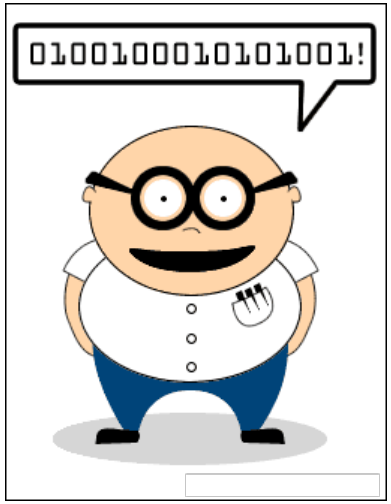
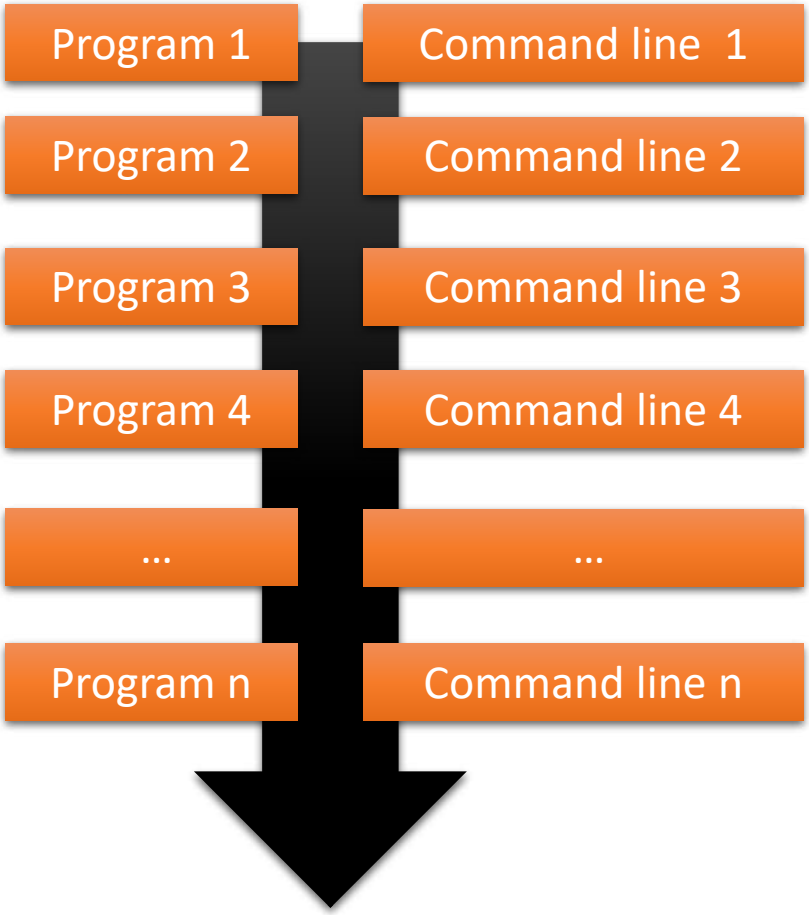
Automatization of NGS data analysis : Galaxy workflows

Stéphanie Le Gras
(slegras@igbmc.fr)

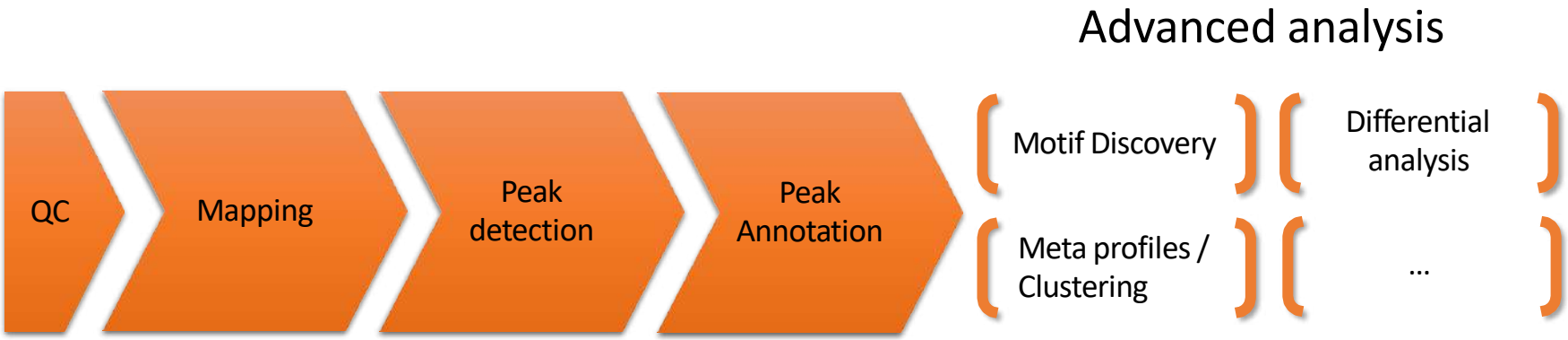
A long time ago…





PIPELINE / WORKFLOW

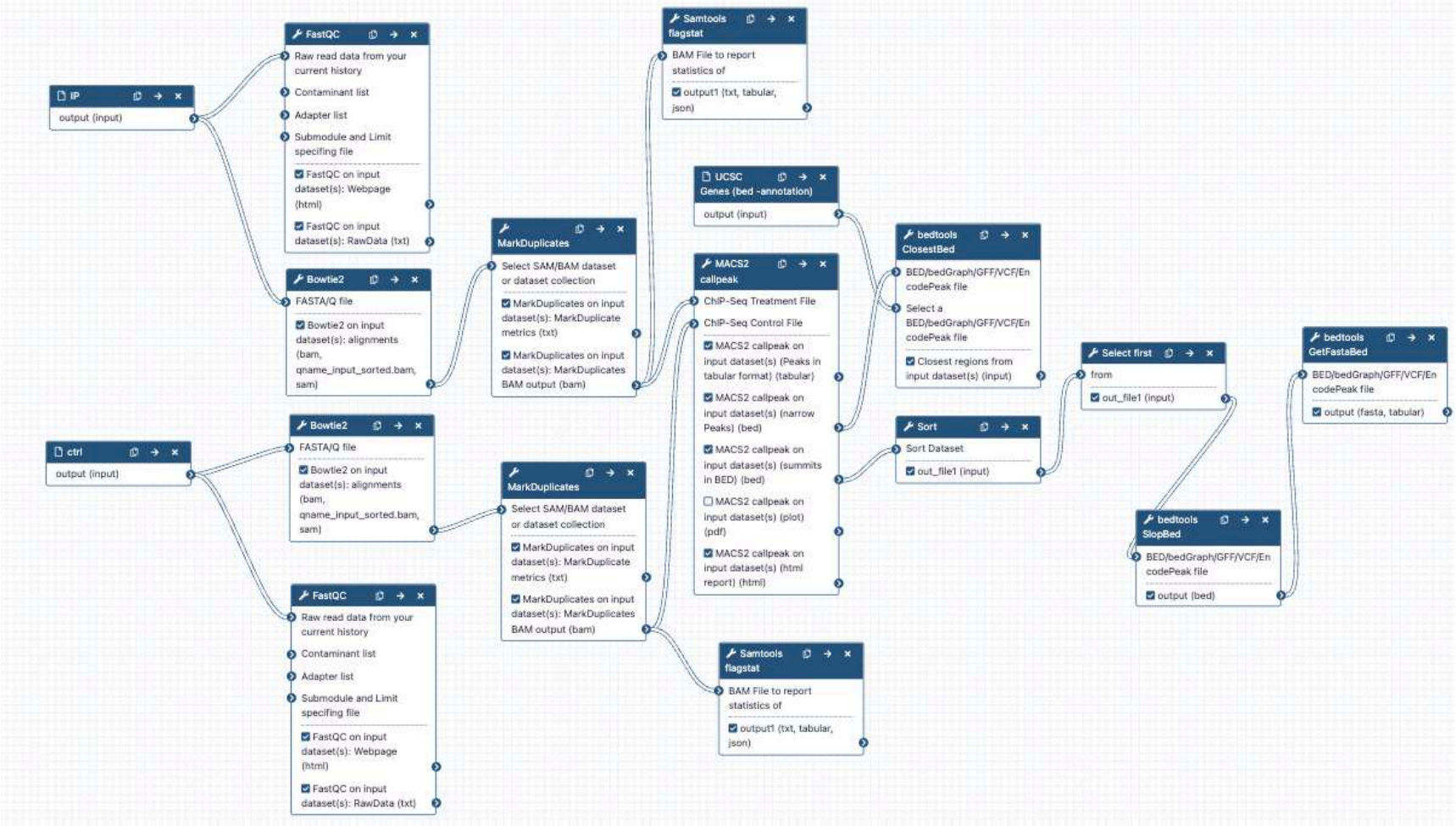
More recently…



During the entire training session..

What if we'd mix all together

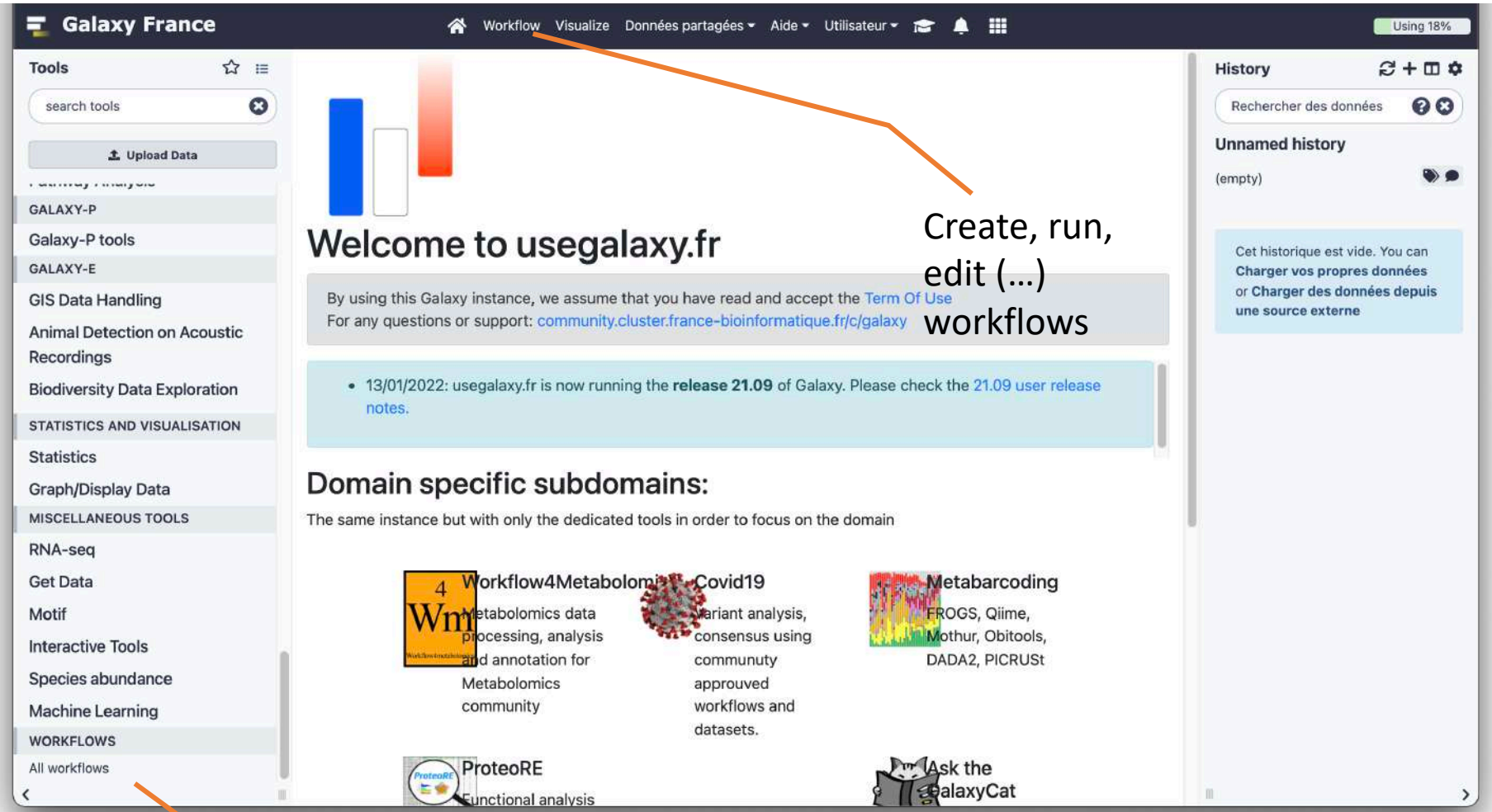
Galaxy workflow



Galaxy workflows

- Workflow:
  - Analysis protocol with several steps (tools)
  - The output of a step is used as the input of the next next so file formats between two steps should be compatible!
- Workflows are often made general so that they can be run on various datasets
- Some of the parameters are pre-defined while others are set at runtime

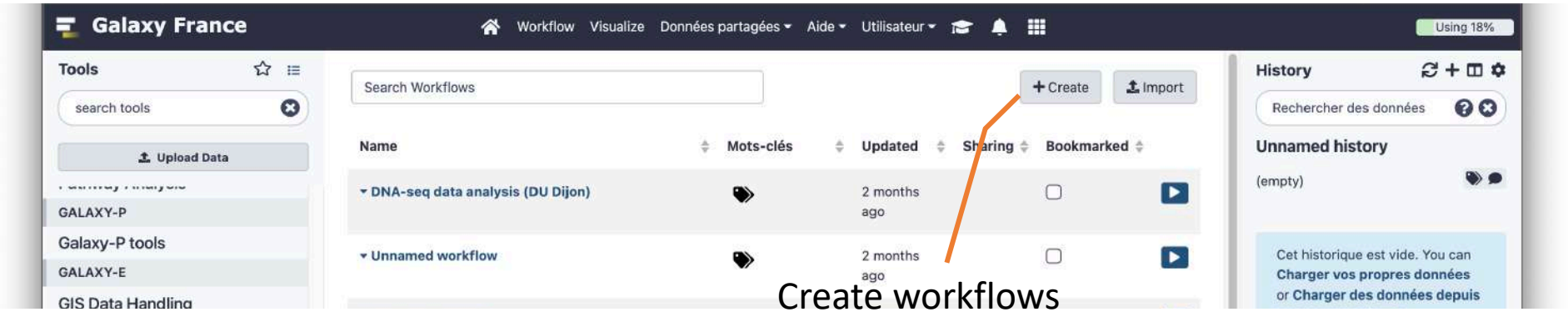
Workflows

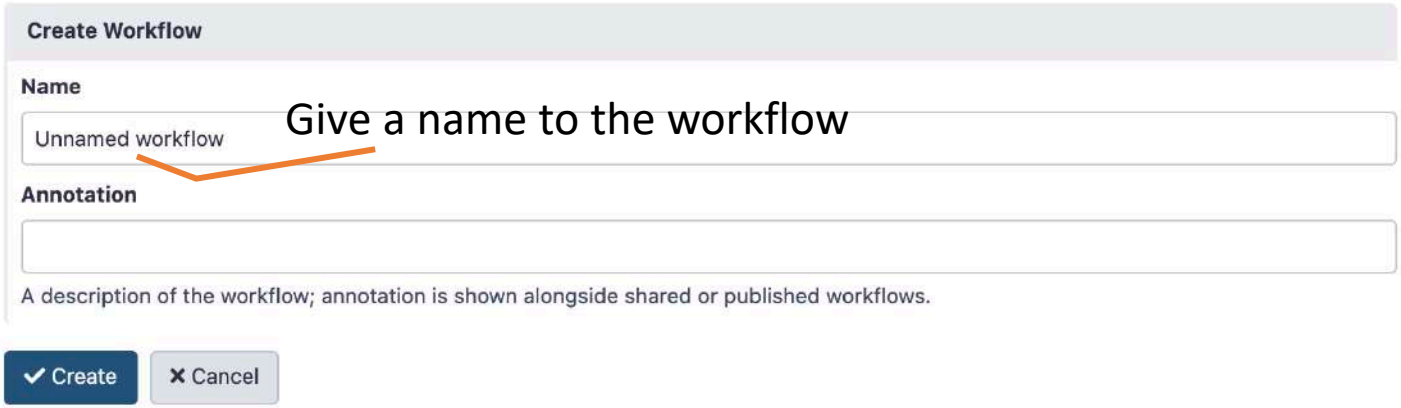


Create, run, edit (…) workflows

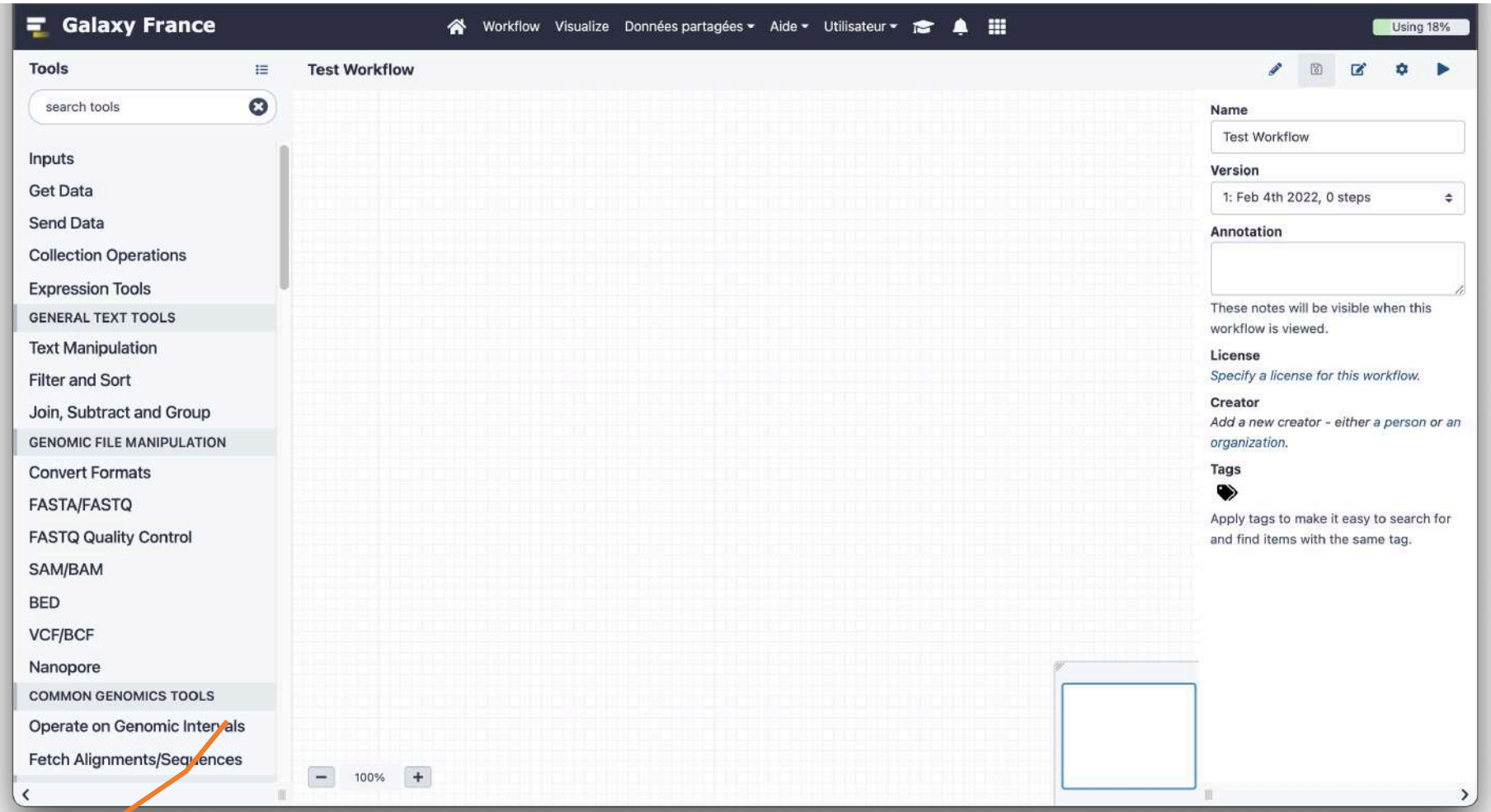
Run workflows

Workflows


Galaxy France
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 18%
Tools
search tools
Upload Data
GALAXY-P
Galaxy-P tools
GALAXY-E
GIS Data Handling
Search Workflows
+ Create
Import
Name
Mots-clés
Updated
Sharing
Bookmarked
DNA-seq data analysis (DU Dijon)
2 months ago
Unnamed workflow
2 months ago
History
Rechercher des données
Unnamed history
(empty)
Cet historique est vide. You can Charger vos propres données or Charger des données depuis
Create workflows

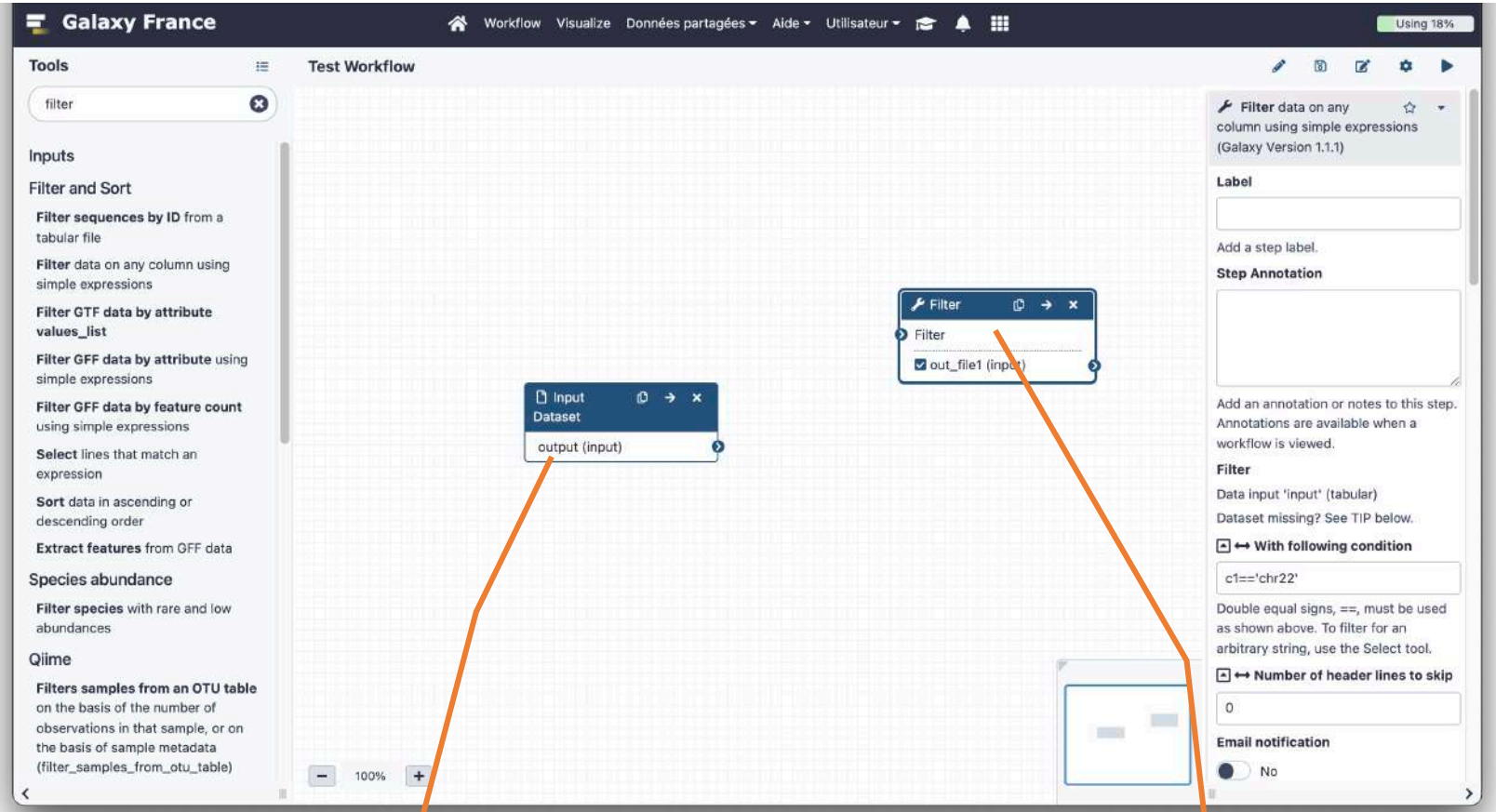

Create Workflow
Name
Unnamed workflow
Give a name to the workflow
Annotation
A description of the workflow; annotation is shown alongside shared or published workflows.
Create
Cancel

Workflow creation



Add tools or input datasets to the workflow

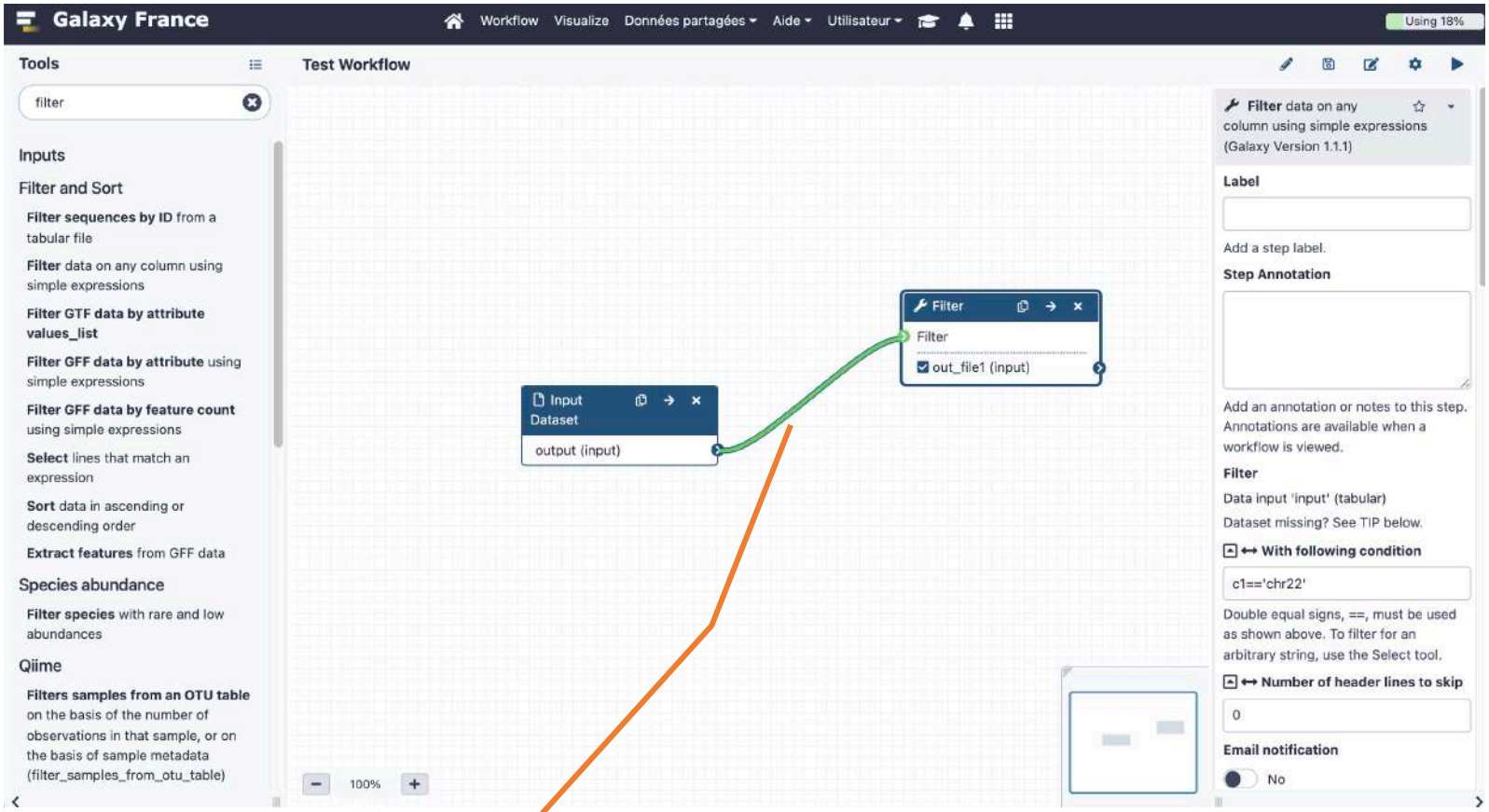
Workflow creation



Input dataset.
Most of the time, a workflow starts with an input dataset to which analyses are applied. In Galaxy, the file format of the input dataset will be limited to the input file format of the subsequent step

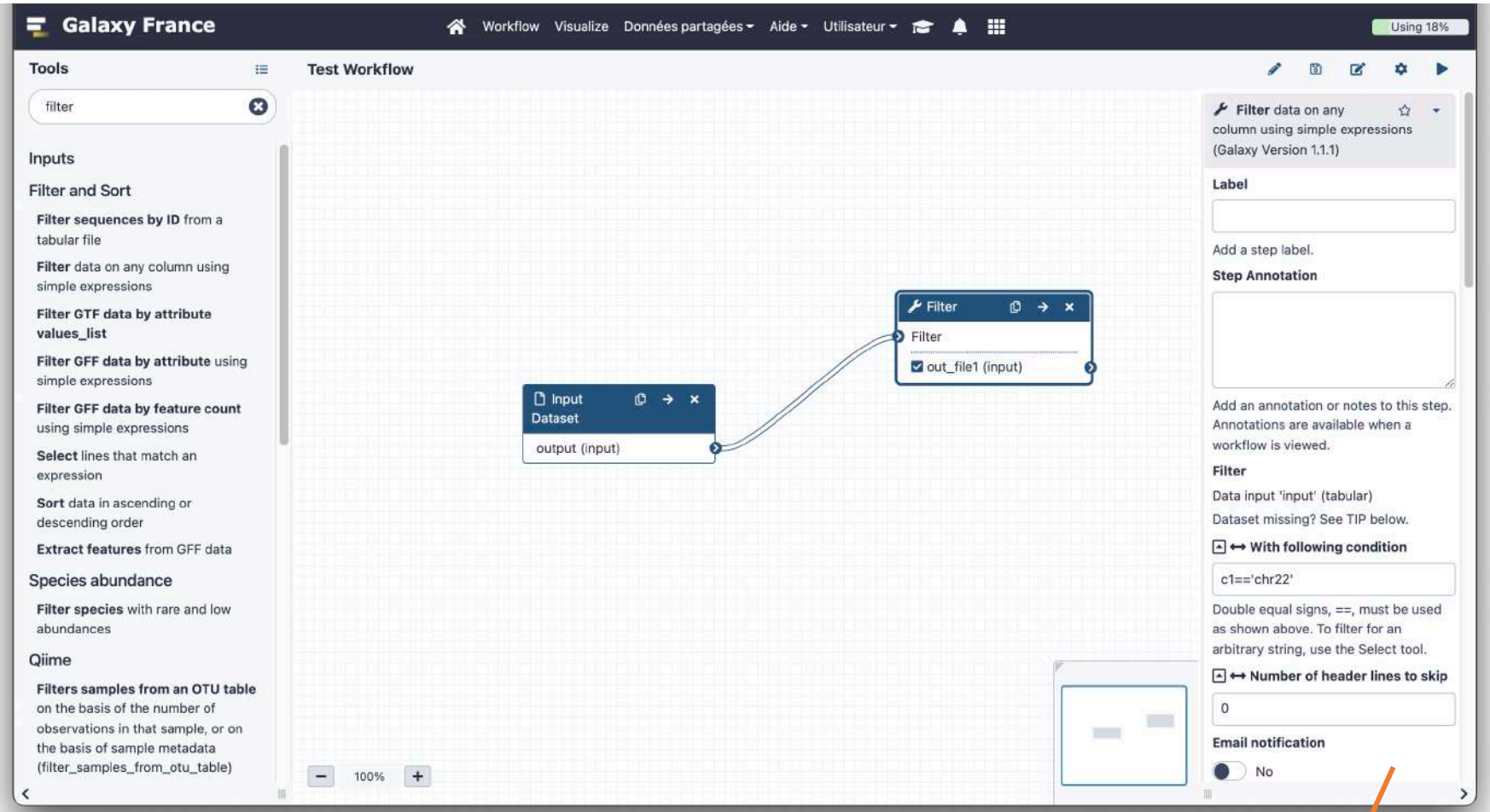
Tool to be run

Workflow creation



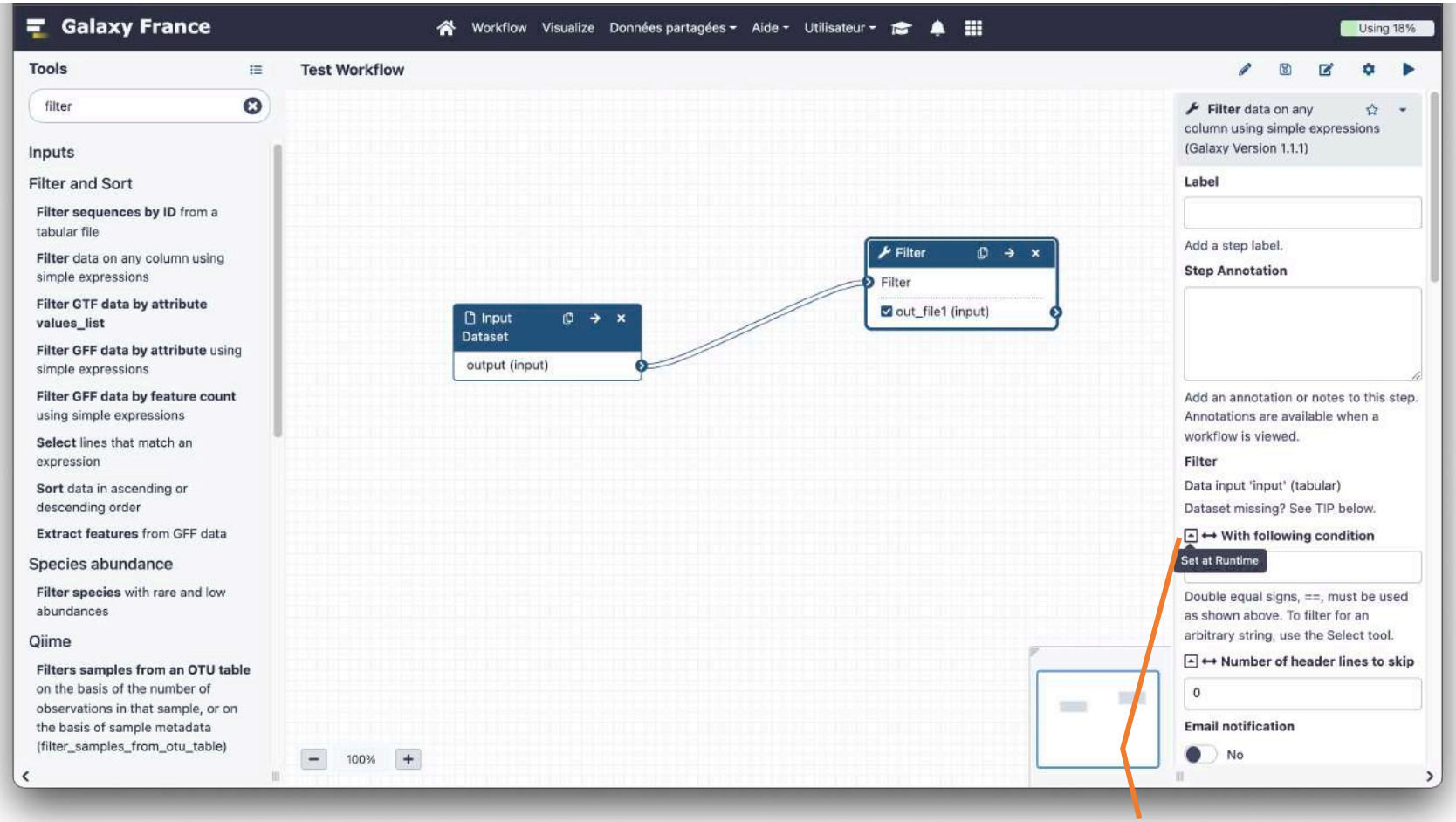
If two steps can be linked together, the link between the two boxes is green

Workflow creation



Pre-configure tool parameters and configure parameters to be set at run time

Workflow creation

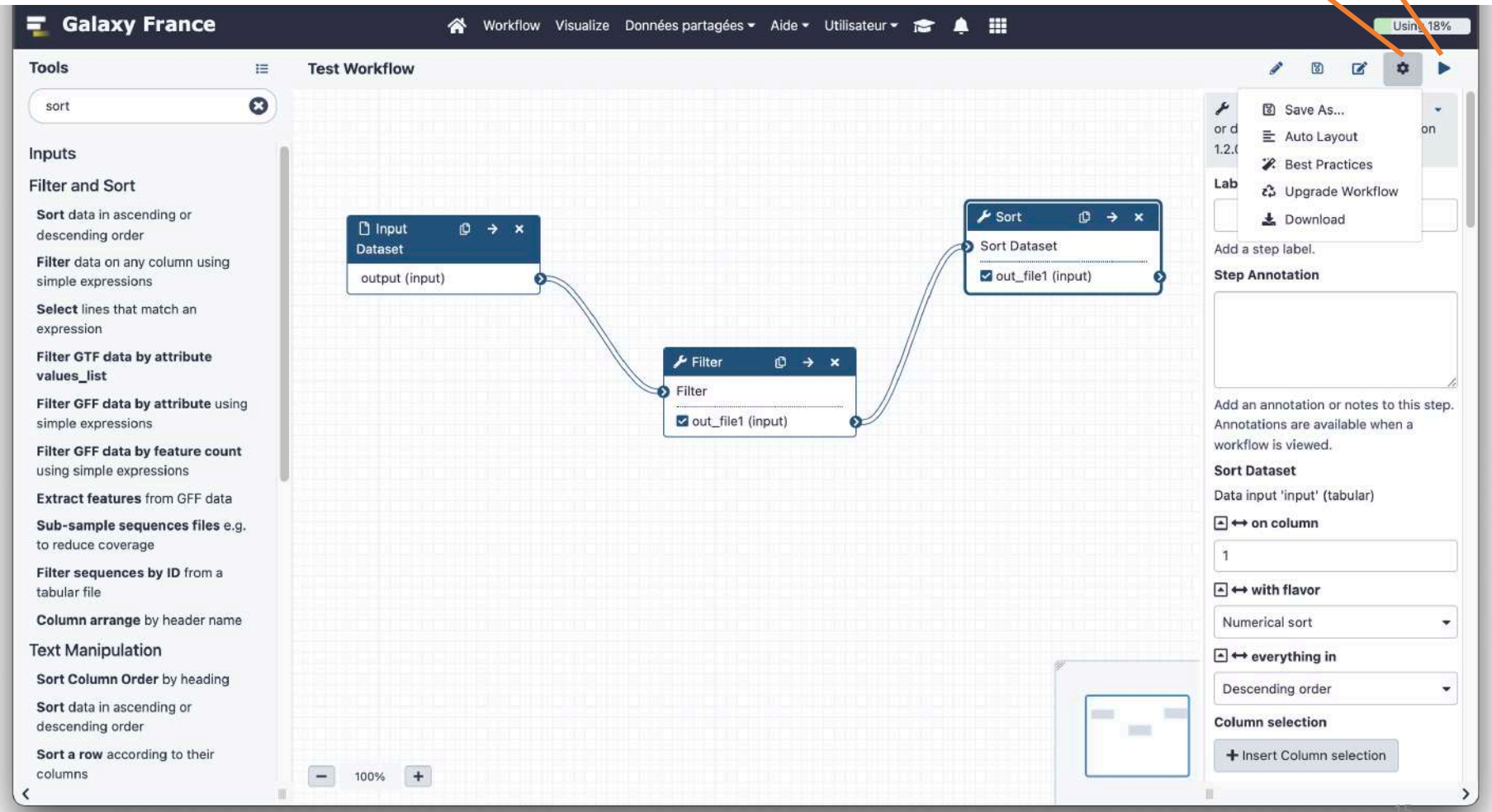


Click to get the parameter to be set at runtime

Workflow creation

Run workflow

Save



Run workflow

Set input file(s).
Found in current history!

Run workflow

Set parameters

Galaxy France

Workflow | Visualize | Shared Data | Help | User

Using 24%

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
Mapping

Workflow: BED Ensembl to BED UCSC

Run Workflow

History Options

Send results to a new history

No

1: Input dataset

1: hg38_ens105.bed

2: Replace -1 by - (Galaxy Version 1.0.0)

3: Replace + by 1 (Galaxy Version 1.0.0)

4: Add prefix chr (Galaxy Version 1.0.0)

5: Remove header line (Galaxy Version 1.0.0)

6: Replace chrMT by chrM (Galaxy Version 1.0.0)

7: Sort by coordinates (Galaxy Version 2.30.0)

History

search datasets

Unnamed history

1 shown

2.63 MB

1: hg38_ens105.bed

NGS analysis automatization: Galaxy workflows

(Questions and answers)

Exercise 1: Create your first workflow

We are going to create a workflow that modify a BED file generated with Ensembl/Biomart to make it compatible with BEDTools. To do this we are going to :

- Change the strand column that contains [1, -1] into [+, -].
- Change chromosome names and add the prefix « chr ».
- Change the name of chromosome mitochondrial from MT to M.
- Apply this pipeline to the file we extracted from Ensembl/Biomart: hg38_ens105.bed.

Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Gene stable ID	Gene name	Strand
1	1211340	1214153	ENSG00000186827	TNFRSF4	-1
1	1203508	1206592	ENSG00000186891	TNFRSF18	-1
1	1471765	1497848	ENSG00000160072	ATAD3B	1
1	1249777	1251334	ENSG00000260179		-1

chr1	11869	14409	ENSG00000223972	DDX11L1	+
chr1	14404	29570	ENSG00000227232	WASH7P	-
chr1	17369	17436	ENSG00000278267	MIR6859-1	-
chr1	29554	31109	ENSG00000243485	MIR1302-2HG	+
chr1	30366	30503	ENSG00000284332	MIR1302-2	+

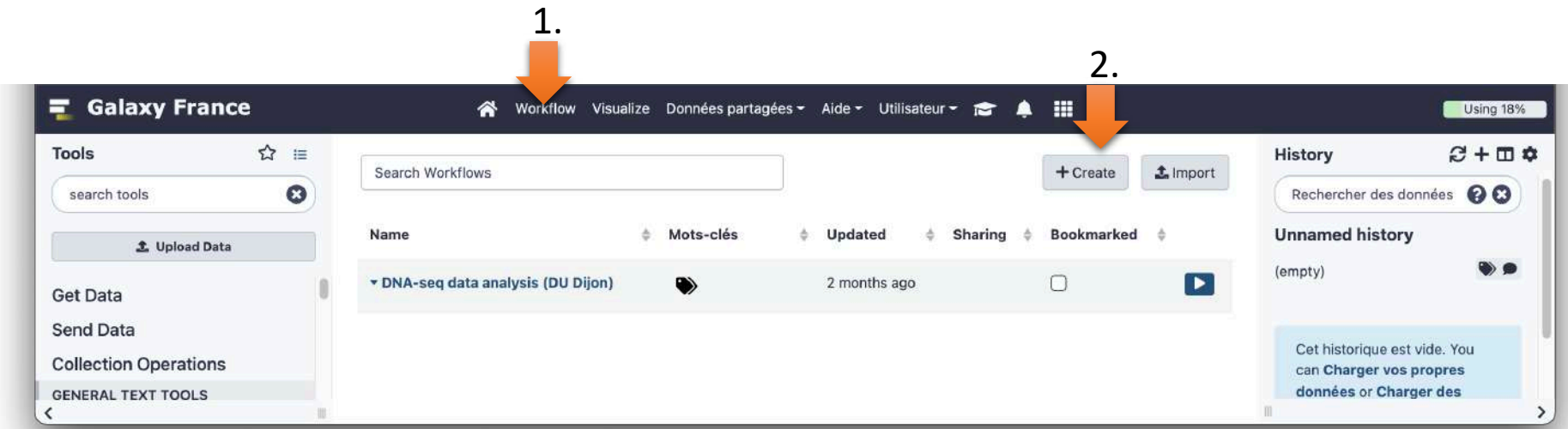
Exercise 1: Create your first workflow

Steps:

1. Create an empty workflow named « BED Ensembl to BED UCSC ».
2. Add the following steps:
    1. An input file
    2. Use the tool **Regex Replace** to turn trailing « -1 » into « - »
    3. Use the tool **Regex Replace** to turn trailing « 1 » into « + »
    4. Use the tool **Regex Replace** to add the prefix chr at the beginning of lines (in front of chromosome names)
    5. Use the tool **Remove beginning** to remove the first line of the dataset (the one with header
    6. Use the tool **Regex Replace** to turn « chrMT » into « chrM »
    7. Change the format of the dataset into a bed file
    8. Use the tool **bedtools SortBED** to sort the file by chromosome then by start position.
3. Save the workflow
4. Run the workflow
    1. Import the file hg38_ens105.bed into Galaxy.
    2. Apply the workflow « BED Ensembl to BED UCSC » on the file hg38_ens105.bed in a new history names « hg38_ens105: BED Ensembl to BED UCSC »

Exercise 1: Create your first workflow

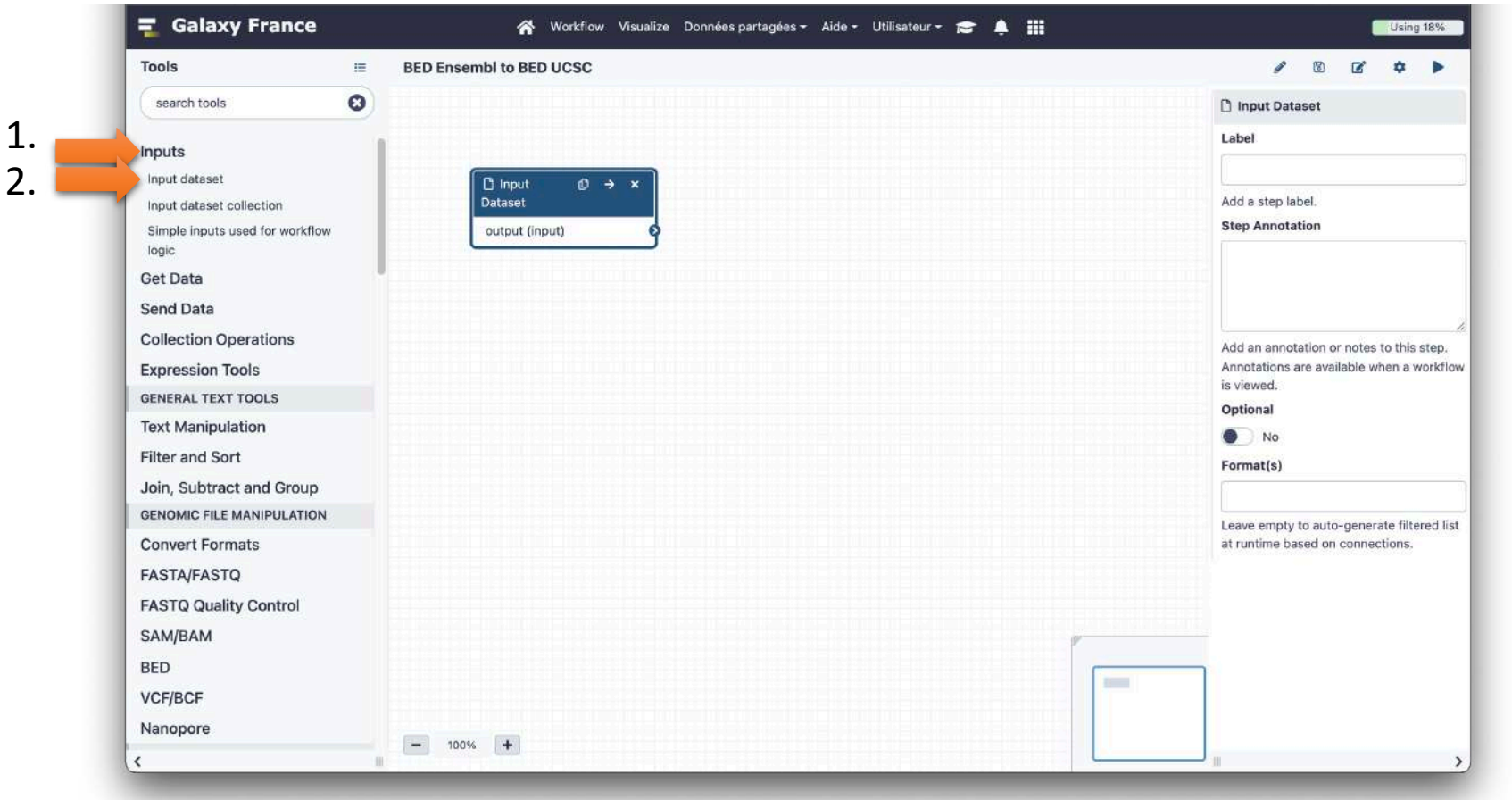
1. Create an empty workflow named « BED Ensembl to BED UCSC ».



Exercise 1: Create your first workflow

2.1.

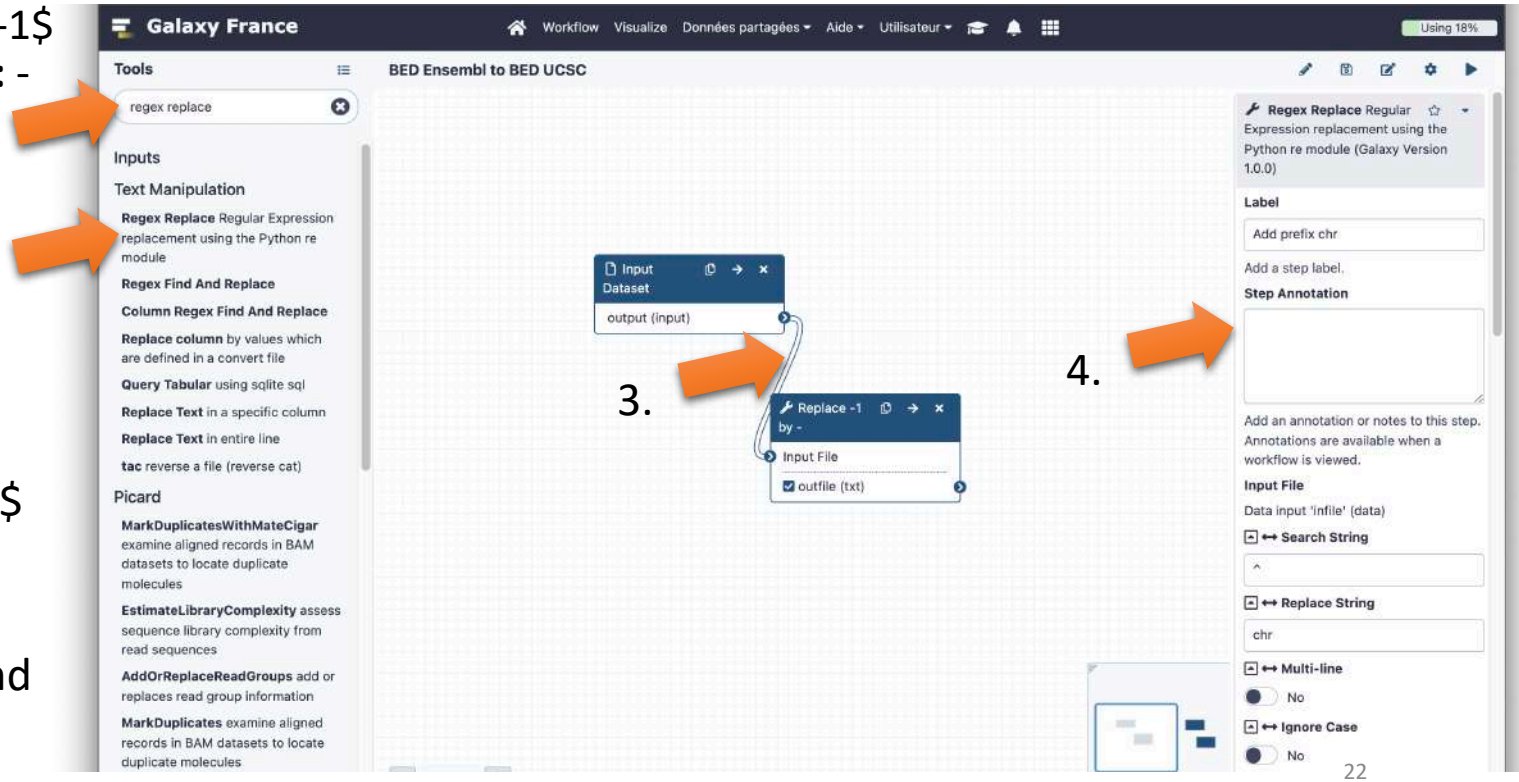
- Click on Inputs (1)
- Click on Input dataset (2) to add a box with an input dataset



Exercise 1: Create your first workflow

2.2.

- Search the tool « regex replace » in the tool panel (1)
- Click on **Regex Replace** to add a box for the tool (2)
- Link the two boxes « Input Dataset » and « Regex Replace » (3)
- Click on the box « Regex Replace » to edit the parameters: (4)
  - **Label**: Replace -1 by -
  - **Search String:** -1$
  - **Replace String:** -



Hint: the symbol -1$ means trailing -1. Here we want to replace -1 at the end of the line by a -.

Exercise 1: Create your first workflow

2.3.

- Search the tool regex replace in the tool panel (if needed)
- Click on **Regex Replace** to add a box for the tool
- Click on the box « Regex Replace » to edit the parameters:
  - **Label**: Replace 1 by +
  - **Search String:** 1$
  - **Replace String:** +
- Link the two boxes « Replace -1 by - » and « Replace 1 by + » *(the new one)*

Exercise 1: Create your first workflow

2.4.

- Search the tool regex replace in the tool panel (if needed)
- Click on **Regex Replace** to add a box for the tool
- Click on the box « Regex Replace » to edit the parameters:
  - **Label**: Add prefix chr
  - **Search String:** ^
  - **Replace String:** chr
- Link the two boxes « Replace 1 by + » and « Add prefix chr » *(the new one)*

Hint: the symbol « ^ » means the beginning of the line. Here we add the prefix « chr » at the beginning of the line in front of the chromosome names.

Exercise 1: Create your first workflow

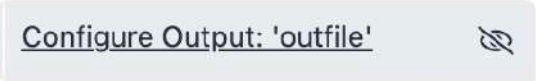
2.5.

- Search the tool remove beginning in the tool panel (if needed)
- Click on **Remove beginning** to add a box for the tool
- Click on the box « Remove beginning » to edit the parameters:
  - **Label**: Remove header line
  - **Remove first:** 1
- Link the two boxes « Add prefix chr » and « Remove header line » *(the new one)*

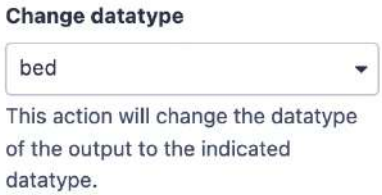
Exercise 1: Create your first workflow

2.6.

- Search the tool regex replace in the tool panel
- Click on **Regex Replace** to add a box for the tool
- Click on the box « Regex Replace » to edit the parameters:
  - **Label**: Replace chrMT by chrM
  - **Search String:** ^chrMT
  - **Replace String:** chrM
  - Click on Configure Output: 'outfile'



    - **Change datatype:** bed



- Link the two boxes « remove header line » and « Replace chrMT by chrM » *(the new one)*
- **Here we ask that the output format after this step is turned into a BED file rather than a text file (by default). This is done to fit the input of next step.**

Exercise 1: Create your first workflow

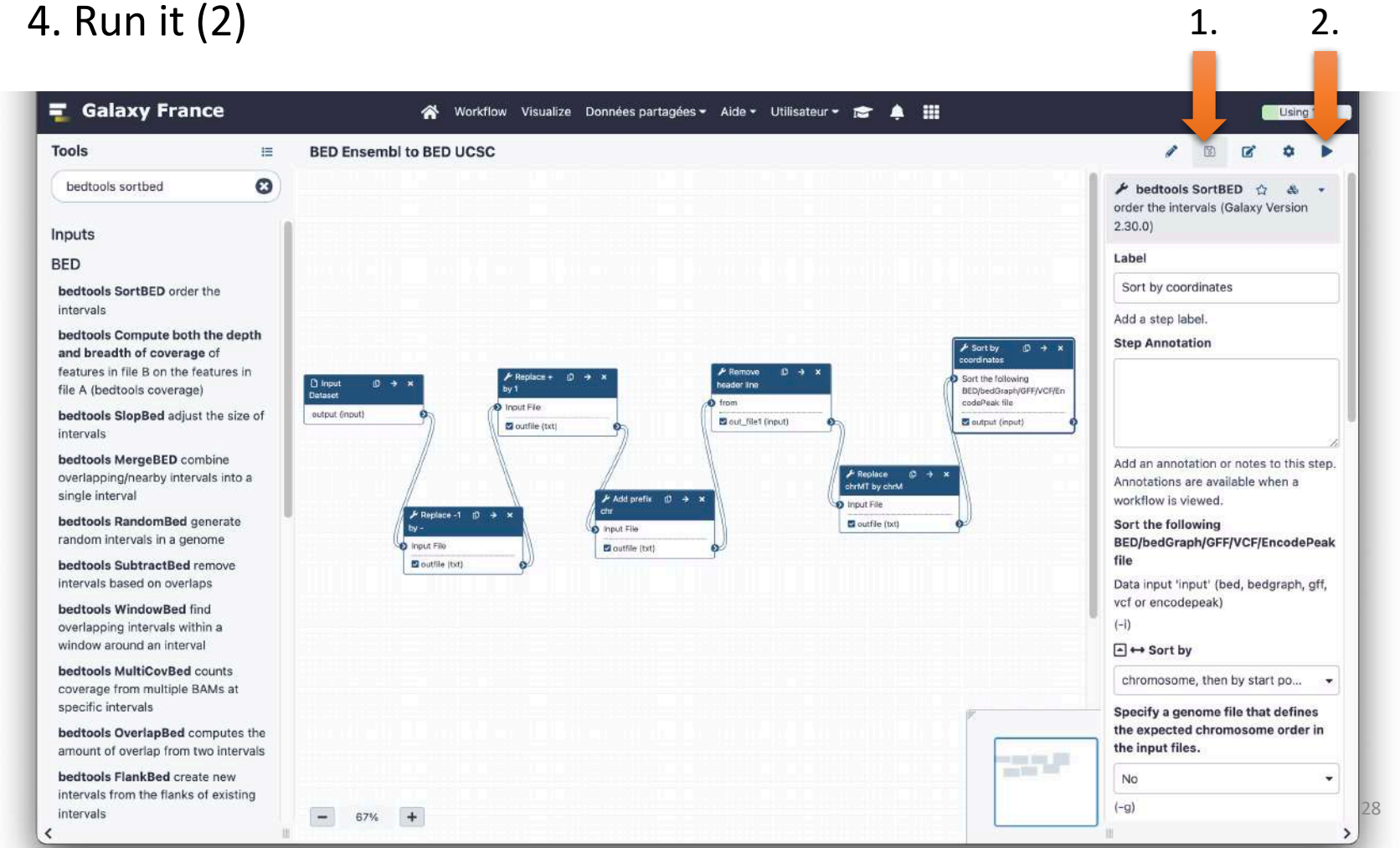
2.7.

- Search the tool bedtools sortbed in the tool panel (if needed)
- Click on **bedtools SortBed** to add a box for the tool
- Click on the box « bedtools SortBed » to edit the parameters:
  - **Label**: Sort by coordinates
- Link the two boxes « Replace chrMT by chrM » and « Sort by coordinates » *(the new one)*

Exercise 1: Create your first workflow

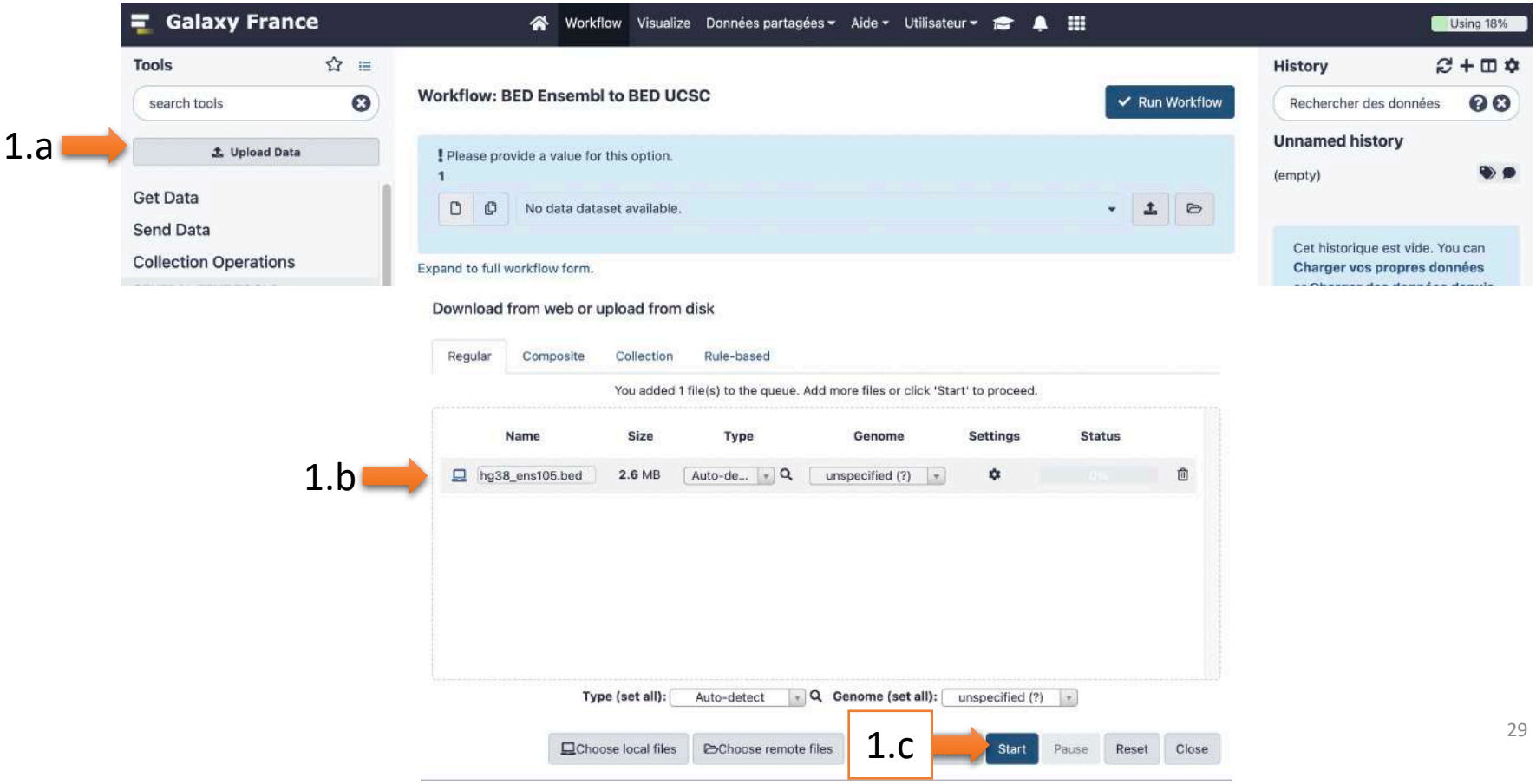
3. Save the workflow (1)

4. Run it (2)



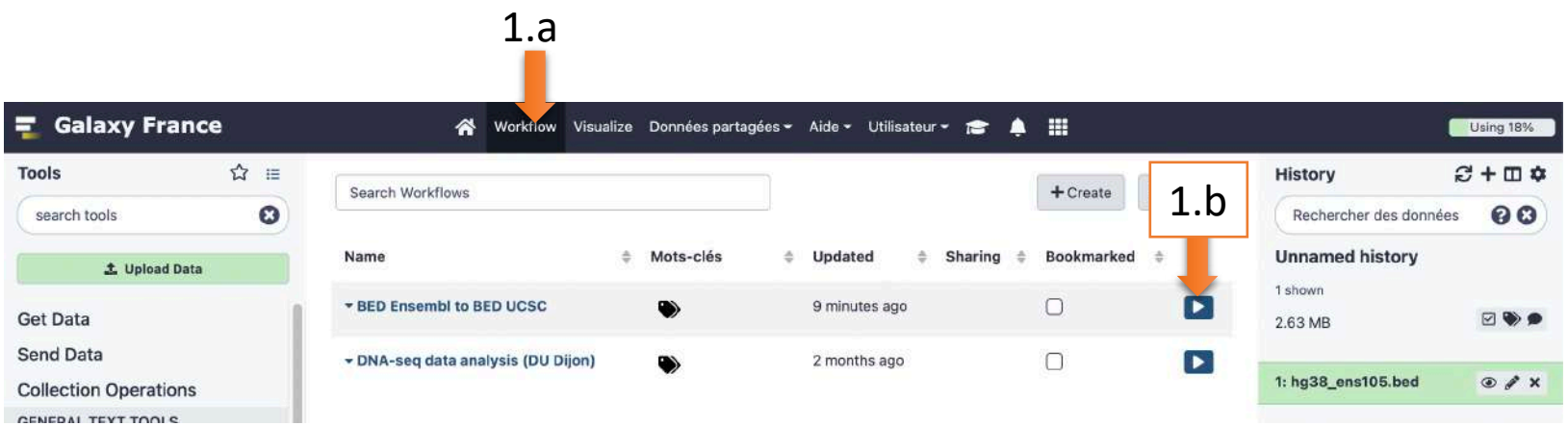
Exercise 1: Create your first workflow

4.1. Upload the file hg38_ens105.bed (1). If you don't have it anymore, find it in the directory ensembl.



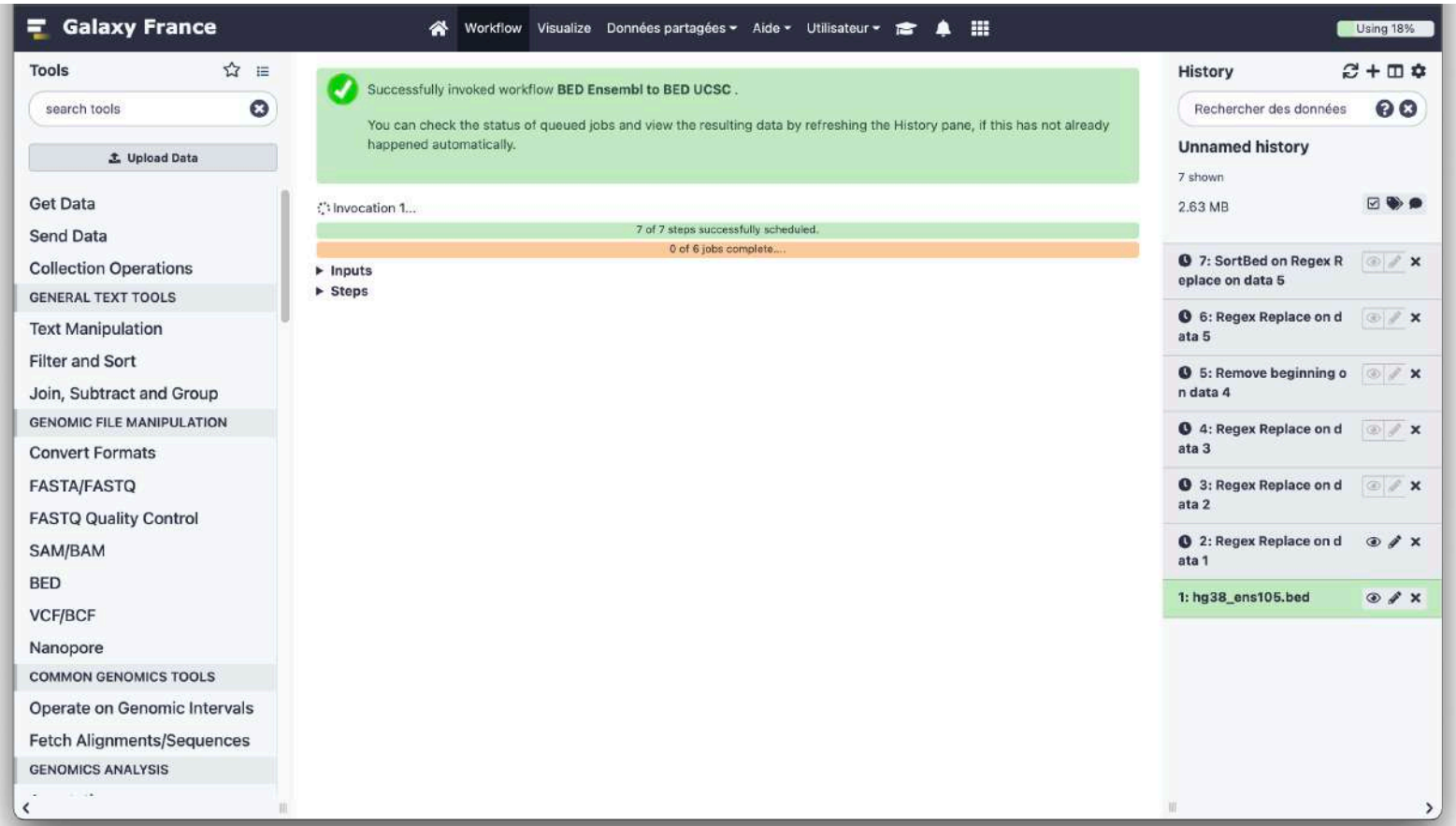
Exercise 1: Create your first workflow

4.2. Run the workflow (1)



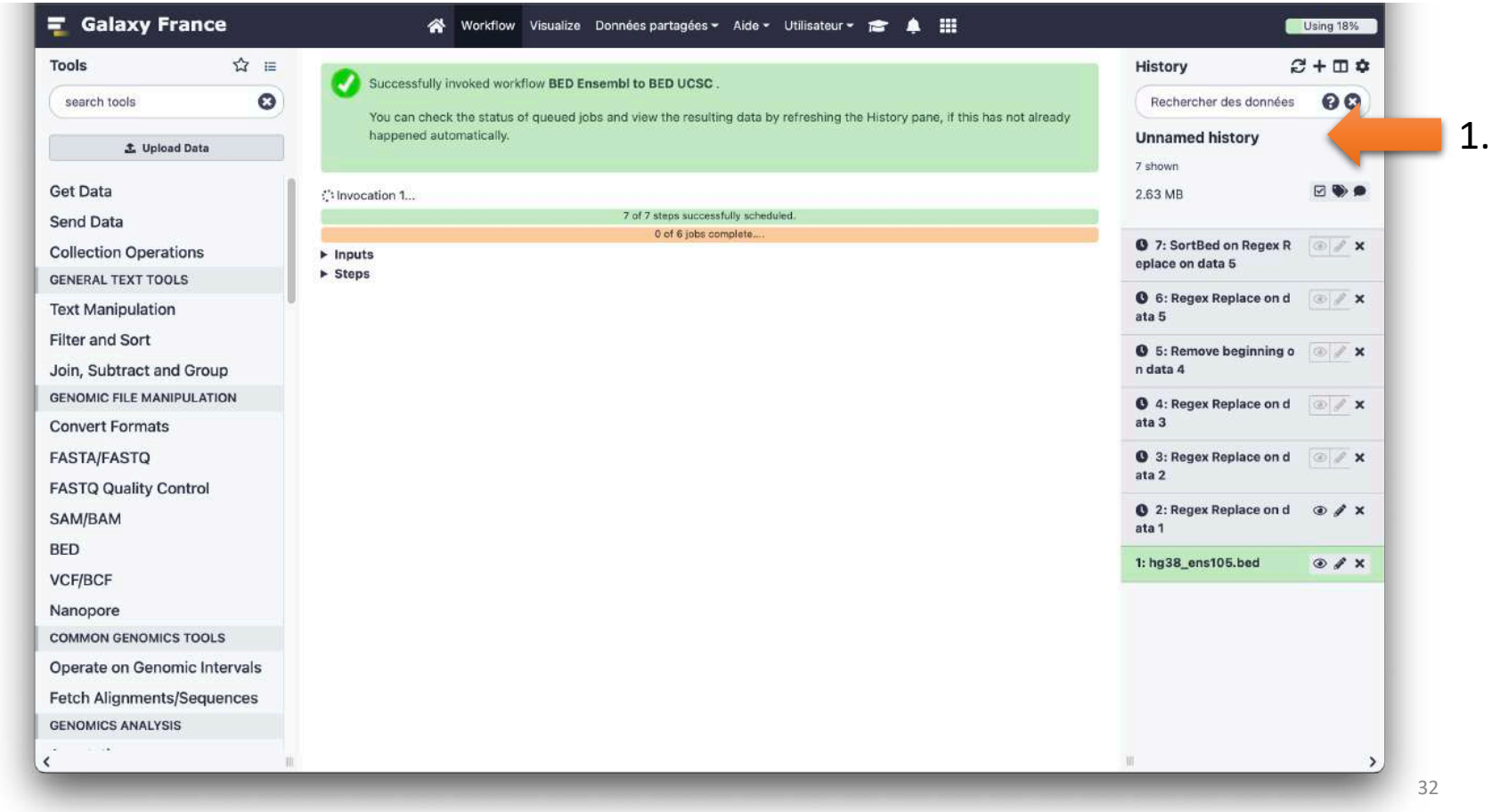
Exercise 1: Create your first workflow

The workflow is running



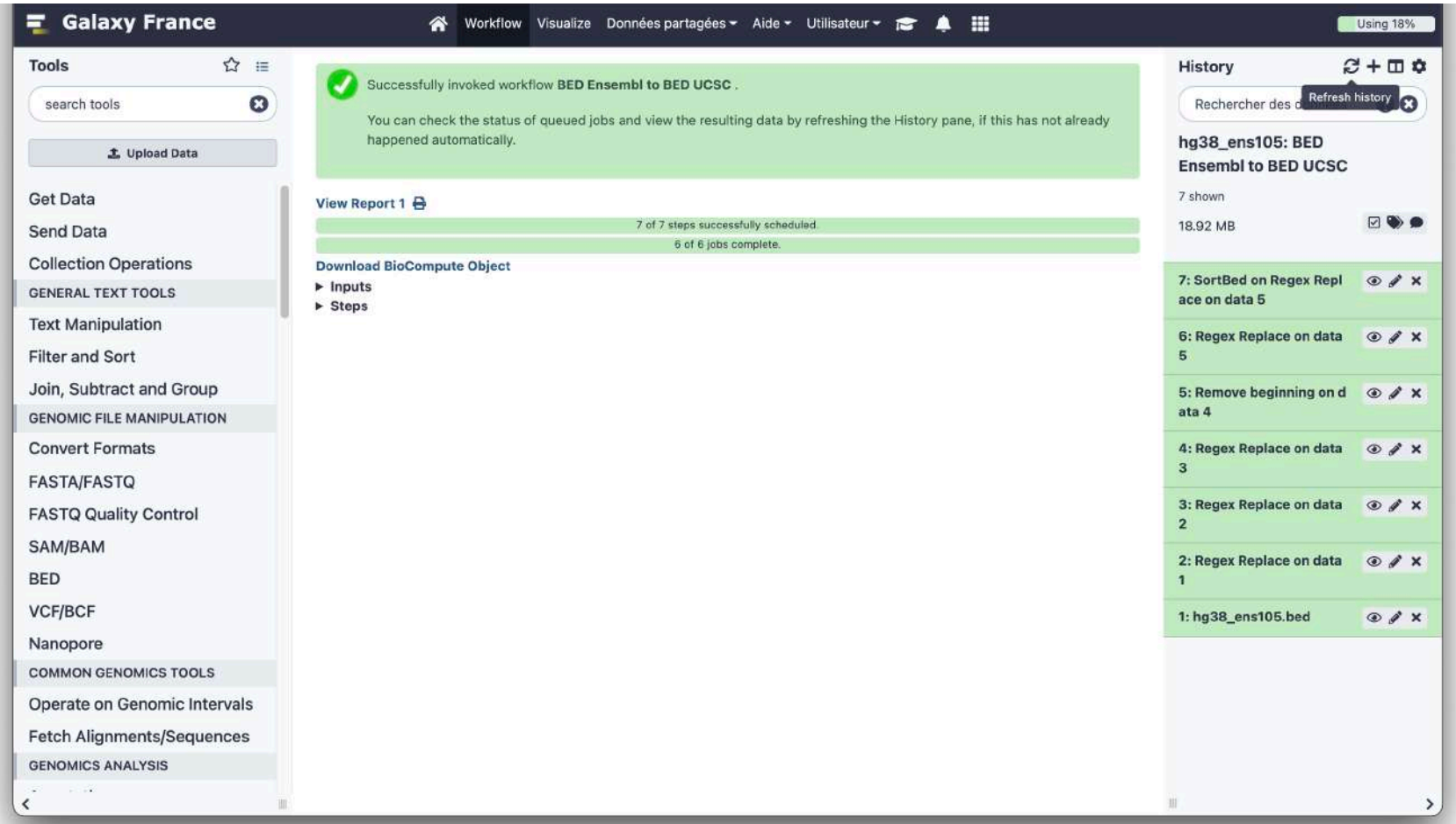
Exercise 1: Create your first workflow

Rename your history to « hg38_ens105: BED Ensembl to BED UCSC » (1)



Exercise 1: Create your first workflow

It’s all done!



Exercise 2: Extract a workflow out of an history

We want to create a workflow to automatically analyze chIP-seq data in Galaxy the same way we did during the ChIP-seq data analysis training. We are going to edit it so that instead of starting from aligned reads, it will start from raw reads (fastq file). We'll add 2 steps per input file:

- A quality control step
- A mapping step

Steps:

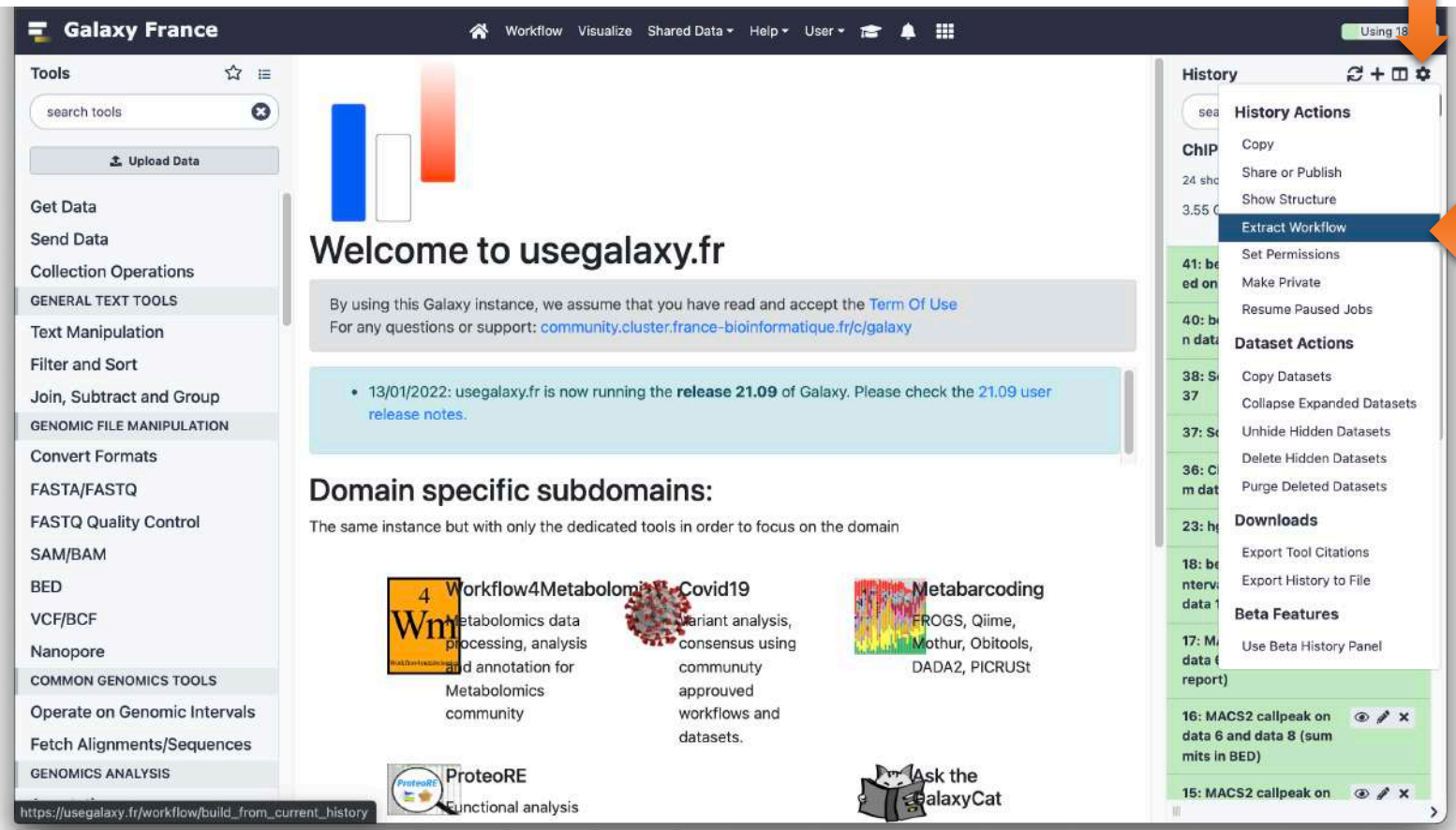
1. Extract a workflow from your history "ChIP-seq data analysis".
    1. Select the steps to keep in the workflow
    2. Name it "ChIP-seq data analysis workflow"
2. Edit the workflow
    1. Remove the link between the Input files, Mark duplicates and samtools flagstat.
    2. Add a step with **Fastqc** after the input file (do it for the two input files)
    3. Add a step with Bowtie 2 between the input file and Mark duplicates (do it for the two input files)
    4. Make the workflow working for any genome.
3. Save the workflow
4. Test it!

Exercise 2: Extract a workflow out of an history

1.1.

- Switch to the history « ChIP-seq data analysis »
- Extract the workflow from your history (1)

1.a.

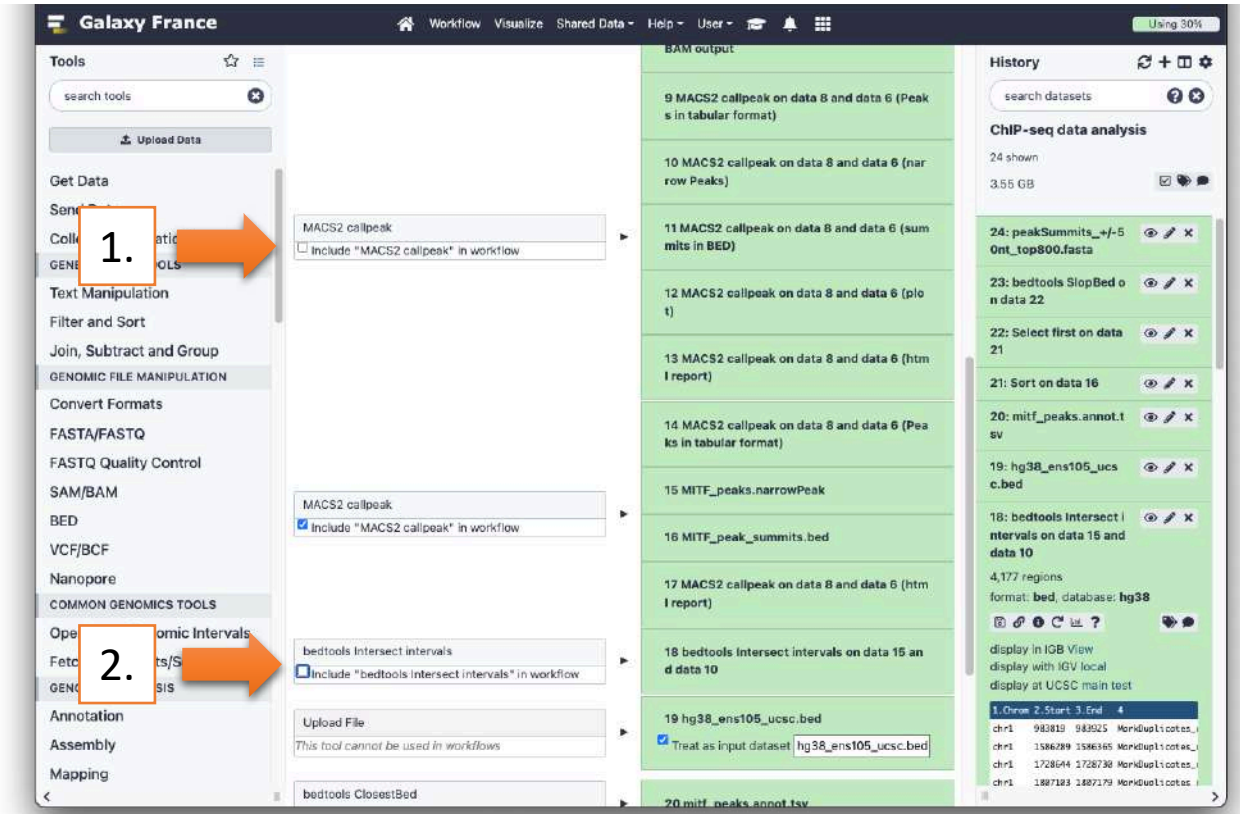


1.b.

Exercise 2: Extract a workflow out of an history

1.1

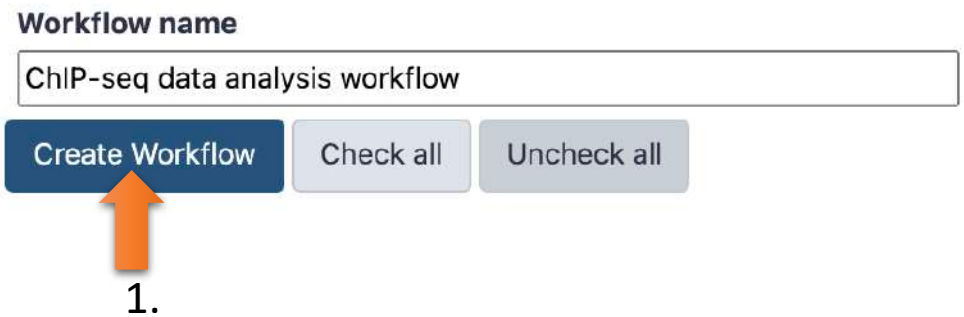
- Change the name to « ChIP-seq data analysis workflow”
- Select the steps to keep: keep all steps but:
  - the first run of MACS (1)
  - Bedtools intersect intervals (2)



Exercise 2: Extract a workflow out of an history

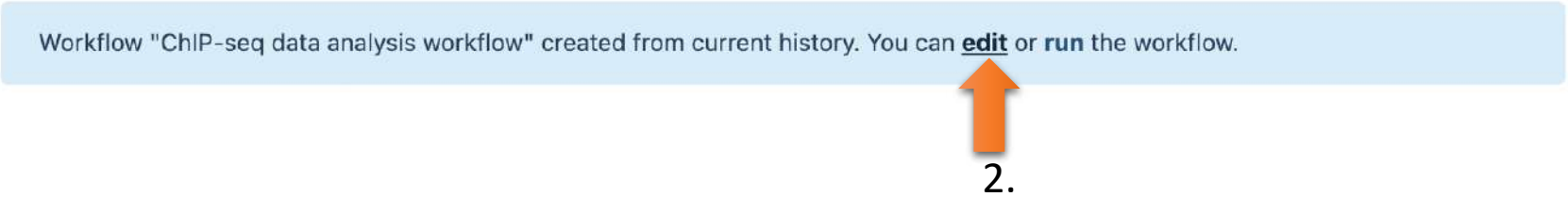
1.1.

- Click on Create Workflow (1)



1.2.

- Click on edit to edit the workflow (2)

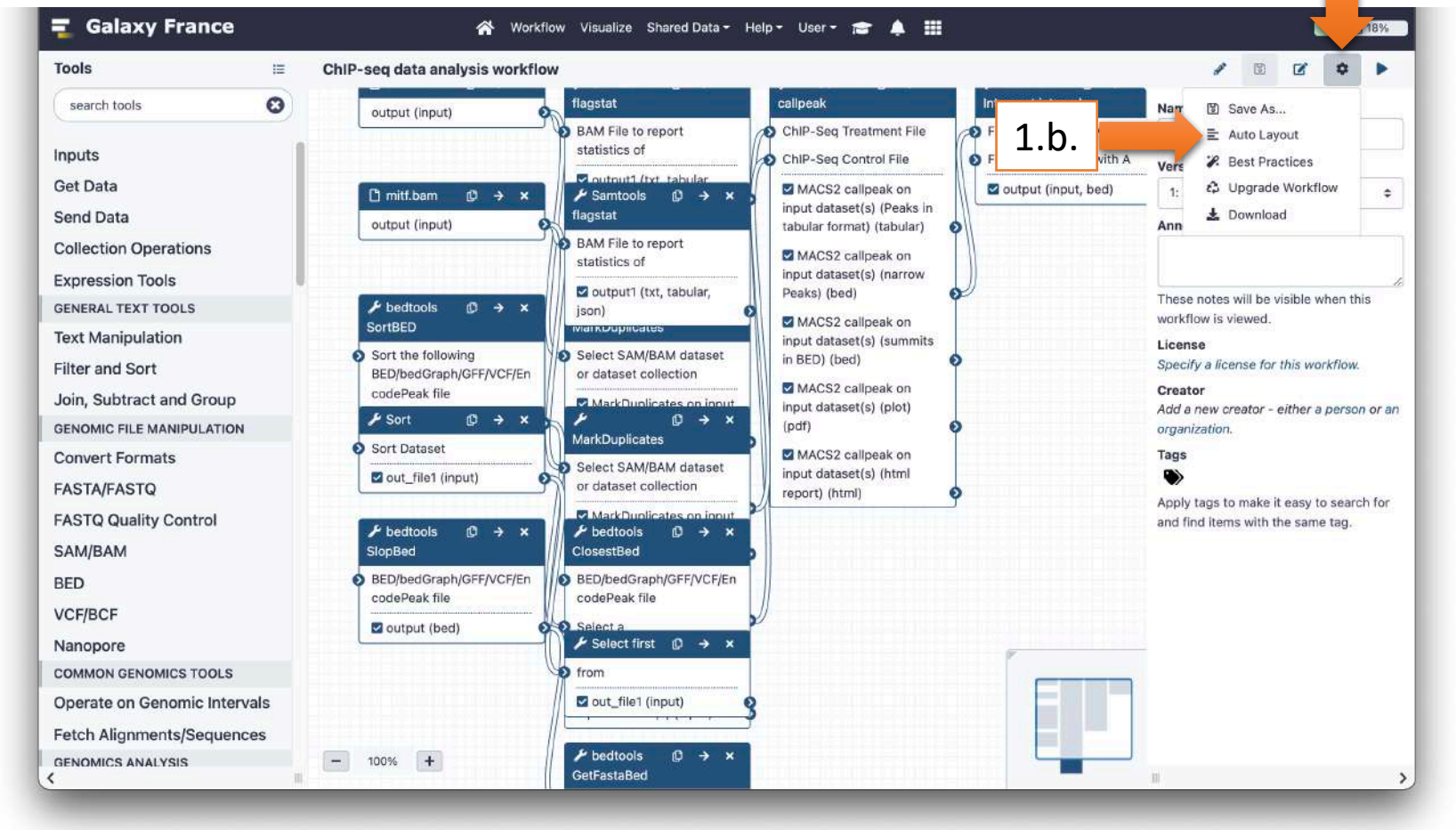


Exercise 2: Extract a workflow out of an history

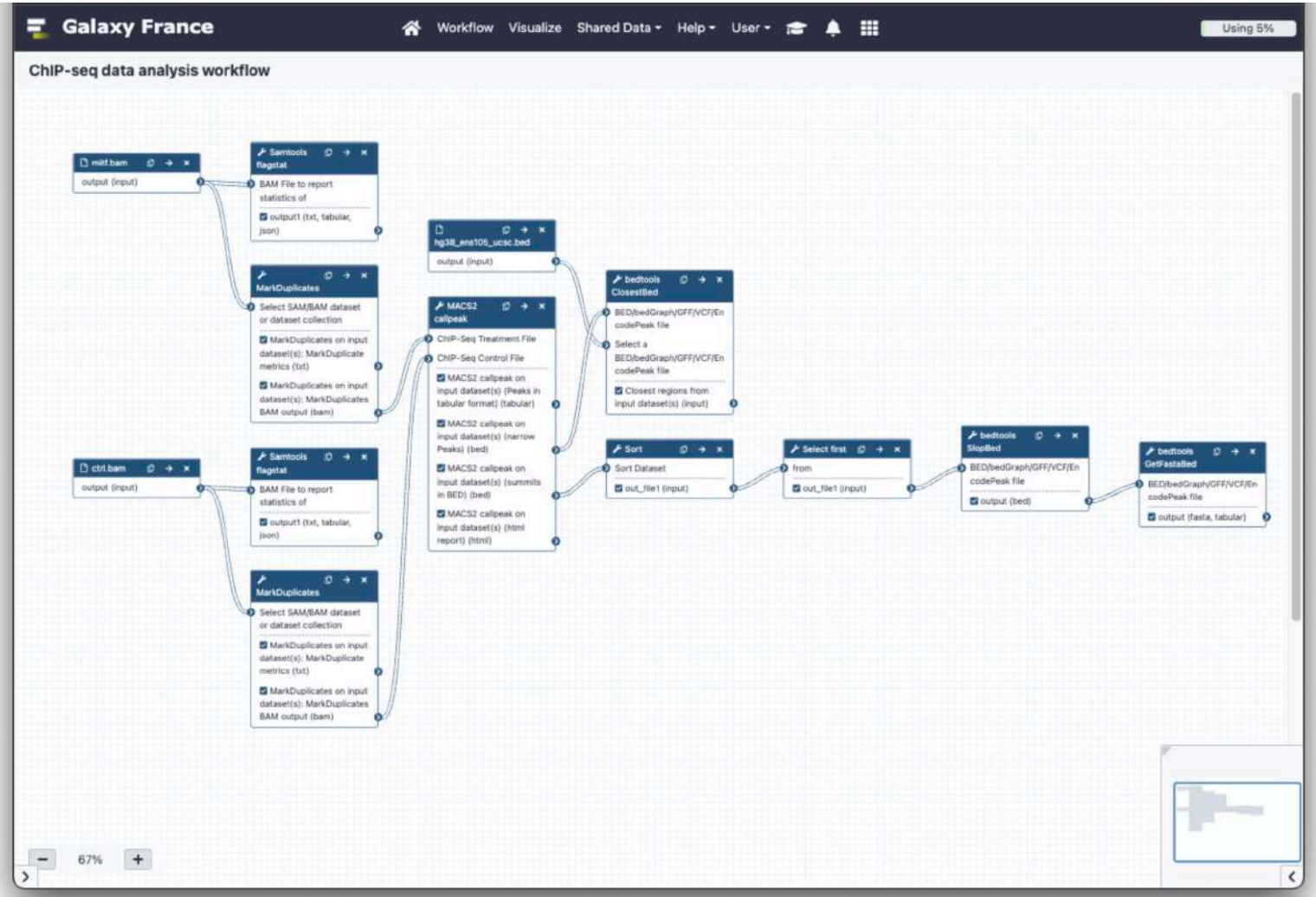
1.2.

- Re-order the boxes (1).

1.a.



Exercise 2: Extract a workflow out of an history

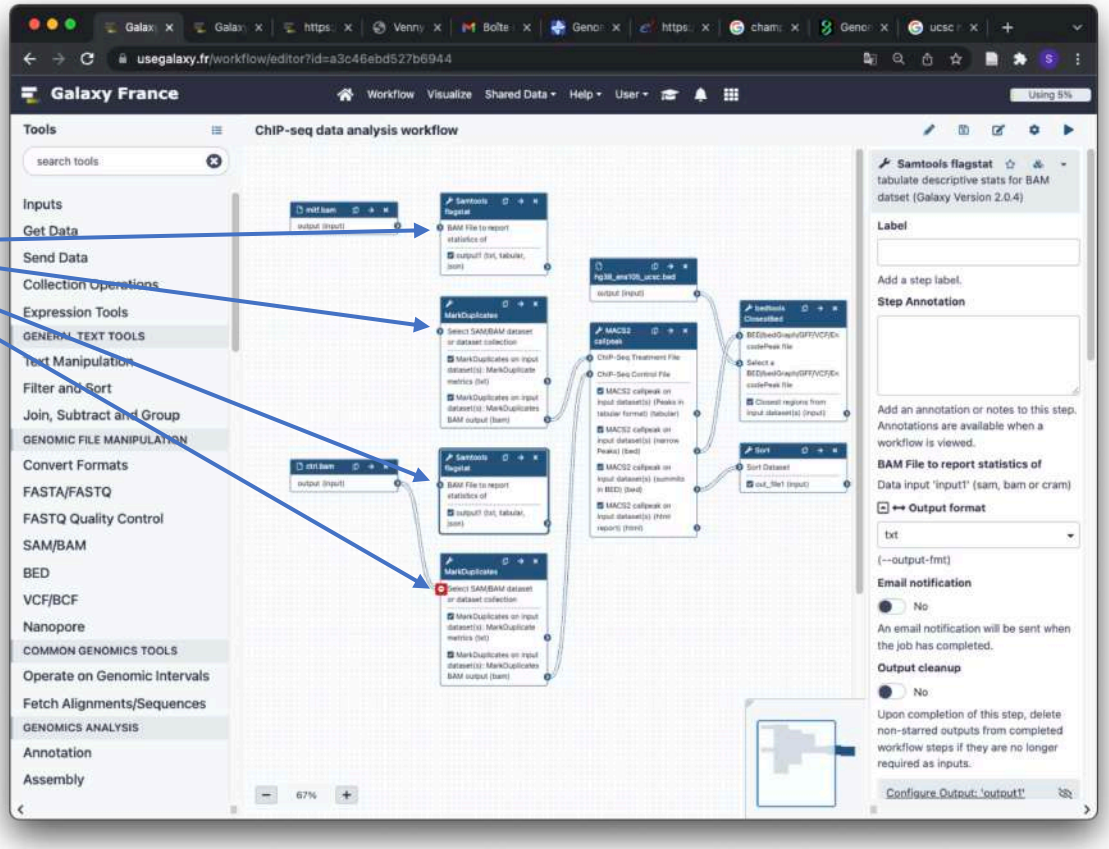


Let's edit it and make it general so that it'll work on many datasets.

Exercise 2: Extract a workflow out of an history

2.1.

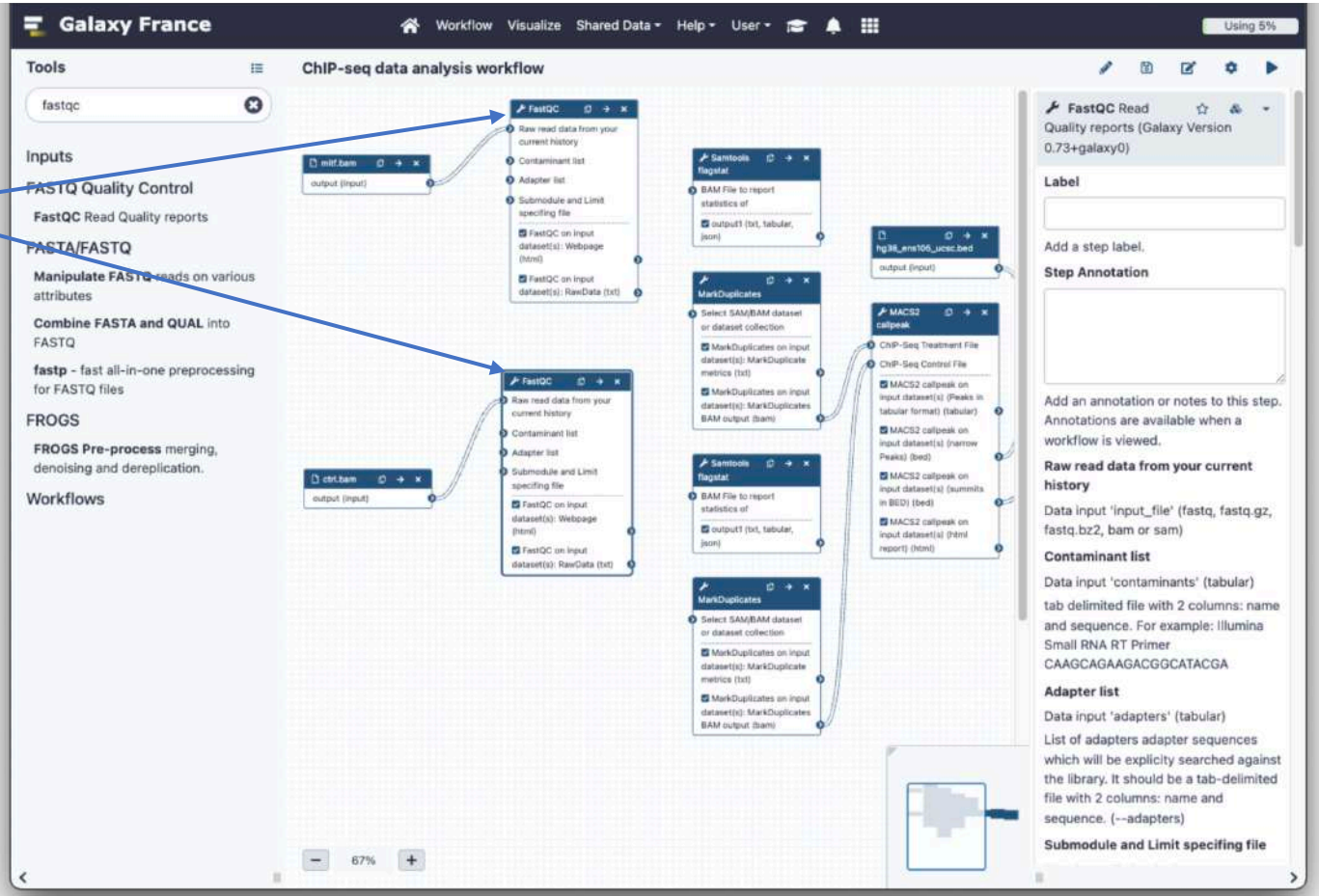
- Remove the links between the inputs files and the steps after



Exercise 2: Extract a workflow out of an history

2.2

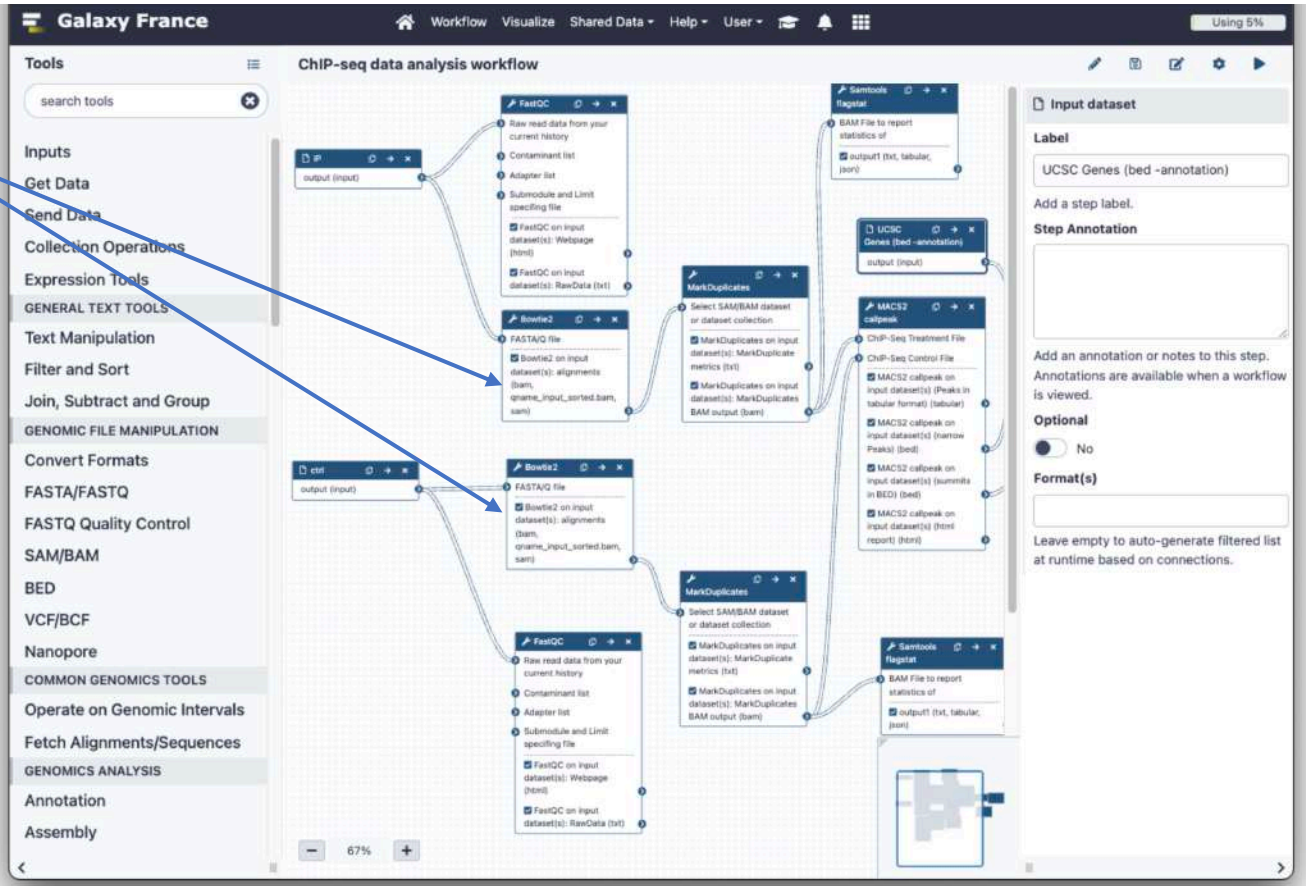
- Add 1 fastqc step per input file



Exercise 2: Extract a workflow out of an history

2.3.

- Add a step with Bowtie2 for each of the input file.

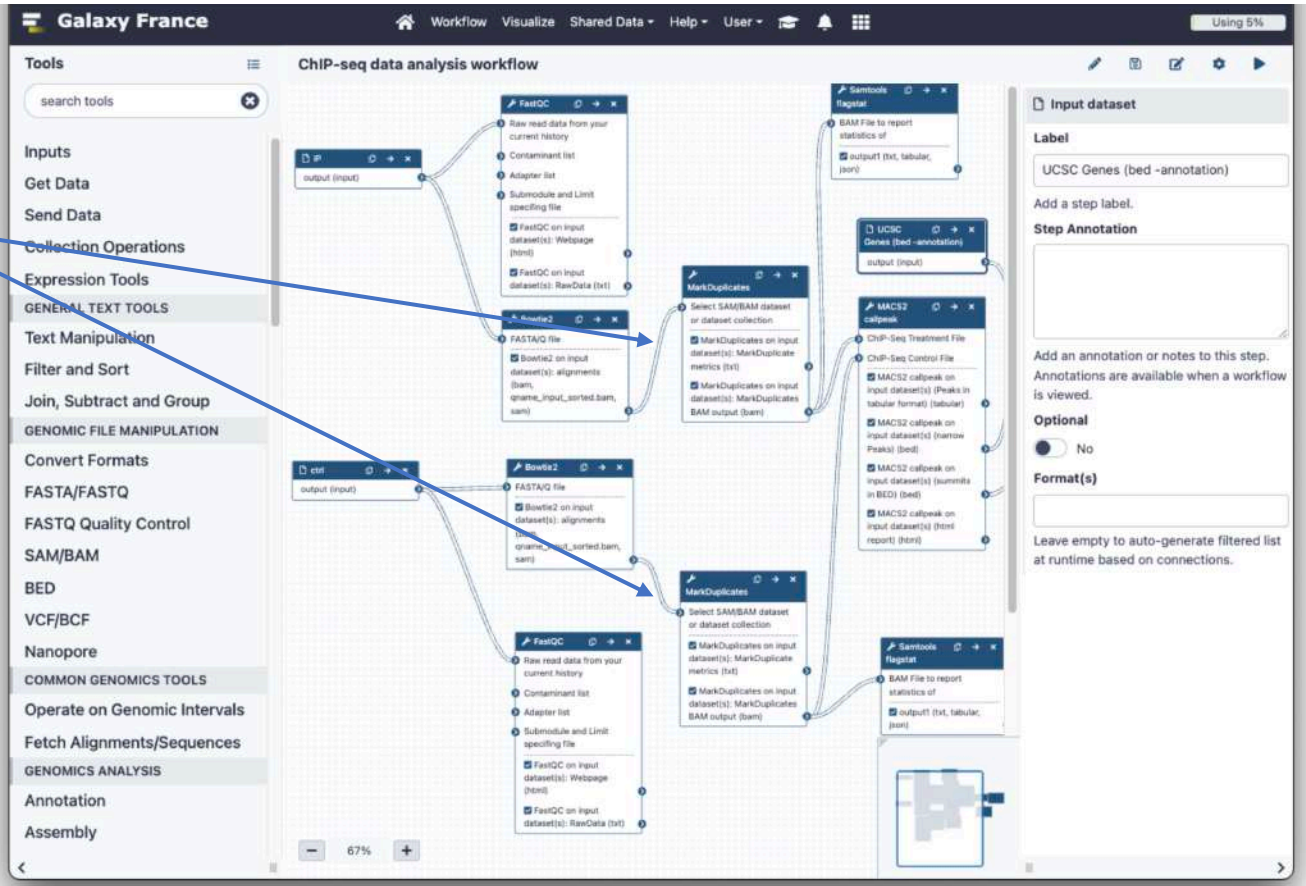


Exercise 2: Extract a workflow out of an history

2.3.

Edit the links:

- Bowtie2 output is linked to MarkDuplicates

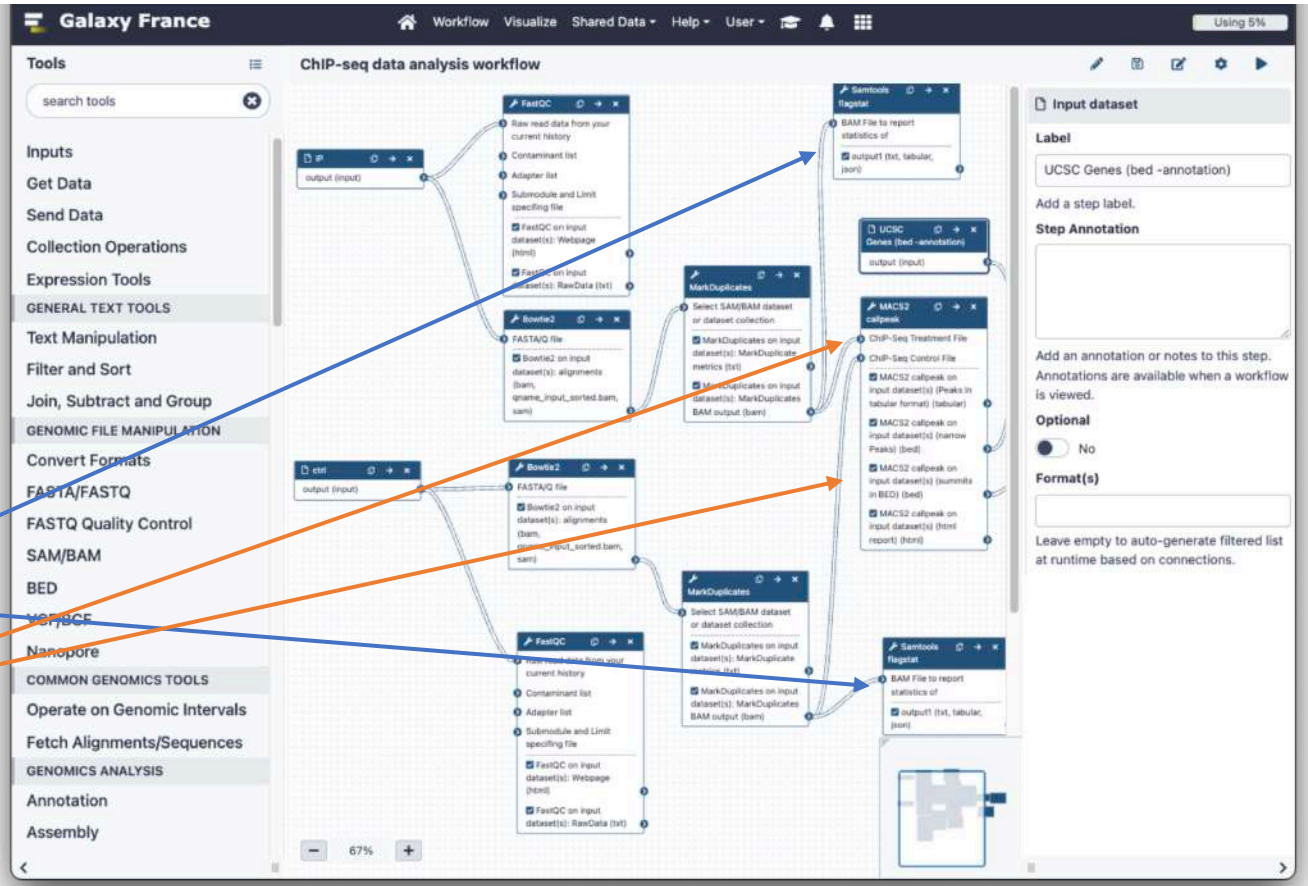


Exercise 2: Extract a workflow out of an history

2.3.

Edit the links:

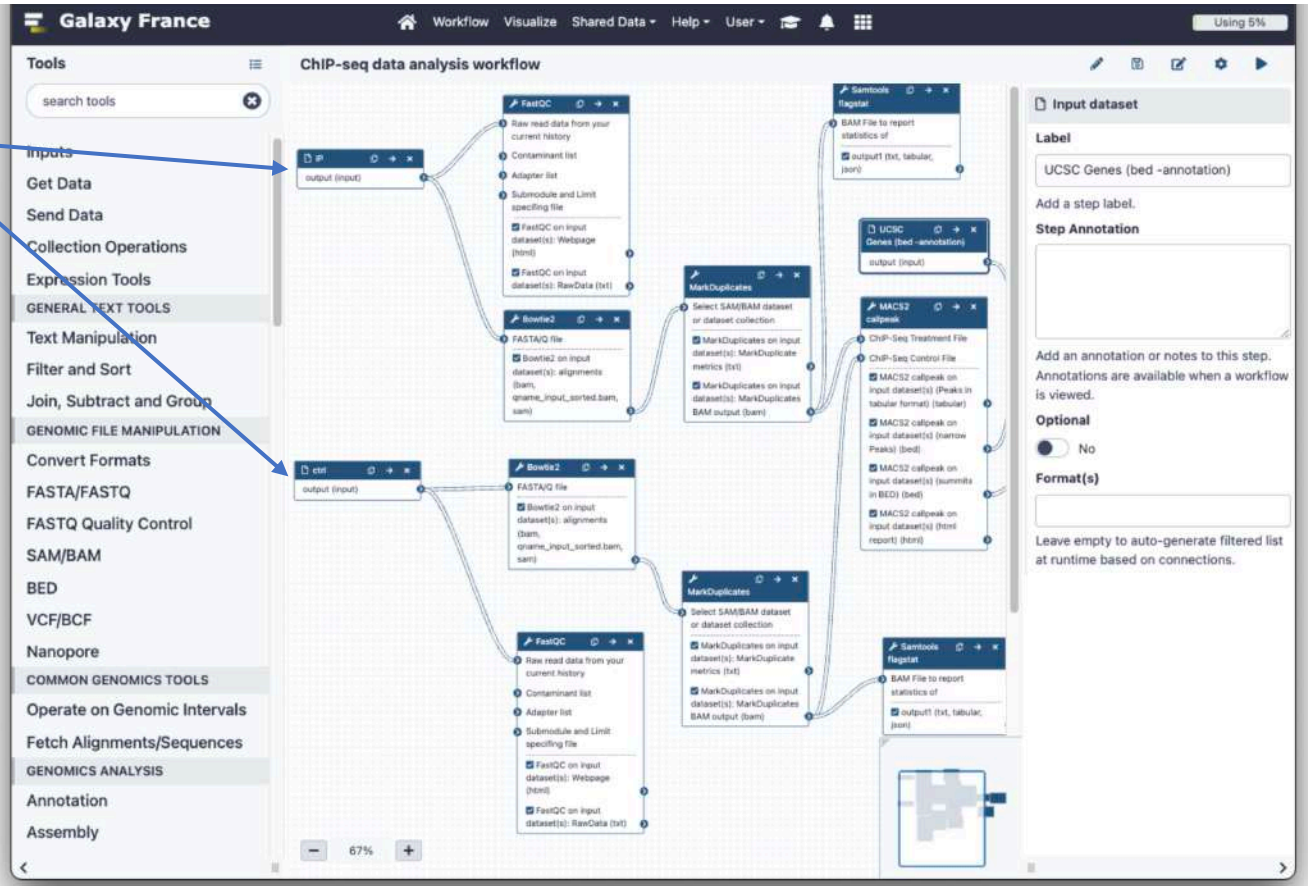
- Bowtie2 output is linked to MarkDuplicates
- MarkDuplicates output is linked to :
  - Samtools flagstat
  - MACS2



Exercise 2: Extract a workflow out of an history

2.3.

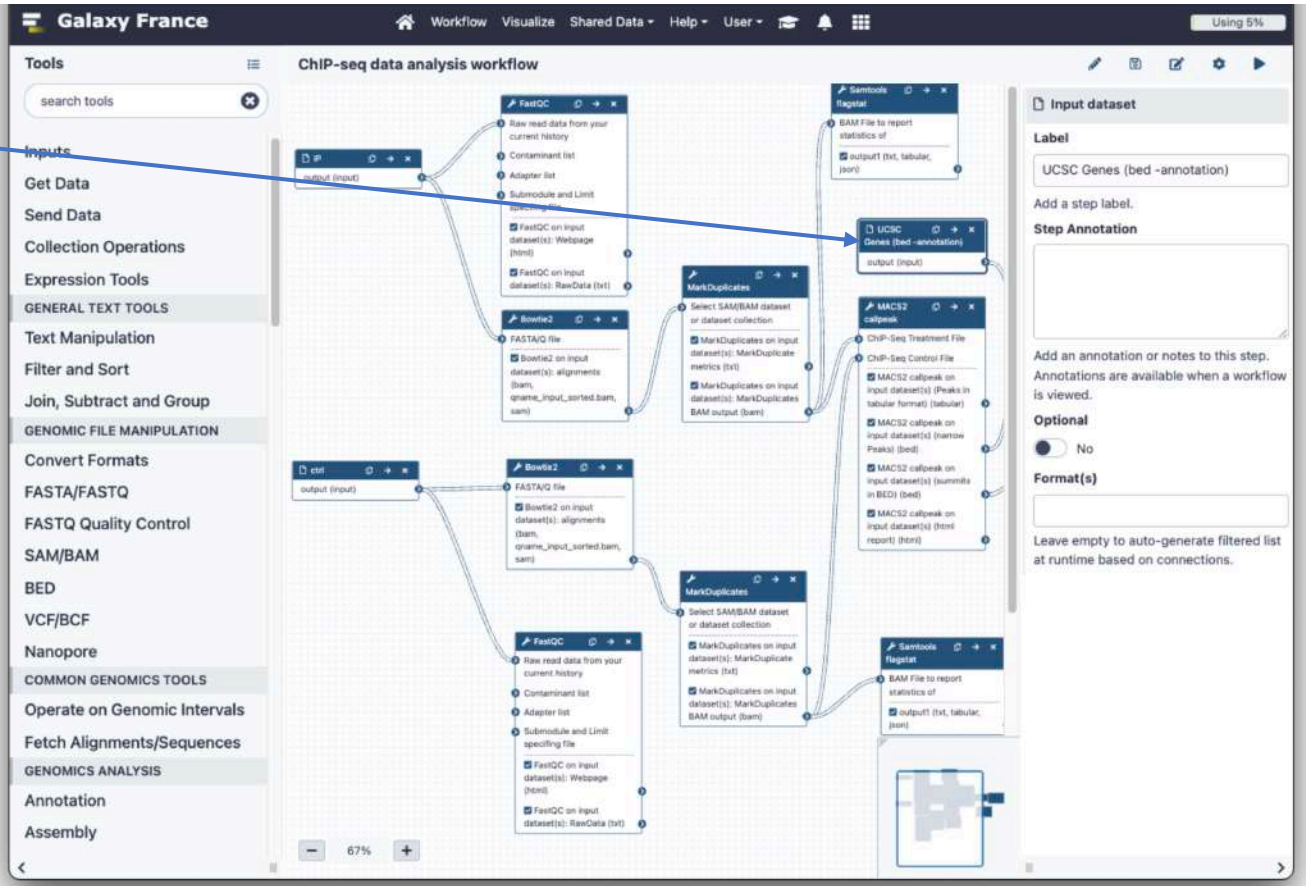
- Change the name of the two inputs:
  - Mitf.bam -> IP
  - Ctrl.bam -> Ctrl



Exercise 2: Extract a workflow out of an history

2.3.

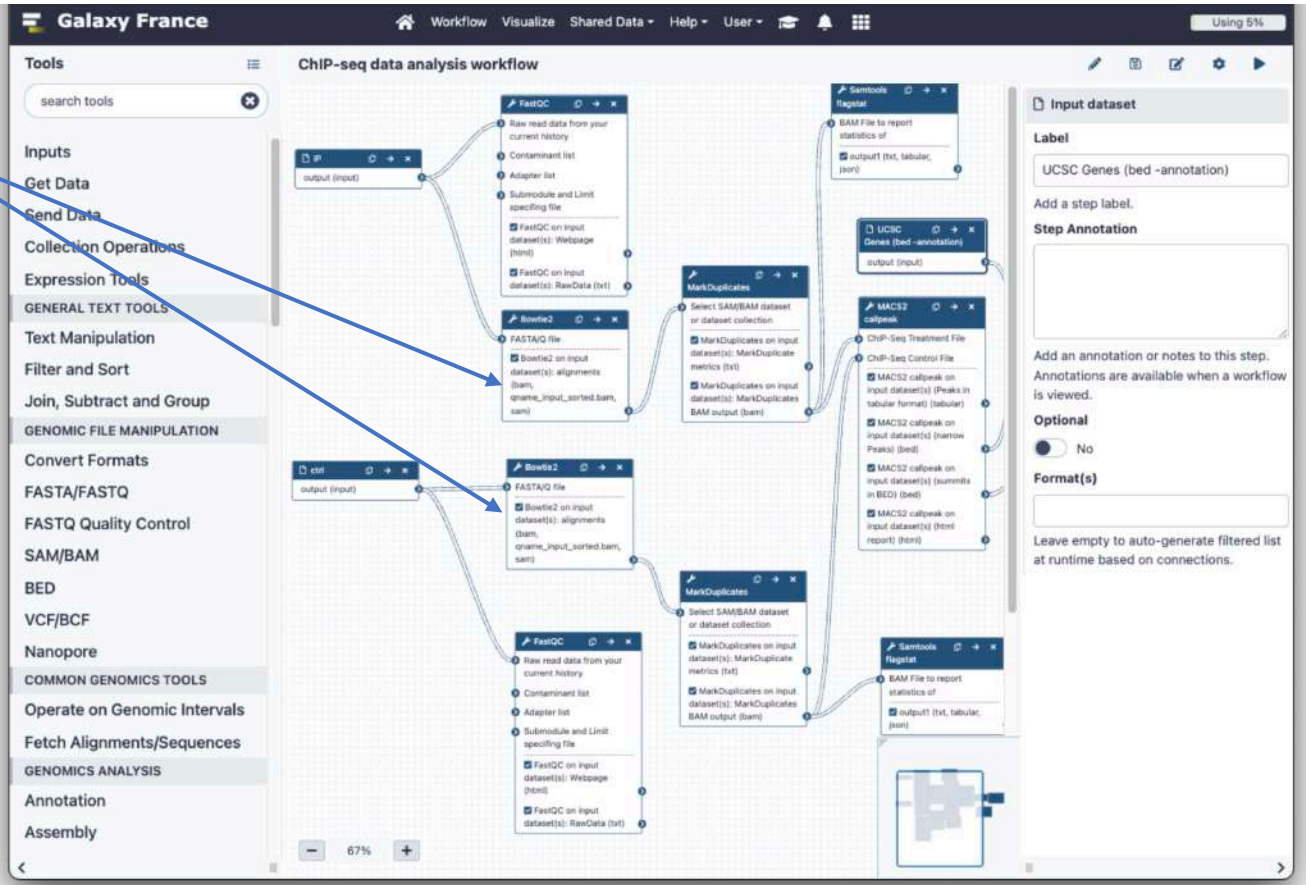
- Change the name of the file used to annotate peaks:
  - hg38_ens105_ucsc.bed -> UCSC Genes (bed -annotation)



Exercise 2: Extract a workflow out of an history

2.4.

- Edit Bowtie2 parameters:
  - **Will you select a reference genome from your history or use a built-in index?:** Use a built-in genome index
  - Select a reference genome



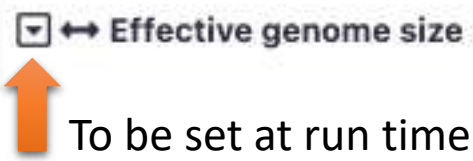
Select reference genome

To be set at run time

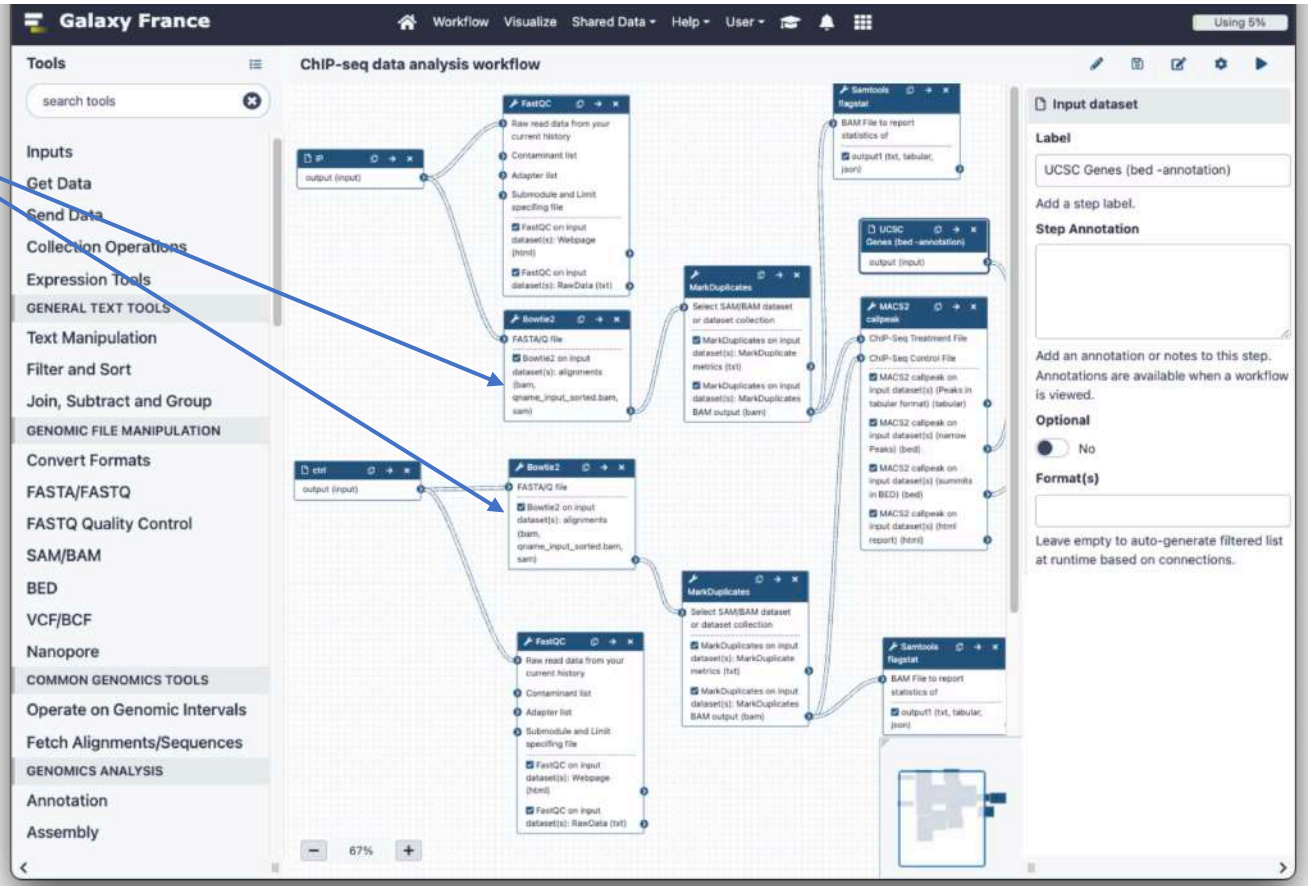
Exercise 2: Extract a workflow out of an history

2.4.

- Edit MACS2:
  - **Build Model:** Build the shifting model
  - **Effective genome size:** User defined



To be set at run time

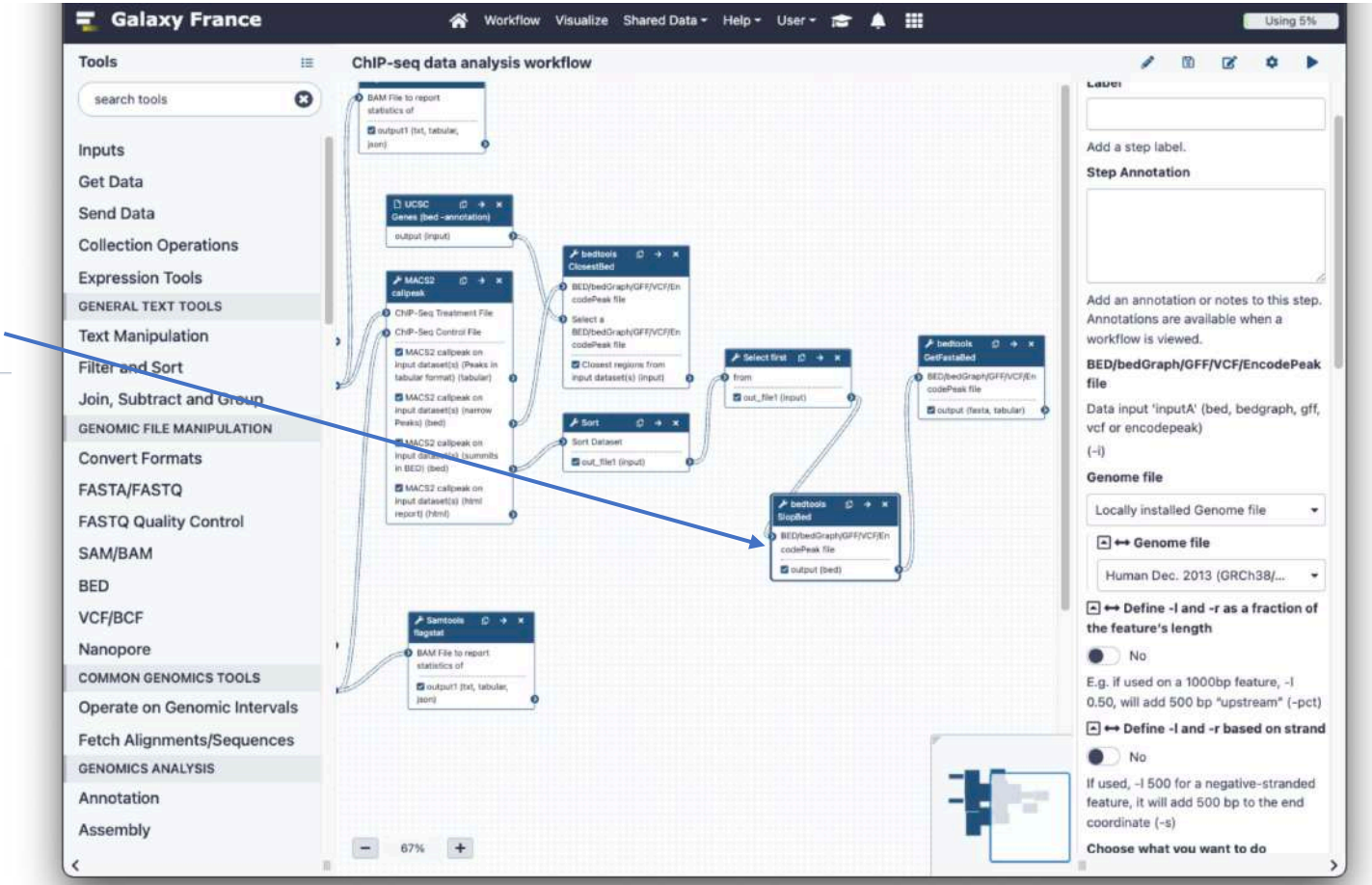


Exercise 2: Extract a workflow out of an history

2.4.

- Edit Bedtools SlopBed:
  - **Genome file:** Locally installed Genome file

To be set at run time

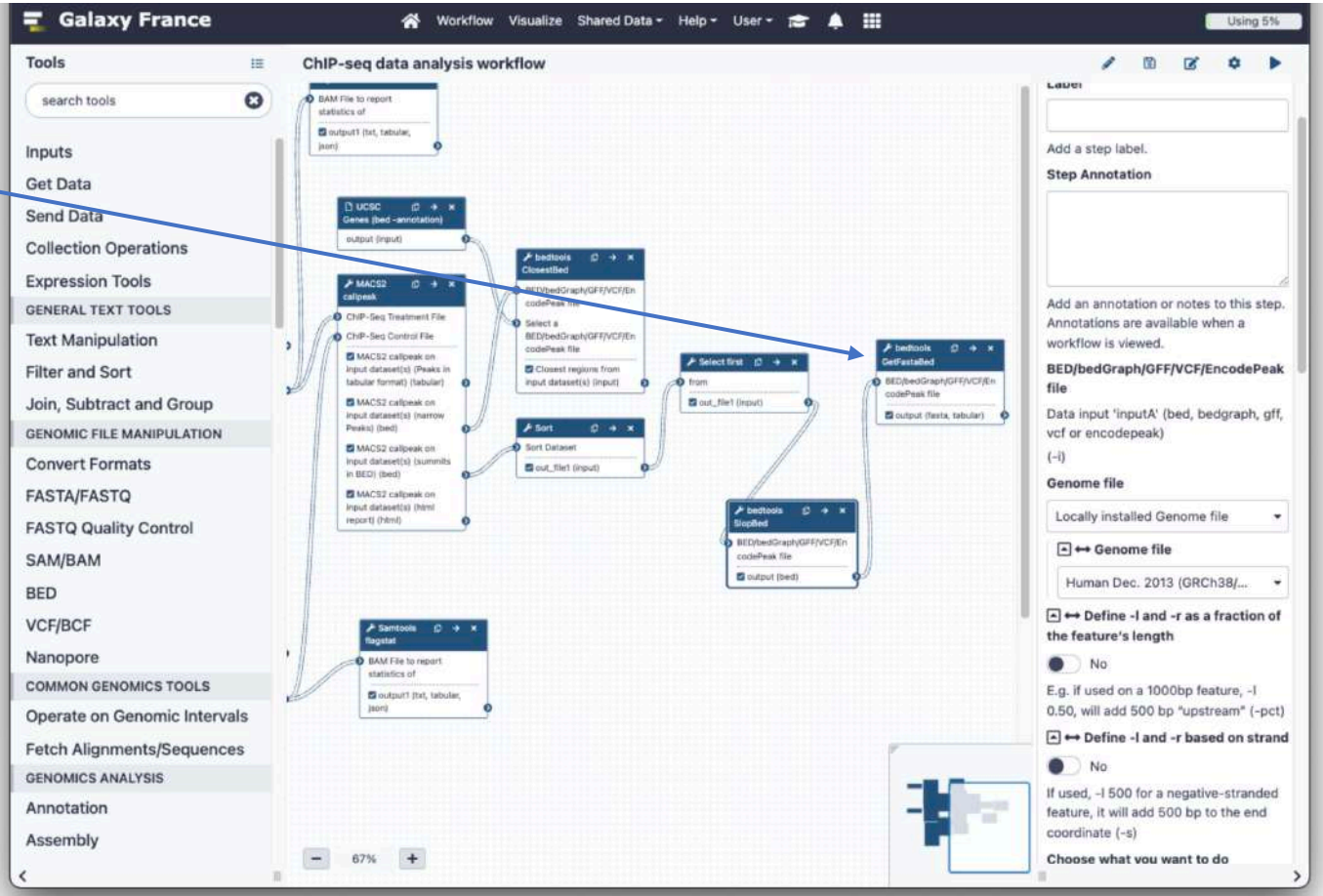


Exercise 2: Extract a workflow out of an history

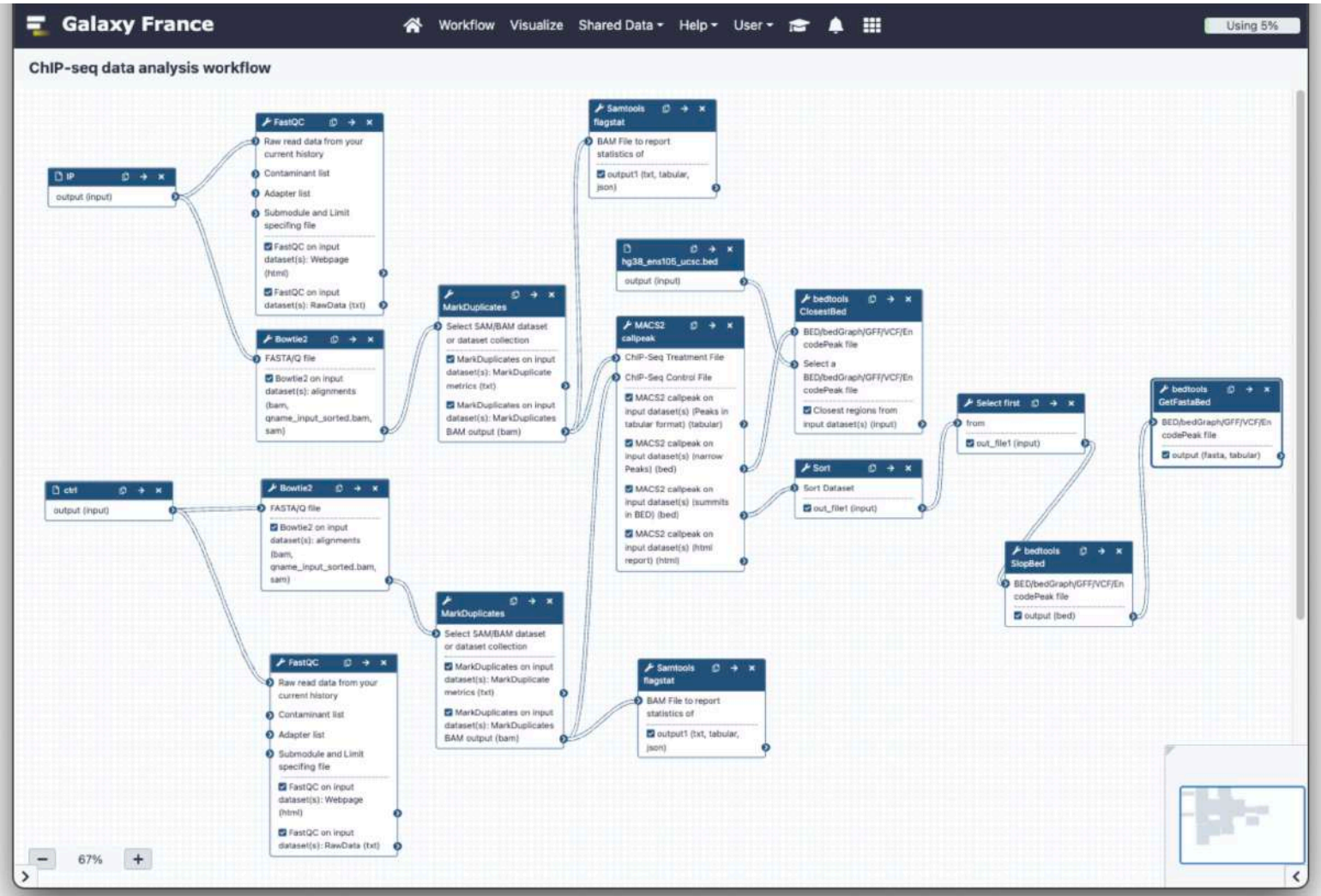
2.4.

- Edit Bedtools getFastaBed:
  - **Choose the source for the FASTA file:** Server indexed files

To be set at run time



Exercise 2: Final Workflow



Exercice 2: Try it out

- Input files (from imported history « NGS data analysis training Strasbourg »):
  - 26:Mitf_chr10.fastq
  - 27:Ctrl_chr10.fastq
  - 25:hg38_ens105_ucsc.bed
- Parameters:
  - Genome: use hg38 each time it is requested (Bowtie2, bedtools SlopBed, bedtools GetFastaBed)
  - MACS2: Effective Genome size: 107037937 (80% of chr10 length).

RNAseq workflow ?

Quality

FastQC

FASTQ Groomer

Mapping

STAR

Possibility to use a GTF on our Galaxy server as Gene Annotation Mode

RSeQC |

Quantification

HTSeq-count or **featureCounts**

Possibility to use a GTF on our Galaxy server as Gene Annotation Model

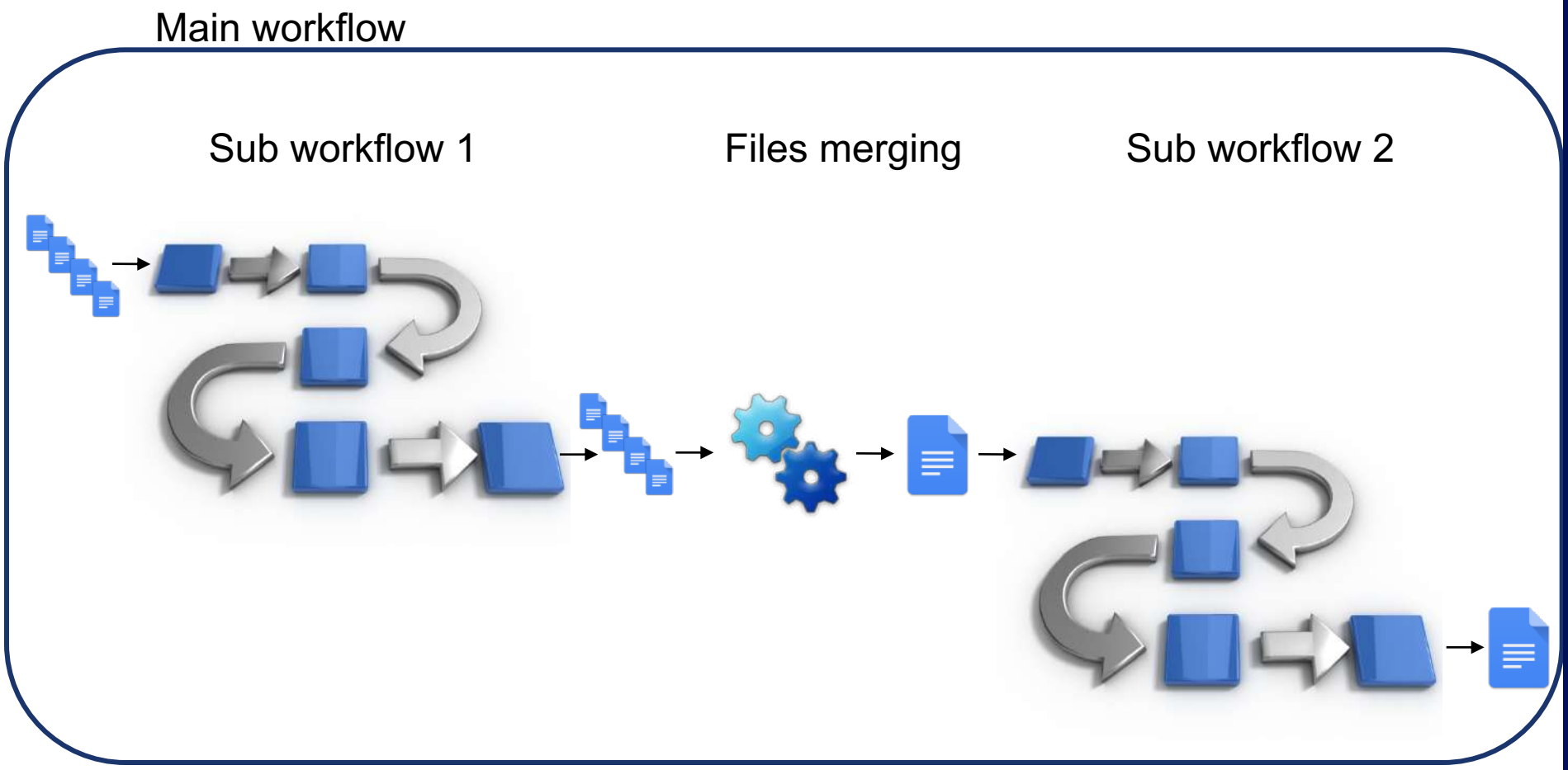
Statistical analysis

SARTools

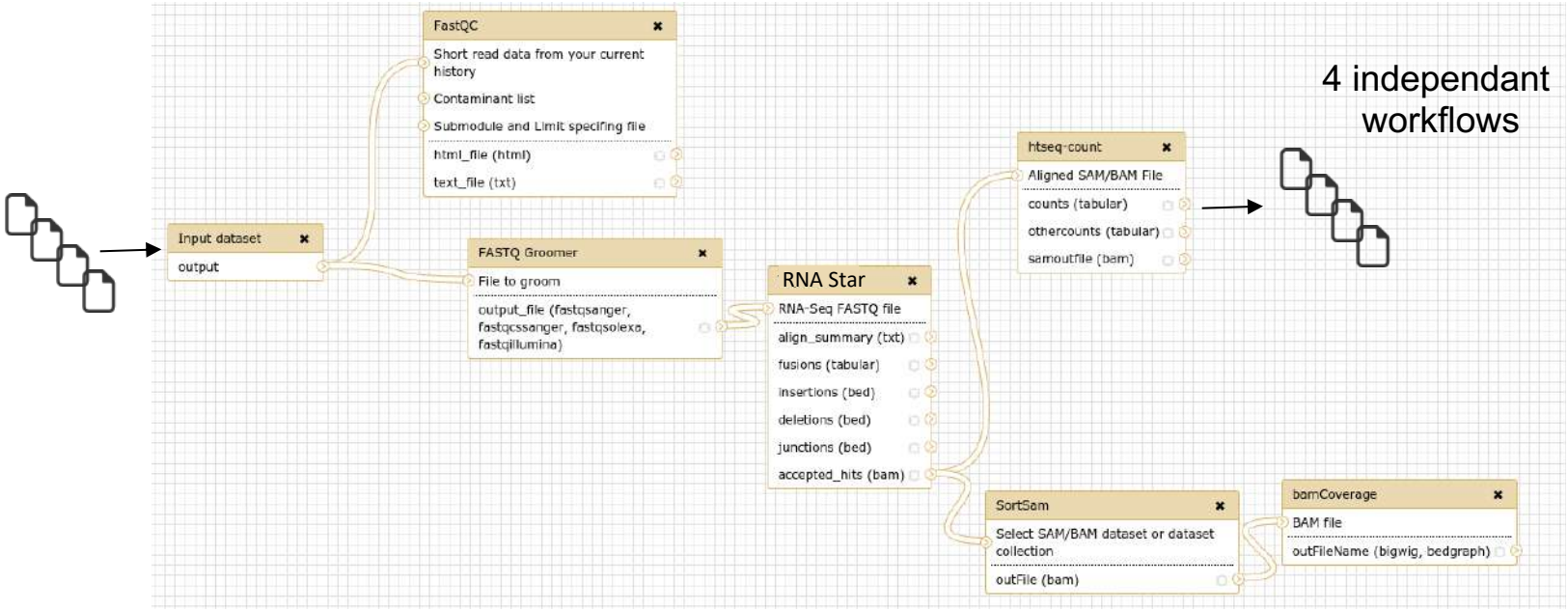
RNAseq Data Annotation
(own script)

Problem : all steps can't be in a same workflow

RNAseq workflow : limits



RNAseq workflow : limits



HTSeq-count outputs

→ Merge → SARTools