



École de l'ADN Nîmes

Initiation aux NGS

Galaxy

Christian Siatka  
Directeur Général de l'École de l'ADN  
Professeur - Université de Nîmes  
siatka@ecole-adn.fr



Galaxy

NGS Analysis

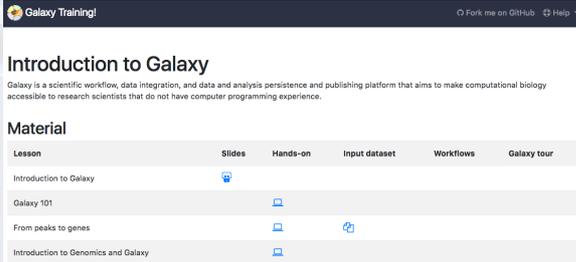
<http://galaxyproject.org/>  
<https://galaxyproject.org/use/usegalaxy-eu/>

Galaxy Training Network

École de l'ADN Nîmes

## Programme

- Qu'est-ce que Galaxy?
- Galaxy pour les bioinformaticiens
- Galaxy pour les biologistes
- Utiliser Galaxy pour l'analyse NGS
- Visualisation et exploration de données NGS avec IGV



**Galaxy Training!** Fork me on GitHub Help

### Introduction to Galaxy

Galaxy is a scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience.

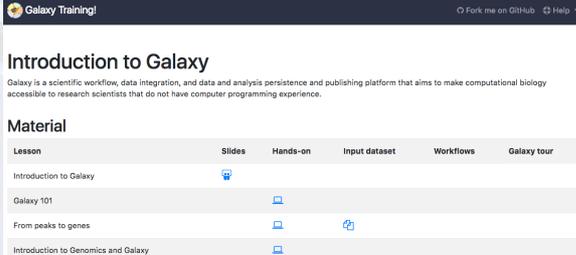
#### Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour
Introduction to Galaxy					
Galaxy 101					
From peaks to genes					
Introduction to Genomics and Galaxy					

École de l'ADN

## Programme

- **Qu'est-ce que Galaxy?**
- Galaxy pour les bioinformaticiens
- Galaxy pour les biologistes
- Utiliser Galaxy pour l'analyse NGS
- Visualisation et exploration de données NGS avec IGV



**Galaxy Training!** Fork me on GitHub Help

### Introduction to Galaxy

Galaxy is a scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience.

#### Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour
Introduction to Galaxy					
Galaxy 101					
From peaks to genes					
Introduction to Genomics and Galaxy					

École de l'ADN

## Galaxy:

### une plateforme d'analyse de génome en libre accès sur le Web

- Galaxy est un **framework open-source** permettant d'intégrer différents outils de calcul et d'analyses de bases de données dans un espace de travail cohérent.
- Un service basé sur le Web, intégrant de nombreux outils et ressources populaires pour la génomique comparative.
- Une application complètement autonome pour construire vos propres style d'analyses Galaxy.

5

## Interface Galaxy

Latest Training Materials

The Galaxy community maintains a wide variety of trainings at <https://training.galaxyproject.org>. Check out the latest metagenomics tutorial by @bebatut and @shiltemann.

150 days

Eat, get fat, and be merry

**News**

Galaxy Release 18.01 – Performance, uWSGI, Collection usability, new BAM datatypes, Experimental job caching

March 2018 News of the Galaxy! – GCCBOSC Abstracts, Registration, Housing & Kenotes; Europe, Africa, blog, pubs, servers, jobs, ...

**Events**

Galaxy Africa – An opportunity to learn from leading bioinformaticists, systems administrators and engineers about Galaxy and accessible, reproducible analysis of biological data

Galaxy : Traitement de données de séquences par Galaxy – détection de SNP, analyse de données RNA-

**@galaxyproject**

Galaxy Project Retweeted

**EMBL-EBI Training**

@EBITraining

Learn how to build a #metabolomics workflow in @galaxyproject in this webinar from Etienne Thévenot. Featuring the @PhnmiH2018 systems and data from @MetaboLights 18. [ow.ly/vXYJ309tlm](#)

**OPEN CHAT**

## Galaxy interface web principale

The screenshot shows the Galaxy web interface. The main content area features a large heading "Running Your Own Understanding how Galaxy works" with the subtitle "An in-depth tutorial". Below this is a tweet from the Galaxy Project (@galaxyproject) mentioning a new GTN tutorial by Maria Doyle, @BelindaPhipson & @hdashnow. The right sidebar displays a "History" section with a search bar and a list of recent jobs, including "Bowtie2 on data 21: aligned reads (sorted BAM)" and "MultiQC on data 29, data 27, and others: Stats".

## Programme

- Qu'est-ce que Galaxy?
- **Galaxy pour les bioinformaticiens**
- Galaxy pour les biologistes
- Utiliser Galaxy pour l'analyse NGS
- Visualisation et exploration de données NGS avec IGV

Galaxy Training! Fork me on GitHub Help

### Introduction to Galaxy

Galaxy is a scientific workflow, data integration, and data analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience.

#### Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour
Introduction to Galaxy					
Galaxy 101					
From peaks to genes					
Introduction to Genomics and Galaxy					

**Galaxy:**  
**la plate-forme instantanée d'intégration d'outils et de gestion de données sur le Web en libre accès**

- Module Open Source téléchargeable pouvant être utilisé dans des laboratoires individuels
- Modulaire
  - Ajouter de nouveaux outils
  - Intégrer de nouvelles sources de données
  - Facile à utiliser dans votre propre environnement (après qqe demi journées d'utilisation)
- Facile à exécuter votre propre serveur "Galaxy privé"

9

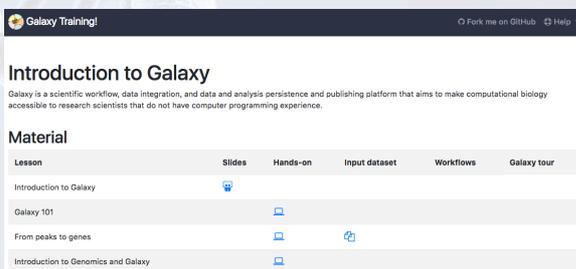
**Galaxy**

<https://galaxy.bioinfo.ucr.edu/root>

The screenshot shows the Galaxy Platform Directory website. The main heading is "Galaxy Platform Directory: Servers, Clouds, and Deployable Resources". Below this, there is a large banner that says "125+ platforms for using Galaxy". The banner contains a grid of various logos representing different tools and services, including ELIXIR, INNOVAGEN, metaNei, plant, Quantifit, RNAcom, SCDI, and others. Below the banner, there are tabs for "UseGalaxy", "All", "Public Servers", "Academic Clouds", "Commercial Clouds", "Containers", and "VMs". At the bottom, there is a section for "UseGalaxy Resources" and an "OPEN CHAT" button.

## Programme

- Qu'est-ce que Galaxy?
- Galaxy pour les bioinformaticiens
- **Galaxy pour les biologistes**
- Utiliser Galaxy pour l'analyse NGS
- Visualisation et exploration de données NGS avec IGV



Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour
Introduction to Galaxy					
Galaxy 101					
From peaks to genes					
Introduction to Genomics and Galaxy					

École de l'ADN

## Galaxy – “le site” pour l'analyse du génome

- Analyser
  - Récupérez des données directement à partir des bases de données standard ou téléchargez les directement les vôtres.
  - Manipulez de manière interactive les données génomiques à l'aide d'un ensemble d'outils, validés, pratiques, complets et en pleine expansion.
  - Galaxy est conçu pour fonctionner avec de nombreux types de données différentes.
- Visualiser
  - Trackster est l'environnement de visualisation et d'analyse visuelle de Galaxy.
  - See more details ([Link](#))
- Publier et partager
  - Résultats et enregistrement d'analyse pas à pas (Bibliothèques de données et historiques)
  - pipelines (Workflows) personnalisables
  - Protocoles complets (Pages)

École de l'ADN

12

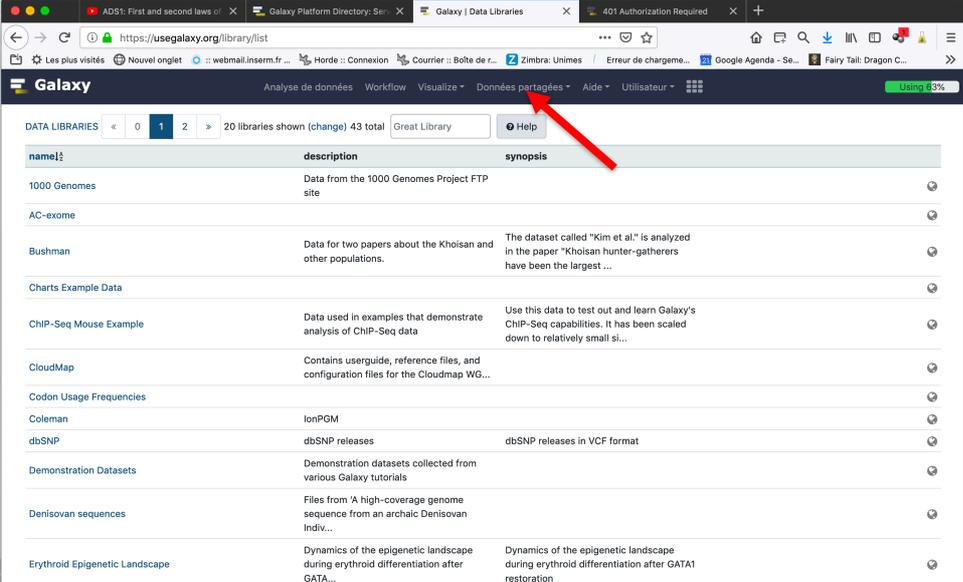
## Outils et ressources

- Datasources ( Données)
  - Télécharger un fichier depuis votre ordinateur
  - Navigateur de table UCSC
  - BioMart, modENCODE fly server
- Suites d'outils
  - Manipulation de sequences
  - Convertiseur de formats
  - NGS
  - graphiques
  - Plus...

13

## Data Libraries (données partagées)

- Datasets accessibles depuis Galaxy ou peuvent être téléchargés.



The screenshot shows the Galaxy Data Libraries page with a table of datasets. A red arrow points to the 'Help' button in the top right corner of the table.

name	description	synopsis
1000 Genomes	Data from the 1000 Genomes Project FTP site	
AC-exome		
Bushman	Data for two papers about the Khoisan and other populations.	The dataset called "Kim et al." is analyzed in the paper "Khoisan hunter-gatherers have been the largest ..."
Charts Example Data		
ChIP-Seq Mouse Example	Data used in examples that demonstrate analysis of ChIP-Seq data	Use this data to test out and learn Galaxy's ChIP-Seq capabilities. It has been scaled down to relatively small si...
CloudMap	Contains userguide, reference files, and configuration files for the Cloudmap WG...	
Codon Usage Frequencies		
Coleman	IonPGM	
dbSNP	dbSNP releases	dbSNP releases in VCF format
Demonstration Datasets	Demonstration datasets collected from various Galaxy tutorials	
Denisovan sequences	Files from 'A high-coverage genome sequence from an archaic Denisovan indiv...	
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during erythroid differentiation after GATA...	Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration

## Workflows

- Le Workflow définit les étapes d'un processus.
- Les Workflows sont des analyses destinées à être exécutées, chaque fois avec différents jeux de données fournies par l'utilisateur.

The screenshot shows the Galaxy platform interface. The 'Workflows' tab is highlighted in the top navigation bar. The main content area displays a table of workflows:

Name	Tags	Owner	# of Steps	Published	Show in tools panel
Workflow constructed from history 'INSERM TRAINING'		You	22	No	<input type="checkbox"/>
Workflow constructed from history 'INSERM TRAINING'		You	0	No	<input type="checkbox"/>
Workflow constructed from history 'INSERM TRAINING'		You	22	No	<input type="checkbox"/>
SavedAs_Workflow constructed from history 'Cs training 23'		You	22	No	<input type="checkbox"/>
Workflow constructed from history 'Cs training 23'		You	22	No	<input type="checkbox"/>
Genome Annotation		You	12	No	<input type="checkbox"/>
aspergillus		You	0	No	<input type="checkbox"/>
CS ngs-wes-illumina-hg10-bwa-freebayes-snpfilt-annovar-2018.01		You	27	No	<input type="checkbox"/>
CS RNASeq-DESeq2 VCS		You	13	No	<input type="checkbox"/>
Imported: Test Fasta - Fasta V3		You	2	No	<input type="checkbox"/>

The right sidebar shows a 'History' section with a list of workflow steps:

- 38: Bowtie2 on data 21: aligned reads (sorted BAM)
- 37: Bowtie2 on data 20: aligned reads (sorted BAM)
- 36: Bowtie2 on data 19: aligned reads (sorted BAM)
- 35: Bowtie2 on data 18: aligned reads (sorted BAM)
- 31: MultiQC on data 29, data 27, and others: Webpage
- 30: MultiQC on data 29, data 27, and others: Stats

## Pages

- Les pages sont des documents qui expliquent les étapes et le raisonnement dans un historique ou un flux de travail particulier.

The screenshot shows the Galaxy platform interface. The 'Données partagées' tab is highlighted in the top navigation bar. The main content area displays a table of published pages:

Title	Annotation	Owner	Community Rating	Community Tags
First using Galaxy	Record greenhand	zyx1088	★★★★★	
demonstration page		j-nomic-s	★★★★★	
Deepools dataset for Pergola	Deepools dataset for Pergola	toniher	★★★★★	pergola, deepools
NOS Assignment		haze194	★★★★★	ngs, assignment
NOS Assignment, Kaja		kaya489	★★★★★	
test french encoding file		ylebras	★★★★★	
Identification of SNPs in single-cell whole genome sequencing of brain cells of Alzheimer's disease patients	Identification of SNPs in DNA samples	mkerdawwy	★★★★★	dna, alzheimer, p05
Cancer Analyses	Page describes data and workflows in Goecks et al.'s 2014 paper on integrated cancer genomics with Galaxy.	jeremy	★★★★★	

The right sidebar shows a 'History' section with a list of workflow steps:

- 38: Bowtie2 on data 21: aligned reads (sorted BAM)
- 37: Bowtie2 on data 20: aligned reads (sorted BAM)
- 36: Bowtie2 on data 19: aligned reads (sorted BAM)
- 35: Bowtie2 on data 18: aligned reads (sorted BAM)
- 31: MultiQC on data 29, data 27, and others: Webpage
- 30: MultiQC on data 29, data 27, and others: Stats

# Histories

- Histories sont toutes les étapes du processus et le réglage utilisé.
- Histories peuvent être importés dans votre session et réexécutés tels quels ou modifiés.

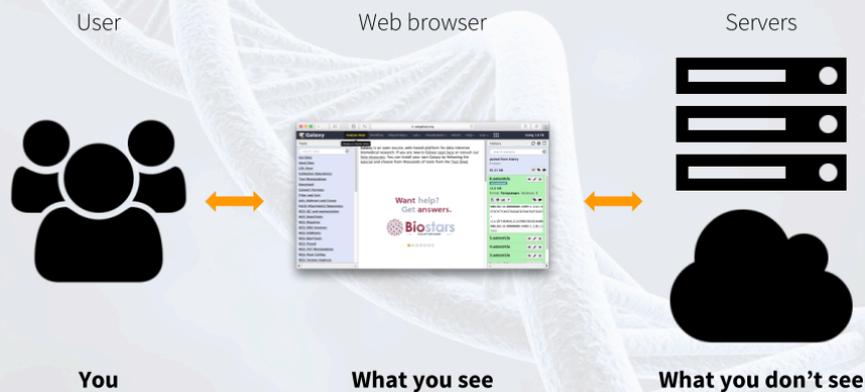
The screenshot shows the Galaxy web interface. The 'Published Histories' section is active, displaying a table of history entries. A red arrow points to the 'Données partagées' dropdown menu. The table lists various history entries with their names, annotations, owners, community ratings, and tags. The right-hand panel shows a detailed view of a history entry, including its name, annotation, and a list of associated data files.

Name	Annotation	Owner	Community Rating	Community Tags
RNA-Seq reads to counts		omar-samir	★★★★★	
WES2imputedVCF		miqrom	★★★★★	wesbamssantoolsgrc37.75
PE-bioinformatics-assessment-data	ngs data gen med MSC assessment data	pmeaton	★★★★★	ngs illumina wex wes
WES fastq	Dartelabs WES fastq	miqrom	★★★★★	sample
NGS ASSESSMENT DATA BIOINFORMATICS	ngs data gen med MSC assessment data	P...	★★★★★	ngs illumina wex wes
GM-2018-P06		prevorovsky	★★★★★	
Practical lessons		michaelsimova	★★★★★	
Robyn Santo Thesis		profbiot	★★★★★	
Kiera Lawlor		profbiot	★★★★★	

17

# User Account

- Un compte n'est pas requis pour accéder aux plications principales publiques ou de test de Galaxy



18

## Programme

- Qu'est-ce que Galaxy?
- Galaxy pour les bioinformaticiens
- Galaxy pour les biologistes
- **Utiliser Galaxy pour l'analyse NGS**
- Visualisation et exploration de données NGS avec IGV

Galaxy Training! Fork me on GitHub Help

### Introduction to Galaxy

Galaxy is a scientific workflow, data integration, and data analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience.

#### Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour
Introduction to Galaxy					
Galaxy 101					
From peaks to genes					
Introduction to Genomics and Galaxy					



## Programme

- **Prise en main simple !!!**
- **Utiliser Galaxy pour l'analyse NGS**

**Retrieved Exons**  
1. Exons

**Retrieved SNPs**  
2. SNPs

3. Join on data 2 and data 1

4. Group on data 3

5. Sort on data 4

6. Select first on data 5

7. Join two datasets on data 6 and data 1

8. Cut on data 7





## NGS Data

- Raw: read de séquençage (FASTQ)
- Obtenus à partir de:
  - Alignements contre le génome de référence (SAM/BAM)
  - Annotations
    - GFF/GTF (the GFF (General Feature Format) format consists of one line per feature, each containing 9 columns of data, plus optional track definition lines. The following documentation is based on the Version 2 specifications. The GTF (General Transfer Format) is identical to GFF version 2)
    - BED (The BED format consists of one line per feature, each containing 3-12 columns of data, plus optional track definition lines)

The screenshot shows the Ensembl website's documentation for the BED File Format. The page title is "BED File Format - Definition and supported options". It states: "The BED format consists of one line per feature, each containing 3-12 columns of data, plus optional track definition lines." Below this, there are links for "Required fields", "Optional fields", "Track lines", and "BedGraph format". A section titled "Required fields" notes: "The first three fields in each feature line are required:".

21

## FASTQ Format

**A FASTQ fichier utilise normalement quatre lignes par séquence.**

- La ligne 1 commence par un caractère '@' et est suivie d'un identificateur de séquence.
- La ligne 2 correspond à la séquence brute.
- La ligne 3 commence par un caractère '+', est éventuellement suivie par le même identificateur de séquence.
- La ligne 4 code les valeurs de qualité "phred" pour la séquence est la ligne 2, chaque valeur représente la probabilité d'erreur en caractère ASCII.

```
@SRR064154.208 HWUSI-EAS627_1:8:1:2:1681 length=38
ANGANNNGGACTTTGAAAAGAGAGTCAAAGAGTGCTTG
+
?!08!!!3C?BCBB<BCBB?BBACABBBBBBB@CABAB
```

22



## Format GFF et GTF

- General feature format (GFF)

```
browser position chr22:10000000-10025000
browser hide all
track name=regulatory description="TeleGene(tm) Regulatory Regions"
visibility=2
chr22 TeleGene enhancer 10000000 10001000 500 + . touch1
chr22 TeleGene promoter 10010000 10010100 900 + . touch1
chr22 TeleGene promoter 10020000 10025000 800 - . touch2
```

- Gene Transfer format (GTF)
  - L'attribut list doit commencer par 2 attributs obligatoires.
  - Gene\_id\_value, transcript\_id\_value

```
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
```

## BED format

- Une façon très "flexible" de définir les lignes de données dans la ligne d'annotation.

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

### BCF/VCF format

Le Variant Call Format (VCF) est la nouvelle norme pour le stockage des données de variante. Conçu à l'origine pour les SNP et les INDEL courts, il fonctionne également pour les variations structurelles.

VCF est un format de fichier texte stocké de manière compressée. Il contient des lignes de méta-informations, une ligne d'en-tête, puis des lignes de données contenant chacune des informations sur une position dans le génome.

# BCF/VCF format

Le Variant Call Format (VCF) est le nouveau norme pour le stockage des

The screenshot displays the Galaxy VCF viewer interface. At the top, it shows the Galaxy logo and navigation options. The main area is divided into a sidebar on the left with tool categories like 'Get Data', 'Send Data', 'Text Manipulation', and 'Filter and Sort'. The central panel shows a table of variants with columns: Chrom, Pos, ID, Ref, and Qual. Below the table, there is a detailed view of a specific variant, including its genomic coordinates and quality scores. The bottom right corner of the screenshot shows the page number '27'.

## Les outils de base disponibles pour l'analyse NGS

- Préparer, contrôler la qualité et manipuler les reads FASTQ
- Mapping
- SAMTools
- Analyses de SNP et INDEL
- Analyses de RNA
- Peak calling / CHIP-seq

The screenshot shows the Galaxy tool menu with a search bar and a list of tool categories. The categories listed include: Get Data, Send Data, Lift-Over, Collection Operations, Text Manipulation, Datamash, Convert Formats, Filter and Sort, Join, Subtract and Group, Fetch Alignments/Sequences, NGS: QC 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, NGS: Gemini, NGS: Assembly, and NGS: Chromosome Conformation.

## Analyses NGS avec Galaxy

- **Présentation générale de Galaxy et Interface**
- Importer des Data in Galaxy
- Analyser les Data dans Galaxy
  - Quality Control
  - Mapping Data
- Historique et workflow
- Sequences et format d'alignment
- Entraînement sur "Galaxy" !!!




## Commencer avec Galaxy

Data intensive biology for everyone

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

Welcome to the Galaxy Community Hub, where you'll find community curated documentation of all things Galaxy.

**News**

[Galaxy Release 18.01](#) – Performance, uWSGI, Collection usability, new BAM datatypes, Experimental job caching

[March 2018 News of the Galaxy!](#) – GCCBOSC Abstracts, Registration, Housing & Kenotes; Europe, Africa, blog, pubs, servers, jobs, ...

**Events**

[Galaxy Africa](#) – An opportunity to learn from leading bioinformaticists, systems administrators and engineers about Galaxy and accessible, reproducible analysis of biological data

[Galaxy : Traitement de données de séquences par Galaxy](#) – détection de SNP, analyse de données RNA-

**@galaxyproject**

Galaxy Project Retweeted

**EMBL-EBI Training**  
@EBItraining

Learn how to build a #metabolomics workflow in @galaxyproject in this webinar from Etienne Thévenot. Featuring the @Phnml1000 platform and data from @MetaboLights18

[OPEN CHAT](#)

## Concepts de Galaxy

**Obtain data** from many data sources including the UCSC Table Browser, BioMart, WormBase, or your own data.

**Prepare data** for further analysis by rearranging or cutting data columns, filtering data and many other actions.

**Analyze data** by finding overlapping regions, determining statistics, phylogenetic analysis and much more

The screenshot shows the Galaxy web interface with several tool panels. On the left, there's a 'Tools' sidebar with categories like 'Get Data', 'Text Manipulation', and 'Join'. The main area shows a 'Filter' tool configuration with a 'Filter' dropdown set to 'UCSC Genes' and a 'Filter' field containing 'chr2'. Below it, there's a 'Join' tool configuration with 'First query' and 'Second query' dropdowns. The right side shows a 'History' panel with a list of jobs and their outputs.

## Galaxy : Interface Sections

utilisateur

enregistrement

Login

The screenshot shows the Galaxy main workspace. At the top, there's a navigation bar with 'utilisateur', 'enregistrement', and 'Login' buttons. Below it, there's a 'Tools' sidebar on the left. The main area shows a dataset view with a 'Search tools' dropdown and a 'Get Data' section. The dataset view shows a list of genomic coordinates and a 'Menu de résultats et d'analyse de données' button. On the right, there's a 'History' panel with a list of jobs and their outputs.

Tous les liens pour télécharger et analyser les datas

Menu de résultats et d'analyse de données

Historique de votre activité et de vos étapes d'analyses

## Analyses NGS avec Galaxy

- Présentation générale de Galaxy et Interface
- **Importer des datas in Galaxy**
- Analyser les datas dans Galaxy
  - Quality Control
  - Mapping Data
- Historique et workflow
- Sequences et format d'alignment
- Entraînement sur "Galaxy" !!!





## Les Datas dans Galaxy



The screenshot shows the Galaxy web interface. A blue arrow points to the 'Importer les Datas' button in the top-left navigation menu. The main content area features a 'Try Galaxy on the Cloud' banner and a tweet from EMBL-EBI Training. The right-hand sidebar shows a 'History' panel with a list of recent data uploads, including '47: Cd20\_Rep1' (29.6 MB) and '46: Gene Annotations for chr19' (108.7 MB).

## Les Datas dans Galaxy

The screenshot shows the Galaxy web interface with the 'Upload File' dialog box open. The dialog box has a search bar for species, a 'Drop files here' area, and buttons for 'Choose local file', 'Choose FTP file', and 'Paste/Fetch from URL'. A list of species is shown on the right, with 'Human Mar. 2006 (hg18)' highlighted. Blue arrows point to 'Upload File', 'Species', and 'Executer' buttons.

## Les Datas dans Galaxy

Voir les variantes  
d'enregistrement de datas  
en ligne directement

36

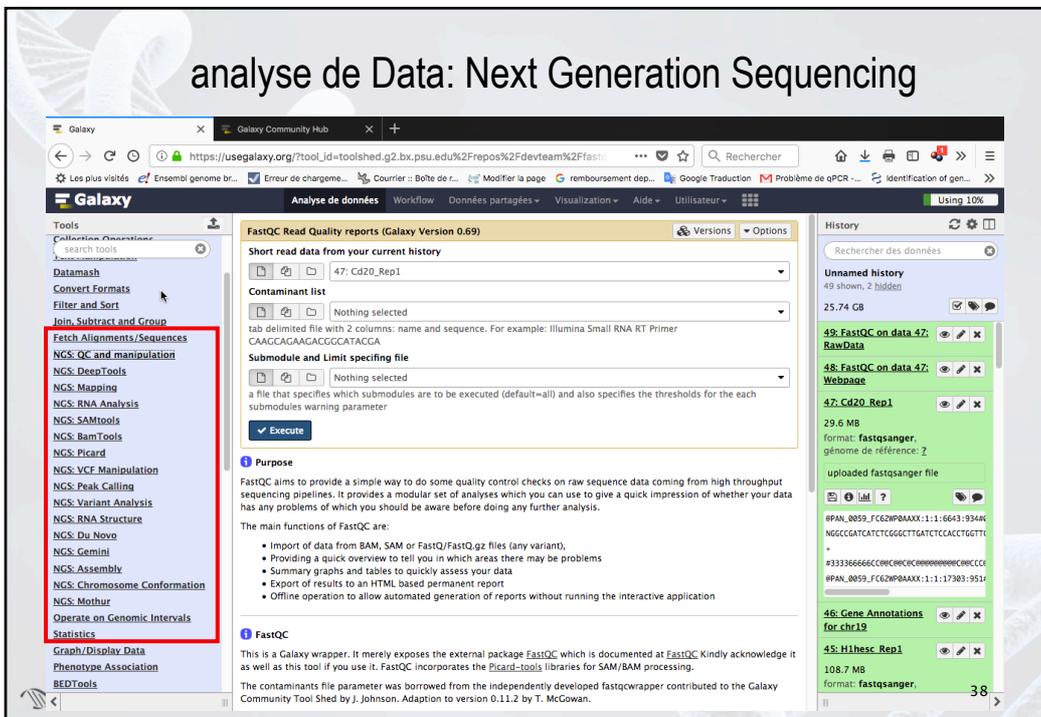
## Analyses NGS avec Galaxy

- Présentation générale de Galaxy et Interface
- Importer des Data in Galaxy
- **Analyser les datas dans Galaxy**
  - Quality Control
  - Mapping Data
- Historique et workflow
- Sequences et format d'alignment
- Entraînement sur "Galaxy" !!!





## analyse de Data: Next Generation Sequencing



The screenshot displays the Galaxy web interface for the FastQC tool. The main content area shows the tool configuration for 'FastQC Read Quality reports (Galaxy Version 0.69)'. The configuration includes a 'Short read data from your current history' dropdown set to '47: Cd20\_Rep1', a 'Contaminant list' dropdown set to 'Nothing selected', and a 'Submodule and Limit specifying file' dropdown also set to 'Nothing selected'. An 'Execute' button is visible at the bottom of the configuration area.

The left-hand navigation menu is visible, with the 'Fetch Alignments/Sequences' section highlighted in red. This section includes the following tools:

- NGS: QC 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
- NGS: Gemini
- NGS: Assembly
- NGS: Chromosome Conformation
- NGS: Motthur
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- BEDTools

The right-hand history panel shows a list of recent jobs, including '49: FastQC on data 47: RawData', '48: FastQC on data 47: Webpage', '47: Cd20\_Rep1', '46: Gene Annotations for chr19', and '45: Hihesc\_Rep1'. The '47: Cd20\_Rep1' job is currently selected and highlighted in green.

## analyse de Data: Next Generation Sequencing

**FASTQC file manipulation, like format conversion, summary statistics, trimming reads, filtering reads by quality score...**

**Purpose**  
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ/FASTQ.gz files (any variant).
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

**FastQC**  
This is a Galaxy wrapper. It merely exposes the external package FastQC which is documented at [FastQC](#). Kindly acknowledge it as well as the Galaxy wrapper. FastQC incorporates the Picard-tools libraries for SAM/BAM processing. The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by J. Johnson. Adaption to version 0.11.2 by T. McCowan.

## Analyzing Data: Next Generation Sequencing

**NGS: SAM Tools**

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

**Filter pileup**

Select dataset:

which contains:

Do not consider read bases with quality lower than:

Do not report positions with coverage lower than:

Only report variants:

Convert coordinates to intervals:

Print total number of differences:

Print quality and base string:

**What it does**  
Allows one to find sequence variants and/or sites covered by a specified number of reads with bases above a set quality threshold. The tool works on six and ten column pileup formats produced with `samtools pileup` command. However, it also allows you to specify columns in the input file manually. The tool assumes the following:

- the quality scores follow phred33 convention, where input qualities are ASCII characters equal to the phred quality plus 33.
- the pileup dataset was produced by the `samtools pileup` command (although you can override this by setting column assignments manually).

**Types of pileup datasets**  
The descriptions of the following pileup formats are largely based on information that can be found on the [SAMTools](#) documentation page. The 6- and 10-column variants are described below.

**Six column pileup**

## analyse de Data: Next Generation Sequencing

NGS SAM tools,  
PICARD,  
RNA analysis  
Assemblage  
mapping

**Tools**

- Calibration Operations
- Search tools
- Contaminant list
- Datamash
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- NGS: QC 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
- NGS: Gemin
- NGS: Assembly
- NGS: Chromosome Conformation
- NGS: Motif
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- BEDTools

**FastQC Read Quality Control**

Short read data file

Contaminant list

Subprocess and Limit specifying file

Execute

**Purpose**

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ/FastQ.gz files (any variant),
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

**FastQC**

This is a Galaxy wrapper. It merely exposes the external package `FastQC` which is documented at `FastQC`. Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the `Picard-tools` libraries for SAM/BAM processing.

The contaminants file parameter was borrowed from the independently developed `fastqcwrapper` contributed to the Galaxy Community Tool Shed by J. Johnson. Adaption to version 0.11.2 by T. McGowan.

**History**

25.74 GB

49: FastQC on data 42: RawData

48: FastQC on data 42: Webpage

47: Cd20\_Rep1

29.6 MB

format: fastqsanger,

génomme de référence: 2

uploaded fastqsanger file

46: Gene Annotations for chr19

45: H1hesc\_Rep1

108.7 MB

format: fastqsanger.

## Analyses NGS avec Galaxy

- Présentation générale de Galaxy et Interface
- Importer des Data in Galaxy
- Analyser les datas dans Galaxy
  - Quality Control
  - Mapping Data
- **Historique et workflow**
- Sequences et format d'alignement
- Entraînement sur "Galaxy" !!!

**Galaxy**  
PROJECT

École de l'ADN

## History: History Options

**Historique**

Répertoire les historiques enregistrés et les historiques partagés. Travailler sur l'historique actuel, créer un nouveau, cloner, partager, créer un flux de travail, définir des autorisations, afficher des ensembles de données supprimés ou supprimer l'historique.

The screenshot shows the Galaxy interface with the 'History' panel on the right. The panel lists various datasets, including 'FastQC on data 47: RawData', 'FastQC on data 47: Webpage', and 'Cd20\_Rep1'. A blue arrow points from the word 'Historique' to the 'History' panel. A blue callout box contains text in French describing the history options.

## Workflow

Crée un workflow, permet à l'utilisateur de répéter l'analyse en utilisant différents jeux de données.

The screenshot shows the Galaxy interface with the 'Workflow' panel on the left. The panel lists various workflows, including 'Regional Variation', 'FASTA manipulation', and 'Multiple Alignments'. A blue callout box contains text in French describing the workflow creation process.

## Workflow

Créer un workflow,  
permet à l'utilisateur de répéter l'analyse  
en utilisant différents jeux de données.

## Il n'y a plus qu'à...

**TRAINING**

COACHING    TEACHING    KNOWLEDGE

SKILLS    LEARN    DEVELOPMENT

EXPERIENCE

# Il n'y a plus qu'à...

**Sequence analysis**  
Analyses of sequences

**Requirements**  
Before diving into this topic, we recommend you to have a look at:

- [Galaxy introduction](#)

**Material**

Lesson	Slides	Hands-on	Interactive dataset	Workflows	Galaxy tour
Quality Control					
Mapping					
Genome Annotation					
RAD-Seq Reference-based data analysis					
RAD-Seq de-novo data analysis					
RAD-Seq to construct genetic maps					
Genome annotation with Prokka					

47

# Il n'y a plus qu'à... 1

**Quality Control**

**Overview**

- ❏ **Questions**
  - How to control quality of NGS data?
  - What are the quality parameters to check for each dataset?
  - How to improve the quality of a sequence dataset?
- ❏ **Objectives**
  - Manipulate FastQ files
  - Control quality from a FastQ file
  - Use FastQC tool
  - Understand FastQC output
  - Use tools for quality correction
- ❏ **Requirements**
  - [Galaxy introduction](#)
- 🕒 **Time estimation:** 1h

**Introduction**

During sequencing, errors might be introduced, such as the incorporation of ambiguous nucleotides. These are due to the technical limitations of each sequencing platform. Sequencing errors might bias the analysis, ultimately leading to a misinterpretation of the data.

## Il n'y a plus qu'à... 1

Dans ce tuto, nous allons vérifier la qualité de deux jeux de données pour nous assurer que les données sont correctes avant d'en déduire des informations supplémentaires.

Cette étape est la base de toute analyse de pipeline telle que RNA-Seq, ChIP-Seq ou toute autre analyse OMIC reposant sur des données NGS.

Les étapes du contrôle qualité sont similaires pour tout type de données de séquençage



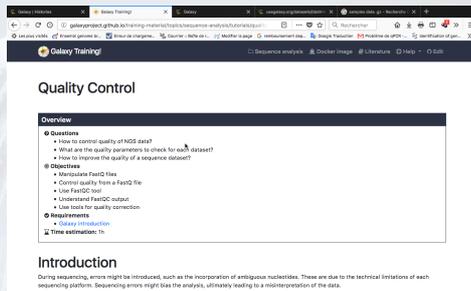
Quality checking with FastQC

↓

Improvement of the quality of the sequences with Trim Galore

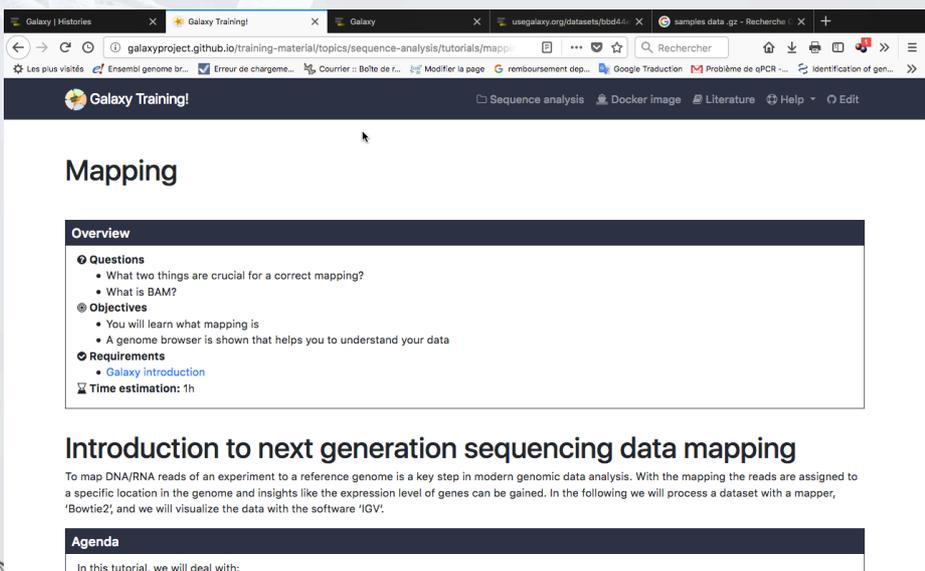
↓

Quality re-checking with FastQC



49

## Il n'y a plus qu'à... 2



50

## Il n'y a plus qu'à... 2

### Introduction au mapping de données de séquençage NGS

Mapper des lectures d'ADN / ARN d'une expérience à un génome de référence est une étape clé dans l'analyse des données génomiques modernes.

Avec la cartographie, les lectures sont assignées à un emplacement spécifique dans le génome et des aperçus comme le niveau d'expression des gènes peuvent être obtenus. Dans la suite, nous allons traiter un ensemble de données avec un mappeur, **'Bowtie2'**, et nous allons visualiser les données avec le logiciel **'IGV'**.



**Mapping**

**Overview**

**Questions**

- What two things are crucial for a correct mapping?
- What is BAM?

**Objectives**

- You will learn what mapping is
- A genome browser is shown that helps you to understand your data

**Requirements**

- Galaxy introduction

**Time estimation:** 1h

**Introduction to next generation sequencing data mapping**

To map DNA/RNA reads of an experiment to a reference genome is a key step in modern genomic data analysis. With the mapping the reads are assigned to a specific location in the genome and insights like the expression level of genes can be gained. In the following we will process a dataset with a mapper, 'Bowtie2' and we will visualise the data with the software 'IGV'.

**Agenda**

In this tutorial, we will deal with:



<http://software.broadinstitute.org/software/igv/download>

51

## Il n'y a plus qu'à... 3

### Genome Annotation

#### Overview

##### Questions

- First question addressed during the tutorial
- Second question addressed during the tutorial

##### Objectives

- First learning objectives of the tutorial
- Second learning objectives of the tutorial

##### Requirements

- Galaxy introduction

**Time estimation:** 1h/1d

### Introduction

Genome annotation is the process of attaching biological information to sequences. It consists of three main steps:

- identifying portions of the genome that do not code for proteins
- identifying elements on the genome, a process called gene prediction, and
- attaching biological information to these elements.

#### Agenda



52

