LIÈGE université Library

https://lib.uliege.be

https://matheo.uliege.be

DÉTECTION DE MARQUEURS DE LA LEUCÉMIE LYMPHOÏDE AIGÜE

Auteur : Dehaye, Jordan
Promoteur(s) : PALMEIRA, Léonor
Faculté : Faculté des Sciences
Diplôme : Master en bioinformatique et modélisation, à finalité approfondie
Année académique : 2019-2020
URI/URL : http://hdl.handle.net/2268.2/9911

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

DÉTECTION DE MARQUEURS DE LA LEUCÉMIE LYMPHOÏDE AIGÜE

Année académique 2019 - 2020

Mémoire de fin d'études effectué dans le cadre de l'obtention du Master en Bioinformatique et Modélisation, à finalité, à l'Université de Liège

Jordan Dehaye

Travail effectué au Laboratoire de Génétique Humaine du CHU de Liège
Promoteur : Leonor Palmeira PhD.
Co-promoteur : Marc Hanikenne PhD.

Je remercie toutes les personnes qui m'ont encadré
et soutenu durant toute cette année de mémoire.
Merci à Leonor, Inès et Benoit pour leur accueil au laboratoire
ainsi que le suivi et l'enseignement qu'ils m'ont apporté,
de même que leurs encouragements tout au long de ce travail.

Table des matières

Liste des figures

Liste des tableaux

Relevé bibliographique

Description de la leucémie lymphoïde aigüe

La leucémie est un cancer du sang qui prend naissance dans la moelle osseuse. La moelle osseuse est constituée de cellules souches hématopoïétiques immatures en capacité de se multiplier et de se transformer. Ces cellules souches donnent naissance à deux grandes lignées qui sont la lignée myéloïde et la lignée lymphoïde. La lignée myéloïde permet la formation des érythrocytes, des plaquettes, des granulocytes et des monocytes. La lignée lymphoïde quant à elle est impliquée dans la formation des lymphocytes T et B. (voir Figure 1)

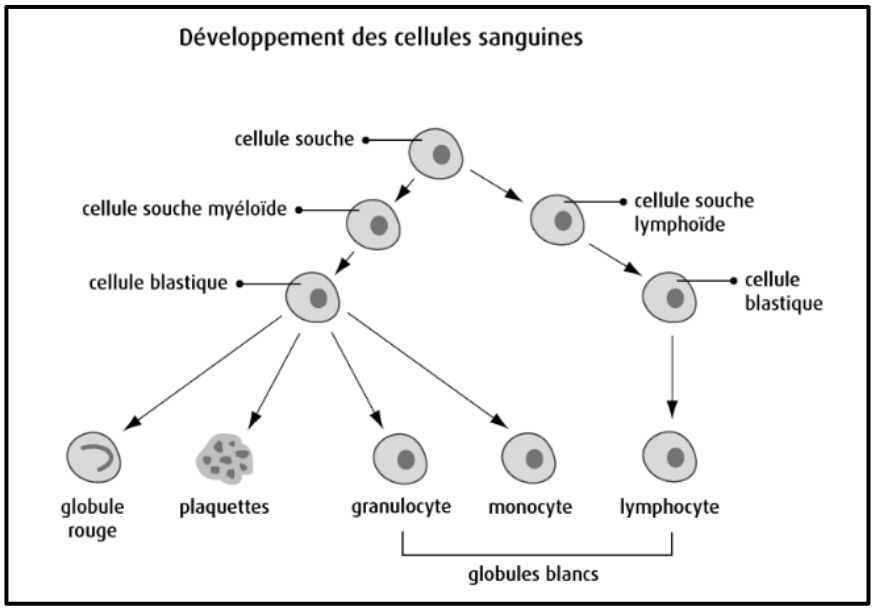


Figure 1: Différenciation des cellules sanguines depuis les cellules souches de la moelle osseuse, jusqu'à leur maturité.

Durant la maturation des cellules souches hématopoïétiques, des variations génétiques peuvent apparaître et donc provoquer l'apparition de maladies telles que la leucémie. On peut dès lors distinguer quatre grands types de leucémies[1] qui sont les suivantes :

- La leucémie lymphoïde aigüe (LLA)
- La leucémie myéloïde aiguë (LMA)
- La leucémie lymphoïde chronique (LLC)
- La leucémie myéloïde chronique (LMC)

Dans le cadre de ce mémoire, nous nous concentrons spécifiquement sur la *Leucémie Lymphoïde Aigüe* (LLA). Il s'agit d'une augmentation anormale de blastes au niveau du système sanguin. Elle peut toucher les cellules lymphocytaire B ou T.[2]

Dans la LLA, des cellules lymphoblastiques s'accumulent d'abord dans la moelle osseuse, puis bloquent la production des cellules sanguines normales avant d'être déversées dans la circulation sanguine.[3] Les cellules leucémiques atteignent, par la circulation sanguine, le foie, la rate, les ganglions, le cerveau et les testicules où elles peuvent subsister. Toutefois, les cellules de la LLA peuvent s'accumuler n'importe où dans l'organisme. Elles peuvent se disséminer dans les couches de tissus qui recouvrent le cerveau et la moelle épinière (méningite leucémique). La multiplication de ces cellules anormales peut ainsi induire, par remplacement des cellules blastiques sanguines normales dans la moyenne une anémie, une insuffisance hépatique et rénale ainsi que des lésions dans d'autres organes.[4]

Prévalence de la leucémie lymphoïde aigüe

La LLA est une forme de leucémie touchant les cellules lymphocytaires B ou T. Les cellules NK ne sont que très rarement touchées. La LLA représente 75% de tous les cas de leucémie touchant les enfants de moins de 15 ans. Ce type de leucémie présente un pronostic plus favorable chez les patients en bas âge. Dans le cas du développement d'une LLA chez l'enfant, dans 85% des cas, ce sont les lymphocytes B qui sont responsables de l'apparition d'une LLA suite à une fusion de gènes, et dans 75% des cas chez les adultes.[5] Bien que l'origine génétique des LLA soit connue, la raison de ces mutations est encore incomprise. A l'heure actuelle, les chercheurs privilégient les origines épigénétiques (rayonnement ionisant, agent mutagène, …).[6]

Transcrits de fusion

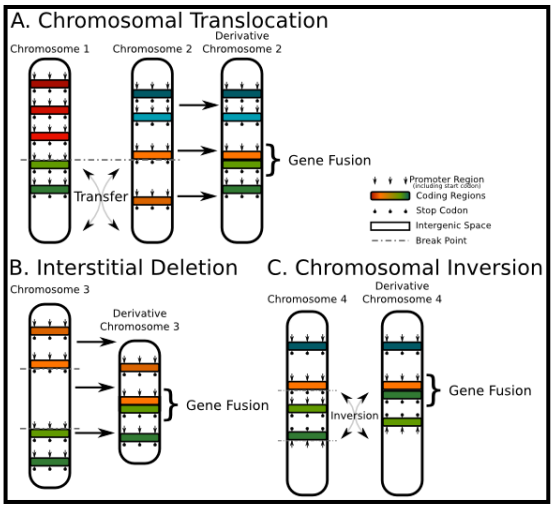


Figure 2 : Origines d'une fusion de gènes. (source : https://en.wikipedia.org/wiki/Fusion_gene)

Deux types d'altérations peuvent être à l'origine de l'apparition d'une LLA. Le plus souvent ce sont des « gènes de fusion » qui sont impliqués, et dans une moindre mesure, la LLA apparait suite à la

relocalisation d'un oncogène dans un emplacement où il est alors sur-exprimé.[7] La fusion consiste en un rapprochement de deux gènes séparés en temps normal, soit sur le même chromosome, soit sur deux chromosomes différents. La fusion peut se produire de différentes manières, suite à une translocation, une délétion, ou une inversion de gènes (voir Figure 2). L'événement de translocation est à l'origine de la plupart des occurrences de LLA. Cette altération génétique survient lors de la réplication cellulaire. Un gène peut être excisé de son locus initial pour ensuite être inséré dans un autre locus, cela peut se produire avec ou sans réciprocité. Dans de rares cas, la duplication d'un gène peut aussi être à l'origine de l'apparition d'une LLA. [8,9,10]

L'exemple de gènes de fusion le plus connu est BCR-ABL1 (découvert en 1973) qui correspond à ce que l'on nomme le « *chromosome de Philadelphie* », nommée t(9;22)(q34;q11), selon la nomenclature ISCN. Cette nomenclature signifie qu'en absence de translocation, le gène ABL1 se situe sur le bras long (q) du chromosome 9 en position 34 et le gène BCR se trouve sur le bras long (q) du chromosome 22 en position 11.[11] Cette fusion provoque, dans 95% des cas, l'apparition d'une Leucémie Myéloïde Chronique, mais elle est également à l'origine de la majorité des cas de LLA en soutenant 25% des cas d'apparition de LLA chez l'adulte et 5% des cas chez l'enfant.[12]

Des douzaines de gènes de fusions ont été mis en évidence ces trente dernières années, par plusieurs méthodes de biologie moléculaire. Différents traitements ont pu être développés pour certaines d'en elles. Dans le cas des leucémies, certains traitements sont d'origine médicamenteuses, et d'autres, de type invasifs consistent en une greffe de moelle osseuse.[13]

Cependant, les techniques de biologie moléculaire utilisées en diagnostic ne permettent de mettre en évidence que les fusions de gènes pour lesquelles on a pu produire des sondes complémentaires spécifiques en utilisant comme template les ARNm qu'elles expriment. Ceci constitue un frein majeur à la détection directe de nouvelles fusions de gènes.

Détection de fusion de gènes par la biologie moléculaire

FISH

La technique cytogénétique de FISH (*Fluorescence in situ hybridization*) est l'une des premières méthodes de génétique moléculaire à avoir été développée. Son principe se base sur la complémentarité de séquence entre une sonde et une cible. Deux sondes spécifiques biotinylées et capables de se lier à un site spécifique d'un gène sont mises en présence d'une préparation chromosomique. La liaison à l'ADN est mise en évidence grâce à un anticorps anti-biotine couplé

à une molécule fluorescente (différente pour les deux sondes).[14] Si les deux gènes impliqués dans la fusion sont présents dans leur intégralité et non altérés, les deux sondes se lieront à leurs séquences complémentaires spécifiques et seront détectables au microscope à fluorescence, à deux emplacements distincts sur leurs chromosomes respectifs. Dans le cas d'une fusion, les deux points situés en temps normal sur deux chromosomes différents, se retrouvent côte à côte.

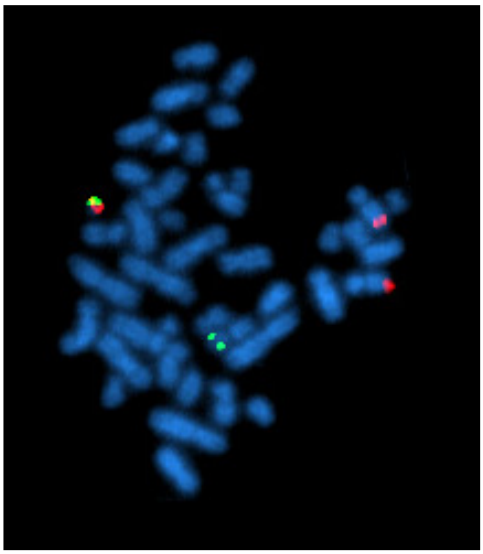


Figure 3: Image obtenue par la technique de FISH d'une fusion de gène (BCR--ABL1) (source : https://fr.wikipedia.org/wiki/Hybridation_in_situ_en_fluorescence#/media/Fichier:Bcrablmet.jpg)

La méthode de FISH est rapide et fiable mais, étant une technique ciblée, il n'est pas possible de tester la présence de nombreux types de fusions en une seule fois : seules quelques fusions à la fois peuvent être détectées sur une préparation. La FISH a une résolution de 1,5Mb pour la détection des translocations.[15] Enfin, elle requiert l'utilisation d'un appareil de microscopie à fluorescence qui reste relativement onéreux (jusqu'à 120 000 $).[16]

Le développement des méthodes de séquençage et plus particulièrement du séquençage de deuxième génération, ont permis d'augmenter la résolution des analyses, ainsi que d'évaluer en une seule fois plusieurs types de fusion.

Séquençage

Séquençage de première génération

Le séquençage consiste à déterminer la séquence nucléotidique d'un fragment ou de l'entièreté d'un génome. Le séquençage de première génération est aussi connu sous le nom de « méthode Sanger », reconnue comme le standard de référence. Cette technique utilise des didésoxyribonucléotides (ddNTP) couplés à des sondes fluorescentes. Ils sont séparés dans quatre compartiments et mis en présence des trois autres bases fonctionnelles. Les ddNTPs empêchent la polymérisation continue de la séquence par absence de terminaison 3'-OH (« chain-termination method »). Lorsqu'elles sont ajoutées à la séquence complémentaire du brin d'ADN en cours de séquençage, la polymérisation est stoppée. L'analyse sur gel d'électrophorèse révèle la composition complète de la séquence étudiée par l'union des bandes obtenues pour chacun des 4 compartiments. Cette méthode présente l'avantage d'avoir un *error-rate* de l'ordre de 0.001%.[17] Cependant, bien que les gels d'électrophorèses aient été remplacés par un système d'électrophorèse capillaire, cette méthode reste contraignante, même si elle est toujours très répandue, car elle reste inégalée en termes de qualité.

Next Generation Sequencing (NGS)

L'évolution des techniques de séquençage automatisées et des méthodes de biologie moléculaire ont permis le développement du séquençage de deuxième génération (NGS). La méthode NGS la plus utilisée à l'heure actuelle est connue sous le nom de « sequencing by synthesis (SBS) », que l'on associe généralement à la firme *Illumina*. Contrairement à la méthode Sanger où l'ajout de

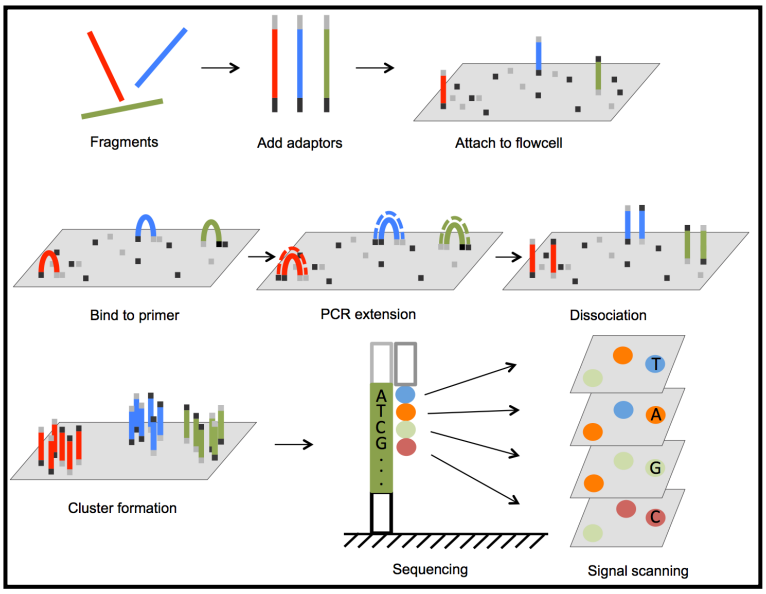


Figure 4: Principe d'amplification par pont et du séquençage par synthèse (SBS) (source Next Generation Sequencing in Aquatic Models, Yuan Lu & all. 2015)

ddNTPs stoppe totalement la polymérisation ultérieure d'un brin, la méthode SBS offre l'avantage de pouvoir continuer le séquençage du brin. Chaque nucléotide destiné à la synthèse du brin complémentaire au fragment à séquencer est fonctionnel et lié à une sonde lumineuse (différente pour les quatre bases). Une fois ajoutée en 3' du brin synthétisé, la base libère sa sonde, qui émet un flash lumineux capté par plusieurs caméras. La libération de la sonde permet également à la base d'être fonctionnelle pour continuer la réaction de polymérisation. Le flash lumineux est capté par plusieurs caméras (voir Figure 4).[18,19]

Comme l'illustre la figure, les fragments sont fixés et amplifiés (*bridge amplification*) sur le flowcell pour générer des clusters de fragments identiques, qui seront ensuite séquencés.[20]

Pour cela, l'ADN est extrait des cellules et fragmenté en séquences pouvant aller jusqu'à 500 pb. Après sélection des fragments sur base de leur taille, les fragments (« inserts ») sont réparés en ajoutant, à leurs extrémités, des adaptateurs qui vont permettre la fixation au flowcell et contiendront également une séquence qui permettra l'initiation de la polymérisation (grâce à l'utilisation d'un primer spécifique). Son avantage principal est son débit (environs ≈4-5Gb par run en 27h en continu avec le séquenceur de type « MiSeq »).[21] Il existe deux protocoles de séquençage en SBS, le Single-end et le Paired-end. Le premier consiste à séquencer l'insert par un seul côté, et le second, à séquencer l'insert par les deux côtés. La différence dans leur choix correspond à l'objectif de l'expérimentation. Le Paired-end permet un meilleur ancrage lors du mapping. Dans le cadre de la détection de fusions de gènes, il permet de mapper la séquence sur le génome et de détecter une fusion présente dans l'insert mais absente des reads en eux-mêmes . En effet, si le point de cassure des gènes fusionnés correspond à un emplacement non séquencé de l'insert situé entre les deux paires séquencées, la fusion sera tout de même détectée, étant donné qu'un read mappera sur l'un des deux gènes d'origine de la fusion et le read complémentaire mappera sur l'autre gène. Par opposition, l'utilisation du Single-end ne permettrait pas de déduire le point de cassure dans le cas d'un fragment qui ne serait pas séquencé à 100%, car chaque read mapperait sur l'un des deux gènes impliqués, mais par sur le second gène. Malgré un error-rate de 0.1%[22], et un nombre de bases par reads séquencés inférieur à celui de Sanger (± 150 bases pour *Illumina* « short read sequencing », contre 900 bases pour Sanger), la méthode Illumina reste la plus répandue, surtout grâce à son coût (1000$ pour un génome entier contre 4000$ pour une analyse « haute qualité » avec la méthode Sanger)[22,23]. La méthode de séquençage Illumina présente en revanche un inconvénient qui est de ne pouvoir séquencer que de l'ADN. Il est donc impossible de séquencer directement le transcriptome d'une cellule. Cependant, ce problème peut être contourné grâce à la méthode RNAseq.

RNAseq

L'intérêt du séquençage ARN (*RNAseq*) est d'offrir un instantané de l'entièreté du transcriptome des cellules à une moment précis. Le RNAseq permet de déterminer le niveau d'expression des différents gènes en une seule fois, sans connaitre au préalable quel gène est précisément exprimé. Cependant, l'analyse de transcriptome complet séquence une quantité très importante d'ARNs présentant souvent un intérêt limité selon l'étude. En effet, 85% des ARNs cellulaires sont ribosomiques, alors que les ARNm ne représentent que 3-5% des ARNs codant pour des protéines.[24] De ce fait, deux défauts sont mis à jour, le premier étant qu'une étape d'extraction des ARNm est nécessaire pour éviter de séquencer des ARNs ultra-majoritaires qui ne présenteraient qu'un intérêt limité. Le second défaut est lié à la différence d'expression des gènes. Tous les ARNm n'étant pas exprimés de la même manière, il est alors possible qu'un ARN présent en grande quantité puisse « éclipser » la présence d'autres ARNm, la puissance statistique pour détecter un événement précis sera alors limitée.[25]

Pour remédier à ces problèmes, des méthodes de détections ciblées ont été mises au point pour sélectionner spécifiquement les ARNm que l'on souhaite séquencer. L'une d'entre elle est la méthode *Target Anchored MultiPlex.*

Méthode Target AMP

La méthode *Target Anchored Multiplex PCR* (AMP) est une méthode d'analyse destinée à produire rapidement des librairies enrichies et ciblées pour le NGS (voir Annexe 1).[26] Le principe est similaire à celui du RNAseq, à la différence que, dans le cas de la méthode ciblée, seuls les ARNs comprenant une séquence précise sont amplifiés et séquencés. Cette méthode offre différents avantages par rapport au RNAseq classique. D'abord elle ne nécessite pas d'extraire les ARNm de l'échantillon. En effet, l'utilisation de primer spécifiques aux séquences recherchées permet d'amplifier exclusivement ces séquences et ainsi, éviter qu'elles soient masquées par d'autres séquences présentes initialement en plus grande quantité. De plus, le séquençage ciblé offre la possibilité d'augmenter la précision des analyses, tout en diminuant les coûts par la diminution de la quantité d'éléments à séquencer. La méthode AMP utilise également un système de « Molecular Barcode » (MBC) qui sont uniques et spécifiques à chaque fragment et permettent d'effectuer les comptages des fragments supportant la fusion présents au départ dans l'échantillon. Ces MBCs permettent la déduplication de données et la correction d'erreurs (pour les reads ayant le même barcode, une séquence consensus peut être générée et permet d'éliminer les erreurs de séquençage).

Cette technique est donc tout à fait adaptée dans le cadre du diagnostic de la présence de transcrits de fusion. En effet, comme pour la méthode de FISH le ciblage de séquence permet de mettre spécifiquement en évidence la présence d'une fusion. Mais, comparativement à la méthode de FISH, la méthode *Target AMP* offre une sensibilité plus importante. Effectivement, la technique de FISH manque de sensibilité pour les tests à haut volume de détection (la détection de plusieurs fusions sur un même patient nécessite plusieurs préparations histologiques) et peut passer à côté d'une fusion si celle-ci présente également une micro-délétion. De plus, les tests immuno-histochimiques nécessitent d'utiliser des anticorps de haute qualité pour révéler la fusion.[27]
Aussi, dans la méthode de FISH, deux loci seulement sont visualisés dans une cellule, il est donc facile de les manquer.[27] Enfin, l'état compact de la chromatine peut conduire à localiser deux régions cibles l'une à côté de l'autre alors que ce n'est pas réellement le cas.[14] Ces problèmes sont résolus grâce au séquençage via la méthode *Target AMP*. Cependant, le défaut de cette technique se révèle dans le cas d'une nouvelle fusion de gènes de séquence inconnue, il sera impossible de la détecter à l'aide des sondes si elles n'ont pas été préalablement ciblées. Il est donc possible, dans le cas d'une recherche de gènes de fusion, d'obtenir un résultat de type *faux-négatif* si la fusion recherchée ne correspond pas à une fusion déjà connue. Un autre défaut de la technique *Target AMP* est inhérent à la méthode de séquençage à proprement parler. Effectivement, les appareils de séquençages NGS ne sont capables de séquencer que de l'ADN. De cette manière, les fragments d'ARNm sélectionnés doivent obligatoirement subir une étape de rétro-transcription pour régénérer le brin d'ADN initial. Des erreurs peuvent dès lors apparaitre, de l'ordre de une pour 15-27 000 bases.[28]

Third Generation Sequencing (TGS)

Le problème du séquençage NGS ne pouvant séquencer que de l'ADN, est solutionné grâce au séquenceurs TGS, qui peuvent séquencer directement l'ARN sans nécessité de passer par une phase de rétro-transcription. Parmi les technologies de séquençage de troisième génération nous nous focalisons ici sur le nanopore, qui est capable de séquencer une molécule d'ADN ou d'ARN, même si elle est présente en une seule copie. Nonobstant le fait que son taux d'erreur se situait autour des 20% en 2018[29], il a pu être abaissé à 1,5% en séquençant également le brin complémentaire.[30] Le principe du nanopore (Figure 5) est d'utiliser une membrane percée de trous où sont placées individuellement des structures protéiques en forme de tonneau, les *pores*.
Le fragment à séquencer glisse d'un coté à l'autre de la membrane, à travers le pore, grâce à un champs électrique. Chaque base qui passe au travers du pore modifie le champ électrique.

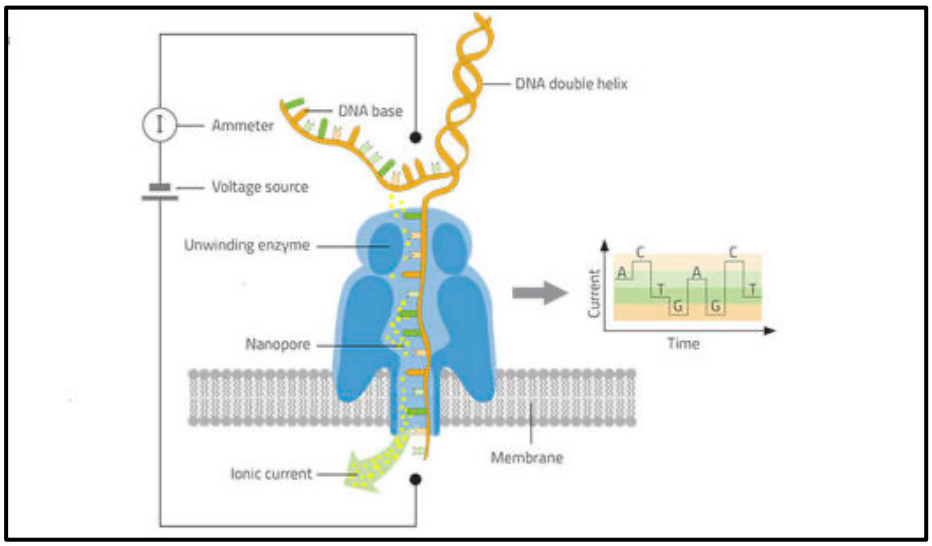


Figure 5: Principe du nanopore (source: https://www.cirad.fr/)

L'intensité de cette variation diffère en fonction du nucléotide incorporé.[31] Ce type de technologie de séquençage se démocratise un peu plus chaque jour (1500$ pour actuellement un appareil MinION, de la taille d'un téléphone). De plus le nanopore permet de générer des reads très longs de 13-20 kb (pouvant même aller jusqu'à 134 kbp, contre 900 bp pour le NGS, 100 pour le Sanger !)[32] à la vitesse de 500 bases par seconde.[33]

La bioinformatique pour la détection de gènes de fusion

En 1990, le séquençage complet d'un génome humain a nécessité 13 ans de travail et couté 2.7 milliards de dollars. Aujourd'hui, le séquençage d'un génome complet peut être obtenu en seulement quelques jours.[34] Les séquenceurs NGS ont offert des possibilités techniques en termes d'études des génomes et de l'expression des gènes, mais ont également fait exploser la quantité d'informations (informatiques et biologiques) renvoyées. Le séquençage d'un génome humain complet contient typiquement de l'ordre de 100 Go de données.[35] Ces données peuvent bien entendu être compressées pour leur stockage, mais la quantité de données à analyser reste très importante et la bioinformatique rencontre trois challenges importants : celui de développer des algorithmes qui s'adaptent à cette abondance de données ; celui de trouver des solutions softwares et hardwares pour permettre le stockage de ces données ; et celui d'adapter la puissance de calcul d'un laboratoire aux données et aux algorithmes existants. De ce fait, la bioinformatique arrive désormais au premier plan des besoins d'un laboratoire dans les étapes d'analyse des données de séquençage. En effet, la bioinformatique est une discipline qui nécessite des connaissances pointues dans plusieurs domaines disciplinaires autrefois séparés (statistiques, informatique,

génétique moléculaire) et le transfert de ces compétences au sein des laboratoires de biologie moléculaire se fait plus lentement que la vitesse à laquelle des données s'accumulent.

Au-delà du séquençage de génome, le RNAseq offre des perspectives pour la compréhension de l'expression des gènes. Les exemples incluent les réponses cellulaires à des stimuli physiologiques, les effets de perturbations expérimentales sur des gènes et des voies spécifiques, ou le dysfonctionnement de la régulation génique dans les maladies.[36]

Une analyse RNAseq suit typiquement la série d'étapes suivantes : les ARNs d'intérêt (souvent les ARNm) sont extraits, fragmentés, transformés en cDNA destiné au NGS et enfin préparés via l'ajout d'adaptateurs, de barcodes, … Lors du séquençage de la librairie, plusieurs échantillons à la fois sont séquencés durant un même run de séquençage. L'étape de démultiplexage permet de réassigner chaque read à son échantillon et de générer les fichiers FASTQ nécessaires à l'analyse bioinformatique. Dans la plupart des cas, les reads doivent subir plusieurs étapes supplémentaires avant d'être alignées sur le génome. Les données sont alors alignées sur un génome de référence pour connaitre l'origine des reads.

Un problème qui peut se poser lors d'une analyse RNAseq, est consécutif au fait que les séquences sont discontinues. La plupart des outils de mapping sont capables de mapper une séquence complète sur un génome. Cependant, la question de la détection des transcrits épissés et des fusions de gènes nécessite de pouvoir détecter les reads dont une partie mappe sur un gène, et le reste, sur une autre partie de ce gène ou un autre gène.

Dans le cadre de ce mémoire, nous avons choisi d'utiliser STAR ainsi que STAR-Fusion pour la détection et l'analyse des fusions présentes dans nos données. STAR-Fusion est une version adaptée de STAR, qui utilise des options de STAR programmées par défaut et qui correspondent aux options les plus adaptées pour la détection des fusions. Cet outil est reconnu comme étant l'un de ceux présentant la meilleure sensibilité et vitesse d'exécution.[37]

STAR-Fusion permet non seulement de détecter les fusions en les mappant sur un génome de référence, mais également de confirmer leur présence en assemblant en contigs, les régions géniques correspondant aux fusions détectées. Puis il mappe à nouveau les reads sur ces contigs. Si une paire de reads provient bien d'une fusion de gènes, cette paire s'aligne alors parfaitement sur le contig et elle est validée.[38] L'assemblage des reads en contigs utilise TrinityFusion (qui est également un outil de détection de fusion, cependant dans le cas d'une utilisation spécifique et individuelle pour la détection de fusion, il présente une haute spécificité, pour une sensibilité inférieure à celle de STAR-Fusion).[37] Cette association STAR-Fusion / TrinityFusion permet donc de détecter rapidement et valider efficacement les fusions potentielles mises à jour en RNAseq.

De plus, STAR a été développé non seulement pour le traitement de reads de taille courte (10bp et moyenne (200 bp), mais il montre également un potentiel pour la détection de fusions en utilisant des longs reads, pouvant aller jusqu'à plusieurs kilobases.[39] Les outils STAR et STAR-Fusion

présentent donc les qualités recherchées pour une analyse RNAseq pour la détection de transcrits de fusion sur des données NGS (Illumina) et TGS (Nanopore).
Le développement d'un pipeline bioinformatique utilisant les outils STAR et STAR-Fusion, permettrait la détection de transcrits de fusion via les techniques NGS pour les analyses RNAseq sur les ARNm non ciblés, ou à termes, une analyse RNAseq non ciblée utilisant des méthodes TGS telles que le nanopore et effectuée directement sur les ARNm totaux. Cette nouvelle méthode de mise en évidence de transcrit de fusion, en se basant essentiellement sur une analyse bioinformatique des séquences ARNm totales, pourrait être utilisée comme outil de diagnostic rapide et qui présenterait une fiabilité plus importante que les méthodes biochimiques actuelles dont la limitation principale est l'obligation de la connaissance préalable des séquences recherchées.

Approche méthodologique

Méthodologie biologique et séquençage

Quatre échantillons de sang dont les patients présentent un diagnostic de leucémie et positifs à différentes fusions de gènes ont été choisis pour cette validation. Les ARN des cellules lymphoïdes ont été isolés et amplifiés en utilisant la méthode « target » (*Archer® AMP FusionPlex ALL kit*) qui cible des fusions de gènes connues et qui intègrent des Molecular Barcodes (MBC) à chaque fragment d'ARN réparé, de manière à identifier de façon précise les reads dupliqués (PCR). Les librairies de séquençage des ARNm sélectionnés ont ensuite été produites par « bridge amplification » selon la méthode décrite par Archer®. Le séquençage Paired-end (en 2x250 et 2x150) a été réalisé par la technique SBS Illumina sur un séquenceur MiSeq pour obtenir, pour chaque échantillon, deux fichiers FASTQ contenant les paires de reads pour chaque échantillon.

Méthodologie bioinformatique

L'analyse bioinformatique a été réalisée à partir de containers Singularity et de modules sur le Cluster du GIGA/CHU (SLURM version 14.11.11). Nous avons d'abord évalué la qualité des reads grâce à l'outil FastQC (version 0.11.5) de manière à détecter la présence d'éléments ne faisant pas partie des séquences d'intérêt telles que les barcodes et les adaptateurs Illumina par l'analyse de séquences sur-représentées. Suite à cette analyse, nous avons utilisé l'outil UrQt (version 1.0.17) pour trimmer les queues polyT et faciliter le mapping, ainsi que l'outil Trimmomatic (version 0.32) qui nous a permis d'éliminer le début et la fin des reads dont la qualité est faible, en utilisant l'option `-TruSeq3-PE-2.fa :2 :30 :10` pour exciser les adaptateurs *Illumina* mis en évidence par

FastQC, puis l'option `-MINLEN 20` pour éliminer les reads dont la longueur est inférieure à 20 bases. A la fin du trimming, l'option `-MINLEN 10` a été ajoutée à Trimmomatic pour éliminer les reads restant, dont la taille était inférieure à 10 bases après trimming. Nous avons, par ailleurs, développé un script Perl de manière à déplacer les 8 bases de l'identifiant du barcode (MBC) dans le nom de chaque read (séparation par un « _ ») et aussi éliminer les 13 bases de la séquences commune suivant les barcodes. Les 21 premiers caractères de la ligne qualité, sont bien entendus aussi éliminés pour chacun des reads. (voir Annexe 2)

Une seconde analyse FastQC a permis de confirmer l'augmentation de la qualité des reads de chaque échantillon.

Nous avons utilisé le génome de référence humain GRCh38_gencode_v32_CTAT_lib_Dec062019 pour mapper les reads grâce à l'outil STAR-Fusion (version .1.8.1) du *Trinity Cancer Transcriptome Analysis Toolkit (CTAT)* qui permet de détecter les fusions dans les données de transcriptome en utilisant les options `-FusionInspector -inspect` et `-validate`, faisant ainsi appel à l'outil Fusion Inspector (version 2.3.1) Celui-ci permet de générer un assemblage *de novo* de « mini-contigs » au départ des séquences chromosomiques sur lesquelles les fragments ont été mappés. Il aligne à nouveau les séquences originales sur ces contigs. Une fusion avérée produira ainsi un alignement parfait et les reads supportant chaque fusion sont alors disponibles dans les fichiers BAMs d'output. Grâce à un outil Perl in-house en combinaison avec Samtools, nous avons compté les reads tronqués supportant directement la fusion (fichier « finspector.junction_read.bam ») ainsi que les paires encadrant une fusion (fichier « finspector.spanning_read.bam »), de ce fait, chaque read d'une même paire mappe sur un gène/chromosome différent. Nous avons par ailleurs, au fur et à mesure de l'analyse des fusions, utilisé IGV (version 2.8.7) pour visualiser les fichiers BAMs des différentes étapes.

Pour finir, et de manière à valider les fusions détectées, nous avons ensuite comparé nos résultats aux résultats obtenus par le logiciel officiel de la firme Archer®, « Archer® Analysis (version 6.0) ». Ce logiciel est utilisé en routine par nos collègues de diagnostic au CHU au sein de leur analyse LLA accréditée ISO15189 par BELAC. Pour effectuer la comparaison, comme les données traitées par le logiciel Archer® Analysis sont mappées sur le génome de référence GRCh37/hg19, nous avons utilisé l'outil LiftOver (sur l'UCSC Genome Browser) pour convertir les emplacements des gènes mappés au niveau du génome de référence hg19 vers le génome de référence hg38 utilisé lors de notre analyse.

Résultats

Première analyse FastQC

La taille moyenne et le nombre total des reads de chaque fichier, représentant chacune des paires de reads séquencés dans chaque échantillon, sont repris dans le tableau N°1.

Nom des fichiers	Taille des reads en b	Nombre total de reads
Sample-1_R1	238 - 251	765563
Sample-1_R2	239 - 251	765563
Sample-2_R1	138 - 151	2113116
Sample-2_R2	138 - 151	2113116
Sample-3_R1	139 - 151	1823441
Sample-3_R2	136 - 151	1823441
Sample-4_R1	238 - 251	841089
Sample-4_R2	239 - 251	841089

Tableau 1 : Noms des échantillons (colonne 1), taille des reads des données brutes dans chaque fichier FASTQ en bases (colonne 2), nombre total de reads des données brutes de chaque fichier FASTQ (colonne 3).

L'analyse FastQC a permis de démontrer la présence des séquences d' « Adaptateurs Universels *Illumina* », à la fin des reads, ce qui rend plus difficile la détection et l'analyse des reads chimériques (STAR-Fusion ne détecte alors que 1.46 % de reads chimériques sur l'échantillon S1 et 4.20 % sur l'échantillon S2). Nous avons donc effectué un nettoyage des adaptateurs et en avons profité pour éliminer les extrémités des reads qui présentaient une faible qualité de séquençage (voir Figure 6). Les reads sont ensuite traités avec un script Perl in-house de manière à éliminer les MBCs des reads et scores qualité et à déplacer cette information dans le nom (*header*) de chaque read.

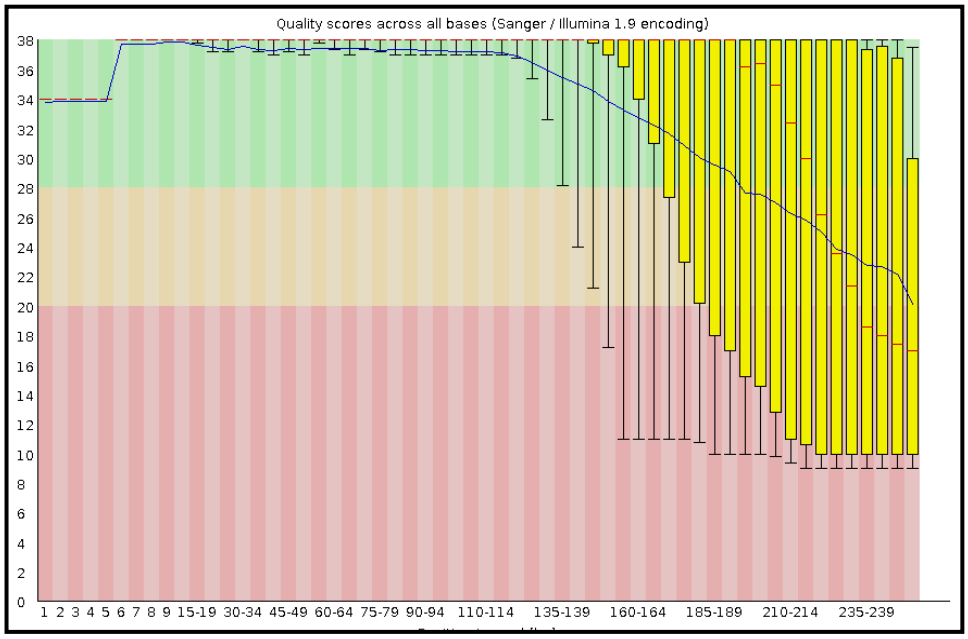


Figure 6 : Résultat de FastQC pour l'analyse de la qualité des reads du fichier FASTQ Sample-1_R1. La qualité de la séquence décroit au fur et à mesure du séquençage. L'axe des abscisses représente le score de Phred (>30 = qualité > 99,9%) pour chaque base, en ordonnée, pour les 200 000 premiers reads du fichier FASTQ.

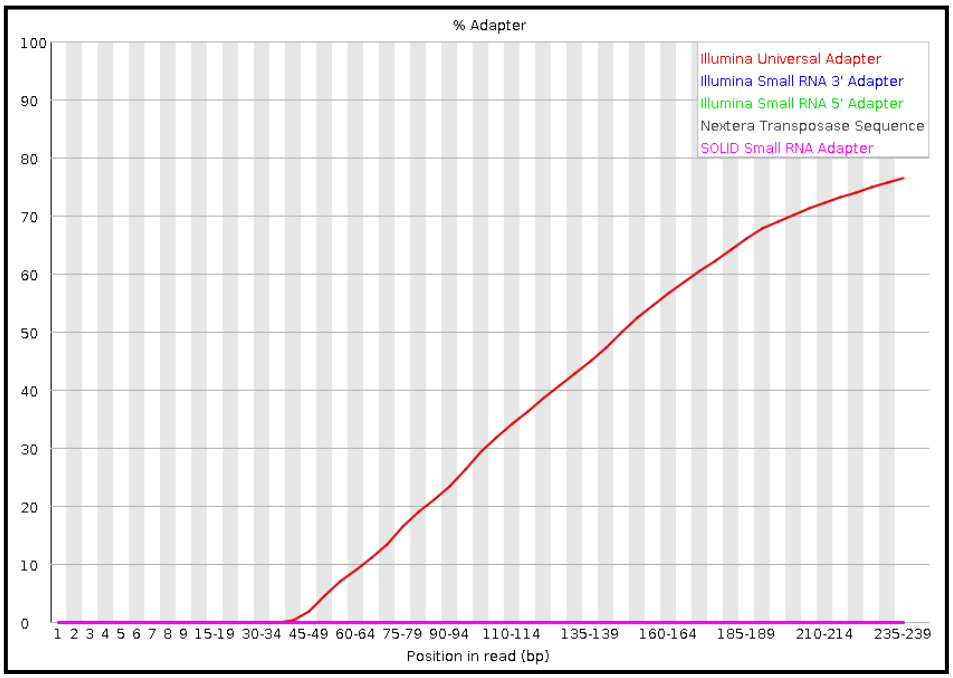


Figure 7: Exemple de la présence d'Adaptateurs Universel Illumina suite à une analyse FastQC sur le fichier FASTQ Sample-1_R1. L'axe des ordonnées indique la position générale (en bases) des adaptateurs sur les reads. En ordonnée le pourcentage d'adaptateur par rapport la progression de l'analyse sur la séquence.

Suite à la préparation des reads, une nouvelle analyse avec FastQC indique que les adaptateurs universel *Illumina* ont bien été éliminés.

Il en résulte aussi que la qualité des reads a grandement augmenté (voir Figure 8).

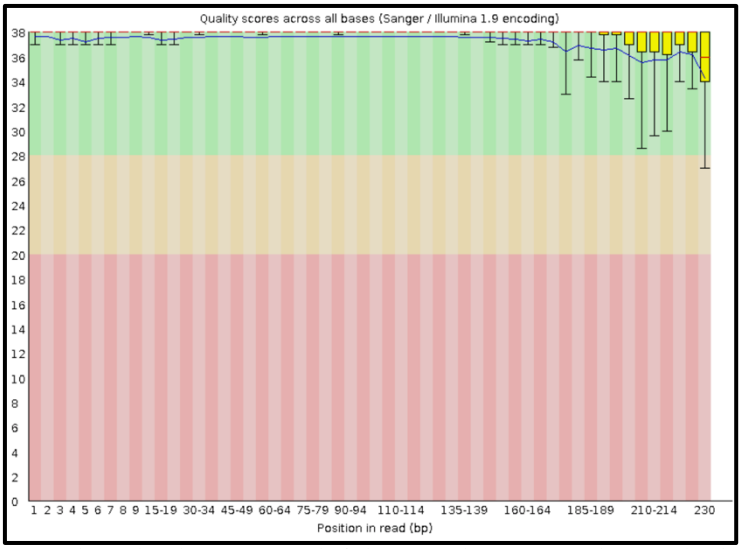


Figure 8: Résultats de FastQC pour la qualité des séquences du fichier Sample-1_R1. avec le score de Phred en absisce pour la totalité des bases (en ordonnée) du fichier FASTQ. A gauche la qualité des reads avant le nettoyage et à droite la qualité qui a augmenté après le nettoyage.

Fichiers FASTQ	Nombre de reads au départ	Nombre de reads restant	Pourcentage de reads conservés
Sample-1	765563	740557	96.7%
Sample-2	2113116	2070156	97.9%
Sample-3	1823441	1783686	97.8%
Sample-4	841089	814436	96.8%

Tableau 2: Nombre de reads des fichiers FASTQ pour chacun des 4 échantillons pairés R1 et R2, avant l'utilisation des outils de trimming UrQt et Trimmomatic (colonne 2), après l'utilisation de UrQt et Trimmomatic (colonne 3) et le pourcentage de reads conservés (colonne 4), entre le nombre de reads au départ et le nombre de reads après l'étape de trimming.

Le pourcentage de séquences conservées pour chacun des 4 échantillons après le nettoyage se situe entre 96,7 et 97,9%.

Détection des fusions de gènes

Les résultats statistiques concernant les reads Paired-end des 4 échantillons mappés par STAR-Fusion sur le génome de référence humain hg38.

Échantillons	Pourcentage de reads mappés
Sample-1	66.42 %
Sample-2	44.91%
Sample-3	45.07 %
Sample-4	68.16 %

Tableau 3 : Pourcentage de reads mappés pour chacun des 4 échantillons.

Résultats Archer®

Chaque échantillon a été préalablement analysé par le logiciel de détection de traitement Archer® *Analysis* qui a permis la mise en évidence des fusions ciblées. Les fusions détectées sont reprises dans le Tableau 4. Chaque échantillon présente les noms des 2 gènes impliqués, excepté l'échantillon *Sample-2*, qui présente une fusion (IKZF1) avec un seul gène, en plus de la fusion BCR--ABL1. La fusion IKZF1 est le résultat d'une délétion dans le gène IKZF1.

Échantillons	Fusions détectées par le logiciel Archer® Analysis
Sample-1	KMT2A--MLLT1
Sample-2	IKZF1 / BCR--ABL1
Sample-3	BCR--ABL1
Sample-4	RUNX1--RUNX1T1

Tableau 4: Fusions détectées par le logiciel Archer® Analysis sur les 4 échantillons. L'échantillon 2 présente un gène de fusion à un seul nom, contrairement aux autres fusions qui sont indiquées par les 2 partenaires impliqués dans la fusion.

Le Tableau 4 sert de référence pour la validation des fusions détectées par STAR-Fusion.

Validation des fusions détectés

Différentes fusions ont été mises en évidence par STAR-Fusion et validées par FusionInspector (cf. :Tableau 5).

	Fusions détectée par STAR-Fusion (en gras celles détectées aussi par Archer®)	Nombre de Junction reads (FusionInspector)	Nombre de Spanning reads (FusionInspector)	Nombre de reads totaux supportant la fusion	Archer® Spanning reads (SS)
Sample-1	**KMT2A--MLLT1**	223	22	245	177
Sample-2	**BCR--ABL1**	283	72	355	826
	IKZF1--ACTR2	1	0	1	462
	ETV6--UBRN	0	0	0	/
	ETV6--YBX3	2	1	3	/
	AC119673.2--ABL2	1	0	1	/
	CTNNB1--CHD1	1	0	1	/
	DNTT--ZNF608	2	0	2	/
	HPS4--KMT2A	2	0	2	/
	SUP98--STIM1	0	0	0	/
	BCR--LEF1-AS1	1	1	2	/
	ETV6--NAP1L1	1	1	2	/
Sample-3	**BCR--ABL1**	682	129	811	392
	ETV6--YBX3	4	0	4	/
	CXCR4--YBL6	0	0	0	/
	IER2--ABL2	1	0	1	/
	PTMA--CHD1	3	0	3	/
	DDX5--CHD1	2	0	2	/
	ELL--KLF2	2	0	2	/
	IKZF1--AC005520.1	2	0	2	/
	IKZF1--NCOR2	1	0	1	/
	IKZF1--SNX2	1	0	1	/
	NUP98--STIM1	1	0	1	/
	PAX5--MEPCE	1	0	1	/
	SOAT1--ABL2	1	0	1	/
	TMP4--CHD1	0	0	0	/
	IKZF1--SET	1	2	3	/
	BLNK--DNTT	1	1	2	/
	DLEU2--CHD1	1	1	2	/
	ETV6--AC046134.2	1	1	2	/
	ETV6--LRP6	1	2	3	/
	KLHL18--PTK2B	1	1	2	/
	LEF1--CHD1	1	1	2	/
	PAX5--SFPQ	1	1	2	/
	ZNF700--ABL1	1	1	2	/
Sample-4	**RUNX1--RUNX1T1**	524	36	560	341
	OAZ1--KLF2	7	1	8	/
	IKZF1--GALK2	1	0	1	/

	INTS9--PTK2B	1	0	1	/
	RASSF1--KLF2	2	0	2	/

Tableau 5: Fusions de gènes détectées par STAR-Fusion et validées par FusionInspector pour chacun des 4 échantillons (colonne 2), le nombre de reads se trouvant directement sur la fusion la fusion (colonne 3), le nombre de reads entourant la fusion (colonne 4), le nombre de reads total supportant la fusion (colonne 5), et le nombre de reads totaux supportant la fusion mis en évidence par le logiciel Archer® Analysis (colonne 6). Le nombre de reads supportant chaque fusion a été calculé à l'aide d'un script bash in-house.

La colonne 2 du Tableau 5 révèle pour l'échantillon 1, une seule fusion qui correspond à celle mise en évidence par le logiciel Archer®. STAR-Fusion détecté une fusion principale (BCR--ABL1) pour l'échantillon 2. Dix fusions potentielles supplémentaires ont aussi été détectées par STAR-Fusion. L'échantillon 3 montre la présence d'une fusion correspondant à celle détectée par Archer®, ainsi que 22 fusions potentielles. L'échantillon 4 présente une fusion correspondant à celle détectée par le logiciel Archer®, mais également 4 fusions potentielles. Le nom et le nombre de reads supportant chaque fusion a été compté par un script bash in-house (voir Annexe 3), pour déterminer le poids de chaque fusion détectée. Nous pouvons séparer les fusions « principales » détectées à la fois par STAR-Fusion avec un très grand nombre de reads supportant la fusion, des fusions « potentielles », détectées à la fois par STAR-Fusion avec un très faible nombre de reads.
Des fusions détectées par Archer®, seule la fusion IKZF1 de l'échantillon 2 n'est pas détectée par STAR-Fusion. Le logiciel a détecté une fusion IKZF1--ACTR2 qui ne correspond pas à une délétion dans le gène IKZF1.
La modification dans STAR-Fusion des paramètres de distance entre 2 emplacements de mapping n'a pas permis de détecter la délétion dans le gène IKZF1.
Nous inférons ainsi plusieurs fusions pour chacun des 4 échantillons, dont 4 des 5 fusions attendues.

Calcul de sensibilité et précision

La sensibilité et la précision sont calculées en utilisant les éléments du Tableau 5. Un threshold a été utilisé pour ne sélectionner que les fusions qui présentent un nombre de reads supportant la fusion égal ou supérieur à 15.

Sensibilité VP/(VP+FN)	Précision VP/(VP+FP)	
4/(4+1) = 0.8	4/(4+0) = 1	avec seuil de 15 reads
4/(4+1) = 0.8	4/(4+35) = 0.10	sans seuil

Tableau 6: Calculs de sensibilité et précision par rapport aux événements détectés par STAR-Fusion.

En se basant sur les fusions détectées par STAR-Fusion, en comparaison avec celles détectées par le logiciel de référence Archer® Analysis, on détermine pour STAR-Fusion, une sensibilité de

80% mais une précision de 100%, dans le cas de l'utilisation d'un threshold de 15 reads. Dans le cas où le threshold n'est pas utilisé, la sensibilité augmente à 100%, mais abaisse la précision à 0,1%.

Comparaison des emplacement gènes de fusion détectés par le kit LLA FusionPlex et STAR-Fusion

Le logiciel Archer® Analysis utilise le génome hg19 pour mapper les séquences sélectionnées. Il fournit les résultats des emplacements des gènes chimériques mis en évidence par le kit. STAR-Fusion utilise, dans notre cas, le génome hg38 pour mapper les reads des échantillons. La comparaison des séquences a été faite en utilisant l'outil LiftOver. Seules les fusions détectées parallèlement par le kit et STAR-Fusion sont présentées dans le Tableau 7.
Les emplacements sélectionnés sont ceux qui présentent le plus de reads qui supportent donc le mieux la fusion (*spanning* et *junction* reads). Les positions des cassures sont exactement identiques à celles détectées par le logiciel Archer® Analysis.

Échantillon	Fusion détectée	Emplacements *Archer® hg19*	Emplacement STAR-Fusion hg38	Liftover hg19 -> hg38
Sample-1	KMT2A--MLLT1	chr11:118355690,chr19:6270770	chr11:118484975,chr19:6270759	chr11:118484975,chr19:6270759
Sample-2	BCR--ABL1	chr22:23524426,chr9:133729451	chr22:23182239,chr9:130854064	chr22:23182239,chr9:130854064
Sample-3	BCR--ABL1	chr22:23632600,chr9:133729451	chr22:23290413,chr9:130854064	chr22:23290413,chr9:130854064
Sample-4	RUNX1--RUNX1T1	chr21:36231771,chr8:93029591	chr21:34859474,chr8:92017363	chr21:34859474,chr8:92017363

Tableau 7: Gènes à l'origine des fusions détectées à la fois par STAR-Fusion et le kit FusionPlex LLA (colonne 2), l'origine des gènes détectés par le logiciel Archer® via le kit FusionPlex avec un mapping sur le génome hg19 (colonne 3), l'origine des gènes détectés par STAR-Fusion (colonne 4) et la conversion par LiftOver des emplacement des gènes sur le génome hg19 vers le génome hg38 (colonne 5).

Dans l'échantillon 2, STAR-Fusion avait également mis en évidence la présence d'une fusion potentielle IKZF1--ACTR2. Le gène IKZF1 est impliqué dans un événement de fusion mis en évidence par le kit Archer® et correspond à une délétion au niveau du chromosome 7. STAR-Fusion a détecté la présence de reads s'alignant en partie sur le gène IKZF1 à l'emplacement chr7:50399918 sur le génome hg19, et qui correspond à l'emplacement chr7:50467616 sur le génome hg38.

Comptage des Molecular Barcodes

Les molecular barcodes (MBCs) ont été comptabilisés pour déterminer le nombre de reads différents qui supportent la fusion, en utilisant un script bash in-house (voir Annexe 3). Chaque read possédant en principe un barcode, il en résulte que chacun des barcodes est présent une à cinq fois lors de la détection des fusions pour chacun des échantillons (selon les échantillons). On utilise le fichier `junction.reads` où se trouvent les reads situés sur la fusion et qui contient le plus de séquences validées par STAR-Fusion. En comparaison, le fichier `spanning.reads` contient bien des MBCs présentant chacun 2 à 4 occurrences par barcodes, les occurrences contenues dans le fichier `spanning.reads` sont toujours présentes en nombre paire, étant donné qu'ils entourent la fusion par paire.

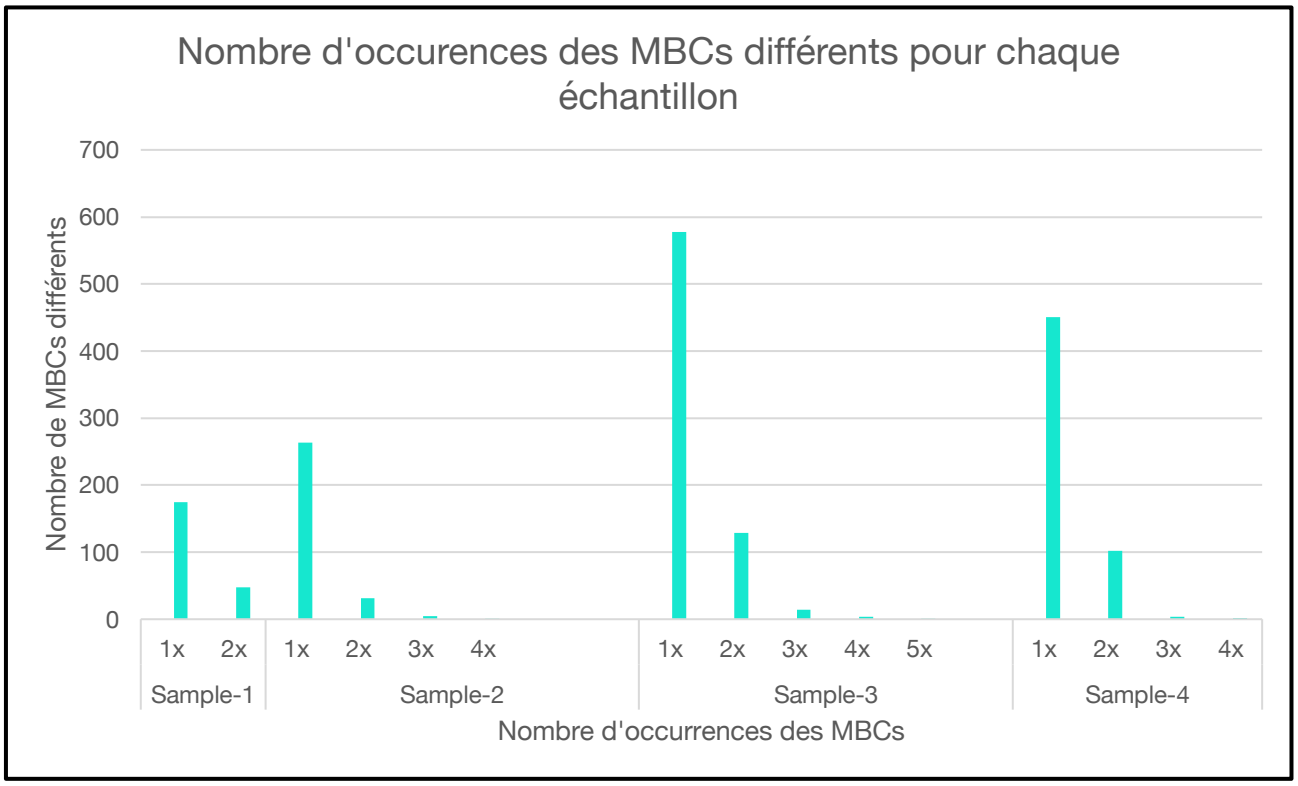


Figure 9: Nombre de MBCs différents et leurs occurrences dans les fichiers junction.reads pour les 4 échantillons.

Discussion

Nous avons été capables de détecter, au moyen d'un pipeline utilisant STAR-Fusion, quatre événements de fusions connus sur les cinq attendus, pour les échantillons des patients ayant été diagnostiqués comme porteurs d'au moins une fusion de gènes provoquant une leucémie lymphoïde aigüe.

La sensibilité de notre méthode est ainsi de 80%. Le pipeline développé dans le cadre de notre étude est, de ce fait, correctement adapté pour détecter, sans a priori la présence de fusions de gènes sur des données RNAseq *Illumina*. En revanche, la détection de fusion par STAR-Fusion a jusqu'à maintenant, bénéficié d'une sélection de gènes ciblée. Dans le cas d'une détection de gènes effectuée sur des échantillons RNAseq n'ayant pas subi de sélection de fragments préalables, la précision, ainsi que la sensibilité pourraient diminuer car nous pourrions avoir beaucoup plus d'artefacts techniques de par la nature de la technique. En effet, STAR-Fusion a également détecté la présence d'une série de fusions « potentielles » pour chacun des échantillons, qui se trouvent être des faux-positifs. Cependant, ces fusions ne sont que très peu supportées : on a pu aisément définir un threshold qui semble cohérent sur base de ces quelques tests, et ainsi amener pratiquement à 0, le nombre de fusions artefactuelles détectées par STAR-Fusion, en comparaison avec les fusions avérées mises en évidence par le logiciel Archer® Analysis. Grâce au threshold, on passe alors d'une précision de 10% à une précision de 100%, les fusions potentielles non ciblées sont éliminées des résultats, pour ne renvoyer que les fusions avérées. Néanmoins, la sensibilité de l'outil STAR-Fusion est de 80% lorsque les résultats obtenus ne sont pas soumis à un threshold de 15 reads, et elle reste à 80% lorsque les résultats sont filtrés avec un threshold de 15 reads supportant la fusion. Ceci est très encourageant pour la détection de fusions. Cette valeur pourrait cependant baisser si le pipeline développé ici est utilisé sur un plus grand nombre d'échantillons. En effet, plus la taille de l'échantillon augmente, et plus la valeur mesurée pour la sensibilité va être proche de la réalité, elle pourrait donc augmenter ou, plus probablement, baisser un peu.

Star-Fusion n'a pas permis de détecter spécifiquement la fusion IKZF1 issue d'une délétion au niveau du chromosome 7 et détectée par le logiciel Archer® Analysis. Cependant, il a pu associer son origine au niveau du chromosome 7, mais pas sa terminaison sur le même chromosome. Ceci pourrait provenir du fait que la calibration de l'outil STAR-Fusion semble adaptée pour la détection de fusions dont les reads mappent sur deux chromosomes différents. La distance minimale entre deux emplacements pour qu'une fusion soit détectée par STAR-Fusion, est de 100 000 bases[40]. Cette distance, dans le cas de la délétion d'une partie du gène IKZF1 à l'origine de la fusion, est de 100 263 bases (distance détectée par le logiciel Archer® Analysis). Étant donné que STAR-Fusion utilise différents points (distance entre les deux emplacements, nombre de reads supportant la fusion, taille de la séquence mappée sur chacun des emplacements) pour valider les fusions afin

d'augmenter sa précision, il est probable qu'il n'ait pas pu détecter cette fusion/délétion précisément et ait mappé les reads correspondant en grande partie sur le gène IKZF1, et pour la fin du read, l'ait plutôt associée à un autre gène qui correspondait de manière plus adaptée selon ses critères. Cependant, STAR-Fusion a été en mesure de détecter et de valider la présence de reads correspondant à une section du gène IKZF1. Néanmoins, STAR-Fusion possède différentes options de détection qui permettent de diminuer la distance nécessaire pour estimer qu'une fusion est bien présente. En revanche, la modification de la distance minimale entre 2 emplacements de mapping pour un même read pourrait diminuer la précision des résultats et ainsi augmenter l'apparition de faux-positifs. Lors des essais de modification des distances entre 2 emplacements de mapping pour la détection de la fusion dans le gène IKZF1, STAR-Fusion n'a pas été en mesure de détecter cette fusion. Différents tests sur un nombre d'échantillons plus élevés seront nécessaires pour déterminer les valeurs des options permettant de détecter les délétions.

STAR-Fusion a détecté la présence de plusieurs fusions potentielles pour trois des quatre échantillons. Mais, celles-ci n'étant supportées que par un nombre très faible de reads, la présence de ces fusions potentielles peut se situer dans les artéfacts liés aux manipulations techniques (extraction, PCR, …), ou être d'origine biologique.

Une fusion artéfactuelle peut résulter d'un problème lors des étapes de préparation de la librairie. Les étapes de préparation utilisent la transcription inverse, dans ce cas, lors de la phase d'élongation, la RT-PCR peut parfois sauter d'un segment d'ARN en phase de rétro-transcription vers un autre un fragment d'ARN homologue, ce qui conduit à la formation d'un « faux transcrit chimérique ». Également, un fragment d'ARN présent dans le milieu peut faire office de primer aléatoire et provoquer l'amplification involontaire d'un segment. Dans ces deux cas, c'est le nombre de reads supportant la fusion potentielle qui va déterminer la validation de la fusion.[41]

Dans le cas d'une fusion artéfactuelle d'origine biologique, elle peut être le résultat de cis-splicing dans lequel l'ARN a été épissé et correspond alors à une séquence semblable à celle d'un gène de fusion. Puisque, dans notre cas, la sélection des fragments est faite via un primer spécifique, le fragment artéfactuel devrait être présent en quantité très faible en comparaison des fragments chimériques véritables. Ce qui permet alors de les éliminer en tant que fusion probable, grâce à l'utilisation d'un threshold. Certains événements de splicing peuvent aussi conduire à la formation d'un transcrit alternatif tronqué. Un ARN semblable à celui issu d'une fusion de gènes, peut avoir été produit lorsque la transcription du gène ne s'est pas terminée à l'endroit prévu et a donc débordé sur le gène suivant. Dans ce cas, STAR-Fusion peut détecter ce genre d'anomalie si les deux gènes ont été correctement annotés.[41] Les reads ne sont alors pas validés comme étant issus d'une fusion et n'apparaissent pas dans le fichier BAM de validation et d'alignement.

On remarque que l'utilisation des MBCs présente un résultat particulier. Plus des trois quart des reads supportant la fusion pour chaque échantillon présente un MBC unique. Nous nous serions attendus à un nombre de MBCs plus élevés, qui justifierait leur utilisation. Il ne s'agit pas d'un

problème en soi, car un faible nombre de répétitions MBCs indique plutôt une bonne qualité de la librairie. Nous effectuerons un retour vers les techniciens pour les informer de ce point.
Un autre élément interpelant se trouve dans l'analyse des reads situés directement sur la fusion. Lors de l'analyse des fichiers `junction_reads` des échantillons, il apparait qu'il contient un nombre considérable de reads supportant directement la fusion elle-même. Ce qui amène à la réflexion que les fragments ciblés durant la phase de PCR pourraient être trop petits, puisqu'ils se situent majoritairement sur la fusion. Les reads sont par ailleurs *overlappant* (chevauchant), car nous avons identifié les deux paires d'un même insert dans le fichier `junction_reads` et l'information qu'ils apportent quant aux origines de la fusion, est limitée. Dans le cas d'une manipulation correctement calibrée, les reads ne devraient pas se chevaucher et donc les *spanning reads* entourant la fusion devraient être en nombre bien plus conséquent qu'ils ne le sont actuellement. De ce fait, augmenter la taille de l'insert permettrait d'obtenir une augmentation du nombre de reads entourant la fusion (*spanning reads*), car cela augmente la probabilité d'obtenir des reads pairés qui mappent chacun sur un gène à l'origine de la fusion. L'emplacement exact de la fusion ne nécessite que très peu de *junction reads* (voir Tableau 5), ainsi beaucoup plus de *spanning reads* pourraient supporter la fusion car le séquençage contiendrait plus d'information (deux reads uniques + une distance entre reads) à la place d'information redondante (de l'ordre de 200bp). Un insert plus grand, permet d'éviter le séquençage complet de l'insert et d'utiliser la puissance du Paired-end sequencing pour la détection des fusions. L'outil bioinformatique STAR-Fusion détecterait directement les origines des gènes de fusion comme cela a été démontré lors de notre travail.
Enfin, nous avons également mis en évidence que, dans le cadre de notre étude, le génome hg38 est tout à fait adapté à la détection des fusions, en comparaison du génome hg19 utilisé par Archer® Analysis. Le génome hg38 étant plus complet, car plus récent, il peut être utilisé pour des analyses ultérieures plus approfondies, comme les variations ponctuelles des séquences (SNP)[42] ou la détection d'un transcrit chimérique fictif.

Conclusion

Le pipeline mis au point a bien permis de nettoyer les reads provenant d'un séquençage des ARNm ciblé par la méthode AMP sur des d'ADNc par la technique de « sequencing by synthesis » et de mettre en évidence la présence la présence de transcrits issus d'une fusion de gènes. Pour chacun 4 des 5 échantillons analysés, le pipeline utilisant STAR-Fusion a permis de détecter des gènes de fusion identiques à ceux mis en évidence par le logiciel de la firme Archer®. Il a également détecté, pour chacun des échantillons, la présence d'autres fusions potentielles. Les fusions potentielles ont été éliminées grâce à une sélection basée sur un threshold de 15 reads supportant la fusion. Cette méthode permet d'obtenir une sensibilité de détection de 80% dans le cas d'une sélection

ciblée des séquences issues d'une fusion de gènes. De plus, en utilisant un threshold de 15 reads, la précision passe de 10% à 100% de fusions détectées tout en gardant une sensibilité identique. Concernant les MBCs, la méthode utilisée dans le cas de notre étude a montré que leur utilisation ne semble pas pertinente dans le cadre de la manipulation AMP. Aussi, la taille des inserts semble être trop petite pour du séquençage Paired-end, ce qui induit un coût de séquençage élevé pour le peu d'information qu'ils peuvent apporter. L'augmentation de la taille de l'insert serait judicieux pour augmenter le nombre de reads entourant la fusion et ainsi détecter de façon plus robuste les origines du gène chimérique.

La méthode a également démontré que l'utilisation du génome hg38 pour la validation de la méthode est acceptée pour la comparaison avec les résultats obtenus avec le logiciel Archer® Analysis, puisque les emplacement des séquences obtenues après le mapping sont identiques pour hg19 et hg38 en utilisant l'outil de conversion *LiftOver*.

Dans le cadre d'un développement plus approfondi de notre méthode, si nous travaillons sur une méthode non ciblée, l'avantage sera de pouvoir détecter n'importe quelle fusion sans connaissance a priori des partenaires de fusion. Ceci nécessite de définir la couverture de séquençage nécessaire pour capturer suffisamment de reads entourant la fusion pour pouvoir la détecter, ce qui dépend de l'abondance du transcrit dans le transcriptome et du pourcentage d'infiltration tumorale. [44,45]

Bibliographie

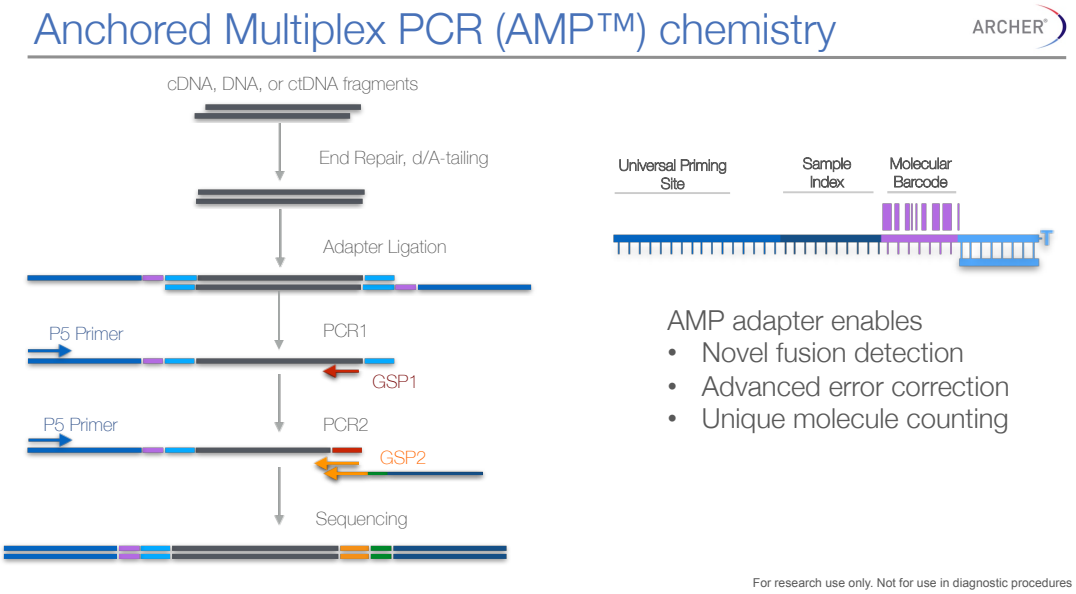
1. Redaelli A, Stephens JM, Laskin BL, Pashos CL, Botteman MF. The burden and outcomes associated with four leukemias: AML, ALL, CLL and CML. *Expert Rev Anticancer Ther*. 2003;3(3):311-329. doi:10.1586/14737140.3.3.311
2. Puckett Y, Chan O. Acute lymphocytic leukemia. *xPharm Compr Pharmacol Ref*. 2007;13(10):1-5. doi:10.1016/B978-008055232-3.60841-4
3. Richard A. Larson MD. Leucémie lymphoblastique aiguë. *Livret Acute Lymphoblastic Leuk*.
4. Sayyed AH, Aleem A, Al-Katari MS, et al. Acute lymphoblastic leukemia presenting with liver infiltration and severe lactic acidosis. *Am J Case Rep*. 2018;19:453-457. doi:10.12659/AJCR.907383
5. Acute C, Leukemia L, Health P, Library N, Institutes N. Childhood Acute Lymphoblastic Leukemia Treatment (PDQ ®) General Information About Childhood Acute Lymphoblastic Leukemia (ALL). 2015;(Md):1-103.
6. Fusion I, Using T, Generation N, Review SA. Identifying Fusion Transcripts Using Next Generation Sequencing Advanced Review. 2017;7(6):811-823. doi:10.1002/wrna.1382.Identifying
7. Roy-Tourangeau M. Stratégie de détection des anomalies chromosomiques dans les leucémies aiguës pédiatriques. *Univ Montréal*.
8. López-Nieva P, Fernández-Navarro P, Graña-Castro O, et al. Detection of novel fusion-transcripts by RNA-Seq in T-cell lymphoblastic lymphoma. *Sci Rep*. 2019;9(1):1-11. doi:10.1038/s41598-019-41675-3
9. Avet-Loiseau H. Fish analysis at diagnosis in acute lymphoblastic leukemia. *Leuk Lymphoma*. 1999;33(5-6):441-449. doi:10.3109/10428199909058449
10. Schichman SA, Caligiuri MA, Gu Y, et al. ALL-1 partial duplication in acute leukemia. *Proc Natl Acad Sci U S A*. 1994;91(13):6236-6239. doi:10.1073/pnas.91.13.6236
11. Gonon-Demoulian R, Goldman JM, Nicolini FE. Historique de la leucémie myéloïde chronique: Un paradigme de traitement du cancer. *Bull Cancer*. 2014;101(1):56-67. doi:10.1684/bdc.2013.1876
12. Leoni V, Biondi A. Tyrosine kinase inhibitors in BCR-ABL positive acute lymphoblastic leukemia. *Haematologica*. 2015;100(3):295. doi:10.3324/haematol.2015.124016
13. Parker BC, Zhang W. Fusion genes in solid tumors: An emerging target for cancer diagnosis and treatment. *Chin J Cancer*. 2013;32(11):594-603. doi:10.5732/cjc.013.10178
14. Bishop R. Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Biosci Horizons*. 2010;3(1):85-95. doi:10.1093/biohorizons/hzq009
15. Romana Serge MV. Cytogénétique moléculaire. In: *Collège National Des Enseignants et Praticiens de Génétique Médicale*. ; 2011.
16. Gozzetti A, Le Beau MM. Fluorescence in situ hybridization: Uses and limitations. *Semin Hematol*. 2000;37(4):320-333. doi:10.1053/shem.2000.16443
17. Victoria Wang X, Blades N, Ding J, Sultana R, Parmigiani G. Estimation of sequencing error rates in short reads. *BMC Bioinformatics*. 2012;13(1):1-12. doi:10.1186/1471-2105-13-185
18. Slatko BE, Gardner AF, Ausubel FM. Overview of Next Generation Sequencing technologies (and bioinformatics) in cancer. *Mol Biol*. 2018;122(1):1-15. doi:10.1002/cpmb.59.Overview
19. Lu Y, Shen Y, Warren W, Walter R. Next Generation Sequencing in Aquatic Models. *Next Gener Seq - Adv Appl Challenges*. 2016. doi:10.5772/61657
20. Flexibility RL, Depth CT. Genomic sequencing - illumina. *Specif Sheet*. 2010. doi:10.1385/0-89603-248-5:169
21. Illumina. HiSeq Systems. *Specif Sheet*. 2011:1-4. http://res.illumina.com/documents/products/datasheets/datasheet_hiseq_systems.pdf.
22. Glenn TC. Field guide to next-generation DNA sequencers. *Mol Ecol Resour*.

2011;11(5):759-769. doi:10.1111/j.1755-0998.2011.03024.x
23. NHGRI. The Cost of Sequencing a Human Genome. *Natl Hum Genome Res Inst*. 2016:19. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost%0Ahttps://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost%0Ahttps://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.
24. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of Cell Growth. *Science (80-)*. 2010;330(6007):1099-1102. doi:10.1126/science.1192588
25. Łabaj PP, Kreil DP. Sensitivity, specificity, and reproducibility of RNA-Seq differential expression calls. *Biol Direct*. 2016;11(1):1-12. doi:10.1186/s13062-016-0169-7
26. Zheng Z, Liebers M, Zhelyazkova B, et al. Anchored multiplex PCR for targeted next-generation sequencing. *Nat Med*. 2014;20(12):1479-1484. doi:10.1038/nm.3729
27. Markey FB, Ruezinsky W, Tyagi S, Batish M. Fusion FISH imaging: Single-molecule detection of gene fusion transcripts in situ. *PLoS One*. 2014;9(3). doi:10.1371/journal.pone.0093488
28. Arezi B, Hogrefe HH. Escherichia coli DNA polymerase III ε subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT-PCR procedures. *Anal Biochem*. 2007;360(1):84-91. doi:10.1016/j.ab.2006.10.009
29. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. *Dev Growth Differ*. 2019;61(5):316-326. doi:10.1111/dgd.12608
30. Kaire Loit , Kalev Adamson , Mohammad Bahram , Rasmus Puusepp , Sten Anslan , Riinu Kiiker, ReinA, B Drenkhan LT. Relative performance of Oxford Nanopore MinION vs. Pacific Biosciences Sequel third-generation sequencing platforms in identification of agricultural and forest pathogens. 2019.
31. Montel F. Séquençage de l'ADN par nanopores: Résultats et perspectives. *Medecine/Sciences*. 2018;34(2):161-165. doi:10.1051/medsci/20183402014
32. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res*. 2018;28(2):266-274. doi:10.1101/gr.221184.117
33. Ben McNally1,*, Alon Singer1,*, Zhiliang Yu1, Yingjie Sun1, Zhiping Weng2, and Amit Meller1 +. Optical recognition of converted DNA nucleotides for single- molecule DNA sequencing using nanopore arrays. *Bone*. 2008;23(1):1-7. doi:10.1038/jid.2014.371
34. Gyles C. The DNA revolution. *Proc Inst Civ Eng - Eng Hist Herit*. 2018;172(3):92-93. doi:10.1680/jenhh.2019.172.3.92
35. The Genomic Data Challenges Of The Future. 2018;The Medica(July):2017.
36. Fei Ji RIS. RNA-seq: Basic Bioinformatics Analysis. *Curr Protoc Mol Biol 2018 Oct ; 124(1) e68 doi101002/cpmb68*. 2018;176(3):139-148. doi:10.1016/j.physbeh.2017.03.040
37. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*. 2019;20(1):1-16. doi:10.1186/s13059-019-1842-9
38. Haas B, Dobin A, Stransky N, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*. 2017:120295. doi:10.1101/120295
39. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
40. Guide W. Local Run Manager RNA Fusion. (June 2018).
41. Peng Z, Yuan C, Zellmer L, Liu S, Xu N, Liao DJ. Hypothesis: Artifacts, including spurious chimeric RNAs with a short homologous sequence, caused by consecutive reverse transcriptions and endogenous random primers. *J Cancer*. 2015;6(6):555-567. doi:10.7150/jca.11997
42. Pan B, Kusko R, Xiao W, et al. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*. 2019;20(Suppl 2). doi:10.1186/s12859-019-2620-0
43. Boer JM, den Boer ML. BCR-ABL1-like acute lymphoblastic leukaemia: From bench to

bedside. *Eur J Cancer*. 2017;82:203-218. doi:10.1016/j.ejca.2017.06.012
44. Maher CA, Palanisamy N, Brenner JC, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A*. 2009;106(30):12353-12358. doi:10.1073/pnas.0904720106
45. Palanisamy N, Monzon F, Lowery-nordberg M. Transcriptome Sequencing To Detect Gene Fusions in Cancer Microarrays for Clinical Cancer Diagnostics. *Science (80-)*. 2011;458(7234):225. doi:10.1038/nature07638.Transcriptome

Annexes

Annexe 1



Annexe 2

```
#!/usr/bin/env perl

# avoid boilerplate
use Modern::Perl '2011';
use Smart::Comments;
use Path::Class 'file';

unless (@ARGV == 2) {
        die << "EOT";
        Usage: $0 <infile.fastq> <outfile.fastq>
        This tool remove the common sequence and the barcode from and Illumina unzipped fastq
file
        and put the barcode of the read, at the end of the header.
        It requieres a fastq file as input and the name of the output file to be created.
        Example: $0 S1_R1.fastq final.fastq
EOT
}


my $infile = shift;                     # Lit le nom du fichier d'entrée
my $outfile = shift;                    # Lit le nom du fichier de sortie
my $sout = "";                          # Déclare la variable de stockage des données en traitement

open(FH, '<', $infile) or die "Could not open file '$infile' $!";

my $ProcessedLine = 0;                  # Suivre si on a déjà traité la ligne
my $NextIsQuality = 0;                  # Suivre si la prochaine ligne est celle de qualité

while (<FH>) {                          # Lit le fichier depuis le début jusqu'à la fin
my($in) = $_;                           # Stocke le fichier dans une variable liste
$ProcessedLine = 0;

 if (($in =~ m/\@M00/) && ($ProcessedLine == 0)) {                      # Si une ligne contient la
                                                                        chaine « @M00: »
                                                                        enregistrer qu'on
                                                                        démarre un nouveau read
chomp $in;                                                              # Ne pas prendre en compte le
                                                                        retour à la ligne

        $in =~ s/\s/-/g;                        # remplacer les blanc par «-»
        $sout .= $in .= "";                     # La variable $out est remplie avec $in
        $NextIsQuality = 0;                     # Passer à la ligne qualité qui est mise à 0 (FALSE)
        $ProcessedLine = 1;                     # Sortie de la ligne processée qui est alors à 1 (TRUE)
}
   elsif (($in =~ m/^[ACGTN]*$/) && ($NextIsQuality == 0) && ($ProcessedLine == 0)) {

# Si la ligne commence par les caractères ACGT, 21x; utiliser les 21 premiers caractères comme $in

my $first8 = substr $in, 0, 8;          # Déclarer une variable contenant les 8 premiers caractères
        my $substring = substr ($in, 21);       # Elimine les 21 premiers caractères de la séquence
```

```
        $sout .= "_$first8\n";                  # Ajoute les 8 caractères du barcode à la fin du header
        $sout .= $substring;                    # Ajoute de séquence moins les 21 premiers caractères

        $ProcessedLine = 1;
}
   elsif (($in =~ m/^\+$/) && ($ProcessedLine == 0)) {      # Si la ligne ne contient qu'un "+"
   $sout .= $in;                                            # Ajouter la ligne  « + » au fichier out
        $NextIsQuality = 1;
        $ProcessedLine = 1;
}
   elsif (($NextIsQuality == 1) && ($ProcessedLine == 0)){      # Si on est à la ligne de qualité,
                                                                elle ne match pas avec les 3
                                                                autres
        my $substring = substr ($in, 21);
        $sout .= $substring;
        $NextIsQuality = 0;
        $ProcessedLine = 1;
}
}
close(FH);
open(FH, '>', $outfile);
print FH $sout;
close(FH);
print "final_out.fastq  done\n";                # On imprime une indication comme quoi tout
                                                s'est bien passé. Le fichier final est créé grâce au
                                                unless du début.
```

Annexe 3

Extraction des noms des différentes fusions validées dans un fichier BAM et comptage de leur nombre :

```
samtools view finspector.junction_reads.bam | awk '{print $3}' | sort -u
| wc -l

samtools view finspector.spanning_reads.bam | awk '{print $3}' | sort -u
| wc -l
```

Comptage du nombre de reads supportant la fusion [Fusion-name] :

```
for FusionName in "[Fusion-name]" ; do
     echo $FusionName
     samtools view finspector.junction_reads.bam | grep $FusionName |awk
'BEGIN {FS="_"}  {print $2}' | awk '{print $1}' | sort | uniq -c | wc -l
done


for FusionName in "[Fusion-name]" ; do
     echo $FusionName
     samtools view finspector.spanning_reads.bam | grep $FusionName |awk
'BEGIN {FS="_"}  {print $2}' | awk '{print $1}' | sort | uniq -c | wc -l
done
```

Comptage du nombre MBCs dans un fichier BAM :

```
samtools view finspector.junction_reads.bam | awk 'BEGIN {FS="_"} {print
$2}' | awk '{print $1}' | sort | uniq -c | wc -l

samtools view finspector.spanning_reads.bam | awk 'BEGIN {FS="_"} {print
$2}' | awk '{print $1}' | sort | uniq -c | wc -l
```