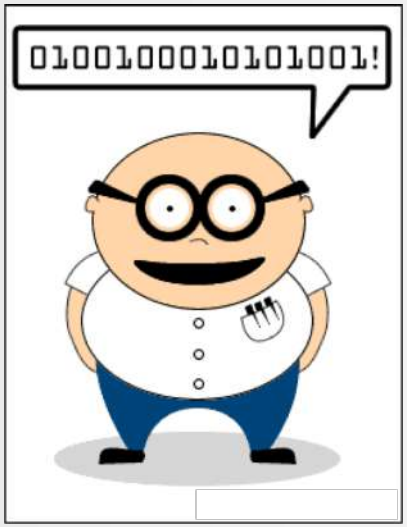
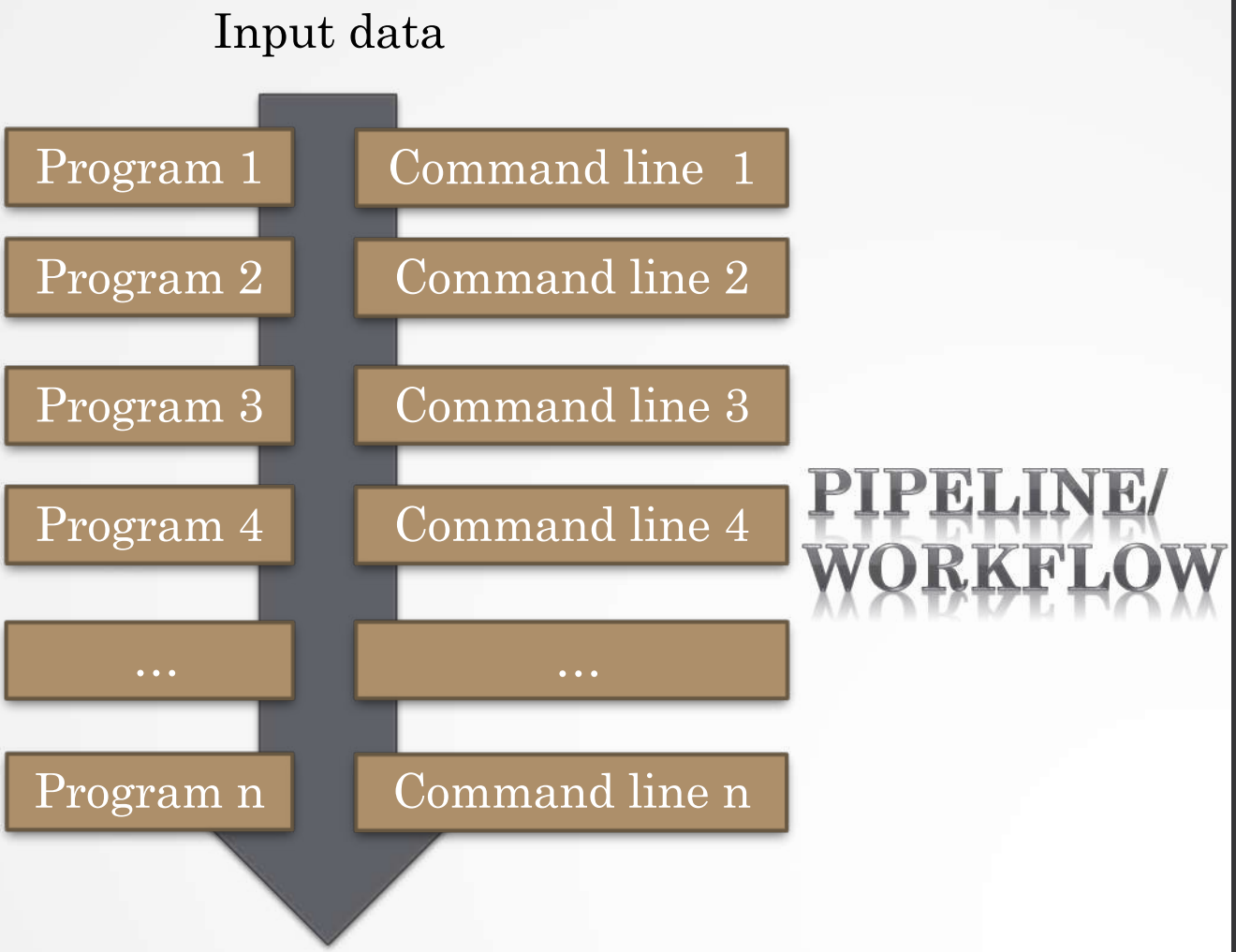
NGS analysis automatization: Galaxy workflows

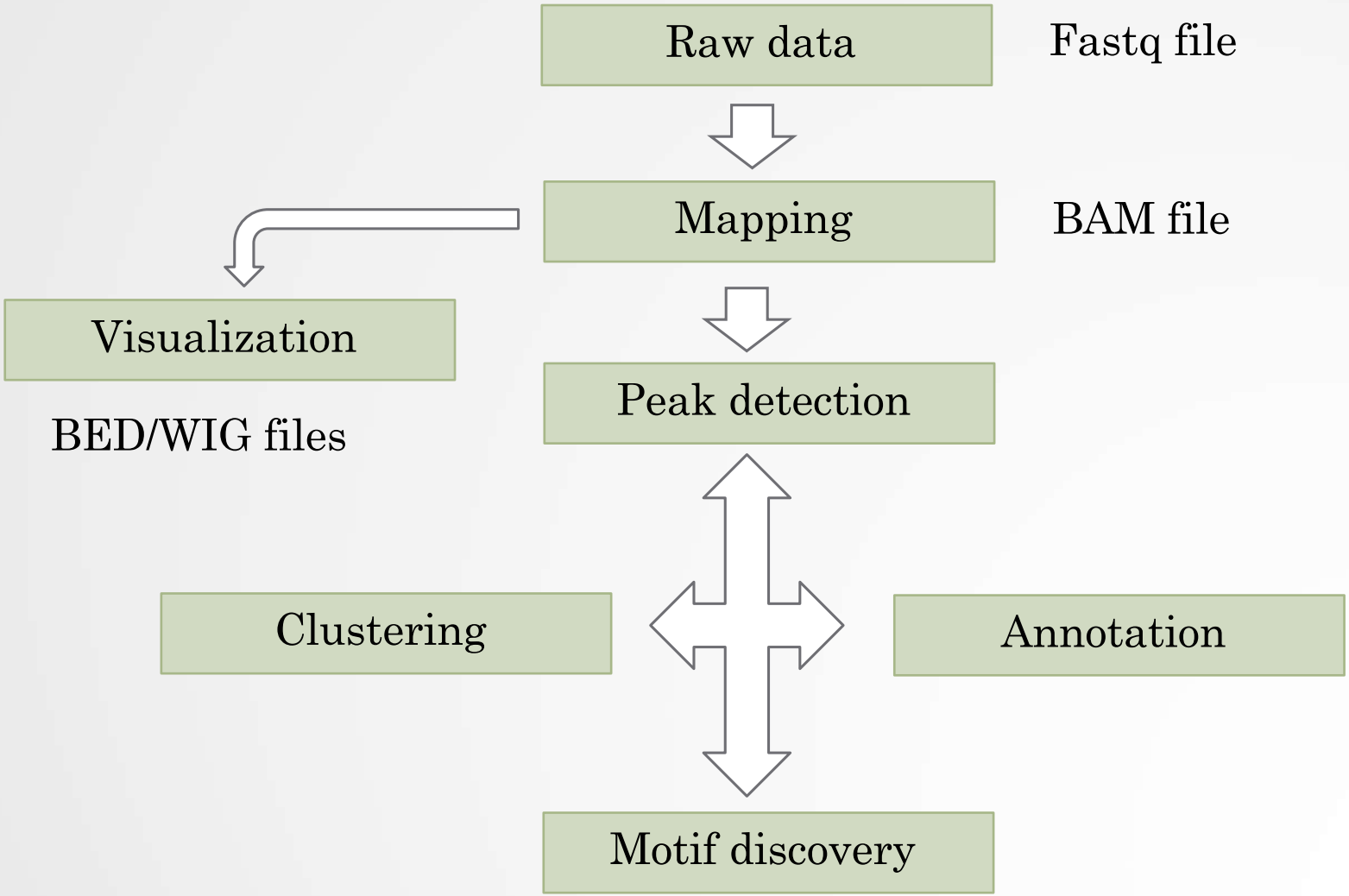
Stéphanie Le Gras
(slegras@igbmc.fr)

A long time ago…





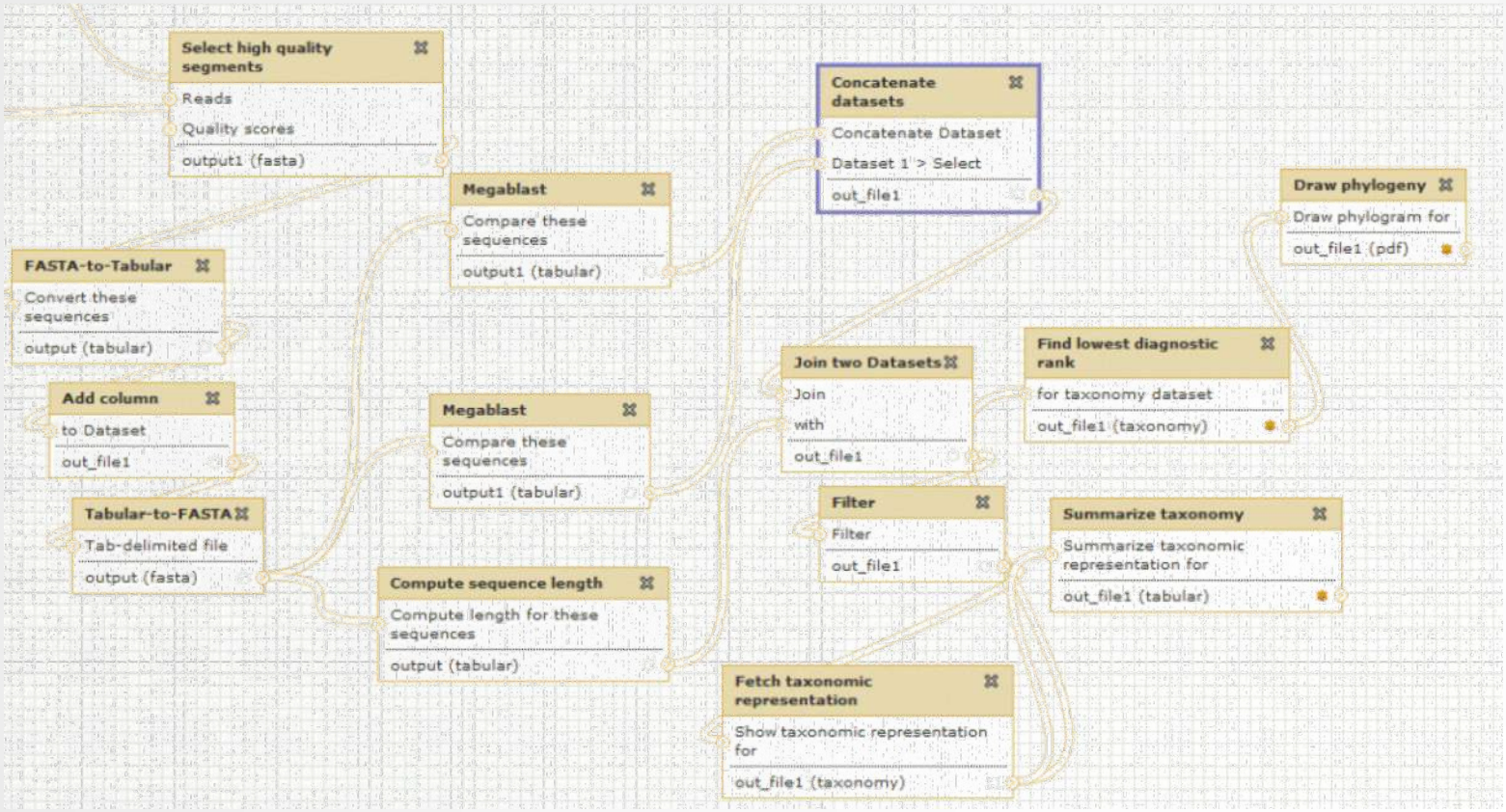
More recently…



During the entire training session..

What if we'd mix all together

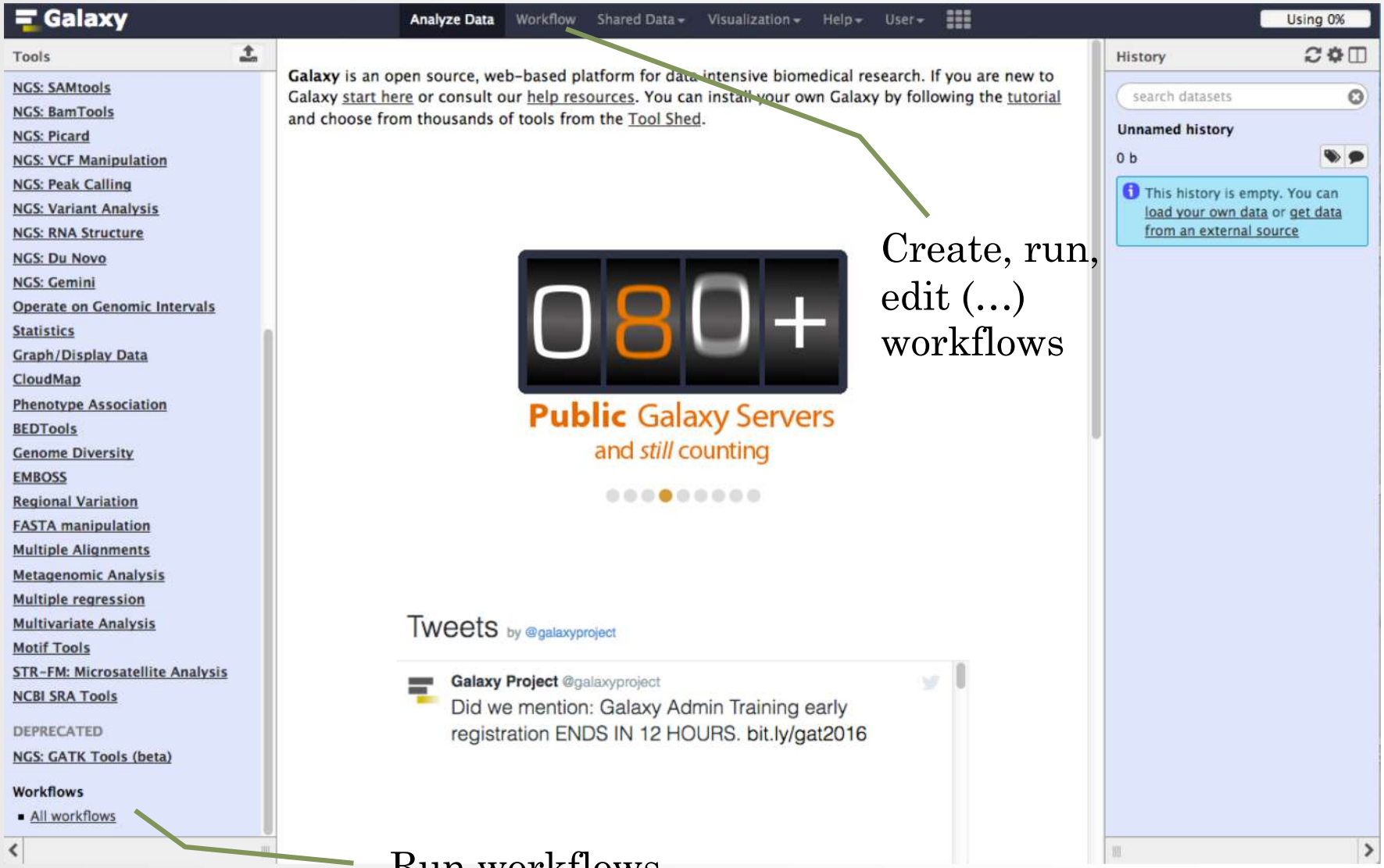
Galaxy workflow



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

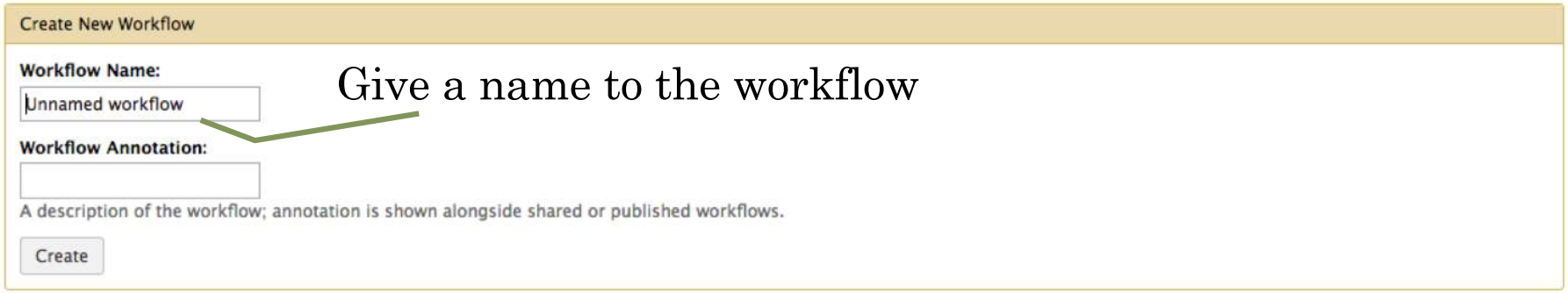


Create, run, edit (…) workflows

Run workflows

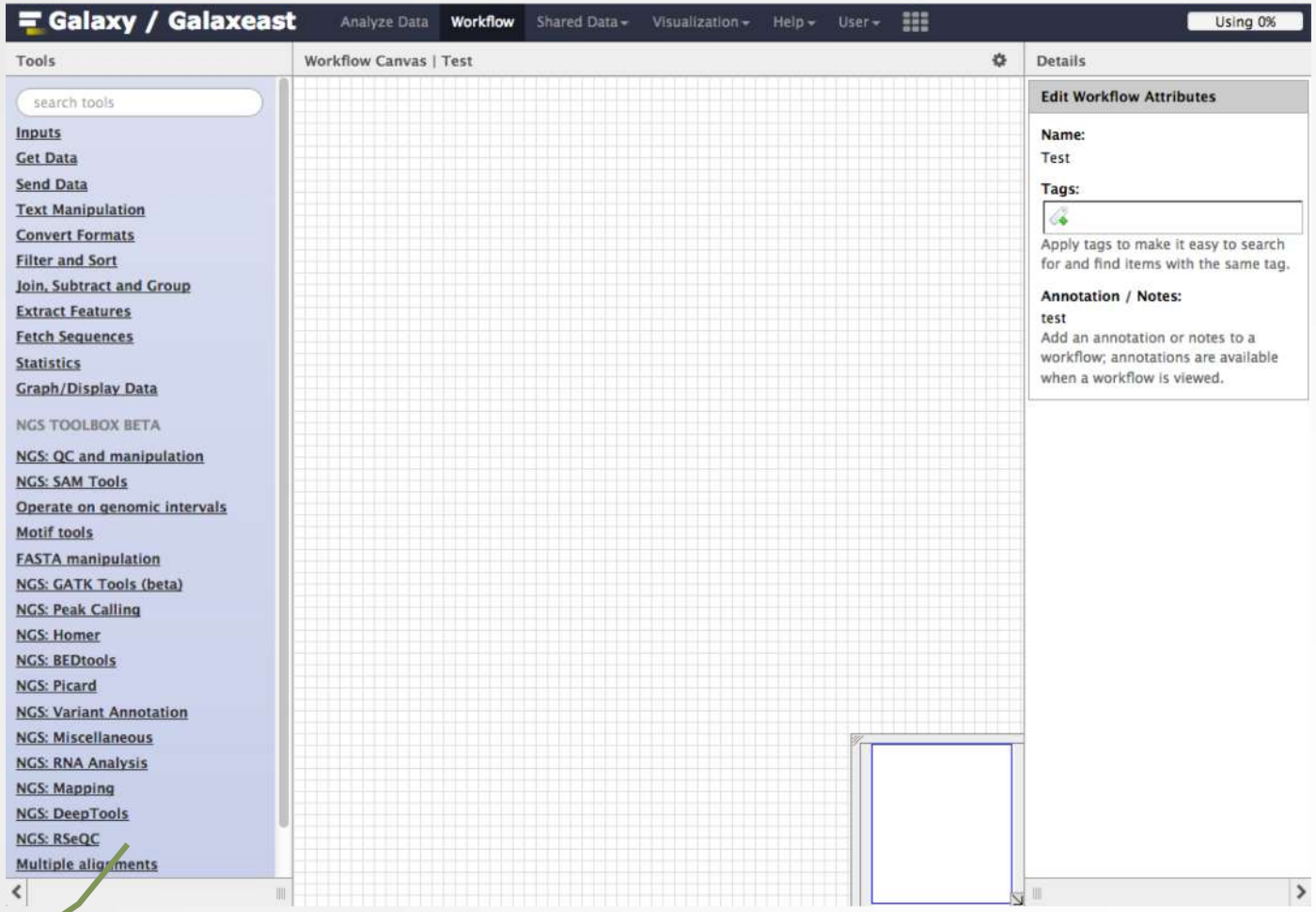
Workflows

Create workflows



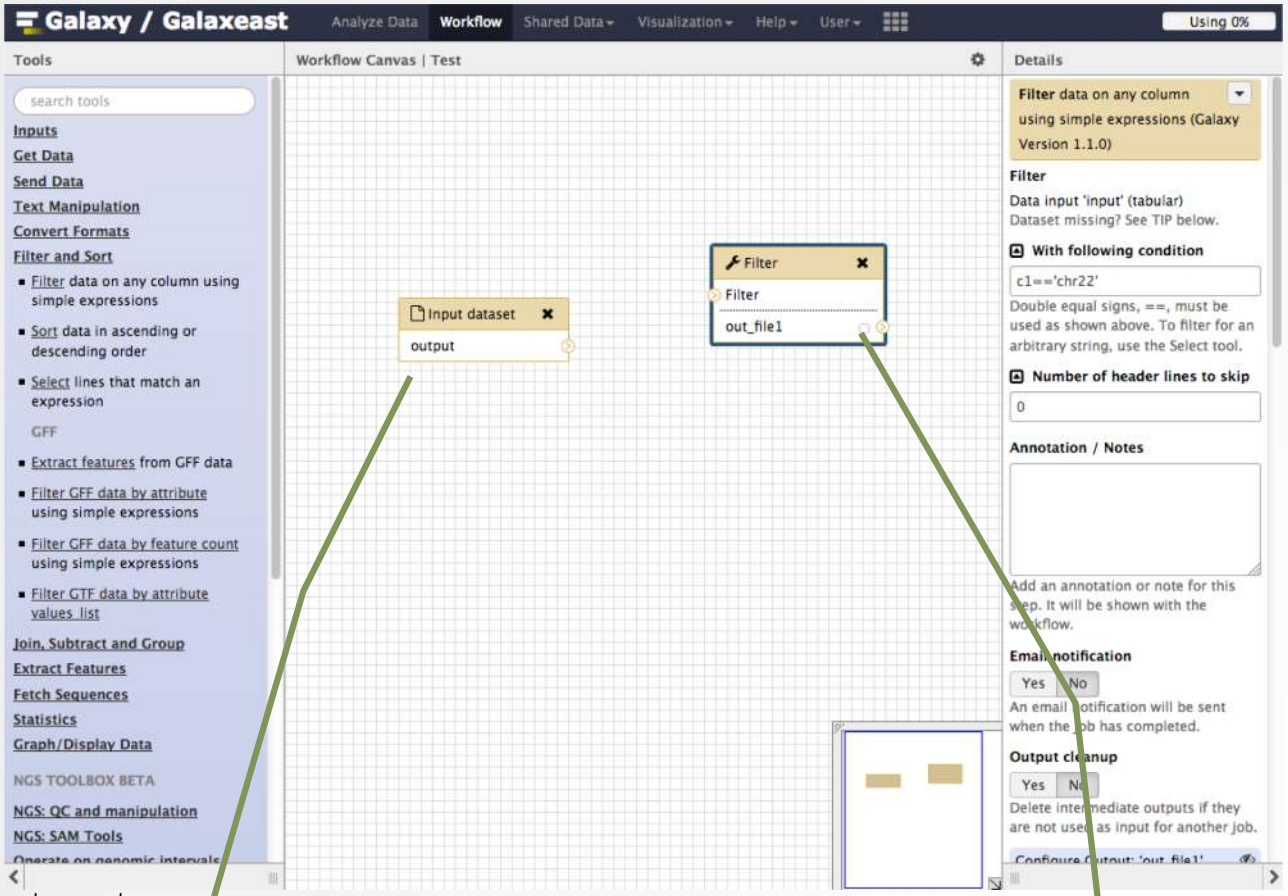
Give a name to the workflow

Workflow creation



Add tools or input datasets to the workflow

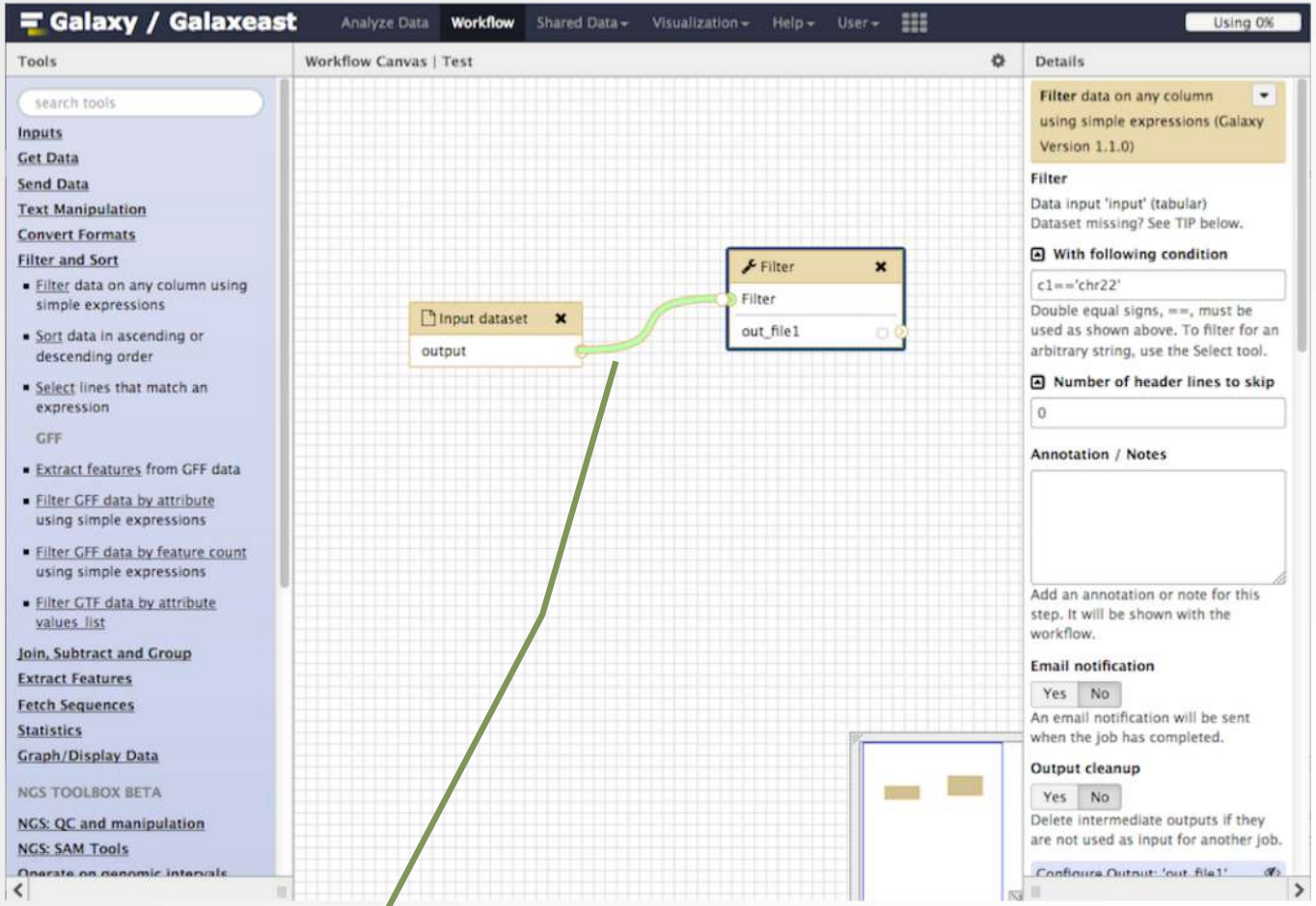
Workflow creation



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

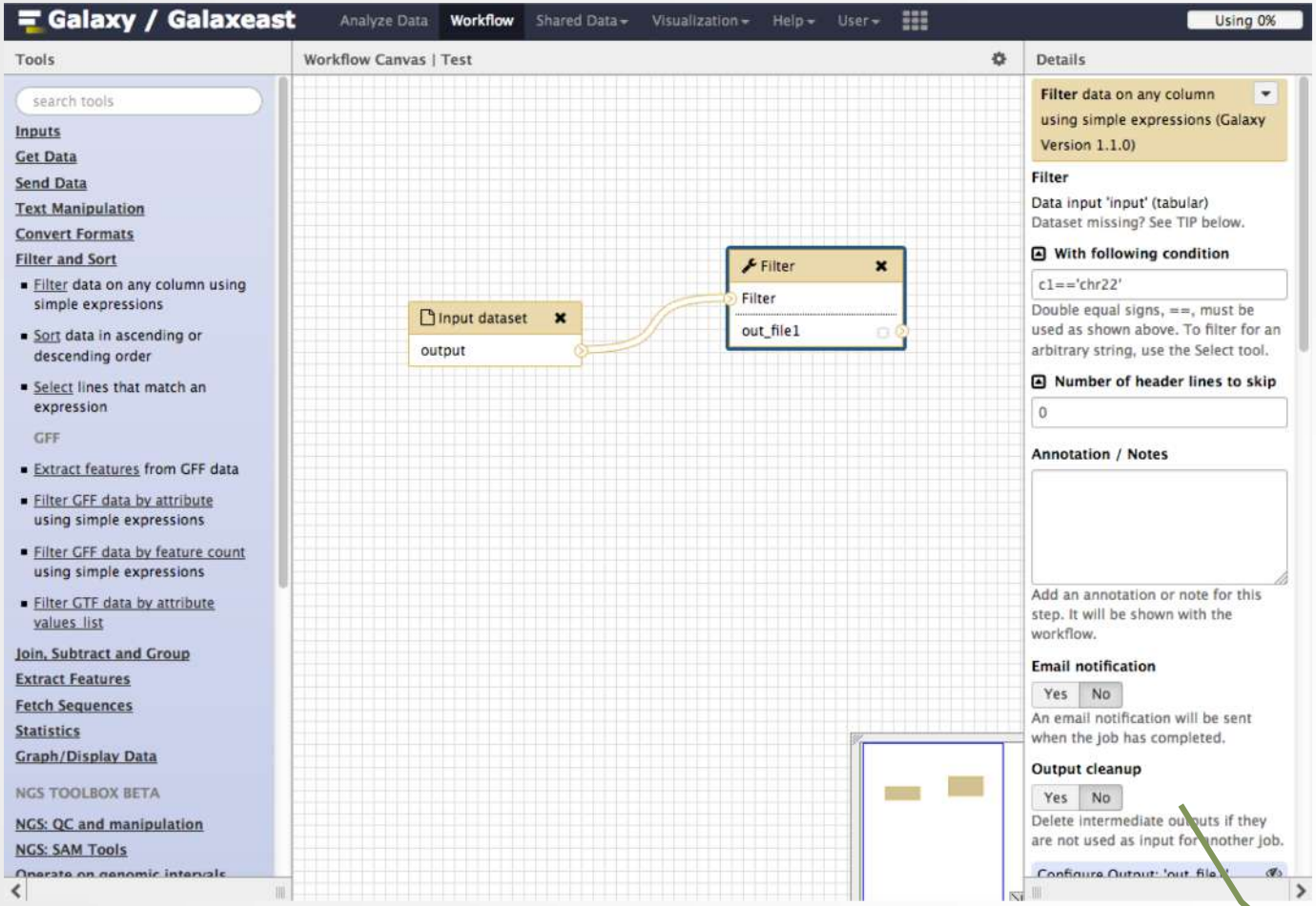
Tool to be run

Workflow creation



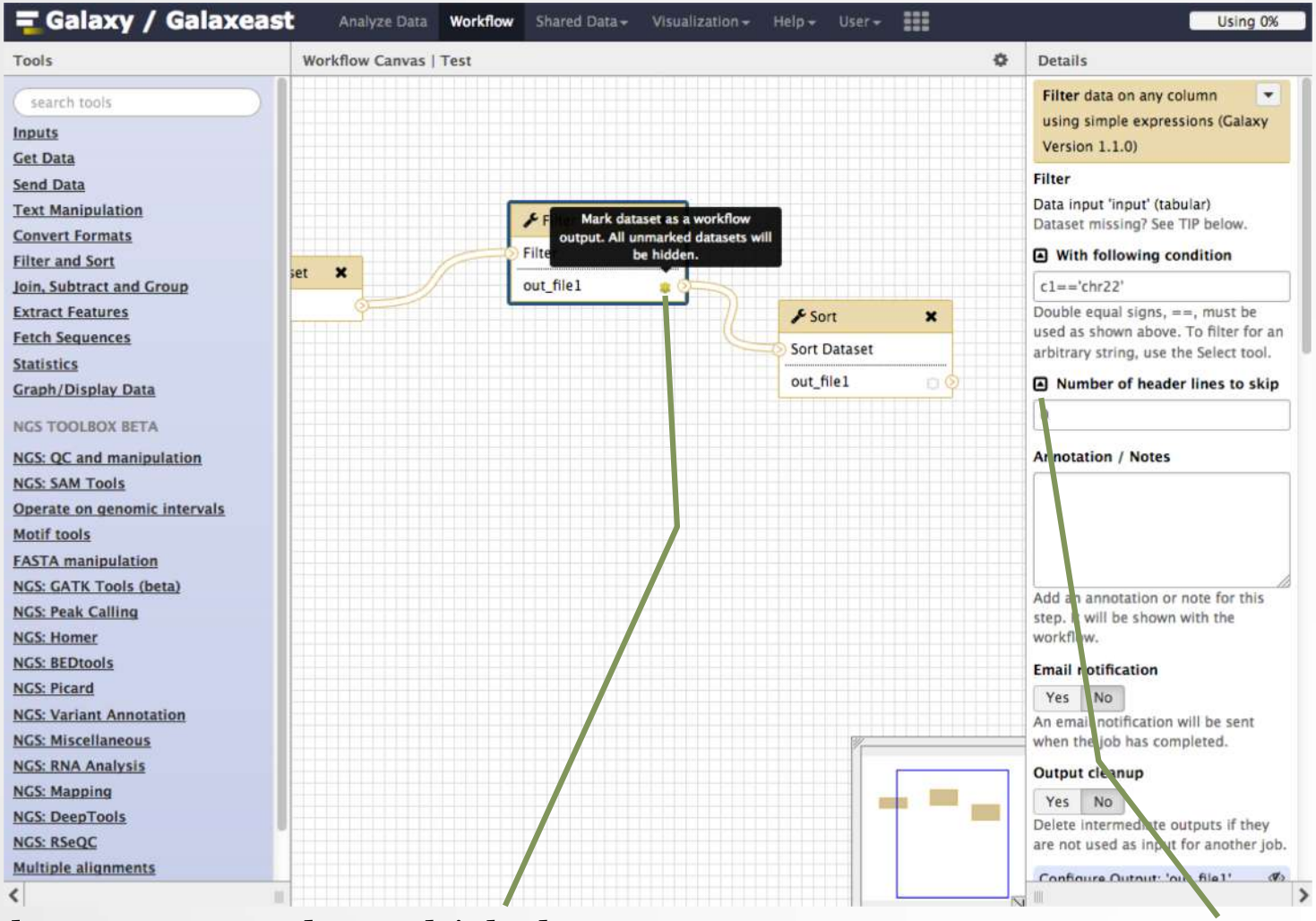
If two steps can be linked together, the link between the two boxes is green

Workflow creation



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

Workflow creation

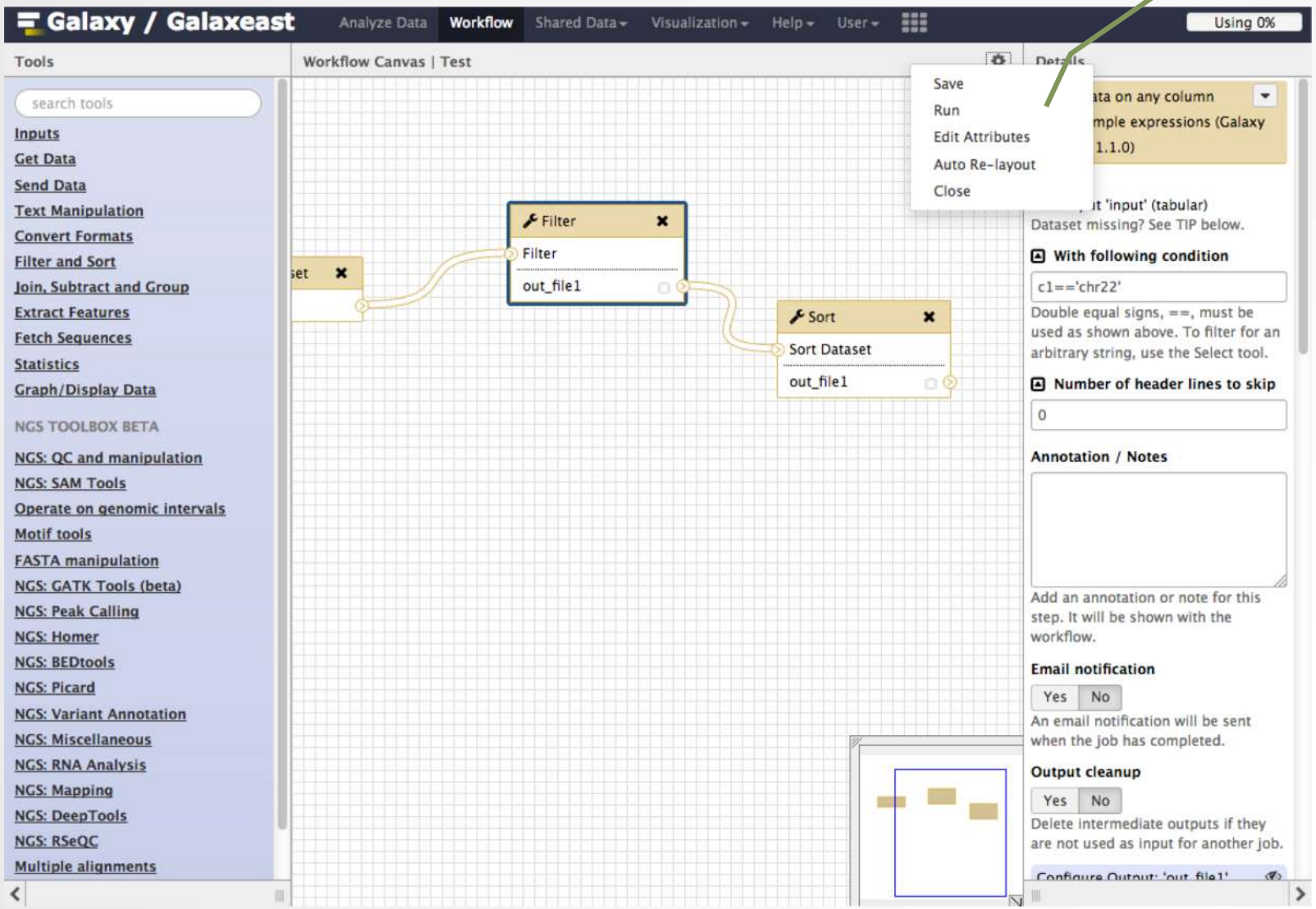


Click on star to select which datasets will be displayed in the history generated when running of the workflow

Click to get the parameter to be set at runtime

Workflow creation

Save, run workflows



Run workflows

Set input file(s)

Galaxy / Galaxeast

Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 34%

Tools

search tools

Get Data
Send Data
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Statistics
Graph/Display Data

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: SAM Tools
Operate on genomic intervals
Motif tools
FASTA manipulation
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: Homer
NGS: BEDtools
NGS: Picard
NGS: Variant Annotation
NGS: Miscellaneous
NGS: RNA Analysis
NGS: Mapping

Running workflow "chip workflow"

Expand All | Collapse

Step 1: Input dataset

Input Dataset

4: chr10_ctr2_1.fastq.gz

type to filter

Step 2: Map with Bowtie for Illumina (version 1.1.3)

Step 3: MACS (version 1.4.2)

Step 4: homer_annotatePeaks (version 0.0.5)

Homer peaks OR BED format

Output dataset 'output_bed_file' from step 3

Genome version

tair10

Extra options

Action:

Hide output 'out_log'.

Send results to a new history

Run workflow

History

search datasets

test

1 shown, 3 deleted

120.7 MB

4: chr10_ctr2_1.fastq

format: fastqsanger, database: hg19

Set parameters

Run workflow

NGS analysis automatization: Galaxy workflows

(answers to questions)

Exercise: your workflows for NGS data analysis

We want to create a workflow to automatically analyze chIP-seq data in Galaxy.

1. Based on what you've learned during the courses, what would be the steps to implement in the workflow? The workflow must handle two input datasets: a treatment and a control (fastq files)

2. Implement the workflow into Galaxy

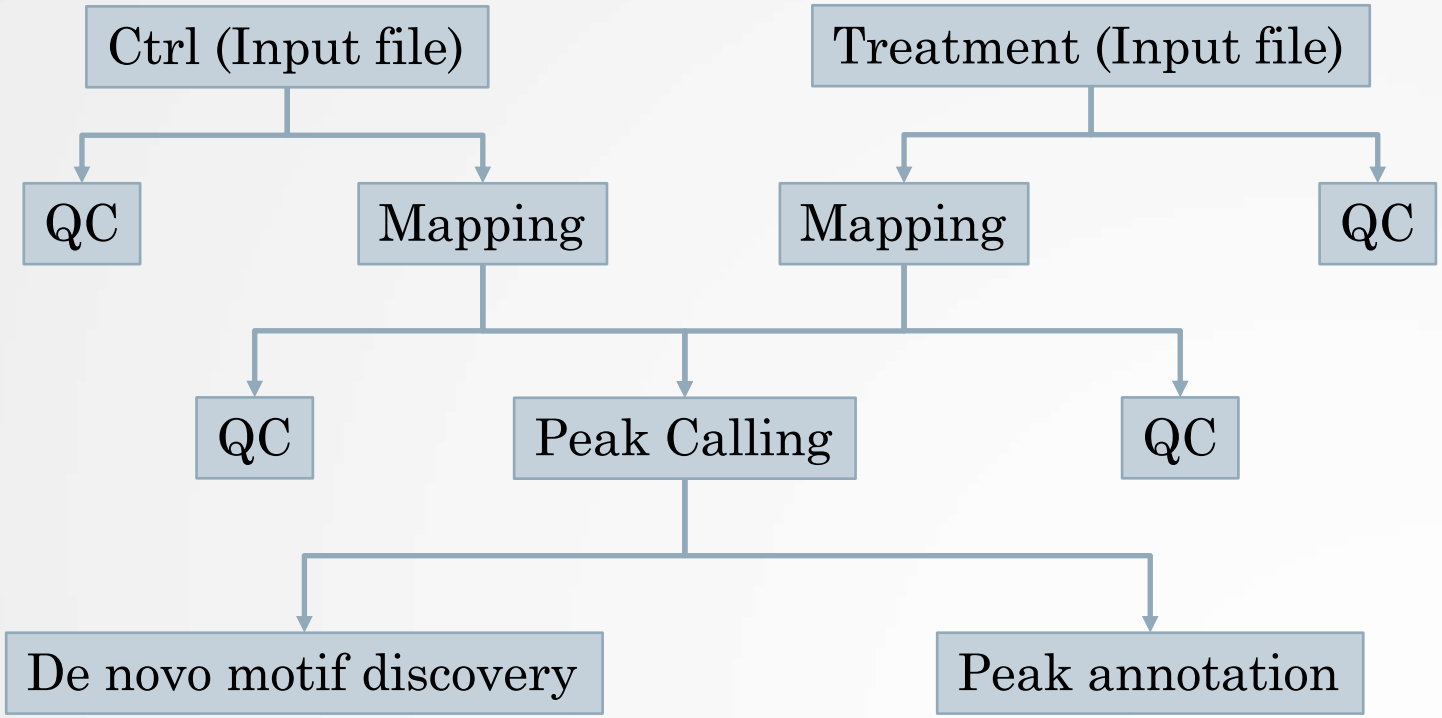
3. Import all datasets from the data library NGS data analysis training > ChIPseq > workflow. Run the workflow on the data

We also want to create a workflow for automatic analysis of RNA-seq data in Galaxy

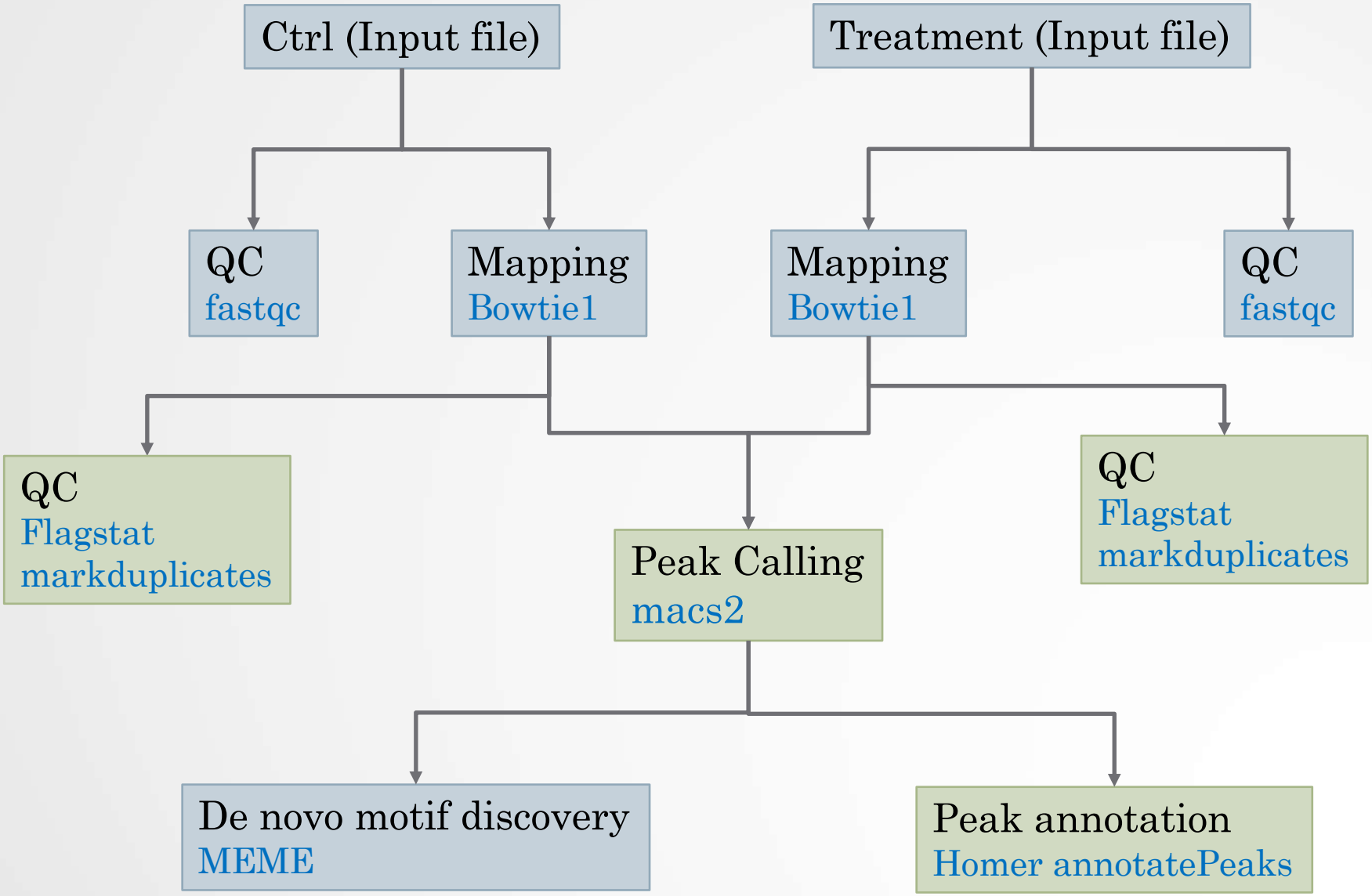
4. What would be the steps, what limitation do you see in implementing RNA-seq data in Galaxy?

Exercise: your workflows for NGS data analysis

1.



Exercise: your workflows for NGS data analysis

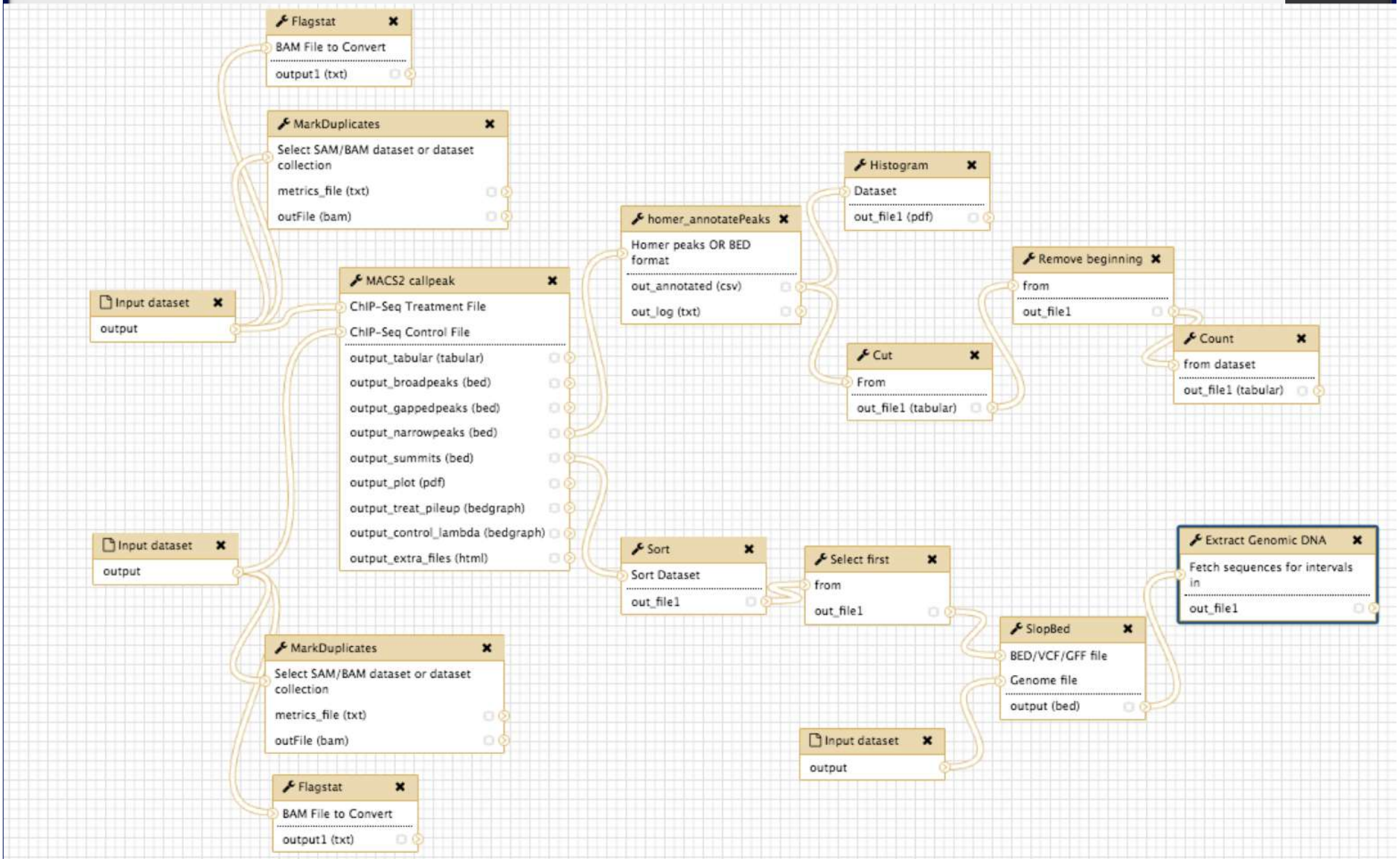


Exercise: your workflows for NGS data analysis

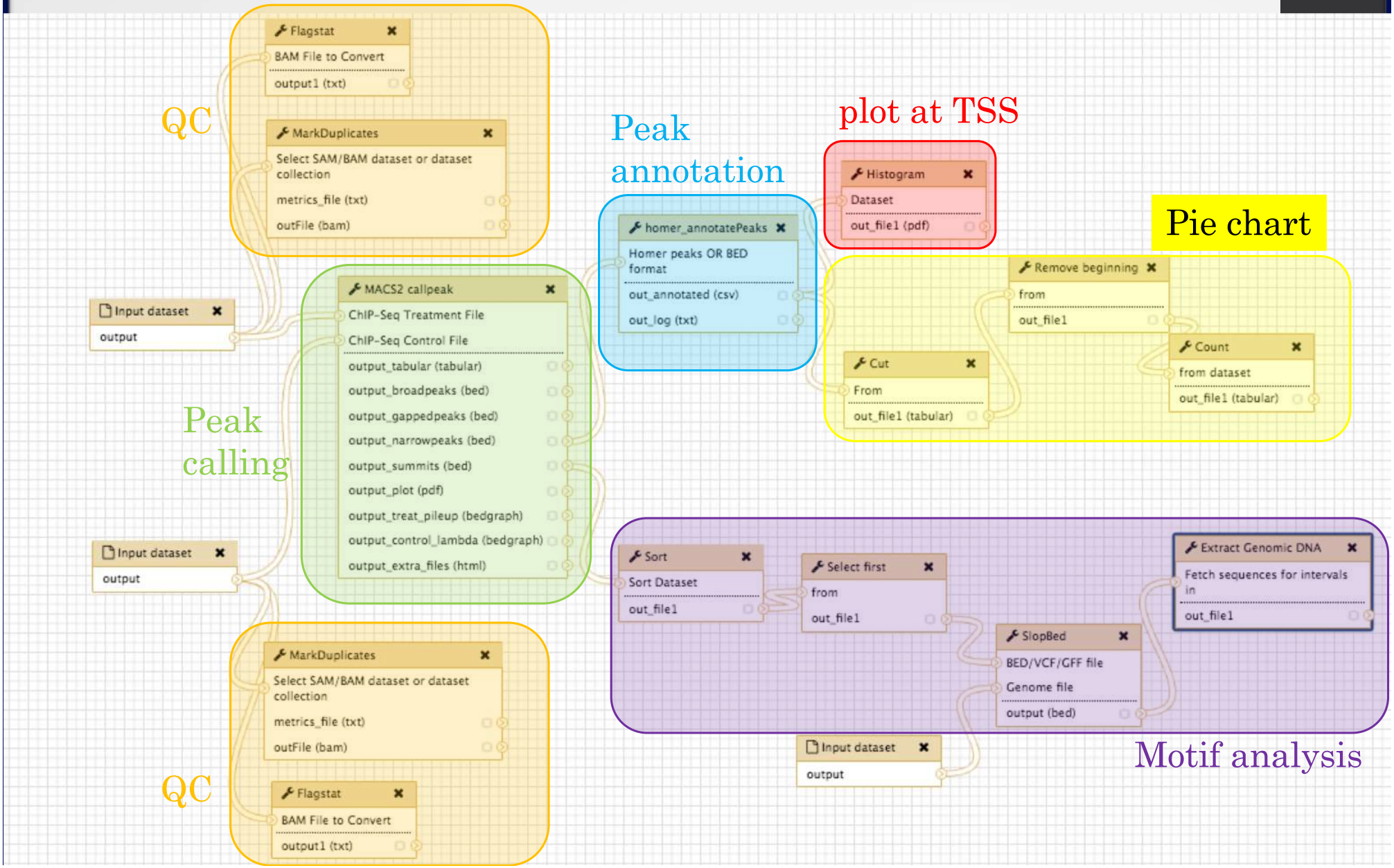
- 2.
  - Go to the history in which you analyzed chIP-seq data (history named "ChIP-seq data analysis")
  - Click on ⚙
  - Select Extract Workflow
  - Enter the workflow name: "ChIP-seq data analysis"
  - Adapt the workflow steps to the needs:
    - Keep the second MACS2 run

  Hint: You can give a name to input datasets to know what kind of file/data is expected to run the workflow. Name the input datasets:
    - "Treatment" for IP input sample
    - "Control" for control sample
    - "Chrom length" for chromosome length dataset
  - Click on Create Workflow
  - Then to edit the workflow:
    - Click on Workflow (top menu)
    - Click on ChIP-seq data analysis > Edit

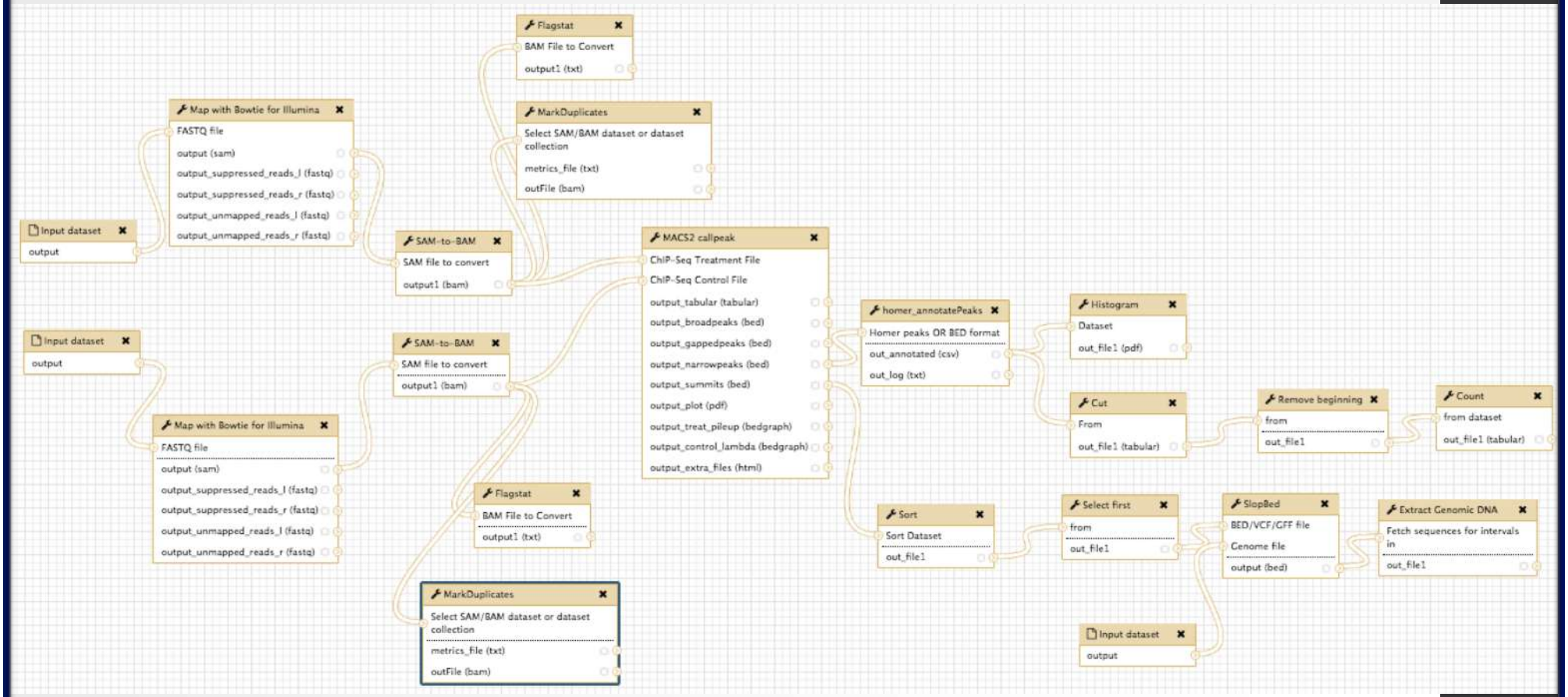
Exercise (before editing)



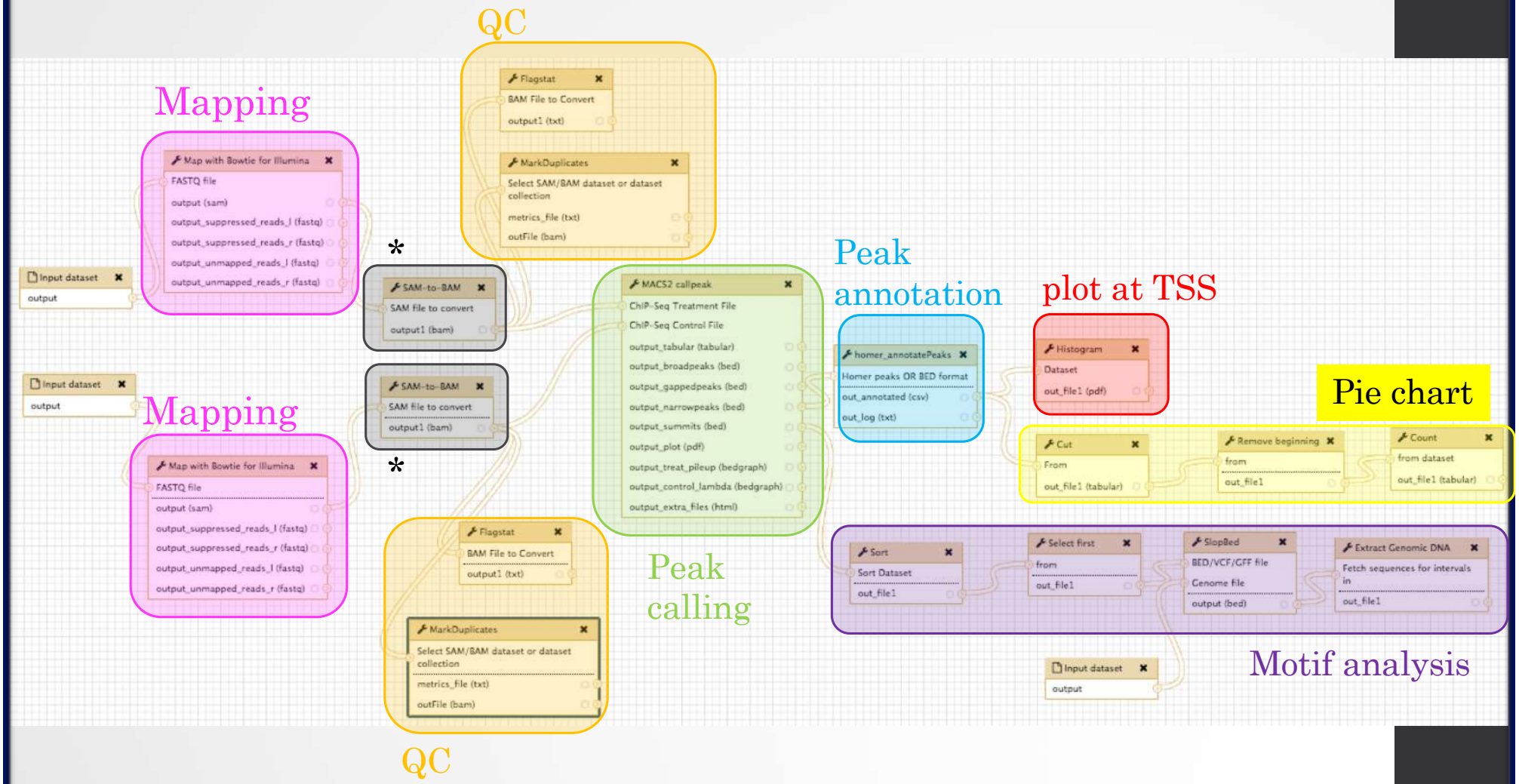
Exercise (before editing)



Exercise (after editing)

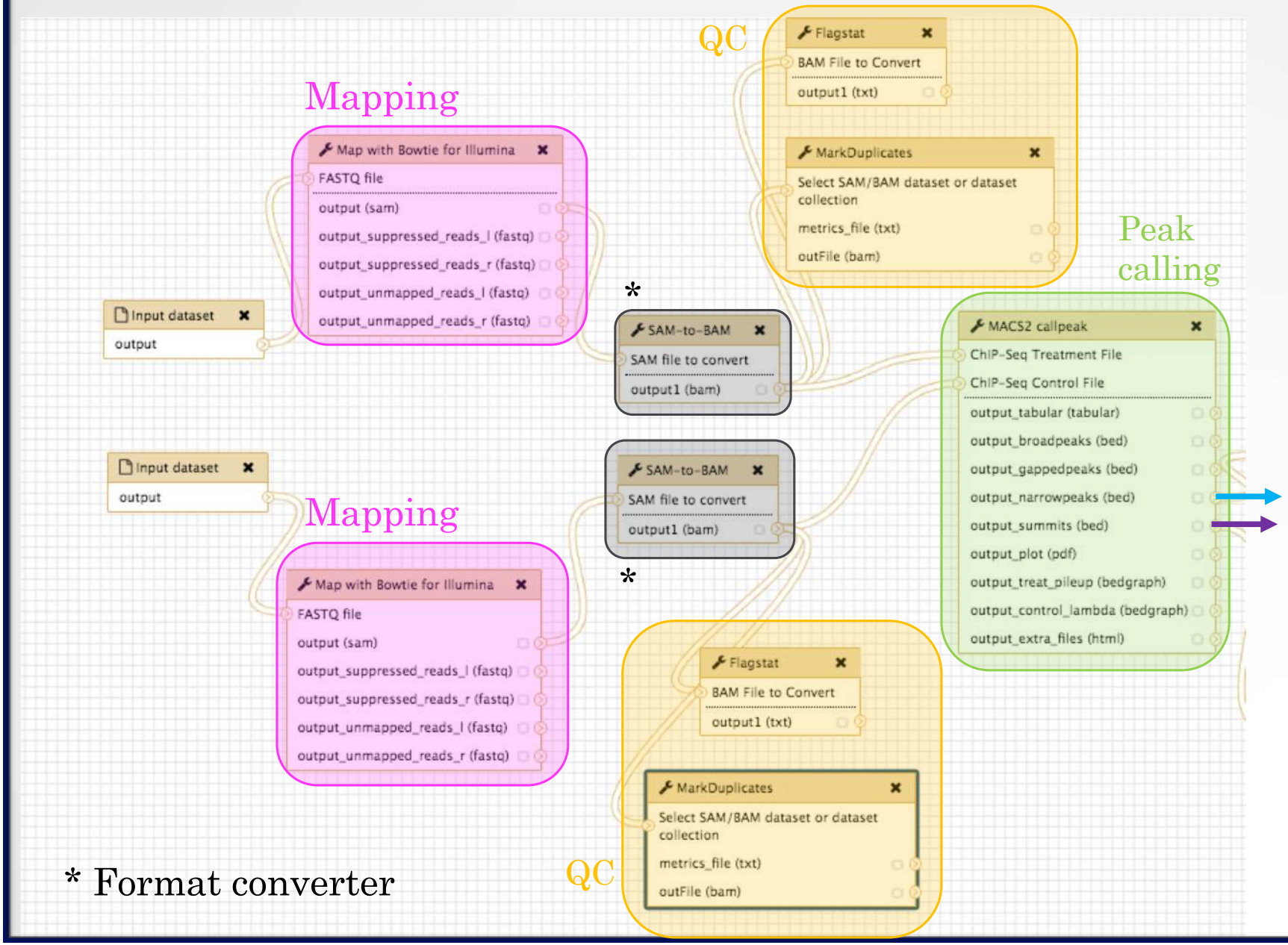


Exercise (after editing)



* Format converter

Exercise (after editing - ZOOM)



* Format converter

Exercise: your workflows for NGS data analysis

- Bowtie 1 parameters:
  - Select a reference genome : **set at runtime**

  - Bowtie settings to use: **Full parameters list**
  - Whether or not to make Bowtie guarantee that reported singleton alignments are 'best' in terms of stratum and in terms of the quality values at the mismatched positions (--best): **Use best**
  - Whether or not to report only those alignments that fall in the best stratum if many valid alignments exist and are reportable (--strata): **Use strata**
  - Suppress all alignments for a read if more than n reportable alignments exist (-m): **1**

  Hint: Do it for the two alignment steps

- SAM-to-BAM
  - Reference Genome: set at run time

- MACS2
  - Build Model: **Build the shifting model**

Exercise: your workflows for NGS data analysis

- Homer annotatePeaks
  - Genome version: set at run time
- Select the box of the tool **cut**
  - Click on **configure Output: out_file1**
  - Change datatype: **interval**
- MEME parameters:
  - Options Configuration: **Advanced**
  - Number of different motifs to search: **2**
  - Min width of motif to search: **6**
  - Max width of motif to search: **12**
  - E-value to stop looking for motifs : **1**
  - I certify that I am not using this tool for commercial purposes: **Yes**
- Click on ⚙ and Save

Exercise: your workflows for NGS data analysis

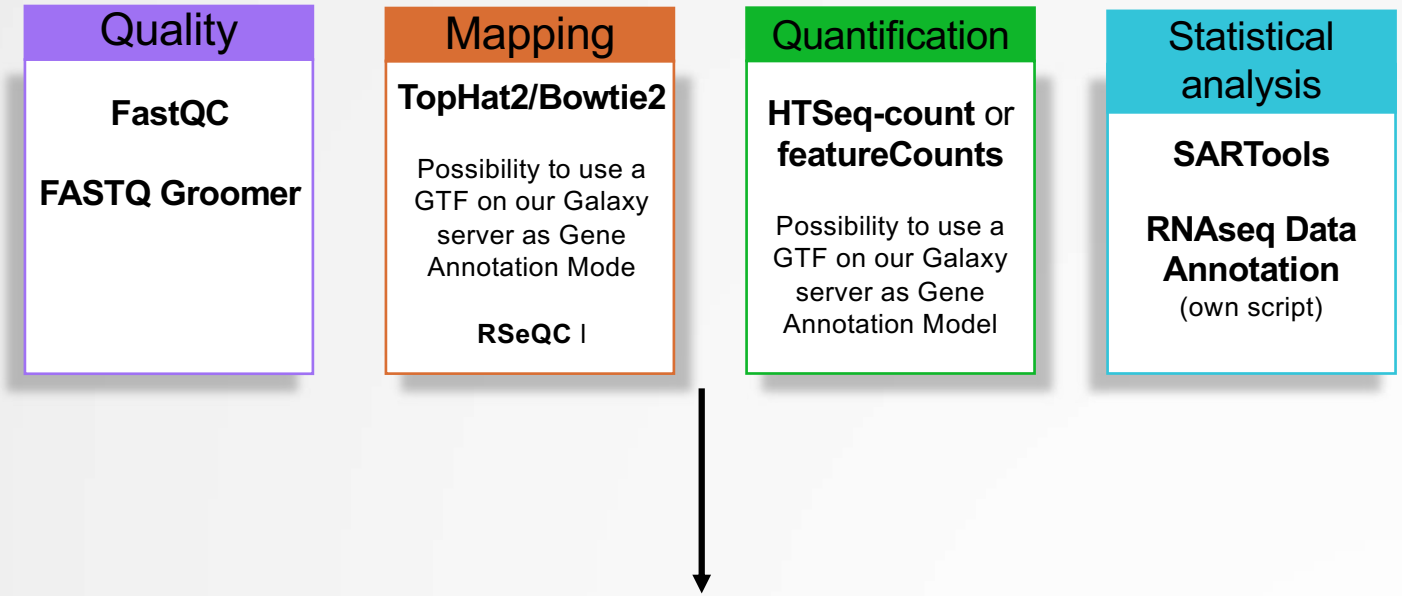
- 3.
  - Click on Analyze Data (top menu)
  - Go to Shared data > Data Libraries > NGS data analysis training > ChIPseq > workflow and add the two datasets to your history.
    - Import all data
    - Click on the button To history
    - Add the datasets to the new history “ChIP-seq test workflow”
  - Click on Workflow (top menu)
  - Click on the workflow “ChIP-seq data analysis” and select Run.
    - Treatment: chr10_mitf_2.fastq
    - Control: chr10_ctrl2_1.fastq
    - Chrom length: hg38.len
    - Step 4: Map with Bowtie for Illumina:
      - Select a reference genome: hg38
    - Step 5: Map with Bowtie for Illumina
      - Select a reference genome: hg38
    - Step 13: Homer annotatePeaks
      - Genome version: hg38
    - Click on Run workflow

Exercise: your workflows for NGS data analysis

- 3.
  - Click on the workflow “ChIP-seq data analysis” and select Run.
    - Treatment: chr10_mitf_2.fastq
    - Control: chr10_ctrl2_1.fastq
    - Chrom length: hg38.len
    - Step 4: Map with Bowtie for Illumina:
      - Select a reference genome: hg38
    - Step 5: Map with Bowtie for Illumina
      - Select a reference genome: hg38
    - Step 6: Sam-to-BAM
      - Using reference genome: hg38
    - Step 7: Sam-to-BAM
      - Using reference genome: hg38
    - Step 13: Homer annotatePeaks
      - Genome version: hg38
    - Click on Run workflow

Exercise: your workflows for NGS data analysis

- 4.



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

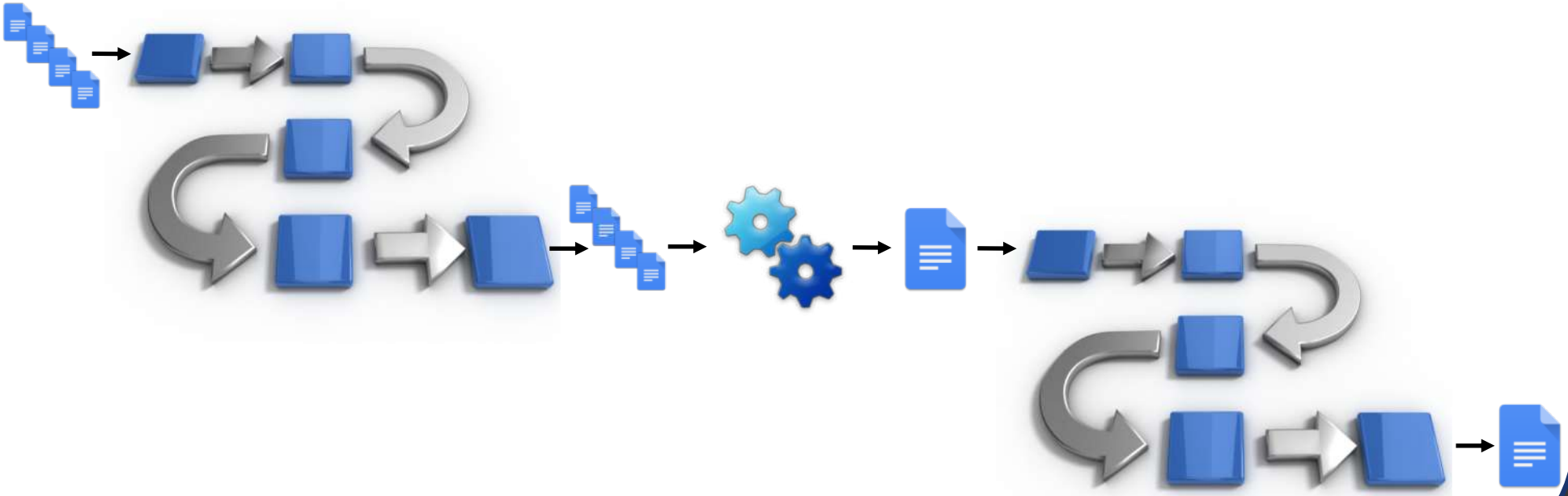
RNAseq workflow : limits

Main workflow

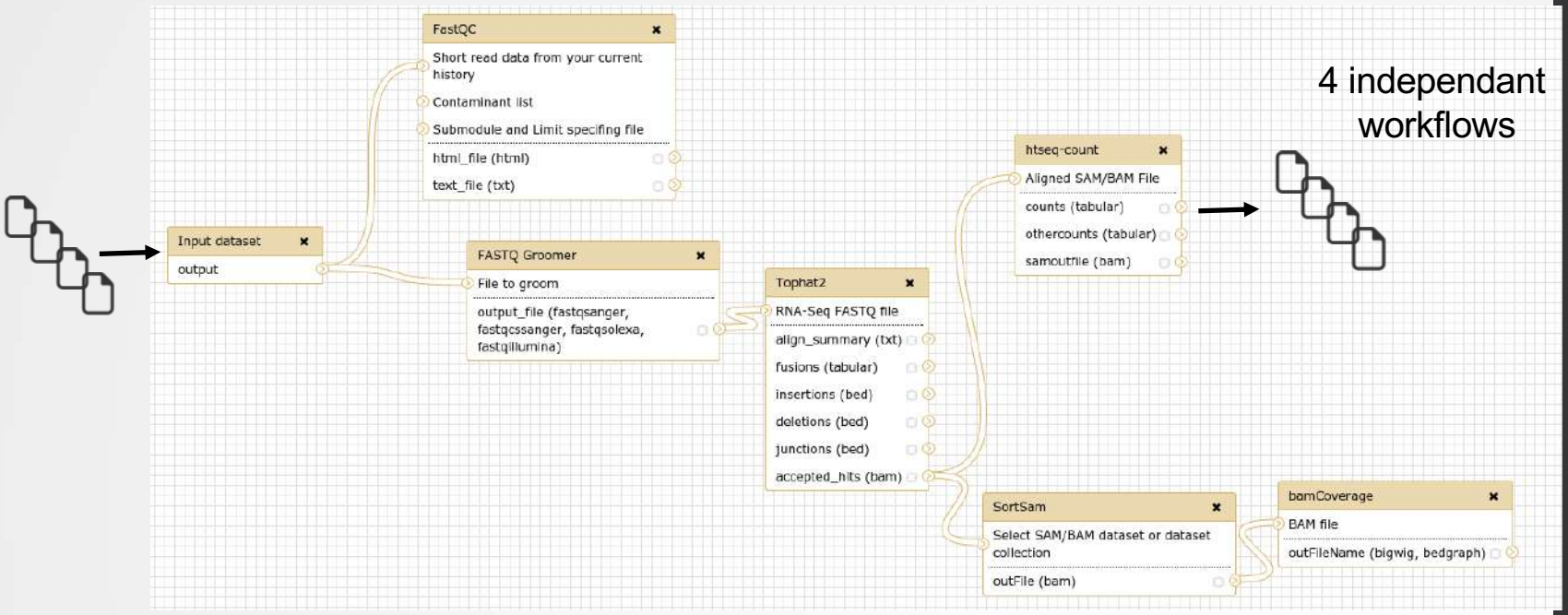
Sub workflow 1

Files merging

Sub workflow 2



RNAseq workflow : limits



HTSeq-count outputs

→ Merge → SARTools