TP Mapping

1 Contexte Biologique

Vous disposez de données issues d'une expérience de séquençage de transcriptome humain. Dans cette expérience, on cherche à comprendre quels gènes sont exprimés dans la lignée celullaire SKNSH en présence d'un traitement par acide rétinoïque. Cette lignée cellulaire a été établie en 1970 à partir de cellules métastatiques de moelle osseuse, d'une enfant de 4 ans. Ce sont des cellules de neuroblastome qui ont une morphologie de type épitheliale. Le traitement par acide rétinoïque induit une differentiation, et les cellules adoptent un phénotype neuronal, caractérisé par une excroissance notable des neurites (axones, dendrites). Cette lignée cellulaire est donc particulièrement utilisée dans les études visant à caractériser les voies de signalisation impliquées dans la différentiation neuronale.

2 Connexion

Pour effectuer vos analyses, vous disposez d'un accès à la machine ngs-provisoire Vous pouvez vous y connecter par ssh avec la commande suivante :
ssh -X etudiant1@ngs-provisoire Pour copier des données depuis votre machine vers la machine ngs4bim :
scp mydata.txt etudiant1@ngs-provisoire :/home/etudiant1/monrepertoire

Les données à analyser se trouvent ici :
/data/data/ENCODE/SKNSH

Pour tester votre pipeline, vous travaillerez sur les fichiers 1M. Votre rapport devra etre fait sur les fichiers 10M (10 millions de lignes, soit 2.5 millions de lectures)

Vous rendrez par binomes un rapport de 6 pages qui présentera votre analyse de données, et les choix que vous avez faits. L'objectif est d'établir une liste de gènes dont l'expression a changé après traitement par acide rétinoique, et de clarifier le niveau de confiance que vous pouvez accorder aux gènes de cette liste.

3 Mapping

À l'aide d'un mapper de votre choix (gem, bwa, bowtie ou Star), vous alignerez vos données contre le génome humain. Des index précalculés du génome humain se trouvent ici :
`/data/Public/GEMIndex`, `/data/Public/bwaIndex`

Dans vos analyses, vous répondrez aux questions suivantes :
Question 1 : Expliquez le format de sortie de votre mapper. Faites varier en par-

ticulier le nombre de mismatches autorisés. À quoi correspond un mismatch d'un point de vue biologique ? Quel est le bon nombre de mismatches à autoriser ?

Question 2 : Existe-t-il des lectures qui ne s'alignent pas sur le génome de référence ? Des lectures qui s'alignent à une seule position ? À plusieurs positions ? Quantifiez chaque catégorie et synthétisez vos résultats dans un tableau facile à lire.

Question 3 : Quelle stratégie pouvez-vous mettre en place pour le cas des lectures qui s'alignent à plusieurs endroits ? Discutez de l'impact attendu de cette stratégie sur la détection de gènes différentiellement exprimés

Question 4 : Implémentez au moins deux stratégies de résolution des alignements multiples. Pour quantifier les gènes, vous pouvez utiliser soit HTSeq soit BedTools. Dans les deux cas, il vous faut fournir une annotation du génome de référence. Prenez par exemple Gencode (`/data/data/Ensembl/gencode.v24.basic.annotation_genes.gtf`).

La commande BEDTools est intersectbed -a reads.bed -b gencode.gtf -c

Question 5 : Trouvez trois gènes qui n'ont pas le meme niveau d'expression suivant la stratégie implémentée. Discutez les en détail. Combien de gènes sont dans ce cas ? Qu'ont-ils en commun ?

Question 6 : Établir une liste de gènes différentiellement exprimés en présence d'acide rétinoïque. Discutez de l'exhaustivité de cette liste.