Vall d'Hebron
Institut de Recerca

DATA FORMATS IN NGS
INTRODUCTION TO GALAXY

Bioinformàtica per a la Recerca Biomèdica

Mireia Ferrer[1], Álex Sánchez[1,2]
Esther Camacho[1], Angel Blanco[1,2]
1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR
2 Departament de Genètica, Microbiologia i Estadística, UB

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · Instituto de Salud Carlos III

Vall d'Hebron
Institut de Recerca

1. Data formats used in NGS
2. Introduction to Galaxy

2. Introduction to Galaxy

- An open, web-based platform integrating many popular tools and resources for intensive biomedical research.
- **What can be done?**
  - Obtain data from many data sources like UCSC Table Browser, Biomart, WormBase, or your own data
  - Prepare data for further analysis by rearranging or cutting data columns, filtering data and many other options
  - Analyze data by finding overlapping regions, determining statistics, preprocessing NGS data and much more
  - Share data and workflows

2. Introduction to Galaxy

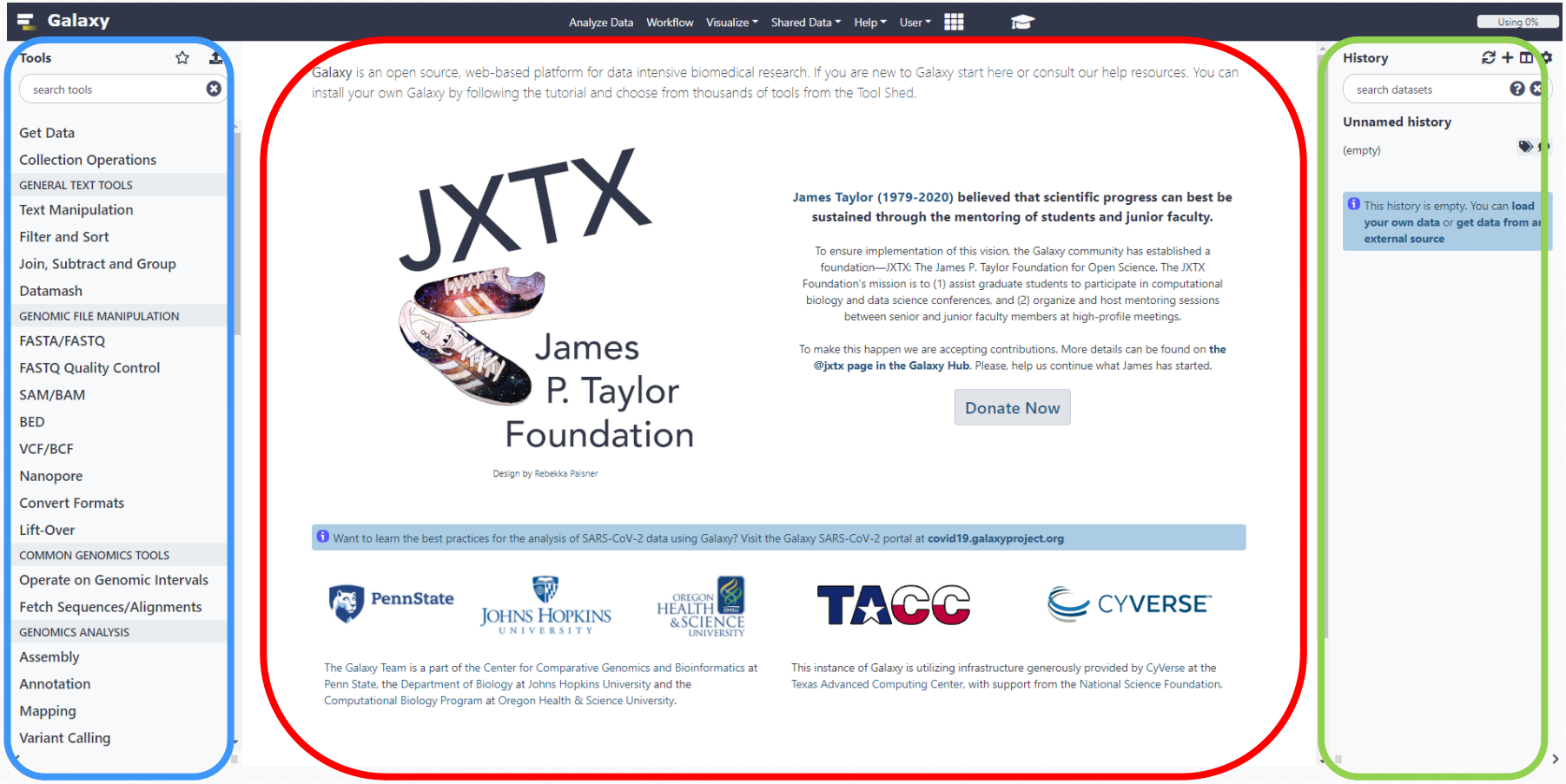
Vall d'Hebron Institut de Recerca

The Galaxy page is divided into three panels:

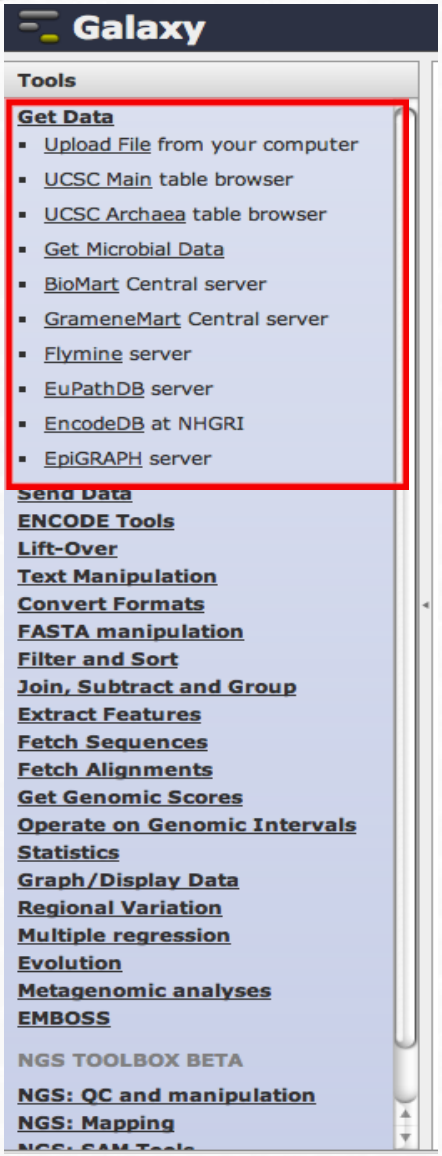
Tools for uploading, processing and analysis

Viewing panel (menus, data, results)

History of analysis steps and datasets



2. Introduction to Galaxy



Tools for data analysis

Get Data
- From databases (UCSC Table Browser, ...)
- From uploaded files
- From urls

Text manipulation

Filter and Sort

Operate on Genomic Intervals

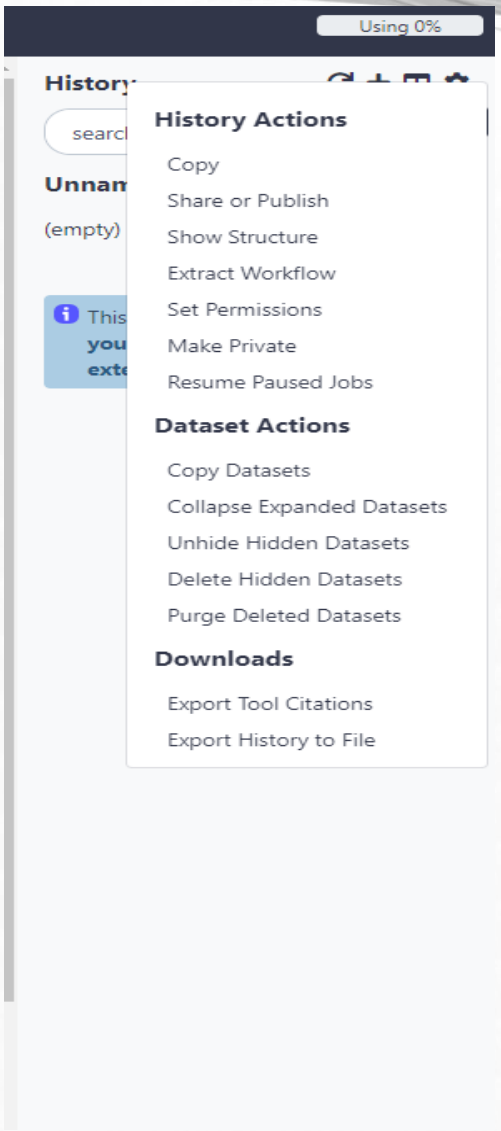
FASTA manipulation

NGS analysis
- QC
- Fastq file pre-processing
- Read Alignment / Mapping
- SAM tools

2. Introduction to Galaxy

Histories

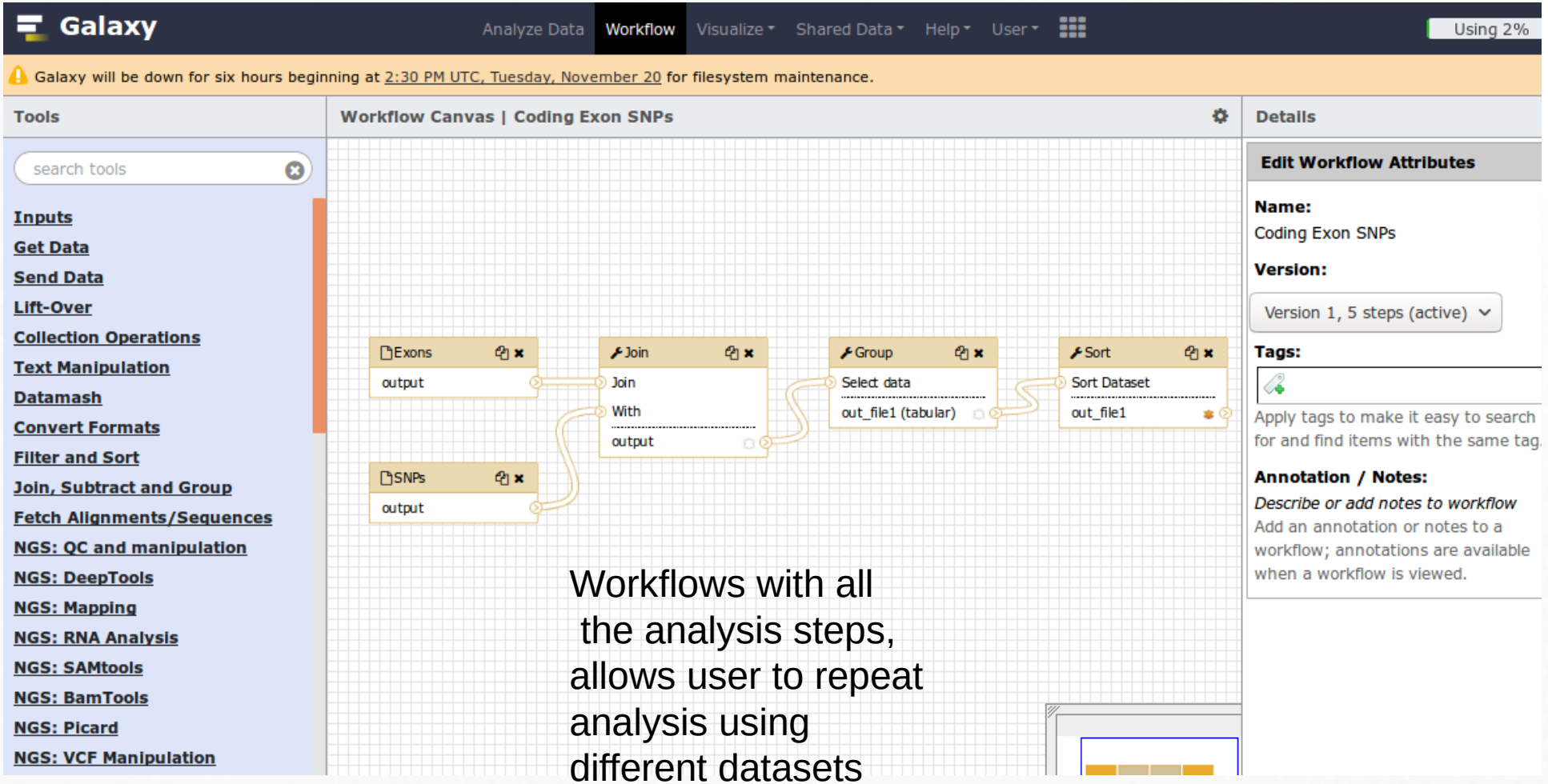
List saved histories and shared histories.
Work on Current History, create new, clone, share, create workflow, set permissions, show deleted datasets or delete history.



2. Introduction to Galaxy

Vall d'Hebron
Institut de Recerca

Workflows

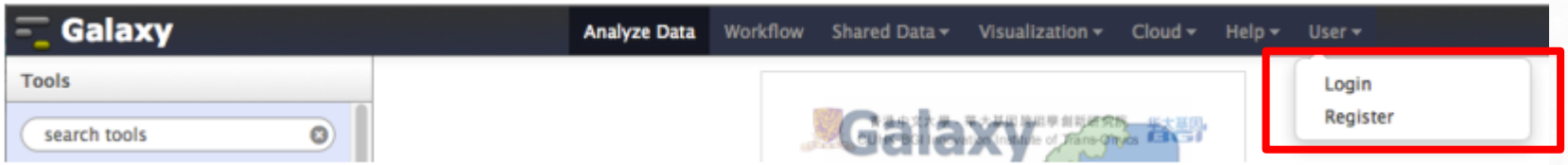


Workflows with all the analysis steps, allows user to repeat analysis using different datasets

2. Introduction to Galaxy

Register for a Galaxy account

This will only take a moment, and will allow all the work that you do to persist between sessions and allow you to name, save, share, and publish Galaxy histories, workflows, datasets and pages. It allows you to store up to 250GB of data on this public server.



https://usegalaxy.eu/

2. Introduction to Galaxy

Training Infrastructure as a Service

We want to help you conduct your training seminars. You provide the training, we provide you training infrastructure *at no cost*.

Why use UseGalaxy.eu training infrastructure?

- Free
- Private queue, no wait times
- No Galaxy Maintenance
- No Galaxy Administration
- Official Galaxy Training Materials guaranteed to work

Simply fill out the infrastructure request form and we'll get back to you shortly.

Find out more

After registration in European Galaxy server

https://usegalaxy.eu/join-training/ueb_bi2021

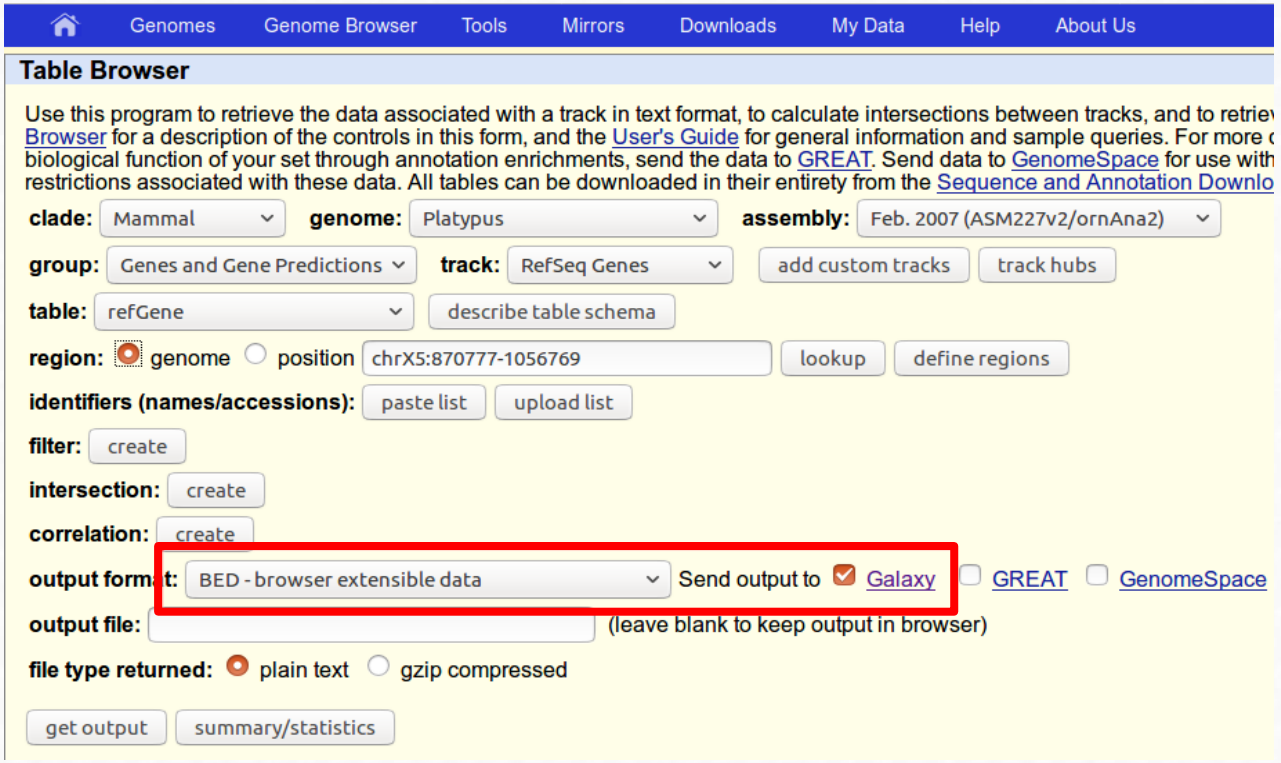
Vall d'Hebron Institut de Recerca

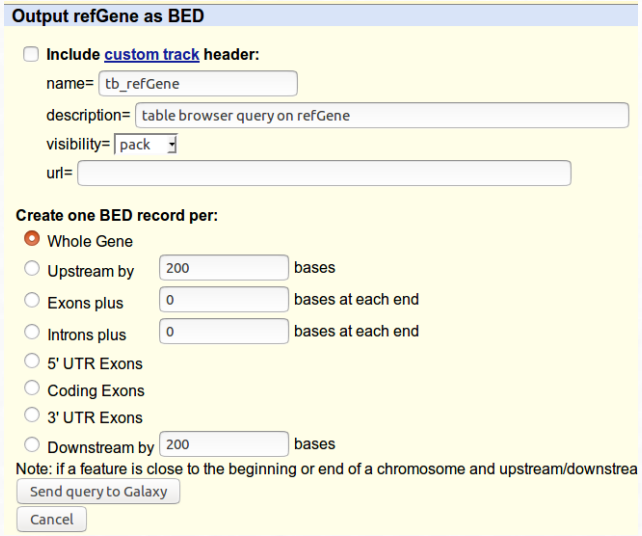
Importing data into Galaxy

1. From database queries (eg. UCSC): obtain a BED-formatted dataset of all RefSeq genes from platypus.

   Get Data > UCSC Main – Table Browser tool

   Set genome, RefSeg Genes, and BED output format (send to Galaxy)



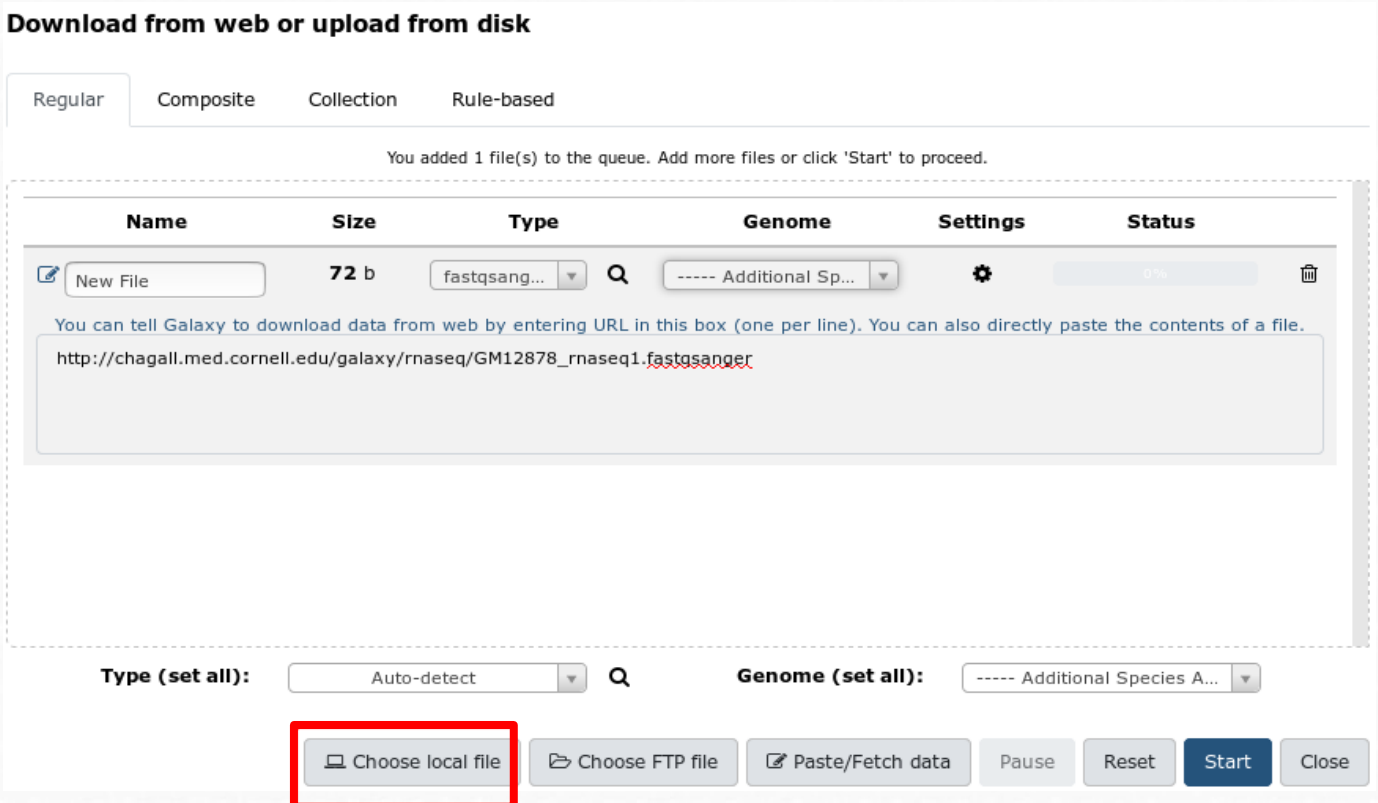


Vall d'Hebron
Institut de Recerca

Importing data into Galaxy

2. From a File on your computer / FTP file:

Get Data > Upload File



Vall d'Hebron
Institut de Recerca

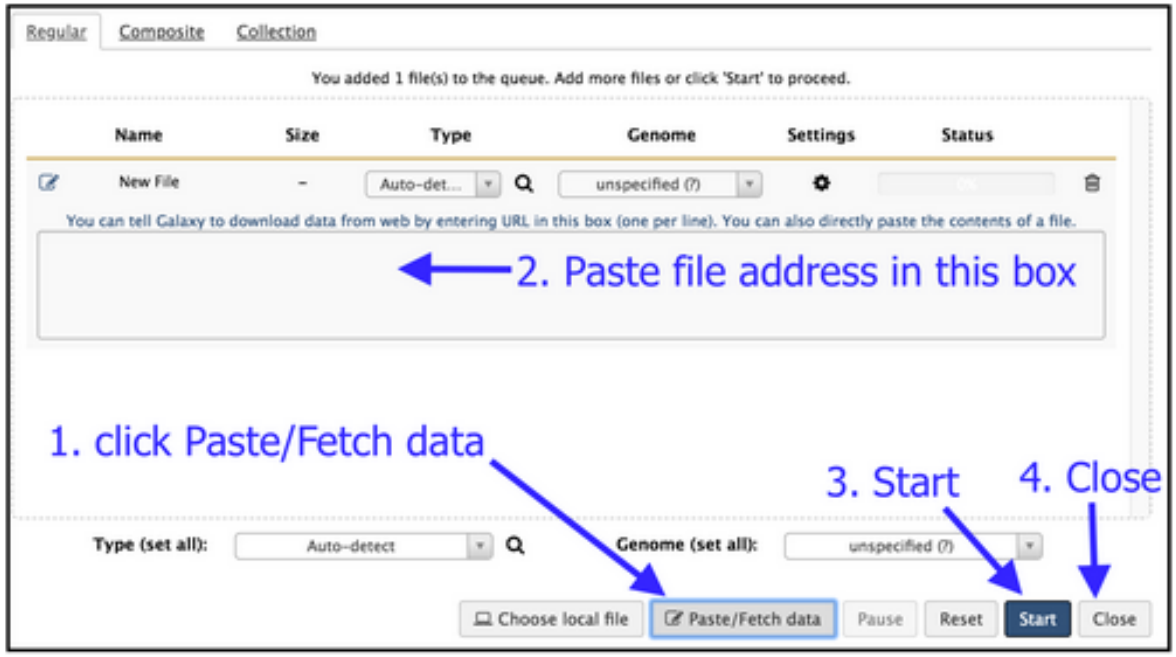
Importing data into Galaxy

3. From a website:

   Get Data > Upload File

   Copy this URL into the text-entry box:

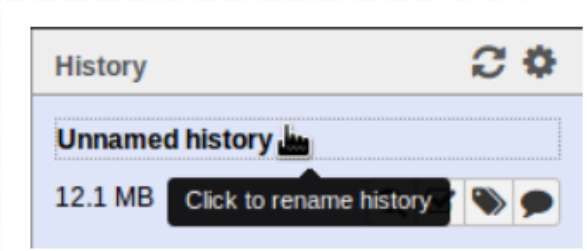
   url: https://zenodo.org/record/582600/files/mutant_R1.fastq

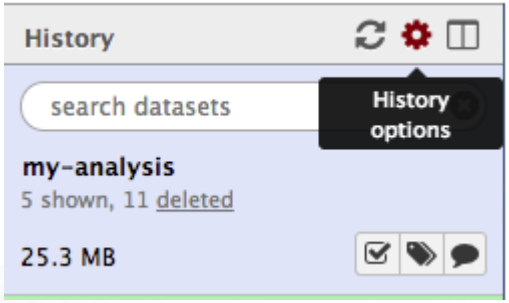


Vall d'Hebron Institut de Recerca

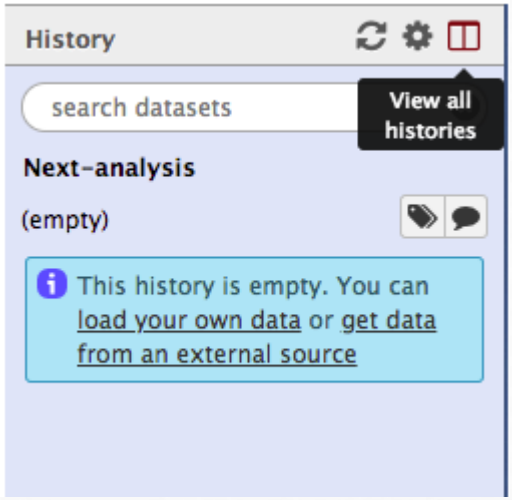
Managing histories

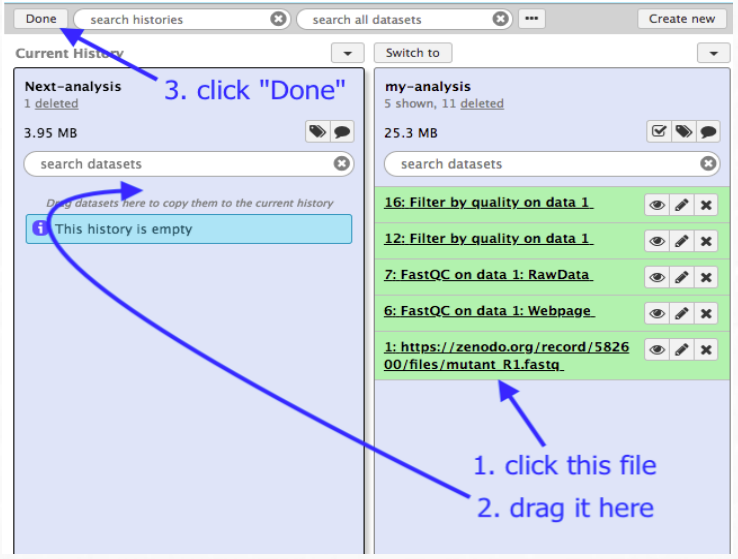
- Name your current history
- Create new history and rename it





- Manage datasets and histories:
- View all histories
- Drag files between histories (new history must be set to current)





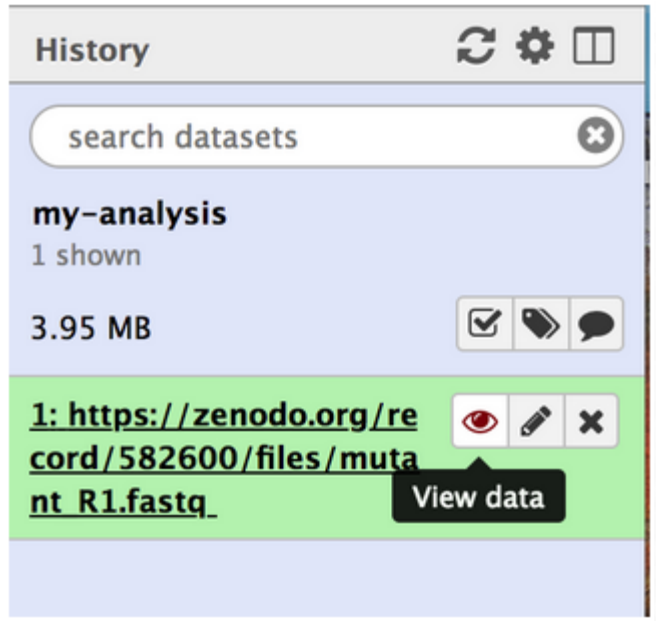
Vall d'Hebron
Institut de Recerca

Visualizing

- You can view content by clicking the eye icon on any step in your history.

The mutant_R1.fastq file contains DNA sequencing reads from a bacteria, in FASTQ format:

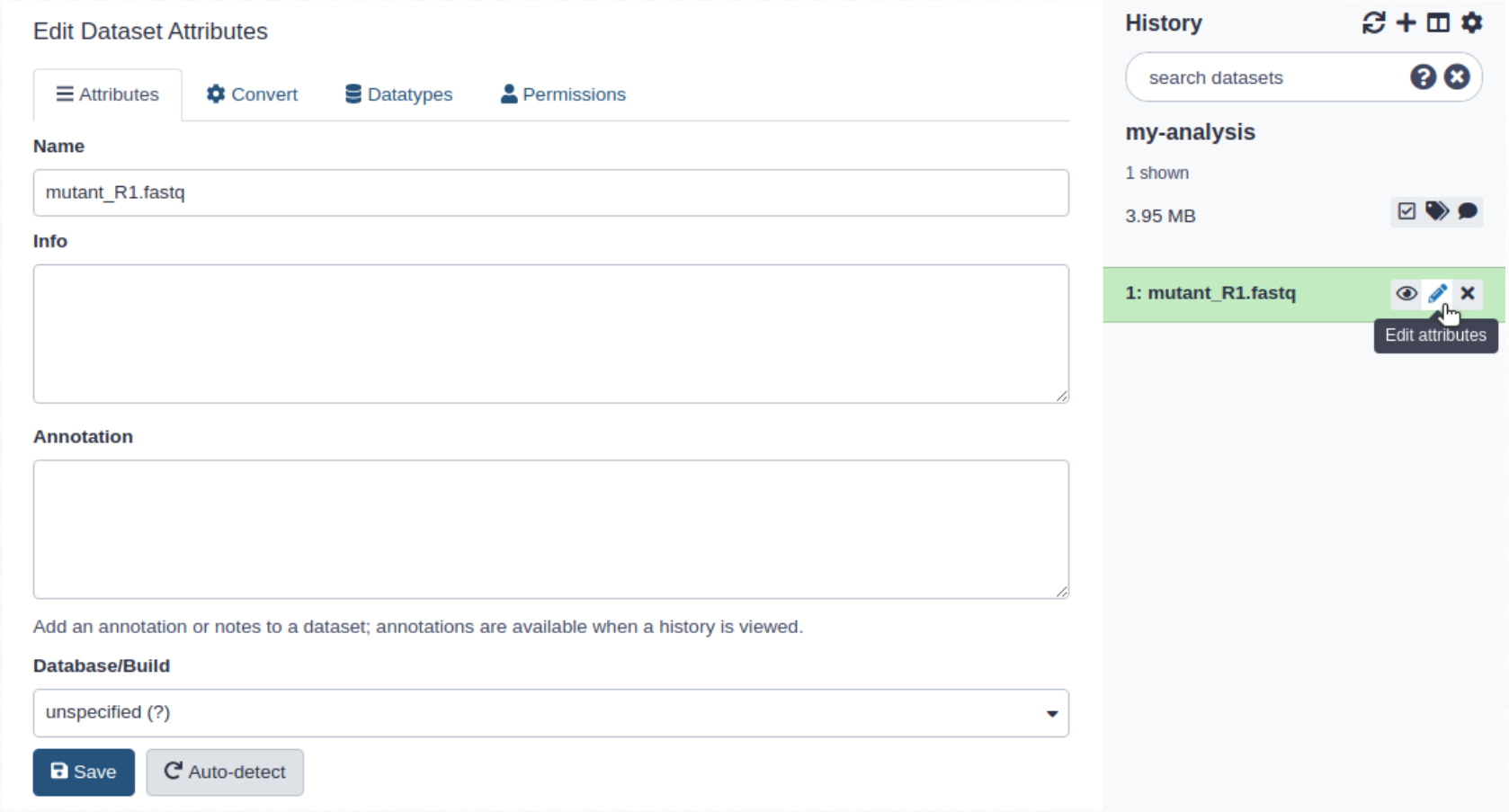
```
@mutant-no_snps.gff-24960/1                                    read 1 sequence
AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAG
+                                                              read 1 quality scores
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIHFIIHGICI
@mutant-no_snps.gff-24958/1
CAAAGTCGTTGGTCATATAAAAAACCGCGTACAGTCAACTATAGATACAATCAAGATAAACTCATGCACAGATTG
+
?A????@?DDDABDE9FGGGFGICFHIIIBGHIIIGICHHIFH=IHAFIHIHHHHIFCIIEIHAIFGIHIDDIHE
@mutant-no_snps.gff-24956/1
TATAAATTCAACTTTGCAACAGAACCATCTAATCTTCAACAAACTGGCCCGTTTGTTGAACTACTCTTTAATAAA
+
?????BBADD5DDDDDGFGCFFEECFBBICIII,IIHIICHIHIIFHHHHHIHIIIIIIIAHHIHHH5FHDHHHH
```



Vall d'Hebron
Institut de Recerca

Editing basic attributes

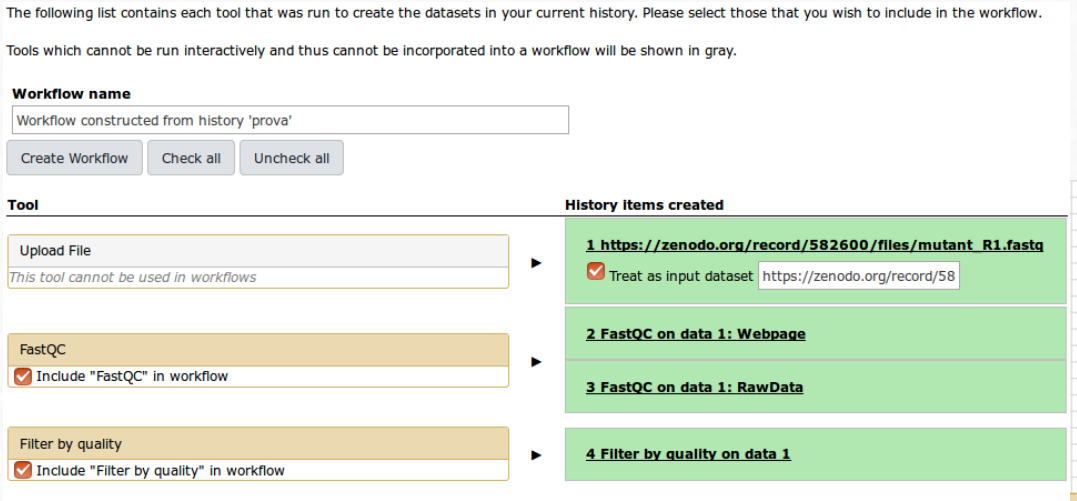
- You can edit several basic attributes by clicking the pencil icon on any step in your history

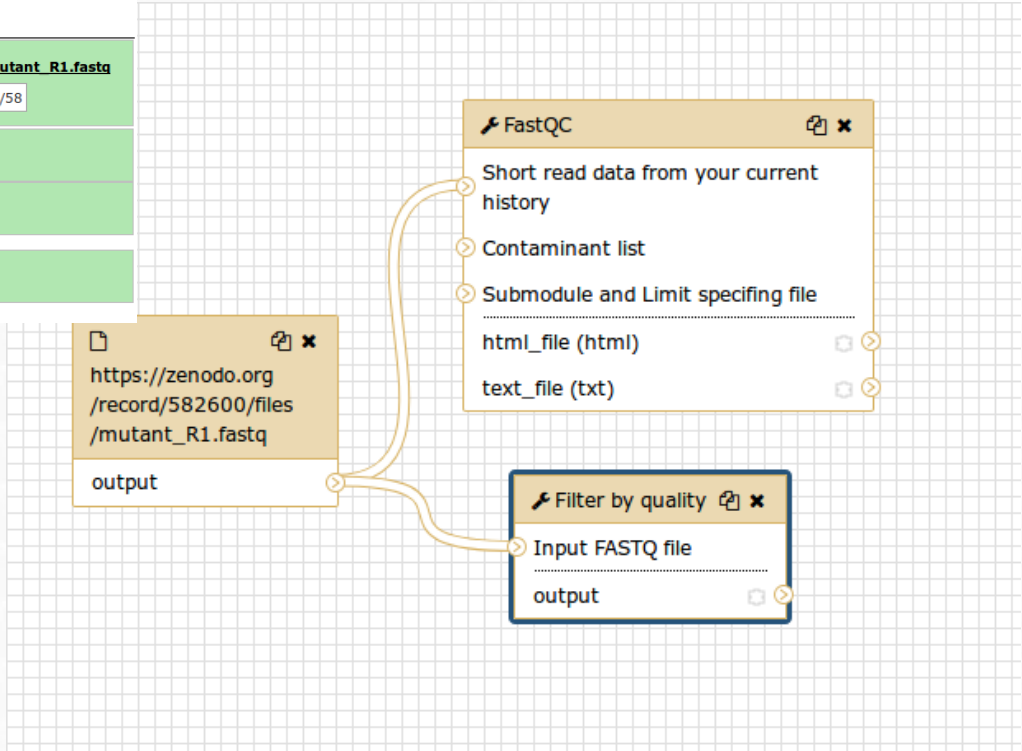


Vall d'Hebron Institut de Recerca

Create workflow from history

- From history options: Export workflow





Vall d'Hebron Institut de Recerca

Galaxy Workflows

- In Galaxy, a Workflow is a defined set of ‘tasks’ that can be stored and executed on demand in an automated fashion.
- A workflow is composed of :
  - any number of tools and dataset operations available on the ‘Tools’ panel (*what to do and with what data*).
  - the relationships among them and their specific run parameters (*how to do it*).
- Very useful:
  - Time saving
  - Less error-prone (no need to set any step and parameter again and again manually)
  - Increased repeatability
  - Increased reproducibility

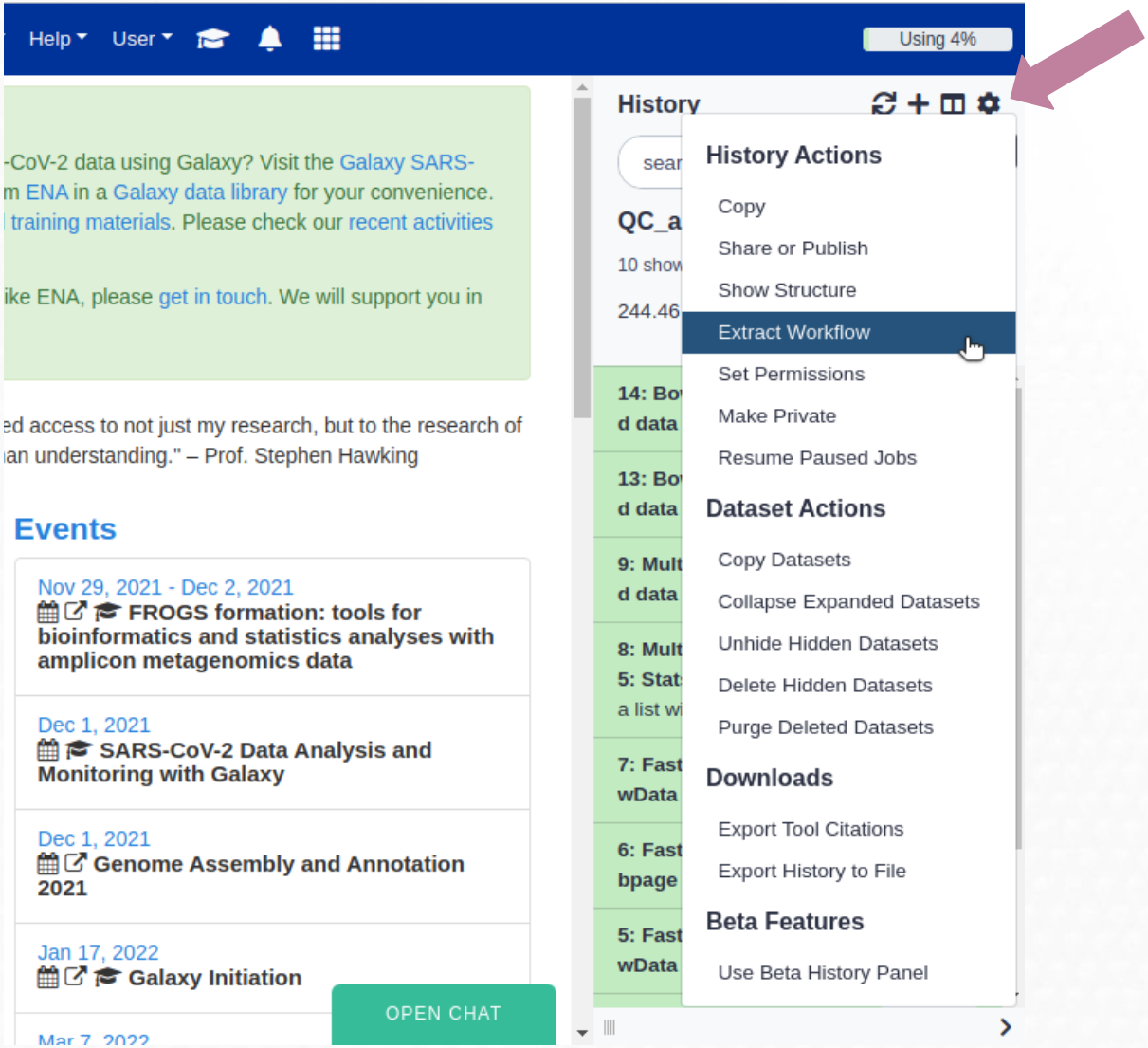
Access your stored workflows:

2. Introduction to Galaxy

Galaxy Workflows

Easy to create:

- **From an existing history**
- Using the integrated visual editor

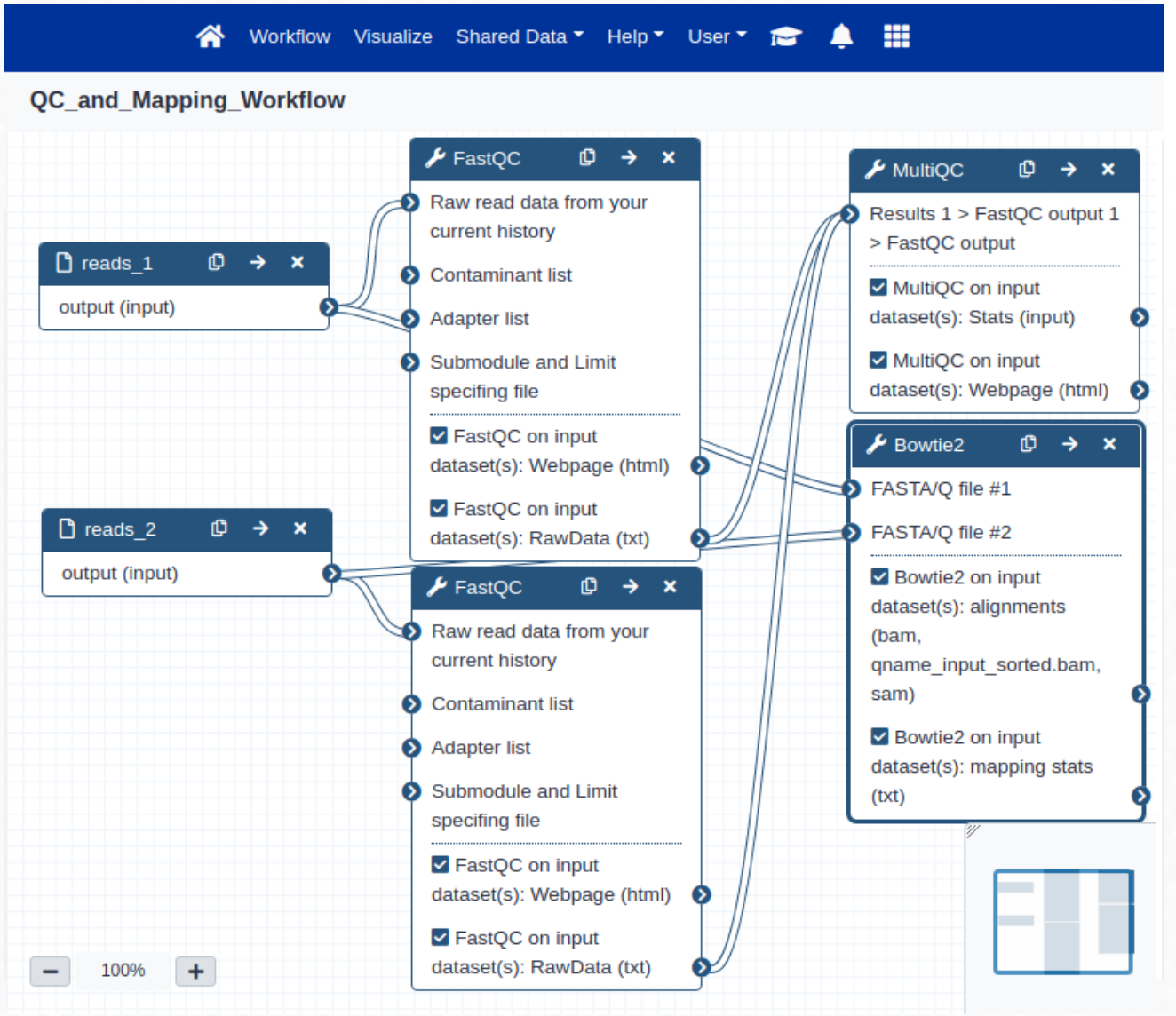


2. Introduction to Galaxy

Galaxy Workflows

Easy to create:

- From an existing history
- **Using the integrated visual editor**



2. Introduction to Galaxy

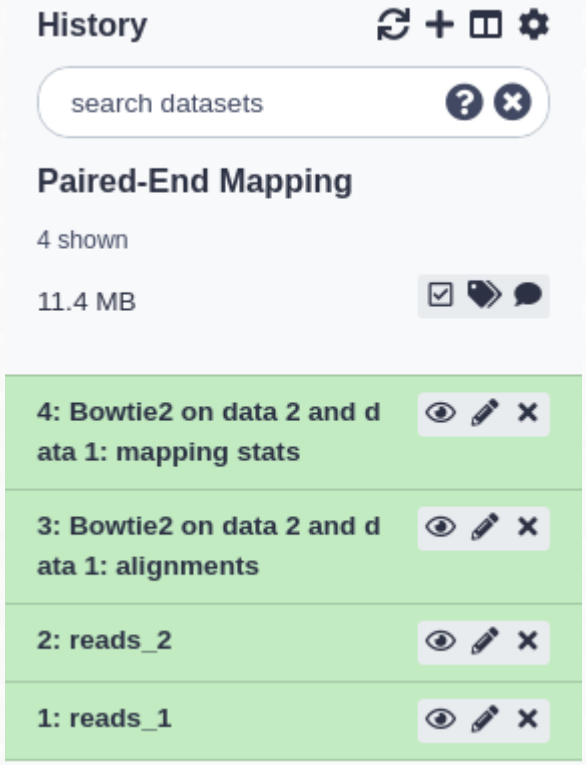
Micro Hands On: Create a Workflow for mapping paired end reads

1. Create a new history and name it ‘Paired-End Mapping’
2. Import the following files containing paired-end reads:
   - https://zenodo.org/record/1324070/files/wt_H3K4me3_read1.fastq.gz
   - https://zenodo.org/record/1324070/files/wt_H3K4me3_read2.fastq.gz
3. Change their names to ‘reads_1’ and ‘reads_2’ respectively
4. On the Tools panel, find a tool named ‘**Bowtie2**’ and click on it. This tool will map our reads to a reference genome.
5. Set the following parameters for Bowtie2 on the central panel:
   - “*Is this single or paired library*”: **Paired-end**
   - “*FASTA/Q file #1*”: **reads_1**
   - “*FASTA/Q file #2*”: **reads_2**
   - “*Do you want to set paired-end options?*”: **No**
   - “*Will you select a reference genome from your history or use a built-in index?*”: **Use a built-in genome index**
   - “*Select reference genome*”: **Mouse (Mus musculus): mm10**
   - “*Select analysis mode*”: **Default setting only**
   - “*Save the bowtie2 mapping statistics to the history*”: **Yes**
6. Click ‘Execute’

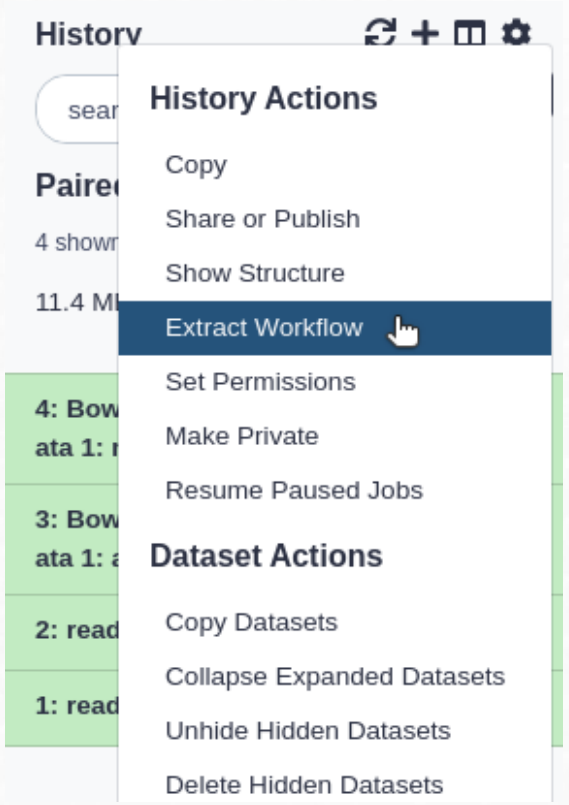
2. Introduction to Galaxy

Micro Hands On: Create a Workflow for mapping paired end reads

After the mapping process is finished, you should have a history like this:

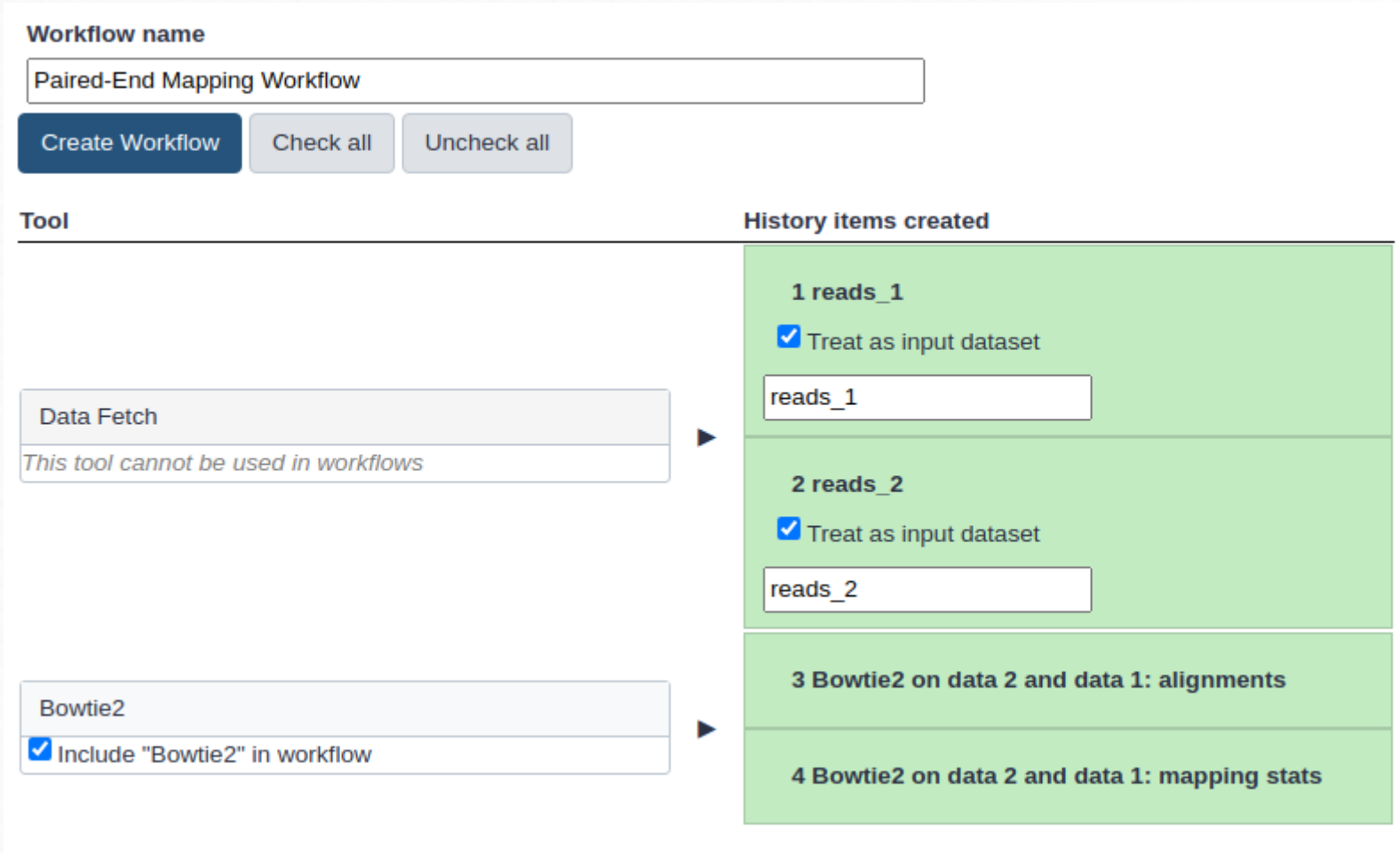


Now we 'extract' a Workflow from this history:



Micro Hands On: Create a Workflow for mapping paired end reads

Change the name to ‘Paired-End Mapping Workflow’ and click ‘Create Workflow’:



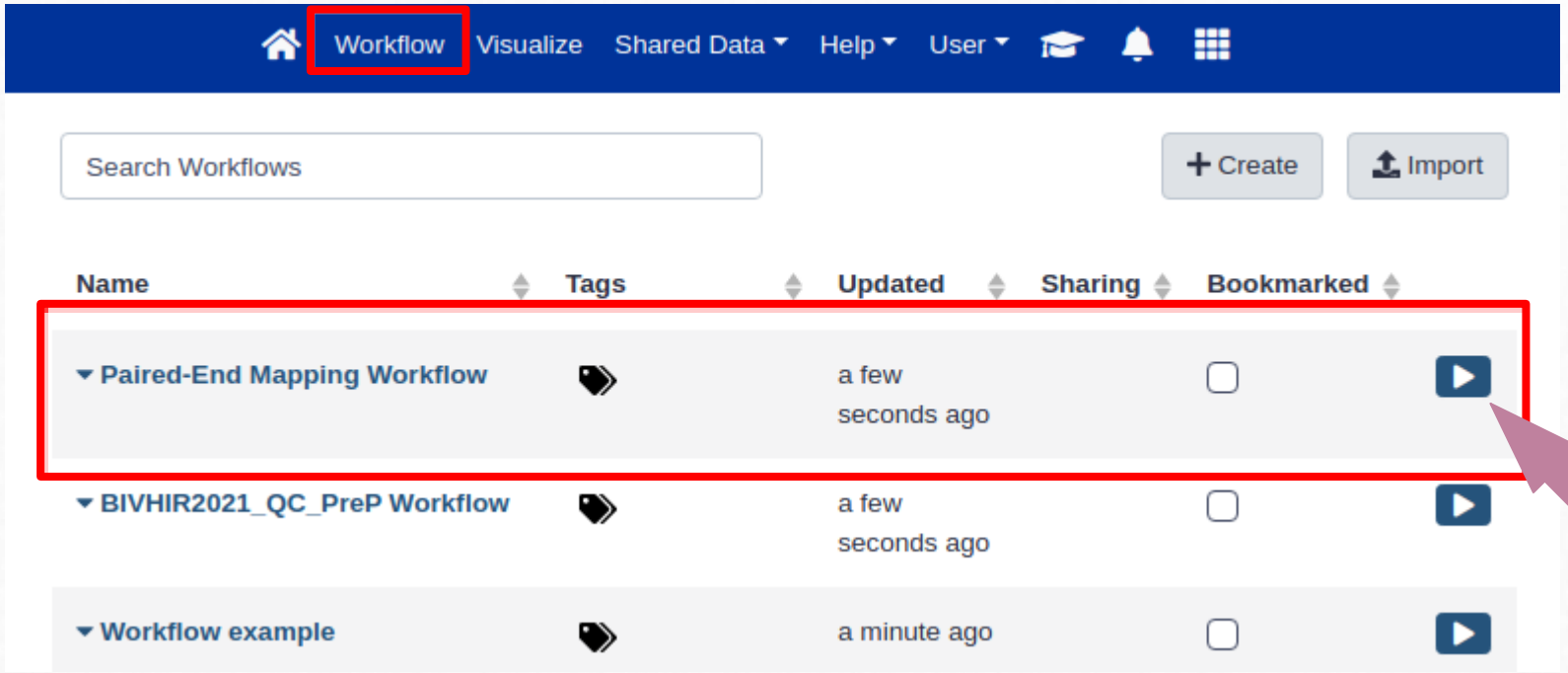
Micro Hands On: Create a Workflow for mapping paired end reads

Now we are going to run this newly created workflow using a diferent set of paired-end reads:

1. Create a new history and name it with a distinctive name
2. Import the following files containing paired-end reads:
   - https://zenodo.org/record/3243160/files/father_R1.fq.gz
   - https://zenodo.org/record/3243160/files/father_R2.fq.gz
3. Rename them to ‘father_R1.fq.gz’ and ‘father_R2.fq.gz’ respectively (if they are not automatically named like that)
4. Go to the ‘Workflow’ section on the top main menu. You should see your newly created Workflow listed.
5. Click on the arrow icon to run the workflow.

2. Introduction to Galaxy

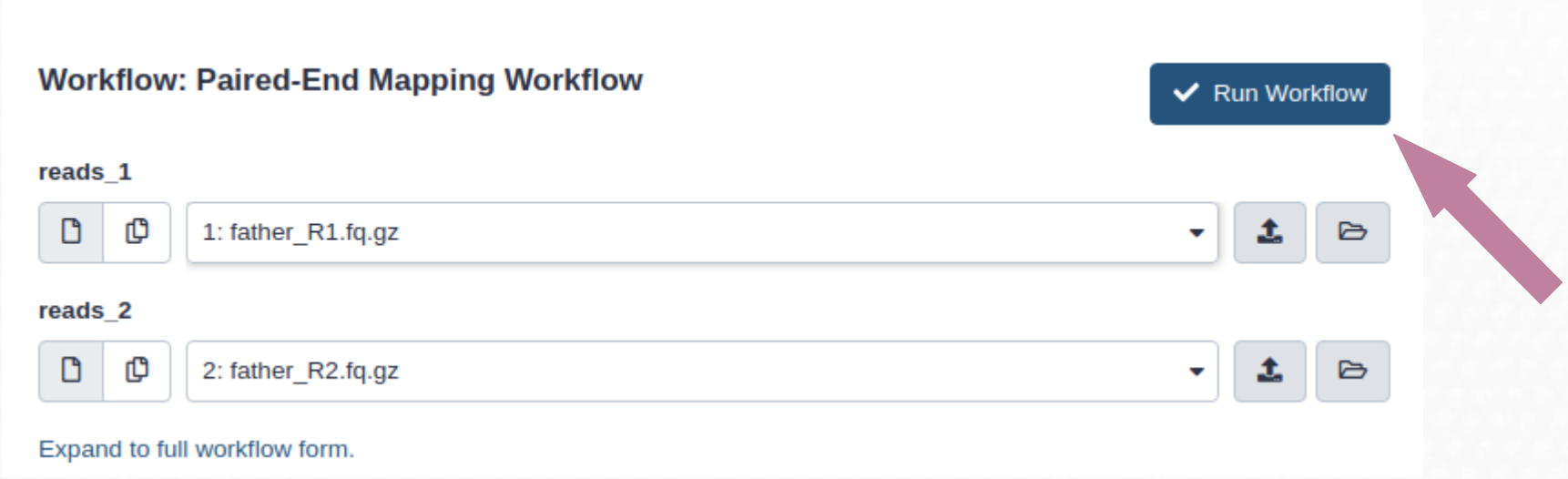
Micro Hands On: Create a Workflow for mapping paired end reads



Vall d'Hebron
Institut de Recerca

Micro Hands On: Create a Workflow for mapping paired end reads

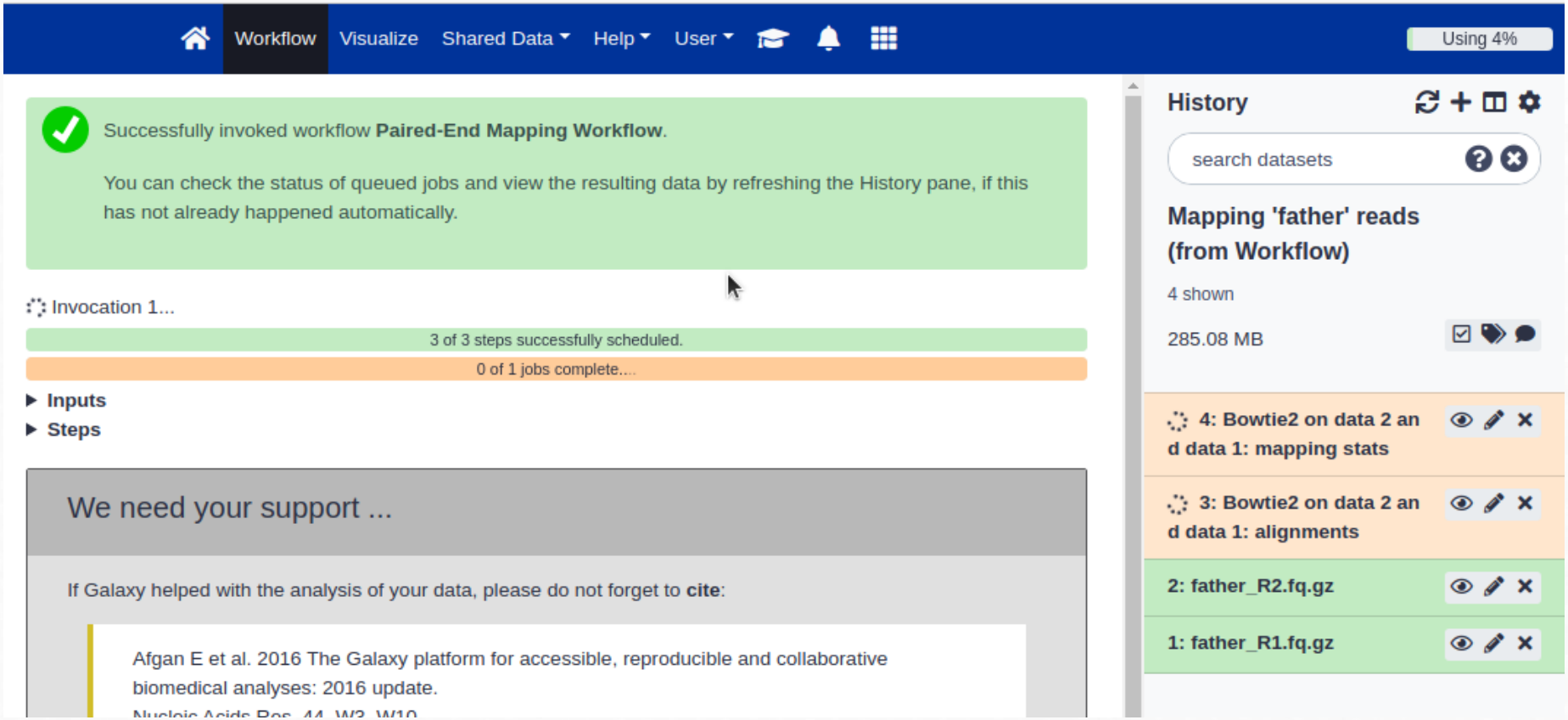
1. Set the inputs for running your workflow to the new reads:
   - "*reads_1*": **father_R1.fq.gz**
   - "*reads_2*": **father_R2.fq.gz**
2. Click 'Run Workflow'



Vall d'Hebron
Institut de Recerca

Micro Hands On: Create a Workflow for mapping paired end reads

Your workflow is running!



2. Introduction to Galaxy

Vall d'Hebron
Institut de Recerca

- https://galaxyproject.org/learn/

Learn Galaxy

There are many approaches to learning how to *use* Galaxy. The most popular is probably to just dive in and use it. Galaxy is simple enough to use that you can do many analyses just by exploring the interface. However, you may miss much of the power this way.

Have you created or know of a resource that is useful for teaching with Galaxy? Then please share it! This will help others and also help get the word out about your resource. Use this Google form to describe your resource. **Also**: consider joining Galaxy Training Network and contributing your tutorial as described here!

Tutorials by Galaxy Training Network

Thanks to a large group of wonderful contributors there is a constantly growing set of tutorials maintained by the Galaxy Training Network. These include:

Introductory Tutorials

- Introduction to Galaxy Analyses
- Data Manipulation
- User Interface and Features

Scientific Analyses

- Assembly
- Computational chemistry
- Ecology
- Epigenetics
- Genome Annotation
- Imaging
- Metabolomics
- Metagenomics
- Proteomics
- Sequence analysis
- Statistics and machine learning
- Transcriptomics
- Variant Analysis