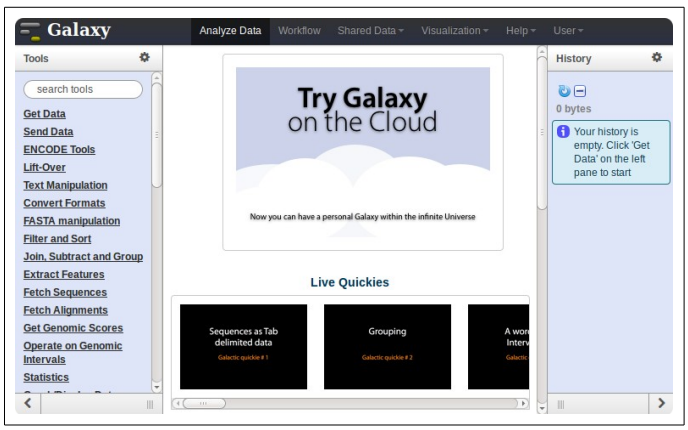
- Galaxy -

Initiation à la plateforme Galaxy

- EXERCICES -



Galaxy plateforme de traitements informatiques et bioinformatiques accessible depuis l'url : http://sigenae-workbench.toulouse.inra.fr/

Objectifs :

Cette formation a pour objectif de vous familiariser à l'utilisation de votre workbench Galaxy (http://galaxy-workbench.toulouse.inra.fr).

Vous découvrirez notamment comment :

- Traiter des fichiers sans utiliser de ligne de commande
- Lancer des traitements bioinformatiques sans Linux

Pour réaliser l'ensemble de ces exercices, vous avez besoin :

- De vous connecter à la plateforme Galaxy en utilisant les login et mot de passe de votre compte « genotoul » : http://galaxy-workbench.toulouse.inra.fr
- Des fichiers disponibles sur NG6 et des supports disponibles sur : http://genoweb.toulouse.inra.fr/~formation/1_Galaxy_Initiation/

Vous pouvez utiliser vos identifiants et mots de passe de votre compte sur la plateforme bioinfo de Toulouse, ou bien utiliser un des comptes disponibles le temps de la formation :

- *Logins :* anemone arome aster bleuet camelia capucine chardon clematite cobee coquelicot cosmos cyclamen
- *Password :* **Demander au formateur.**

Rappel : Ces comptes ne sont valables que le temps de la formation. Nous vous demandons d'utiliser un compte personnel si vous avez besoin de traiter ou stocker des données.

Pour répondre à vos questions:

- Mail : sigenae-support@listes.inra.fr
- Une FAQ et un manuel utilisateur sont disponibles depuis la page d'accueil de l'instance Sigenae de Galaxy.
- Les formations de la plateforme Bioinfo Genotoul sont disponibles sur http://sig-learning.toulouse.inra.fr

En fin de formation, penser à nettoyer votre compte de formation (« Delete permanently ») de l'ensemble des « histories » créés.

Exercice n°1 : Connexion à Galaxy, exploration de l'interface, téléchargement de datasets.

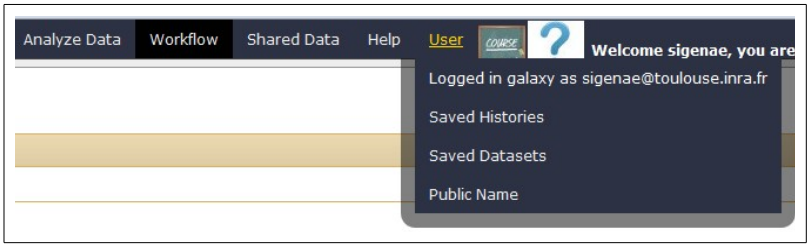
Connexion à la plateforme Galaxy

Vous pouvez accéder à votre plateforme Galaxy (en précisant votre login et mot de passe « genotoul ») à l'adresse suivante : http://galaxy-workbench.toulouse.inra.fr

Explorer l'interface

Depuis la barre du menu principal, vous avez accès aux onglets suivants :

- **Analyse Data** : Pour télécharger vos fichiers de données privées, et utiliser des modules de traitements.
- **Workflow** : Liste vos workflows archivés.
- **Shared Data** : Accès aux bibliothèques de données, ainsi qu'aux historiques et workflows publiés.
  - Data Libraries
  - Published Histories
  - Published Datasets
- **Help** :
  - Support
  - Galaxy Wiki
  - Video tutorials
  - How to cite Galaxy
- **User** :
  - Logged in galaxy as sigenae@toulouse.inra.fr
  - Saved Histories
  - Saved Datasets
  - Public Name



Note : La documentation autour de Galaxy est très aboutie, explorer le menu « help » et notamment la rubrique « Video tutorials »...

Afin de vous permettre une meilleure prise en main de l'interface Galaxy, nous vous encourageons à rechercher les outils à l'aide du menu « Options » - « Show Tool Search » disponible dans la partie « Tools » tout à gauche de l'interface.

Import de données

1 Téléchargement de fichier avec copie sur le serveur (non recommandé)

Télécharger, avec « Upload File », les fichiers « 454.fastq » via l'url http://genoweb.toulouse.inra.fr/~formation/15_FROGS/ (choisir le mois correspondant).

Il est possible de renseigner plusieurs URL en sautant une ligne entre chaque URL puis exécuter l'outil.

Renommer le dataset en « 454.fastq » puis renommer l'historique en « historique 454 »

L'ensemble des outils permettant l'import dans Galaxy est disponible dans la section « 1- Upload your data => Get data »

L'outil « **Upload File** » télécharge en copiant votre fichier sur le serveur Galaxy. Cette copie diminue votre quota Galaxy.

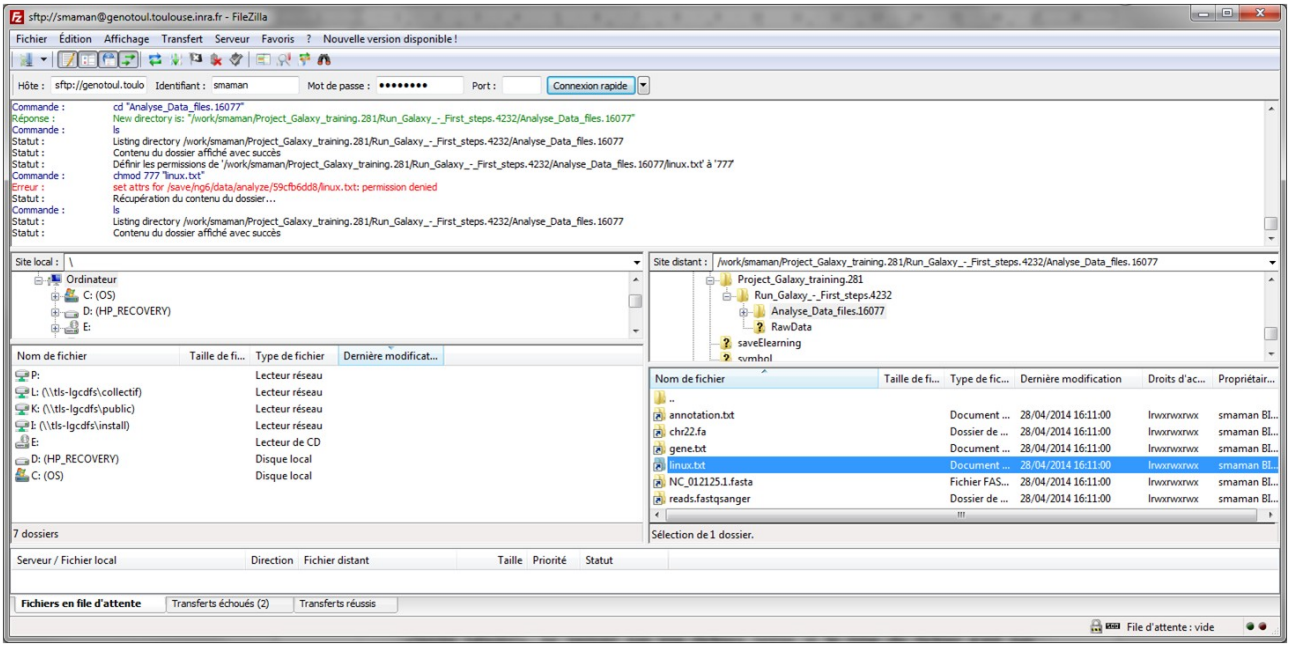
Vos fichiers de données téléchargés apparaîtront dans votre historique courant et seront automatiquement archivés dans « User / Saved Datasets ».

2 Télécharger des données depuis vers ordinateur sur votre /work puis dans Galaxy

Tout d'abord, veuillez récupérer les fichiers sur votre ordinateur :
- Liste des fichiers à récupérer sur votre ordinateur :
Depuis http://genoweb.toulouse.inra.fr/~formation/15_FROGS/ : veuillez récupérer les fichiers sampleA_R1.fastq et sampleA_R2.fastq, multiplex.fastq et barcode.tabular
Depuis http://genoweb.toulouse.inra.fr/~formation/1_Galaxy_Initiation/Data/, veuillez récupérer l'ensemble des fichiers disponibles.
- Pour chaque fichier, clic droit sur le nom du fichier pour obtenir l'adresse de téléchargement, puis « enregistrer la cible du lien sous » (« Copy link location »), puis enregistrer chaque fichier sur votre ordinateur.

Depuis FileZilla ou WinSCP, vous pouvez accéder à vos fichiers sur le serveur genotoul avec les paramètres suivants :
• Hôte : genotoul.toulouse.inra.fr
• Identifiant : Votre login sur genotoul
• Mot de passe : Votre mot de passe sur genotoul
• Port : 22

Parcourer votre ordinateur à gauche et le serveur Genotoul à droite.



A droite, dans Genotoul, créer un repertoire galaxy/ dans votre /work, puis déplacer l'ensemble des fichiers dans ce répertoire galaxy/ nouvellement créé.

A l'aide d'un clic droit sur le répertoire /galaxy, veillez à ce que le répertoire galaxy/ ai bien les permissions « x » (execution) pour tous. De même pour votre /work. Puis vérifier, de la même manière, que l'ensemble des fichiers contenus dans galaxy/ (uniquement) aient bien les droits « r »(lecture).

Si cela n'est pas fait, l'outil d'upload de Galaxy ne sera pas en mesure d'accéder ni de lire les fichiers que vous souhaitez télécharger et sur lesquels vous allez travailler. Après l'éxécution de l'outil « upload » de Galaxy, la dataset sera rouge (donc en erreur).

Créer deux nouveaux historiques :
- un nouvel historique renommé « historique R1R2 ».
- un autre nouvel historique renommé « TP ini Galaxy ».
- un troisième nouvel historique renommé « historique multiplex ».

Utiliser l'outil « Upload File from Genotoul » afin créer le lien dans votre historique galaxy. Cette méthode de téléchargement de fichiers dans Galaxy entamme beaucoup moins votre quota que la méthode exposée précédemment.

L'historique « TP ini Galaxy » comprendra l'ensemble des fichiers récupérés depuis http://genoweb.toulouse.inra.fr/~formation/1_Galaxy_Initiation/Data/
L'historique « historique R1R2 » comprendra les fichiers sampleA_R1.fastq et sampleA_R2.fastq récupérés depuis http://genoweb.toulouse.inra.fr/~formation/15_FROGS/
L'historique « historique multiplex » comprendra les deux fichiers multiplex.fastq et barcode.tabular récupérés depuis http://genoweb.toulouse.inra.fr/~formation/15_FROGS/

L'outil « Upload local file from filesystem path Upload data to history without copying on server » vous permet de créer un lien symbolique, depuis votre work, sur le serveur Galaxy, sans avoir besoin de copier vos données sur le serveur Galaxy.
Grâce à cet outil, vous économisez de l'espace disque et optimisez votre quota sur Galaxy.

Important – les droits : Les droits d'exécution sur le répertoire et de lecture sur les fichiers sont nécessaires pour que vos données puissent être accessibles dans Galaxy. (chmod +x REPERTOIRE et chmod +r FICHIER)

Chemin d'accès à linux.txt : Le chemin doit être complet (nom du fichier compris) et pointer sur le work (et non sur le /save ou le /home) afin que le cluster puisse, par la suite, travailler sur ce fichier.

Important – les formats de fichier : Les outils Galaxy qui prennent en entrée des fichiers « textes tabulés », ne verront pas vos fichiers textes si le type du fichier n'est pas correctement spécifié (format « tabular »).

3 Télécharger des données de l'UCSC : « UCSC Main table browser »

Télécharger l'annotation (gènes RefSeq) du chromosome 1 bovin (btau4), paramètres :

- Clade : Mammal
- Genome : Cow
- Assembly : Oct. 2007
- Group : Genes and Genes Prediction Tracks
- Track : RefSeq Genes
- Table : refGene
- Region – position : chr1:1-161106243 (enter « chr1 » puis cliquer sur « lookup »)
- Output format : BED – browser extensible data
- Sélectionner « Send Output to Galaxy » puis cliquer sur « Get Output » et « Send query to Galaxy ».

Visualiser le nouveau dataset, et notamment ses propriétés (« database »). Que remarquez vous ?

Explorer les liens disponibles pour ce dataset.

Relancer le téléchargement en modifiant le Output format : GTF – gene transfert format
Comparer les deux fichiers GTF et BED.

4 Télécharger des données compressées avec l'outil « FROGS Upload archive from your computer »

Créer un nouvel historique et renommer le en « historique contiged ».

A l'aide de cet outil « FROGS Upload archive from your computer », veuillez récupérer l'archive « 100spec_90000seq_9samples.tar.gz » disponible depuis cette URL :
http://genoweb.toulouse.inra.fr/~formation/15_FROGS/

Renommer la dataset Galaxy : « 100spec_90000seq_9samples.tar.gz ».

Exercice n°2: Utilisation d'outils de traitement de fichiers (équivalent aux commandes Linux)

Outils de traitement de fichiers

- En utilisant l'outil « Add column to an existing dataset » ajouter une colonne « chr1 » au fichier « linux.txt »
  - Ajouter la colonne
  - Renommer le dataset obtenu en « linux_add »
- Trier numériquement le fichier « linux_add » par ordre descendant sur la première colonne
  - Outil « Sort data in ascending or descending order »
  - Renommer le fichier généré en « linux_add_sort »
- Filtrer le fichier « linux_add_sort » pour ne conserver uniquement les lignes commençant

par 1, 2 ou 3
  - Deux outils possibles :
    - « Select lines that match an expression »
    - « Filter data on any column using simple expressions »
  - Renommer le fichier généré en « linux_add_sort_filter »
- Joindre, soustraire et grouper
  - Joindre les fichiers « annotation.txt » et « gene.txt », en utilisant l'outil « Join two Datasets side by side on a specified field », sur la colonne « gene »
  - Renommer le fichier obtenu en « annot_gene.txt ».

Outils bioinformatiques

- A partir des deux datasets BED et GTF « UCSC Main on Cow: refGene (chr1:1-161106243) » précédemment importés, extraire du génome la séquence de chacun des gènes.
  Utilisation de l'outil « Extract Genomic DNA » de la section « Fetch sequences » (output data type « Interval »)
- Comparer le nombre de lignes dans les nouveaux datasets avec ceux d'origine. Pourquoi cette différence ?
- Convertir le dataset obtenu à partir du BED en multi-fasta.
  Utilisation de l'outil « Tabular-to-FASTA converts tabular file to FASTA format »

- Calculer le %GC des gènes (outil "Compute GC content")
  Utilisation de l'outil « Compute GC content »
- Calculer la longueur de chaque gène
  Utilisation de l'outil « Compute sequence length »
- Produire un fichier tabulé de trois colonnes : GeneName<tab>Lenght<tab>GC%

- Régions promotrices. Construire un multi-fasta des régions promotrices.
  - A partir du fichier d'annotation BED (sans la séquence) utilser l'outil « Get flanks » pour extraire les régions en amont de chaque gène (longueur 1kb avec un offset de 100pb)
  - Produire le multi-fasta

Exercice n°3: Création et partage de datasets, d'historiques et de workflows.

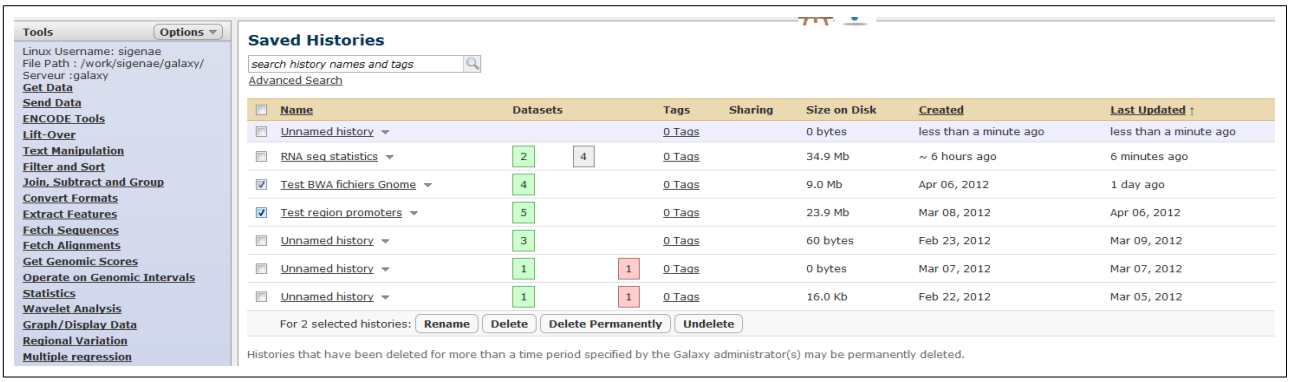
Notions d'historique

Traitements archivés dans un historique

Au fur et à mesure que vous faites appel aux différents outils au sein de votre interface depuis le menu « Analyse Data », l'ensemble des étapes sont enregistrées dans un historique qui est automatiquement archivé dans « User / Saved Histories » et que vous pouvez ensuite, si besoin, partager dans « Shared Data / Published Histories ».

Gérer ses historiques

Depuis le menu « User » / « Saved Histories », vous avez la possibilité de gérer vos historiques (delete, delete permanently, rename, undelete) en cliquant sur l'intitulé de l'historique.
Remarque, lors de votre connexion au workbench Galaxy, un « current history » est automatiquement créé.



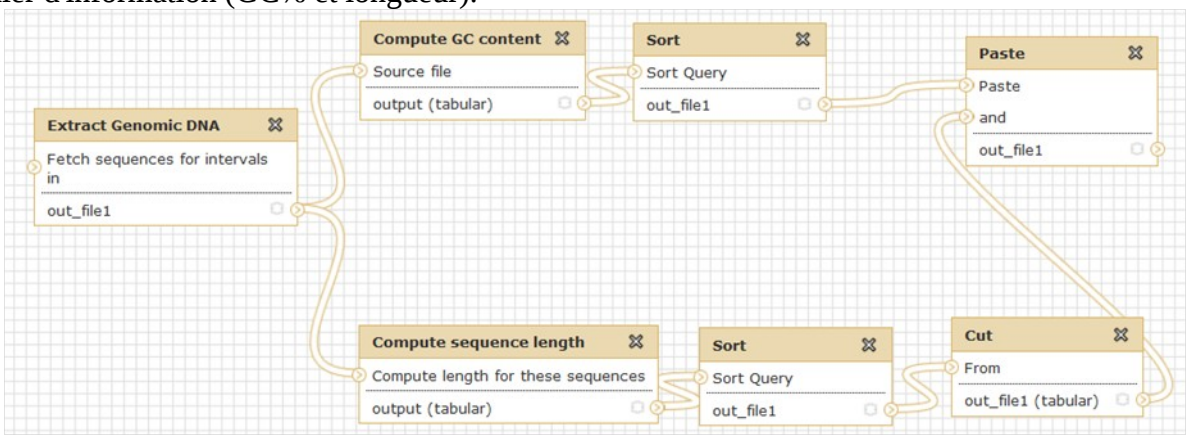
Exercice

- Créer un nouvel historique (nommer le en le préfixant par votre login) et ajouter (copie) un ou plusieurs de vos datasets
- Partager ce nouvel historique avec votre voisin
- Copier et modifier l'historique de votre voisin. Est-ce que cette modification impacte l'historique d'origine sur votre interface Galaxy ? Sur l'interface Galaxy de votre voisin ?

Notions de workflow : convertir un historique en workflow.

Convertir un historique en workflow

Créer un workflow à partir des traitements bioinformatiques précédemment réalisés. Soit un workflow permettant, à partir d'une annotation (format BED), de générer un multi-fasta ainsi qu'un fichier d'information (GC% et longueur).



Les principales étapes :

- « History panel » Options → « Extract workflow »
- Sélectionner les bons datasets
- Créer le workflow

Création de workflow :

- A partir de rien : Menu « Workflow » puis « Create a new workflow »
- A partir d'un historique : « History panel » Options → « Extract workflow »

Comme pour les historiques, il est possible de partager des workflows.

Récupérer ses fichiers dans son work/

Avec FileZilla, créer un répertoire GALAXY dans votre work : /work/user/GALAXY/
Donner les droits d'écriture à votre répertoire GALAXY/

Puis, depuis votre interface Galaxy, utiliser l'outil «Download my Galaxy dataset on Genotoul (work) » de la section « Download Data » pour récupérer quelques datasets de votre historique Galaxy dans votre work/
Quelle est la différence entre votre work et votre save ?

Conseils pour gérer au mieux votre quota

Pour vous aider à gérer votre espace de travail, veuillez vous connecter à la plateforme d'auto-formations en ligne http://sig-learning.toulouse.inra.fr, vous inscrire à la session « Galaxy », puis lire le chapitre « GOOD PRATICE or How to be a good Galaxy user ? »