Amélioration de l'accès aux résultats biologiques :

Séquençage ARN et Application Shiny

Master 1 Bioinformatique

Université de Rennes 1

2019 - 2020

David GALLIEN & Gabin COUDRAY

UMR CNRS 6290
Institut de Génétique et Développement de Rennes (IGDR)

Annabelle MONNIER
Equipe Oncogénomique translationnelle

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) ……
Etudiant (e) en ……

Déclare être pleinement informé (e) que le plagiat de documents ou d'une partie de documents publiés sous toute forme de support (y compris l'internet), constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Signature

Laboratoire de Génomique Médicale
BMT-HC - CHU Pontchaillou

2 rue Henri le Guilloux
35033 Rennes Cedex
FRANCE

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr
TÉL. 33 (0)2 99 28 92 54

UMR CNRS 6290
Institut de Génétique et Développement de Rennes (IGDR)

Annabelle MONNIER
Equipe Oncogénomique translationnelle

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) ……
Etudiant (e) en ……

Déclare être pleinement informé (e) que le plagiat de documents ou d'une partie de documents publiés sous toute forme de support (y compris l'internet), constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Signature

Laboratoire de Génomique Médicale
BMT-HC - CHU Pontchaillou

2 rue Henri le Guilloux
35033 Rennes Cedex
FRANCE

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr
TÉL. 33 (0)2 99 28 92 54

Table des matières

Introduction

Contexte

Aujourd'hui, la plupart des chercheurs n'ont en général pas le temps d'utiliser les outils permettant une analyse des données générées par les nouvelles technologies. C'est pour cela qu'il est important de leur offrir la possibilité d'avoir accès à des outils facilitant l'analyse et leur permettant d'être plus efficaces. Le but principal de notre projet est de mettre en place une application qui pourra aider ces scientifiques à explorer leurs résultats sous l'environnement R. Pour cela le package Shiny nous permet de créer une application web interactive.

Néanmoins, afin de créer une application interactive, nous avons besoin de quelques choses à montrer. Dans le domaine de la recherche médicale qui est le principal axe de recherche, les techniques de séquençage de seconde génération sont souvent utilisées. Ces techniques génèrent de nombreuses données qui ont besoin d'être explorées et analysées. Nous avons donc décidé de concentrer notre travail sur le séquençange de l'ARN (RNA-seq). Cette technique de séquençage de seconde génération a pour principal but de détecter des expressions différentielles entre des types cellulaires de différentes conditions.

Le RNA-seq est un nouveau moyen permettant un séquençage de l'ARN plus rapide que les techniques qui existaienet précédemment comme la méthode de Sanger. Le but principal du RNA-seq est d'étudier l'expression différentielle de gènes entre différentes conditions. Le séquançage de l'ARN a été cité pour la première en 2008, et depuis, le nombre de publication contenant des données de RNA-seq augmentent d'années en années. Ce genre d'analyses utilises les technologies de séquançage de nouvelles génération (NGS) comme Illumina, Roche 454 ou encore Ion Torrent.

Une analyse RNA-seq présente 3 grandes étapes :

- Fragmentation aléatoire des ARN matures
- Amplification de ces fragments par PCR
- Séquençage de ces fragments donnant des millions de reads

Le nombre de reads obtenu est proportionnel à l'abondance des ARN dans la cellules. Ces reads sont stockés dans des fichiers au format fastQ et leur qualité est estimé grâce à des outils spécifiques. Ensuite, chaque read est mappé sur le génome de référence de l'organisme étudié. Après ce mapping, des fichiers BAM sont obtenus. Dans ces fichiers, chaques ligne représante un alignement d'un read. Pour finir, un comptage des reads pour chaque position est realisé afin de remplir une table de comptage permettant l'analyse des données RNA-seq.

Analyse de données RNA-seq

Les étapes de l'analyse RNA-seq présentées ci-dessous sont inspirées du mode d'emploi d'analyse de données RNA-seq sur le site bioinfo-fr.net.

Etape 1 : Les données

Premièrement, pour obtenir le jeu de données, l'ARN est extrait des cellules et l'ARNm est isolé grâce à sa queue poly-adénylée. Une fois extrait, l'ARNm est fragmenté et subit une reverse transcription en ADNc. Ensuite, l'ADNc est séquencé grâce aux NGS. Aujourd'hui, le plus utilisé est la technologie Illumina qui utilise une amplification clonale et un séquençage par synthèse. Le séquençage peut être "single end" (chaque read est indépendant) ou "paired-end" (les reads sont pairés). Après le séquençage, des millions de reads sont obtenus.

Etape 2 : Contrôle qualité

A la sortie du séquenceur on retrouve des fichiers fastQ. Ce genre de fichier est composé de blocs de 4 lignes représentants un read. Grâce aux fichiers fastQ, la qualité du séquençage peut être estimée à l'aide de programmes comme FastQC.

Etape 3 : Mapping

Cette partie de l'analyse consiste à aligner tous les reads sur le génome de l'organisme étudié. Un read est mappé sur la régions du génome qui lui est la plus similaire. Cette étape permet d'obtenir des fichiers BAM dans lesquels chaque ligne correspond à un read. De plus, la moyenne du nombre de reads mappés sur une régions est appelée la profondeur.

Etape 4 : Quantification

Le nombre de reads est un témoin de l'abondance d'ARN dans la cellule. Ainsi, il est possible d'estimer le niveau d'expression d'un gène. C'est pourquoi il est important de compter les reads mappés pour chaque gène. Le but de cette étape est de remplir une table de comptage afin de pouvoir la manipuler facilement avec R par exemple.

Etape 5 : Statistiques

Différents résultats statistiques peuvent être obtenus ainsi que des graphiques ou encore des carte de densité pour les exons. Il est important de normaliser les données et de comparer les p-value ajustées obtenues après différents tests. Tout ceci permet d'établir une liste de gène différentiellement exprimés.

Objectifs

Aujourd'hui, de plus en plus d'études utilisent le séquençage de l'ARN, il en résulte de plus en plus de données à analyser. Pour essayer de répondre à cette problématique nous avons décidé d'élaborer une application interactive grâce au package RShiny. Cette application a pour but d'aider à l'analyse de données RNA-seq le plus profondément possible en répondant aux plus de questions possibles et permettre une visualisation intuitive des résultats. Le but de ce projet est de nous permettre d'en apprendre plus sur cette nouvelle technique de séquençage de l'ARN et son analyse. Cela nous permettra aussi d'apprendre à utiliser différents packages disponibles sous R comme Shiny pour la conception de l'application ou Markdown pour la rédaction du rapport.

Matériels et méthodes

Packages

Analyse RNAseq

DESeq2

gplot et ggrepel

Tidyverse

NMF

Création de l'application

Shiny

Shinydashboard

Autres

Rédaction du rapport

Markdown

LaTeX

Jeu de données

Nous allons tout d'abord procéder à une analyse de l'expression différentielle de gènes en utilisant le package DESeq2. Pour cela, nous avons récupéré un jeu de données de l'étude RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells de Himes BE, Jiang X, Wagner P, et al. Les glucocorticoïdes sont utilisés pour traiter l'asthme et le but de cette étude est de comprendre le mécanisme dans les muscles lisses des voies respiratoires en utilisant la technologie RNA-seq.

Cette expérience rassemble 8 échantillons : 4 traités avec du dexamethasone (glucocorticoïde synthétique) et 4 échantillons contrôles sans traitements. Nous avons donc une table de comptage de reads dans laquelle on trouve le nombre de reads mappés sur chaque gène pour chaque échantillon. De plus, nous avons un fichier d'annotation des gènes contenant des informations sur tous les gènes.

Résultats

Analyse de données RNA-seq

Application RShiny

Conclusion et perspectives

Bibliographie

1. Himes BE, Jiang X, Wagner P, et al. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. PLoS One. 2014;9(6):e99625. Published 2014 Jun 13. doi: 10.1371/journal.pone.0099625

2. Koch CM, Chiu SF, Akbarpour M, et al. A Beginner's Guide to Analysis of RNA Sequencing Data. Am J Respir Cell Mol Biol. 2018;59(2):145-157. doi: 10.1165/rcmb.2017-0430TR

3. Julien Delafontaine. (2013, septembre 11). Analyse de données RNA-seq : mode d'emploi. Consulté à l'adresse https://bioinfo-fr.net/lanalyse-de-donnees-rna-seq-mode-demploi

4. Bioconductor - Home. (2003). Consulté à l'adresse https://www.bioconductor.org/

5. Wickham, H. (2016). Mastering Shiny. Consulté à l'adresse https://mastering-shiny.org/

Résumé

Abstract

Annexe 1 : Script R pour l'analyse RNAseq via DESeq2

Annexe 2 : Scripts R pour l'application Shiny

User Interface

Server

Fonctions