Sorbonne Université

Année Universitaire 2021-2022

RAPPORT DE STAGE

Vers un portrait moléculaire complet d'une tumeur unique reposant uniquement sur l'ARN

Effectué dans le cadre de l'obtention du diplôme

M2 Bioinformatique & Modélisation (BIM-BMC)

Présenté et soutenu par Nikita LAGRANGE

Préparé au sein de l'I2BC, Equipe Séquence, Structure et Fonction des ARN (SSFA)

sous la direction de Daniel GAUTHERET, Professeur

17 juin 2022

Remerciements

Je tiens à remercier le Professeur Daniel Gautheret pour m'avoir accordé sa confiance depuis le début du stage. Il a été toujours présent et a su répondre à mes diverses interrogations. Ses connaissances à la fois dans la biologie moléculaire du cancer et dans les outils d'analyse des données biologiques m'ont guidé tout au long de ce projet. Je le remercie également pour son invitation à mon premier séminaire scientifique où j'ai pu exposer notre nouvel outil à la communité scientifique des k-mers. Marville a été une très belle expérience !

Je remercie le Docteur Mélina Gallopin pour ses conseils en modélisation statistique. Antoine pour son aide à l'utilisation du cluster et de son outil précieux d'annotation. Hugues pour la génération des k-mers gnomAD. Haoliang pour ses travaux précurseurs sur l'analyse des données de cancer. Claire et Fabrice pour les aspects quotidiens du fonctionnement du laboratoire.

Table des matières

Résumé

Ces dernières années la biologie a vu croître de manière significative les données expérimentale grâce à l'impulsion des projets de séquençage haut-débit (NGS). La plupart de ces données sont analysées en cohorte et une large partie de ces données n'est pas pleinement exploitée. Ce stage s'inscrit dans la thématique de la médecine personnalisée et tire parti de ces données accumulées. Le cancer est une maladie génétique dont de nombreuses données transcriptomiques sont disponibles. Ce stage a étudié l'adénocarcinome du poumon à l'aide de données RNA-Seq. Une nouvelle méthode d'analyse de données RNA-Seq a été développé. Elle se base sur les connaissances accumulées du génome et du transcriptome humain. Cette méthode a pour objectif d'exclure des fragments d'ARN « normaux » qui sont référencés soit dans des annotations, type GENCODE, soit dans des bases de données de variations, type gnomAD, soit dans des libraires RNA-Seq de tissus normaux. Pour cela plusieurs outils ont été utilisé. Ces outils sont basés sur le concept de k-mer, un moyen computationnellement efficace et exhaustif pour analyser de grand fichier de séquences. Ce protocole permet d'exclure jusqu'à 99.85 % des séquences présentes dans un fichier de séquençage du transcriptome d'une tumeur. Les fragments d'ARN restant constituent la brique de base d'un portrait moléculaire d'un patient. En effet une annotation de ces fragments a permis de retrouver des mutations impliquées dans les gènes drivers de l'adénocarcinome du poumon. De plus des fragments d'immunoglobulines, de lncRNA, de séquences intergéniques, des séquences répétées, des virus sont présents au sein de ce pool de fragments d'ARN. Nous avons à disposition tout un paysage transcriptomique qui est propre à une tumeur et à un individu.

Mots-clefs : Médecine personnalisée, RNA-Seq, Cancer, Transcriptomique, k-mer, Index

I – Introduction

1. Le cancer une maladie multigénique

Le cancer fait partie des maladies dites multigénique ou polygénique, c'est-à-dire affectant plusieurs gènes. En effet la cause d'un cancer est la survenu de mutations dans un ensemble de gènes spécifique. Une mutation est définie comme tout modification d'une partie de l'ADN au sein d'une cellule. On distingue deux types de mutations : - les mutations germinales qui affectent les cellules germinales (gamètes) et seront transmises à la descendance et les mutations somatiques qui affectent les autres cellules du corps et ne seront pas transmises à la descendance. La mutation la plus commune au cancer est la mutation somatique qui représente environ 90 % des cas de cancers, ceux sont les cancers dits sporadiques. Les cancers dus aux mutations germinales représentent 10 % des cas et sont appelés les cancers héréditaires (Anand et al., 2008). Ce stage se focalisera uniquement sur les cancers dus aux mutations somatiques.

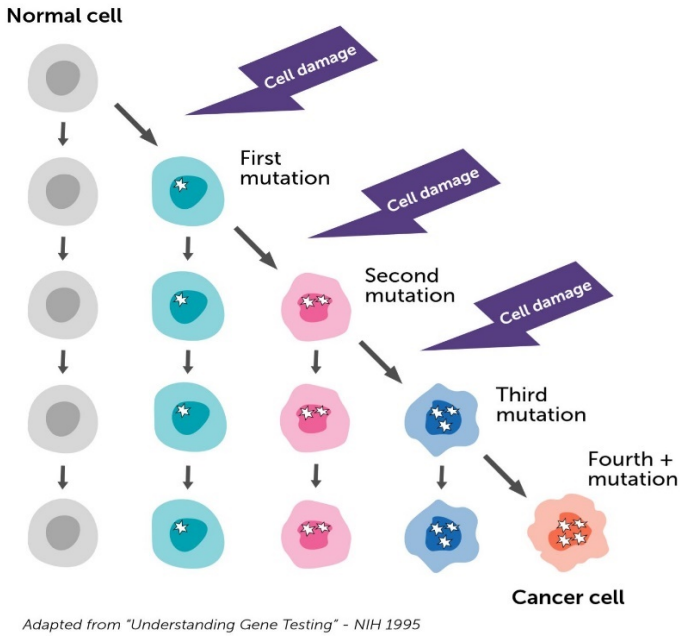

Figure 1 – Schéma de la cascade mutationnel à l'origine d'une cellule cancéreuse sporadique

Les mutations sont un phénomène courant. Plusieurs facteurs favorisent l'apparition de mutations : les facteurs environnementaux (tabagisme, UV, alcool, produits chimiques …), le vieillissement, les virus. La plupart de ces mutations seront réparées par les systèmes de réparation de l'ADN. Les mutations restantes sont généralement neutres, elles n'auront pas d'impact sur le fonctionnement de la cellule. Enfin certaines mutations seront à l'origine de l'apparition du cancer. Il est important de rappeler qu'une mutation ne conduira pas au cancer. Seule une combinaison de mutations dans un ensemble de gènes (**Figure 1**) sera susceptible de

conduire la cellule dans un état cancéreux. Cet ensemble de gènes peut être répartit en trois catégories fonctionnelles :

- Gène suppresseur de tumeurs (p53, BRCA1/2, PTEN …)
- Proto-oncogène (KRAS, NRAS, c-myc …)
- Gène de réparation de l'ADN (ATM, PRKDC, POLE, …)

Les gènes suppresseurs de tumeurs sont généralement impliqués dans la régulation négative de la prolifération cellulaire. Une mutation dans ces gènes, généralement par perte de fonction, va permettre aux cellules de proliférer de manière incontrôlée.
Les proto-oncogènes interviennent dans la régulation de l'embryogenèse ou de la croissance cellulaire. Une mutation dans ces gènes aura comme conséquence l'expression d'une forme anormale de ces gènes qui induira l'état cancéreux d'une cellule.
Enfin une mutation dans des gènes de réparation de l'ADN ne permettra plus d'assurer l'intégrité du génome au sein d'une cellule.
Cet ensemble de gènes est couramment appelé les gènes *drivers*, car une mutation dans ces gènes va augmenter la prolifération cellulaire, caractéristique d'une cellule cancéreuse.

2. La transcriptomique : moyen d'étude de la diversité génétique

Le transcriptome correspond à l'ensemble des transcrits présent dans une cellule. Une technique d'étude du transcriptome est le RNA-Seq, le séquençage haut-débit des transcrits. Le résultat

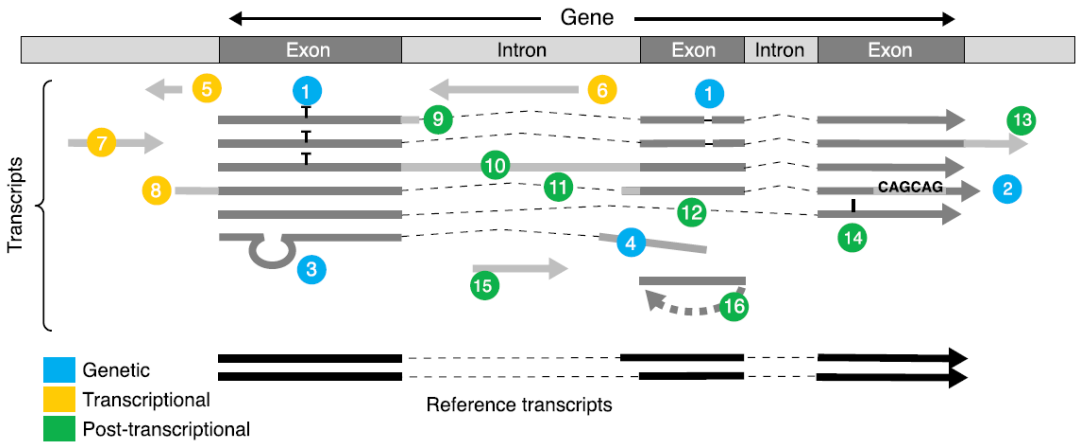

Figure 2 – Schéma de l'univers des transcrits capturés par RNA-Seq (Morillon and Gautheret, 2019)

de ce séquençage est un fichier FASTQ. Une séquence nucléique de ce fichier est appelée lecture (*read*). Ainsi cette technologie permet de capturer à la fois la séquence de ces transcrits et leurs abondances au sein d'une cellule. La **figure 2** montre les différents types de transcrits qui peuvent être séquencés par RNA-Seq. Tout d'abord les mutations somatiques telles que les polymorphismes d'un seul nucléotide (SNV), les indels, les séquences répétées, les transposons,

les gènes de fusions. Les altérations au niveau de la transcription : transcrit antisens, site d'initiation de la transcription alternatif. Et les altérations post-transcriptionnelles : épissage alternatif (rétention d'intron, saut d'exon …), ARN circulaire. Tout une diversité de séquences est présente au niveau du transcriptome. On estime qu'il y a 4 fois plus de transcrits annotés (Morillon and Gautheret, 2019) que de gènes annotés. Ce qui montre que la technique de RNA-Seq permet de capturer une plus grande diversité de séquence que la technologie classique de séquençage de génome ou de mesure d'expression par puce à ADN. De plus le transcriptome est le premier phénotype observable d'une cellule car il révèle les activités de régulation et donc le fonctionnement de la cellule. Ainsi cette technologie est un moyen d'étude particulièrement adaptée pour l'étude du cancer.

3. Analyse du transcriptome : l'approche par k-mers

Un k-mer est une sous-chaîne de caractère de taille k. Ce concept a été appliqué à l'analyse de séquences biologiques. En effet le k-mer présente plusieurs avantages. Il permet d'analyser les variations locales d'une séquence à la résolution du nucléotide, c'est une méthode sans référence (pas d'a priori sur les données à analyser), et c'est un concept computationnellement efficace. Dernièrement plusieurs outils ont été développé sur l'utilisation de k-mers pour analyser le transcriptome. DE-kupl est l'un de ces outils (Audoux et al., 2017). Il permet de capturer l'ensemble des variations locales des transcrits présente au sein d'une collection de librairies RNA-Seq. Pour cela DE-kupl extrait tous les k-mers qui sont différentiellement exprimés entre deux conditions puis assemble ces k-mers en des contigs qui seront annotés. Une autre manière d'appliquer les k-mers aux données NGS est de construire une table de comptage où chaque k-mer d'une librairie est associée à son compte. Jellyfish est un outil (Marçais and Kingsford, 2011) qui construit rapidement cette table de comptage. Il peut également effectuer des requêtes de k-mer contre la table de comptage.
De plus le k-mer est à la base du graphe de De Bruijn (**Figure 3**). Il s'agit d'un graphe dirigé qui représente un ensemble de séquences. Cet ensemble de séquences est décomposé en un ensemble de k-mers. Les nœuds de ces graphes sont l'ensemble des k-mers et les arrêtes correspondent au chevauchement de longueur k-1 entre les k-mers. Ainsi ce graphe est utilisé pour l'assemblage de novo de génome ou de transcriptome (Bankevich et al., 2012) ainsi que de la détection de variant dans les données NGS (Sacomoto et al., 2012).

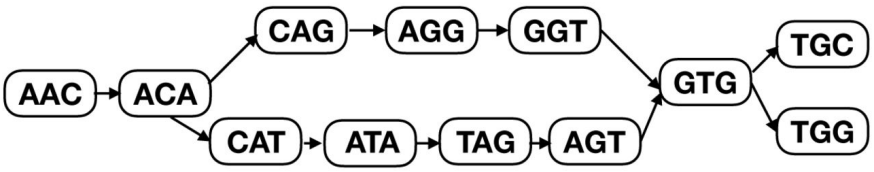

Figure 3 – Exemple de graphe de De Bruijn, avec taille du k-mer k = 3. Deux séquences alternatives sont possibles : 1 – AACAGGTGC (chemin du haut), 2 – AACATAGTGG (chemin du bas) **(**CC BY-SA 4.0)

4. Vers une médecine personnalisée

La médecine personnalisée est un domaine émergent qui a pour objectif de prendre en compte plus l'individu que la maladie dans le processus de soin. En effet ces dernières années, nous avons vu que chaque individu possède sa particularité au niveau moléculaire. Ainsi certains patients peuvent répondre favorablement à un traitement alors que d'autres non. Les traitements actuels sont vus pour traiter une maladie et ne sont pas conçus spécifiquement pour un individu. Un traitement personnalisé ciblerait les cellules aberrantes, au lieu de toutes les cellules en division comme dans un traitement systémique. Le bénéfice principal est la réduction des effets secondaires par rapport à un traitement traditionnel.
La médecine personnalisée tire parti du progrès des séquenceurs haut-débit pour personnaliser le traitement.

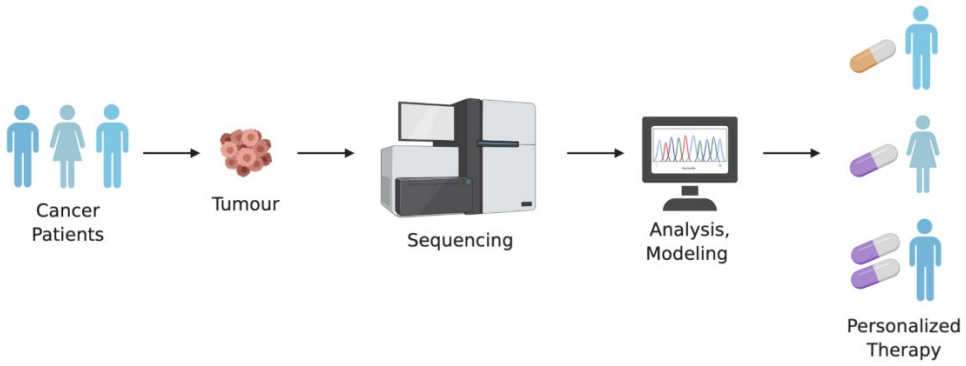

Figure 4 – Schéma de la philosophie de la médecine personnalisée **(**CC BY-SA 4.0)

En effet la **figure 4** illustre ce concept. Nous souhaitons à partir d'un échantillon d'une tumeur d'un individu effectuer un séquençage RNA-Seq puis analyser ces données à l'aide de protocole de bio-informatique afin d'extraire de l'information pertinente. Cette information peut-être les séquences des gènes drivers mutés, l'environnement immunitaire de la tumeur ou toutes mutations susceptibles d'être à l'origine de la tumeur. A partir de cette analyse, des médicaments peuvent être développé spécifiquement pour un individu.

Des approches de diagnostic et de pronostic par RNA-Seq ont développés dans l'étude des cancers et des maladies héréditaires (Frost et al., 2020; Hong et al., 2020; Kremer et al., 2017). Toutefois elles utilisent des approches conventionnelles restreintes au transcrits annotés. Potentiellement ces approches excluent de leurs analyses les transcrits qui ne sont pas annotés, ce qui peut être le cas pour des transcrits issus d'événement mutationnel. A notre connaissance un seul groupe (Cmero et al., 2021) tente d'analyser le transcriptome d'un patient par une méthode sans référence. Mais ils ne traitent pas l'aspect mutationnel. Nous proposons dans ce projet une approche réellement exhaustive avec l'objectif d'identifier toutes les anomalies d'un échantillon.

5. Problématique du stage

La question biologique qui a guidé ce stage est la suivante : comment distinguer un transcrit « normal » d'un transcrit spécifique d'une tumeur ? En effet une cellule tumorale exprime à la fois des gènes qui sont impliqués dans le fonctionnement de la cellule et des gènes qui vont être propre à la cellule tumorale. Parmi les gènes dits « normaux » des mutations peuvent être présentes sans impacter le fonctionnement de la cellule. Il s'agit du polymorphisme qui est à la fois propre à un individu et présent au sein d'une population. Cet ensemble de gènes « normaux » brouille le signal du transcriptome spécifique d'une tumeur. C'est pourquoi il important d'éliminer cet ensemble afin d'analyser plus finement l'information moléculaire d'une cellule cancéreuse. De plus nous souhaitons développer un protocole qui permet d'analyser individuellement un fichier de séquençage dans une optique de médecine personnalisée. Notre objectif est de proposer un protocole d'analyse d'un fichier RNA-Seq d'une tumeur qui va exclure l'ensemble des transcrits « normaux » à l'aide de méthodes basées sur les k-mers. Ainsi le résultat souhaité de ce protocole est un ensemble de transcris qui correspondra le plus fidèlement possible aux transcrits propres d'une tumeur.

II – Matériel et méthodes

1. Protocole d'exclusion des fragments d'ARN « normaux »

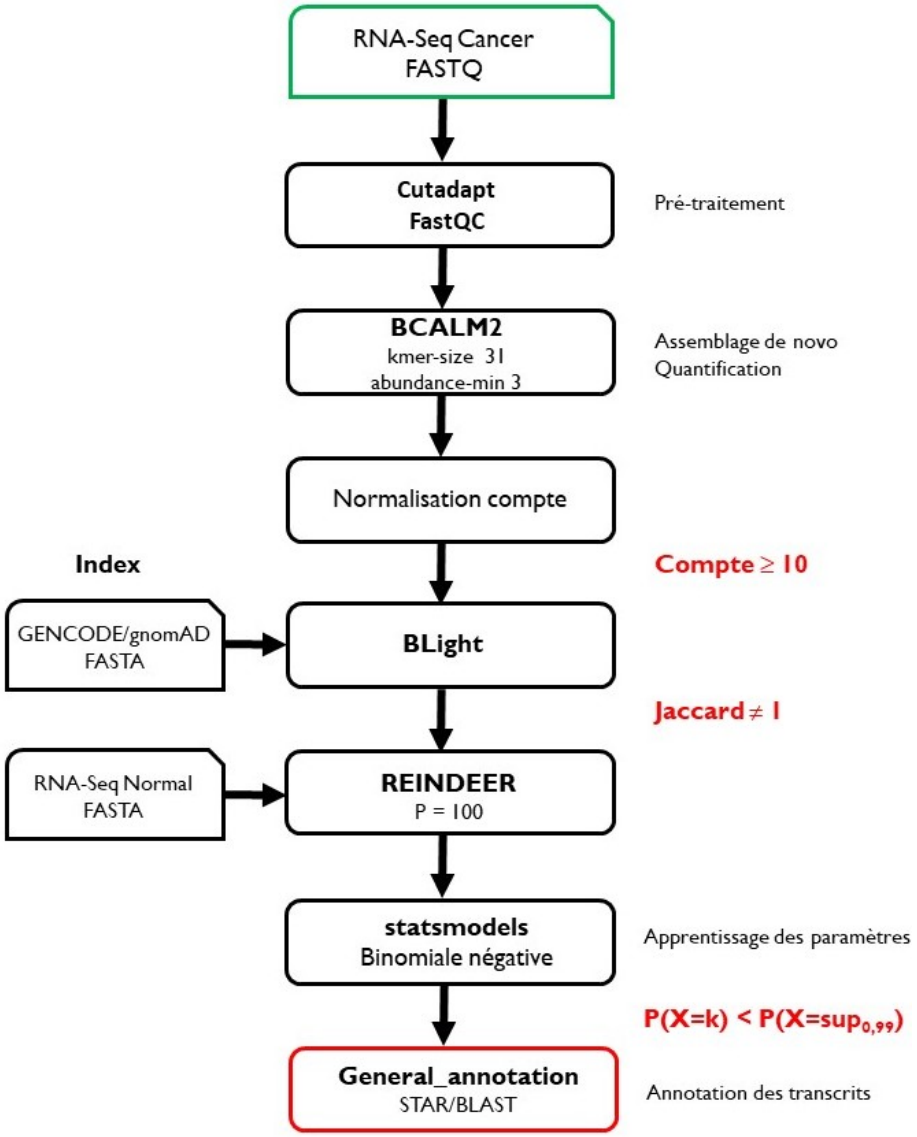

Figure 1 – Pipeline d'exclusion des fragments d'ARN « normaux »

Notre protocole (**Figure 1**) prend en entrée un fichier FASTQ issu d'un RNA-Seq d'un échantillon tumoral. Le protocole d'exclusion des fragments d'ARN « normaux » se décompose en trois grandes étapes :

1- Pré-traitement du FASTQ
2- Assemblage de novo et quantification des fragments d'ARN « normaux »
3- Exclusion des fragments d'ARN « normaux » (filtre : texte rouge)

A la fin du protocole, plusieurs analyses sont possibles. Nous avons décidé d'annoter les fragments d'ARN restants.

2.. Jeu de données

Les données analysées ont pour origine la publication de (Seo et al., 2012). Ces données sont issues de séquençage haut-débit Illumina HiSeq 2000 de type RNA-Seq de tissus pulmonaires sains et cancéreux au sein d'une population coréenne. Les librairies ont été construites à l'aide d'un séquençage en paire (« paired end ») et non-orienté (« unstranded ») sur des fragments à séquencer de taille 100 nt. Le cancer étudié est l'adénocarcinome du poumon, l'un des types de cancer du poumon le plus commun. Nous avons à disposition 77 échantillons normaux et 77 échantillons cancéreux appariés. C'est-à-dire que pour chaque individu nous avons un échantillon du tissu tumoral et un échantillon du tissu normal adjacent. Les fichiers FASTQ du séquençage sont en libre accès sous l'identifiant ERP001058 au sein du Sequence Read Archive (SRA). Nous avons également 59 échantillons normaux de poumons issus du projet TCGA (Cancer Genome Atlas Research Network et al., 2013).

3. Pré-traitement des fichiers RNA-Seq

Les séquences des adaptateurs utilisées lors de la préparation de la librairie sont présentes dans les fichiers FASTQ. Il est donc important de les éliminer. L'outil Cutadapt (Martin, 2011) est dédié à cette tâche. Les séquences des adaptateurs sont données lors de la préparation des librairies par Illumina.
Afin de calculer le nombre total de lectures et la taille moyenne d'une lecture dans une librairie l'outil FastQC est utilisé. Un fichier html est fournie en sortie de FastQC. MultiQC (Ewels et al., 2016) permet d'agréger un ensemble de fichiers html de FastQC en un unique fichier html, ce qui facilite l'analyse des résultats.

4. Assemblage de novo de transcriptome

Une fois les fichiers RNA-Seq pré-traités, il faut assembler les lectures et les quantifier. Plusieurs méthodes ont été développées spécifiquement soit pour l'assemblage soit pour la quantification des séquences soit pour la combinaison de ces deux objectifs. La plupart des outils développés repose sur le principe d'un alignement sur un transcriptome de référence. Notre choix s'est porté sur un outil qui n'utilise pas une référence et qui peut à la fois quantifier l'abondance d'une séquence et assembler le fichier RNA-Seq. BCALM2 (Bruijn CompAction in Low Memory) répond à cette exigence. BCALM2 (Chikhi et al., 2016) construit un graphe de De Bruijn compacté à partir de données de séquençage. Contrairement à un graphe de De

Bruijn, sa forme compactée présente des nœuds de différentes longueurs (≥ k) et les nœuds sont des unitigs (concaténation de k-mers chevauchant, correspondant à des fragments d'ARN (assemblage local)). L'algorithme BCALM2 est spécialement optimisé pour la construction de ce graphe. Il peut prendre en entrée soit un fichier FASTQ soit un fichier FASTA et en sortie donne l'ensemble des unitigs du graphe de de Bruijn compacté sous format FASTA (**Figure 2**). Chaque ligne (unitig) correspond à un nœud du graphe et l'en-tête donne la quantification de la séquence ainsi que sa longueur.

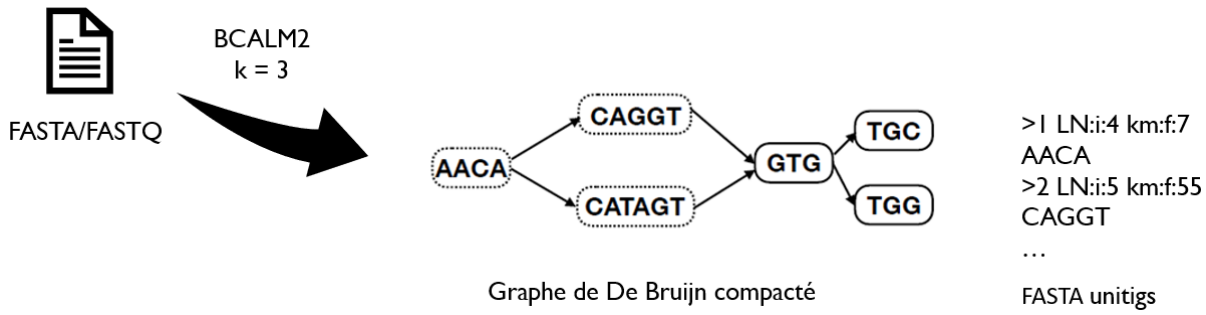

Figure 2 – Schéma du fonctionnement de BCALM2, avec en paramètre une taille de k-mer k = 3. Le champ LN:i correspond à la longueur d'un unitig et le champ km:f à l'abondance de l'unitig

Nous pouvons contrôler la taille du k-mer pour la construction du graphe. Il a été choisi une taille de k = 31, une valeur standard en bio-informatique. De par son compromis entre spécificité du k-mer avec le génome humain et sa complexité computationnelle. De plus un seuil de compte permet de filtrer l'abondance d'un k-mer au sein du jeu de données d'entrée. Afin d'éliminer potentiellement les k-mers présentant des erreurs de séquençage ou des k-mers qui ne sont pas assez représentatif de l'ensemble de k-mers. Nous avons appliqué un seuil de 3. C'est-à-dire que seul les k-mers présents au moins 3 fois dans le fichier d'entrée FASTA ou FASTQ sont conservés.

Par la suite les dénominations fragment d'ARN, transcrits et unitigs sont équivalentes.

5. Normalisation des comptes

Une étape essentielle dans l'analyse de données transcriptomiques est la normalisation de l'abondance des transcrits. En effet divers paramètres peuvent influencer la quantification des transcrits. Notamment la profondeur de séquençage. La profondeur ou la couverture est définie comme le nombre de fois qu'une séquence va être séquencer par le séquenceur. Par exemple une couverture de 30 signifie qu'on s'attend pour une séquence donnée que le séquenceur l'ait séquencé au moins 30 fois. Donc plus la profondeur sera élevée, plus l'abondance d'une séquence sera élevée. Cependant ce paramètre est difficilement contrôlable. La profondeur peut varier entre les différentes librairies et cela conduit à un biais dans l'expression d'une même

séquence mais présente dans deux librairies ayant des profondeurs différentes. C'est pourquoi la normalisation est importante pour supprimer ce biais. Un moyen indirect de normaliser est de compter le nombre de lectures au sein d'une librairie et de compter la longueur moyenne d'une lecture. Il a été défini le facteur de normalisation suivant :

$$N = \frac{10^9}{(L_{R1} - k + 1) \times Nb_{R1} + (L_{R2} - k + 1) \times Nb_{R2}}$$

Avec $L_{R1/2}$: la longueur moyenne d'une lecture au sein du FASTQ 1/2 (« paired end »)

Avec $Nb_{R1/2}$: le nombre total de lecture au sein du FASTQ 1/2 (« paired end »)

Avec k : la taille du k-mer

Ainsi une librairie présentant une couverture élevée aura un nombre total de séquence plus important et donc un facteur de normalisation faible qui aura pour effet de diminuer l'abondance estimée par BCALM2 et réciproquement.
Le nombre total et la longueur moyenne d'une lecture ont été calculés lors de l'étape de pré-traitement.
L'abondance d'un fragment d'ARN estimée par BCALM2 a été multipliée par le facteur de normalisation associé à sa librairie.

6. Premier filtre : seuil de compte

Afin de réduire le nombre de séquences à analyser, un premier filtre a été mis en place. Il s'agit d'un filtre de compte. Nous avons décidé d'appliquer un seuil de 10. Seules les séquences dont les comptes sont supérieurs ou égaux à 10 sont conservées. Cela permet d'exclure les séquences issues d'erreurs de séquençages qui ont pu passer l'étape de pré-traitement des données ainsi que des erreurs d'assemblage. En effet dans ces deux cas ces deux événements sont relativement rares, donc on s'attend que les séquences présentant ces erreurs aient des comptes faibles. Par ailleurs on souhaite conserver à la fin de notre protocole les séquences les plus significatives chez un patient. Ce filtre permet de s'affranchir de séquences faiblement exprimées, qui peuvent être assimiler à du bruit de fond.

7. Construction d'une banque de transcrits référencés et de transcrits polymorphiques

Le transcriptome de référence d'*Homo sapiens* a été obtenu sur GENCODE 39 (Frankish et al., 2021) sous format FASTA. Les séquences polymorphiques au sein de la population ont été obtenues à l'aide du consortium gnomAD v2.1.1 (Genome Aggregation Database) (Genome Aggregation Database Production Team et al., 2020) à partir de 125 748 données d'exomes. A

cela est ajouté les variants de l'ADN mitochondrial issu de la version 3.1 de gnomAD. gnomAD à partir de ces données d'exomes effectue une identification de variant en excluant dans son analyse les variations propres au cerveau, au cancer et aux pathologies. Seules les variations de type polymorphique et potentiellement non impliquées dans des maladies sont conservées. gnomAD applique également un filtre de qualité sur ces variants. Nous avons gardé uniquement les variations ayant passé ce filtre et dont les fréquences alléliques sont supérieures à 1 %. Ce seuil est couramment utilisé afin de distinguer les variations susceptibles d'être pathologiques, fréquence allélique inférieure à 1 %, des variations polymorphiques, fréquence allélique supérieure à 1 %.

Ces variations sont au format VCF (Variant Call Format). Ce format nous donne la localisation de la variation au sein du génome (chromosome et position), la séquence locale de référence et la séquence locale allélique. Cependant nous souhaitons convertir le format VCF en un format FASTA. Pour cela nous avons utilisé l'outil SeqTailor (Zhang et al., 2019) spécialement dédié à cette conversion. Cet outil prend en entrée un fichier VCF et en sortie nous obtenons un fichier FASTA qui contient à la fois la séquence de référence et la séquence alternative. Un paramètre de cet outil est la fenêtre dans laquelle nous souhaitons intégrer la variation. Pour cela nous avons choisi une fenêtre de 61 nt en amont de la variation et 61 nt en aval de la variation.

Par la suite nous avons concaténé les deux fichiers FASTA : GENCODE et gnomAD. Ce fichier FASTA constitue la brique de base de notre première banque de référence.

Afin de faire des requêtes rapides de présence de séquence contre cette banque, nous avons utilisé l'outil BLight (Marchet et al., 2021). BLight est une structure de données associative similaire à une table de hachage (i.e. dictionnaire). Elle permet d'indexer un ensemble de k-mers. Chaque k-mer dans l'index est associé à un identifiant unique. Ainsi une séquence requête est décomposé en un ensemble de k-mers et chaque k-mer de cet ensemble est interrogé individuellement contre l'index constitué de notre référence GENCODE plus gnomAD. Si le k-mer est présent dans l'index alors est retourné l'identifiant du k-mer dans l'index sinon l'absence de k-mer est notifié par un -1 (**Figure 3**). BLight permet de faire un million de requêtes en une seconde avec un seul CPU.

Deux étapes sont nécessaires pour le fonctionnement de BLight. La première étape est la construction de l'index. BLight est initialement conçu pour prendre en entrée un fichier FASTA où un ensemble de k-mers est présent de manière unique. Pour cela nous utilisons BCALM2 avec pour entrer le fichier FASTA concaténé GENCODE/gnomAD avec un seuil de compte minimal de 1 et une taille de k-mer de 31. Ainsi l'ensemble de nos séquences sont représentées au sein du graphe. L'index est construit par la suite avec une taille de k-mer 31.

La deuxième étape consiste à la requête de séquence contre l'index.

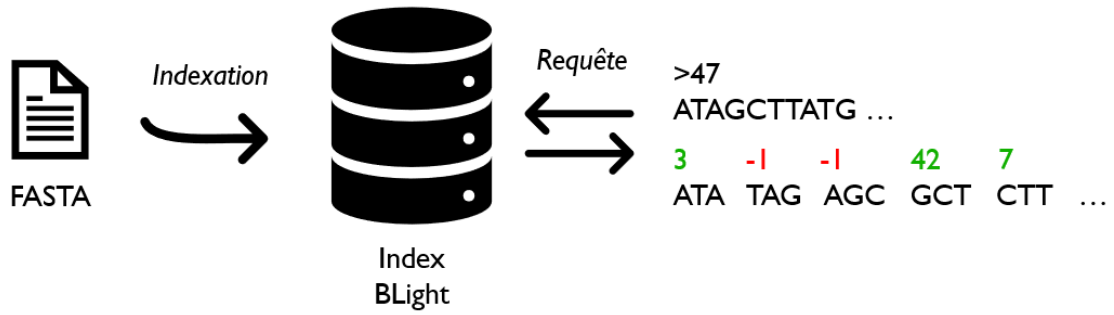

Figure 3 – Schéma du fonctionnement de BLight. La sortie de BLight est un vecteur d'entiers d'une taille correspondant au nombre de k-mers présent au sein de la séquence requête. -1 (rouge) le k-mer est absent de l'index, sinon (vert) le k-mer est présent dans l'index. L'index est construit avec k = 3.

8. Analyse des résultats de requêtes : index de Jaccard

La sortie d'une requête correspond à un vecteur de la taille du nombre de k-mers présent au sein d'une requête. Une séquence de taille L et une taille k de k-mer a $L - k + 1$ k-mers. Par exemple une séquence de taille 100 nt et une taille de k-mer 31 aura $100 - 31 + 1 = 70$ k-mers. Donc une requête de cette séquence aura pour sortie une liste de longueur 70, où si le k-mer est présent dans la banque de référence la valeur sera différente de -1 sinon la valeur est de -1. A partir de ce vecteur, il est possible de déduire une valeur représentant l'appartenance de cette séquence à cette banque. Il s'agit de l'index de Jaccard.

Sa valeur est comprise entre 0 et 1. Si la valeur est de 0 la séquence est absente de la banque, si la valeur est de 1 la séquence est présente dans la banque. Entre 0 et 1 exclue, une partie seulement de la séquence est présente dans la banque.

De manière plus formelle, soit A l'ensemble de k-mers de la séquence requête et soit B l'ensemble de k-mers de l'index de référence l'index de Jaccard J se définit comme :

$$J(A,B) = \frac{|A \cap B|}{|A|}$$

$$J(A,B) \in [0,1]$$

C'est le rapport entre le nombre de k-mers de l'intersection de l'ensemble A et B sur le nombre de k-mers de l'ensemble A. En effet si l'ensemble des k-mers de A est contenu dans B alors ce rapport est de 1. Nous pouvons voir cet index de Jaccard comme un alignement de séquence et cette valeur est un témoin indirect de la similarité entre la séquence de requête et la banque de référence.

Notre objectif est d'étudier uniquement les séquences non référencées. Donc seul les séquences qui ont un index de Jaccard différent de 1 seront conservées par la suite.

9. Construction d'une banque de librairies RNA-Seq

REINDEER (REad Index for abuNDancE quERy) (Marchet et al., 2020) est un outil qui indexe des k-mers et leurs abondances au sein de collections de jeux de données (e.g. RNA-seq ou métagénomique). Lors d'une requête contre cet index, en sortie est donnée l'abondance de cette requête au sein de chacune des librairies indexées.

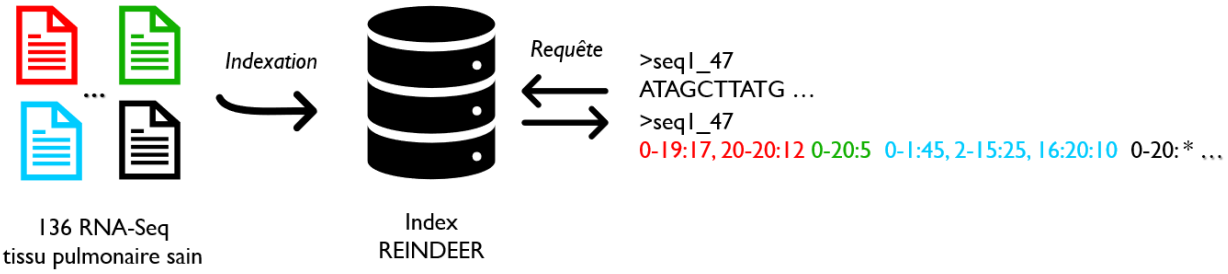

Figure 4 - Schéma du principe de fonctionnement de REINDEER. Séquence seq1 de taille 51 nt (k =31, 21 k-mers) et de compte 47. Chaque colonne de la sortie de la requête correspond à une abondance d'une plage de k-mers au sein d'une librairie indexée. 0-19 :17 les k-mers 0 à 19 ont une abondance de 17 dans la première librairie indexée, 0-20 : * les k-mer 0 à 20 (la requête) sont absents de la quatrième librairie indexée.

Comme pour BLight, REINDEER est défini en deux étapes : l'indexation et la requête (**Figure 4**). Lors de l'indexation, l'algorithme prend en entrée des graphes de De Bruijn compacts de chacune des librairies à indexer.

On utilise BCALM2 pour construire ces graphes. Nous souhaitons construire un index de référence de librairies provenant d'échantillons normaux. Nous avons à disposition comme rappelé au paragraphe 1 des librairies de tissus normaux de poumons qui sont au nombre de 136. Par la suite nous avons généré les graphes de De Bruijn compactés sur ces 136 fichiers FASTQ avec en paramètre une taille de k-mer de 31 et une abondance minimale de 3.

L'index REINDEER a été construit sur ces 136 graphes de De Bruijn avec une taille de k-mer de 31.

L'étape de requête est effectuée avec en entrée un fichier FASTA. La sortie de la requête correspond à une matrice : la première colonne est pour l'en-tête des séquences du fichiers FASTA et les 136 colonnes suivantes sont pour les résultats de la requête des 136 graphes de De Bruijn.

Une requête du fichier FASTA est décomposée en un ensemble de k-mers de taille 31 et cet ensemble est interrogé contre l'index REINDEER. Un seuil de chevauchement de l'ensemble des k-mers de la requête sur le graphe de De Bruijn de l'index de REINDEER est un paramètre

contrôlable. Nous avons décidé d'utiliser un pourcentage de 100 % de chevauchement de l'ensemble de k-mers sur le graphe de De Bruijn assemblée de l'index REINDEER. Ainsi si l'ensemble de k-mers d'une requête n'est pas présent dans une librairie (colonne) alors en sortie nous avons le symbole '*'. Nous avons considéré que cette absence est similaire à une abondance nulle (0). Si l'ensemble des k-mers est présent dans une librairie alors en sortie nous avons une plage de k-mers. Afin d'avoir une valeur unique, nous avons fait la moyenne pondérée de cette plage d'abondance. Soit C_i le compte associé à une requête dans une librairie (colonne) spécifique i, on a :

$$C_i = \frac{\sum N_{k-mers(i)} \times A_{k-mers(i)}}{\sum N_{k-mers(i)}}$$

Avec $N_{k-mers(i)}$: le nombre de k-mers au sein d'une plage

$A_{k-mers(i)}$: l'abondance de la plage de k-mers

L'abondance de la plage de k-mers est normalisée avec le facteur de normalisation associée à la librairie i selon le protocole du paragraphe 2.

Ainsi pour chaque requête donnée nous avons un vecteur de 136 valeurs associées à l'abondance de cette requête au sein des 136 librairies construites sur des tissus pulmonaires normaux.

Pour l'analyse des échantillons normaux, le même processus est appliqué à l'exception que la colonne correspondant à l'échantillon est exclue de la distribution. Le vecteur est donc de taille 135 pour ces échantillons.

5. Modélisation de la distribution de l'abondance

Le vecteur de 136 valeurs d'abondance associé à une requête peut être vu comme une distribution d'abondance de la requête au sein des 136 librairies. Ainsi il est possible de modéliser cette distribution à l'aide d'une loi statistique. De manière formelle un compte associé à une séquence est une valeur discrète positive. La loi binomiale négative est une loi couramment utilisée pour modéliser la distribution de données de compte dispersé. Ce qui est généralement le cas pour les données de type RNA-Seq.

On note X une variable aléatoire modélisant la distribution de compte :

$$X \sim NB(n, p),$$

$$P(X = k) = \binom{k + n - 1}{k} p^n q^k \forall k = 0,1, \ldots$$

- n est le nombre d'échecs à observer
- p est la probabilité de succès de chaque expérience de Bernoulli

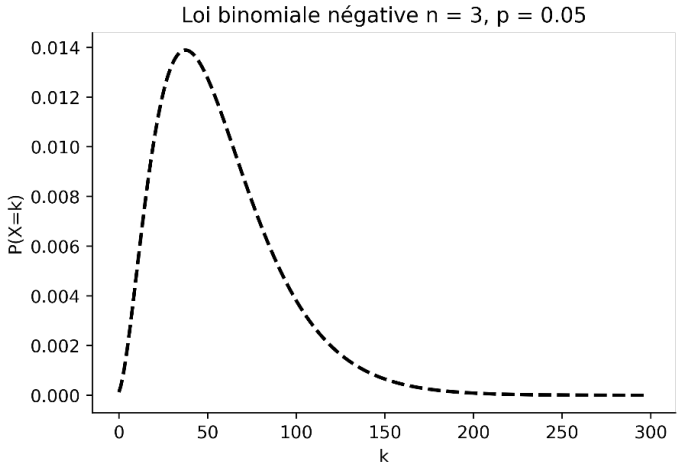

Figure 5 – Exemple d'une distribution d'une loi binomiale négative. Pour une meilleure visualisation nous avons tracé une ligne pointillée reliant les points

Dans le cadre de données RNA-Seq conventionnelles (table d'expression de gènes) un succès correspond à l'observation d'une lecture issue d'un gène donné et un échec correspond à l'observation d'une lecture non issue du gène donné.

Ainsi cette loi est construite à partir de deux paramètres n et p. Ces deux paramètres peuvent être estimer à partir de l'échantillon de 136 comptes associées à chaque requête. On utilise le concept de maximum de vraisemblance pour estimer ces paramètres. Le package Python statsmodels implémente le maximum de vraisemblance pour l'apprentissage des paramètres d'un modèle binomial négatif.

Pour chaque requête, les paramètres p *et* n sont appris sur la distribution de compte avec les valeurs par défaut de la fonction *fit* de statsmodels.

6. *Filtre d'appartenance à une distribution binomiale négative*

En application de la section précédente, un troisième filtre de transcrits est la probabilité d'appartenance d'un compte de transcrit à la distribution de compte associée de la requête REINDEER. En effet l'apprentissage des paramètres p et n permet de calculer cette probabilité. Soit X la variable aléatoire modélisant la distribution de compte associée à la requête et qui suit une loi binomiale négative, nous pouvons calculer la probabilité $P(X = k)$ avec k le compte associé à la séquence requêtée.

Ainsi si un compte n'est pas associé à cette distribution, la probabilité sera proche de 0. Si un compte est associé à cette distribution, la probabilité sera différente de 0. Les valeurs de

probabilités étant souvent faibles, il a été décidé de passer au logarithme népérien afin d'éviter les problèmes d'approximation.

Ce filtre présente deux facettes. Il prend en compte à la fois la présence d'un transcrit dans la distribution des comptes des 136 librairies normale et son abondance par rapport à cette distribution. Il a été décidé de choisir un seuil qui correspond à la borne supérieure qui contient 99 % de la distribution. Seul les séquences présentant une valeur strictement inférieure à ce seuil et qui sont surexprimées par rapport la distribution sont conservées par la suite.

7. Annotation des transcrits filtrés

Les transcrits ayant passé la succession des 3 filtres : 1 – Filtre de compte, 2- Filtre Index de Jaccard, 3- Filtre de probabilité d'appartenance à la distribution de la binomiale négative sont par la suite annotés. Par annotation, nous entendons de connaître de quel gène est issu le transcrit, sa position, sa particularité notamment au niveau du type de mutation (indels, SNP, épissage alternatif, transcrit chimérique). Un outil a été développé spécialement pour cet objectif d'annotation au sein du laboratoire. General Annotation prend en entrée un fichier tsv où une colonne correspond à la séquence du transcrit et une colonne à l'en-tête du transcrit. En retour est donné un fichier tsv qui résume l'annotation de l'ensemble des transcrits donnés.

Cet outil se base sur STAR (Dobin et al., 2013) qui est un aligneur particulièrement efficace en temps de calcul. En effet l'annotation de séquence est faite en alignant la séquence sur un génome de référence. Dans notre cas le génome de référence est celui d'*Homo sapiens* hg38 Ensembl (Howe et al., 2021).

De plus cet outil peut fusionner des sorties de BLAST (Altschul et al., 1990). Nous avons effectué un BLAST sur des banques de transcrits de bactéries et d'archées, de virus, de séquences répétées et d'immunoglobulines.

La banque de transcrits de bactéries et d'archées est issue de la version 18 du jeu de données RDP (Cole et al., 2014) qui contient la séquence 16S rRNA de 20 712 bactéries et 601 archées. La banque de virus est issue du jeu de données VIRTUS (Yasumizu et al., 2021) qui contient 762 séquences de virus humains. La banque de séquences répétées est issue de la base de données DFAM (Hubley et al., 2016) qui contient 2 157 séquences. La banque d'immunoglobulines provient de la base de données IMGT (Lefranc et al., 2015) qui contient 2 622 séquences.

III – Résultats et discussion

1. Chiffre sur les données

	Nombre	Taille FASTQ	Total reads	Taille moyenne d'un read
Cancer	77	661 GB	6.4 G	93 nt
Normal	136	710 GB	11 G	74 nt

Table 1 – Colonnes : nombre d'échantillons RNA-Seq, taille de l'ensemble des échantillons, nombre total de reads de l'ensemble des échantillons, taille moyenne d'un read dans un échantillon

L'ensemble des données RNA-Seq (patients et contrôles normaux) à traiter correspond à environ 1.4 TB en termes de stockage. Ces données contiennent 17 milliards de reads. La taille moyenne d'un read est d'environ 90 nt pour les échantillons cancers et 74 nt pour les échantillons normaux. Cette moyenne est légèrement inférieure du fait des échantillons normaux TCGA qui ne sont pas préparés de la même manière que les échantillons Seo.

2. Normalisation

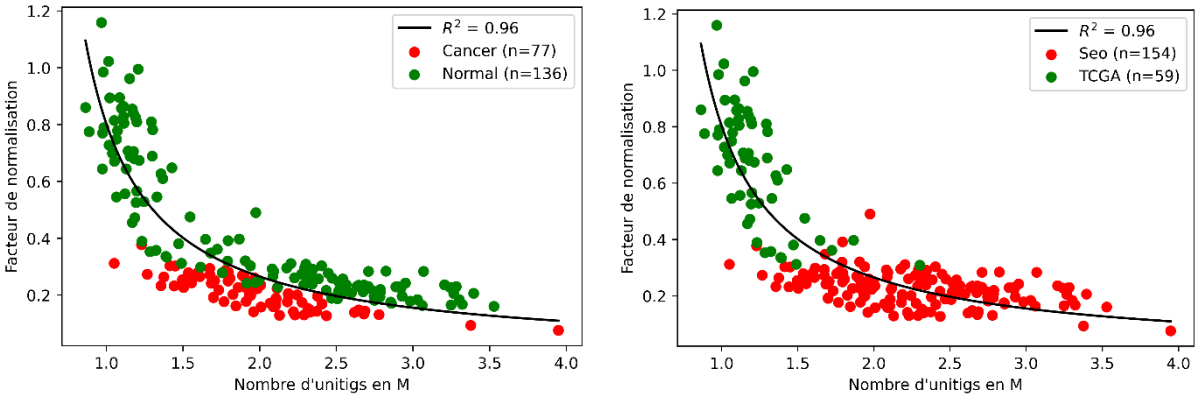

Figure 1 – Nuage de points du nombre d'unitigs issus de BCALM2 en fonction du facteur de normalisation calculé sur les fichiers FASTQ. Ce nuage de point est modélisé par une fonction inverse de type $\frac{a}{x+b}$ où a et b sont des constantes réelles

Le nombre d'unitigs issus de BCALM2, avec en entrée l'ensemble des données RNA-Seq disponibles, varie de 865 000 à 4 millions. La médiane est de 2 millions d'unitigs. En observant le facteur de normalisation, qui est dépendant du nombre de reads au sein d'un fichier FASTQ, et le nombre d'unitigs, on remarque une relation non linéaire de type inverse. En effet plus le nombre de reads est élevé plus le facteur de normalisation est faible et donc plus le nombre d'untigs sera élevée. Et inversement, moins il y a de reads plus le facteur de normalisation est proche de 1 et donc moins il y a d'unitigs.

Pour étudier plus finement cette relation nous avons identifié les points qui appartiennent soit à des échantillons cancers ou normaux, soit à des échantillons Seo ou TCGA (**Figure 1**). On remarque que la dispersion des valeurs n'a pas pour origine les conditions biologiques. Un échantillon normal et cancer peut avoir le même facteur de normalisation. Mais que cette dispersion est due à la technique de séquençage. En effet on observe une limite claire entre les échantillons Seo et TCGA. Les échantillons Seo ont un nombre de reads et une longueur moyenne plus élevés que les échantillons TCGA. Cela montre l'importance de la normalisation des comptes afin de s'affranchir de ce biais expérimental.

3. Filtre compte

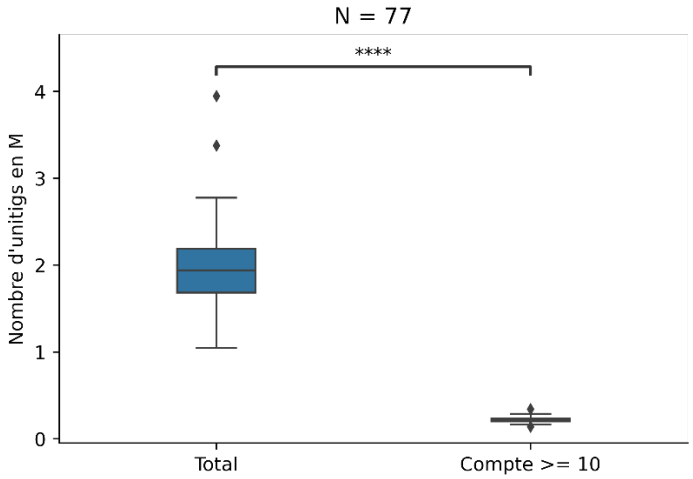

Figure 2 – Boxplot du nombre d'unitigs dans l'ensemble (N=77) des échantillons cancers au départ et après application du filtre de compte. La différence est testée par un test de Student apparié, **** : p <= 1e-04

Le nombre médian d'unitigs BCALM2 sur les échantillons cancers est de 2 millions. Après application du filtre de compte, seul les séquences dont l'abondance est supérieure ou égale à 10 sont conservées, le nombre médian d'unitigs est de 217 000. Soit une réduction de 89 % des séquences de départ (**Figure 2**).

4. Index BLight GENCODE/gnomAD

Le nombre total de transcrit annoté par GENCODE 39 est de 244 939 séquences. Après application du filtre qualité de gnomAD et du filtre de la fréquence allélique (> 1%) sur les variations issues de gnomAD, nous avons 169 458 variations exoniques ainsi que 10 850 variations mitochondriales. Cet ensemble de référence est donc constituée de 425 247 séquences. La décomposition en k-mers de taille 31 de cet ensemble à l'aide de l'outil BCALM2 fournit un ensemble de 2 649 839 unitigs. Cet ensemble d'unitigs est indexé par la suite par

BLight. L'indexation est très rapide, de l'ordre de la minute. L'index contient 162 701 116 k-mers de taille 31 et fait 382 MB en mémoire.

6. Index REINDEER d'échantillons RNA-seq normaux

L'index REINDEER est construit sur les 136 RNA-Seq normaux de poumons. L'indexation a duré environ 4 heures, et l'index est stocké sur 12 GB. L'index contient 494 323 372 k-mers de taille 31.

5. Filtre de Jaccard

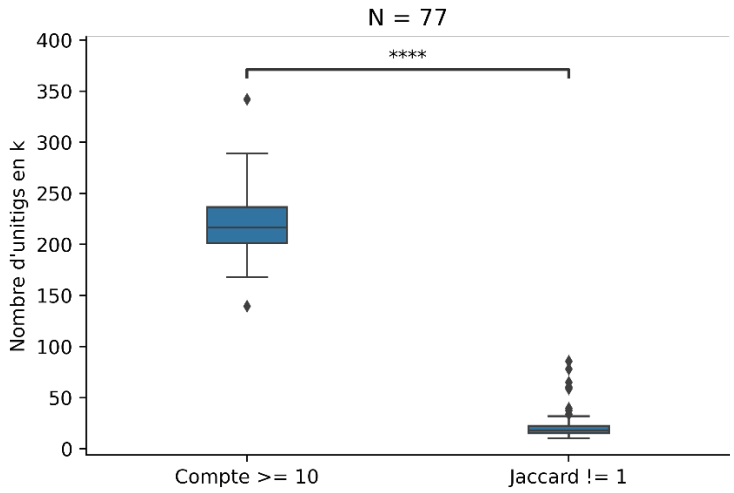

Figure 3 – Boxplot du nombre d'unitigs dans l'ensemble (N=77) des échantillons cancers après le filtre de compte et après application du filtre de Jaccard. La différence est testée par un test de Student apparié, **** : p <= 1e-04

Après application du filtre de Jaccard, seules les séquences qui ont un index de Jaccard inférieure à 1 sont conservées, sur les séquences ayant passé le filtre de compte le nombre médian d'unitigs est de 18 010. Soit une réduction de 92 % de séquences (**Figure 3**).

7. Filtre binomiale négative

Après application du filtre de la binomiale négative, seule les séquences qui ont une abondance qui s'écarte de la borne supérieure de 99 % de la distribution des comptes normaux sont conservées, sur les séquences ayant passé le filtre de Jaccard, le nombre médian d'unitigs est de 2 955. Soit une réduction de 84 % (**Figure 4**).

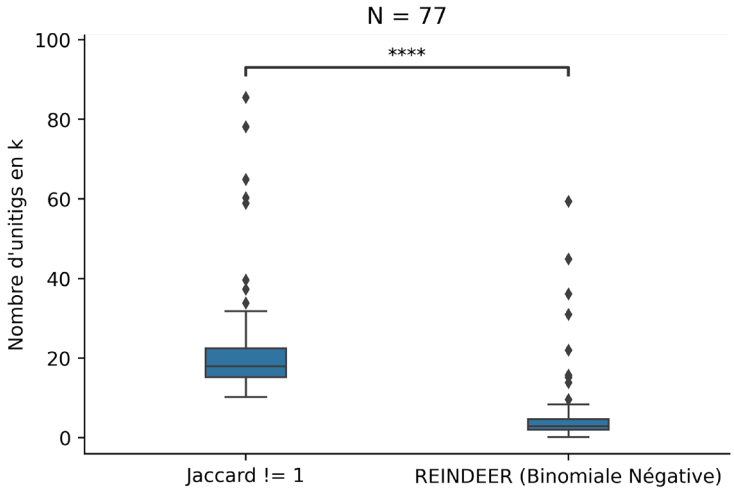

Figure 4– Boxplot du nombre d'unitigs dans l'ensemble (N=77) des échantillons cancers après le filtre de Jaccard et après application du filtre de la binomiale négative. La différence est testée par un test de Student apparié, **** : p <= 1e-04

8. Résumé de l'exclusion des fragments d'ARN normaux

	Total unitigs BCALM2	Total unitigs filtre 1	Total unitigs filtre 2	Total unitigs filtre 3	Médiane séquences restantes	Médiane % séquences éliminées
Cancer	152 M	17 M	2 M	454 k	2 955	99.85 %
Normal	269 M	38 M	4 M	435 k	1 372	99.92 %

Table 2 – Colonnes : nombre total d'unitigs en sortie de BCALM2 dans l'ensemble des échantillons, nombre total d'unitigs à l'issu du filtre de compte (filtre 1)/filtre de Jaccard (filtre 2)/filtre binomiale négative (filtre 3) dans l'ensemble des échantillons, médiane du nombre de séquences restantes à l'issu de l'application des 3 filtres dans un échantillon, médiane de la proportion de séquences restantes à la fin du protocole par rapport au nombre d'unitigs de départ de BCALM2 dans un échantillon

La succession des trois filtres : filtre de compte, filtre de l'index de Jaccard et filtre de la binomiale négative a permis de réduire drastiquement le nombre de séquences présente en entrée. Nous sommes passés d'un nombre médian de 2 millions de séquence au départ pour arriver à un nombre médian de 5 000 séquences. Ce qui représente une réduction médiane de 99.85 % des séquences en entrée.

En termes de performance de réduction, l'index de Jaccard est le meilleur. Il permet de réduire 92 % des séquences. Ce qui montre que la majorité des transcrits exprimés dans un tissu est référencée dans les banques de transcrits référencés et de variations polymorphiques.

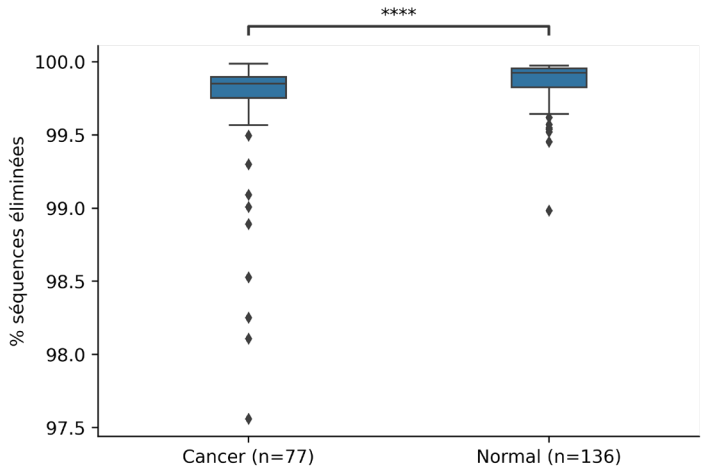

Figure 5 – Boxplot du pourcentage de séquences éliminées à la fin du protocole par rapport au nombre initial d'unitigs à analyser. La différence est testée par un test de Student, **** : p <= 1e-04

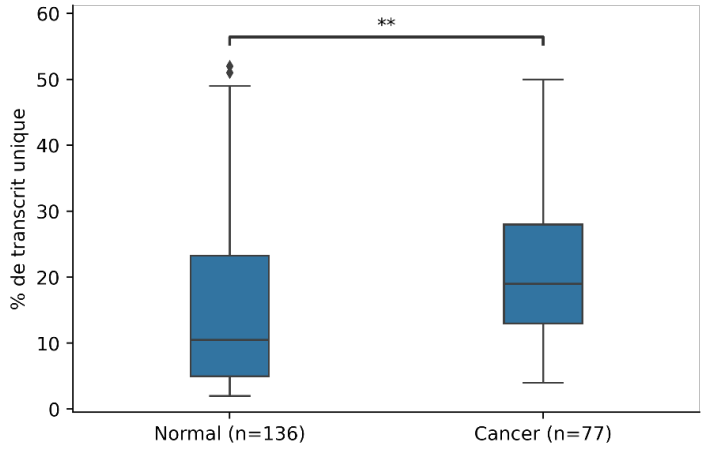

Figure 6– Boxplot de la proportion de transcrits unique à un échantillon RNA-Seq (retrouvé seulement dans cet échantillon). La différence est testée par un test de Student, ** : p = 2.1^{e}-03

Comme contrôle de notre analyse, nous avons appliqué ce protocole sur les 136 échantillons normaux de poumons. Nous éliminons plus de transcrits dans l'ensemble normal (médiane : 99.92 %) que dans l'ensemble cancer (médiane : 99.85 %) à la fin du protocole (**Figure 5**). Par ailleurs l'ensemble normal présente plus d'unitigs au départ que l'ensemble cancer, à la fin de l'analyse l'ensemble cancer a plus de transcrits que l'ensemble normal (**Table 2**). Ce qui montre la spécificité de ce protocole pour l'élimination des transcrits « normaux ».

De plus si nous regardons la proportion de transcrits propre à un transcriptome analysé. C'est-à-dire que le transcrit est présent uniquement dans une librairie et absent dans les 136 librairies normales pour le transcriptome de cancer et les 135 librairies normales pour le transcriptome normale. Nous avons une proportion de ces séquences plus élevées dans les échantillons cancer (médiane : 19 %) que dans les échantillons normaux (médiane : 10.5 %, **Figure 6**). Ainsi notre

protocole d'exclusion des transcrits normaux permet de capturer les transcrits les plus significatifs d'une tumeur.

9. Analyse de l'annotation

Par la suite nous avons annoté les séquences ayant passé la succession des trois filtres.
On note l'absence dans 17 échantillons cancers de gènes drivers (22 % des échantillons) connus de l'adénocarcinome du poumon. Sur l'ensemble des transcrits les gènes drivers ne représentent que 0 à 2 % des transcrits. Dans les 60 échantillons où des transcrits de gènes drivers sont présents, 6 échantillons ont l'ensemble des transcris qui sont totalement absent des 136 librairies normales de poumons, 30 échantillons ont l'ensemble des transcrits surexprimés qui se retrouvent au moins une fois dans les 136 libraires, 24 échantillons restant ont en moyenne 34 % des transcrits drivers qui sont totalement absent des librairies. 50 % des transcrits drivers sont des introns, 29 % sont des exons et 21 % sont des exons et des introns (rétention d'intron). ALK est présent dans 29 % des 60 échantillons. Il s'agit du gène driver le plus représenté au sein de ces 60 échantillons cancers (**Figure 7**).
La grande majorité des transcrits ALK sont annotés comme soft-clipped. C'est-à-dire que soit l'extrémité 5' soit l'extrémité 3' soit les deux ne s'alignent pas avec le génome de référence. En effet STAR présente une faiblesse pour l'alignement de séquence inférieure à 100 nt (Sun et al., 2016). Mais derrière l'annotation soft-clipped il peut s'agir d'un indel, d'un transcrit chimérique ou d'une erreur de séquençage. L'erreur de séquençage est une hypothèse exclue car la portion qui n'est pas alignée est de taille suffisamment élevée (supérieure à 10 nt).

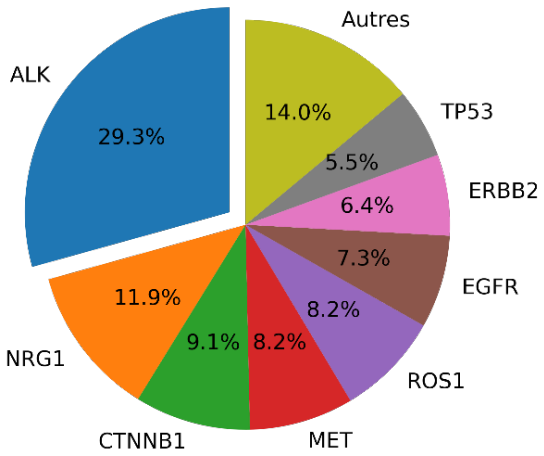

Figure 7 – Proportions des gènes drivers de l'adénocarcinome du poumon dans 60 échantillons cancer Autres proportion ≤ 5 % (EML4, PDGFRA, FGFR1, MYC, KRAS, NRAS, DDR2, PTEN, RET)

Par exemple une micro-délétion dans l'exon 19 du gène EGFR est une mutation couramment observée dans l'adénocarcinome du poumon (Seo et al., 2012). Nous identifions cette délétion dans un échantillon. Cette dernière est annotée comme soft-clipped en 3'.

L'annotation soft-clipped en 5' est retrouvé dans 28 % des transcrits drivers et 7 % en 3'. La mutation majoritaire dans les drivers est le SNV (43 %). Nous avons également des événements d'épissage alternatif (28 %), d'insertion (8 %) et de transcrit chimérique (2 %). A titre de comparaison les transcrits exoniques qui ne sont pas des drivers présentent 99 % de SNV et 1 % de soft-clipped en 5'. Ce qui montre que les gènes drivers ont une plus large diversité mutationnelle que les autres gènes.

Dans les séquences non annotées par STAR, BLAST trouve une séquence probablement apparentée à la bactérie *Acidimicrobium* dans l'ensemble des échantillons analysées (n=77). Il s'agit d'un faux positif, cette bactérie est hyperthermophile et a été retrouvé dans des sources d'eaux chaudes (Clum et al., 2009). Sur la banque de virus, BLAST trouve des transcrits apparentés au papillomavirus (HPV6, HPV69, HPV141, HPV84). Le papillomavirus HPV6 serait associé dans la cancérogénèse de l'adénocarcinome du poumon (Ragin et al., 2014).

En termes de transcrits non codant nous retrouvons des longs ARN non codant comme LINC00342 (8 échantillons) qui serait impliqué dans la métastase de l'adénocarcinome du poumon (Su et al., 2022).

Les transcrits exoniques autres que les drivers sont des gènes propres au fonctionnement de la cellule pulmonaire : les gènes mucines qui produisent le mucus sont présent 14 échantillons (MUC6, MUC3A, MUC16), et les gènes du collagène dans 9 échantillons (COL1A1, COL3A1)

Les séquences répétées retrouvés principalement sont AluSx#SINE (14 échantillons), HERVH#LTR (10 échantillons), ALR#Satellite (12 échantillons).

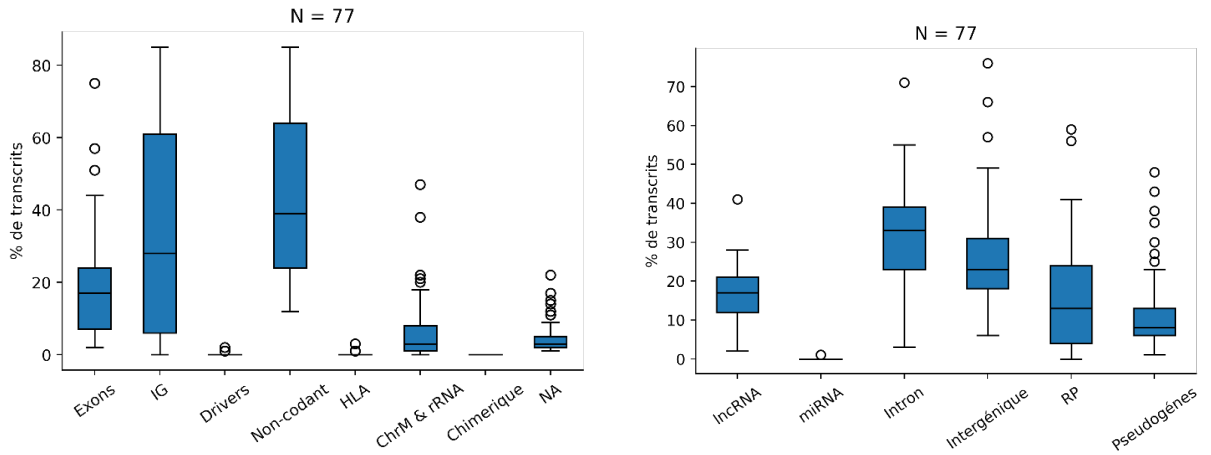

Figure 8 – Boxplot de la répartition des transcrits dans les principales catégories d'annotations (gauche) et dans les catégories de type non-codant (droite) dans les échantillons cancers. IG : Immunoglobulines, HLA (Human Leucocyte Antigen), NA (non annoté), RP (répétition)

A titre de comparaison avec les échantillons normaux, les échantillons cancers présentent une proportion plus élevée de transcrits non-codant (médiane 39 % contre 22 %, **Figure 8 & 9**) avec notamment plus de séquences répétées (médiane 13 % contre 5 %, **Figure 8 & 9**). Les immunoglobulines sont plus présentes dans les échantillons cancers (médiane 28 % contre 12 %, **Figure 8 & 9**) que les échantillons normaux. De plus il y a moins de séquences exoniques dans les échantillons cancers (médiane 17 % contre 48 %, **Figure 8 & 9**). Ces différences sont testées avec un test de Student ($p < 0.05$).

En effet les cellules cancéreuses sont caractérisées par une dérégulation de l'expression du génome non codant et plus particulièrement des séquences répétées. Enfin le système immunitaire va s'activer lors de l'apparition de cellule cancéreuse dans l'organisme.

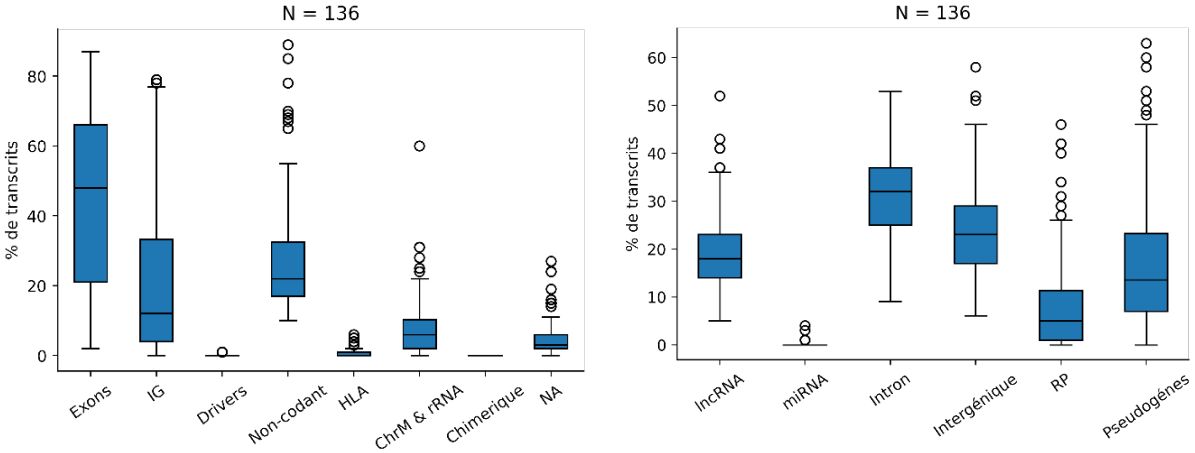

Figure 9 – Boxplot de la répartition des transcrits dans les principales catégories d'annotations (gauche) et dans les catégories de type non-codant (droite) dans les échantillons normaux. IG : Immunoglobulines, HLA (Human Leucocyte Antigen), NA (non annoté), RP (répétition)

On note l'absence de gènes drivers dans 51 des échantillons normaux (37.5 % des échantillons), parmi les transcrits identifiés le profil mutationnel est légèrement différent que dans les échantillons cancers. Nous avons plus de SNV (56 % contre 43 %), une absence de transcrits chimériques et moins d'événements d'épissage alternatifs (3 % contre 28 %).

En termes de transcrits exoniques nous retrouvons les gènes du surfactant (SFTPC, SFTPA, 61 échantillons). Les gènes de mucine (3 échantillons) et du collagène (3 échantillons) sont très peu présent par rapport aux échantillons cancer.

Enfin la répétition principale est AluSx#SINE (47 échantillons) mais on retrouve très peu les répétitions HERVH#LTR et ALR#Satellite (2 échantillons).

IV – Conclusion et perspectives

1. Conclusion

Lors de ce projet, nous avons développé une nouvelle méthode d'analyse de données RNA-Seq. Cette méthode propose d'exclure des transcrits dits « normaux » dans un transcriptome d'un échantillon de cellules cancéreuses. Pour cela une succession de filtre a été créé. Un premier filtre classique, dit filtre de compte, permet d'éliminer tous les transcrits faiblement exprimé (abondance inférieure à 10). Ainsi les transcrits issus d'erreurs de séquençage ou d'assemblage ou faiblement exprimés sont exclus du protocole. Cela permet de réduire 89 % d'un ensemble de transcrit de départ. Le deuxième filtre se base sur les connaissances acquises de l'annotation du transcriptome humain (GENCODE) et des grands projets de séquençage de cohortes (gnomAD) dans l'objectif de décrire la variabilité du génome humain. Une base de données de l'ensemble de ces informations a été construite. Elle contient 162 701 116 k-mers unique de taille 31. Cette banque peut être interrogé de manière rapide et un indicateur, index de Jaccard, permet de quantifier l'appartenance d'une séquence à cette banque. Seules les séquences qui ont index différent de 1 (absence totale des transcrits dans cette banque) sont conservées. Cela permet d'éliminer 92 % des séquences ayant passé le filtre de compte. Enfin le troisième filtre est basé sur l'indexage de jeux de données RNA-Seq « normaux ». Cet indexage nous donne deux informations : la présence d'une séquence et son abondance au sein des librairies indexées. Ainsi une banque de 136 échantillons RNA-Seq « normaux » a été créé, elle contient 494 323 372 k-mers de taille 31. Nous pouvons effectuer des requêtes contre cet index et modéliser la distribution de compte obtenu par une loi binomiale négative. Cette modélisation nous permet de calculer un seuil basé sur la borne supérieure de 99 % de cette distribution. Ainsi seule les séquences qui ont une probabilité inférieure à ce seuil et qui sont surexprimées par rapport à la distribution sont conservées. Nous éliminons 84 % des séquences ayant passé le filtre de Jaccard. L'ensemble de ces filtres éliminent 99.85 % des transcrits de départ. En comparaison, ce protocole appliqué sur les échantillons normaux élimine plus de transcrits qu'avec les échantillons cancers. De plus la proportion de transcrits unique à un échantillon est plus élevé dans les échantillons cancers que dans les échantillons normaux.

Ce qui montre que cette analyse permet effectivement d'exclure les transcrits « normaux » et de conserver les transcrits propres à une tumeur.

Un portrait moléculaire des transcrits restant peut-être dresser. Nous détectons des mutations dans les gènes drivers ainsi qu'un univers de transcrits (immunoglobulines, virus, lncRNA, répétitions …) qui semblent décrire fidèlement l'environnement moléculaire d'une tumeur.

L’environnement moléculaire d’un tissu sain présente des différences, notamment dans le type de répétition, du nombre de séquences exonique et d’immunoglobulines.

2. Perspectives

Nous pouvons étendre l’index REINDEER à une plus grande collection de librairies RNA-Seq « normales ». En effet le projet GTEx (The GTEx Consortium, 2020) a pour objectif de récolter des données transcriptomiques (RNA-Seq) à partir de tissus sains de donneurs post-mortem. Ces donneurs ne présentent aucune forme de cancer. Ce qui permet d’augmenter la puissance de cet index. En effet un inconvénient des données transcriptomiques « normales » est la possible présence de cellules cancéreuses qui vont venir brouiller le signal des transcrits normaux. Si nous augmentons la taille de cet index avec ces données, nous pourrons s’attendre à un filtre plus précis et une exclusion plus importante des transcrits « normaux ».
Le protocole développé est généralisable à d’autres questions biologiques fondamentales. Par exemple dans quelle mesure le transcriptome diffère d’un individu à autre, d’un organe à autre, d’un tissu normal d’un tissu pathologique ? Nous pouvons imaginer pour la question de comparer le transcriptome de deux organes différents de développer un index REINDEER construit sur des données d’un RNA-Seq d’un organe spécifique et d’un index BLight construit sur des gènes spécifiquement exprimés dans cet organe et effectué le protocole avec comme en entrée des données RNA-Seq d’un autre organe. Ainsi on masquerait le transcriptome catalogué d’un organe et nous conserverions le transcriptome propre d’un organe.
Pour l’aspect médical, ce protocole identifie chez un patient des altérations principalement au niveau de l’ARN qui pourraient être de nouvelles cibles pour de futurs traitements personnalisés.

V – Bibliographie

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Anand, P., Kunnumakkara, A.B., Kunnumakara, A.B., Sundaram, C., Harikumar, K.B., Tharakan, S.T., Lai, O.S., Sung, B., and Aggarwal, B.B. (2008). Cancer is a preventable disease that requires major lifestyle changes. Pharm Res *25*, 2097–2116. https://doi.org/10.1007/s11095-008-9661-9.

Audoux, J., Philippe, N., Chikhi, R., Salson, M., Gallopin, M., Gabriel, M., Le Coz, J., Drouineau, E., Commes, T., and Gautheret, D. (2017). DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. Genome Biol *18*, 243. https://doi.org/10.1186/s13059-017-1372-2.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

Chikhi, R., Limasset, A., and Medvedev, P. (2016). Compacting de Bruijn graphs from sequencing data quickly and in low memory. Bioinformatics *32*, i201–i208. https://doi.org/10.1093/bioinformatics/btw279.

Clum, A., Nolan, M., Lang, E., Del Rio, T.G., Tice, H., Copeland, A., Cheng, J.-F., Lucas, S., Chen, F., Bruce, D., et al. (2009). Complete genome sequence of Acidimicrobium ferrooxidans type strain (ICPT). Stand. Genomic Sci. *1*, 38–45. https://doi.org/10.4056/sigs.1463.

Cmero, M., Schmidt, B., Majewski, I.J., Ekert, P.G., Oshlack, A., and Davidson, N.M. (2021). MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. Genome Biol *22*, 296. https://doi.org/10.1186/s13059-021-02507-8.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res *42*, D633-642. https://doi.org/10.1093/nar/gkt1244.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). GENCODE 2021. Nucleic Acids Research *49*, D916–D923. https://doi.org/10.1093/nar/gkaa1087.

Frost, F.G., Cherukuri, P.F., Milanovich, S., and Boerkoel, C.F. (2020). Pan-cancer RNA-seq data stratifies tumours by some hallmarks of cancer. J Cell Mol Med *24*, 418–430. https://doi.org/10.1111/jcmm.14746.

Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Wang, Q., Pierce-Hoffman, E., Cummings, B.B., Alföldi, J., Francioli, L.C., Gauthier, L.D., Hill, A.J., O'Donnell-Luria, A.H., et al. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun *11*, 2539. https://doi.org/10.1038/s41467-019-12438-5.

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., and Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. J Hematol Oncol *13*, 166. https://doi.org/10.1186/s13045-020-01005-x.

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Research *49*, D884–D891. https://doi.org/10.1093/nar/gkaa942.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res *44*, D81–D89. https://doi.org/10.1093/nar/gkv1272.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun *8*, 15824. https://doi.org/10.1038/ncomms15824.

Lefranc, M.-P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., Carillon, E., Duvergey, H., Houles, A., Paysan-Lafosse, T., et al. (2015). IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Research *43*, D413–D422. https://doi.org/10.1093/nar/gku1056.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011.

Marchet, C., Iqbal, Z., Gautheret, D., Salson, M., and Chikhi, R. (2020). REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. Bioinformatics *36*, i177–i185. https://doi.org/10.1093/bioinformatics/btaa487.

Marchet, C., Kerbiriou, M., and Limasset, A. (2021). BLight: Efficient exact associative structure for k-mers. Bioinformatics btab217. https://doi.org/10.1093/bioinformatics/btab217.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.

Morillon, A., and Gautheret, D. (2019). Bridging the gap between reference and real transcriptomes. Genome Biol *20*, 112. https://doi.org/10.1186/s13059-019-1710-7.

Ragin, C., Obikoya-Malomo, M., Kim, S., Chen, Z., Flores-Obando, R., Gibbs, D., Koriyama, C., Aguayo, F., Koshiol, J., Caporaso, N.E., et al. (2014). HPV-associated lung cancers: an international pooled analysis. Carcinogenesis *35*, 1267–1275. https://doi.org/10.1093/carcin/bgu038.

Sacomoto, G.A.T., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., and Lacroix, V. (2012). KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics *13 Suppl 6*, S5. https://doi.org/10.1186/1471-2105-13-S6-S5.

Seo, J.-S., Ju, Y.S., Lee, W.-C., Shin, J.-Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.-O., Shin, J.-Y., et al. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res *22*, 2109–2119. https://doi.org/10.1101/gr.145144.112.

Su, H., Yu, S., Sun, F., Lin, D., Liu, P., and Zhao, L. (2022). LINC00342 induces metastasis of lung adenocarcinoma by targeting miR-15b/TPBG. Acta Biochim Pol https://doi.org/10.18388/abp.2020_5697.

Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., and Kocher, J.-P.A. (2016). Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Brief Bioinform bbw069. https://doi.org/10.1093/bib/bbw069.

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

Yasumizu, Y., Hara, A., Sakaguchi, S., and Ohkura, N. (2021). VIRTUS: a pipeline for comprehensive virus analysis from conventional RNA-seq data. Bioinformatics *37*, 1465–1467. https://doi.org/10.1093/bioinformatics/btaa859.

Zhang, P., Boisson, B., Stenson, P.D., Cooper, D.N., Casanova, J.-L., Abel, L., and Itan, Y. (2019). SeqTailor: a user-friendly webserver for the extraction of DNA or protein sequences from next-generation sequencing data. Nucleic Acids Research *47*, W623–W631. https://doi.org/10.1093/nar/gkz326.