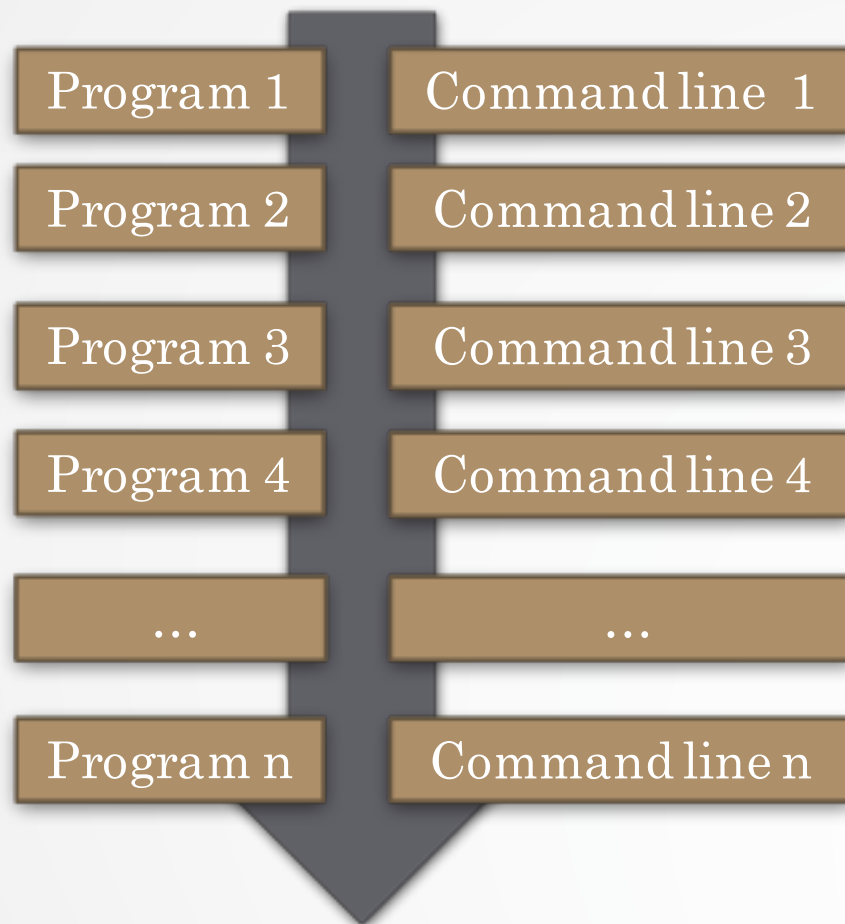
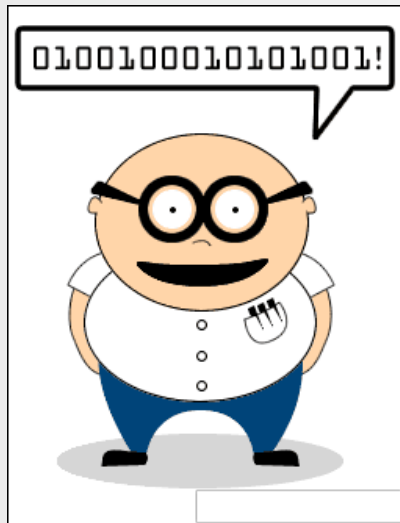


NGS analysis automatization: Galaxy workflows

Stéphanie Le Gras
(slegras@igbmc.fr)

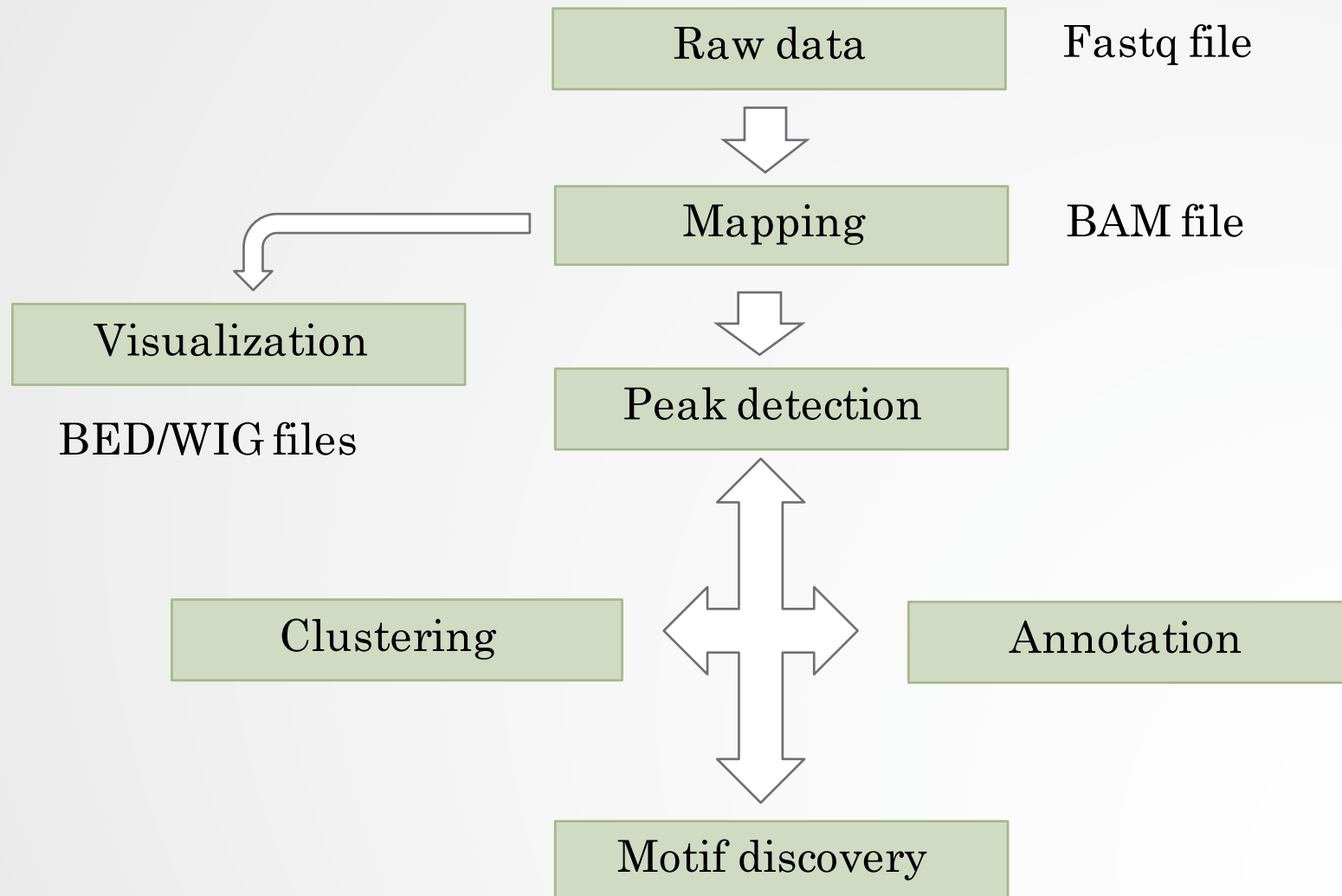
A long time ago...

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.

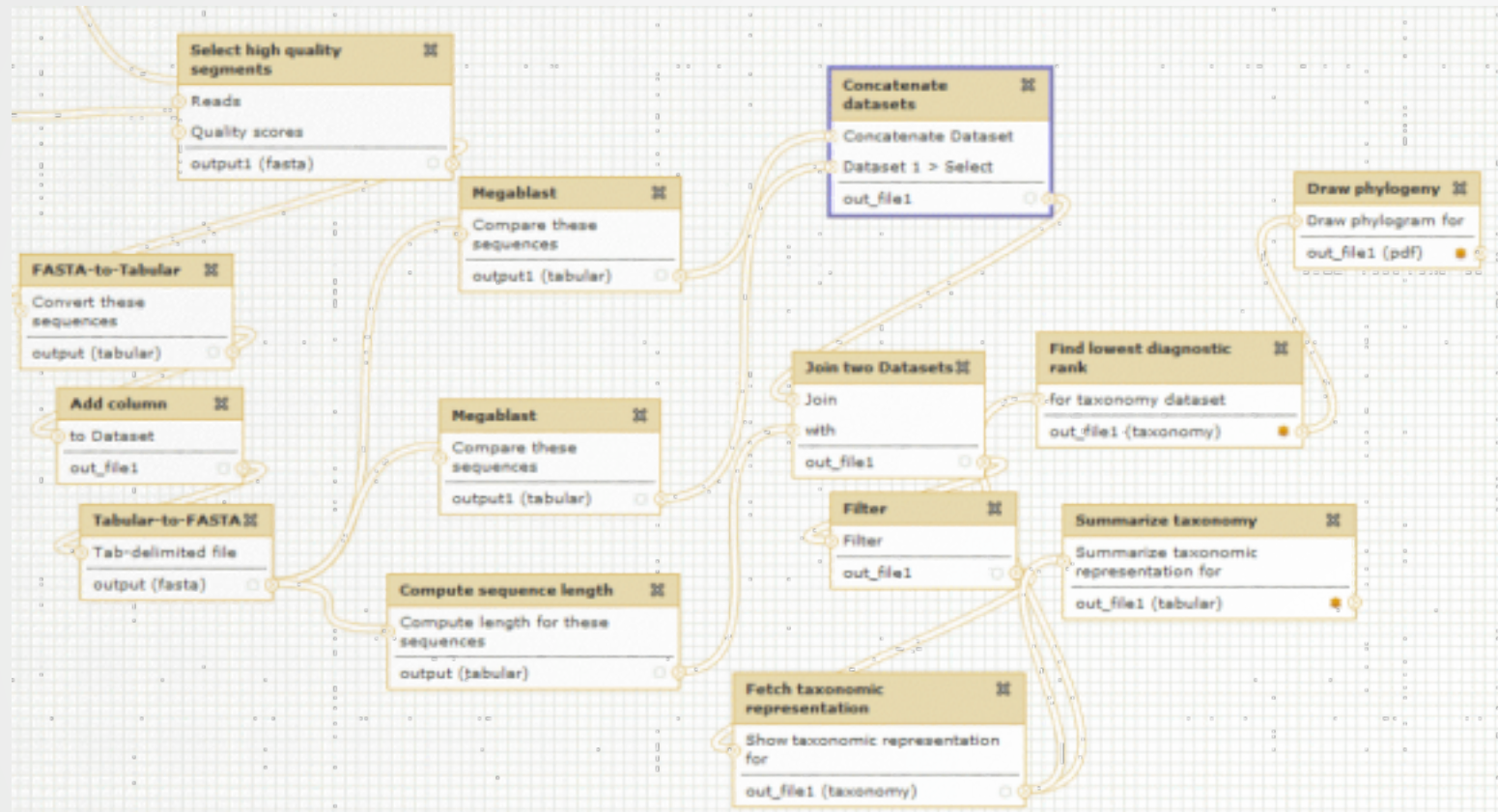
Galaxy

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

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 like NGS: SAMtools, NGS: BamTools, NGS: Picard, etc. 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 'Public Galaxy Servers and still counting' banner with a '080+' logo. A 'Tweets' section shows a tweet from the Galaxy Project: 'Did we mention: Galaxy Admin Training early registration ENDS IN 12 HOURS. bit.ly/gat2016'. On the right, a 'History' panel shows 'Unnamed history' with '0 b' 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:

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

The screenshot displays the Galaxy/Galaxeast interface. On the left, a 'Tools' sidebar lists various categories like 'Inputs', 'Text Manipulation', and 'Filter and Sort'. The 'Filter and Sort' section is expanded, showing options for filtering data on any column using simple expressions. The main 'Workflow Canvas' shows two tools: 'Input dataset' and 'Filter'. The 'Filter' tool is selected, and its configuration panel is open on the right. The configuration panel includes a dropdown for 'Filter data on any column using simple expressions (Galaxy Version 1.1.0)', a 'Filter' section with a condition 'c1=='chr22'', a 'Number of header lines to skip' field set to 0, and sections for 'Email notification' and 'Output cleanup'. A green line points from the text 'Input dataset.' to the 'Input dataset' tool, and another green line points from the text 'Tool to be run' to the 'Filter' tool.

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 / Galaxeast interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main area is divided into three panels: 'Tools', 'Workflow Canvas | Test', and 'Details'. The 'Tools' panel on the left lists various categories like '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', and 'Operate on genomic intervals'. The 'Workflow Canvas' shows two steps: 'Input dataset' (output: 'output') and 'Filter' (output: 'out_file1'). A green line connects the 'output' of the 'Input dataset' step to the 'Filter' step. The 'Details' panel on the right shows the configuration for the 'Filter' step, including the condition 'c1=='chr22'' and the output name 'out_file1'. The 'Filter' step is also configured with 'With following condition', 'Number of header lines to skip' (0), 'Email notification' (Yes/No), and 'Output cleanup' (Yes/No).

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 workspace 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', and 'Graph/Display Data'. The 'Filter and Sort' category is expanded, showing various filtering options.

Workflow Canvas | Test: A workflow is being constructed on a grid. It starts with an 'Input dataset' tool (output: 'output') connected to a 'Filter' tool (output: 'out_file1').

Details: The configuration panel for the 'Filter' tool is shown. 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''. A note below states: '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:** A text input field containing '0'.
- Annotation / Notes:** A text area for adding notes.
- Email notification:** Radio buttons for 'Yes' and 'No'. A note states: 'An email notification will be sent when the job has completed.'
- Output cleanup:** Radio buttons for 'Yes' and 'No'. A note states: 'Delete intermediate outputs if they are not used as input for another job.'

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

Workflow creation

The screenshot displays the Galaxy workflow editor. The central 'Workflow Canvas' shows a workflow with two tools: 'Filter' and 'Sort'. A tooltip over the 'Filter' tool reads: 'Mark dataset as a workflow output. All unmarked datasets will be hidden.' A callout box points to a star icon on the 'Filter' tool, explaining its function. The right sidebar shows the configuration for the 'Filter' tool, including a condition 'c1=='chr22'' and options for email notification and output cleanup.

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

Set

Filter

Sort

Filter

out_file1

Sort Dataset

out_file1

Save

Run

Edit Attributes

Auto Re-layout

Close

Filter on any column

With following condition

c1==chr22

Number of header lines to skip

0

Annotation / Notes

Email notification

Output cleanup

Configure Output: 'out_file1'

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

NGS: DeepTools

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

16

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?