Programme de la formation Bioinformatique pour les NGS - ATGC Montpellier

Plateforme ATGC NGS

21 juin 2016

Ce document donne un programme détaillé préliminaire de la formation. Il est susceptible de changer, notamment en fonction de la rapidité d'avancement, mais les thèmes prévus restent. Les cours et TP seront donnés par les membres de la plateforme et des intervenants extérieurs. Le principe est d'alterner des cours qui présentent les notions, les outils, et les travaux dirigés de manipulation de ces outils sur ordinateur. La formation aborde les analyses avec et sans génome de référence.

1 Linux : shell et lignes de commandes (Jour 1)

Introduction aux commandes Linux pour permettre d'exécuter les outils d'analyse

- lancement d'un terminal
- répertoire et fichiers
- déplacement, copie, renommage, suppression de fichiers
- redirection
- compression

2 Introduction, technologies NGS et format de données (Jour 1)

- Présentation rapide des technologies, des formats de données, des manipulations de base
- Type de séquençage : single, paired-end, mate pairs, orienté ou non
- Survol des types d'expériences NGS et des analyses bioinformatiques
- Exemples de pipeline d'analyse
- Exécution en ligne de commande
- Extraction de données (wget, SRAToolkit)
- Filtrage et nettoyage des données avec cutadapt et FastQC

3 Mapping pour le génomique, analyse des résultats, et visualisation (Jour 2)

- Mapping : définition et méthodes ; rôle et importance des index de génomes
- Les variantes du mapping et leur spécificité
- Présentation des principaux outils : BWA, Bowtie, CRAC.
- TP avec BWA ou Bowtie (sans épissage) et CRAC (avec épissage)
- Difficultés et pièges, évaluation, et mappabilité
- Analyse post mapping : calculer les couvertures (samtools)
- Visualisation

4 Détection de variants génomiques (Jour 2)

- Outils d'analyse après mapping pour la détection de variants
- Format de fichier VCF pour le stockage de variants
- Présentation GATK, samtools, VarScan

— TP identification et sélection des variants après mapping (samtools)
— Production d'un fichier de sortie VCF

5 Assemblage génomique et correction de reads (Jour 3)

— Les principes des méthodes d'assemblage : gloutonne, avec graphe de chevauchements, ou avec graphe de De Bruijn
— Principaux outils : CAP3, Velvet, Minia
— Correction : besoin, principes, méthodes, outils
— Cas des reads courts et reads longs
— Méthode et pipeline pour reads longs et courts par graphe de chevauchements : nettoyage (cutadapt), correction (LoRDEC), évaluation données (FastQC), assemblage (CAP3), évaluation (QUAST), (éventuellement clustering)
— TP pipeline approche par graphe de chevauchements
— Méthode pour reads courts : approche par graphe de De Bruijn ; TP Velvet/Minia

6 Analyse transcriptome (Jour 4)

— RNA-seq : applications ; analyse qualitative et quantitative.
— Mapping pour RNA-seq : génome ou transcriptome de référence
— Identifier les variants d'épissage, des candidats d'ARN de fusion ou d'ARN chimères
— Utilisation en (ré-)annotation de génome
— Cours/TP Comparaison de deux conditions, expression différentielle :

1. Mise en forme et chargement des données
2. Normalisation et filtrage
3. Matrice de design
4. Tests statistiques : calcul du facteur de variation (fold change) et de la P-valeur
5. Extraction des résultats

7 Transcriptome sans génome de référence (Jour 4)

— Mots-clés : Assemblage de novo, RNA-seq, épissage alternatif, SNP, indels
— Description : KisSplice est un assembleur local de transcriptome. Il permet d'identifier et quantifier des SNPs, indels et évènements d'épissage alternatif à partir de données RNA-seq. KissDE est un paquet R permettant de tester si un variant (génomique ou d'épissage) est enrichi dans une condition . Dans le cas où un génome de référence est disponible, le paquet KisSplice2RefGenome permet d'annoter les variants découverts par KisSplice.
— Le TP sera l'occasion de manipuler les différentes briques de cette suite logicielle, en utilisant un jeu de données issu du projet ENCODE correspondant au séquençage du transcriptome d'une lignée cellulaire SKN-S-H (neuroblastome) traitée ou non par acide rétinoïque. En fin de TP, on obtient une liste de gènes différentiellement épissés en présence d'acide rétinoïque. Au cours du TP, on discutera les avantages/inconvénients de l'approche assemblage local par rapport à l'assemblage global (Trinity) et par rapport à une approche de mapping (TopHat).

8 Retour sur assemblage et analyses post-mapping - Discussion (Jour 5)

On reviendra sur les points difficiles de cette formation en particulier l'assemblage et les analyses post-mapping. La dernière partie sera consacrée à une discussion avec les participants.