Analyse de données multi-omiques (Séance 5/7)

Gaëlle LELANDAIS et Fabrice CONFALONIERI

(gaelle.lelandais@universite-paris-saclay.fr – fabrice.confalonieri@universite-paris-saclay.fr)

Ce contenu est mis à disposition selon les termes de la licence [Creative Commons BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)

Objectif(s) de la séance

Pour rappel, votre objectif est de caractériser le régulon DdrO, chez la bactérie radio-résistante *Deinococcus radiodurans*. Pour cela, nous répétons les analyses de données multi-omiques présentées dans un article scientifique (Eugénie *et al.*, 2021). A l'issue de la séance précédente, vous avez obtenu des listes de gènes candidats pour être différentiellement exprimés (contexte de présence / absence de la protéine DdrO). Ces listes sont fondées sur vos choix, personnels et individuels. Au cours de cette séance, nous analysons les données ChIPseq permettant de localiser à l'échelle du génome, les sites de fixation à l'ADN de la protéine DdrO. Ainsi, les étapes d'une analyse de données ChIPseq seront présentées, avec une attention particulière portée à l'étape de « peak calling » (outils MACS et bPeaks).

Organisation de la séance

Cette séance se décompose en plusieurs activités. Elles sont détaillées ci-dessous. Les temps de réalisation des activités sont indicatifs et peuvent être ajustés en fonction de vos besoins. Pour une bonne cohérence générale des apprentissages, chacune des tâches est à réaliser, pendant la séance, ou à l'issue de la séance. **À tout moment ne pas hésiter à solliciter l'équipe pédagogique pour d'éventuelles clarifications.**

❖ **Activité 1 : Pipeline d'analyse de données ChIPseq**

Résumé

Au cours de cette activité, les grandes étapes d'une analyse de données ChIPseq sont discutées. Une attention particulière est donnée à l'étape de « peak calling » dont l'objectif est de détecter les régions génomiques fixées par la protéine d'intérêt.

Travail à réaliser

☐	Relire les diapositives du cours. Les notions abordées sont-elles claires ? Solliciter l'équipe pédagogique si des clarifications sont nécessaires.
☐	Réaliser l'alignement des séquences (fichiers FASTQ) sur le génome de référence, en utilisant l'outil Bowtie2. Renommer les fichiers BAM afin de bien identifier les échantillons (IP1, IP2, IP3, MOCK et INPUT). Réaliser une visualisation avec IGV et vérifier la présence d'un « pic » dans une région génomique exposée dans l'article de référence.
☐	Un des réplicats IP est moins utile pour les analyses. Lequel et pourquoi ? Solliciter l'équipe pédagogique si des clarifications sont nécessaires.

Ressource(s) complémentaire(s)

- Historique partagé avec les fichiers FASTQ : <https://usegalaxy.fr/u/gaellelelandais/h/chipseq-ddro---fichiers-fastq>

❖ **Activité 2 : Détection des régions génomiques fixées par DdrO**

Résumé

La détection des « pics » est de loin l'étape la plus difficile d'une analyse de données ChIPseq. En fonction de l'outil bioinformatique choisi, les résultats obtenus peuvent être très différents. Au cours de cet atelier, nous utiliserons le logiciel MACS2, disponible sur Galaxy. Les résultats obtenus avec le logiciel bPeaks (package R) vous seront partagés. Ils seront utilisés dans la suite des analyses.

Travail à réaliser

- Comparer les échantillons IP1, IP2 aux échantillons contrôles MOCK et INPUT (4 comparaisons à réaliser). Combien de pics sont détectés ? Est-ce cohérent avec les résultats exposés dans l'article de référence ?
- Exporter les fichiers BED obtenus et les ajouter sur la visualisation IGV réalisée précédemment (Activité 1). Les pics observés sont-ils détectés ? Soumettre votre image sur eCampus (note incluse dans le CC).

❖ **Activité 3 : Comparaison des régions détectées avec celles conservées dans la publication**

Résumé

Dans la publication originale, la détection des pics a été réalisée avec le logiciel bPeaks. C'est un programme qui s'utilise avec les logiciels R et RStudio et qui a été développé pour la détection des pics dans les génomes de petite taille.

Travail à réaliser

- Les régions retenues dans la publication originale sont données dans le fichier « FinalPeaks_DdrO_Paper.bed ». Comparer ces résultats avec ceux que vous avez obtenus avec le logiciel MACS. Pour cela, vous pouvez réaliser une visualisation avec IGV et également utiliser l'outil « Intersect », disponible sur Galaxy.

Ressource(s) complémentaire(s)

- Vidéo de présentation de la mise en application de la méthode bPeaks, à partir des fichiers d'analyse de données ChIPseq obtenus sur Galaxy : <https://youtu.be/FYyXKE0bECY>

Pour la suite des analyses (recherche des motifs, séance 6/7), vous pourrez travailler avec vos propres fichiers de résultats ou bien avec ceux qui ont été utilisés pour la réalisation de la publication.