Next-Generation Sequencing
Analyse de variants génotypiques

Thomas Dugé de Bernonville

December 12, 2018

Ce TD a pour objectif de vous initier à la détection de SNP (Single Nucleotide Polymorphism) pouvant être à l'origine de désordres phénotypiques chez l'Humain. Vous travaillerez sur des données de séquençage haut débit de type WES (Whole Exome Sequencing). L'interface Galaxy sera utilisée pour:

1. **Vérifier la qualité des lectures**
2. **Aligner les lectures sur la séquence de référence Hg38**
3. **Effectuer une sortie des SNP identifiés.**

Les SNP detectés seront alors confrontés à des données obtenues sur la population humaine pour déterminer si votre échantillon possède des SNP indicateurs de désordres phénotypiques.

1 Les bases de données NGS

Il existe plusieurs bases de données archivant les données issues de NGS ainsi que les assemblages réalisés. Les deux plus importantes sont:

- NCBI: Sequence Read Archive (SRA - `http://www.ncbi.nlm.nih.gov/sra/`)
- EBI: European Read Archive (ENA - `http://www.ebi.ac.uk/ena`)

Beaucoup d'accessions sont communes entre les deux. Ces bases de données sont utilisées pour déposer des données publiées ou non. Elles servent à garder une trace des travaux antérieurs, mais offrent également la possibilité de les retraiter. Les accessions présentes dans le SRA et l'ENA existent sous 4 niveaux:

1. Numéro de run: SRR ou ERR
2. Numéro d'expérimentation: SRX ou ERX
3. Numéro d'échantillon: SRS ou ERS
4. Numéro de projet: SRP ou ERP

Les données utilisées pour la séance sont tirées du run SRR866988. Pour récupérer des informations concernant cet échantillon, ouvrez un navigateur web et entrez l'accession correspondante dans l'ENA. **Indiquez:**

- Description de l'échantillon
- Technologie
- Single end ou Paired end
- Nombre de reads
- Projet
- Description du projet

Pour info: vous pouvez récupérer un certain nombre d'informations sur l'état de la base ENA, à ce lien `http://www.ebi.ac.uk/ena/about/statistics`.

2 Préparation des données

Nous allons à présent analyser une partie des lectures obtenues dans le projet comprenant le run SRR866988. L'analyse des données NGS peut constituer un réel challenge informatique car des séquençages à forte profondeur génère des fichiers relativement lourd (plusieurs Go). De plus, la manipulation des nombreuses lectures de courtes taille générées en Illumina ou en SOLiD nécessite généralement une puissance informatique convenable (utilisation de stations de travail 8/16 CPUs ou machines de calcul) ainsi que plusieurs heures à plusieurs jours de calcul. Pour des raisons de temps et de limite de ressources, vous travaillerez donc sur des fichiers de petite taille. Pour cela, vous allez utiliser l'interface Galaxy en libre accès sur Internet pour des demandes limitées en taille. Cette interface graphique regroupe un grand nombre de programmes fonctionnant normalement en ligne de code sous Linux ou MacOs, rendant leur utilisation plus intuitive.

Commencez par créer un dossier sur le bureau où vous stockerez tous vos fichiers de travail.

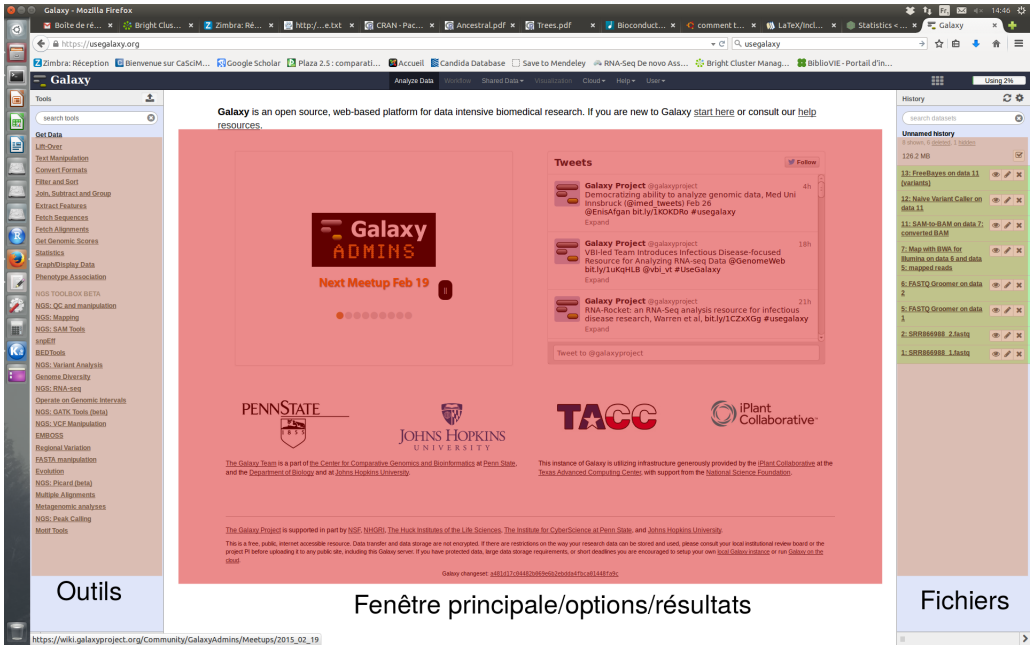
2.1 Découverte de Galaxy: chargement et contrôle qualité des échantillons

Feuille de route Les étapes que vous allez réaliser sur Galaxy sont les suivantes et seront décrites dans les paragraphes suivants:

1. Chargement des données (outil "File Upload")
2. Contrôle qualité (outil "FastQC")
3. Conversion du format Fastq (outil "Fastq Groomer")
4. Alignement au génome de référence hg19 (avec "Bowtie2")
5. Création d'un fichier de couverture (avec "BedTools")
6. Visualisation dans IGV
7. Détection de SNP (avec "FreeBayes") et génération d'un fichier VCF
8. Assemblage de donnée NGS (outil "VelvetOptimiser")

N'hésitez pas à vous référer à cette liste si vous vous sentez perdus.

Vous pouvez vous connecter à des serveurs Cloud en libre accès: `http://galaxy-qld.genome.edu.au/galaxy` ou `http://galaxy-tut.genome.edu.au/galaxy`. L'interface se compose de trois fenêtres.



Pour info: la création de compte permet des utilisations plus poussées. Pour un usage vraiment intensif, l'idéal est une installation locale et personnalisée.

Les différents outils triés par catégorie sont placés tout à gauche. La partie centrale permet l'affichage du paramétrage des programmes ainsi que les résultats des analyses. Enfin, les données chargées ainsi que les fichiers de résultats sont indiqués à droite. **Commencez par charger vos données: téléchargez les deux fichiers à l'adressse suivante (`http://bbv-ea2106.sciences.univ-tours.fr/index.php/web-links/60-td-betd`) puis chargez les dans Galaxy via l'outil "GetData".**

Visualizez un des fichiers en cliquant dessus. Les données NGS sont écrites dans le format **fastq**. Alors que le format **fasta** ne contient que la séquence (nucléotidique ou protéique) précédée de sa désignation (commençant forcément par le caractère '>'), le format **fastq** contient en plus un autre vecteur de taille identique à la séquence. Ce vecteur correspond à l'encodage Phred de la mesure de la qualité du séquençage. A chaque base séquencée est attribué un score révélant la qualité du séquençage. C'est une caractéristique importante qui permettra notamment de s'assurer que le processus de séquençage n'a pas rencontré de problème. **Lancer le contrôle qualité FastQC sur chacun des fichiers fastq en utilisant les paramètres par défaut (NGS toolbox, NGS: QC and manipulation.**

Le programme **FastQC** analyse la qualité des données de séquençage en se focalisant sur plusieurs points. Parmi les principaux, il faut regarder la répartition de la qualité des lectures donnée dans le fichier fastq. Cette répartition est donnée par un ensemble de boxplots représentée à chacune des bases. **FastQC** rapporte également la distribution du %G+C ainsi qu'une analyse de la composition en k-mers. Ces k-mers sont des mots de longueurs k qui sont construits selon la composition des lectures. (Dans le cours, nous avons vu que le fonctionnement des assembleurs *de novo* repose notamment sur la génération d'une bibliothèque de tous les mots possibles de longueur k extraits du contenu en lectures. *Il faut retenir que cette analyse basée sur les k-mers est une solution informatique moins coûteuse en ressources que la manipulation directe des lectures.* Vous pouvez regarder la répartition des k-mers donnée par **FastQC**. Le programme renseigne également sur la surabondance de certains k-mers qui pourrait témoigner de la présence de séquences contaminantes.

2.2 Mapping des lectures sur l'assemblage Hg38 du génome humain

Conversion des fichiers fastq. Certains fichiers fastq peuvent présenter des problèmes lorsqu'ils sont utilisés tels quel, ce qui est dû à des différences notamment dans l'encodage des vecteurs qualité, ainsi que des conflits entre versions. L'outil **FastQ Groomer** va permettre de rendre compatibles les fichiers que vous avez chargé dans l'environnement de travail. L'outil se trouve dans la rubrique *NGS: QC and manipulation* dans le panneau de gauche. Indiquez le fichier que vous souhaitez convertir et renseigner le format original, **Sanger & Illumina 1.8+** (qui est notamment le format en sortie d'un séquenceur type Illumina GIIa).

Mapping. Vous pouvez à présent accéder aux outils de mapping des lectures présentes dans vos fichiers. **L'objectif est d'aligner l'ensemble des lectures sur un assemblage (génome ou transcriptome) de référence, soit pour assembler une nouvelle séquence soit pour détecter des variants nucléotidiques**. Dans notre cas, nous voulons détecter la présence de SNP (Single Nucleotide Polymorphism) dans notre échantillon par rapport au génome humain. Cette étape est un premier pas dans le diagnostic de désordres phénotypiques. Comme vu en cours, il existe de nombreux outils, les plus populaires reposant sur l'utilisation d'un algorithme Burrows-Wheeler comme **BWA** et **Bowtie**. Toutefois, l'utilisation de BWA devient de plus en plus rare et nous utiliserons Bowtie2 qui est le plus utilisé à l'heure actuelle. Les champs à renseigner sont:

- La séquence de référence: assemblage humain hg38
- Design single-end ou paired-end
- Les fichiers d'entrée: forward pour fichier **_1**, reverse pour fichier **_2**

Attention: pour les fichiers d'entrée, veillez à bien indiquer les fichiers convertis par FastQ Groomer plutôt que les fichiers originaux.

Le fichier de résultats au format **SAM** (Sequence Alignment/Map) ou **BAM** (format compressé) contient les valeurs d'alignement de chaque lecture à la séquence de référence. Les premières lectures dans ce fichier sont celles qui n'ont pas pu être cartographiées. En parcourant le fichier, vous trouverez les coordonnées des lectures correctement alignées. Les caractéristiques de chaque colonne sont décrites lorsque vous ouvrez la page du programme dans Galaxy. Les programmes de détection des SNP utilisent plutôt le format binaire du SAM, le format **BAM**. Assurez-vous que le fichier soit bien au format **BAM**, sinon utilisez l'outil de conversion **SAM-to-BAM** dans la rubrique *NGS: SAM Tools*. Renseignez correctement le fichier d'entrée (le fichier SAM) ainsi que la séquence de référence utilisée (hg38).

Visualisation du mapping. *Pour faciliter la visualisation, vous allez tout d'abord lancer le calcul de la couverture du génome de référence par votre échantillon. Dans BedTools, utilisez le programme* ***Create a BedGraph of genome coverage*** *sur le fichier BAM obtenu précédemment, en précisant de ne pas rapporter les zones sans couverture. Les données de mapping et de la couverture peuvent être visualisées à l'aide de certains logiciels tels que Integrated Genomics Viewer (IGV, http://www.broadinstitute.org/software/igv/download). Téléchargez le fichier binaire au format zip; dans le terminal (Ctrl+alt+T), mettez vous dans le dossier et lancer l'application java (java -jar igv.jar). Vous pouvez aussi cliquer dans l'objet* ***Genome Coverage BedGraph*** *sur Display with IGV web current. Lorsque IGV se lance, la première étape consiste à charger le génome de référence (hg38) en haut à gauche. Ensuite sur Galaxy, téléchargez les résultats de l'alignement au format BAM (dataset et direcbam_index). Etant donné que nos fichiers de départs ne contiennent que peu de lectures, il faut zoomer directement dans IGV sur une zone contenant des données. L'utilisation du fichier de couverture (au format BED) obtenu précédemment est ici très utile pour rechercher ces zones contenant des lectures alignées. Vous pouvez aussi rentrer des coordonnées directement dans IGV, par exemple 11:110660192.*

3 Analyse de mutations

A présent que vous avez cartographié les lectures issues du séquençage de votre échantillon, vous allez pouvoir déterminer si cet échantillon contient des variations nucléotidiques par rapport à la séquence de référence et interroger des bases de donnéees pour déterminer si ces variations peuvent avoir un impact phénotypique. Vous allez donc:

1. Détecter les variants: générer un fichier VCF
2. Analyser l'importance des variations

3.1 Génération du fichier Variant Call Format (VCF)

Pour détecter des variations nucléotidiques par rapport à la séquence de référence, il suffirait simplement de comparer la séquence des lecteurs à celle de référence. Il est cependant possible que des variations observées dans les lectures soient en réalité dues à des erreurs de séquençage. Dans ce cas, il se

peut que le nucléotide observé comme muté ait un score de qualité plutôt faible dans le fichier fastq. Si l'on dispose d'une profondeur de séquençage suffisant, on pourra également observer que les autres lectures cartographiées au même endroit sur la séquence de référence ne présentent pas de mutations. Ainsi, il existe des programmes qui vont tenir compte de ces différents critères (qualité du séquençage et nombre de lectures présentant la même mutation) pour calculer une probabilité sur la vraisemblance de cette mutation.

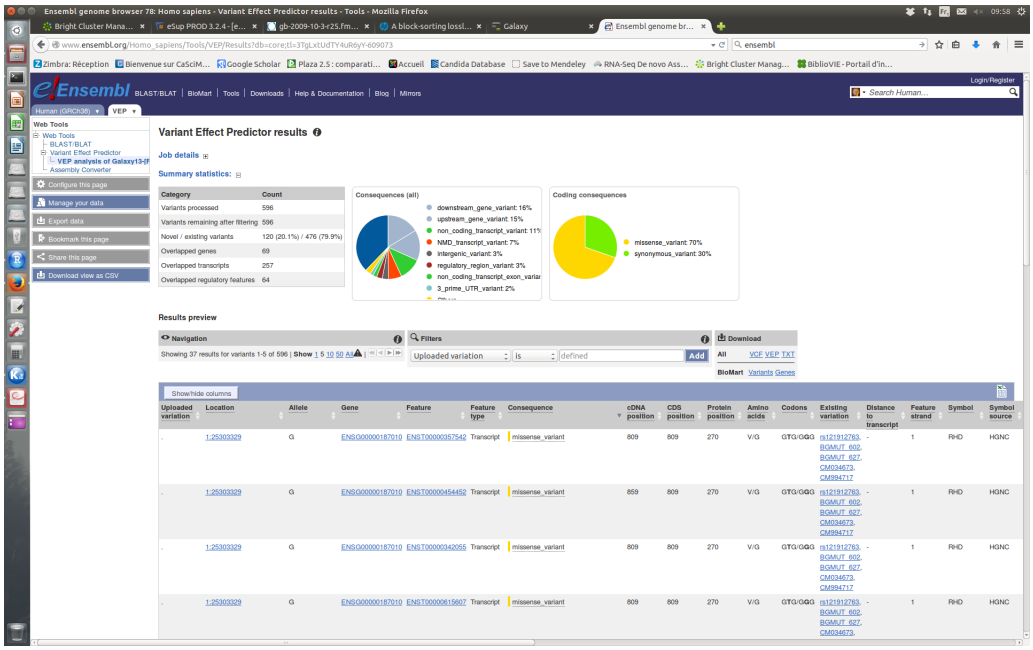
Les fichiers VCF correspondent à une structure uniformisée pour répertorier des données génotypiques dans une analyse donnée (qui peut posséder plusieurs échantillons). Chaque fichier VCF possède des métadonnées (description de l'analyse, ...), une ligne d'en-têtes et une variation par ligne. Il existe plusieurs outils dans la catégorie *NGS: Variant Analysis* sous Galaxy. Nous allons utiliser **FreeBayes** qui va répertorier les différents variants génotypiques de notre échantillon et attribuer une probabilité à leur existence. **FreeBayes** utilise en entrée le fichier BAM généré après l'alignement. Il faut également renseigner la séquence de référence utilisée (hg38).

Une fois l'analyse **FreeBayes** terminée, vous pouvez visualisez le résultat. Le fichier débute bien par un ensemble de métadonnées puis affiche ensuite les différents variants dans un format uniformisé. FreeBayes permet de détecter aussi bien des SNP que des insertions/délétions (indels). La colonne $QUAL$ rapporte la valeur du test statistique pour la probabilité de l'occurence du variant. Cette sortie étant plutôt indigeste, nous allons maintenant procéder à l'analyse des variants présentées dans ce fichier.

3.2 Prediction des variants génotypiques

L'étape suivant l'identification des variations dans votre échantillon par rapport à la séquence de référence consiste à comparer ces variations avec le catalogue des variations déjà caractérisées. Pour rappel, de grandes études ont été initiées dès le début des années 2000 pour cataloguer un grand nombre de SNP puis ensuite les associer à des traits phénotypiques. Vous allez utiliser votre fichier VCF dans l'outil **Variant Predictor (Ve!P)** disponible sur le serveur Ensembl (outils bioinformatiques mis à disposition par EMBL-EBI et le Wellcome Trust Institute). Enregistrer le fichier VCF à partir de galaxy et chargez-le dans **Ve!P** (`http://www.ensembl.org/Homo_sapiens/Tools/VEP`).

Le serveur **Ve!P** présente la comparaison des variants génotypiques de votre échantillon avec ceux déjà référencés. Un résumé des variations est présenté en début de page (catégories, conséquences,...).



Dans votre cas, le nombre de variants reste assez simple à analyser avec les moyens dont nous disposons. Ceci ne serait pas le cas si vous aviez traité l'ensemble des données WES pour un voir plusieurs échantillons en même temps. Plutôt que d'afficher l'ensemble des variants découverts, nous allons les filtrer par importance. Toutes les mutations n'ont bien entendu pas le même impact. Celles présentes dans les introns et les séquences non codantes par exemple ont moins de chances de provoquer un désordre phénotypique. Vous obtiendrez une description détaillée des différentes variations et leurs importances relatives à cette adresse: `http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences`. Utilisez l'outil de filtre pour rechercher les variants dont la *conséquence* est *missense_variant* (mutations non-synonymes). Pour visualiser l'ensemble de ce type de variants, n'oubliez pas d'utiliser ensuite l'outil de navigation pour tout afficher. Le premier variant observé correpsond à une mutation déjà référencée qui remplace une thréonine par une arginine (T/R, codon: ACA/AGA) à la position 68 de la séquence codante du gène ENSG00000149300. Cette mutation porte le numéro d'accession *rs7124407*.

Uploaded variation	Location	Allele	Gene	Feature	Feature type	Consequence	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variation	Distance to transcript	Feature strand	Symbol	Symbol source
.	11:111924361	G	ENSG00000149300	ENST00000276601	Transcript	missense_variant, splice_region_variant	164	68	23	T/R	ACA/AGA	COSM4145309, rs7124407	-	1	C11orf52	HGNC

En cliquant sur le numéro d'accession rs, vous aurez accès à un grand nombre d'information sur l'éventuelle implication de la mutation sur le phénotype, la répartition dans la population humaine, le déséquilibre de liaison avec les autres allèles,... Vous pouvez obtenir ces informations en rentrant ce même numéro sur le NCBI, à cette adresse: `http://www.ncbi.nlm.nih.gov/gquery/`. Vous accéderez par ce lien à la base de données **dbSNP**.

3.3 Assemblage des données génomiques

A titre indicatif, vous pouvez lancer l'assembleur Velvet (*dans les outils* ***NGS:Assembly***) sur les fichiers fastq. Cet assembleur teste différentes longueurs de k pour les k-mers (cf cours). Ceci permet d'homogénéiser les contigs qui résulteront de l'assemblage, permettant à la fois une bonne reconstruction des séquences de courte et longue tailles. L'analyse vous donne un premier fichier "log" qui contient les informations de couverture sur les différents contigs reconstruits et un deuxième fichier correspondant aux séquences des contigs. A partir de cet assemblage, il est possible de réaliser des analyses de séquences homologues par BLAST par exemple.

4 Résumé et pour aller plus loin...

Vous êtes maintenant familiarisés avec la détection de variants génotypiques dans des données de séquençage haut débit. Vous devriez être capables de retrouver les différentes étapes menées depuis l'acquisition des données:

1. Contrôle qualité des données
2. Conversion si nécessaire
3. Mapping à la séquence de référence
4. Création d'un fichier VCF
5. Recherche de variants potentiellement graves.
6. Assemblage de données issues de NGS

Pour cela, vous avez utilisé l'interface **Galaxy** et l'outil de prédiction **Ve!P**. Ce sont les bases d'un diagnostic utilisant les technologies NGS.

Il existe depuis quelques années de nombreuses entreprises proposant de séquencer votre génome ou de détecter des SNP pour découvrir vos origines génétiques ou bien faire du pré-diagnostic. C'est le cas notamment de 23andme (`https://www.23andme.com/`). L'article suivant vous donnera quelques pistes de réflexions sur l'usage de ces données...