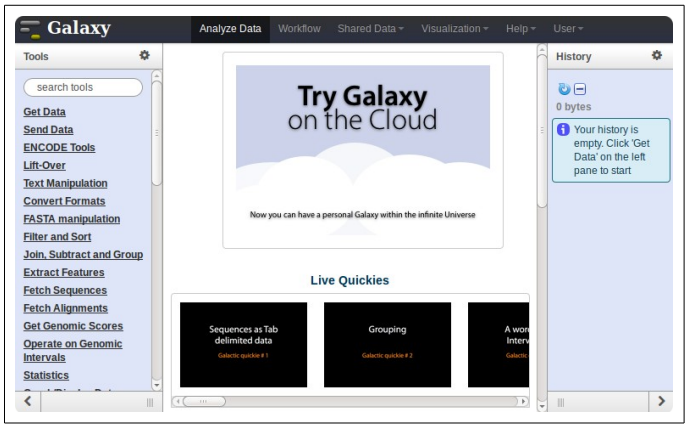
- Galaxy -

Initiation à la plateforme Galaxy

- EXERCICES -



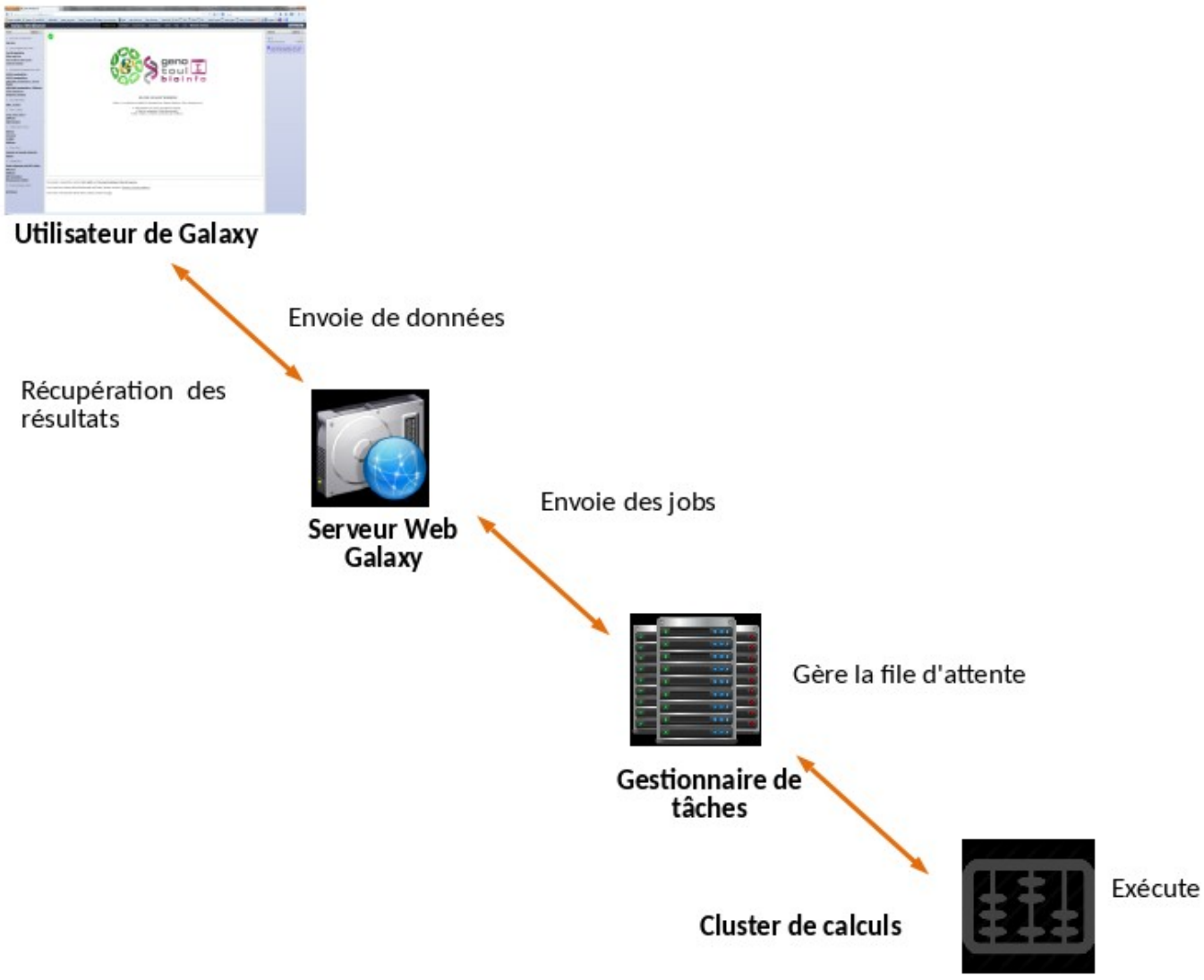
Galaxy plateforme de traitements informatiques et bioinformatiques accessible depuis l'url : http://sigenae-workbench.toulouse.inra.fr/

Quelques mots sur Galaxy

Galaxy a été créé par l'équipe américaine "Galaxy project" :

- Le Center for Comparative Genomics and Bioinformatics - Penn State,
- Des départements "Biology" et "Mathematics and Computer Science" de l'Université d'Emory.

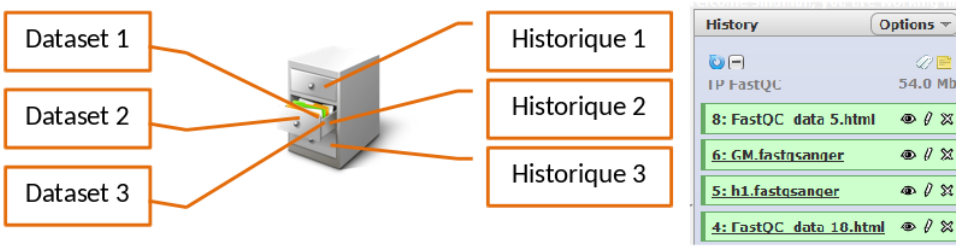
La communauté autour de cet outil est active. Nous utiliserons l'instance Sigenae/BioInfo Genotoul de Galaxy.



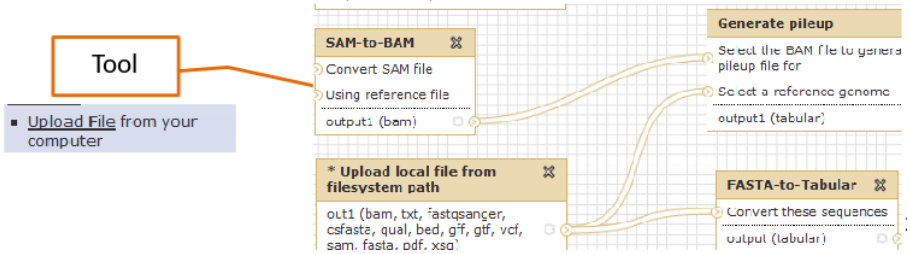
Nous utiliserons un vocabulaire spécifique à Galaxy :

Un **DATASET** est un fichier de données (fichiers d'entrée, fichiers résultats) :

Votre **HISTORIQUE** est un « répertoire » qui « liste » l'ensemble de vos fichiers de donnée (fichiers d'entrée, fichier résultat) utilisés ou générés par un **TOOL** :



Votre **WORKFLOW** est un ensemble : fichiers, outils, traitements.



Objectifs :

Cette formation a pour objectif de vous familiariser à l'utilisation de votre workbench Galaxy (http://galaxy-workbench.toulouse.inra.fr).
Cette formation est destinée aux personnes souhaitant traiter des données (bio)informatiques sans connaissances spécifiques en informatique (sans avoir à connaître Linux et la ligne de commande).

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.

Si votre mot de passe genotoul est trop complexe, n'hésitez pas à interpeller un formateur pour qu'ils vous le change

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 (petit écrou).

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

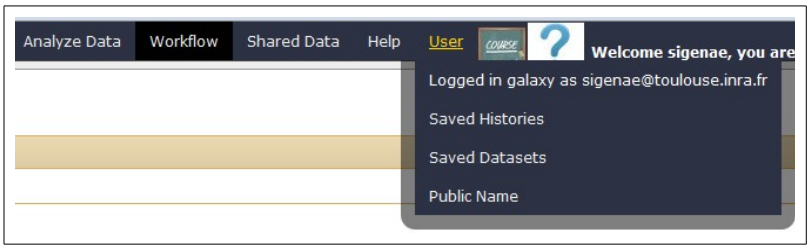
Connexion à la plateforme Galaxy

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 Préparer vos historiques de travail

Depuis le menu « History » à droite, cliquez sur « History options » (petit écrou), choisissez « Create new » et créer ainsi, un après l'autre, plusieurs historiques.
Afin de gérer au mieux son espace de travail et le suivi du TP, il est important de renommer vos historiques de la manière suivante en cliquant sur « Unnamed history » :
- « historique R1R2 »
- « TP ini Galaxy »
- « historique multiplex »
- « historique 454 »
- « historique contiged »

Entrer le nom de l'historique et liquer sur « Entrée » pour valider.

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

Dans l'« **historique 454** », nous allons télécharger, avec « Upload File », le fichier « 454.fastq » disponibles depuis Internet via l'url http://genoweb.toulouse.inra.fr/~formation/15_FROGS/ (choisir le mois correspondant à votre formation).

Pour récupérer un fichier disponible sur Internet, veuillez utiliser l'outil « Upload File » de Galaxy (outil disponible dans la section « Get Data »).

Cliquer sur « Paste/Fetch » data. Puis copier/coller l'URL COMPLETE d'accès au fichier pour le télécharger dans Galaxy. Exemple d'URL complète : http://genoweb.toulouse.inra.fr/~formation/15_FROGS/April2016/454.fastq

Une URL complète contient le chemin d'accès au fichier plus le nom et l'extension du fichier concerné. Pour récupérer une URL complète rapidement et sans faire d'erreur, nous vous conseillons de faire un clic droit sur le nom du fichier et de choisir « copier l'adresse du lien ». Puis un « coller » (Ctrl V) permet de récupérer cette URL complète. Puis lancer le téléchargement. Le fichier sera bientôt disponible dans votre fenêtre « History ».

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

Pour gérer vos fichiers de données, nous vous conseillons de renommer chaque dataset. Ainsi, veuillez s'il vous plaît renommer le FASTQ téléchargé en « 454.fastq ».

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 méthode d'upload n'est pas recommandée car elle impacte considérablement votre quota.

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

3 *Télécharger des données depuis votre ordinateur sur votre /work puis dans Galaxy*

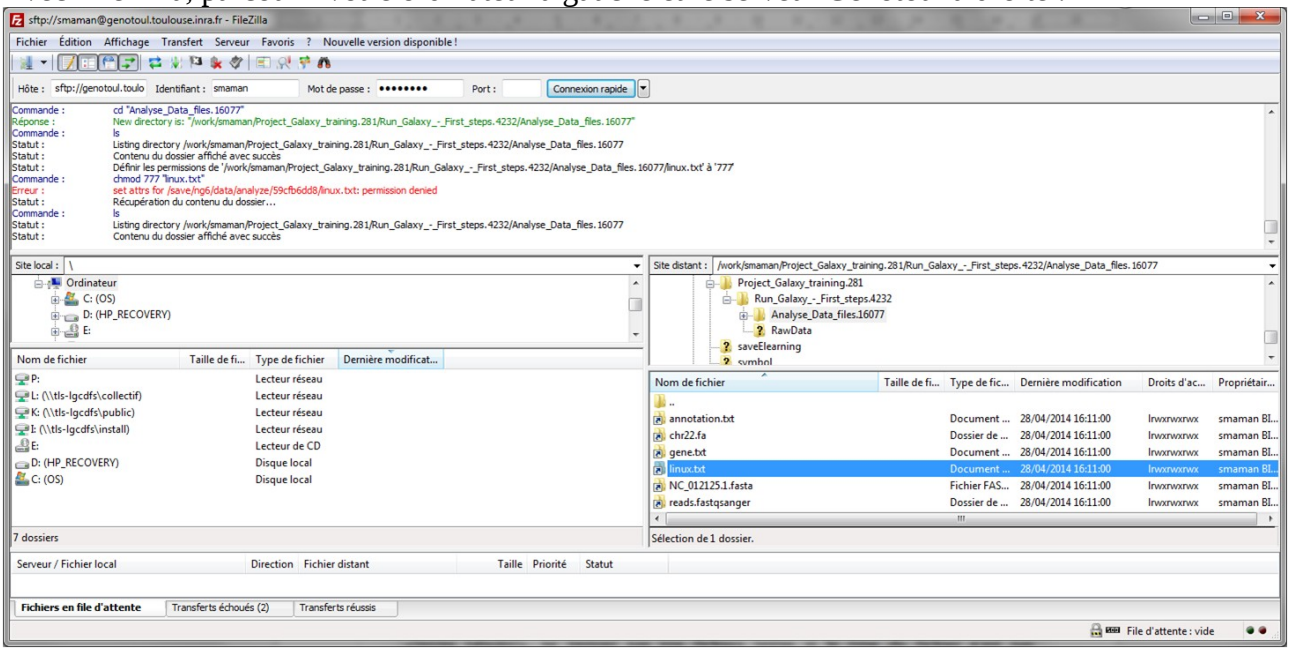
Dans l'« **historique multiplex** », nous allons récupérer les fichiers sampleA_R1.fastq et sampleA_R2.fastq, multiplex.fastq et barcode.tabular sur votre ordinateur.
Ces fichiers sont disponibles sur Internet depuis http://genoweb.toulouse.inra.fr/~formation/15_FROGS/

Pour chaque fichier, veuillez faire un 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.

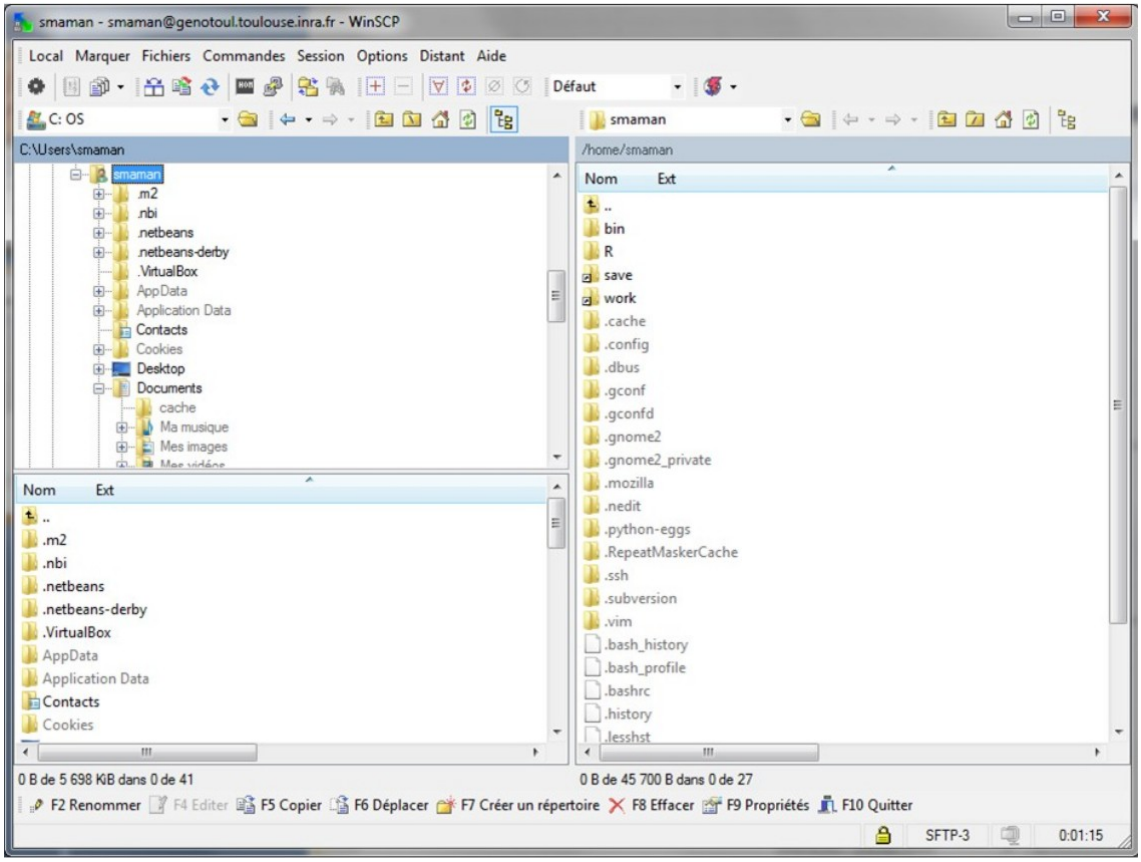
Il est conseillé de ne pas ouvrir ou modifier vos fichiers sous Windows car ce dernier ajoute des caractères spéciaux qui ne sont pas pris en charger par le cluster de calculs.

Nous allons ensuite transférer ces fichiers de votre ordinateur à Galaxy. Pour ce faire, veuillez ouvrir le logiciel WinSCP (ou FileZilla). Cet outil vous permettre de voir le contenu de vos répertoires sur Genotoul et de gérer les permissions sur ces répertoires et fichiers.
Pour vous connecter à Genotoul, veuillez compléter les paramètres suivants :
• Hôte : genotoul.toulouse.inra.fr
• Identifiant : Votre login sur Genotoul
• Mot de passe : Votre mot de passe sur Genotoul
• Port : 22

Avec FileZilla, parcourir votre ordinateur à gauche et le serveur Genotoul à droite :

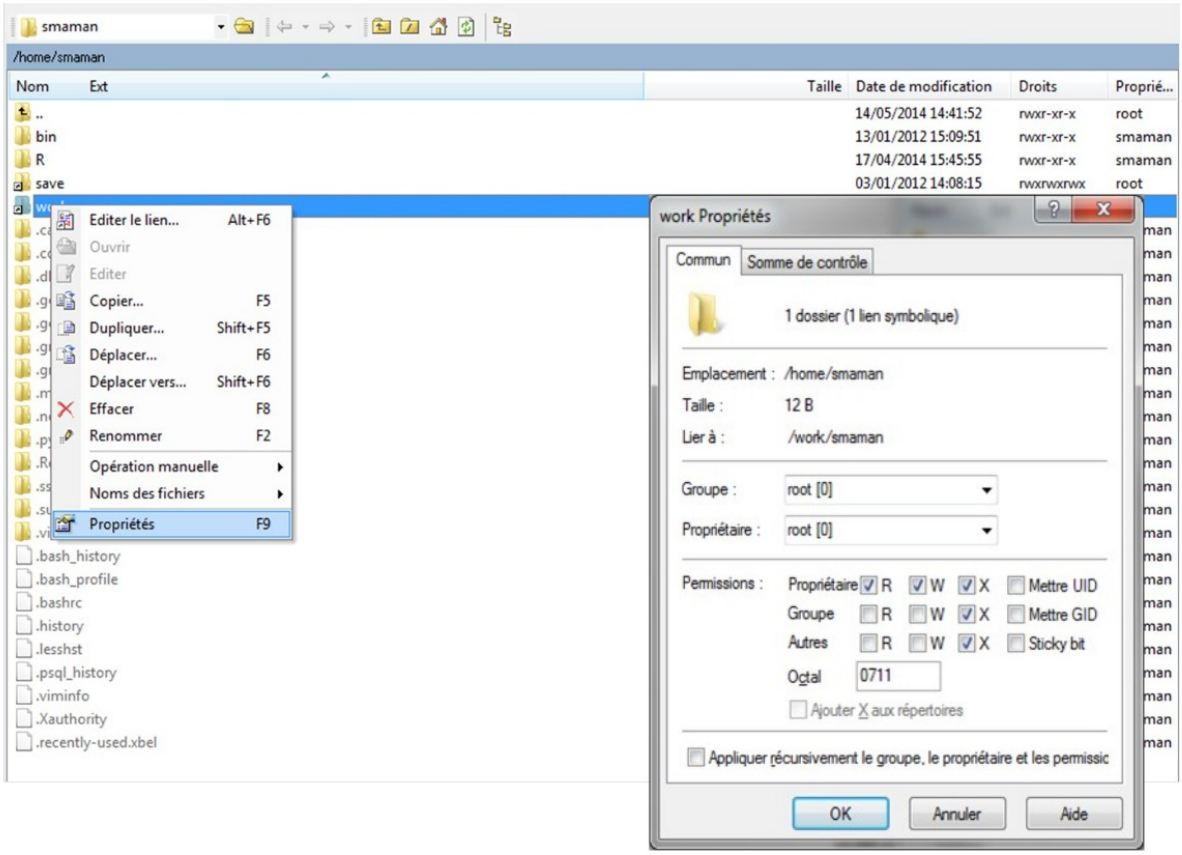


Avec WinSCP :

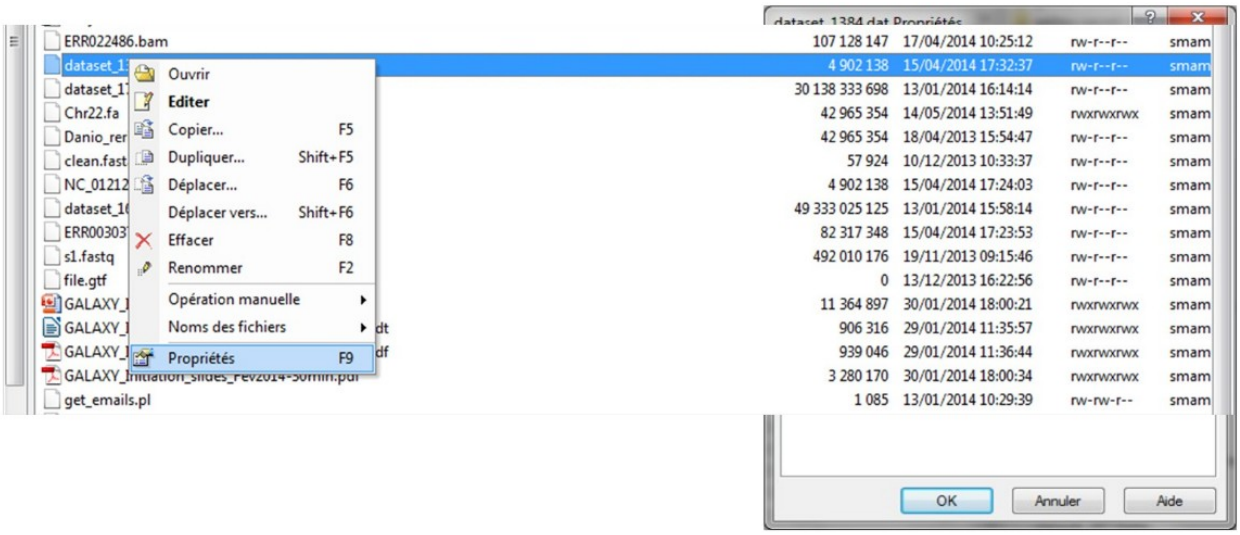


A droite, dans Genotoul, créer un répertoire galaxy/ dans votre /work, puis déplacer l'ensemble des fichiers dans ce répertoire galaxy/ nouvellement créé.

A l'aide d'un clic droit sur le répertoire /galaxy, veillez à ce que le répertoire galaxy/ ai bien les permissions « x » (exécution) 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).
Pour modifier les droits sur votre /work/username/ , clic droit sur votre /work/username/, puis « Droits d'accès au fichier », puis donner les droits d'exécution (X) sur votre /work.



De même pour chacun des fichiers à récupérer, pour modifier leurs droits, clic droit sur le nom du fichier, puis « Droits d'accès au fichier », puis donner les droits de lecture au fichier concerné.



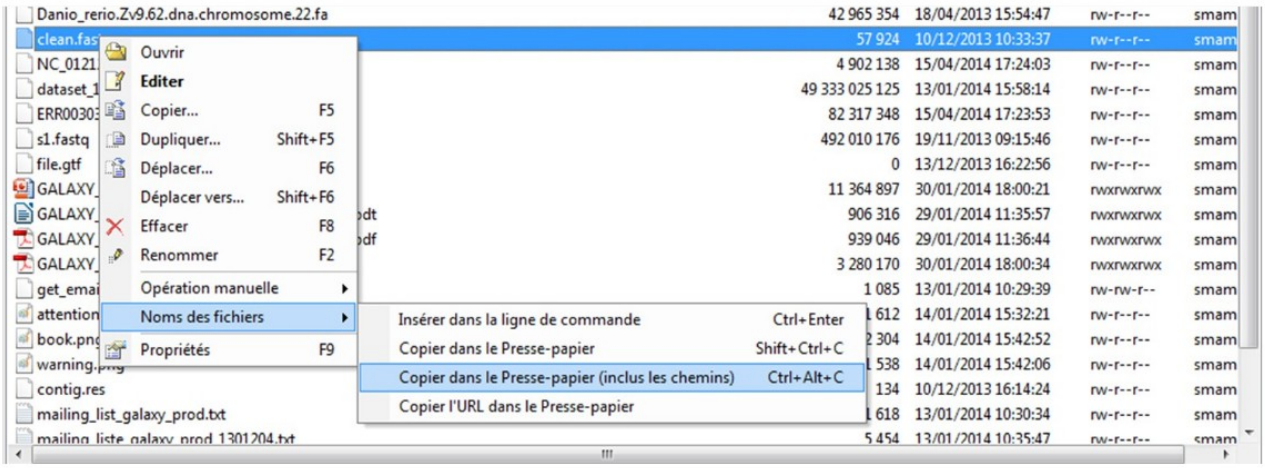
Si cela n'est pas fait, l'outil d'upload de Galaxy ne sera pas en mesure d'accéder ni de lire les fichiers que vous souhaitez télécharger et sur lesquels vous allez travailler. Après l'exécution de l'outil « upload » de Galaxy, la dataset sera rouge (donc en erreur).

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

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.

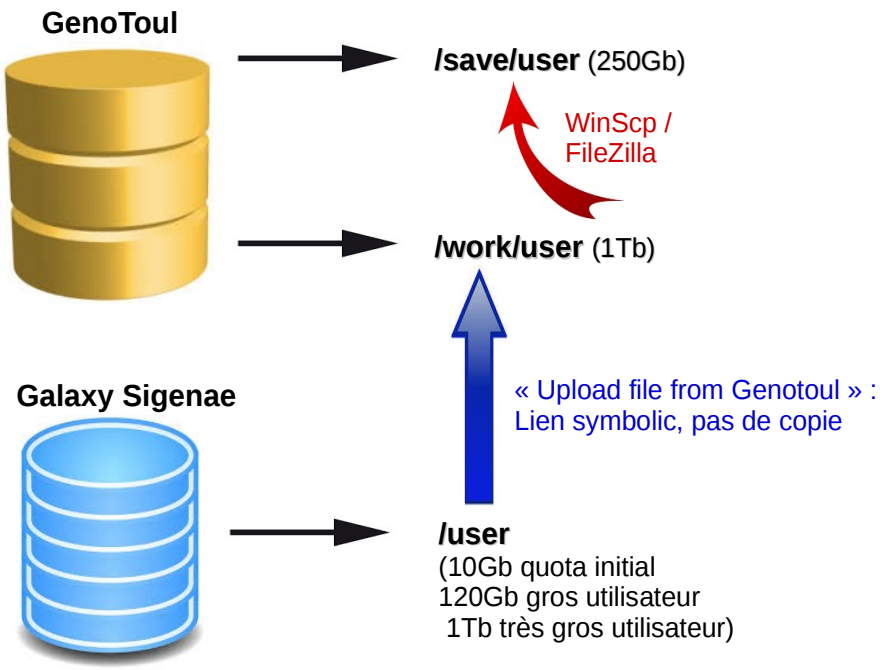
L'historique « historique R1R2 » comprendra les fichiers sampleA_R1.fastq et sampleA_R2.fastq
L'historique « historique multiplex » comprendra les deux fichiers multiplex.fastq et barcode.tabular

Le chemin d'accès au fichier (« Path to file ») 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.
Avec WinSCP, il est possible de récupérer ce chemin complet avec un clic droit sur le nom du fichier, « Noms des fichiers » puis « Copier dans le Presse-papier (inclus les chemins) ».



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 »).

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

Dans l'« **historique contiged** », à l'aide de cet outil « FROGS Upload archive from your computer », veuillez récupérer l'archive « 100spec_90000seq_9samples.tar.gz » disponible depuis cette URL (choisir le mois de la formation) :
http://genoweb.toulouse.inra.fr/~formation/15_FROGS/

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

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

Outils de traitement de fichiers

- Dans l'historique « TP ini Galaxy », 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 »

Analyse de la qualité de vos séquences

“FastQC is a quality control tool for high throughput sequence data.” http://www.bioinformatics.bbsrc.ac.uk/

Lancer FastQC sur chacun des fichiers FASTQ.

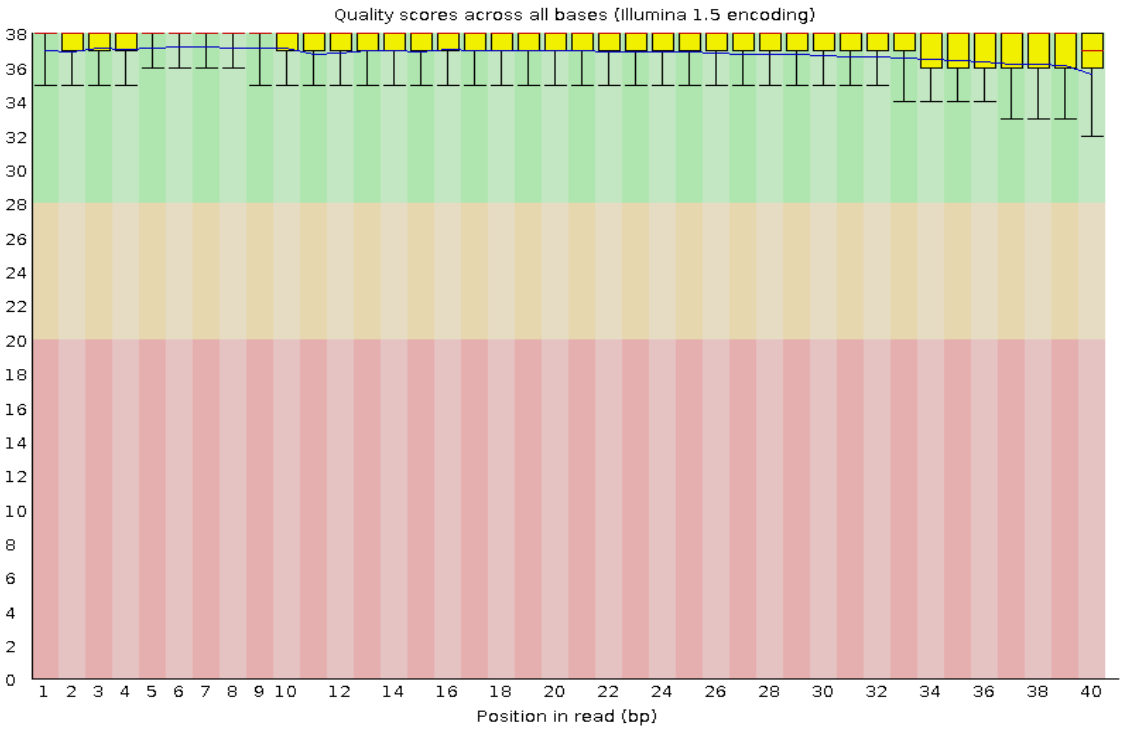
FastQC est un outil permettant d'obtenir un contrôle qualité des données de séquences brutes issues de divers séquenceurs. FastQC fournit une série de modules permettant d'identifier d'éventuels problèmes avec les données, avant de commencer les analyses.

Les fonctions principales de FastQC sont :
- l'import de fichiers BAM, SAM ou FastQ
- fournir un aperçu rapide de la cause de problèmes possibles avec les données brutes
- exporter les résultats sous format HTML

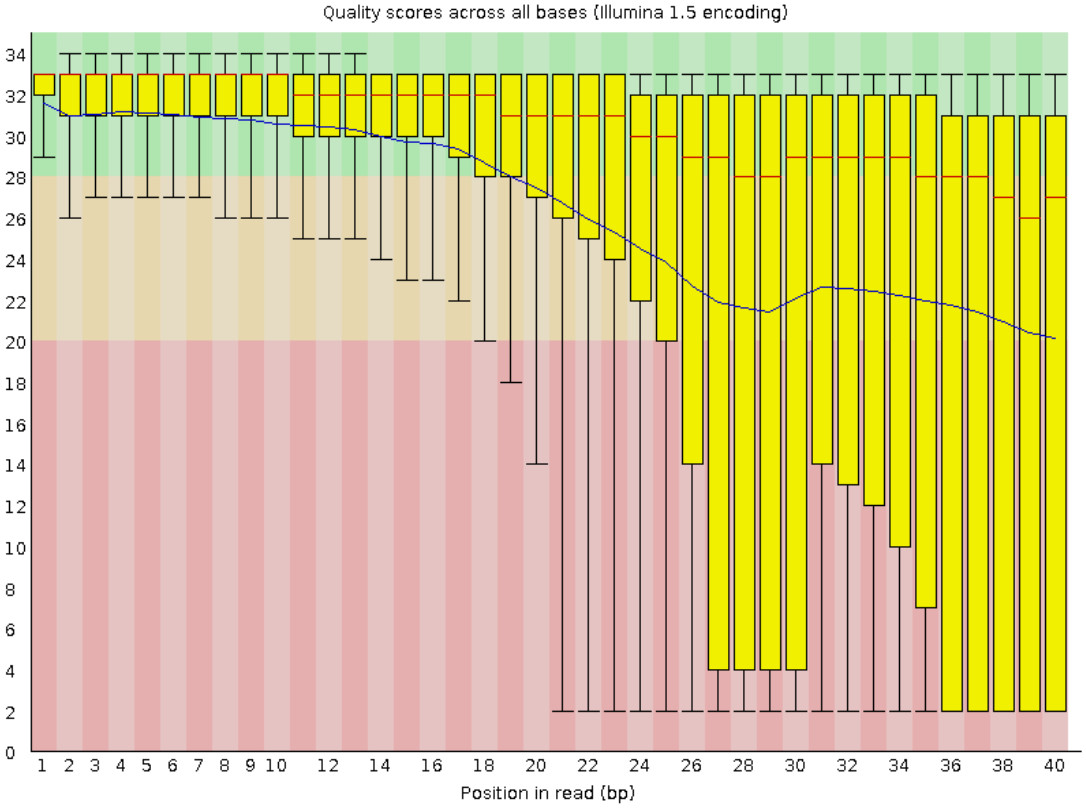
Ci-dessous sont affichés des exemples de résultats obtenus avec FastQC.

Qualité des séquences par base :

Bonne qualité :

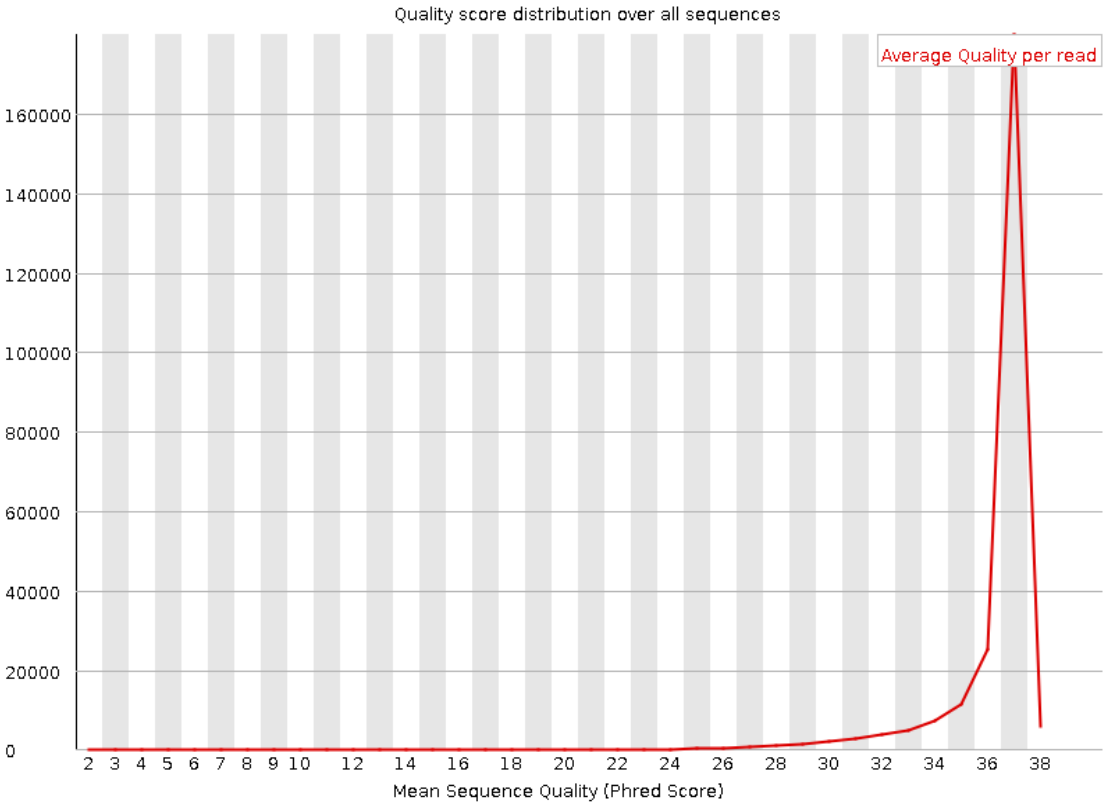


Mauvaise qualité :

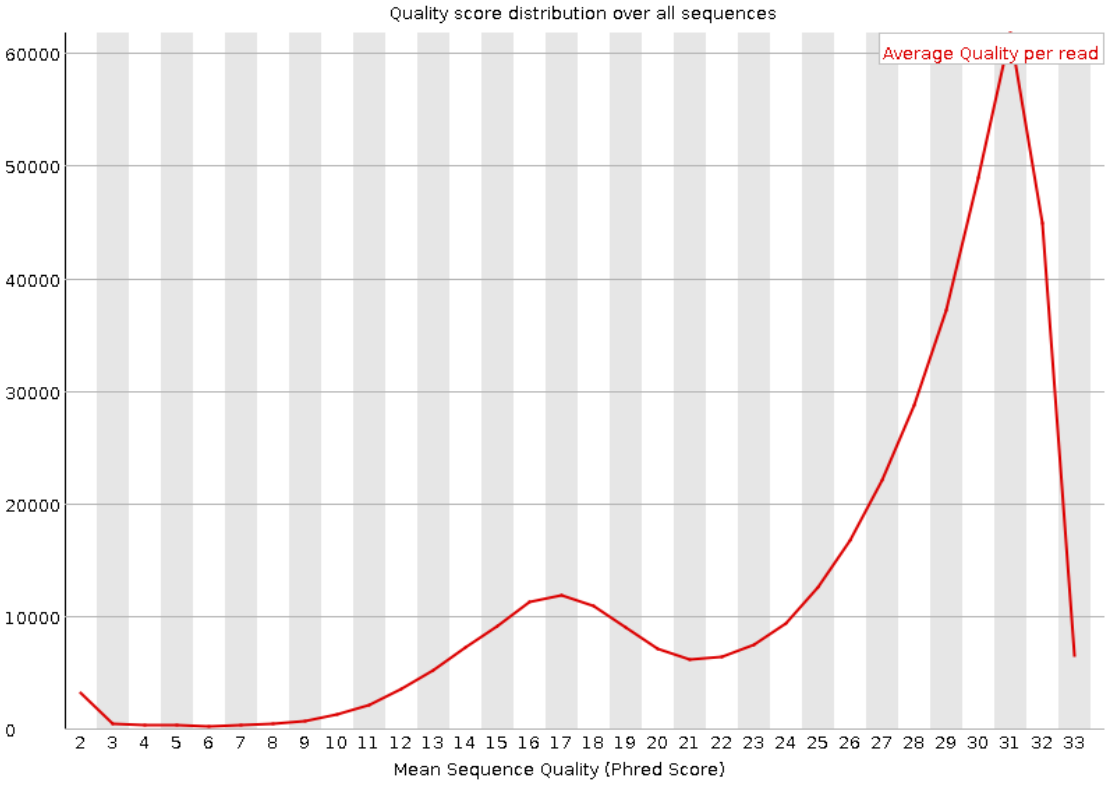


Scores de qualité par séquence

Bonne qualité :

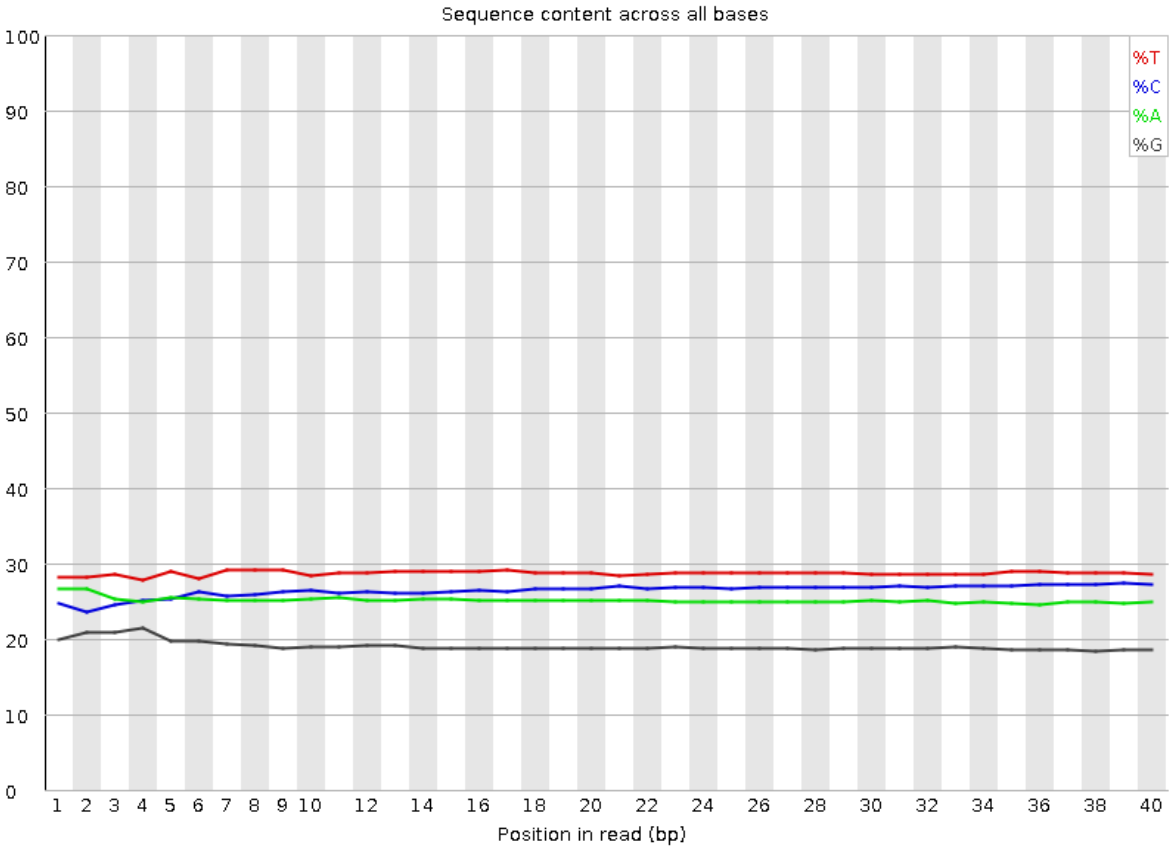


Mauvaise qualité :



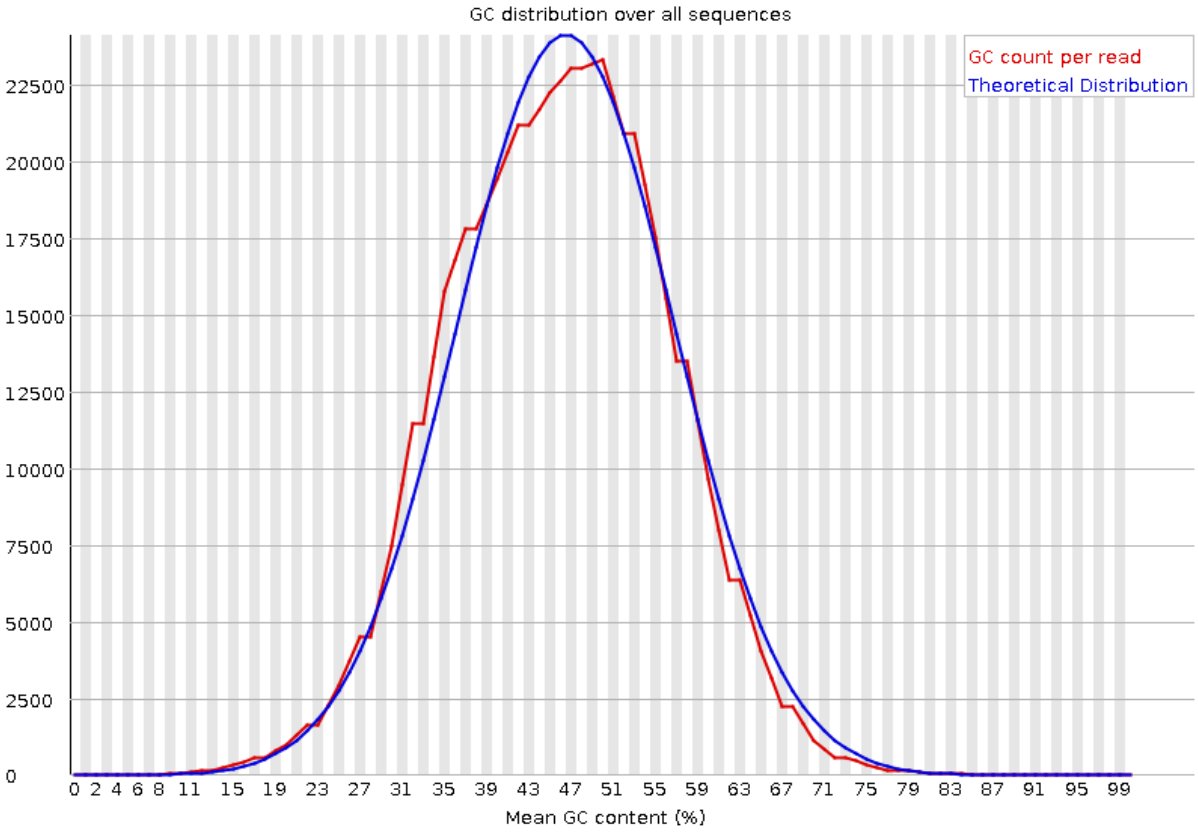
Proportion des bases par séquences

Bonne qualité :



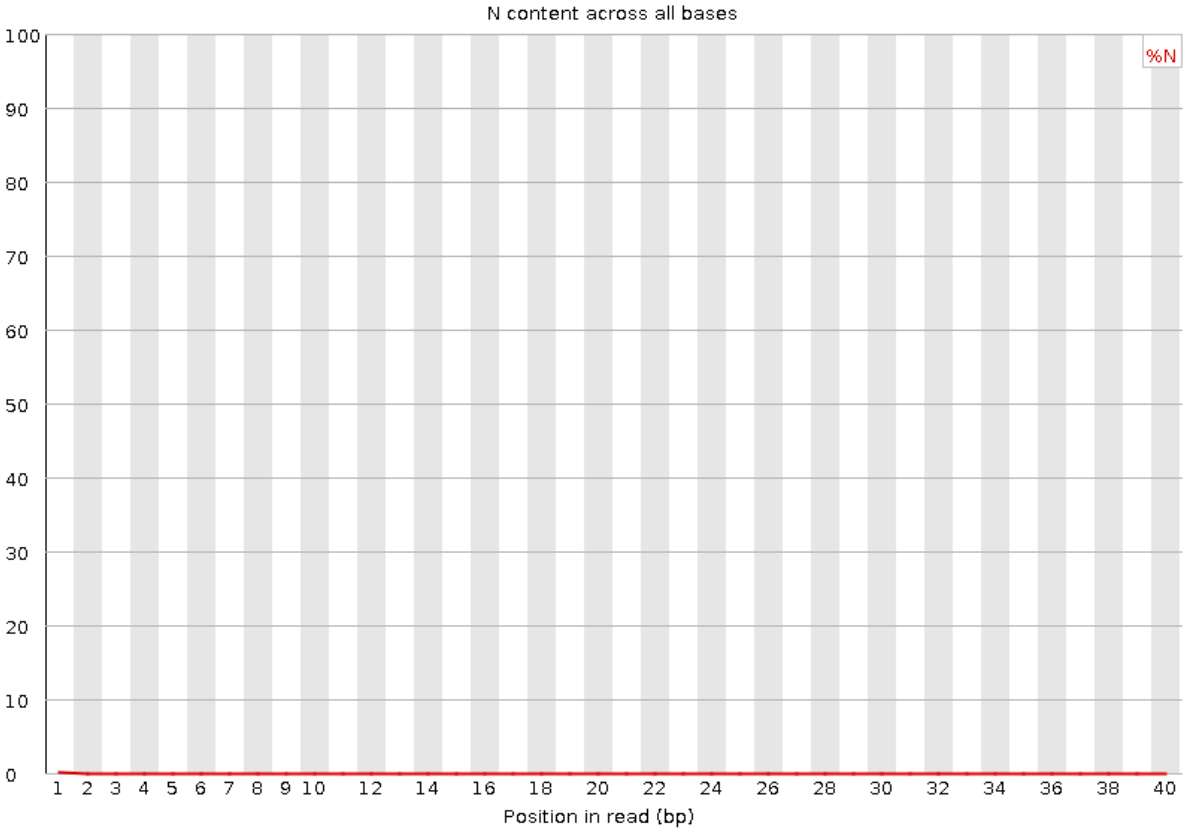
Contenu en GC par séquence

Bonne qualité :



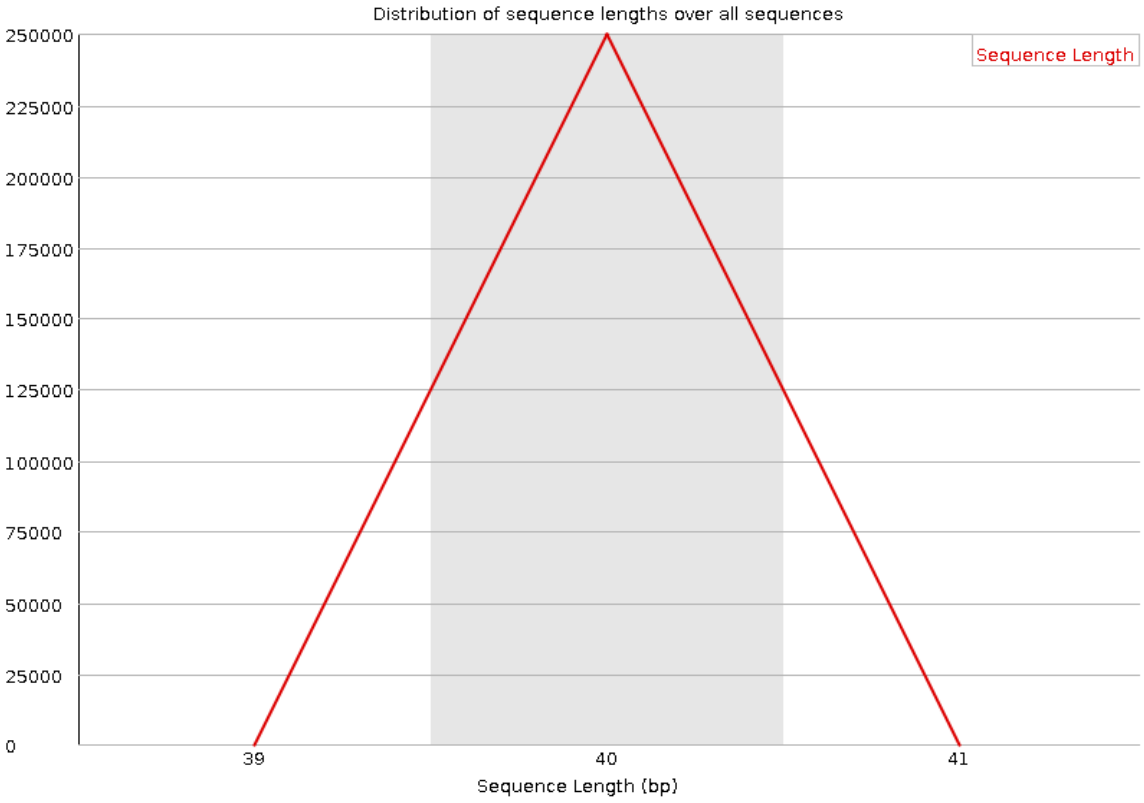
Contenu en « N » des séquences

Bonne qualité :



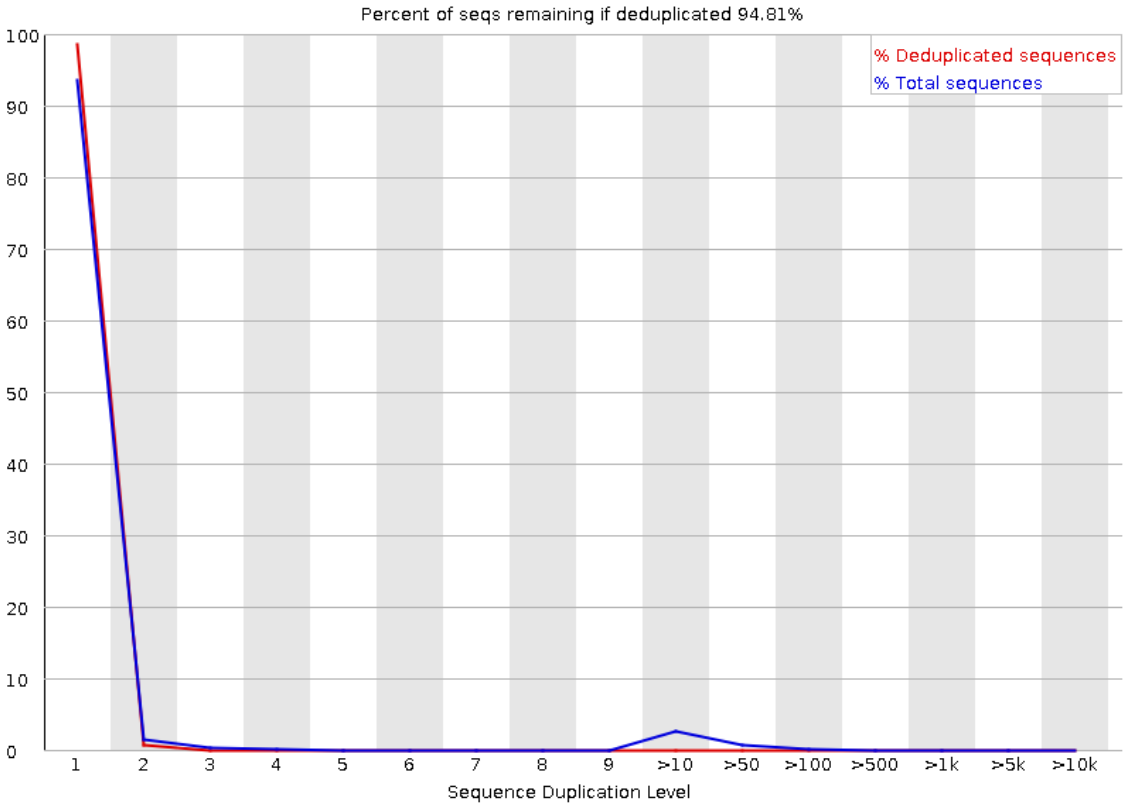
Distribution des tailles de séquences

Bonne qualité (Illumina, séquences de même taille attendues) :

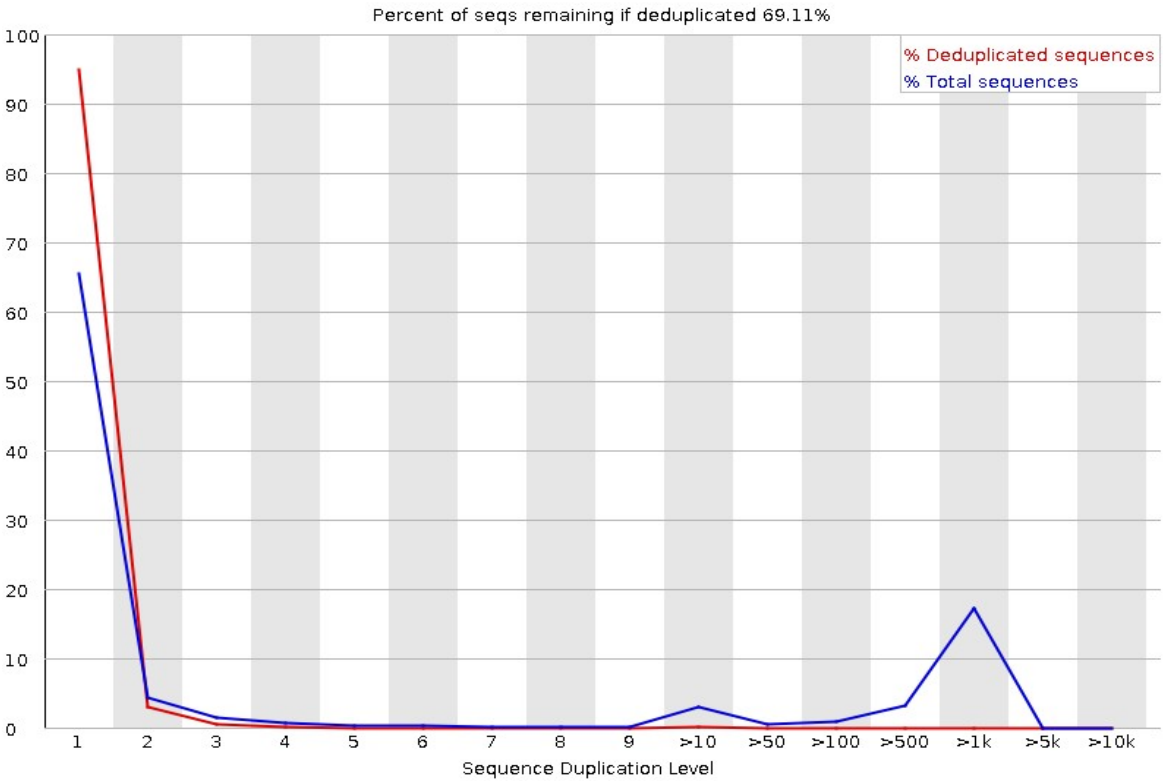


Niveau de duplication des séquences

Bonne qualité :

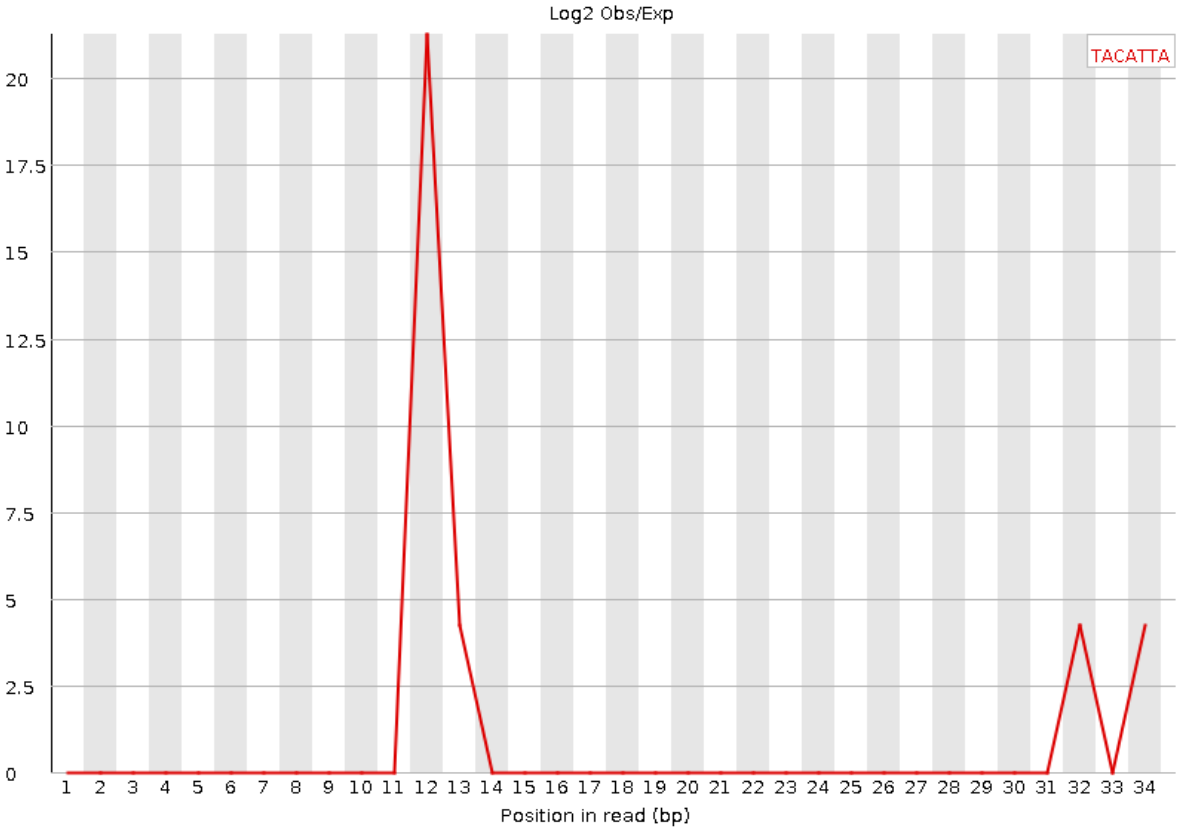


Qualité questionnable (warning) :

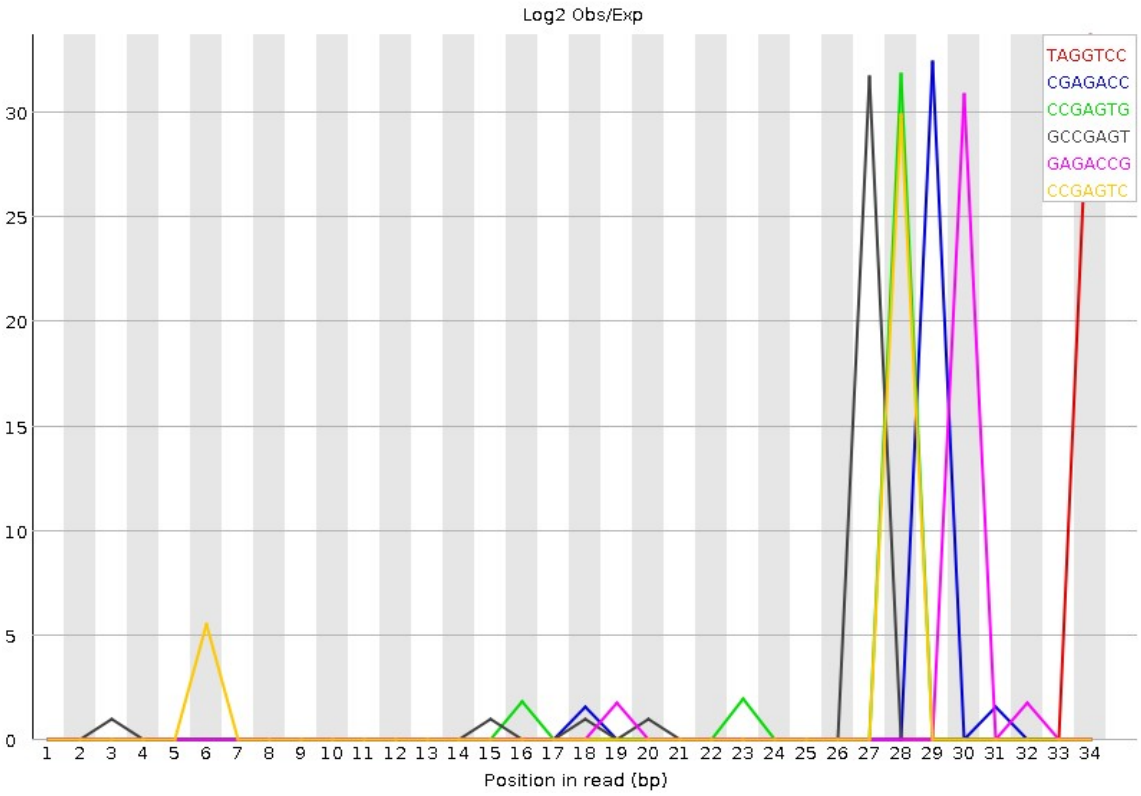


Contenu en Kmer

Bonne qualité :



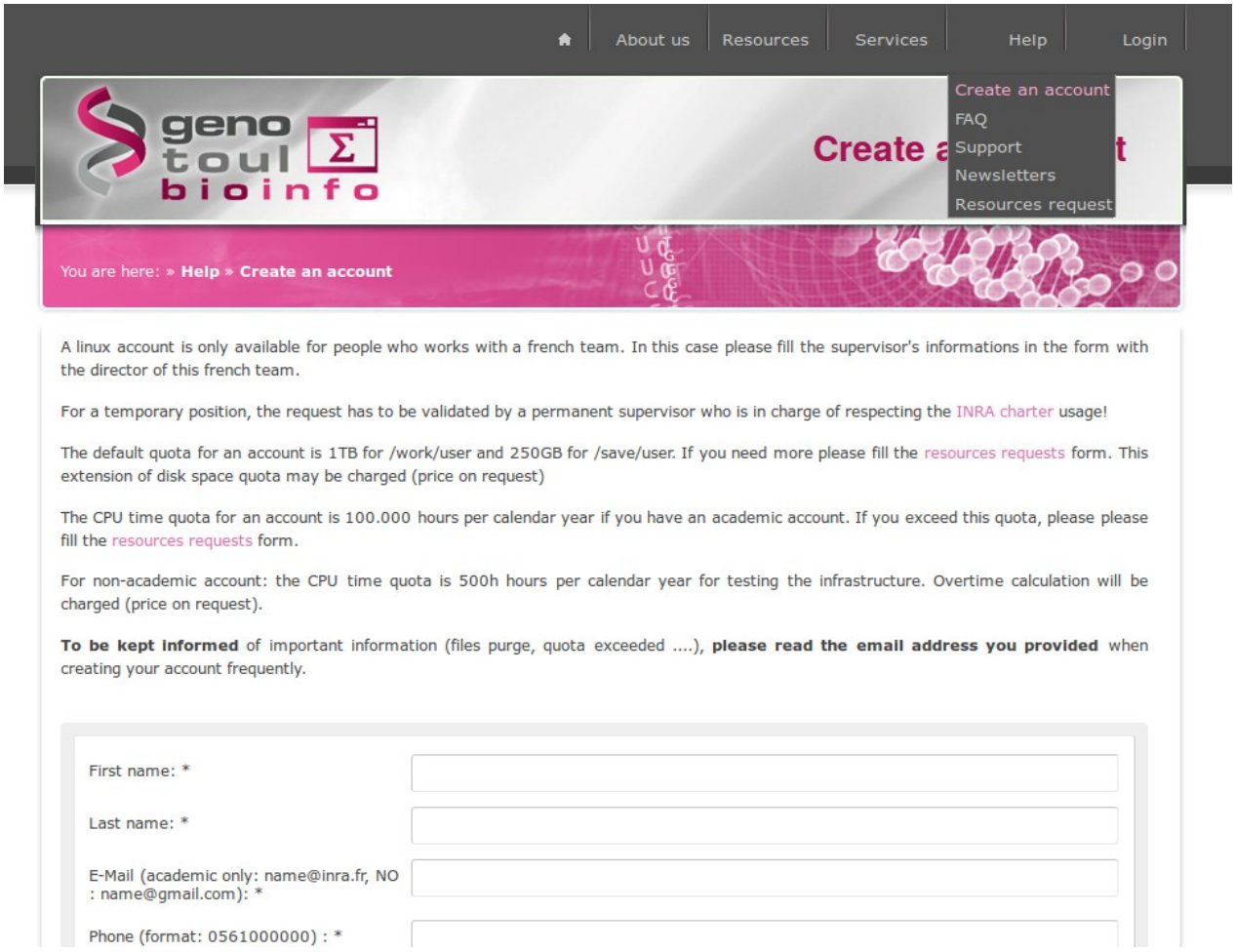
Qualité questionnable (warning) :



Vous trouverez la démarche à suivre en détail dans le site d' « e-learning » ainsi qu'une explication des différents modules de FastQC (5 - FastQC : quality control tool & reports interpertation) comme indiqué ci-dessous.

Pour accéder à l'espace E-LEARNING :

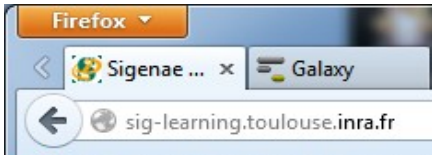
Si vous n'avez pas encore de compte GenoToul, vous pouvez en créer un sur le site de genotoul : http://bioinfo.genotoul.fr/index.php, vous trouverez un formulaire d'ouverture de compte (onglet "help", puis "create an account")



En plus du cluster de calculs et de l'instance galaxy (http://galaxy-workbench.toulouse.inra.fr/), ce compte vous permet d'avoir accès à la plateforme de e-learning (http://sig-learning.toulouse.inra.fr/).

Pour accéder aux formations en e-learning, veuillez suivre la démarche suivante :

1- Taper l'adresse de « sig-learning » : http://sig-learning.toulouse.inra.fr/



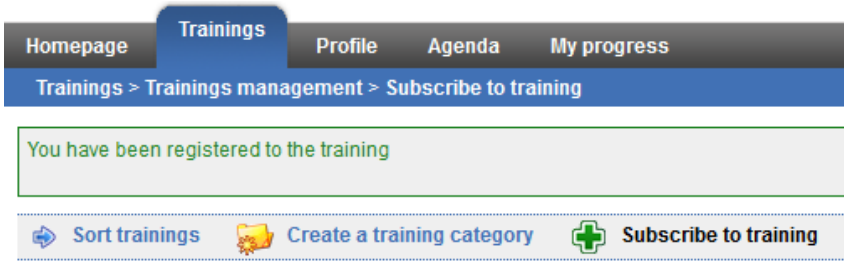
2- Authentification :

3- Ajouter des formations à votre compte :

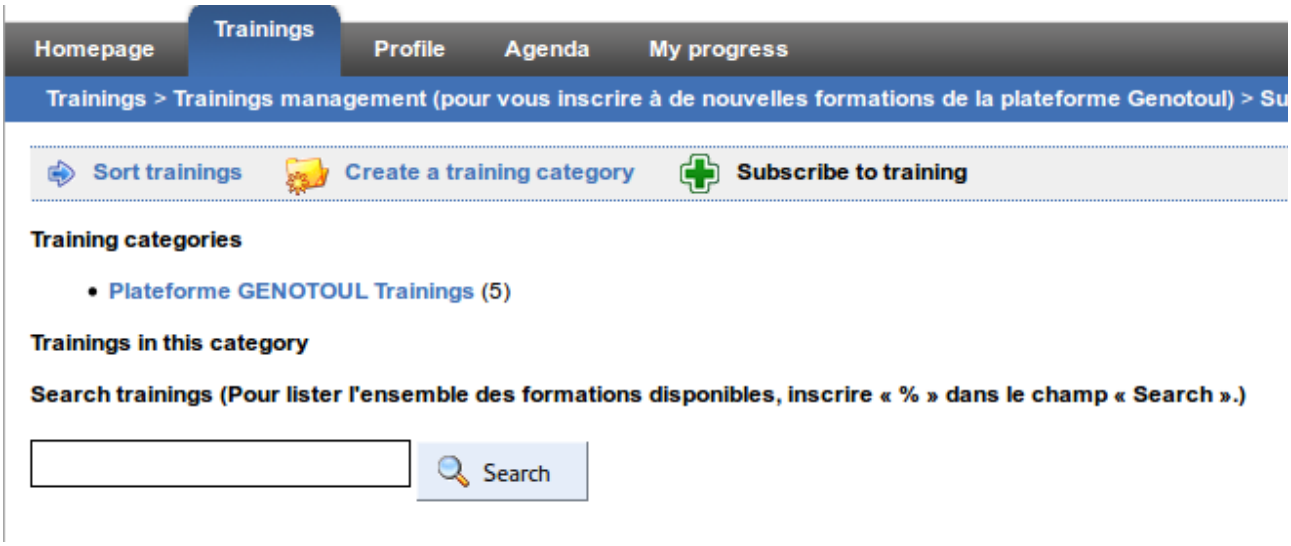
Une fois connecté sur la plateforme d'e-learning, pour accéder aux formations proposées, veuillez suivre cette démarche :
- depuis l'onglet "trainings",
- à droite au niveau de "My account", cliquez sur "Trainings management (pour vous inscrire à de nouvelles formations de la plateforme Genotoul)"
- et enfin + " Subscribe to training "

Onglet training pour accéder à vos formations :

Il vous est possible de vous inscrire directement en ligne à une formation : « Trainings » « Trainings management » puis « Subscribe to training » :



L'inscription s'effectue via une recherche de la formation par mots clés :



Dans "Search" , indiquer "%" pour lister l'ensemble des formations disponibles puis cliquer sur le tableau vert des formations qui vous intéressent pour vous y inscrire.

Voici donc la liste des formations actuellement disponibles :

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

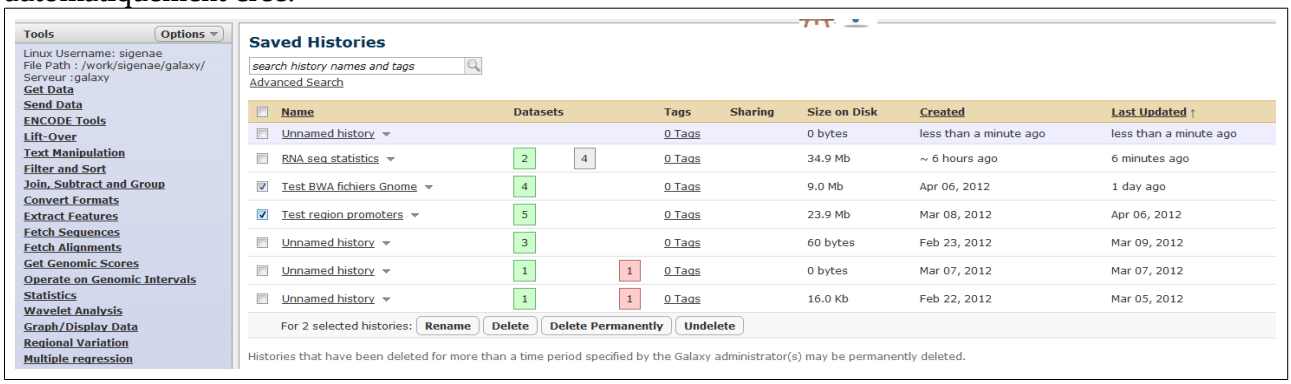
Notions d'historique

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.

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.

Sauvegarder une chaîne de traitements

Etape 1 : Depuis le menu « Options » de votre historique en cours, choisir « Extract Workflow ».
Un workflow est ainsi généré automatiquement et disponible depuis le menu « Workflow ».
Etape 2 : Exporter ensuite ce workflow en cliquant sur le flèche noir à côté de l'intitulé du workflow et choisir « Download or Export ».
Etape 3 : Download to File
Veuillez cliquer sur « Download workflow to file so that it can be saved or imported into another Galaxy server. »
Puis enregistrer ce fichier sur votre PC
Il vous sera ensuite possible de le ré-importer dans votre instance Galaxy.

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 ? »

Astuce :
Le /work est purgé régulièrement des fichiers non utilisés de plus de 120 jours mais les liens symboliques ne sont pas purgés.
Vous pouvez donc créer un lien symbolique de votre /save vers votre /work en cochant la case « Dupliquer avec une copie locale temporaire ».

Quotas Galaxy :
- Première utilisation : 10 Gb
- Gros utilisateur : 120 Gb
- Très gros utilisateur : 1000 Gb

Galaxy gère très mal les caractères spéciaux et les accents.

Remerciements

Fonds Européen
de Développement Régional

RÉGION
MIDI-PYRÉNÉES