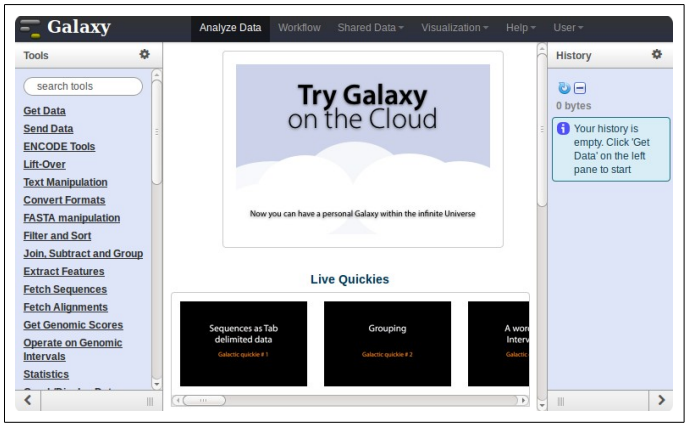
- Galaxy -

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, aster, bleuet, iris, muguet, narcisse, pensee, rose, tulipe, violette, lilas, pervenche, laurier, lavande, lis, capucine, coquelicot, geranium, liseron, arome, chardon
- *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.
- 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.

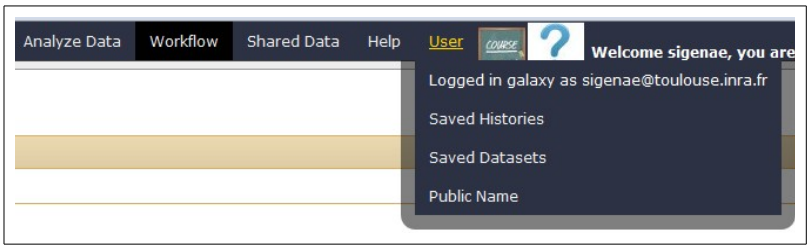
Connexion à la plateforme Galaxy

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 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 », « chr22.fa » et « gene.txt » disponibles via l'url http://genoweb.toulouse.inra.fr/~formation/1_Galaxy_Initiation/Data/

Renommer les datasets.

L'ensemble des outils permettant l'import dans Galaxy est disponible dans la section « Upload your data «

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

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

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

2 Télécharger des données de l'UCSC via l'outil « 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%

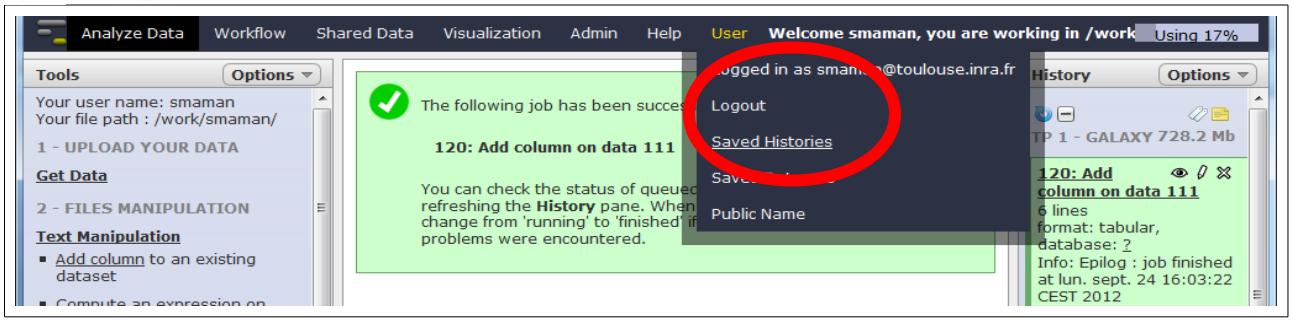
- Régions promotrices. Construire un multi-fasta des régions promotrices.
  - A partir du fichier d'annotation BED (sans la séquence) utilser l'outil « Get flanks » pour extraire les régions en amont de chaque gène (longueur 1kb avec un offset de 100pb)
  - Produire le multi-fasta

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

Notions d'historique

Traitements archivés dans un historique

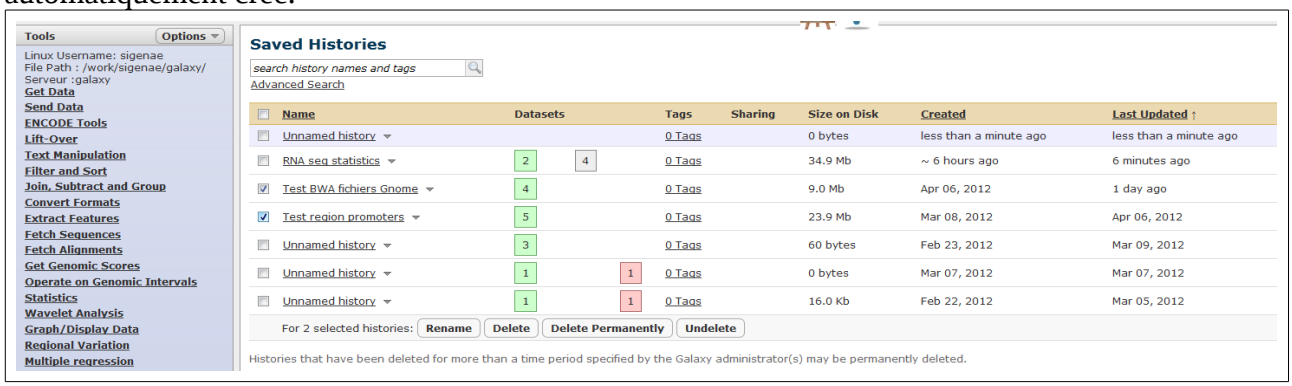
Au fur et à mesure que vous faites appel aux différents outils au sein de votre interface depuis le menu « Analyse Data », l'ensemble des étapes sont enregistrées dans un historique qui est automatiquement archivé dans « User / Saved Histories » et que vous pouvez ensuite, si besoin, partager dans «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



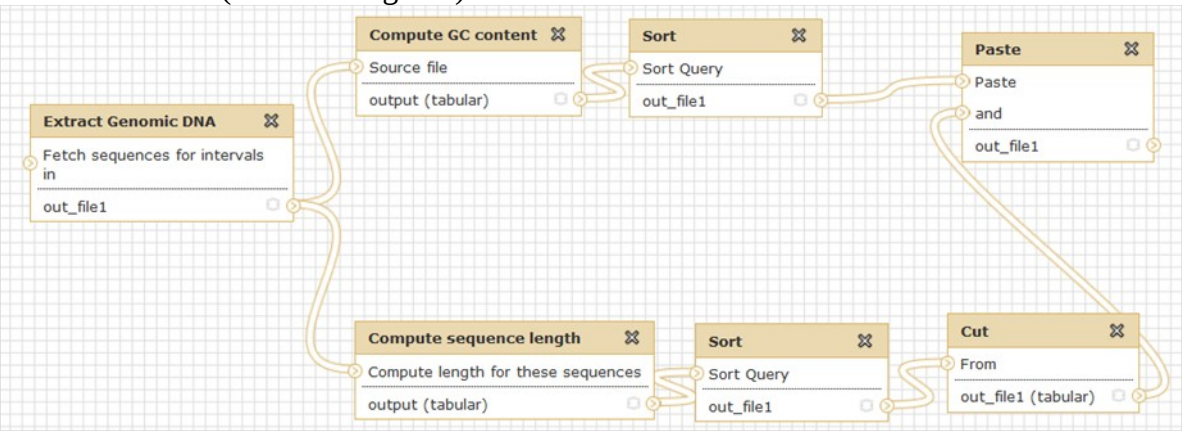
Exercice

- 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
- Importer l'historique de votre voisin
- Modifier votre historique
- Réimporter l'historique modifié de votre voisin
- Supprimer l'historique que vous partagez
- Supprimer les historiques importés

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

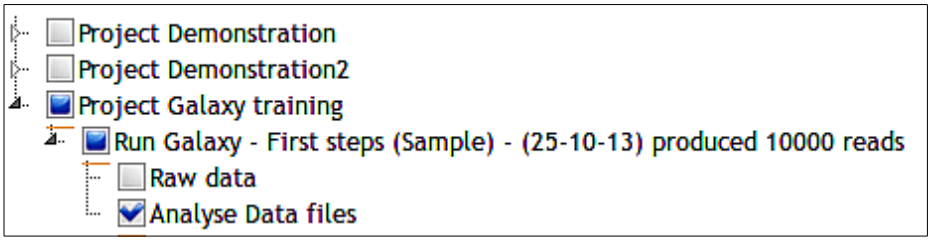
Création de workflow :

- 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.

Import de données par téléchargement sans copie sur le serveur (recommandé)

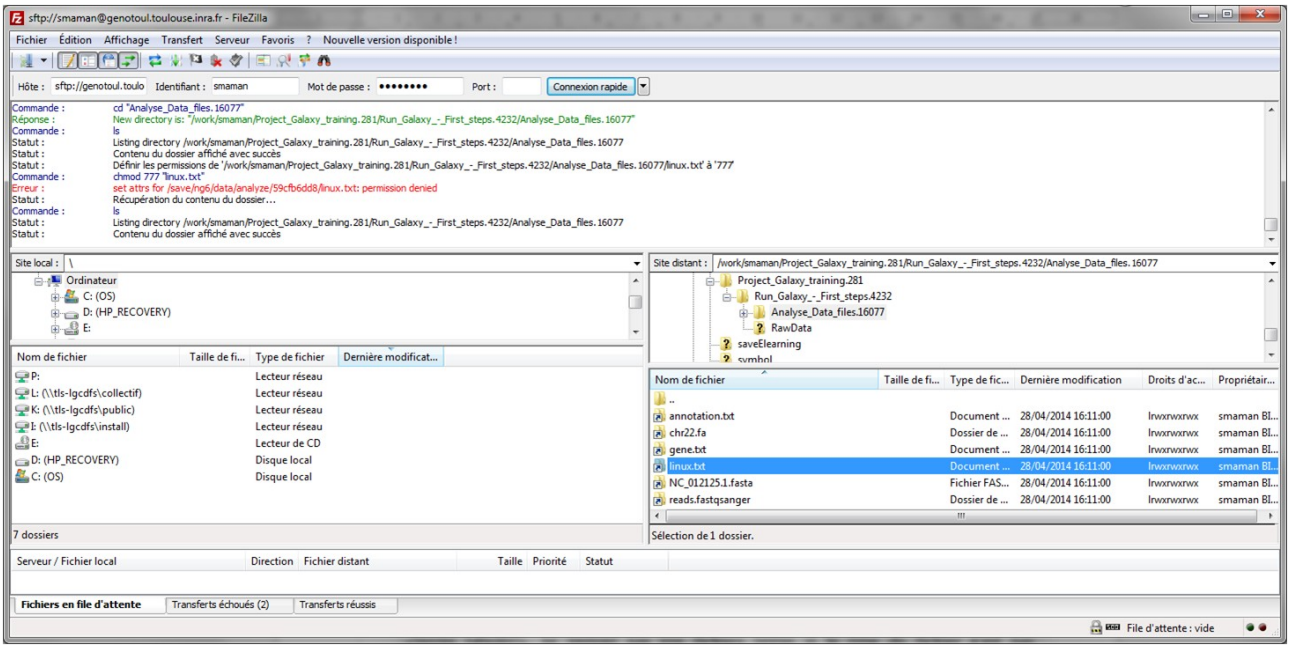
A partir la plateforme NG6 (http://ng6.toulouse.inra.fr), projet public « Galaxy training », créer les liens symboliques dans votre espace de travail sur le serveur genotoul :



Depuis FileZilla (ou WinSCP), vous pouvez accéder à vos fichiers sur le servuer 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

Parcourez votre espace de travail (logiciel ftp) sur le server genotoul afin de noter le chemin vers le fichier (lien) « chr22.fa » (ex : /work/smaman/Project_Galaxy_training.281/Run_Galaxy_-_First_steps.4232/Analyse_Data_files.16077/chr22.fa).



Utiliser l'outil « **Upload file** from genotoul » afin créer un lien vers le fichier « chr22.fa » dans votre historique galaxy.

L'outil « Upload file from genotoul » 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 »).

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, dans votre /work/YourUserName/.
Lister les fichiers sauvegardés.

A l'aide de WinSCP, veuillez déplacer vos fichiers de votre /work à votre /save :
1- Options / Remote Panel / Tree
2- Puis Duplicate / Cocher « Duplicate via local temporary copy »
Et dans le chemin, remplacer « work » par « save » avec le chemin souhaité sur le « save ».

Avec FileZilla, ce déplacement de fichier nécessite de passer par votre ordinateur donc de potentiellement dégrader vos fichiers. FileZilla est donc déconseillé pour cette étape.

Pourquoi ce déplacement de fichier est important ?

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 ? »*