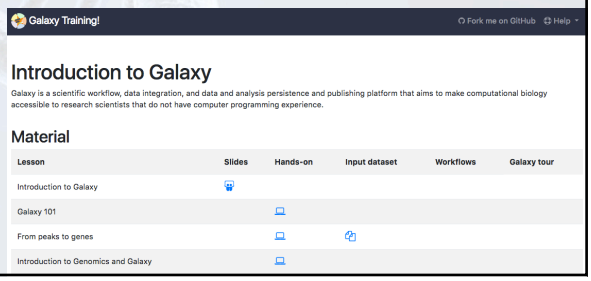

École de l'ADN Nîmes
Initiation aux NGS
Galaxy
Christian Siatka
Directeur Général de L'Ecole de l'ADN
Professeur - Université de Nîmes
siatka@ecole-adn.fr
1


Galaxy
NGS Analysis
http://galaxyproject.org/
https://galaxyproject.org/use/usegalaxy-eu/
Galaxy
Galaxy Training Network
Training Resources Directory
Trainer Directory
Compute Platform Advice
Best Practices
Mailing List
Events
Interested?
Training Resources Directory
OPEN CHAT
École de l'ADN
2

Programme

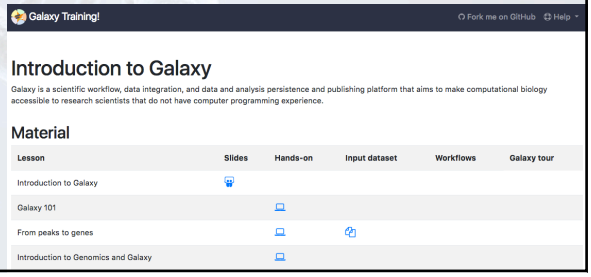
- Qu'est-ce que Galaxy?
- Galaxy pour les bioinformaticiens
- Galaxy pour les biologistes
- Utiliser Galaxy pour l'analyse NGS
- Visualisation et exploration de données NGS avec IGV



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



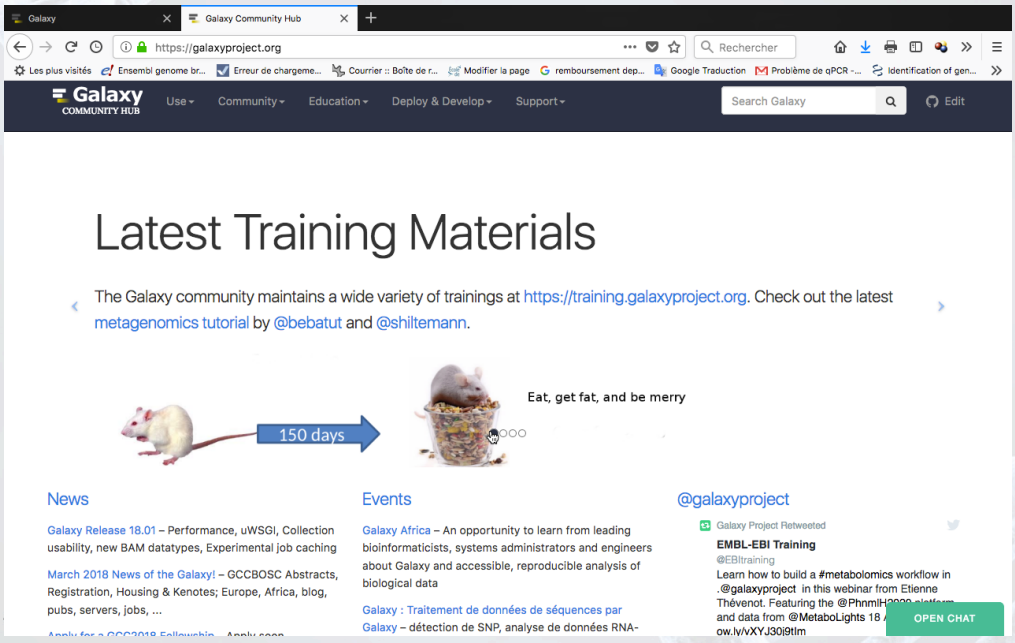
École de l'ADN

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

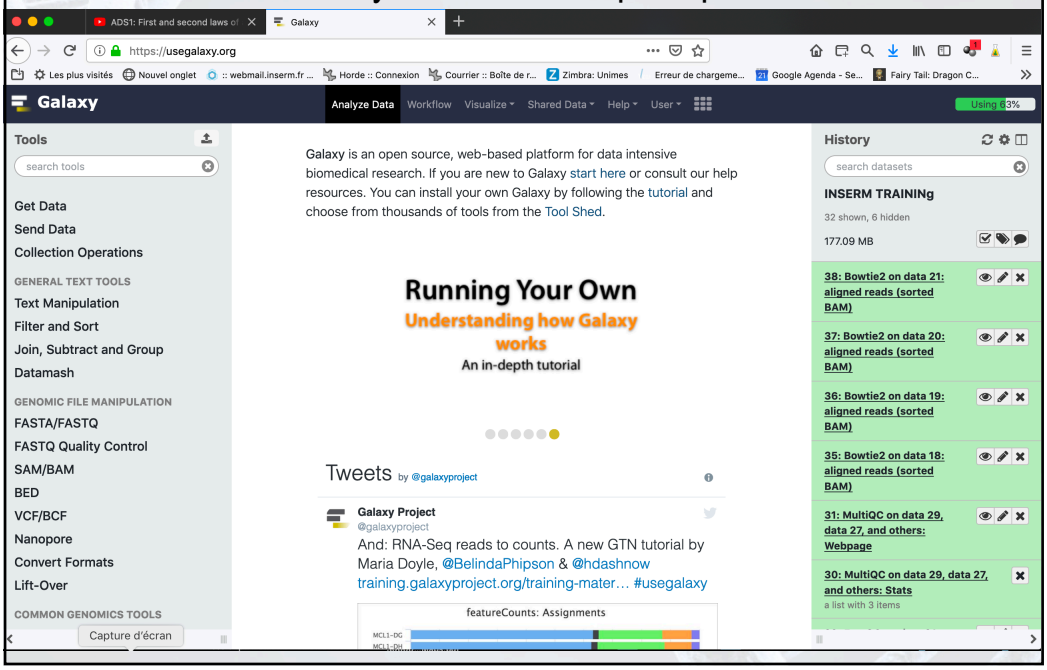
- Galaxy est un **framework open-source** permettant d'intégrer différents outils de calcul et d'analyses de bases de données dans un espace de travail cohérent.
- Un service basé sur le Web, intégrant de nombreux outils et ressources populaires pour la génomique comparative.
- Une application complètement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

5

Interface Galaxy

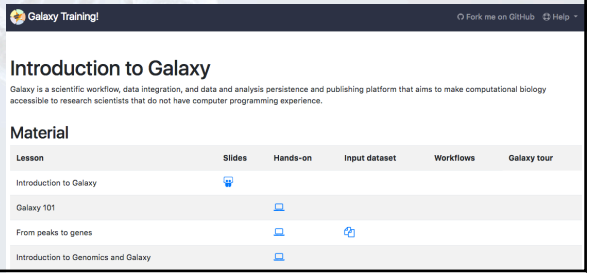


Galaxy interface web principale



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

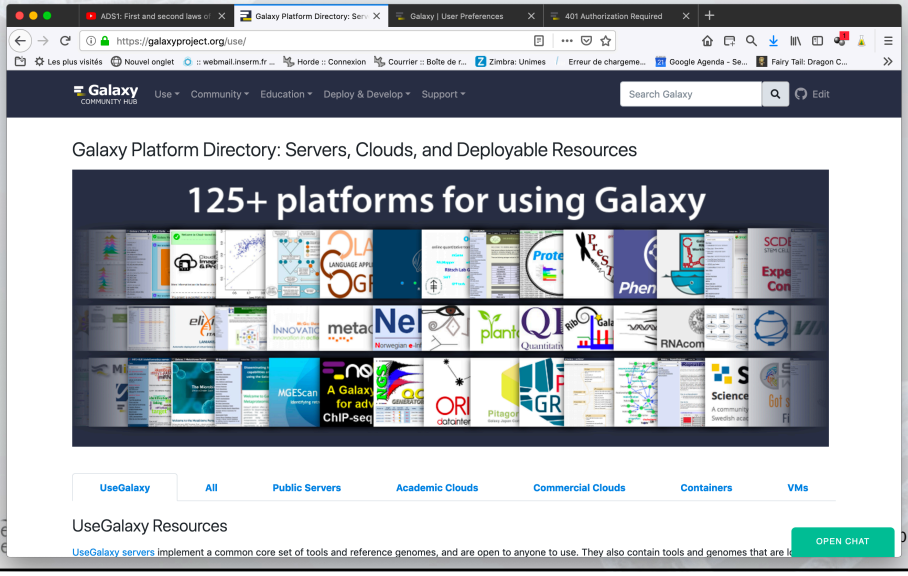


Galaxy:
la plate-forme instantanée d'intégration d'outils et de gestion de données sur le Web en libre accès

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

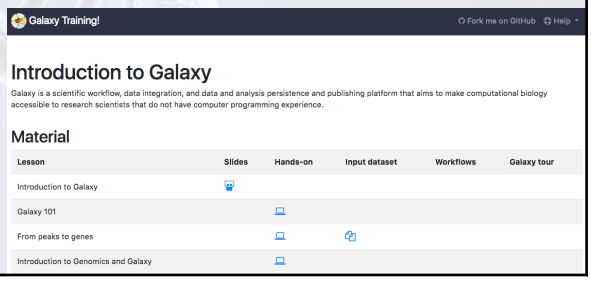
Galaxy

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



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 – "le site" pour l'analyse du génome

- Analyser
  - Récupérez des données directement à partir des bases de données standard ou téléchargez les directemet 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) personalisables
  - Protocoles complets (Pages)

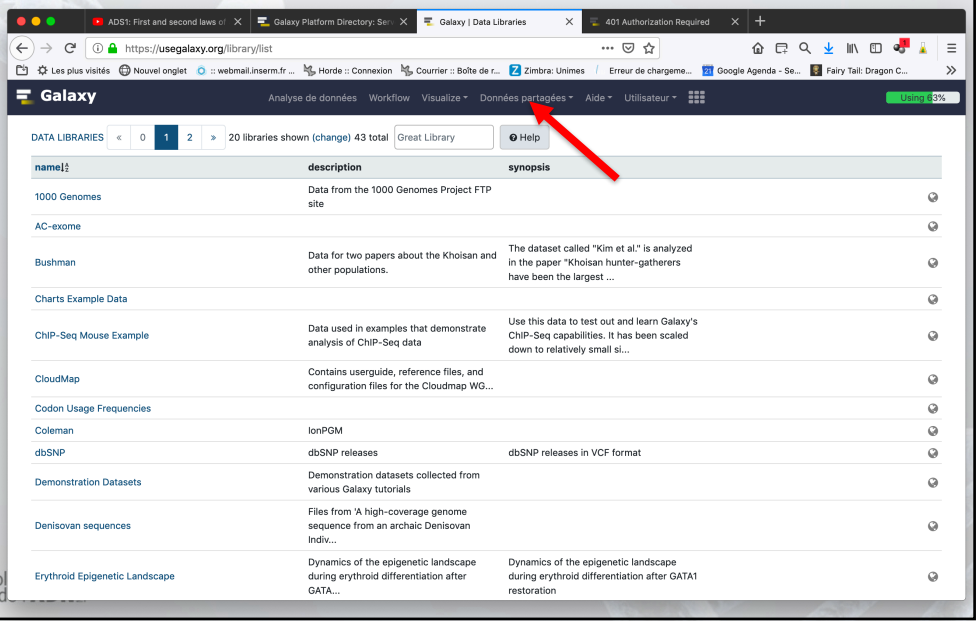
12

Outils et resources

- 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…

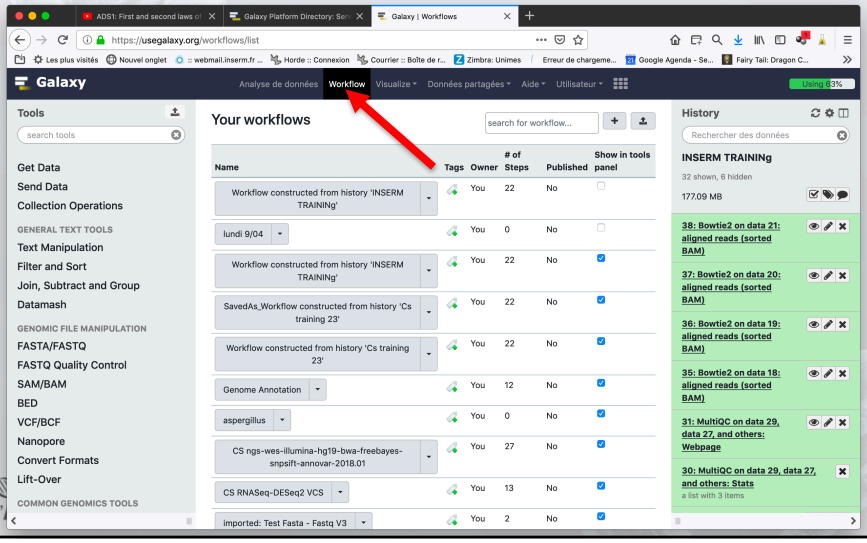
Data Libraries (données partagées)

- Datasets accessibles depuis Galaxy ou peuvent etre téléchargées.



Workflows

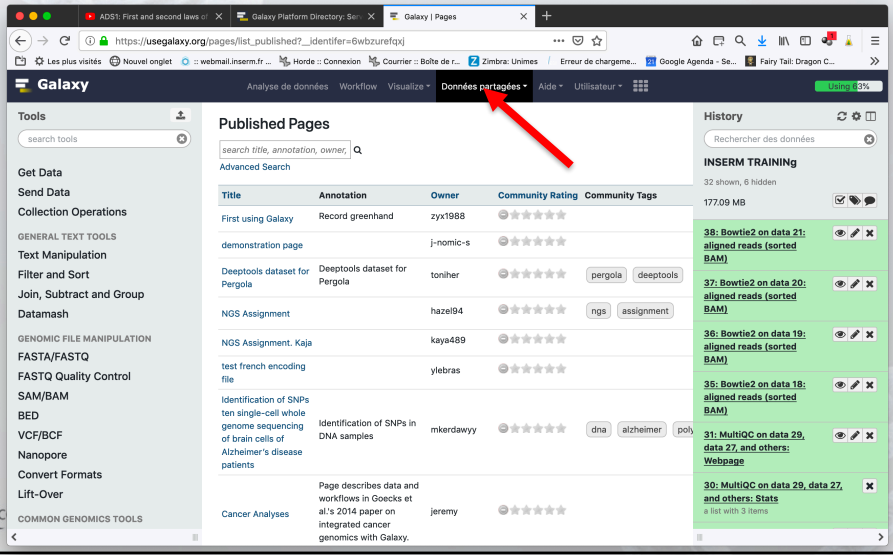
- Le Workflow définie les étapes d'un processus.
- Les Workflow sont des analyses destinées à être exécutées, chaque fois avec différents jeux de données fournies par l'utilisateur.



15

Pages

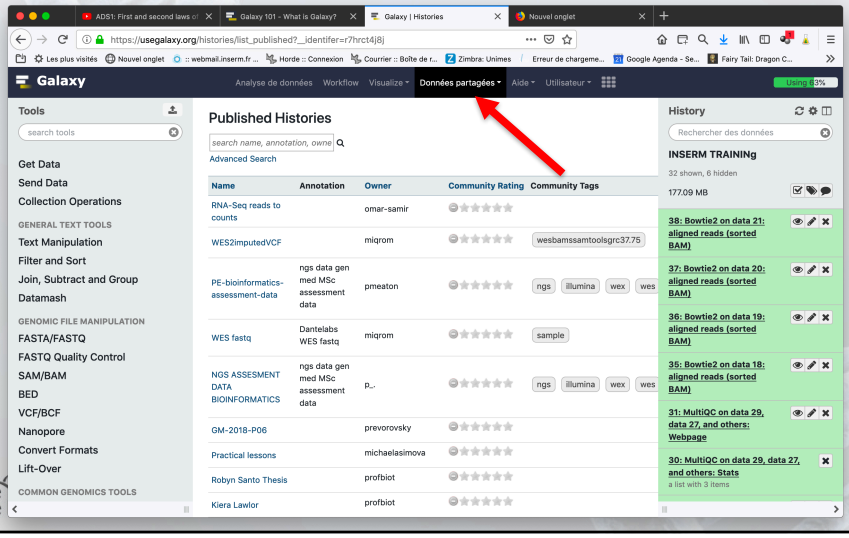
- Les pages sont des documents qui expliquent les étapes et le raisonnement dans un historique ou un flux de travail particulier.



16

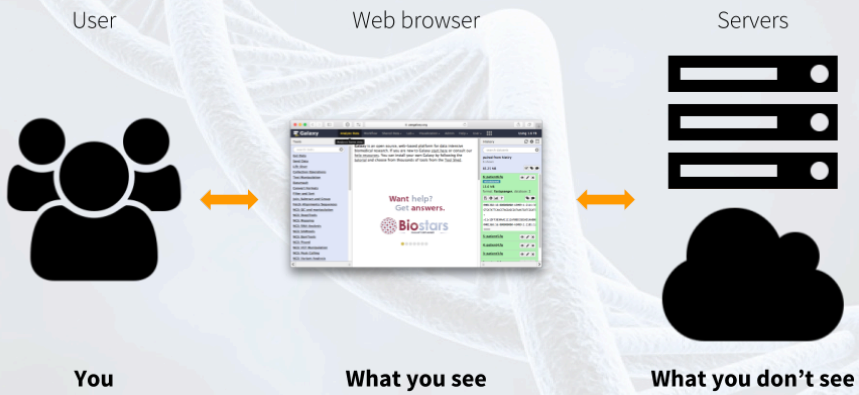
Histories

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



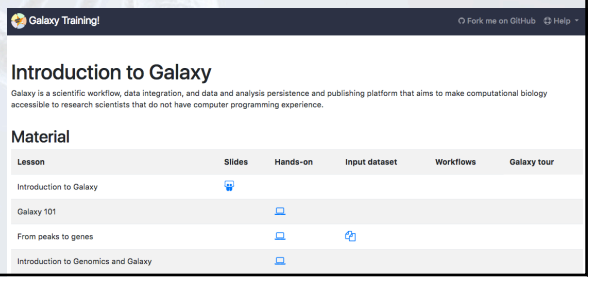
User Account

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



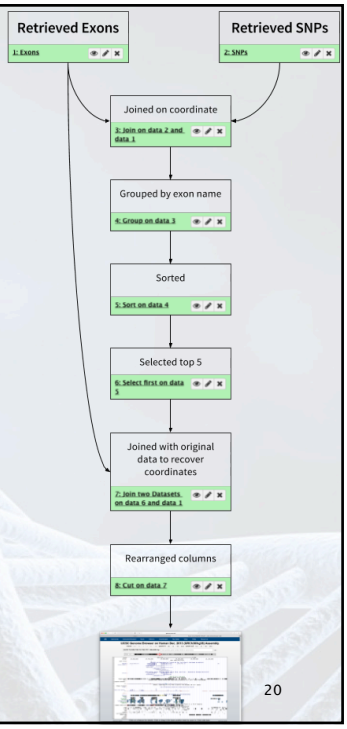
Programme

- 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



Programme

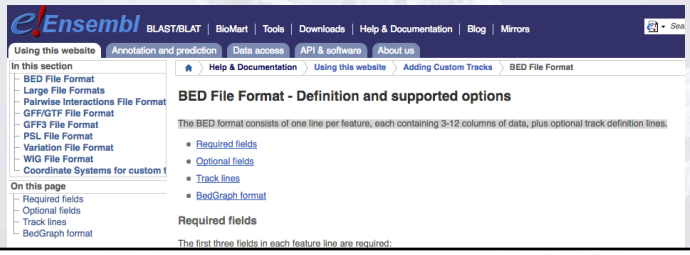
- **Prise en main simple !!!**
- **Utiliser Galaxy pour l'analyse NGS**



20

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)



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 caratère ASCII.

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

FASTQ Quality Score

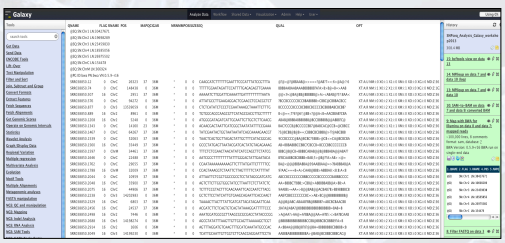
- Le score de qualité représente la probabilité d'erreur d'un basecall.
- Dans le fichier fastq, le score de qualité est souvent représenté en utilisant l'alphabet ASCII.
- Par exemple, un score Phred de 40 peut être représenté :
- En sangers + 33 par le caractère ASCII "I" (40 + 33 = ASCII # 73),
- En Illumina + 40 par le caractère ASCII "h" (40 + 64 = ASCII # 104).
- La gamme de scores dépendra de la technologie de sequençage, mais sera généralement notée jusqu'à 40.

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
   |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger, Illumina 1.8+     Phred+33,  raw reads typically (0, 41)
 X - Solexa                    Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+             Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+             Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
```


Le format SAM/BAM

- SAM : Sequence Alignment/Map format.
- Pour plus de details: http://samtools.sourceforge.net/SAM1.pdf
- SAM est synonyme de format d'alignement de séquence / carte. C'est un format de texte délimité par TAB composé d'un en-têtesection, qui est facultative, et une section d'alignement.
- S'il est présent, l'en-tête doit être antérieur aux alignements.Les lignes d'en-tête commencent par `@ ', alors que les lignes d'alignement ne le font pas. Chaque ligne d'alignement a 11 champs obligatoires d'informations d'alignement essentielles telles que la position de mapping et le nombre variable de champs facultatifs pour utiliser des informations spécifiques à l'aligneur.



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 facon 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, binary variant call format, est la version binaire de VCF. Il conserve

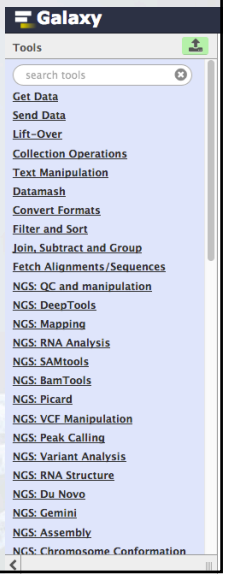
BCF/VCF format

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

entre BCF et VCF est similaire à celle entre BAM et SAM.

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



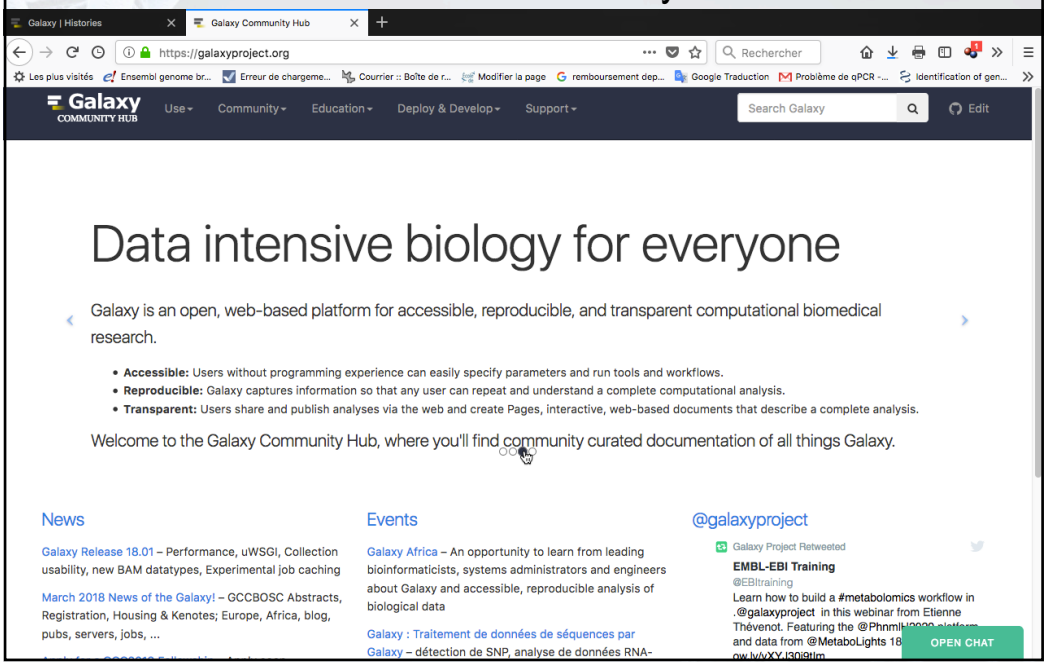
Analyses NGS avec Galaxy

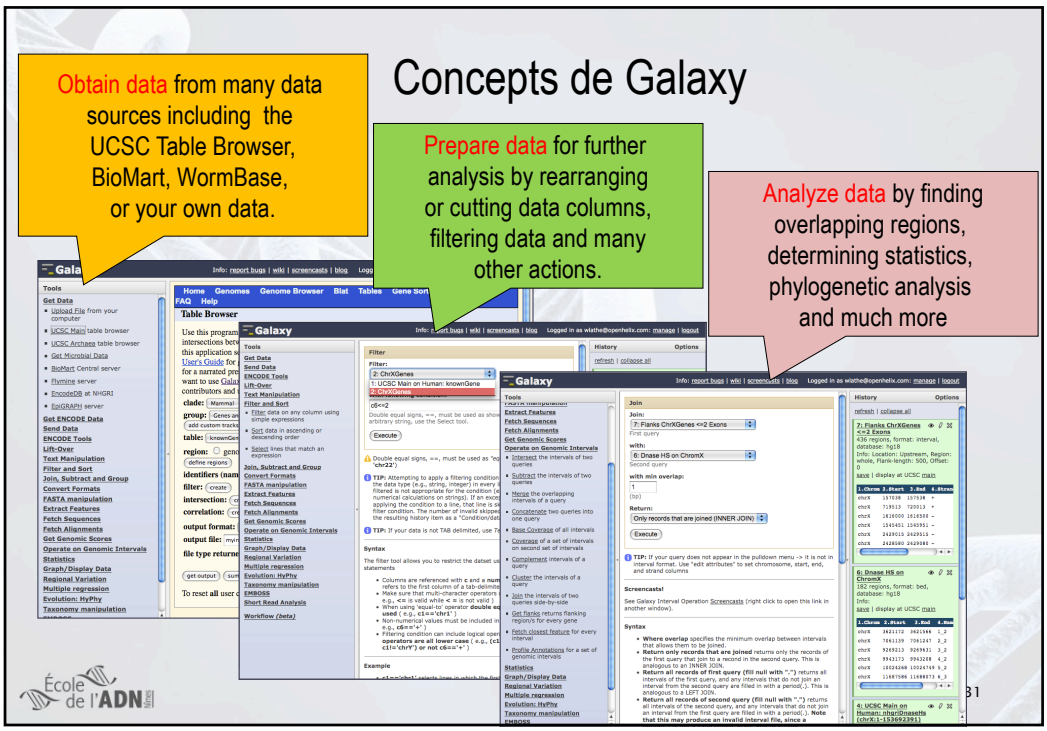
- **Présentation générale de Galaxy et Interface**
- Importer des Data in Galaxy
- Analyser les Data dans Galaxy
  - Quality Control
  - Mapping Data
- Historique et workflow
- Sequences et format d'alignment
- Entrainement sur "Galaxy" !!!

École de l'ADN

Galaxy PROJECT

Commencer avec Galaxy




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
École de l'ADN
31


Galaxy : Interface Sections
utilisateur
enregistrement
Login
Register
Tous les liens pour telecharger et analyser les datas
Menu de résultats et d'analyse de données
Historique de votre activité et de vos etapes d'analyses
École de l'ADN
32

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
- Entrainement sur "Galaxy" !!!

École de l'ADN

Galaxy PROJECT

Les Datas dans Galaxy

Les Datas dans Galaxy

Les Datas dans Galaxy

Voir les variantes
d'enregistrement de datas
en ligne directement

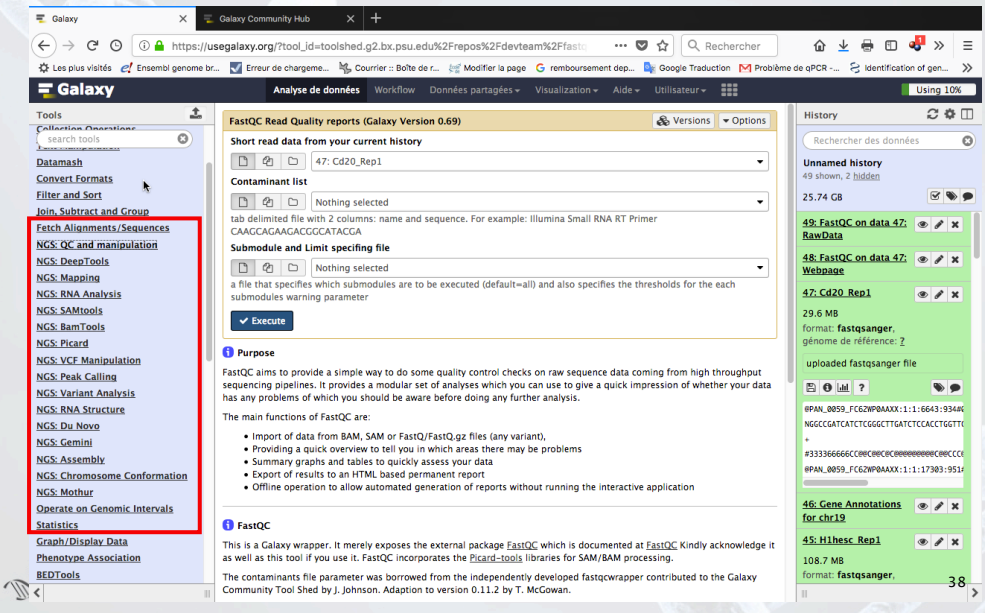
Analyses NGS avec Galaxy

- 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
- Entrainement sur "Galaxy" !!!

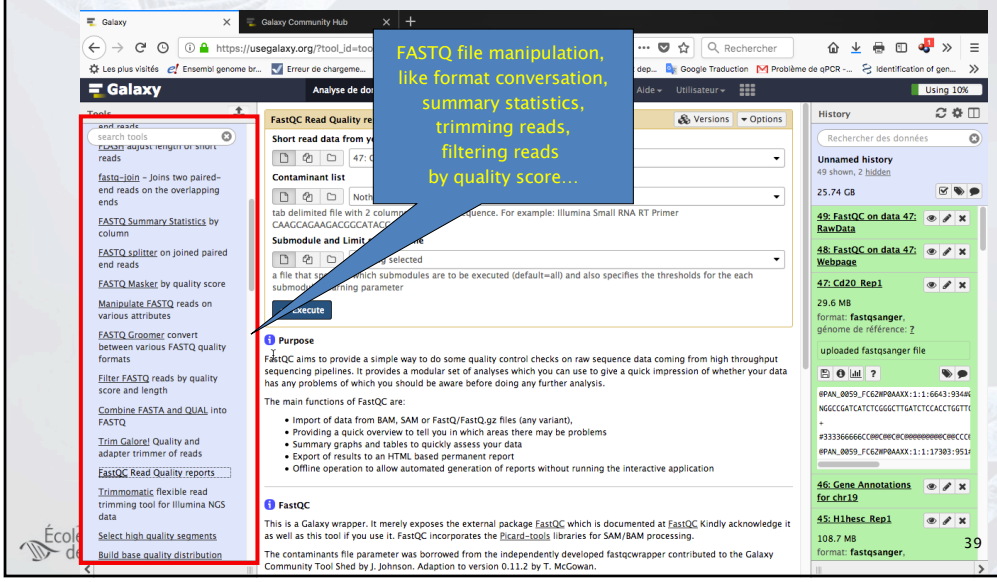
École de l'ADN

Galaxy PROJECT

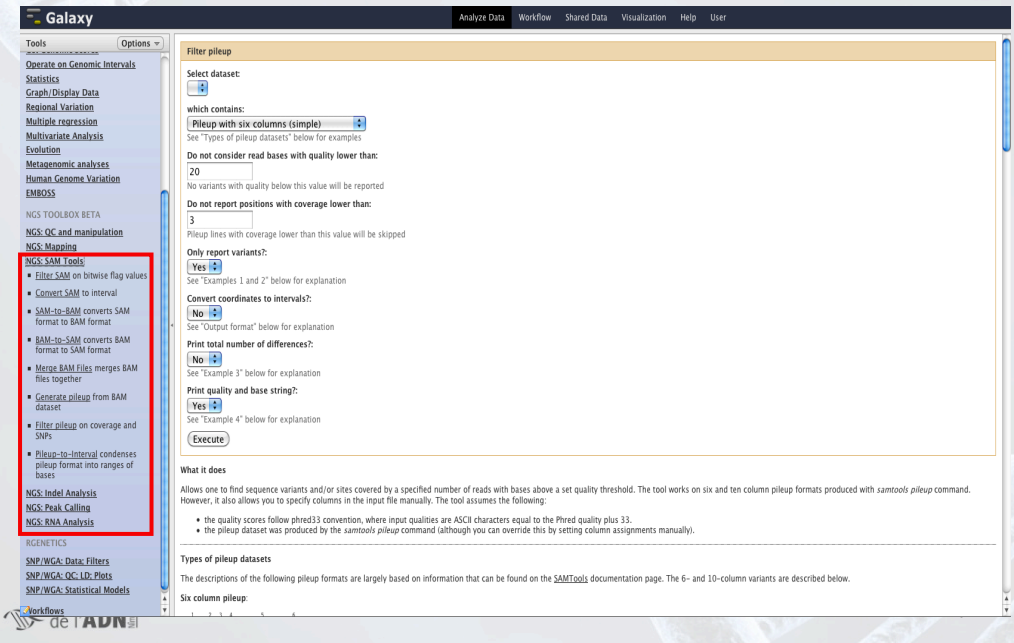
analyse de Data: Next Generation Sequencing



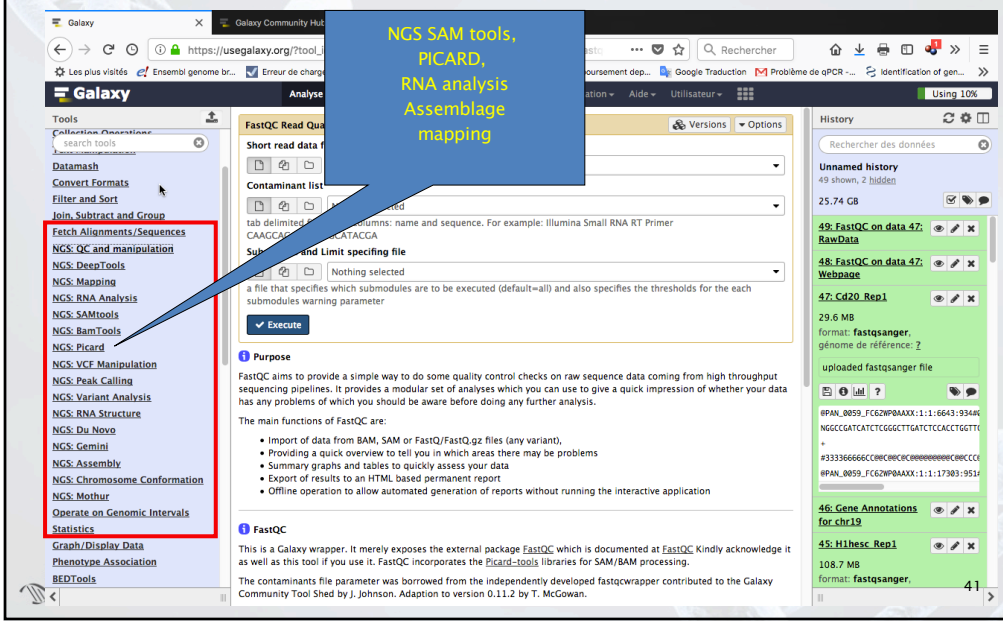
analyse de Data: Next Generation Sequencing



Analyzing Data: Next Generation Sequencing



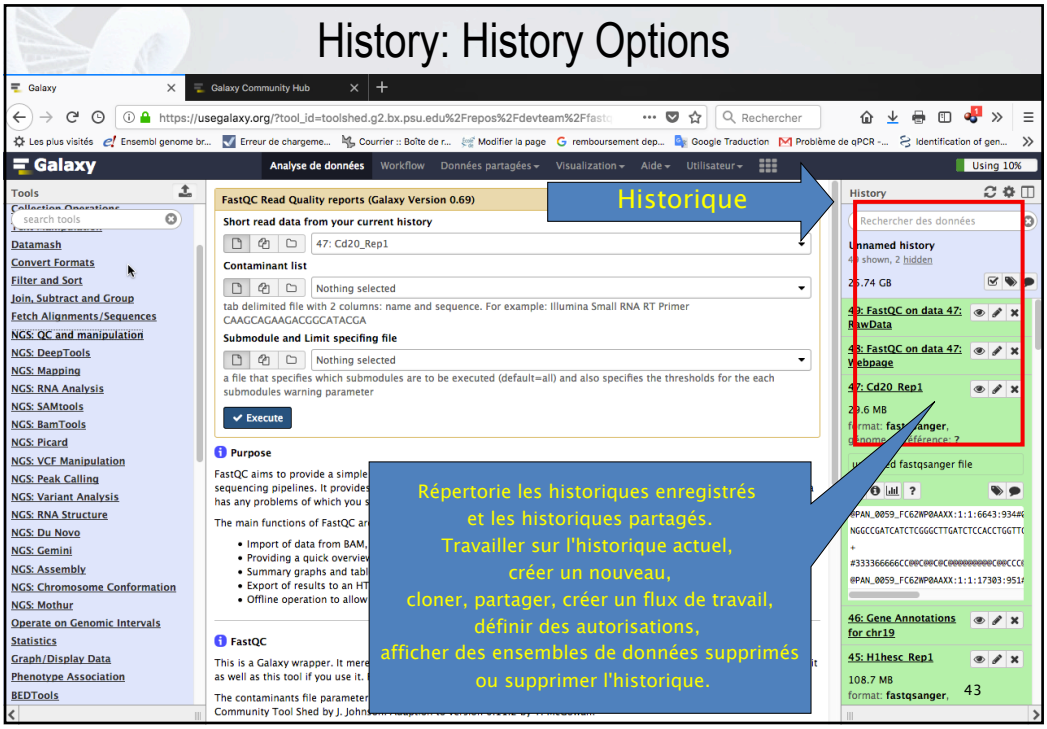
analyse de Data: Next Generation Sequencing



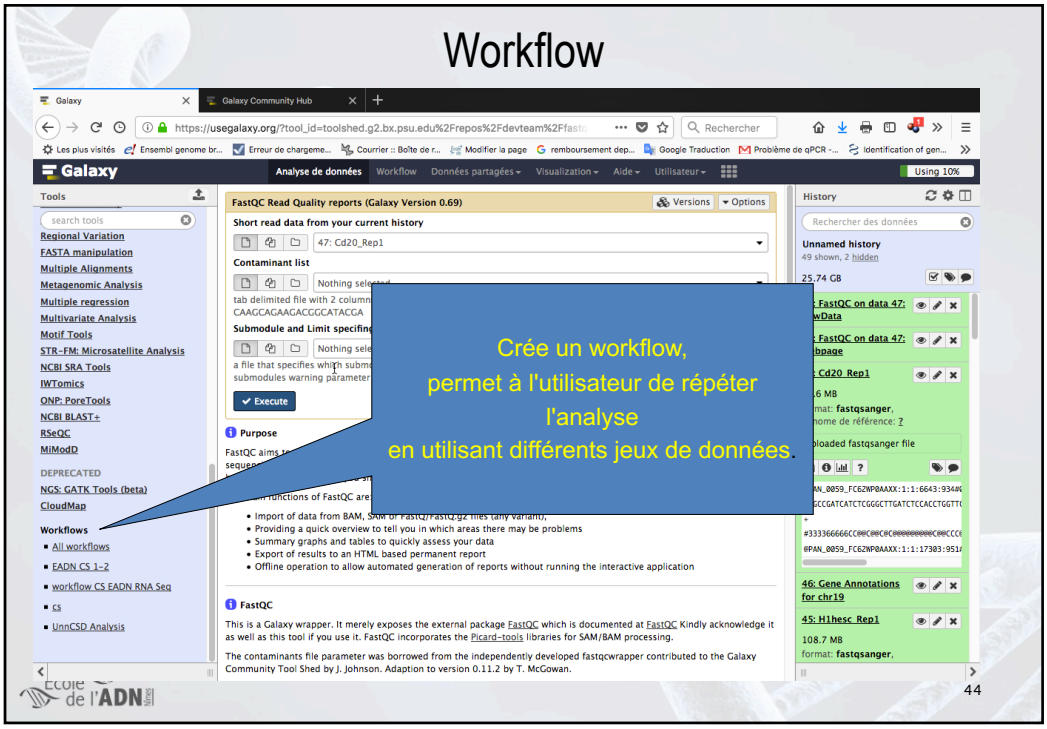
Analyses NGS avec Galaxy

- 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
- Entrainement sur "Galaxy" !!!

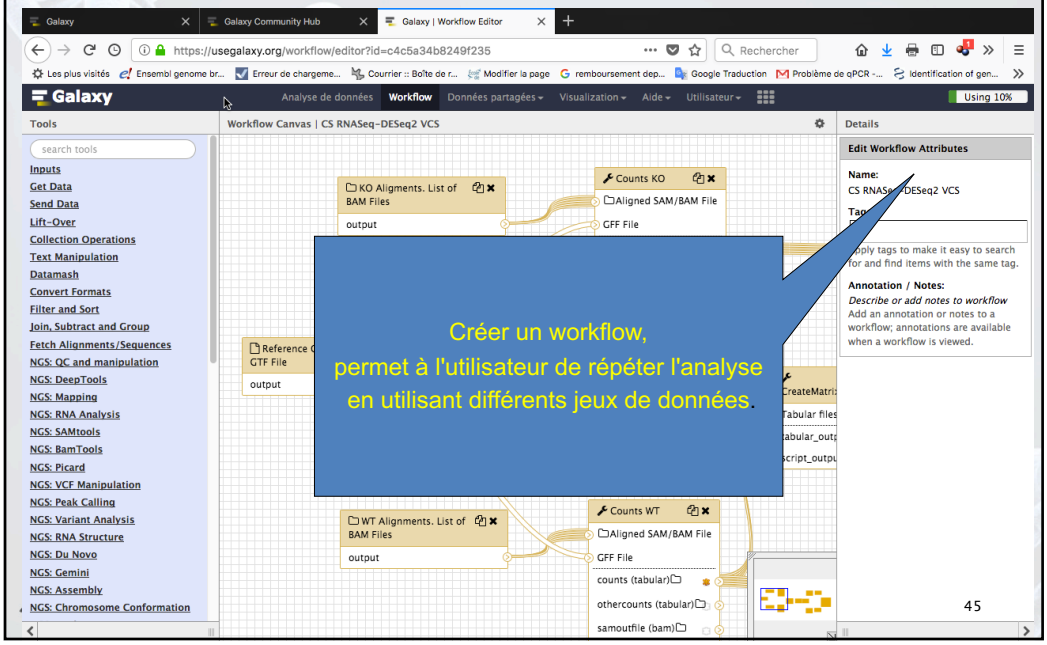
History: History Options



Workflow



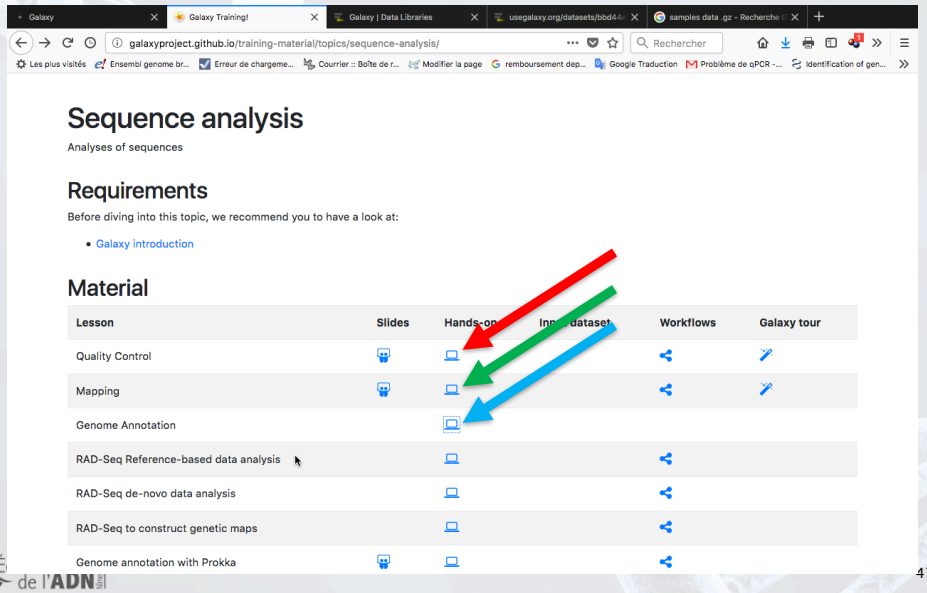
Workflow



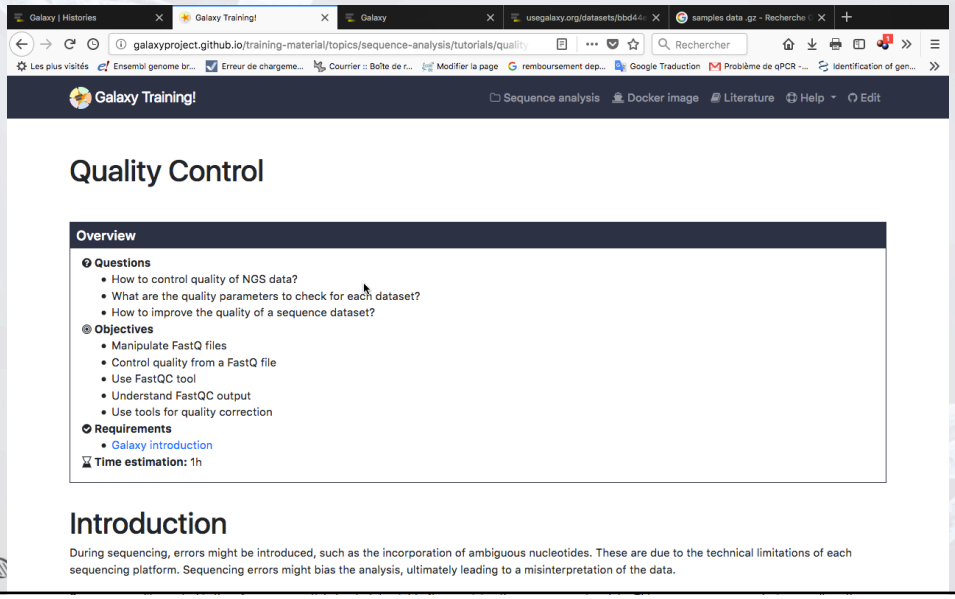
Il n'y a plus qu'à…

École de l'ADN
46

Il n'y a plus qu'à…



Il n'y a plus qu'à… 1

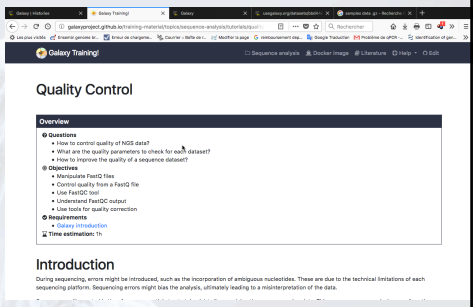


Il n'y a plus qu'à... 1

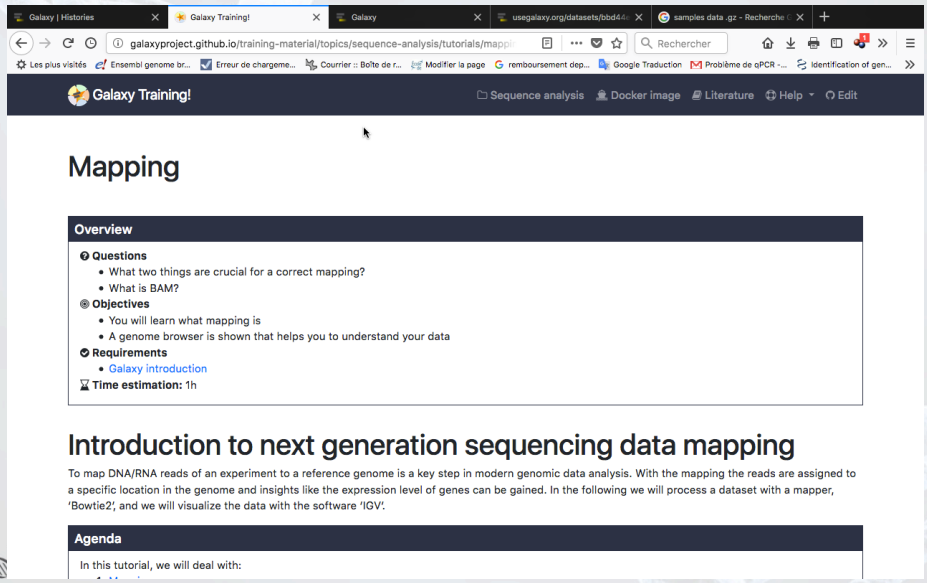
Dans ce tuto, nous allons vérifier la qualité de deux jeux de données pour nous assurer que les données sont correctes avant d'en déduire des informations supplémentaires.

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



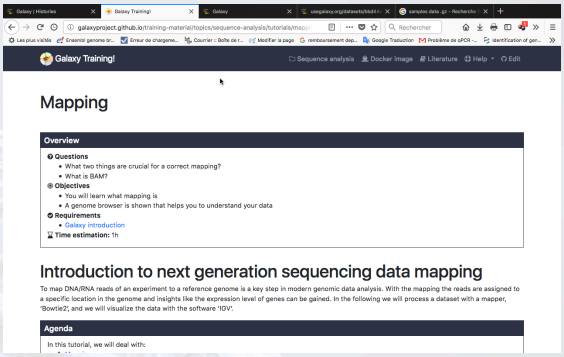
Il n'y a plus qu'à... 2



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'**.

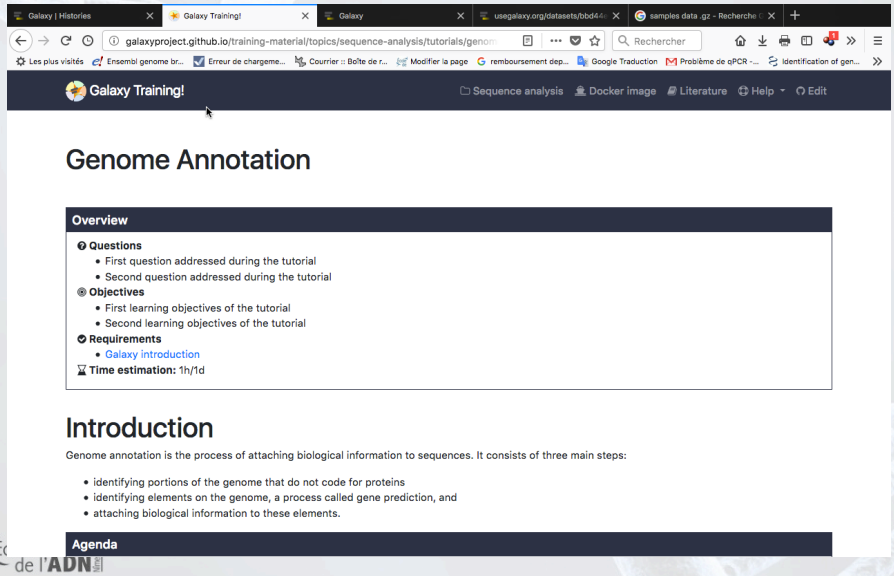


Integrative Genomics Viewer

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

51

Il n'y a plus qu'à…3



52


That's all Folks!
An Ecole de l'ADN
CARTOON