TP Module 11 : Réalisation d'une analyse d'exome

Titouan Guillou

Présentation de Galaxy.eu

Galaxy.eu est une plateforme informatique en ligne dédiée à l'analyse de données biomédicales. Il a été conçu pour fournir aux scientifiques et aux chercheurs un environnement intuitif et accessible pour la gestion, l'analyse et la visualisation de leurs données de recherche. Galaxy.eu utilise une interface web conviviale pour fournir un accès à une large gamme d'outils d'analyse de données, sans la nécessité d'installer des logiciels sur les ordinateurs personnels des utilisateurs.Galaxy.eu est soutenu par une communauté internationale de développeurs et de scientifiques, qui travaillent ensemble pour développer et maintenir la plateforme. Cela garantit que Galaxy.eu est constamment mis à jour avec les derniers outils d'analyse de données et les méthodes les plus performantes.

Le séquençage

Il existe différentes techniques de séquençage :

- Sanger (méthode des années 80) utilisés en consultation pour petit débit de traitement
- SNP arrays (puce qui interroge des milliers de positions dans notre génome). Cela va aider à faire de portraits génétiques pour des fins de loisirs par exemple

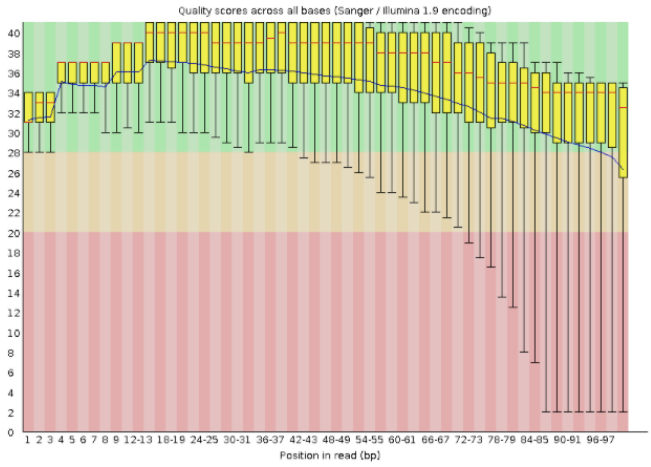
Nous allons traiter différentes données au cours de ce TP :

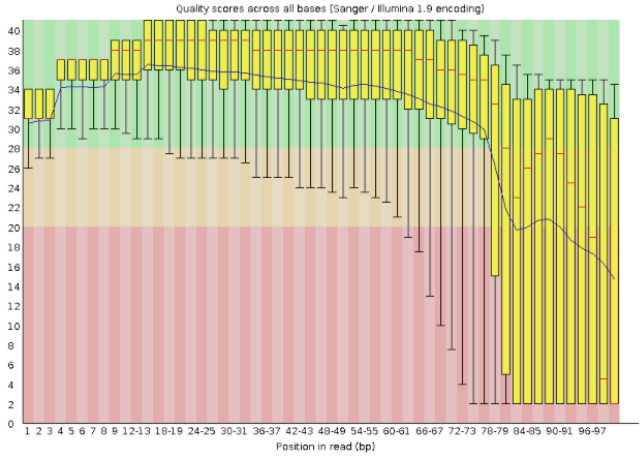
- Des séquences d'ARN de tissue normaux (sang) : normal R1 &R2
- Des séquences d'ARN de tissue tumoral (cellule du cancer du poumon) : tumor R1 &R2

En génétique somatique, on extrait de l'ADN dans des tissus normaux et dans la tumeur et on compare dans le but de trouver des mutations.

1ère étape : Nous chargeons 4 tables de données, à partir de fichiers FASTQ : normal_R1.fatsq ; normal_R2.fatsq ; tumor_R1.fatsq ; tumor_R2.fatsq.

2ème étape : On convertit les fichiers en Fastqc pour obtenir des informations sur les différents reads des 4 fichiers. On obtient plusieurs reads pour chaque fichier car on prend plusieurs cellules avec chacune son ADN. La tumeur est un mélange de plusieurs cellules (cellules sanguines, cellules tumorales...) donc tous les reads ne porte pas tous la mutation. On remarque néanmoins que la qualité des reads des fichiers R2 baisse vers la fin.





Measure	Value
Filename	normal_R1_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	45323

Measure	Value
Filename	normal_R2_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	45323

⇒L'objectif après cette première analyse est de créer une pipeline sur une des classes (« tumor » ou « normal ») et de ensuite juste la sauvegarder pour l'appliquer sur la classe pas encore étudiée. J'ai choisir de commencer par la classe « normal »/

3ème étape : Nous commençons par un traitement « Trimmomatic » des fichiers normal_R1.fastq et normal_R2.fastq. Trimmomatic est un outil utilisé pour préparer les données de séquençage en vue d'analyses ultérieures. Il est utilisé pour éliminer les séquences de mauvaise qualité et les régions inutiles des lectures de séquençage pour améliorer la qualité générale des données. Les données nettoyées par Trimmomatic sont plus fiables et peuvent conduire à des résultats plus précis lors des analyses ultérieures, telles que l'assemblage de génome ou la découverte de variants.
On obtient un fichier Fastq, où l'on peut vérifier leur gain de qualité en réalisant un fastqc sur ces fichiers.

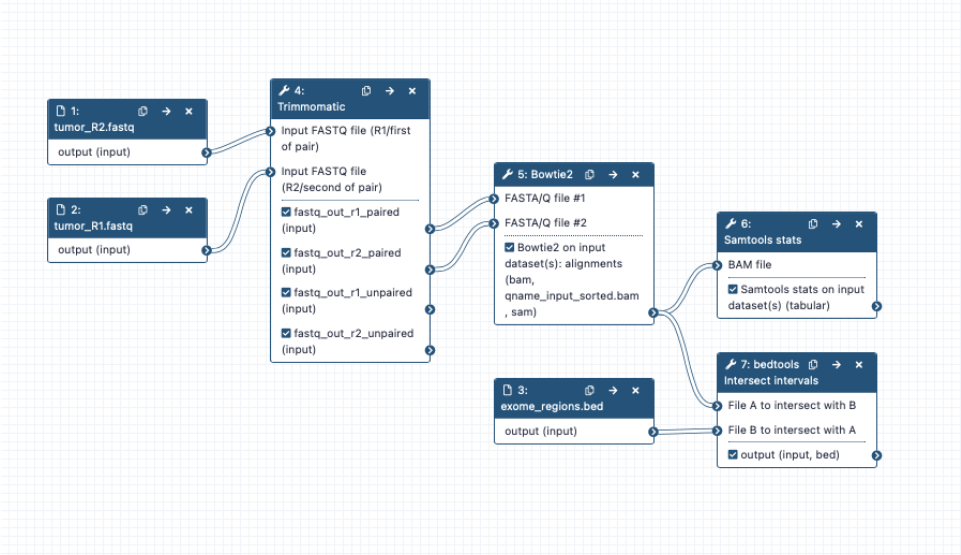
4ème étape : On réalise un traitement Bowtie sur les fichiers obtenus Trimmomatic on normal_R1.fastq et Trimmomatic on normal_R2.fastq. On sélectionne Paired-End et on fait bien attention à choisir le génome de référence Hg19 (Homo Sapiens). Bowtie est un outil d'alignement de séquences qui peut être utilisé pour aligner les lectures de séquençage sur un référentiel de génome. Les fichiers Trimmomatic normal R1 et Trimmomatic normal R2 représentent les lectures d'un échantillon paire-à-fin, chacune provenant d'une extrémité différente du fragment d'ADN. En utilisant Bowtie avec "Paired end" sur ces deux fichiers, les lectures peuvent être alignées de manière à reconstruire les fragments originaux d'ADN. Cela est important pour les analyses qui nécessitent une information sur la structure des fragments d'ADN, telles que l'analyse d'exome.
On obtient un fichier BAM, exploitable sur IGV et une database sur hg19.

5ème étape : On veut alors traiter ce fichier BAM en utilisant la « boite à outils » Samtools. Samtools est utilisé pour effectuer des analyses statistiques sur les données d'alignement, telles que le calcul des profils de couverture, le filtrage des alignements et la compression des données. Avec Samtools, on peut extraire des informations qualitatives sur les alignements, telles que les statistiques de qualité, les statistiques de qualité par base, les statistiques de qualité par cycle et les statistiques de qualité globale pour les lectures alignées. Cela permet aux utilisateurs de s'assurer que les données d'alignement sont de qualité suffisante pour les analyses ultérieures.

SN	raw total sequences:	89438	# excluding supplementary and secondary reads
SN	filtered sequences:	0	
SN	sequences:	89438	
SN	is sorted:	1	
SN	1st fragments:	44719	
SN	last fragments:	44719	
SN	reads mapped:	46794	
SN	reads mapped and paired:	4220	# paired-end technology bit set + both mates mapped
SN	reads unmapped:	42644	
SN	reads properly paired:	2598	# proper-pair bit set
SN	reads paired:	89438	# paired-end technology bit set
SN	reads duplicated:	0	# PCR or optical duplicate bit set
SN	reads MQ0:	695	# mapped and MQ=0
SN	reads QC failed:	0	
SN	non-primary alignments:	0	
SN	supplementary alignments:	0	
SN	total length:	7917550	# ignores clipping
SN	total first fragment length:	3958775	# ignores clipping
SN	total last fragment length:	3958775	# ignores clipping
SN	bases mapped:	3981792	# ignores clipping
SN	bases mapped (cigar):	3981792	# more accurate
SN	bases trimmed:	0	
SN	bases duplicated:	0	
SN	mismatches:	7099	# from NM fields
SN	error rate:	1.782866e-03	# mismatches / bases mapped (cigar)
SN	average length:	88	
SN	average first fragment length:	89	
SN	average last fragment length:	89	
SN	maximum length:	101	
SN	maximum first fragment length:	101	
SN	maximum last fragment length:	101	
SN	average quality:	36.5	
SN	insert size average:	234.7	

6ème étape : On procède ensuite à l'intersection des intervalles Bedtools, avec le fichier bed des « exome_regions ». Bedtools est un outil utilisé pour effectuer des analyses génomiques sur les données d'alignement. Il peut être utilisé pour effectuer des comparaisons entre les régions d'exome et les régions du génome qui ont été couvertes par les lectures de séquençage. Bedtools peut être utilisé pour déterminer les régions du génome qui ont été couvertes par un certain nombre de lectures de séquençage et produire une carte de la couverture génomique pour les différents échantillons.
On obtient un fichier Bam, visionnable sur IGV.

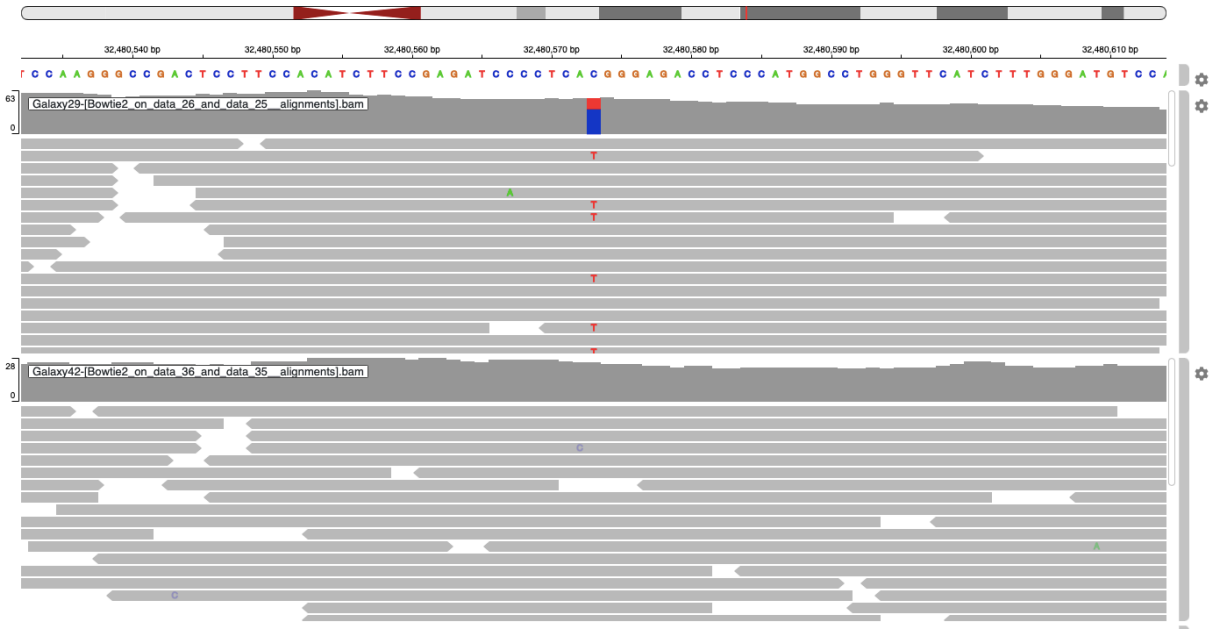
A la fin de cette étape on a donc fini de créer notre workflow.

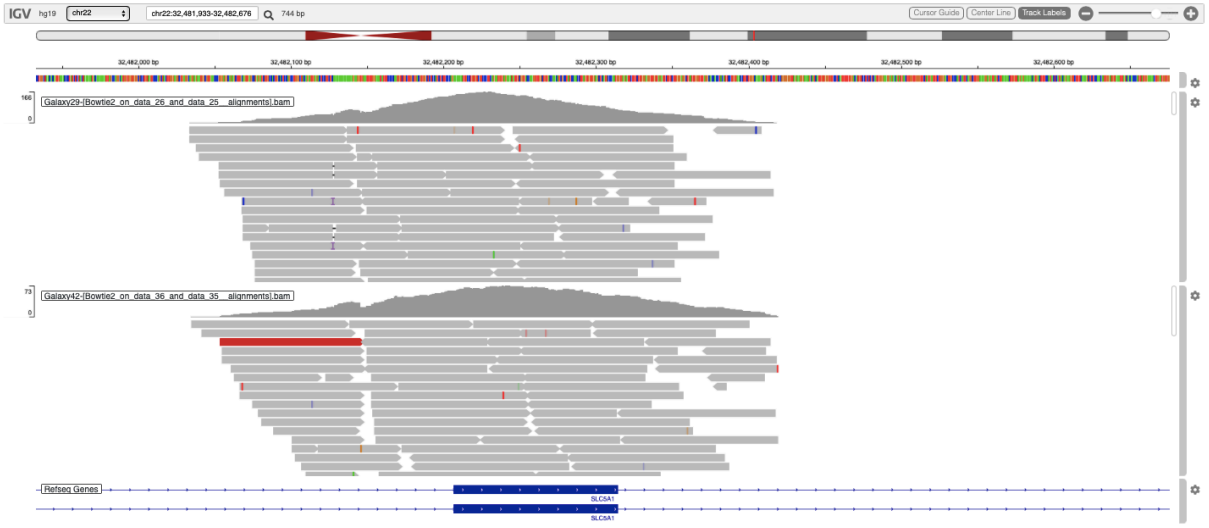


On le fait donc maintenant tourner avec nos fichiers « tumor_R1 » et « tumor_R2 ».
Après cela on peut alors procéder au Varscan. On choisit bien le géome hg19 en tant que référence et nous comparons NORMAL BAM interesect et TUMA BAM intersect. On choisit « Do not perform posterior filtering » car on ne veut pas effectuer de filtrage postérieur sur les données. Le filtrage postérieur consiste à utiliser des algorithmes pour éliminer les données qui sont considérées comme peu fiables en raison de critères tels que la qualité de la séquence, la profondeur de couverture ou les erreurs de séquençage.

On vérifie la sortie de Varscan, avec la présence de SOMATIC sur certains chromosomes.
Le terme "SOMATIC" fait référence à une variation qui n'est présente que dans une seule cellule ou un seul tissu d'un individu, contrairement à une variation qui est présente dans toutes les cellules d'un individu. On peut donc ainsi trouver les variations somatiques.

Nous pouvons en visualiser certaines sur le logiciel IGV.


32,480,540 bp
32,480,550 bp
32,480,560 bp
32,480,570 bp
32,480,580 bp
32,480,590 bp
32,480,600 bp
32,480,610 bp
Galaxy29-[Bowtie2_on_data_26_and_data_25__alignments].bam
Galaxy42-[Bowtie2_on_data_36_and_data_35__alignments].bam


IGV
hg19
chr22
chr22:32,481,933-32,482,676
744 bp
Cursor Guide
Center Line
Track Labels
32,482,000 bp
32,482,100 bp
32,482,200 bp
32,482,300 bp
32,482,400 bp
32,482,500 bp
32,482,600 bp
Galaxy29-[Bowtie2_on_data_26_and_data_25__alignments].bam
Galaxy42-[Bowtie2_on_data_36_and_data_35__alignments].bam
Refseq Genes
SLC5A1
SLC5A1