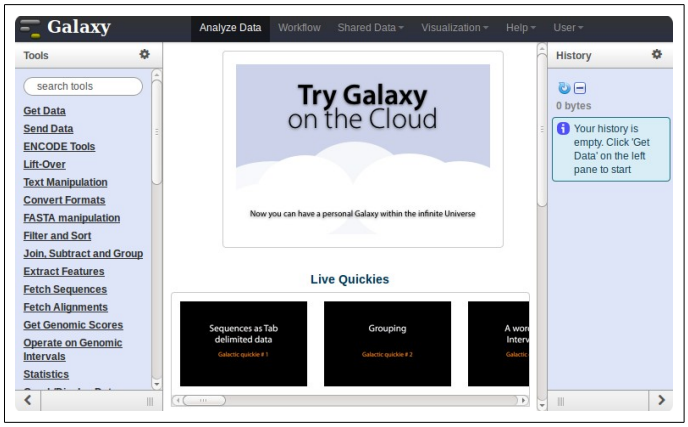
- Galaxy -

Initiation à la plateforme Galaxy

- EXERCICES -



Galaxy plateforme de traitements informatiques et bioinformatiques

Objectifs :

Cette formation a pour objectif de vous familiariser à l'utilisation du workbench Galaxy de l'instance SIGENAE (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

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:* Cosmos cyclamen dahlia digitale geranium gerbera glaieul hortensia iris jacinthe
- pervenche rose tulipe violette renoncule reine sauge trefle
- *Password:* **f1o2r3!**

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.
- Certaines formations de la plateforme Bioinfo Genotoul sont disponibles sur http://sig-learning.toulouse.inra.fr

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 – Workflow - Shared Data – Help - User

L'onglet Visualization n'est pas fonctionnel pour le moment.

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 des fichiers avec copie sur le serveur (non recommandé)

- Télécharger, avec «Upload File», les fichiers «reads.fastqsanger», «NC_012125.1.fasta», «annotation.txt», «linux.txt» et «gene.txt» disponibles via l'url http://genoweb.toulouse.inra.fr/~formation/1_Galaxy_Initiation/Data/
- Renommer les datasets.

Pour obtenir l'adresse de téléchargement, faites un clic droit sur le lien de téléchargement, puis « Copy link location ».

2 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.

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 : Partage

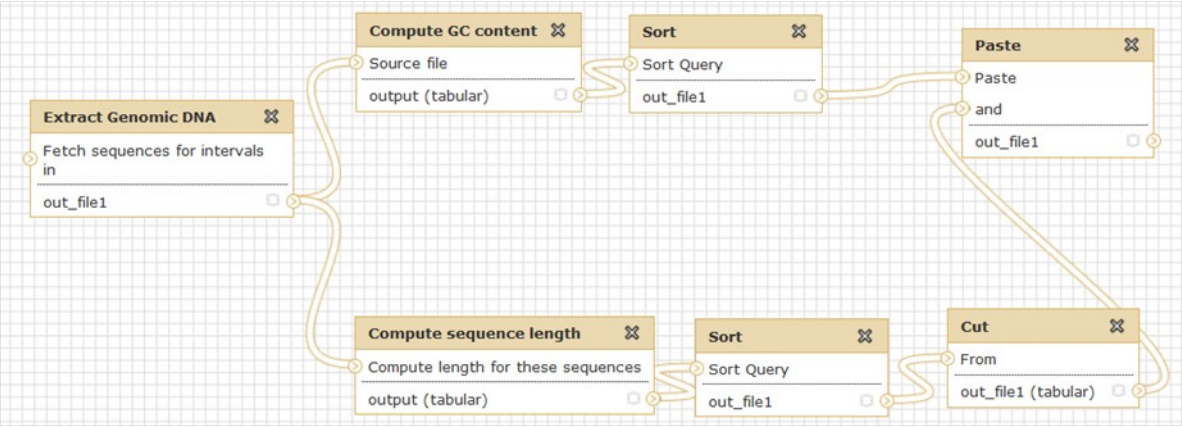
Partage d'historiques

- Lister les historiques partagés sur votre compte à partir du menu "Shared data" / "Published histories".
- Partager un historique avec votre voisin: «User / Saved Histories» → «Share or Publish»

Exercice n°4: Workflow

Notions de workflow : convertir un historique en workflow.

Il vous est demandé de 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.

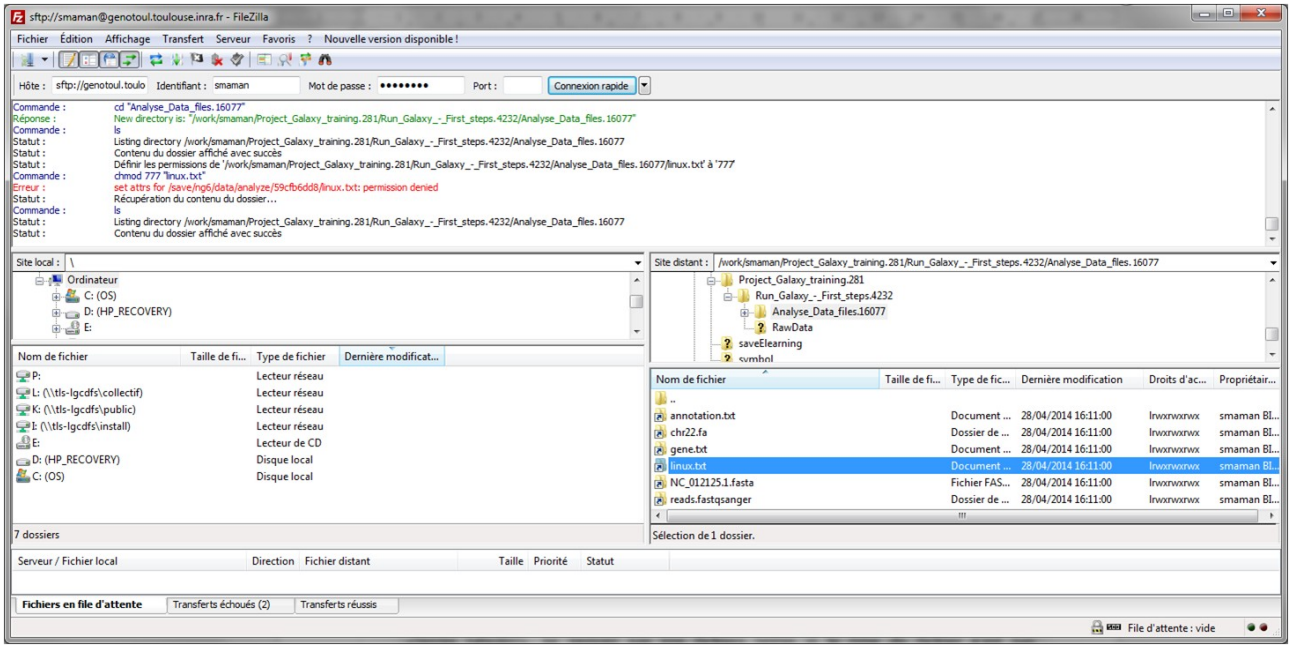
Exécuter un workflow

- Lancer le workflow sur l'annotation (gène RefSeq) du chromosome 2 bovin (btau4).
- Sauver les datasets générés dans votre compte sur Genotoul.
- Lister les fichiers sauvegardés.
- Sauvegarder les fichiers de sortie sur votre poste.

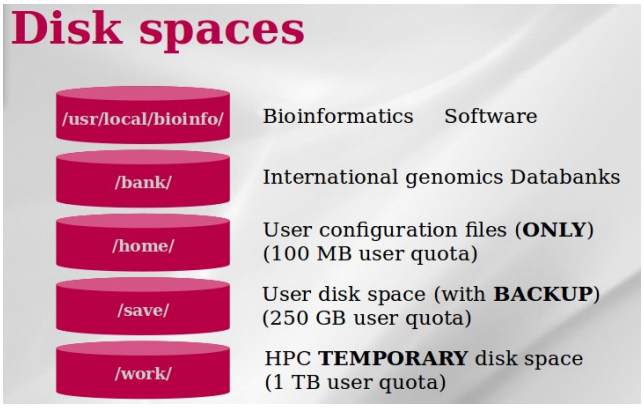
Exercice n°5: Travailler avec les fichiers disponibles sur votre compte genotoul

Téléchargement sans copie sur le serveur (recommandé)

- Grâce au logiciel FileZilla présent sur votre poste de travail, copiez un fichier sur /work/your_user/.



Architecture genotoul :



Le fichier est maintenant stocké sur votre espace de travail dédié sur le cluster.

- Utiliser l'outil « **Upload local file from filesystem path** Upload data to history without copying on server » afin créer le lien dans votre historique galaxy.

L'outil «Upload local file from filesystem path» 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.

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: 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.

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 »).