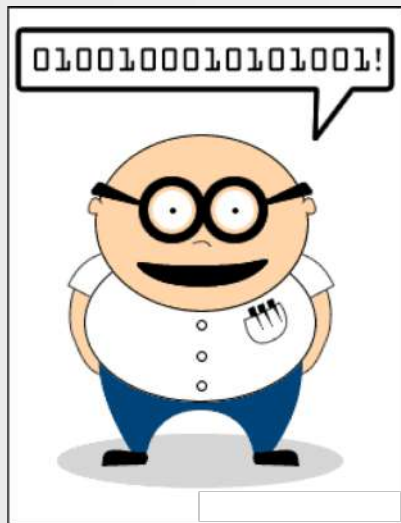


# NGS analysis automatization: Galaxy workflows

Stéphanie Le Gras  
([slegras@igbmc.fr](mailto:slegras@igbmc.fr))

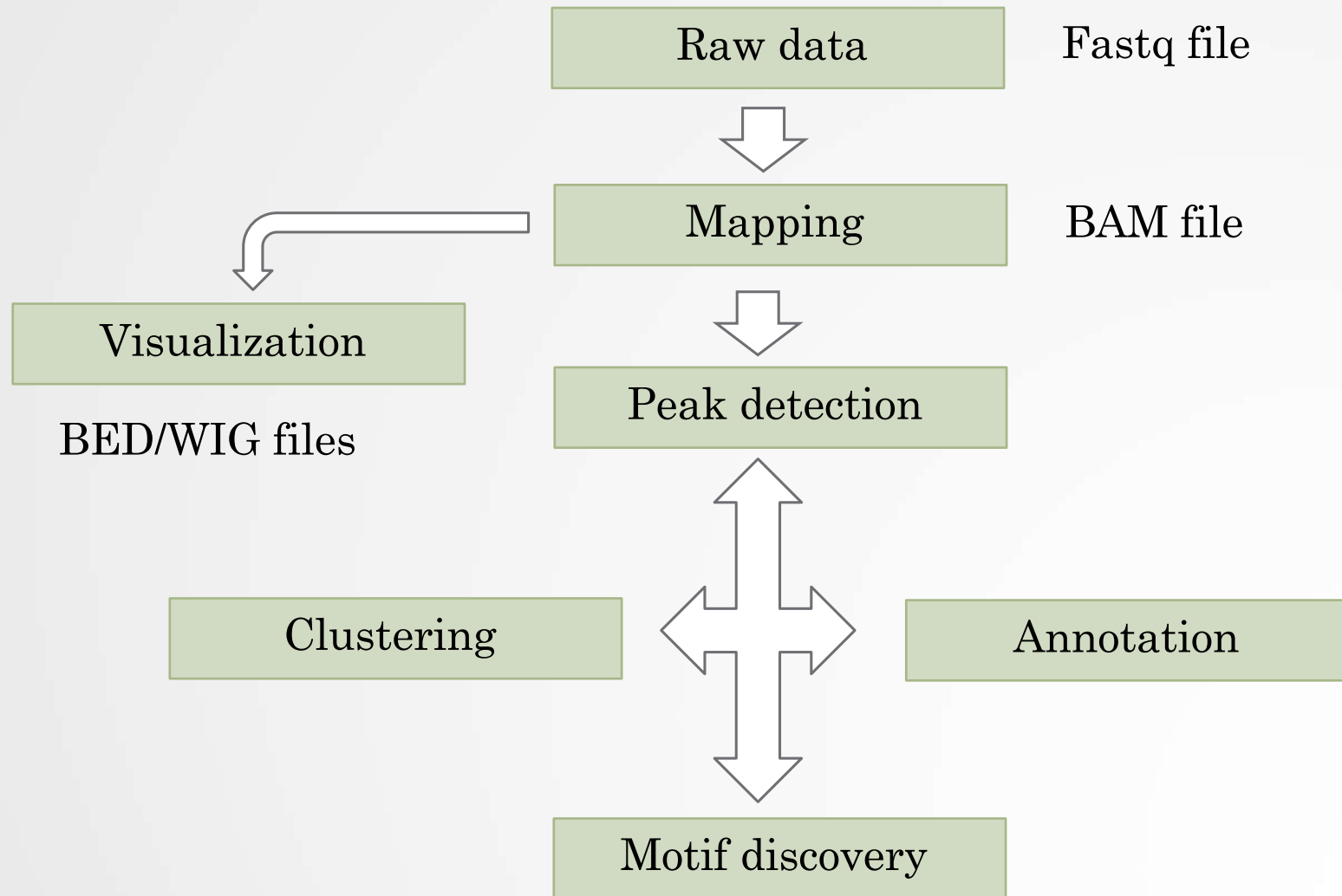
# A long time ago...

Input data



**PIPELINE/  
WORKFLOW**

# More recently...



During the entire training session..



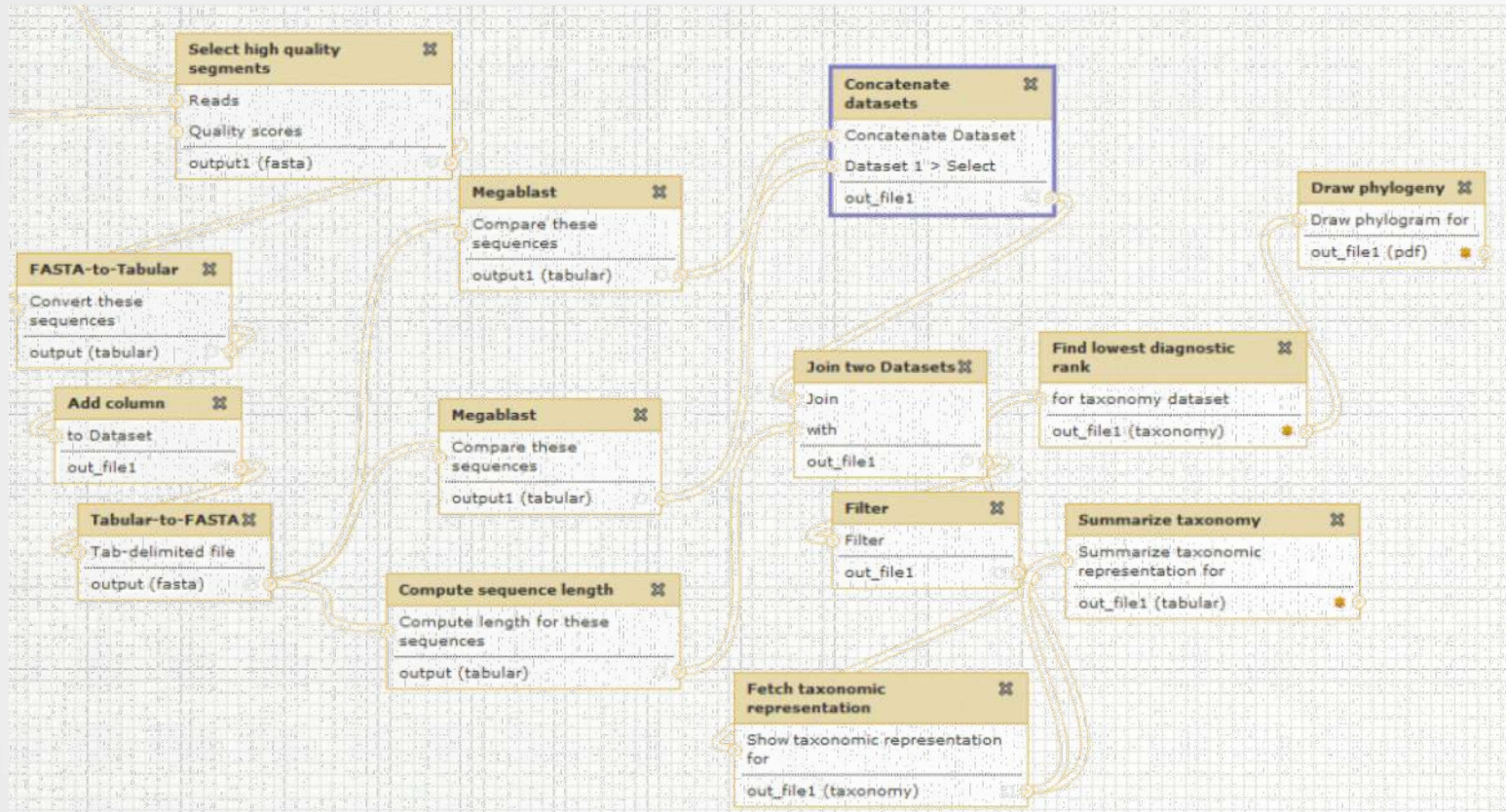
# Galaxy

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. A tweet from Galaxy Project @galaxyproject reads: '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'.

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](http://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](#)

Tools  
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)  
Workflows  
All 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

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 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', 'Text Manipulation', 'Filter and Sort', and 'NGS TOOLBOX BETA'. The 'Workflow Canvas' shows a workflow with two steps: 'Input dataset' and 'Filter'. The 'Filter' tool is highlighted with a green box, and a green line points from the text 'Tool to be run' to it. The 'Details' panel on the right shows the configuration for the 'Filter' tool, including the condition 'c1==chr22' and the output file '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

The screenshot displays the Galaxy / Galaxeast interface for creating a workflow. The main area is the 'Workflow Canvas | Test', which shows a grid with two steps: 'Input dataset' and 'Filter'. A green link 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 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', and 'User'. The main area is divided into three panels: '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', and 'Graph/Display Data'. A search bar is located at the top of this panel.

The 'Workflow Canvas | Test' panel shows a workflow graph on a grid. It consists of two tools: 'Input dataset' (output: 'output') and 'Filter' (output: 'out\_file1'). A green arrow points from the 'Filter' tool in the canvas to the 'Details' panel.

The 'Details' panel for the 'Filter' tool shows the following configuration options:

- 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'
- 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.

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'), 'Number of header lines to skip', 'Annotation / Notes', 'Email notification' (Yes/No), and 'Output cleanup' (Yes/No). A green arrow points from the 'Annotation / Notes' section to a small preview window at the bottom right.

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

Workflow Canvas | Test

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

Set

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 is titled "Running workflow 'chip workflow'" and contains four steps:

- Step 1: Input dataset**: Shows an "Input Dataset" dropdown menu with the selected file "4: chr10\_ctr2\_1.fastq.gz". Below it is a search box labeled "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)**: Shows "Homer peaks OR BED format" and "Output dataset 'output\_bed\_file' from step 3". It includes a "Genome version" dropdown set to "tair10", an "Extra options" field with a checkmark icon, and an "Action:" section with the text "Hide output 'out\_log'".

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

Annotations with green lines point to the "Input Dataset" dropdown (labeled "Set input file(s)"), the "Genome version" dropdown (labeled "Set parameters"), and the "Run workflow" button (labeled "Run workflow").

The left sidebar contains a "Tools" section with a search box and various tool categories like "Get Data", "Text Manipulation", "NGS TOOLBOX BETA", etc. The right sidebar shows a "History" section with a search box and a list of datasets, including the current one: "4: chr10\_ctr2\_1.fastq" in format "fastqsanger" on 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 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?