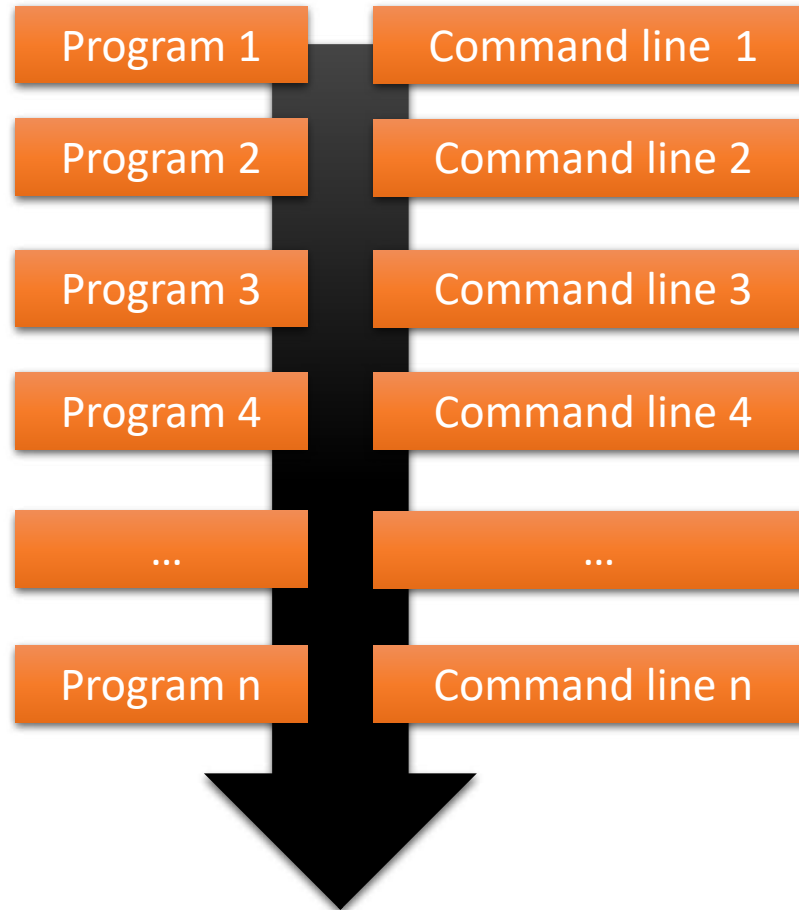
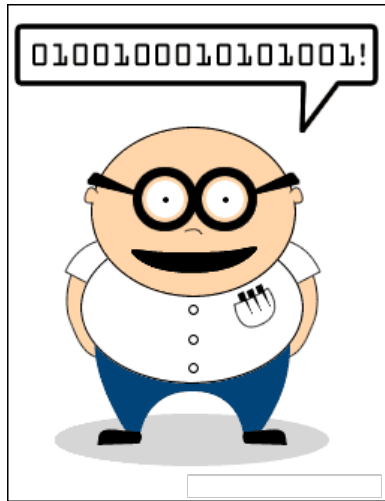


# Automatization of NGS data analysis : Galaxy workflows

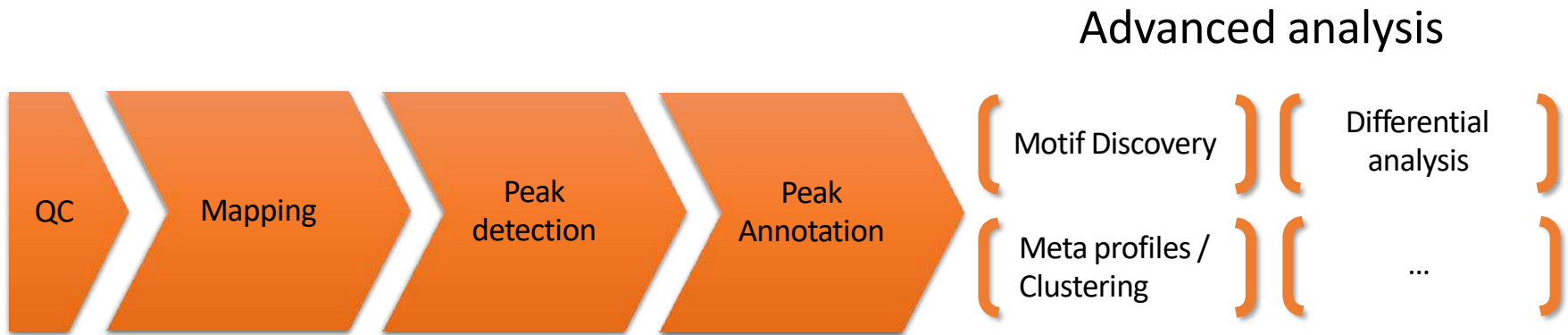
Stéphanie Le Gras  
([slegras@igbmc.fr](mailto:slegras@igbmc.fr))

# A long time ago...



**PIPELINE /  
WORKFLOW**

# More recently...



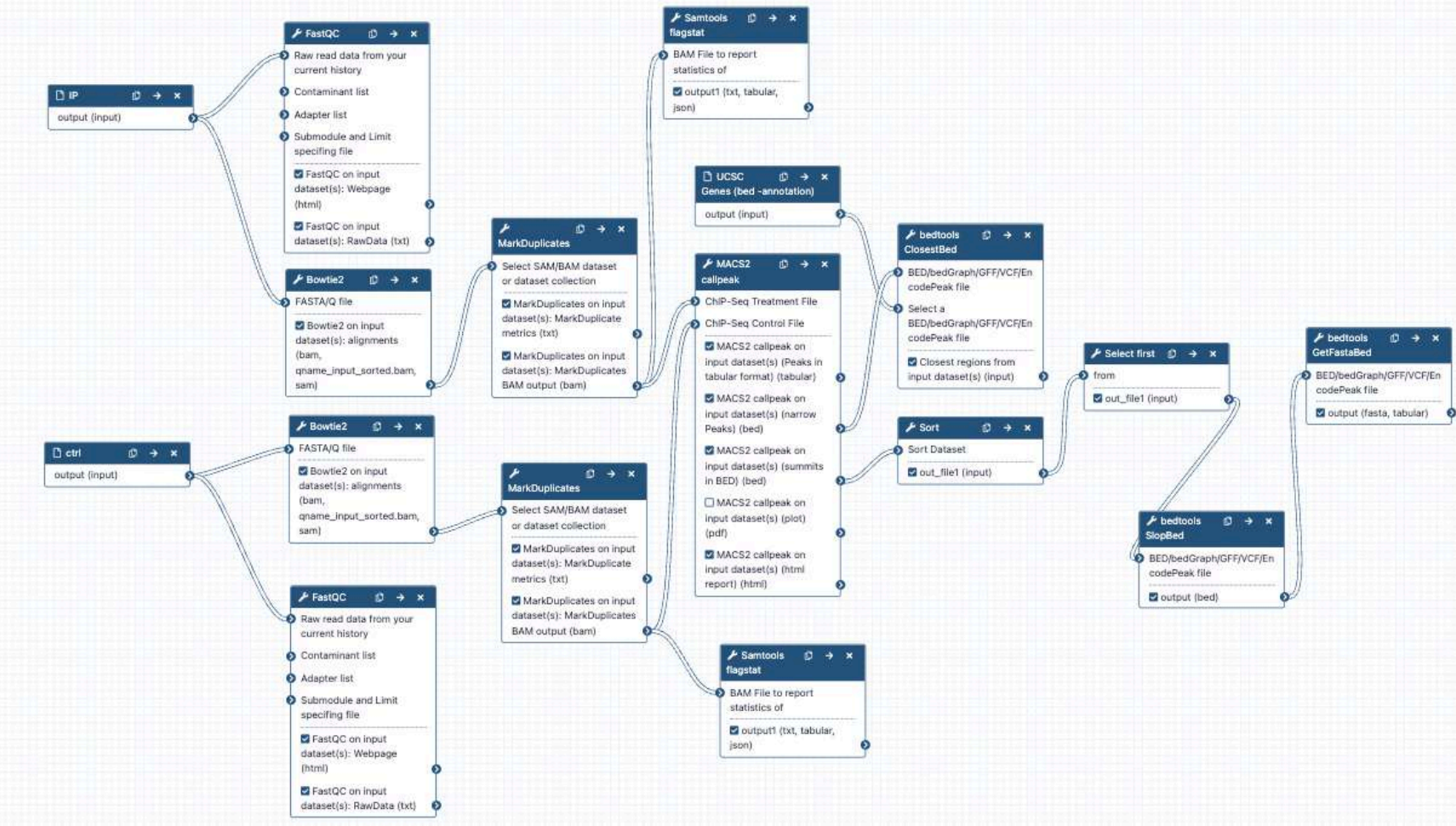
During the entire training session..



# 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

The screenshot shows the Galaxy France web interface. The top navigation bar includes 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a user profile icon. The left sidebar contains a 'Tools' section with a search bar and a list of categories: GALAXY-P, GALAXY-E, GIS Data Handling, Animal Detection on Acoustic Recordings, Biodiversity Data Exploration, STATISTICS AND VISUALISATION, MISCELLANEOUS TOOLS, and WORKFLOWS. The 'WORKFLOWS' category is highlighted, and 'All workflows' is listed below it. The main content area features a 'Welcome to usegalaxy.fr' message, a 'Term Of Use' link, and a news item dated 13/01/2022. Below this, the 'Domain specific subdomains:' section lists several workflow cards: 'Workflow4Metabolomics' (Metabolomics data processing, analysis and annotation for Metabolomics community), 'Covid19' (Variant analysis, consensus using community approved workflows and datasets), 'Metabarcoding' (FROGS, Qiime, Mothur, Obitoools, DADA2, PICRUST), 'ProteoRE' (Functional analysis), and 'Ask the GalaxyCat'.

Create, run,  
edit (...)  
workflows

Run workflows



# Workflows

The screenshot shows the Galaxy France interface. At the top, there is a navigation bar with 'Galaxy France' and various menu items. Below this, there is a 'Tools' sidebar on the left and a main content area. The main content area has a search bar for workflows and a table of existing workflows. The table has columns for Name, Mots-clés, Updated, Sharing, and Bookmarked. Two workflows are listed: 'DNA-seq data analysis (DU Dijon)' and 'Unnamed workflow', both updated 2 months ago. An orange arrow points from the '+ Create' button to the text 'Create workflows' below the table. On the right, there is a 'History' sidebar showing 'Unnamed history' (empty) and a message: 'Cet historique est vide. You can Charger vos propres données or Charger des données depuis'.

Name	Mots-clés	Updated	Sharing	Bookmarked
▼ DNA-seq data analysis (DU Dijon)		2 months ago		<input type="checkbox"/>
▼ Unnamed workflow		2 months ago		<input type="checkbox"/>

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

The screenshot displays the Galaxy France interface for creating a workflow. The top navigation bar includes 'Galaxy France', 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a 'Using 18%' indicator. The left sidebar, titled 'Tools', contains a search bar and a list of tool categories: 'Inputs', 'Get Data', 'Send Data', 'Collection Operations', 'Expression Tools', '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', and 'Fetch Alignments/Sequences'. The main workspace is a grid titled 'Test Workflow'. On the right, a metadata panel includes fields for 'Name' (Test Workflow), 'Version' (1: Feb 4th 2022, 0 steps), 'Annotation', 'License' (Specify a license for this workflow.), 'Creator' (Add a new creator - either a person or an organization.), and 'Tags' (Apply tags to make it easy to search for and find items with the same tag.). An orange arrow points from the text below to the 'Operate on Genomic Intervals' tool in the sidebar.

Add tools or input datasets to the workflow

# Workflow creation

The screenshot displays the Galaxy France interface for creating a workflow. On the left, the 'Tools' sidebar is visible, listing various tools under categories like 'Filter and Sort', 'Species abundance', and 'QIIME'. The central workspace shows a workflow diagram with two tools: 'Input Dataset' and 'Filter'. An orange arrow points from the text 'Input dataset.' to the 'Input Dataset' tool. Another orange arrow points from the text 'Tool to be run' to the 'Filter' tool. The configuration panel for the 'Filter' tool is open on the right, showing options for filtering data based on simple expressions, with a condition 'c1=='chr22'' entered.

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

The screenshot shows the Galaxy France interface for creating a workflow. On the left, a 'Tools' panel lists various filtering options. The main workspace, titled 'Test Workflow', contains two tool boxes: 'Input Dataset' and 'Filter'. A green arrow connects the output of the 'Input Dataset' tool to the input of the 'Filter' tool. The 'Filter' tool's configuration panel is open on the right, showing options for labeling, step annotation, and filtering conditions. The condition 'c1==chr22' is entered in the 'With following condition' field. The 'Number of header lines to skip' is set to 0, and 'Email notification' is set to 'No'.

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

# Workflow creation

The screenshot shows the Galaxy France interface for creating a workflow. On the left, a 'Tools' panel lists various tools under categories like 'Filter and Sort', 'Species abundance', and 'Qiime'. The central workspace, titled 'Test Workflow', shows a workflow with two steps: 'Input Dataset' and 'Filter'. The 'Filter' step is pre-configured with the parameter 'out\_file1 (input)'. On the right, a configuration panel for the 'Filter' tool is visible, showing options for 'Label', 'Step Annotation', 'With following condition' (set to 'c1==chr22'), and 'Number of header lines to skip' (set to 0). The 'Email notification' toggle is currently set to 'No'.

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



# Workflow creation

The screenshot displays the Galaxy France interface for creating a workflow. On the left, a 'Tools' panel is filtered for 'filter', showing various options under categories like 'Filter and Sort' and 'Species abundance'. The main workspace, titled 'Test Workflow', shows a workflow diagram with two steps: 'Input Dataset' (output: 'output (input)') and 'Filter' (input: 'out\_file1 (input)'). The 'Filter' tool's configuration panel is open on the right, showing options for 'Label', 'Step Annotation', and 'Filter' conditions. A red arrow points to the 'With following condition' section, specifically to the 'Set at Runtime' checkbox, which is currently unchecked. Below this, there are input fields for 'Number of header lines to skip' (set to 0) and 'Email notification' (set to No).

Click to get the parameter to be set at runtime

# Workflow creation

Run workflow

Save

The screenshot displays the Galaxy France workflow editor interface. The main workspace shows a workflow titled "Test Workflow" with three steps connected by arrows:

- Input Dataset**: The first step, with the output labeled "output (input)".
- Filter**: The second step, with the output labeled "out\_file1 (input)".
- Sort Dataset**: The third step, with the output labeled "out\_file1 (input)".

The left sidebar contains a "Tools" section with a search bar containing "sort". Below the search bar, there are two main categories of tools:

- Filter and Sort**:
  - Sort data in ascending or descending order
  - Filter data on any column using simple expressions
  - Select lines that match an expression
  - Filter GTF data by attribute values\_list
  - Filter GFF data by attribute using simple expressions
  - Filter GFF data by feature count using simple expressions
  - Extract features from GFF data
  - Sub-sample sequences files e.g. to reduce coverage
  - Filter sequences by ID from a tabular file
  - Column arrange by header name
- Text Manipulation**:
  - Sort Column Order by heading
  - Sort data in ascending or descending order
  - Sort a row according to their columns

The right sidebar shows a configuration panel for the "Sort Dataset" step. It includes a "Save As..." menu, a "Step Annotation" field, and configuration options for the sort operation:

- Sort Dataset**: Data input 'input' (tabular)
- on column**: 1
- with flavor**: Numerical sort
- everything in**: Descending order
- Column selection**: + Insert Column selection

At the top right of the interface, there are icons for "Save" and "Run workflow", which are highlighted by orange arrows from the text labels above. The top navigation bar includes "Workflow", "Visualize", "Données partagées", "Aide", "Utilisateur", and a "Using 18%" indicator.

# Run workflow

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

Galaxy France

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

Run workflow

Set parameters



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	+

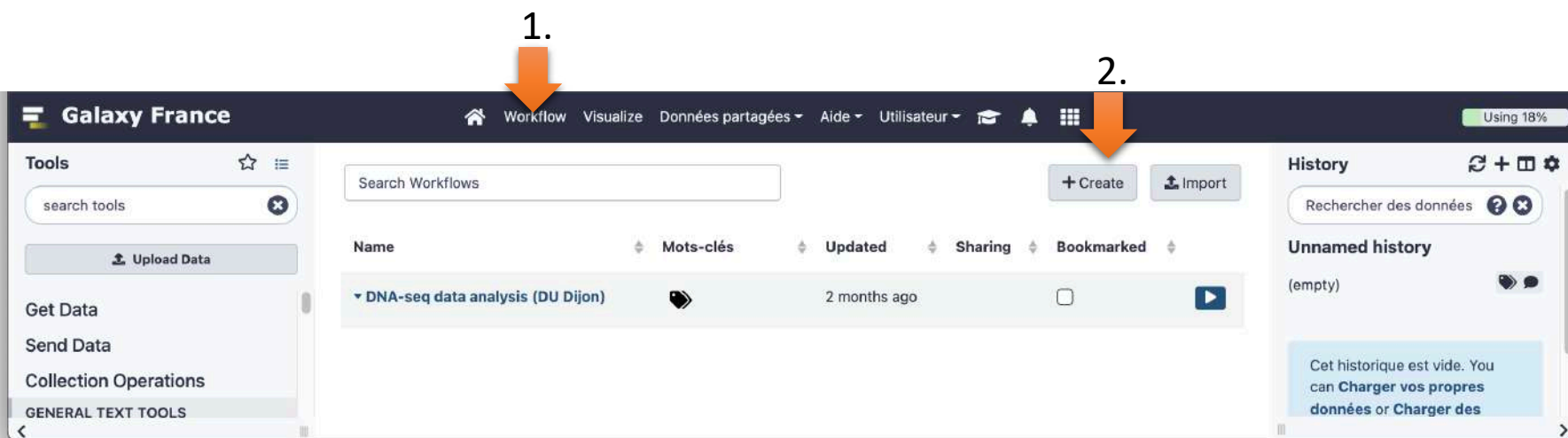
# Exercise 1: Create your first workflow

## 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

The screenshot displays the Galaxy France web interface. On the left, a 'Tools' sidebar is visible with a search bar and a list of categories. Two orange arrows point to the 'Inputs' category (labeled '1.') and the 'Input dataset' option (labeled '2.'). The main workspace, titled 'BED Ensembl to BED UCSC', contains a workflow step box labeled 'Input Dataset' with an 'output (input)' field. On the right, a configuration panel for the 'Input Dataset' step is shown, including fields for 'Label', 'Step Annotation', and 'Format(s)'. The top navigation bar includes 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a 'Using 18%' indicator.

# 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:** -

1.

2.

3.

4.

**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

Change datatype

This action will change the datatype of the output to the indicated datatype.

- 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

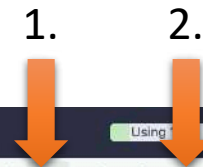
## 2.7.

- 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)



The screenshot shows the Galaxy France interface with a workflow titled "BED Ensembl to BED UCSC". The workflow consists of several steps: "Input Dataset", "Replace + by 1", "Remove header line", "Sort by coordinates", "Replace chrMT by chrM", "Add prefix chr", and "Replace -1 by -". The "Tools" panel on the left lists various bedtools tools. The "bedtools SortBED" tool is highlighted in the right-hand panel, showing its configuration options: "Label" (Sort by coordinates), "Step Annotation", "Sort by" (chromosome, then by start po...), and "Specify a genome file that defines the expected chromosome order in the input files." (No).

# 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](#).

**1.a** → Upload Data

**1.b** →

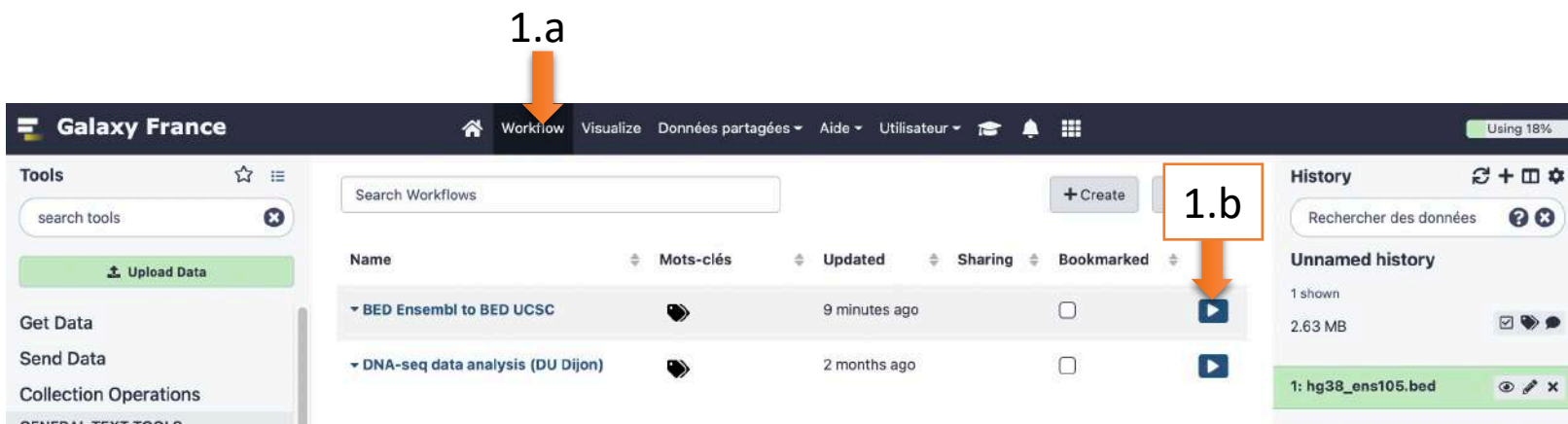
Name	Size	Type	Genome	Settings	Status
hg38_ens105.bed	2.6 MB	Auto-de...	unspecified (?)		0%

**1.c** → Start

# Exercise 1: Create your first workflow

## 4.2. Run the workflow (1)

1.a



Galaxy France

Workflow Visualize Données partagées Aide Utilisateur Using 18%

Tools

search tools

Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Search Workflows

+ Create

Name	Mots-clés	Updated	Sharing	Bookmarked	
BED Ensembl to BED UCSC		9 minutes ago		<input type="checkbox"/>	
DNA-seq data analysis (DU Dijon)		2 months ago		<input type="checkbox"/>	

History

Rechercher des données

Unnamed history

1 shown

2.63 MB

1: hg38\_ens105.bed

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur Using 18%

Tools

search tools

Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Workflow: BED Ensembl to BED UCSC

1

1: hg38\_ens105.bed

Expand to full workflow form.

1.c

Run Workflow

History

Rechercher des données

Unnamed history

1 shown

2.63 MB

1: hg38\_ens105.bed

# Exercise 1: Create your first workflow

The workflow is running

The screenshot displays the Galaxy France web interface. At the top, the navigation bar includes 'Galaxy France', 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a 'Using 18%' indicator. On the left, a 'Tools' sidebar lists various categories like 'Get Data', 'Send Data', and 'GENERAL TEXT TOOLS'. The main workspace shows a green notification: 'Successfully invoked workflow BED Ensembl to BED UCSC. 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.' Below this, a progress bar for 'Invocation 1...' indicates '7 of 7 steps successfully scheduled' and '0 of 6 jobs complete...'. The right-hand 'History' panel, titled 'Unnamed history', lists seven steps: '1: hg38\_ens105.bed', '2: Regex Replace on data 1', '3: Regex Replace on data 2', '4: Regex Replace on data 3', '5: Remove beginning on data 4', '6: Regex Replace on data 5', and '7: SortBed on Regex Replace on data 5'. The first step is highlighted in green.

# Exercise 1: Create your first workflow

Rename your history to « hg38\_ens105: BED Ensembl to BED UCSC » (1)

The screenshot displays the Galaxy France web interface. The top navigation bar includes 'Galaxy France', 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and 'Using 18%'. The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: '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', and 'GENOMICS ANALYSIS'. The main workspace shows a green notification: 'Successfully invoked workflow BED Ensembl to BED UCSC . 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.' Below this, a workflow invocation is shown with a progress bar indicating '7 of 7 steps successfully scheduled' and '0 of 6 jobs complete...'. The right sidebar features a 'History' panel with a search bar and a list of history entries. An orange arrow points to the 'Unnamed history' entry, which is labeled '1.'.

History

Rechercher des données

Unnamed history

7 shown

2.63 MB

7: SortBed on Regex Replace on data 5

6: Regex Replace on data 5

5: Remove beginning on data 4

4: Regex Replace on data 3

3: Regex Replace on data 2

2: Regex Replace on data 1

1: hg38\_ens105.bed



# Exercise 1: Create your first workflow

It's all done!

The screenshot displays the Galaxy France web interface. At the top, the navigation bar includes 'Galaxy France', 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a 'Using 18%' indicator. The left sidebar contains a 'Tools' section with a search bar and an 'Upload Data' button, followed by categories like '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', and 'GENOMICS ANALYSIS'. The main content area features a green success message: 'Successfully invoked workflow BED Ensembl to BED UCSC . 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.' Below this is a 'View Report 1' section with a progress bar indicating '7 of 7 steps successfully scheduled' and '6 of 6 jobs complete.' The right sidebar shows the 'History' pane with a search bar and a 'Refresh history' button. The workflow history is listed as follows:

Step	Tool	Actions
7	SortBed on Regex Replace on data 5	View Edit Delete
6	Regex Replace on data 5	View Edit Delete
5	Remove beginning on data 4	View Edit Delete
4	Regex Replace on data 3	View Edit Delete
3	Regex Replace on data 2	View Edit Delete
2	Regex Replace on data 1	View Edit Delete
1	hg38_ens105.bed	View Edit Delete

# 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.

The screenshot shows the Galaxy France web interface. On the right side, the 'History' panel is open, displaying a list of actions. The 'Extract Workflow' option is highlighted in blue. An orange arrow labeled '1.a.' points to the 'History' panel header, and another orange arrow labeled '1.b.' points to the 'Extract Workflow' option. The main content area shows a 'Welcome to usegalaxy.fr' message and a list of domain-specific subdomains including 'Workflow4Metabolomics', 'Covid19', 'Metabarcoding', 'ProteoRE', and 'Ask the GalaxyCat'.

# 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)

The screenshot displays the Galaxy France interface. The 'Tools' panel on the left lists various genomic tools. Two tools are highlighted with orange boxes and arrows:

- 1.** MACS2 callpeak (with checkbox 'Include "MACS2 callpeak" in workflow')
- 2.** bedtools Intersect intervals (with checkbox 'Include "bedtools Intersect intervals" in workflow')

The central workflow panel shows a sequence of 24 steps, including '9 MACS2 callpeak on data 8 and data 6 (Peaks in tabular format)', '10 MACS2 callpeak on data 8 and data 6 (narrow Peaks)', '11 MACS2 callpeak on data 8 and data 6 (summits in BED)', '12 MACS2 callpeak on data 8 and data 6 (plot)', '13 MACS2 callpeak on data 8 and data 6 (html report)', '14 MACS2 callpeak on data 8 and data 6 (Peaks in tabular format)', '15 MITF\_peaks.narrowPeak', '16 MITF\_peak\_summits.bed', '17 MACS2 callpeak on data 8 and data 6 (html report)', '18 bedtools Intersect intervals on data 15 and data 10', '19 hg38\_ens105\_ucsc.bed', and '20 mitf\_peaks.annot.tsv'. The 'History' panel on the right shows the 'ChIP-seq data analysis' workflow with 24 steps, 3.55 GB of data, and a 'Using 30%' indicator.

# Exercise 2: Extract a workflow out of an history

1.1.

- Click on Create Workflow (1)

**Workflow name**

1.

1.2.

- Click on edit to edit the workflow (2)

Workflow "ChIP-seq data analysis workflow" created from current history. You can [edit](#) or [run](#) the workflow.



2.

# Exercise 2: Extract a workflow out of an history

1.2.

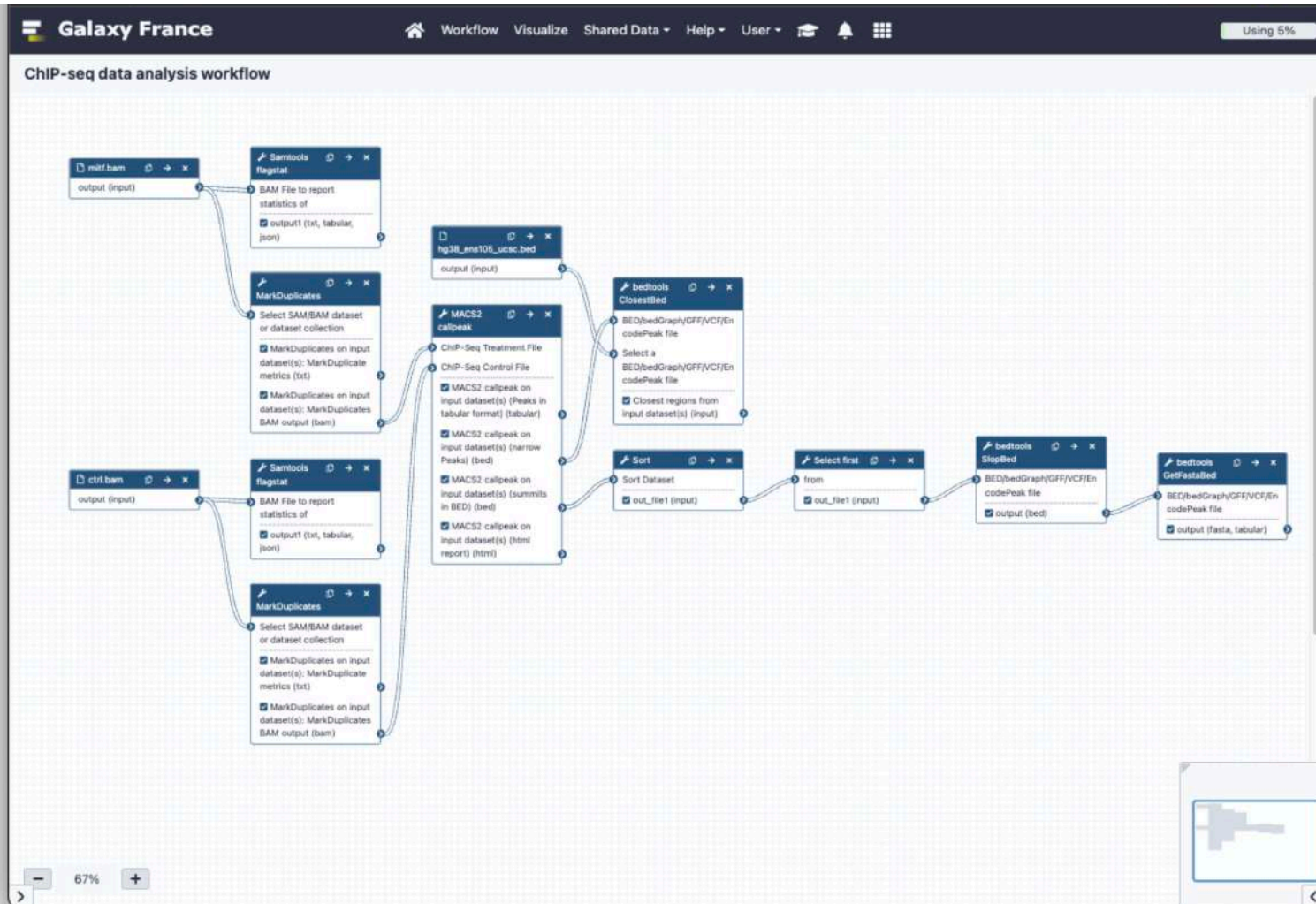
- Re-order the boxes (1).

1.a.

The screenshot displays the Galaxy France web interface. The main workspace shows a workflow titled "ChIP-seq data analysis workflow" with several tool boxes connected by arrows. The tools include "flagstat", "callpeak", "bedtools", "Sort", "MarkDuplicates", "Select first", and "GetFastaBed". A settings menu is open on the right side, showing options like "Save As...", "Auto Layout", "Best Practices", "Upgrade Workflow", and "Download". An orange arrow labeled "1.a." points to the settings menu, and another orange arrow labeled "1.b." points to the "callpeak" tool box.



# 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

The screenshot displays the Galaxy workflow editor interface. The main workspace shows a workflow titled "ChIP-seq data analysis workflow" with several steps connected by arrows. The steps include: "samtools flagstat", "MarkDuplicates", "MACS2 callpeak", "Sort Dataset", and "samtools coverage". The workflow starts with an input "out1.bam" and "out2.bam" leading to a "samtools flagstat" step. This is followed by a "MarkDuplicates" step, then a "MACS2 callpeak" step, a "Sort Dataset" step, and finally a "samtools coverage" step. The output of the "samtools coverage" step is "out3.bed".

Blue arrows point from the text "Remove the links between the inputs files and the steps after" to the connections between the input files and the first "samtools flagstat" step, and between the "MarkDuplicates" and "MACS2 callpeak" steps. The "samtools flagstat" step has a red 'x' icon next to it, indicating it is selected for removal.

The right sidebar shows the configuration for the "samtools flagstat" step, including options for "Label", "Step Annotation", "Output format", "Email notification", and "Output cleanup".



# Exercise 2: Extract a workflow out of an history

## 2.2

- Add 1 fastqc step per input file

The screenshot displays the Galaxy France interface for a workflow titled "CHIP-seq data analysis workflow". The workflow consists of several steps: two "FastQC" steps, two "Samtools flagstat" steps, two "MarkDuplicates" steps, and one "MACS2 callpeak" step. The "FastQC" steps are connected to input files "mif1.bam" and "chr1.bam". The "Samtools flagstat" steps are connected to the "FastQC" steps. The "MarkDuplicates" steps are connected to the "Samtools flagstat" steps. The "MACS2 callpeak" step is connected to the "MarkDuplicates" steps. The interface includes a "Tools" panel on the left with a search bar containing "fastqc". The right panel shows the configuration for the "FastQC Read Quality reports" tool, including fields for "Label", "Step Annotation", and "Raw read data from your current history".

# Exercise 2: Extract a workflow out of an history

## 2.3.

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

The screenshot displays the Galaxy France interface for a ChIP-seq data analysis workflow. The workflow is titled "ChIP-seq data analysis workflow" and is composed of several interconnected steps:

- FastQC**: Raw read data from your current history. Contaminant list, Adapter list, Submodule and Limit specifying file. FastQC on input dataset(s): Webpage (html), FastQC on input dataset(s): RawData (txt).
- Bowtie2**: FASTAQ file. Bowtie2 on input dataset(s): alignments (bam, genome\_input\_sorted.bam, sam).
- MarkDuplicates**: Select SAM/BAM dataset or dataset collection. MarkDuplicates on input dataset(s): MarkDuplicate metrics (txt), MarkDuplicates on input dataset(s): MarkDuplicates BAM output (bam).
- MACS2 callpeak**: ChIP-Seq Control File. MACS2 callpeak on input dataset(s): Peaks in tabular format (tabular), MACS2 callpeak on input dataset(s) (narrow Peaks) (bed), MACS2 callpeak on input dataset(s) (summits in BED) (bed), MACS2 callpeak on input dataset(s) (html report) (html).
- Samtools flagstat**: BAM File to report statistics of. output (txt, tabular, json).

The "Tools" panel on the left lists various tool categories, including "GENERAL TEXT TOOLS", "GENOMIC FILE MANIPULATION", "FASTA/FASTQ", "FASTQ Quality Control", "SAM/BAM", "BED", "VCF/BCF", "Nanopore", "COMMON GENOMICS TOOLS", "Operate on Genomic Intervals", "Fetch Alignments/Sequences", "GENOMICS ANALYSIS", "Annotation", and "Assembly".

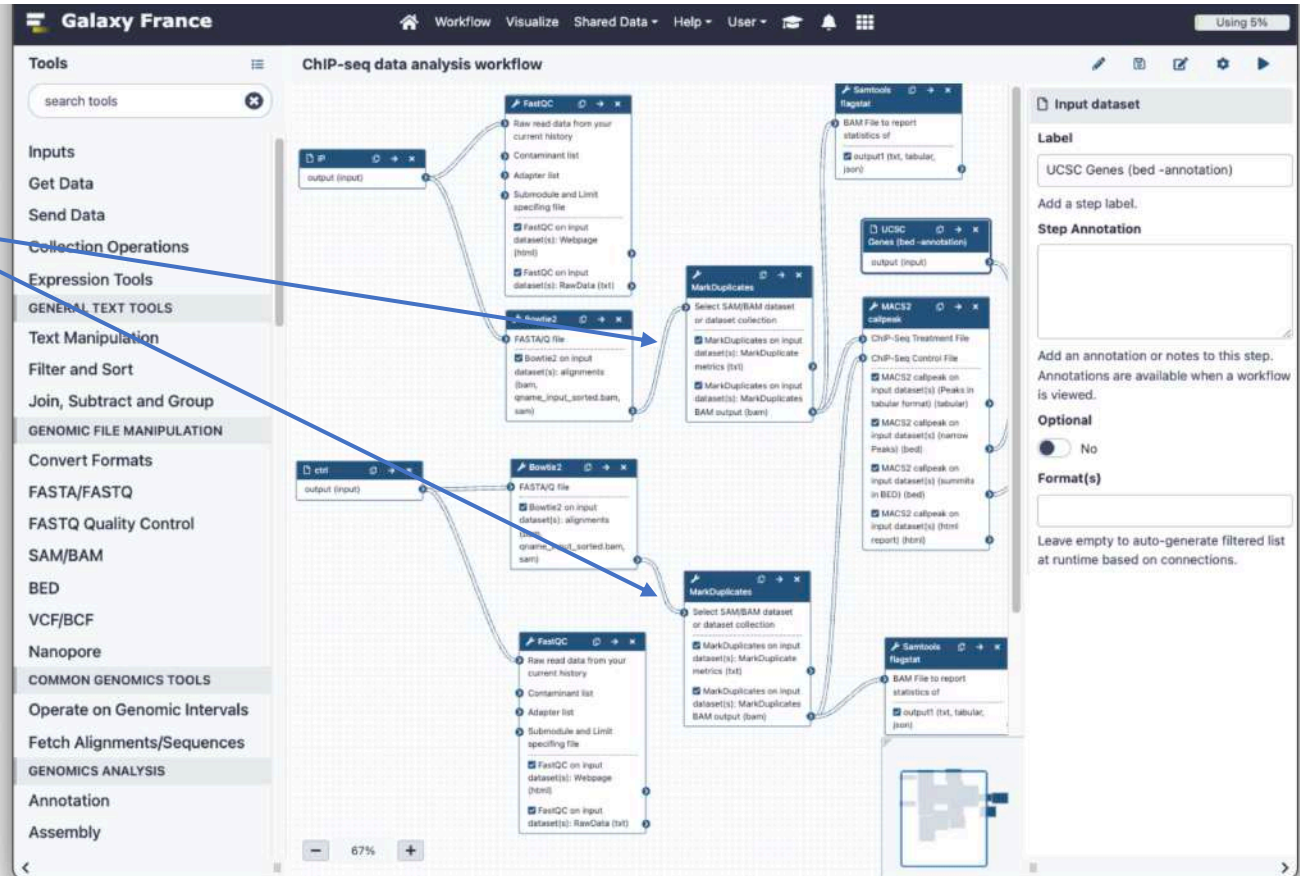
The "Input dataset" panel on the right shows the "UCSC Genes (bed - annotation)" dataset. The "Step Annotation" section is empty, and the "Optional" section is set to "No".

# Exercise 2: Extract a workflow out of an history

2.3.

Edit the links:

- Bowtie2 output is linked to MarkDuplicates

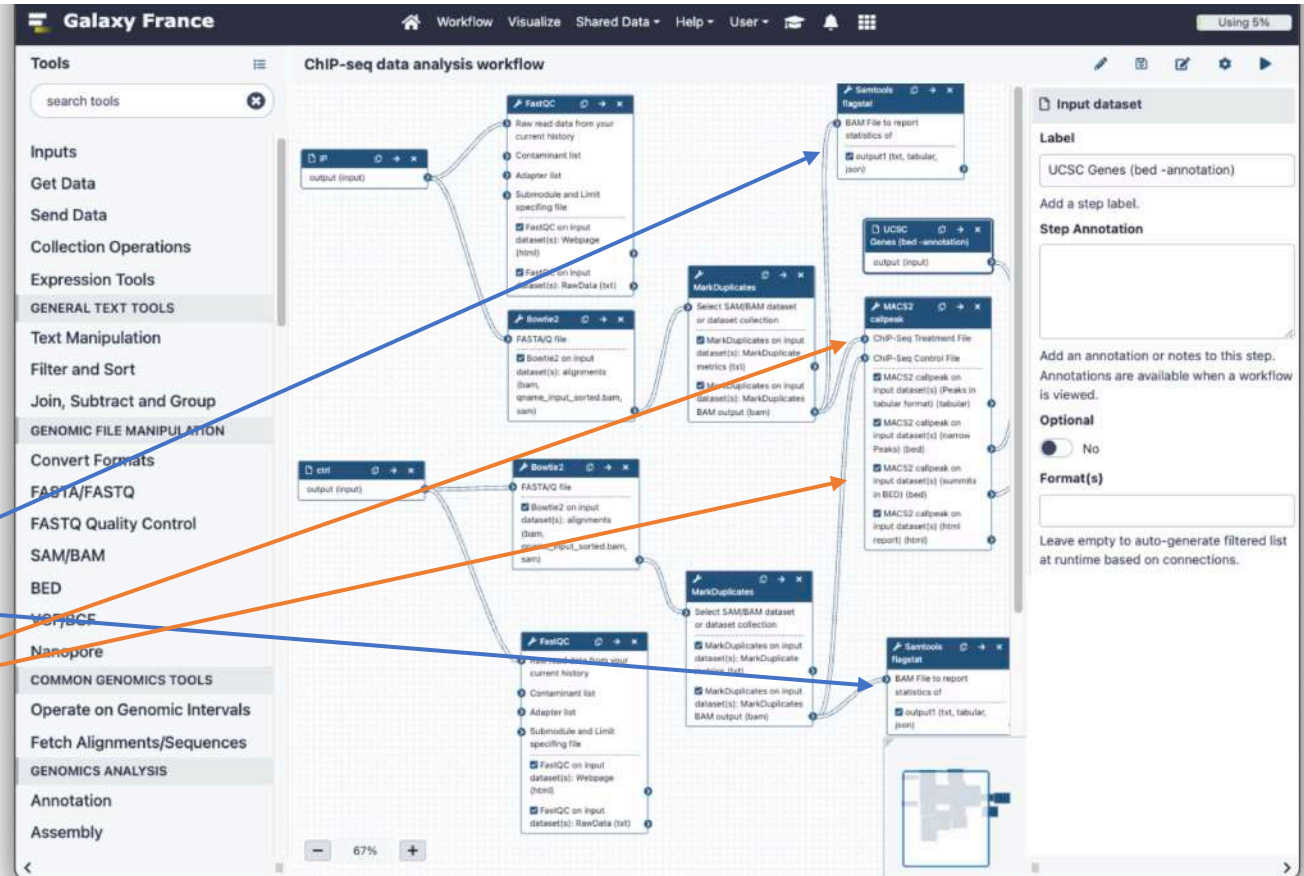


# Exercise 2: Extract a workflow out of an history

## 2.3.

Edit the links:

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

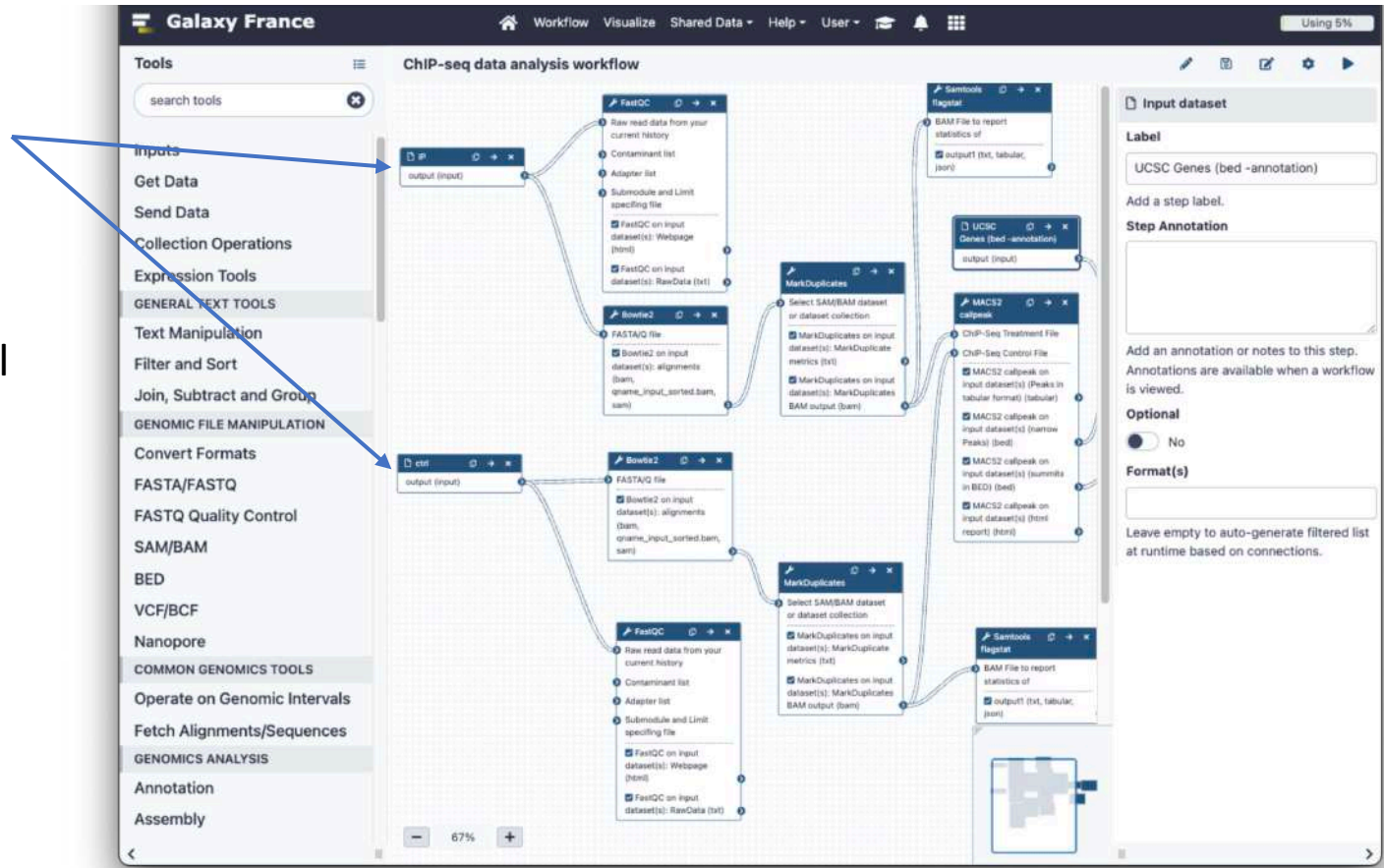




# Exercise 2: Extract a workflow out of an history

## 2.3.

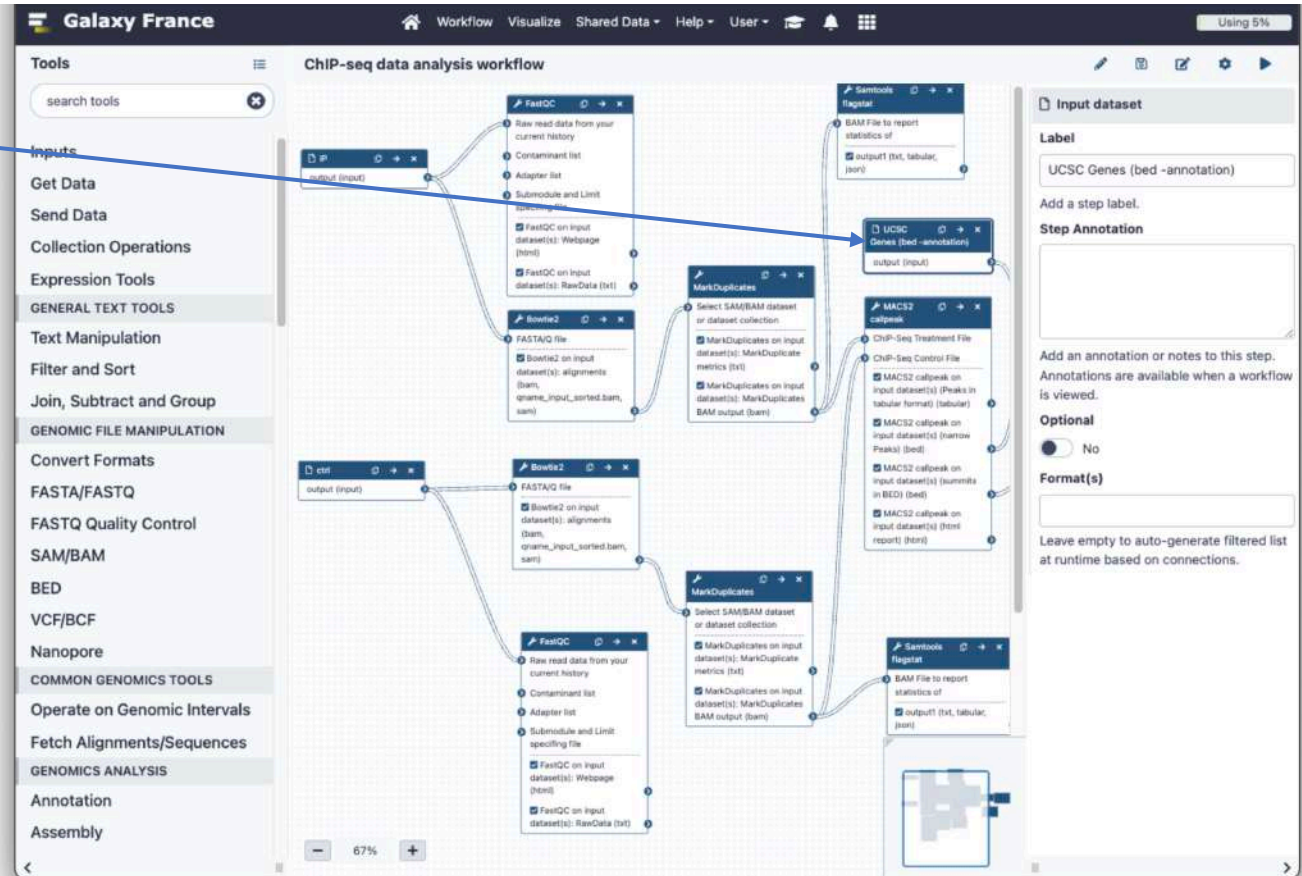
- Change the name of the two inputs:
  - Mitf.bam -> IP
  - Ctrl.bam -> Ctrl



# Exercise 2: Extract a workflow out of an history

## 2.3.

- Change the name of the file used to annotate peaks:
  - hg38\_ens105\_ucs.c.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

The screenshot displays the Galaxy France interface for a ChIP-seq data analysis workflow. The workflow consists of several steps: FastQC (Raw read data from your current history), Bowtie2 (FASTAQ file), MarkDuplicates (Select SAM/BAM dataset or dataset collection), and MACS2 (MACS2 callpeak on input dataset(s)). A blue arrow points from the text 'Will you select a reference genome from your history or use a built-in index?: Use a built-in genome index' to the 'Select reference genome' dropdown menu in the Bowtie2 step configuration. Another blue arrow points from the text 'Select a reference genome' to the same dropdown menu. The interface also shows a 'Tools' sidebar on the left and a 'Step Annotation' panel on the right.

 Select reference genome

 To be set at run time

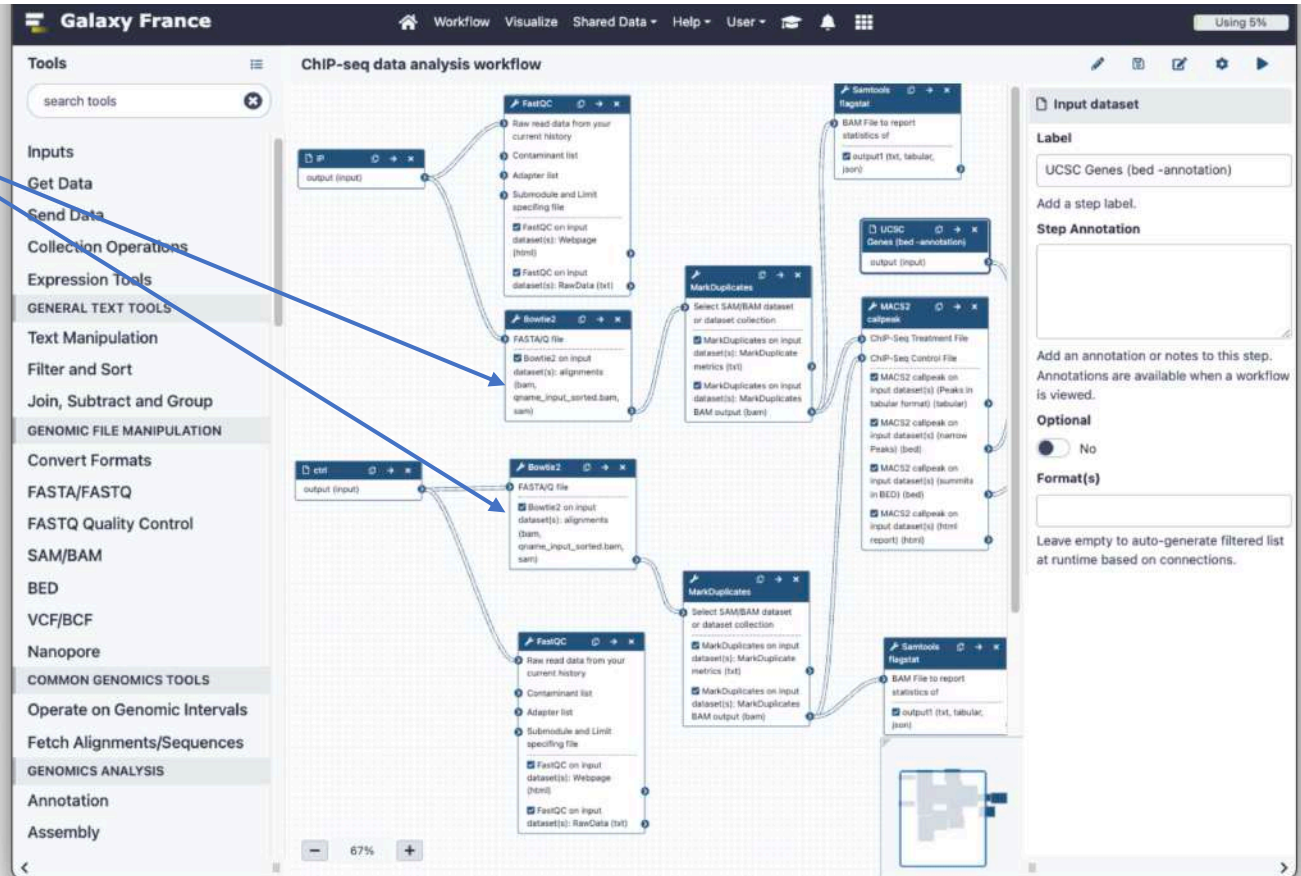
# Exercise 2: Extract a workflow out of an history

2.4.

- Edit MACS2:

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

☑ ↔ Effective genome size  
↑ 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

↔ Genome file



To be set at run time

The screenshot displays the Galaxy France interface for a 'CHIP-seq data analysis workflow'. The workflow consists of several tools connected in a sequence: 'BAM File to report statistics of', 'UCSC Genes (bed-annotation)', 'MACS2 callpeak', 'bedtools ClosestBed', 'Sort', 'Select first', 'bedtools GetFastBed', and 'bedtools SlopBed'. The 'bedtools SlopBed' tool is highlighted with a blue arrow pointing from the 'Genome file' checkbox in the left sidebar. The sidebar lists various tool categories, including 'GENOME FILE MANIPULATION' and 'COMMON GENOMICS TOOLS'. The right sidebar shows the configuration for the 'bedtools SlopBed' tool, with the 'Genome file' dropdown set to 'Locally installed Genome file' and the 'Define -l and -r as a fraction of the feature's length' option selected.

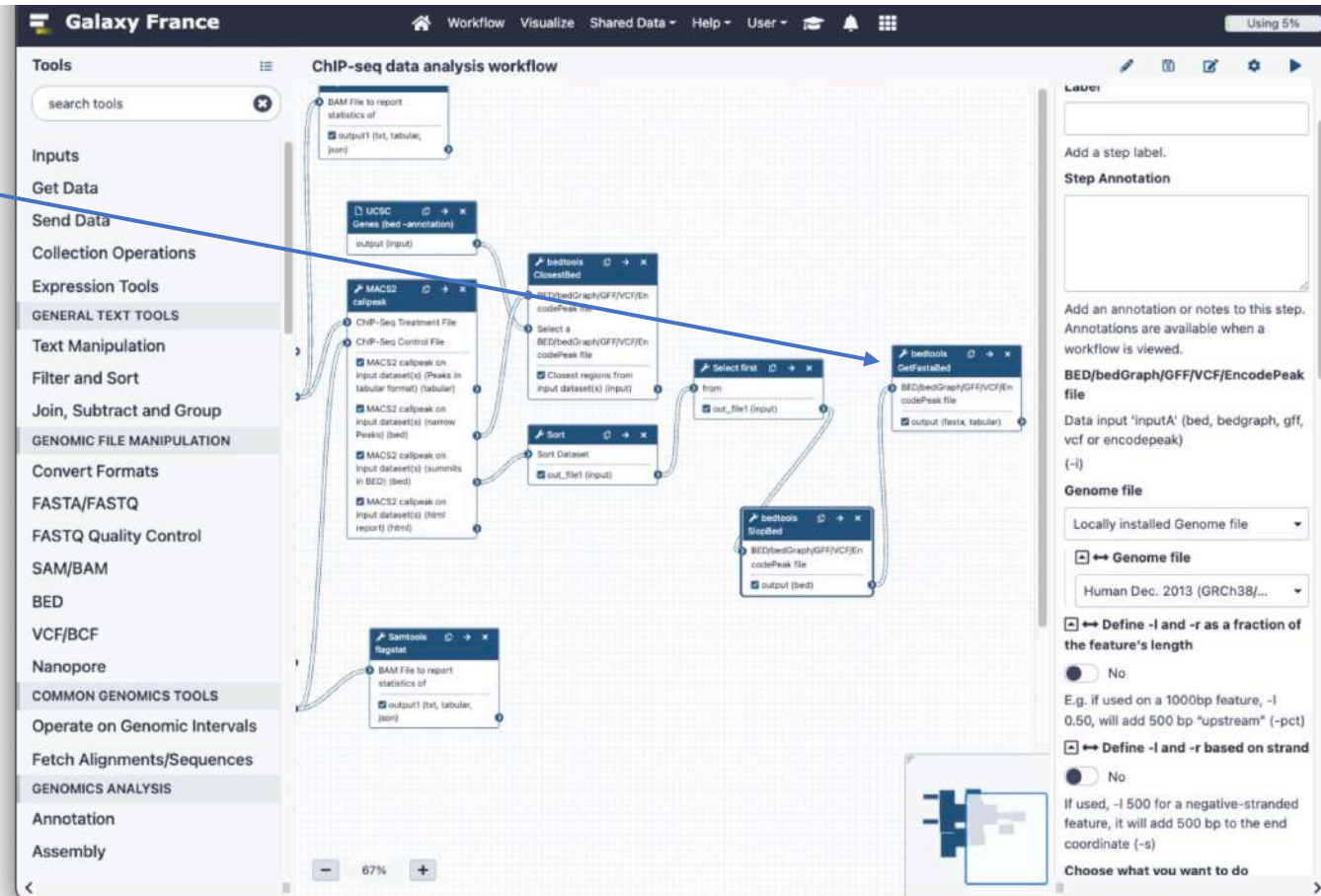
# Exercise 2: Extract a workflow out of an history

2.4.

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



To be set at run time



**Tools**

search tools

**Inputs**

Get Data

Send Data

Collection Operations

Expression Tools

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

**CHIP-seq data analysis workflow**

BAM File to report statistics of

UCSC Genes (bed -annotation)

MACS2 callpeak

bedtools ClosestBed

Sort

Select first

bedtools BedBed

bedtools GetFastaBed

**bedtools GetFastaBed**

BED/bedGraph/GFF/VCF/EncodePeak file

Data input 'inputA' (bed, bedgraph, gff, vcf or encodepeak)

(-l)

**Genome file**

Locally installed Genome file

Genome file

Human Dec. 2013 (GRCh38/...

Define -l and -r as a fraction of the feature's length

No

E.g. if used on a 1000bp feature, -l 0.50, will add 500 bp "upstream" (-pct)

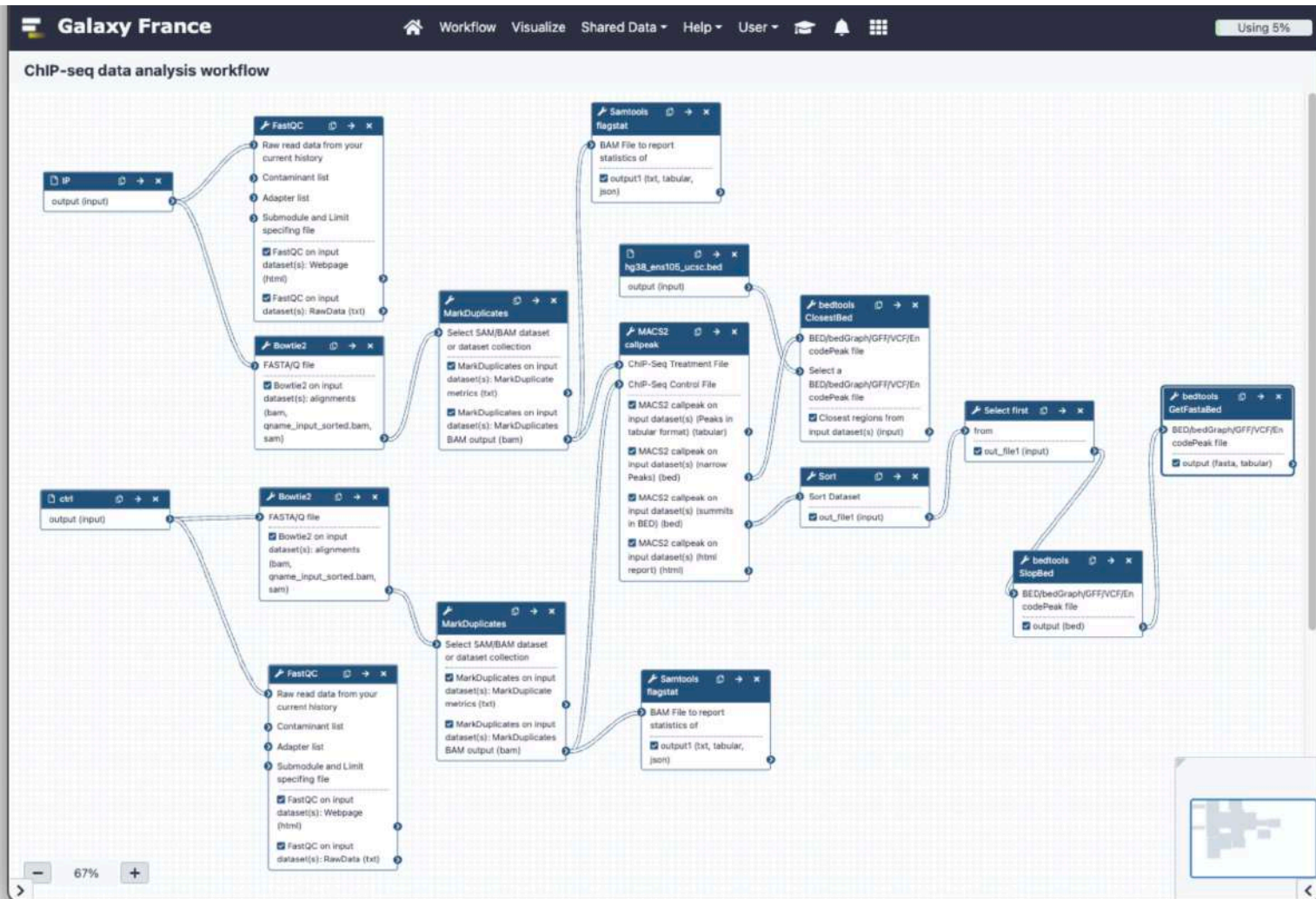
Define -l and -r based on strand

No

If used, -l 500 for a negative-stranded feature, it will add 500 bp to the end coordinate (-s)

Choose what you want to do

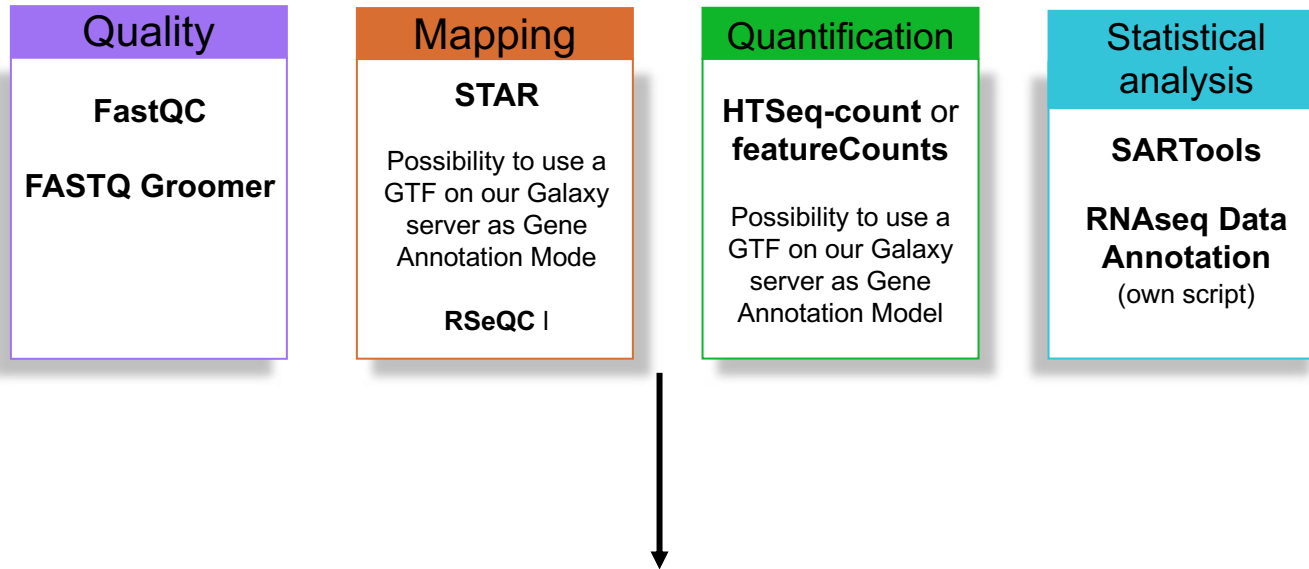
# Exercise 2: Final Workflow



## Exercise 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 ?



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

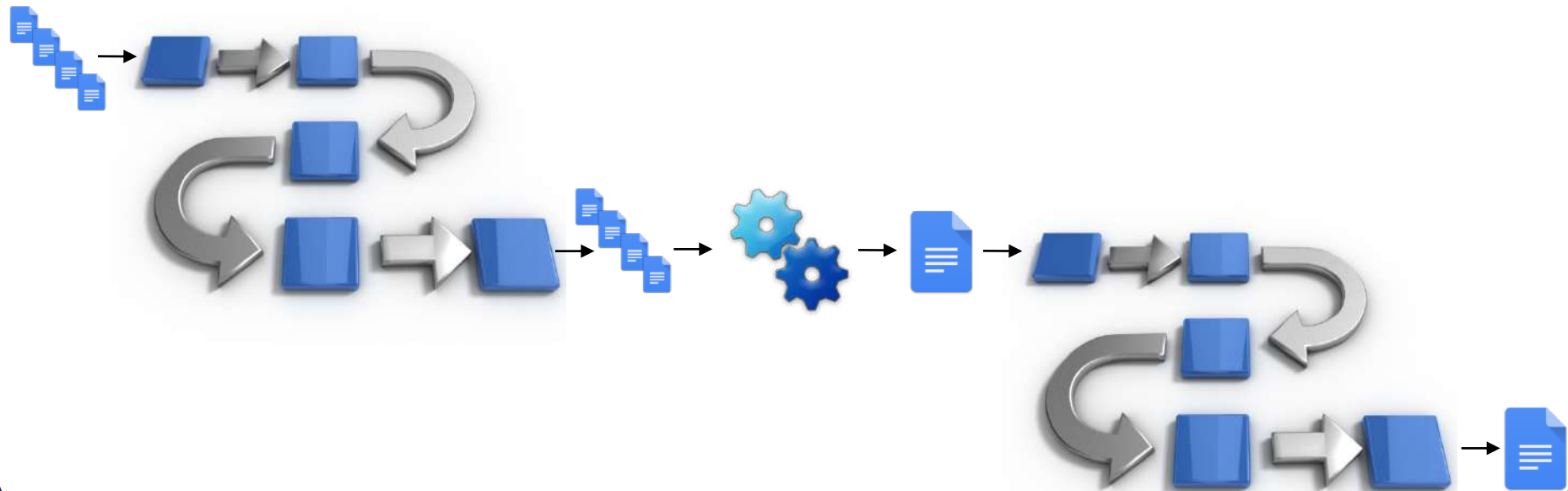
# RNAseq workflow : limits

Main workflow

Sub workflow 1

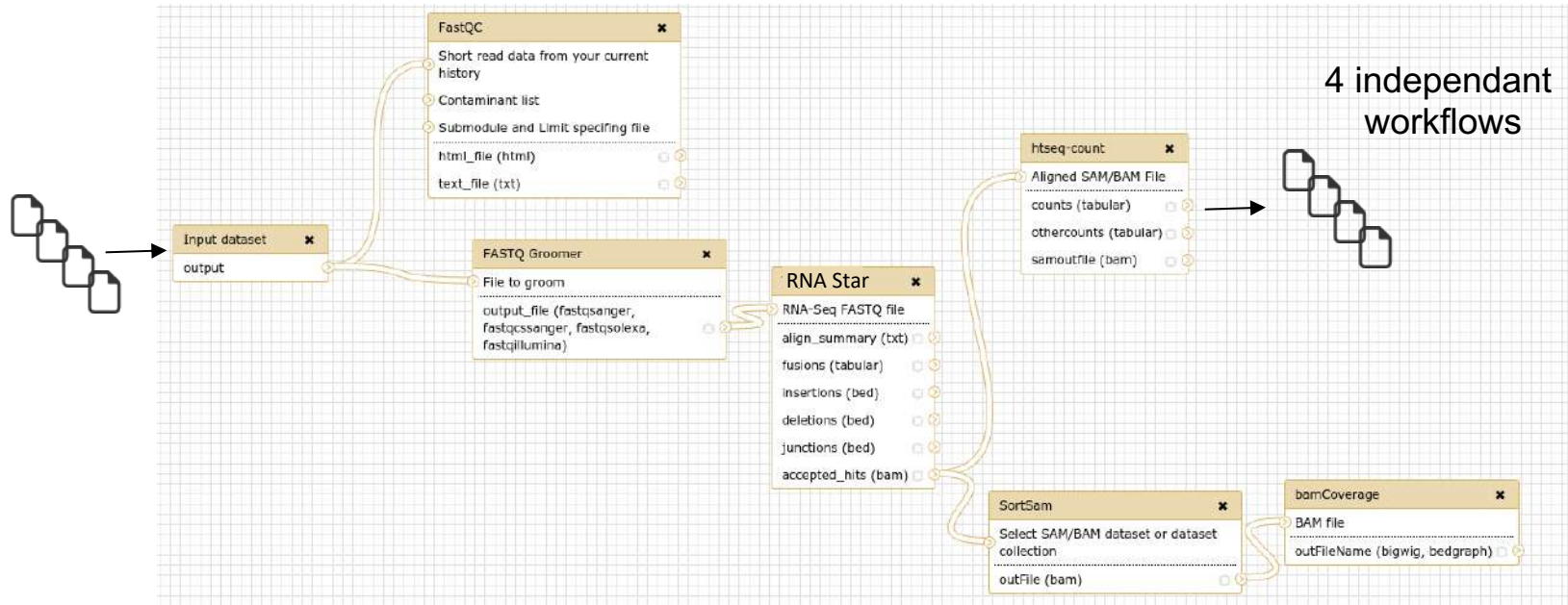
Files merging

Sub workflow 2





# RNAseq workflow : limits



HTSeq-count outputs



Merge



SARTools