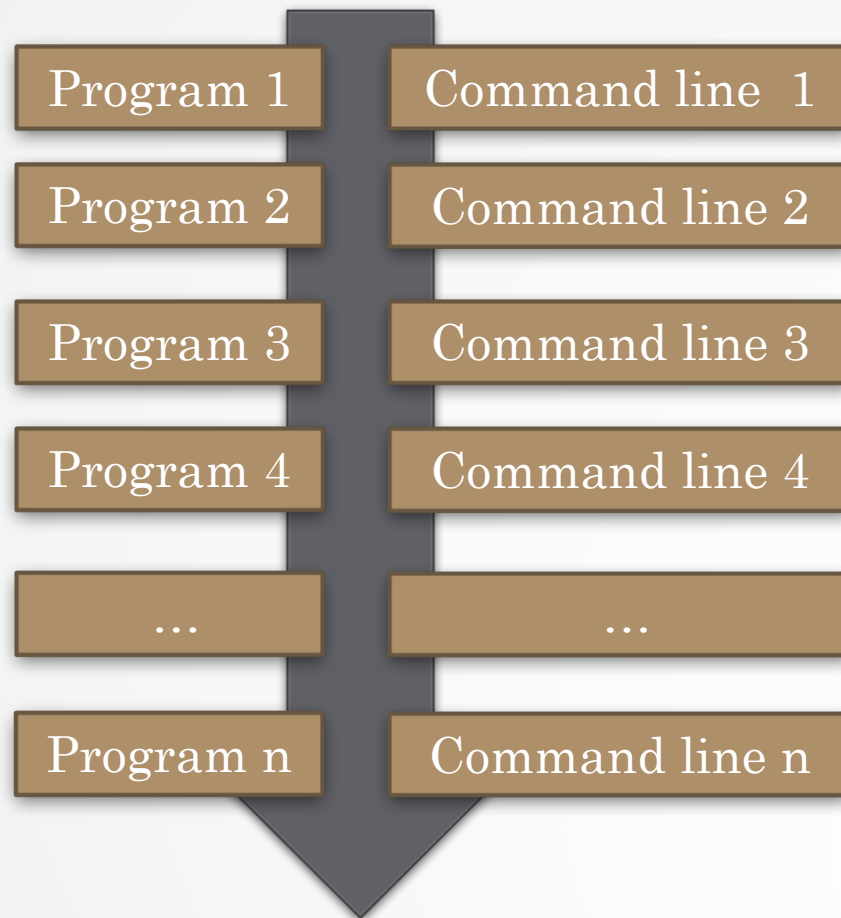
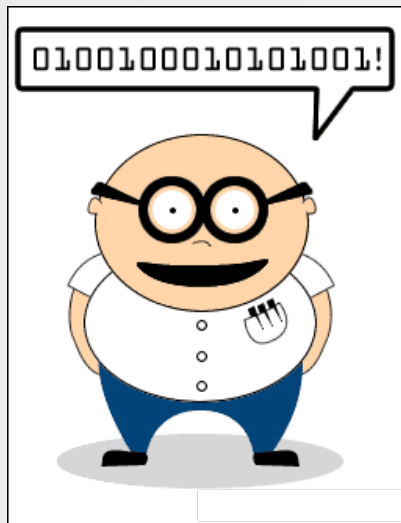


NGS analysis automatization: Galaxy workflows

Stéphanie Le Gras
(slegras@igbmc.fr)

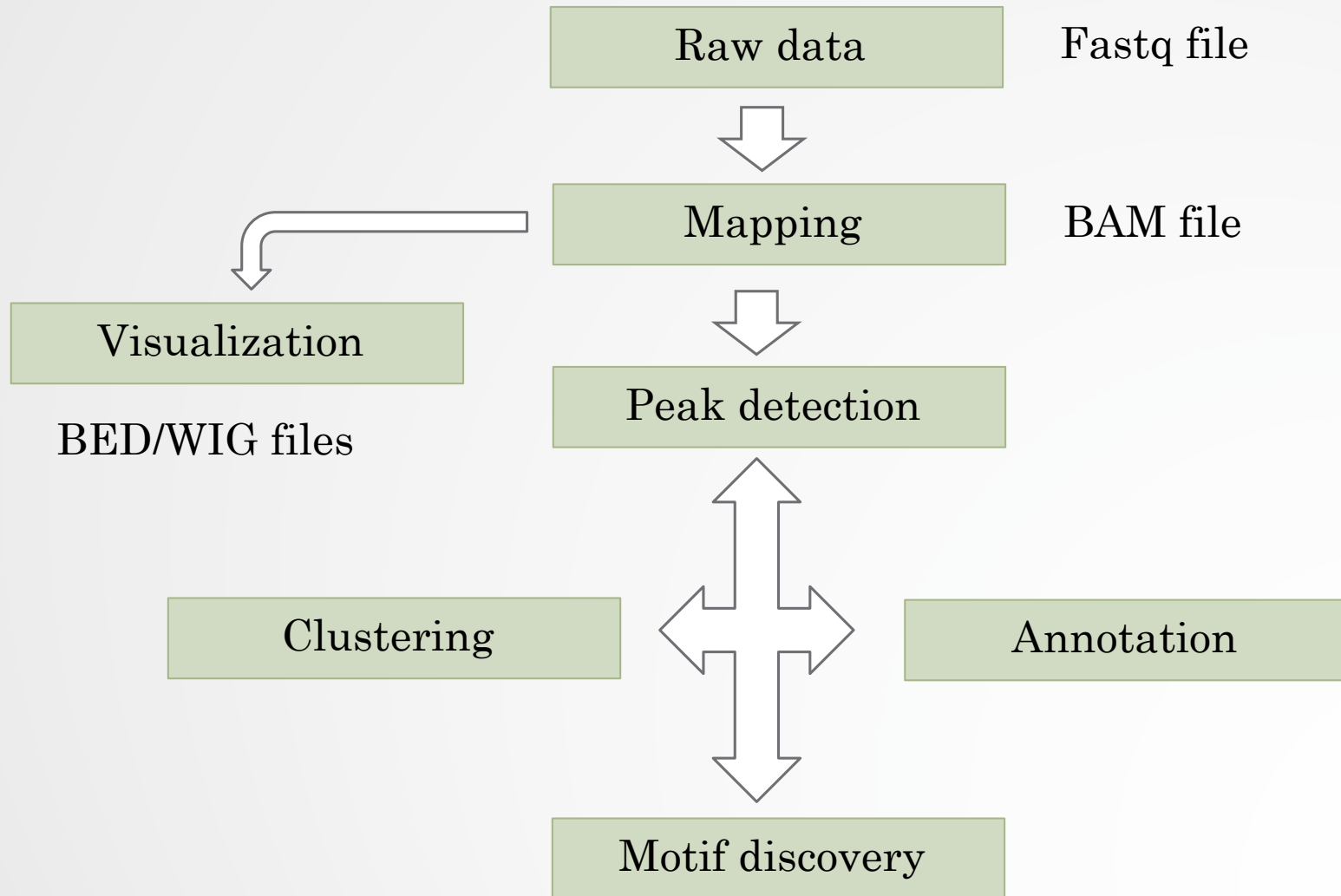
A long time ago...

Input data



**PIPELINE/
WORKFLOW**

More recently...



During the entire training session..

The logo icon consists of three horizontal bars of varying lengths stacked vertically, with a yellow-to-white gradient bar at the bottom.

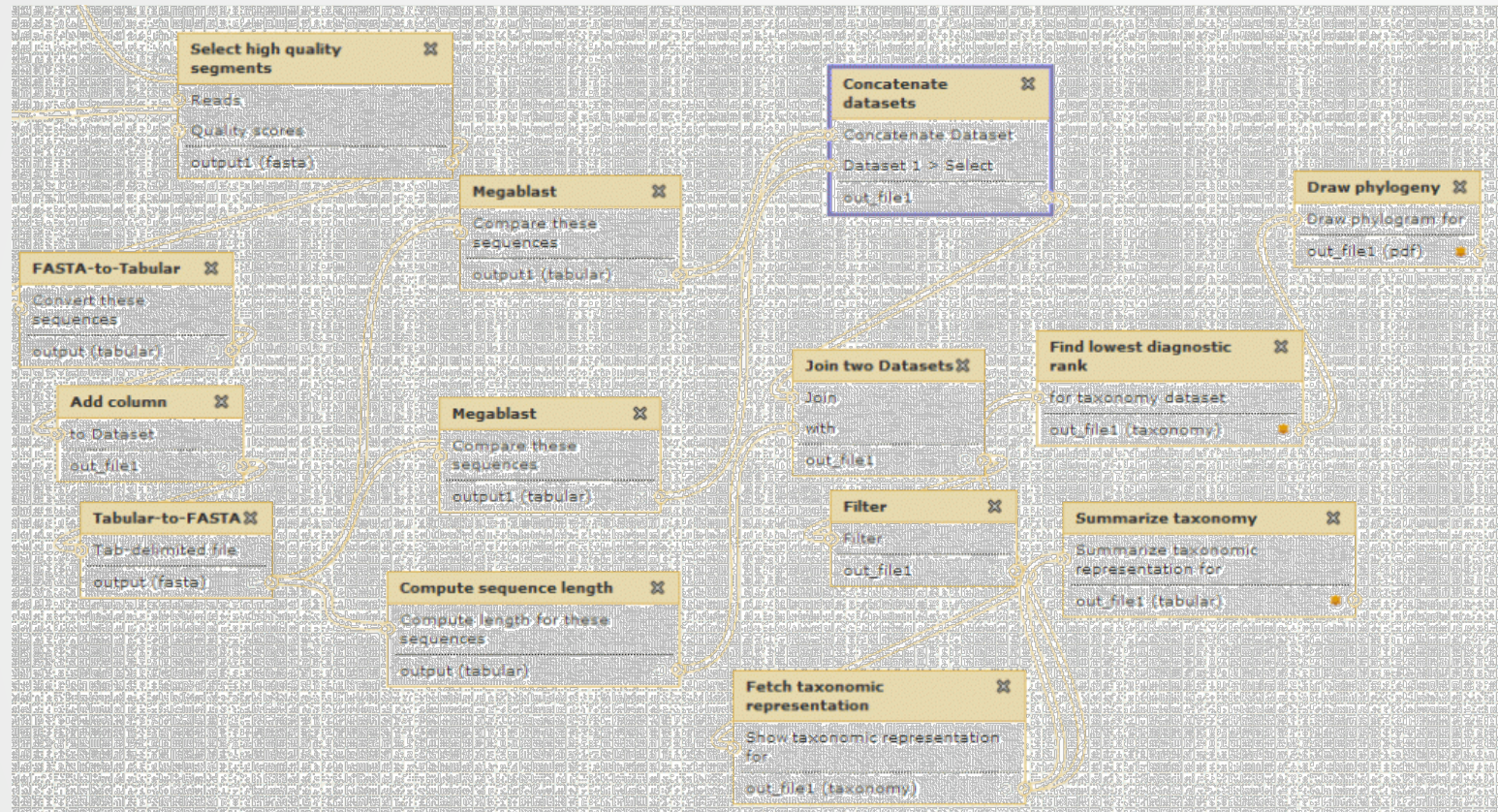
Galaxy

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

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](#).

080+
Public Galaxy Servers
and *still* counting

Tweets by @galaxyproject

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

History
search datasets
Unnamed history
0 b
This history is empty. You can [load your own data](#) or [get data from an external source](#)

Using 0%

Run workflows

Create, run,
edit (...) 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:

Unnamed workflow

Give a name to the workflow

Workflow Annotation:

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

Create

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, indicating a link. 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, number of header lines to skip, and output cleanup options.

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'.

Tools: A search bar and a list of tool categories are visible, including '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'.

Workflow Canvas | Test: A grid-based workspace showing a workflow. An 'Input dataset' tool (output: 'output') is connected to a 'Filter' tool (output: 'out_file1').

Details: The configuration panel for the 'Filter' tool. It includes the following options:

- Filter data on any column using simple expressions (Galaxy Version 1.1.0):** A dropdown menu.
- Filter:** Data input 'input' (tabular). Dataset missing? See TIP below.
- With following condition:** A text input field containing 'c1=='chr22''.
- Number of header lines to skip:** A text input field containing '0'.
- Annotation / Notes:** A text area for adding notes.
- Email notification:** Radio buttons for 'Yes' and 'No'.
- Output cleanup:** Radio buttons for 'Yes' and 'No'.

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

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 left sidebar lists various tool categories such as 'Inputs', 'Text Manipulation', 'Filter and Sort', and 'NGS TOOLBOX BETA'. The central 'Workflow Canvas | Test' area shows a workflow with a 'Filter' tool and a 'Sort' tool. A tooltip for the 'Filter' tool states: 'Mark dataset as a workflow output. All unmarked datasets will be hidden.' A green arrow points to a star icon on the 'Filter' tool. Another green arrow points to a gear icon on the 'Sort' tool. The right sidebar shows the configuration for the 'Filter' tool, including a search bar, a dropdown menu, and various options like 'With following condition' and 'Number of header lines to skip'.

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

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

Filter

out_file1

Sort

Sort Dataset

out_file1

Save

Run

Edit Attributes

Auto Re-layout

Close

Details

Filter

Filter on any column

Simple expressions (Galaxy 1.1.0)

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)

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: [edit icon]
 - 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

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?