Biogenouest

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

ABiMS

27/05/2024

Galaxy

Initiation

Loraine Guéguen
Erwan Corre

Credits to Gildas Le Corguillé, Galaxy Training Network

v3.2

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

cnrs

Setup

- Connect to local computer
- Get slides and datasets: https://abims.sb-roscoff.fr/training/courses
- Login on Galaxy France: https://usegalaxy.fr
- Visit https://usegalaxy.fr/join-training/init2024/
  - ✓ reserved computing resources

Schedule

- Schedule
  - 09:00 - 12:00
  - 13:30 - 17:00

- Short round table

Learning objectives

- Learning objectives:
  - Familiarize yourself with the basics of Galaxy
  - Learn how to import data
  - Learn how to run tools
  - Learn how histories work
  - Learn how to create a workflow
  - Learn how to share your work
  - Understand and master dataset collections

Program

- Introduction
- Data import
- Tools
- Dataset
- History
- Workflow
- Share

Sequence files manipulation

INTRODUCTION / PROBLEMATIC

Introduction / problematic

- In biomedical research, high-throughput technologies produce **large datasets**.
- How to perform analyses of these data **without bioinformatics skills** ?
    - Assemble transcript sequences de novo
    - Determine the gene expression
    - Build a phylogenetic tree
    - Predict subcellular targeting for proteins
    - Identify and quantify metabolites detected by LC-MS
    - …

Introduction / problematic

Command line tools represent the majority of scientific tools

+ good parameters completeness
+ can be executed on high performance computers
+ automation, workflows ...
– minimum linux knowledge is required
– cruel lack of ergonomics

Introduction / problematic

Command line tools represent the majority of scientific tools

+ good parameters completeness
+ can be executed on high performance computers
+ automation, workflows ...
– minimum linux knowledge is required
– cruel lack of ergonomics

```
login@sbr4-p023:~$ ssh -Y bioinfo.sb-roscoff.fr

[...]

login@slurm0:~$ cd /shared/projects/my_project/finalresult

login@slurm0:~$ head -n 2 ../input/query.fasta

>Ec-00_000010.1 Ferric reductase, NAD binding (657) ;mRNA; r:150-6731
MGSQCWQGVLNHGNHFHDIYMVCPDDPTQWCNQFSFESGGPTPNQYRLRMIGIFMGLVCSSHFAIILVPVSRD

login@slurm0:finalresult$ module load blast/2.9.0

login@slurm0:finalresult$ srun --mem 100G --cpus-per-task 4 blastp -query
../input/query.fasta -db /shared/bank/uniref90/current/blast/uniref90 -out query.out -outfmt
6 -evalue 1e-6
```

Introduction / problematic

Command line tools represent the majority of scientific tools

+ good parameters completeness
+ can be executed on high performance computers
+ automation, workflows ...
– minimum linux knowledge is required
– cruel lack of ergonomics

```
login@sbr4-p023:~$ ssh -Y bioinfo.sb-roscoff.fr

[...]

login@slurm0:~$ cd /shared/projects/my_project/finalresult

login@slurm0:~$ head -
```

```
>Ec-00_000010.1 Ferric                                    r:15
MGSQCWQGVLNHGNHFHDIYMV                                   CSSHF          D

login@slurm0:finalresu

login@slurm0:finalresult$ srun --mem 100G --cpus-per-task 4 blastp -query
../input/query.fasta -db /shared/bank/uniref90/current/blast/uniref90 -out query.out -outfmt
6 -evalue 1e-6
```

Introduction / problematic

MR. GEEK

```
login@sbr4-p023:~$ ssh -Y bioinfo.sb-roscoff.fr

[...]

login@slurm0:~$ cd /shared/projects/my_project/finalresult

login@slurm0:~$ head -n 2 ../input/query.fasta

>Ec-00_000010.1 Ferric reductase, NAD binding (657) ;mRNA;
r:150-6731
MGSQCWQGVLNHGNHFHDIYMVCPDDPTQWCNQFSFESGGPTPNQYRLRMIGIFMGLVCSSHF
IILVPVSRD

login@slurm0:finalresult$ module load blast/2.9.0

login@slurm0:finalresult$ srun --mem 100G --cpus-per-task 4
blastp -query ../input/query.fasta -db
/shared/bank/uniref90/current/blast/uniref90 -out query.out
-outfmt 6 -evalue 1e-6
```

MR. HAPPY
by Roger Hargreaves

INTRODUCTION / GALAXY

Introduction / Galaxy

Introduction to Galaxy

Andrea Bagnacani | Bérénice Batut | Saskia Hiltemann | Anne Pajon
Nicola Soranzo | Helena Rasche | Christopher Barnett | Michele Maroni
Anne Fouilloux | Nadia Goué | Olha Nahorna | Dave Clements
Add Contributions!

Updated: Oct 25, 2021
Or view the plain-text slides without JS
Tip: press P to view the presenter notes

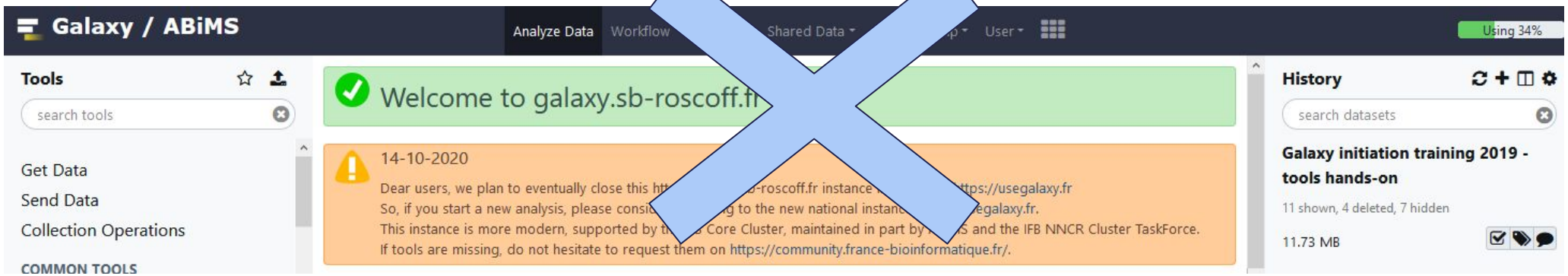
https://training.galaxyproject.org/training-material/topics/introduction/slides/introduction.html

usegalaxy.fr

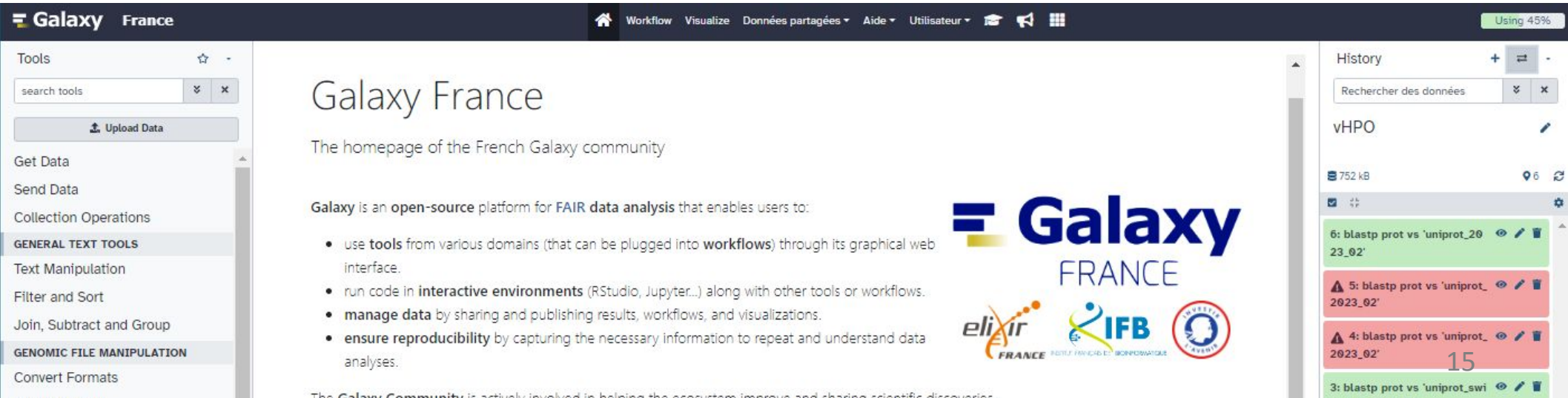
GALAXY FRANCE

Galaxy ABiMS --> Galaxy France

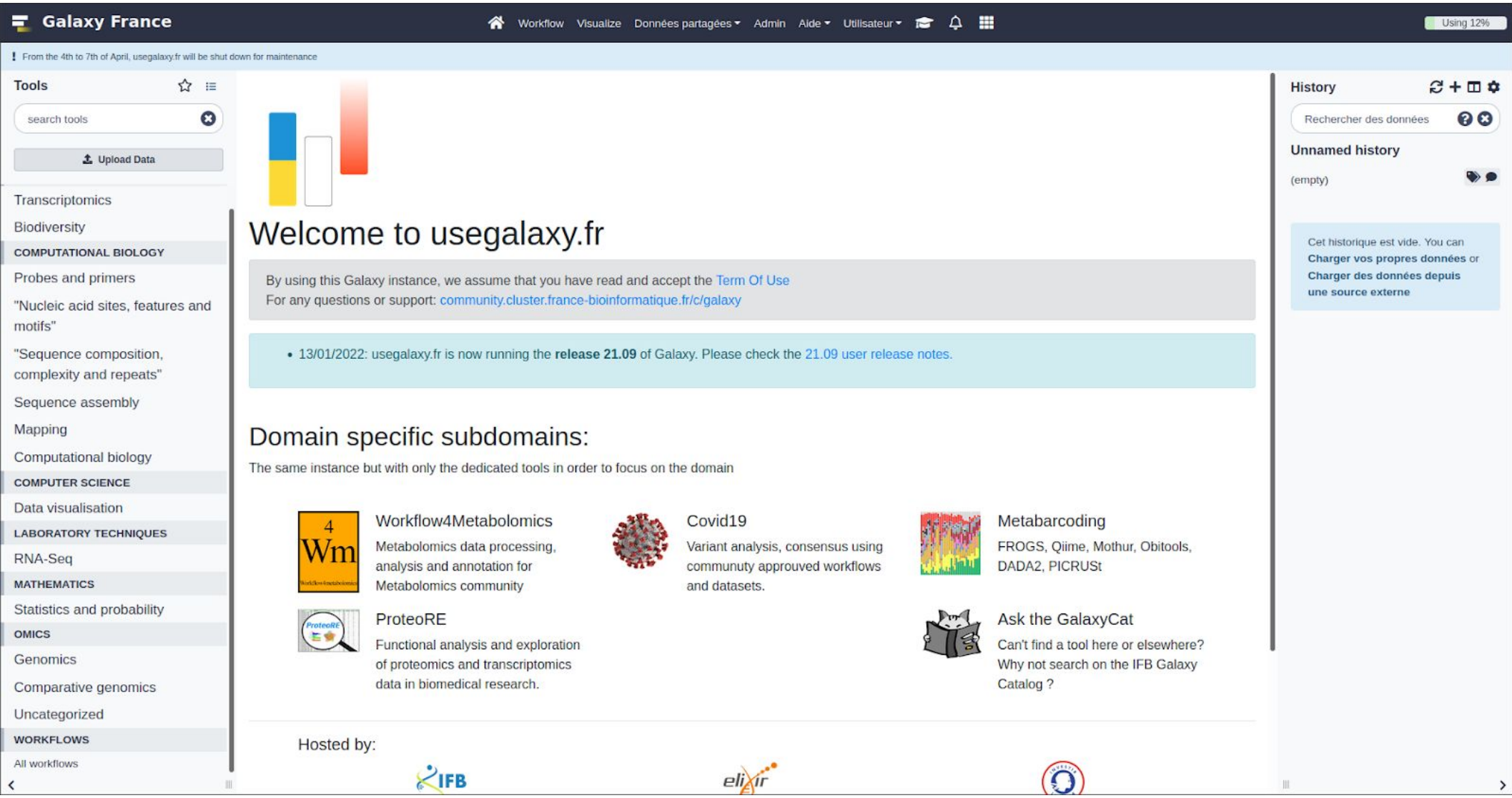
https://galaxy.sb-roscoff.fr/



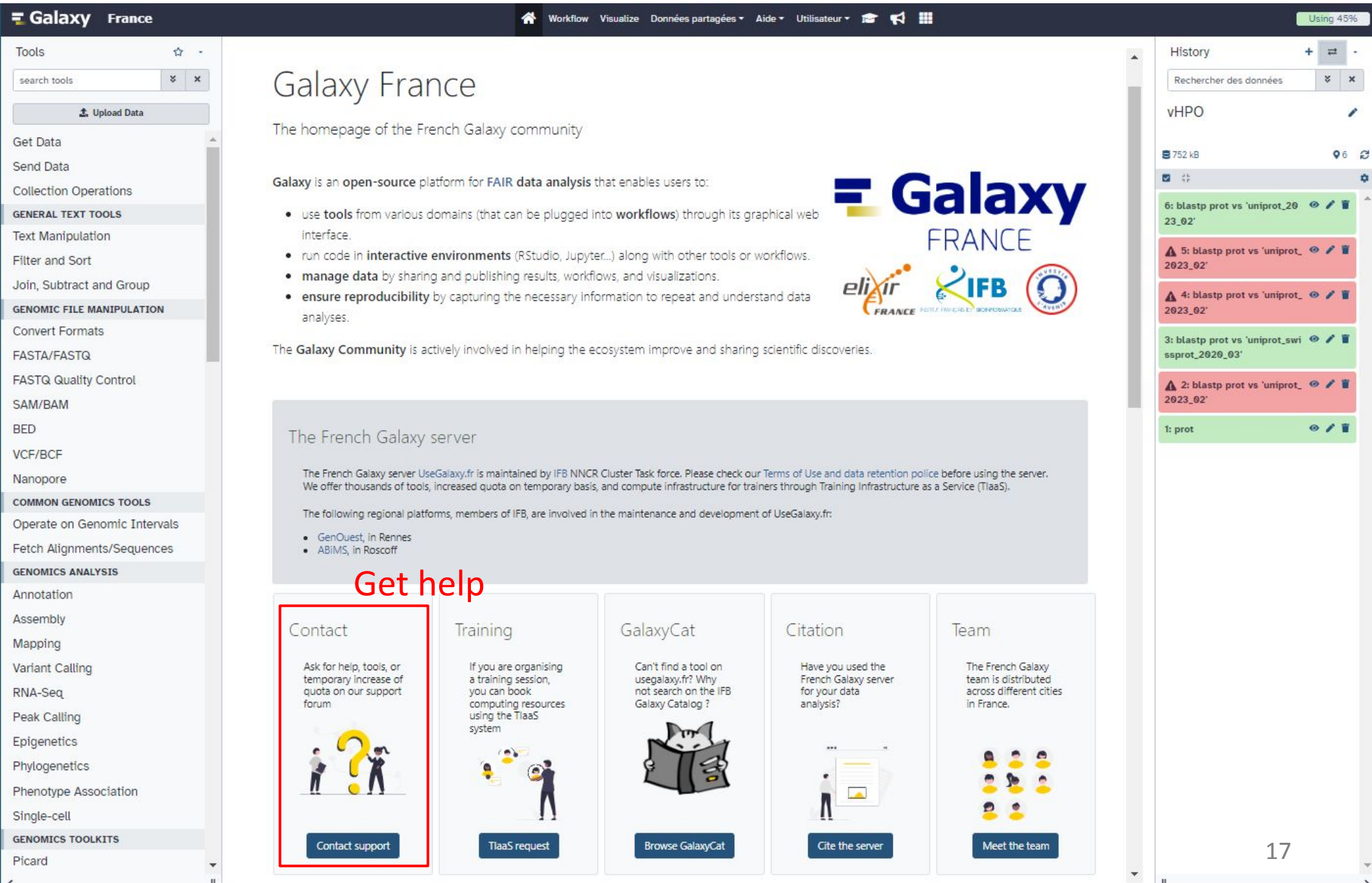
https://usegalaxy.fr/



Galaxy France: https://usegalaxy.fr



Galaxy France



Get help: https://community.france-bioinformatique.fr/c/galaxy/8

Hands-on

DATA IMPORT & FIRST TOOL

https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html

Galaxy Training! | Introduction to Galaxy Analyses | Learning Pathways | Help | Settings | Search Tutorials

Galaxy Basics for genomics

Authors: Saskia Hiltemann, Nicola Soranzo, Clemens Blank, Anton Nekrutenko, Björn Grüning, Anne Pajon, Helena Rasche

Overview

Questions:
- Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?

Objectives:
- Familiarize yourself with the basics of Galaxy
- Learn how to obtain data from external sources
- Learn how to run tools
- Learn how histories work
- Learn how to create a workflow
- Learn how to share your work

Time estimation: 1 hour
Level: Introductory
Supporting Materials: Datasets | Workflows | FAQs | Recordings | Available on these Galaxies
Published: Dec 19, 2016
Last modification: May 3, 2024
License: Tutorial Content is licensed under Creative Commons Attribution 4.0 International License. The GTN Framework is licensed under MIT
PURL: https://gxy.io/GTN:T00186
Rating: 4.1 (10 recent ratings, 141 all time)
Revision: 131

Stop after step "Find exons with the most SNPs"

Setting the stage: Exons and SNPs
Get your workspace ready
Analysis
Histories and workflows: A brief introduction
Share your work

This tutorial aims to familiarize you with the Galaxy user interface. It will teach you how to perform basic tasks such as importing data, running tools, working with histories, creating workflows, and sharing your work.

Comment: Results may vary

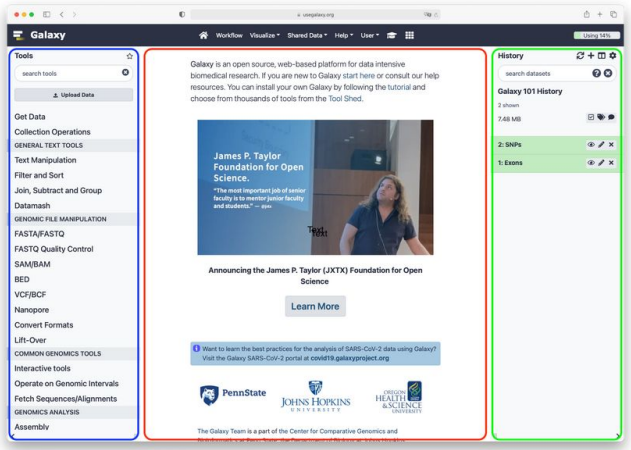
Your results may be slightly different from the ones presented in this tutorial due to differing versions of tools, reference data, external databases, or because of stochastic processes in the algorithms.

Agenda

TOOLS & HISTORY

Tools & history

Main Galaxy interface



Home page divided into 3 panels

https://training.galaxyproject.org/training-material/topics/introduction/tutorials/introduction/slides.html#7

DATA IMPORT

Data import

Getting data into Galaxy

Anne Pajon Dave Clements Maria Doyle Saskia Hiltemann
Add Contributions!

Updated: Jul 9, 2021
Or view the plain-text slides without JS
Tip: press P to view the presenter notes

https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/get-data/slides.html

Tips & tricks

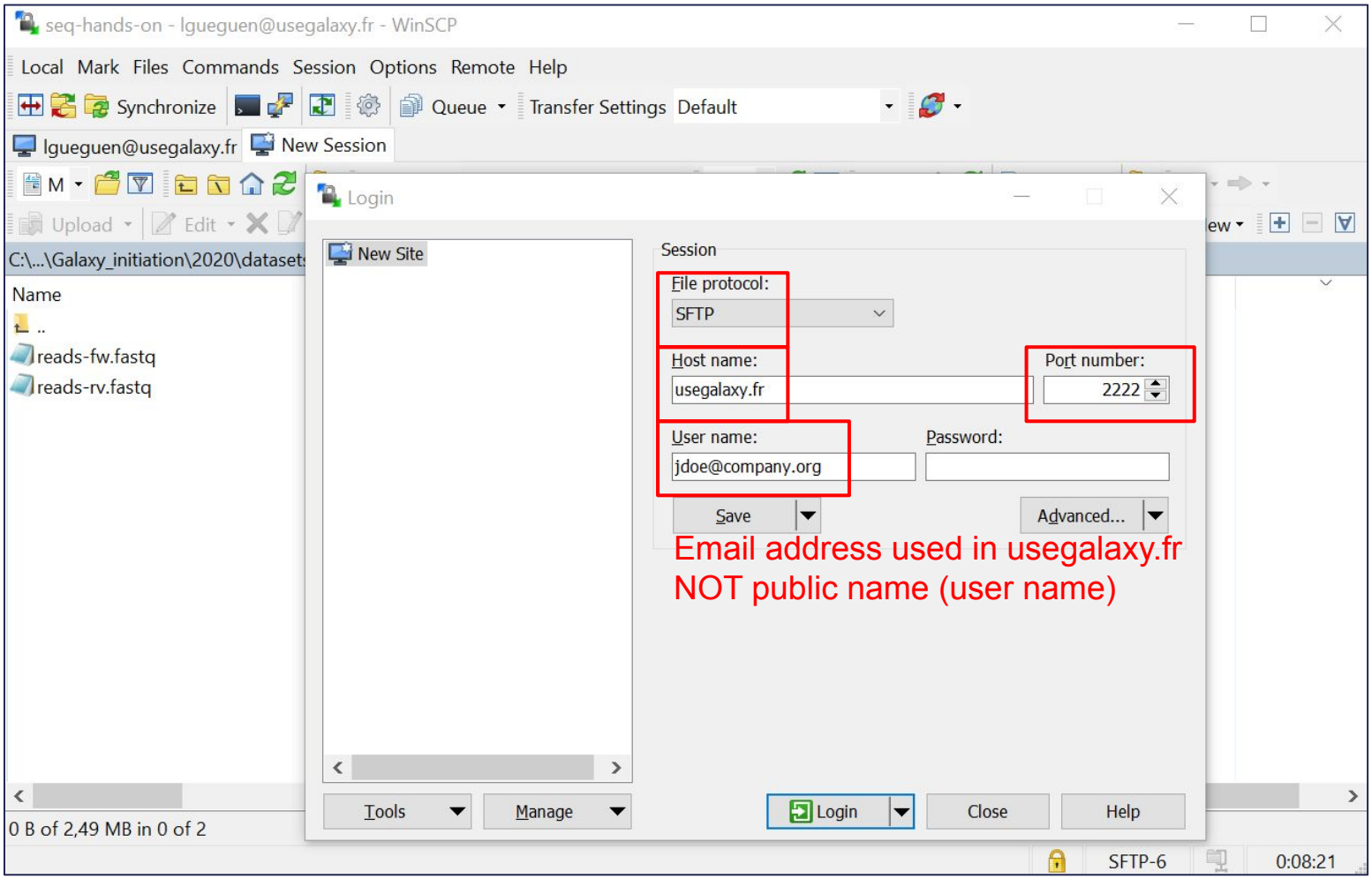
DATA IMPORT

Data import using FTP

STEP 1: CHOOSE A FTP CLIENT

Data import using FTP

Example with WinSCP for usegalaxy.fr

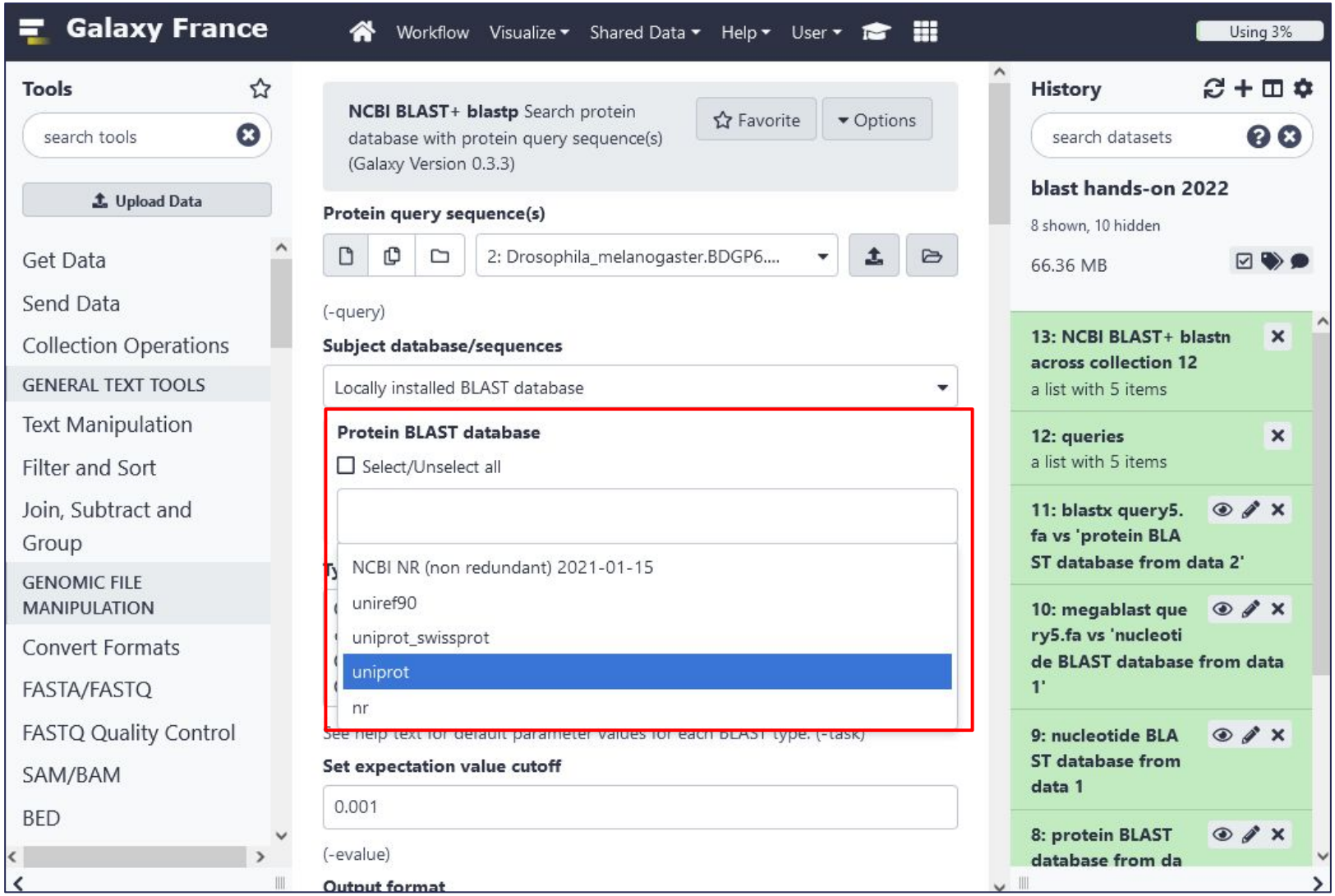


Command line: https://community.france-bioinformatique.fr/t/ftp-usegalaxy-fr/1453

Data import

For HUGE public resources: genome, databank …

--> Make a request to the support team

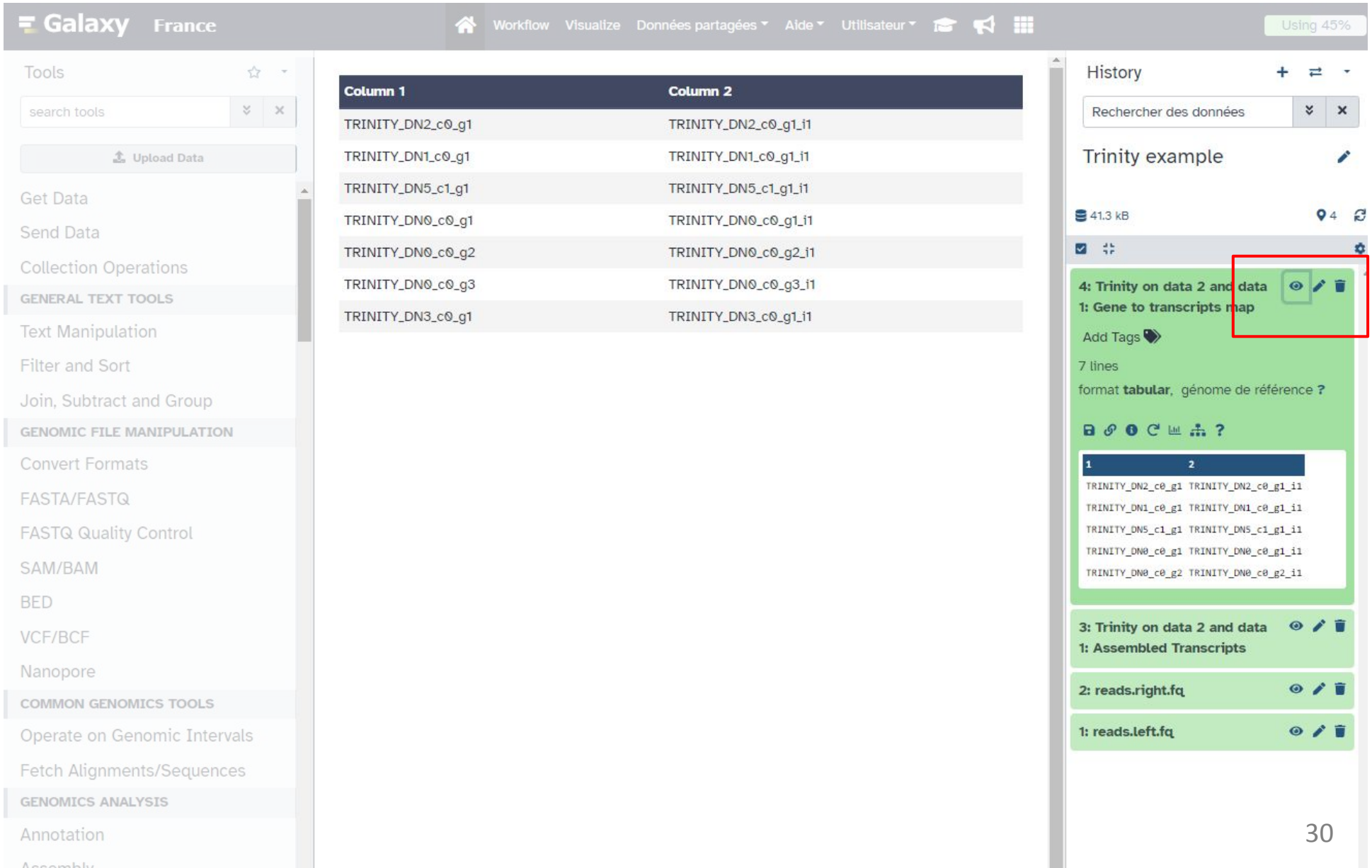


Icons & buttons

DATASET

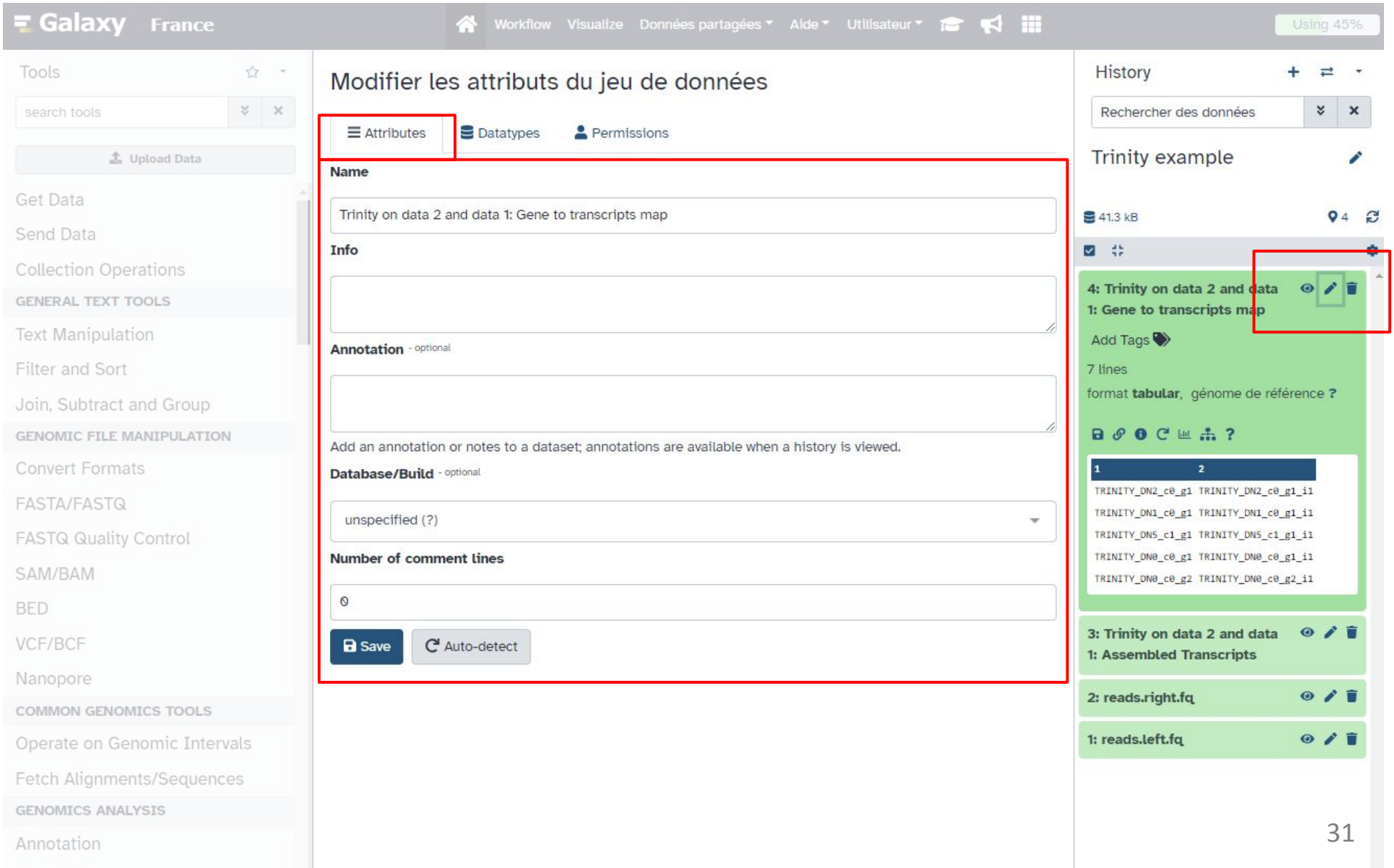
Dataset

Dataset display : text, tabular, pdf, picture, html …



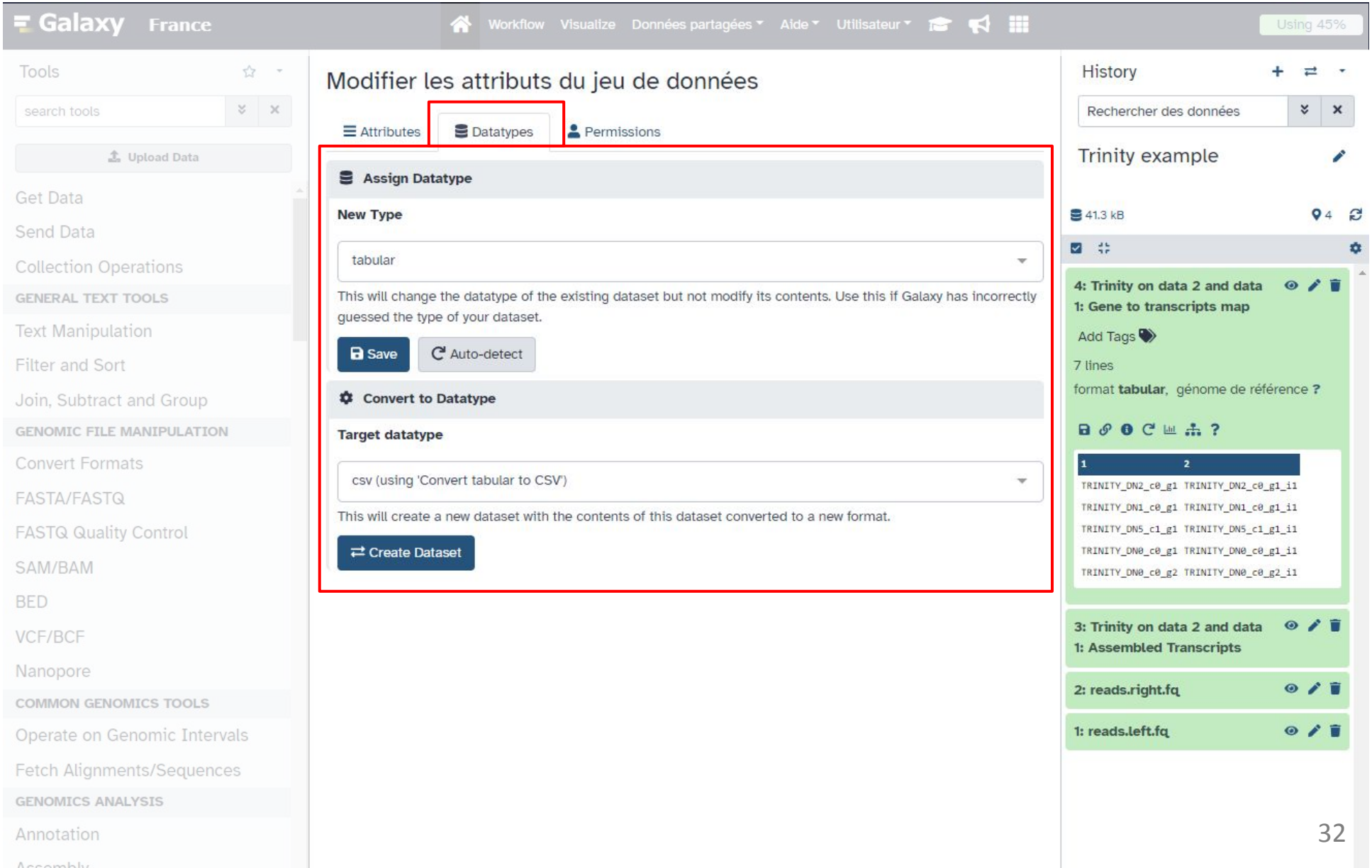
Dataset

Rename, annotate, change datatype…

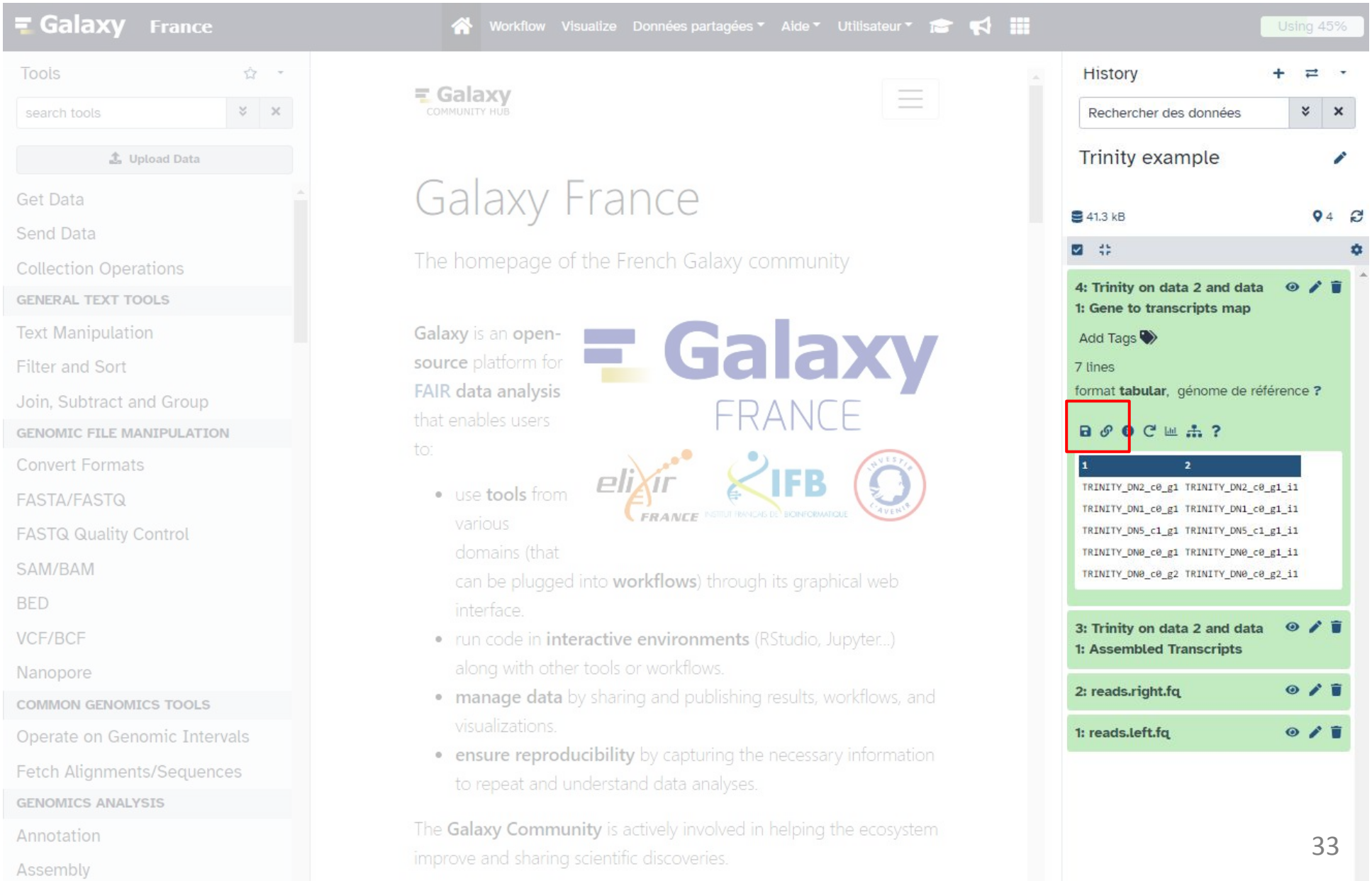


Dataset

Rename, annotate, change datatype…

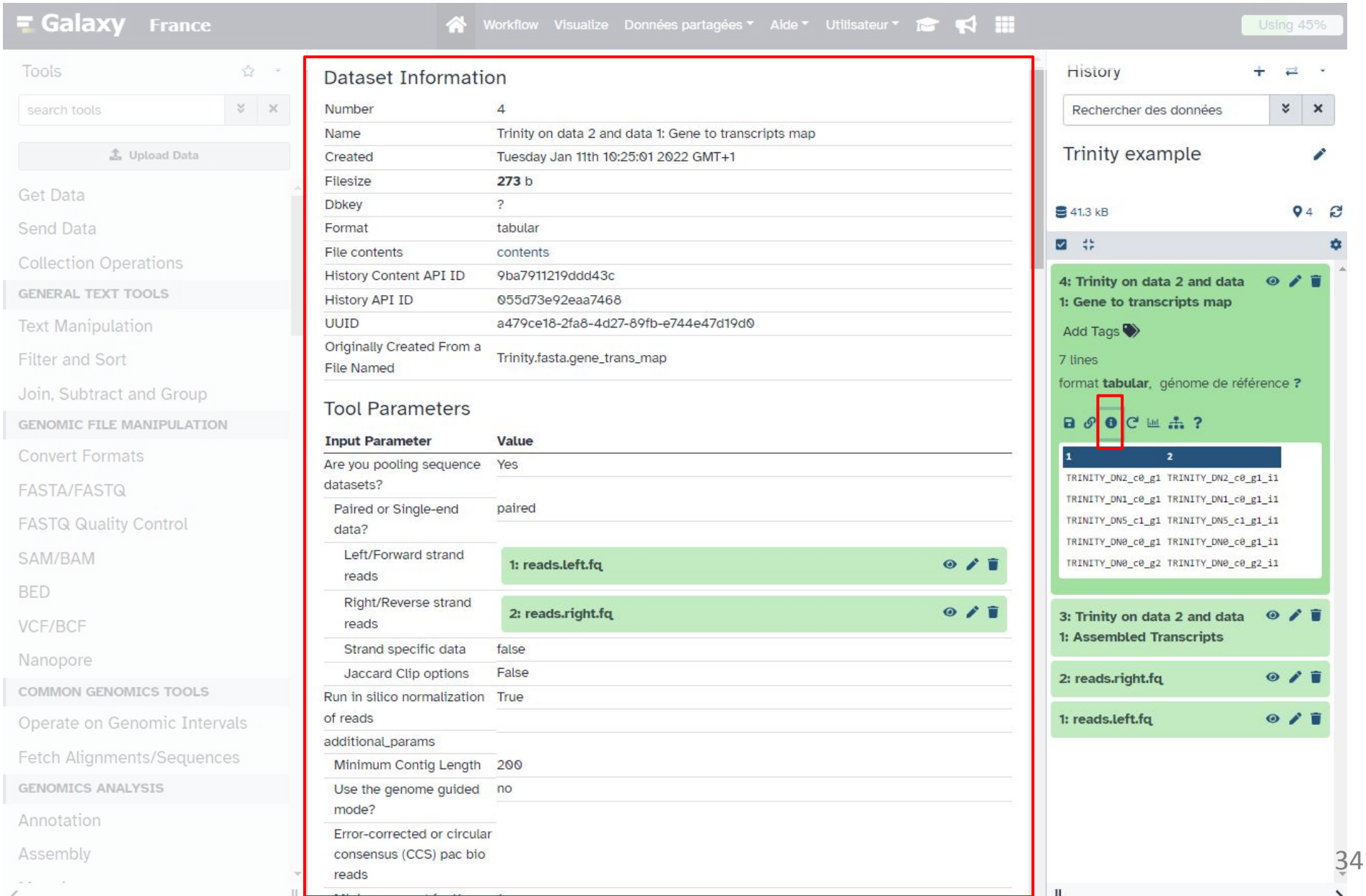


Download ; link to data

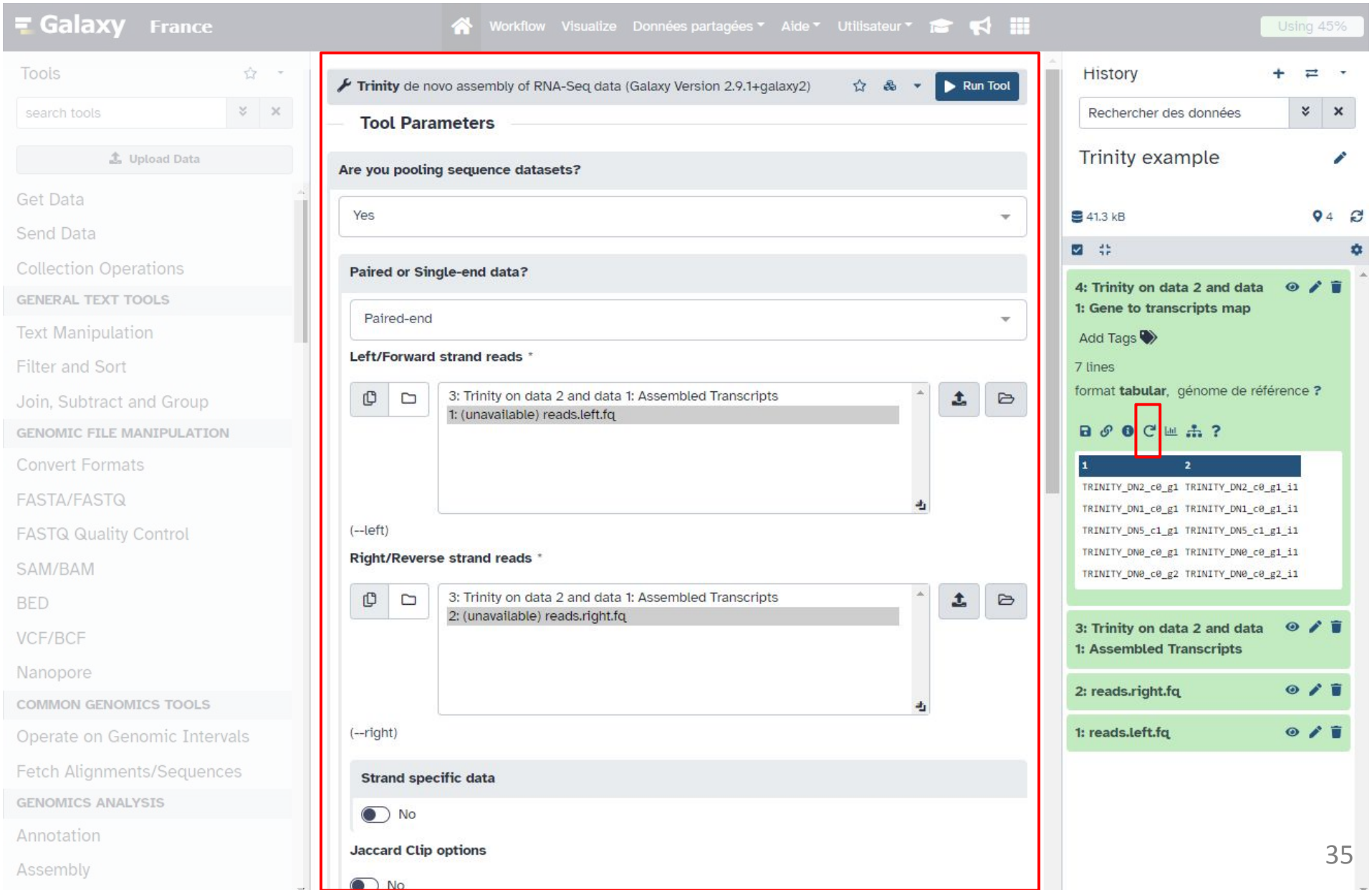


Dataset

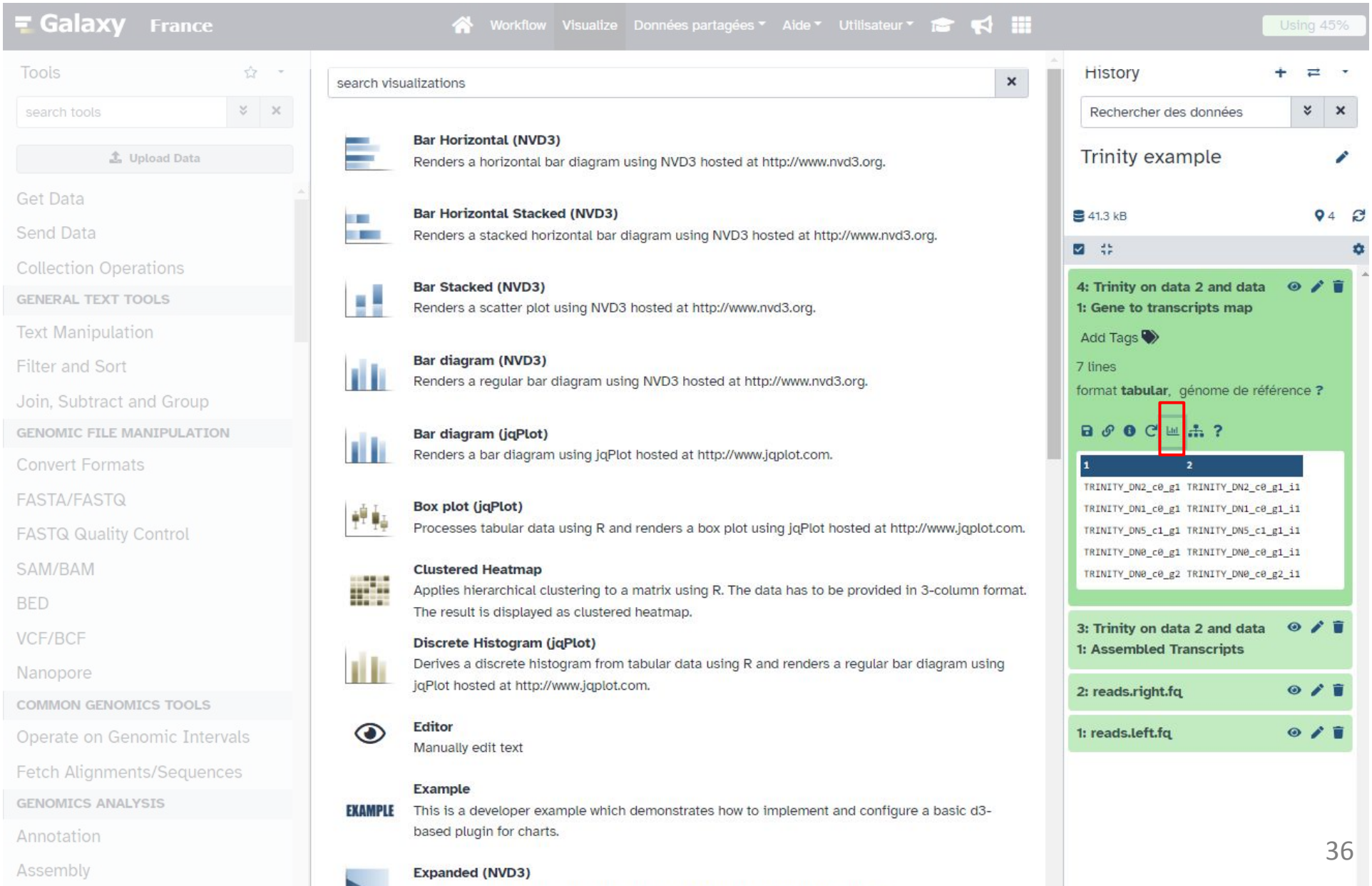
Informations



Re-run a job


Galaxy France
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 45%
Tools
search tools
Upload Data
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Alignments/Sequences
GENOMICS ANALYSIS
Annotation
Assembly
Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.9.1+galaxy2)
Run Tool
Tool Parameters
Are you pooling sequence datasets?
Yes
Paired or Single-end data?
Paired-end
Left/Forward strand reads *
3: Trinity on data 2 and data 1: Assembled Transcripts
1: (unavailable) reads.left.fq
(--left)
Right/Reverse strand reads *
3: Trinity on data 2 and data 1: Assembled Transcripts
2: (unavailable) reads.right.fq
(--right)
Strand specific data
No
Jaccard Clip options
No
History
Rechercher des données
Trinity example
41.3 kB
4
4: Trinity on data 2 and data 1: Gene to transcripts map
Add Tags
7 lines
format tabular, génome de référence ?
1 2
TRINITY_DN2_c0_g1 TRINITY_DN2_c0_g1_i1
TRINITY_DN1_c0_g1 TRINITY_DN1_c0_g1_i1
TRINITY_DN5_c1_g1 TRINITY_DN5_c1_g1_i1
TRINITY_DN0_c0_g1 TRINITY_DN0_c0_g1_i1
TRINITY_DN0_c0_g2 TRINITY_DN0_c0_g2_i1
3: Trinity on data 2 and data 1: Assembled Transcripts
2: reads.right.fq
1: reads.left.fq

Visualisation tools



Dataset - Visualisations

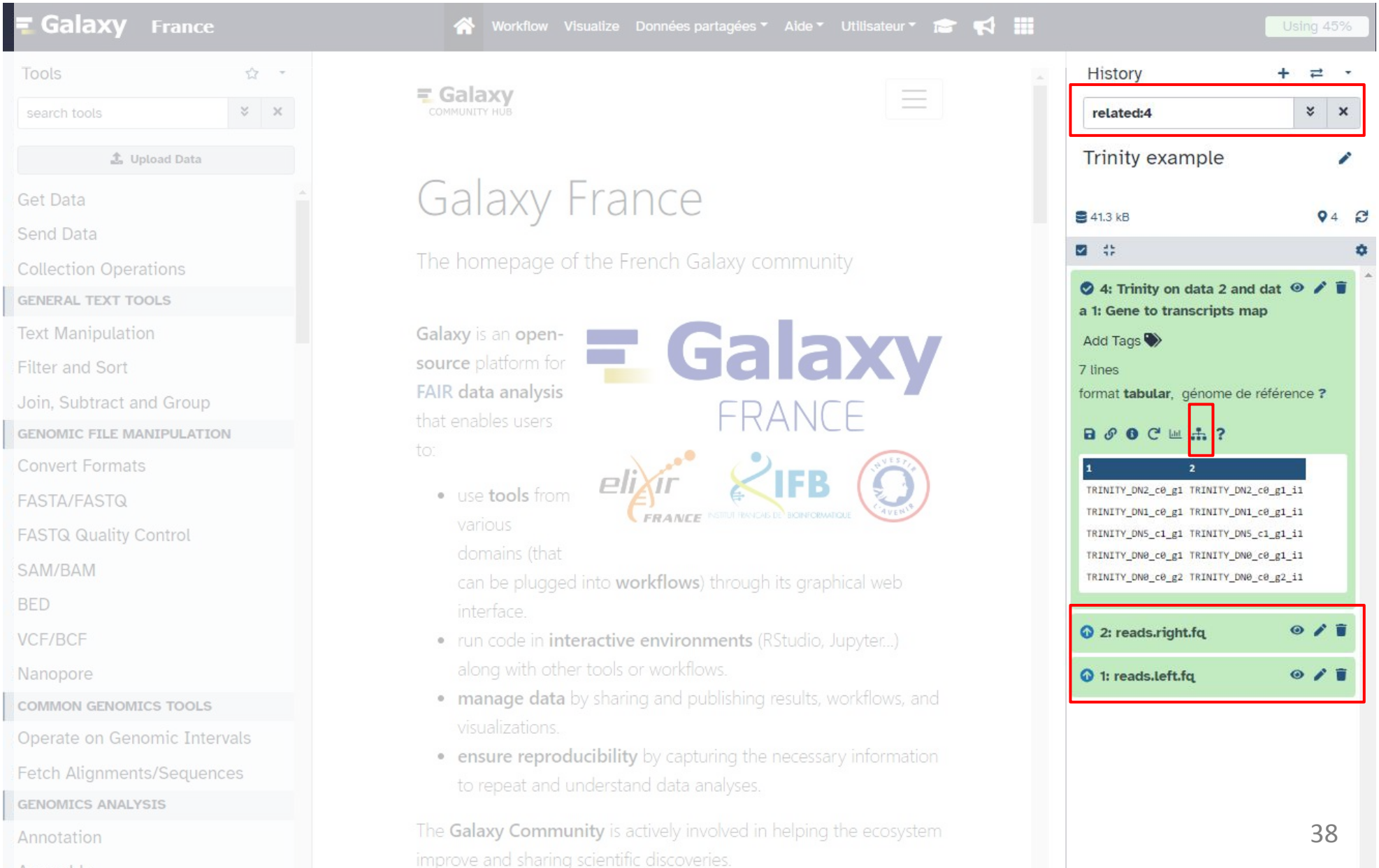
Visualisations in Galaxy

Helena Rasche Saskia Hiltemann Add Contributions!

Updated: Jul 9, 2021
Or view the plain-text slides without JS
Tip: press P to view the presenter notes

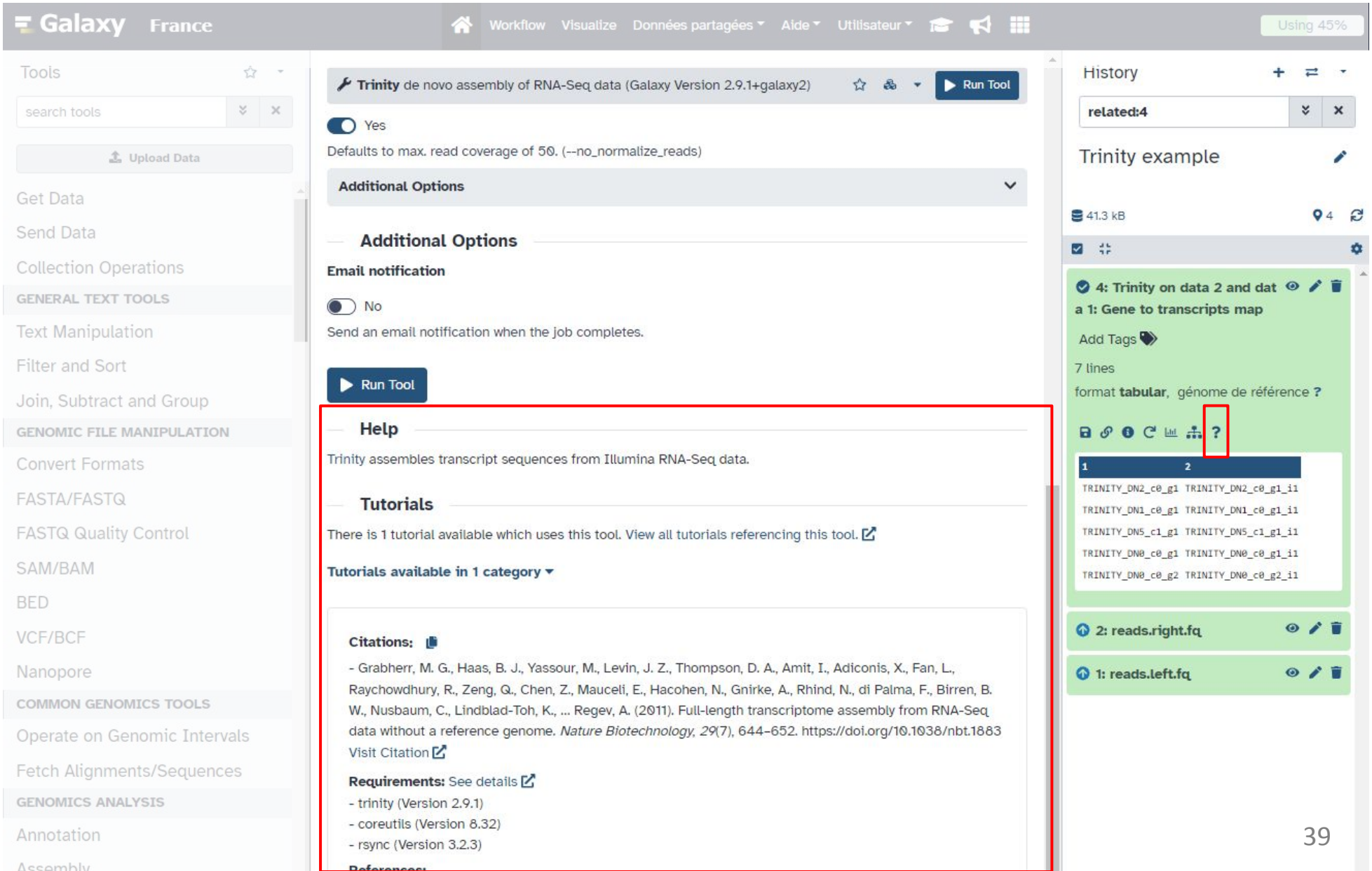
https://training.galaxyproject.org/training-material/topics/visualisation/slides/introduction.html

Show related items

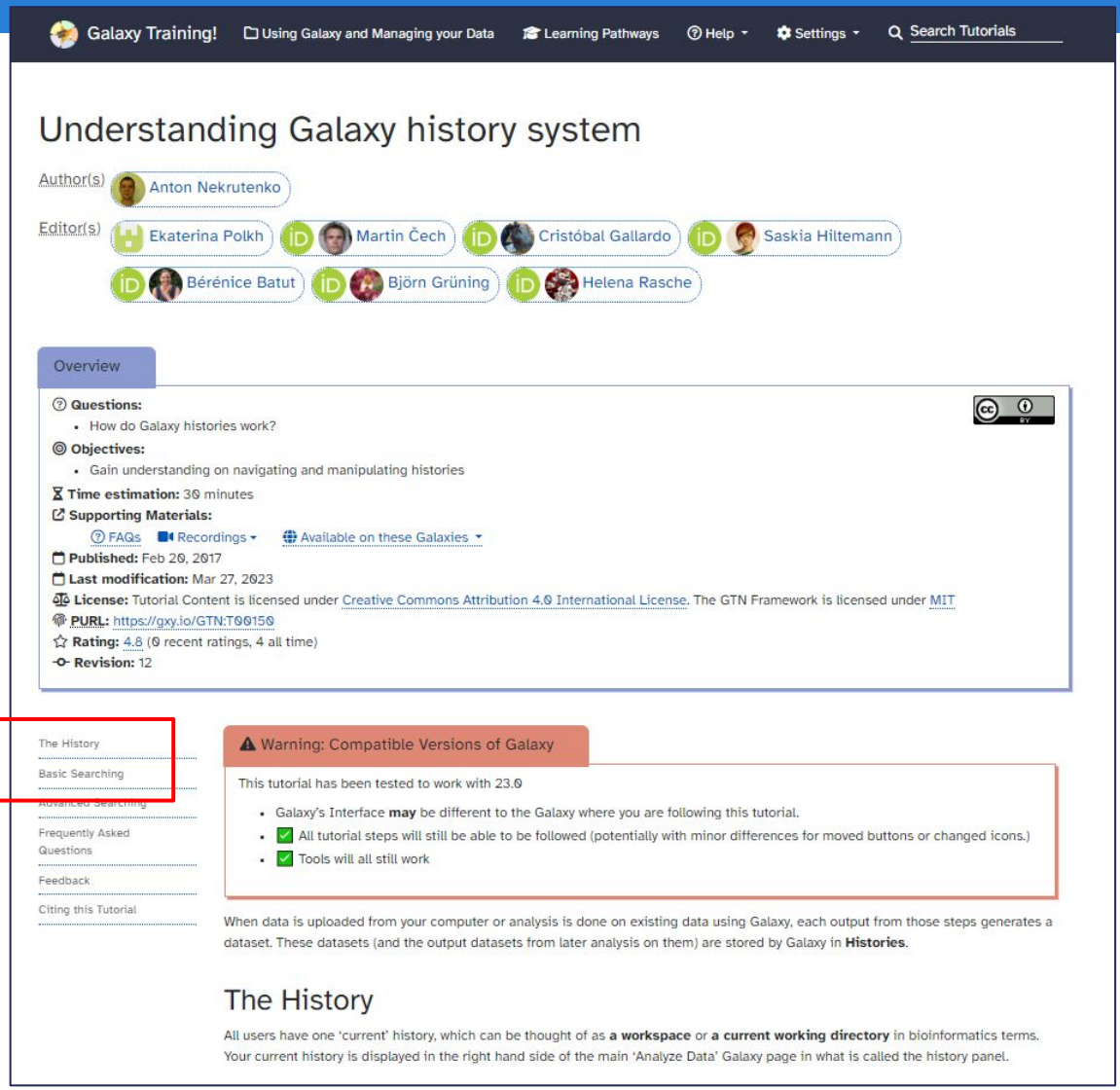


Dataset

Help



History & datasets



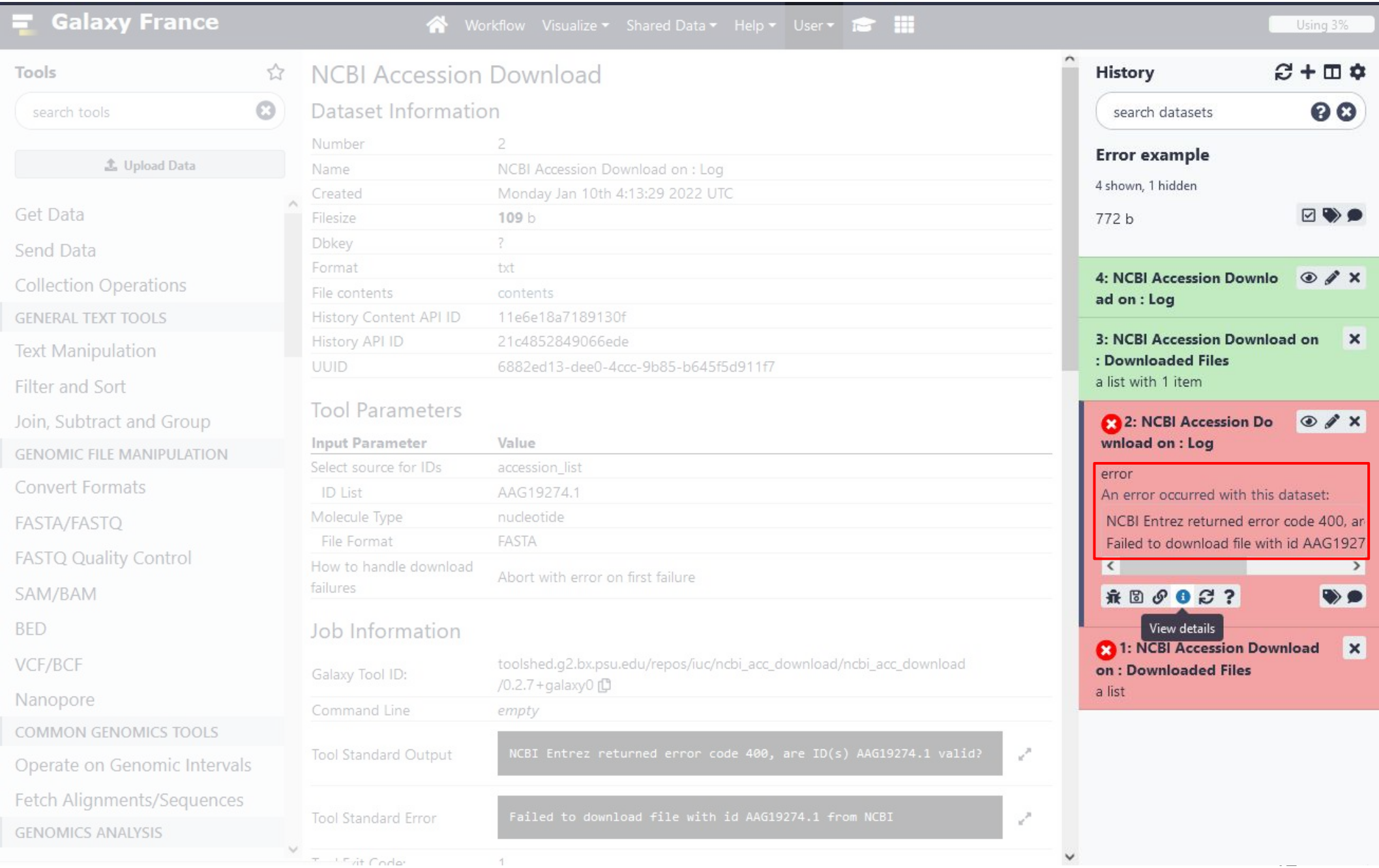
https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/history/tutorial.html

Supplementary material on deleting data:
https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/download-delete-data/tutorial.html#deleting-data-from-galaxy

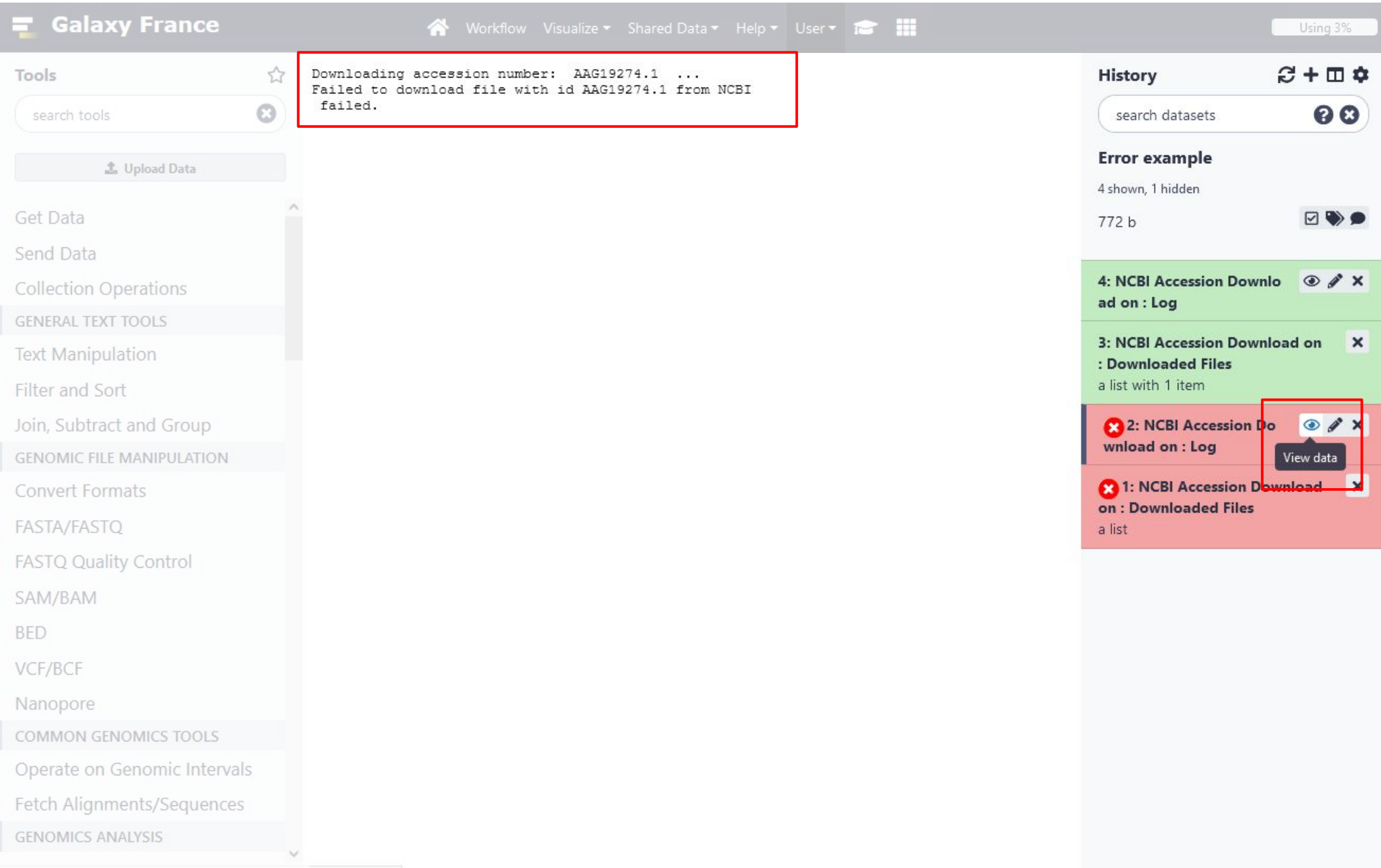
Tips & tricks

Handle tool errors

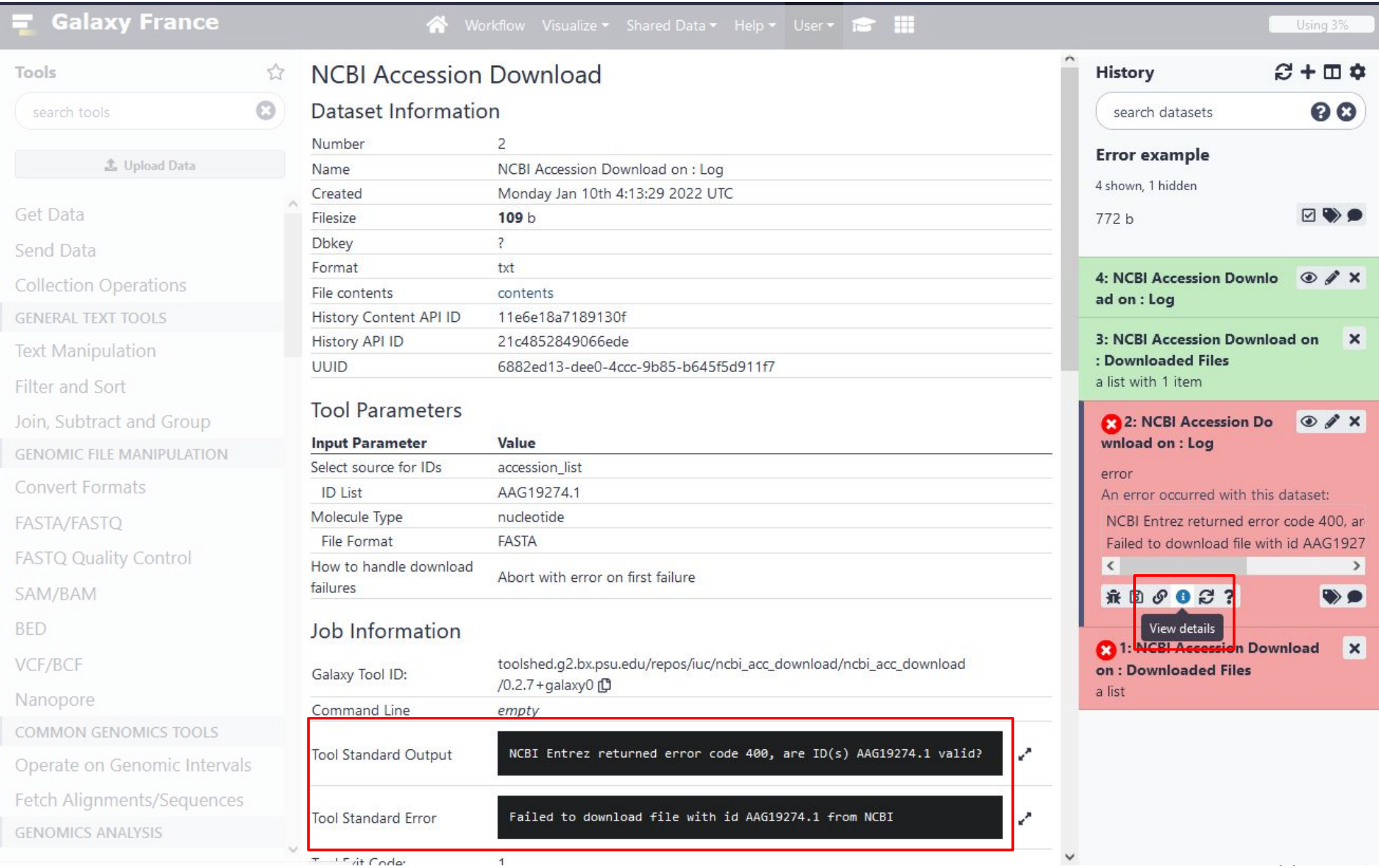
Tools - Handle errors



Tools - Handle errors



Tools - Handle errors



Tools - Handle errors

Galaxy France | Workflow | Visualize | Shared Data | Help | User | Using 3%

Tools

search tools

Upload Data

Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Alignments/Sequences
GENOMICS ANALYSIS

Dataset Error Report

An error occurred while running the tool **toolshed.g2.bx.psu.edu/repos/iuc/ncbi_acc_download/ncbi_acc_download/0.2.7+galaxy0**.

Details

Execution resulted in the following messages:

```
Fatal error: Exit code 1 ()
```

Tool generated the following standard error:

```
Failed to download file with id AAG19274.1 from NCBI
```

Troubleshooting

There are a number of helpful resources to self diagnose and correct problems.
Start here: **My job ended with an error. What can I do?**

Issue Report

Please provide detailed information on the activities leading to this issue:

Report

History

search datasets

Error example

4 shown, 1 hidden

772 b

4: NCBI Accession Download on : Log

3: NCBI Accession Download on : Downloaded Files
a list with 1 item

2: NCBI Accession Download on : Log
error
An error occurred with this dataset:
NCBI Entrez returned error code 400, ar
Failed to download file with id AAG1927

View or report this error

1: NCBI Accession Download on : Downloaded Files
a list

Tools - Handle errors

Galaxy France | Workflow | Visualize | Shared Data | Help | User | Using 3%

Tools

search tools

Upload Data

Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Alignments/Sequences
GENOMICS ANALYSIS

Dataset Error Report

An error occurred while running the tool **toolshed.g2.bx.psu.edu/repos/iuc/ncbi_acc_download/ncbi_acc_download/0.2.7+galaxy0**.

Details

Execution resulted in the following messages:

```
Fatal error: Exit code 1 ()
```

Tool generated the following standard error:

```
Failed to download file with id AAG19274.1 from NCBI
```

Troubleshooting

https://galaxyproject.org/support/tool-error/

There are a number of helpful resources to self diagnose and correct problems.
Start here: **My job ended with an error. What can I do?**

Issue Report

Please provide detailed information on the activities leading to this issue:

Report

History

search datasets

Error example

4 shown, 1 hidden

772 b

4: NCBI Accession Download on : Log

3: NCBI Accession Download on : Downloaded Files
a list with 1 item

2: NCBI Accession Download on : Log
error
An error occurred with this dataset:
NCBI Entrez returned error code 400, ar
Failed to download file with id AAG1927

View or report this error

1: NCBI Accession Download on : Downloaded Files
a list

Tools - Handle errors

Galaxy France | Workflow | Visualize | Shared Data | Help | User | Using 3%

Tools

search tools

Upload Data

Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Alignments/Sequences
GENOMICS ANALYSIS

Dataset Error Report

An error occurred while running the tool **toolshed.g2.bx.psu.edu/repos/iuc/ncbi_acc_download/ncbi_acc_download/0.2.7+galaxy0**.

Details

Execution resulted in the following messages:

```
Fatal error: Exit code 1 ()
```

Tool generated the following standard error:

```
Failed to download file with id AAG19274.1 from NCBI
```

Troubleshooting

There are a number of helpful resources to self diagnose and correct problems.
Start here: **My job ended with an error. What can I do?**

Issue Report

Please provide detailed information on the activities leading to this issue:

Report

History

search datasets

Error example

4 shown, 1 hidden

772 b

4: NCBI Accession Download on : Log

3: NCBI Accession Download on : Downloaded Files
a list with 1 item

2: NCBI Accession Download on : Log
error
An error occurred with this dataset:
NCBI Entrez returned error code 400, ar
Failed to download file with id AAG1927

View or report this error

1: NCBI Accession Download on : Downloaded Files
a list

For usegalaxy.fr, create a new request on
https://community.france-bioinformatique.fr/c/galaxy

Hands-on

TOOLS

https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html

Galaxy Training! | Introduction to Galaxy Analyses | Learning Pathways | Help | Settings | Search Tutorials

Galaxy Basics for genomics

Authors: Saskia Hiltemann, Nicola Soranzo, Clemens Blank, Anton Nekrutenko, Björn Grüning, Anne Pajon, Helena Rasche

Overview

Questions:
- Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?

Objectives:
- Familiarize yourself with the basics of Galaxy
- Learn how to obtain data from external sources
- Learn how to run tools
- Learn how histories work
- Learn how to create a workflow
- Learn how to share your work

Time estimation: 1 hour
Level: Introductory
Supporting Materials: Datasets | Workflows | FAQs | Recordings | Available on these Galaxies
Published: Dec 19, 2016
Last modification: May 3, 2024

PURL: https://gxy.io/GTN:T00186
Rating: 4.1 (10 recent ratings, 141 all time)
Revision: 131

Run next steps

Setting the stage: Exons and SNPs
Get your workspace ready
Analysis
Histories and workflows: A brief introduction
Share your work

This tutorial aims to familiarize you with the Galaxy user interface. It will teach you how to perform basic tasks such as importing data, running tools, working with histories, creating workflows, and sharing your work.

Comment: Results may vary

Your results may be slightly different from the ones presented in this tutorial due to differing versions of tools, reference data, external databases, or because of stochastic processes in the algorithms.

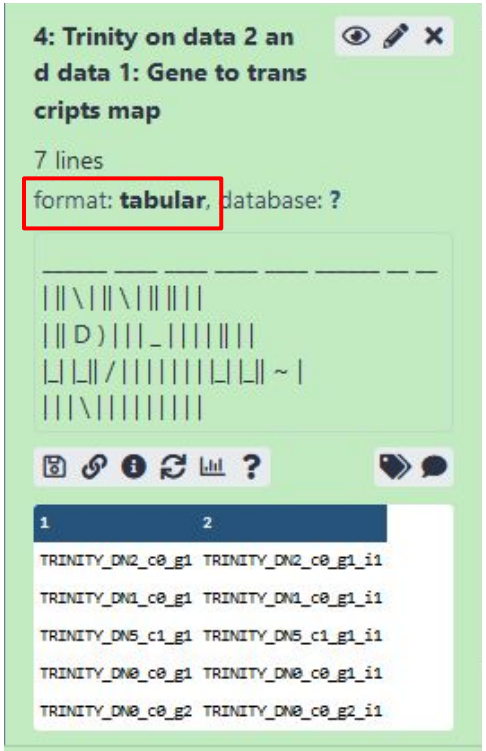
Agenda

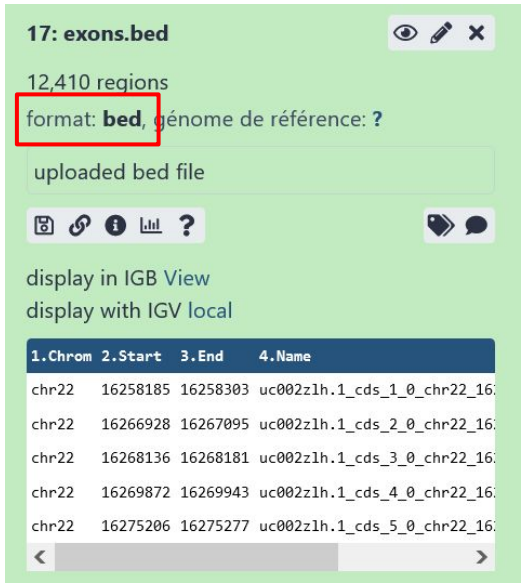
Datatypes: a short overview

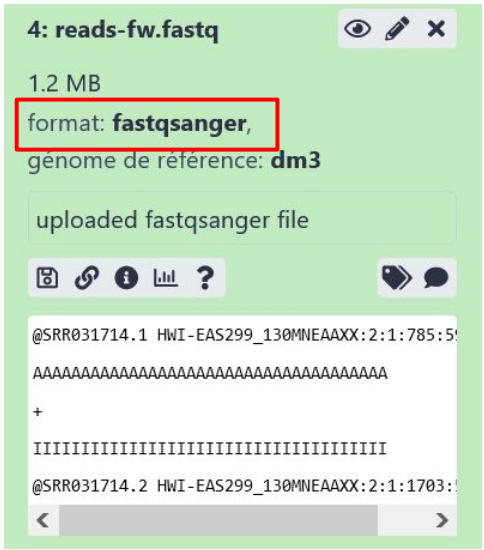
DATASET

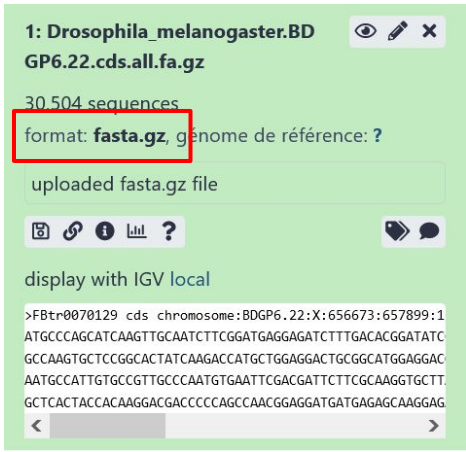
Dataset - Datatypes

- Every Galaxy dataset is associated with a datatype.
- Datasets produced by a tool have their datatype assigned by the tool









Dataset - Datatypes

Common text formats:

- ***txt***: plain text ('.txt')
- ***tabular***: tab delimited ('.tab', '.txt', etc.)

```
          wt_37_2  wt_37_3  wt_37_1
TR24|c0_g1_i1     90.00    67.00    85.00
TR2779|c0_g1_i1   186.00   137.00   217.00
TR127|c1_g1_i1    9.00 23.00    16.00
```

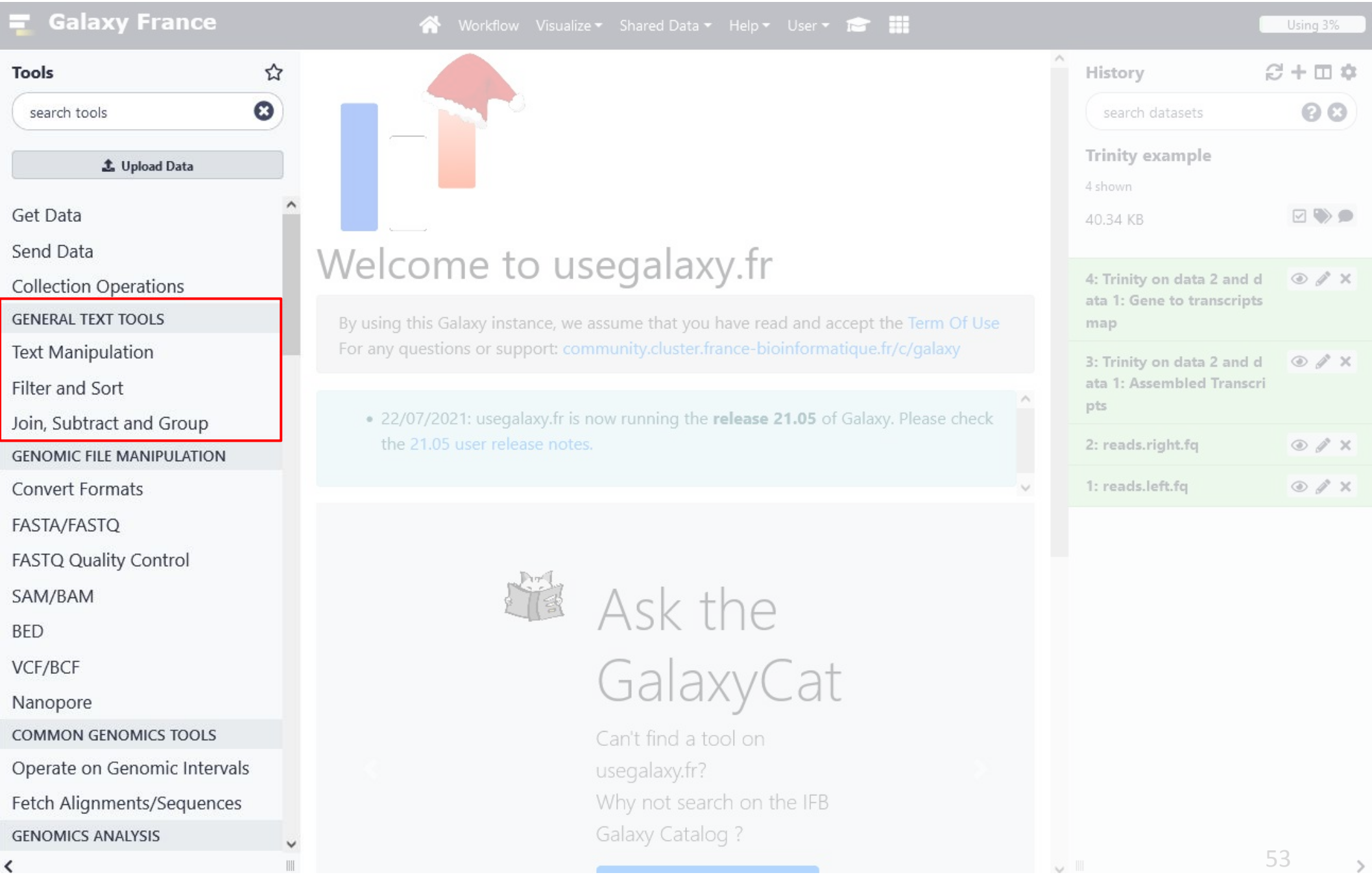
- ***csv***: comma-separated values ('.csv')

```
Year,Make,Model
1997,Ford,E350
2000,Mercury,Couga
r
```

- ***html***: standard language for web pages

```
<!DOCTYPE html>
<html>
  <head>
    <title>This is a
title</title>
  </head>
  <body>
    <p>Hello world!</p>
  </body>
</html>
```

Dataset - Datatypes



Dataset - Datatypes

Common binary formats:

- ***data***: generic binary format
- ***zip, tar***: archives
- ***pdf, png, jpg, bmp, tiff, gif***: images
- ***rdata***: statistical computing program R
- ***bam*, wig, bigwig**: sequence alignment

Dataset - Datatypes

Sequence file formats:

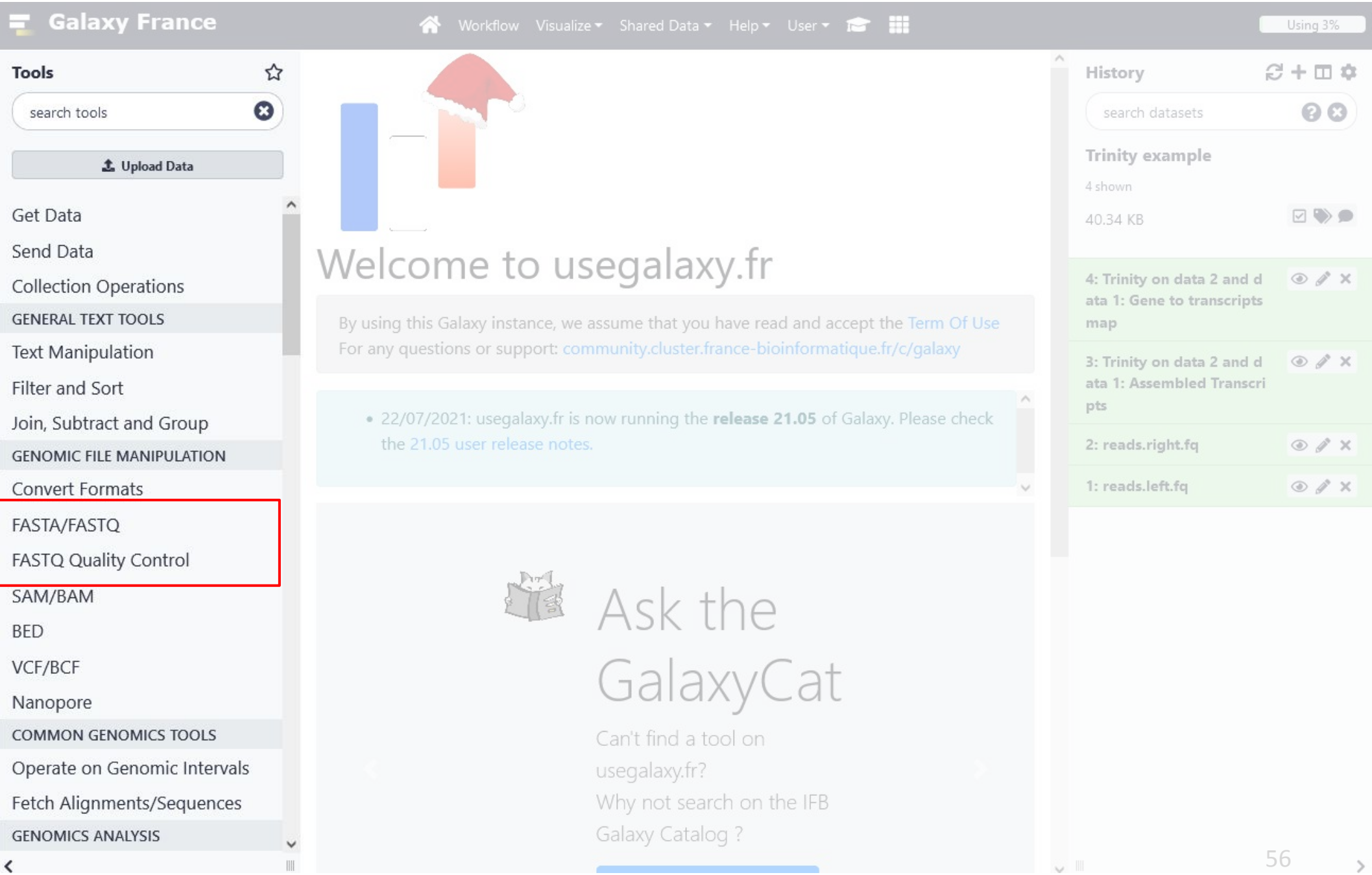
- ***fasta***: a single-line description with '>', followed by lines of sequence data ('.fasta', '.fas')

```
>sequence1
atgcgtttgcgtgcatgcgtttgcgtgcatgcgtttgcgtgcatgcgtttgcgtg
c
atgcgtttgcgtgc
>sequence2
tttcgtgcgtatagtttcgtgcgtatagtttcgtgcgtatagtttcgtgcgtata
g
tggcgcggt
```

- ***fastq***: sequence + quality score ('.fastq', '.fq')

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
T
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC6
5
@SEQ_ID2
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
T
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC6
```

Dataset - Datatypes



Dataset - Datatypes

Sequence file formats:

- ***gff3, bed, genbank***: sequence + annotations

bed

```
track name=pairedReads description="Clone Paired Reads"
useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

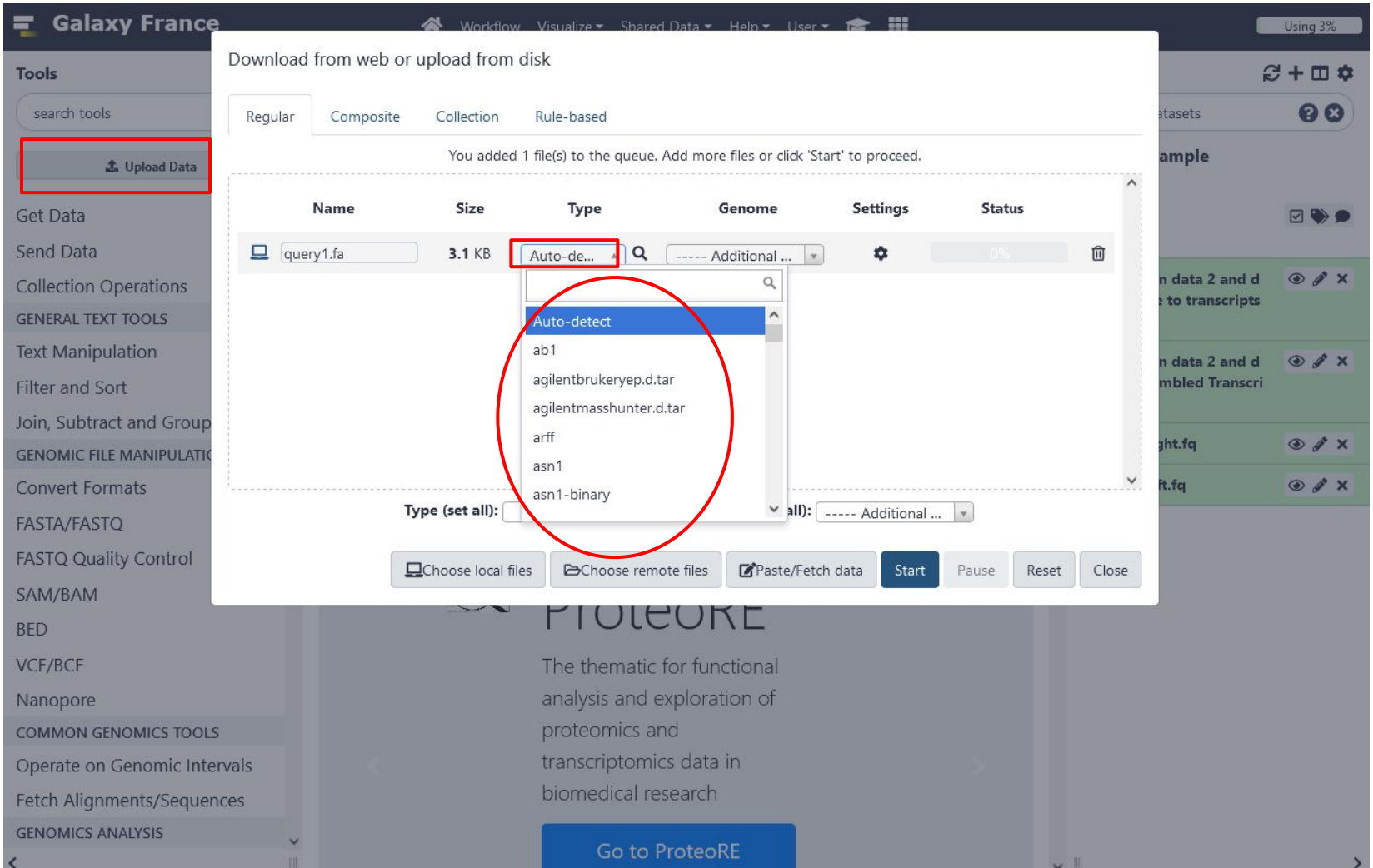
gff3

```
##gff-version 3
ctg123  .  exon  1300  1500  .  +  .
ID=exon00001
ctg123  .  exon  1050  1500  .  +  .
ID=exon00002
ctg123  .  exon  3000  3902  .  +  .
ID=exon00003
##FASTA
>ctg123
cttctgggcgtacccgattctcggagaacttgccgcaccattccgcctt
g
tgttcattgctgcctgcatgttcattgtctacctcggctacgtgtggct
g
…
```

https://genome.ucsc.edu/FAQ/FAQformat.html#format3

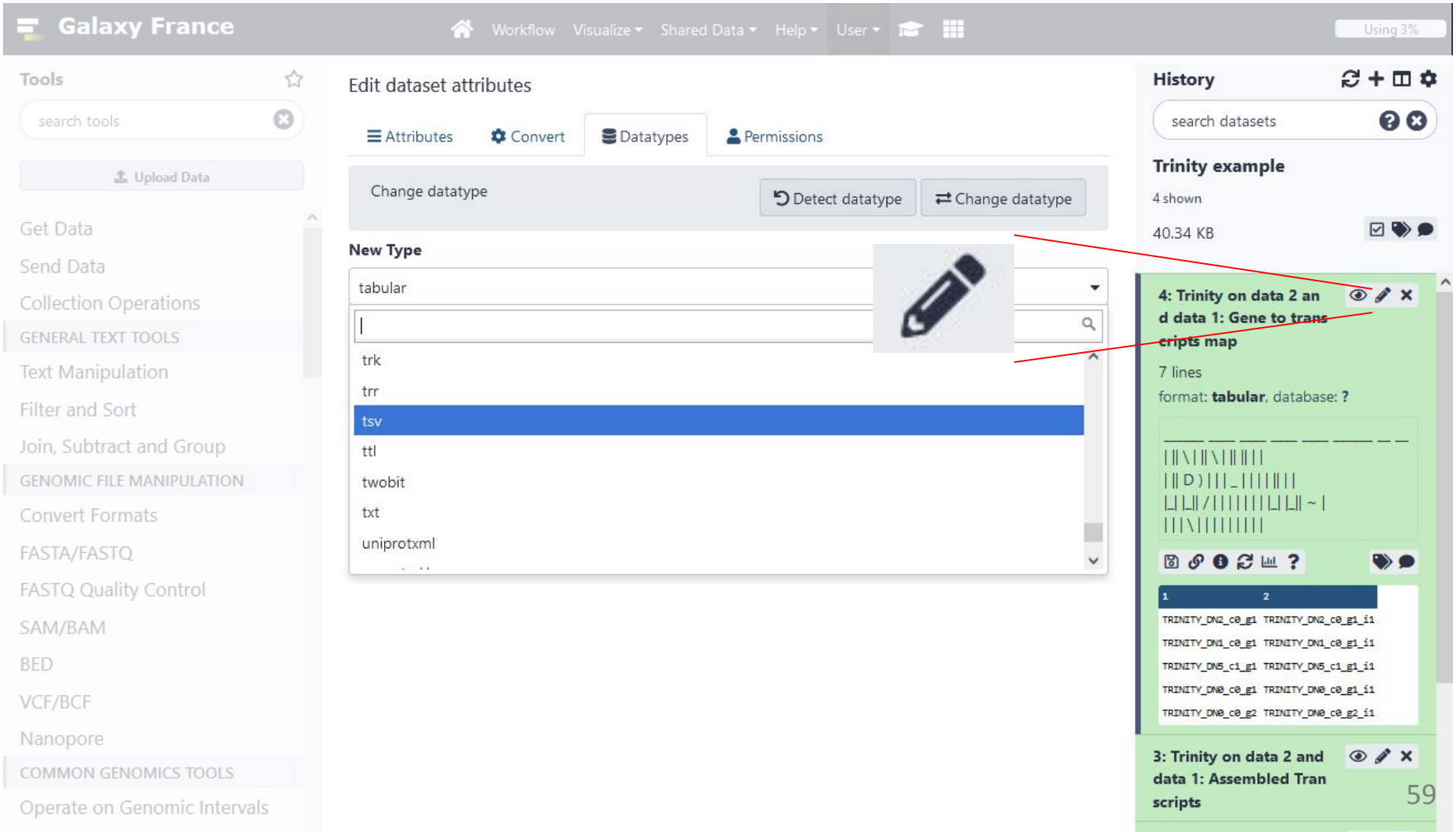
Dataset - Datatypes

- On upload, datatype can be detected or user specified.



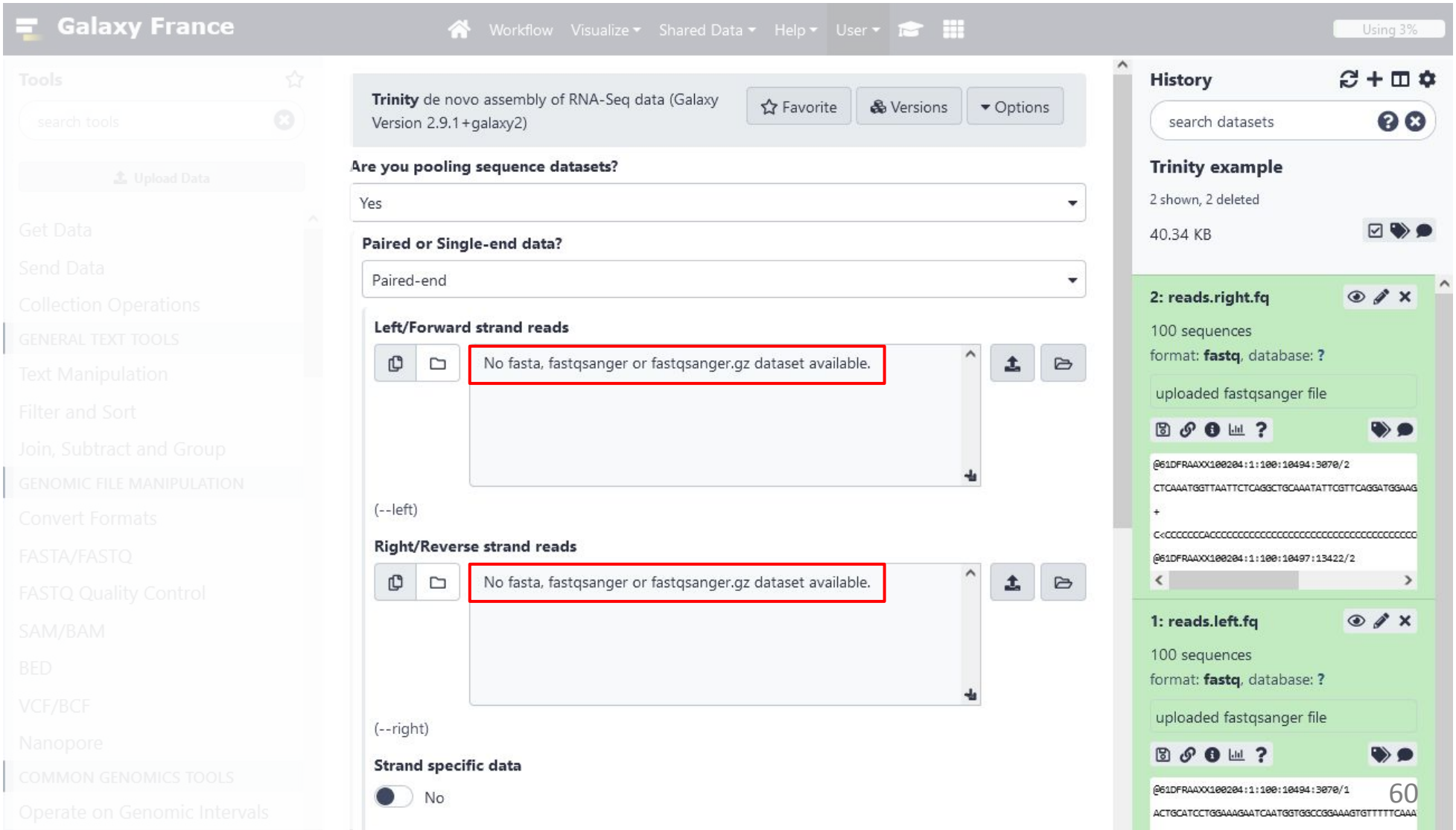
Dataset - Datatypes

- Change the datatype of a dataset



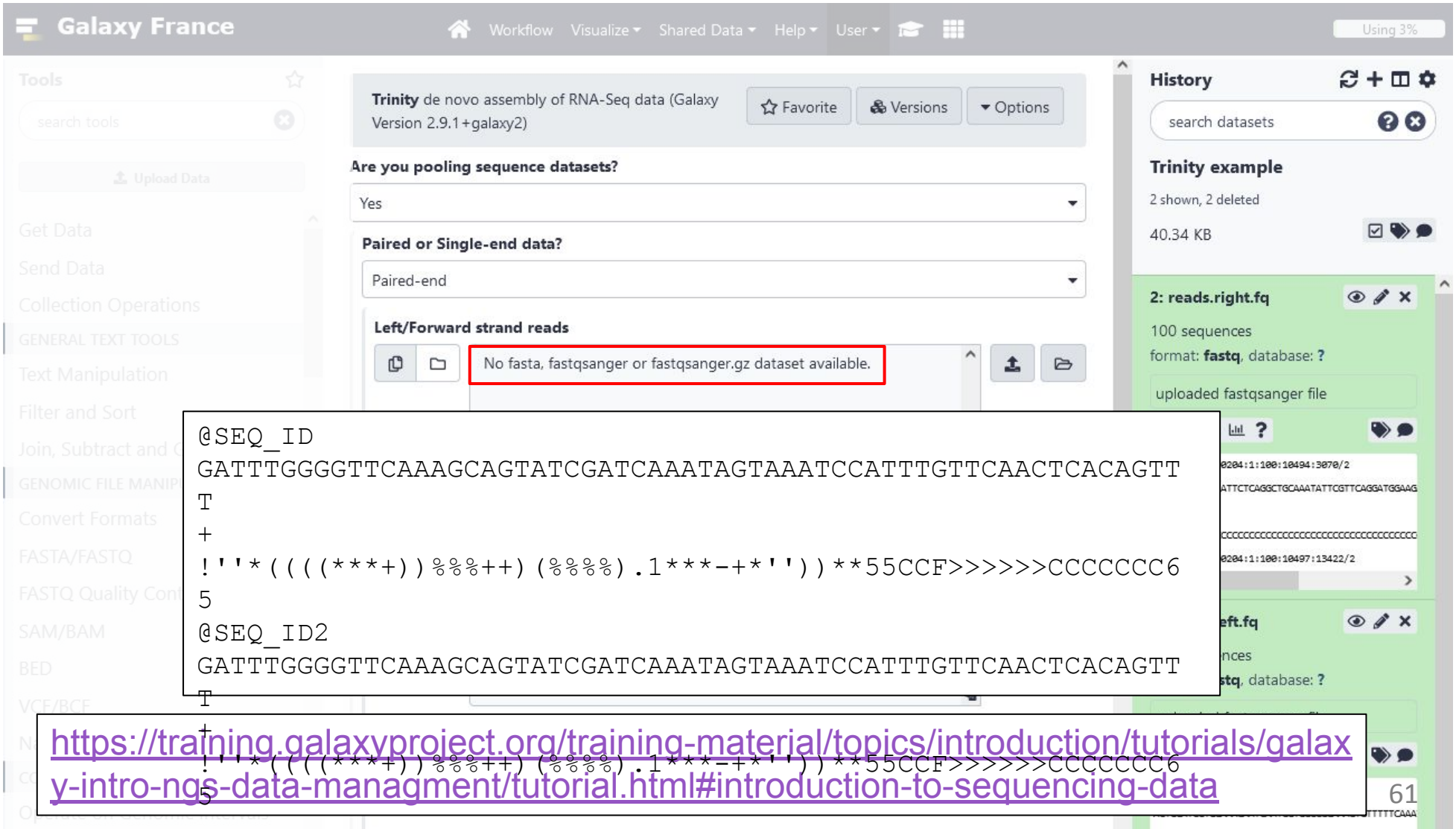
Dataset - Datatypes

- Many tools will only accept input datasets with the appropriate datatype assigned.



Dataset - Datatypes

- Many tools will only accept input datasets with the appropriate datatype assigned.



https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-ngs-data-managment/tutorial.html#introduction-to-sequencing-data

Hands-on

TOOLS (sequence files manipulation)

Tools (sequence files manipulation) – Hands-on

Part 1:

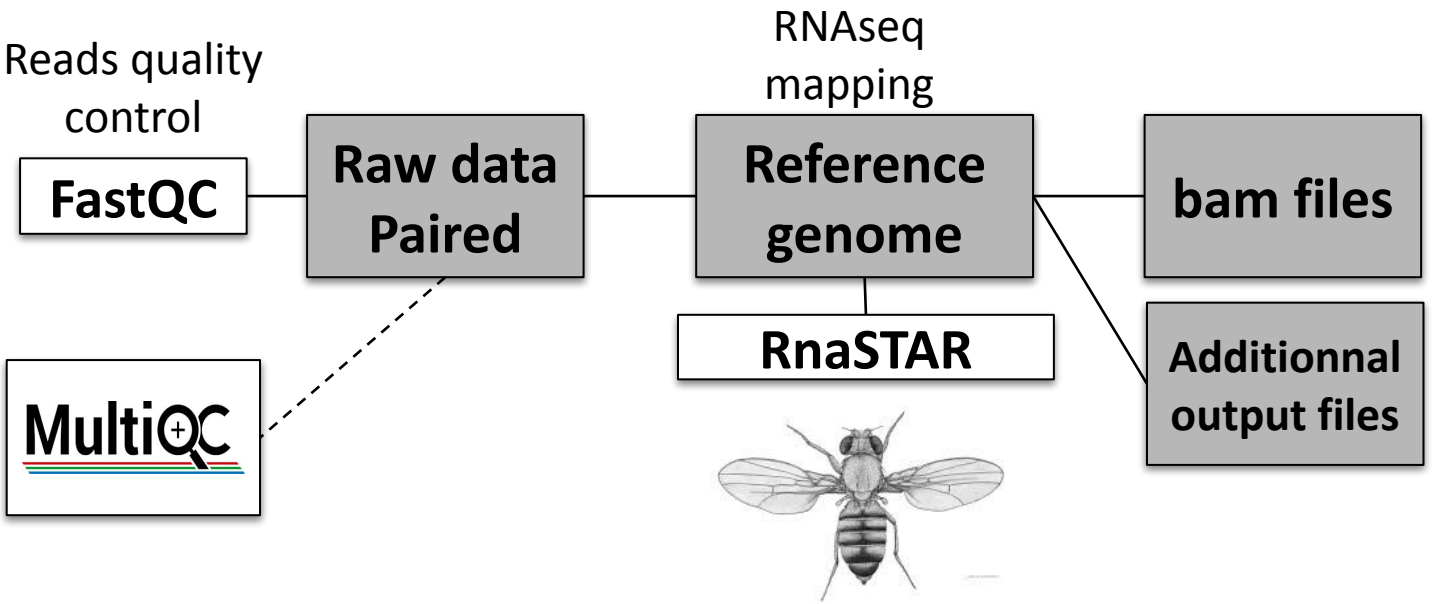
You have sequencing data from your favorite species. You want to check the quality of your sequences and to map on the reference genome.

Case of Drosophila melanogaster (dm)

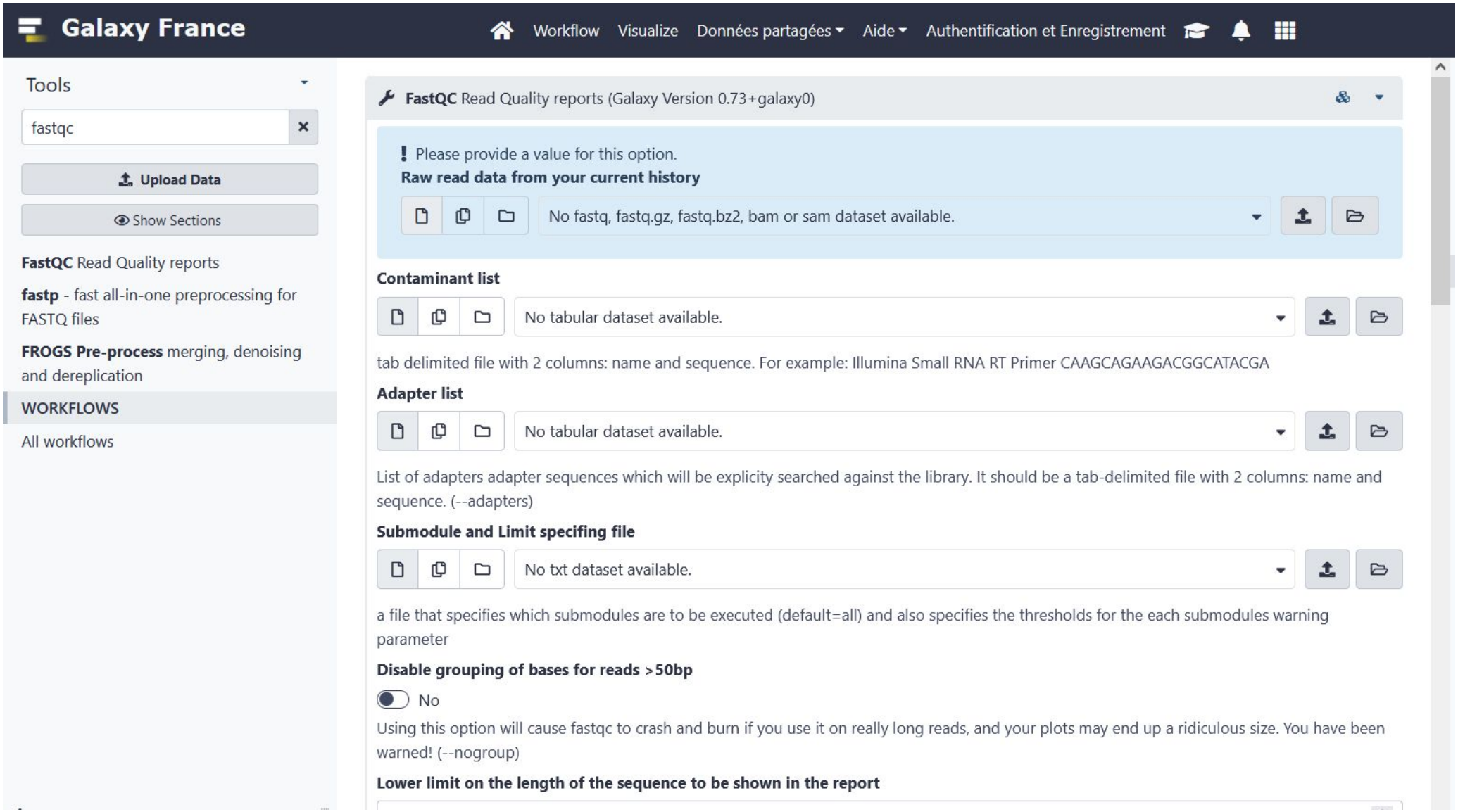
Part 1:

You have sequencing data from your favorite species. You want to check the quality of your sequences and to map on the reference genome.

Case of Drosophila melanogaster (dm)



Run Fastqc


Galaxy France
Workflow
Visualize
Données partagées
Aide
Authentification et Enregistrement
Tools
fastqc
Upload Data
Show Sections
FastQC Read Quality reports
fastp - fast all-in-one preprocessing for FASTQ files
FROGS Pre-process merging, denoising and dereplication
WORKFLOWS
All workflows
FastQC Read Quality reports (Galaxy Version 0.73+galaxy0)
Please provide a value for this option.
Raw read data from your current history
No fastq, fastq.gz, fastq.bz2, bam or sam dataset available.
Contaminant list
No tabular dataset available.
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA
Adapter list
No tabular dataset available.
List of adapters adapter sequences which will be explicity searched against the library. It should be a tab-delimited file with 2 columns: name and sequence. (--adapters)
Submodule and Limit specifing file
No txt dataset available.
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter
Disable grouping of bases for reads >50bp
No
Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)
Lower limit on the length of the sequence to be shown in the report

Part 1:

You have sequencing data from your favorite species. You want to check the quality of your sequences and to map on the reference genome.

Case of Drosophila melanogaster (dm)

1. Look into the description of the tool FastQC, what is its purpose ?
2. Does it takes compressed files ? ".gz"
3. Which encoding of the file (in FastQC results, basic statistics) ? => edit if needed the datatype of your reads files
4. Look at mapping tools, map your reads on genome with RNA STAR.
5. What is the default parameter of the option "Maximum ratio of mismatches to mapped length" in RNA STAR ?
6. How to map on another reference genome or assembly release ?

Tools (sequence files manipulation) – Hands-on

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases
5. Run BLAST against the CDS database
6. Run BLAST against the protein database

Tools (sequence files manipulation) – Hands-on

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases *Blast --> NCBI BLAST+ makeblastdb*
5. Run BLAST against the CDS database
6. Run BLAST against the protein database

Tools (sequence files manipulation) – Hands-on

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases *Blast --> NCBI BLAST+ makeblastdb*
5. Run BLAST against the CDS database *Blast --> NCBI BLAST+ blastn*
6. Run BLAST against the protein database

Tools (sequence files manipulation) – Hands-on

Part 2:

You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases *Blast --> NCBI BLAST+ makeblastdb*
5. Run BLAST against the CDS database *Blast --> NCBI BLAST+ blastn*
6. Run BLAST against the protein database *Blast --> NCBI BLAST+ blastx*

DATASET COLLECTION

Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

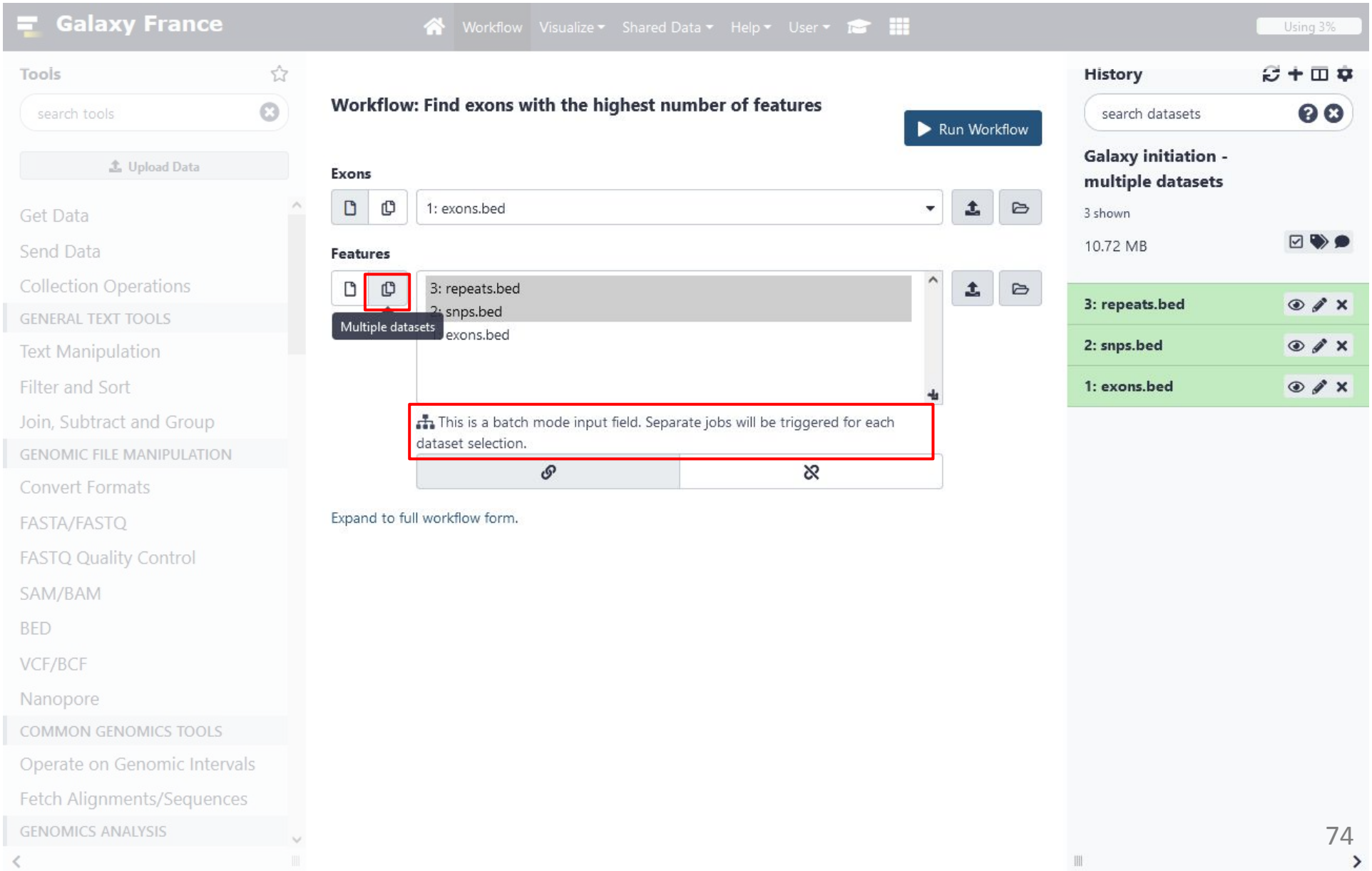
Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input

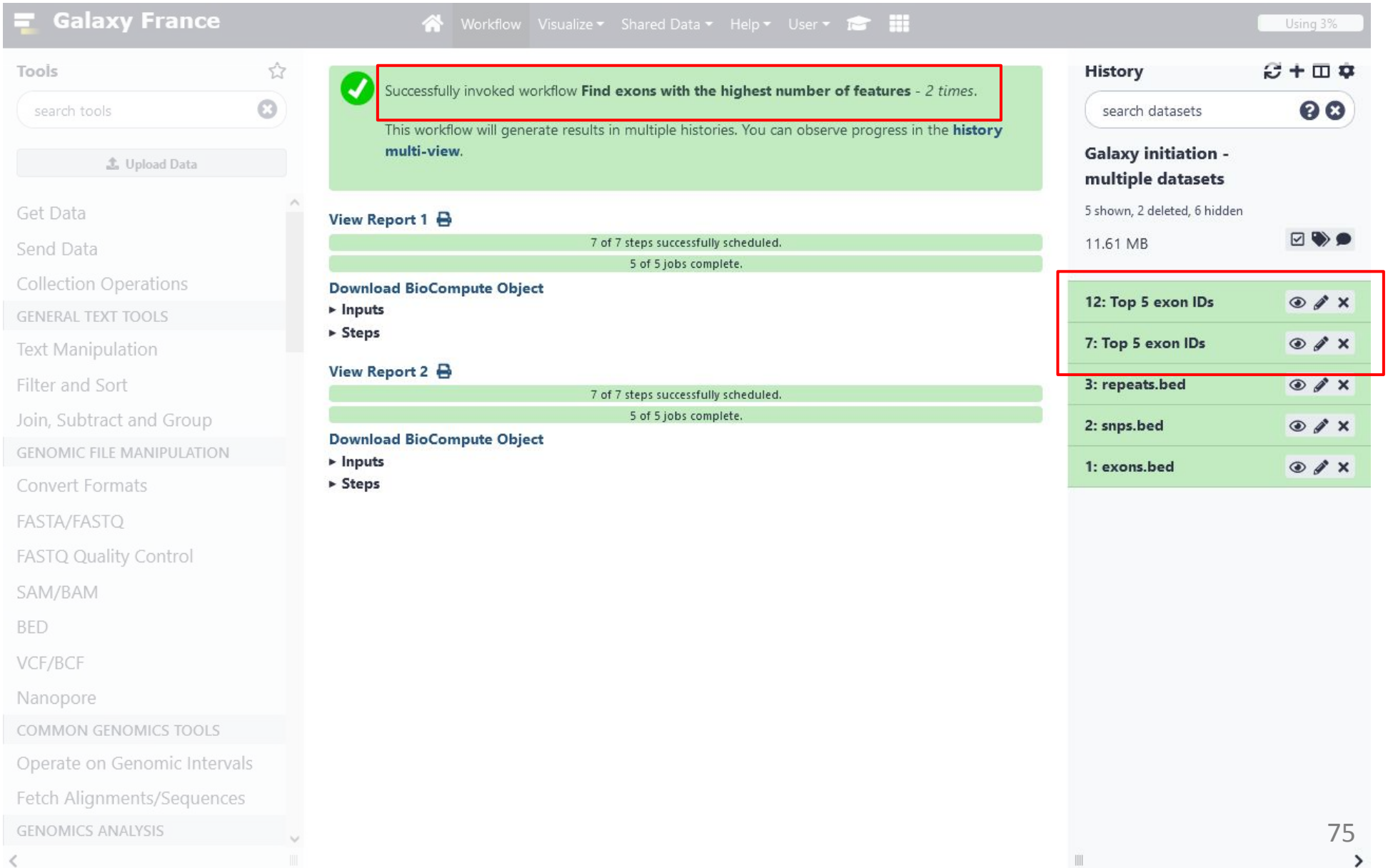
Dataset collection

Select multiple datasets as input



Dataset collection

Select multiple datasets as input



Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input
- Solution 2: create a dataset collection (any number of datasets bundled as a **single entity**, i.e. minimize clutter)

Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

- Solution 1: select multiple datasets as input
- Solution 2: create a dataset collection (any number of datasets bundled as a **single entity**, i.e. minimize clutter)
  - Dataset list: set of files of the same type
  - Dataset pairs: pairs of read files (forward, reverse)
  - List of dataset pairs

Dataset collection

- Problematic: you have a large numbers of datasets to send through the same analysis

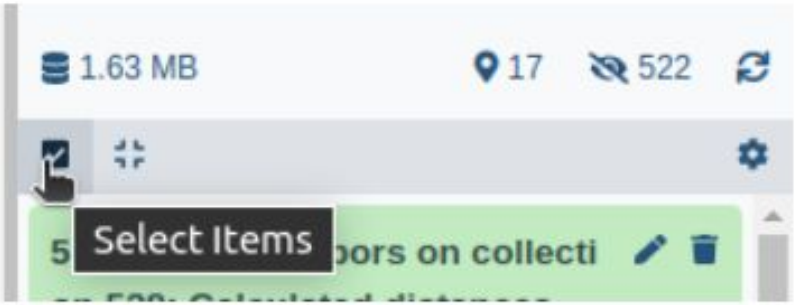
- Solution 1: select multiple datasets as input
- Solution 2: create a dataset collection (any number of datasets bundled as a **single entity**, i.e. minimize clutter)
  - Dataset list: set of files of the same type
  - Dataset pairs: pairs of read files (forward, reverse)
  - List of dataset pairs

- Galaxy runs the tool automatically on each dataset using the same settings

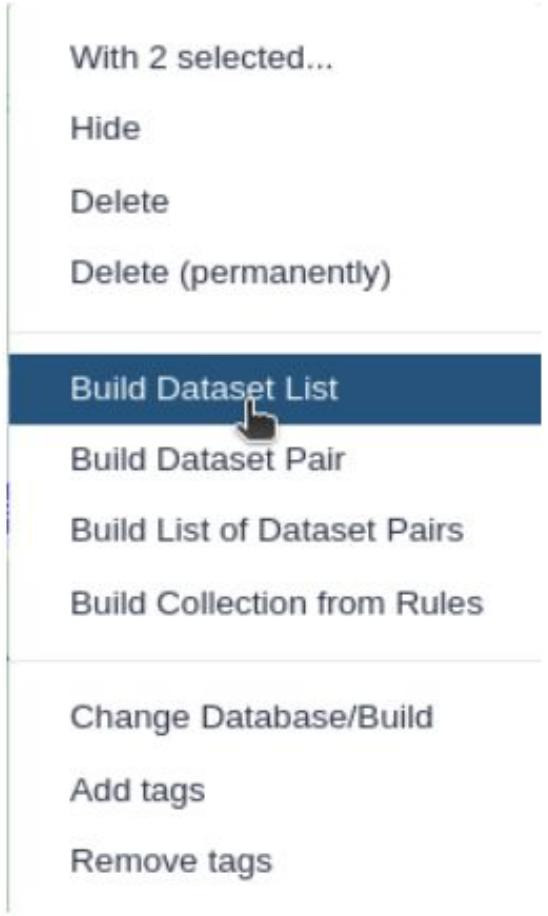
Dataset collection

Creating a collection from datasets in your history. Renaming. Tagging.

- Click on ☑ **Select Items** at the top of the history panel



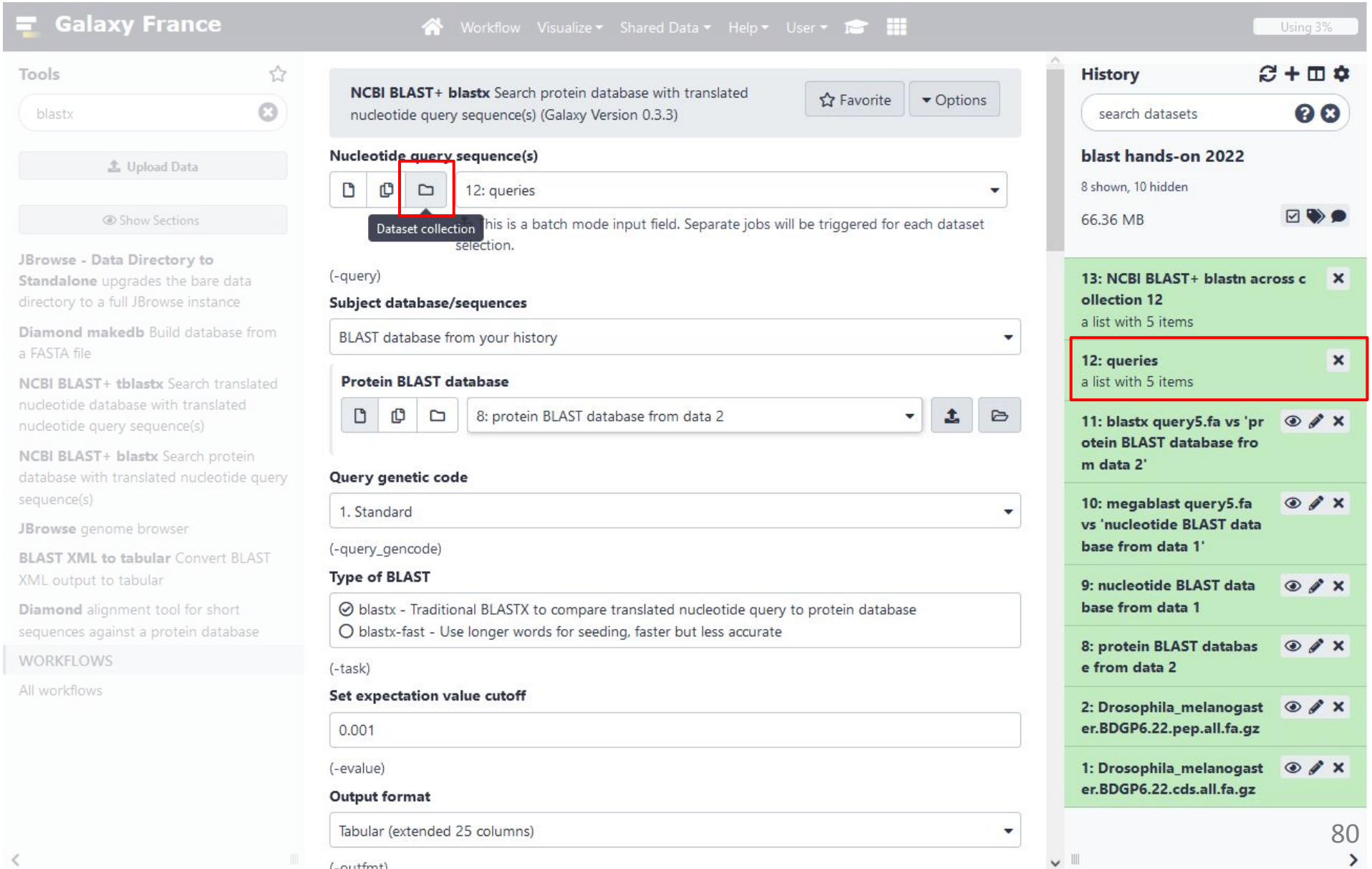
- Check all the datasets in your history you would like to include
- Click **n of N selected** and choose **Build Dataset List**
- Enter a name for your collection
- Click **Create collection** to build your collection
- Click on the checkmark icon at the top of your history again



https://training.galaxyproject.org/training-material/faqs/galaxy/#collections

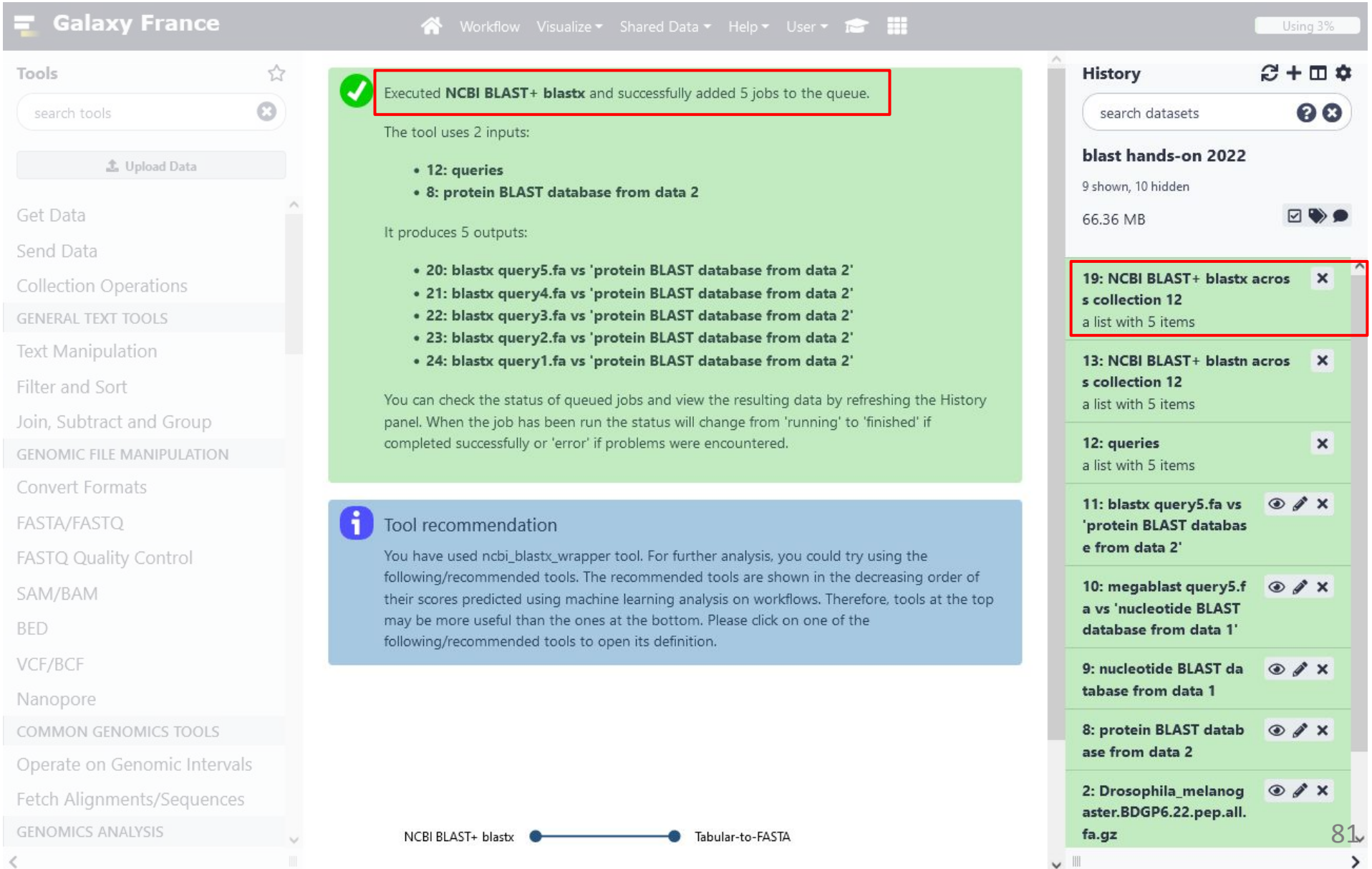
Dataset collection

Use a collection as input



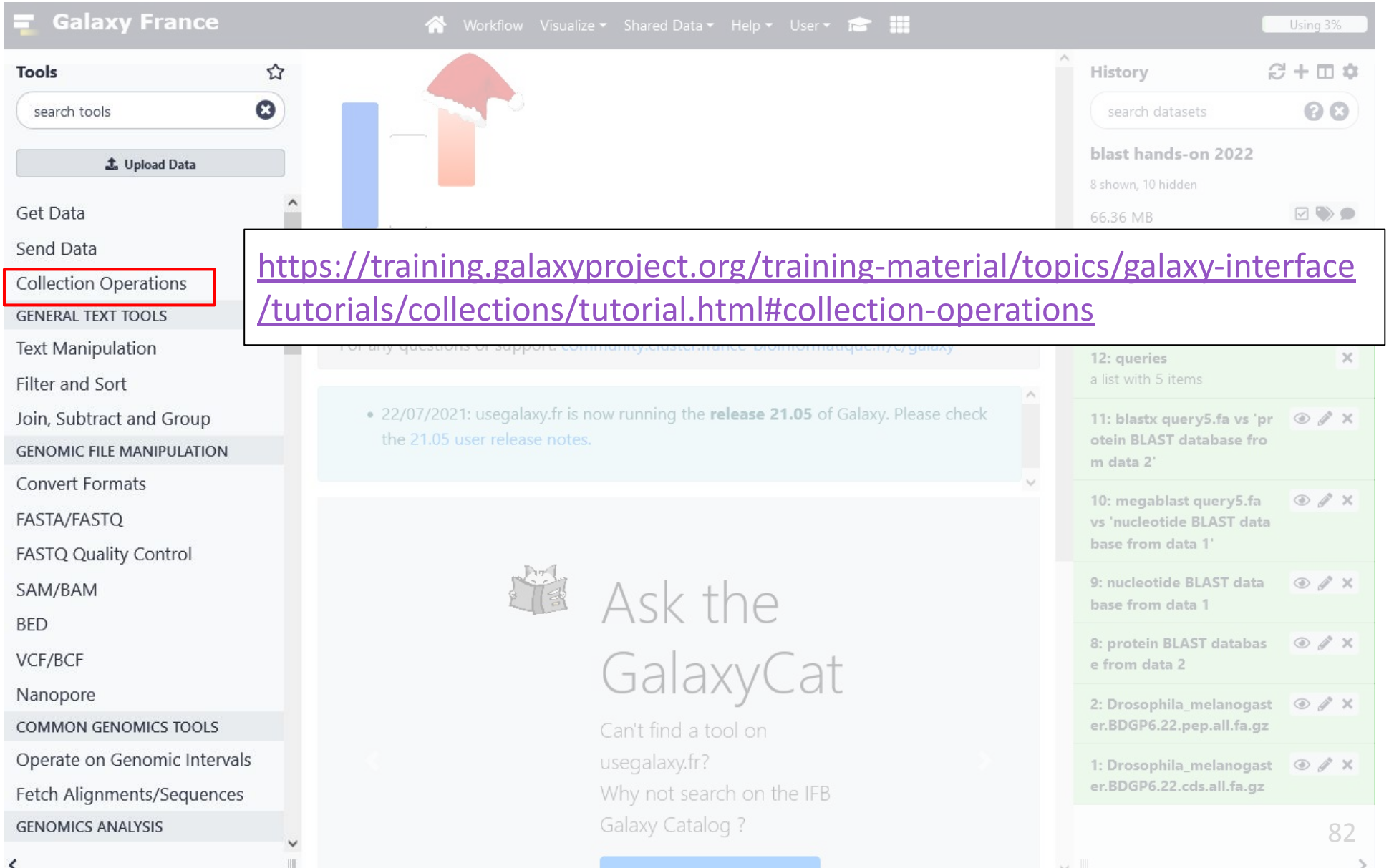
Dataset collection

Use a collection as input



Dataset collection

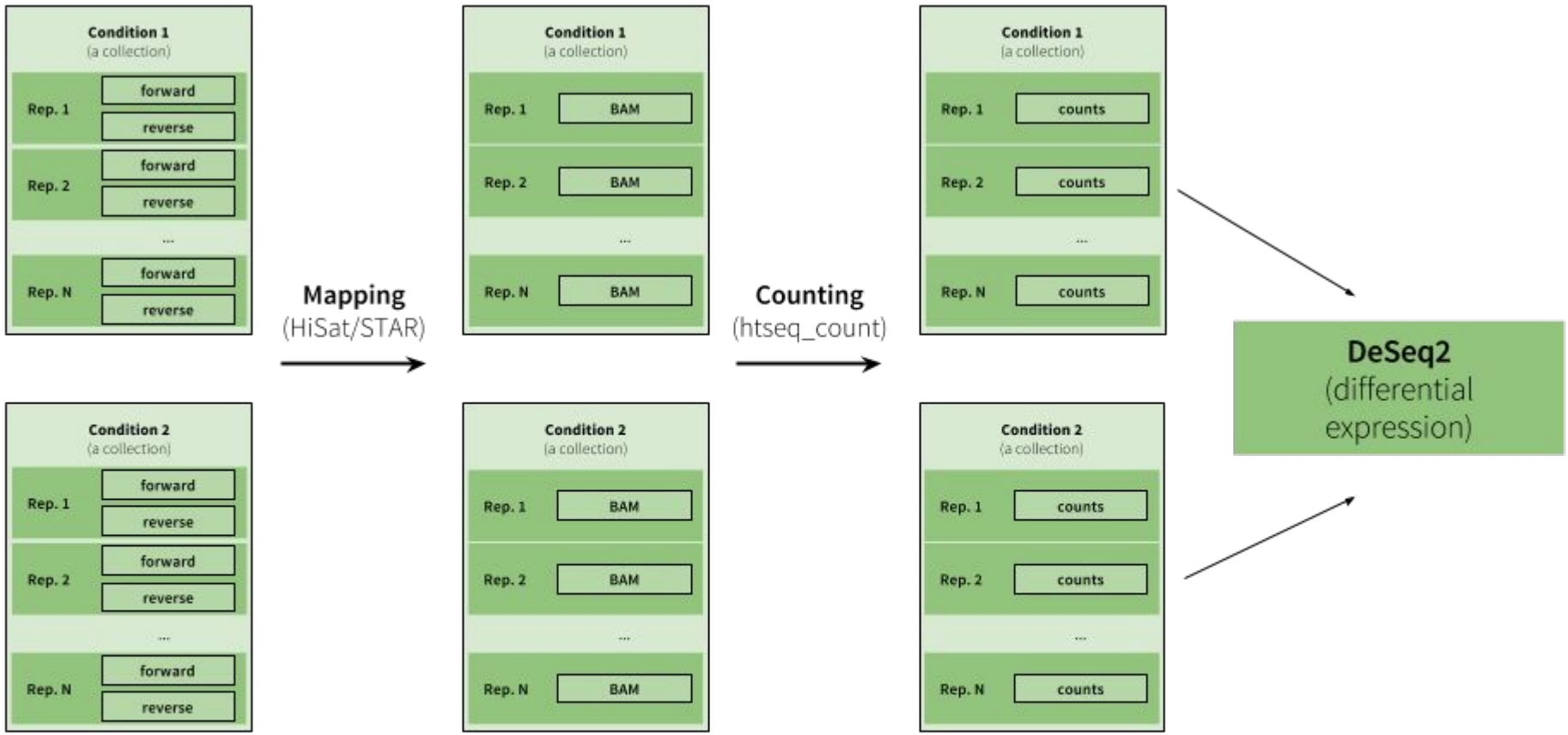
Use dataset collection tools:



https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/collections/tutorial.html#collection-operations

Dataset collection

Use-case: RNA-seq analysis



FYI: Interesting resource on dataset collections

https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/collections/tutorial.html

Using dataset collections

Authors: Anton Nekrutenko | Add Contributions!

Overview

Questions:
- How to manipulate large numbers of datasets at once?

Objectives:
- Understand and master dataset collections

Time estimation: 30 minutes

Level: Intermediate

Supporting Materials:

Datasets | GTN Video Library

Last modification: Nov 16, 2021

Here we will show Galaxy features designed to help with the analysis of large numbers of samples. When you have just a few samples - clicking through them is easy. But once you've got hundreds - it becomes very annoying. In Galaxy we have introduced **Dataset collections** that allow you to combine numerous datasets in a single entity that can be easily manipulated.

Getting data

First, we need to upload datasets. Cut and paste the following URLs to Galaxy upload tool (see a **Tip** on how to do this below).

```
https://zenodo.org/record/5119008/files/M117-bl_1.fq.gz
https://zenodo.org/record/5119008/files/M117-bl_2.fq.gz
https://zenodo.org/record/5119008/files/M117-ch_1.fq.gz
https://zenodo.org/record/5119008/files/M117-ch_2.fq.gz
https://zenodo.org/record/5119008/files/M117C1-bl_1.fq.gz
https://zenodo.org/record/5119008/files/M117C1-bl_2.fq.gz
```

Hands-on

COLLECTION

Collection – Hands-on

Part 2:

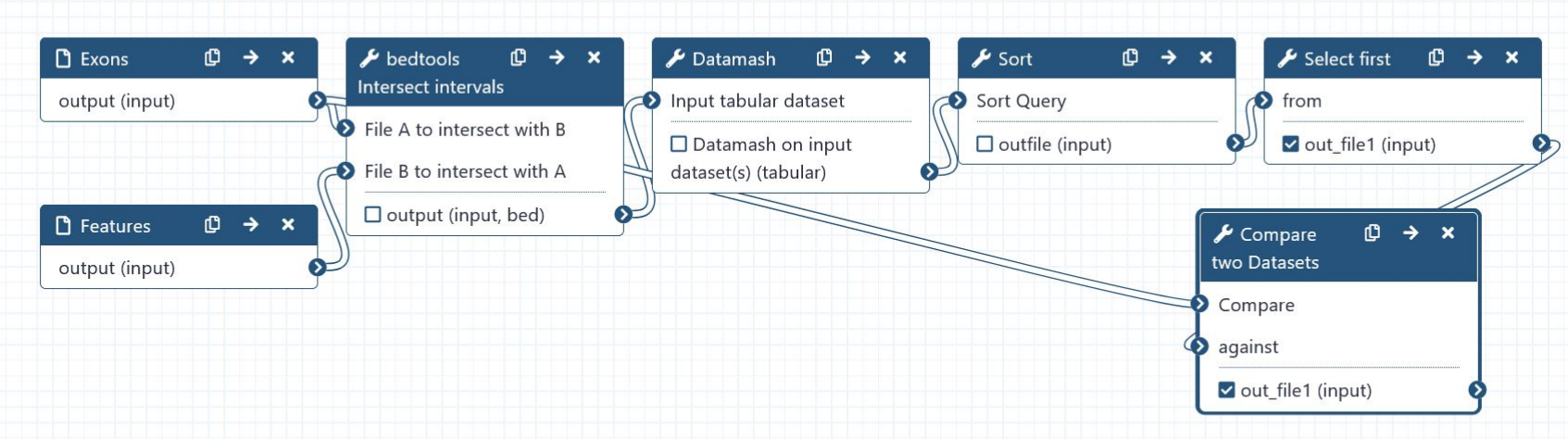
You have new sequences that you want to compare with the gene and protein databases from your favorite species (BLAST).

Case of Drosophila melanogaster (dm)

1. Create new history
2. Import CDS and peptide sequences databases
3. Import query sequences
4. Make BLAST databases
5. Run BLAST against the CDS database
6. Run BLAST against the protein database
7. Create a dataset list with all the query sequences
8. Run BLAST against the CDS database on the dataset list

WORKFLOW

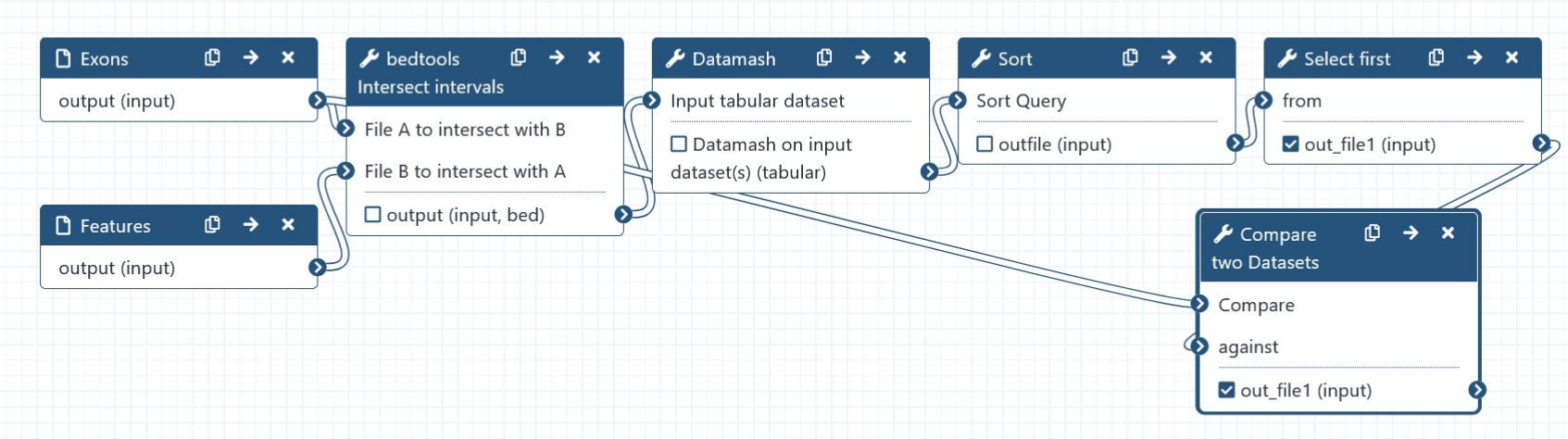
Workflow



What?

- A sequence of tool operations and parameters

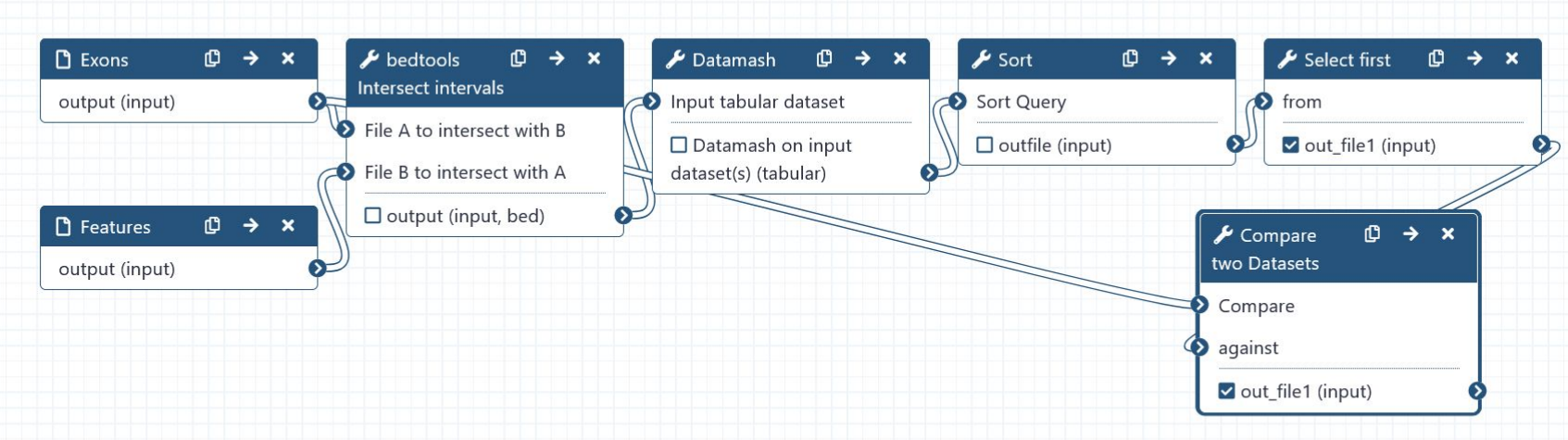
Workflow



Why?

- Re-run the same analysis on different input data sets
- Change parameters before re-running a similar analysis
- Make use of the workflow job scheduling (jobs are submitted as soon as their inputs are ready)
- Share workflows for publication and with the community

Workflow

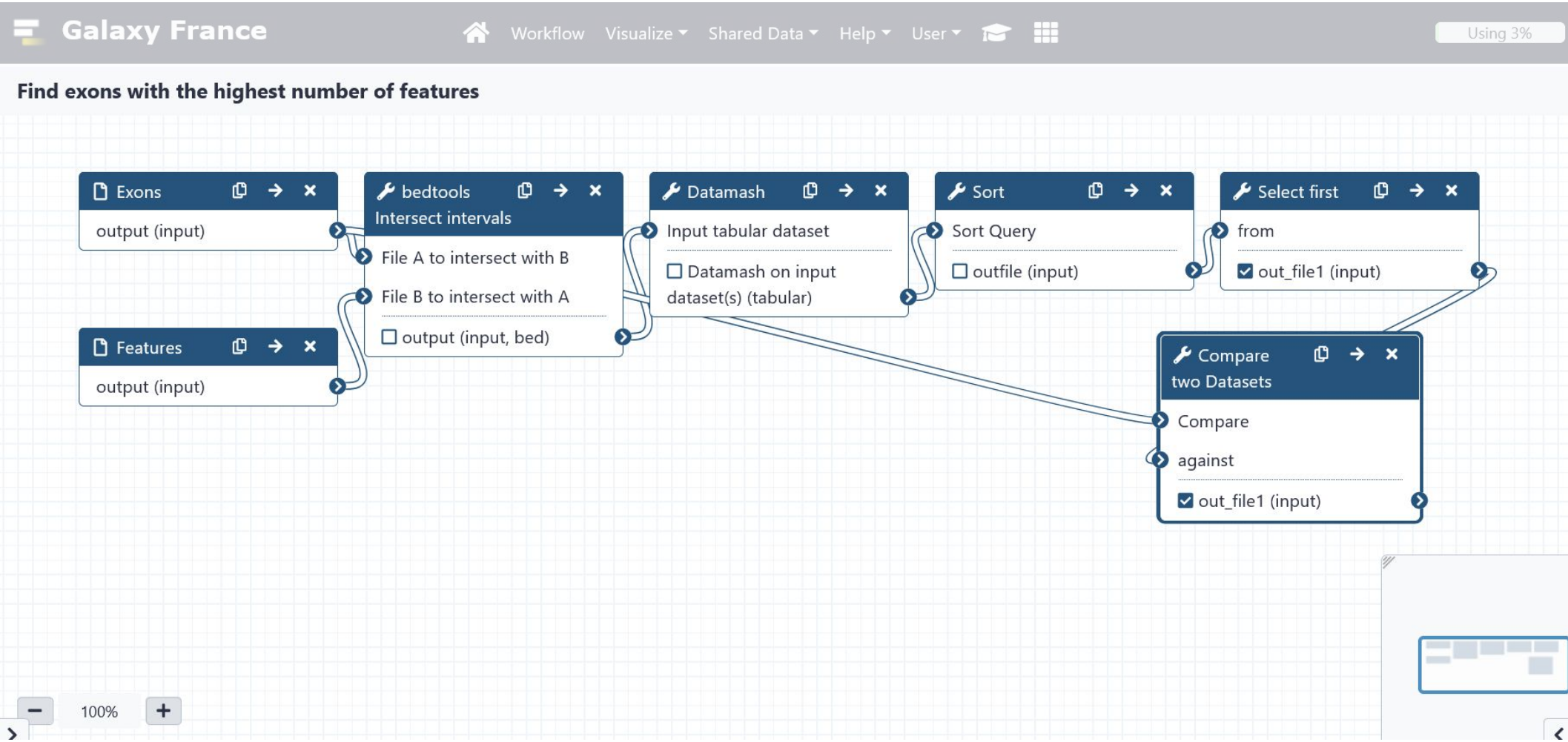


How?

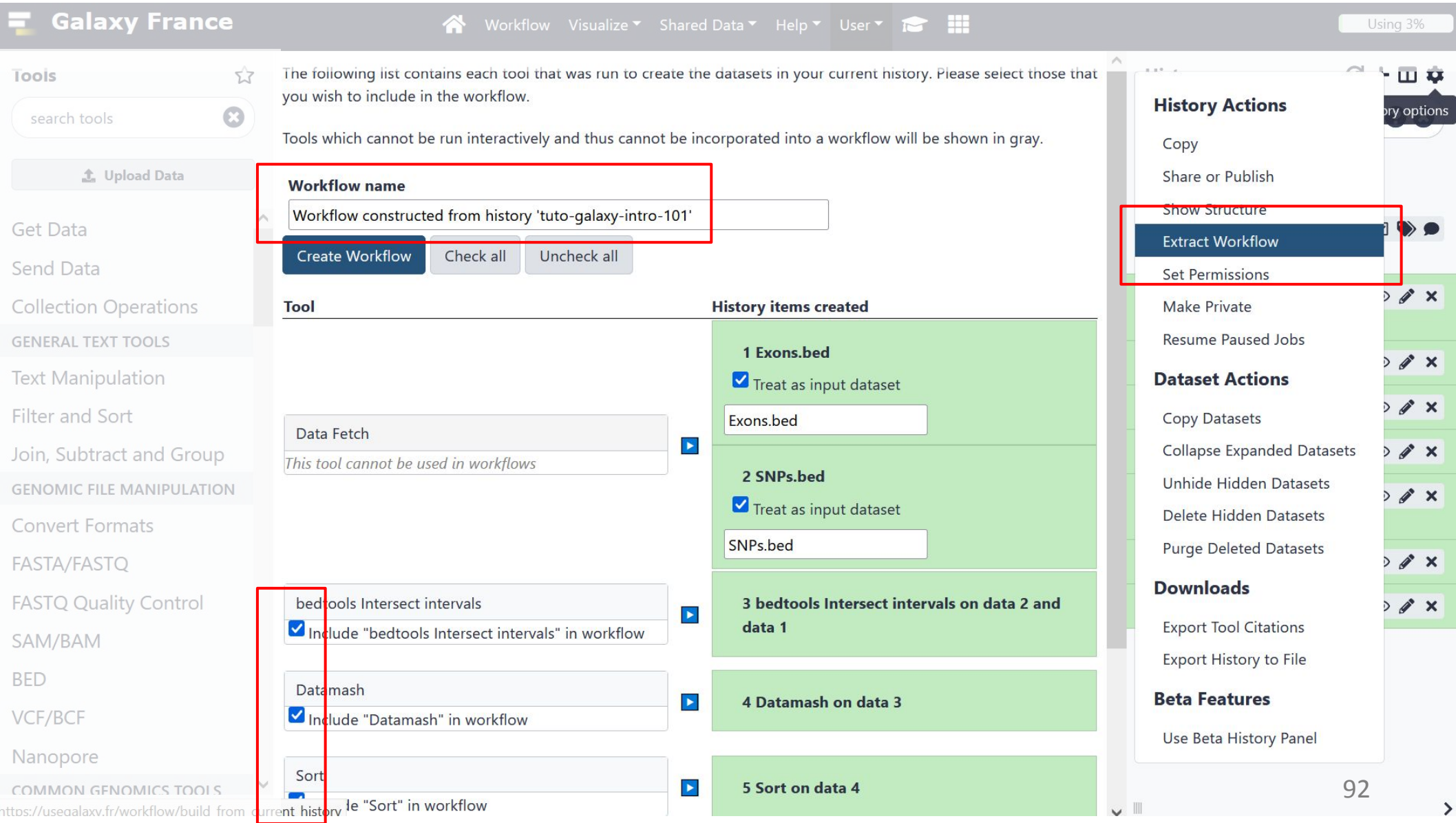
- Extracted from a history
- Built manually by adding and configuring tools using the canvas
- Imported using an existing shared workflow

Workflow

Our workflow with Galaxy

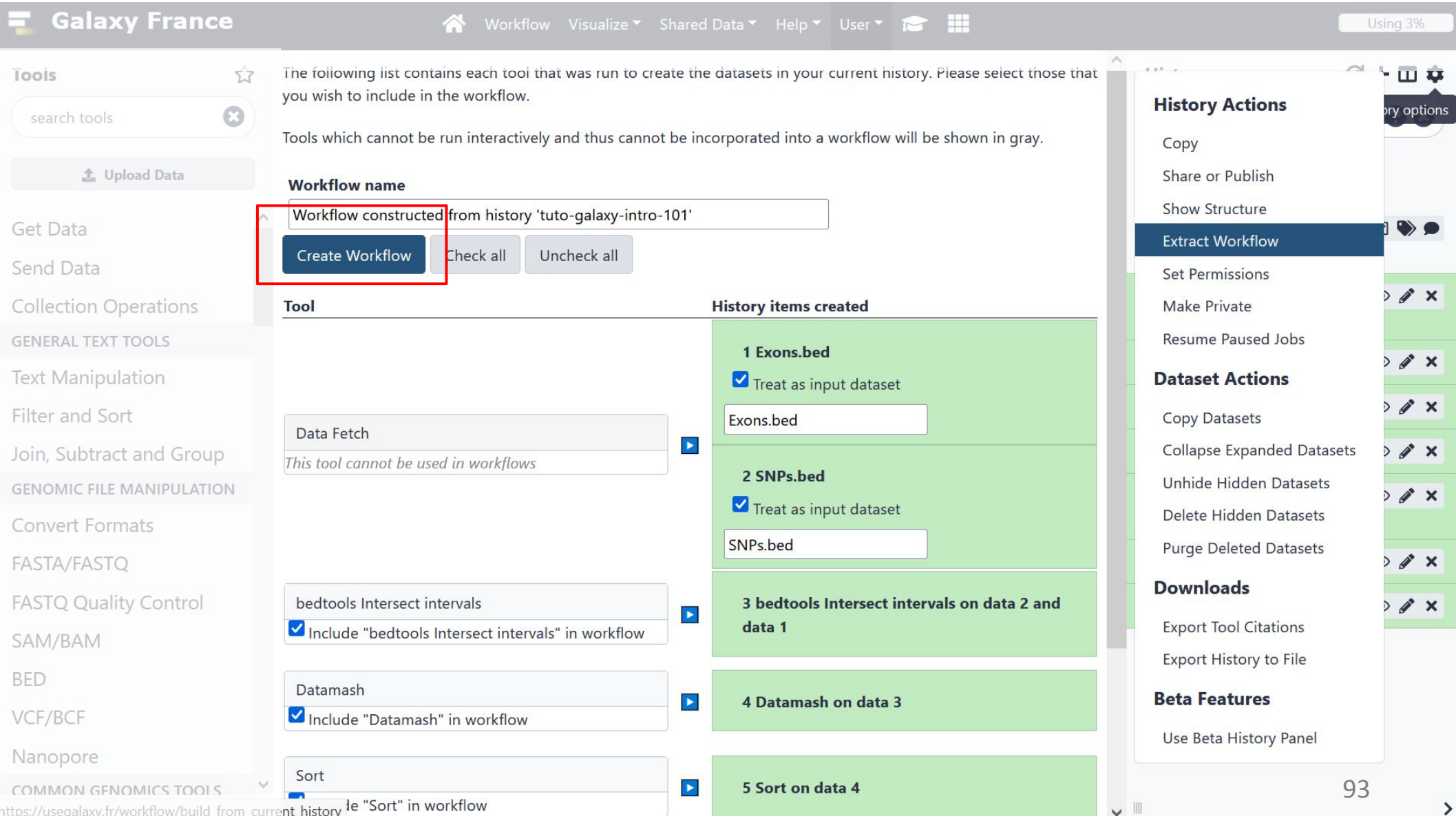


From history



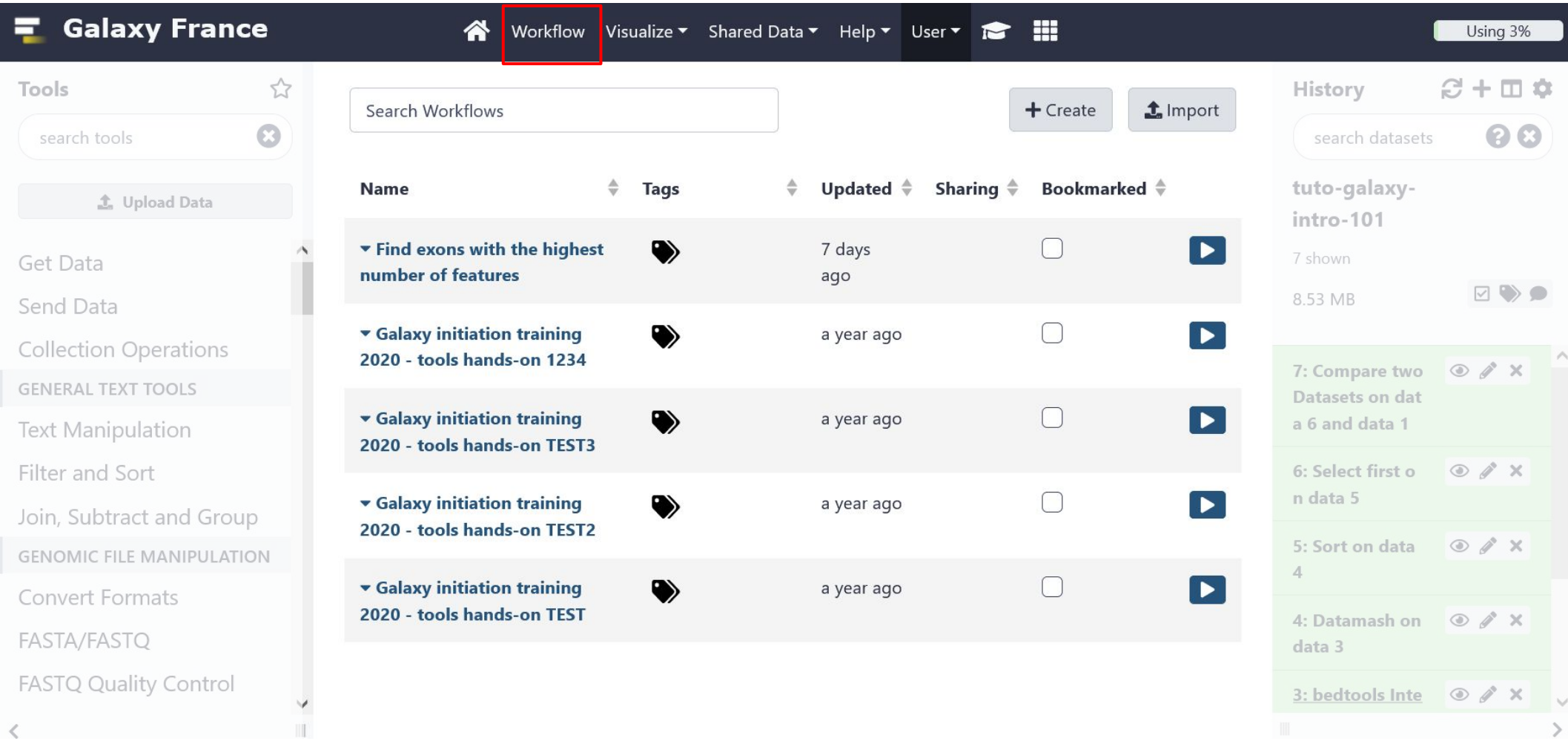
Workflow

From history



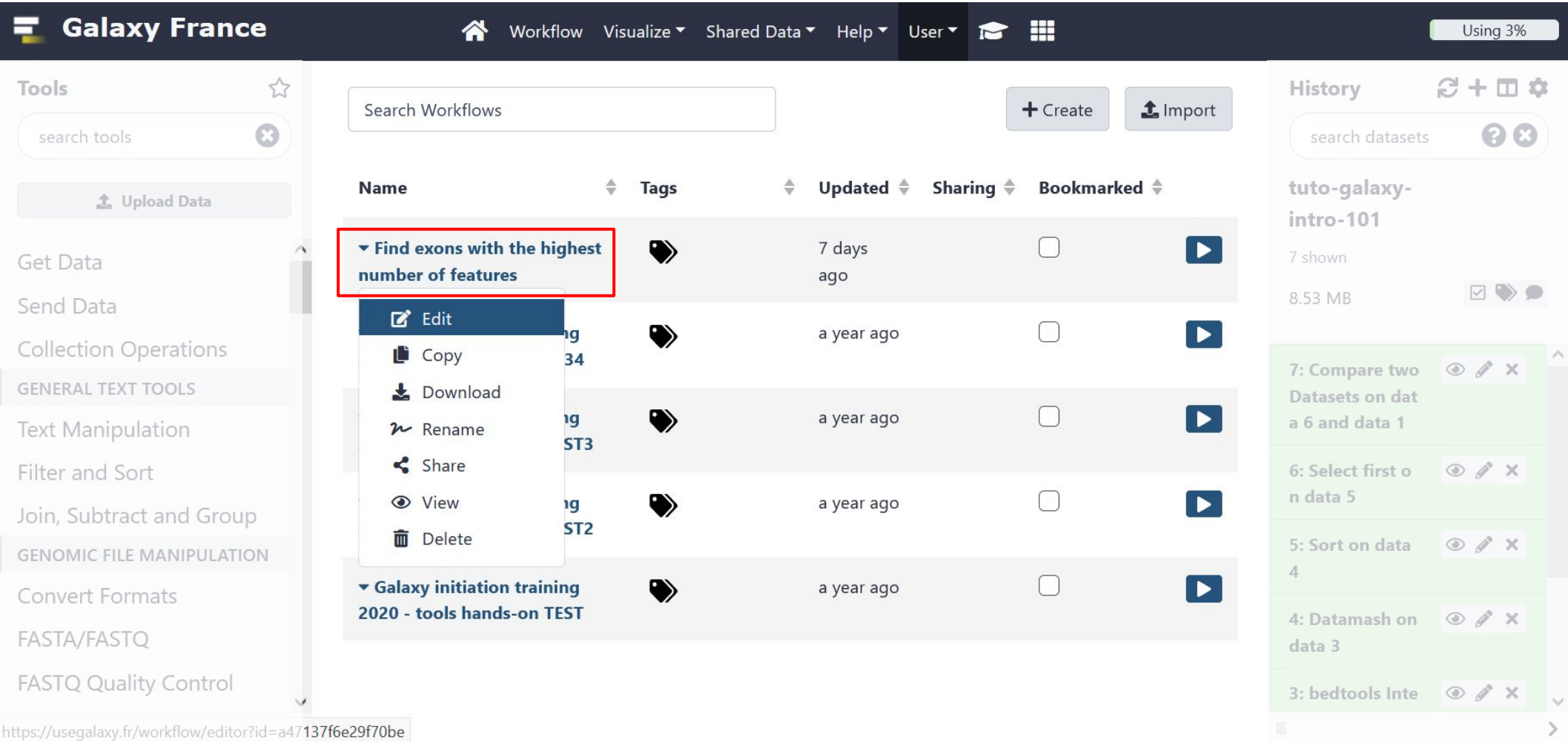
Workflow

Workflow manager



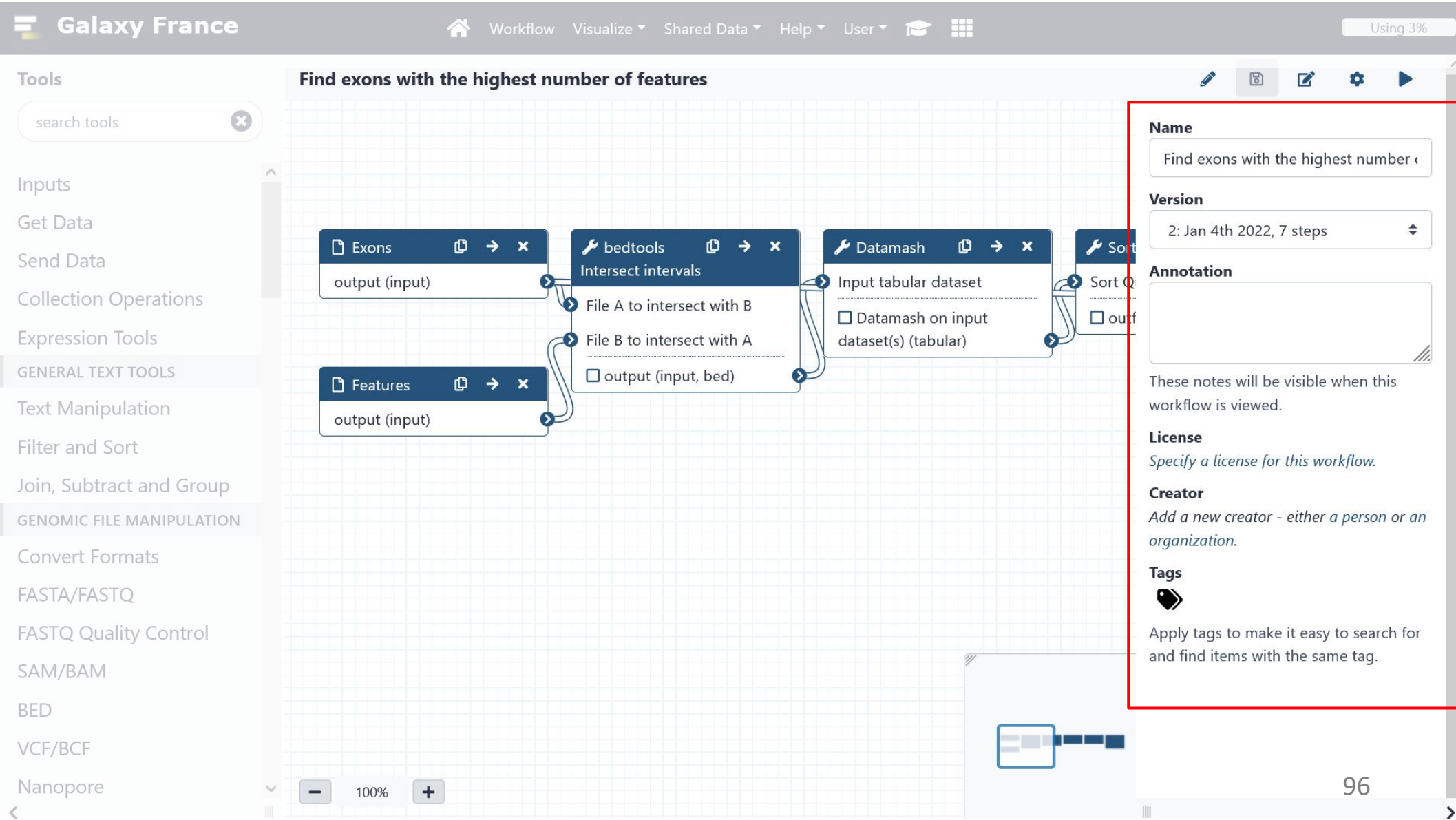
Workflow

Workflow manager



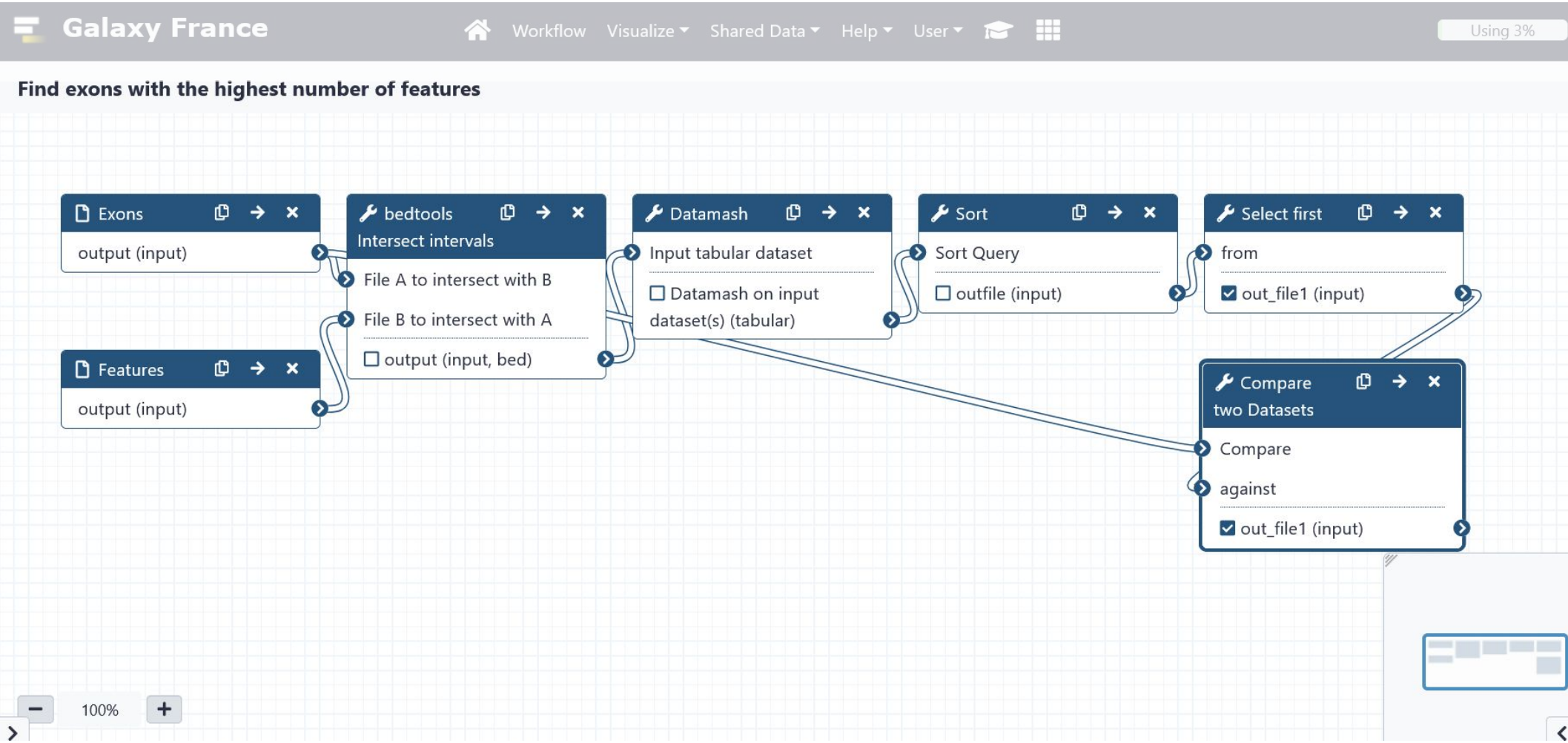
Workflow

Edit a workflow: attributes



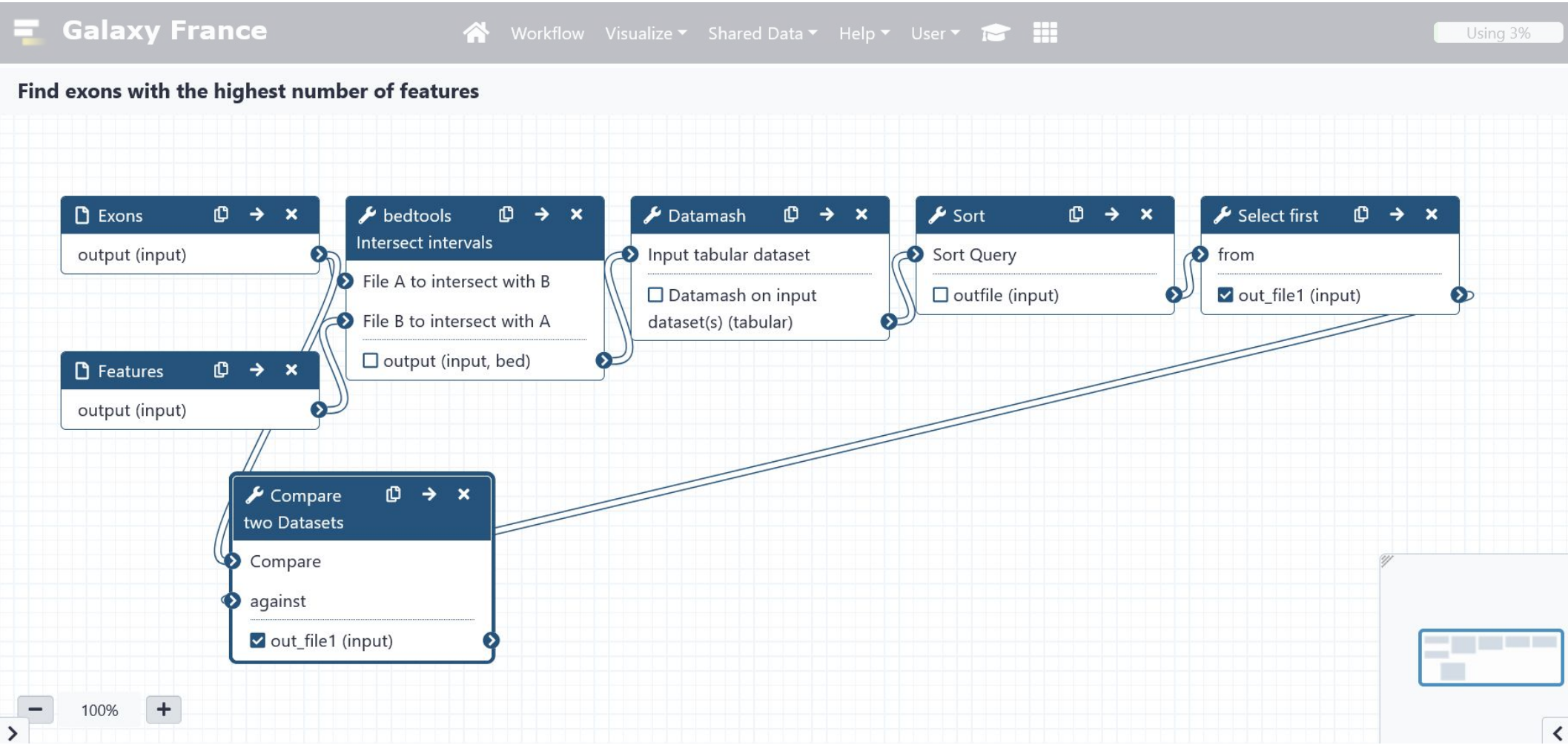
Workflow

Edit a workflow: drag and drop



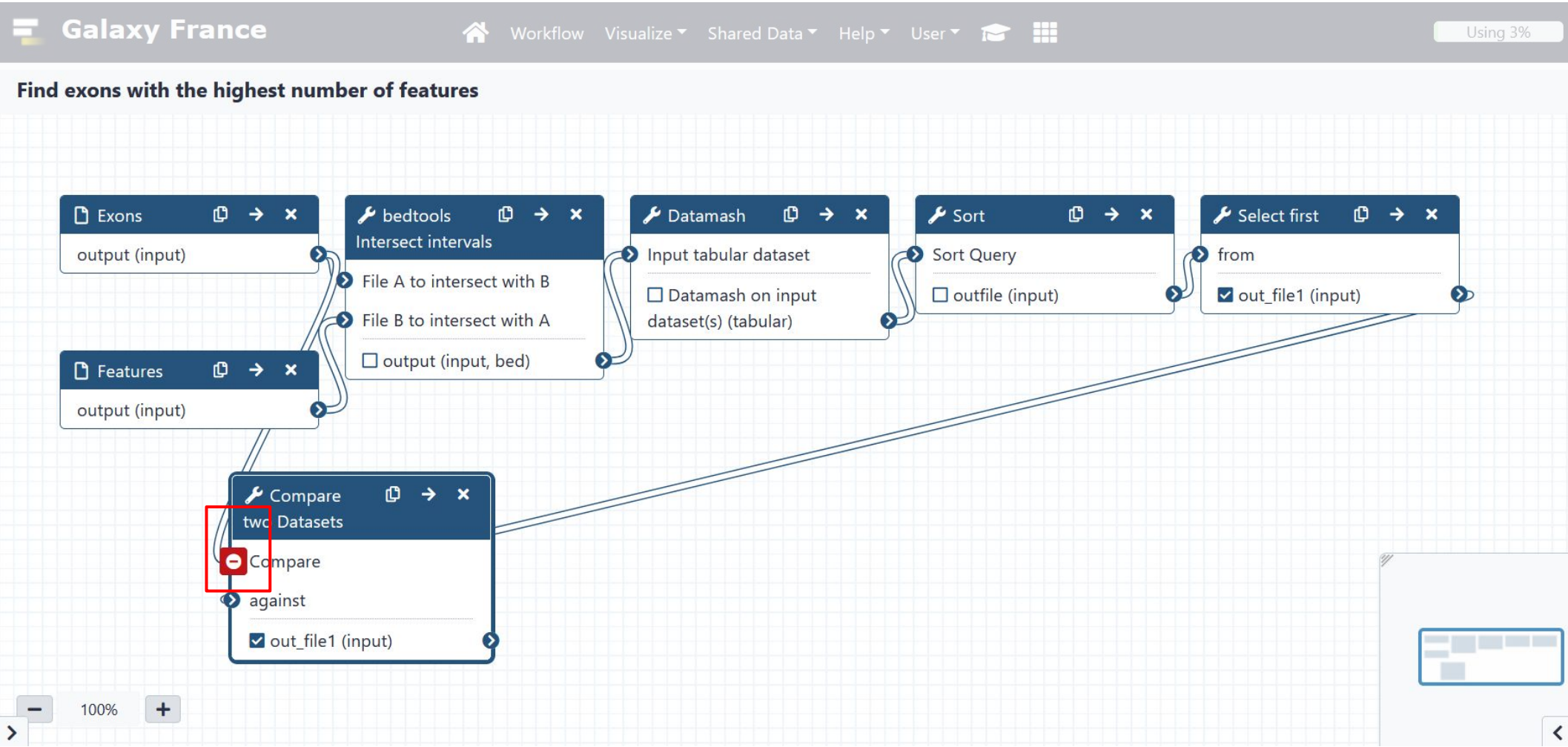
Workflow

Edit a workflow: drag and drop

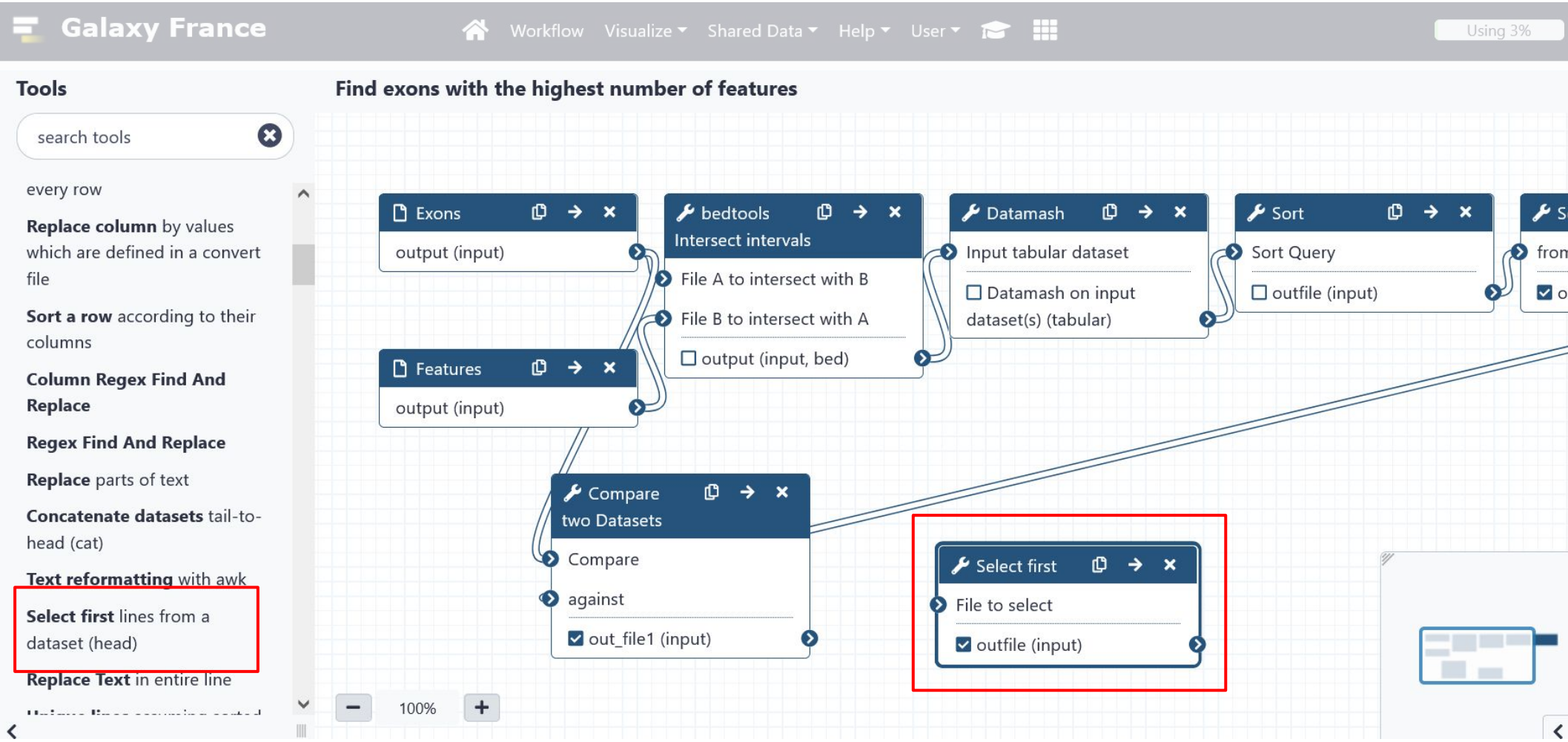


Workflow

Edit a workflow: delete a noodle

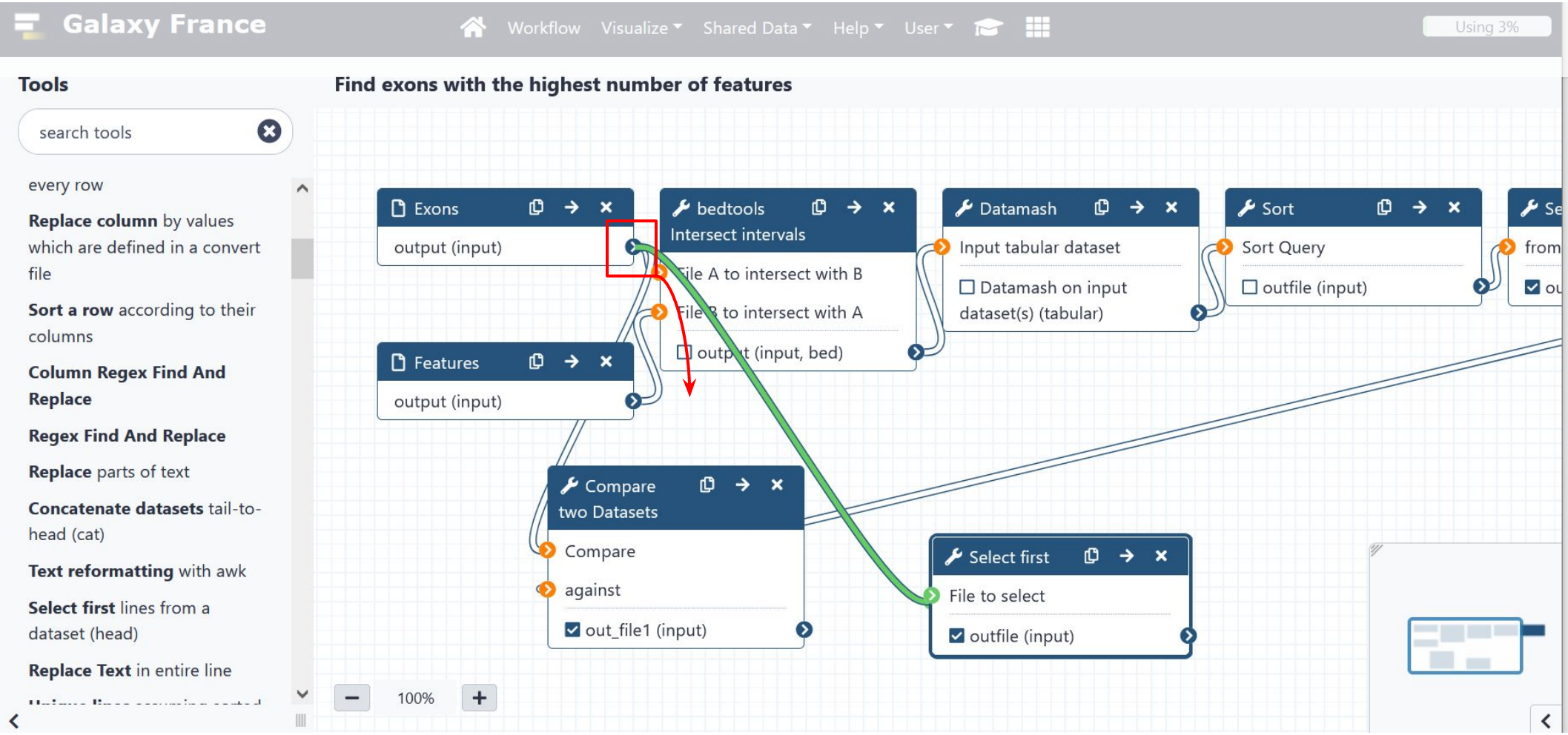


Edit a workflow: add a tool

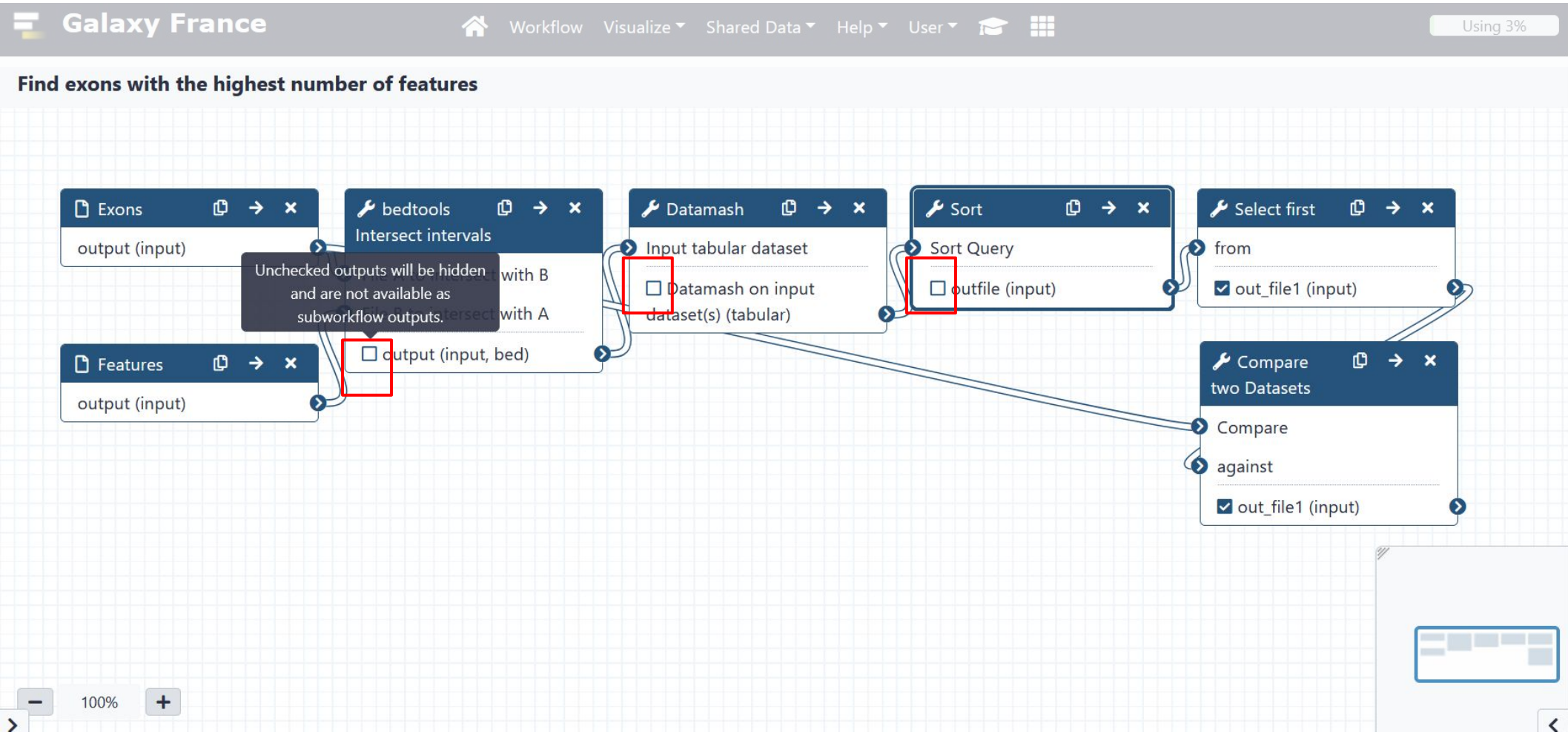


Workflow

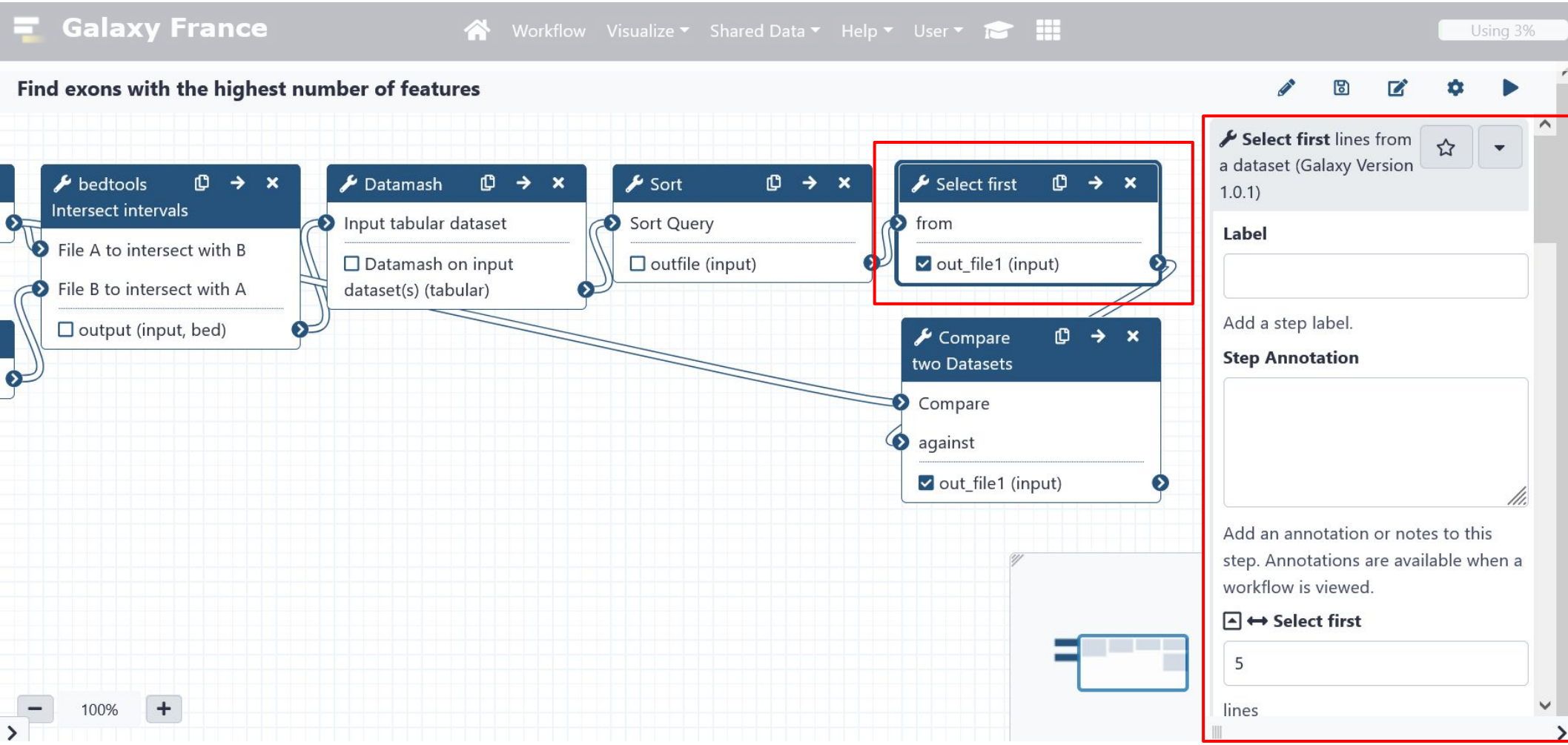
Edit a workflow: add a noodle



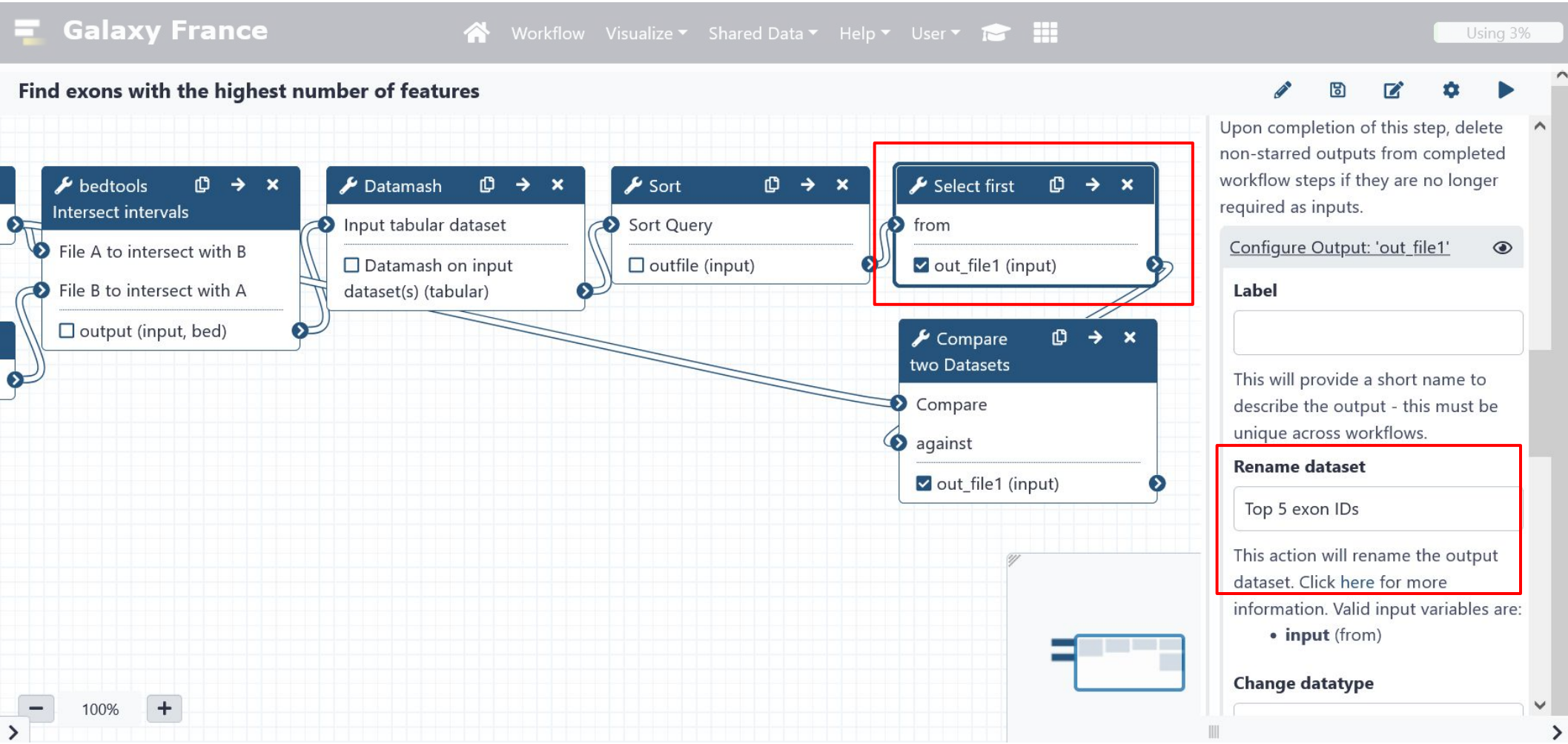
Edit a workflow: hide intermediate steps



Edit a workflow: edit step

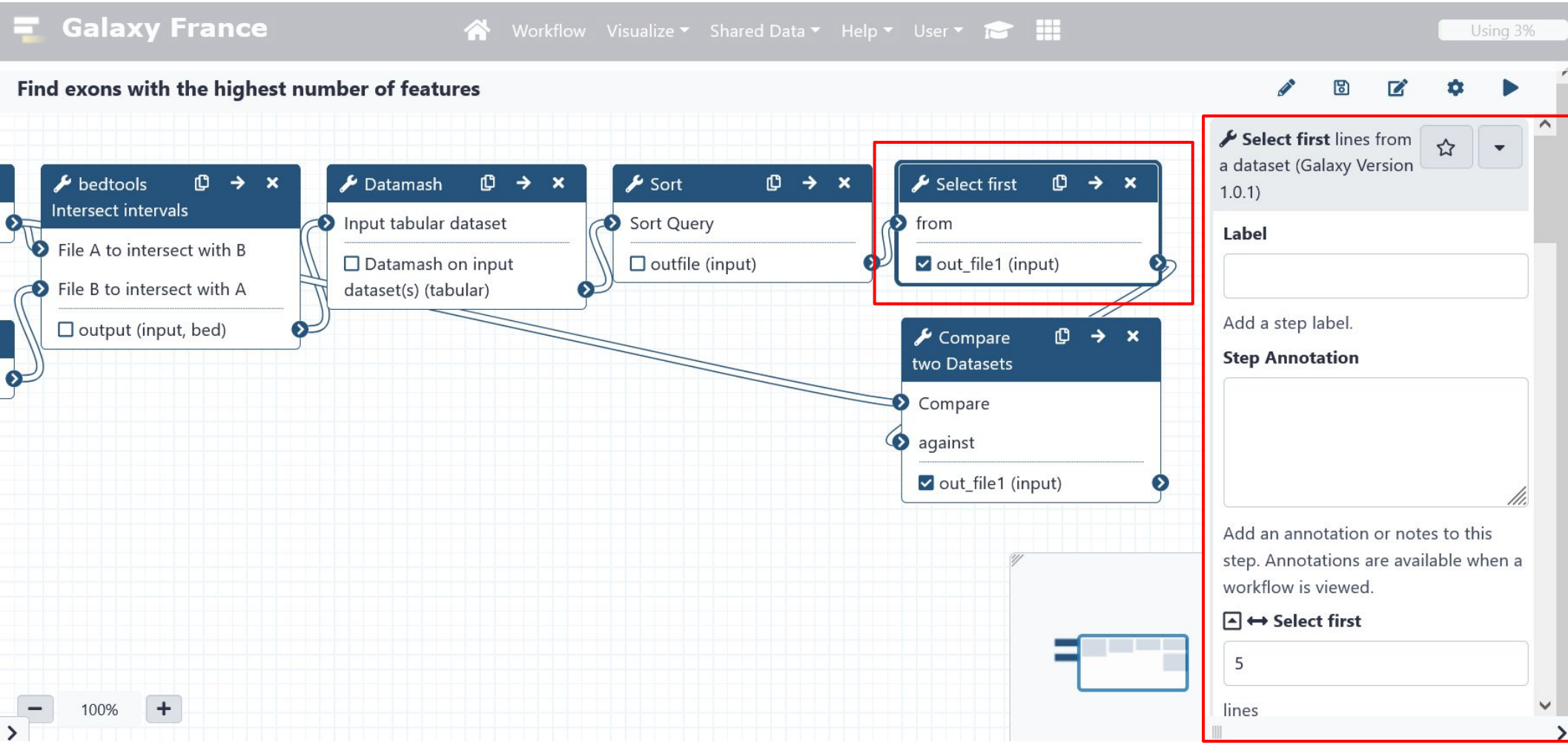


Edit a workflow: rename the outputs

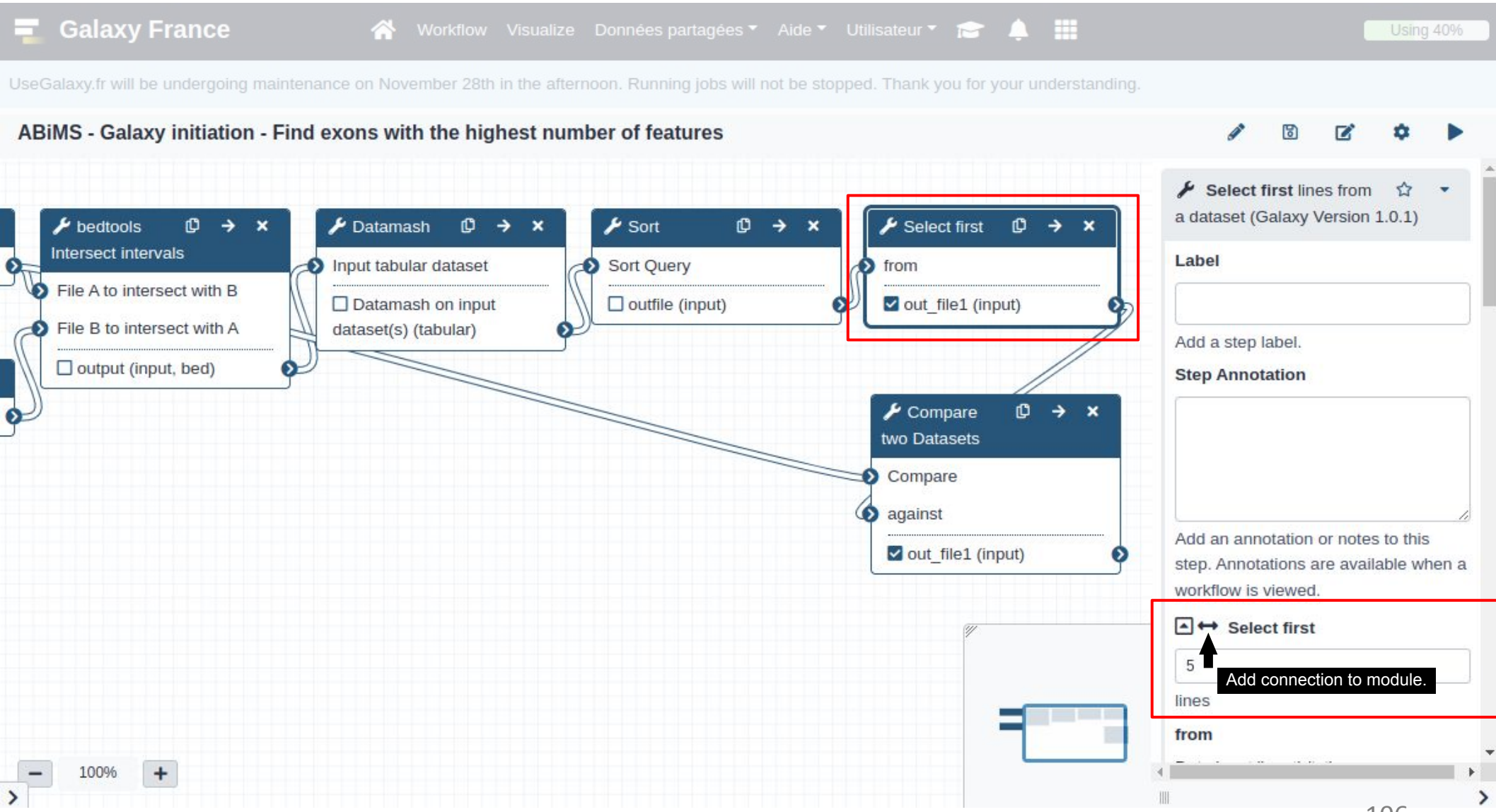


Workflow

Edit a workflow: use workflow parameters to make your workflow more versatile

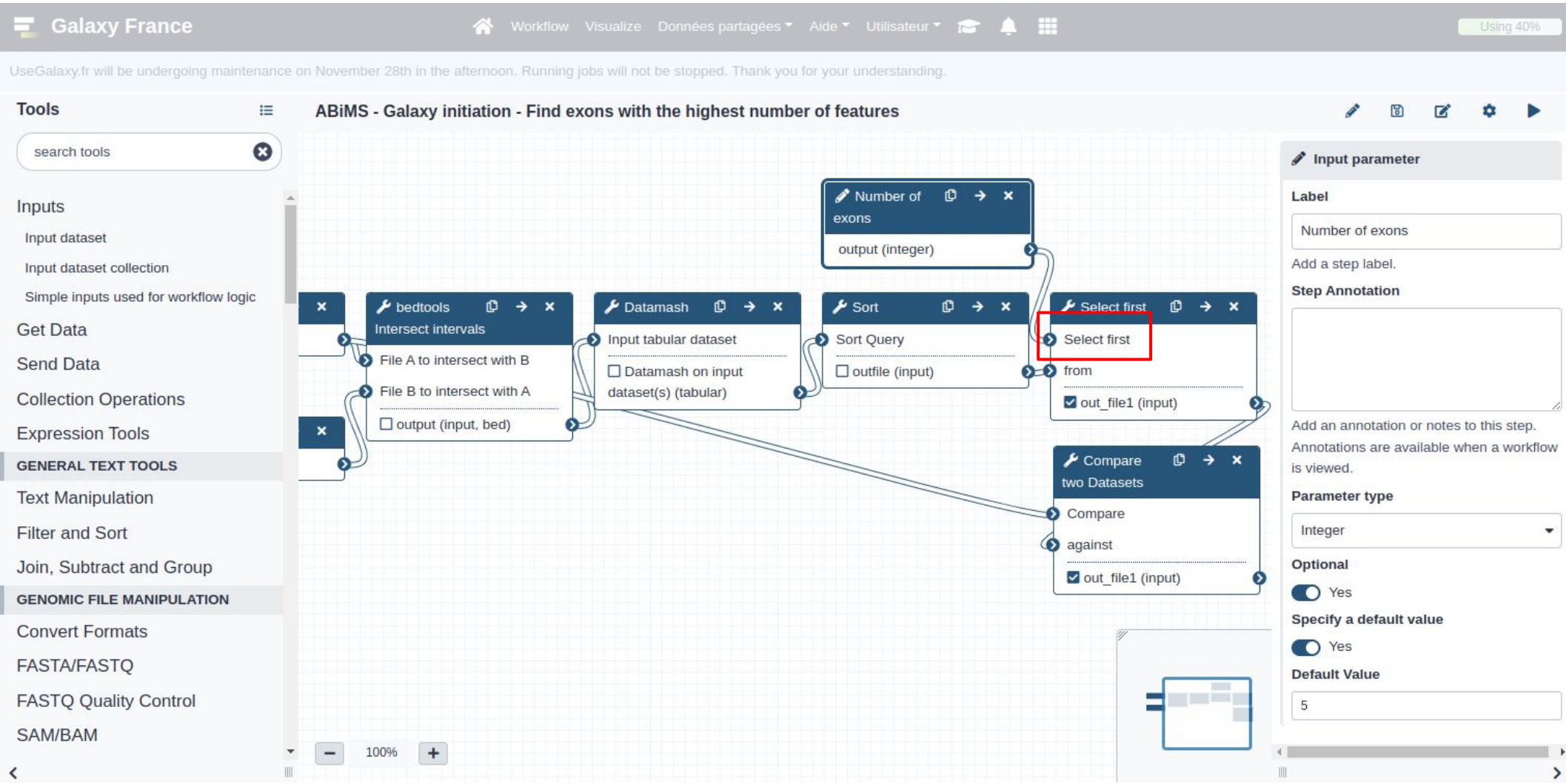


Edit a workflow: use workflow parameters to make your workflow more versatile



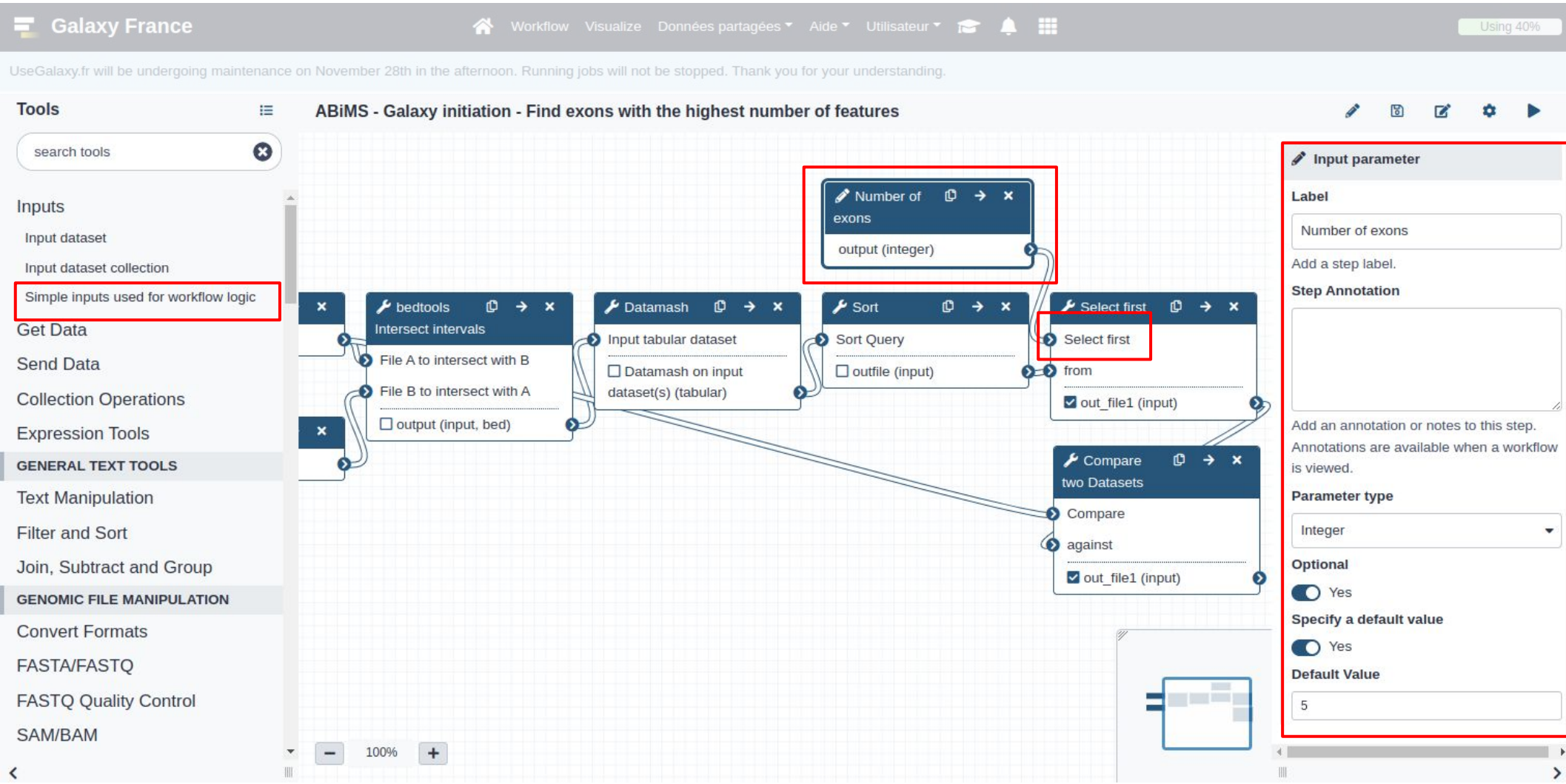
Workflow

Edit a workflow: use workflow parameters to make your workflow more versatile

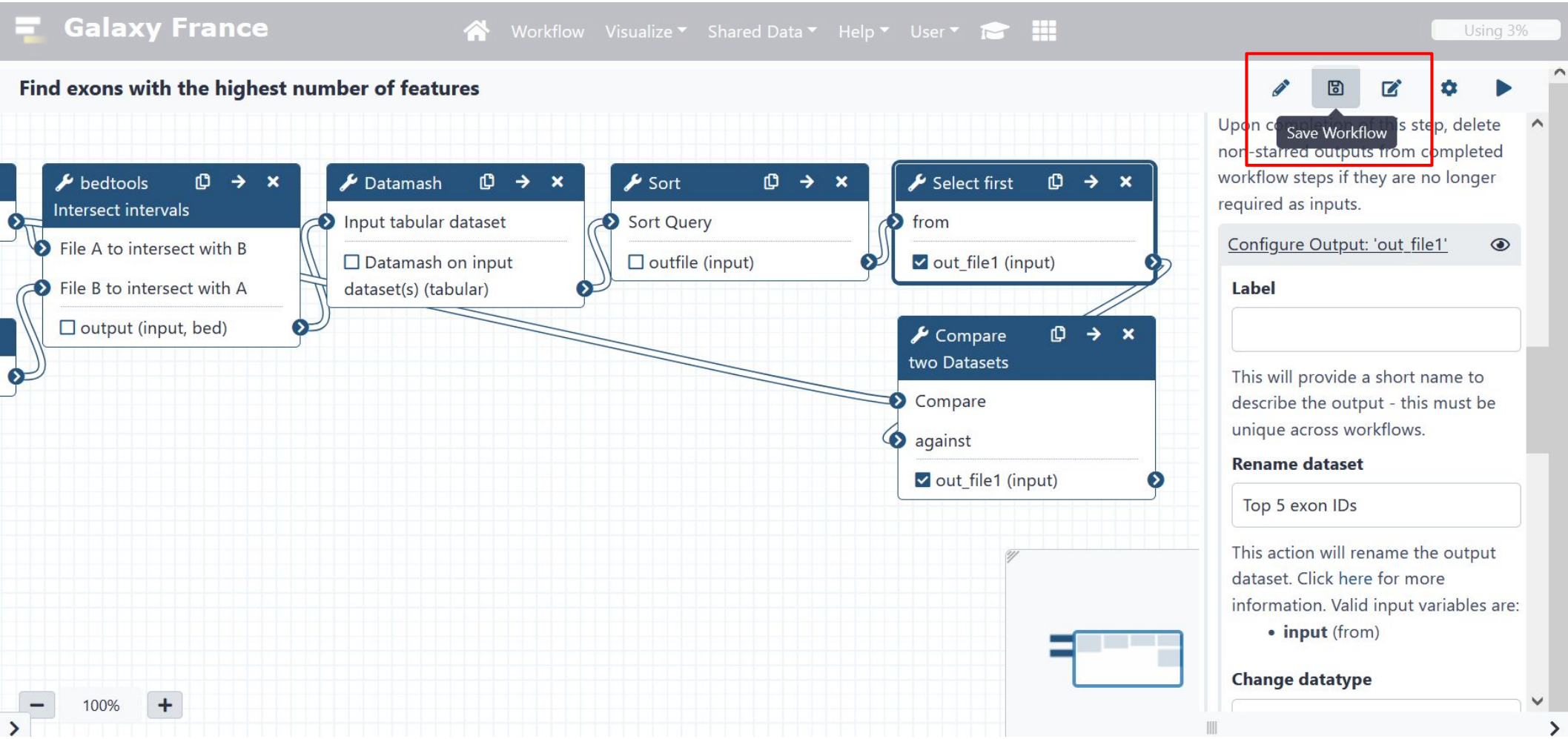


Workflow

Edit a workflow: use workflow parameters to make your workflow more versatile

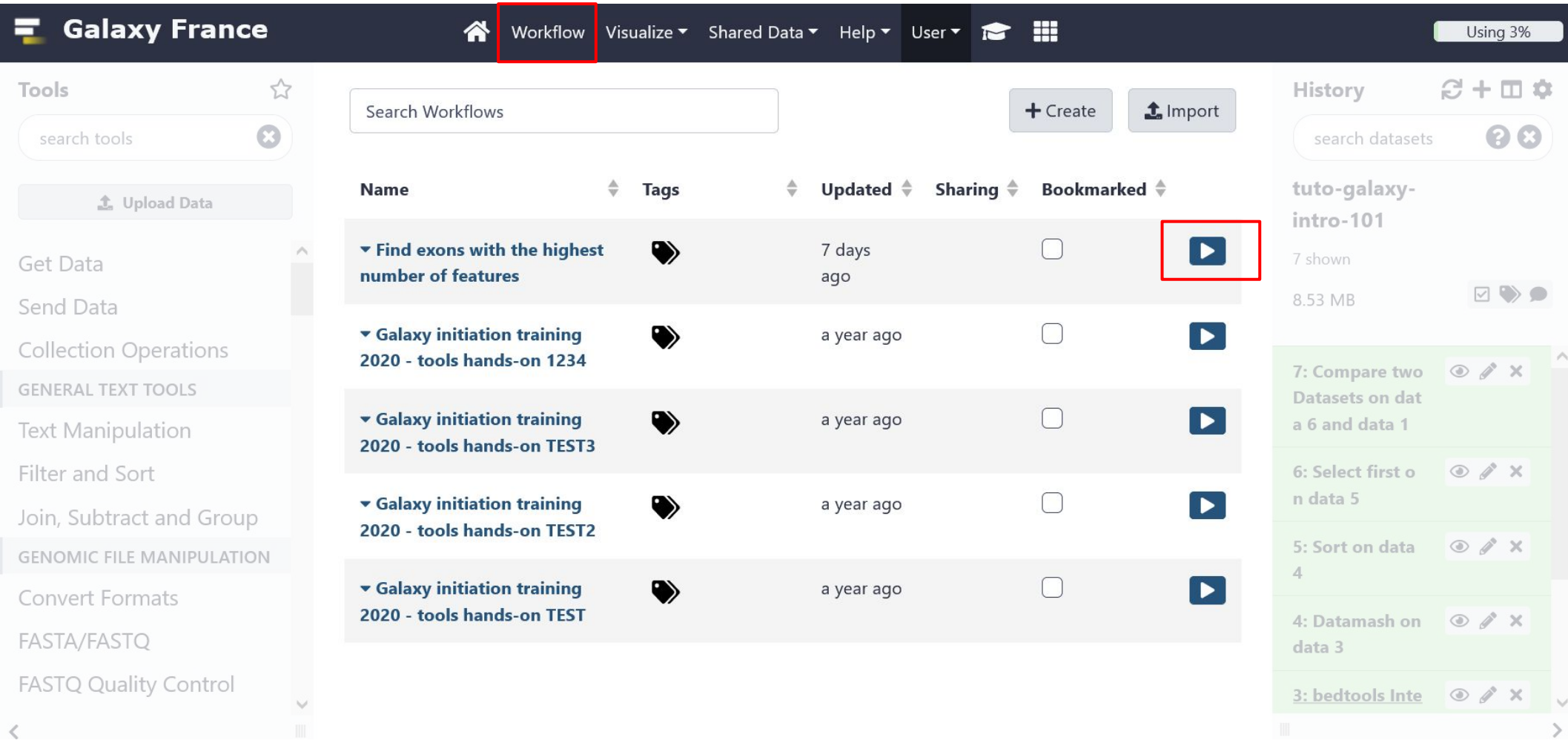


Save

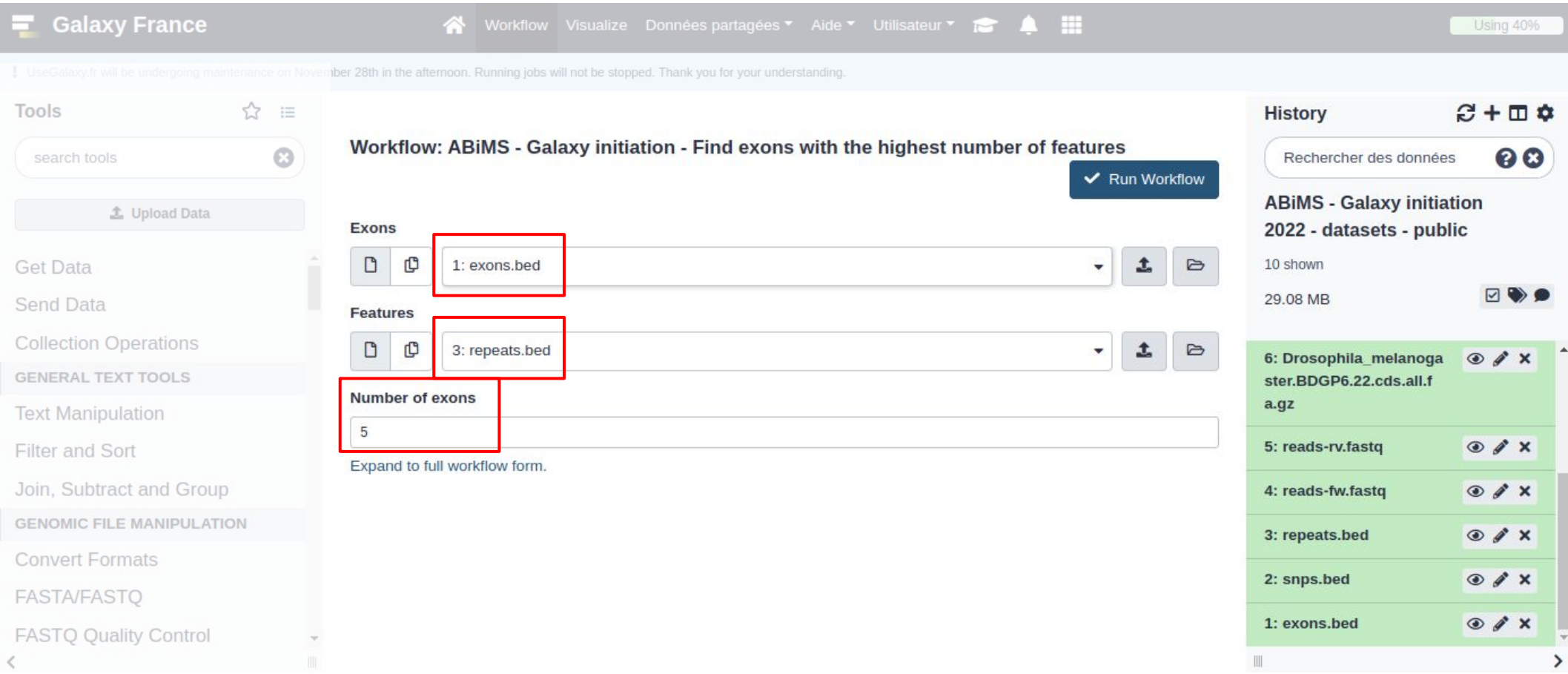

Galaxy France
Workflow
Visualize
Shared Data
Help
User
Using 3%
Find exons with the highest number of features
Save Workflow
bedtools Intersect intervals
File A to intersect with B
File B to intersect with A
output (input, bed)
Datamash
Input tabular dataset
Datamash on input dataset(s) (tabular)
Sort
Sort Query
outfile (input)
Select first
from
out_file1 (input)
Compare two Datasets
Compare
against
out_file1 (input)
Upon completion of this step, delete non-starred outputs from completed workflow steps if they are no longer required as inputs.
Configure Output: 'out_file1'
Label
This will provide a short name to describe the output - this must be unique across workflows.
Rename dataset
Top 5 exon IDs
This action will rename the output dataset. Click here for more information. Valid input variables are:
input (from)
Change datatype
100%

Workflow

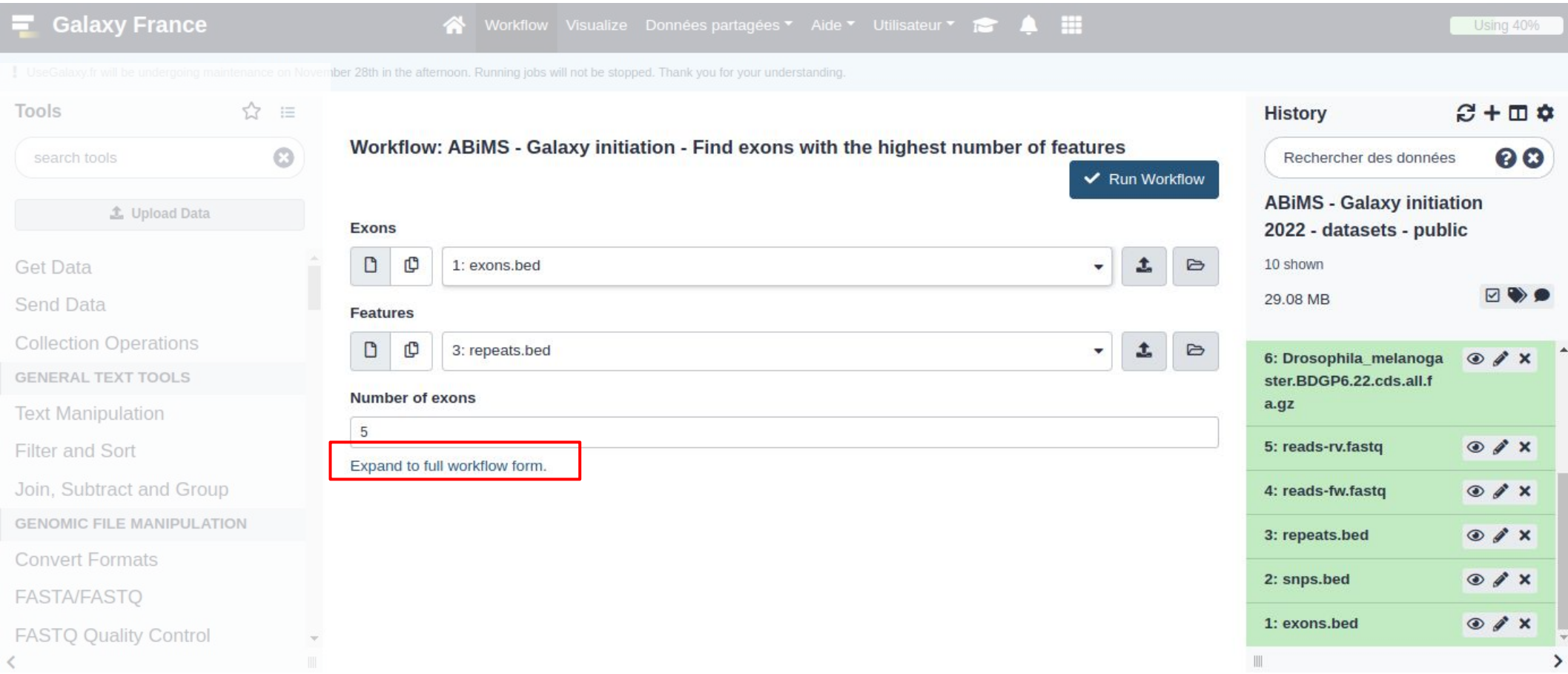
Run a workflow



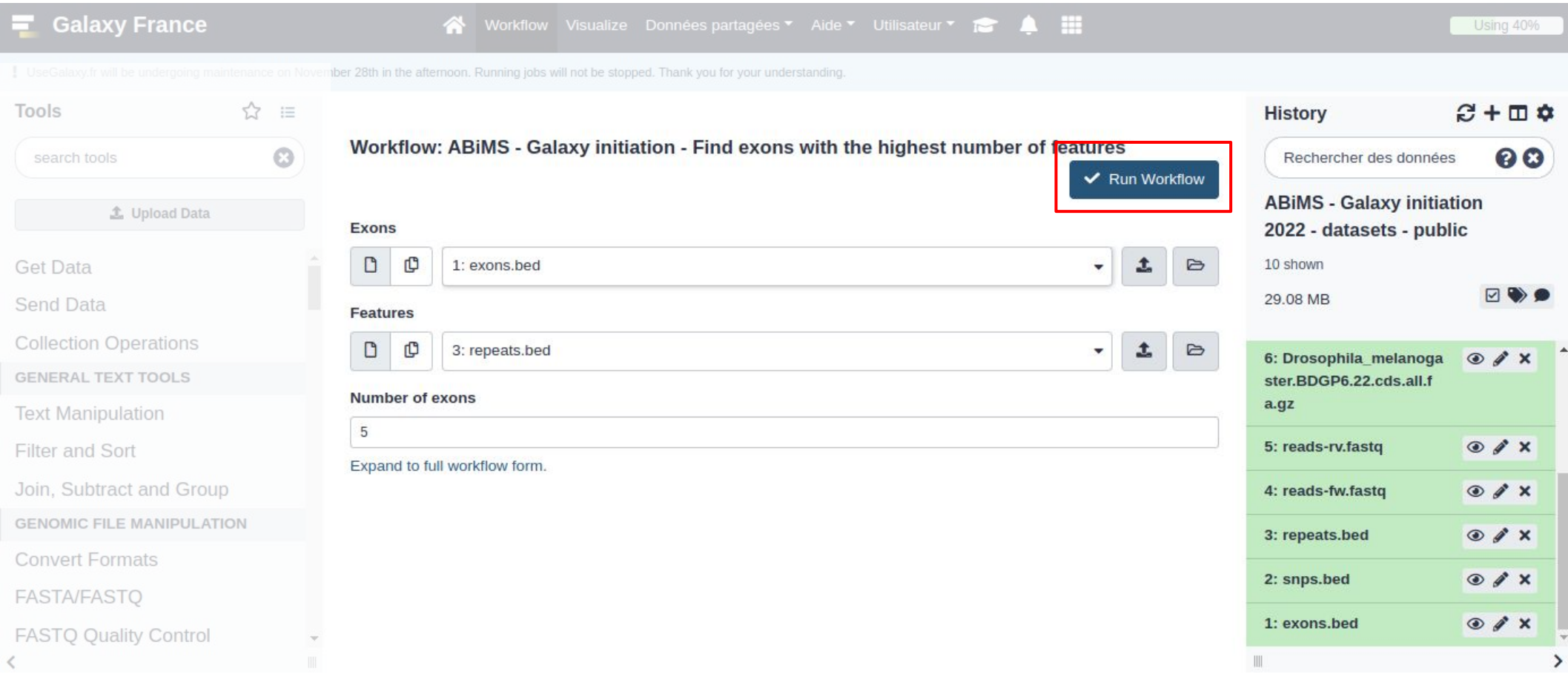
Run a workflow



Run a workflow

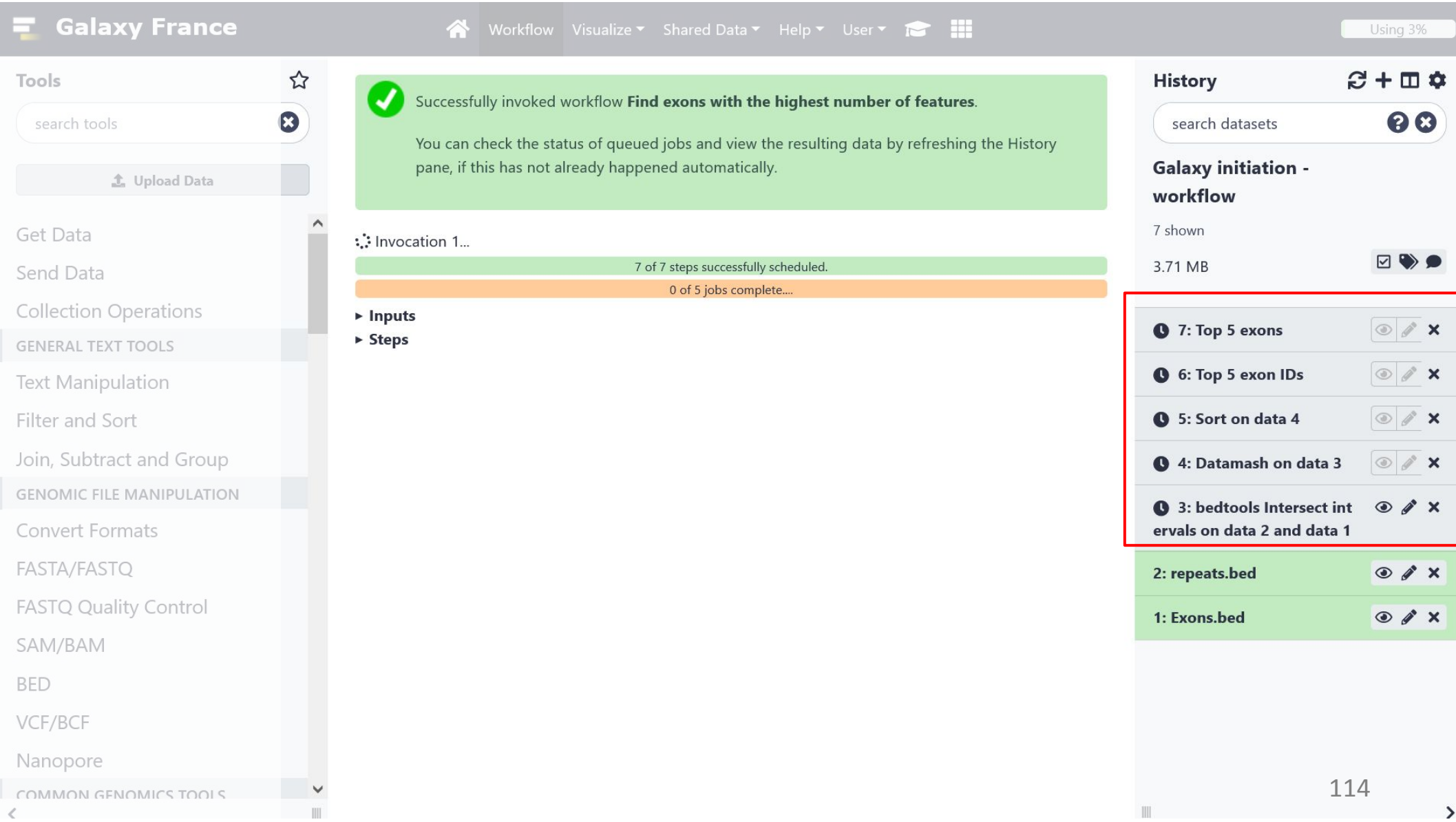


Run a workflow

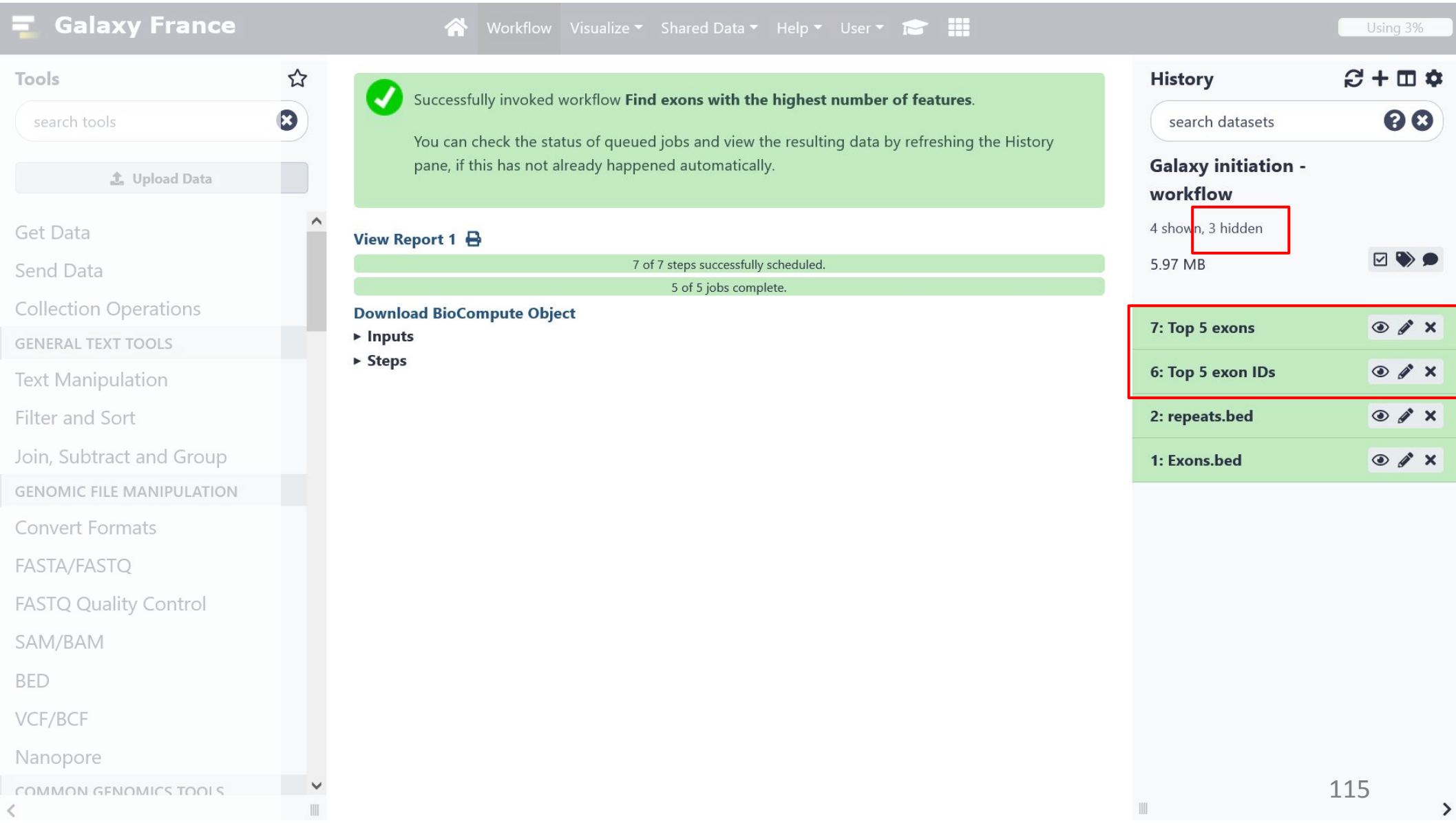


Workflow

Run a workflow

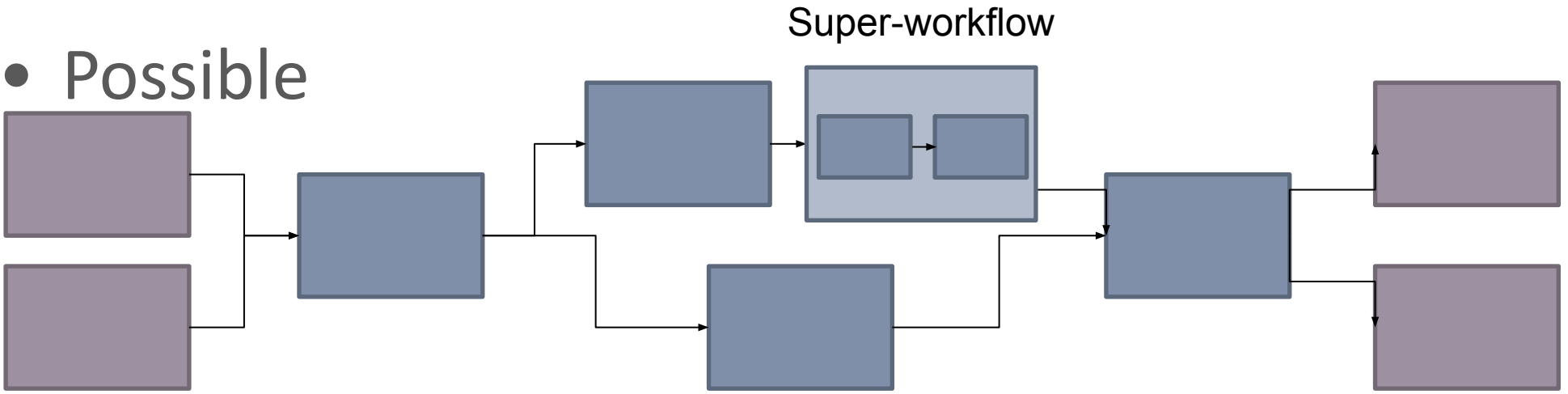


Run a workflow

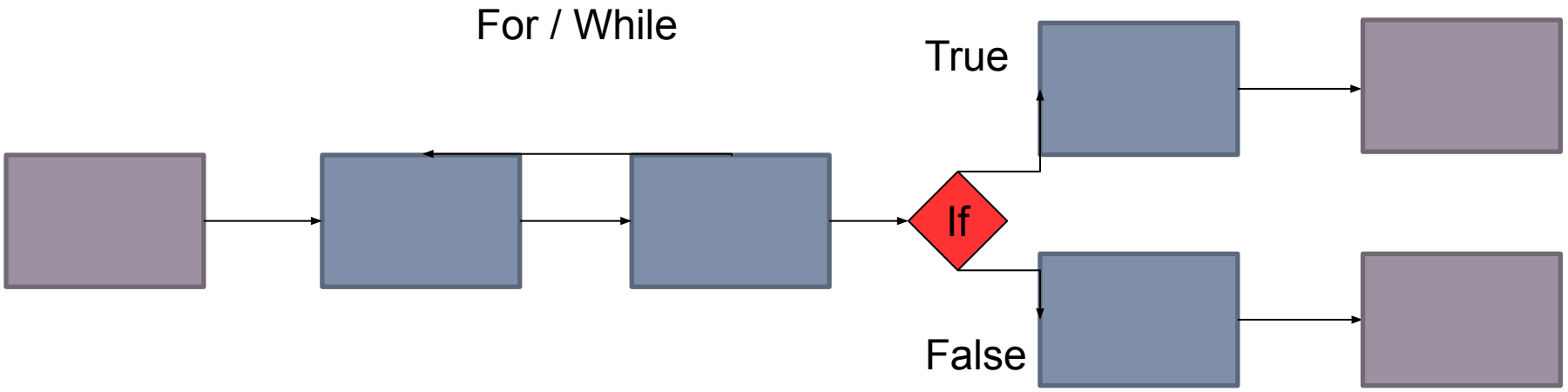


Workflow

- Possible



- Impossible (until now)



Last but not least!

SHARE

Share everything you do in Galaxy - histories, workflows, and visualizations

- Directly using a Galaxy account's email address on the same instance
- Using a web link, with anyone who knows the link
- Using a web link and publishing it to make it accessible to everyone from the Shared Data menu

Sharing your history:
https://training.galaxyproject.org/training-material/faqs/galaxy/histories_sharing.html

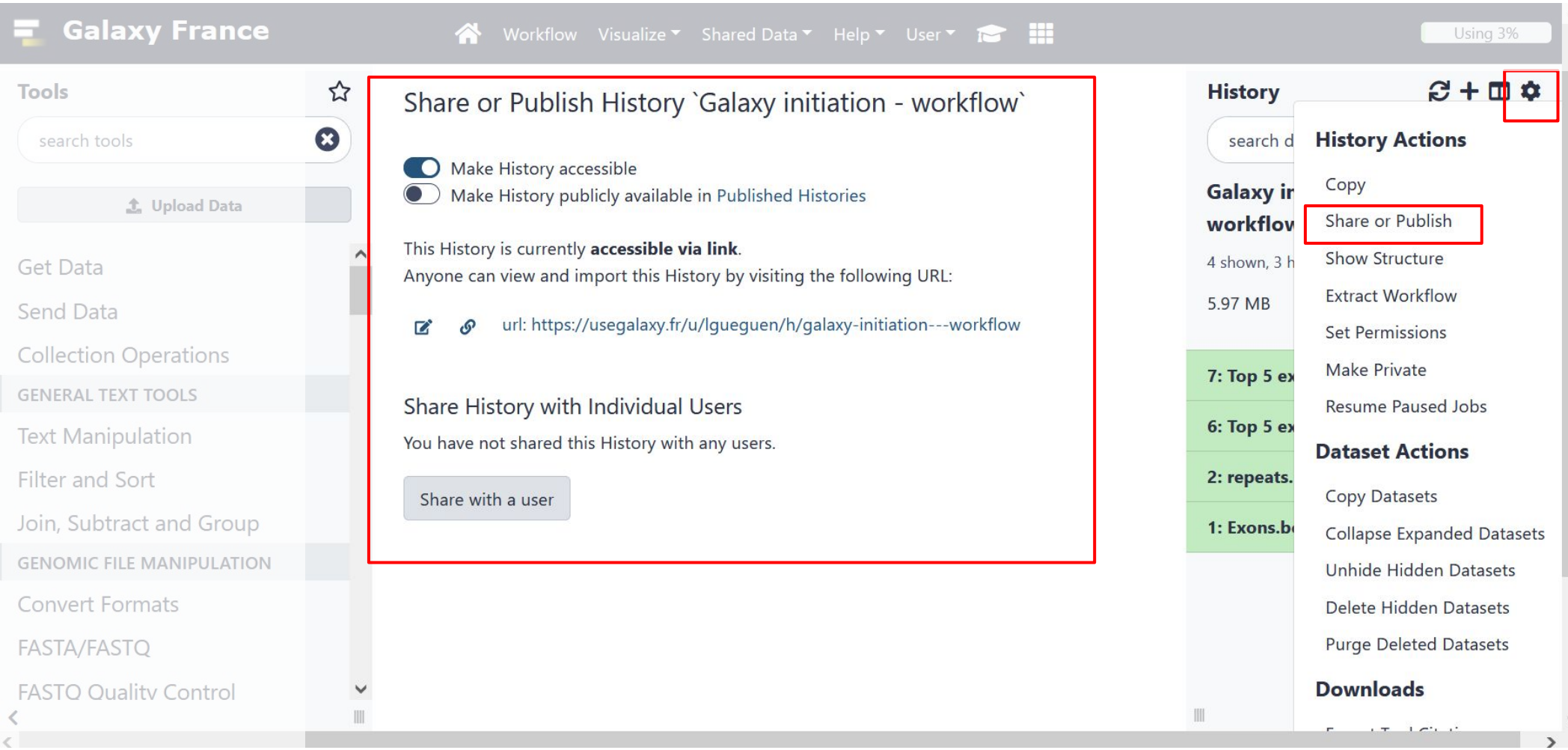
Share everything you do in Galaxy - histories, workflows, and visualizations

- Directly using a Galaxy account's email address on the same instance
- Using a web link, with anyone who knows the link
- Using a web link and publishing it to make it accessible to everyone from the Shared Data menu

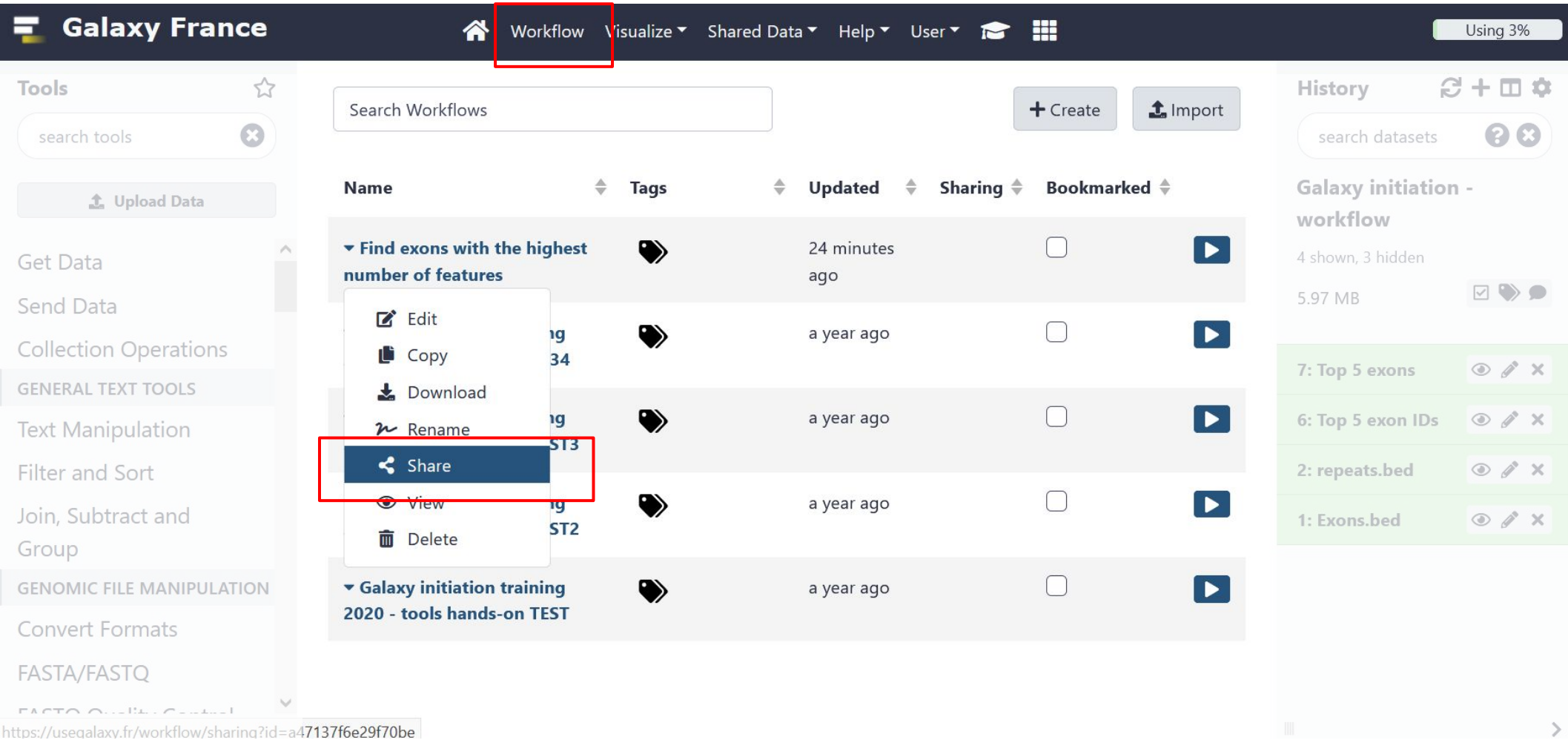
Sharing your history:
https://training.galaxyproject.org/training-material/faqs/galaxy/histories_sharing.html

- Sharing tools, scripts and wrappers: toolshed

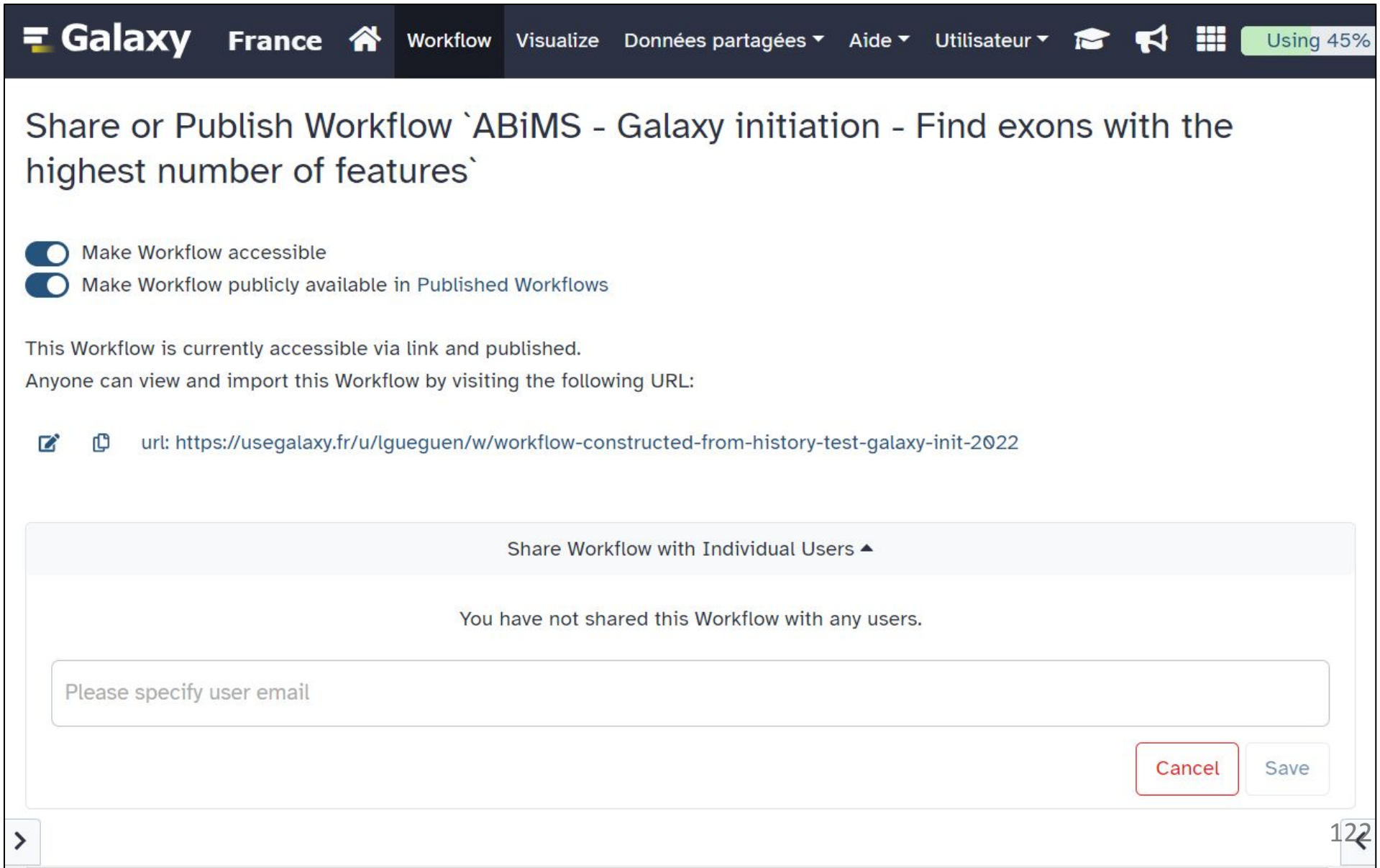
History


Galaxy France
Workflow
Visualize
Shared Data
Help
User
Using 3%
Tools
search tools
Upload Data
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
FASTQ Quality Control
Share or Publish History `Galaxy initiation - workflow`
Make History accessible
Make History publicly available in Published Histories
This History is currently accessible via link.
Anyone can view and import this History by visiting the following URL:
url: https://usegalaxy.fr/u/lgueguen/h/galaxy-initiation---workflow
Share History with Individual Users
You have not shared this History with any users.
Share with a user
History
search d
Galaxy ir
workflow
4 shown, 3 h
5.97 MB
7: Top 5 ex
6: Top 5 ex
2: repeats.
1: Exons.b
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
Purge Deleted Datasets
Downloads

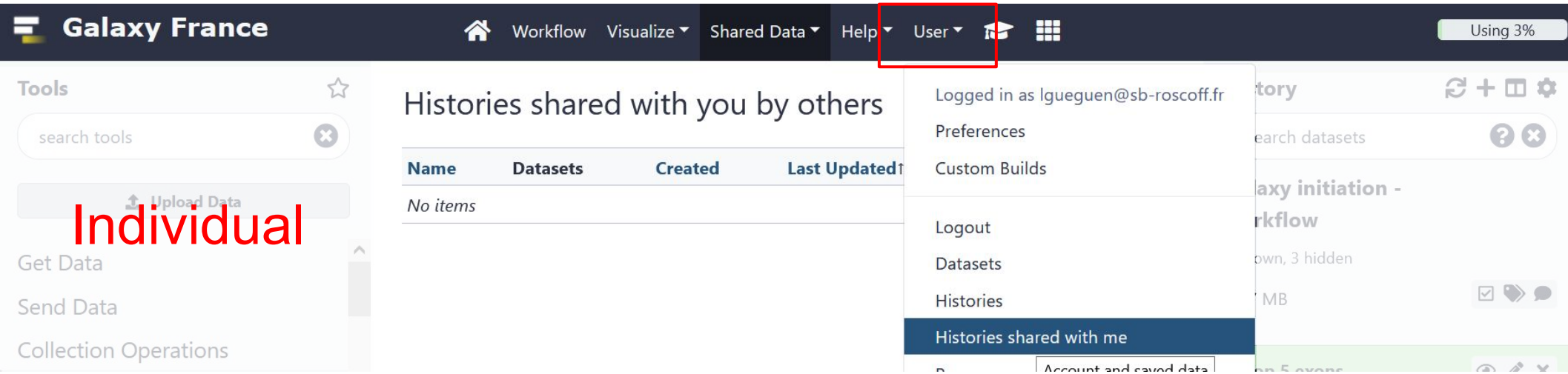
Workflow

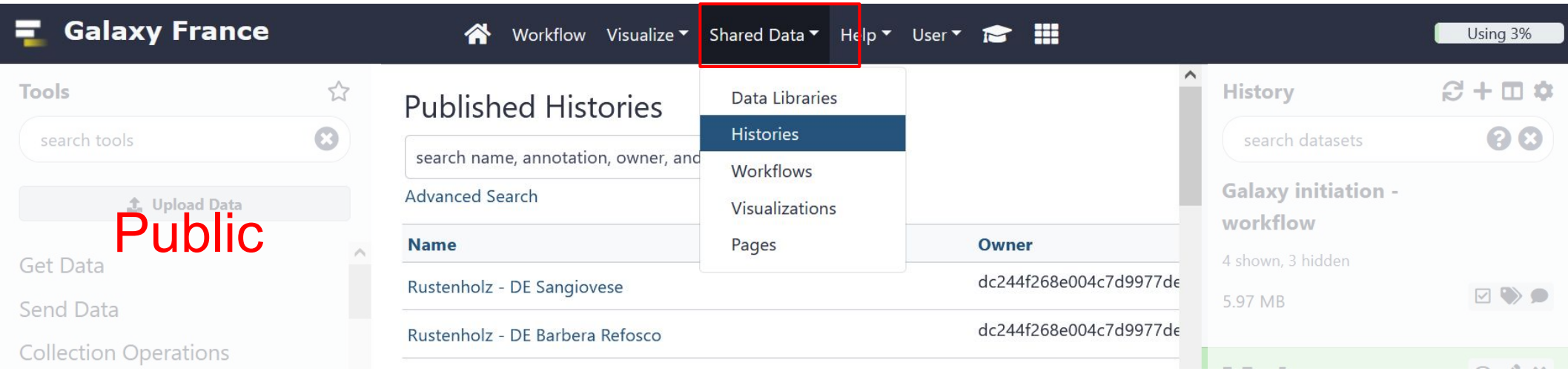

Galaxy France
Workflow
Visualize
Shared Data
Help
User
Using 3%
Tools
search tools
Upload Data
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
Search Workflows
+ Create
Import
Name
Tags
Updated
Sharing
Bookmarked
Find exons with the highest number of features
24 minutes ago
a year ago
a year ago
a year ago
Galaxy initiation training 2020 - tools hands-on TEST
a year ago
Edit
Copy
Download
Rename
Share
View
Delete
History
search datasets
Galaxy initiation - workflow
4 shown, 3 hidden
5.97 MB
7: Top 5 exons
6: Top 5 exon IDs
2: repeats.bed
1: Exons.bed
https://usegalaxy.fr/workflow/sharing?id=a47137f6e29f70be

Workflow: mode

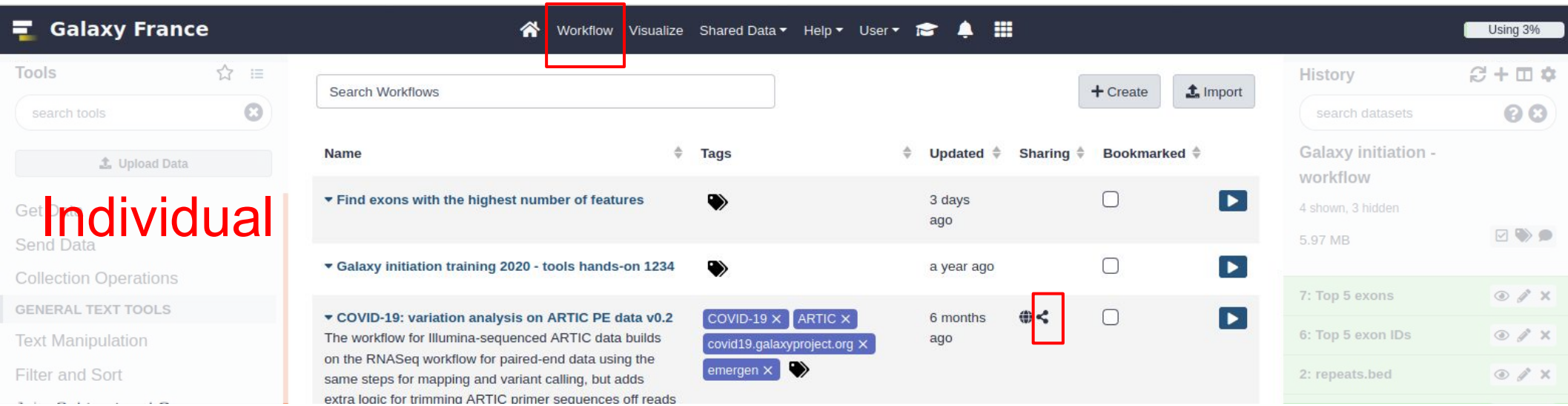


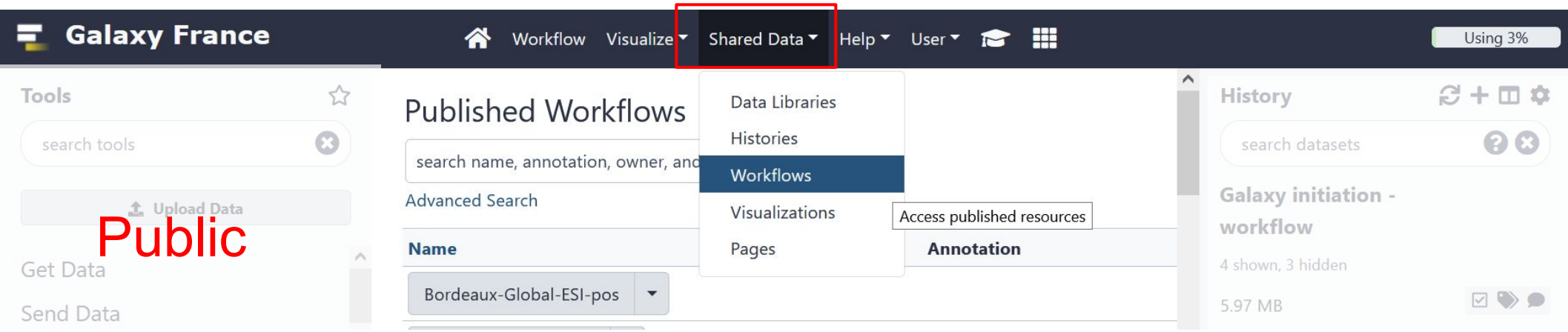
- Get shared histories



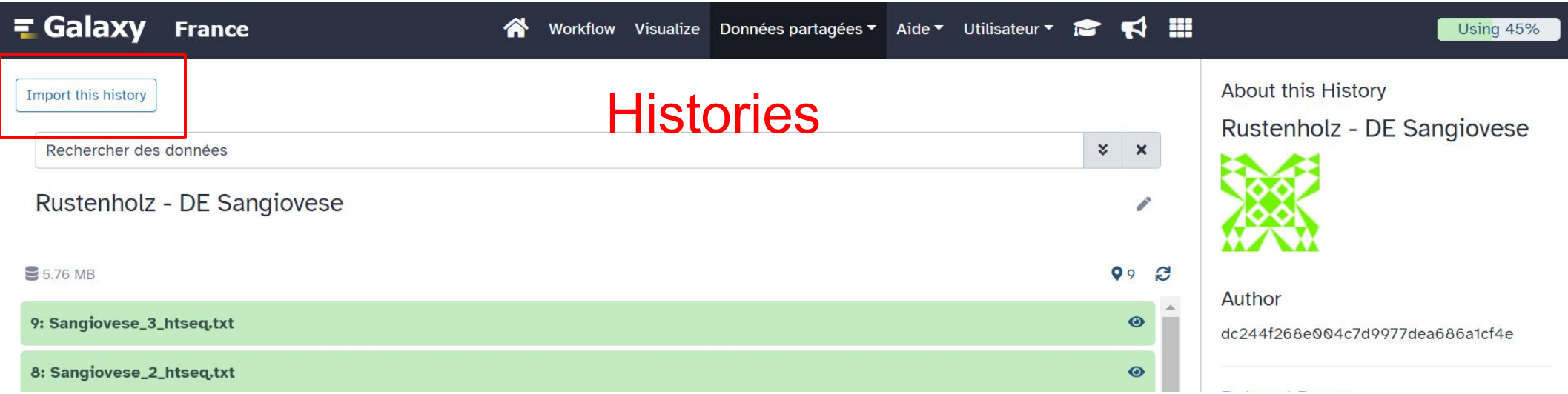


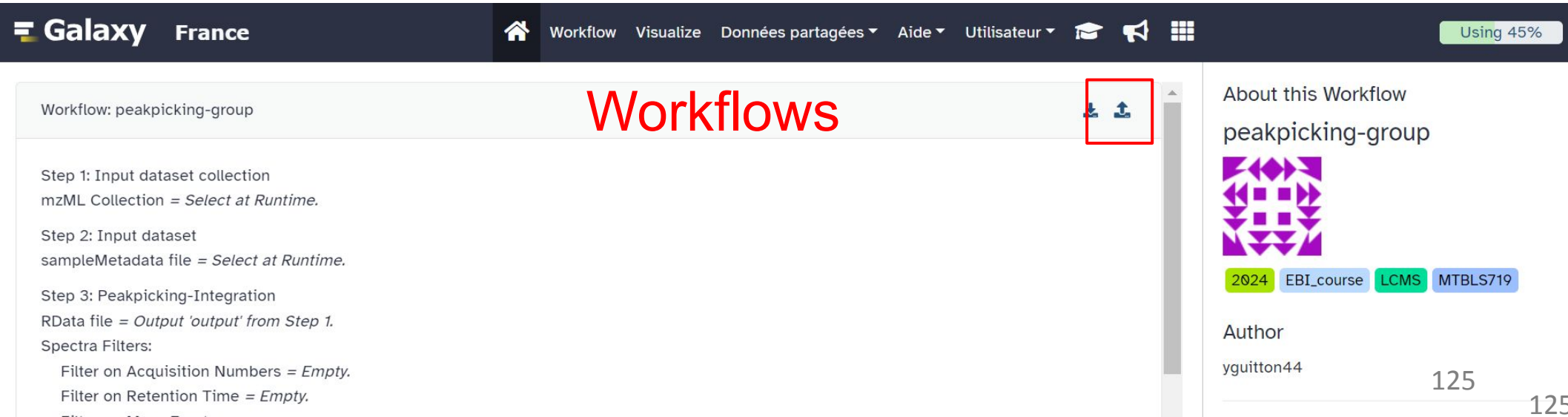
- Get shared workflows





- Import shared





125

Hands-on

WORKFLOW AND SHARE

https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html

Galaxy Basics for genomics

Authors: Saskia Hiltemann, Nicola Soranzo, Clemens Blank, Anton Nekrutenko, Björn Grüning, Anne Pajon, Helena Rasche

Overview

Questions:

- Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?

Objectives:

- Familiarize yourself with the basics of Galaxy
- Learn how to obtain data from external sources
- Learn how to run tools
- Learn how histories work
- Learn how to create a workflow
- Learn how to share your work

Time estimation: 1 hour
Level: Introductory
Supporting Materials: Datasets | Workflows | FAQs | Recordings | Available on these Galaxies
Published: Dec 19, 2016
Last modification: May 3, 2024

PURL: https://gxy.io/GTN:T00186
Rating: 4.1 (10 recent ratings, 141 all time)
Revision: 131

This tutorial aims to familiarize you with the Galaxy user interface. It will teach you how to perform basic tasks such as importing data, running tools, working with histories, creating workflows, and sharing your work.

Comment: Results may vary

Your results may be slightly different from the ones presented in this tutorial due to differing versions of tools, reference data, external databases, or because of stochastic processes in the algorithms.

Agenda

CONCLUSION

Key points on Galaxy

- Easy-to-use graphical user interface for often complex command-line tools
- Keeps a full record of your analysis in a history
- Workflows enable you to repeat your analysis on different data
- Galaxy can connect to external sources for data import and visualization purposes
- Galaxy provides ways to share your results and methods with others

Best practices

- Regularly free space with "Purge Deleted Datasets "or" Delete Permanently "
- On usegalaxy.fr, if you need support (issue, request for a tool…), please open a subject on https://community.france-bioinformatique.fr/c/galaxy/8

Now, choose your favorite Galaxy!

FIND A GALAXY SERVER

Galaxy servers

130+ platforms for using Galaxy: https://galaxyproject.org/use

[Servers, clouds, deployable resources]

Galaxy servers

130+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

usegalaxy.fr, usegalaxy.org, usegalaxy.eu, usegalaxy.org.au

Galaxy servers

130+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

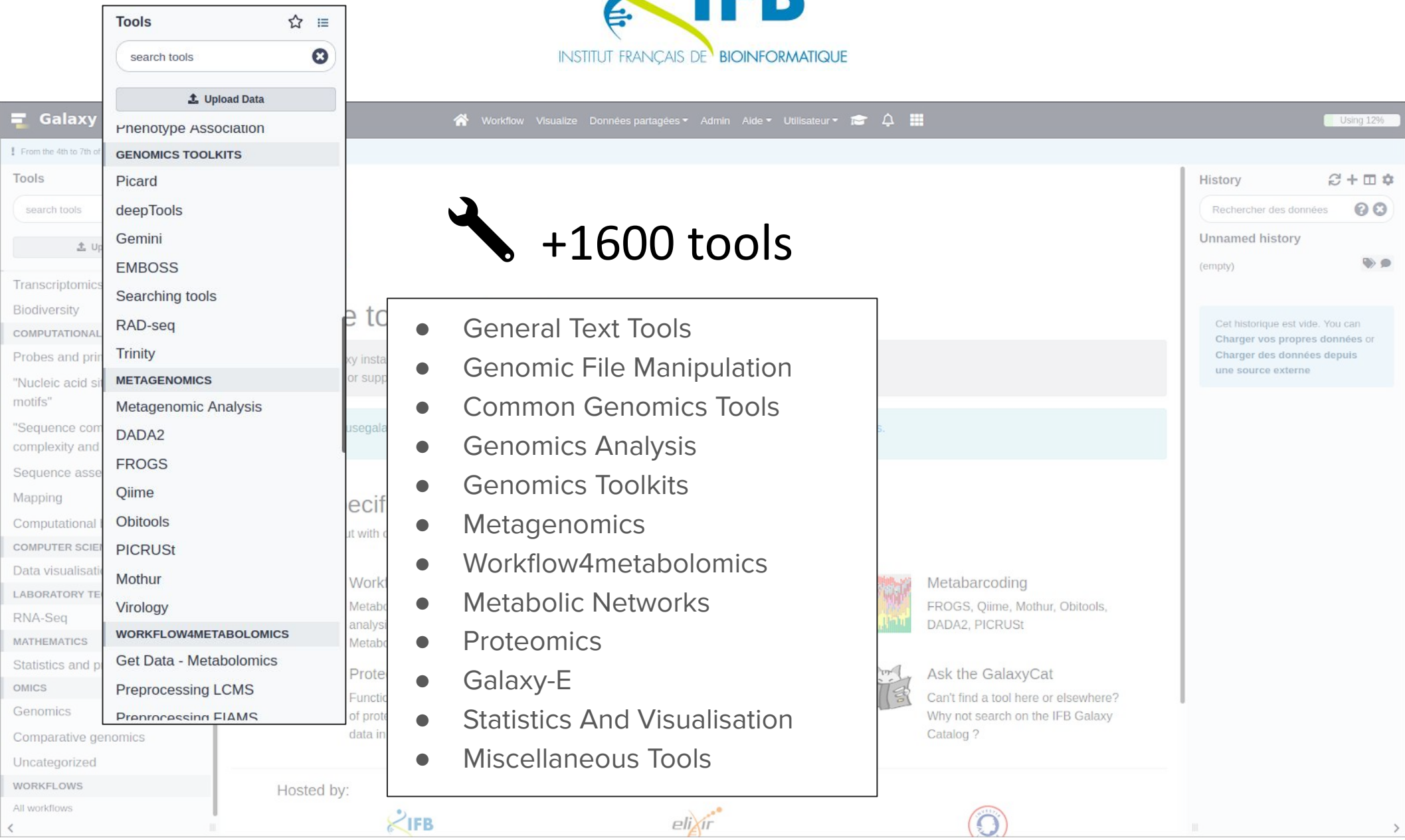
GENERAL PURPOSE:

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

usegalaxy.fr, usegalaxy.org, usegalaxy.eu, usegalaxy.org.au

	Quota		Concurrent jobs	
	registered	unregistered	registered	unregistered
usegalaxy.fr	100 GB	10 GB	25	2
usegalaxy.org	250 GB	5 GB	6	1
usegalaxy.eu	250 GB			

Galaxy France: https://usegalaxy.fr



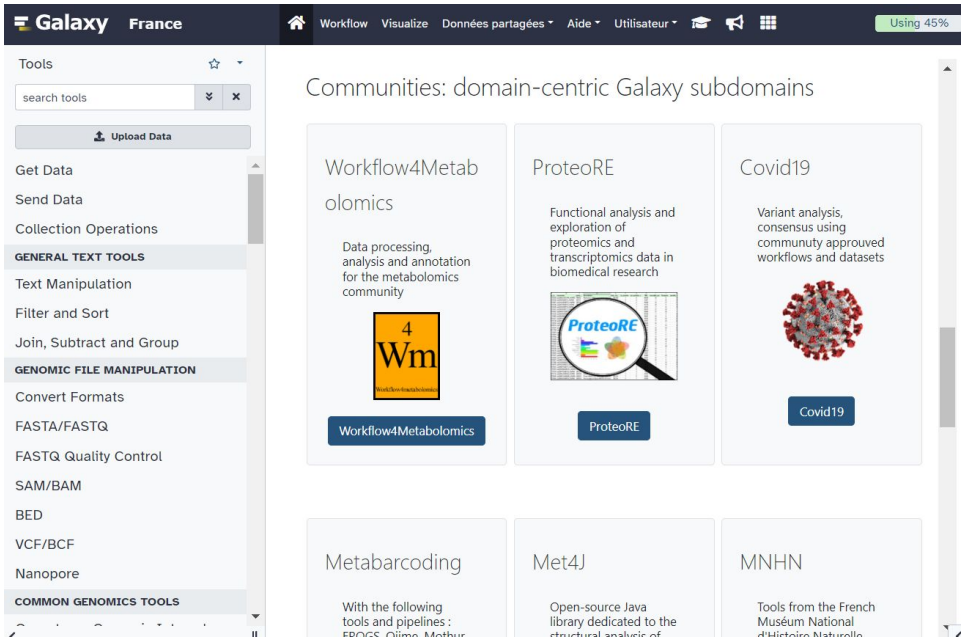
+1600 tools

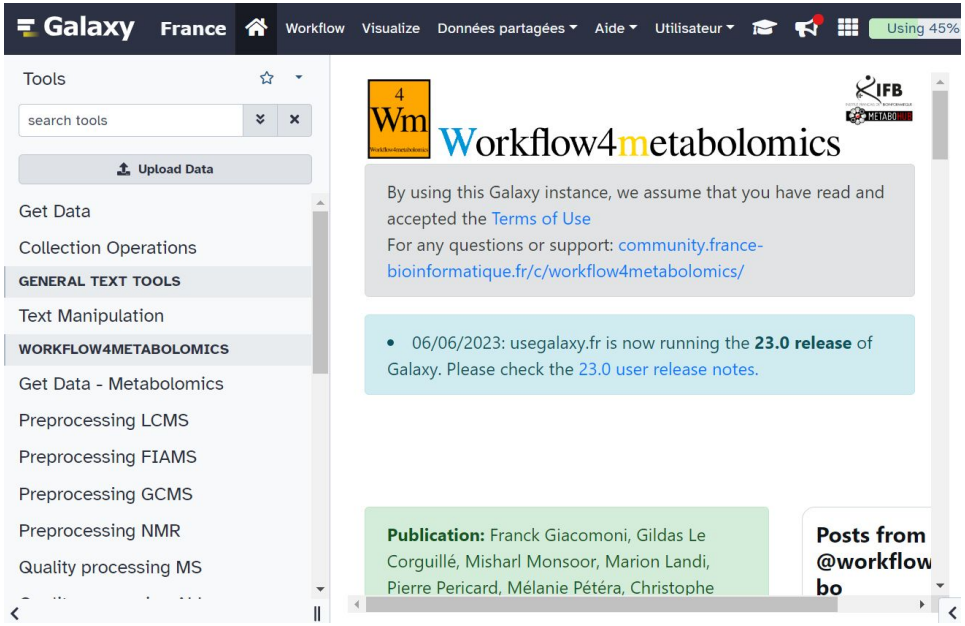
- General Text Tools
- Genomic File Manipulation
- Common Genomics Tools
- Genomics Analysis
- Genomics Toolkits
- Metagenomics
- Workflow4metabolomics
- Metabolic Networks
- Proteomics
- Galaxy-E
- Statistics And Visualisation
- Miscellaneous Tools

Galaxy France: https://<sub-domain>.usegalaxy.fr

7 sub-domains:

- ecology
- metabarcoding
- proteore
- covid19
- workflow4metabolomics
- mnhn
- met4j





Focus on your analysis

- Personalized welcome page
- Filtered tool list

Thanks to IFB

Galaxy servers

130+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

usegalaxy.fr, usegalaxy.org, usegalaxy.eu, usegalaxy.org.au

DOMAIN SPECIFIC:

- ecology
- metabarcoding
- proteore
- workflow4metabolomics
- …

SBR-tools:

ChIP-seq:

Galaxy servers

130+ platforms for using Galaxy: https://galaxyproject.org/use
[Servers, clouds, deployable resources]

GENERAL PURPOSE:

usegalaxy.fr, usegalaxy.org, usegalaxy.eu, usegalaxy.org.au

DOMAIN SPECIFIC:

- ecology
- metabarcoding
- proteore
- workflow4metabolomics
- …

SBR-tools: ABiMS

ChIP-seq: GalaxEast, nebula.ouro.fr

Catalog of French Galaxy tools:
http://galaxycat.france-bioinformatique.fr

To go further

- On your own:
  - Training materials:

    https://galaxyproject.github.io/training-material
  - Interactive tours of Galaxy:

    https://usegalaxy.fr/tours

- Training courses:

Training	What ?	Where ?	When ?
RNAseq analysis with Galaxy	RNAseq	Roscoff, France	?
Galaxy Community Conference (GCC)	General purpose (data-intensive biology and Galaxy)	Brno, Czech Republic	24-29 June 2024
Workflow4Experimenters	Metabolomics	Archamps, France	April 2024

Please complete the evaluation questionnaire!

END

Training evaluation

Thank you for completing the training evaluation questionnaire:

http://abims.sb-roscoff.fr/evaluation_formation