Programme de la formation Bioinformatique pour les NGS Lyon

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 TD seront donnés par les membres du Laboratoire de Biométrie et Biologie Evolutive. 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 Unix (Jour 1)

Introduction aux commandes Linux pour permettre d'exécuter les outils d'analyse (TD)
— 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
— TD en exécution en ligne de commande
— Extraction de données (wget, SRAToolkit)
— Filtrage et nettoyage des données (UrQt, cutadapt)
— Controle de la qualite des donnees (FastQC)

3 Mapping pour le génomique, détection de variants 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.
— Difficultés et pièges, évaluation, et mappabilité
— TD mapping
— Alignement de lectures sur un génome de référence
— Analyse post mapping : calculer les couvertures (samtools)
— Visualisation (IGV)
— Présentation d'outils d'analyse après mapping pour la détection de variants (GATK, samtools, BEDtools)
— TD détection de variants
— Détection et sélection de variants
— Production de fichier VCF pour le stockage de variants
— Discussion autour des paramètres et différents jeux de données (reliquats biologiques, PoolSeq, espèces polyploïdes…)

4 Assemblage génomique et correction de reads (Jour 3)

— Les principes des méthodes d'assemblage
— Principaux assemblers : IDBA, Ray, SOAP denovo, SPAdes
— Correction : besoin, principes, méthodes, outils

— Cas des reads courts (paired-ends et mate pairs) et reads longs
— Méthode et pipeline pour reads longs et courts par graphe de chevauchements : nettoyage (cutadapt, UrQt), correction (LoRDEC), évaluation données (FastQC), assemblage de contigs et de scaffolds de paired-end reads (IDBA, Ray, SOAP denovo) et de mate-pairs (SSPACE), évaluation des assemblages (N50calc, BUSCO, QUAST), éventuellement mapping sur un génome proche pour évaluation de la synténie (SynMap)
— **TD assemblage de novo** à partir de reads courts paired-ends (approche par graphe de De Bruijn ; IDBA)
— Rappels de statistiques : Vocabulaire de base (variables aléatoires, données discrètes et continues), Lois de distribution (binomiale, Poisson, binomiale négative; normale), Test simples, Tests multiples (Bonferroni, FDR)

5 Analyse de transcriptome avec génome de référence (Jour 4)

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

6 Transcriptome sans génome de référence (Jour 5)

— 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 **TD** 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 TD, on obtient une liste de gènes différentiellement épissés en présence d'acide rétinoïque. Au cours du TD, 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).

7 Workshop autour de jeux de données (Jour 5)

Les participants travailleront par groupe avec le soutien des intervenants pour définir les pipelines d'analyse à recommander pour des jeux de données particuliers (qui peuvent être fournis par les participants eux-mêmes).