

Développement d'un pipeline d'analyse de données de RNA-seq

Contexte et enjeux

Les maladies rares sont définies par une prévalence inférieure à 1 personne sur 2000 en population générale. 80% de ces maladies sont d'origine génétique ce qui correspond à environ 7000 maladies rares touchant 3% des nouveau-nés. Il s'agit donc de maladies très hétérogènes nécessitant des moyens importants pour le diagnostic. En effet, la prise en charge des patients atteints de maladies rares implique de multiples analyses pour aboutir au diagnostic de leur pathologie. Malgré l'avènement récent du séquençage d'ADN haut-débit (exome, génome) un diagnostic moléculaire n'est confirmé que pour 50% des patients. Les patients sont donc considérés en errance diagnostique dans l'attente d'une confirmation moléculaire. La mise en place de technologies de séquençage haut débit (NGS) a permis de résoudre l'étape d'identification des variations de l'ADN. Mais afin de réduire l'errance diagnostique il faut augmenter le rendement de ces analyses en permettant de prouver l'effet causal d'une variation dans la pathologie considérée. Notre laboratoire est référent pour l'étude de plusieurs de ces pathologies rares (déficience intellectuelle, neurodégénérescences, mucoviscidose, albinisme, syndrome de Rubinstein-Taybi, filaminopathie A, cardiopathies congénitales).

L'objectif global du laboratoire de génétique moléculaire du CHU est l'amélioration constante des stratégies mises en œuvre pour le diagnostic moléculaire des patients atteints de maladies rares, grâce à l'adaptation de technologies issues de la recherche en génomique fonctionnelle.

Nous développons des outils d'analyse d'expression des gènes impliqués dans les pathologies pour lesquelles nous sommes référents. L'analyse par RNA-Seq a déjà été mise en place dans le laboratoire avec une capture d'ARN polyA (Kit Illumina Stranded) puis séquençage sur l'automate Illumina Next-Seq 550.

Ce stage s'intégrera dans le projet de recherche clinique « RID (Résoudre les impasses diagnostiques dans les maladies rares : tests fonctionnels et nouvelles stratégies d'études moléculaires des variants géniques issus du séquençage nouvelle génération.) » financé par le CHU de Bordeaux. Ce financement permettra une analyse en RNA-Seq pour 25 patients afin de valider le pipeline développé.

A ce jour le laboratoire dispose d'un pipeline d'analyse d'exome sous docker et d'une première version d'un pipeline d'analyse RNAseq, également sous docker. L'ensemble des données générées est sauvegardé dans une base de données SQL et exploité via une interface web (en cours de développement).

Objectifs

L'objectif de ce stage est dans un premier temps de prendre en main le pipeline d'analyse de données de RNA-seq du projet. Il s'agira ensuite d'exploiter ces données afin d'identifier de potentiels transcrits aberrants ou encore de mettre en évidence une expression mono-allélique. Pour cela un workflow d'analyse d'expression différentielle sera élaboré et implémenté pour l'approche quantitative et la définition de profils d'expression dans le sang, tissu facilement accessible pour le diagnostic.