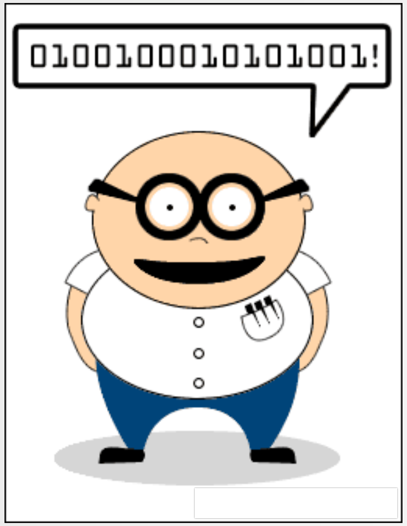
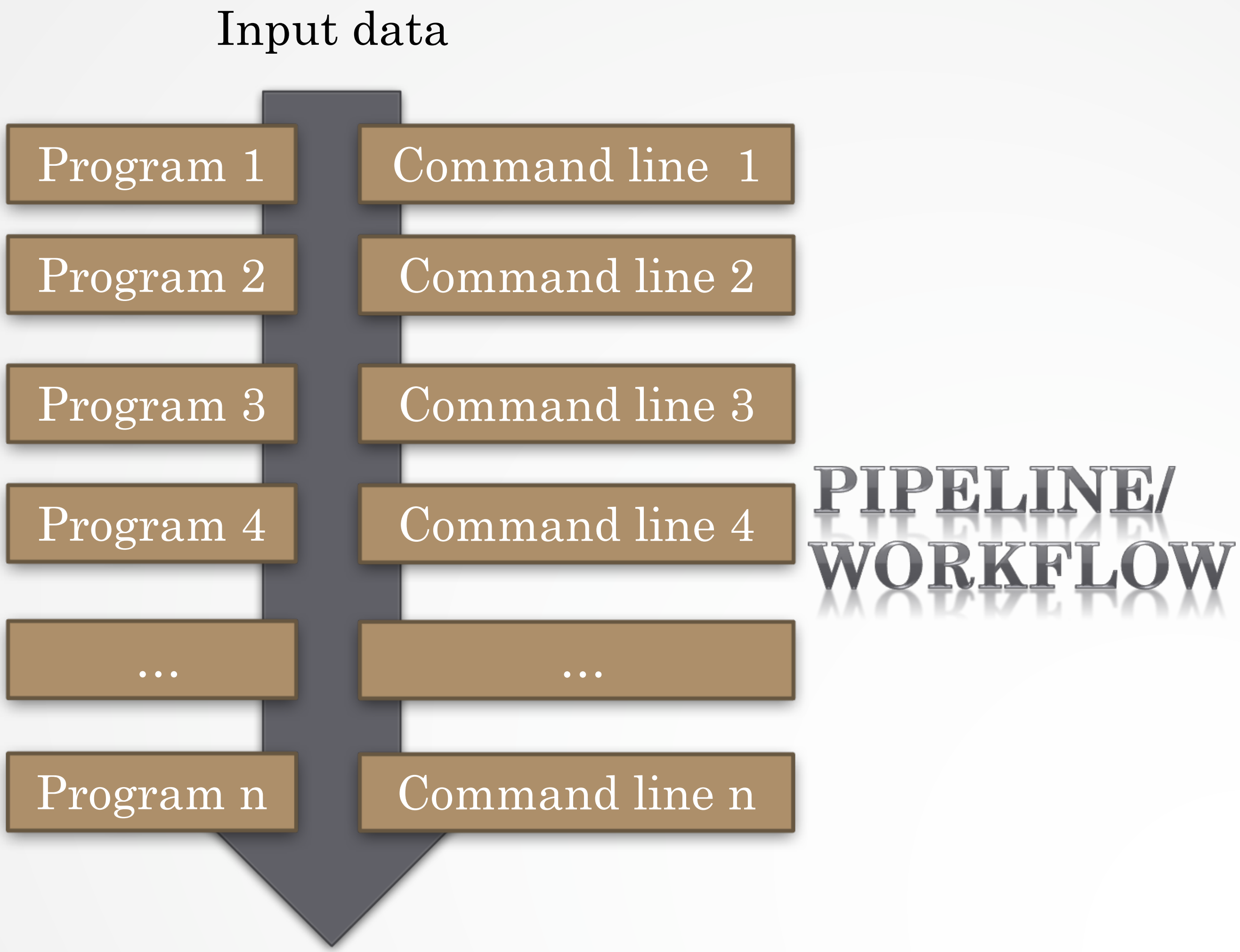
NGS analysis automatization: Galaxy workflows

Stéphanie Le Gras
(slegras@igbmc.fr)

A long time ago…





Input data

Program 1 | Command line 1

Program 2 | Command line 2

Program 3 | Command line 3

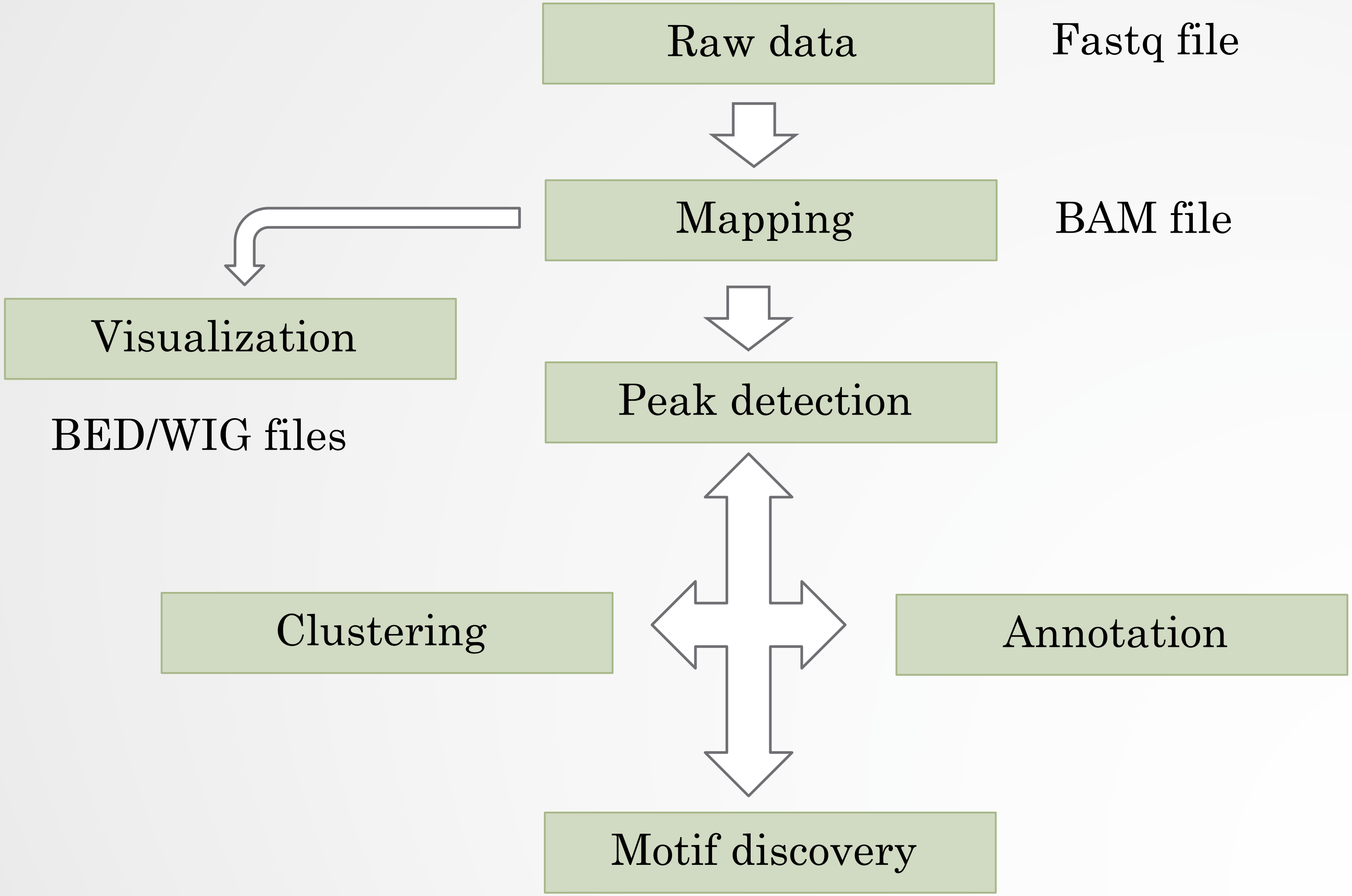
Program 4 | Command line 4

… | …

Program n | Command line n

PIPELINE/ WORKFLOW

More recently…



During the entire training session..

Galaxy

PROJECT

What if we'd mix all together

0100100010101001!

Input data

Program 1

Program 2

Program 3

Program 4

…

Program n

Command line 1

Command line 2

Command line 3

Command line 4

…

Command line n

PIPELINE/ WORKFLOW

Raw data

Fasta file

Mapping

BAM file

Peak detection

Visualization

BED/WIG files

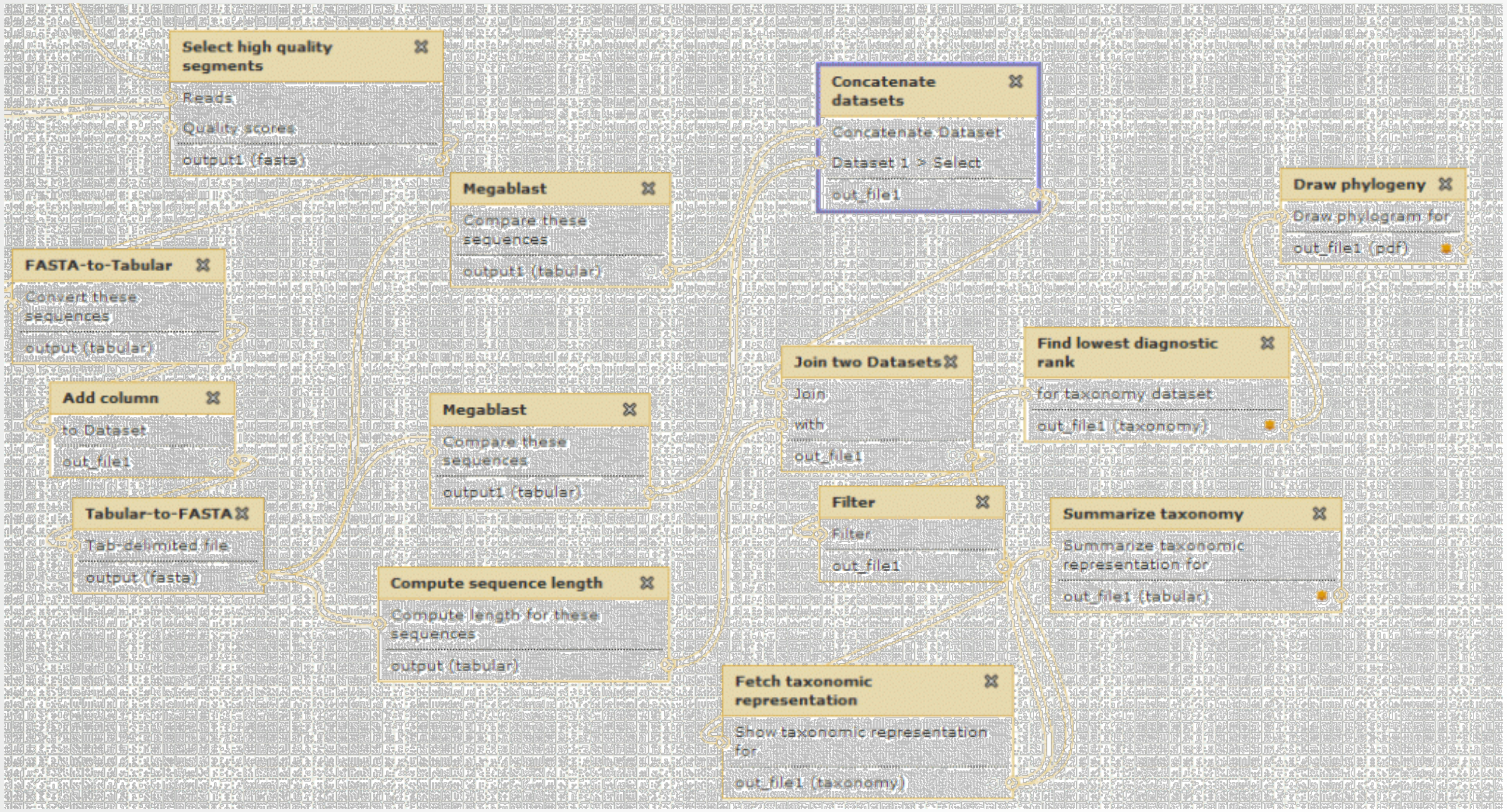
Clustering

Annotation

Motif discovery

Galaxy
PROJECT

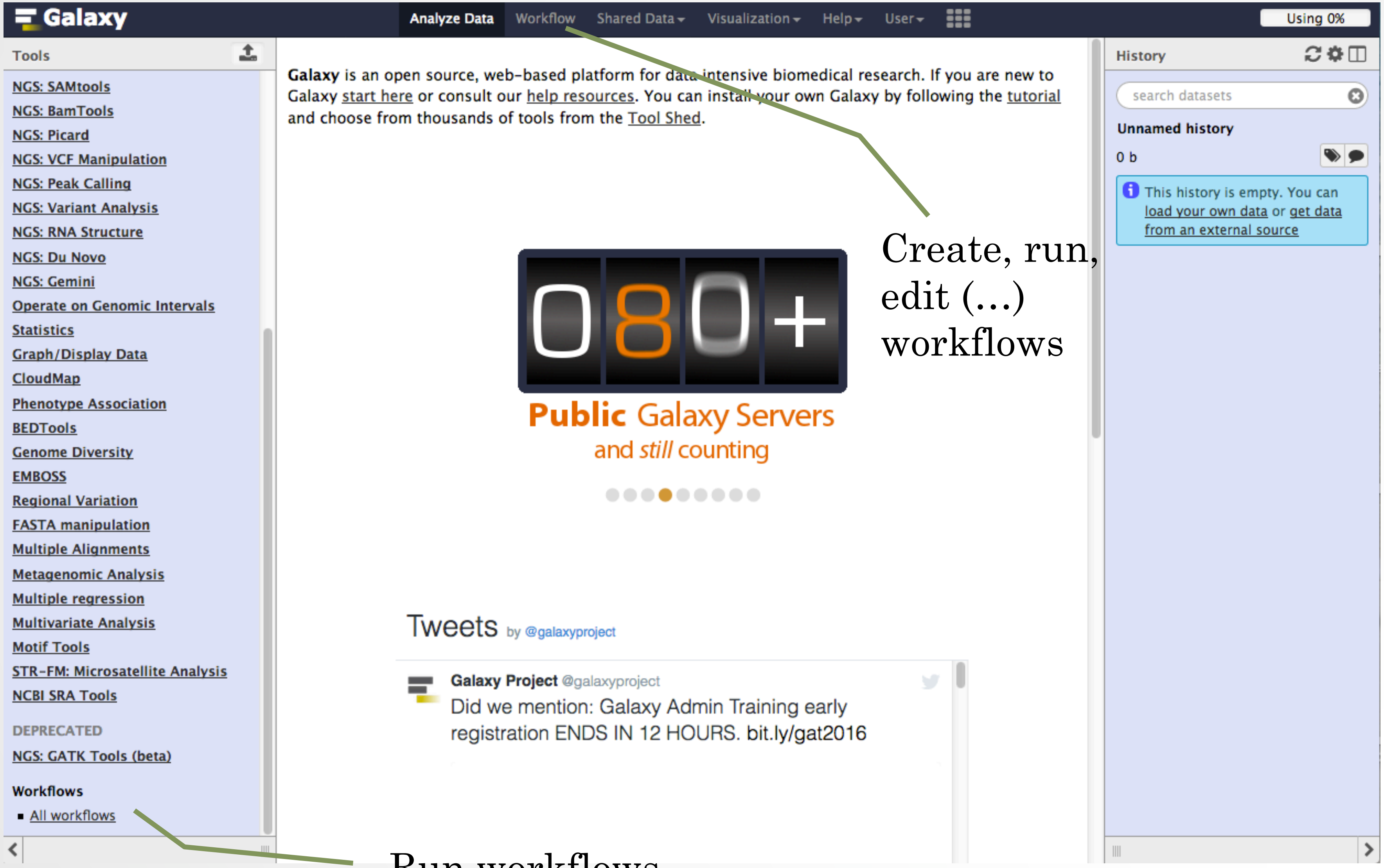
Galaxy workflow



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



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.

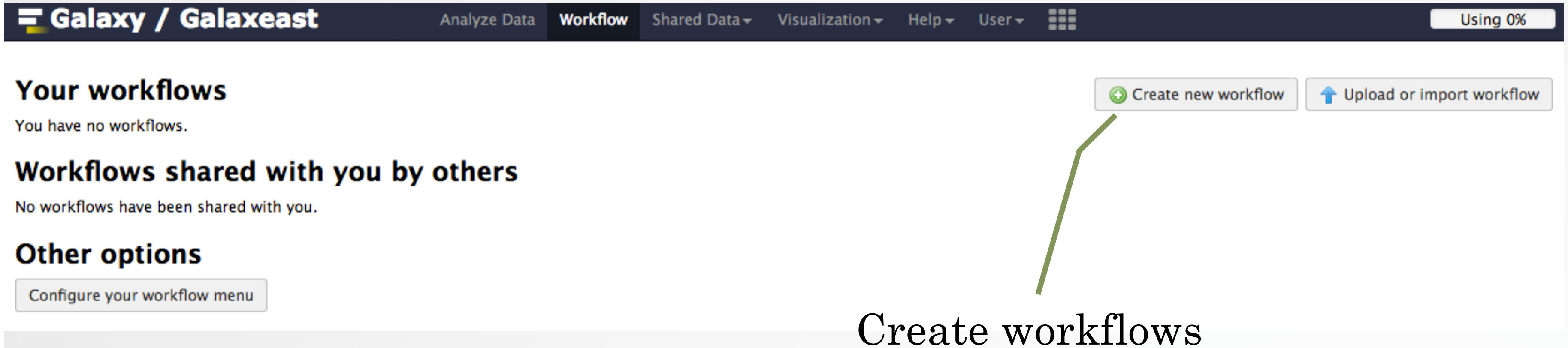
This history is empty. You can load your own data or get data from an external source

Did we mention: Galaxy Admin Training early registration ENDS IN 12 HOURS. bit.ly/gat2016

Create, run, edit (…) workflows

Run workflows

Workflows



Your workflows

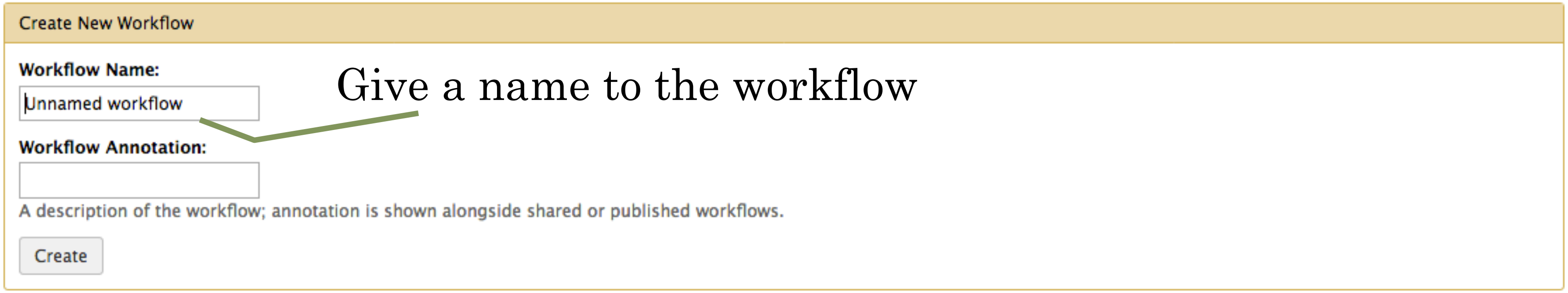
You have no workflows.

Workflows shared with you by others

No workflows have been shared with you.

Other options

Create workflows


Create New Workflow

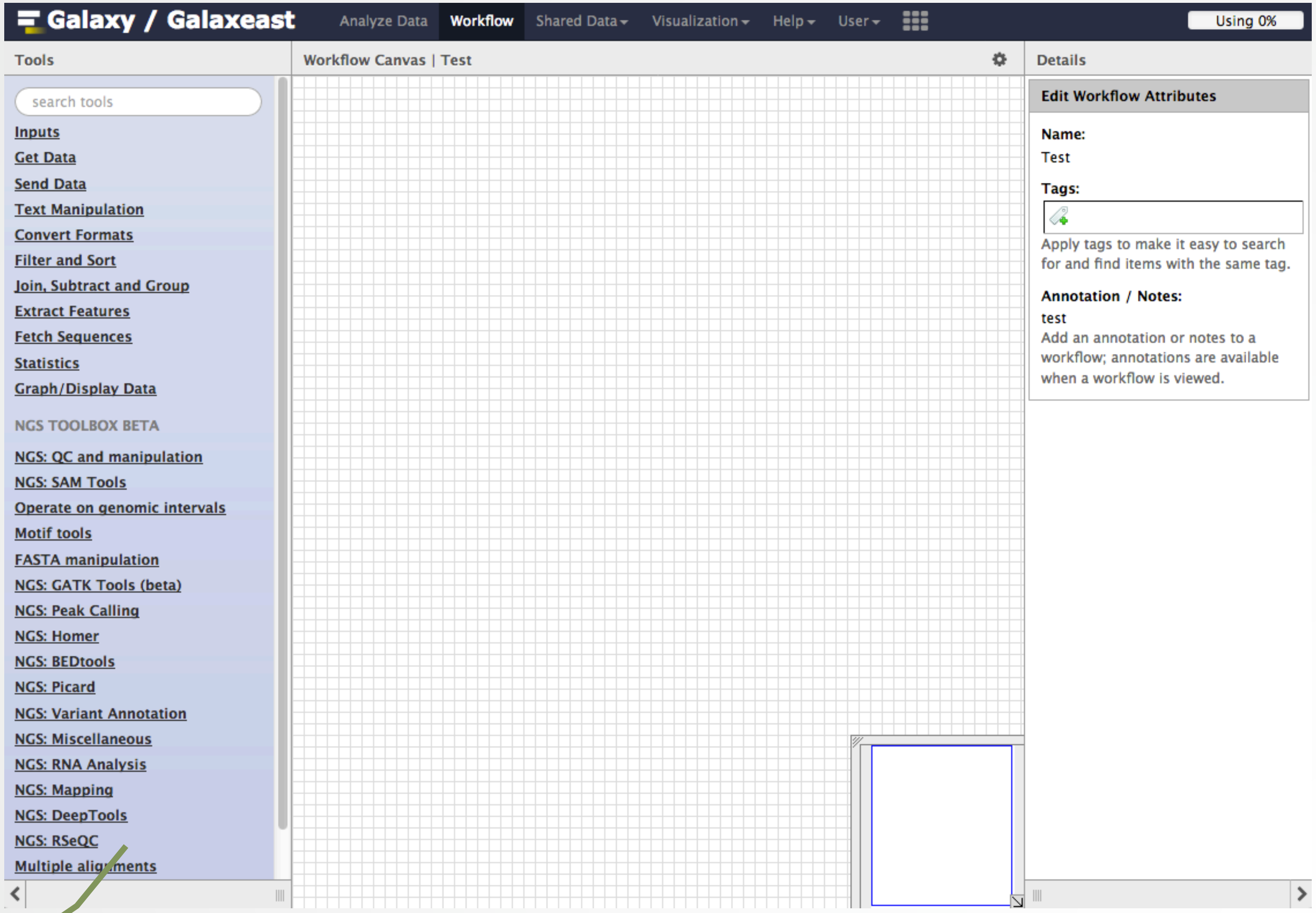
Workflow Name: Unnamed workflow

Workflow Annotation:

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

Give a name to the workflow

Workflow creation



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

Edit Workflow Attributes

Name:
Test

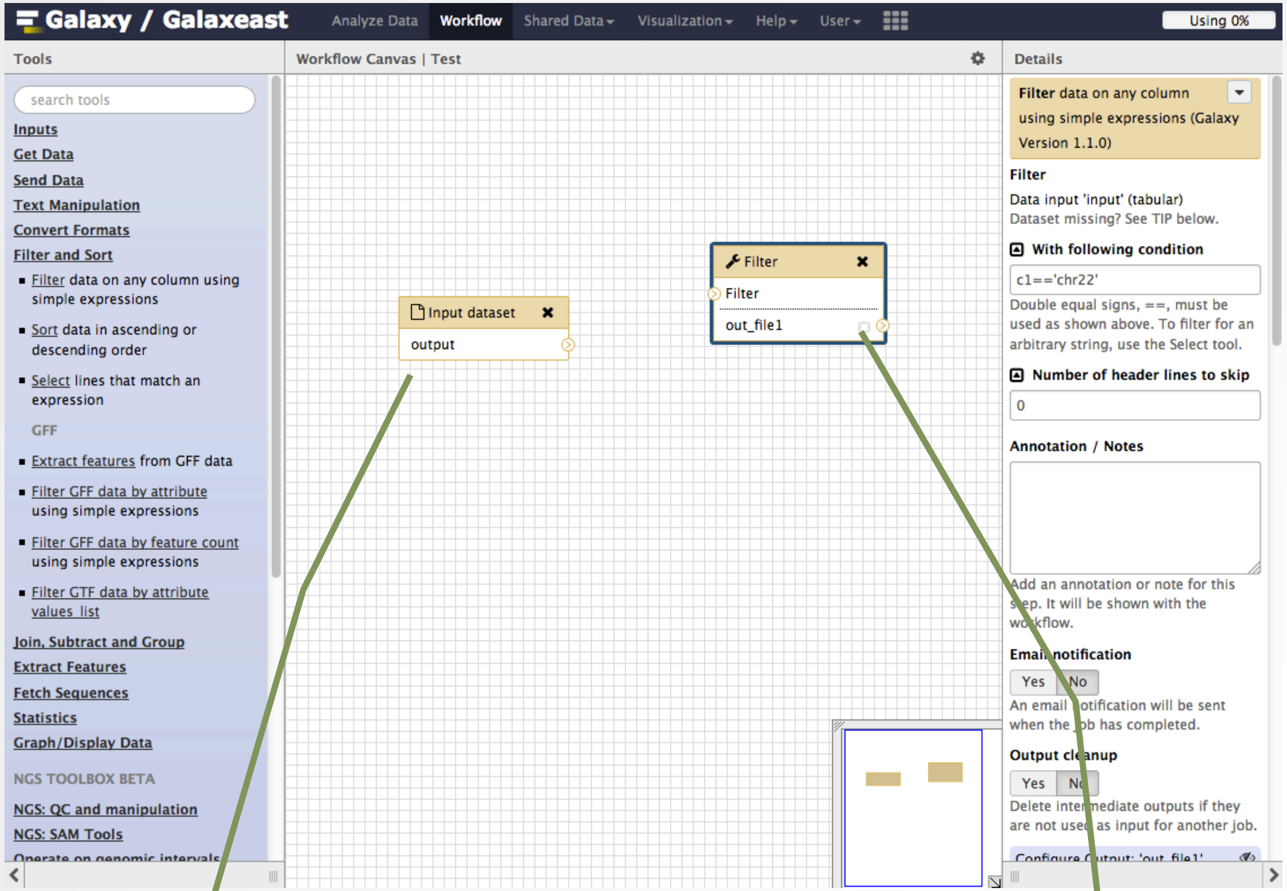
Tags:

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



Filter data on any column using simple expressions (Galaxy Version 1.1.0)

Filter

Data input 'input' (tabular)
Dataset missing? See TIP below.

With following condition

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Annotation / Notes

Add an annotation or note for this step. It will be shown with the workflow.

Email notification

Yes No

An email notification will be sent when the job has completed.

Output cleanup

Yes No

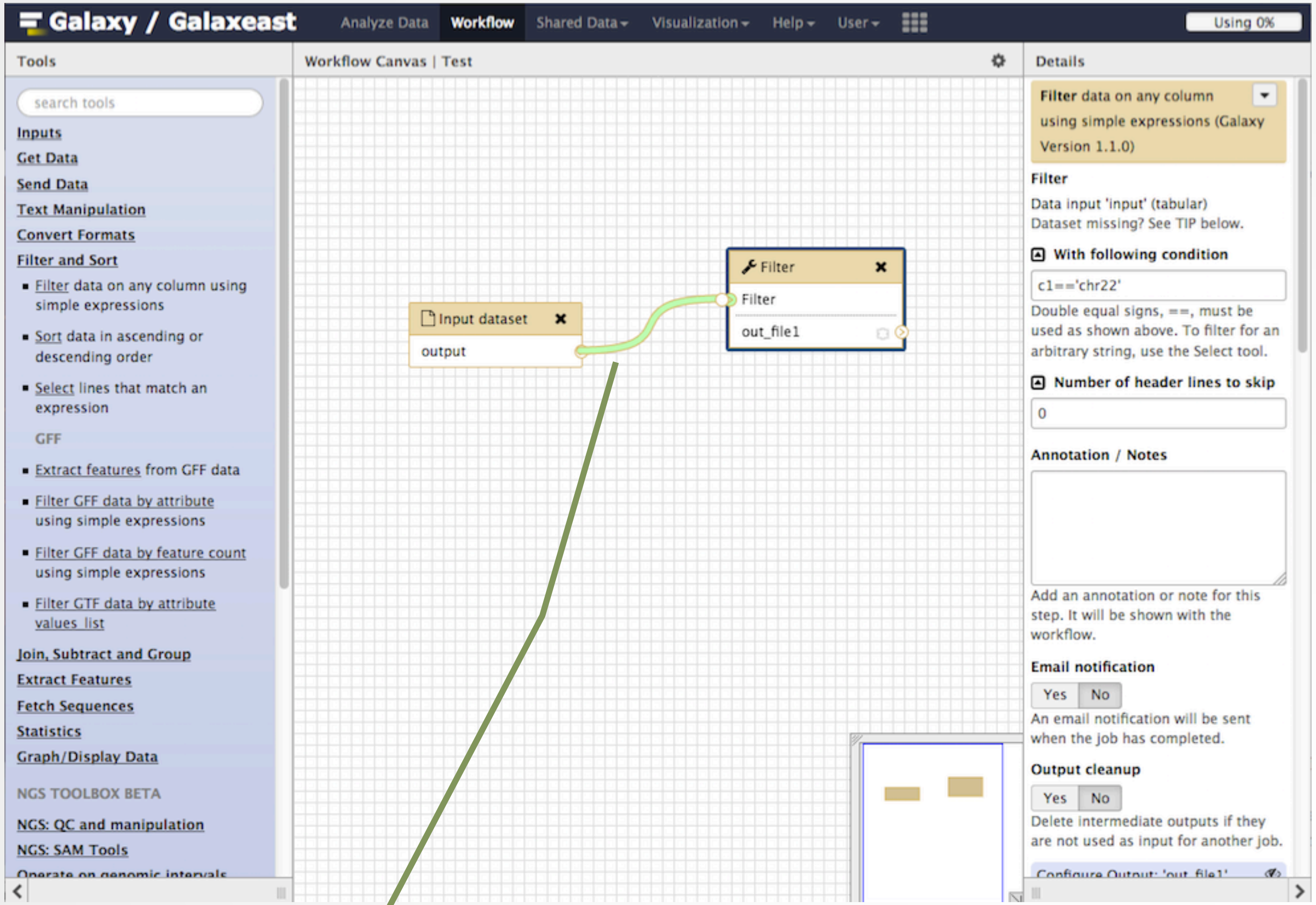
Delete intermediate outputs if they are not used as input for another job.

Configure Output: 'out_file1'

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



Filter data on any column using simple expressions (Galaxy Version 1.1.0)

Filter

Data input 'input' (tabular)
Dataset missing? See TIP below.

With following condition

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Annotation / Notes

Add an annotation or note for this step. It will be shown with the workflow.

Email notification

Yes No

An email notification will be sent when the job has completed.

Output cleanup

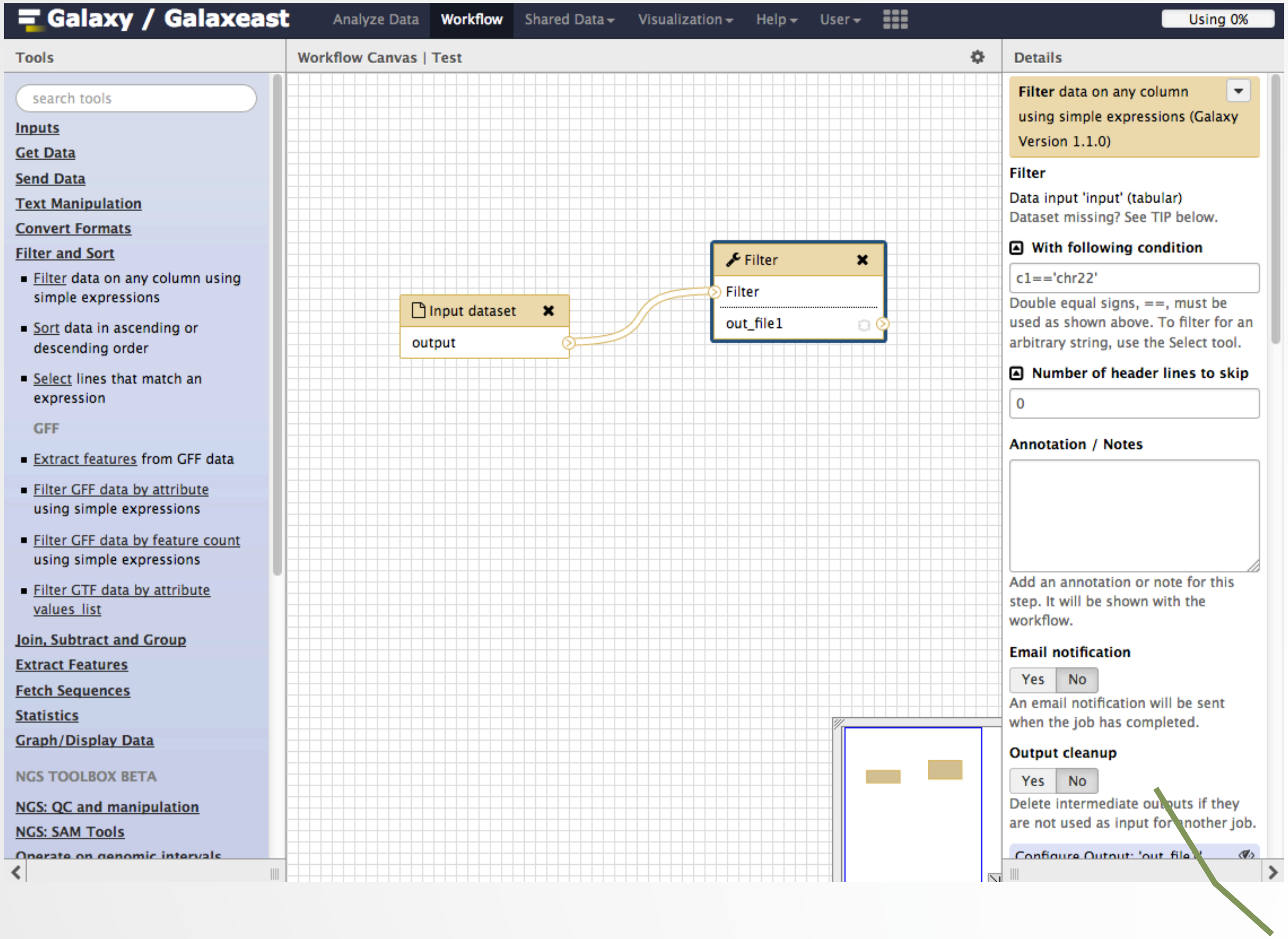
Yes No

Delete intermediate outputs if they are not used as input for another job.

Configure Output: 'out_file1'

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

Workflow creation



Filter data on any column using simple expressions (Galaxy Version 1.1.0)

Filter

Data input 'input' (tabular)
Dataset missing? See TIP below.

With following condition

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Annotation / Notes

Add an annotation or note for this step. It will be shown with the workflow.

Email notification

Yes No

An email notification will be sent when the job has completed.

Output cleanup

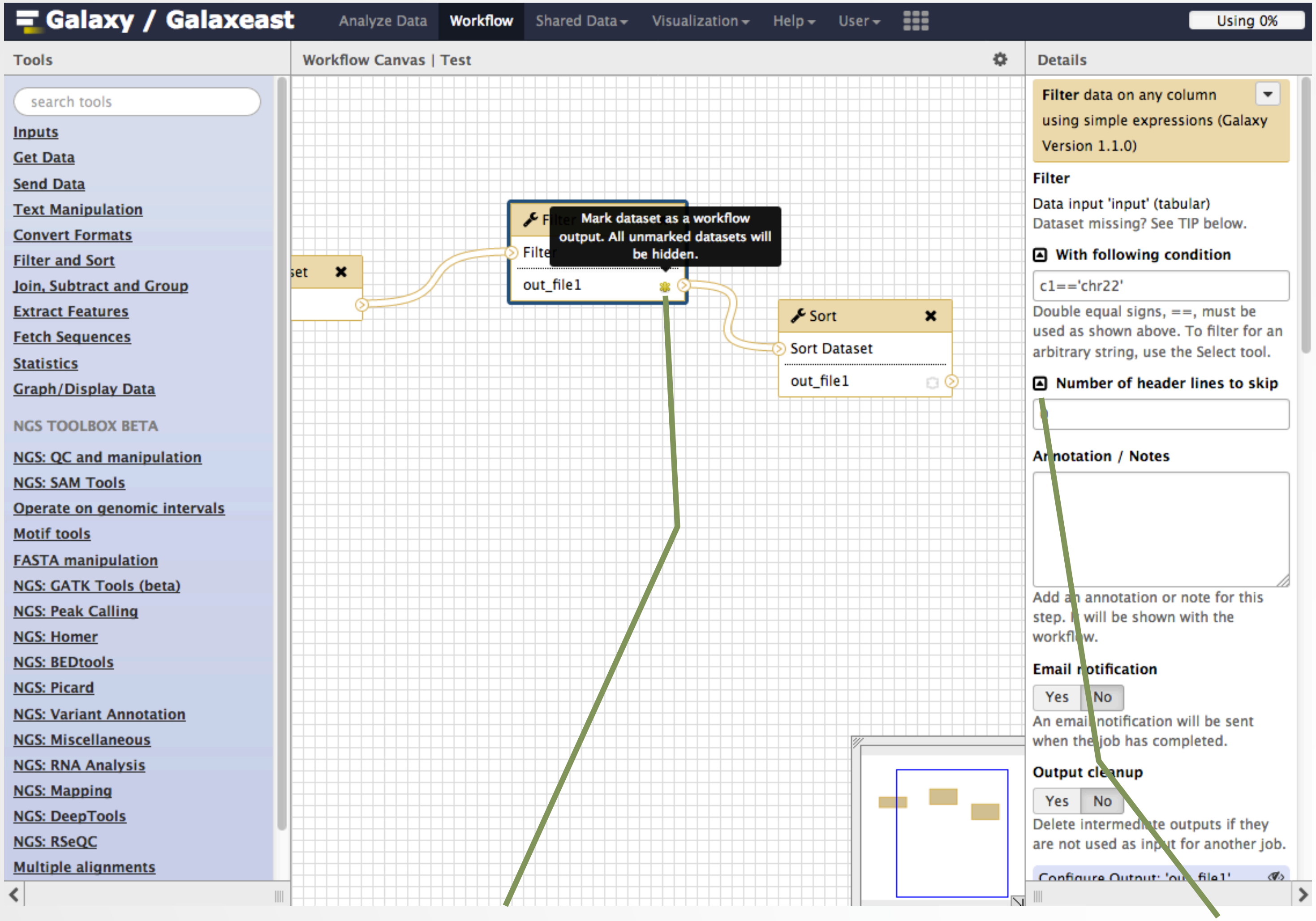
Yes No

Delete intermediate outputs if they are not used as input for another job.

Configure Output: 'out_file1'

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

Workflow creation



Mark dataset as a workflow output. All unmarked datasets will be hidden.

Filter data on any column using simple expressions (Galaxy Version 1.1.0)

Filter

Data input 'input' (tabular)
Dataset missing? See TIP below.

With following condition

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Annotation / Notes

Add an annotation or note for this step. It will be shown with the workflow.

Email notification

Yes No

An email notification will be sent when the job has completed.

Output cleanup

Yes No

Delete intermediate outputs if they are not used as input for another job.

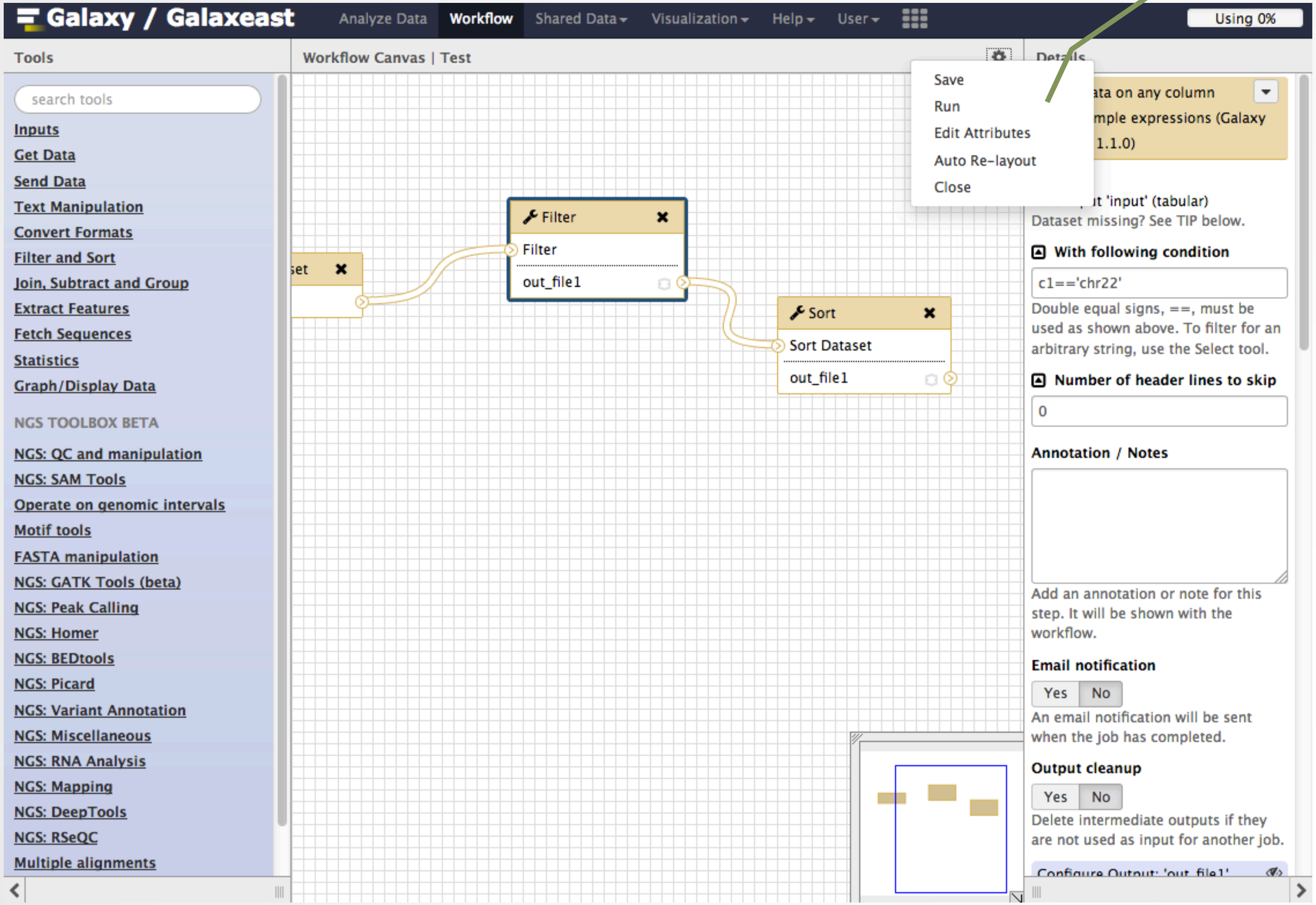
Configure Output: 'out_file1'

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



ata on any column
mple expressions (Galaxy
1.1.0)

t 'input' (tabular)
Dataset missing? See TIP below.

With following condition

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Annotation / Notes

Add an annotation or note for this step. It will be shown with the workflow.

Email notification

Yes No

An email notification will be sent when the job has completed.

Output cleanup

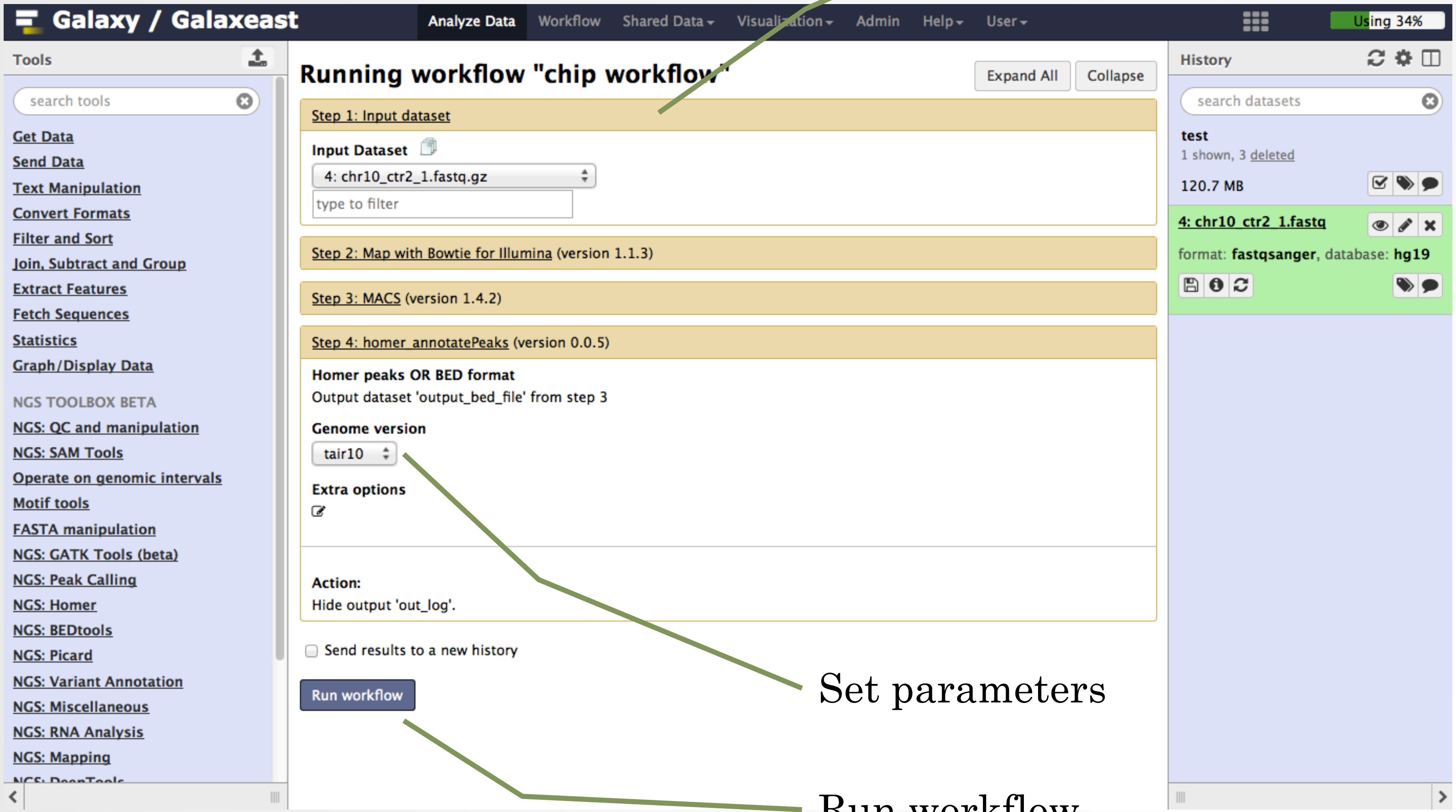
Yes No

Delete intermediate outputs if they are not used as input for another job.

Configure Output: 'out_file1'

Run workflows

Set input file(s)



Running workflow "chip workflow"

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

Action:
Hide output 'out_log'.

☐ Send results to a new history

test
1 shown, 3 deleted

120.7 MB

4: chr10_ctr2_1.fastq

format: **fastqsanger**, database: **hg19**

Set parameters

Run workflow

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 the datasets (chr10_ctr2_1.fastq and chr10_mitf_2.fastq) from the data library CNRS 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?