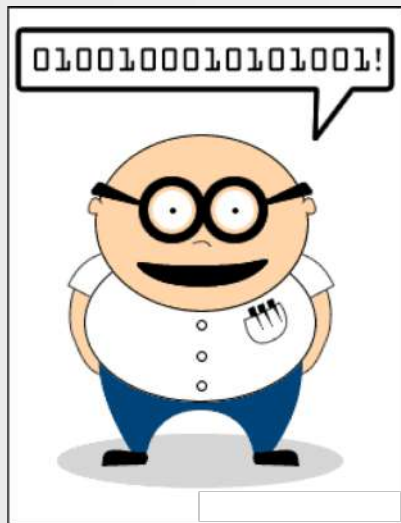


NGS analysis automatization: Galaxy workflows

Stéphanie Le Gras
(slegras@igbmc.fr)

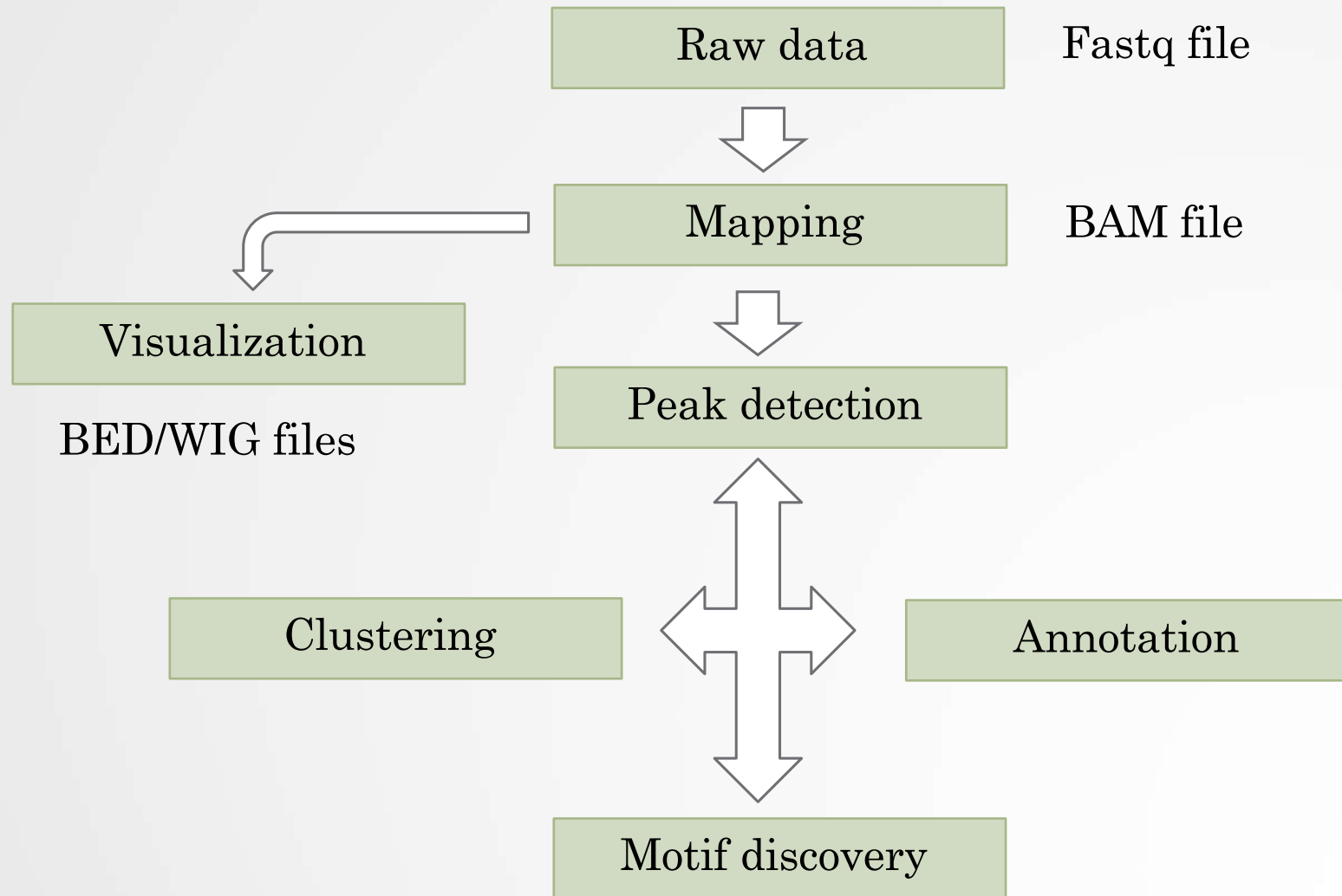
A long time ago...

Input data



**PIPELINE/
WORKFLOW**

More recently...



During the entire training session..



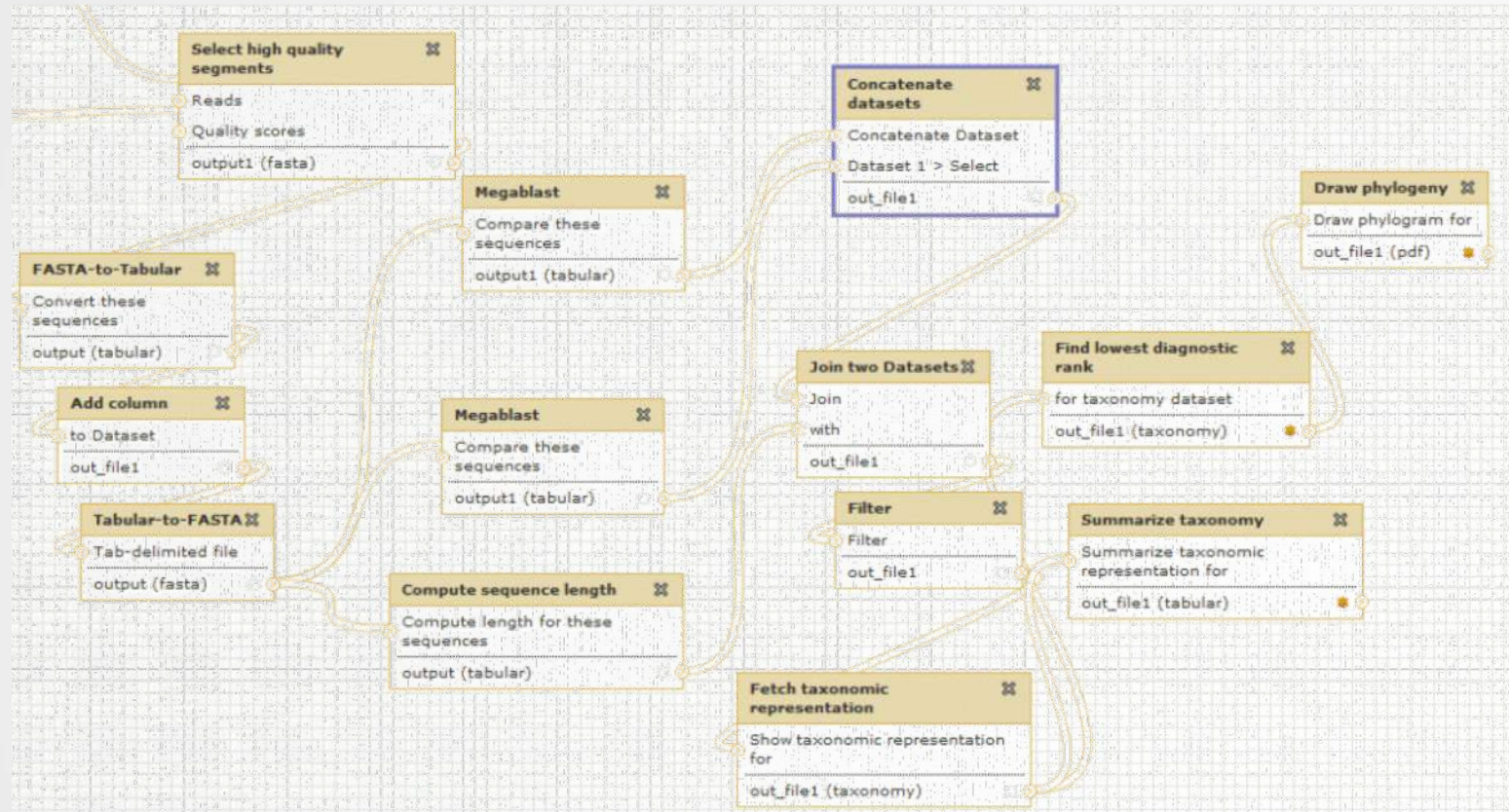
Galaxy

PROJECT

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 web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Workflow' tab is selected. On the left, a 'Tools' sidebar lists various categories such as NGS: SAMtools, NGS: BamTools, NGS: Picard, NGS: VCF Manipulation, NGS: Peak Calling, NGS: Variant Analysis, NGS: RNA Structure, NGS: Du Novo, NGS: Gemini, Operate on Genomic Intervals, Statistics, Graph/Display Data, CloudMap, Phenotype Association, BEDTools, Genome Diversity, EMBOSS, Regional Variation, FASTA manipulation, Multiple Alignments, Metagenomic Analysis, Multiple regression, Multivariate Analysis, Motif Tools, STR-FM: Microsatellite Analysis, NCBI SRA Tools, DEPRECATED, NGS: GATK Tools (beta), and Workflows. The 'Workflows' section is expanded to show 'All workflows'. The main content area displays a welcome message: 'Galaxy is an open source, web-based platform for data-intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.' Below this is a logo for '080+ Public Galaxy Servers and still counting' and a 'Tweets' section by @galaxyproject. The right sidebar shows 'History' with a search bar and a message: 'This history is empty. You can load your own data or get data from an external source'. A green arrow points from the 'Workflow' tab to the text 'Create, run, edit (...) workflows'. Another green arrow points from the 'All workflows' link in the sidebar to the text 'Run workflows'.

Create, run,
edit (...) workflows

Run workflows

Workflows

Your workflows

You have no workflows.

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu

Create new workflow Upload or import workflow

Create workflows

Create New Workflow

Workflow Name:

Workflow Annotation:

A description of the workflow; annotation is shown alongside shared or published workflows.

Create

Give a name to the workflow

Workflow creation

Galaxy / Galaxeast Analyze Data Workflow Shared Data Visualization Help User Using 0%

Tools

search tools

Inputs

- 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
- NGS: DeepTools
- NGS: RSeQC
- Multiple alignments

Workflow Canvas | Test

Details

Edit Workflow Attributes

Name:
Test

Tags:

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

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

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

The screenshot displays the Galaxy / Galaxeast interface for creating a workflow. The main canvas shows two steps: 'Input dataset' and 'Filter'. A green line connects the 'output' of the 'Input dataset' step to the 'Filter' step. The 'Filter' step is configured with the condition 'c1==chr22'. The 'Details' panel on the right shows the configuration for the 'Filter' step, including the condition 'c1==chr22', the number of header lines to skip (0), and options for email notification and output cleanup.

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

Workflow creation

The screenshot displays the Galaxy workflow editor interface. The top navigation bar includes 'Galaxy / Galaxeast', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and 'Using 0%'. The main area is divided into three sections: 'Tools', 'Workflow Canvas | Test', and 'Details'.

The 'Tools' panel on the left lists various tool categories such as 'Inputs', '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', and 'NGS: SAM Tools'. The 'Filter and Sort' section is expanded, showing several filter-related tools.

The 'Workflow Canvas | Test' section shows a workflow with two tools: 'Input dataset' and 'Filter'. The 'Input dataset' tool has an output named 'output' which is connected to the 'Filter' tool. The 'Filter' tool has an output named 'out_file1'.

The 'Details' panel on the right is for the 'Filter' tool. It shows the tool's name, version, and description. The 'With following condition' section is checked, and the condition 'c1=='chr22'' is entered. The 'Number of header lines to skip' is set to 0. There are also sections for 'Annotation / Notes', 'Email notification', and 'Output cleanup'.

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

Workflow creation

The screenshot shows the Galaxy workflow editor interface. The top navigation bar includes 'Galaxy / Galaxeast', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and 'Using 0%'. The left sidebar contains a 'Tools' panel with a search bar and various tool categories like 'Inputs', 'Text Manipulation', 'Statistics', and 'NGS TOOLBOX BETA'. The central 'Workflow Canvas | Test' area shows a workflow with a 'Filter' tool connected to a 'Sort Dataset' tool. A tooltip over the 'Filter' tool reads: 'Mark dataset as a workflow output. All unmarked datasets will be hidden.' The right 'Details' panel shows configuration options for the 'Filter' tool, including 'Filter data on any column using simple expressions (Galaxy Version 1.1.0)', 'Filter' (Data input 'input' (tabular)), 'With following condition' (c1='chr22'), and 'Number of header lines to skip'. A green arrow points from the 'Number of header lines to skip' field to a 'Configure Output' button at the bottom of the details panel.

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

The screenshot displays the Galaxy / Galaxeast workflow editor. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main area is the 'Workflow Canvas | Test', which shows a workflow with three steps: 'Set', 'Filter', and 'Sort'. The 'Filter' step is currently selected, and a context menu is open over it, showing options: 'Save', 'Run', 'Edit Attributes', 'Auto Re-layout', and 'Close'. A green arrow points from the text 'Save, run workflows' to the 'Run' option in the menu. The right-hand side features a configuration panel for the selected 'Filter' tool, with sections for 'With following condition' (containing the expression 'c1==chr22'), 'Number of header lines to skip' (set to 0), 'Annotation / Notes', 'Email notification' (Yes/No buttons), and 'Output cleanup' (Yes/No buttons). The bottom of the configuration panel shows 'Configure Output: out_file1'.

Run workflows

Set input file(s)

The screenshot displays the Galaxy web interface for running a workflow. The main panel shows the workflow steps:

- Step 1: Input dataset**
 - Input Dataset: 4: chr10_ctr2_1.fastq.gz
- 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: [checkbox]
 - Action: Hide output 'out_log'.

At the bottom of the workflow configuration, there is a checkbox for "Send results to a new history" and a "Run workflow" button.

The History panel on the right shows the dataset "4: chr10_ctr2_1.fastq" with format "fastqsanger" and 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.

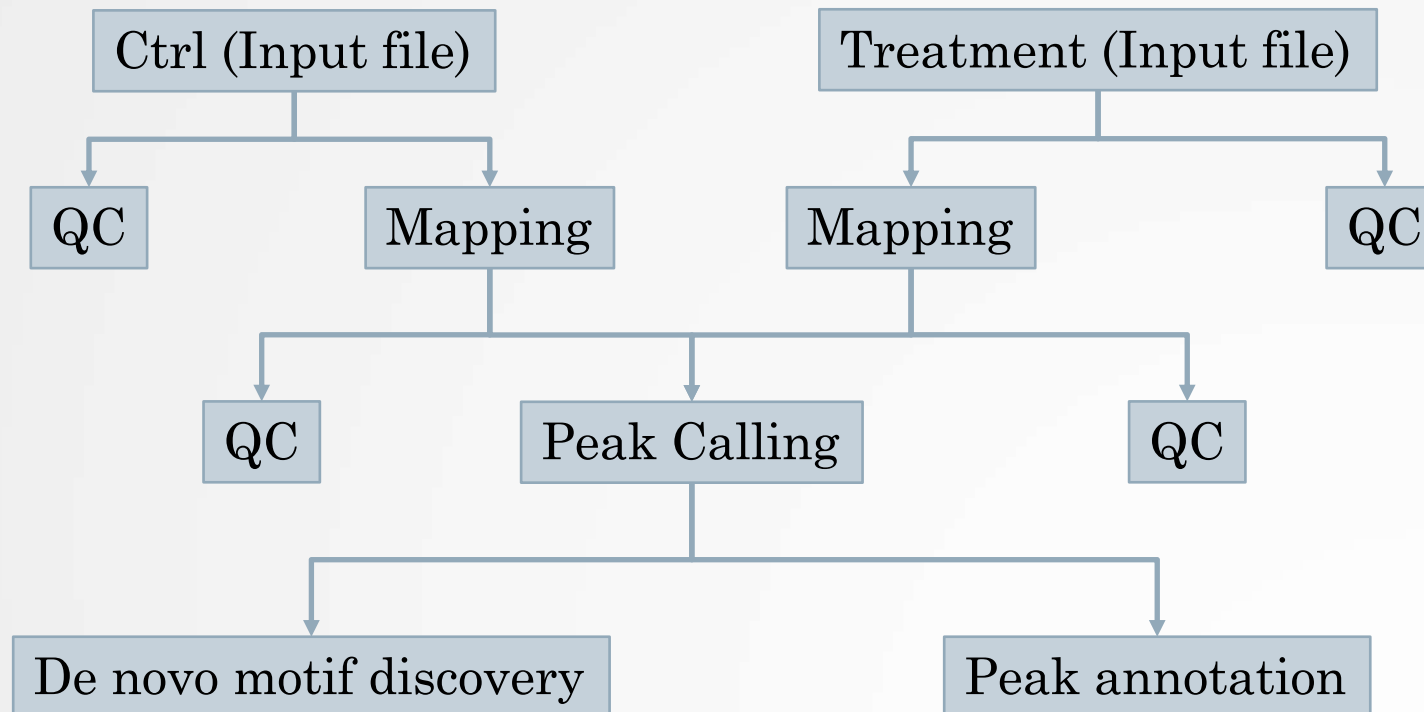
1. Based on what you've learned during the courses, what would be the steps to implement in the workflow? The workflow must handle two input datasets: a treatment and a control (fastq files)
2. Implement the workflow into Galaxy
3. Import all datasets from the data library NGS data analysis training > ChIPseq > workflow. Run the workflow on the data

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

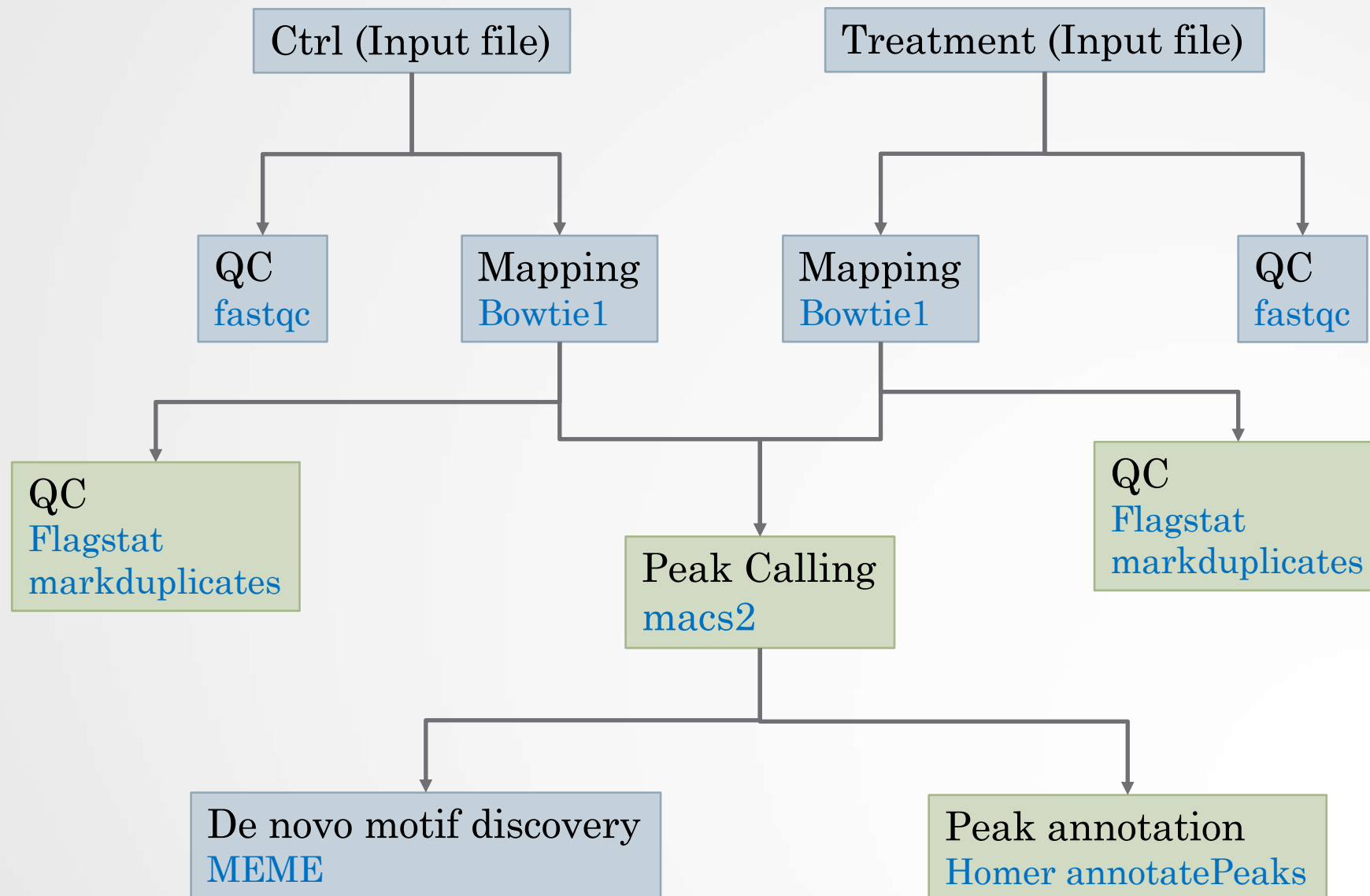
4. What would be the steps, what limitation do you see in implementing RNA-seq data in Galaxy?

Exercise: your workflows for NGS data analysis


1.



Exercise: your workflows for NGS data analysis



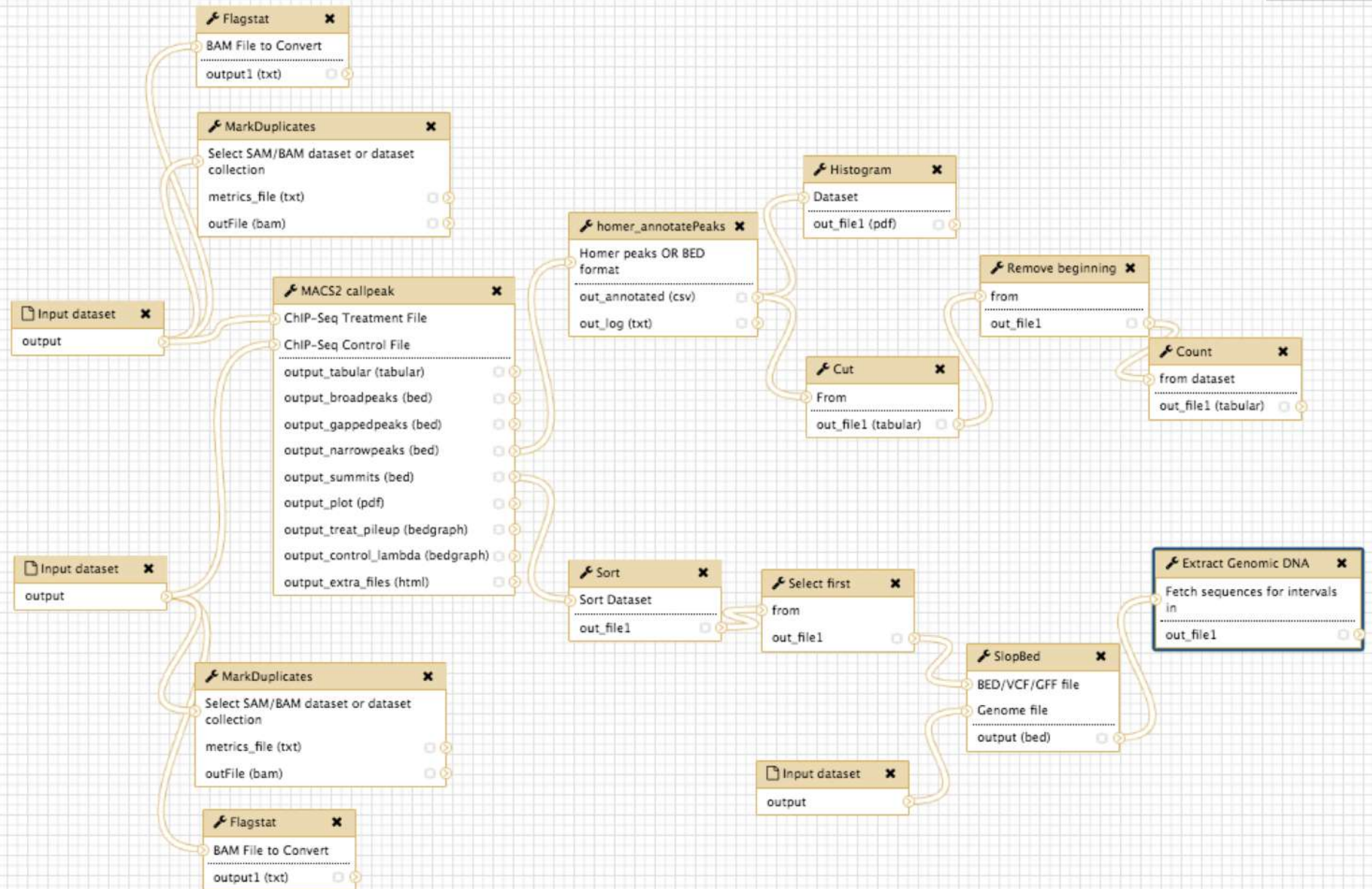
Exercise: your workflows for NGS data analysis

- 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

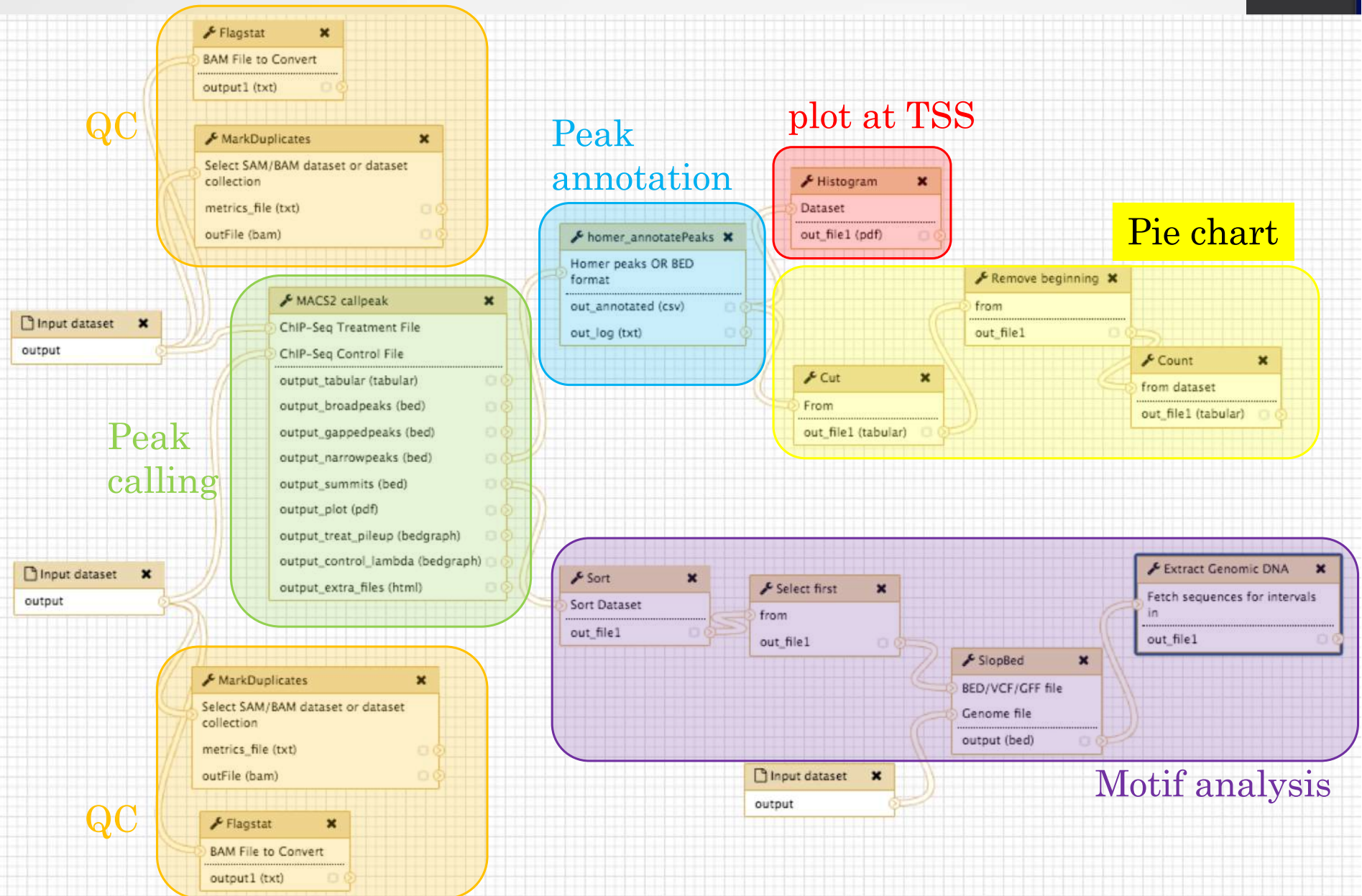
Hint: You can give a name to input datasets to know what kind of file/data is expected to run the workflow. Name the input datasets:

- “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

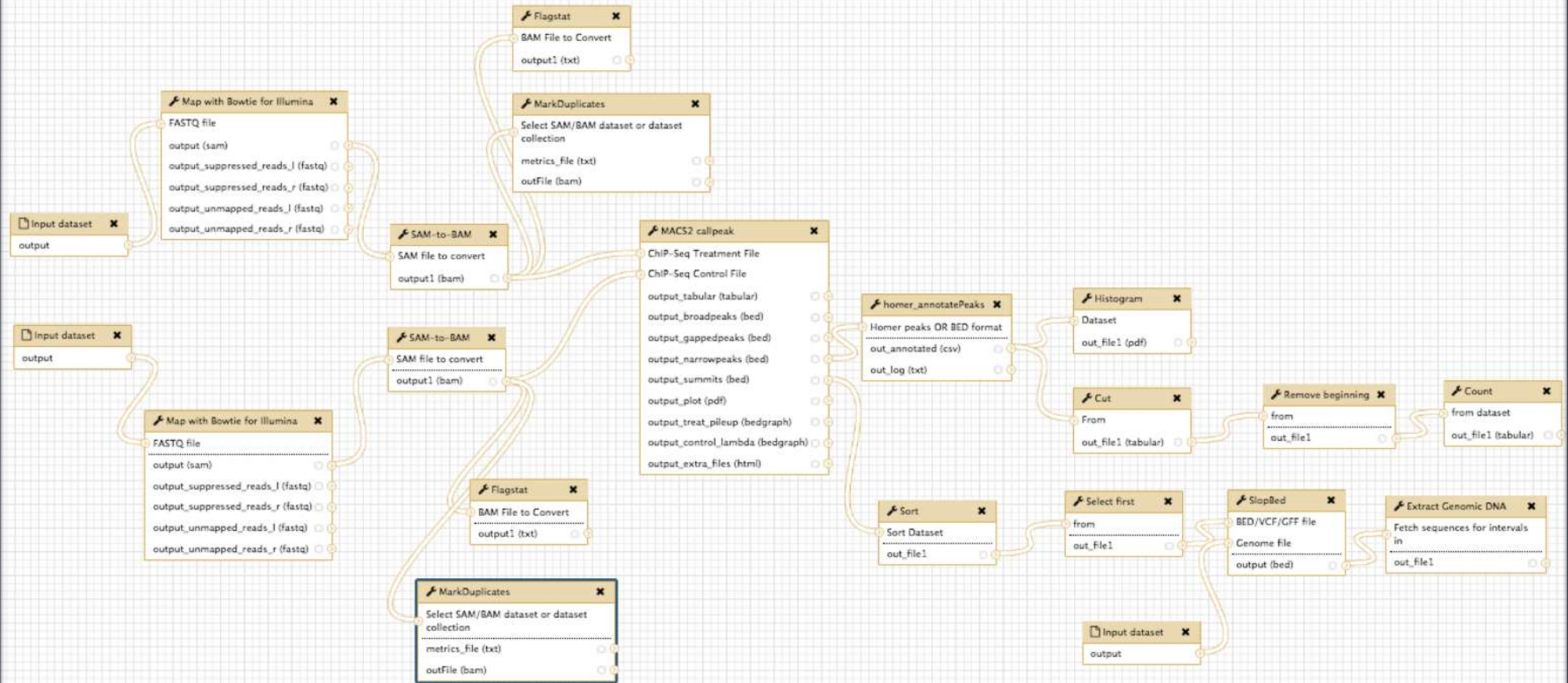
Exercise (before editing)



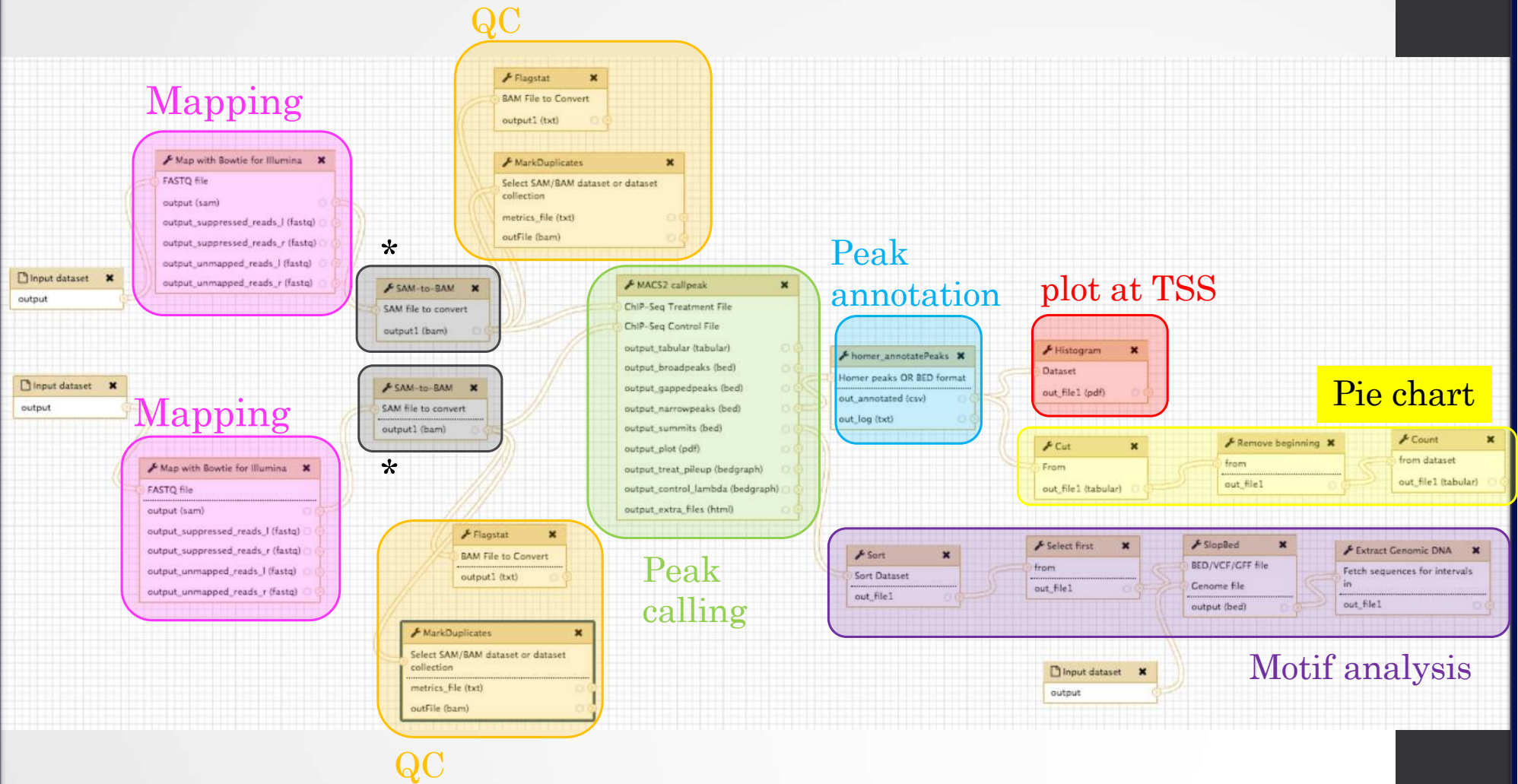
Exercise (before editing)



Exercise (after editing)

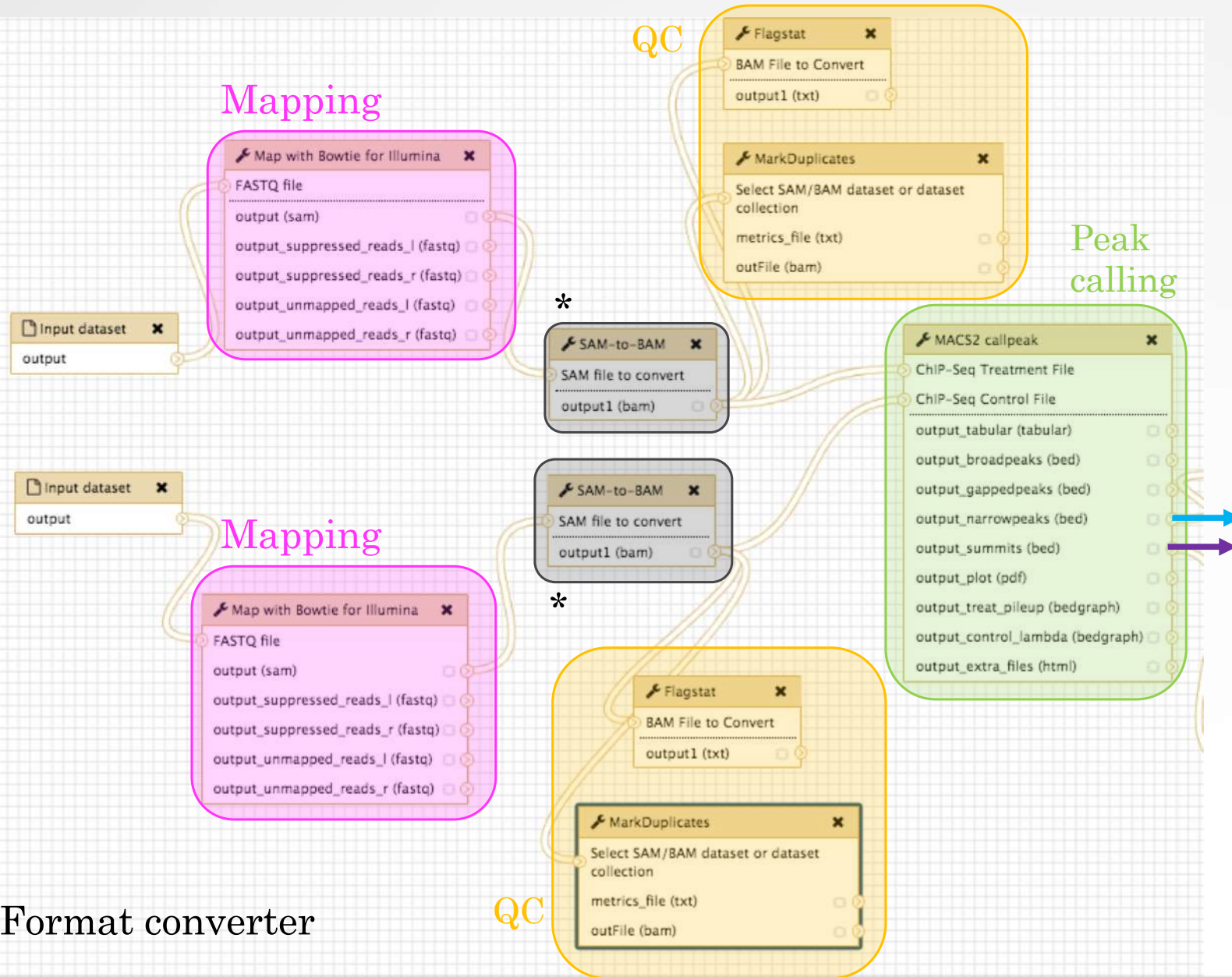


Exercise (after editing)



* Format converter

Exercise (after editing - ZOOM)



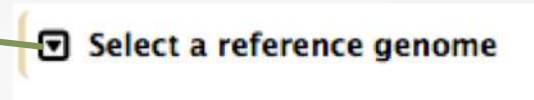
* Format converter

QC

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

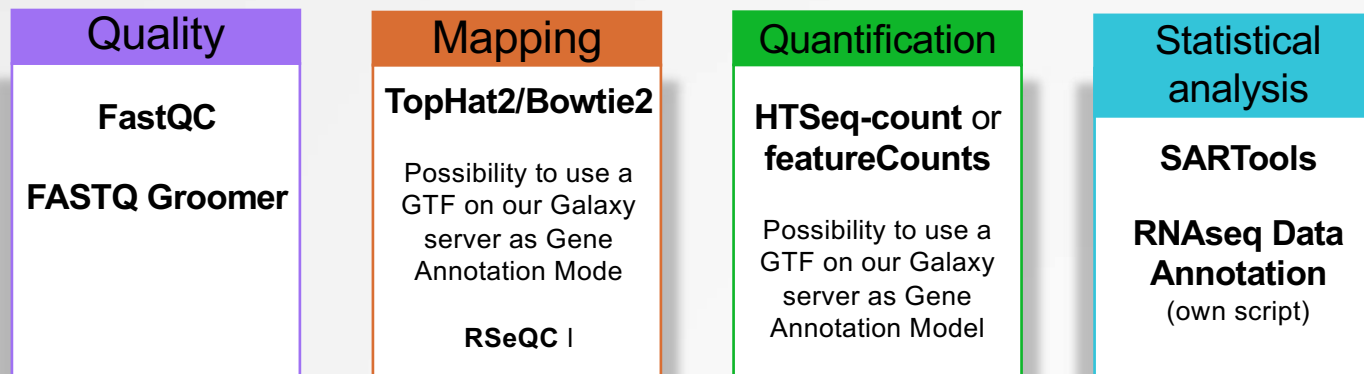
- 3.
 - Click on Analyze Data (top menu)
 - Go to Shared data > Data Libraries > NGS data analysis training > ChIPseq > workflow and add the two datasets to your history.
 - Import all data
 - Click on the button To history
 - Add the datasets to the new history “ChIP-seq test workflow”
 - Click on Workflow (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

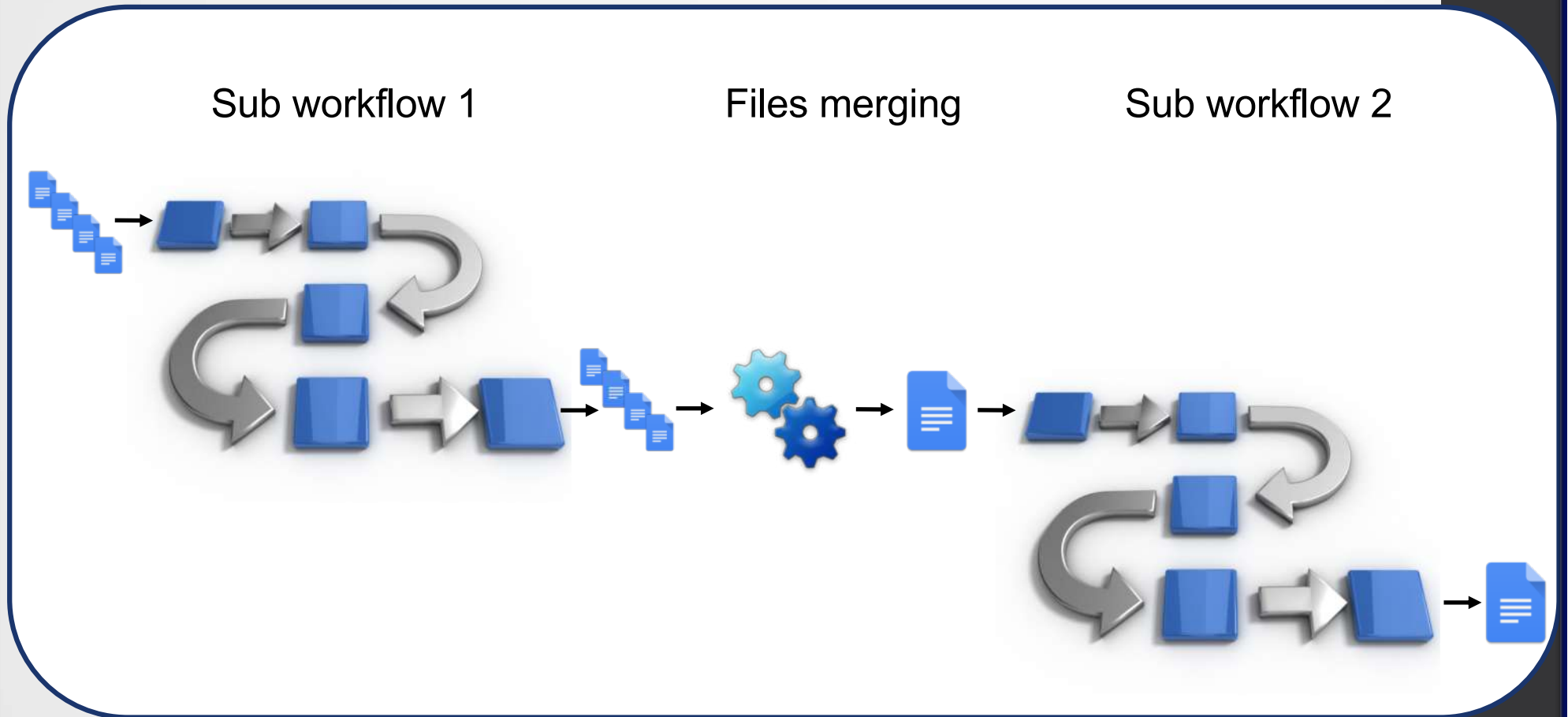
• 4.



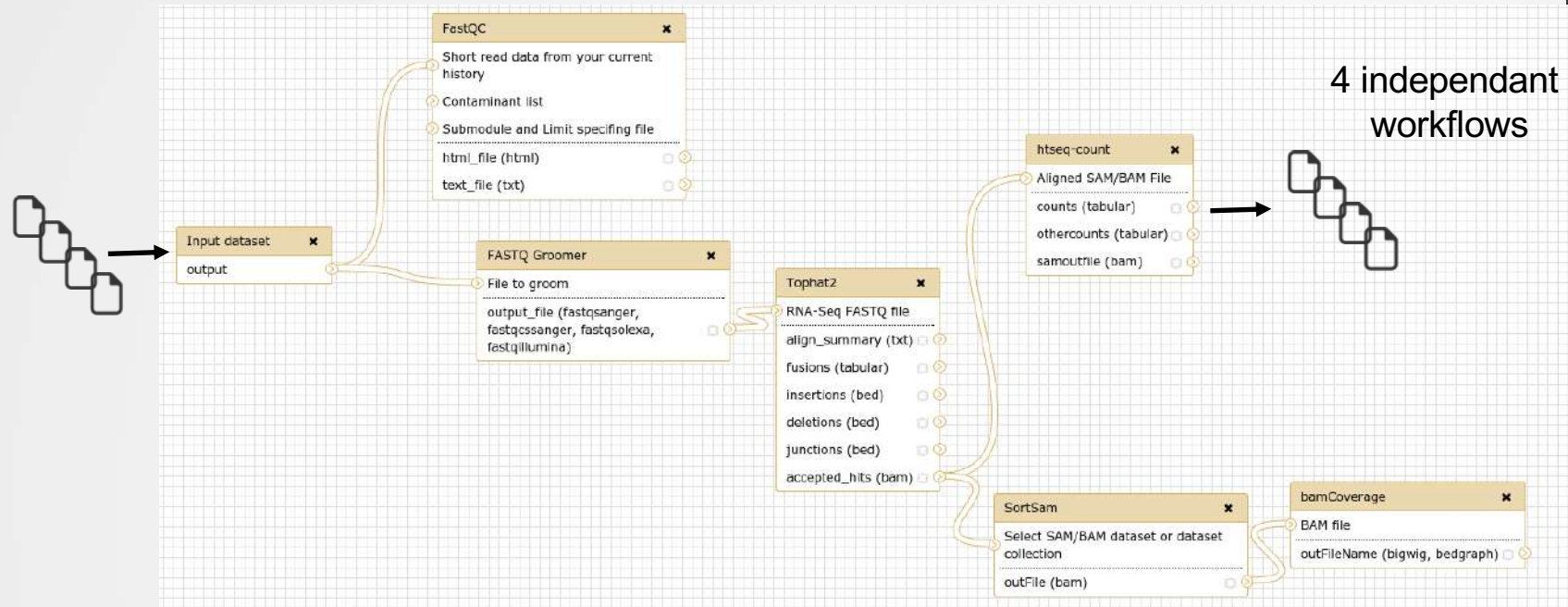
Problem : all steps can't be in a same workflow

RNAseq workflow : limits

Main workflow



RNAseq workflow : limits



HTSeq-count outputs



Merge



SARTools