

IBT 432 Aplikasi Bioinformatika

Analisis komparasi genomika I: Pengenalan terhadap Galaxy

Riza Arief Putranto

Rencana Perkuliahan

- ~~1. Kontrak belajar dan pengenalan bioinformatika aplikatif~~
- ~~2. Database sekuen dan analisis genomika~~
- ~~3. Anotasi sekuen ke genom – Praktik~~
4. Analisis komparasi genomika I
5. Analisis komparasi genomika II
6. Analisis komparasi genomika III
7. Analisis komparasi genomika – Praktik
8. Protein modelling I
9. Protein modelling II
10. Protein modelling III
11. Protein modelling - Praktik
12. Visualisasi protein modelling
13. Visualisasi protein modelling - Praktik
14. Presentasi mahasiswa

What is Galaxy?

Galaxy

- ❑ **Web-based platform** for computational biomedical research
 - ✓ Developed at Penn State, Johns Hopkins and G. Washington universities with substantial outside contribution
 - ✓ Open source under Academic Free License
- ❑ More than **4,900 citations**
- ❑ More than **80 public Galaxy servers**
 - ✓ Many more non-public
 - ✓ Both general-purpose and domain-specific

Core values

❑ Accessibility

- ✓ **Users without programming experience** can easily upload/retrieve data, run complex tools and workflows, and visualize data

❑ Reproducibility

- ✓ Galaxy **captures information** so that any user can understand and repeat a complete computational analysis

❑ Transparency

- ✓ Users can **share or publish** their analyses (histories, workflows, visualizations)
- ✓ Pages: online Methods for your paper

Pages: interactive, web-based documents that describe a complete analysis.

User interface

Main Galaxy interface

The screenshot displays the Galaxy web interface with three main panels:

- Tools Panel (Left):** Contains a search bar and a list of tool categories such as "Get Data", "Send Data", "Lift-Over", "Collection Operations", "Filter and Sort", "Join, Subtract and Group", "Fetch Alignments/Sequences", "NGS: OC and manipulation", "NGS: DeepTools", "NGS: Mapping", "NGS: RNA Analysis", "NGS: SAMtools", "NGS: BamTools", "NGS: Picard", "NGS: VCF Manipulation", "NGS: Peak Calling", "NGS: Variant Analysis", "NGS: RNA Structure", "NGS: Du Novo", and "NGS: Gemini". A blue box labeled "Tools" is overlaid on this panel.
- Main Panel (Center):** Features a purple box labeled "Main" and a black box with yellow text that reads: "Looking to learn? New comprehensive tutorials on: Diploid variant calling, Reference based RNAseq, Processing multiple samples, Introduction to NGS technologies, Galaxy 101, parts 1 & 2". Below this is a tweet from @galaxyproject about proteogenomics research, including a diagram of a workflow. A blue box labeled "Main" is overlaid on the top of this panel.
- History Panel (Right):** Shows a list of recent jobs with a green box labeled "History" overlaid on top. The jobs listed include "34: Cuffdiff on data 12, data 7, and data 23: transcript FPKM tracking", "33: Cuffdiff on data 12, data 7, and data 23: transcript differential expression testing", "32: Cuffdiff on data 12, data 7, and data 23: gene FPKM tracking", "31: Cuffdiff on data 12, data 7, and data 23: gene differential expression testing", and "30: Cuffdiff on data 12, data 7, and data 23: TSS groups FPKM tracking".

The browser address bar shows "https://usegalaxy.org" and the system tray at the bottom indicates the date "11-Dec-17".

Homepage divided into three panels

Top menu

Analyze Data Workflow Shared Data ▾ Visualization ▾ Cloud ▾ Help ▾ User ▾

- ❑ **Analyze Data** - go back to the 3-panels homepage
- ❑ **Workflow** - access existing workflows or create new one using the editable diagrammatic pipeline
- ❑ **Shared data** - access data libraries, histories, workflows, visualizations and pages shared with you
- ❑ **Visualization** - create new track browser and access your saved visualisations
- ❑ **Help** - links to Galaxy Biostar (Q&A), Galaxy Community Hub (Wiki), and Interactive Tours
- ❑ **User** – your preferences and saved histories, datasets, and pages

Tool interface

The screenshot displays the Galaxy web interface. At the top, the browser address bar shows 'https://usegalaxy.org'. The navigation bar includes 'Galaxy' and various menu items like 'Analyse de données', 'Workflow', and 'Utilisateur'. On the left, the 'Tools' panel is visible, with a search bar and a list of tools. The 'Get Data' section is highlighted with a red box, containing 'Upload File from your computer' and several other tools like 'UCSC Main table browser'. The main content area features a text block about Galaxy, a 'Looking to learn?' banner with tutorial links, and a tweet from the Galaxy Project. On the right, the 'History' panel shows a list of recent jobs, including 'Workshop HPC NGS' and several 'Cuffdiff on data' jobs. The bottom of the image shows the Windows taskbar with various application icons and the system tray.

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](#).

Looking to learn?
New comprehensive tutorials on:
Diploid variant calling
Reference based RNAseq
Processing multiple samples
Introduction to NGS technologies
Galaxy 101, parts 1 & 2

Tweets by @galaxyproject

Galaxy Project Retweeted
NCIP @NCL_NCIP
Find out how to use the Galaxy #bioinformatics platform to create a customizable database for your #proteogenomics research in @AACR's special open-access issue: bit.ly/2BqZwLI

History

Rechercher des données

Workshop HPC NGS
31 shown, 25 deleted, 2 hidden
276.65 MB

35: Cuffdiff on data 12, data 7, and data 23: cummeRbund SQLite database

34: Cuffdiff on data 12, data 7, and data 23: transcript FPKM tracking

33: Cuffdiff on data 12, data 7, and data 23: transcript differential expression testing

32: Cuffdiff on data 12, data 7, and data 23: gene FPKM tracking

31: Cuffdiff on data 12, data 7, and data 23: gene differential expression testing

30: Cuffdiff on data 12, data 7, and data 23: TSS groups FPKM tracking

Tools

- ❑ **Each tool is a text file describing:**
 - ✓ input datasets, parameters, commands, and outputs
 - ✓ help, tests, citations, dependency requirements
- ❑ **Free tool store: Galaxy Tool Shed**
 - ✓ thousands of tools already available
 - ✓ every software can be embedded
 - ✓ if a tool is not available, ask the Galaxy community for help!
 - ✓ only a Galaxy admin can install tools
- ❑ **New versions can be installed without removing old ones to ensure**
 - ✓ reproducibility

History


❑ Location of all your analyses

- ✓ collects all **datasets** produced by tools you run
- ✓ collects all **operations** performed on your data

❑ At the heart of Galaxy's reproducibility

For **each dataset**, the history tracks:

- ✓ name, format, size, creation time, datatype-specific
- ✓ metadata
- ✓ tool id and version, inputs, parameters
- ✓ standard output (stdout) and error (stderr)
- ✓ state (waiting, running, success, failed)
- ✓ hidden, deleted, purged




The screenshot shows the 'History' panel in Galaxy. At the top, there is a search bar labeled 'Rechercher des données'. Below it, the panel is titled 'Workshop HPC NGS' and shows '31 shown, 25 deleted, 2 hidden' and a size of '276.65 MB'. The main area contains a list of analyses, each with a number, a description, and icons for viewing, editing, and deleting. The analyses listed are:

- 35: Cuffdiff on data 12, data 7, and data 23: cummeRbund SQLite database
- 34: Cuffdiff on data 12, data 7, and data 23: transcript FPKM tracking
- 33: Cuffdiff on data 12, data 7, and data 23: transcript differential expression testing
- 32: Cuffdiff on data 12, data 7, and data 23: gene FPKM tracking
- 31: Cuffdiff on data 12, data 7, and data 23: gene differential expression testing
- 30: Cuffdiff on data 12, data 7, and data 23: TSS groups FPKM tracking





Multiple histories

- ❑ You can have as many histories as you want
 - ✓ each history should correspond to a different analysis
 - ✓ and should have a meaningful name

Saved Histories

[Advanced Search](#)

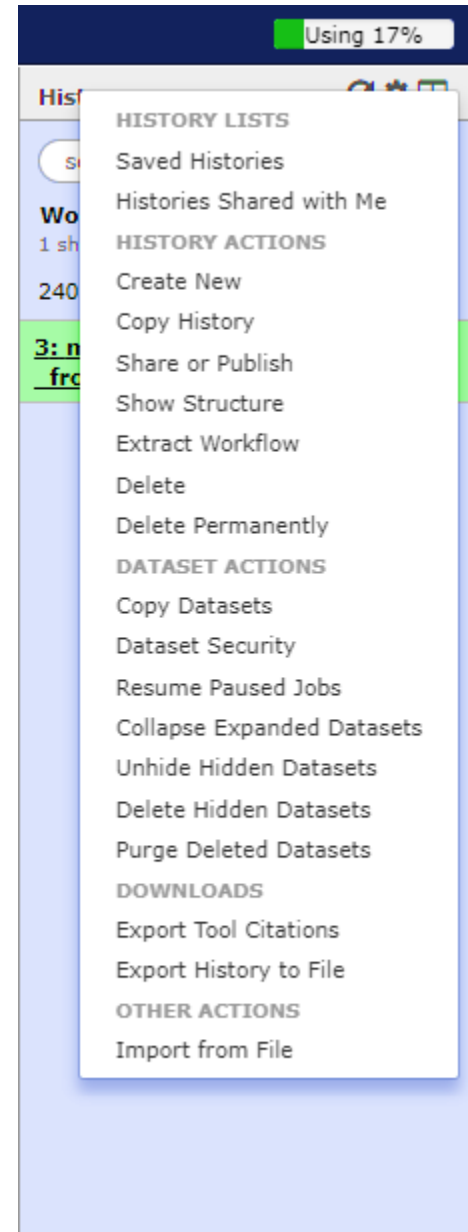
<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated↑	Status
<input type="checkbox"/>	Workshop HPC Cuffdiff+CummRbund	4			2.1 MB	~1 hour ago	~1 hour ago	
<input type="checkbox"/>	Workshop HPC NGS	31			276.6 MB	~2 hours ago	~1 hour ago	current history
<input type="checkbox"/>	Hevea Genome RRIM 600	2			1.3 GB	May 19, 2016	Oct 21, 2016	
<input type="checkbox"/>	Oil Palm Gano RNAseq	4			7.9 GB	Mar 15, 2016	Mar 15, 2016	

For 0 selected items:

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.

History option menu

- ❑ History behavior is controlled by the **History options** (gear icon)
- ❑ **Create New history** will not make your current history disappear
- ❑ To list all your histories, choose **Saved Histories**
- ❑ You can **Copy Datasets** from one history to another
 - ✓ saves disk space and your quota



Loading data

Importing data

- Copy/paste from a **file**
- Upload data from a **local computer**
- Upload data from **internet**
- Upload data from **online databases**
 - ✓ UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Import from **Shared Data** (libraries, histories, pages)
- Upload data from **FTP** (>2GB)



Data types

- ❑ When uploading, **datatype** can be automatically detected or assigned by user
- ❑ For datasets produced by a tool, the datatype is assigned by the tool
- ❑ Tools only accept input datasets with the appropriate datatypes
- ❑ You can **change** the **datatype** of a dataset in 2 ways:
 - ✓ Edit Attributes -> Datatype
 - ✓ Edit Attributes -> Convert Formats



Reference genomes

- Genome build specifies** which genome assembly a dataset is associated with e.g. mm10, hg38...
- Genome build can be **automatically detected** or **assigned** by user
- User can define their own custom genome build
- New genome assemblies can be added by the site Galaxy admin

Workflows

Workflow interface

The screenshot displays a workflow interface for RNA-Seq analysis, titled "Workflow Canvas | RNA-Seq Tutorial v.2". The interface is divided into several main sections:

- Tools Panel (Left):** A sidebar containing a search bar and categorized tool lists. Categories include "COMMON TOOLS", "MICROBIOLOGY", "VARIANT CALLING", "METAGENOMICS", "NGS: RNA ANALYSIS", "CHIP-SEQ", and "OTHER TOOLS".
- Workflow Canvas (Center):** A grid-based workspace where workflow steps are represented as nodes. The nodes are:
 - Input dataset:** Multiple nodes providing "output" to various processing steps.
 - FastQC:Read QC:** Nodes that take "Short read data from your current history" and a "Contaminant list" (html_file) as input.
 - FASTQ Trimmer:** A node that takes a "FASTQ File" (output_file) as input.
 - TopHat2:** Two nodes that take "RNA-Seq FASTQ file, forward reads" and "RNA-Seq FASTQ file, reverse reads" as input, along with a "reference genome" selection. The top node is highlighted with a blue border.
 - Cufflinks:** Two nodes that take a "SAM or BAM file of aligned RNA-Seq reads" as input and produce various output files like "genes_expression (tabular)", "transcripts_expression (tabular)", and "total_map_mass (txt)".
 - Cuffmerge:** A node that takes a "GTF file produced by Cufflinks" and "Reference Annotation" as input to produce "merged_transcripts (gtf)".
- Details Panel (Right):** A panel for the selected "TopHat2" tool, showing:
 - Tool:** TopHat2
 - Version:** 0.6
 - Is this library mate-paired?:** Paired-end
 - RNA-Seq FASTQ file, forward reads:** Data input 'input1' (fastqsanger)
 - RNA-Seq FASTQ file, reverse reads:** Data input 'input2' (fastqsanger)
 - Mean Inner Distance between Mate Pairs:** 110
 - Std. Dev for Distance between Mate Pairs:** 20
 - Report discordant pair alignments?:** Yes
 - Use a built in reference genome or own from your history:** Use a genome from history
 - Select the reference genome:** Data input 'ownFile' (fasta)
 - TopHat settings to use:** Use Defaults
 - Specify read group?:** No
 - Edit Step Actions:** Includes "Rename Dataset" and "align_summary" with a "Create" button.
 - Edit Step Attributes:** Includes an "Annotation / Notes" field.

Workflows

- ❑ Can be **extracted** from a history
 - ✓ Allow to easily convert an existing history into an analysis workflow
- ❑ Can be **built manually** by adding and configuring tools using the workflow canvas
- ❑ Can be **imported** using an existing shared workflow

Why would you want to create workflows?



- ❑ **Re-run** the same analysis on different input data sets
- ❑ **Change parameters** before re-running a similar analysis
- ❑ Make use of the workflow **job scheduling**
 - ✓ jobs are submitted as soon as their inputs are ready
- ❑ Create **sub-workflows**: a workflow inside another workflow
- ❑ **Share workflows** for publication and with the community

Data sharing



- ❑ You can **share** your Galaxy items - histories, workflows, visualizations, and pages - with other people in three different ways:
 - ✓ Directly using a Galaxy account's email addresses on the same instance
 - ✓ Using a web link, with anyone who knows the link
 - ✓ Using a web link and publishing it to make it accessible to everyone from the Shared Data menu
- ❑ Tools are shared using the **free tool store**: **Galaxy Tool Shed** (<https://toolshed.g2.bx.psu.edu/>)

Galaxy Tool Shed

5257 valid tools on Oct 18, 2017

Search

- [Search for valid tools](#)
- [Search for workflows](#)

Valid Galaxy Utilities

- [Tools](#)
- [Custom datatypes](#)
- [Repository dependency definitions](#)
- [Tool dependency definitions](#)

All Repositories

- [Browse by category](#)

Available Actions

- [Login to create a repository](#)

Repositories by Category

Name	Description	Repositories
Assembly	Tools for working with assemblies	97
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	51
Combinatorial Selections	Tools for combinatorial selection	8
Computational chemistry	Tools for use in computational chemistry	51
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	12
Convert Formats	Tools for converting data formats	91
Data Export	Tools for exporting data to various destinations	2
Data Managers	Utilities for Managing Galaxy's built-in data cache	43
Data Source	Tools for retrieving data from external data sources	69
Entomology	Tools that involve insect studies	1
Epigenetics	Tools for analyzing Epigenetic/Epigenomic datasets	18
Fasta Manipulation	Tools for manipulating fasta data	87
Fastq Manipulation	Tools for manipulating fastq data	67

Data visualization

- ❑ **Charts**
- ❑ Each datatype can have some **visualizations associated**
- ❑ Track browser called **Trackster**
 - ✓ To visualize genomic data in a tightly integrated way

Community

- ❑ Be part of an active and friendly community
- ❑ Get support and your questions answered on **Galaxy Biostars**
(<https://biostar.usegalaxy.org/>)
- ❑ Access community curated documentation on **Galaxy Community Hub** (<https://www.galaxyproject.org/>)
- ❑ Learn more about Galaxy for scientists and for developers and admins on **Galaxy Training Community**
(<https://galaxyproject.github.io/training-material/>)

It was **still** **the** **fourth** **course**, don't
get **dizzy** **yet**

