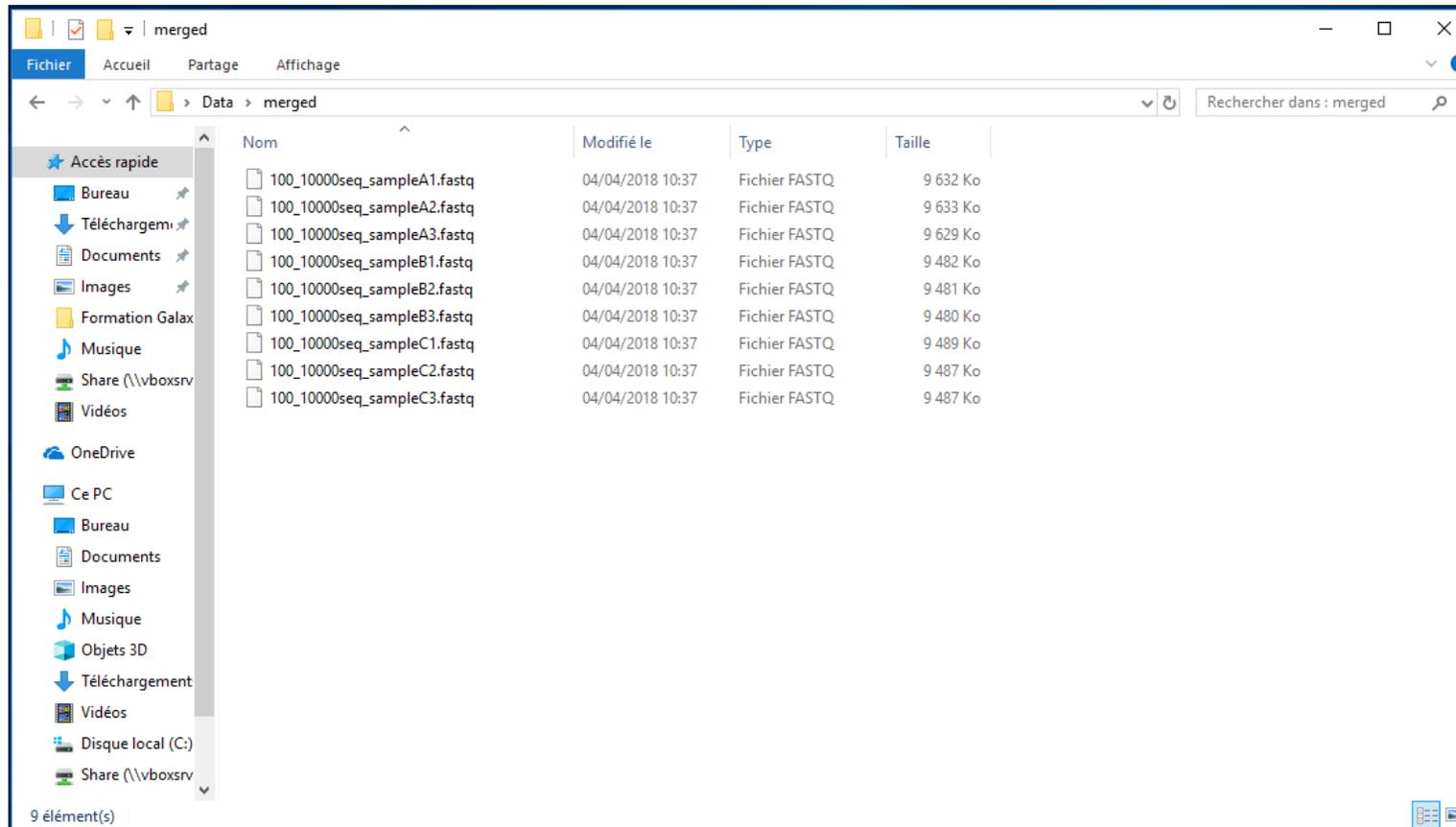
Create a Tar file with 7-Zip

- Select all files and right click on it > 7-Zip > Add to archive . . .
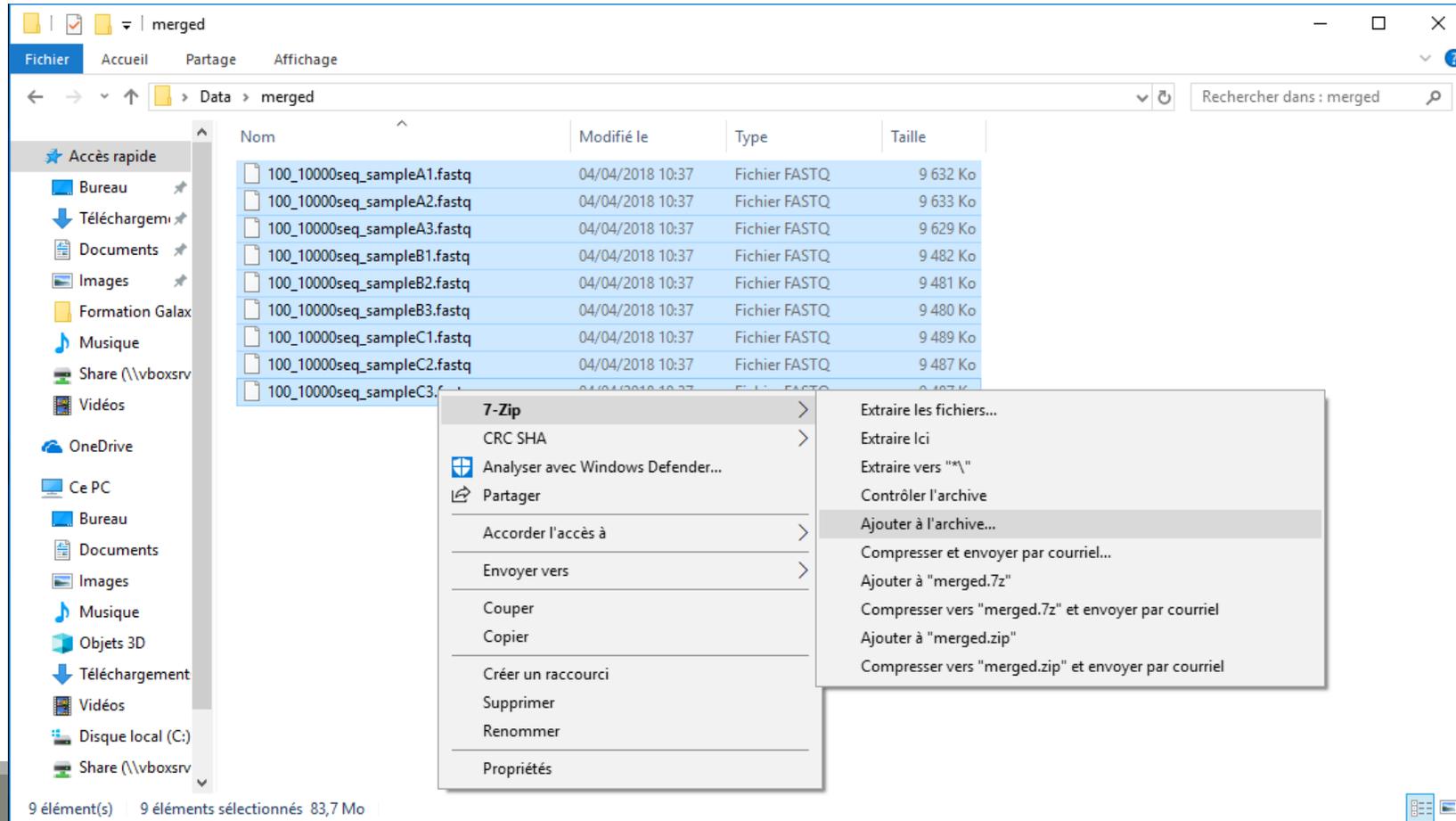
Create a Tar file with 7-Zip

- Change the name of the archive to:

« 100_10000seq_sample.tar »

- Choose « tar » as archive format .
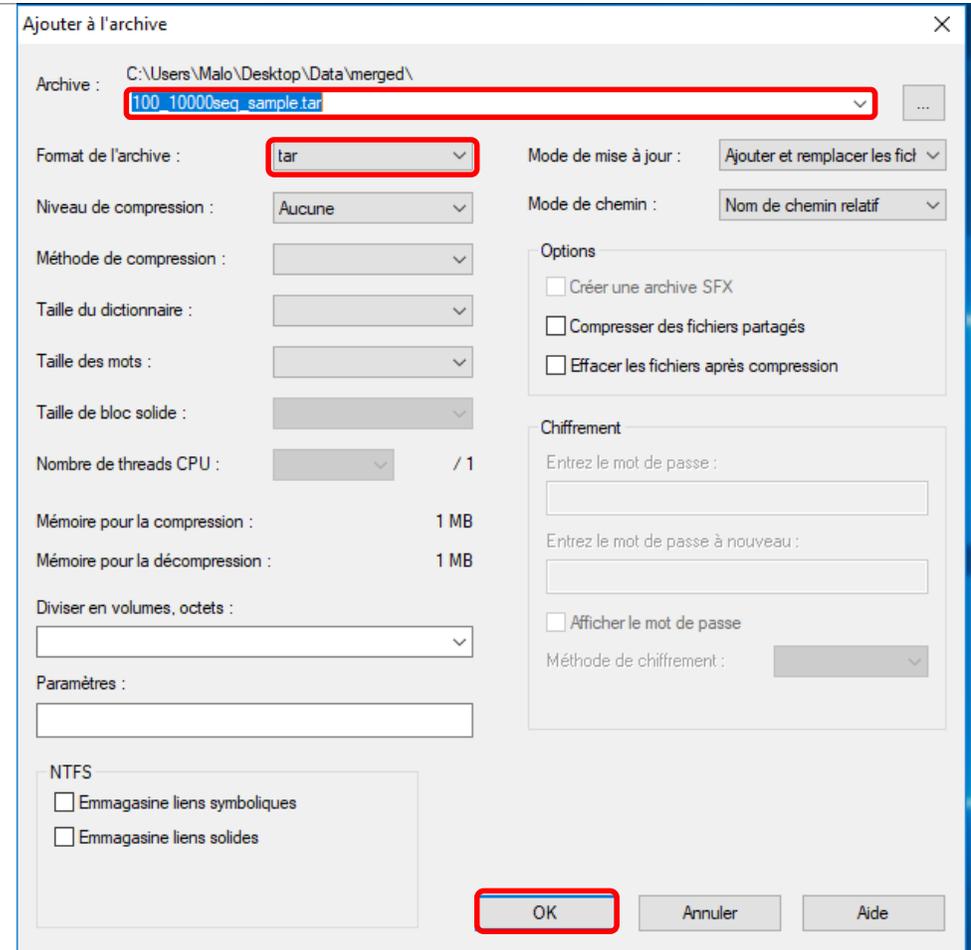
- Click on OK .

Create a Tar file with 7-Zip

- Select the newly created archive and right click on it > 7-Zip > Add to archive . . .
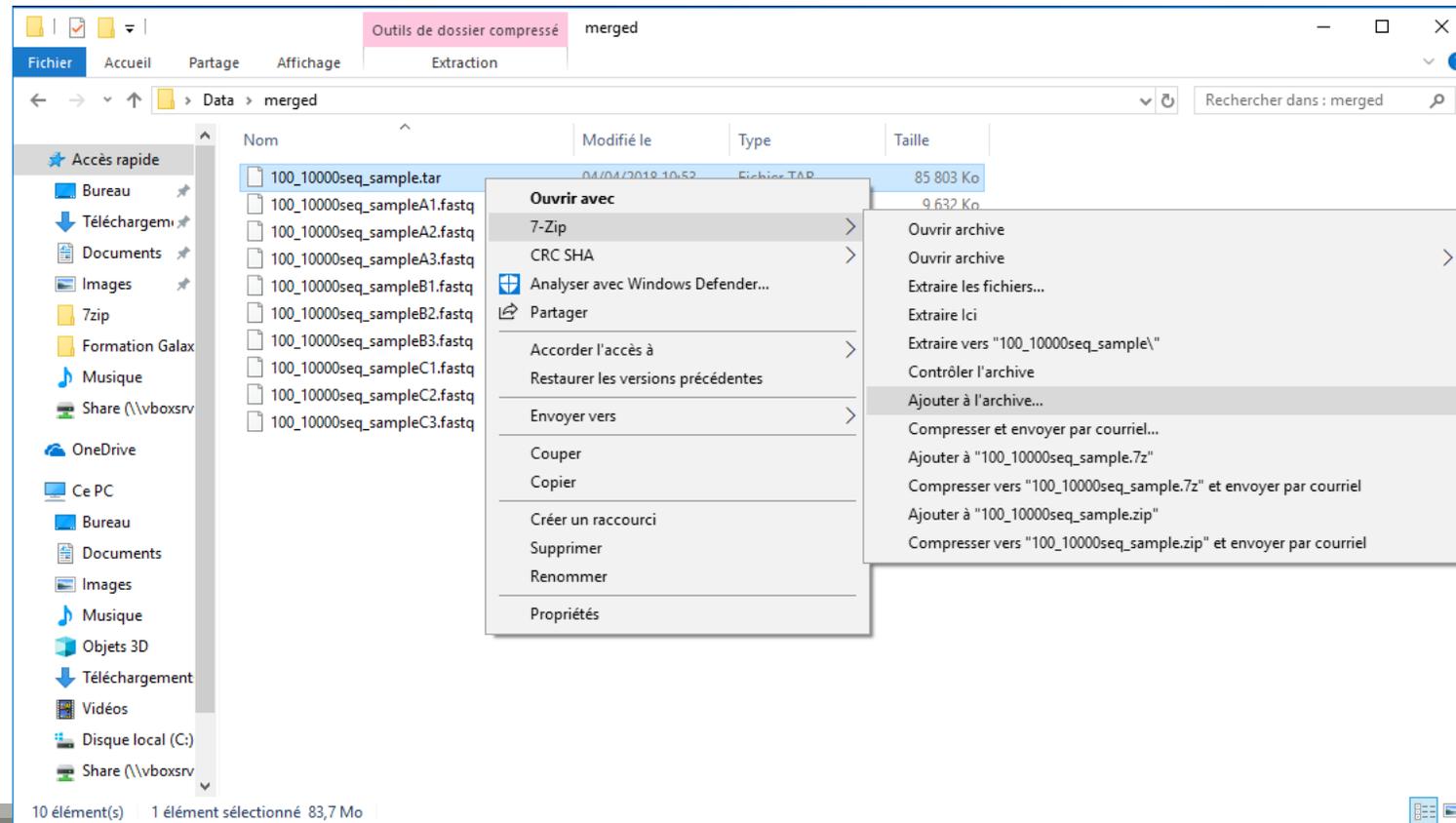
Create a Tar file with 7-Zip

- Choose « gzip » as archive format .
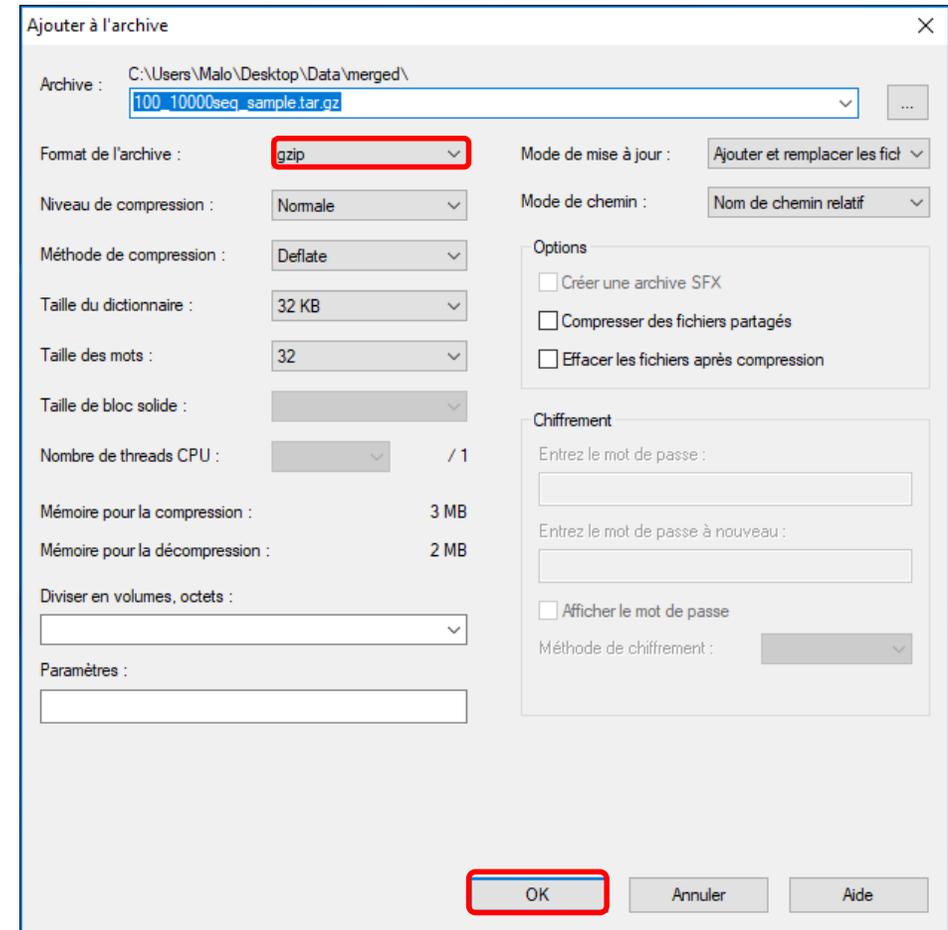
- Click on OK .

Create a Tar file with 7-Zip
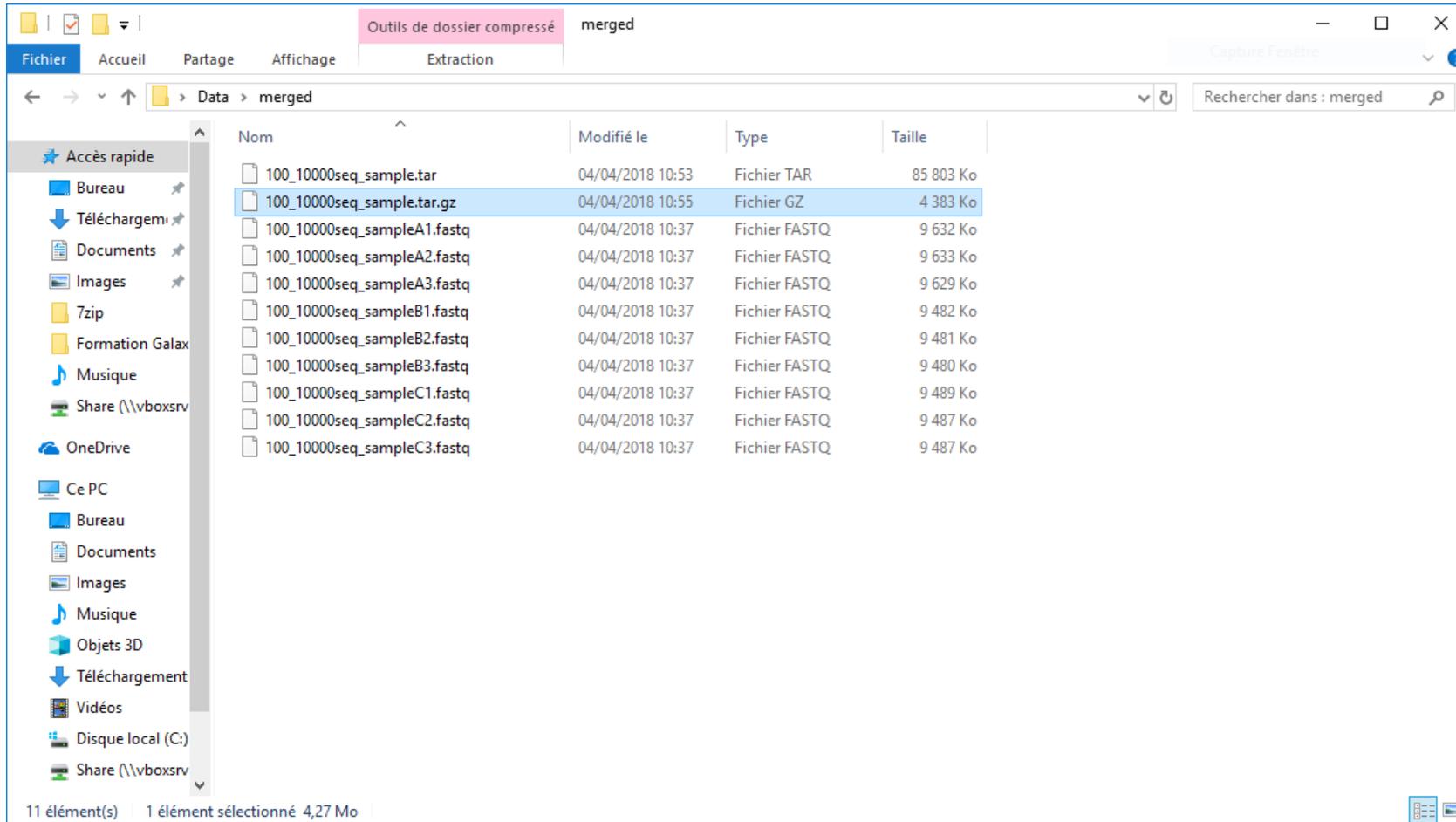
Your Turn!
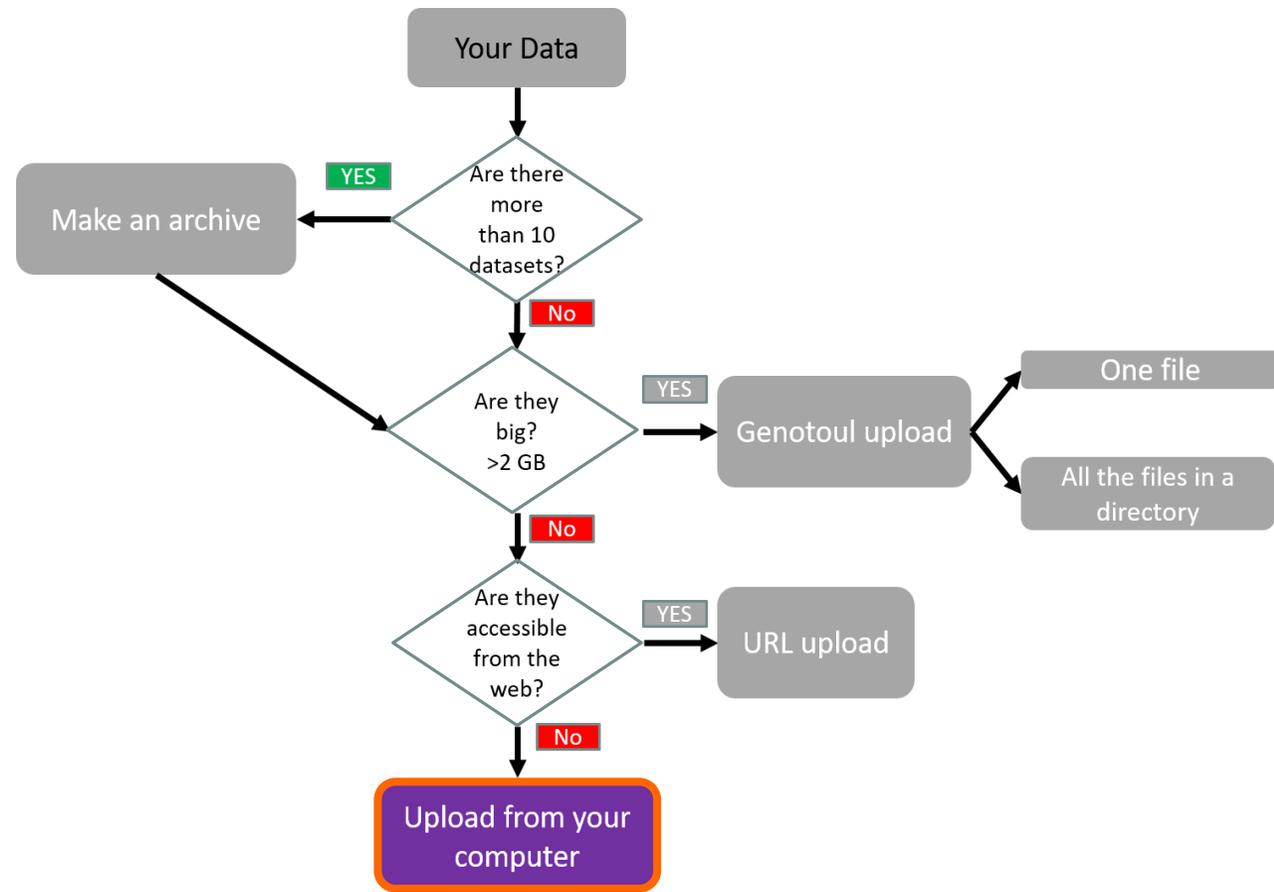
UPLOAD FILES FROM YOUR COMPUTER

Upload local files

- In Galaxy, your current history must be multiplex.

- Click on the « Get Data » tool.

- Upload file from your computer.

⚠ For files smaller than 2 GB
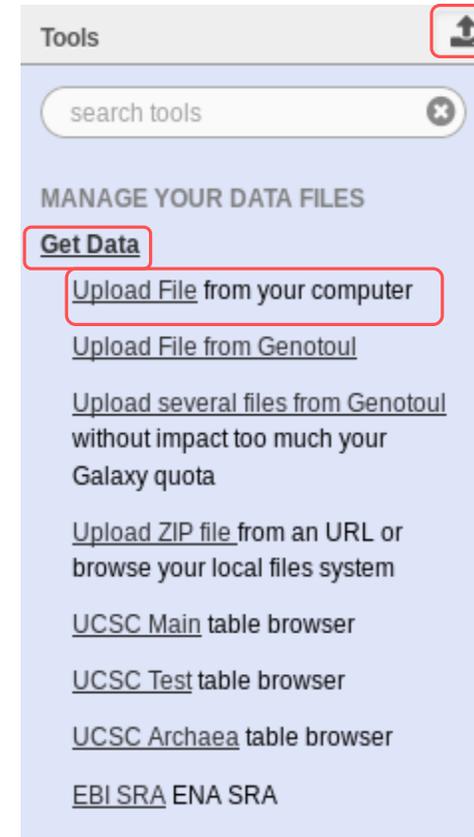
Upload local files

Download from web or upload from disk

Regular   Composite

Drop files here

Type (set all): Auto-detect 🔍   **Genome (set all):** unspecified (?)

Click here to choose a file on your hard drive.

Upload one of the two files in the multiplex directory.

Choose local file Paste/Fetch data Pause Reset Start Close

Upload local files

The datatype of a file indicates

the content of it.

.fasta = fasta

.fastq = fastq

.tar = tar

Select the type corresponding

to your file.

Do no not trust blindly

the auto-detect!

Download from web or upload from disk

Regular   Composite

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status	
🖥 barcode_forward.tabular	260 b	Auto-det... ▾ 🔍	unspecified (?) ▾	⚙		🗑
🖥 multiplex.fastq	2.1 MB	Auto-det... ▾ 🔍	unspecified (?) ▾	⚙		🗑

1. Select the type of file (Do not leave on Auto-Detect!).

2. Select your other files the same way.

3. Begin upload.

Type (set all): Auto-detect ▾ 🔍   **Genome (set all):** unspecified (?) ▾

🖥 Choose local file   ✎ Paste/Fetch data   Pause   Reset   Start   Close

Upload local files
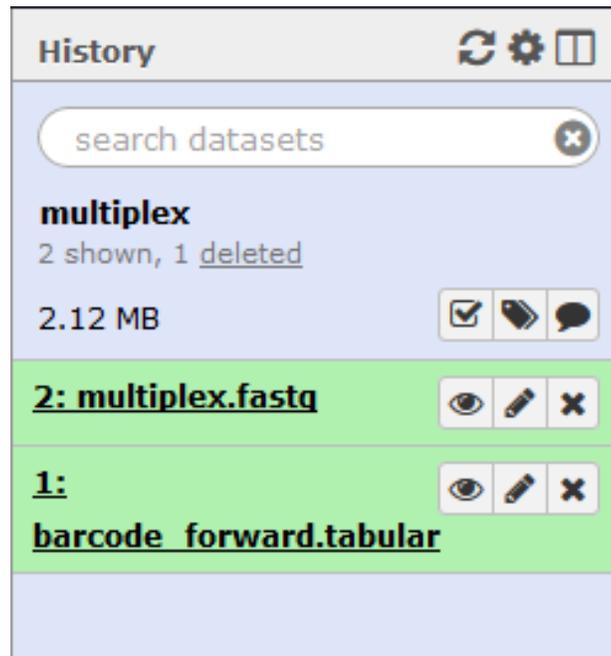
Your Turn!

UPLOAD FILE FROM AN URL

Your Data

Are there more than 10 datasets?
- YES → Make an archive
- No ↓

Are they big? >2 GB
- YES → Genotoul upload → One file / All the files in a directory
- No ↓

Are they accessible from the web?
- YES → URL upload
- No ↓

Upload from your computer
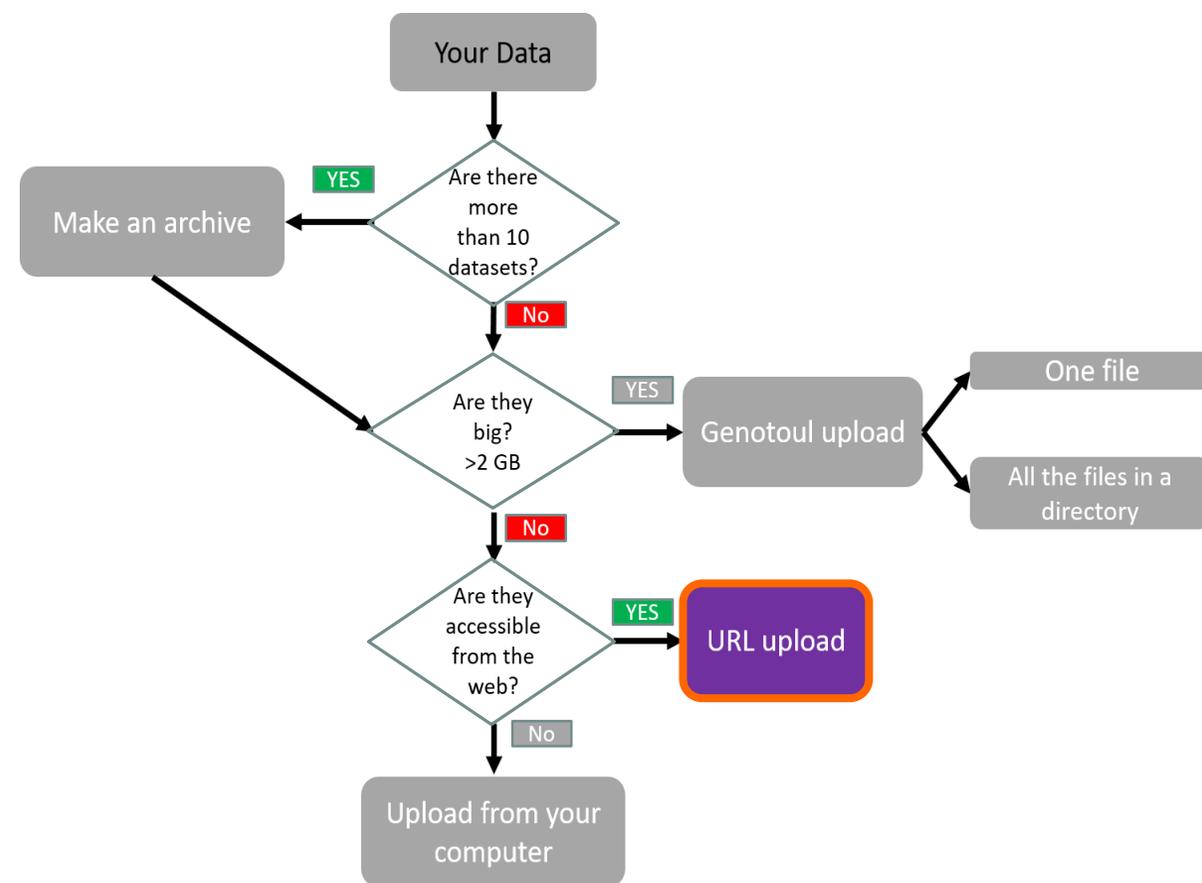
Upload file from URL

1. Switch to 454 history as current history.

2. Go to **Get Data** > **Upload File from your computer**

3. Click on **Paste/Fetch Data**

4. Copy the address of the file:
   http://genoweb.toulouse.inra.fr/~formation/15_FROGS/FROGS_ini/DATA/454.fastq

5. Change the type!

6. Click on Start.

7. You can put one address per line for multiple uploads.
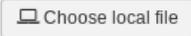
Upload file from URL

Your Turn!

Your Data

Are there more than 10 datasets?
YES → Make an archive
No

Are they big? >2 GB
YES → Genotoul upload
- One file
- All the files in a directory
No

Are they accessible from the web?
YES → URL upload
No

Upload from your computer
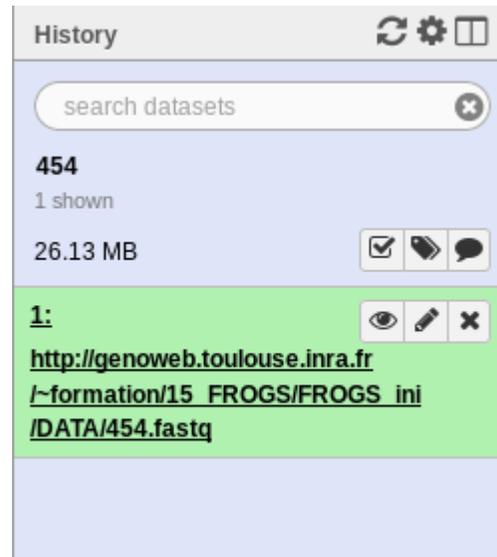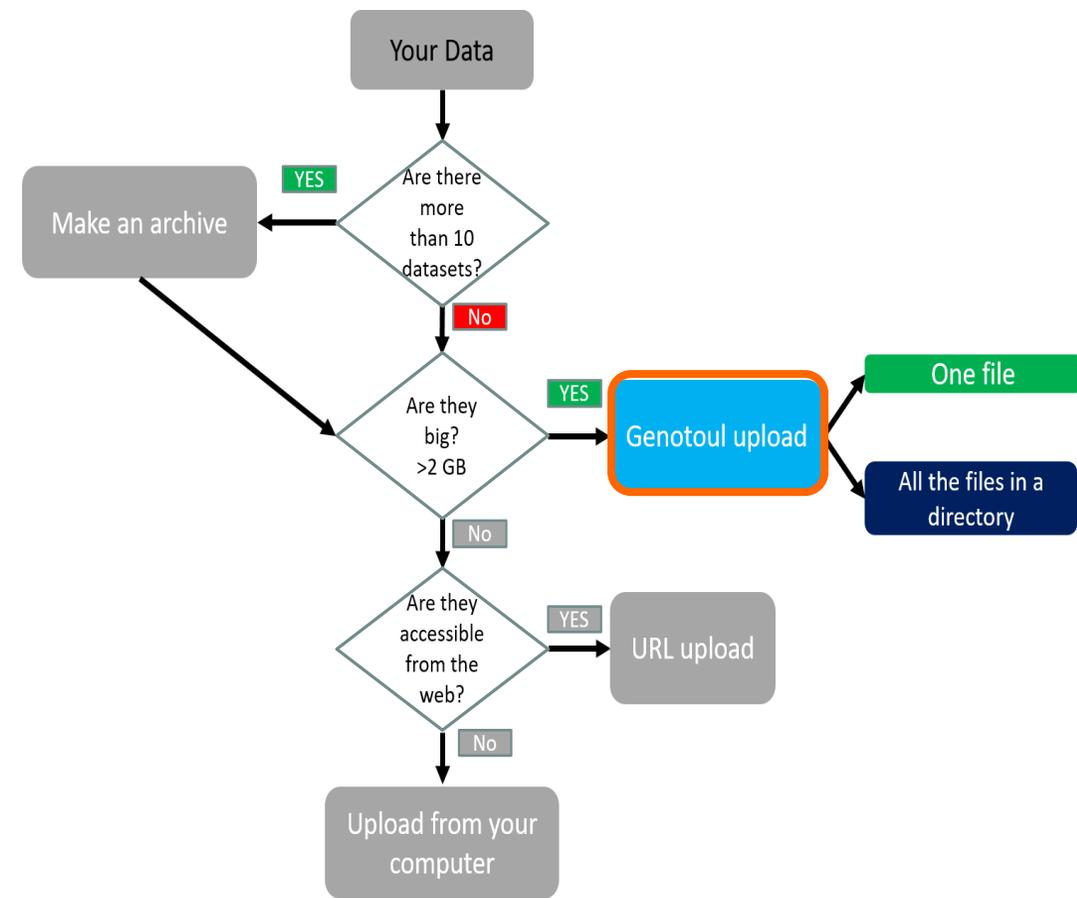
Objectives
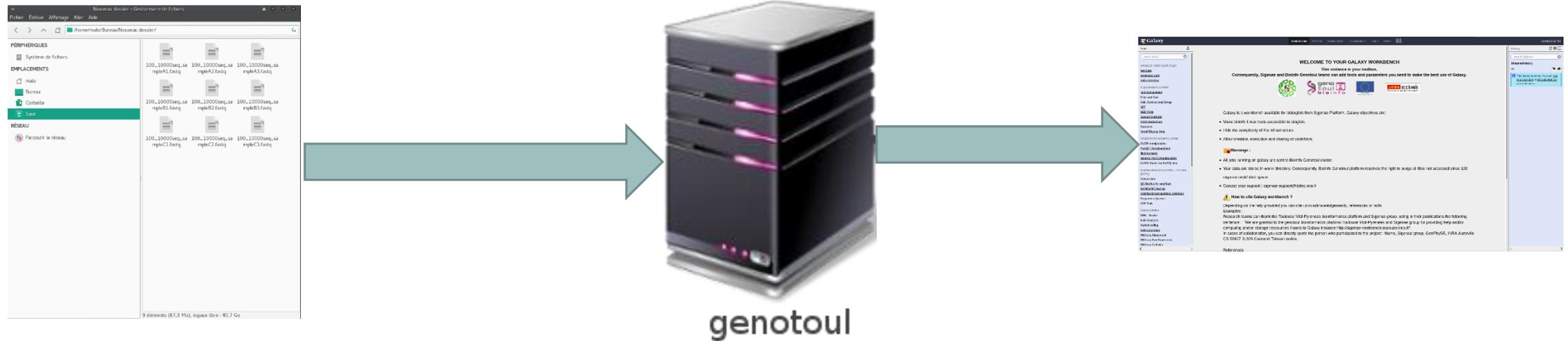
Transfer your files to your Genotoul account and link the file to Galaxy.

Preparation

- Open the data directory on your desktop.

- Launch Filezilla.

  - Filezilla is a FTP client *i.e.* can transfer files to a distant server.

  - You can download it on : https://filezilla-project.org/

Browse to the data directory on your desktop.

To connect to Genotoul click on this icon.

- Host: genotoul.toulouse.inra.fr

- Port: 22

- Protocol: SFTP

- User: your Genotoul login

- Logon Type: Ask for password

- Click on connect.

- Password: your password

Double click on work directory to access it.

- Create new directory with right click > Create directory.
  - Name it « Formation ».
  - Go inside this directory next.

- Check that all boxes for execute and read are checked.
  - Recurse this action to all files and subdirectories.
  - That allows Galaxy to access your files on Genotoul.
    - Click on Ok.

Upload files from Genotoul

Switch to merged history.

Next go to Get Data > Upload File from Genotoul.

Path to file

/work/mleboulch/Formation

Path must be like : /work/USERNAME/somewhere/afile

Path to file

/work/mleboulch/Formation/100_10000seq_sample.tar.gz

Path must be like : /work/USERNAME/somewhere/afile

You must add « / » and the name of the file at the end of the text.

After executing, the file is imported into Galaxy.

Upload files from Genotoul

Galaxy

Analyze Data   Workflow   Shared Data ▾   Visualization ▾   Help ▾   User ▾

Using 813.2 MB

Tools

search tools

MANAGE YOUR DATA FILES

Get Data
  Upload File from your computer
  Upload File from Genotoul
  Upload several files from Genotoul without impact too much your Galaxy quota
  Upload ZIP file from an URL or browse your local files system
  UCSC Main table browser
  UCSC Test table browser
  UCSC Archaea table browser
  EBI SRA ENA SRA
Download 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

Upload several files from Genotoul without impact too much your Galaxy quota (Galaxy Version 1.0.1)   ▾ Options

Path to your directory which contains several files

/work/mleboulch/Formation/temp

Path must be like : /work/USERNAME/somewhere/

✔ Execute

ℹ **What it does**

This program allows you to use a file stored in your genotoul work directory and optimize Galaxy work space by creating symlinks.

Path to file

This must be an absolute path to a file located in your genotoul work directory. The path must start with **/work/YOUR_USER_NAME/directory**

valid path : /work/LinuxUserName/directory
invalid path : /home/LinuxUserName/work/directory

⚠ To use this tool and to maintain the confidentiality of yours directories:

1. Create a "galaxy" directory in your work : mkdir galaxy

2. chmod a+x /work/LinuxUserNam

Example : drwxr-x--x 4

ℹ **Thanks to the fact that this t**

For example, if your data to down

Add "x" rights to "others" on /work

It is not useful that "others" ha

Thus, Galaxy can access and rea

Version Galaxy Tool : V1.0

Versions of bioinformatics tools us

Contacts (noms et emails) : sigen

E-learning available : Yes.

Please cite :

Depending on the help provided you can cite us in acknowledgements, references or both.

Examples : Acknowledgements We wish to thank the SIGENAE group for ....

History

search datasets

temp

0 b

ℹ This history is empty. You can load your own data or get data from an external source

- Go back to Galaxy and switch to temp history.

- Go to **Get Data** > **Upload Several files from Genotoul**

  - Paste the address into field.

- Be careful all the files from the directory will be uploaded!

Click on execute.

All the files from the directory are uploaded.

Upload by Genotoul
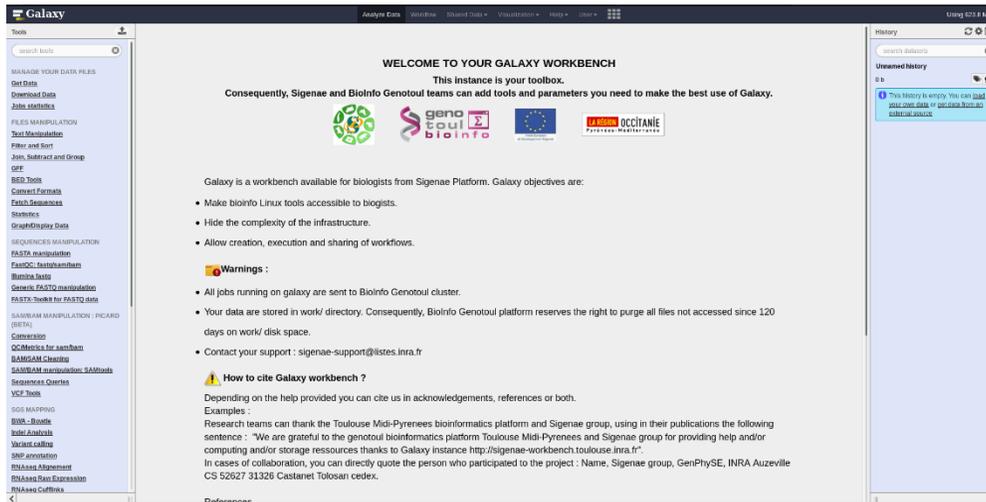
1TB of space

/work: 1TB (1024GB) of space
/save: 250GB of space

- This method allows you to have more disk space and to upload bigger files.

- Files not used in the last 180 days can be purged at any time.

Share a History

Share a history

- You can share a history with another Galaxy user:

  - For working with your colleague.

  - For support, so we can help you better and faster.

- You can import shared history to your account too.

Your Turn!

IMPORT A SHARED HISTORY TO YOUR ACCOUNT

Import a shared history

Click on the wheel and click here.

Your Turn!

SHARE A HISTORY WITH YOUR NEIGHBOUR

Share a history

- Switch to the history you want to share.
- Click on the wheel and click here.

History

HISTORY LISTS

Saved Histories

Histories Shared with Me

HISTORY ACTIONS

Create New

Copy History

Share or Publish

Show Structure

Galaxy user emails with which to share histories

malo.leboulch@inra.fr

malo.leboulch@inra.fr   a-separat

Submit

- You must click on the address in the dropdown menu.
- Click on submit.
- The history will appear in your co-worker's history.

Manipulate datasets

Your Turn!

RENAME A DATASET

Rename a dataset

- Switch to 454 history.

Click here to display attributes and change the name.

Rename a dataset

Attributes   Convert Format   Datatype   Permissions

Edit Attributes

Name:

http://genoweb.toulouse.inra.fr/~formatio

Change the name here and call it « 454.fastq ».

Info:

uploaded fastq file

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

unspecified (?)

Save

Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Rename a dataset

Attributes   Convert Format   Datatype   Permissions

Edit Attributes

Name:

http://genoweb.toulouse.inra.fr/~formatio

Info:

uploaded fastq file

If you put the wrong datatype in during the upload, you can change it here.

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

unspecified (?)

Save

Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Rename a dataset

- Do the same with the merged history:

  - Switch to the merged history.

 - Change the name of the file to «100_10000seq_sample.tar.gz ».

Your Turn!

DELETE A DATASET

Delete a dataset

- Switch to temp history.
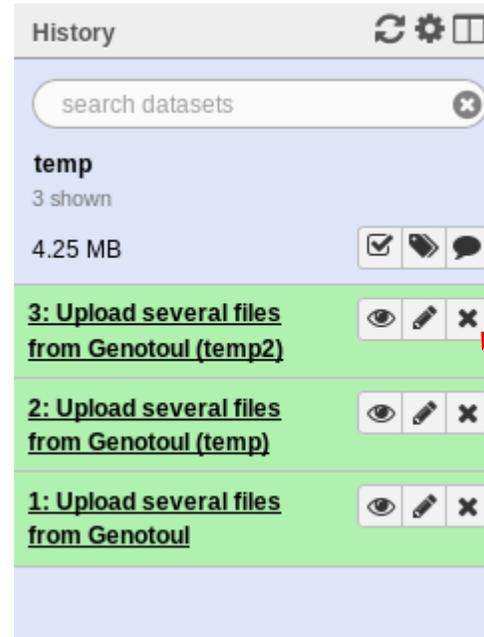
Click here to delete a dataset.

Delete a dataset

Click here.

Delete this dataset permanently.

Nothing is deleted in Galaxy until you delete it permanently (=purge).

How to delete a history?

- Stay in the temp history.

- Click on the wheel.

- Click on delete.

- A deleted item on Galaxy is recoverable.

- To definitively delete it:
  click on « Delete Permanently ».
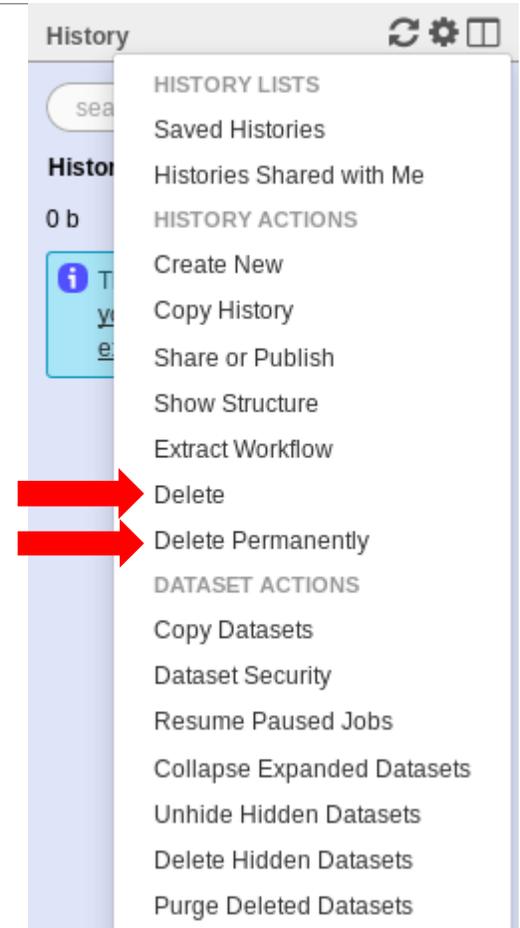
One word about data collections

- Used to perform the same analysis on multiple files.

- Gather multiple datasets in one collection.

Don't work with FROGS!

Don't work with FROGS!

B.

C.

D.

E.

F.

G.
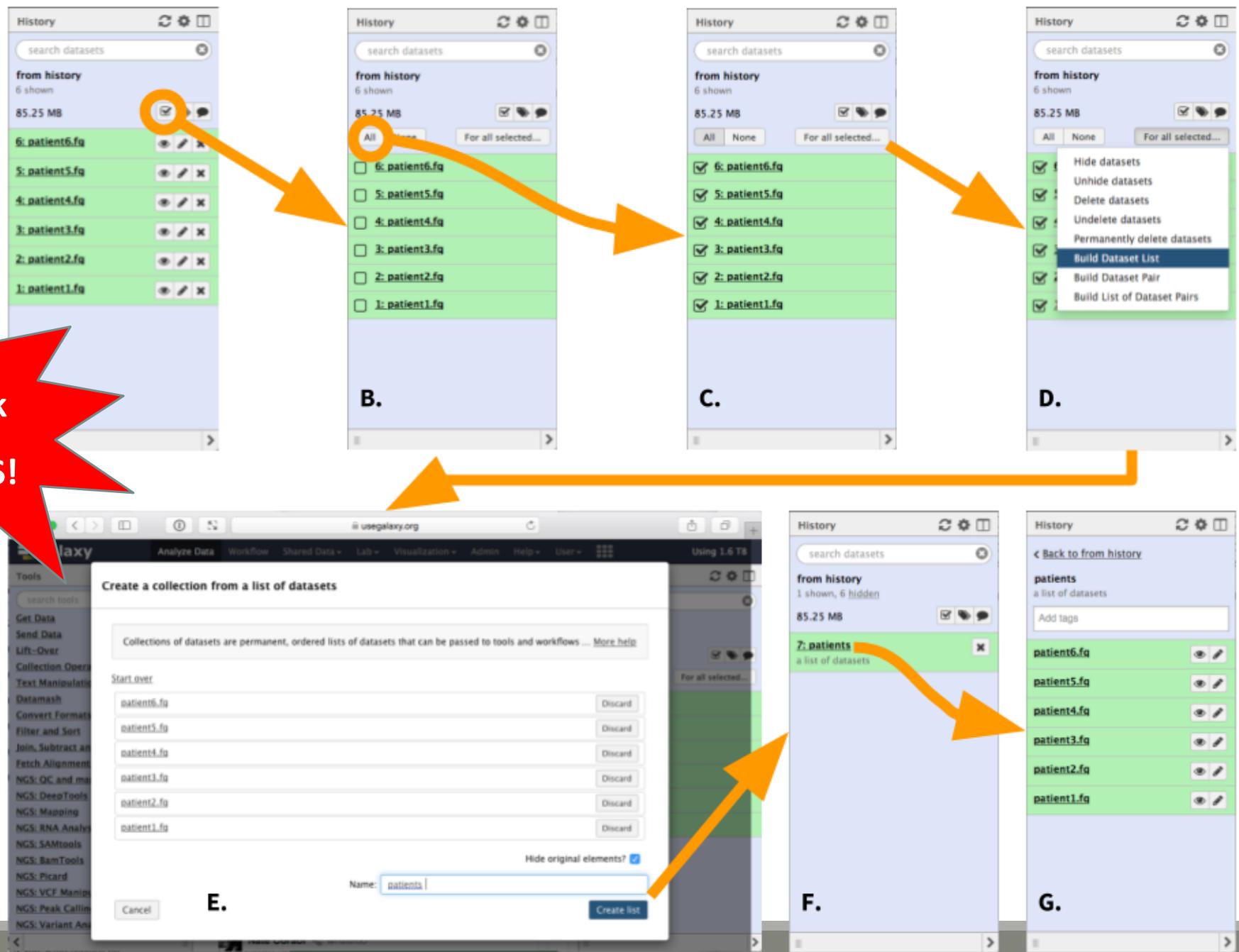
Others tools available on galaxy

- Text Manipulation

- Filter and Sort

- FASTA manipulation

Galaxy support

- Mail: support.sigenae@inra.fr

- If you need more training about bioinformatics and Galaxy, please connect to Sigenae e-learning platform: http://sig-learning.toulouse.inra.fr/

How to cite Galaxy?

- Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. *"Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences."* Genome Biol. 2010 Aug 25;11(8):R86.

- Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. *"Galaxy: a web-based genome analysis tool for experimentalists"*. Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.

- Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. *"Galaxy: a platform for interactive large-scale genome analysis."* Genome Research. 2005 Oct; 15(10):1451-5.

How to cite Genotoul Galaxy workbench?

▪ Research teams can thank the Toulouse Midi-Pyrenees bioinformatics platform and Sigenae group, using in their publications the following sentence : "We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees and Sigenae group for providing help and/or computing and/or storage resources thanks to Galaxy instance http://sigenae-workbench.toulouse.inra.fr".

▪ In cases of collaboration, you can directly quote the person who participated to the project : Name, Sigenae group, GenPhySE, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex.