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
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

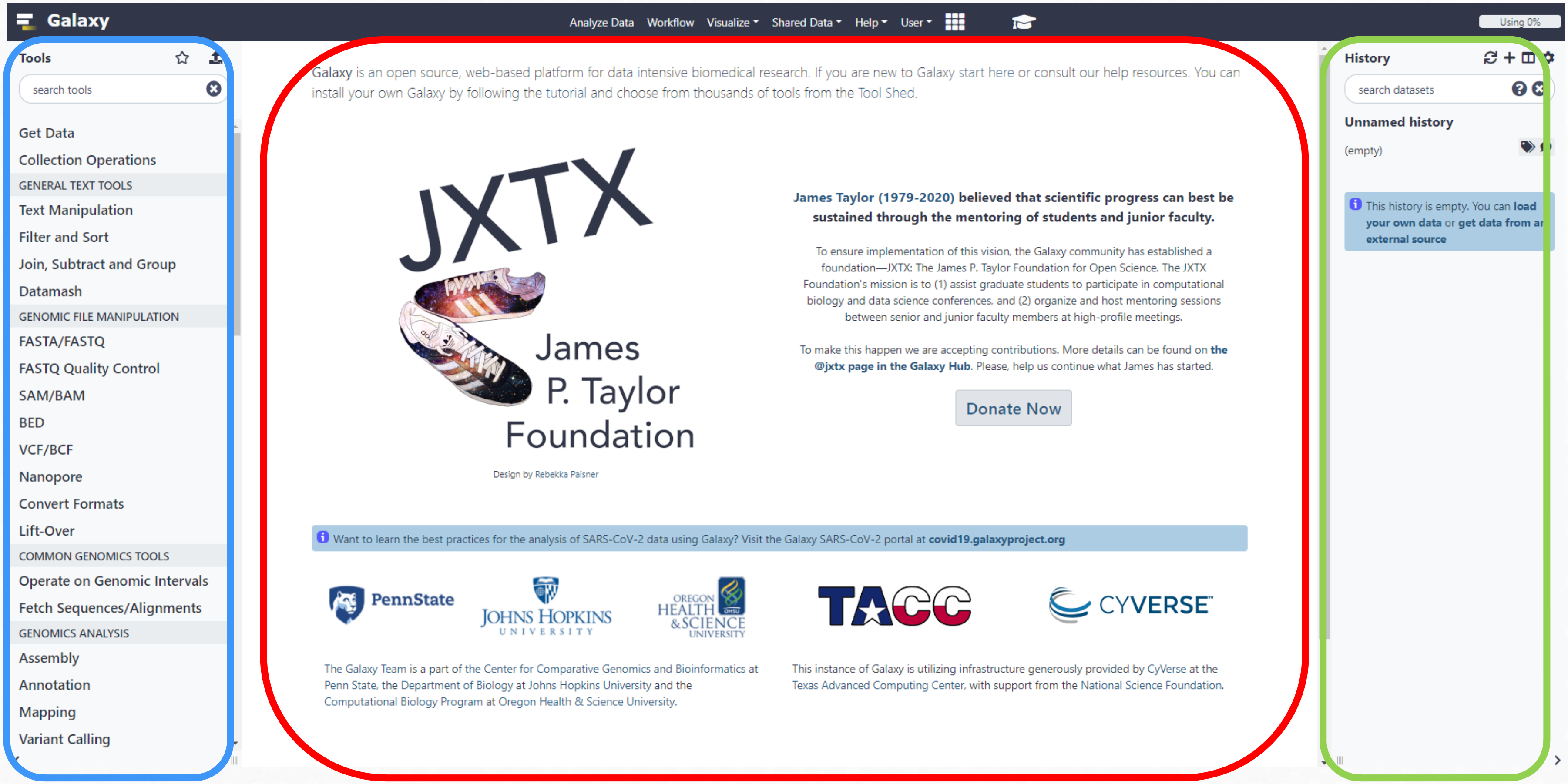
2. Introduction to Galaxy

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



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.

James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.

To ensure implementation of this vision, the Galaxy community has established a foundation—JXTX: The James P. Taylor Foundation for Open Science. The JXTX Foundation's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting contributions. More details can be found on **the @jxtx page in the Galaxy Hub**. Please, help us continue what James has started.

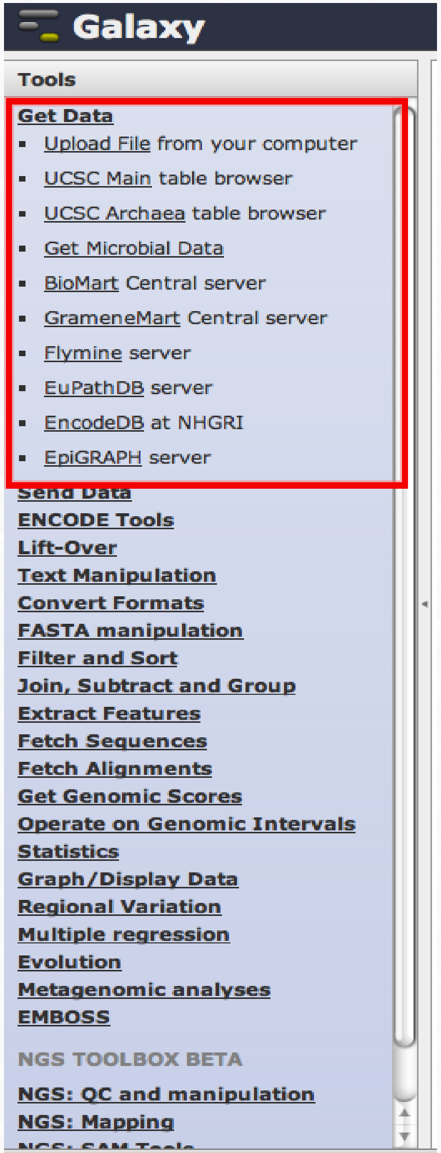
Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at **covid19.galaxyproject.org**

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at Oregon Health & Science University.

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

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

2. Introduction to Galaxy



Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- Get Microbial Data
- BioMart Central server
- GrameneMart Central server
- Flymine server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Evolution
Metagenomic analyses
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools

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

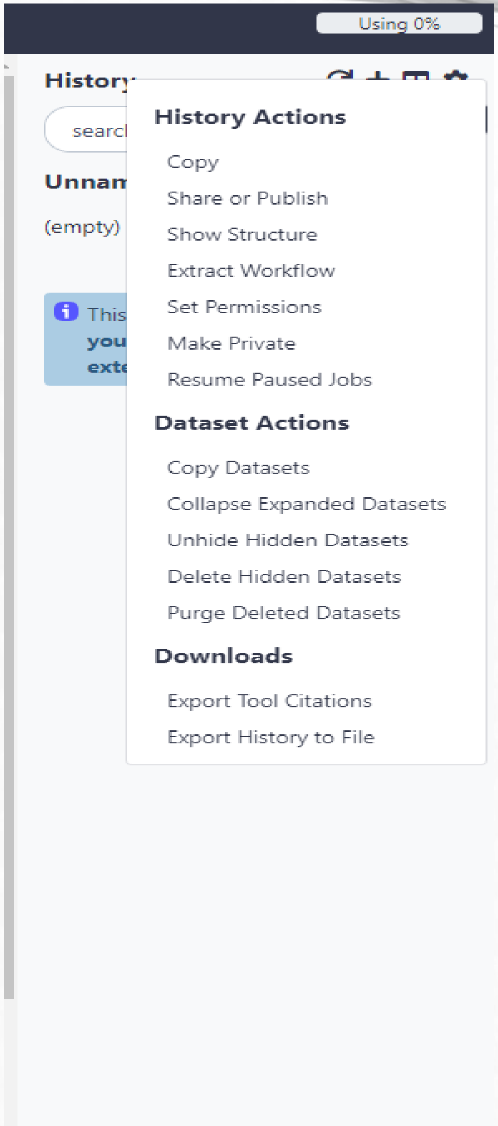
NGS analysis

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

2. Introduction to Galaxy

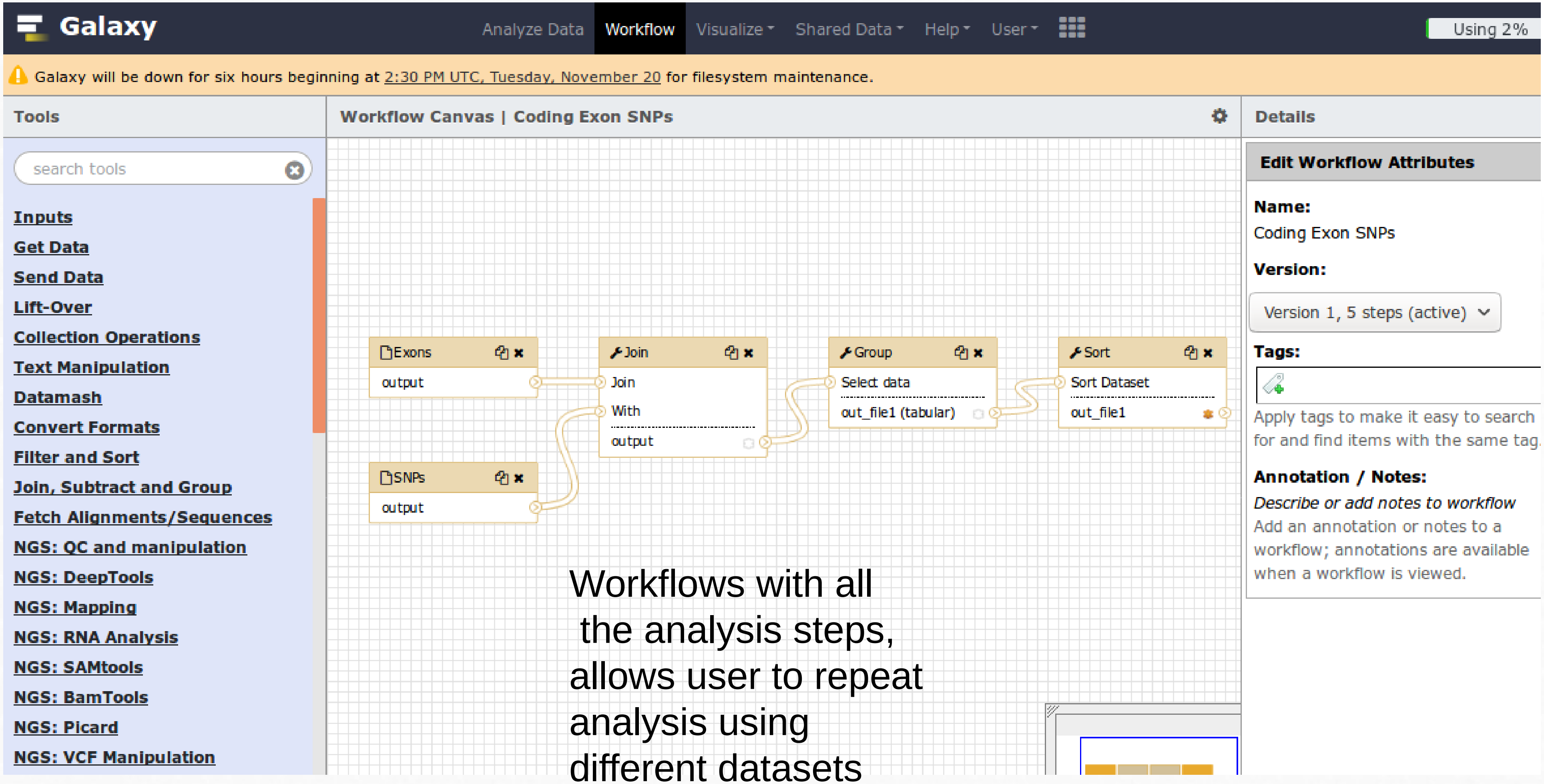
Histories

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

Workflows



Galaxy | Analyze Data | Workflow | Visualize | Shared Data | Help | User | Using 2%

Galaxy will be down for six hours beginning at 2:30 PM UTC, Tuesday, November 20 for filesystem maintenance.

Tools

search tools

Inputs
Get Data
Send Data
Lift-Over
Collection Operations
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
NGS: Picard
NGS: VCF Manipulation

Workflow Canvas | Coding Exon SNPs

Details

Edit Workflow Attributes

Name:
Coding Exon SNPs

Version:
Version 1, 5 steps (active)

Tags:

Apply tags to make it easy to search for and find items with the same tag.

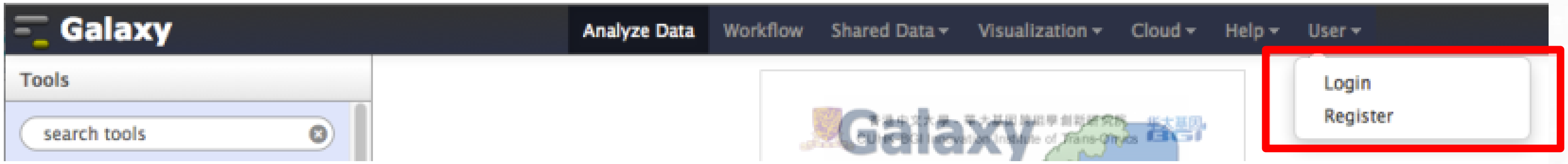
Annotation / Notes:
Describe or add notes to workflow
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

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

2. Introduction to Galaxy

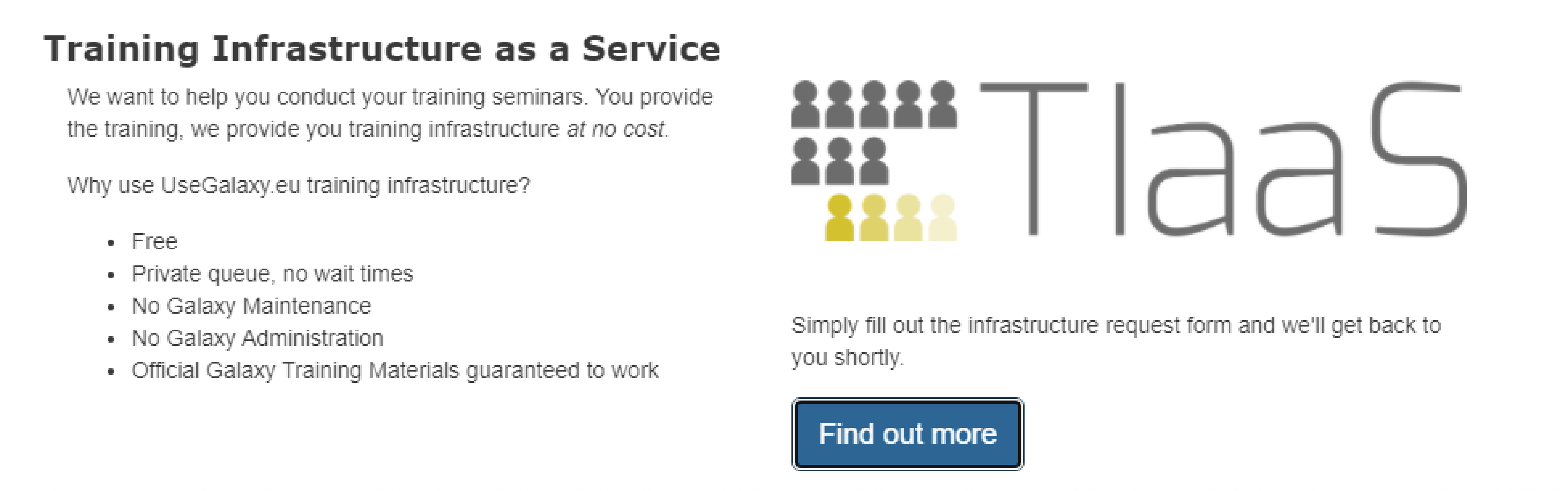
Register for a Galaxy account

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.

After registration in [European Galaxy server]

<https://usegalaxy.eu/join-training/ueb_bi2022>

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)

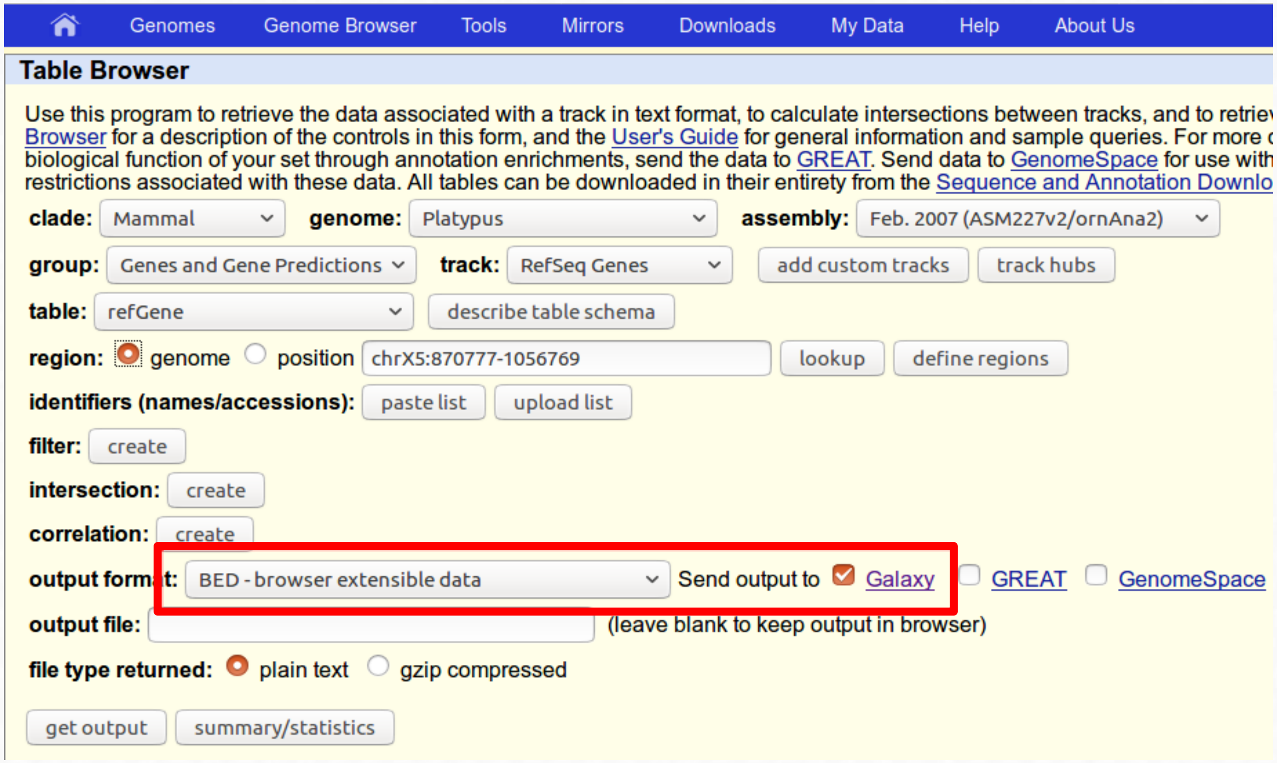


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retriev
Browser for a description of the controls in this form, and the User's Guide for general information and sample queries. For more d
biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with
restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downlo

clade: Mammal **genome:** Platypus **assembly:** Feb. 2007 (ASM227v2/ornAna2)

group: Genes and Gene Predictions **track:** RefSeq Genes add custom tracks track hubs

table: refGene describe table schema

region: ● genome ○ position chrX5:870777-1056769 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

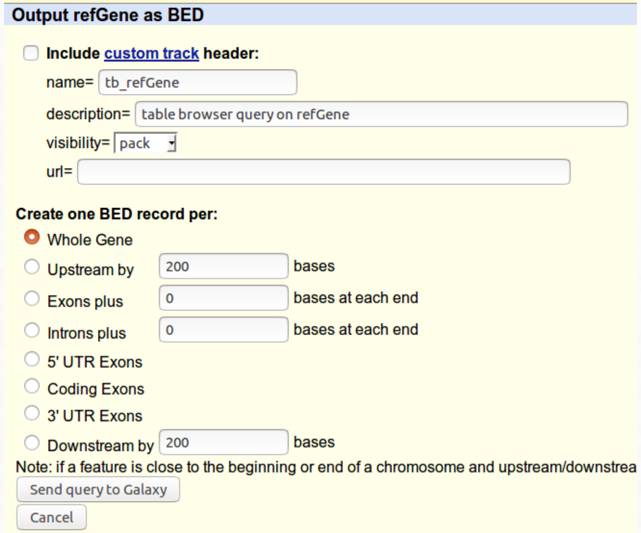
correlation: create

output format: BED - browser extensible data Send output to ☑ Galaxy □ GREAT □ GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics



Output refGene as BED

□ **Include custom track header:**

name= tb_refGene

description= table browser query on refGene

visibility= pack

url=

Create one BED record per:

● Whole Gene
○ Upstream by 200 bases
○ Exons plus 0 bases at each end
○ Introns plus 0 bases at each end
○ 5' UTR Exons
○ Coding Exons
○ 3' UTR Exons
○ Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstrea

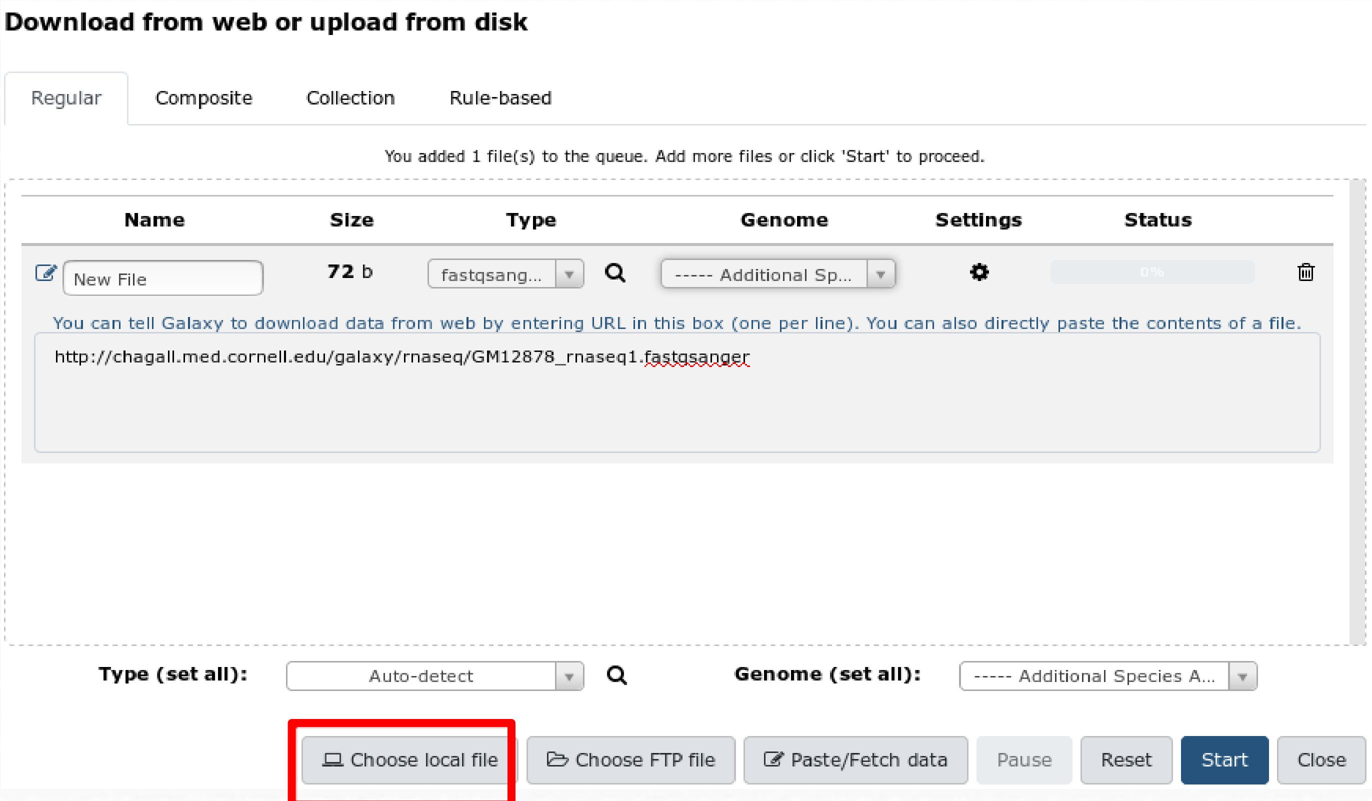
Send query to Galaxy

Cancel

Importing data into Galaxy

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

Get Data > Upload File



Download from web or upload from disk

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

Name	Size	Type	Genome	Settings	Status
New File	72 b	fastqsang...	----- Additional Sp...		0%

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

```
http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878_rnaseq1.fastqsanger
```

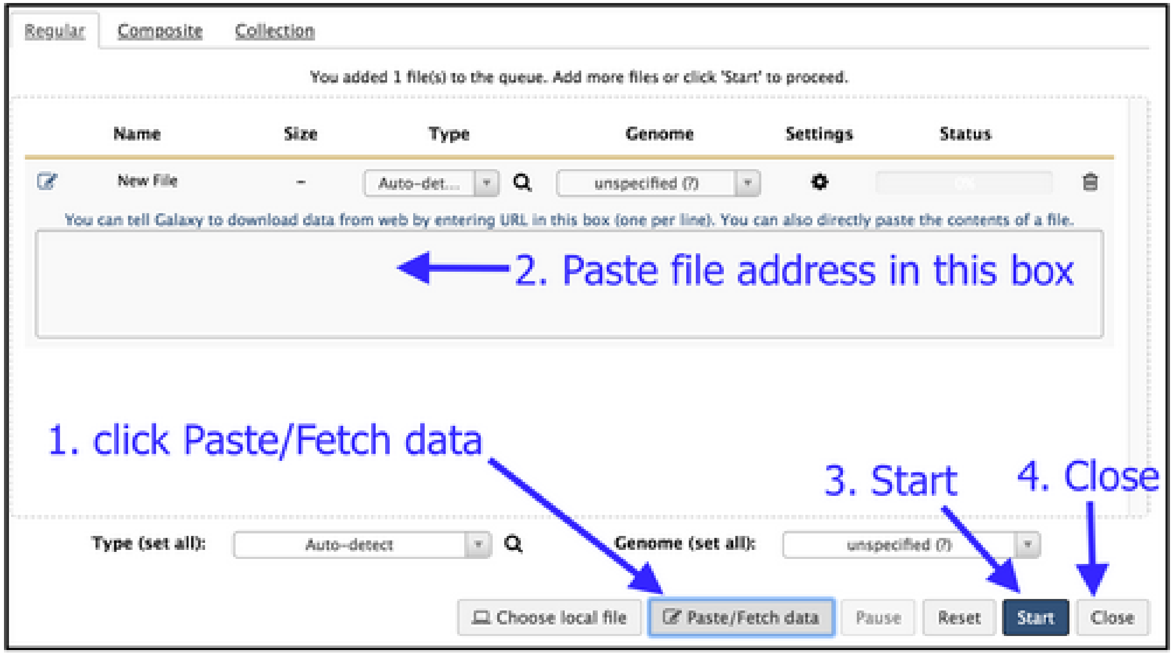
Type (set all): Auto-detect

Genome (set all): ----- Additional Species A...

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



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

Name	Size	Type	Genome	Settings	Status
New File	-	Auto-det...	unspecified (?)		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

2. Paste file address in this box

1. click Paste/Fetch data

3. Start

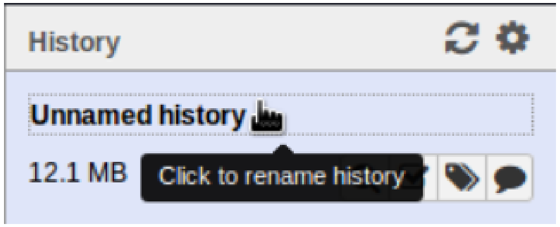
4. Close

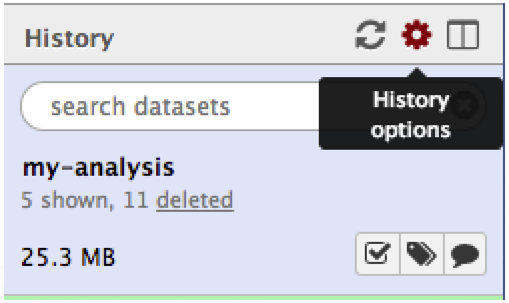
Type (set all): Auto-detect

Genome (set all): unspecified (?)

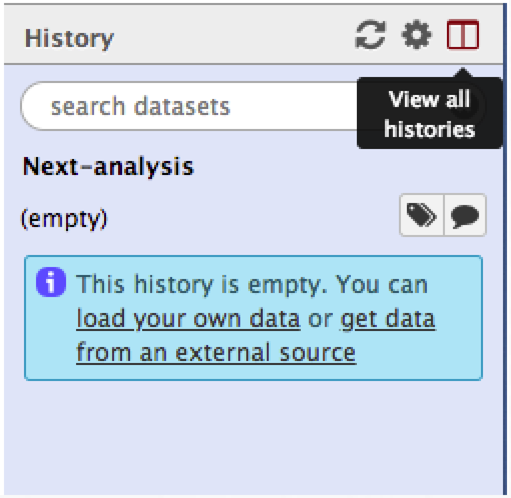
Managing histories

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

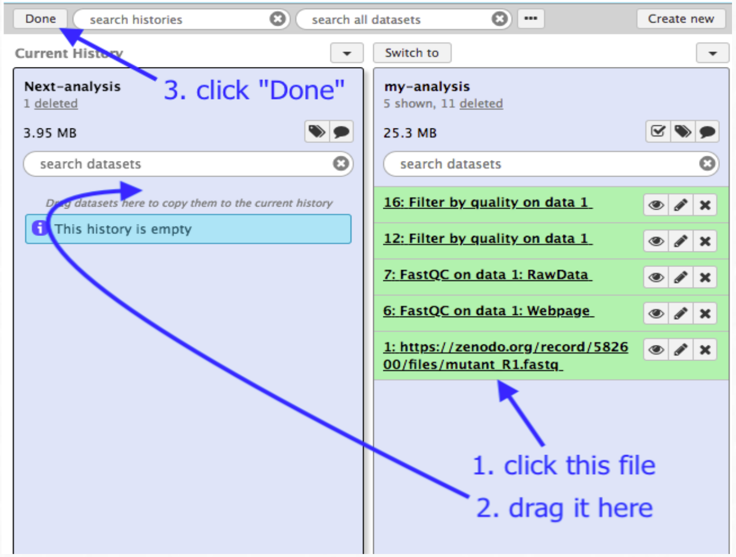




- Manage datasets and histories:
- View all histories
- Drag files between histories (new history must be set to current)



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


Done
search histories
search all datasets
Create new
Current History
Switch to
Next-analysis
1 deleted
3.95 MB
search datasets
my-analysis
5 shown, 11 deleted
25.3 MB
search datasets
16: Filter by quality on data 1
12: Filter by quality on data 1
7: FastQC on data 1: RawData
6: FastQC on data 1: Webpage
1: https://zenodo.org/record/582600/files/mutant_R1.fastq

Drag datasets here to copy them to the current history

This history is empty

1. click this file

2. drag it here

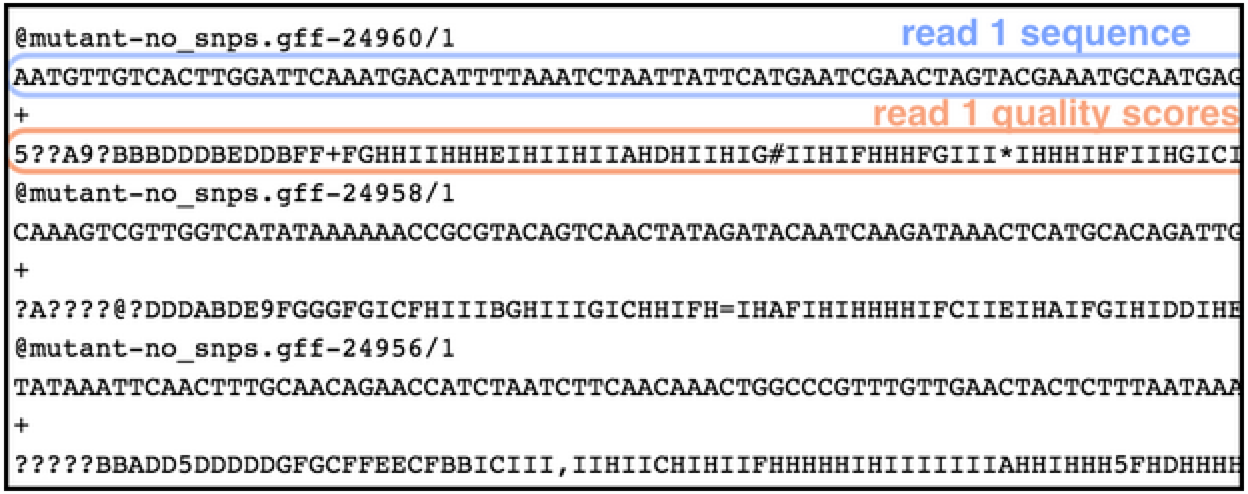
3. click "Done"

2. Introduction to Galaxy

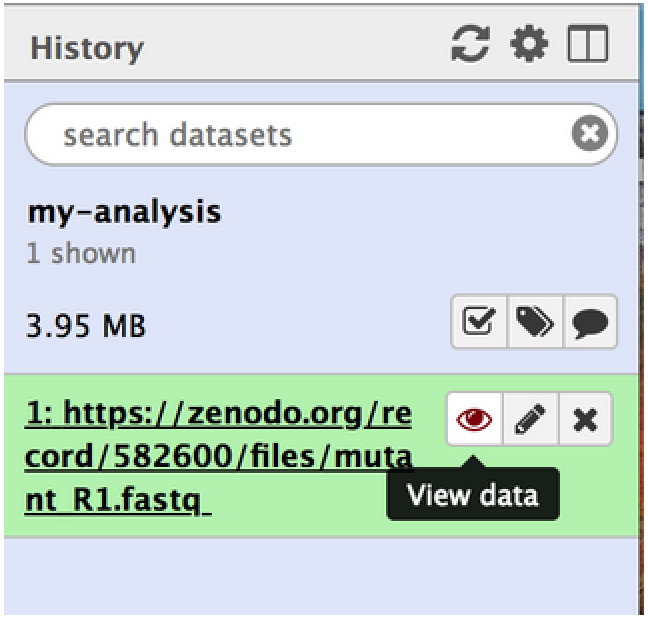
Visualizing

- 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:


read 1 sequence
read 1 quality scores

```
@mutant-no_snps.gff-24960/1
AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAG
+
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
```



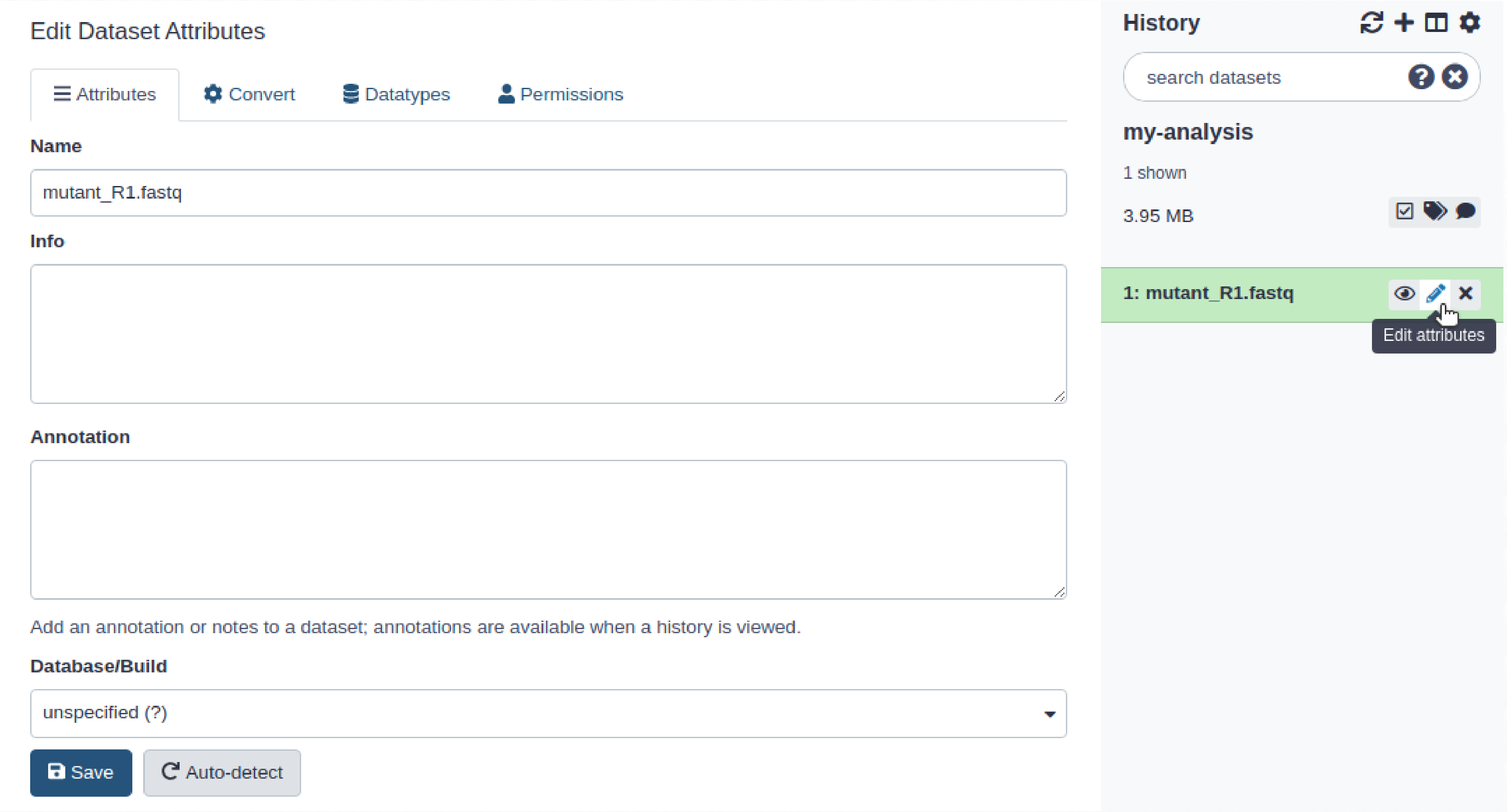
my-analysis
1 shown

3.95 MB

1: https://zenodo.org/record/582600/files/mutant_R1.fastq_

Editing basic attributes

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



Edit Dataset Attributes

Name

mutant_R1.fastq

Info

Annotation

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

Database/Build

unspecified (?)

my-analysis

1 shown

3.95 MB

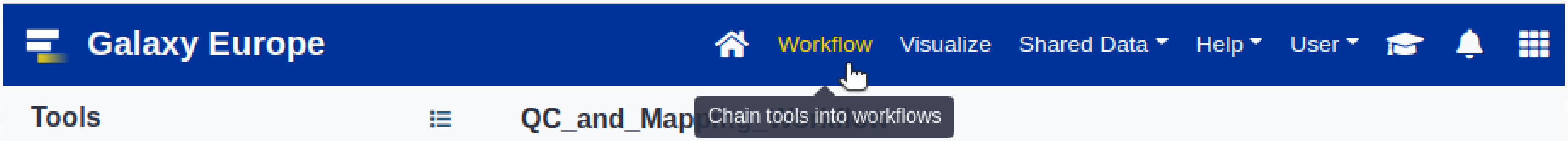
1: mutant_R1.fastq

2. Introduction to Galaxy

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:

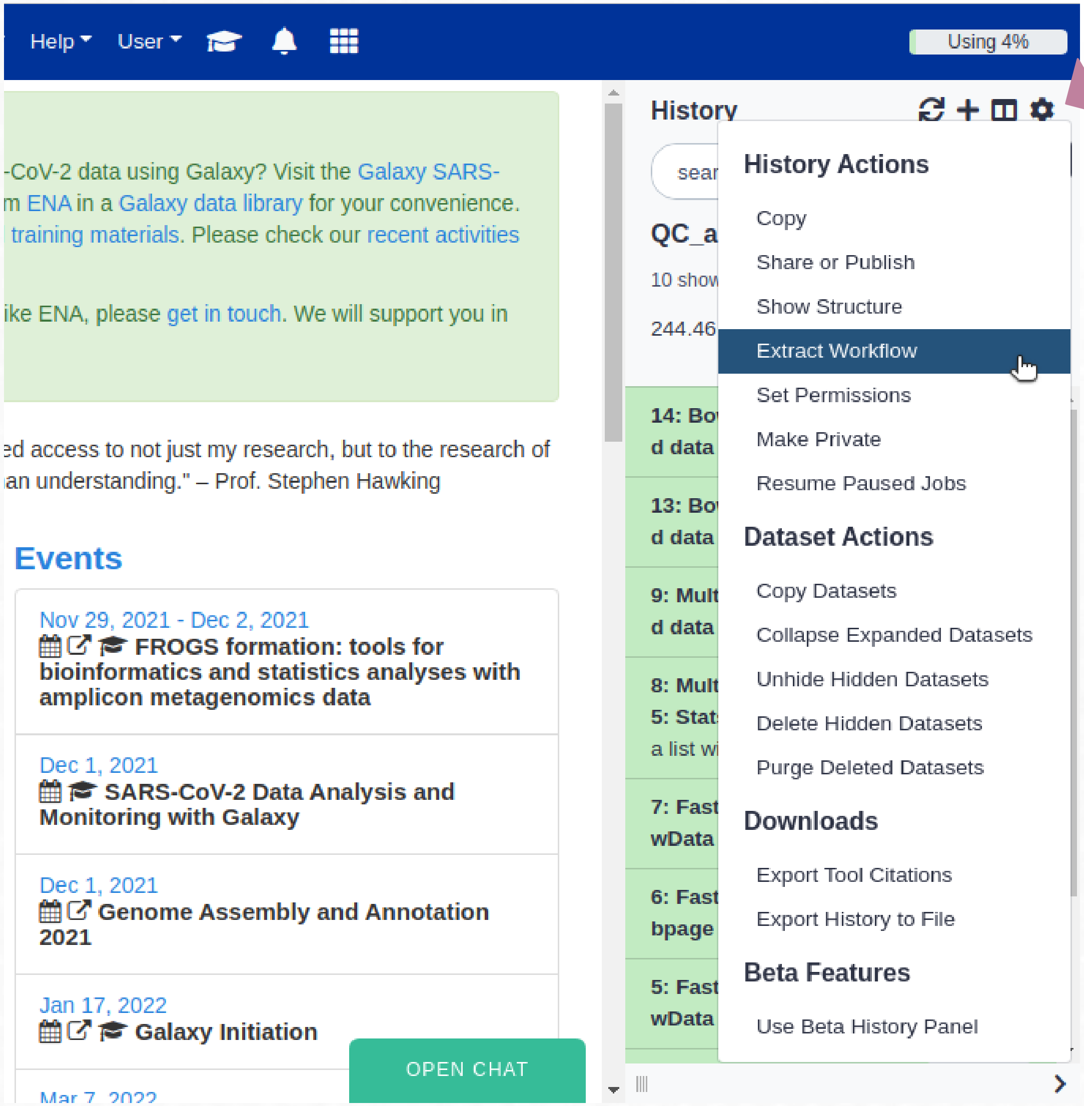


2. Introduction to Galaxy

Galaxy Workflows

Easy to create:

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



-CoV-2 data using Galaxy? Visit the Galaxy SARS-
m ENA in a Galaxy data library for your convenience.
training materials. Please check our recent activities

ike ENA, please get in touch. We will support you in

ed access to not just my research, but to the research of
an understanding." – Prof. Stephen Hawking

Events

Nov 29, 2021 - Dec 2, 2021
FROGS formation: tools for bioinformatics and statistics analyses with amplicon metagenomics data

Dec 1, 2021
SARS-CoV-2 Data Analysis and Monitoring with Galaxy

Dec 1, 2021
Genome Assembly and Annotation 2021

Jan 17, 2022
Galaxy Initiation

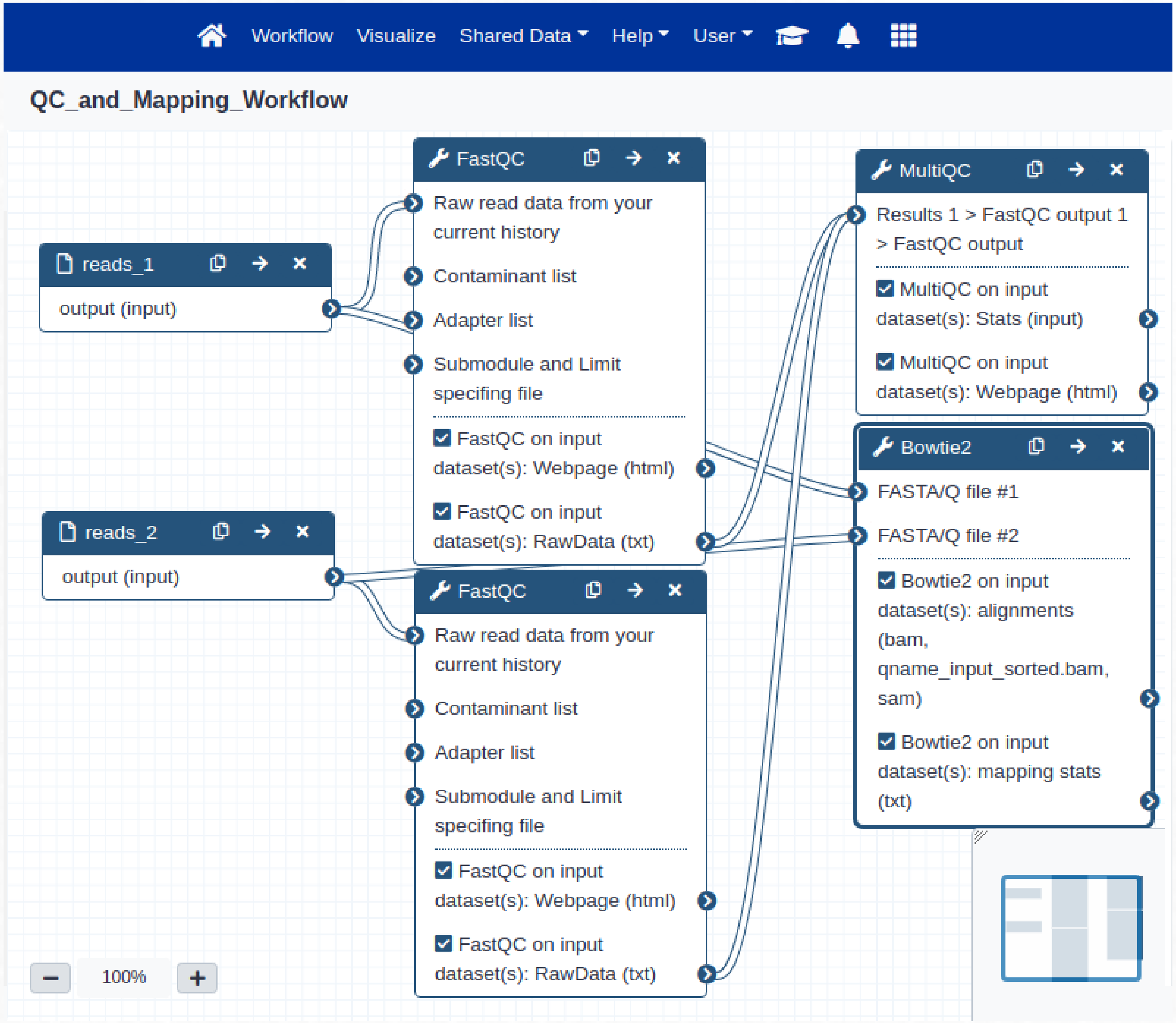
Mar 7, 2022

2. Introduction to Galaxy

Galaxy Workflows

Easy to create:

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


Workflow
Visualize
Shared Data
Help
User
QC_and_Mapping_Workflow
reads_1
output (input)
reads_2
output (input)
FastQC
Raw read data from your current history
Contaminant list
Adapter list
Submodule and Limit specifing file
☑ FastQC on input dataset(s): Webpage (html)
☑ FastQC on input dataset(s): RawData (txt)
FastQC
Raw read data from your current history
Contaminant list
Adapter list
Submodule and Limit specifing file
☑ FastQC on input dataset(s): Webpage (html)
☑ FastQC on input dataset(s): RawData (txt)
MultiQC
Results 1 > FastQC output 1 > FastQC output
☑ MultiQC on input dataset(s): Stats (input)
☑ MultiQC on input dataset(s): Webpage (html)
Bowtie2
FASTA/Q file #1
FASTA/Q file #2
☑ Bowtie2 on input dataset(s): alignments (bam, qname_input_sorted.bam, sam)
☑ Bowtie2 on input dataset(s): mapping stats (txt)
100%

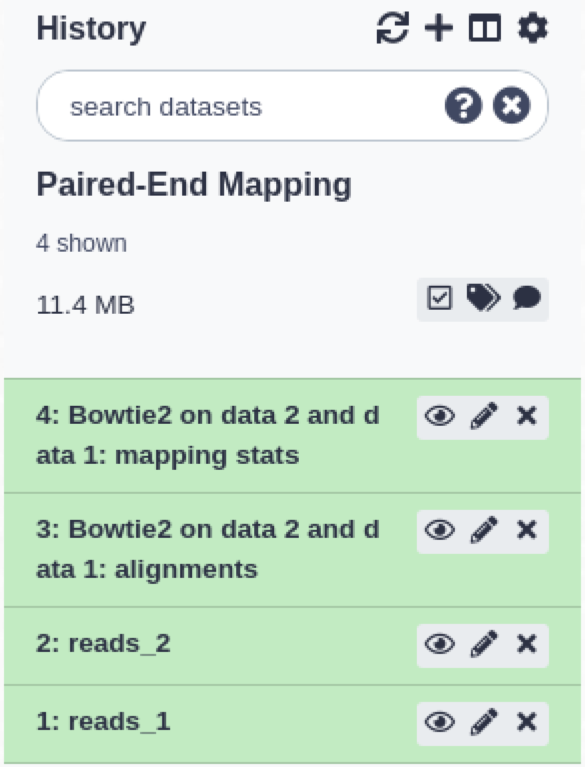
2. Introduction to Galaxy

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’

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

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



History

search datasets

Paired-End Mapping

4 shown

11.4 MB

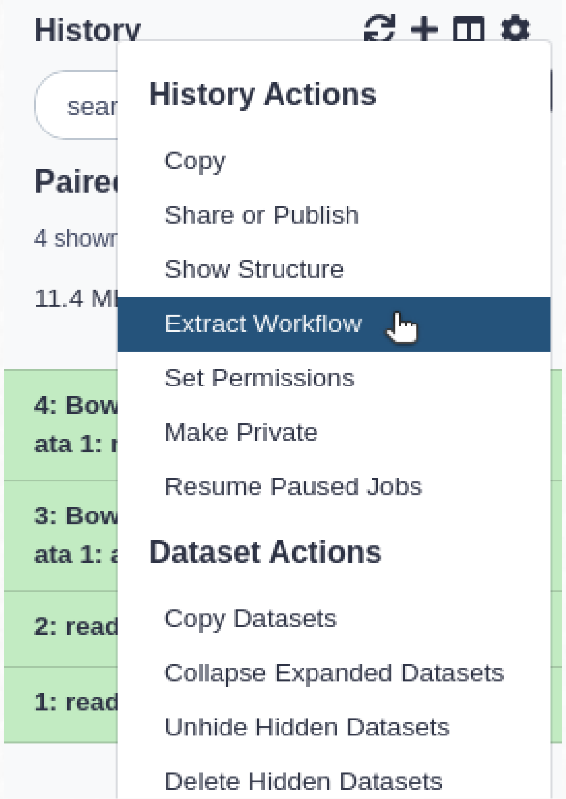
4: Bowtie2 on data 2 and data 1: mapping stats

3: Bowtie2 on data 2 and data 1: alignments

2: reads_2

1: reads_1

1. Now we ‘extract’ a Workflow from this history:



History

sear

Paired

4 shown

11.4 MB

4: Bow
ata 1: r

3: Bow
ata 1: a

2: read

1: read

History Actions

Copy

Share or Publish

Show Structure

Extract Workflow

Set Permissions

Make Private

Resume Paused Jobs

Dataset Actions

Copy Datasets

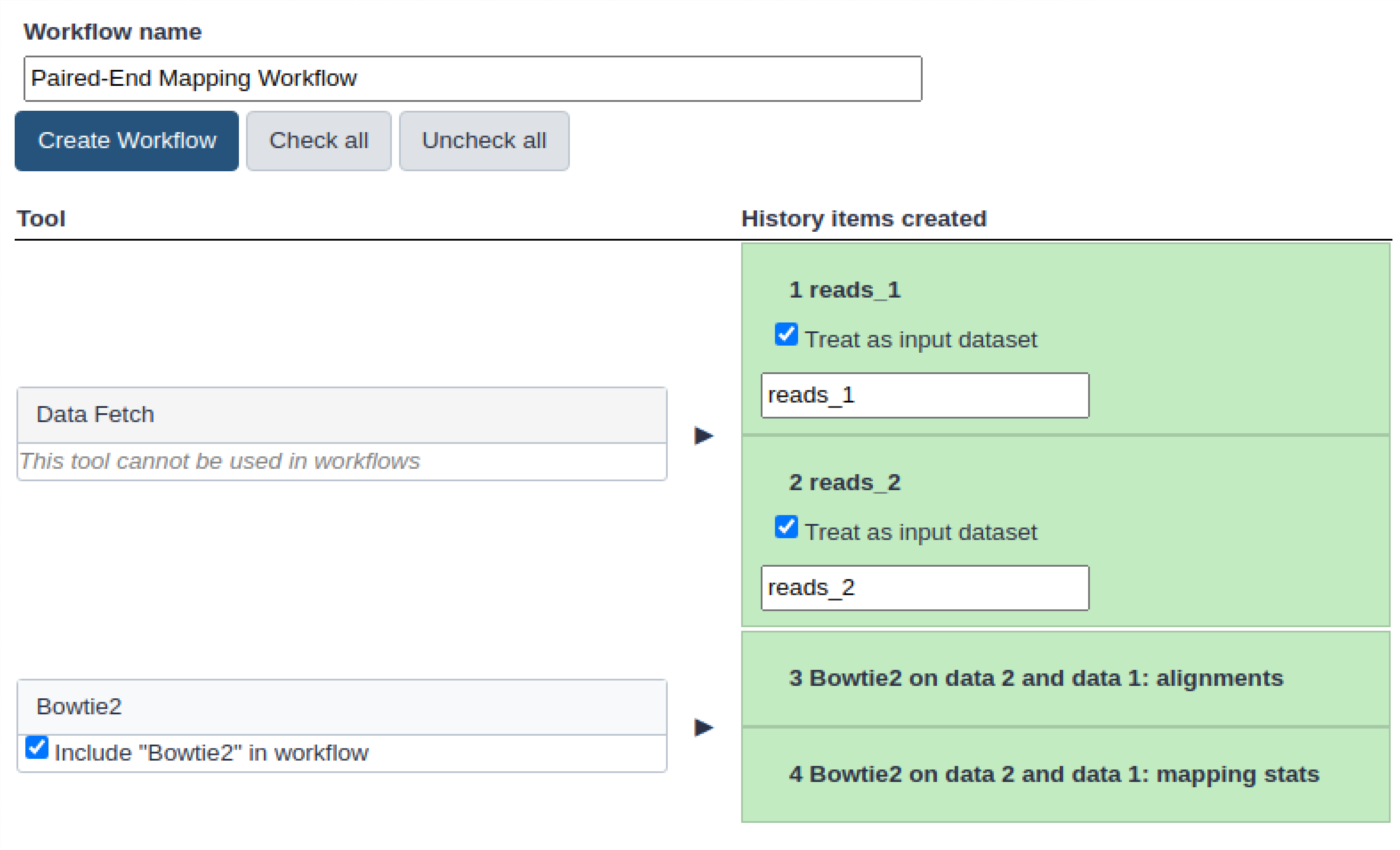
Collapse Expanded Datasets

Unhide Hidden Datasets

Delete Hidden Datasets

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

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



Workflow name

Paired-End Mapping Workflow

Tool	History items created
Data Fetch *This tool cannot be used in workflows*	**1 reads_1** ☑ Treat as input dataset reads_1 **2 reads_2** ☑ Treat as input dataset reads_2
Bowtie2 ☑ Include "Bowtie2" in workflow	**3 Bowtie2 on data 2 and data 1: alignments** **4 Bowtie2 on data 2 and data 1: mapping stats**

2. Introduction to Galaxy

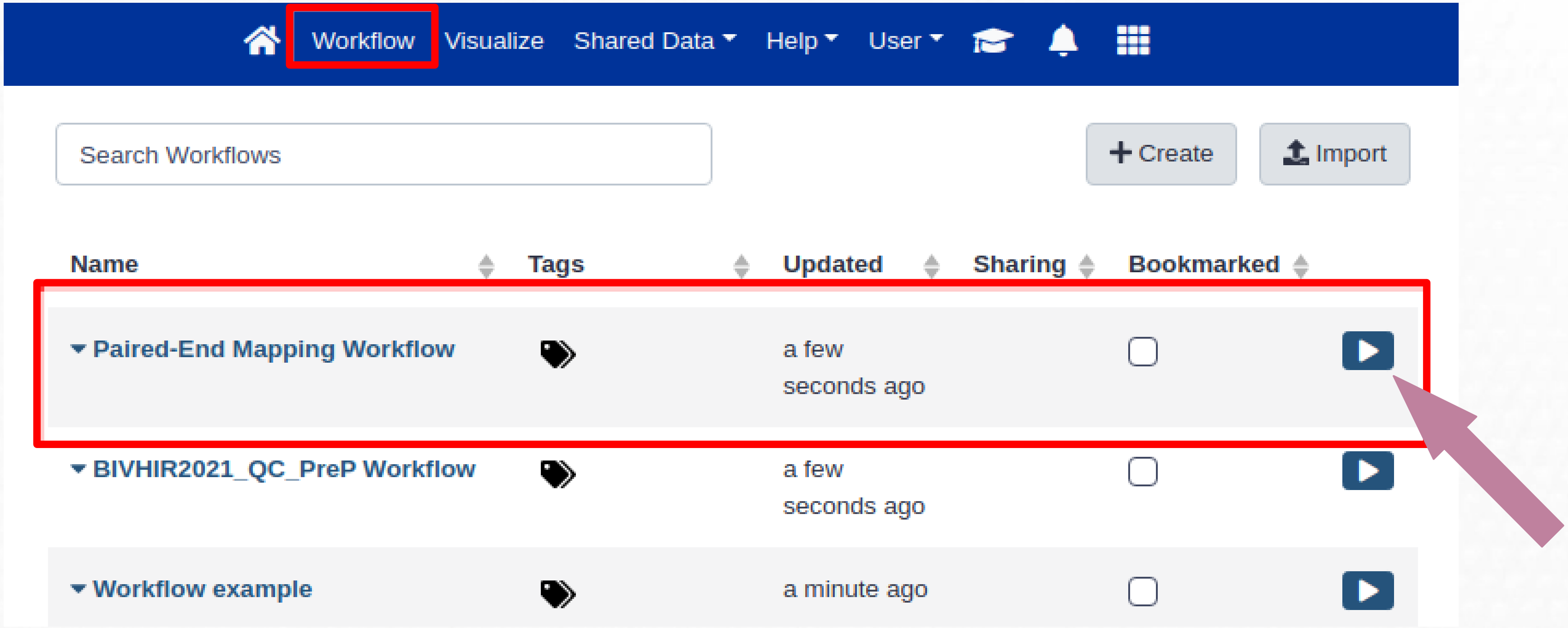
Micro Hands On: Create a Workflow for mapping paired end reads

1. 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

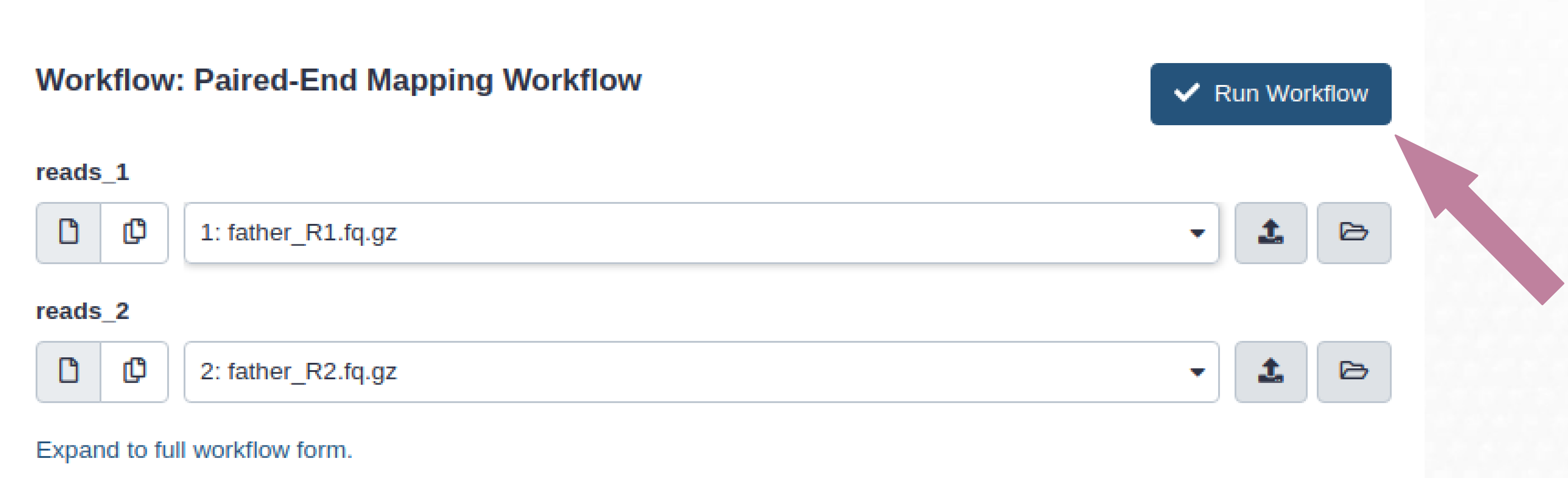
Micro Hands On: Create a Workflow for mapping paired end reads



Name	Tags	Updated	Sharing	Bookmarked	
▾ Paired-End Mapping Workflow		a few seconds ago		□	▶
▾ BIVHIR2021_QC_PreP Workflow		a few seconds ago		□	▶
▾ Workflow example		a minute ago		□	▶

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’



Workflow: Paired-End Mapping Workflow

reads_1: 1: father_R1.fq.gz

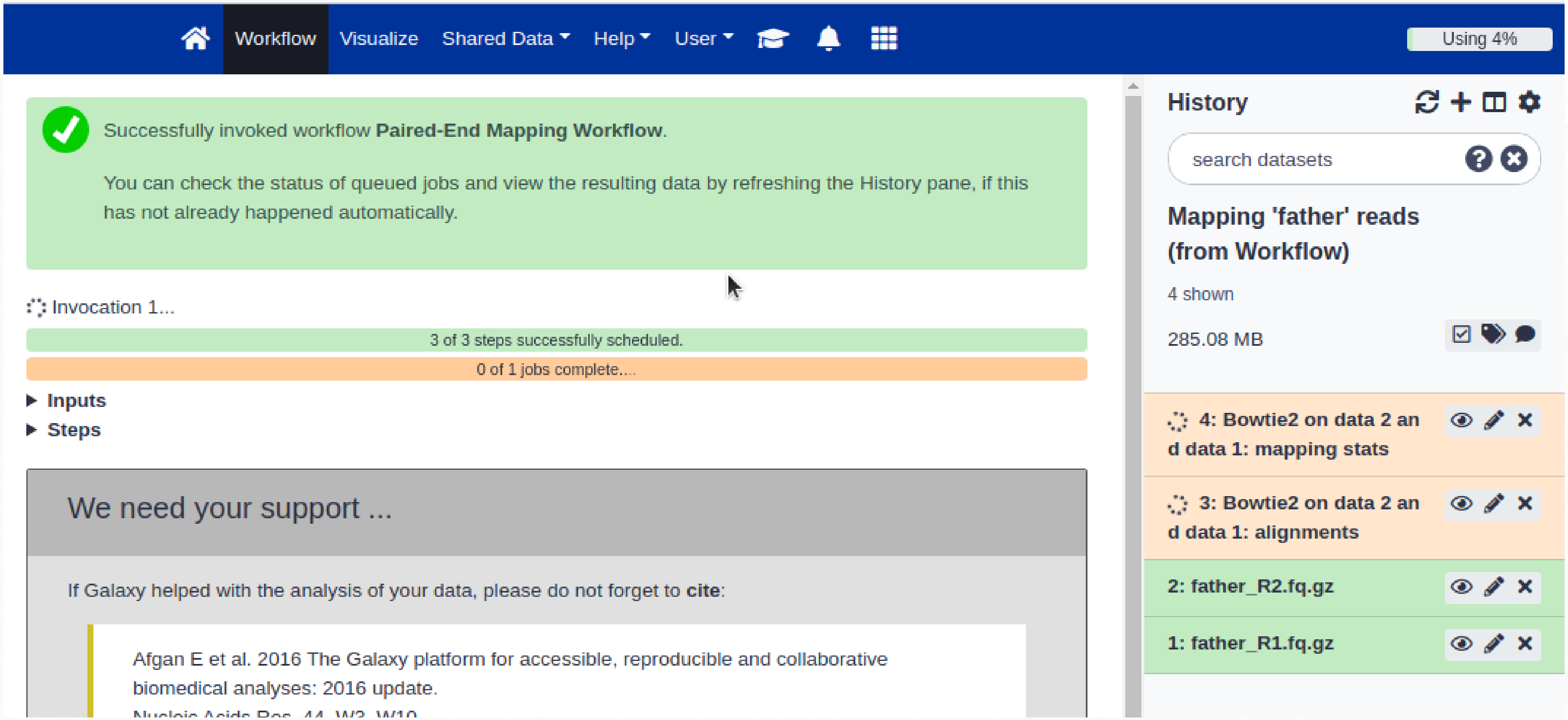
reads_2: 2: father_R2.fq.gz

Expand to full workflow form.

2. Introduction to Galaxy

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

Your workflow is running!



Successfully invoked workflow **Paired-End Mapping Workflow**.

You can check the status of queued jobs and view the resulting data by refreshing the History pane, if this has not already happened automatically.

Invocation 1...

3 of 3 steps successfully scheduled.

0 of 1 jobs complete....

▶ **Inputs**

▶ **Steps**

We need your support ...

If Galaxy helped with the analysis of your data, please do not forget to **cite**:

Afgan E et al. 2016 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44, W3–W10

History

search datasets

Mapping 'father' reads (from Workflow)

4 shown

285.08 MB

4: Bowtie2 on data 2 and data 1: mapping stats

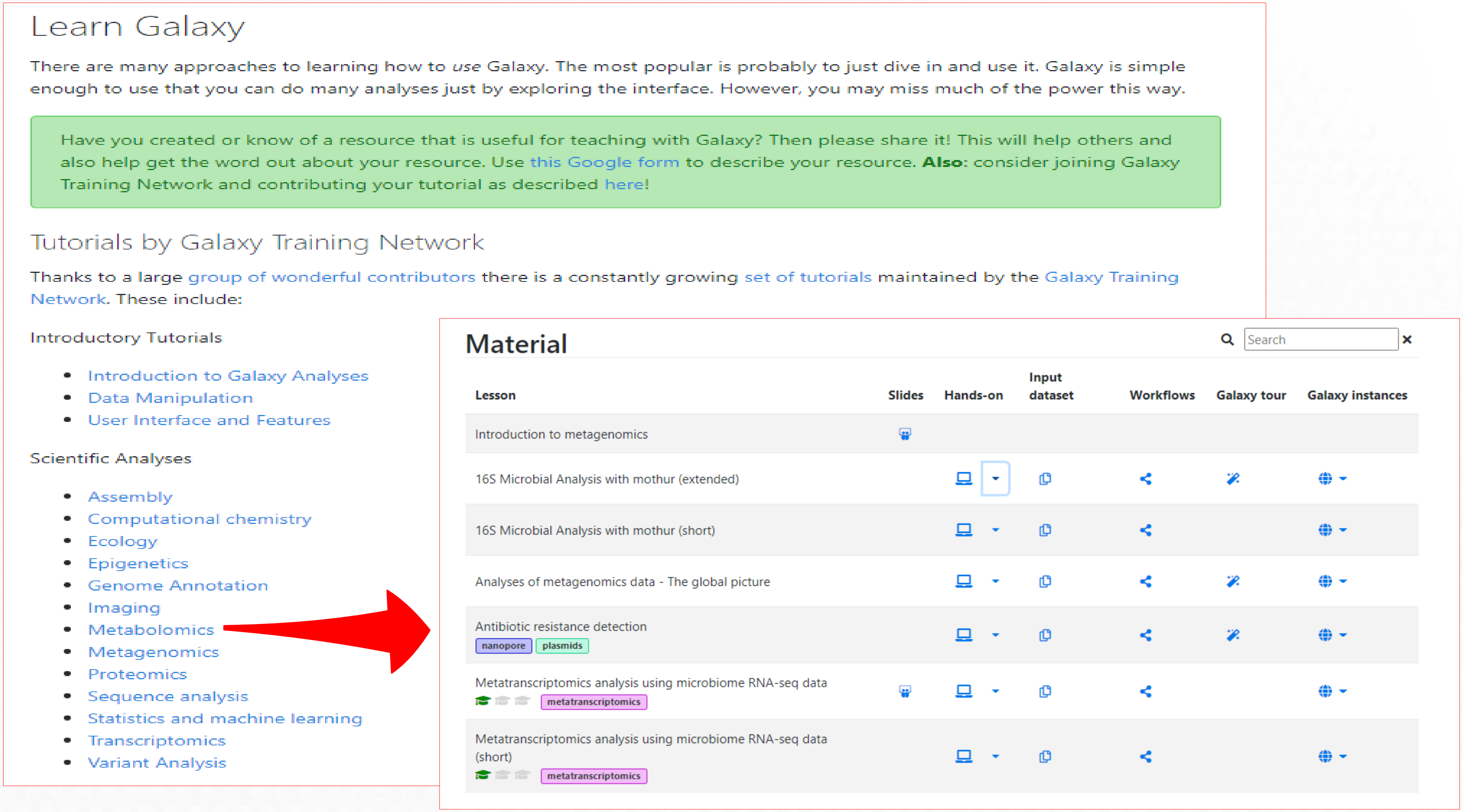
3: Bowtie2 on data 2 and data 1: alignments

2: father_R2.fq.gz

1: father_R1.fq.gz

2. Introduction to Galaxy

- 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

Material

Search ✕

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to metagenomics	✓					
16S Microbial Analysis with mothur (extended)		✓ ▾	✓	✓	✓	✓ ▾
16S Microbial Analysis with mothur (short)		✓ ▾	✓	✓		✓ ▾
Analyses of metagenomics data - The global picture		✓ ▾	✓	✓	✓	✓ ▾
Antibiotic resistance detection nanopore plasmids		✓ ▾	✓	✓	✓	✓ ▾
Metatranscriptomics analysis using microbiome RNA-seq data metatranscriptomics	✓	✓ ▾	✓	✓		✓ ▾
Metatranscriptomics analysis using microbiome RNA-seq data (short) metatranscriptomics		✓ ▾	✓	✓		✓ ▾