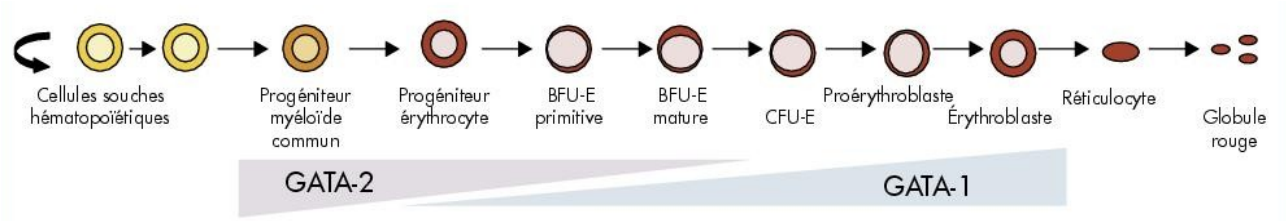

Pour ce TD, nous allons utiliser un jeu de données d'une expérience de ChIP-Seq effectuée sur la protéine CTCF (régulateur transcriptionnel) dans une lignée cellulaire murine G1E-ER4. Le jeu de données est restreint principalement à des reads s'alignant sur le chromosome 19.

Cette lignée cellulaire est GATA1 natif nulle, et exprime stablement une fusion des protéines GATA1 (facteur de transcription important pour les érythroblastes) et ER4 (domaine de liaison au récepteur de l'œstradiol). L'étude de cette lignée permet de mieux comprendre la régulation des gènes impliqués dans le développement des érythroblastes (érythropoïèse). L'ajout d'œstradiol aux cellules permet leur différenciation.

Pour analyser ces jeux de données, nous avons utilisé une plateforme en ligne : Galaxy. La plateforme Galaxy propose une "constellation" d'outils pour analyser, manipuler et visualiser des données génomiques, sans avoir besoin de connaissance en programmation. Elle est développée par The Center for Comparative Genomics and Bioinformatics. Cette plateforme est disponible sur le site de l'UCSC.

Nous avons généré au préalable tous les résultats issus de Galaxy. Tout se trouve dans l'archive TDNGS_ChIP.zip.

En un premier temps, nous avons importé les jeux de données dans Galaxy. Pour cela il suffit de cliquer sur les 2 liens ci-dessous.

Jeux de données condition avec œstradiol au format Sanger
1_G1E_ER4_FASTQ1.fastq (ChIP-Seq)
4_G1E_ER4_input_FASTQ2.fastq (Contrôle)

Étude du jeu de données ChIP-Seq, avec oestradiol

Étape 1:

Nous allons vérifier la qualité en générant un résumé statistique du jeu de données. Pour cela, il faut utiliser "NGS: QC and Manipulation > FASTQ Summary Statistics", et choisir notre jeu de données 1.

Nous avons stocké le résultat dans le fichier 1_G1E_ER4_CTCF_summary_statistique.txt. Ouvrez le fichier. (Vous pouvez aussi utiliser FastQC sur Galaxy, allez à la rubrique "NGS: QC and Manipulation > FastQC:Read QC". Cela génère des résultats en html visibles dans Galaxy en cliquant sur le petit œil (display data on browser)).

Les reads ont quelle longueur ? Quelle est la qualité médiane et moyenne à la dernière position ? Combien de reads sont "processés" ?

Étape 2:

Une fois la qualité vérifiée, nous pouvons aligner nos reads sur un génome de référence (ici la lignée dont sont issus les reads est murine). Nous avons utilisé "NGS: Mapping > Map with Bowtie for Illumina", en sélectionnant notre jeu de données à aligner (**1_G1E_ER4_FASTQ1.fastq)** et le génome de référence que l'on souhaite.

Cette étape a été effectuée pour éviter de long temps de calcul. Ouvrez le fichier 2_sample_G1E_ER4_CTCF_FASTQ1_aligned_by_bowtie.txt qui représente le header + les 10 premiers reads alignés par bowtie (échantillon du fichier SAM).

Sur quel génome et quelle version du génome sont alignés les reads ? Combien de reads sont alignés (sur le brin sens et antisens)? Sur quel chromosome ?

Étape 3:

Les reads sont maintenant alignés (3_G1E_ER4_CTCF_FASTQ1_aligned_by_bowtie.sam). Grâce aux informations stockées dans le fichier SAM, nous avons extrait les pics de reads avec l'outil MACS, "NGS: Peak Calling > MACS". Les paramètres sont ceux de base et sans contrôle, excepté la taille des reads qu'il faut changer avec la valeur trouvée lors de l'Étape 1. Deux résultats sont générés, un rapport html sur le "peak calling" et un fichier de coordonnées génomiques « .bed ».

Ouvrez le dossier 3_MACS_on_G1E_ER4_CTCF_FASTQ1(html_report). Ouvrez le fichier html et observez les différents fichiers accessibles.

Combien de pics sont trouvés ? Sont ils tous sur le chromosome 19 ? Pourquoi ?

Étude du jeu de données ChIP-Seq + contrôle, avec oestradiol

Étape 4:

Lors de cette étape, nous avons répété les opérations effectuées préalablement jusqu'à l'étape 2, mais sur un fichier différent : 4_G1E_ER4_input_FASTQ2.fastq. Pour le "peak calling", on réitère l'opération avec les paramètres de base (reads size à changer). Mais ici nous avons mis un ChIP-Seq Controle File : 4_G1E_ER4_input_FASTQ2_aligned_by_bowtie.sam.

On récupère de nouveau deux résultats. Quel effet a eu l'ajout d'un contrôle dans le résultat du « calling » des pics ? Pourquoi à votre avis ?

Lors de ces 4 étapes, nous avons analysé les jeux de données issus d'expériences induites par l'œstradiol. Nous allons maintenant effectuer ces mêmes étapes avec des jeux de données d'expériences non induites par l'œstradiol, puis utiliser Galaxy pour identifier les sites de liaison présents dans les deux états de développement cellulaire (différencié ou non).

Étude du jeu de données ChIP-Seq + contrôle, sans oestradiol

Pour importer les jeux de données des conditions sans œstradiol, au format Sanger

5_G1ECTCF_FASTQ3.fastq (ChIP-Seq)
5_G1Einput_FASTQ4.fastq (Contrôle)

Étape 5:

Comme précédemment, nous avons aligné les reads de l'expérience et du contrôle ("NGS: Mapping > Map with Bowtie for Illumina" sur la référence mm9), puis nous avons utilisé MACS ("NGS: Peak Calling > MACS") comme lors de l'étape 4, avec un contrôle sans oublier de changer la taille des reads. On récupère le rapport html et le fichier de coordonnées génomiques « .bed ».

Identification des pics spécifiques selon la condition avec ou sans œstradiol (cellules différenciées ou non)

Étape 6:

Nous avons ici utilisé la fonction "Operate on Genomic Intervals > Subtract".

En première entrée (case "Subtract"), on sélectionne le dernier set de pics créé (Expérience sans œstradiol, 5_MACS_on_G1ECTCF_FASTQ3_with_controleFASTQ4.bed).

En second (case "From"), on sélectionne le premier set de pics. (Expérience avec œstradiol, 4_MACS_G1E_ER4_CTCF_FASTQ1_with_controleFASTQ2.bed). Le résultat ne contient que les pics spécifiques de la lignée différenciée. Combien y a-t-il de pics ?

Étape 7:

Nous avons refait l'étape 6 mais en inversant les listes de pics. Combien de pics sont spécifiques de la lignée indifférenciée ?

Étape 8:

Nous avons utilisé l'outil "Operate on Genomic Intervals > Intersect" pour trouver les pics communs aux deux états cellulaires. Combien de pics sont communs ?

Les listes de pics vont nous permettre d'interpréter les résultats biologiques. Les étapes suivantes à ce TD sont :
- L'annotation des pics (quel gène?)
- La recherche de références bibliographiques et fouille des bases de données
- Est-ce une cible d'autres facteurs de transcription ?
- ...
Mais ceci est une autre histoire...