Développement de nouvelles approches bioinformatiques pour l'analyse omique de l'ADN tumoral libre circulant des lymphomes

Pierre-Julien Viailly

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité ASPECTS MOLECULAIRES ET CELLULAIRES DE LA BIOLOGIE

Préparée au sein de l'Université de Rouen Normandie

Développement de nouvelles approches bioinformatiques pour l'analyse omique de l'ADN tumoral libre circulant des lymphomes

Présentée et soutenue par
PIERRE-JULIEN VIAILLY

Thèse soutenue le 17/12/2021 devant le jury composé de		
MME MARY CALLANAN	PROFESSEUR DES UNIV - PRATICIEN HOSP., CENTRE HOSPITALIER DIJON-BOURGOGNE	Rapporteur du jury
MME MARIE-HÉLÈNE DELFAU LARUE	PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE PARIS-EST CRETEIL	Rapporteur du jury
MME HÉLÈNE DAUCHEL	MAITRE DE CONFERENCES, Université de Rouen Normandie	Membre du jury
M. PIERRE SUJOBERT	PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE LYON 1 CLAUDE BERNARD	Membre du jury
M. BRUNO TESSON	INGENIEUR,	Membre du jury
M. FABRICE JARDIN	PROFESSEUR DES UNIV - PRATICIEN HOSP., Université de Rouen Normandie	Directeur de thèse

Thèse dirigée par FABRICE JARDIN, GENOMICS AND PERSONALIZED MEDICINE IN CANCER AND NEUROLOGICAL DISORDERS

UNIVERSITÉ DE ROUEN NORMANDIE

ED 497 NBISE
Normande de Biologie Intégrative, Santé, Environnement

iRiB

RÉSUMÉ

Français

Développement d'outils bioinformatiques pour l'analyse de l'ADN tumoral libre circulant des lymphomes

En France, le lymphome est le 6e cancer le plus fréquent avec chaque année environ 15 000 nouveaux cas diagnostiqués et près de 4 500 décès. Derrière cette maladie se cache en réalité une très grande hétérogénéité tant sur le plan clinique que phénotypique. Le développement des approches d'immunohistochimie, de cytogénétique et l'avènement récent des séquenceurs de nouvelle génération permettent une caractérisation toujours plus précise de cette maladie via la quantification de plusieurs biomarqueurs à partir de la tumeur. L'intégration de ces différentes sources de données a permis une meilleure classification des lymphomes aujourd'hui scindés en plusieurs dizaines d'entités distinctes.

Le concept de biopsie liquide, qui regroupe un ensemble d'examens réalisés à partir de fluides biologiques tels que le plasma, est devenu un enjeu majeur de ces dernières années. La biopsie liquide, en permettant une détection non invasive des biomarqueurs issus de la tumeur à différents temps de la prise en charge du patient, permet de suivre l'évolution de la maladie et pourrait permettre à plus ou moins moyen terme de proposer aux patients le bon diagnostic, le bon traitement et au bon moment de la maladie via le développement des thérapies ciblées.

Les travaux de ce mémoire visent à présenter les différents développements bioinformatiques menés afin de mieux caractériser les biopsies liquides par séquençage à haut-débit. Différents algorithmes, intégrant ou non des barcodes moléculaires, seront détaillés et associés à des exemples d'application en conditions réelles. Un état de l'art antérieur au développement des nouveaux outils sera présenté et leurs limites seront discutées.

Mots-clés Lymphome, Biopsie Liquide, Séquençage de nouvelle génération

Anglais

Development of new bioinformatics algorithms for the analysis of cell-free DNA for the management of lymphoma

Lymphoma is the sixth most common cancer in France with 15,000 new cases diagnosed and 4,500 deaths each year. Lymphomas are very heterogeneous diseases both in clinical and phenotypic aspects. The development of immunohistochemistry, cytogenetics and more recently next generation technologies gives a more precise picture of this disease by allowing the measure of several biological parameters from the tumor tissue to guide the diagnosis. The integration of these different data sources allows a better classification of lymphomas which are today divided into several dozen distinct entities in the last WHO classification.

The concept of liquid biopsy, which consists in extracting a set of biological features no longer from the tumor but from biological fluids such as plasma, has become a major research axis. Liquid biopsy allows to extract biological parameters from the tumor tissue of origin at different stages of the disease to help with disease monitoring or to adapt treatments in case of relapse for example.

The present work consisted in developing new bioinformatics algorithms to improve the analysis of lymphoma cell-free DNA samples using various high-throughput sequencing technologies. Unique molecular barcodes combined with the use of these new algorithms will be discussed. Each new program will be put into perspective with previously published algorithms.

Keywords Lymphoma, Liquid biopsy, Next-generation sequencing technologies

Laboratoire d'accueil

INSERM 1245 - Equipe 2 - *Génomique des lymphomes et des tumeurs et le développement de marqueurs personnalisés du cancer*

Centre Henri Becquerel,

Rue d'Amiens, 76038 Rouen CEDEX 1

REMERCIEMENTS

A mon directeur de thèse, le Pr. Fabrice JARDIN, pour la confiance qu'il m'a témoignée tout au long de mon cursus universitaire et pour la richesse scientifique des projets portés dans notre unité de recherche qui ont conduit *in fine* à l'écriture de ce manuscrit.

A ma co-directrice de thèse, Hélène DAUCHEL, pour le suivi de mon parcours de jeune bioinformaticien depuis ma licence jusqu'à la soutenance de ce doctorat. Cette épopée a démarré avec FunEVA il y a maintenant quelques années et m'a conduit à découvrir la recherche en bioinformatique. Sans nos nombreuses discussions autour de l'annotation des variants, nul doute que je n'aurai pas écrit ce manuscrit aujourd'hui !

A mon jury de thèse, pour avoir accepté la lourde tâche d'évaluer ce travail malgré les emplois du temps chargés de chacun. Au **Pr Marie-Hélène DELFAU LARUE** et au **Pr Mary CALLANAN** pour nos échanges sur les analyses du ctDNA et pour toute l'effervescence autour du projet FrenchConnect !

Au **Dr Bruno TESSON**, infatigable bioinformaticien qui m'accompagne dans les projets multicentriques portés par le CALYM et le LYSA et dans de nombreux projets de recherche. Merci pour nos discussions toujours aussi enrichissantes !

Au Pr Pierre SUJOBERT et au Pr Jean-Philippe JAIS, pour la qualité constante de nos échanges tout au long de ma thèse au travers de mon Comité de Suivi Individuel et de nos thématiques de recherche communes.

A tous les enseignants et aujourd'hui collègues du master bioinformatique de l'Université de Rouen et plus particulièrement à Thierry LECROQ, Nicolas VERGNE, Caroline BERARD, Laurent MOUCHARD et Arnaud LEFEBVRE.

Au Pr Hervé TILLY, sans qui toute cette aventure n'aurait pas existé. Merci de m'avoir fait confiance pour intégrer le Centre Henri Becquerel.

Au Dr Philippe RUMINY, pour nos débats animés autour de la biologie du lymphome et... de l'infobiologie ! Merci pour tous nos échanges autour de cet ouvrage obscur appelé WHO… et à tout le savoir que tu me transmets avec passion depuis mon arrivée à Centre !

A Sylvain MARESCHAL, qui m'a fait découvrir la bioinformatique de terrain lors de ma première année de master et durant mon alternance. Merci de m'avoir initié aux joies de R !

A Xavier RENAULT, Emmanuel SEUTRE, Christian LEPECQ, Corinne LAROSE et toutes celles et tous ceux qui m'ont transmis la passion des sciences depuis le plus jeune âge.

A Sydney, Fanny, Liana, Vincent S., Vincent C., Elodie, Mathieu, Marie-Delphine, Vinciane et tant d'autres… A tous les étudiants en médecine, pharmacie, sciences ou bioinformatique passés dans mon bureau tout au long de ces années.

A l'ensemble du personnel technique du département de biologie moléculaire, qui font un travail remarquable au quotidien pour la prise en charge des patients, et tout particulièrement à Shirley, Emilie, Laëtitia, Colas, Pauline, Axel, Soizic…

A mes parents, à ma compagne, à mon frère, à ma famille et à mes amis, qui m'ont toujours soutenu personnellement dans tout ce que j'ai souhaité entreprendre.

ABRÉVIATIONS

BQSR	Base Quality Score Recalibration
BWA	Burrows-Wheeler Aligner
CAPP-Seq	Cancer Personalized Profiling by Deep Sequencing
CDR	Complementarity Determining Region
cfDNA	cell-free DNA
CGH	Comparative Genomic Hybridization
CHOP	Cyclophosphamide, Hydroxyadriamycine (doxorubicine), Oncovin (vincristine) et Prednisone
CNV	Copy Number Variation
CNV	Copy Number Variation
COO	Cell-of-Origin
ddPCR	droplet digital PCR
DSCS	Double-Stranded Consensus Sequence
EBV	Epstein-Barr Virus
fdDNA	fetal-derived cfDNA
FFPE	Formalin-Fixed Paraffin-Embedded
FISH	Fluorescent In-Situ Hybridization
GATK	Genome Analysis ToolKit
GCB	Germinal Center B-cell like
GSP	Gene Specific Primer
iDES	Integrated Digital Error Suppression
IF	Impact Factor
Ig	Immunoglobuline
Ig-HTS	Immunoglobulin High-Throughput Sequencing
IHC	ImmunoHistoChimie
indels	insertions/deletions
IPI	International Prognostic Index
ISP	Ion Sphere Particle
LCM	Lymphome à Cellules du Manteau
LPS	Linear Predictor Score
LYSA	LYmphoma Study Association
MALT	Mucosa-Associated Lymphoid Tissue
MZL	Marginal Zone Lymphoma
NGS	Next Generation Sequencing
NMF	Nonnegative Matrix Factorization
OMS	Organisation Mondiale de la Santé
PCNSL	Primary Central Nervous System Lymphoma
PCR	Polymerase Chain Reaction
PMBL	Primary Mediastinal B-cell Lymphoma
qPCR	Quantitative Polymerase Chain Reaction
R-CHOP	CHOP + Rituximab
RNA-seq	(whole) RiboNucleic Acid sequencing
RTMLPA	Reverse-Transcriptase Multiplex Ligation dependent Probe Amplification
SAM	Sequence Alignment Map
SERS	Surface Enhanced Raman Spectroscopy
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SSCS	Single-stranded consensus sequence
ssDNA	single strand DNA
Tam-Seq	Tagged-amplicon deep Sequencing
TEC	Targeted Error Correction
UMI	Unique Molecular Identifier
UP	Universal Primer
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization

TABLE DES MATIÈRES

TABLE DES FIGURES

AVANT PROPOS

Le présent mémoire vise à détailler l'ensemble des travaux réalisés autour du développement de nouvelles approches bioinformatiques pour l'analyse de données de séquençage à haut-débit (NGS) dans des échantillons d'ADN tumoral circulant de patients atteints de lymphome.

Nous reviendrons dans un premier temps sur une description des dernières avancées concernant le diagnostic du lymphome et la nécessité grandissante d'intégrer différentes sources de données moléculaires hétérogènes afin d'en préciser le sous-type. Ce mémoire sera l'occasion de présenter l'évolution de la biologie moléculaire dans les laboratoires de diagnostic depuis les premières techniques de séquençage jusqu'à l'avènement récent des technologies de séquençage à haut-débit. Le fonctionnement des deux principales plateformes de séquençage sera présenté de manière exhaustive depuis les signaux d'acquisition des séquenceurs jusqu'au traitement bioinformatique des données en passant par l'étape cruciale de préparation des librairies. Enfin, le concept de biopsie liquide sera introduit et remis en perspective dans le cadre des lymphomes.

Dans un second temps, nous présenterons les différents développements bioinformatiques ayant conduit à la publication de nouveaux algorithmes pour améliorer le traitement des données de séquençage avec plus de précision dans le contexte des biopsies liquides : LowVarFreq et UMI-VarCall, deux outils dédiés à la détection des variants ; mCNA visant à améliorer la qualité de détection des remaniements du nombre de copies de gène ; et enfin UMI-Gen, un simulateur de données de séquençage intégrant des barcodes moléculaires. Chacun des algorithmes sera remis en perspective vis à vis des autres outils disponibles dans la littérature et détaillé de manière exhaustive. Des exemples d'application sur des échantillons de biopsie liquide seront présentés.

Enfin, ce mémoire sera l'occasion de discuter des perspectives de la biopsie liquide dans la prise en charge des patients atteints de lymphomes dans les années à venir et du chemin qu'il reste à parcourir tant sur le plan technologique que bioinformatique pour y parvenir.

I. INTRODUCTION

Ce chapitre vise à introduire les différents concepts en lien les travaux menés dans le cadre de cette thèse.

Nous présenterons les différents concepts associés aux cancers des cellules sanguines et de leurs précurseurs appelés hémopathies malignes. Nous nous concentrerons en particulier sur les hémopathies impactant des cellules immunitaires appelées lymphocytes, ou globules blancs, pour lesquelles des échantillons ont été analysés. Ces lymphocytes peuvent être à l'origine de l'apparition de cancers appelés lymphomes. Nous nous intéresserons plus particulièrement d'une part à un sous-type de lymphome non-hodgkinien (LNH) que sont les lymphomes B diffus à grandes cellules et d'autre part au lymphome de Hodgkin (LH). Nous introduirons pour chacune de ces entités les principaux outils utilisés au diagnostic. Les analyses morphologiques, immunohistochimiques, génétiques et transcriptomiques sur du matériel biologique provenant de la tumeur des patients sont aujourd'hui autant de sources de données utilisées au diagnostic afin de caractériser au mieux cette maladie très hétérogène.

Ce chapitre sera aussi l'occasion de présenter l'évolution des technologies de séquençage depuis la chimie Sanger et les séquenceurs de première génération jusqu'au développement plus récent des séquenceurs de nouvelle génération. Le fonctionnement des deux principales technologies de séquençage à haut-débit sera détaillé ainsi que l'étape cruciale pré-séquençage de préparation des librairies.

Si la prise en charge des patients reposait jusqu'à présent, sur le plan de la biologie, principalement sur l'analyse des biomarqueurs identifiés dans la tumeur à la suite d'une biopsie, le concept de « biopsie liquide » est un domaine de la recherche qui semble particulièrement prometteur et intéressant dans les années à venir. Il regroupe un ensemble d'examens biologiques réalisés non plus sur une biopsie mais à partir d'un fluide biologique comme un échantillon sanguin ou encore l'urine du patient. Nous détaillerons dans cette introduction les différentes applications possibles de l'analyse de ces biopsies liquides, leurs caractéristiques et les méthodes d'analyse applicables dans le contexte des lymphomes.

A. Hétérogénéité des lymphomes

A.1. Généralités

En France, le lymphome est le 6e cancer le plus fréquent. Il représente 3% de la totalité des cancers avec près de 15 000 nouveaux cas et près de 4 500 décès chaque année. Son incidence est en augmentation du fait notamment du vieillissement de la population et de notre exposition toujours plus importante à des facteurs environnementaux néfastes tels que certains pesticides ou polluants [1].

Les lymphomes sont des proliférations cancéreuses agressives touchant des cellules immunitaires appelées lymphocytes. Dans des conditions physiologiques, les lymphocytes, ou « globules blancs », participent à la lutte contre les infections. Ils proviennent de la maturation de précurseurs dans des organes lymphoïdes tels que la moelle osseuse, la rate ou encore les ganglions lymphatiques retrouvés à différents endroits de notre corps (figure 1). Les lymphocytes, une fois produits et matures, sont libérés dans la circulation sanguine où ils accomplissent leur action effectrice.

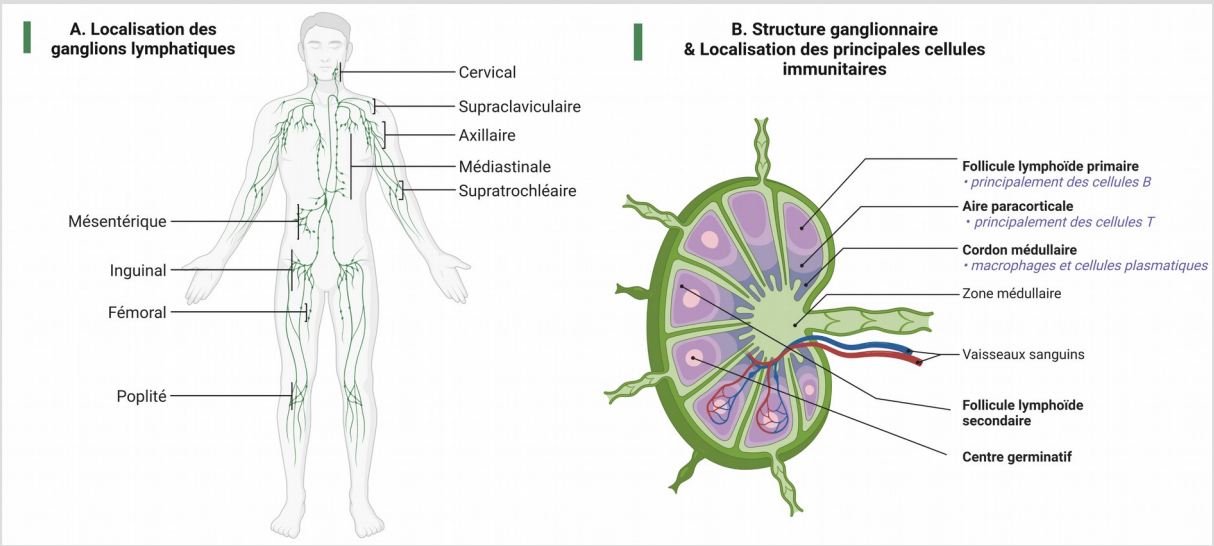


Figure 1: Localisation et structure des ganglions lymphatiques.

La figure représente en (A) les principales localisations des ganglions lymphatiques dans le corps humain et en (B) la structure d'un ganglion lymphatique et la localisation des cellules immunitaires.

Il existe deux grands types de lymphocytes : les lymphocytes B et T. Les lymphocytes B ayant terminés leur cycle de maturation deviennent des B mémoires ou des plasmocytes capables de produire des anticorps en réponse à une infection par des bactéries, à l'exposition

à certaines toxines ou encore pour lutter contre certaines cellules tumorales. Les lymphocytes T jouent quant à eux, par exemple, un rôle dans l'élimination des cellules de l'organisme ayant été infectées par des virus ou ayant échappées à l'action des lymphocytes B.

Les hémopathies malignes regroupent un ensemble très hétérogène de cancers des cellules immunitaires et de leurs précurseurs. Parmi cet ensemble, on distingue les leucémies, les syndromes myélodysplasiques et les lymphomes (figure 2). Ces trois grandes catégories d'hémopathies malignes se divisent elles même en un très grand nombre de maladies très distinctes sur le plan clinique en fonction du stade de différenciation des cellules.

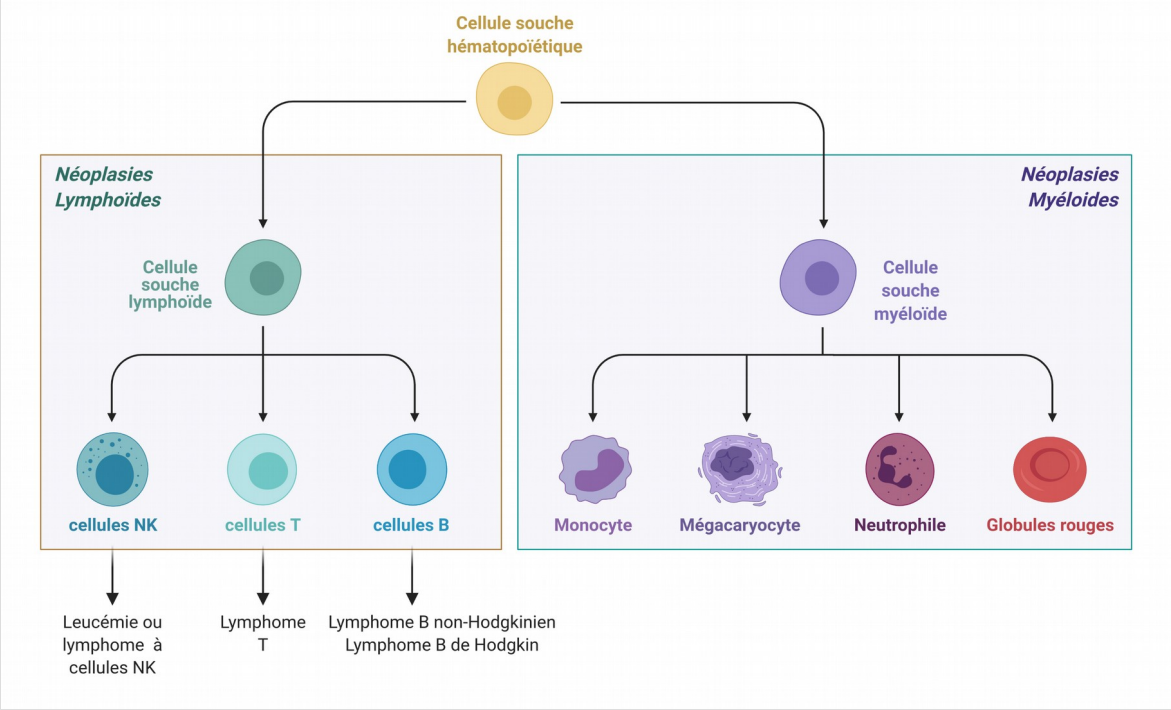


Figure 2: Différenciation des cellules immunitaires

Ce schéma représente les stades de différentiation des cellules immunitaires depuis la cellule souche hématopoïétique. Les cellules de la lignée lymphoïde, en fonction de leur différentiation et de leurs anomalies génétiques, peuvent conduire à l'apparition de lymphomes spécifiques. Les myélomes ont pour origine des descendants de la cellule souche myéloïde.

Classification histologique à partir des biopsies tissulaires

La classification des lymphomes est très complexe. Dans les années 1990, celle-ci reposait principalement sur l'étude histologique des tissus tumoraux à partir de biopsies ganglionnaires et entraînait des difficultés aux pathologistes et aux cliniciens pour poser un diagnostic précis tant cette maladie est très hétérogène. En 1993, une première classification est proposée par un groupe d'hématopathologistes (the International Lymphoma Study Group) à Berlin [2]. Ce

groupe de travail avait pour objectif d'harmoniser les pratiques des différents pathologistes ayant souvent recours à de nombreux schémas de classification différents afin de poser un diagnostic.

Des critères morphologiques et immunophénotypiques[1] semblent suffisant pour classer la plupart des grandes familles de lymphomes. Néanmoins, il n'existe pas de marquage en immunohistochimie (IHC) spécifique d'une entité et les pathologistes ont recours à des combinaisons de critères morphologiques et immunohistochimiques afin de poser un diagnostic. Il existe aussi, au sein d'un même sous-type de lymphome, des différences de marquages en IHC en fonction des biopsies et des anticorps ce qui complique encore plus la reproductibilité de ces classifications. Le recours aux techniques d'immunohistochimie nécessite enfin des prélèvements de qualité conservant l'intégrité des tissus afin d'apprécier à la fois l'intensité des immunomarquages mais aussi de la structure des ganglions (figure 1.B).

Marqueurs moléculaires au diagnostic

Les marqueurs génétiques jouent un rôle de plus en plus important dans la classification des lymphomes. Des études génétiques, notamment sur les réarrangements des gènes des immunoglobulines (Ig) des lymphocytes B, sont des outils utilisés au diagnostic afin de déterminer une clonalité.

Les anticorps, ou Ig, sont composés de chaînes de protéines lourdes et légères. Chaque type d'immunoglobuline contient une partie constante (C) et une partie variable (V) entrecoupée ou non par une région dite de diversité (D) et une région de jonction (J). Les gènes codant ces chaînes légères ou lourdes se situent à différents endroits du génome. La chaîne lourde (μ, δ, γ1, γ2, γ3, γ4, α1, α2, ε) est codée par des gènes localisés sur le chromosome 14, la chaîne légère kappa (κ) par un locus sur le chromosome 2 et la chaîne légère lambda (λ) par des gènes localisés sur le chromosome 22. Ces segments sont recombinés dans la moelle osseuse lors du premier processus de recombinaison somatique (ou recombinaison VDJ).

La diversité des gènes des régions V, D et J permet de créer une grande variété d'Ig qui sont nécessaires à la reconnaissance de l'immense variété d'antigènes étrangers (figure 3). Le locus de la chaîne lourde contient par exemple une cinquantaine de gènes V, une trentaine de gènes D et six gènes de jonction J. Cette diversité est aussi présente dans les gènes codants

1 immunophénotypique: l'immunophénotypage est un test basé sur l'utilisation d'anticorps marqués (immunohistochimie, IHC) dirigés contre certains marqueurs exprimés à la surface des cellules immunitaires. Il est employé dans le but de recenser des types cellulaires différents.

pour les chaînes légères même si ces chaînes ont la particularité de ne pas avoir de région D. Un deuxième processus qui a lieu dans les organes lymphoïdes secondaires appelée processus d'hypermutation somatique conduit à l'introduction de mutations sur les segments variables des chaînes légères et lourdes augmentant ainsi considérablement le nombre d'anticorps différents possibles (N-diversité).

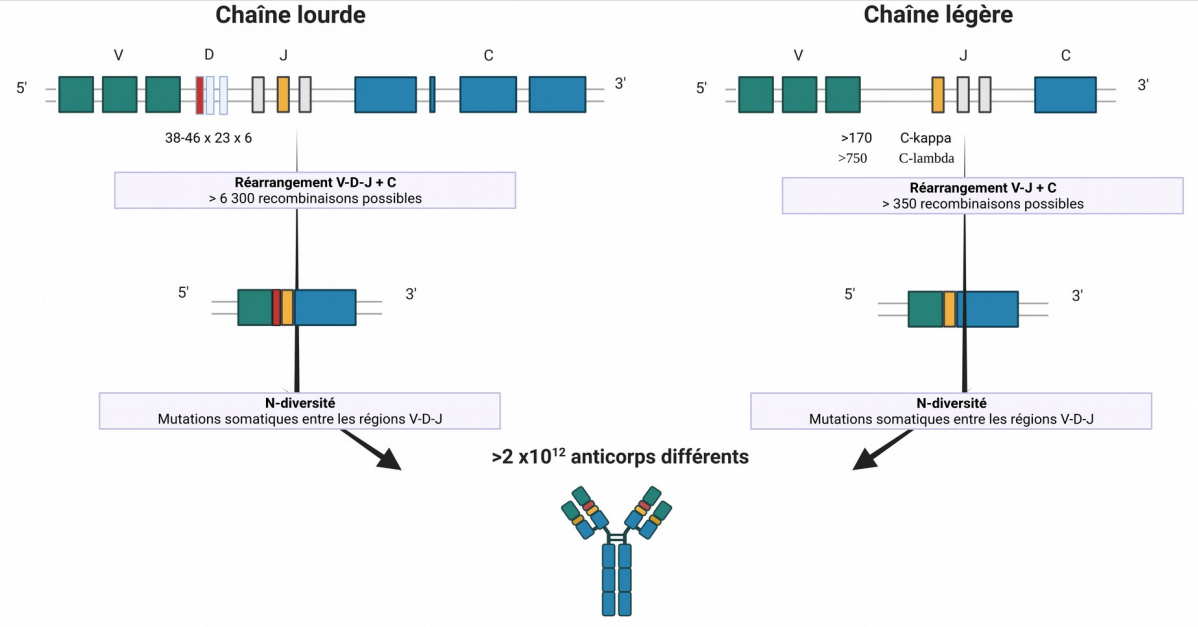


Figure 3: Recombinaison V-D-J des immunoglobulines

La chaîne lourde est composée de régions variables (V), de diversité (D), de jonction (J) et d'une constante (C). Le réarrangement de ces loci va rapprocher ces différentes régions afin de conduire à la formation d'un segment VDJ fonctionnel codant pour la chaîne lourde de l'anticorps. Le même processus conduit à la création de la chaîne légère de l'anticorps.

L'analyse des réarrangements VDJ permet ainsi de déterminer si une prolifération de lymphocytes dans un ganglion provient ou non d'un seul clone, c'est à dire d'un ensemble de cellules partageant le même réarrangement, ou d'une population très variée de lymphocytes et donc d'une réaction immunitaire dans des conditions physiologiques. Ces réarrangements sont une signature du clone tumoral. L'étude européenne collaborative BIOMED-2 a mis au point un protocole de PCR multiplexe pour standardiser la détection de ces réarrangements au diagnostic ainsi que certaines translocations chromosomiques comme la t(14;18)[2] et t(11;14) [3].

2 t(14;18) : nomenclature utilisée en cytogénétique afin de décrire, parallèlement à la formule chromosomique du caryotype, une translocation entre le chromosome 14 et le chromosome 18 dans cet exemple.

Le développement des approches de séquençage, détaillées dans la section I.B.3, ont permis de s'intéresser aux principales variations génétiques acquises dans ces pathologies. On distingue les variations génétiques germinales, appelées polymorphismes (SNP, Single Nucleotide Polymorphism), qui sont bénignes et présentes dans toutes les cellules d'un individu, et les anomalies génétiques acquises dans un sous-ensemble de cellules tumorales, appelées mutations somatiques. Ce sont ces dernières qui sont généralement étudiées et recherchées dans le cadre d'un cancer. Citons par exemple le cas de la mutation L265P du gène *MYD88* [4] permettant de discriminer entre autre au sein des LDGCB le sous-type ABC (Activated B-cell like), où la mutation est retrouvée fréquemment, du sous-type GCB (Germinal Center B-cell like) [5], ou encore la mutation V600E sur le gène *BRAF* [6], [7] retrouvée dans les leucémies à tricholeucocytes (HCL, hairy cell leukemia).

Les techniques à haut-débit de caractérisation de l'ADN telles que la CGH (Comparative Genomic Hybridization), les puces à ADN (SNP-array) ou encore le séquençage à haut-débit ont conjointement permis de mettre en évidence la présence de variations de nombre de copie de gènes (Copy Number Variation of genes, CNV). Des anomalies génétiques acquises dans les génomes tumoraux peuvent entraîner le gain, l'amplification ou la délétion de segments chromosomiques plus ou moins larges conduisant généralement à une augmentation ou à une baisse de l'expression des gènes de ces segments. Certains CNV sembleraient jouer un rôle pronostic comme les délétions des gènes *CDKN2A* et *CDKN2B [8]*.

L'apport des analyses transcriptomiques, c'est à dire de la mesure de l'expression des gènes du génome (ARNm), a permis d'affiner encore un peu plus certaines classifications dans des pathologies pour lesquelles la morphologie n'est pas suffisante. Ces signatures transcriptomiques peuvent permettre de distinguer différentes entités de lymphomes [9]–[11], d'identifier des marqueurs pronostiques comme l'impact de la co-expression des gènes *MYC* et *BCL2* sur la survie dans les lymphomes B [9] ou enfin de créer des modèles statistiques afin de prédire les risques de rechute [12].

Ces nouvelles sources de données complémentaires de l'anatomopathologie, conduisent de plus en plus au développement d'approches intégratives qui mettent en lumière de nouvelles entités. Elles ont conduit à l'adoption d'une nouvelle classification des hémopathies malignes par l'OMS (WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues) dont la dernière édition révisée date de 2017 [13]. Cette dernière laisse une place de plus en plus importante aux analyses à haut-débit, et notamment de la transcriptomique, dans les critères de classification des hémopathies malignes.

Nous nous intéresserons dans les prochaines sections plus particulièrement à trois familles de lymphomes B : le lymphome diffus à grandes cellules B (LDGCB), le lymphome primitif du médiastin (PMBL) et le lymphome de Hodgkin (LH).

A.2. Lymphome diffus à grandes cellules B

Le LDGCB est un sous-type de lymphome faisant parti de la famille des « lymphomes non hodgkiniens » (LNH). Derrière ce terme de LNH se cache en réalité une grande diversité de lymphomes sur le plan clinique, phénotypique et anatomopathologique avec une trentaine d'entités différentes. Ces entités dépendent principalement du type de lymphocytes à l'origine de la cellule tumorale. On retrouvera ainsi une majorité de lymphomes B (85 % des LNH) et plus rarement des lymphomes T.

Certaines de ces entités sont des lymphomes dits indolents, comme le lymphome folliculaire (FL), le lymphome du tissu lymphoïde associé aux muqueuses (MALT) ou encore le lymphome de la zone marginale (MZL). D'autres seront qualifiés de lymphomes agressifs. C'est le cas notamment de l'entité la plus fréquente qu'est le LDGCB décrite dans la suite de cette section.

Hétérogénéité et classification des lymphomes diffus à grandes cellules B

Le LDGCB est le sous-type de LNH le plus courant et représente à lui seul près de 30 % des cas de LNH diagnostiqués. Ce type de lymphome agressif croît rapidement dans les nœuds lymphoïdes et touche souvent la rate, le foie, la moelle osseuse et d'autres organes à des stades plus avancés. Son développement débute habituellement dans les ganglions du cou ou de l'abdomen. Les LDGCB apparaissent la plupart du temps *de novo*, c'est à dire sans antécédent connu de lymphome. Néanmoins, ils peuvent aussi découler d'une transformation d'un lymphome moins agressif comme par exemple le LF ou le MZL [11]–[12].

Des études morphologiques, moléculaires, biologiques et cliniques ont subdivisé les LDGCB en différentes entités. Néanmoins, il existe un certain nombre de cas pour lesquels la classification des biopsies représente un réel challenge tant l'hétérogénéité biologique est importante dans cette maladie (figure 4).

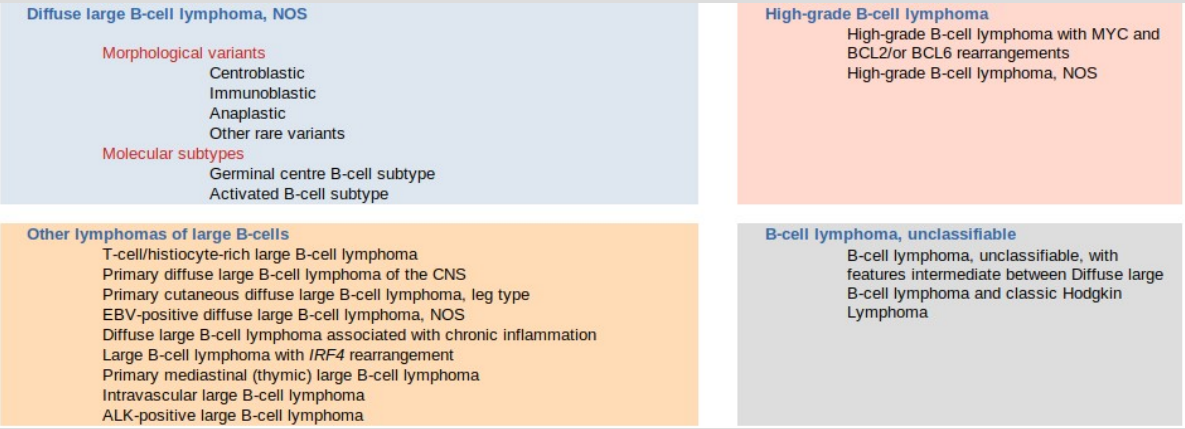


Figure 4: Classification des LDGCB selon la classification OMS dans sa version révisée de 2017 [13].

L'arrivée du traitement par immunochimiothérapie R-CHOP, basée sur une combinaison de molécules (rituximab, cyclophosphamide, doxorubicin, vincristine et prednisone), a constitué un réel progrès dans la prise en charge thérapeutique des patients atteints de LDGCB avec des taux de survie à 5 ans autour de 65 % [16]–[18]. Cette approche de traitement uniforme dans cette maladie hétérogène, pour les patients en rechute ou réfractaires en première ligne de R-CHOP, nécessite aujourd'hui d'établir des classifications plus précises afin d'identifier d'éventuelles nouvelles thérapeutiques ciblées dans un contexte de médecine personnalisée.

Classification moléculaire des lymphomes diffus à grandes cellules B

Des études de transcriptomique à haut débit par puces ont démontré que les LDGCB sont constitués d'un groupe très hétérogène de LNH qui avait été initialement regroupé sur des critères morphologiques, immunophénotypiques et d'agressivité en une seule et unique entité [13]. Dans les années 1990, elles ont permis de mettre en évidence que les LDGCB étaient en réalité constitués de deux sous-types très différents sur le plan clinique, biologique et moléculaire [19], [20].

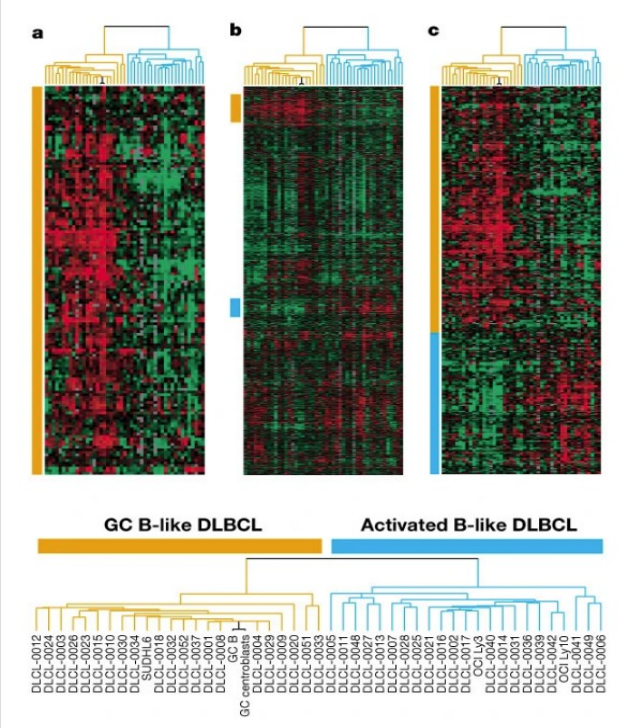


Figure 5: Classification hiérarchique GCB/ABC sur puces Lymphochip. (Figure extraite de Alizadeh et al, Nature 2000 [22])

Cette figure présente les groupes de gènes différentiellement exprimés entre sous-types GCB (en marron) et ABC (en bleu) dans la cohorte de LDGCB.

La première étude est basée sur l'utilisation d'une puce transcriptomique à façon baptisée « lymphochip », technologie déjà capable de mesurer l'expression de plus de 17 000 ARN messagers (ARNm) [21]. Appliquées sur une centaine d'échantillons, ces puces ont permis la toute première comparaison des profils d'expression génique des principales hémopathies B (LDGCB, FL, LCM, LLM) [22]. La classification hiérarchique appliquée à ces profils d'expression a permis de distinguer deux sous-ensembles homogènes de patients surexprimant ou non des ARNm du centre germinatif du ganglion (figure 5). La comparaison au profil d'expression de lymphocytes B normaux a ensuite conduit à préciser ces deux sous-groupes de LDGCB : ceux ayant un profil similaire aux « lymphocytes issus du centre germinatif » (Germinal Center B-Cell like, GCB) et ceux dont le profil d'expression se rapproche des « lymphocytes périphériques activés » (Activated B-cell like, ABC). Cette distinction entre les sous-types ABC et GCB s'est avérée pronostique dans cette série de 40 patients.

En 2002, cette même observation est réalisée sur une cohorte de 240 patients par A. Rosenwald [23] à partir de la mesure de l'expression des 100 gènes les plus discriminants de

la série de lymphochips de A. Alizadeh : le sous groupe ABC est de moins bon pronostic que le sous groupe GCB avec respectivement 35 % et 60 % de survie des patients traités par CHOP, ancêtre du R-CHOP, à 5 ans. Lors de cette étude, un troisième groupe de patients de « type 3 », c'est à dire non-GCB et non-ABC selon la signature d'expression, est apparu.

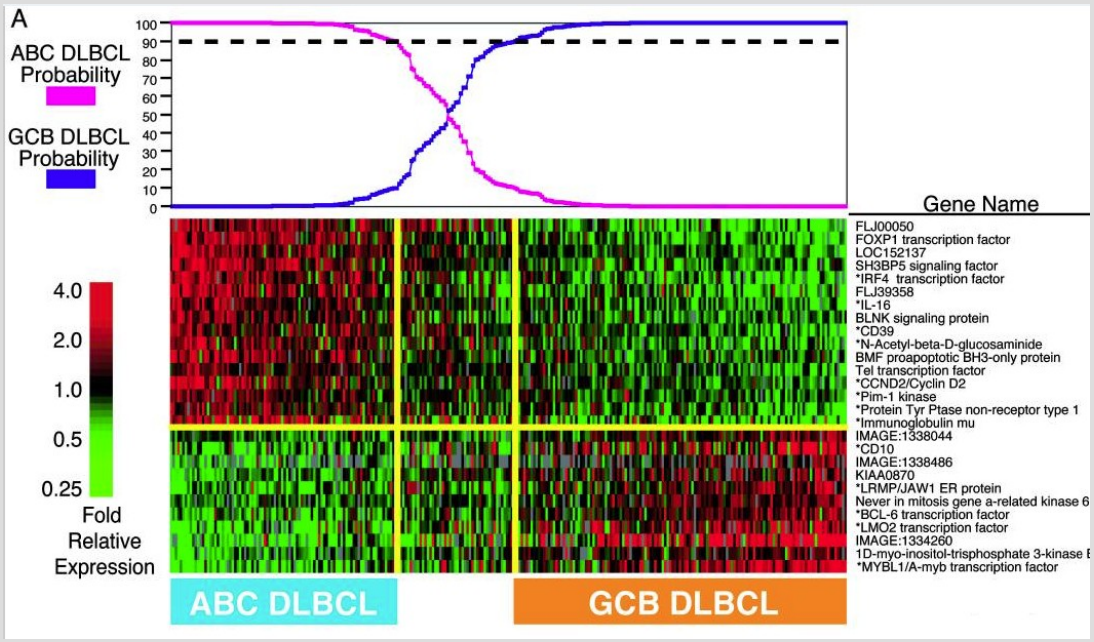


Figure 6: Modèle de classification GCB/ABC de l'étude de Wright et al. [24]

Cette figure représente la probabilité d'appartenance au groupe ABC ou GCB en fonction du niveau de l'expression de la signature GCB/ABC dans une cohorte de LDGCB. Un groupe de patients non classés apparaît dans ce modèle au centre de la *heatmap*. La liste des gènes est visible à droite de la figure.

Devant la difficulté à accéder à des puces Lymphochip, l'équipe de G. Wright cherche l'année suivante à reproduire la classification GCB/ABC en se basant cette fois-ci sur l'utilisation de puces pan-transcriptomiques commercialisées par la société Affymetrix [24]. Devant l'instabilité inhérente aux méthodes de classifications hiérarchiques, son équipe développe une méthode probabiliste de classification des échantillons (figure 6). C'est la toute première fois qu'un modèle de classification, basée sur l'utilisation d'un score linéaire de prédiction (Linear predictor score, LPS), détermine la probabilité d'appartenance à l'une des deux entités de LDGCB (GCB ou ABC) à partir de la mesure de l'expression de gènes. Ce modèle confirmera une nouvelle fois l'impact pronostique de ces deux sous-types moléculaires sous CHOP et sera considéré dans les années suivantes comme le *gold standard* pour la classification des LDGCB. L'équipe de G. Lenz confirmera en 2008 l'impact

pronostique de la signature GCB/ABC dans une cohorte de patients traités cette fois-ci par R-CHOP à la fois sur la survie globale et sur la survie sans progression à 5 ans [25].

Classification immunohistochimique des lymphomes diffus à grandes cellules B

Les classifications des LDGCB par puces transcriptomiques n'étaient pas facilement transposables en routine car elles nécessitaient un savoir-faire non négligeable pour la manipulation des puces et un appareillage spécifique. De plus, le rendu du diagnostic dans les hémopathies malignes repose toujours sur l'intervention des services d'anatomopathologie et c'est donc tout naturellement que ces signatures d'expression ont été transformées en des algorithmes de classification plus simples et basés sur une technique bien maîtrisée par les anatomopathologistes : l'IHC. L'IHC est une technique permettant de mesurer l'intensité de marquage d'un élément cellulaire par un anticorps comme un récepteur exprimé à la surface d'une cellule. Cette technique, couplée à la visualisation au microscope des tissus, permet une identification des cellules de la coupe mais aussi de mesurer certains marqueurs spécifiques à une pathologie comme le CD15 et le CD20 exprimés par les cellules de Reed-Sternberg dans la maladie de Hodgkin.

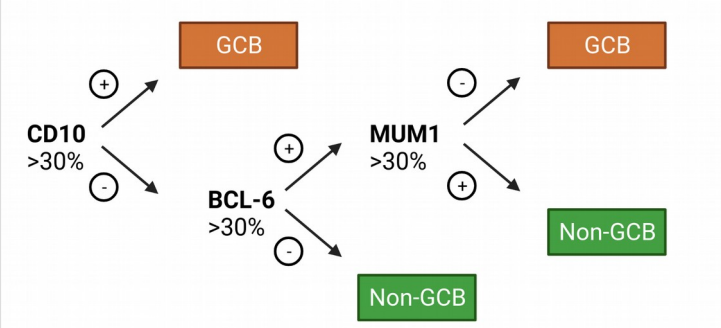


Figure 7: Algorithme immunohistochimique de Hans.

Cet algorithme repose sur l'immunomarquage de 3 marqueurs (CD10, BCL6 et MUM1) afin de classifier les patients en GCB/nonGCB. Les critères de décision reposent sur un pourcentage de cellules marquées par l'anticorps.

Christine P.Hans et al proposent une première retranscription des signatures d'expression génique en un algorithme immunohistochimique. Cet algorithme de Hans vise à identifier via l'immunomarquage de 3 marqueurs (CD10, BCL6 et MUM1) le sous-type GCB. Les autres entités (ABC et non-classés/type 3) sont regroupés dans un même sous-groupe non-GCB (figure 7).

D’autres algorithmes immunohistochimiques ont vu le jour comme l’algorithme de Hans modifié, de Nyman, de Muris, de Choi, de Choi modifié, de Tally, de Natkunam ou encore de Visco-Young. En 2013, une étude comparative de tous ces algorithmes démontre une concordance très pauvre avec seulement 4% des tumeurs classées GCB et 21 % comme non-GCB/ABC par l’ensemble des algorithmes (figure 8) [26].

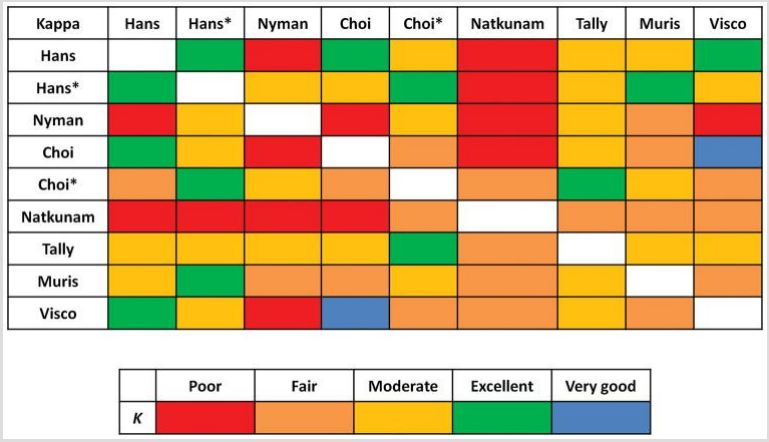


Figure 8: Concordance entre les algorithmes immunohistochimiques pour la classification GCB/non-GCB. Figure extraite de l’étude de R. Coutinho et al [26].

Cette figure représente la concordance de la classification GCB/non GCB selon plusieurs algorithmes immunohistochimiques.

La classification de Hans reste aujourd’hui la référence en terme de classification immunohistochimique [10]. Elle est cependant concurrencée par des approches transcriptomiques à moyen-débit telles que la RT-PCR dépendante de ligation (ligation-dependent RT-PCR, LD-RTPCR) [9], [27] permettant la mesure simultanée de l’expression d’une vingtaine de gènes sur un séquenceur capillaire Sanger, la signature Lymph2Cx proposée sur la plateforme NanoString [28] ou encore la publication récente de modèles d’intelligence artificielle couplés à la mesure en LD-RTPCR de plus de 200 gènes [8] sur des séquenceurs de nouvelle génération (NGS).

Entre 10 et 15 % des LDGCB restent non-classés [22]–[25], [29] avec ces approches transcriptomiques. Cependant, elles ont pour avantage d’offrir une classification robuste et reproductible entre différentes plateformes, là où l’algorithme de Hans nécessite une véritable expertise d’anatomopathologistes spécialistes pour l’appréciation des immunomarquages. Aujourd’hui, l’OMS recommande en priorité d’utiliser ces techniques transcriptomiques pour classifier les LDGCB, ou à défaut si ces techniques ne sont pas disponibles, de le faire par immunohistochimie [13].

La fréquence relative des sous-types GCB et ABC dépend principalement de la localisation géographique des cohortes analysées, de leur âge médian et de la méthode de classification utilisée mais on compte globalement 60 % de LDGCB GCB et 40 % de LDGCB ABC [30]. Déterminer ces sous-types au diagnostic est devenu aujourd'hui un enjeu majeur dans un certain nombre d'essais cliniques. En effet, des données préliminaires suggéreraient que seuls les patients de sous-type ABC bénéficieraient d'un avantage à l'ajout de bortezomib, de lenalidomide ou d'ibrutinib au traitement par R-CHOP [31]–[37]. Démontrer l'efficacité de ces thérapies ciblées passera nécessairement par la classification la plus rigoureuse possible des patients dans les essais cliniques avec des méthodes transposables au diagnostic.

Apport du séquençage à haut-débit dans la classification des lymphomes diffus à grandes cellules B

Cette section vise à décrire l'apport des données de séquençage pour la caractérisation des LBDGCB. Une description plus complète du fonctionnement des technologies de séquençage de première et de deuxième générations sera présentée dans le chapitre I.B.

Le séquençage d'exomes (Whole Exome Sequencing, WES), c'est à dire de l'ensemble des gènes, et de génomes complets (Whole Genome Sequencing, WGS) ont permis d'explorer les profils mutationnels dans les hémopathies malignes [5], [25]–[27].

La série de Morin RD et al publiée dès 2010 dans Nature rapporte déjà le profil mutationnel de près de 31 LDGCB [40] et entre autre la découverte de mutations récurrentes du gène *EZH2*. En 2011, une nouvelle étude de L. Pasqualucci et al enrichit cette première description avec la découverte des délétions des gènes *CREBBP* et *EP300* [41] dans une cohorte plus de 100 DLBCL analysés par puces. En 2011, L. Staudt et al décrivent pour la toute première fois la mutation L265P du gène *MYD88* à partir du séquençage en RNA-seq de quatre lignées cellulaires de sous-type ABC et du séquençage des parties codantes de *MYD88* sur une cohorte de validation de 380 LDGCB [42].

A

Characteristic	Frequency		
	ABC DLBCL	GCB DLBCL	PMBL
Rearrangements			
BCL2	<5%	40%	0%
BCL6	25–30%	15%	0%
MYC, single hit	5–8%	5–8%	0%
CD274/*PDCD1LG2* (also called *PDL1*/*2*)	Rare	Rare	20%
CIITA	Rare	Rare	38%
TBL1XR1	0%	5%	0%
Copy-number aberrations			
1p36.32 deletion (*TNFRSF14*)	10%	30%	Rare
2p16 gain/amplification (*REL*)	Rare	30%	60–75%
3q27 gain/amplification	45%	15–20%	Rare
6q21 deletion (*PRDM1*)	45%	25%	n/a
9p21 deletion (*CDKN2A*)	40%	20%	Rare
9p24.1 gains/amplification (*CD274*/*PDCD1LG2*)	Uncommon	Uncommon	60–75%
18q21.3 gain/amplification (*BCL2*)	55%	15%	Rare
Recurrent mutations[a]			
EZH2	Rare	20–25%	n/a
GNA13	Rare	25%	n/a
KMT2D (also called *MLL2*)	35%	40%[f]	n/a
TP53	25%	20%	n/a
MEF2B	5%	15–20%	n/a
SGK1	5–10%	15–20%	n/a
CREBBP	10%	30%	n/a
TNFRSF14	Rare	30%	n/a
SOCS1	Uncommon	10–15%	40%
PTPN1	n/a	n/a	20%
STAT6	Rare	5%	35%
CARD11	10–15%	10–15%	n/a
CD79B	20–25%	Uncommon	n/a
MYD88	35%	Uncommon	n/a
PRDM1	15%	Rare	n/a
B2M	15–20%	20–25%	n/a
CD58	10%	10%	n/a

B

	ABC *n* = 81	GCB *n* = 83	PMBL *n* = 18	other *n* = 33	FDR
STAT6	0%	14%	72%	6%	**6.8e−14**
XPO1	1%	1%	39%	3%	**4.8e−10**
SOCS1	6%	16%	56%	12%	**2.5e−05**
BCL2	1%	24%	0%	3%	**2.5e−05**
CIITA	12%	10%	56%	9%	**3.1e−05**
TNFAIP3	15%	11%	61%	15%	**3.1e−05**
CD79B	25%	2%	0%	3%	**3.2e−05**
PIM1	33%	8%	0%	6%	**3.3e−05**
GNA13	9%	12%	50%	12%	**2.8e−04**
CD58	6%	10%	39%	6%	**1.2e−03**
CREBBP	6%	31%	11%	24%	**1.2e−03**
B2M	9%	18%	50%	24%	**1.2e−03**
EZH2	0%	18%	6%	9%	**1.8e−03**
TNFRSF14	2%	17%	0%	24%	**1.8e−03**
MFHAS1	1%	10%	28%	9%	**3.9e−03**
MYD88	28%	10%	0%	15%	**4.7e−03**
ITPKB	9%	16%	39%	9%	**1.4e−02**
PRDM1	16%	6%	0%	3%	5.1e−02
NOTCH2	2%	10%	0%	15%	7.6e−02
IRF4	14%	5%	11%	0%	8.7e−02
MEF2B	12%	23%	11%	6%	1.4e−01
BRAF	0%	0%	0%	3%	2.1e−01
FOXO1	4%	12%	6%	3%	2.1e−01
KMT2D	41%	46%	17%	42%	2.2e−01
CARD11	14%	7%	0%	9%	3.5e−01
NOTCH1	7%	1%	6%	6%	3.7e−01
CD79A	0%	2%	0%	0%	4.5e−01
TP53	19%	16%	11%	6%	4.6e−01
CDKN2B	0%	1%	0%	3%	5.4e−01
ID3	5%	2%	6%	9%	5.5e−01
MYC	5%	10%	6%	3%	5.5e−01
CDKN2A	2%	1%	0%	0%	7.4e−01
TCF3	1%	2%	0%	0%	7.4e−01
EP300	15%	14%	17%	18%	9.6e−01

Figure 9: Fréquence des principales anomalies génétiques dans les LDGCB à partir de la classification de l'OMS (A) [13] et de la série de 215 LDGCB de S Dubois et al [50].

L'accessibilité grandissante des technologies de séquençage à haut-débit [41], [43]–[47] a permis de consolider ces connaissances sur les profils mutationnels et la dernière actualisation de la classification OMS rapporte la fréquence des principales anomalies (figure 9.A). On

estime avoir décrit aujourd'hui la plupart des mutations récurrentes retrouvées dans les LDGCB. En effet, plus de 1300 exomes, génomes ou transcriptomes de LDGCB sont publiquement accessibles et directement interrogeables en exploratoire sur des banques de données telles que cBioportal [48], [49].

Des études de séquençage de plus grandes cohortes de patients ont été conduites par notre équipe à partir d'un panel d'une trentaine de gènes appelé LymphoPanel dans le but de préciser la fréquence et la répartition des altérations en fonction des différents sous-types de LDGCB [5], [50] (figure 9.B.). Ce panel couvre les principales régions génomiques d'intérêt extraites des séries d'exomes et des données de la littérature. On constate que pour bon nombre de gènes la fréquence des altérations dépend du sous-type de LDGCB. Par exemple, les mutations des gènes *EZH2* et *GNA13* sont retrouvées presque exclusivement dans le sous-type GCB, tandis que les gènes *CARD11*, *MYD88* et *CD79B* sont caractéristiques du sous-type ABC [5], [40], [43], [50], [51]. A l'inverse, d'autres marqueurs sont retrouvés mutés à la fois dans le sous-type ABC et dans le sous-type GCB comme *TP53*.

On retrouve aussi par CGH et analyses WGS/WES des variations de nombre de copies de segments chromosomiques (CNV) qui sont elles aussi inégalement distribuées entre GCB/ABC [52], [53]. Les LDGCB GCB portent fréquemment des gains ou des amplifications des loci 2p16 et 8q24 et des délétions des loci 1p36 (*TNFRSF14*) et 10q23. Dans le sous-type ABC, on retrouve des gains des loci 3q27, 11q23 et 18q21 et des délétions des régions et 9p21 [52]–[59]. Certaines de ces délétions emportent des gènes d'intérêt comme *CDKN2A/B* ou *TP53* et sembleraient avoir un impact pronostique défavorable [8]. On retrouve dans près de 30 % des cas de LDGCB des translocations dans la région 3q27 impliquant le gène *BCL6* [60]–[64]. Ces réarrangements chromosomiques seraient plus fréquents dans le sous-type ABC [30], [62], [65]. On retrouve aussi des translocations t(14;18) impliquant le gène *BCL2* dans 20 à 30 % des LDGCB et des réarrangements du gène *MYC* qui sont présents chez 8 à 14 % des patients. La cooccurrence des translocations de *MYC* et de *BCL2*/*BLC6* constituent dans la dernière révision de la classification OMS une nouvelle entité à part de LDGCB de mauvais pronostic dits « double-hit » ou « triple-hit » [66], [67].

Classification intégrative des lymphomes diffus à grandes cellules B

Toutes ces nouvelles données conduisent aujourd'hui au développement d'approches intégratives afin d'affiner les classifications existantes.

Nous avons vu dans les sections précédentes que le développement des analyses transcriptomiques a permis de mettre en évidence la présence de différents sous-types de LDGCB : ABC, GCB et 10 à 15 % de non-classés. Si cette dichotomie ABC/GCB est aujourd'hui bien établie et n'est plus contestée sur le plan biologique, elle ne se traduit pas nécessairement par des résultats très probants sur le plan clinique dans des cohortes de patients traités sous R-CHOP [68]–[70].

Durant les dix dernières années, beaucoup d'études se sont intéressées à caractériser une seule source d'anomalies soit par NGS pour l'analyse des mutations [5], [44], [46], [50], [55], [71], [72] ou soit par CGH pour détecter les variations de nombre de copies de gènes (CNV) [52], [53]. Les données de séquençage de nouvelle génération et de variants structuraux ont conduit à une caractérisation détaillée des profils génétiques des LDGCB et aux premières études intégratives visant à les intégrer conjointement avec la classification GCB/ABC [73]–[75].

Une première étude de A. Reddy et al propose une classification intégrative basée sur des données d'exomes et de transcriptomes de 1001 LDGCB afin de construire un modèle prédisant la survie en fonction des anomalies détectées et de la surexpression d'un certain nombre de gènes [75]. Ce modèle est comparé à la classification cell-of-origin (COO, GCB/ABC), à l'IPI[3] (International Prognostic Index) et la co-expression des gènes *MYC* et *BCL2*. En appliquant une méthode de classification par propagation d'affinité (affinity propagation clustering, APC) [76], les auteurs ont mis en évidence 31 clusters non chevauchants regroupant des marqueurs génétiques corrélés. Beaucoup de ces clusters sont liés à la classification COO, c'est à dire à des groupes de gènes différentiellement exprimés ou mutés entre les sous-types *GCB* et *ABC*. Néanmoins, deux nouvelles classes de gènes ont émergé dans cette étude. La première classe capture des groupes de gènes liés aux cellules immunitaires présentes dans l'environnement (lymphocytes T, cellules myéloïdes et NK) et aux cellules stromales. La seconde emporte un ensemble de gènes liés aux processus d'oncogenèse comme la prolifération, la transcription, la traduction, la réplication ou encore au contrôle du cycle cellulaire. Le modèle supervisé de prédiction de la survie a conduit les auteurs à stratifier les patients en un groupe à haut risque et un autre groupe à bas risque de rechute (figure 10). Il surclasse les index pronostics habituellement utilisés tels que l'IPI, la

3 L'Index Pronostique International (IPI) est un score clinique reposant sur l'âge, le stade de la maladie (Ann Arbor), le dosage sanguin de la lactate déshydrogénase (LDH), l'indice de performance OMS évaluant les capacités physiques des patients et enfin le nombre de sites extra-ganglionnaires impactés par la maladie. Ce score IPI est pronostic et est le score utilisé à l'heure actuelle au diagnostic par les cliniciens.

COO ou la co-expression des facteurs *MYC* et *BCL2*. Le modèle proposé dans cette étude reste cependant très contesté puisque certains paramètres du modèle influençant le pronostic, comme les mutations des gènes *ZFAT* et *NF1*, n'ont jamais été retrouvés dans les séries ultérieures.

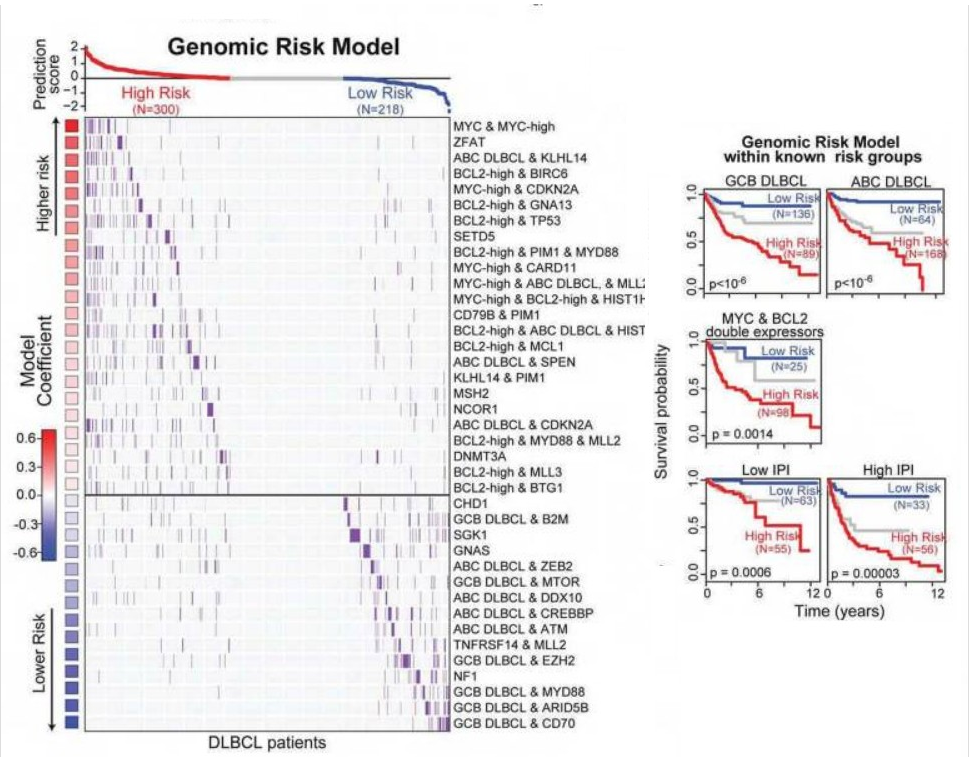


Figure 10: Modèle de prédiction sur la survie dans l'étude de A. Reddy et al.

La première partie de la figure montre les différents facteurs contribuant à la construction du modèle de prédiction pour chacun des groupes. La partie droite met en évidence les courbes de survie des deux groupes au sein de facteurs pronostics déjà identifiés comme l'IPI, la COO et la co-expression de *MYC* et de *BCL2*.

Une deuxième étude conduite par R. Schmitz et al en 2018 vise à intégrer différentes sources de données dans une cohorte de 572 LDGCB [74]. 556 cas ont été séquencés en WES et/ou en séquençage ciblé sur un panel de 530 gènes fréquemment mutés dans les LDGCB. Des profils de CGH ont été obtenus sur 560 cas afin d'identifier les remaniements de nombre de copies et la classification COO a été réalisée à partir de données de RNA-seq sur 562 cas.

L'intégration de toutes ces données a conduit à identifier 4 groupes génétiques, appelés *clusters*, de LDGCB : le cluster « MCD » basé sur la cooccurrence des mutations des gènes *MYD88* (L265P) et *CD79B*, le cluster « BN2 » basé sur les fusions du gènes *BCL6* et les

mutations de *NOTCH2*, le cluster « N1 » basé sur les mutations du gènes *NOTCH1* et le groupe « EZB » basé sur les mutations du gènes *EZH2* et les translocations de *BCL2*.

Cette étude a montré qu'il existait bien une hétérogénéité des profils génétiques dans les LDGCB ayant un impact sur la survie avec d'un côté un groupe de patients BN2 et EZB de bon pronostic et de l'autre les patients MCD et N1 de mauvais pronostic (figure 11). En revanche, cette nouvelle classification qui est proposée ne classe que 46,6 % des LDGCB de cette série ce qui pose des questions sur son applicabilité dans les essais cliniques et en routine.

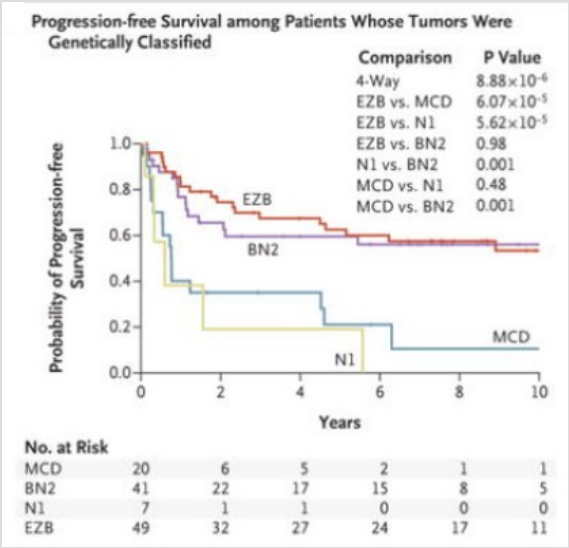


Figure 11: Courbes de survie des LDGCB classés en EZB, N1, BN2 ou MCD selon l'étude de R. Schmitz [74].

Cette courbe de survie représente la probabilité de rechute des patients dont les anomalies génétiques les associent à l'un des groupes EZB, N1, BN2 ou MCD. Les groupes MCD et N1 ont un risque accru de rechute à deux ans.

Une étude publiée quelques mois plus tard par Chapuy et al sur une cohorte de 304 LDGCB tente elle aussi de définir de nouveaux groupes en identifiant des gènes significativement mutés dans une série de 304 WES via l'algorithme MutSig2CV [77].

On retrouve dans celle-ci une description des mutations du gène suppresseur de tumeur *TP53*, de gènes impliqués dans la régulation épigénétique (*KMT2D*, *CREBBP* et *EP300*), des gènes des voies du BCR, des TLR et de NF-kB (*CD79B*, *MYD88*, *CARD11* et *TNFAIP3*), des gènes de la voie RAS (*KRAS*, *BRAF*) et des gènes impliqués dans les voies immunitaires (*B2M*, *CD58*, *CD70* et *CIITA).* A ces gènes viennent s'ajouter de nouveaux candidats sur les voies du BCR et des TLR comme *PTPN6*, *LYN*, *HVCN1*, *PRKCB et TLR2*,

des gènes impliqués dans la régulation des histones ou encore le ligand de PD1 *CD274* (ou *PDL1*).

A partir des mutations détectées et des données de CNV et de translocations, les auteurs ont identifié 5 clusters (C1-C5) regroupant une grande majorité des échantillons de cette série :

- le premier cluster C5 regroupe 64 LDGCB ayant un gain 18q induisant la surexpression de *BCL2*. On observe chez ces patients un nombre important de mutations des gènes *MYD88* (L265P) et *CD79B* fréquemment retrouvés dans le sous-type ABC. 96 % des patients de ce groupe sont d'ailleurs de sous-type ABC.
- le cluster C1 regroupe 56 LDGCB ayant des variations structurelles sur le gène *BCL6* et des mutations de *NOTCH2*. On retrouve aussi des mutations fréquentes des gènes de la voie NF-kB. La majorité des cas de ce cluster sont eux aussi de sous-type ABC.
- le cluster C3 regroupe 55 LDGCB ayant des variants structuraux impliquant le gène *BCL2* et des altérations des gènes de l'épigénétique tels que *KMT2D* (*MLL2*), *CREBBP* et *EZH2*. On retrouve dans ce cluster des altérations du gène *PTEN* que ce soit par la présence de mutations tronquantes ou de délétions focales 10q23.31. Ce cluster regroupe une majorité de LDGCB de sous-type GCB.
- le cluster C4 regroupe 51 LDGCB de sous-type GCB ayant des mutations des gènes *CD83*, *CD58*, *CD70*, *RHOA*, *GNA13*, *CARD11* ou encore *BRAF* et *STAT3*.
- le cluster C2 regroupe 64 LDGCB ayant des inactivations bi-alléliques fréquentes du gène *TP53* et des délétions 9p21 (*CDKN2A*). Ce cluster a la particularité d'être indépendant de la COO en regroupant à la fois des patients de sous-type ABC et GCB.

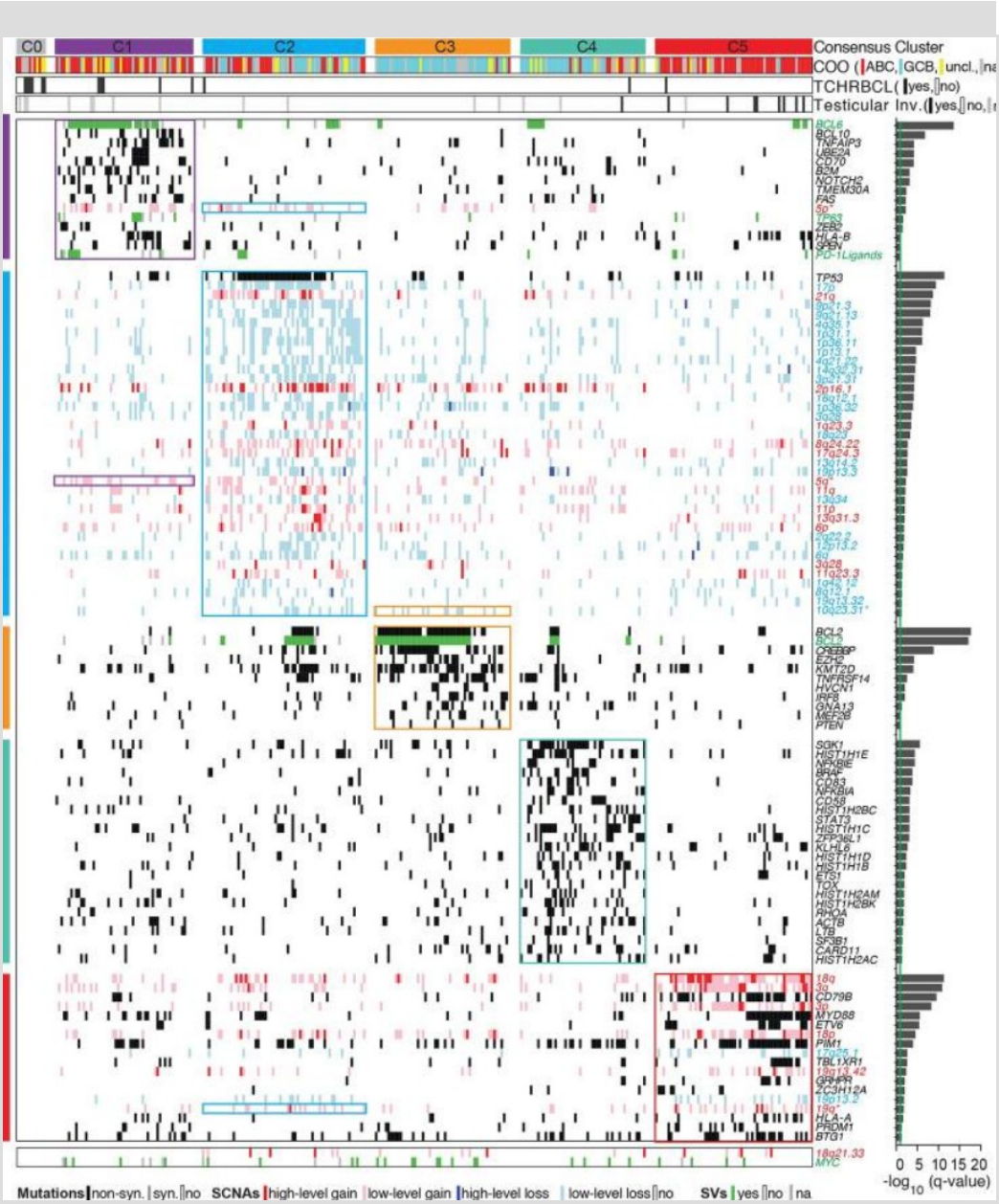


Figure 12: Profils mutationnels de 304 LDGCB à partir des données de séquençage d'exomes de l'étude conduite par Chapuy et al [73].

On retrouve en haut de la figure les différents groupes C0 à C5 et à droite les anomalies associées et leur fréquence au sein de la série.

Enfin, on retrouve comme dans l'étude précédente de R. Schmitz et al [74] un cluster C0 regroupant les patients non-classés. Cette fraction de LDGCB non classés est bien moins importante que dans les études précédentes puisque que le cluster C0 n'agrège que 12 LDGCB, soit 4% de la série. Cependant, on remarque que les groupes proposés ne sont pas clairement indépendant sur la classification hiérarchique réalisée (figure 12) et que les

altérations retrouvées dans ces groupes ne sont pas exclusives. La frontière entre les différentes entités identifiées par les *clusters* est parfois floue.

Figure 13: Analyse intégrative des données de séquençage, de transcriptome, de FISH et d'immunohistochimie sur une cohorte de 223 LDGCB (Dubois & al [77]).

La figure présente les résultats du clustering NMF (Nonnegative Matrix Factorization) sur la cohorte de patients. On retrouve la liste des anomalies génétiques associées à chaque cluster avec le nom du gène et le type d'anomalie. On retrouve aussi la classification GenClass (MCD, BN2, N1, EZB, STS) précédemment décrite.

Notre équipe a conduit en 2019 une analyse intégrative des données de transcriptome par puce Affymetrix, de séquençage à haut-débit, de CGH, de FISH et d'IHC de 223 LDGCB de l'étude prospective, multicentrique et randomisée de l'essai clinique LNH-03B du LYSA [78]. Une analyse en composantes indépendantes (independent component analysis ou ICA) sur les données transcriptomiques a permis de mettre en évidence 38 composantes décrivant la variabilité d'expression dans cette cohorte de LDGCB (figure 13). Beaucoup de ces composantes étaient liées à des signatures déjà décrites comme la COO, la signature stromale ou encore l'expression de *MYC*. La corrélation avec les autres sources de données, notamment les mutations somatiques, a permis d'expliquer le rationnel biologique sous-jacent derrière

quelques unes de ces composantes transcriptomiques. De façon intéressante, une composante liée aux gains du locus 19q13 s'est avérée corrélée significativement à la probabilité de rechute et à la survie dans cette cohorte. La comparaison des différentes composantes de l'ICA de cette étude aux groupes MCD, BN2, N1 et EZB précédemment décrits par R. Schmitz et al montre des recouvrements partiels [74].

Toutes ces études soulignent l'importance d'appréhender l'hétérogénéité des LDGCB au regard de toutes les découvertes liées aux technologies à haut-débit.

A.3. Lymphome primitif du médiastin

Le lymphome primitif du médiastin (Primary mediastinal large B-cell lymphoma, PMBL) est un lymphome B agressif qui compte pour 2 à 3 % des formes de LNH. On retrouve ce sous-type de lymphomes chez les sujets plutôt jeunes avec une médiane d'âge au diagnostic autour de 35 ans. Il impacte plus souvent les femmes avec un sex-ratio de 2:1 [79]–[81]. Les patients sont diagnostiqués dans plus de 80 % à des stades précoces de la maladie [81].

Sur le plan clinique, les personnes atteintes de PMBL ont souvent une masse volumineuse dans le thorax causant des symptômes en fonction de l'évolution de son volume. On retrouve entre autre un essoufflement, de la toux, des douleurs thoraciques et pour les masses importantes un blocage partiel de la veine cave supérieure. La maladie est très souvent localisée avec des envahissements autour du médiastin mais sans implication d'autres ganglions distants. Certaines formes rares de PMBL non médiastinaux sont décrits dans la littérature [82]–[85] mais leur fréquence reste probablement sous-estimée car seules les techniques de transcriptomique peuvent permettre de les mettre en évidence et celles-ci ne sont pas toujours réalisées au diagnostic. La prise en charge de ces patients repose sur un traitement par chimiothérapie associé ou non à de la radiothérapie. Ces lymphomes sont globalement de bon pronostic.

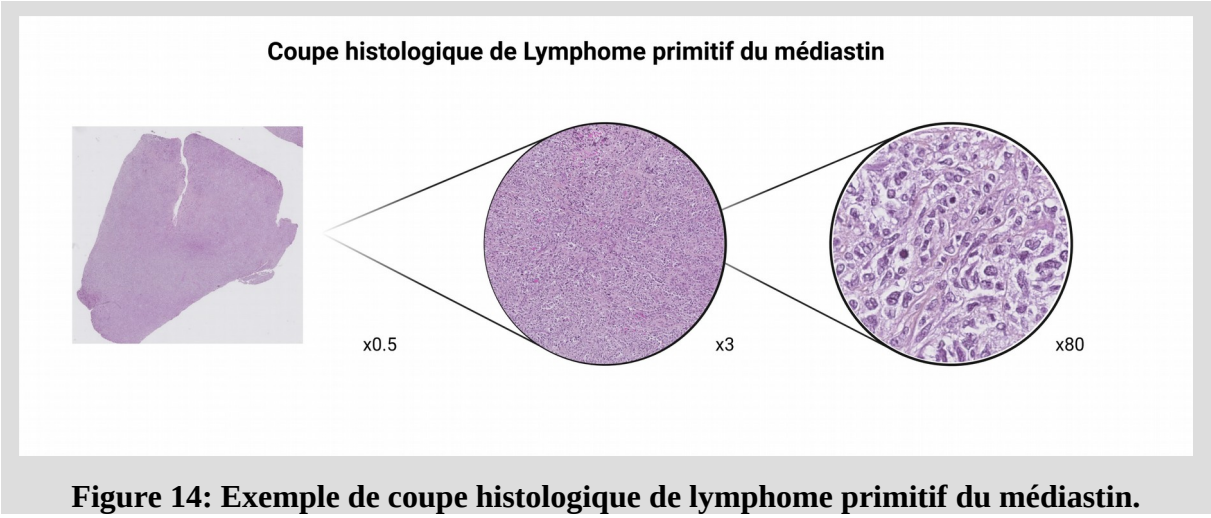


Figure 14: Exemple de coupe histologique de lymphome primitif du médiastin.

L'aspect morphologique du PMBL est très variable en microscopie (figure 14). On retrouve de larges cellules associées souvent à de la fibrose interstitielle [80], [86], [87]. Il arrive parfois que les lymphocytes tumoraux ressemblent aux cellules de Reed-Sternberg qui sont caractéristiques des lymphomes de Hodgkin (voir chapitre A.4). Rarement, ces lymphomes sont appelés « grey-zone » lorsqu'ils combinent à la fois les traits caractéristiques d'un PMBL et d'un LH [87], [88]. Des cas de rechutes de patients atteints de LH de forme scléro-nodulaire en PMBL ont aussi été rapportés dans la littérature [88], [89] ce qui démontre bien la difficulté de classification des cas frontières.

Sur le plan immunohistochimique, les PMBL expriment les marqueurs classiques de la lignée B (*CD19*, *CD20*, *CD22* et *CD79A*). Le marqueur *CD30* est présent dans une grande majorité des cas [89]. Les PMBL sont presque toujours EBV négatif [80]. A l'inverse du LDGCB, on retrouve dans plus de 70 % des cas une expression des marqueurs *CD23*, *MAL* et des gènes *PD1/PDL-1* [91]–[93].

Sur le plan moléculaire, on retrouve dans ce sous-type de lymphome des réarrangements fonctionnels des gènes des immunoglobulines avec cependant une forte charge de mutations [94]. Les réarrangements des gènes *BCL2*, *BCL6* et *MYC* sont très rares [95]. Des réarrangements et mutations de *CIITA* ont été rapportés dans plus de 50 % des PMBL [96], [97] entraînant une diminution du nombre de molécules du complexe majeur d'histocompatibilité de classe II (CMH II) à la surface des lymphocytes B tumoraux. Les gains et amplifications du locus 9p24.1 sont présents dans plus de 70 % des cas de PMBL [98] et entraînent une surexpression des gènes *PDL1* et *PDL2* [99], [99]. On retrouve aussi dans près de 50 % des cas des gains du locus 2p16.1.

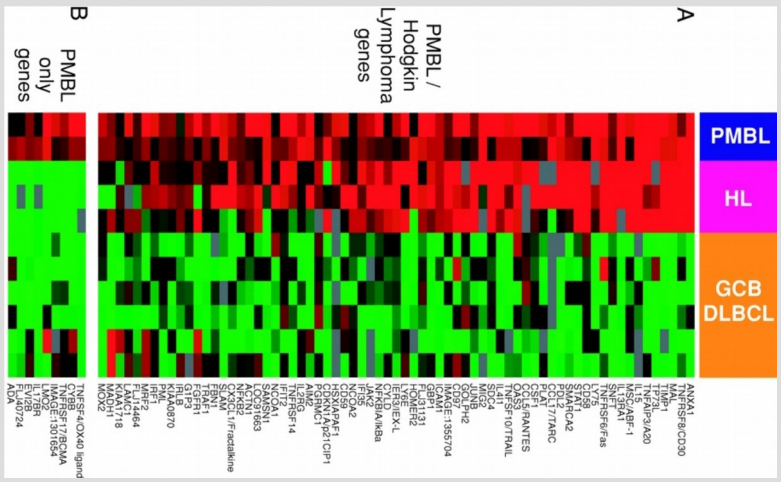


Figure 15: Signature transcriptionnelle comparant des lignées cellulaires de PMBL, de LH et de DLBCL [83].

A. Signature partagée par les lignées cellulaires de PMBL et de LH. **B.** Signature spécifique des lignées de PMBL.

Les PMBL ont une activation constitutive de la voie NF-κB [100] pouvant être reliée aux délétions fréquentes du gène *TNFAIP3* dans 60 % des cas [50], [101]. On retrouve aussi des mutations fréquentes sur les gènes *SOCS1*, *STAT6*, *PTPN1*, *ITPKB*, *MFHAS1* et *XPO1* [50], [102]. Les PMBL ont une signature transcriptomique distincte des autres entités de lymphomes B mais qui partage certaines similarités en terme d'expression avec le LH classique [83], [84] (figure 15).

A.4. Lymphome de Hodgkin

Le lymphome hodgkinien (LH), ou maladie de Hodgkin, doit son nom au médecin Thomas Hodgkin à l'origine de sa découverte au XIXème siècle. Il représente 10% de tous les lymphomes diagnostiqués et touche principalement le sujet jeune, avec une moyenne d'âge au diagnostic de 25 ans pour les hommes et de 22 ans pour les femmes. Il touche aussi plus rarement les personnes de plus de 80 ans.

Le LH peut être divisé en deux grandes familles : le LH classique (Classic Hodgkin Lymphoma, cHL), qui compte pour 95% des cas, et le LH nodulaire à prédominance lymphocytaire (Nodular lymphocyte perdominant Hodgkin Lymphoma).

Lymphome de Hodgkin classique

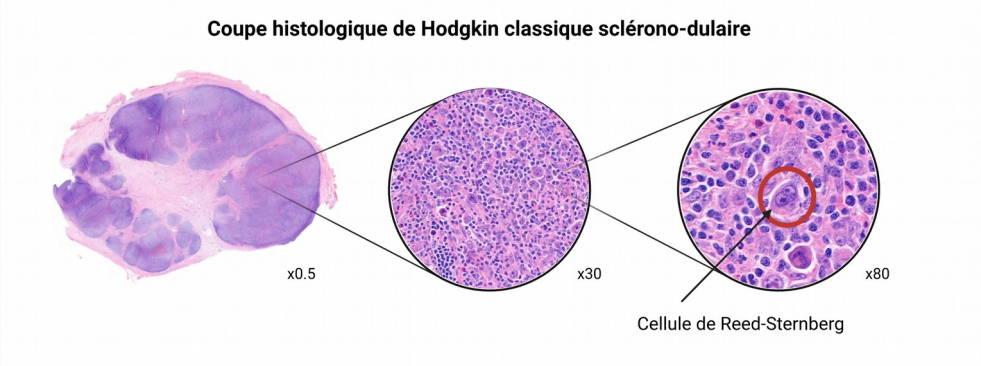


Figure 16: Exemple de coupe histologique de lymphome de Hodgkin scléro-nodulaire.

La cellule tumorale, ou cellule de Reed-Sterberg, ne représente qu'une faible proportion des cellules du tissu. Ces images proviennent du service d'anatomopathologie du Centre Henri Becquerel.

D'un point de vue morphologique, le LH classique est caractérisé par la présence de cellules B anormales appelées cellules de Reed-Sternberg (RS), qui prennent l'aspect de grosses cellules avec un ou ou plusieurs noyaux (figure 16).

Des études transcriptomiques ont montré que plusieurs gènes ont un niveau d'expression dérégulé dans les cellules de RS [103]–[109]. On retrouve une activation anormale du facteur de transcription NF-kB conduisant à une dérégulation des processus de prolifération, de survie cellulaire [110]–[112] et de la voie de signalisation JAK/STAT [113].

Ces cellules de RS sont retrouvées associées à un mélange de cellules immunitaires et, en fonction de cet infiltrat réactionnel, on distingue plusieurs entités :

- une forme scléro-nodulaire dans laquelle la cellule de RS est entourée de lymphocytes normaux et où on retrouve des tissus cicatriciels (Nodular sclerosis classif Hodgkin lymphoma, NSCHL)
- une forme à cellularité mixte, associée à une infection à l'EBV dans 75 % des cas [114], [115] et caractérisée par la présence d'un grand nombre de cellules de RS dans les ganglions lymphatiques impactés (Mixed cellularity classic Hodgkin Lymphoma, MCCHL)
- une forme riche en lymphocytes dans laquelle on ne retrouve que très peu de cellules de RS typiques (Lymphocyte-rich classic Hodgkin lymphoma, LRCHL)

- une forme à déplétion lymphocytaire dans laquelle on ne retrouve que très peu de lymphocytes normaux à l'inverse des cellules de RS qui se retrouvent abondamment dans le tissu tumoral (Lymphocyte-depleted classic Hodgkin lymphoma, LDCHL)

Sur le plan génétique, l'émergence des technologies de séquençage à haut débit associée à des techniques de microdissection permettant d'enrichir les librairies en cellules de RS a permis de mieux appréhender les anomalies génétiques caractéristiques des LH. On retrouve des anomalies génétiques impactant principalement la voie NF-kB (*NFKBIA*, *NFKBIE*, *TNFAIP3*) et la voie JAK/STAT (*JAK2*, *SOCS1*, *PTPN1*, *PTPN2*, *STAT6*) [101], [116]–[124]. On retrouve aussi des altérations des gènes *B2M* et *CIITA* [117], [118] et des anomalies récurrentes du gène *XPO1* qui code pour une protéine impliquée dans le transport nucléaire [102], [125]. Néanmoins, le faible contingent tumoral dans cette pathologie rend en pratique les analyses génétiques à partir de biopsies des LH complexes à mettre en œuvre.

Lymphome de Hodgkin nodulaire à prédominance lymphocytaire

La maladie de Hodgkin nodulaire à prédominance lymphocytaire (nodular lymphocyte predominant HL, NLPHL), antérieurement appelée paragranulome de Poppema, est une entité rare. Elle représente moins de 10 % de tous les cas diagnostiqués de LH [126] et touche principalement les hommes entre 30 et 52 ans. Sur le plan clinique, la NLPHL est une maladie évoluant lentement et qui est globalement de bon pronostic même si les rechutes ne sont pas rares après traitement. La survie globale des patients à 10 ans est de supérieure à 80 % pour les stades I et II [127]. Dans certains pays, comme en France, les stades I de la maladie ne sont pas traités après chirurgie du ganglion atteint [128].

Le diagnostic de NLPHL est complexe de part sa ressemblance morphologique et phénotypique aux LRCHL ou aux lymphomes à grandes cellules B riches en lymphocytes T. La distinction entre ces trois pathologies, bien que complexe, est nécessaire car la prise en charge thérapeutique et les pronostics sont différents en fonction de ces entités.

D'un point de vue morphologique, la NLPHL est caractérisée par une prolifération de cellules tumorales appelées « pop-corn » dispersées sur un fond cellulaire constitué principalement de petits lymphocytes B regroupés en nodules. Ces cellules sont positives à un certain nombre de marqueurs immunohistochimiques comme le CD20, OCT2, CD75, CD79a, PAX5 et CD45 [129]–[131]. Des études transcriptomiques ont montré que le programme B des lymphocytes tumoraux est conservé [132], ce qui n'est pas le cas des cellules de RS dans la forme classique de LH. On retrouve une infection des cellules tumorales par le virus de

l’EBV dans 3 à 5 % des cas [133]. De nouveaux marqueurs d’expression, tels que les facteurs de transcription *BOB1* [104], pourraient dans l’avenir être une aide supplémentaire au diagnostic différentiel.

Sur le plan génétique, les cellules tumorales de RS présentent une clonalité avec des réarrangements souvent fonctionnels des gènes d’immunoglobulines [134], [135], détectables uniquement après microdissection car les cellules de RS sont rares dans les biopsies des patients. On retrouve dans près de la moitié des cas des réarrangements de *BCL6* [136, p. 6], [137], [138]. Les gènes *PAX5*, *PIM1*, *RHOH* et *MYC* sont retrouvés mutés dans près de 80 % des cas [139]. Des mutations sont aussi retrouvées dans la moitié des cas sur les gènes *SGK1*, *DUSP22* et *JUNB [140]*. L’analyse des anomalies génétiques dans le LH passe nécessairement, en l’absence de techniques lourdes de microdissection, par des approches très sensibles comme le séquençage à haut-débit.

B. Séquençage de l'ADN

Le chapitre précédent a été l'occasion d'explorer l'importante hétérogénéité des lymphomes et plus spécifiquement des LH et des LNH. Si le diagnostic repose toujours sur l'anatomopathologie et l'immunohistochimie, de nouvelles approches moléculaires semblent jouer un rôle de plus en plus important dans leur classification. Les différentes études présentées précédemment, basées sur l'analyse du transcriptome et des profils mutationnels à partir de biopsies tissulaires, ouvrent des perspectives intéressantes pour la caractérisation des tumeurs au diagnostic.

La recherche de mutations prend une place grandissante dans les laboratoires de diagnostic afin de mieux caractériser les tumeurs. Nous nous intéresserons dans ce chapitre au développement des méthodes de séquençage de première génération et seconde génération permettant de rechercher des mutations au diagnostic.

B.1. Structure de l'ADN

L'épopée débute en 1944 avec le canadien Oswald Avery qui découvrit pour la toute première fois de quoi sont composés nos chromosomes : l'acide désoxyribonucléique (ADN) [141]. Edwin Chargaff, biochimiste lui aussi canadien, démontre quelques années plus tard que cet ADN est en réalité composé d'une succession de nucléotides : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Il démontre de plus que si la composition en nucléotides de l'ADN semble varier d'une espèce à une autre, les rapports A/T et G/C sont quasiment égaux et proches de 1. Cette dernière observation fût la toute première description de la complémentarité des nucléotides.

En 1953, un article publié dans le journal Nature du généticien américain James Watson et du physicien britannique Francis Crick décrit pour la toute première fois la structure en trois dimensions de l'ADN : une double hélice enroulée autour d'un axe [142]. Cette découverte, rendue possible par les travaux de diffraction des rayons X de Maurice Wilkins et de Rosalind Franklin, conduira à l'obtention d'un prix Nobel de médecine en 1962. Cette caractérisation de la structure de l'ADN bouleversera l'histoire de la biologie et de la génétique. Elle permettra entre autre de comprendre les principaux mécanismes à l'origine du fonctionnement de nos cellules.

B.2. Séquençage de première génération de l'ADN

Le séquençage de l'ADN vise à déterminer l'ordre d'enchaînement des quatre nucléotides d'un brin d'ADN. Véritables pionniers de la biologie moléculaire, Walter Gilbert aux États-Unis et Frederick Sanger au Royaume-Uni développèrent deux premières méthodologies de séquençage qui les conduiront à la co-obtention d'un prix Nobel en 1980 pour leurs travaux.

Les approches de Sanger et de Gilbert sont pourtant diamétralement opposées : l'une repose sur la synthèse enzymatique du brin complémentaire d'un ADN simple brin tandis que l'autre se base sur les propriétés de dégradation chimique de l'ADN en utilisant les réactivités sélectives des différents nucléotides afin d'en déduire sa séquence.

La méthode de séquençage Sanger [143] est de nos jours toujours considérée comme la méthode de référence en terme de qualité de séquençage. Elle vise à synthétiser un brin complémentaire à partir d'un brin matrice et d'une amorce servant à initier la polymérisation de l'ADN. L'élongation de cette amorce est ensuite réalisée par une ADN polymérase thermostable. Des désoxyribonucléotides (dATP, dCTP, dGDP, dTTP) sont ajoutés au mélange afin de permettre l'action de l'ADN polymérase. On ajoute une faible concentration de l'un des didésoxyribonucléotides marqués par un traceur radioactif (ddATP, ddCTP, ddGTP ou ddTTP). L'incorporation des didésoxyribonucléotides, qui ne possèdent pas d'extrémité 3'-OH libre, empêche la poursuite de l'élongation du brin synthétisé.

Le séquençage d'un même fragment d'ADN est donc réalisé en répétant quatre fois cette réaction en parallèle avec les quatres didésoxyribonucléotides différents marqués par un traceur radioactif. Les fragments générés sont de tailles différentes et sont séparés par électrophorèse sur gel de polyacrylamide afin de reconstituer la séquence du fragment d'intérêt (figure 17).

Dans les années qui ont suivies, un très grand nombre d'améliorations ont été apportées au séquençage Sanger avec entre autre le remplacement du marquage radioactif par une révélation basée sur l'utilisation de fluorochromes ou encore le développement de l'électrophorèse capillaire [144]–[149]. (figure 17.4). Ces deux avancées conduiront au développement des premier séquenceurs dits de « première génération » capables d'automatiser les réactions de séquençage. Le séquençage Sanger est encore très largement

utilisé dans les laboratoires de diagnostic car il permet de tester rapidement certaines mutations récurrentes (hotspots) avec une bonne précision et surtout un coût de séquençage très modéré.

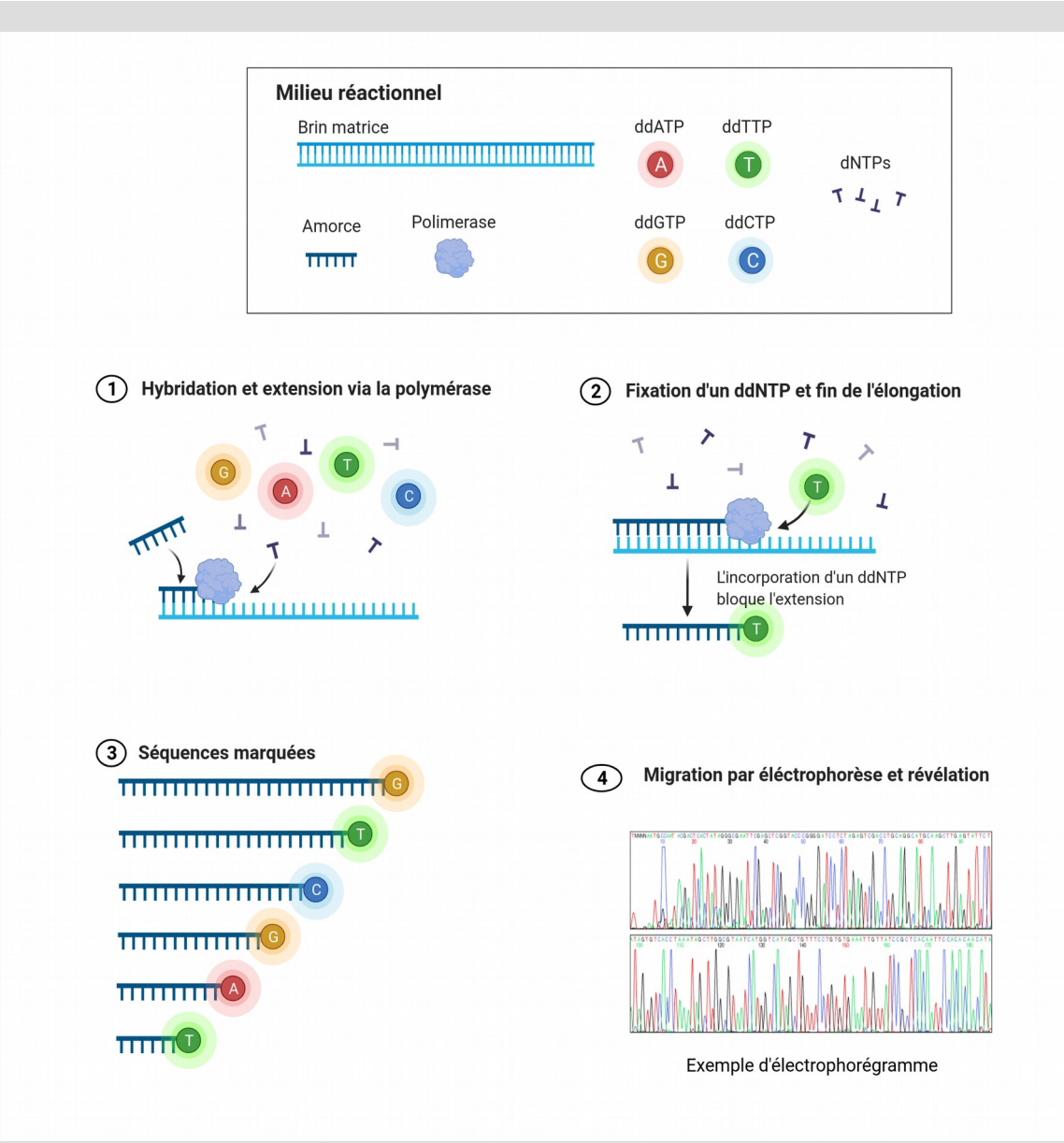


Figure 17: Principe du séquençage selon la chimie Sanger.

B.3. Séquençage de nouvelle génération de l'ADN

La technologie de séquençage par la chimie Sanger a inspiré très largement le développement des nouvelles technologies de séquençage à haut débit par synthèse et notamment des séquenceurs Illumina et Ion Torrent. Nous détaillerons dans cette section les

différentes méthodes de construction de librairie de séquençage ainsi que le fonctionnement des séquenceurs de nouvelle génération.

Les nouvelles technologies de séquençage (Next-generation sequencing, NGS) font référence à l'exploration profonde, à haut débit et de façon massivement parallèle d'échantillons biologiques (ADN, ARN, protéines…). Elles ont entraîné une diminution très importante des coûts de séquençage puisque l'obtention d'un séquençage complet de trois génomes humains peut se réaliser aujourd'hui en trois jours sur des séquenceurs Illumina tels que l'HiSeq X pour un coût proche des 1000 dollars par génome, là où le projet initial du séquençage du génome humain aura coûté près de 3 milliards de dollars sur 13 ans.

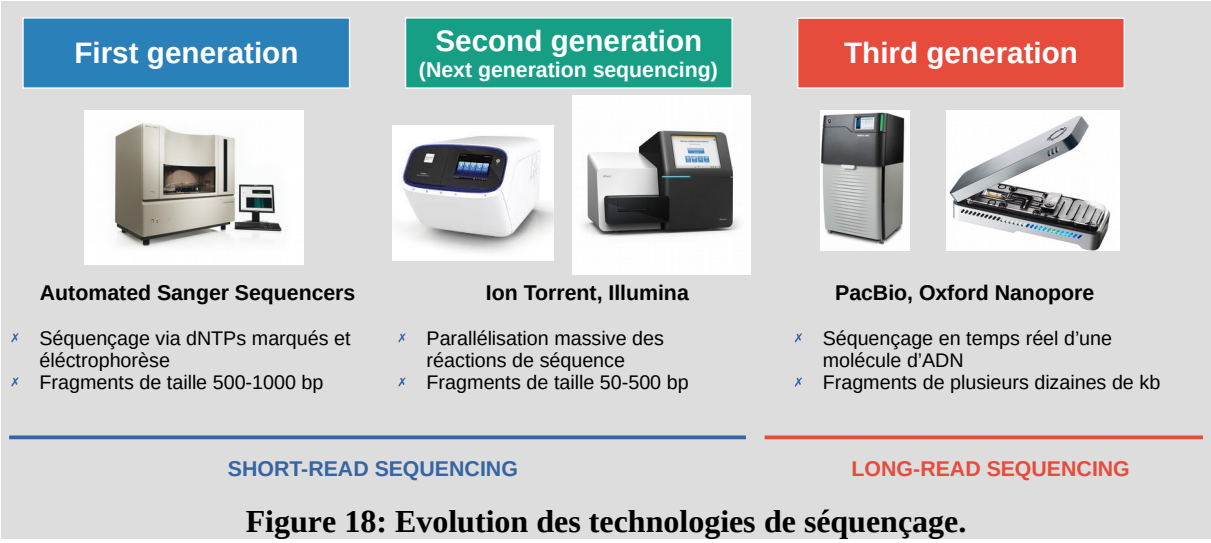


Figure 18: Evolution des technologies de séquençage.

Les séquenceurs de nouvelle génération sont devenus la technologie de choix pour l'analyse de génomes, d'exomes ou de transcriptomes car elles offrent la possibilité d'obtenir plusieurs centaines de gigaoctets (Go) de données en une seule expérimentation.

Les technologies NGS peuvent en réalité se découper en deux grandes familles : le séquençage de deuxième génération qui propose le séquençage de lectures courtes (short-read sequencing) et le séquençage de troisième génération produisant des lectures longues (long-read sequencing) (figure 18). Les séquenceurs de troisième génération, tels que les technologies Nanopore, permettent le séquençage de longues macromolécules en une seule séquence contiguë. Véritable révolution sur le plan technologique, son utilisation est néanmoins très limitée en cancérologie car son taux d'erreur reste trop élevé.

Les laboratoires biomédicaux sont aujourd'hui massivement équipés de séquenceurs de deuxième génération, et plus spécifiquement des technologies de type Ion Torrent ou Illumina qui se partagent la plus grande part du marché. Si ces deux gammes de séquenceurs reposent

sur des technologies bien différentes pour déterminer la séquence, elles nécessitent toutes deux la construction de librairies de séquençage à partir des échantillons biologiques.

Construction des librairies

La première étape nécessaire au séquençage à haut-débit de deuxième génération est la construction de librairies de séquençage. Cette étape vise à amplifier, par PCR, le matériel biologique extrait (ADN ou ARN) avant de réaliser son séquençage.

Nous introduirons brièvement dans cette section trois grandes familles de méthode de construction de librairies : la génération d'amplicons par PCR multiplexe (AmpliSeq), l'enrichissement des régions d'intérêt par des sondes de capture (Agilent) ou via l'utilisation d'un seul primer spécifique (Qiagen).

Préparation des librairies par PCR multiplexe

Le séquençage par amplicons est basé sur le multiplexage massif d'amorces pour amplifier des régions d'intérêt pouvant aller de quelques gènes à quelques centaines de gènes (figure 19). Certains produits commerciaux, comme le kit AmpliSeq Exome RDY développé par la société Thermofisher, sont capables de multiplexer jusqu'à 24 000 amorces par réaction.

Ces technologies de préparation de librairie sont particulièrement utilisées pour le séquençage ciblé puisque les protocoles sont relativement simples et rapides à mettre en place. Elles sont compatibles avec peu de quantité d'ADN au départ (10 ng) et sont généralement applicables à une grande variété d'échantillons biologiques. En fonction de la nature des échantillons à analyser, l'ADN des échantillons peut être plus ou moins fragmenté. Il est donc nécessaire d'adapter la taille des amplicons en fonction des applications de sorte à ce que les deux amorces spécifiques de l'amplification d'une région d'intérêt puissent se fixer sur le fragment d'ADN ciblé.

Il est aussi possible d'utiliser différents types d'adaptateurs pour basculer d'une technologie de séquençage à une autre avec un même pool d'amorces. Les coûts de ces technologies sont eux aussi très abordables et fonction du nombre d'amorces synthétisées, et donc du nombre de régions ciblées.

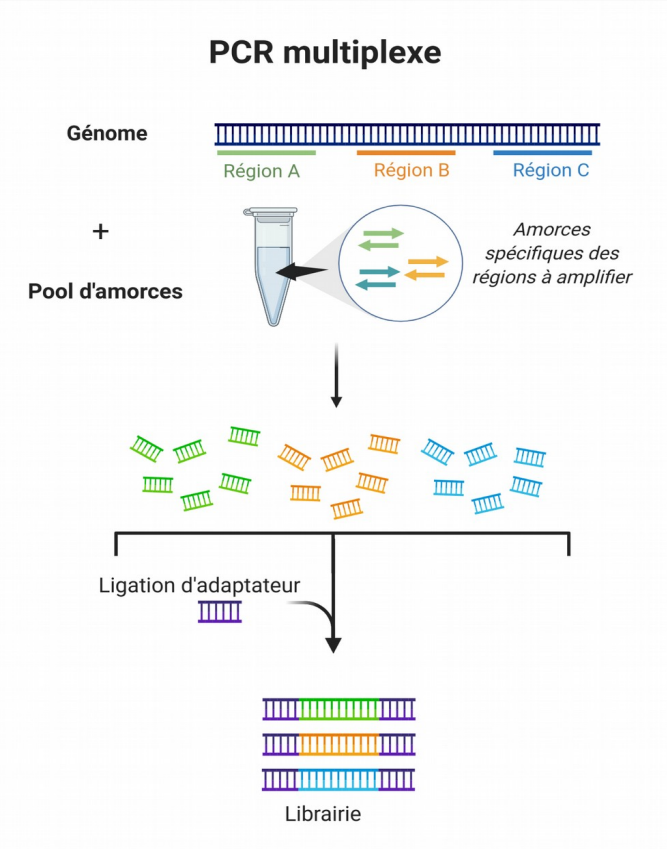


Figure 19: *C*onstruction des librairies par PCR multiplexe.

Les amorces spécifique des régions génomiques ciblées sont hybridées sur l'ADN extrait de l'échantillon biologique. Une étape d'amplification par PCR permet d'amplifier les régions ciblées puis d'ajouter les adaptateurs du support de séquençage. Ces séquences d'adaptateurs sont spécifiques de la technologie de séquençage choisie.

Préparation des librairies par capture

Les technologies de préparation de librairie basées sur la capture reposent sur l'utilisation d'amorces d'ARN, ou plus rarement d'ADN, qui sont biotinylées (Figure 20). Les séquences de ces sondes sont complémentaires de régions génomiques d'intérêt dont on souhaite déterminer la séquence.

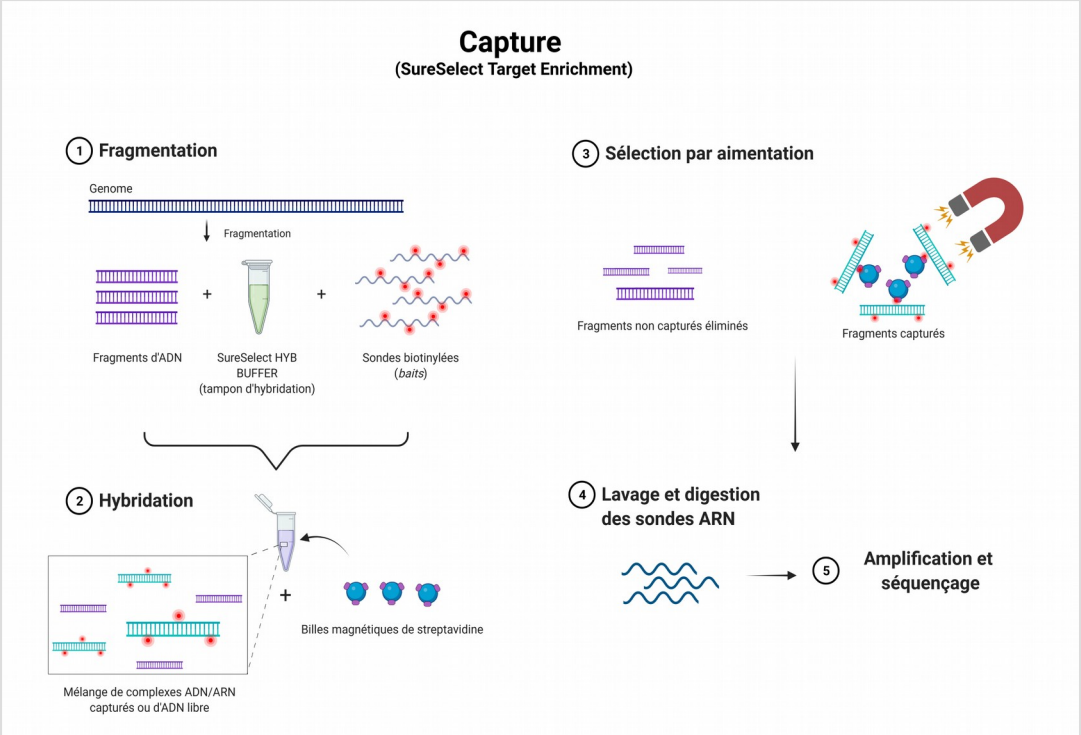


Figure 20: Préparation de librairie par capture.

Après fragmentation de l'ADN extrait de l'échantillon, les sondes biotinylées sont hybridées sur les régions d'intérêt. L'étape d'aimantation permet d'éliminer les fragments d'ADN qui ne sont pas ciblés de sorte à n'amplifier que les régions d'intérêt.

La matrice d'ADN est dans un premier temps fragmentée puis les fragments d'ADN sont mis en présence des sondes de capture dans un tampon d'hybridation. Des billes magnétiques couplées à la streptavidine sont alors utilisées pour se lier aux sondes biotinylées, sondes qui sont elles mêmes liées aux fragments d'ADN d'intérêt. Ces fragments sont alors isolés des autres par aimantation puis lavage. Les sondes ARN biotinylées sont enfin dégradées par digestion enzymatique de sorte à n'obtenir que les fragments d'ADN pour la construction ultérieure des librairies de séquençage.

En fonction des solutions commerciales utilisées, les adaptateurs permettant la fixation au support de séquençage sont ajoutés soit en amont de l'étape de sélection des régions génomiques soit en aval au moment de la PCR d'amplification de la librairie.

Préparation des librairies par extension d'une amorce unique

La principale limitation des technologies de PCR multiplexe pour la préparation des librairies de séquençage réside dans la nécessité de trouver deux amorces spécifiques autour

de la région d'intérêt pour pouvoir l'amplifier. Cela implique de trouver deux amorces qui soient suffisamment spécifiques pour initier l'amplification mais aussi de placer ces amorces sur des sites dépourvus d'insertions, de délétions ou de SNP afin de ne pas diminuer leur potentiel d'hybridation. Si par exemple une délétion somatique se trouve sur un site de fixation de l'une des deux amorces, alors cette région ne pourra pas être amplifiée et en conséquence l'anomalie génétique sous-jacente ne pourra pas être détectée.

Afin de palier à cette contrainte, la société Qiagen a développé une solution appelée « QIAseq Targeted DNA panel » dans laquelle une seule amorce spécifique de la région d'intérêt est nécessaire.

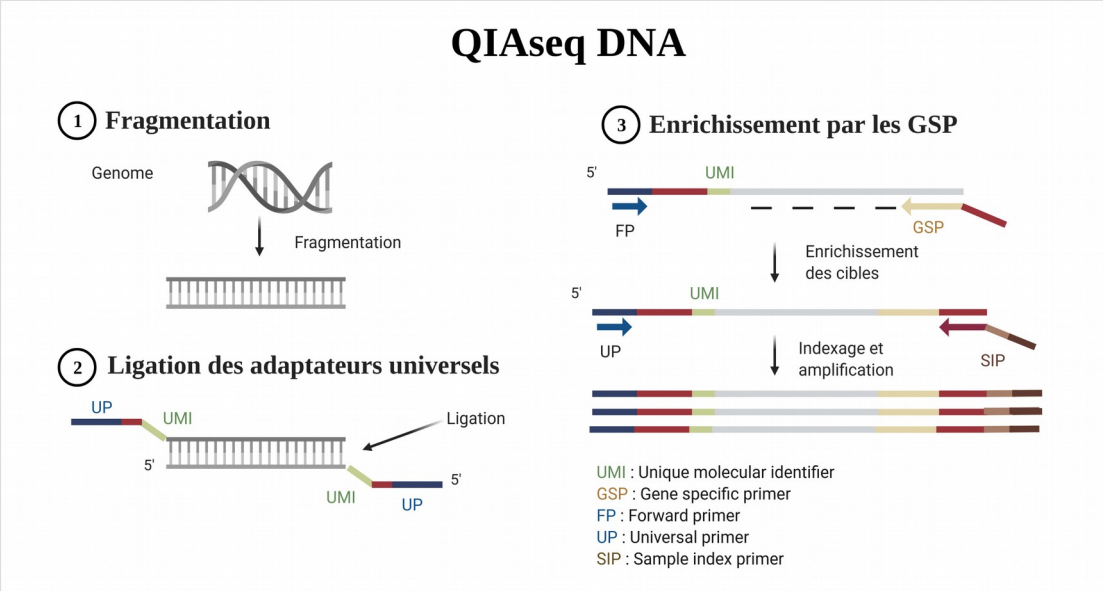


Figure 21: Chimie QIAseq développée par la société Qiagen.

L'ADN extrait de l'échantillon biologique est d'abord fragmenté puis les extrémités sont réparées (1). Une étape de ligation de l'amorce universelle contenant une séquence universelle et une séquence aléatoire (UMI) est réalisée (2). Cette construction liguée à l'ADN fragmenté est la même pour l'ensemble des fragments. Enfin, une PCR permet d'amplifier les régions d'intérêt via l'utilisation d'amorces spécifiques des régions d'intérêt (Gene Specific Primers, GSP) (3).

L'ADN double brin de l'échantillon est d'abord fragmenté puis des adaptateurs comprenant un primer universel est amené en 5' des fragments d'ADN par ligation (figure 21). Les primers spécifiques des régions d'intérêt, appelés Gene Specific Primers (GSP), sont hybridés. Seules les régions ayant reçues à la fois le primer universel par ligation et l'hybridation d'un GSP sont alors amplifiées. Il est ainsi possible de créer plusieurs amorces

autour des régions d'intérêt afin d'assurer une couverture de 2 ou 3 GSP par région et ainsi limiter le risque de non fixation des amorces.

Les librairies QIAseq ont la particularité d'introduire des séquences appelées « Unique molecular identifier » (UMI). Ces barcodes moléculaires constituent des séquences aléatoires ajoutées en même temps que les adaptateurs universels avant toute étape d'amplification de la librairie. Les UMI sont de taille 12 et permettent une identification unique des fragments d'ADN après l'étape d'enrichissement via les GSP. Le rôle des UMI est décrit dans la section I.D.

Technologies de séquençage

Les technologies de séquençage de seconde génération, qui permettent le séquençage de fragments courts entre 120 pb et 600 paires de base (pb), sont celles les plus utilisées dans les laboratoires de diagnostic. Elles permettent de réaliser des séquençages plus ou moins ciblés allant de quelques cibles et/ou mutations hotspots jusqu'au séquençage d'exomes. Les prélèvements analysés sont en effet très souvent fragmentés (biopsie incluse en paraffine, fragments circulants) et ne nécessitent pas de séquençage de fragments longs via des technologies de troisième génération comme nanopore par exemple.

On retrouve principalement deux technologies :

- les séquenceurs Ion Torrent développés par la société Thermofisher, qui repose sur la pHmétrie pour déterminer la séquence des fragments d'ADN des librairies

- la société Illumina qui développe une gamme de séquenceurs utilisant la détection de fluorescence via un séquençage par synthèse.

Technologie Ion Torrent

Le séquençage de la technologie Ion Torrent repose sur la détection d'ions hydrogènes durant la polymérisation de l'ADN. C'est une méthode de séquençage par synthèse durant laquelle un brin complémentaire est synthétisé à partir d'un brin matrice fixé sur des billes. Brièvement, cette technologie nécessite tout d'abord d'hybrider les librairies de séquençage sur des billes appelées « Ion Sphere Particule » (ISP). Cette étape nécessite un dosage très précis des librairies de sorte à ce que chaque ISP ne reçoive qu'un fragment unique de la librairie. Une fois chaque fragment ligué à chaque ISP, une étape d'amplification clonale est réalisée de sorte à tapisser chaque ISP du même fragment d'ADN afin d'accroître le signal

lors de la réaction de séquençage. Le séquenceur ne sera pas en mesure d'analyser les ISP vides ou ayant reçues plusieurs fragments de librairie différents (polyclonales) (figure 22).

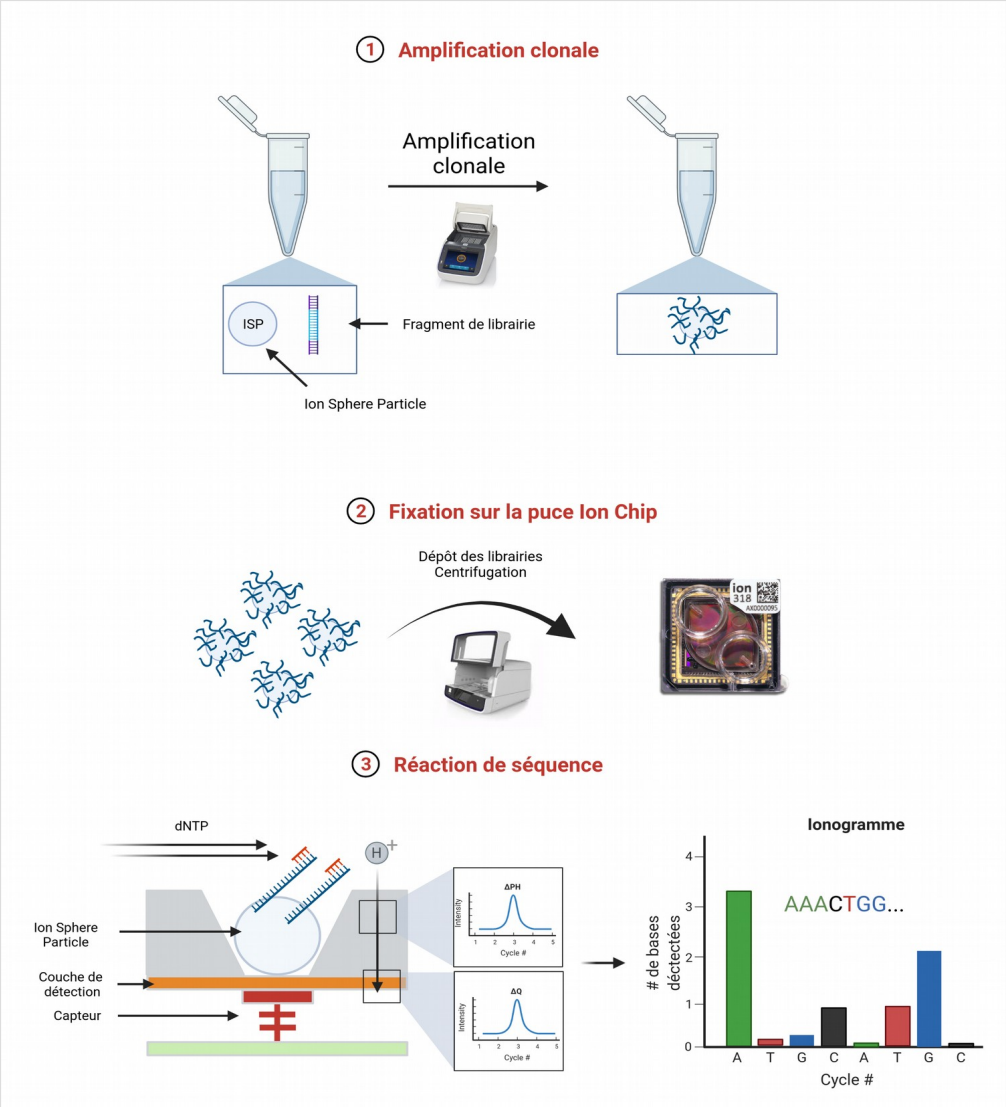


Figure 22: Technologie de séquençage Ion Torrent.

Les fragments d'ADN de l'échantillon sont hybridés puis amplifiés sur des billes appelées ISP. La concentration des ISP est adaptée à la quantité d'ADN utilisé lors de la construction des librairies de sorte à ce que chaque ISP n'amplifie qu'un seul fragment de librairie (amplification clonale). Les ISP sont ensuite chargées sur le support de séquençage (Ion Chip) afin de réaliser la réaction de séquence et l'acquisition des signaux.

Une fois cette étape d'amplification clonale réalisée, les ISP sont déposées sur le support de séquençage qui est une puce tapissée de puits appelée « Ion Chip ». Chaque puits, de par ses caractéristiques physiques, ne peut accueillir qu'une seule ISP. Afin d'assurer une répartition homogène des ISP à la surface de la puce, une étape de centrifugation est réalisée.

Cela permet d'exploiter pleinement les puits de chaque puce, puits qui dictent directement la quantité de données générées.

	 314	 316	 318	 IP1/IP2/IP3
Nombre de puits	1,3 M	6,3 M	22,3 M	165 M- 1,2 milliard
Nombre de séquences	300 000	1,6 M	4,6 M	124 - 496M
Taille des séquences	~ 400 pb	~300 pb	~320 pb	~640 pb
Durée	2,4 H	3,1 H	4,5 H	4H
Coût	$400	$500	$800	$1000

Figure 23: Ion Torrent - Tableau comparatif des différentes puces IonChip.

La capacité massivement parallèle de séquençage de cette technologie est directement liée à la quantité de puits analysables. Il existe plusieurs types de puce avec des nombres de puits différents à leur surface afin de générer plus ou moins de données en fonction des applications souhaitées. Les puces 314, 316 et 318 sont réservées au séquenceur PGM (Personal Genome Machine) tandis que les puces de plus grosses capacités IP1, IP2 et IP3 sont réservées au séquenceur Proton (figure 23).

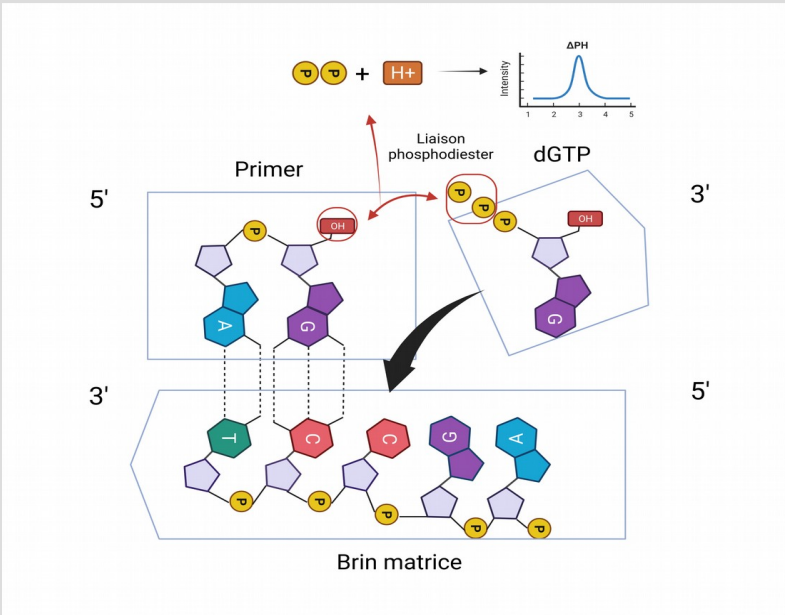


Figure 24: Ion Torrent - Réaction de séquence

L'ensemble des brins matrices, fixés indépendamment sur chaque ISP, sont ensuite séquencés en parallèle. Le séquenceur injecte à la surface de la puce successivement chaque dNTP. Si celui-ci est complémentaire du brin matrice, il est incorporé à la séquence via une liaison phosphodiester qui conduira à la libération d'un ion hydrogène H+ (figure 24).

Ce relargage de H+ dans la solution conduit à une variation locale du pH au niveau du puits et un capteur convertit celle-ci en variation d'intensité électrique. S'il y a une répétition de plusieurs nucléotides sur le brin matrice et donc une incorporation de plusieurs dNTP lors du même cycle, alors le capteur libérera un signal proportionnel au nombre d'H+ libérés. A noter que si ce signal reste proportionnel au nombre de dNTP incorporés au brin matrice, la technologie Ion Torrent a tendance à engendrer un nombre d'erreurs important dans les régions d'homopolymères longs du fait d'une saturation du signal d'acquisition.

La technologie Ion Torrent offre principalement des séquenceurs de faible et moyen débits avec une première génération de séquenceur appelés Ion Proton et Ion PGM et une deuxième génération de machines plus automatisées baptisées Ion S5. Ces séquenceurs ont tous la particularité d'être adaptés au séquençage de panels de gènes et à quelques exomes tout au plus. Ils offrent aux utilisateurs une interface particulièrement intuitive avec le séquenceur et un serveur de traitement bioinformatique des données appelé Torrent Server. Celui-ci permet à la fois un traitement primaire des données en temps réel, c'est à dire la conversion des signaux d'acquisition du séquenceur en des séquences nucléotidiques, mais permet aussi leur alignement et la détection des variants via une suite logiciel appelée « Torrent Suite ». Ces séquenceurs sont donc un moyen simple pour les laboratoires de diagnostic ne disposant pas de ressources bioinformatiques de s'équiper et de traiter leurs données.

Le traitement bioinformatique des données d'acquisition jusqu'à la détection des variants est détaillé dans le chapitre II.

Technologie Illumina

Les séquenceurs Illumina utilisent la chimie de séquençage par synthèse qui fonctionne en plusieurs étapes : la fixation de la librairie sur le support de séquençage, une étape d'amplification par pontage des fragments pour former des « clusters » et enfin le séquençage.

Les librairies de séquençage sont déposées aléatoirement sur des supports en verre appelés « flowcells » via des adaptateurs. Ces adaptateurs synthétiques sont des séquences nucléotidiques incluant différents segments :

- une séquence complémentaire des oligonucléotides fixés sur la flowcell qui permet d'attacher les séquences de la librairie au support de séquençage

- une séquence appelée barcode qui permet de réattribuer la séquence au bon échantillon lors de l'étape ultérieure de démultiplexage

- un site de fixation de l'amorce de séquençage permettant l'initiation de la réaction de séquençage par synthèse

L'amplification des fragments de la librairie est une amplification dite par pontage durant laquelle les fragments d'ADN liés à la flowcell créent des groupes de séquences nucléotidiques identiques appelés « clusters ». Cette étape d'amplification clonale vise à augmenter le signal de fluorescence afin de faciliter l'acquisition des signaux par le système optique du séquenceur lors de l'étape de séquençage.

A la fin des cycles d'amplification, tous les brins anti-sens (reverse) sont clivés et éliminés de la flowcell par lavage afin de n'obtenir que des clusters composés de séquences sens (forward). Le primer de séquence se fixe ensuite sur le site de fixation de l'amorce de séquençage et une polymérase ajoute les quatre dNTP fluorescents au brin d'ADN à synthétiser en respectant la complémentarité des bases. Le fluorochrome lié permet de bloquer l'élongation de la séquence de sorte à ce qu'un seul nucléotide marqué puisse s'incorporer par cycle. Chaque base a des propriétés d'émission différentes permettant ainsi d'établir quel dNTP s'est fixé aux séquences composant chaque cluster. Le clivage du fluorochrome est ensuite réalisé, libérant ainsi le blocage de la synthèse et l'élongation suit ainsi son cours lors d'un nouveau cycle d'acquisition.

Si la chimie Illumina incluait dans un premier temps une chimie basée sur 4 couleurs pour le marquage des bases, le lancement des technologies NextSeq et MiniSeq ont introduit une nouvelle chimie basée sur l'utilisation de deux couleurs uniquement. Les nucléotides incorporés sont alors séparés selon la présence d'une des deux couleurs, des deux couleurs ou de l'absence de couleur lors de la réaction de séquence (absence de marquage des G).

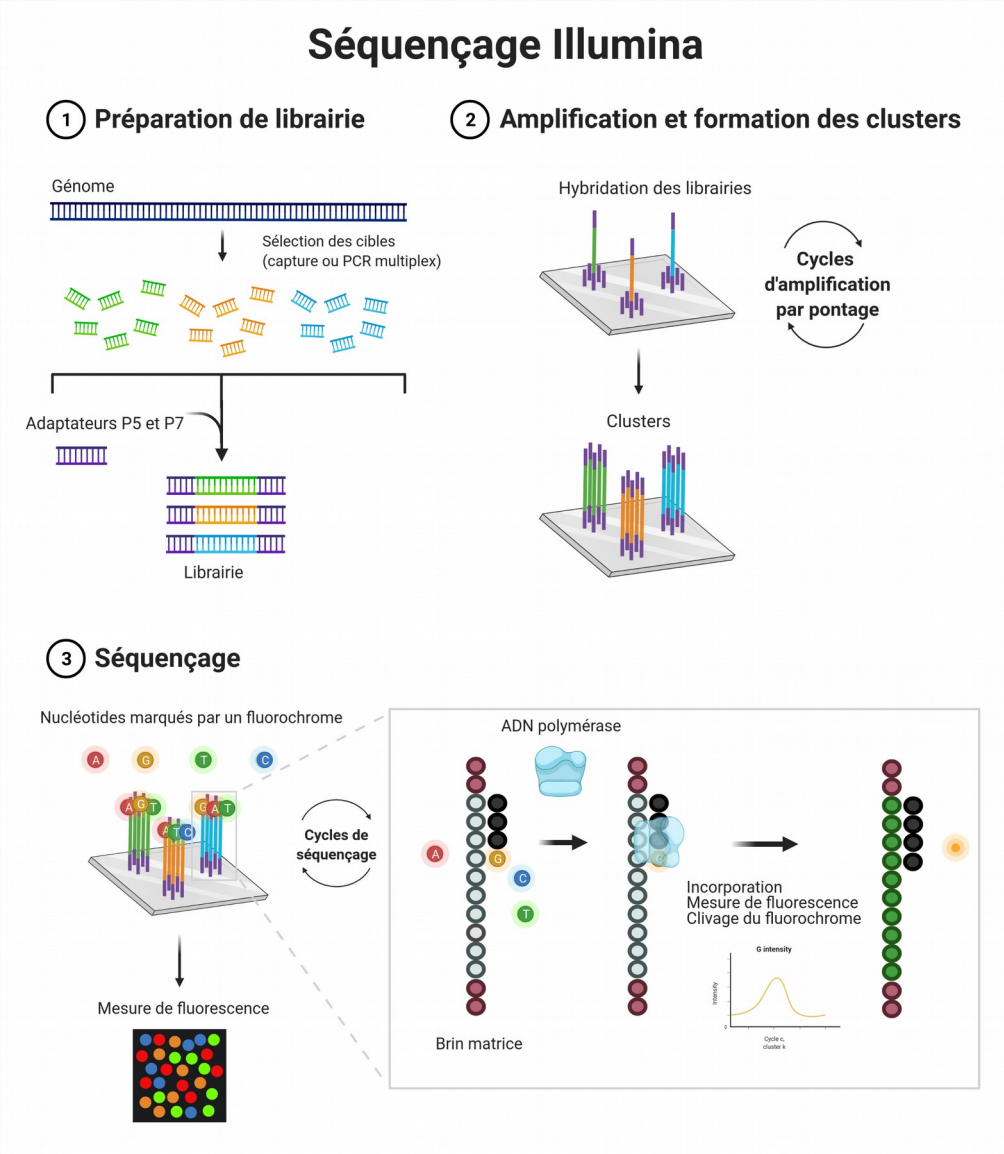


Figure 25: Séquençage Illumina

Les technologies Illumina permettent de reproduire la réaction d'amplification par pontage de sorte à séquencer dans un second temps le fragment anti-sens. Pour se faire, le séquenceur réalise l'étape d'amplification par pontage en éliminant cette fois-ci les brins sens afin de conserver que les brins anti-sens. Chaque cluster de séquences est alors analysé selon la même réaction de séquence que précédemment, conduisant à une deuxième lecture indépendante de chaque fragment de la librairie. Cette étape de séquençage, appelée « paired-end », conduit à la création d'un second FASTQ par échantillon.

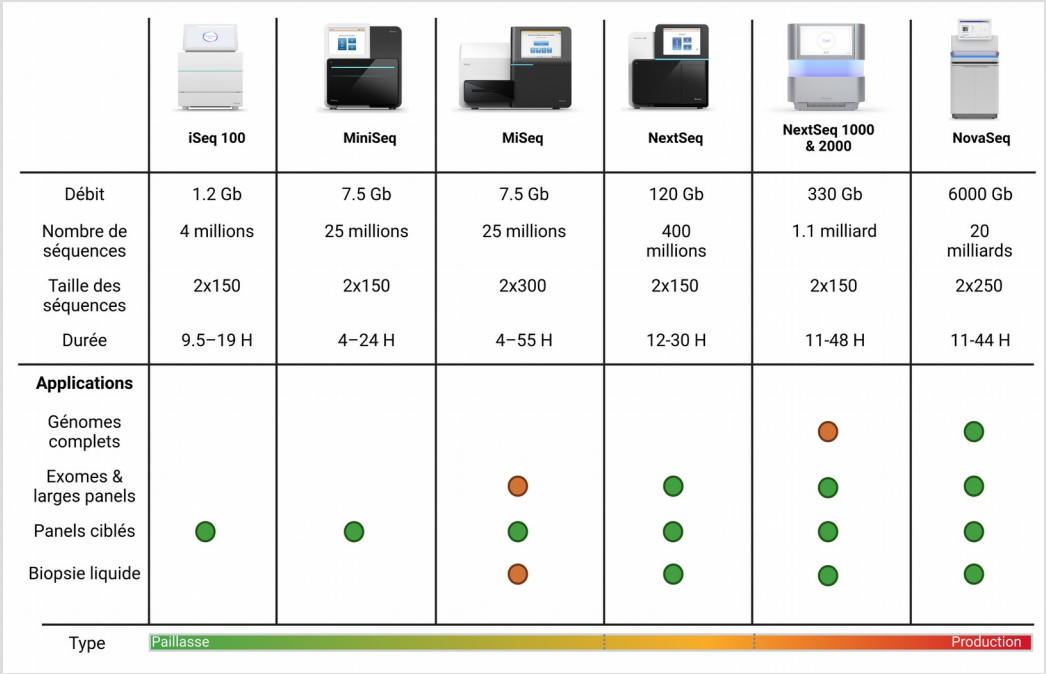


Figure 26: Gamme de séquenceurs Illumina et leurs principales caractéristiques.

Illumina dispose d'une gamme de séquenceurs relativement étendue qui peut être décomposée en deux grands types de produits : les séquenceurs dits de « paillasse » et les séquenceurs de « production » (figure 26). On retrouve principalement dans les laboratoires de diagnostic dédiés à la biologie moléculaire en cancérologie des séquenceurs de paillasse de type NextSeq 500, MiSeq ou MiniSeq. Le système NextSeq 500 permet de générer de 20 à 120 Gb de données en une seule analyse et assure une grande flexibilité à travers un nombre important d'applications allant du séquençage profond de panels ciblés de gènes jusqu'au séquençage de quelques exomes. Il permet de lire des séquences sur des longueurs allant de 75 à 300 pb en fonction des flowcells et des kits de construction de librairie utilisés. Il est tout particulièrement adapté au séquençage profond d'échantillons pauvres en contingent tumoral sur des panels de gènes et permet un gain de sensibilité important en comparaison du système MiSeq. Le système MiSeq sera en effet dédié presque exclusivement à des applications de séquençage ciblé dès lors que la profondeur souhaitée n'est pas trop importante et les panels de séquençage relativement restreints. Il permet de générer entre 0.5 à 15 Gb de données par expérience avec des lectures comprises entre 60 et 600 pb en fonction là encore des consommables utilisés et des librairies construites.

Enfin, la société Illumina développe et commercialise des séquenceurs dits de production incluant le NextSeq 550, les NextSeq 1000/2000 ou encore le NovaSeq 6000. Ils permettent de générer de 400 millions de séquences pour le NextSeq 550 jusqu'à 20 milliards de

séquences pour le NovaSeq en une seule flowcell. Si des premiers centres hospitaliers s'équipent aujourd'hui de NextSeq 550 avec le développement d'analyses nécessitant de générer beaucoup de données, les séquenceurs NextSeq 1000/2000 et NovaSeq 5000/6000 sont principalement retrouvés dans des grands centres de séquençage centralisant des analyses pour des projets de recherche européens et internationaux ou qui traitent des volumétries importantes d'échantillons par semaine.

C. Le concept de biopsie liquide

L'analyse informatique des données de santé, et plus particulièrement des données omiques, couplée à l'émergence de nouveaux traitements ciblés, conduisent à une médecine de plus en plus personnalisée. Celle-ci a pour objectif de caractériser toujours mieux les tumeurs via l'identification de biomarqueurs mais aussi de suivre leur évolution afin de proposer aux patients le bon diagnostic, le bon traitement et au bon moment de l'évolution de la maladie.

Si la prise en charge des patients reposait jusqu'à présent, sur le plan de la biologie, principalement sur l'analyse des biomarqueurs identifiés dans la tumeur à la suite d'une biopsie, le concept de « biopsie liquide » est un domaine de la recherche qui semble particulièrement prometteur et intéressant dans les années à venir. Il regroupe un ensemble d'examens biologiques réalisés non plus sur une biopsie mais à partir d'un fluide biologique comme un échantillon sanguin ou encore de l'urine.

Cette nouvelle approche possède des avantages non négligeables pour la prise en charge des patients. Premièrement, elle pourrait permettre de remplacer dans certaines conditions des examens invasifs et parfois douloureux (biopsie, ponction, chirurgie) par une simple prise de sang afin de poser le diagnostic. De plus, là où généralement une seule biopsie est réalisée au moment du diagnostic de la maladie, la biopsie liquide permet de répéter les analyses tout au long de la prise en charge du patient afin de mesurer l'évolution de la maladie. Les cancers sont par définition des maladies qui évoluent au cours du temps et qui entraînent bien souvent des lésions multiples, comme par exemple l'atteinte de plusieurs ganglions à différents endroits du corps dans les lymphomes. Les prélèvements répétés par biopsie liquide pourraient permettre aux biologistes et aux médecins de disposer d'une photographie de l'ensemble de ces populations tumorales à un moment précis de la prise en charge du patient.

Enfin, la réalisation d'une biopsie n'est pas toujours possible notamment dans des maladies de localisation particulière comme le lymphome cérébral [150]. L'ADN extrait de la biopsie n'est pas non plus la source biologique la plus informative dans certaines pathologies comme le lymphome de Hodgkin (voir chapitre I.A.4) où la cellule tumorale de Reed-Sternberg ne représente que 0,1 à 2 % de la masse tumorale [118], [151]. Dans ces conditions particulières le séquençage de l'ADN extrait à partir de biopsie liquide pourrait être une véritable alternative au recours à la biopsie. D'autres fluides biologiques que le sang peuvent être analysées en fonction de la localisation de la tumeur comme l'urine dans le cancer de la vessie

ou encore le liquide cérébro-spinal dans les tumeurs cérébrales [150], [152]. Le plasma reste cependant la fluide le plus largement étudié.

Nous décrirons dans ce chapitre la découverte de l'ADN tumoral circulant, ses principales caractéristiques et une remise en perspective des principales technologies publiées pour l'analyse des biopsies liquides. Les informations apportées dans ce chapitre ont conduit à la publication d'une revue [153] en 2021 dans le journal *Pharmaceuticals*. La publication est visible en annexe V.A.

C.1. Découverte de l'ADN circulant

L'ADN circulant (cell free circulating DNA, cfDNA) désigne des fragments d'ADN qui ne sont plus contenus dans le noyau des cellules mais dans des fluides biologiques comme le plasma, l'urine ou le liquide cérébro-spinal.

Les acides nucléiques du plasma sanguin chez l'Homme,

par P. MANDEL et P. MÉTAIS.

Sujet	Sexe	Age	Affection	P phospho-protéine mg.	P ribonu-cléique mg.	P desoxyri-bonu-cléique mg.	P total acides nucléiques mg.
1	F	42	Normal	0	5,0	1,2	6,2
2	F	22	»	0	4,0	0,4	4,4
3	H	24	»	0	5,2	1,3	6,5
4	F	27	»	0	4,7	0,3	5,0
5	F	20	»	0	3,7	0,8	4,5
6	H	48	»	0	4,6	1,3	5,9
7	H	45	»	0	4,5	0,6	5,1
8	F	26	»	0	5,0	0,2	5,2
9	F	37	»	0	4,8	0,6	5,4
10	H	39	»	0	5,0	0,9	5,9
11	H	62	Insuffis. card.	0	3,8	0,7	4,5
12	H	62	»	0	3,8	0,45	4,25
13	H	42	»	0	5,1	0,9	6,0
14	F	33	Endocard. maligne	0	3,35	0,65	4,0
	»	»	»	0	3,5	0,8	4,3
15	H	19	Goître	0	5,6	0,4	6,0
16	F	5	Basedow	0	3,6	0,3	3,9
17	H	48	Diabète	0	3,6	0,4	4,0
18	H	61	»	0	3,5	0,4	3,9
19	H	48	Cirrhose	0	5,3	1,2	6,5
20	F	52	Ictère	0	3,6	0,4	4,0
21	H	48	Goutte	0	3,5	1,0	4,5
	»	»	»	0	2,66	0,8	3,46
22	H	33	Goutte	0	5,5	0,5	6,0
	»	»	»	0	4,75	0,75	5,5
23	H	26	Néphrite	0	3,75	0,7	4,45
24	H	37	Tuberculose	0	3,5	0,45	3,95
25	F	23	Grossesse 7e m.		7,65	1,35	9,0
	»	»	»		7,25	1,00	8,25

Figure 27: La première identification d'ADN extracellulaire circulant à partir de prélèvements sanguins par Mandel et Metais en 1948.

Le cfDNA fût décrit pour la toute première fois en 1948 par Mandel et Metais [154] à partir de plasmas d'individus sains ou atteints d'affections non cancéreuses (diabète, tuberculose). Par la suite, des études démontrèrent que la concentration en cfDNA était plus importante dans des conditions pathologiques comme les maladies auto-immunes [155] et chez les patients atteints de cancers [156].

En 1989, Philippe Anker et Maurice Stroun de l'Université de Genève furent les premiers à démontrer que les fragments de cfDNA portent les mêmes caractéristiques que l'ADN issu des cellules tumorales [157]. Enfin, David Sidransky et son équipe de recherche américaine publient en 1991 dans Science la toute première description des mutations du gène *TP53* dans le cancer de la vessie à la fois sur la biopsie tumorale mais aussi sur des échantillons d'urine de trois patients à partir de nouvelles techniques de PCR et de clonage [158]. Au cours de cette étude, cette équipe soulève déjà l'implication que pourraient avoir de tels résultats pour la prise en charge individuelle des patients. Cette recherche d'anomalies génétiques ciblées dans l'ADN circulant s'est ensuite accentuée avec la publication d'essais sur les mutations des gènes *NRAS* et *KRAS* ou encore des amplifications du gène *HER-2* [159]–[161]. C'est à l'issue de ces premières études qu'apparaît pour la première fois la notion d'ADN tumoral circulant. Cette notion de part tumorale de l'ADN circulant (circulating tumor DNA, ctDNA) parmi l'ensemble des fragments d'ADN circulant (cfDNA) sera reprise dans toutes les études ultérieures.

C.2. Propriétés

Il existe de nombreuses variétés de cfDNA parmi lesquels le cfDNA dérivé de l'ADN mitochondrial (cell-free mitochondrial DNA, mtDNA), celui dérivé de la tumeur (ctDNA) ou encore celui dérivé par exemple du fœtus (fetal-derived cfDNA, fdDNA).

Si les sources de cfDNA sont donc diverses, les fragments circulants semblent partager des propriétés communes. Ils sont tout d'abord dilués dans le plasma, majoritairement double brin et fragmentés entre 170 et 500 paires de bases (pb) [162]–[164]. Cette taille de 170 pb correspond au nombre de bases enroulées autour d'un cœur d'histones, appelé nucléosome, et donc protégées de la dégradation. La taille des fragments augmente en fonction du nombre de nucléosomes sur lesquels ils sont fixés : 300pb pour les dinucléosomes et 500 pb pour les trinucléosomes [165], [166]. Ces tailles plus importantes correspondent à des fragments provenant de la dégradation de l'ADN cellulaire par apoptose [167]. Les molécules d'ADN sont présentes de façon transitoire dans le sang avec une demie vie rapportée variant de 30 minutes à 2 heures selon les études [168]–[171]. La préparation de librairies de séquençage simples brins (single stranded DNA, ssDNA) a permis de mettre en évidence la présence de fragments plus courts entre 40 et 100 pb [172]. Enfin, certaines études ont rapporté la présence de fragments beaucoup plus longs (ultra-long cfDNA) avec des tailles rapportées de 80 000 pb chez des individus [173]. Les mécanismes de relargage de ces très grands fragments ne sont pas encore parfaitement décrits.

La concentration en cfDNA fluctue beaucoup que ce soient dans des conditions physiologiques chez des individus sains (~5-20 ng/ml) ou des conditions pathologiques (>500 ng/ml dans le cancer de l'œsophage) [174]. Cette concentration en cfDNA est corrélée au stade du cancer, avec des concentrations plus élevées dans les stades avancés, et à la taille de la tumeur. La part tumorale de cet ADN circulant (ctDNA) varie entre 0,01 % et plus de 90 % dans le sang [175].

La cinétique du ctDNA lors du suivi des patient est corrélée au pronostic, où une baisse importante de sa concentration est associée à un meilleur pronostic. Son augmentation au cours du traitement est souvent prédictif d'un échec des traitements de première ligne et peut être le témoin de l'apparition de clones tumoraux résistants [176]–[179]. La détection du ctDNA durant le suivi des patients, pour prédire par exemple précocement la rechute ou la réponse au traitement, reste un réel défi sur le plan technologique dans la mesure où la fraction d'ADN circulant issu de la tumeur peut représenter moins de 0,01 % de l'ADN circulant total extrait [180], [181].

Le développement des nouvelles technologies de séquençage qui sont de plus en plus sensibles permettent de détecter des anomalies somatiques avec des fréquences alléliques (VAF) de plus en plus faibles non seulement pour la détection des anomalies au diagnostic mais aussi pour le suivi de la maladie résiduelle à différents temps de la prise en charge du patient. Néanmoins, de nombreux paramètres peuvent influencer la qualité de détection des anomalies dans le cfDNA tant sur le plan pré-analytique que sur le traitement informatique des données séquencées.

C.3. Conditions pré-analytiques

Quelque soit la méthodologie ou la technologie d'analyse du cfDNA, un certain nombre de paramètres pré-analytiques vont considérablement impacter la qualité de détection des anomalies.

Des précautions toutes particulières doivent être prises concernant la dégradation de ces échantillons fragiles et notamment leur contamination par de l'ADN génomique suite à la lyse de cellules sanguines. Plus ce phénomène de lyse sera important et plus la proportion de ctDNA sera faible au regard de tout l'ADN extrait de l'échantillon biologique.

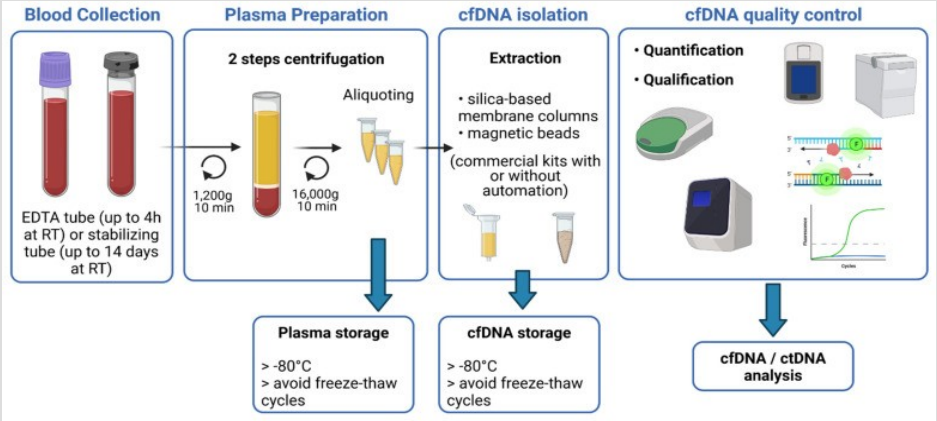


Figure 28: Représentation schématique du traitement des échantillons sanguins pour l'extraction du cfDNA [153].

La nature des anticoagulants ou des agents stabilisants contenus dans le tube dans lequel le sang est stocké avant l'extraction, le volume de plasma, la préservation du plasma ou encore le choix du kit d'extraction sont des éléments déterminants afin d'obtenir suffisamment de matériel pour pouvoir effectuer les analyses ultérieures. La figure 28 présente les principales étapes du traitement des échantillons sanguins jusqu'à l'extraction de l'ADN. Afin de diminuer le phénomène de lyse cellulaire des cellules sanguines, il est recommandé d'utiliser des tubes de collection (BCT, blood collection tube) qui procurent une stabilité suffisante du plasma. Plusieurs études ont comparé différents BCT, et notamment les tubes EDTA qui sont traditionnellement utilisés, aux nouveaux BCT permettant de stocker le plasma plus longtemps comme les tubes Streck (cfDNA BCT), Roche Diagnostics (Cell-free DNA collection tube), Qiagen (PAXgene Blood ccfDNA Tube) ou encore les tubes Norgen Biotel Corporation (cf-DNA/cf-RNA Preservative tubes). Ces derniers ont la particularité d'être dotés d'agents de conservation limitant les phénomènes d'hémolyse. Le sang dans les tubes EDTA classiques doit être traité très rapidement là où il est possible avec les autres BCT de les traiter plusieurs heures après le prélèvement en le conservant à température ambiante, sans affecter la qualité des résultats d'analyse [182]–[185]. Si les tubes EDTA classiques peuvent encore être utilisés pour des essais monocentriques, leur utilisation n'est pas compatible avec des essais multicentriques. Si les échantillons sanguins doivent être transportés, les nouveaux BCT sont très vivement recommandés.

Une fois l'échantillon sanguin prélevé, celui-ci doit être centrifugé afin d'en extraire le plasma. Là encore, cette étape est critique car elle affecte directement la concentration de cfDNA extrait. Le protocole aujourd'hui retenu pour prévenir le relargage d'ADN génomique

des cellules sanguines consiste à centrifuger l'échantillon en deux étapes : une première centrifugation lente afin d'éliminer les cellules sanguines (1200g / 4°C / 10min) sans les lyser, puis une seconde centrifugation à haute vitesse afin d'éliminer les débris cellulaires (12 000g-16 000g / 4°C / 10min) [186, p.], [187]. La fraction plasmatique résultant de cette dernière centrifugation peut alors être aliquotée puis stockée à -80°C.

Le rendement de l'extraction du cfDNA à partir du plasma est aussi dépendant des kits d'extraction utilisés. De nombreux kits ont été testés dans différentes études, et en particulier les kits de Qiagen (QIAmp circulating nucleic acid kit/QIAmp min Elute ccfDNA mini kit), Promega (Maxwell RSC ccfDNA plasma kit), Applied Biosystems (Mag MAX cell-free DNA isolation kit) et Norgen Biotek (plasma cell-free circulating DNA purification midi kit). En fonction de ces kits, l'extraction peut être réalisée en colonne ou sur billes magnétiques. Certains sont compatibles avec l'utilisation d'automates. Les études comparatives de ces différents kits concluent tous à la supériorité des kits d'extraction Qiagen, que ce soit par extraction manuelle ou automatisée [188], [189].

Dans la mesure où les fragments de ctDNA ont des caractéristiques bien particulières concernant la longueur théorique des fragments post-extraction, de nombreuses études soulignent l'importance de vérifier la qualité et l'intégrité de l'ADN extrait. En effet, comme nous l'avons évoqué précédemment, les fragments de ctDNA sont très fragmentés avec des tailles comprises entre 20 et 220pb, avec une occurrence plus importante des fragments de 170 bp correspondant à la taille de l'ADN enroulé autour d'un nucléosome [163], [190]. Les méthodes fluorométriques classiques de dosage de l'ADN ne sont pas appropriées pour quantifier la concentration en cfDNA après extraction dans la mesure où elles ne sont pas en mesure de discriminer les fragments courts de cfDNA et les fragments plus longs d'ADN génomique (gDNA). A l'inverse, des méthodes d'électrophorèse capillaire sont en mesure d'évaluer l'abondance des fragments en fonction de leur taille [191], [192]. Des méthodes de qPCR et de PCR digitale (dPCR) sont en mesure d'évaluer l'intégrité et l'amplificabilité des échantillons extraits. En revanche, celles-ci sont biaisées par la présence de gDNA en amplifiant préférentiellement les fragments courts vis à vis des fragments longs et en surestimant, en conséquence, la concentration de cfDNA [193].

Une fois le plasma extrait de l'échantillon sanguin et l'ADN extrait et qualifié, les fragments de cfDNA peuvent être analysés par différentes approches en fonction de la question posée.

C.4. Méthodologies d'analyse

Analysis Type	Technique	Sensitivity (LoD)	Targets	Applications	Advantages	Limitations
qPCR	ARMS-PCR PNA-LNA Clamp PCR COLD PCR	0.01–0.1%	Hotspot mutation	Cancer detection and monitoring, targetable alterations, some assays approved for clinical use	High specificity and sensitivity, cost effective, rapid, ease of use	No multiplexing, limited to detection of known mutations
ddPCR BEAMing	digital PCR	0.01–0.1%	Hotspot mutations, gene fusions, CNV	Cancer detection and monitoring, targetable alterations, some assays approved for clinical use	Up to 5 targets, high sensitivity and specificity, absolute quantification, single molecule analysis, cost effective, rapid, ease of use	Limited multiplexing (number of fluorescent colors), limited to detection of known mutations
PCR coupled to spectrometry	SERS UltraSEEK	0.1–1%	Known mutations	Cancer detection and monitoring, targetable alterations, for research use	Multiplexing capacity	Limited to detection of known mutations
Targeted NGS	Tam-Seq	2%	Known and unknown mutations, indels, CNV, chromosomal rearrangements (capture)	Cancer detection and monitoring, classification, targetable alterations, for research use	High specificity	Amplicon methods by multiplex PCR (depend on fragment size), no error correction
	eTam-Seq	0.02%			Error correction	Amplicon methods by multiplex PCR
	Safe-SeqS	0.01–0.05%			Error correction by SSCS	Amplicon methods by multiplex PCR
	Duplex sequencing	0.0001–0.1%			Error correction by DSCS	Amplicon methods by multiplex PCR
	TEC-Seq	0.05–0.1%			Error correction by SSCS, Hybrid capture method (not dependent on fragment size)	Less comprehensive than WGS or WES
	single primer extension (SPE)	0.5–1%			Amplicon methods by SPE (not dependent on fragment size), error correction by SSCS	Less comprehensive than WGS or WES
	SPE-duplex UMI	0.1–0.2%			Error correction by DSCS	Less comprehensive than WGS or WES
	CAPP-Seq	0.02%			Hybrid capture method (not dependent on fragment size)	Need large input, allelic bias (capture), stereotypical errors (hybridization step), less comprehensive than WGS or WES
	iDES eCAPP-Seq	0.00025–0.004%			Error correction by DSCS and correction of stereotypical errors	Less comprehensive than WGS or WES
	Ig-HTS	0.001%	VDJ rearrangements	Non-invasive monitoring, approved for clinical use	Very high sensitivity	Tissue biopsy needed
WES	Untargeted	5%	Coding regions, intron-exon junctions, promoters, untranslated regions, non-coding DNA of miRNA genes	Cancer detection, monitoring of resistant clones in metastasis, for research use	Mutation discovery and signatures, detection of CNV, fusion genes, rearrangements, predicted neoantigens and Tumor Mutational Burden	Low sensitivity (increasing depth lead to high cost), need bioinformatic expertise
WGS		5–10%	Structural variants (fragmentation pattern, genome-wide CNV, methylation profile)	Cancer localization and origin, early detection (early and late stage), for research use	Shallow sequencing, genome wide profiling, identification of cancer signatures	Expensive, variable sensitivity (low) and specificity, need bioinformatic expertise, lots of data generated

Figure 29: Méthodologies d'analyse du ctDNA et applications possibles [153].

Les technologies de séquençage du ctDNA sont très vastes et doivent être adaptées à la question biologique. L'évaluation du ctDNA peut passer par la recherche de quelques mutations par PCR ou conduire à la recherche d'anomalies sur le génome complet sur des séquenceurs à très-haut débit.

Le choix de la technologie dépendra de la fraction du génome humain ciblée et du niveau de sensibilité et de spécificité souhaité. Plus les cibles seront larges et plus le nombre de lectures séquencées devront être importantes sous peine de perdre en sensibilité lors de l'exploration des librairies de séquençage. Un éventail des différentes méthodes est visible sur la figure 29.

Nous nous concentrerons dans cette section sur les technologies d'analyse basées sur la dPCR et les NGS.

Analyse par PCR digitale

La dPCR consiste à réaliser une réaction de PCR, non plus sur un échantillon d'ADN complet, mais sur des ADN compartimentés par dilution limite de l'échantillon. De ce fait, pour un biomarqueur que l'on souhaite quantifier, chaque compartiment ne possédera que théorie que deux états : positif ou négatif. Ces états dépendront de la mesure de fluorescence de sondes portant des fluorophores dirigées contre l'anomalie somatique à quantifier. La dPCR apporte un gain de sensibilité important vis à vis de la PCR quantitative classique (qPCR) dans la mesure où chaque compartiment fera l'objet d'une PCR indépendante. Initialement réalisée en microplaques dès 1999 [194], la dPCR est aujourd'hui réalisée dans des microgouttelettes dont le volume de réaction de chaque compartiment est de l'ordre du picolitre au nanolitre.

La dPCR a pour objectif de quantifier des anomalies somatiques précisément définies et ponctuelles dans un échantillon biologique. Par exemple, dans les LDGCB, cette technique peut avoir un intérêt au diagnostic pour quantifier les mutations simultanées des gènes *MYD88* et *CD79B* dont le statut mutationnel est corrélé à la réponse à l'ibrutinib [36]. De même, dans les lymphomes cérébraux (PCNSL), la quantification de la mutation *MYD88* L265P dans le liquide cérébro-spinal des patients a montré de meilleurs résultats que par qPCR classique [195], [196]. Cette mutation étant présente dans 85 % des PCNSL et étant spécifique de cette entité, celle-ci peut permettre de confirmer un diagnostic à partir des résultats de dPCR. Néanmoins, la dPCR est limitée par la capacité de multiplexage des sondes fluorescentes (jusqu'à 5 simultanées) [197], [198].

La méthode BEAMing (beads, emulsion, amplification, magnetics) est une technique de dPCR combinant à la fois la PCR en émulsion et la cytométrie de flux pour identifier et quantifier des anomalies somatiques [199]. Diehl et al ont utilisé cette approche pour la quantification de mutations à partir de cfDNA dans une cohorte de patients atteints de cancers colorectaux et ont démontré que la cinétique du ctDNA est corrélée à la réponse au traitement et que la quantification du ctDNA post-chirurgie représente un marqueur de maladie résiduel pertinent pour le suivi des patients [200]. Cette méthodologie BEAMing a été très largement déclinée dans un grand nombre de cancers solides comme le cancer colorectal [201], le cancer du sein [202] ou le cancer du poumon [203].

La dPCR peut aussi permettre le suivi de remaniements structuraux tels que les CNV [204], [205] et de remaniements chromosomiques tels que les translocations. Ces translocations sont particulièrement intéressantes à suivre dans certains sous-types de

lymphomes, comme par exemple la translocation t(14;18) impliquant le gène *BCL2* dans les lymphomes folliculaires ou la translocation t(11;14) impliquant *CCND1* dans les lymphomes du manteau [206], [207].

La dPCR reste néanmoins limitée à la détection d'anomalies très ciblées. Elle nécessite une mise au point de chaque essai ce qui rend cette technologie difficilement applicable lorsque l'on s'intéresse à un panel de mutations ou au séquençage de gènes ayant des mutations perte de fonction non récurrentes.

Analyse par séquençage à haut-débit

Le séquençage ciblé à forte profondeur (TDS, Targeted Deep Sequencing) est lui aussi en pratique limité à un certain nombre de régions mais il peut couvrir l'intégralité d'un gène ou de ses parties codantes. Le TDS est donc adapté pour l'analyse de gènes n'ayant pas de mutations récurrentes.

Nous avons vu dans le chapitre qu'il existait plusieurs méthodes de préparation de librairie parmi lesquelles l'amplification directe par PCR, par capture ou par extension d'une amorce unique. Quelque soit la technologie choisie, la principale problématique du TDS est de descendre au maximum en sensibilité en discriminant de façon précise le bruit de fond de séquençage des mutations réellement présentes dans les échantillons pauvres en fragments d'ADN issus de la tumeur d'origine. Les technologies de séquençage à haut-débit ont par nature un taux d'erreur par base entre 0,1 et 0,5 %. Ces erreurs sont liées à la fois à la chimie du séquençage par synthèse sur les séquenceurs Illumina ou Ion Torrent, mais aussi aux technologies d'acquisition des signaux bruts.

De nouveaux protocoles de préparation des librairies ont permis des avancées majeures pour la détection et la quantification des anomalies [208], [209]. Les approches par TDS en particulier à partir de faibles quantités d'ADN augmentent considérablement le risque de relecture de molécules amplifiées ayant acquises des erreurs de polymérase lors des cycles d'amplification lors de la préparation des librairies. Ainsi, des avancées récentes ont permis l'introduction de barcodes moléculaires uniques, appelés UMI, avant toute étape d'amplification de sorte à pouvoir quantifier réellement chacun des événements observés et de déterminer s'il s'agit ou non d'un faux positif (voir section I.D). Le fait de quantifier le nombre de molécules uniques, et non le nombre de molécules amplifiées, permet une amélioration des profils mutationnels détectés et un gain de sensibilité important [210]–[212].

Nous décrirons brièvement dans la suite de cette section un ensemble de protocoles d'analyse applicables aux biopsies liquides : Tam-Seq (Tagged-amplicon deep sequencing), Safe-SeqS (Safe-Sequencing System), TEC-Seq (Targeted Error Correction sequencing), SPE (Single-Primer Extension) et enfin CAPP-Seq (Cancer Personalized Profiling by Deep Sequencing).

Tagged-amplicon deep sequencing (Tam-Seq)

Tam-Seq est une méthode de construction de librairies par amplicon dont les amorces de séquençage possèdent des barcodes. Une première pré-amplification des échantillons est effectuée suivie d'une amplification sélective des régions d'intérêt. Les adaptateurs de séquençage et les barcodes spécifiques des échantillons sont ensuite attachés à cette construction lors d'une dernière PCR. Cette méthode permet de détecter des mutations dans des échantillons de cfDNA avec une sensibilité et une spécificité supérieure à 95 % à des fréquences alléliques proches de 2 % [213]. Elle a récemment été améliorée (enhanced Tam-Seq, eTam-Seq) sur le plan du design des amorces pour permettre l'analyse d'échantillons très fragmentés d'une part et d'autre part par l'emploi d'un nouvel algorithme pour la détection des SNV et CNV qui permet de mieux considérer le bruit de fond de séquençage [214]. Les auteurs rapportent 94 % de mutations détectées à des VAF comprises entre 0,25 et 0,33 % et une limite de détection de 0,02 %. Enfin, une comparaison à la dPCR montre une bonne concordance démontrant ainsi que cette technique est applicable à la quantification d'anomalies dans le cfDNA [214].

Safe-Sequencing System (Safe-SeqS)

Safe-SeqS est aussi une méthode de préparation de librairie par amplicon décrite initialement par le groupe de Bert Vogelstein [211]. C'est la première approche à introduire des barcodes moléculaires pour améliorer la sensibilité des technologies de séquençage à haut-débit. Un barcode appelé UID (unique identifier) est assigné à chaque fragment d'ADN avant toute étape d'amplification. Une mutation est considérée comme présente si 95 % des séquences post-PCR arborant un même UID sont porteurs de l'anomalie. Cette stratégie offre une certaine souplesse de détection en éliminant les erreurs d'amplification et de séquençage. Les auteurs rapportent une limite de sensibilité de 0,05 %. Safe-SeqS a été particulièrement utilisé pour l'analyse d'échantillons de cfDNA dans des tumeurs solides, notamment dans certaines formes de carcinomes [215] et plus récemment en 2021 dans trois cohortes de patients opérés de cancers colorectaux [216].

Targeted Error Correction sequencing (TEC-Seq)

Comme dans l'approche Safe-SeqS, des barcodes moléculaires sont eux aussi utilisés dans cette approche TEC-Seq afin de discriminer les mutations des artefacts de séquençage. Néanmoins, cette approche combine l'information portée par les barcodes et les positions d'alignement des paires de lectures le long du génome de référence. Ces positions sont utilisées comme des « barcodes endogènes ». Les duplicats sont ainsi identifiés par une clef d'identification « barcode moléculaire-positions ». Cette méthodologie a été appliquée à différents types de cancers solides avec une sensibilité de 100 % et 89 % pour des VAF de 0,2 % et 0,1 %, respectivement [212].

Single Primer Extension (SPE)

Cette méthodologie SPE repose sur la création d'amplicons et est à la base des kits de préparation de librairie distribuée par la société QIAGEN (QIASeq targeted DNA panel kits). Elle utilise une seule amorce spécifique par région, appelée GSP, pour l'amplification des régions d'intérêt ce qui la rend moins dépendante de la taille des fragments que les technologies par amplicon classiques qui nécessitent la fixation de deux amorces. Comme pour les méthodes de capture, une première étape de fragmentation est réalisée suivie d'une étape de réparation des extrémités des fragments générés et par la ligation d'adaptateurs. Les UMI sont portés par ces adaptateurs universels qui servent à l'amplification des régions conjointement avec les GSP (figure 21). Ils sont de taille 12 et sont donc capables de générer 4^{12} combinaisons de clefs d'identification différentes réduisant ainsi drastiquement le risque de redondance [217]. Cette méthodologie a été utilisée au diagnostic afin d'identifier des cibles thérapeutiques ou des mutations ponctuelles [218], [219]. Elle permet aussi la détection de CNV via l'utilisation d'algorithmes spécifiques tels que mCNA [220] développé dans le cadre de cette thèse (chapitre III.C).

Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq)

CAPP-Seq est une technologie ultra-sensible de capture développée spécifiquement pour l'analyse de cfDNA. Il s'agit à la fois d'une méthodologie de construction de panel *in silico* et d'une méthode de préparation de librairie en soit.

La première étape consiste à interroger les bases de données de référence afin de déterminer une liste de mutations décrites dans une pathologie d'intérêt. Cette sélection de cibles conduit au design de sondes oligonucléotidiques biotinylées, appelées « Selector », visant à capturer des régions larges autour des anomalies. Le protocole est optimisé pour la

détection de faibles quantités d'ADN et les recommandations en terme de profondeur de séquençage en font par nature une technique très sensible [221], [222].

CAPP-Seq est une technique capable de détecter de nombreux types d'anomalies : des SNV, des insertions/délétions, des variants structuraux et des variations de nombre de copies de gènes. Initialement développée pour la détection d'anomalies somatiques dans le cancer du poumon, de nombreuses publications dans d'autres pathologies ont démontré son efficacité que ce soit dans les cancers solides ou dans les hémopathies malignes comme les LDGCB, les LH ou les LF [118], [179], [223]–[225]. Quelques travaux autour de l'analyse de ctDNA dans le contexte des lymphomes sont présentés dans la section I.C.5.

C.5. Lymphomes et biopsie liquide

Nous avons vu dans les précédents chapitres que les technologies de préparation de librairie et de séquençage sont nombreuses et tendent toutes à améliorer la qualité d'acquisition des données. Nous nous concentrerons dans cette section sur les principales publications associant biopsies liquides et lymphomes.

Biopsie liquide dans le lymphome diffus à grandes cellules B

Comme nous l'avons vu précédemment, les LDGCB sont en réalité une maladie hétérogène dans laquelle coexistent plusieurs entités : GCB et non-GCB (section I.A.2). L'analyse des profils mutationnels dans ces tumeurs a révélé les principales anomalies génétiques acquises dans les cellules tumorales en fonction du sous-type de LDGCB. Les différentes publications associant biopsie liquide et lymphomes tentent de retrouver ces signatures mutationnelles non plus à partir d'ADN extrait de la tumeur mais d'ADN extrait d'autres fluides biologiques, comme le plasma.

Corrélation entre les profils mutationnels de la tumeur et du ctDNA au diagnostic

En 2015, notre équipe publie une première comparaison du profil mutationnel entre tumeur et plasma au diagnostic d'une cohorte de 12 LDGCB [226]. Le panel de séquençage, baptisé Lymphopanel, a été conçu pour être suffisamment informatif sur un nombre de cibles limité afin de constituer un test réalisable en routine. Il couvre 34 gènes d'intérêt (87kb, 872 amplicons) via la technologie de préparation de librairie AmpliSeq. Le séquençage, réalisé sur séquenceur PGM (Ion Torrent), révèle la présence de mutations dans le cfDNA dans 11 cas sur les 12 séquencés. La concordance globale des profils mutationnels entre tumeur et plasma

appariés dans cette série est de 85 %. Un exemple de profils mutationnels appariés est donné en figure 30.

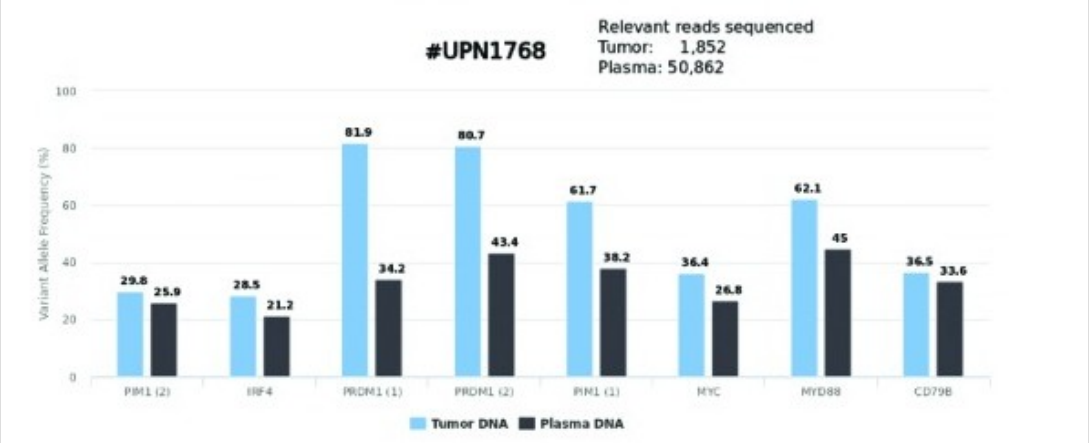


Figure 30: Corrélation entre les profils mutationnels du cfDNA et de la tumeur d'origine chez un patient atteint de LDGCB.

Les fréquences alléliques des variants dans la tumeur d'origine (en bleu) et dans le cfDNA (en noir) montrent une très bonne concordance. Adapté de Bohers *et al* [226].

La présence de mutations circulantes a été à nouveau confirmée par notre équipe dans une cohorte de lymphomes cérébraux [227]. Sur les 25 patients séquencés, 8 ont au moins une mutation détectable dans le plasma. La comparaison entre les profils mutationnels de la tumeur et de la biopsie liquide montre une bonne concordance dans cette petite série de patients (figure 31). Néanmoins, certaines mutations présentes à de faibles VAF dans la tumeur d'origine ne sont pas retrouvées dans le plasma des patients soulevant ainsi un problème de sensibilité.

Parallèlement au séquençage à haut-débit, notre équipe s'est aussi intéressée à la détection des profils mutationnels du cfDNA par dPCR. Trois essais de dPCR ont été développés pour détecter trois mutations récurrentes dans les LDGCB et PMBL : la mutation *XPO1* E571K, la mutation *EZH2* Y641N et enfin la mutation *MYD88* L265P [228]. La comparaison avec les données de séquençage du LymphoPanel sur PGM chez 15 patients et de la plateforme de dPCR QuantStudio3D® (Life Technologies) montre une concordance de 100 % entre les deux technologies et une corrélation presque parfaite sur la fréquence allélique des anomalies ($r=0,99$, $p<0,001$). Par ailleurs, des dilutions sériées de l'ADN muté des lignées cellulaires MedB-1, VAL, et OCI-Ly3 à 50%, 10%, 5%, 1%, 0.5%, 0.1%, 0.05%, et 0% montrent une limite de détection à 0,05 % de la dPCR.

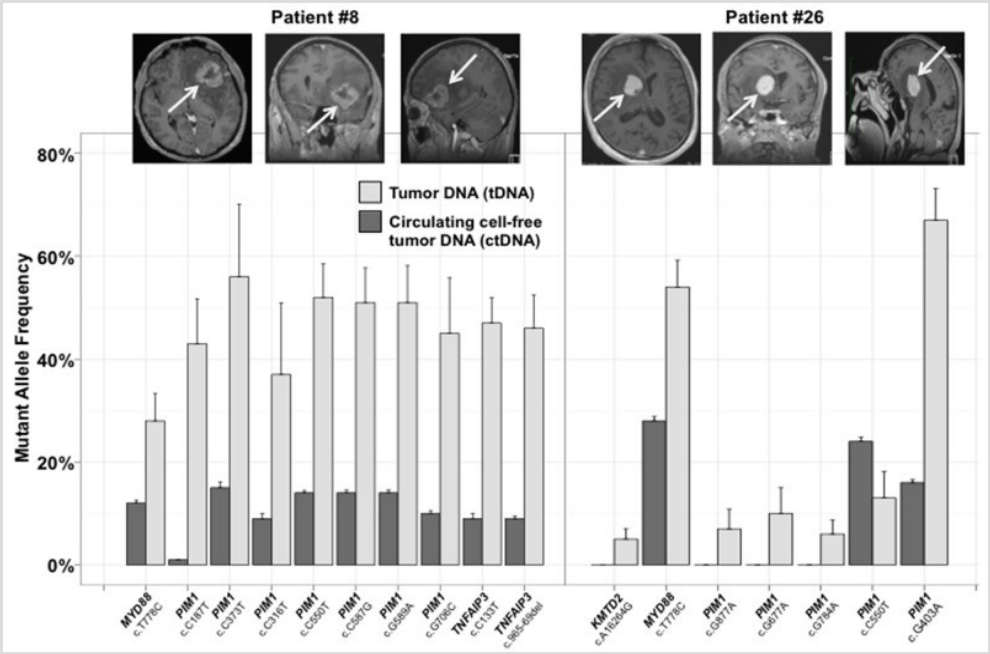


Figure 31: Corrélation entre les profils mutationnels du ctDNA et de la tumeur dans deux cas de lymphomes cérébraux.

Enfin, une dernière étude cette fois-ci prospective est conduite par notre équipe en 2018 afin de valider la valeur clinique de la détection de mutations dans le cfDNA dans une cohorte de 30 LDGCB. Les profils mutationnels du ctDNA et de la tumeur sont comparables dans cette série avec une informativité de 93 %. Dans 5 cas, des mutations additionnelles sont retrouvées dans le ctDNA par rapport à la tumeur.

Cette étude souligne pour la première fois qu'il est possible de détecter des variations de nombre de copies de gènes dans le ctDNA pour les patients ayant une charge tumorale importante dans le plasma avec des VAF importantes (figure 32). Les CNV détectés dans le cfDNA sont comparables à ceux retrouver dans la tumeur chez ces patients. Néanmoins, la positivité des profils est fortement dépendante de la fréquence allélique moyenne des mutations retrouvées dans le plasma et on observe ainsi beaucoup de profils qui ne sont pas exploitables faute de concentration de ctDNA.

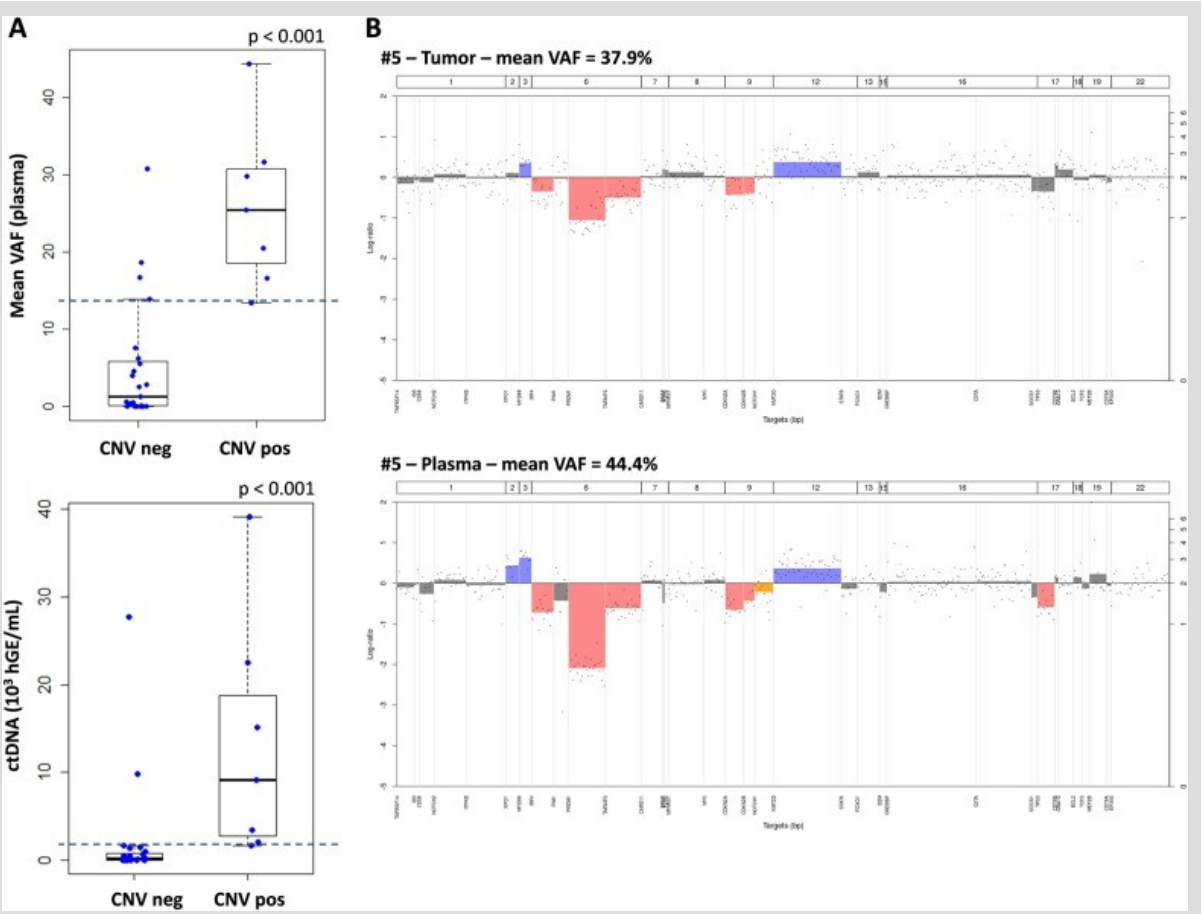


Figure 32: Détection de CNV à partir de données de séquençage du ctDNA dans le LDGCB.

Caractérisation des profils mutationnels par CAPP-seq

L'étude de Scherer *et al* s'intéresse à la détection d'anomalies dans le cfDNA d'une cohorte de 92 patients atteints de LDGCB et de 24 individus sains à partir de la méthodologie CAPP-Seq [224]. Le panel de séquençage ciblé de cette étude couvre environ 240 000 pb et inclut les mutations récurrentes décrites dans la littérature dans le LDGCB, les points de cassure récurrents à l'origine de fusions (*BCL2*, *BCL6*, *MYC* et de *IGH*) et enfin les mutations dans les chaînes lourdes variables des immunoglobulines (*IgVH*) et dans les régions de jonction (*IgJH*). L'informativité dans les tumeurs via ce panel ciblé, c'est à dire le pourcentage d'échantillons ayant au moins une mutation détectable, est de 100 % avec une médiane de 134 anomalies détectées. La concordance avec la FISH pour la détection des translocations dans les tumeurs est de 89 %. L'analyse des plasmas avant traitement montre une informativité de 100 % dans le ctDNA et une spécificité de 99.8 % par comparaison au profil mutationnel de la tumeur au diagnostic.

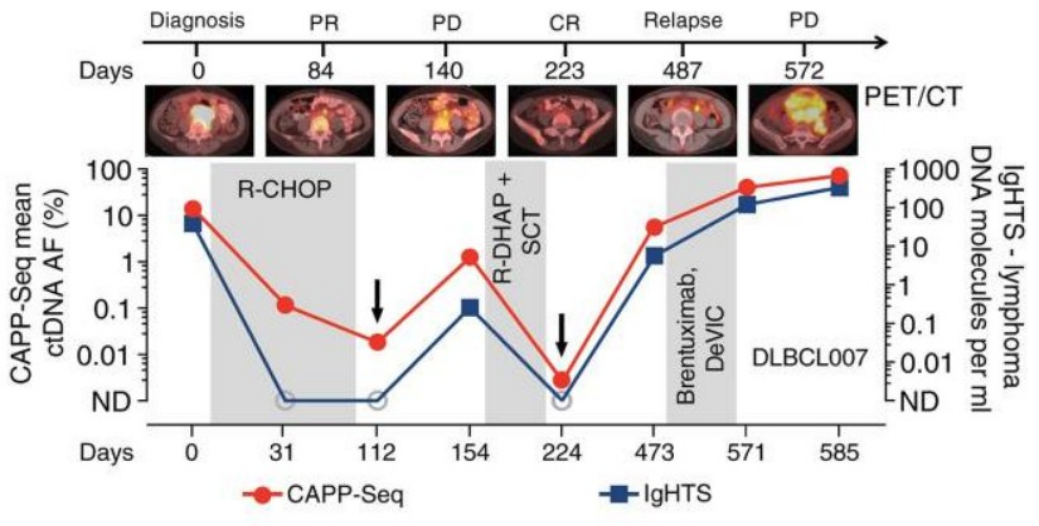


Figure 33: Cinétique de la VAF moyenne dans le ctDNA au cours du traitement.

Le graphique, extrait de Scherer *et al [176]*, décrit l'évolution de la VAF moyenne des anomalies détectées dans le cfDNA d'un patient au cours de son traitement et en lien avec l'imagerie. On observe une décroissance de la VAF moyenne associée à une décroissance du volume tumoral jusqu'au J224 avant d'observer le phénomène inverse lors de la rechute du patient.

La cinétique de la fréquence allélique moyenne des mutations dans le cfDNA au cours du traitement montre une corrélation nette avec l'imagerie où la décroissance du volume tumoral est liée à la diminution de la VAF moyenne dans le plasma et où à l'inverse son augmentation est liée à une augmentation de la charge tumorale (figure 33).

Un cas intéressant de lymphome avec des mutations de résistance à l'Ibrutinib sur le gène *BTK* est rapporté dans cette étude avec l'apparition de deux sous-clones tumoraux résistants (figure 34) au cours du traitement. Cette étude soulève l'impact significatif de la positivité du ctDNA après traitement sur le risque de rechute sur 25 patients. Enfin, un classifieur basé sur les profils mutationnels du ctDNA est construit afin de retrouver la classification COO GCB et non-GCB. Celui-ci montre une concordance intéressante avec l'algorithme immunohistochimique de Hans (figure 7) de 83 % pour les GCB et 94 % pour les non-GCB.

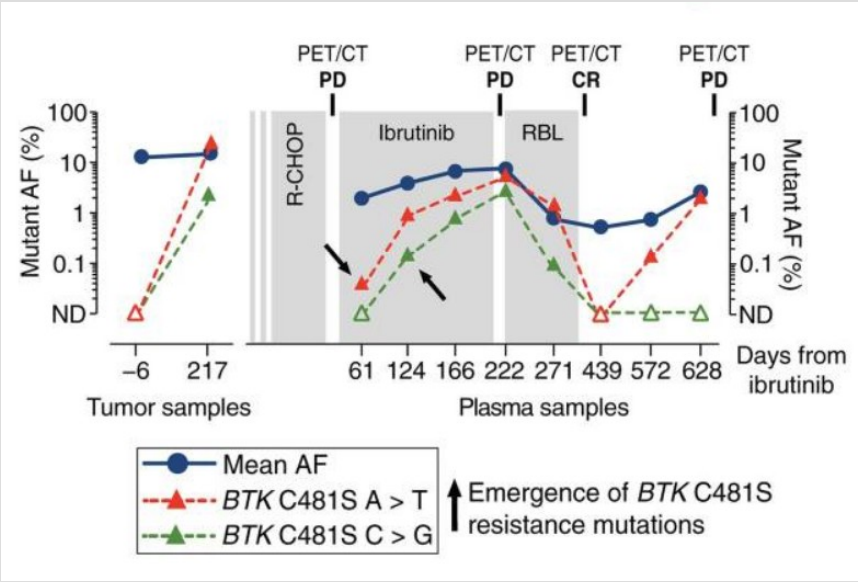


Figure 34: Émergence de sous clones résistants à l'Ibrutinib.

En 2018, Kurtz *et al* décrivent les résultats sur un nouveau panel CAPP-Seq sur plus de 200 LDGCB dans le but de définir des seuils pronostics basés sur la concentration du ctDNA en cours de traitement [179]. Cette nouvelle étude est particulièrement intéressante car elle est multicentrique : les échantillons proviennent de 6 centres hospitaliers. Une informativité de 98 % du cfDNA avant traitement est observée dans cette cohorte.

Les auteurs stratifient une première série des patients ayant une « réponse moléculaire rapide » (early molecular response, EMR) avec une décroissance de 2-log de la concentration en ctDNA après un cycle de chimiothérapie et en « réponse moléculaire majeure » (major molecular reponse, MMR) avec une décroissance de 2,5-log après deux cycles. Les seuils définis dans cette première série d'entraînement pour la classification des patients en EMR et en MMR sont appliqués sur une cohorte de validation. Ces deux classes EMR et MMR sont retrouvées comme étant pronostiques que ce soit sur le risque de rechute (EFS) ou sur le risque de décès (OS) (figure 35).

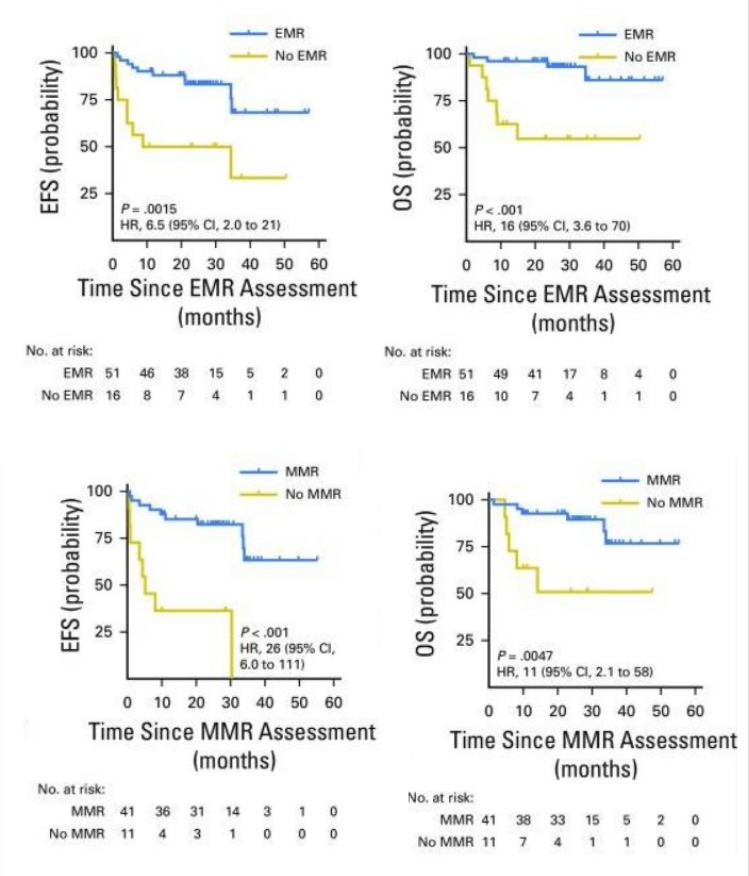


Figure 35: Impact pronostic de la cinétique du ctDNA en cours de traitement.

Les courbes adaptées de Kurtz *et al* [195] représentent l'impact pronostic significatif des classes EMR (early molecular response) et MMR (major molecular reponse) sur la survie sans progression et sur la survie globale.

Biopsie liquide dans le lymphome de Hodgkin

Nous avons vu précédemment que le lymphome de Hodgkin est caractérisé par la présence souvent peu abondante de cellules tumorales dans les biopsies appelées cellules de Reed Sternberg et que ce faible contingent tumoral pose beaucoup de difficultés pour déterminer le profil moléculaire des cellules de RS (section I.A.4). La détection de marqueurs au diagnostic dans des échantillons de cfDNA est donc particulièrement intéressante dans cette pathologie tant la biopsie est difficile à analyser sur le plan génétique. Des techniques de dPCR et de NGS ont permis de mettre en évidence que le cfDNA était informatif dans le LH et que les profils mutationnels entre tumeur et biopsie liquide était tout à fait comparables [118], [125], [229], [230].

La cohorte de cfDNA de LH la plus complète est vraisemblablement la série publiée en 2018 par V. Spina et al. On y retrouve une collection de 80 échantillons au diagnostic pré-traitement avec un suivi longitudinal en cours de traitement, en cas de rechute et en fin de traitement. Les tumeurs micro-disséquées ont elles aussi été séquencées afin de comparer les profils mutationnels de cellules de RS aux profils issus du ctDNA. Au total, ce sont près de 350 échantillons de cfDNA qui ont été séquencés par la technologie CAPP-seq.

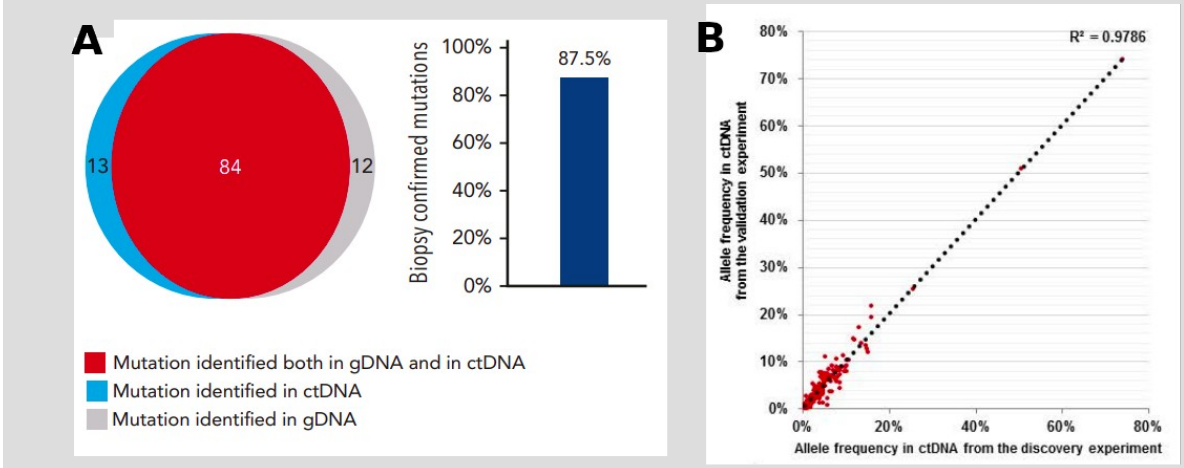


Figure 36: Validation technique de la technologie CAPP-Seq sur les échantillons de LH (adapté de V. Spina et al, Blood 2018).

A. Comparaison des mutations retrouvées entre échantillons tumoraux (gDNA) et de cfDNA appariés au diagnostic. 87.5 % des mutations retrouvées dans la biopsie liquide sont confirmées par l'analyse CAPP-seq de la biopsie appariée.

B. Corrélation entre les fréquences alléliques des mutations non synonymes détectées dans les échantillons de cfDNA entre le réplicat 1 (axe x) et le réplicat 2 (axe y).

La réalisation de duplicats, c'est à dire d'échantillons séquencés plusieurs fois et analysés par le même traitement bioinformatique, montre une concordance des résultats importante (R^2=0.978) validant l'approche CAPP-seq dans cette pathologie (figure 36.B). On retrouve dans cette série une informativité globale du cfDNA supérieure à 80 % au diagnostic (figure 36.A) sur un panel de 191kb couvrant les régions codantes et les bornes d'épissage exoniques de 77 gènes. L'analyse au diagnostic des profils mutationnels du cfDNA montre des résultats concordants avec les données précédemment publiées avec des mutations récurrentes de *STAT6*, *TNFAIP3*, *ITPKB*, *GNA13* ou encore *B2M* (Figure 37).

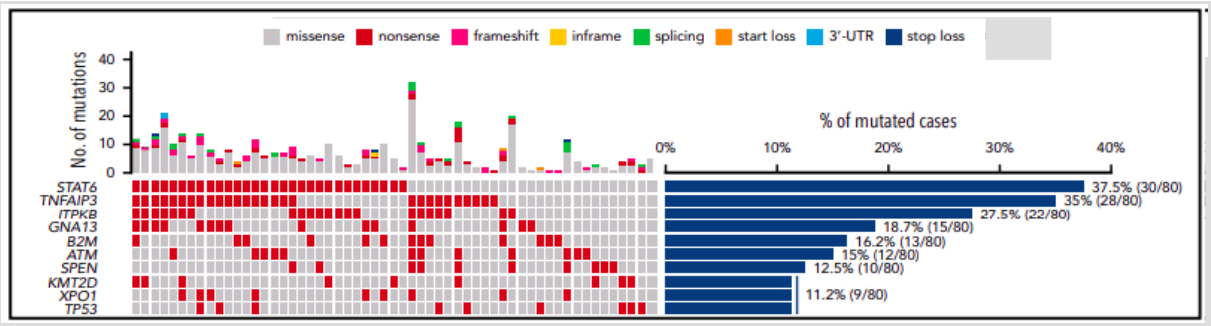


Figure 37: Profils mutationnels de 80 cfDNA de LH (adapté de V. Spina et al, Blood 2018)

La heatmap adaptée de l'étude de V. Spina et al. a été adaptée pour ne laisser apparaître que les 10 gènes les plus mutés dans la série.

D. Apport des barcodes moléculaires uniques

D.1. Biais inhérents au séquençage NGS

Nous avons vu dans les chapitres précédents que le séquençage de nouvelle génération nécessite au préalable une étape d'amplification des matrices d'ADN dont on souhaite déterminer la séquence (section I.B.3).

Cette amplification n'est pas anodine puisqu'elle brise la relation directe entre le nombre de molécules d'ADN initiales extraites de l'échantillon biologique et le nombre de lectures séquencées. Le facteur d'amplification de chaque fragment d'ADN est dépendant de nombreux facteurs tels que la taille de la librairie, le nombre de régions, le pourcentage en GC, la longueur des fragments ou encore l'éventuelle compétition entre deux amorces ciblant une même région (figure 38). Ce facteur d'amplification, dans des librairies sans UMI, n'est pas directement mesurable.

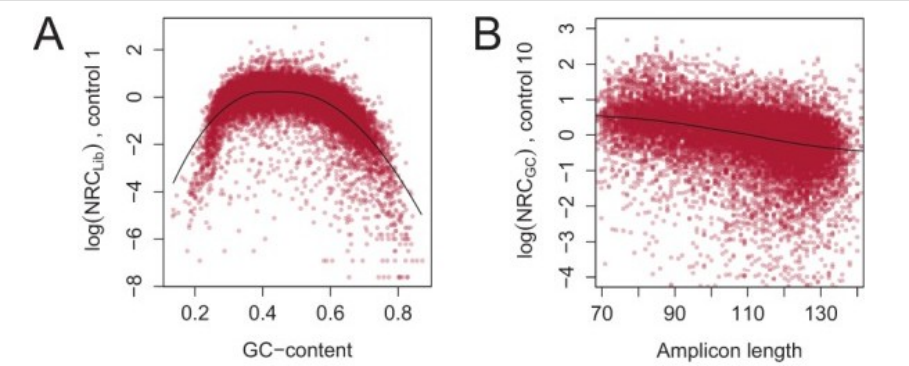


Figure 38: Biais lors du comptage des lectures des amplicons après séquençage NGS.

Cette figure extraite de la publication de l'outil ONCOCNV [288] montre les biais de comptage des lectures séquencées en fonction de paramètres tels que le pourcentage en GC de l'amplicon (A) ou la longueur de l'amplicon (B).

De plus, la détection des anomalies somatiques présentes à de faibles fréquences, à travers l'exploration profonde des échantillons par séquençage, implique de distinguer le signal biologique du bruit de fond technologique. Les erreurs de séquençage sont des facteurs confondants clefs pour maintenir une spécificité acceptable au regard de la recherche de sensibilité nécessaire pour certaines applications comme par exemple la détection non invasive de mutations dans des biopsies liquides.

Le taux d'erreur des appareils de NGS est décrit dans la littérature comme supérieur à 0.1 % [208], [210], [231], [232]. L'analyse bioinformatique primaire des données bruts des

séquenceurs génèrent des scores de qualité par base et par séquence appelés scores PHRED [233], [234]. Ce score reflète la probabilité d'attribution de la bonne base à partir des signaux d'acquisition et a pour but d'aider à l'élimination de ces erreurs. Néanmoins, ce système de score demeure en pratique insuffisant dès lors que l'on s'intéresse aux variations de très faible fréquence (<0,5%) pour déterminer avec précision les mutations réellement présentes dans un échantillon biologique.

Ce chapitre décrit en quoi l'utilisation des UMI dans les librairies de séquençage peut permettre d'améliorer la qualité des résultats.

D.2. Diversité

Les UMI sont des séquences courtes et aléatoires utilisées pour marquer de manière unique chaque fragment d'ADN avant toute étape d'amplification de librairie. L'ajout de ces UMI dépend de la nature des kits de construction de librairie utilisée et des fournisseurs.

La taille des UMI est à adapter au type d'application et à la profondeur de séquençage. La capacité de marquage des fragments d'ADN ciblés est directement proportionnelle à cette longueur avec une diversité théorique comme suit :

$$D = 4^{L}$$

Avec D = diversité, L=longueur de l'UMI

Par ailleurs, les fragments d'ADN sont identifiés à la fois par leur UMI mais aussi par leur séquence : de ce fait, deux fragments d'ADN n'ayant pas la même séquence et qui sont porteurs du même UMI seront considérés comme deux éléments distincts et indépendants par la plupart des algorithmes bioinformatiques. Ce couple UMI et séquence augmente donc encore très largement la diversité d'identification.

Les librairies QIAseq utilisées dans le cadre de cette thèse incluent des UMI de taille 12 et peuvent donc identifier plus de 16 millions de fragments uniques.

D.3. Exemples d'application

Compter individuellement des molécules d'ARN ou d'ADN est une tâche complexe car elles sont particulièrement difficiles à amplifier quantitativement. Les UMI sont une méthode de comptage absolue dès lors que les librairies de séquençage ont été explorées avec suffisamment de profondeur. Avec cette méthode, chaque molécule initiale est identifiée parmi une population de lectures. Compter le nombre de molécules en amont de la préparation

des librairies de séquençage revient à simplement déterminer le nombre d'UMI uniques. Ainsi, les UMI sont particulièrement utilisés dans les analyses nécessitant une quantification précise à partir des données issues du séquençage à haut-débit, comme en RNA-seq par exemple. Ils permettent d'éliminer les duplicats de PCR.

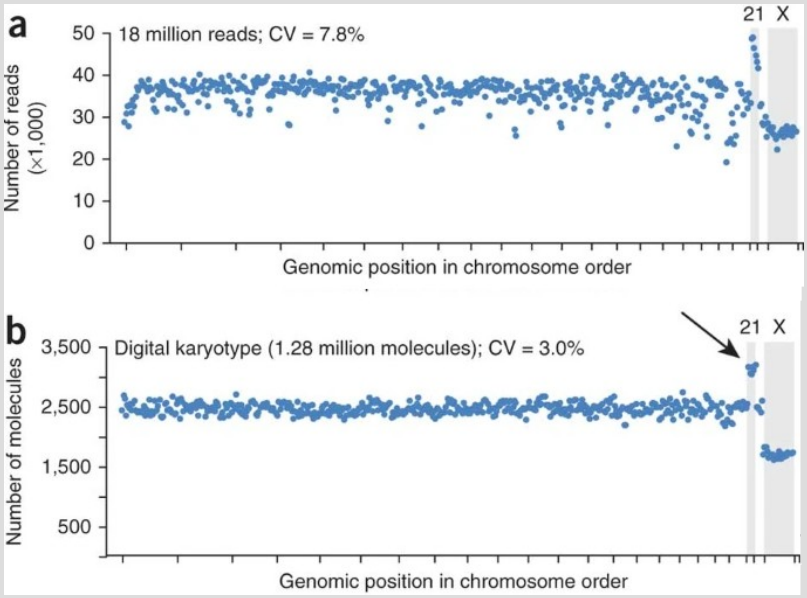


Figure 39: Caryotypage par comptage du nombre de barcodes moléculaires. Adapté de Kivioja et al. (2012, Nature) [235]

Cette figure montre en **a.** le comptage du nombre de lectures alignées par fenêtre de 5 Mb à partir du nombre de lectures. La figure **b.** représente ce même comptage réalisé cette fois-ci à partir du nombre d'UMI uniques par région, et donc du nombre de molécules d'ADN uniques. On observe un bruit de fond bien moins important sur les signaux d'acquisition avec le signal en UMI.

La première application en séquençage d'ADN remonte en 2012 dans l'étude de Kivioja et al dans la revue Nature [235]. Les auteurs se sont intéressés à évaluer la pertinence de l'utilisation des UMI notamment dans des problématiques de caryotypage en mélangeant l'ADN génomique d'un enfant atteint d'un syndrome de Down[4] et de sa mère biologique. Après marquage par les UMI et amplification par PCR, 20 millions de lectures ont été alignées et comptées dans une fenêtre glissante de 5 Mb. Le comptage des lectures n'a pas permis de retrouver significativement la trisomie du chromosome 21 chez l'enfant (figure 39).

4 Le syndrome de Down est un trouble héréditaire présent à la naissance, causé par la présence de copies supplémentaires du chromosome 21. C'est un trouble complexe touchant la santé et le développement des enfants.

En réalisant cette fois-ci le comptage des UMI par région, et non des lectures amplifiées, la trisomie 21 était très clairement visible et le signal d'acquisition beaucoup moins bruité.

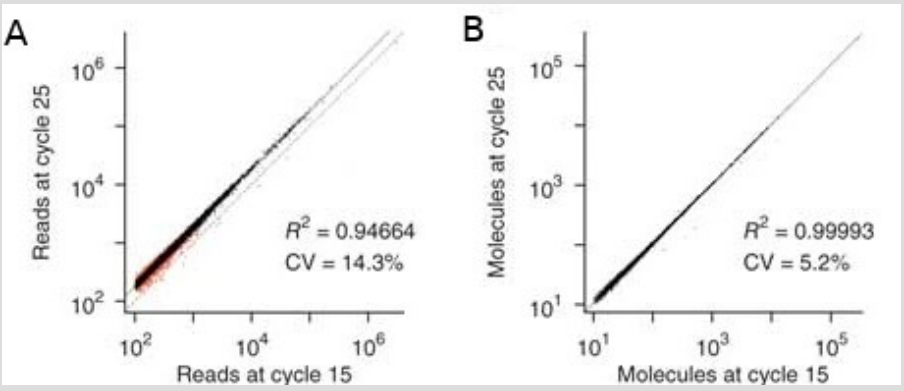


Figure 40: Comparaison de la quantification du nombre de ARNm par comptage des lectures ou des UMI et en fonction du nombre de cycles de PCR. Adapté de Kivioja et al. [235]

La même observation a été conduite sur la reproductibilité des comptages d'ARNm en RNA-seq. Les auteurs ont séquencé un même échantillon en faisant varier le nombre de cycles d'amplification et ont comparé les résultats de comptage. Là encore, les résultats sont significativement améliorés en réalisant les comptes à partir de l'information portée par les UMI (figure 40). Les données de comptage en UMI sont indépendantes du nombre de cycles d'amplification rendant ainsi possible, par exemple, une comparaison plus fiable des niveaux d'expression de différentes échantillons entre eux sans avoir recours à des méthodes de normalisation complexes.

Le choix d'introduire ou non des UMI dans les librairies de séquençage n'est pas anodin car il entraîne des modifications dans la manière de traiter les données. Les UMI doivent être extraits des séquences brutes et leurs séquences stockées dans un format informatique compatible avec les étapes de traitement primaire, secondaire et tertiaire des données. Ces différentes étapes de traitement sont présentées dans le chapitre suivant.

II. TRAITEMENT BIOINFORMATIQUE DES DONNÉES DE SÉQUENÇAGE POUR LA DÉTECTION DES VARIATIONS

Ce chapitre vise à introduire les différents traitements bioinformatiques pour la détection des anomalies somatiques à partir d'une expérimentation de séquençage NGS. Les analyses primaire, secondaire et tertiaire des données sont représentées sur la figure 41.

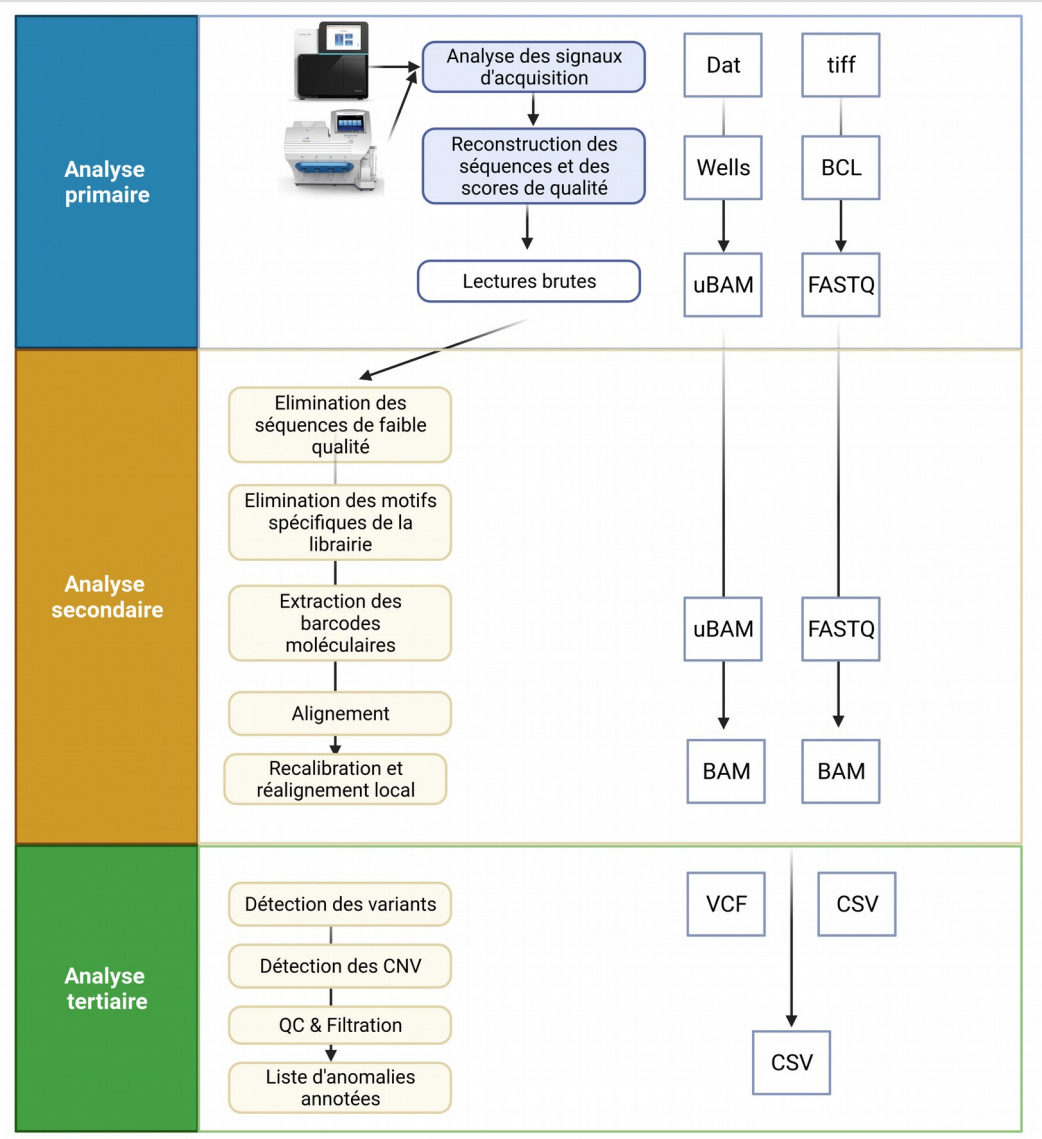


Figure 41: Traitement primaire, secondaire et tertiaire des données de séquençage NGS.

A partir des données brutes du séquenceur, une étape dite d'analyse primaire des données est réalisée afin de convertir les signaux d'acquisition, quelque soit leur nature, en une séquence ordonnée de nucléotides. En fonction des séquenceurs, cette étape implique l'utilisation d'algorithmes différents que nous détaillerons.

Une fois les séquences reconstruites, l'analyse secondaire des données peut être réalisée. Cette étape inclut différents pré-traitements appliqués aux séquences afin d'éliminer les motifs liés à la chimie de préparation de la librairie de séquençage. L'alignement est ensuite réalisé afin d'associer chaque lecture à une position chromosomique. Nous verrons les différents algorithmes intervenant lors de l'analyse secondaire et notamment ceux utilisés dans le cadre du traitement des données au cours de cette thèse : UMI-tools [236], BWA [237] et GATK [238].

Enfin, nous introduirons les concepts autour de l'analyse tertiaire des données et plus spécifiquement la détection des variants et des CNV à partir des lectures alignées. Nous évoquerons le concept d'annotation et de contrôle qualité pour la filtration des variants.

A. Analyse primaire

Le traitement primaire des données consiste à transformer les signaux acquisition des séquenceurs NGS en une séquence ordonnée de nucléotides. Ces signaux d'acquisition, quelque soit la nature de la technologie de séquençage, sont générés de façon massivement parallèle et nécessitent l'utilisation d'algorithmes performants souvent embarqués dans des programmes propriétaires comme bcl2fastq© ou encore le Torrent BaseCaller©. La source brute d'acquisition diffère en fonction de la nature des séquenceurs : il s'agit d'une mesure pH-métrique pour la technologie Ion Torrent ou d'une mesure de fluorescence pour la technologie Illumina.

Lors de la reconstruction des séquences, un score de qualité est attribué à chacune des bases. Ce score PHRED est proportionnel à la qualité des signaux d'acquisition et traduit la probabilité d'identification correcte de la base. PHRED est un algorithme qui a émergé lors du Human Genome Project (HGP), projet durant lequel la comparaison des données de séquençage de différentes plateformes a été nécessaire afin de réaliser pour la première fois l'assemblage de la séquence de référence du génome humain [233], [234]. Historiquement, PHRED évaluait la forme et la résolution des pics d'électrophorèse des séquenceurs de première génération puis, à partir d'immenses tables de correspondance propres à chaque type de séquenceur, ces différentes mesures étaient converties en probabilité d'erreur de séquençage (p) puis en score de qualité (Q) comme suit :

$$Q = -10 \times \log(p)$$

L'algorithme PHRED est toujours implémenté sur les séquenceurs de nouvelle génération. Les scores de qualité PHRED sont utilisés par un très grand nombre d'algorithmes bioinformatiques pour estimer la qualité des séquences, éliminer des portions de séquences de faible qualité, déterminer l'exactitude des séquences assemblées ou encore évaluer le caractère artéfactuel d'une variation génétique observée. C'est un élément indispensable pour discriminer le bruit de fond de séquençage du signal biologique dans des échantillons pauvres en contingent tumoral.

Nous décrirons dans la suite de ce chapitre les différents algorithmes et formats de fichier utilisés pour le traitement primaire des données de séquençage Ion Torrent et Illumina.

A.1. Technologie Ion Torrent

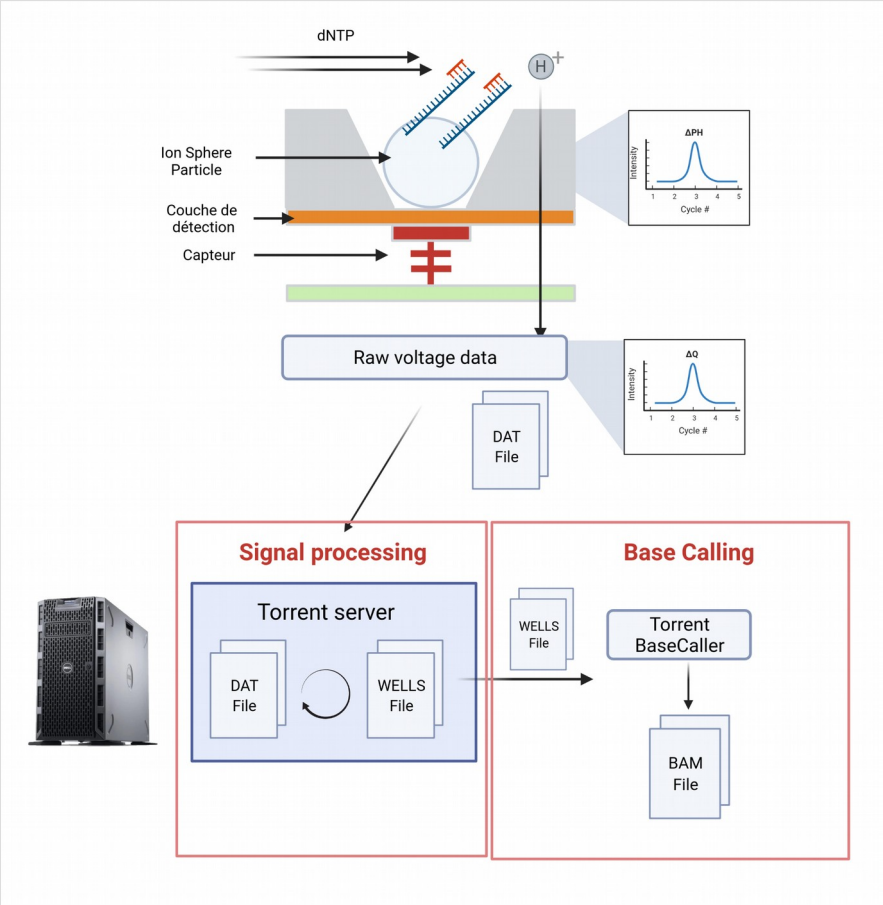


Figure 42: Traitement primaire des données des séquenceurs Ion Torrent

Le traitement primaire des données des séquenceurs Ion Torrent repose sur la transformation des données d'intensité électrique au format DAT. Ces fichiers contiennent les intensités électriques par puits et par cycle au cours du temps. Pour rappel, cette technologie de séquençage repose sur l'injection à chaque cycle d'un dNTP à la surface de la puce puis d'un lavage. Un exemple de signal brut est présenté sur la figure 43. On retrouve un découpage des cycles d'acquisition pour chaque puits en trois phases : une intensité nulle avant le premier cycle, un pic lors de l'incorporation d'un dNTP puis une baisse progressive lors du lavage en fin de cycle. Une intensité résiduelle persiste au cours du temps après l'incorporation du dNTP.

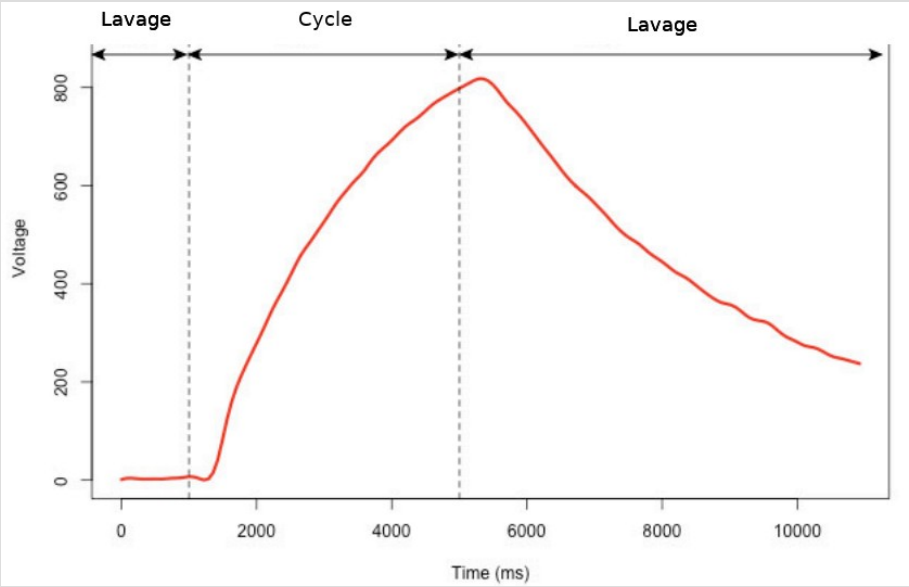


Figure 43: Ion Torrent - Exemple de signal d'acquisition contenu dans un fichier DAT.

La courbe représente la mesure du signal électrique d'un puits au cours d'un cycle et suite à l'incorporation d'une base dans une ISP.

Il est important de noter qu'à cette étape un modèle multi-paramétrique est entraîné à chaque cycle afin de soustraire le bruit de fond d'acquisition. Pour se faire, le programme détecte le puits qui n'a pas incorporé de base le plus proche afin d'extraire l'intensité résiduelle de celui-ci et appliquer une correction par soustraction au puits en cours d'analyse [239]. L'application de ce modèle linéaire de correction aux plusieurs millions de puits par puce et à chaque cycle est l'étape la plus coûteuse en ressources de calcul au regard de tous les traitements bioinformatiques en aval. L'introduction dans les dernières versions de la Torrent Suite (1.4) d'une carte graphique NVIDIA Tesla GPU pour effectuer ces calculs a rendu extrêmement plus rapide cette étape de normalisation des données.

L'étape de *base calling* vise à convertir les signaux d'acquisition corrigés du séquenceur en une base nucléotidique prédite. En amont de cette étape, il n'existe aucun lien entre l'intensité électrique déduite des fichiers DAT/WELLS et le nombre de bases réellement incorporées au cours du cycle. Il est donc nécessaire de calibrer ces signaux d'intensité de sorte à reconstruire la séquence. Lors de la construction des librairies, les fragments d'ADN liés aux ISP intègrent un adaptateur de 7 nucléotides TACGTAC appelé « key sequence » (KS) [239]. Cette séquence synthétique est utilisée afin de réaliser la calibration des signaux électriques lors des 7 premiers cycles de séquençage. Les cycles 0, 2, 3 et 5 dans l'ordre d'injection des dNTPs correspondent à des signaux d'acquisition vides au niveau de la KS et

les cycles 1, 4 et 6 à des signaux pleins d'acquisition d'une seule base. La méthode de normalisation mesure donc l'intensité d'acquisition moyenne des 4 cycles vides afin de déterminer le bruit de fond résiduel, soustrait cette intensité moyenne à tous les puits en cours d'analyse et détermine l'intensité moyenne d'incorporation d'une base à partir des cycles non vides. Finalement, les intensités de tous les puits sont divisées par l'intensité moyenne mesurée pour les cycles pleins d'une base. C'est ainsi que l'algorithme de *base calling* parvient à déterminer si une intensité d'acquisition élevée correspond ou non à l'incorporation de plusieurs nucléotides (2-mer, 3-mer…) durant un même cycle. L'algorithme intègre, au fur et à mesure des cycles, une surveillance des valeurs d'intensité d'incorporation d'une base afin d'adapter ce seuil de détection en cours de séquençage.

Enfin, une correction de phase est intégrée. En effet, il existe deux sources d'erreurs majeures pouvant fausser complètement le résultat de séquençage d'une ISP :

- une extension incomplète (IE) : situation durant laquelle un nucléotide injecté à la surface de la puce devait s'intégrer au brin en cours d'extension fixé sur une ISP mais ne l'a pas été (faux négatif)

- une extension non spécifique (ES) : situation durant laquelle un nucléotide s'est intégré anormalement à l'un des brins en extension d'une ISP de façon non spécifique (faux positif)

Pour rappel, chaque ISP de chaque puits est tapissée d'un ensemble de séquences identiques. Ces phénomènes IE/ES vont conduire à un décalage de phase entre une population de fragments d'une ISP ayant suivie normalement l'incorporation des nucléotides au cours des cycles et une fraction minoritaire d'entre eux ayant un décalage de cycle et venant augmenter le bruit de fond d'acquisition localement au niveau de l'ISP. Ce phénomène, bien que minoritaire statistiquement, s'aggrave à mesure que les cycles de séquençage se poursuivent. C'est l'une des raisons à la limitation en taille des fragments séquençables via cette approche.

Enfin, le logiciel de *base calling* utilise un algorithme glouton afin de trancher l'incorporation ou non d'une base [239]. Un algorithme glouton est un algorithme qui suit le principe de faire, étape par étape, un choix optimum local de paramètres, dans l'espoir d'obtenir un résultat global optimum. Très concrètement, des seuils sont déterminés cycle par cycle afin de statuer sur l'incorporation d'une base (1-mer), de deux bases (2-mer) … Et ces seuils évoluent à mesure de la réaction de séquence.

Le logiciel de *base calling* de la technologie Ion Torrent produit par défaut un fichier BAM non aligné (unaligned BAM, uBAM) pour chaque échantillon séquencé. Il est aussi possible d'installer des modules sur le Torrent Server afin de convertir ce fichier uBAM en un format plus standard comme le format FASTQ.

A.2. Technologie Illumina

Le traitement primaire des données des séquenceurs Illumina repose sur la détection des *clusters* sur le support de séquençage appelé *flowcell* et la transformation des signaux d'intensité de fluorescence de ces clusters en une suite ordonnée de bases.

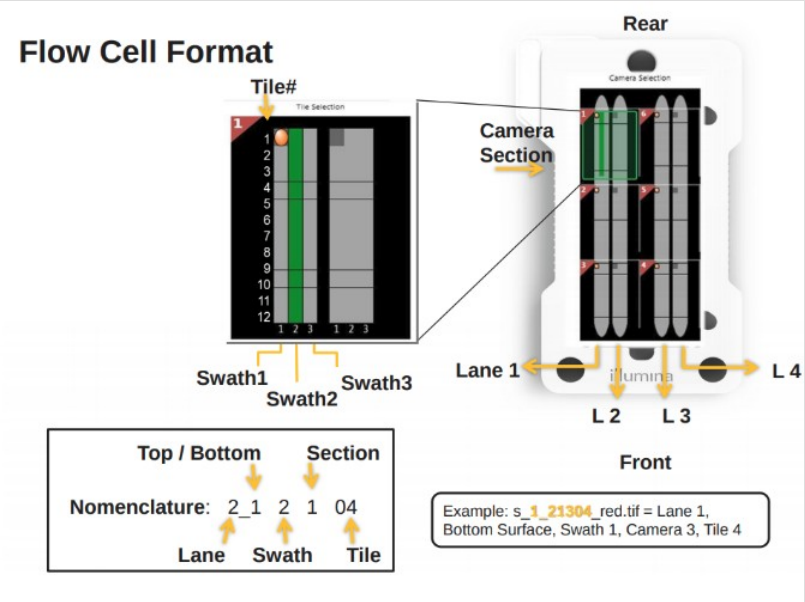


Figure 44: Illumina - Découpage fonctionnel des flowcells.

Cette figure a été adaptée du site de la société Illumina (http://illumina.com).

En fonction de la nature des *flowcells* et du type de séquenceur, celles-ci sont prises en charge par un système de plusieurs caméras réalisant les différentes prises à chaque cycle. Par exemple, les *flowcells* du système MiSeq sont divisées horizontalement en zones appelées « tiles » (tuiles) et en colonnes appelées « lanes ». Chacune des *lanes* est elle-même divisée en bandes plus étroites appelées *swaths* (figure 44). A noter que sur les séquenceurs de production de la gamme Illumina certaines *flowcells* ont la particularité d'être composées de deux faces (*bottom* et *top*). En conséquence, à chaque cycle, le système optique prend deux fois plus d'images par cycle en adaptant son focus pour mesurer la fluorescence de l'une des deux faces puis de la suivante.

Le traitement primaire des données des séquenceurs Illumina est piloté par un logiciel appelé RTA (Real-Time Analysis module). Ce logiciel exécute une suite de dépendances afin de convertir les images en séquences nucléotidiques par cluster à la surface de la *flowcell*. Le traitement repose en réalité sur l'exécution de deux modules : Firecrest et Bustard (figure 45).

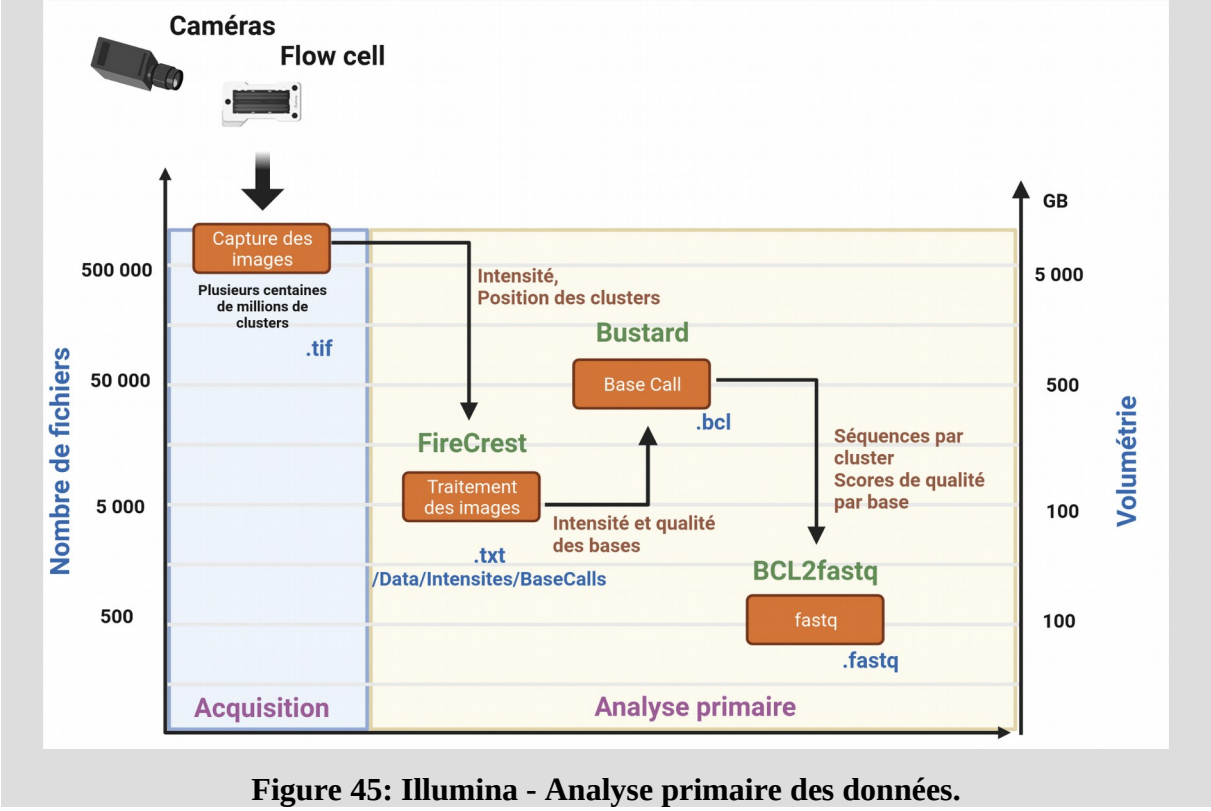


Figure 45: Illumina - Analyse primaire des données.

Firecrest est le module dédié à l'analyse des images haute-résolution au format TIF prises par les caméras des séquenceurs. Les images ont une nomenclature bien spécifique qui inclut les coordonnées d'acquisition à la surface de la *flowcell* (figure 44). Les *clusters* sont identifiés spatialement par le numéro de *Lane*, le numéro de la *Tile* et par leurs coordonnées dans l'espace (X;Y) ou (X;Y;Z) pour les *flowcells* ayant deux faces. Firecrest détermine dans un premier temps les positions des *clusters* à partir des premiers cycles de séquençage et extrait ensuite pour les cycles suivants les intensité de fluorescence pour chaque cluster. Firecrest applique ensuite différents filtres sur les images afin d'accentuer le contraste entre les différents clusters et éliminer le bruit de fond d'acquisition. Finalement, on obtient pour chaque cycle les intensités mesurées.

Bustard est le module utilisé afin de réaliser l'étape de *base calling* à proprement parler. A partir des mesures d'intensité réalisées par Firecrest, Bustard produit des fichiers BCL (binary

base call) par *Lane* et par cycle. Chaque BCL contient pour chaque cluster la base retenue pour le *cluster* ainsi que le score de qualité de la base.

Enfin, un utilitaire développé par la société Illumina appelée BCL2fastq est utilisé afin de transformer ces fichiers BCL en un format plus adéquat pour les analyses ultérieures : le format FASTQ. Ce programme est aussi à l'origine du démultiplexage, c'est à dire à l'attribution des séquences au bon échantillon selon les paramètres stipulés dans le fichier de configuration du run qui liste les barcodes attribués à chaque échantillon (*samplesheet*).

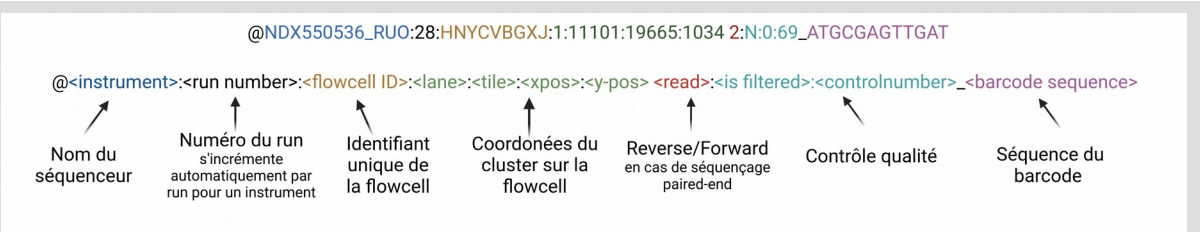


Figure 46: Illumina – Nomenclature de l'identifiant de séquence dans le fichier FASTQ.

Chaque entrée dans le fichier FASTQ est composée de 4 lignes comprenant un identifiant, la séquence nucléotidique, un délimiteur « + » suivi d'un score de qualité donné pour chacune des bases de la séquence au format ASCII.

Chaque identifiant de séquence reprend l'intégralité des caractéristiques extraits des différentes étapes du traitement primaire des données comme la localisation de la séquence sur la *flowcell* ou encore quelques paramètres de qualité (figure 46). Les scores de qualité PHRED par base sont encodés dans une forme compacte au format ASCII. A chaque symbole de cet encodage correspond un Qscore compris entre 0 et 40 dans la dernière actualisation du format.

Une fois les FASTQ générés pour chaque échantillon, les séquences sont prêtes pour l'analyse secondaire de la chaîne de traitement bioinformatique.

B. Analyse secondaire

L'analyse secondaire des données de séquençage consiste à produire des fichiers de séquences alignées contre le génome de référence et à appliquer des algorithmes pour la détection des variants ponctuels et structuraux.

B.1. Pré-traitement des séquences brutes

Généralités

Un pré-traitement des séquences brutes est parfois nécessaire en amont de l'alignement de séquences de sorte à éliminer les portions de séquence résultant de la construction des librairies (adaptateurs internes, UMI…) ou les régions de faible qualité en fin de séquence. Cette étape, appelée *trimming*, est essentielle pour assurer une bonne qualité d'alignement notamment pour des échantillons fortement dégradés, c'est à dire ayant des fragments d'ADN courts dans la librairie de séquençage. La présence de portions de faible qualité ou induite par la chimie de construction de la librairie va entraîner, en fonction de la longueur de ces séquences artefactuelles, un certain nombre de discordances (*mismatch*) lors de l'alignement. Les conséquences peuvent être nombreuses allant de l'apparition de faux variants jusqu'à l'impossibilité pour les algorithmes d'alignement de placer la séquence sur le génome de référence. Il n'existe pas de méthode générale adaptée en tout temps pour réaliser les étapes de *trimming*. Le *trimming* est fonction de la construction de la librairie et de la nature des échantillons analysés.

Il existe de plusieurs algorithmes de *trimming* disponibles dans la littérature. Certains incluent à la fois la suppression des séquences d'adaptateurs et des bases de faible qualité comme Cutadapt [240] ou Trimmomatic [241]. Ces deux programmes sont très largement utilisés dans la communauté bioinformatique. Cutadapt a l'avantage de proposer une implémentation parallélisable, c'est à dire pouvant répartir les tâches de calcul sur différents cœurs d'un ou plusieurs processeurs, mais ne prend pas en charge en une seule exécution les deux FASTQ R1 et R2 produits lors d'un séquençage paired-end. Il doit être lancé indépendamment sur chacun d'entre eux. Trimmomatic de son côté prend en charge les lectures pairées mais fonctionne sur un seul cœur de calcul ce qui le rend difficilement utilisable pour l'exploration profonde d'échantillons. Le temps de calcul est directement proportionnel au nombre de séquences à traiter.

Si les outils sont nombreux, le choix de l'outil de *trimming* à appliquer aux données n'est pas chose aisée. En effet, si chaque algorithme au moment de sa publication offrait les meilleures performances, il n'existe que peu d'études comparant sur un même jeu de données l'efficacité de ces programmes [242]. Les différences observées résultent bien sûr à la fois de la nature des algorithmes implémentés dans ces programmes mais aussi et surtout de tous leurs paramètres d'exécution et de la nature des constructions de librairie.

Les outils de *trimming* sont assez généralistes et ne sont pas toujours adaptés aux constructions de librairie complexes comme celles proposées par Qiagen et nécessitent des développements d'algorithmes locaux. Dans la mesure où ce type de librairie a été très largement utilisé dans les travaux de recherche présentés dans cette thèse, nous décrirons dans la suite de cette section l'implémentation d'un algorithme de *trimming* appliqué à ces librairies.

Pré-traitement des librairies QIAseq

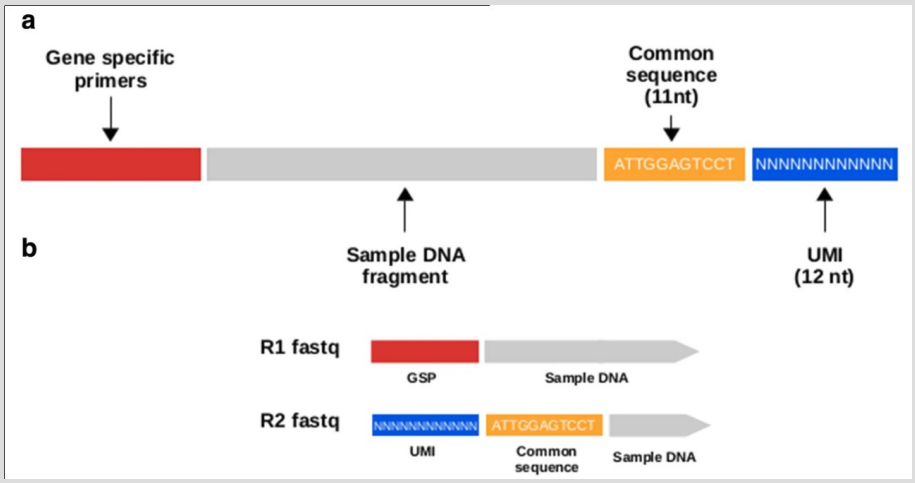


Figure 47: Construction des librairies QIAseq.

a. Représentation de la structure des séquences ; **b.** Représentation de la structure des séquences dans les fichiers FASTQ R1 et R2.

Les séquences brutes dans les FASTQ provenant des librairies QIAseq sont composées d'un GSP, du fragment d'ADN d'intérêt de l'échantillon, d'une séquence dite commune de 11 nucléotides suivie enfin d'un UMI (figure 47). On retrouvera ainsi à l'issu du séquençage dans les FASTQ R1 en 5' la séquence des GSP suivie de la séquence d'ADN à aligner et en R2 en 5' la séquence de l'UMI suivie de la séquence commune (CS) et de la séquence à aligner.

Si la CS est constante et donc facilement identifiable dans la structure des séquences, la séquence du GSP est spécifique de la région ciblée lors de la construction du panel de séquençage. En fonction des panels et du nombre de cibles, ces GSP peuvent représenter plusieurs centaines de séquences synthétiques différentes.

Le *trimming* est spécifique de chaque FASTQ R1 ou R2. Il consiste à éliminer les séquences de GSP et de CS et à extraire et stocker la séquence des UMI dans le nom des séquences dans les fichiers FASTQ. Une difficulté supplémentaire est soulevée par cette

construction de librairie : la séquence de l'UMI n'est lue que dans le FASTQ R2. Dans la mesure où les algorithmes d'alignement identifient les lectures pairées par leur nom dans les FASTQ, le fait d'ajouter la séquence de l'UMI uniquement dans les séquences du FASTQ R2 brise ce lien. Il sera donc nécessaire d'ajouter la séquence de l'UMI déduite du FASTQ R2 dans le nom des lectures du FASTQ R1.

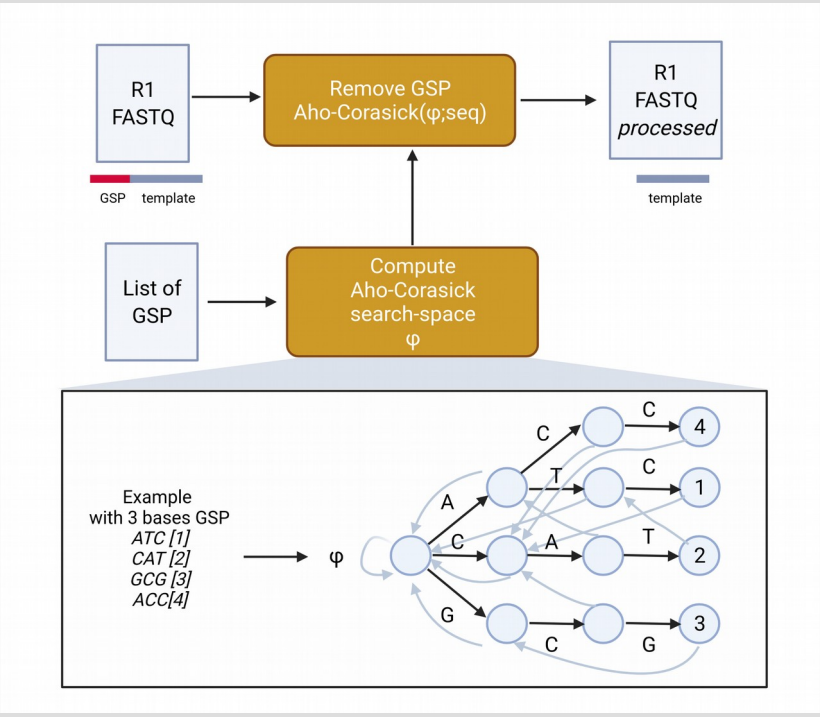


Figure 48: Trimming des librairies QIAseq : traitement des séquences du FASTQ R1.

La première étape de *trimming* consiste à éliminer les GSP des séquences brutes du FASTQ R1 (figure 48). Un fichier comprenant la liste des GSP et leur séquence est fourni par la société Qiagen lors de la commande du kit. Afin de réduire le temps de recherche pour les échantillons ayant une forte profondeur de séquençage, nous avons choisi de ne pas effectuer une recherche naïve visant à comparer chaque séquence du FASTQ à chaque GSP tant que celle-ci n'a pas été trouvée. En effet, la complexité de ce type d'approche est quadratique O(n x g) où n désigne le nombre de séquences du FASTQ et g le nombre de GSP ce qui la rend parfaitement inapplicable pour les panels larges et du séquençage profond. Ainsi, nous avons implémenté cette recherche par la méthode d'Aho-Corasick. L'algorithme d'Aho-Corasick est un algorithme de recherche de motifs dans un texte en complexité linéaire. Une structure de données abstraite, sous forme d'arbre, est créée à partir des séquences des GSP à éliminer. Cette structure de données contient le ou les motifs recherchés en lisant les bases une à une et de sorte à ne lire qu'une seule fois chacune des bases de chaque séquence du

FASTQ. Une fois la séquence du GSP éliminée des lectures du FASTQ R1, un contrôle est réalisé afin de vérifier pour les fragments courts que leur extrémité 3' ne chevauche pas la séquence commune en complément inverse. Le cas échéant, cette portion de contaminant est éliminée.

La deuxième étape vise à traiter les séquences contenues dans le FASTQ R2. L'extraction des UMI de longueur 12 est réalisée par UMI-tools [236]. UMI-tools est un utilitaire capable de prendre en charge un certain nombre de fonctionnalités autour de la gestion des UMI comme par exemple leur extraction ou la fusion des séquences porteurs du même UMI. Dans le cadre du *trimming* des librairies QIAseq, seule la fonction d'extraction visant à soustraire la séquence de l'UMI et à l'écrire dans le nom de chacune des entrées du FASTQ R2 est utilisée. Une fois cette étape réalisée, le FASTQ R2 est parcouru de sorte à extraire pour chaque entrée unique la séquence de l'UMI et à écrire cette séquence dans son entrée homologue dans le FASTQ R1. Durant cet ultime parcours des FASTQ R1 et R2, les séquences dépourvues de CS ou sans GSP sont éliminées tout comme les séquences de moins de 10 paires de base après *trimming*. Elles ont généralement pour origine des duplicats de primers.

A l'issue de ces différents traitements, les librairies QIAseq sont prêtes à être alignées contre le génome de référence.

B.2. Alignement de séquences

L'alignement de séquences consiste à positionner un ensemble de lectures séquencées sur une séquence de référence. Les technologies de séquençage à haut-débit génère une volumétrie de séquences importante qu'il faut pouvoir traiter efficacement en tenant compte de plusieurs facteurs tels que la taille des séquences, le taux d'erreur des séquenceurs et la présence de substitutions, d'insertions ou encore de délétions.

Les premiers aligneurs avaient la particularité de découper les séquences à aligner et d'utiliser une structure de données sous forme de tables de hachage afin de les aligner efficacement. Par exemple, si on considère une séquence comportant deux substitutions, il est possible de la diviser en quatre sous séquences. Dans la mesure où les deux substitutions peuvent subvenir au maximum dans deux de ces sous séquences, cela signifie qu'au moins deux des quatre sous séquences s'aligneront parfaitement avec le génome de référence. En utilisant une structure de données de type clef-valeur via les tables de hachage, il est possible de stocker cette information et de l'utiliser de façon très efficiente pour déterminer les alignements. L'avantage de cette approche est qu'elle permettait d'aligner des séquences

ayant des erreurs de type substitutions puisque que l'alignement ne reposait plus sur un alignement exact de la séquence sur toute sa longueur mais de l'alignement exacte d'une ou plusieurs sous parties.

Les algorithmes les plus efficients sur le plan de la mémoire sont ceux qui hachent les lectures séquencées tels que les programmes RMAP [243], MAQ [244], ZOOM [245] ou encore SeqMap [246]. Ils ont en revanche l'inconvénient de parcourir l'intégralité de la séquence de référence du génome lorsque peu de lectures sont à aligner. D'autres algorithmes pré-traitent la séquence du génome de référence pour parvenir à effectuer l'alignement. C'est le cas des programmes SOAP [247], PASS [248] ou encore ProbeMatch [249]. Ces algorithmes ont pour particularité d'être parallélisables mais au détriment d'une utilisation gourmande en mémoire.

Algorithme BWA

L'algorithme BWA (Burrows–Wheeler Alignment) est probablement le programme bioinformatique le plus utilisé pour le traitement des données de séquençage à haut-débit. BWA repose sur le prétraitement de la séquence du génome de référence afin d'aligner de manière efficiente un grand nombre de séquences courtes sur celui-ci.

BWA repose sur la description en 1992 [250] de la transformée de Burrows–Wheeler (Burrows–Wheeler Transform, BWT). Cette transformée de BWT a entraîné le développement par différents groupes d'algorithmiciens d'aligneurs tels que SOAPv2 [247], Bowtie [251] et BWA-MEM [237]. La transformée BWT et la construction de l'arbre des suffixes d'une séquence X=ATTCGA sont présentées en figure 49.

On appelle alphabet l'ensemble des caractères possibles du texte X à transformer, ici $\sum$=[A, T, G, C]. Chaque mot est terminé par un symbole $ qui n'appartient pas à l'alphabet $\sum$ et qui est lexicographiquement plus petit que l'ensemble des éléments de l'alphabet. Une rotation du mot X est tout d'abord réalisée afin de définir un vecteur de position P définissant les positions de ces rotations. Ce tableau de rotation est ensuite trié dans l'ordre lexicographique. Les positions S(i) des séquences triées sont stockées tout comme la dernière colonne du tableau trié B[i]. Ces deux éléments sont ce qu'on appelle le tableau des suffixes (suffix array, S) et la transformée BWT (B[i]) du texte X.

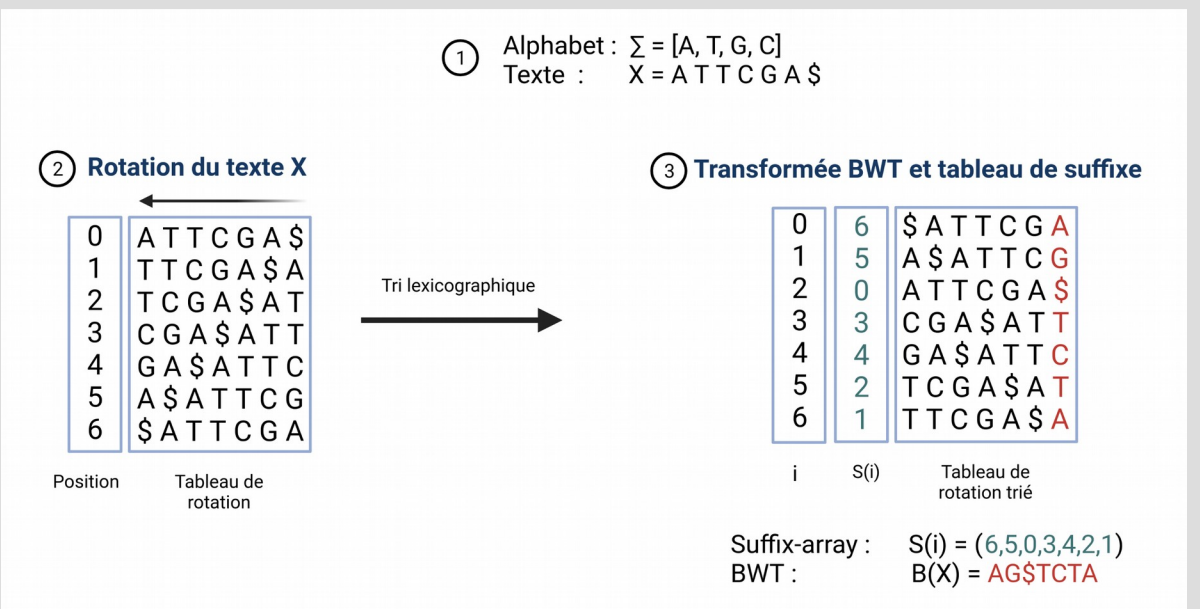


Figure 49: Transformée BWT : exemple d'application sur une séquence nucléique.

L'algorithme BWA nécessite d'indexer la séquence de référence du génome humain selon le même procédé que décrit précédemment avant de réaliser l'alignement des séquences. Cette étape d'indexation est réalisée à partir de la séquence du génome de référence au format FASTA par la commande *bwa index ref.fa*. Cette étape est particulièrement longue en temps de calcul (~3H) et nécessite plusieurs dizaines de Go de RAM. Ces données nécessaires à l'aligneur sont stockées dans différents fichiers de sorte à être réutilisables.

L'indexation de BWA produit en sortie un fichier .bwt contenant la transformée BWT du génome de référence, un fichier .sa contenant le tableau des suffixes ainsi que deux autres fichiers .ann et .amb qui ont pour objectif de stocker l'information sur les bases ambiguës du génome de référence. Ces bases ambiguës sont des positions comportant soit des trous, ou *gaps*, dans la séquence assemblée, soit des bases non définies (N). On retrouve ainsi dans ces fichiers le nombre de bases ambiguës par entrée du fichier FASTA, le nom du chromosome, ainsi que la longueur de la séquence ambiguë. Enfin, un fichier .pac est créé pour compresser et intégrer les données des deux fichiers .pac et .anno.

BWA intègre différents algorithmes d'alignement : BWA-backtrack, BWA-SW et BWA-MEM. Brièvement, l'algorithme BWA-backtrack est réservé à l'alignement de séquences longues. BWA-MEM et BWA-SW présentent des fonctionnalités similaires comme la capacité à aligner des séquences longues et de créer des insertions/délétions lors de l'alignement. Néanmoins, en pratique, BWA-MEM offre des performances d'alignement supérieures pour des séquences de très haute qualité, comme les données des séquenceurs de

nouvelle génération. Les données présentées dans le cadre de cette thèse ont été alignées par BWA-MEM.

Génome de référence

Le génome de référence humain possède différentes versions de sa séquence. En pratique, dans les laboratoires de diagnostic, deux versions coexistent : la version hg19 (ou GRCh37) datant de février 2019 et la version hg38 (ou GRCh38) datant de décembre 2013.

L'assemblage du génome de référence est imparfait : la dernière version hg38 comporte encore environ 250 trous dans des régions particulièrement difficiles à assembler. L'arrivée des technologies de séquençage de troisième génération telles que Nanopore, qui permettent le séquençage long de fragments d'ADN de plusieurs Mb, ont permis de résoudre une centaine de trous dans cette version du génome par rapport à la version antérieure. La première version du génome de référence humain publiée comportait plus de 150 000 régions mal assemblées.

Le génome de référence a de nombreuses limitations dans la mesure où il ne représente qu'une séquence unique pour représenter une diversité génétique importante dans l'espèce humaine. Dans les faits, cette référence n'est constituée que d'un nombre très limité d'individus : 93 % de sa séquence est déterminée à partir de l'ADN extrait du sang de 11 individus seulement. Afin d'avoir une description plus précise des variations génétiques naturelles de ce génome de référence, des banques de populations décrivant la fréquence des polymorphisme comme 1000 Genomes, ExAc ou GnomAD ont vu le jour. Ces banques de données indiquent, à partir du séquençage de plusieurs milliers d'individus, la fréquence des substitutions observées chez des témoins sains par comparaison du génome de référence. Ces fréquences alléliques sont particulièrement utilisées afin d'identifier les polymorphismes des échantillons. En pratique, sont considérés comme polymorphismes l'ensemble des variations génétiques détectées dans un échantillon et qui ont une fréquence dans la population supérieure à 1 %. L'apparition des mutations somatiques dans une population de cellules d'un invidu sont des phénomènes aléatoires : il n'est donc pas rare de trouver des mutations somatiques, c'est à dire restreintes à une partie des cellules d'un échantillon biologique, conduisant à l'apparition de variants décrits comme étant des polymorphismes dans les banques de population.

Fichiers d'alignement

BWA génère à partir des fichiers FASTQ, de la séquence du génome de référence et de ses index, un fichier au format SAM (sequence alignment map, SAM) permettant de stocker les positions d'alignement des séquences. Ce format de stockage des données d'alignement est très répandu et très largement adopté dans la communauté bioinformatique.

Le format SAM comporte deux sections : une section d'entête et une section d'alignement. Afin d'économiser de l'espace disque, une version compressée de ces fichiers SAM appelée BAM (binary alignment map, BAM) existe. Les fichiers BAM contiennent exactement la même information que les fichiers SAM mais au format binaire compressé.

La section d'entête précède dans les fichiers SAM la description des alignements. Les lignes d'entête commencent par le symbole @ afin de les distinguer des données alignées à proprement parler. On retrouve en général dans cette section des traces des différents algorithmes de traitement des fichiers BAM, comme la suite GATK détaillée dans la suite de cette section. La section d'alignement rassemble pour chaque séquence alignée 11 colonnes obligatoires :

- une colonne QNAME avec le nom de la séquence
- une colonne FLAG comportant un descriptif sous forme d'un *bit* décrivant la manière dont l'alignement s'est déroulée pour la lecture en question
- une colonne RNAME faisant référence au nom de la séquence de référence sur laquelle la lecture s'est alignée. Il s'agit en en pratique très souvent du nom du chromosome
- une colonne POS définissant la position d'alignement sur la séquence RNAME
- une colonne MAPQ (MAPping Quality) avec le score d'alignement de la séquence
- une colonne CIGAR décrivant la manière dont la lecture s'est alignée
- une colonne RNEXT stipulant le nom de la lecture appariée en cas de séquençage paired-end
- une colonne PNEXT mentionnant la position de la lecture appariée
- une colonne TLEN avec la longueur de la séquence de référence sur laquelle la lecture s'est alignée
- une colonne SEQ avec la séquence de la lecture déduite du fichier FASTQ

- une colonne QUAL avec le score de qualité par base déduite du fichier FASTQ

Certaines colonnes sont particulièrement intéressantes.

La colonne FLAG détermine la manière dont on peut considérer la séquence alignée après le traitement de l'algorithme d'alignement. Il permet par exemple filtrer les séquences non alignées (FLAG=4), les séquences ne passant pas les scores de qualité selon les recommandations constructeurs (FLAG=512) ou encore les duplicats optiques ou de PCR (FLAG=1024) après traitement des fichiers SAM par la suite GATK.

La colonne MAPQ est elle aussi essentielle pour éliminer les lectures ayant des scores d'alignement faibles, c'est à dire dont la probabilité d'alignement de la séquence n'est pas suffisante pour être confiant sur son positionnement le long de la séquence de référence.

Le code CIGAR est essentiel en cas de doute sur l'alignement d'une lecture. Il permet de déterminer l'effort qu'a du réaliser l'algorithme afin de positionner la lecture à une position de la référence. Il quantifie sur la longueur de la séquence les bases identiques (=), les insertions (I), les délétions (D), les bases soft-clippées (S) et les substitutions (X). Par exemple, le code CIGAR 14M2D31M signifie qu'une séquence de 47 nucléotides s'est alignée sur le génome de référence avec 14 bases identiques, suivies de 2 délétions puis de 31 bases identiques.

Enfin, les fichiers BAM sont rarement utilisés seuls car ils sont globalement assez lourds à traiter sur le plan informatique. Ainsi, ces fichiers BAM sont souvent indexés par la fonction *index* de l'utilitaire *samtools* [252].

Genome Analysis Toolkit

Les fichiers BAM sont ensuite traités par la suite Genome Analysis Toolkit (GATK) dont la dernière version est GATK4 [238], [253], [254]. Développée par la plateforme de Data Sciences du Broad Institute, GATK est une suite d'utilitaire en ligne de commande dédiée à l'analyse des données de séquençage à haut-débit et orientée pour la découverte de tout type de variants. Ainsi, les bioinformaticiens ont la possibilité d'utiliser tout ou partie des outils éprouvés par la communauté bioinformatique.

Nous ne détaillerons pas dans cette section l'intégralité des fonctionnalités de GATK mais uniquement les outils qui ont été utilisés pour le traitement des données dans le contexte de cette thèse : Mark Duplicates et BaseRecalibrator.

Mark Duplicates est un premier utilitaire ayant pour objectif, pour chaque échantillon, d'identifier à partir des fichiers BAM alignés les paires de séquences qui ont vraisemblablement pour origine un même fragment d'ADN. Cet algorithme est particulièrement utile pour les données de séquençage n'intégrant pas d'index moléculaires (UMI) afin d'identifier les duplicats de PCR. Il est important de marquer les duplicats car ces lectures non indépendantes peuvent fausser l'interprétation des fréquences alléliques des variants en venant ajouter artificiellement des allèles mutés ou non mutés. Dans le cadre des librairies intégrant des UMI, l'intérêt principal de cette approche est de pouvoir identifier les duplicats optiques.

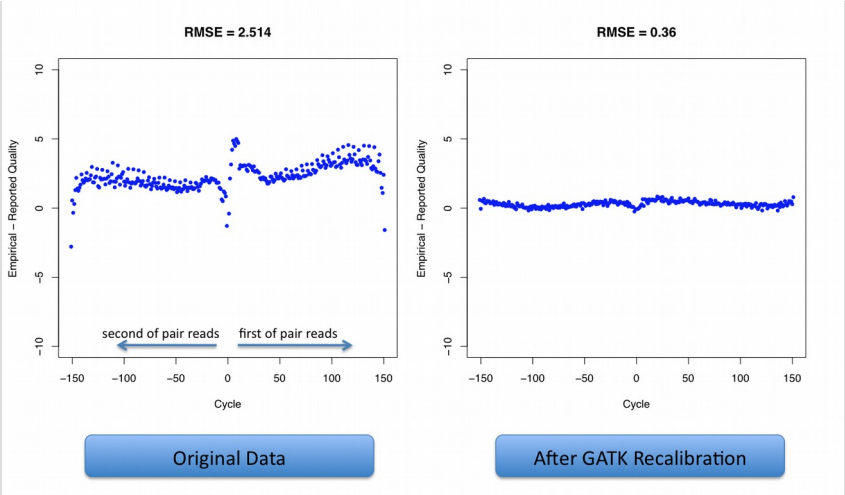


Figure 50: GATK - BQSR : exemple de biais dans l'évaluation des scores de qualité par base en fonction des cycles de séquençage.

Extrait de la documentation en ligne de l'outil BQSR de GATK. Les scores de qualité avant et après requalibration sont reportés pour chaque cycle.

BaseRecalibrator est un utilitaire ayant pour objectif d'utiliser des approches de machine learning pour détecter et corriger des motifs d'erreurs systématiques d'attribution des scores de qualité par base des séquenceurs. Ces biais peuvent provenir de la chimie de préparation de la librairie, du support de séquençage ou bien de l'appareil lui même. Par exemple, il est possible de mettre en évidence qu'une lecture d'une base précédée de AA dans un échantillon est toujours à l'origine d'une déviation du score de qualité de cette base de 7 %. En d'autres termes, si aucune correction n'est apportée, cela signifie que les algorithmes de détection de variants auront une probabilité moindre d'évaluer positivement un variant si il est précédé du

motif AA. Un exemple de mesure résiduelle l'attribution des scores de qualité en fonction du cycle de séquençage est donné en figure 50.

A la fin de ces deux étapes, les fichiers BAM alignés et traités sont prêt à être utilisés pour la suite du traitement des données.

C. Analyse tertiaire

C.1. Détection des variations génétiques et des CNV

La détection des variations ponctuelles dans un échantillon à partir des données de séquençage alignées repose sur l'utilisation d'algorithmes dits de *variant calling*. Ces algorithmes ont pour objectif de compter le nombre de bases mutées à chaque position de l'alignement, à partir des fichiers BAM, et de proposer une liste de variations candidates dont le signal est significativement différent du bruit de fond. Il existe de très nombreux algorithmes de détection des variants reposant sur des modèles mathématiques et des méthodologies parfois très différents.

La détection des variations du nombre de copies de segments génomiques (copy number variation, CNV) repose sur la quantification du nombre de lectures alignées par région ciblée et sur la comparaison de ces valeurs de comptage entre des échantillons dits de référence et un échantillon à tester. De très nombreux algorithmes de détection de CNV existent dans la littérature. Certains sont adaptés aux analyses de grands jeux de données comme des séquençages WES ou WGS, tandis que d'autres ont été spécialement développés pour l'analyse de données de séquençage ciblé de panels de gènes.

Dans la mesure où des développements spécifiques autour de ces deux thématiques ont été conduits dans le cadre de cette thèse, avec le développement d'outils de *variant calling* et d'un algorithme de détection de CNV, de plus amples informations sur ces étapes de l'analyse tertiaire seront données dans le chapitre III.

C.2. Annotation des données

L'annotation des données est une étape essentielle pour aider à l'interprétation des anomalies détectées. Elle consiste à interroger des banques de données afin d'en extraire des ressources et ainsi d'apporter des informations supplémentaires sur les variants ou les CNV. Nous décrirons dans cette section quelques concepts bioinformatiques autour de l'annotation ainsi que quelques exemples de banques de données pour l'annotation des variations génétiques.

Procédure d'annotation

L'étape d'annotation est un processus bioinformatique constitué de plusieurs phases : une étape de requêtage, une étape d'interprétation par la banque puis l'obtention d'une réponse. En fonction de la nature des banques de données, différentes API (Application Programming

Interface) existent et en conséquence complexifient ce processus puisque la procédure de requêtage et les réponse des banques changent tant sur le contenu que sur le format. L'hétérogénéité des formats de fichiers contenant les données d'annotation est importante. On retrouve couramment des fichiers au format CSV, XML ou encore JSON.

Ensembl, développé conjointement par l'*European Bioinformatics Institute* et le *Wellcome Trust Sanger Institute* est interrogeable via une API REST HTTP (Representational State Transfer Application Program Interface). Des scripts développés en PERL peuvent aussi être utilisés afin d'interroger directement les serveurs SQL (Structured Query Language) de *Ensembl* mais ils sont en pratique très peu utilisés pour l'annotation des données de séquençage à haut-débit de par leur lenteur d'exécution.

Du côté américain, le NCBI (National Center for Biotechnology Information) a développé des programmes permettant d'interroger ses banques de données via des requêtes « Entrez ». Cet ensemble de programmes, appelé E-utilities, permet d'effectuer des requêtes HTTP sur les 38 banques de données couvrant une très large variété de données sur les séquences nucléiques et protéiques, les informations sur les gènes et les variants, les structures tri-dimensionnelles des molécules ou encore la littérature biomédicale via PubMed.

Néanmoins, ces deux méthodes d'acquisition de données d'annotation sont en pratique très peu utilisées dans la mesure où les temps d'exécution sont très longs. De nouveaux outils, tels que ANNOVAR [255] ou VEP [256], permettent d'importer localement les informations contenues dans les banques de données sous forme de fichiers plats et d'effectuer l'annotation bien plus efficacement. Le nombre de sources de données est très important ce qui laisse une grande souplesse sur le choix des banques. L'inconvénient majeur de ces approches est la nécessité de mettre régulièrement à jour ces fichiers sources de sorte à avoir l'annotation la plus récente possible à chaque nouvelle version des banques de données. Certaines évoluent lentement avec peu de mises à jour tandis que d'autres sont actualisées tous les mois.

Principales banques de données pour l'annotation des variants

Comme nous l'avons évoqué précédemment, il existe un grand nombre de banques de données. Certaines sont très généralistes en agrégeant plusieurs sources d'information tandis que d'autres sont beaucoup plus spécialisées.

Les banques de population comme 1000 Genomes [257], ExAC [258] ou DGV [259] permettent d'obtenir la liste des polymorphismes (SNV et CNV) connus à partir de cohortes de milliers de témoins. Environ 15 millions de variants sont décrits dans ces banques avec une

fréquence observée dans la population générale supérieure à 1 %. Cette information est particulièrement utilisée pour la filtration des anomalies des patients dans le cadre de la recherche d'anomalies génétiques acquises dans des pathologies cancéreuses. Elle permet notamment de se passer du séquençage apparié d'un échantillon de référence sain pour chaque patient pour éliminer les SNP et les CNV récurrents.

D'autres banques de données telles que dbSNP [260], COSMIC [261] et ClinVar [262] donnent des informations sur les variants telles que la pathologie dans laquelle le variant a déjà été décrit, l'impact de la présence d'une mutation sur les transcrits et sur la protéine, les scores de conservation au cours de l'évolution tels que GERP++ [263] ou PhyloP [264] ou encore la prédiction de l'impact du variant par des algorithmes de prédiction de pathogénicité par différentes approches statistiques (SIFT [265], PolyPhen [266], CADD [267] ou DANN [268]…).

Le choix des banques de données à utiliser pour répondre à une question biologique donnée n'est pas chose aisée et nécessite une connaissance large des différentes banques de données existantes afin de développer des protocoles d'annotation et de filtration cohérents.

III. NOUVEAUX ALGORITHMES DE TRAITEMENT DES DONNÉES DE SÉQUENÇAGE POUR LE CFDNA

Nous avons vu dans les chapitres précédents les avancées récentes dans la classification des lymphomes, dans la nature des échantillons analysés avec le développement des biopsies liquides et enfin les analyses bioinformatiques primaire, secondaire et tertiaire depuis les données brutes des séquenceurs.

Ce chapitre vise à détailler l'ensemble des développements bioinformatiques qui ont été conduits dans le cadre de cette thèse. Il couvrira notamment le développement de quatre nouveaux algorithmes :

- LowVarFreq : un algorithme de détection et de filtration des mutations sur des jeux de données sans barcode moléculaire
- UMI-VarCal : un algorithme de détection des variations génétiques ponctuelles à partir de librairies de séquençage avec UMI
- mCNA : un algorithme de détection des remaniements de nombre de copies de gènes à partir de librairies de séquençage avec UMI
- UMI-Gen : un algorithme de simulation de fichiers d'alignement intégrant des UMI

Chaque programme sera présenté en détails sur le plan algorithmique. Des exemples d'application sur des échantillons tumoraux et de cfDNA seront donnés pour illustrer le fonctionnement de chaque algorithme.

A. Détection des mutations sans UMI : LowVarFreq

A.1. État de l'art

Nous nous sommes intéressés dans un premier temps à la filtration des données de séquençage à haut-débit à partir de librairies n'intégrant pas de barcodes moléculaires (UMI) séquencées sur la technologie Ion Torrent.

Dans ce contexte, la recherche de sensibilité pour permettre la détection de variants de faible fréquence entraîne nécessairement le développement de nouvelles approches afin de filtrer et d'aider à l'interprétation des résultats d'algorithmes de détection de variants (*variant calling*) disponibles dans la littérature. En effet, détecter des variants de faible fréquence revient à diminuer la stringence des différents paramètres des algorithmes de sorte à maximiser les chances de détection d'un événement au détriment d'un nombre très important d'artefacts détectés. Ces artefacts à faible fréquence sont bien souvent liés à la région ciblée qui peut être difficile à séquencer ou à aligner (présence d'homopolymères, régions répétées…), à la chimie utilisée pour préparer la librairie de séquençage ou encore être liés à la dégradation de l'échantillon séquencé.

De plus, nous savons que le recouvrement de la détection des mutations entre différents algorithmes appliqués sur un même jeu de données demeure très imparfait, même sur des données de faible complexité [269]–[271]. Par exemple, en 2015, une étude publiée dans Nature rapporte une concordance de 91 % entre les algorithmes FreeBayes [272], Samtools [252] et GATK-HC sur un même échantillon à partir de plateformes Illumina. D'autres études rapportent des taux de concordance bien inférieurs à 50 % par O'Rawe et al. en 2013 [269] et à 57 % par Cornish et al. [273]. Cette concordance chute drastiquement à 15,5 % sur les séquenceurs Ion Torrent entre Samtools, FreeBayes, GATK-HaplotypeCaller [274] et le Torrent Variant Caller (Figure 51). Le Torrent Variant Caller n'est applicable que sur des jeux de données Ion Torrent puisqu'il a été spécifiquement développé pour cette gamme de séquenceurs. En intégrant des filtres spécifiques, le TVC élimine un grand nombre d'artefacts de séquençage de faible fréquence liés aux signaux d'acquisition de la chimie Ion Torrent, ce qui explique probablement la maigre concordance avec d'autres algorithmes qui eux ne sont pas dédiés cette technologie. Enfin, des algorithmes bioinformatiques ont été développés spécifiquement pour la détection de variants de faible fréquence et peuvent compléter les données fournies par les outils utilisés plus classiquement. C'est le cas par exemple des algorithmes LoFreq [275] ou OutLyzer [276].

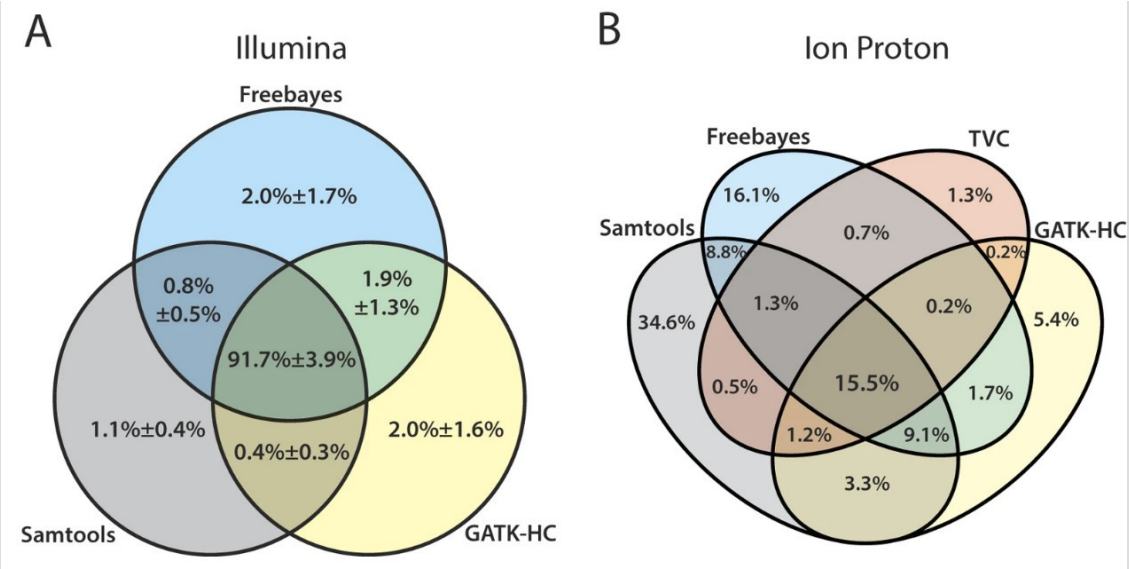


Figure 51: Comparaison des résultats de différents algorithmes de variant calling [269].

La comparaison a été effectuée à partir des données de séquençage Illumina ou Ion Torrent de l'individu NA12878 caractérisé et publié par le consortium Genome in a Bottle (GIAB).

Cette absence de consensus dans le choix des outils, le très grand nombre de paramètres, la recherche de sensibilité pour l'analyse d'échantillons plasmatiques et enfin la difficulté de la filtration des données provenant de plusieurs *variant callers* nous ont conduit à développer une nouvelle approche appelée LowVarFreq.

A.2. Implémentation

LowVarFreq repose sur différents processus bioinformatiques visant à estimer le bruit de fond de séquençage à partir d'échantillons contrôles, à créer un fichier d'alignement *in silico* mimant ce bruit de fond de séquençage, à exécuter plusieurs algorithmes de détection de variants d'intérêt sur ce fichier pour finalement en déduire une liste de faux positifs à soustraire des échantillons de cfDNA à tester.

Extraction du bruit de fond de séquençage

A partir d'échantillons contrôles, c'est à dire de fichiers BAM alignés d'échantillons dépourvus d'anomalies somatiques, LowVarFreq commence par estimer pour chacun d'entre eux le nombre de bases A, T, G, C et d'insertions/délétions présentes à chacune des positions de l'alignement, tout au long des régions ciblées. Ces comptages sont réalisés à partir de l'utilitaire *pileup* de samtools [252].

On cherche ici à extraire le bruit de fond de séquençage dans les échantillons témoins. Pour se faire, pour chaque position p de l'alignement, où p désigne les coordonnées

chromosomiques de la position, les fréquences moyennes du nombre de A, T, G, C et d'insertions/délétions sont déduites des valeurs observées dans le *pileup* afin de construire une matrice N de bruit de fond. On y retrouve le bruit de fond, c'est à dire le nombre de bases générées par position correspondant aux erreurs d'amplification et de lecture du séquenceur, ainsi que les polymorphismes des échantillons contrôles. Ces polymorphismes correspondent nécessairement à deux états dans les échantillons contrôles : hétérozygote dont la fréquence allélique fluctue autour de 50 % de bases mutées ou homozygotes avec une fréquence allélique proche de 100 %. Afin de n'obtenir une matrice N ne contenant que le bruit de fond de séquençage, une étape de filtration est réalisée afin d'éliminer ces polymorphismes.

Elle vise à identifier si au sein de la matrice N certaines positions ont des fréquences alléliques moyennes anormalement élevées (>30%), anormalement discordantes (la fréquence allélique de la position fluctue de plus de 50 % dans les échantillons contrôles) ou référencées dans la banque de population dbSNP à une fréquence dans la population normale à plus de 1 %. Si des positions correspondent à ces critères de filtration, alors le nombre d'événements à la position de l'alignement qui coïncide avec le SNP est gommé de la matrice de comptage. On obtient ainsi une matrice N filtrée dont la fluctuation des fréquences de chaque événement ne peut être expliquée par la présence d'un SNP dans les échantillons contrôles. L'objectif est bien de pouvoir soustraire le bruit de fond de séquençage pour la suite de l'analyse et non les polymorphismes des échantillons testés.

Construction des fichiers SAM témoins

A partir des fréquences de substitutions et d'insertions/délétions stockées dans la matrice N filtrée, deux fichiers BAM sont générés :

- un premier fichier BAM témoin *T* « normal » ne comprenant que des lectures *in silico* parfaitement identiques à la séquence du génome de référence pour chaque région ciblée
- un deuxième fichier BAM « bruité » *B*, généré à partir de *T*, dans lequel chacune des séquences est éditée à chaque position de l'alignement de sorte à introduire des substitutions, des insertions ou des délétions à des fréquences alléliques comparables à celles estimées dans la matrice N

Ces deux étapes sont les plus exigeantes sur le plan des ressources informatiques en temps et en mémoire. En fonction de la largeur des panels et de la profondeur de séquençage

souhaitée dans les fichiers témoins, l'algorithme doit procéder parfois à un grand nombre d'éditions de séquences pour créer ce fichier BAM *B* mimant le bruit de fond de séquençage.

Création des listes d'artefacts

A cette étape du traitement informatique, nous disposons de plusieurs éléments :

- une série de fichiers BAM témoins
- une matrice *N* filtrée gardant la trace des fréquences moyennes observées des bases A, T, G, C et des insertions/délétions à partir des fichiers BAM témoins
- un fichier BAM normal *T* mimant les résultats d'un séquençage sans bruit de fond dans lequel on retrouve pour chaque amplicon un ensemble de séquences parfaitement identiques à la séquence du génome de référence
- un fichier BAM bruité *B* correspondant à un ensemble de séquences reprenant le bruit de fond de séquençage déduit de *N* pour chaque région ciblée

A partir de ces différentes sources de données, LowVarFreq exécute VarScan2, MuTect, LoFreq et OutLyzer afin d'effectuer différentes comparaisons :

- entre chaque fichier BAM témoin et le fichier BAM *B* de sorte à en déduire une liste de variants de faible fréquence qui pourraient être spécifiques d'un des fichiers BAM témoin (faux positifs échantillon-spécifiques)
- entre le fichier BAM bruité *B* et le fichier normal *T*, de sorte à en déduire une liste d'artefacts provenant du bruit de fond de séquençage et qui ne sont pas filtrés par les algorithmes de détection de variants (faux négatifs)

En fonction des algorithmes, les fichiers de variants générés au format VCF (Variant Call Format) sont réconciliés de sorte à générer un dictionnaire de variants dans un format unique. Celui-ci stocke pour chaque variant détecté l'algorithme à l'origine de sa détection, le contexte de détection et son occurrence dans les fichiers BAM témoins. A noter que l'algorithme LoFreq fournit plusieurs fichiers VCF correspondant à des niveaux de stringeance différents (strictes ou relâchés) et que cette information est elle aussi stockée pour servir d'annotation.

La liste d'artefacts potentiels sera à la base du processus d'annotation des variants détectés dans les échantillons de cfDNA analysés sur un panel de séquençage donné.

Analyse d'un échantillon à tester

A partir de la liste d'artefacts probables déduite des étapes de traitement précédentes et du fichier BAM de l'échantillon à tester, les mêmes étapes de *variant calling* sont réalisées. Les résultats de VarScan2, MuTect, LoFreq et OutLyzer sont comparés informatiquement à la liste d'artefacts afin de les annoter. On obtient pour chaque variant candidat :

- la liste des algorithmes ayant considérés le variant comme positif dans l'échantillon à tester
- l'information sur la présence ou non du variant dans la liste des artefacts, ainsi que sa récurrence dans les échantillons contrôles

Dans la mesure où les erreurs de séquençage peuvent être liées à la présence de bases répétées appelées homopolymères, que ce soit pour la technologie de séquençage Illumina ou IonTorrent, une étape d'annotation supplémentaire est réalisée afin de calculer la longueur de l'homopolymère lié au variant candidat. Par exemple, si la mutation est une transition C vers A, LowVarFreq estimera en 5' et en 3' autour de la position génomique candidate le nombre de C présents sur la séquence du génome de référence. L'intérêt de cette approche est de soustraire rapidement les erreurs de lectures récurrentes liées à la composition nucléotidique des régions ciblées.

A.3. Application de LowVarFreq sur le lymphome de Hodgkin

Application bioinformatique

LowVarFreq a été développé afin de rendre possible d'interprétation des résultats de séquençage d'une cohorte de lymphomes de Hodgkin pour lesquels nous disposions à la fois d'ADN extrait de la tumeur des patients mais aussi de plasmas afin d'analyser le cfDNA.

Un panel de 92 amplicons (AmpliSeq) a été créé afin de couvrir 6 gènes fréquemment mutés dans le LH (*XPO1*, *TNFAIP3*, *NFKBIE*, *STATS6*, *B2M* et *PTPN1*). Le séquençage a été réalisé à forte profondeur (>3000X) sur un séquenceur de type PGM. Les analyses bioinformatiques primaires et secondaires ont été réalisés selon les recommandations constructeurs par les logiciels de la Torrent Suite comme précédemment évoquées (chapitre II). Les résultats du Torrent Variant Caller ont été collectés avec des paramétrages relâchés de sorte à maximiser la sensibilité de détection au détriment de la spécificité. Ce processus d'analyse a été appliqué à 24 couples tumeur/plasma dont 12 tumeurs congelées et 12 tumeurs inclus en paraffine (FFPE).

En pratique, le Torrent Variant Caller peine à détecter des variants de façon systématique à des fréquences alléliques inférieures à 3 %. Or, comme nous l'avons vu précédemment, les lymphomes de Hodgkin sont caractérisés par une présence peu abondante de cellules tumorales et l'analyse d'échantillons plasmatiques requiert aussi de descendre à des VAF inférieures à 3 %. Nous avons donc appliqué LowVarFreq afin d'aller au-delà des limitations du Torrent Variant Caller.

Le bruit de fond de séquençage sur ce panel a été estimé à partir des résultats de séquençage de 8 échantillons de sang normaux. La reconstruction du fichier BAM in silico *T* contenant le bruit de fond de séquençage a été réalisée puis ce fichier BAM a été analysé par les différents outils de variant calling précédemment décrits. Les résultats, sur cet échantillon *in silico*, montrent de grosse disparité concernant le nombre de variants détectés en fonction des algorithmes avec 340 variants détectés par le variant caller le moins spécifique (Outlyzer) et seulement 5 variants pour l'algorithme LoFreq en paramétrage stringent.

Le même processus de *variant calling*, appliqué aux échantillons à tester, conduisent à la détection de 182 variants en moyenne par échantillon (min: 29 ; max: 1065). L'annotation des variants à partir des résultats sur les témoins conduit à éliminer les artefacts récurrents mais on observe tout de même dans les échantillons un nombre important de mutations à des fréquences alléliques faibles (<0,5%), variants qui ne sont donc pas expliqués par le bruit de fond estimé à partir des témoins mais par des artefacts échantillon-spécifiques.

En observant la corrélation entre le nombre de variants détectés, la nature de l'échantillon et le taux observé de transition/transversion par échantillon, on constate une nette augmentation du nombre de variants à mesure que le taux de transition/transversion augmente (figure 52). En regardant plus en détails les données, on observe que le taux de transition/transversion est plus important dans les échantillons FFPE que dans les échantillons congelés et dans les échantillons plasmatiques.

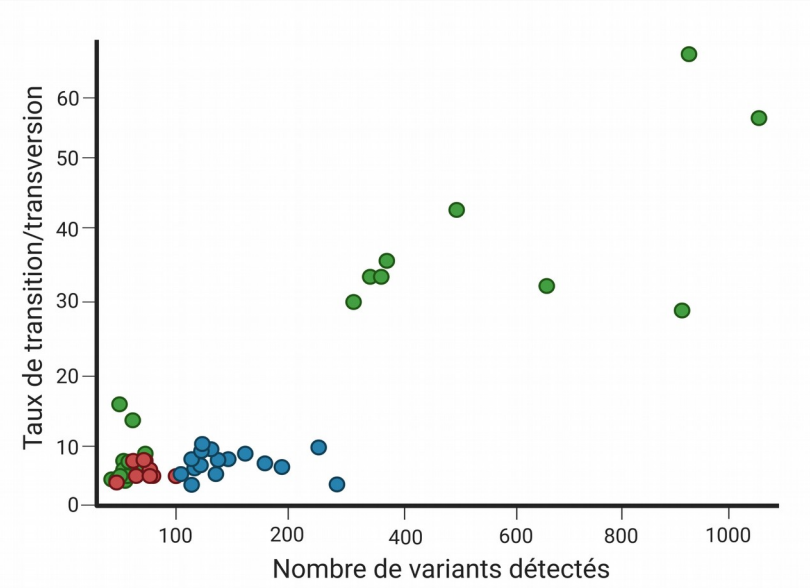


Figure 52: Nombre de variants détectés en fonction du taux de transition/transversion par échantillon.

Les échantillons congelés sont représentés en rouge, les échantillons inclus en paraffine en vert et les échantillons plasmatiques en bleu.

Les fréquences alléliques des variants sont inégalement distribuées entre les transitions et les transversions avec une fréquence allélique moyenne observée significativement plus faible lorsque le variant est une transition (p=2x10-87). Cette observation nous montre que nous sommes bien en présence d'un bruit de fond biologique, probablement dû à une oxydation importante de l'ADN dans les échantillons FFPE conduisant à l'apparition de transitions lors des cycles de PCR. Ce phénomène d'apparition de transitions GC→AT liée à l'utilisation de formaldéhyde est déjà connu mais vient compliquer encore la détection du signal biologique dans les tissus fixés et inclus en paraffine.

Afin de permettre une évaluation objective du bruit de fond d'oxydation, nous avons implémenté un algorithme itératif afin de déterminer un seuil de sensibilité par échantillon dans cette cohorte. Pour se faire, pour chaque échantillon, les variants ont été triés par fréquence allélique croissante et un test d'enrichissement en transition a été réalisé de façon itérative à des seuils de VAF croissantes de sorte à déterminer le seuil à partir duquel on n'observe plus d'enrichissement anormal en transition (figure 53). Cette procédure est réalisée par l'application de tests de Fisher successifs dont la p-valeur est corrigée par FDR (False Discovery Rate).

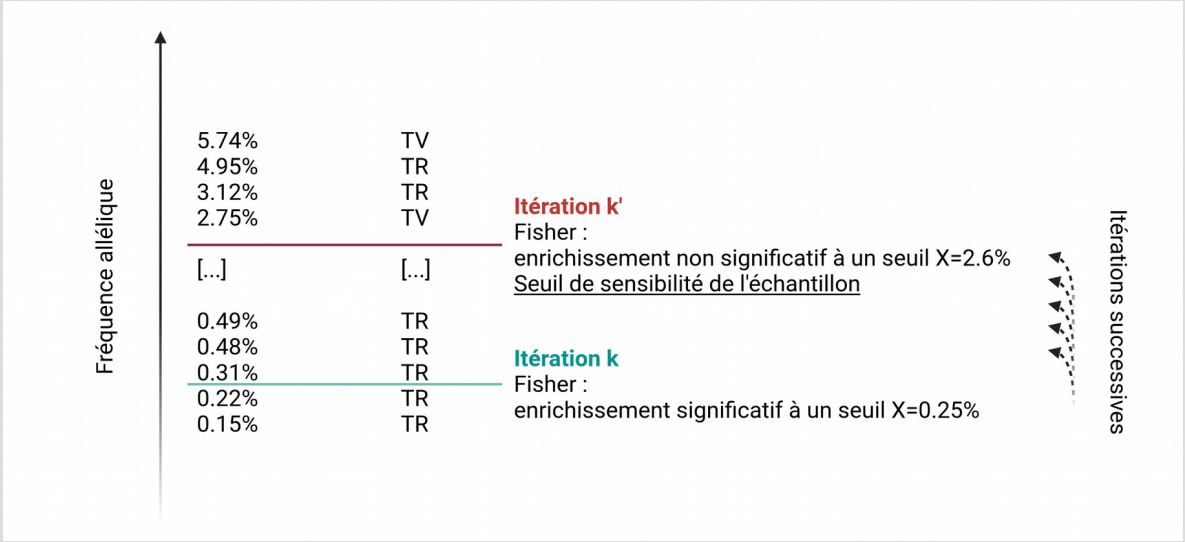


Figure 53: Procédure d'estimation du bruit de fond d'oxydation.

Les variants d'un échantillon dont on souhaite estimer le bruit de fond sont triés par fréquence allélique croissante et selon leur classe Transition (TR) ou Transversion (TV). Un test de Fisher est réalisé de façon itérative à des seuils de fréquences alléliques croissantes afin de déterminer pour l'échantillon son seuil de sensibilité, c'est à dire la fréquence allélique à partir de laquelle on n'observe plus un enrichissement de la classe TR.

Cette approche permet de déterminer, dans nos échantillons, un seuil de sensibilité moyen de 0,96 % (min: 0,01 %, max : 3,2%) permettant d'éliminer en moyenne 73,35 % des variants dans les échantillons (min: 2,96 %, max : 99,44%). On observe là encore des seuils de bruit de fond élevés dans les échantillons FFPE et à l'inverse des seuils relativement bas dans les échantillons plasmatiques. Le nombre de variants filtrés est plus important dans les échantillons ayant un nombre de variants détectés pré-filtres important et dans les échantillons avec un taux transition/transversion fort.

Afin de valider l'approche bioinformatique, c'est à dire de s'assurer que les variants filtrés correspondent bien à des artefacts d'oxydation, nous avons réalisé un nouveau séquençage des échantillons FFPE mais en appliquant cette fois-ci en amont du séquençage le kit de réparation de l'ADN FFPE DNA Repair Mix de la société NEB. Ce kit vise à traiter les ADN extraits de sorte à diminuer le taux d'oxydation de l'ADN. Nous avons ainsi appliqué LowVarFreq dans les mêmes conditions sur ces échantillons FFPE et comparés les fréquences alléliques de chacune des transitions filtrées par l'algorithme avant et après réparation (figure 54).

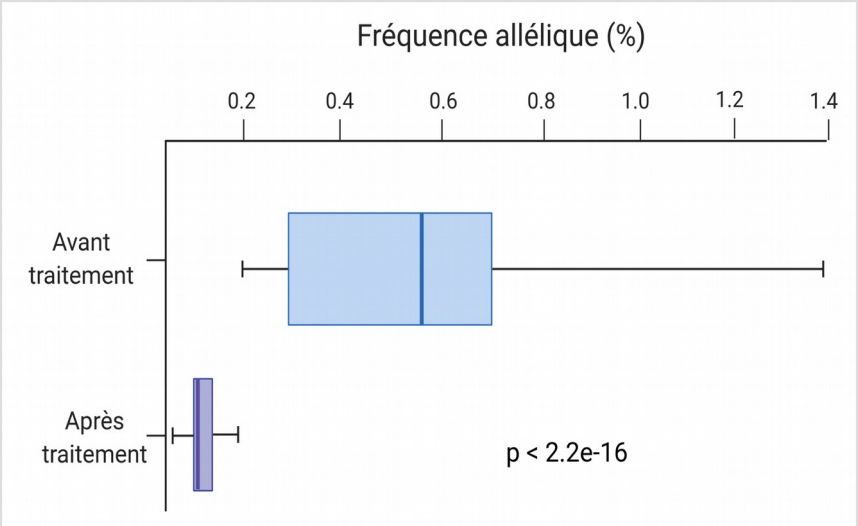


Figure 54: Impact du traitement contre l'oxydation de l'ADN sur les artefacts détectés par LowVarFreq.

Les résultats montrent une diminution très importante de la fréquence allélique des transitions entraînant mécaniquement une meilleure sensibilité. Celle-ci permet l'interprétation de variants à de plus faibles fréquences et augmente les chances de détecter un vrai positif parmi les variants de faibles VAF. Elle confirme aussi par ailleurs que l'algorithme capture bien les artefacts d'oxydation.

Résultats biologiques

La méthodologie bioinformatique et les résultats de cette cohorte de LH ont été publiés dans Leukemia & Lymphoma en 2019 [230]. On retrouve, sur le panel de 6 gènes, une informativité de 54,2 % sur les biopsies et de 47,8 % dans les échantillons plasmatiques. Dans 30,4 % des cas, les profils obtenus dans le cfDNA étaient comparables avec les profils obtenus à partir des biopsies non micro-disséquées.

De façon intéressante, nous avons observé une fréquence allélique moyenne des mutations significativement supérieure dans les échantillons plasmatiques par comparaison avec la biopsie non-micro-disséquée des patients (3,3 % vs 1,8 %, p=.033). Par ailleurs, si on s'intéresse aux couples plasma/tumeur appariés, l'analyse des deux sources d'échantillons permet d'obtenir un profil mutationnel dans 70,8 % des cas, ce qui souligne l'intérêt du cfDNA dans cette pathologie pauvre en cellules tumorales. Une validation de la mutation N417Y du gène *STAT6* par dPCR a permis de mettre en évidence une corrélation parfaite entre les deux technologies avec 5/5 mutations retrouvées via les deux technologies.

L'intégralité des résultats est présentée dans l'article dans la section III.A.5.

A.4. Limitations et perspectives

Si LowVarFreq permet d'objectiver les résultats de séquençage afin de détecter des mutations dans des échantillons pauvres en cellules tumorales ou dans des échantillons de cfDNA, il présente un certain nombre de limitations.

Tout d'abord, le temps d'exécution de l'algorithme est particulièrement long à l'étape de création du fichier BAM bruité à partir des fichiers BAM témoins. L'estimation du bruit de fond de séquençage, qui consiste à moyenner les bases observées à chaque position de l'alignement tout au long du panel ciblé, consiste à traiter autant de fichiers *pileup* qu'il y a de témoins. Dans la mesure où les échantillons sont séquencés à des profondeurs importantes, le temps d'exécution demeure conséquent afin de reproduire et d'éditer les lectures générées *in silico* pour y introduire des mutations pour mimer le bruit de fond observé dans les témoins.

Par ailleurs, LowVarFreq permet d'estimer à partir des fichiers témoins les artefacts récurrents pour un panel de séquençage donné en supposant que ces artefacts seront communs à tous les échantillons séquencés. Néanmoins, nous savons que des artefacts de séquençage spécifiques d'un run peuvent exister et que le bruit de fond biologique est très étroitement lié à la qualité des échantillons analysés. LowVarFreq est en mesure de générer des tailles de lectures fixes au moment de la création du fichier BAM *in silico*. Certains échantillons dégradés ou de cfDNA ont la particularité d'avoir des fragments d'ADN courts (voir section I.C.2) pouvant entraîner l'apparition d'artefacts spécifiques comme des artefacts d'alignement. LowVarFreq n'est pas capable de générer des tailles de lectures variables.

Nous avons observé que LowVarFreq parvenait à identifier les artefacts d'oxydation par évaluation itérative d'un seuil d'enrichissement en transition. Si ce test permet d'objectiver la présence des variants à de faibles VAF, il reste limité à la quantification d'un bruit de fond sans pour autant permettre de le corriger. L'application des seuils de sensibilité de l'algorithme ne permet que de masquer les variants pour éviter de rendre des faux positifs et ne permet pas de discriminer en dessous de ces seuils le signal biologique du bruit de fond d'oxydation. Afin de palier à ce problème, il est souhaitable de développer des approches intégrant des barcodes moléculaires comme nous le verrons dans le chapitre suivant.

A.5. Article

LEUKEMIA & LYMPHOMA
2019, VOL. 60, NO. 2, 498–502
https://doi.org/10.1080/10428194.2018.1492123

Taylor & Francis
Taylor & Francis Group

Check for updates

LETTER TO THE EDITOR

Somatic mutations of cell-free circulating DNA detected by targeted next-generation sequencing and digital droplet PCR in classical Hodgkin lymphoma

Lucile Bessi[a], Pierre-Julien Viailly[a], Elodie Bohers[a], Philippe Ruminy[a], Catherine Maingonnat[a], Philippe Bertrand[a], NasrinSarafan Vasseur[a], Ludivine Beaussire[a], Marie Cornic[a,b], Pascaline Etancelin[a], Vincent Camus[c], Jean-Michel Picquenot[b], Hervé Tilly[a,c], Aspasia Stamatoullas[a,c] and Fabrice Jardin[a,c]

[a]INSERM U1245, Centre Henri Becquerel and Rouen University, Rouen, France; [b]Department of Pathology, Centre Henri Becquerel, Rouen, France; [c]Department of Clinical Hematology, Centre Henri Becquerel, Rouen, France

ARTICLE HISTORY Received 10 March 2018; Revised 26 May 2018; Accepted 13 June 2018

Classical Hodgkin lymphoma (cHL) accounts for 30% of all lymphomas [1]. Currently, 20–25% of cHL cases relapse or are refractory after first-line standard treatments, leading to the need to better understand the underlying mechanisms and to identify new biomarkers that predict adverse outcomes and new targets for targeted therapies.

Over the past decade, the emergence of next-generation sequencing (NGS) technologies has provided new insights into the genomic characterization of cHL [2–4]. NF-κB (*NFKBIA, NFKBIE, TNFAIP3*) [4–6] and JAK/STAT (*JAK2, SOCS1, PTPN1, PTPN2, STAT6*) [2–4,7–11] are the major pathways involved in the lymphomagenesis of cHL. Tumor immune escape is also involved with mutations of *B2M* or *CIITA* [3,4]. More recently, we identified a recurrent mutation of the *XPO1* gene, coding for a protein involved in the nuclear export [12].

However, the scarcity of Hodgkin and Reed–Sternberg (HRS) cells in biopsy samples (0.1 to 10% of total tumor tissue [1]) and the intra-tumor heterogeneity make it difficult to detect somatic mutations in these cells without prior microdissection. In this setting, the use of very sensitive sequencing technologies such as NGS and the use of plasma as a source of tumor DNA may be promising approaches to routinely provide cHL genotyping [13]. For instance, we were able to detect the E571K *XPO1* hotspot from the plasma of cHL patients and to track the mutation during treatment [12]. The concept of the 'liquid biopsy' raises numerous possibilities for the diagnosis and follow-up of patients.

In this study, using routinely applicable NGS and digital droplet polymerase chain reaction (PCR) technologies, we sought to determine whether the pattern of acquired mutations observed in the tissue biopsy DNA of cHL biopsies can also be detected in cftDNA at the time of diagnosis. We analyzed 24 cHL cases from a single center with available matched tissue biopsy DNA and plasma collected before any treatment. Analyses were carried out in accordance with the Helsinki Declaration. All patients provided informed consent for the collection of biological, clinical, and biomarker data.

The main clinical features of the patients are summarized in Table 1. CfDNA was extracted from plasma, and tissue biopsy DNA was extracted from frozen lymph node samples by standard methods or from paraffin-embedded lymph node samples as previously described [12]. The mean cfDNA concentration was 36.2 ng/mL of plasma (range, 17.36–61.2). DNA was sequenced using an Ion Torrent Personal Genome Machine (Life Technologies). Ninety-six nanogram of genomic tissue biopsy DNA and a mean of 12 ng (range, 8.3–15.8) of cfDNA were submitted for NGS using a laboratory-developed Hodgkin panel set designed to identify mutations in *B2M, STAT6, XPO1, NFKBIE, PTPN1,* and *TNFAIP3* genes according to hotspots regions described in previously cited studies. The design covers 6.8 kb and generates 92 amplicons of length inferior to 140bp (Supplementary Table S1). Twenty cycles of amplification for frozen biopsies and twenty-three for cfDNA and FFPE biopsies were performed. The steps of the emulsion PCR and enrichment of the templated Ion Sphere™ Particles from amplified libraries (Ion AmpliSeq™ Library Kit 2.0) were performed with the Ion PGM Hi-Q View Chef™ kit using the Ion Chef™ System (Life Technologies) and then loaded and sequenced on an Ion 318™ v2 Chip (Life Technologies, Waltham, MA).

Bioinformatic analysis of the data was performed for alignment to the reference genome sequence (hg 19), base-calling and quality control using manufacturer's software, and a low-frequency variant detection pipeline

CONTACT Fabrice Jardin fabrice.jardin@chb.unicancer.fr Department of Clinical Hematology, Centre Henri Becquerel, 1 rue d'Amiens, 76038 Rouen, France

Supplemental data for this article can be accessed here.

Table 1. Clinical characteristics and somatic variants (insertion/deletion/single nucleotide variant) detected by sequencing in tissue biopsy DNA and cell-free plasma circulating DN.

									Tissue biopsy DNA			Circulating DNA				
Patient number	Sex	Age (years)	Histologic subtype	Stage	PFS (months)	OS (months)	Gene and mutation	Biopsy nature	VAF (%)	Mean VAF (%)	Mean depth	VAF (%)	Mean VAF (%)	Mean depth	Concentration (ng/mL of plasma)	Circulating tumoral DNA (hGE/mL)
1	M	41	NS	II	80	80	PTPN1 R56W	Frozen	1.68	1.675	11162	0	0	5239	17.36	0
							XPO1 E571K		1.67				NI			
2	M	20	NS	II	10	70	*No mutation found*	Frozen			9111			12237	25.6	NI
3	F	23	NS	II	60	60	STAT6 D419G	Frozen	1.45	1.06	6443	9.25	5.435	11583	28.88	476
							STAT6 N417D		1.45			9.2				
							NFKBIE E250X		0.75			0				
							B2M M1L		0.59			3.29				
4	M	21	NS	IV	52	52	STAT6 K423E	Frozen	0	1.455	8210	0.66	0.81	10558	28	69
							XPO1 E571K		2.91			0.95				
5	M	27	NS	II	52	52	PTPN1 Y176S	FFPE	1.05	1.275	14324	1.46	0.73	11515	35.04	78
							NFKBIE T253fs		1.5			0				
6	M	32	MC		77	77	B2M c.-14G > A	FFPE	0	0	4305	1.27	0.923	5598	31.68	89
							XPO1 F610L		0			0.7				
							TNFAIP3 N399D		0			0.8				
7	M	27	NS	IV	6	28	XPO1 Q626R	FFPE	0.54	0.53	3970	0	0	9500	29.84	0
							PTPN1 c.*54C > T		0.52			0				
8	M	29	MC	III	38	80	No mutation found	FFPE			16857			11779	26.08	NI
9		27	NS		16	80	STAT6 N417Y	FFPE	2.6	1.354	4007	0	0	19074	39.68	0
							STAT6 D523N		1.01			0				
							PTPN1 c.*22C > T		1.1			0				
							PTPN1 S151L		1.02			0				
							TNFAIP3 R71X		1.04			0				
10	F	21	NS	II	57	57	TNFAIP3 N296K	FFPE	1.64	1.665	3416	0	0	9976	30	0
							TNFAIP3 L286X		1.69			0				
11	F	55	NS	II	66	66	No mutation found	FFPE			11235			9969	53.2	NI
12	F	62	NS	IV	13	24	No mutation found	FFPE			10974			9388	53.6	NI
13	M	36	NS	III	64	64	No mutation found	FFPE			9555			7842	61.2	NI
14	F	32	MC	III	57	57	XPO1 E571K	FFPE	2.13	2.297	11450	0	0	8738	41.2	0
							STAT6 N417Y		2.14			0				
							STAT6 D419N		2.62			0				
15	M	23	NS	III	47	47	B2M c.67 + 2T > G	FFPE	1.34	1.35	6572	0	0.363	8242	30.24	33
							PTPN1 A69T		0.98			0				
							STAT6 N417Y		1.73			1.09				
16	M	25	NS	II	64	64	PTPN1 c.*22C > T	FFPE	3.6	3.44	9630	0	1.097	11630	34.48	115
							TNFAIP3 R596G		3.22			0				
							STAT6 N417Y		3.5			3.29				
17	F	41	NS	IV	8	76	No mutation found	Frozen			7329			13262	52	NI
18	F	43	MC	IV	67	67	B2M c.-10T > C	Frozen	0	0	6979	0.58	0.636	7311	39.92	77
							B2M c.-11T > C		0			0.62				
							STAT6 G509D		0			0.6				
							XPO1 C582R		0			0.73				
							PTPN1 S55G		0			0.65				
19	F	28	NS	II	9	69	B2M V113D	Frozen	0.65	0.853	7665	0	0	9693	23.76	0
							B2M c.346 + 2T > A		0.56			0				
							XPO1 E571K		1.35			0				

(*continued*)

Table 1. Continued.

Patient number	Sex	Age (years)	Histologic subtype	Stage	PFS (months)	OS (months)	Gene and mutation	Biopsy nature	Tissue biopsy DNA			Circulating DNA				Circulating tumoral DNA (hGE/mL)
									VAF (%)	Mean VAF (%)	Mean depth	VAF (%)	Mean VAF (%)	Mean depth	Concentration (ng/mL of plasma)	
20	F	25	NS	II	62	62	XPO1 E571K	Frozen	0.36	0.06	6687	2.25	1.252	10220	22.32	85
							NFKBIE T253fs		0			1.25				
							TNFAIP3 Q784X		0			0.95				
							TNFAIP3 D517Y		0			0.85				
							TNFAIP3 c.*36G > A		0			0.95				
							STAT6 N417Y		0			1.26				
21	M	43	LR	II	60	60	TNFAIP3 R569P	Frozen	0	0	5436	2.34	2.34	8551	27.2	193
22	M	30	NS	II	54	54	B2M c.-8G > A	Frozen	0	0	9271	0.93	1.93	8790	27.04	158
							STAT6 N421K		0			2.93				
23	M	35	MC	II	59	59	XPO1 E571K	Frozen	1.58	2	8785	16.2	13.2	10561	59.64	2386
							B2M L7S		2.3			6.2				
							TNFAIP3 C392X		2.12			17.2				
24	M	23	MC	II	48	48	No mutation found	Frozen			4088			9520	51.6	NI

M: male; F: female; NS: nodular sclerosis; MC: mixed cellular; LR: lymphocyte-rich; c.: mutation coordinate on coding DNA; c.*: variation in 3' untranslated region (UTR); fs: frameshift; VAF: variant allelic frequency; NI: not interpretable; hGE/mL: haploid genome equivalents per mL of plasma.

laboratory-developed for this project called LowVarFreq. The approach consisted of combining the use of four low-frequency variant detection algorithms described in the literature (Supplementary Method S1). Artifacts related to DNA degradation were filtered according to a developed in-house method (Supplementary Method S2).

Mutations were detected in 13/24 biopsies (54.2%) and 11/23 plasma samples (47.8%). Sequencing failed for one plasma sample. In 7/23 cHL cases (30.4%), we observed similar or partially similar somatic mutations in cfDNA and matched tissue biopsy DNA (Table 1). The discrepancy observed between tissue biopsy DNA and cfDNA is likely explained by the low variant allele frequency observed in both tissues that not reached in all matched cases, the sensitivity threshold. The median variant allelic frequencies (VAFs) for variants identified in both cfDNA and matched tissue biopsy DNA were 3.3% and 1.8%, respectively ($p = .033$). Overall, we found at least one mutation in tissue biopsy DNA and/or cfDNA in 17/24 (70.8%) cases. Interestingly, for the case #3, the sub-clonal distribution of some mutations detected in tumor tissue, as indicated by the VAF distribution of each individual variant, was also observed in cfDNA.

The concentrations of the cftDNA were expressed in haploid genome equivalents per mL of plasma [hGE/mL] and calculated by multiplying the mean VAF by the concentration of cfDNA [pg/mL of plasma] and dividing by 3.3, as previously described in the publication of Scherer et al. [14]. For seven patients without any mutation detected in both tissue biopsy and cfDNA, we observed a significant amount of cfDNA in 5/7 (#11, #12, #13, #17, and #24). The failure to detect mutation can be explained by an insufficient sensitivity of our 6-gene panel and/or by an increased release of non-tumor circulating DNA (Supplementary Table S2). In contrast to results reported in DLBCL [15], in this small cohort, we did not found any correlation between cftDNA concentrations of the 11 positive plasma sample and disease stage, age at the time of diagnosis, or histological subtypes.

We found a mean of 2.13 mutations per cHL case (range, 0–6) (Table 1). Although the panel targeted only the DNA binding domain of *STAT6*, we identified 12 mutations in 9 (37.5%) patients. These mutations were located on the same allele when the patient was double-mutated. The hotspot N417Y was the most common (5/24 cases, 20.8%). *XPO1* was found to be mutated in 9/24 cHL cases (37.5%) including 6 cases with the E571K mutation (25%). Overall, 7/24 cHL cases (29.2%) harbored *B2M* mutations. The start codon was mutated in only one case, and six mutations were located in 3 'or 5' UTR. The L7S mutation was found in one case. Only the GTAA deletion on exon 1 of *NFKBIE* was targeted by our Hodgkin panel; 2/24 cases (8.3%) had this deletion. A point mutation leading to the appearance of a premature stop codon was identified in a third case. In all, 7/24 cases (29.2%) displayed *PTPN1* mutations without any recurrence, and 7/24 cases (29.2%) harbored *TNFAIP3*

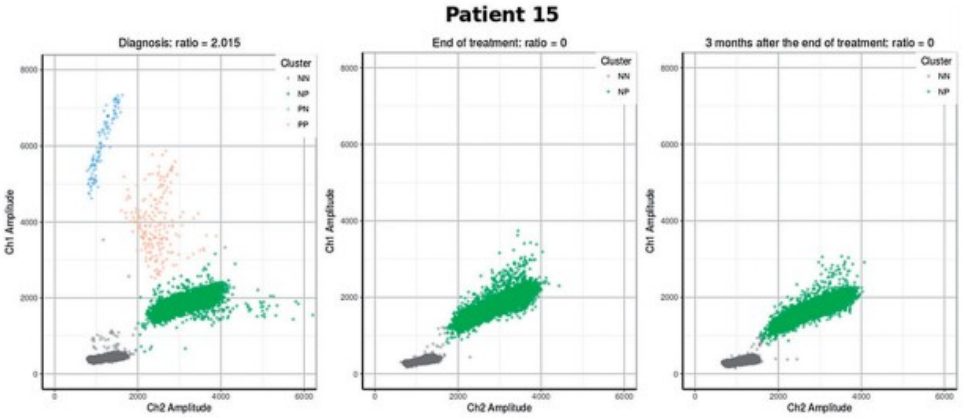


Figure 1. Example of quantification of the STAT6 N417Y mutation with ddPCR in samples from a patient with mutations on NGS and evolution at the end of treatment. Blue cluster (FAM+/VIC−): droplets with mutated sequences. Orange cluster (FAM+/VIC+): droplets with both mutated and wild-type sequences. Green cluster (FAM-/VIC+): droplets with wild-type sequences. Gray cluster (FAM−/VIC−): droplets without amplification.

mutations. In almost all of cases, the mutations were point mutations. Details on percentage of mutated cases depending on the nature of the sample are specified on Supplementary Table S3. In this small series, no statistically significant relationship was observed between the mutational pattern and the stage of the disease.

Because ddPCR is a fast and inexpensive technique to identify recurrent variants, we designed a ddPCR assay targeting the most recurrent mutation found in this cohort, the N417Y mutation of the *STAT6* gene (Supplementary Method S3). The sensitivity of the assay, established using serial dilutions of the L-1236 cell line harboring this mutation, was 0.14%. A null ratio was found for patient #3, carrying the mutation N417D (c.1249A > G), thus demonstrating the specificity of the assay (Supplementary Figure S3). The results between NGS and ddPCR were consistent, as 5/5 NGS-mutated samples (two cases mutated only in the biopsy and three cases mutated in both tissue biopsy and cfDNA, #15, #16, and #20) were also found to be mutated by ddPCR. Of note, the STAT6 N417Y hotspot detectable in three cases by ddPCR in cfDNA at the time of diagnosis was undetectable in 3/3 cases at the end of treatment when a complete remission was obtained (Figure 1 and Supplementary Figure S4). Our results confirm that ddPCR can be used to monitor minimal residual disease (MRD) in cHL patients with recurrent mutations.

To our knowledge, this is only the second study reporting the detection of somatic mutations in cHL from circulating cfDNA by a routine NGS approach. A comparison with previous results recently published by Spina et al. [4] shows very comparable results regarding the rate of mutations detected in cfDNA targeting *STAT6* (30.6%/37.5%), *B2M* (21.7%/16.2%), *XPO1* (21.7%/11.2%), *NFKBIE* (4.3%/6.2%), or *TNFAIP3* (17.4%/35%) (see supplementary Table S3 for details comparison).

Both studies indicate that cfDNA can be used in cHL to detect somatic variants, confirming the concept of 'liquid biopsy' in this type of tumor. However, we identified mutations in cfDNA in only half of cases, requiring an improvement of the method before its usage in routine. This implies an extension of the panel, a better identification of the small in/del (<4 bp), frequently considered like artifacts, and an improvement of pre-analytical steps for limiting DNA degradation. Giving these technical considerations, we started a prospective study with the aim of serially sequencing cfDNA during cHL treatment and follow-up (registered at *clinicaltrials.gov* as NCT02815137). If the current preliminary results are confirmed by the prospective study, new strategies should be proposed for both tailored diagnosis and treatment based on the simple detection and quantification of acquired mutations in the plasma of cHL patients.

Potential conflict of interest: Disclosure forms provided by the authors are available with the full text of this article online at https://doi.org/10.1080/10428194.2018.1492123.

References

[1] Swerdlow SH, Campo E, Harris NL, et al. WHO classification of tumours of haematopoietic and lymphoid tissues. 4th ed. Geneva, Switzerland: World Health Organization; 2008.
[2] Tiacci E, Ladewig E, Schiavoni G, et al. Pervasive mutations of JAK-STAT pathway genes in classical Hodgkin lymphoma. Blood. 2018;131:2454–2465.
[3] Reichel J, Chadburn A, Rubinstein PG, et al. Flow sorting and exome sequencing reveal the oncogenome of

primary Hodgkin and Reed-Sternberg cells. Blood. 2015;125:1061–1072.
[4] Spina V, Bruscaggin A, Cuccaro A, et al. Circulating tumor DNA reveals genetics, clonal evolution and residual disease in classical Hodgkin lymphoma. Blood. 2018;131:2413–2425.
[5] Mansouri L, Noerenberg D, Young E, et al. Frequent NFKBIE deletions are associated with poor outcome in primary mediastinal B-cell lymphoma. Blood. 2016;128:2666–2670.
[6] Schmitz R, Hansmann M-L, Bohle V, et al. TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma. J Exp Med. 2009;206:981–989.
[7] Van Roosbroeck K, Cox L, Tousseyn T, et al. JAK2 rearrangements, including the novel SEC31A-JAK2 fusion, are recurrent in classical Hodgkin lymphoma. Blood. 2011;117:4056–4064.
[8] Mottok A, Renné C, Seifert M, et al. Inactivating SOCS1 mutations are caused by aberrant somatic hypermutation and restricted to a subset of B-cell lymphoma entities. Blood. 2009;114:4503–4506.
[9] Weniger MA, Melzner I, Menz CK, et al. Mutations of the tumor suppressor gene SOCS-1 in classical Hodgkin lymphoma are frequent and associated with nuclear phospho-STAT5 accumulation. Oncogene. 2006;25:2679–2684.
[10] Gunawardana J, Chan FC, Telenius A, et al. Recurrent somatic mutations of PTPN1 in primary mediastinal B cell lymphoma and Hodgkin lymphoma. Nat Genet. 2014;46:329–335.
[11] Kleppe M, Tousseyn T, Geissinger E, et al. Mutation analysis of the tyrosine phosphatase PTPN2 in Hodgkin's lymphoma and T-cell non-Hodgkin's lymphoma. Haematologica. 2011;96:1723–1727.
[12] Camus V, Stamatoullas A, Mareschal S, et al. Detection and prognostic value of recurrent exportin 1 mutations in tumor and cell-free circulating DNA of patients with classical Hodgkin lymphoma. Haematologica. 2016;101:1094–1101.
[13] Vandenberghe P, Wlodarska I, Tousseyn T, et al. Non-invasive detection of genomic imbalances in Hodgkin/Reed-Sternberg cells in early and advanced stage Hodgkin's lymphoma by sequencing of circulating cell-free DNA: a technical proof-of-principle study. Lancet Haematol. 2015;2:e55–e65.
[14] Scherer F, Kurtz DM, Newman AM, et al. Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. Sci Transl Med. 2016;8:364ra155.
[15] Bohers E, Viailly PJ, Dubois S, et al. Somatic mutations of cell-free circulating DNA detected by next-generation sequencing reflect the genetic changes in both germinal center B-cell-like and activated B-cell-like diffuse large B-cell lymphomas at the time of diagnosis. Haematologica. 2015;100:e280–e284.

B. Détection des mutations avec UMI : UMI-VarCal

B.1. État de l'art

Nous avons vu dans le chapitre précédent les limites des approches sans UMI pour la détection d'événements somatiques à de faibles fréquences. Nous nous intéresserons dans cette partie à la détection des mutations dans des données de séquençage QIAseq intégrant des UMI de taille 12 au moment de la préparation de la librairie. Ces séquences aléatoires doivent permettre de discriminer les vrais variants de faible fréquence des artefacts de séquençage ou des erreurs de PCR.

Il existe dans la littérature trois algorithmes de *variant calling* capables de prendre en charge spécifiquement des données de séquençage avec UMI : DeepSNVMiner [277], MAGERI [278] et smCounter2 [279]. Ces trois outils partagent une approche commune visant à corriger les artefacts en quantifiant pour chaque groupe de lectures issues d'un même UMI unique et à chaque position la base majoritairement présente. En effet, en théorie et sans l'absence d'erreurs techniques, toutes les lectures porteuses du même UMI sont censées être parfaitement identiques à chaque position de l'alignement. Une lecture consensus est générée à partir de l'ensemble des lectures de chaque UMI puis différents filtres sont appliqués afin d'extraire une liste de variations candidates.

Pour extraire les listes de variations, les *variant callers* classiques basés sur les comptages des lectures, tout comme DeepSNVMiner et smCounter2, utilisent la fonction *pileup* de l'utilitaire SAMtools permettant d'obtenir à partir des fichiers BAM le nombre d'insertions, de délétions et de substitutions à chaque position de l'alignement. En fonction des outils, des modèles mathématiques sont utilisés afin d'estimer le bruit de fond de séquençage et de détecter les variants dont la fréquence n'est pas compatible avec le bruit de fond.

B.2. Objectifs et approche suivie

Nous avons souhaité proposer une implémentation d'un nouvel algorithme de détection capable de prendre en charge nativement des données de séquençage intégrant des UMI. Cet outil, baptisé UMI-VarCal, a pour particularité de prendre en charge la détection des substitutions, des insertions et des délétions à partir de données de séquençage pairées. Afin d'améliorer les performances de l'algorithme, une nouvelle fonction *pileup* a été

intégralement développée et intégrée dans l'outil de sorte à augmenter ses performances en terme de mémoire et de temps d'exécution pour permettre l'analyse d'échantillons profonds.

Afin de valider les performances de l'outil, plusieurs échantillons *in silico* ont été générés et des variants de fréquence connue ont été insérés dans les fichiers BAM de sorte à constituer une liste de vrai positifs. Les résultats des algorithmes DeepSNVMiner, MAGERI, smCounter2 et UMI-VarCal ont été comparés en terme de sensibilité, de spécificité et de temps d'exécution.

B.3. Implémentation

L'implémentation de UMI-VarCal repose sur 6 étapes de traitement informatique : l'obtention des données de comptage, l'estimation du bruit de fond de séquençage par position, la recherche de variants candidats, une annotation des variants via l'information portée par les UMI, une étape de filtration basée sur le biais de brin et enfin l'élimination des artefacts dans les régions riches en homopolymères.

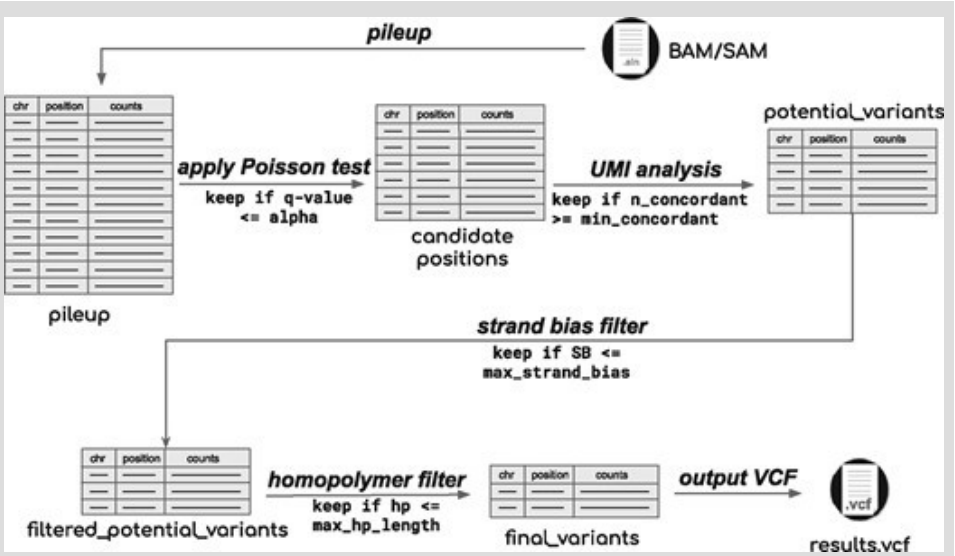


Figure 55: Workflow de l'algorithme UMI-VarCal.

UMI-VarCal prend en entrée un fichier BAM aligné duquel il réalise un *pileup*. Ce *pileup* est filtré par différentes procédures (estimation du bruit de fond, filtration via l'information portée par les UMI, le biais de brin et le nombre d'homopolymères). Les différentes étapes sont décrites en détails dans cette section. Finalement, un fichier VCF est généré dans lequel les variants filtrés sont présents.

Pileup

La première étape d'UMI-VarCal est de réaliser, à partir des séquences alignées, un *pileup* qui consiste à compter pour chaque position de l'alignement le nombre de A, T, G, C, d'insertions et de délétions. Cette matrice de comptage est en réalité limitée à une liste de régions ciblées présentes dans un fichier au format BED. La fonction *pileup* intégrée dans UMI-VarCal a la particularité de lister pour chaque position et pour chaque allèle la liste des séquences des UMI extraites du nom de chaque lecture.

Il est possible, au moment du lancement de l'algorithme, de demander à ce que le fichier *pileup* soit stocké physiquement dans un fichier de sorte à pouvoir exécuter UMI-VarCal avec différents paramètres sans avoir à générer à chaque exécution le *pileup*. Cette fonctionnalité est tout particulièrement utile pour mettre au point les paramètres d'analyse bioinformatiques.

Estimation du bruit de fond par position

Afin de discriminer la présence d'un variant liée à une erreur technique de celle liée à signal biologique, une étape d'estimation du bruit de fond est intégrée à l'algorithme. Nous savons que le bruit de fond n'est pas parfaitement aléatoire et est souvent lié au contexte de détection de la variation. Il est donc intéressant de ne pas généraliser un modèle d'estimation du bruit de fond global pour l'ensemble des positions de l'alignement.

UMI-VarCal utilise les scores PHRED de qualité par base à chaque position qui sont proportionnels à la probabilité d'identification correcte de la base dans chaque lecture. Dans la mesure où chaque base séquencée possède son propre score de qualité, le score de qualité moyen de l'ensemble des bases à cette position de l'alignement peut être calculé. Ce score moyen X_i, où *i* correspond à une position de l'alignement, est converti en probabilité de taux d'erreur comme suit :

$$\varepsilon_i = 10^{-Xi/10}$$

Recherche des variants candidats

Le *pileup* qui contient le nombre de substitutions, d'insertions et de délétions par position est parcouru de sorte à déterminer si un événement alternatif, c'est à dire qui diffère de la base du génome de référence, est présent de manière significative ou non. Pour cela, un test de Poisson est appliqué de sorte à déterminer si les comptages observés dans l'échantillon sont expliqués ou non par la présence d'un bruit de fond à la position de l'alignement. Le bruit de fond est considéré comme étant le nombre de bases alternatives n'étant ni celle du génome de

référence ni celle du candidat à la position. Ce modèle est particulièrement permissif et n'est utilisé que pour filtrer des positions qui ne seraient pas intéressantes à évaluer dans les étapes ultérieures afin de réduire le temps de calcul.

Dans la mesure où beaucoup de tests sont appliqués et peuvent conduire à un grand nombre de faux positifs, une procédure de correction des p-valeurs du test par Benjamini-Hochberg est appliquée. Cette correction maintient un niveau de sensibilité du test en adéquation avec le niveau de sensibilité de détection désiré. Néanmoins, de nombreux faux-positifs passent encore au travers de ce test et nécessitent des niveaux de filtres supplémentaires décrits dans les trois prochaines sections.

Procédure d'évaluation des variants par les UMI

Les couples positions et allèles qui ont passé l'étape de filtration précédente sont toutes évaluées via l'information portée par les UMI. L'étape de *pileup* a permis d'extraire la liste de tous les UMI par allèle et par position de sorte à déterminer et quantifier trois classes de barcodes :

- UMI_{ref} qui quantifie le nombre d'UMI porteurs de l'allèle de référence
- UMI_{alt} qui quantifie le nombre d'UMI porteurs de l'allèle alternatif que l'on souhaite évaluer
- UMI_{noise} qui vise à lister l'ensemble des UMI pour lesquels les lectures associées à chaque UMI ne portent pas le même allèle. Ces UMI sont qualifiés de discordants.

De façon théorique, si un variant est un vrai variant, c'est à dire que l'événement était présent dans un fragment d'ADN initial, alors après ligation d'un UMI unique et amplification du fragment les lectures portant cet UMI sont censées toutes porter la mutation. Les classes UMI_{ref} et UMI_{alt} visent justement à quantifier pour chaque variant candidat ce nombre d'événements. A l'inverse, une erreur de PCR ou de séquençage va conduire à l'introduction d'un artefact seulement dans une sous-population des lectures portant un UMI. L'UMI sera alors qualifié de discordant et quantifié dans la classe UMI_{noise}. Deux exemples sont rapportés sur la figure 56.

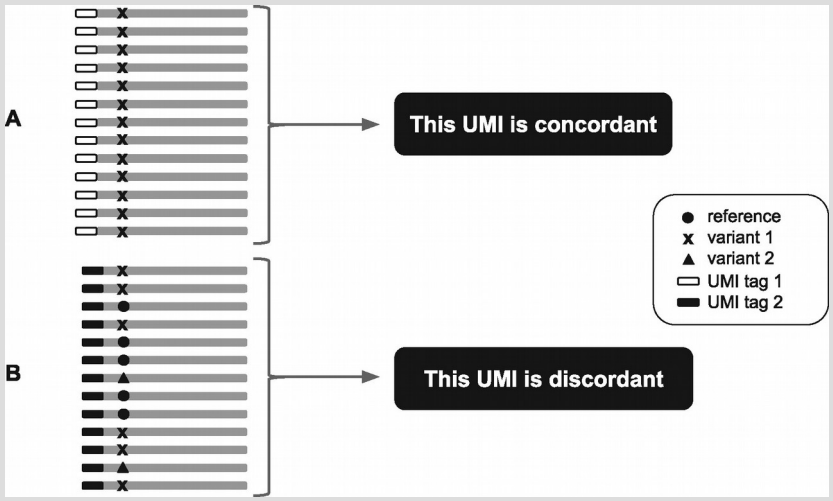


Figure 56: Classification des UMI concordants et discordants.

Ce schéma représente en (A) la présence d'un ensemble de lectures portant le même variant avec le même UMI. L'UMI est donc considéré comme concordant pour l'évaluation de ce variant. On retrouve à l'inverse en (B) un UMI amplifié dont les lectures ne portent pas le même allèle à la même position. Cet UMI, de fait, est considéré comme étant discordant.

Filtration des variants par biais de brin et homopolymères

Une estimation du biais de brin est réalisée sur les variants ayant passés les filtres précédents. Il a été démontré qu'une surreprésentation d'un variant sur l'une des deux lectures *reverse* ou *forward* est une source majeure d'artefacts pour les séquenceurs Illumina [280]. Un score de biais de brin est ainsi calculé de sorte à éliminer les variants se situant dans ces régions . Ce calcul de score est détaillé dans la publication à la fin de ce chapitre. Le seuil de filtration est un paramètre modifiable de l'algorithme.

La qualité de séquençage dans les régions riches en homopolymères est plus faible du fait d'une baisse locale des scores de base PHRED [281]. Pour ces raisons, UMI-VarCal recherche pour chaque variant si celui-ci se trouve dans un homopolymère ou non et filtre par défaut ceux dans des homopolymères de longueur supérieure ou égale à 7.

Implémentation de UMI-VarCal

L'implémentation de toutes ces étapes, depuis le fichier BAM jusqu'à la filtration des variants, a été réalisée en Python3. Afin d'améliorer les performances sur des jeux de données profonds, chacun des modules qui composent le programme ont été compilés via Cython afin

de diminuer les temps d'exécution. UMI-VarCal est exécutable directement en ligne de programme.

Le code source du logiciel est librement disponible à l'adresse https://gitlab.com/vincent-sater/umi-varcal/.

B.4. Performances

UMI-VarCal a été testé et validé sur plusieurs échantillons biologiques et sur des données d'alignement simulées. Nous ne détaillerons dans cette section que les résultats *in silico* pour évaluer les performances informatiques de l'outil, les autres données biologiques étant détaillées dans l'article visible en fin de chapitre.

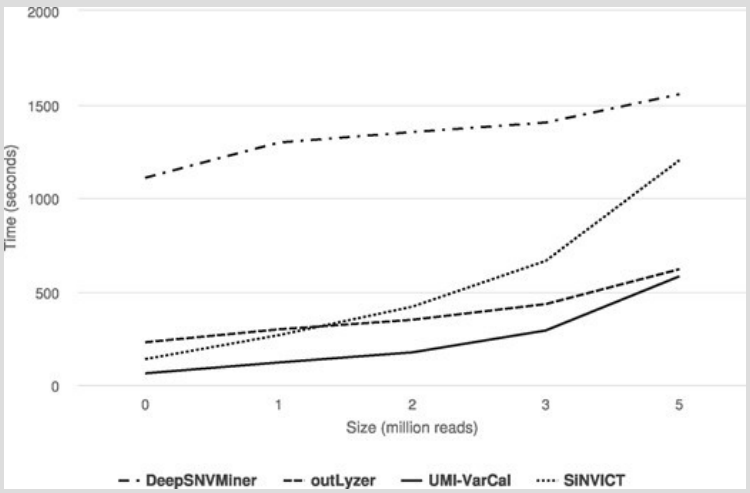


Figure 57: Comparaison des temps d'exécution de UMI-VarCal, DeepSNVMiner, OutLyzer et SINVICT sur un jeu de données simulées.

Le temps d'exécution devient un enjeu majeur en bioinformatique afin de permettre l'analyse de panels de séquençage larges et séquencés profondément notamment pour l'analyse des échantillons de cfDNA. Afin de comparer les performances de UMI-VarCal, DeepSNVMiner, OutLyzer et SINVICT, 5 échantillons *in silico* ont été générés avec respectivement 1, 2, 3, 5 et 10 millions de séquences. Le test repose sur l'exécution 3 fois de chacun des algorithmes sur chacun des fichiers sur un processeur de 2,20 GHz. Les résultats montrent une très bonne performance de UMI-VarCal sur les différents fichiers simulés avec un temps d'exécution inférieur aux autres algorithmes (figure 57).

A noter que les algorithmes outLyzer et UMI-VarCal ont la capacité d'être exécutables sur plusieurs cœurs d'un processeur ce qui entraîne naturellement un temps d'exécution moindre

que les algorithmes non parallélisés, et tout particulièrement sur les gros fichiers. Par ailleurs, si OutLyzer a lui aussi des temps d'exécution intéressants, celui-ci ne tient pas compte de l'information portée par les UMI et se base uniquement sur le nombre de lectures amplifiées afin d'établir une liste de variants. En terme d'utilisation mémoire, UMI-VarCal ne semble pas réellement impacté par la taille du fichier en entrée . En réalité, cette consommation n'est pas facile à évaluer car de nombreux paramètres l'influencent : la largeur du panel séquencé, la profondeur de l'échantillon ou encore le facteur d'amplification de l'échantillon.

B.5. Limitations et perspectives

La détection des variants de faibles fréquences est un enjeu majeur pour l'analyse des échantillons de cfDNA. Les algorithmes de *variant calling* ne tenant pas compte des UMI et qui sont les plus performants nécessitent souvent un séquençage d'un tissu sain apparié afin d'effectuer la comparaison avec un échantillon testé. Cette stratégie est difficilement applicable sur des analyses de cfDNA tant la référence est difficile à déterminer.

Les autres algorithmes ne nécessitant pas d'échantillons appariés comme outLyzer et SINVICT parviennent à descendre relativement bas en terme de sensibilité avec des variants à des fréquences alléliques de 0,5%. Cette recherche de sensibilité se fait en revanche au détriment d'un très grand nombre de faux positifs détectés. UMI-VarCal, par sa capacité à traiter l'information portée par les UMI, parvient à améliorer la filtration de ces faux positifs en éliminant les erreurs d'amplification et les erreurs de séquençage.

Cette approche de détection de variants utilisant les UMI a cependant une limitation majeure : le nombre d'UMI discordants a une très nette tendance à augmenter à mesure que la profondeur de séquençage augmente. UMI-VarCal nécessite d'observer au moins deux à trois lectures par UMI afin d'exploiter pleinement ses capacités. En cas de sur-séquençage, c'est à dire en cas d'augmentation très importante du facteur de relecture de chaque UMI, la probabilité que l'une des lectures de l'UMI intègre une erreur de séquençage augmente. Il est donc important de bien adapter la profondeur de séquençage moyenne désirée à la quantité d'ADN utilisée au moment de la préparation de la librairie de séquençage. Plus la quantité d'ADN utilisée est importante et plus le nombre d'UMI lus par position sera normalement important et donc plus la profondeur de séquençage devra être importante. A l'inverse, il n'est pas recommandé de sur-séquencer des échantillons avec un faible rendement d'extraction.

B.6. Article

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Sequence analysis

UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries

Vincent Sater [1,*], Pierre-Julien Viailly [2,3], Thierry Lecroq [1], Élise Prieur-Gaston [1], Élodie Bohers [2,3], Mathieu Viennot [2,3], Philippe Ruminy [2,3], Hélène Dauchel [2,3], Pierre Vera [1,2] and Fabrice Jardin [2,3]

[1] Normandie Univ, UNIROUEN, LITIS EA 4108, 76000 Rouen, France and
[2] Department of Pathology, Centre Henri Becquerel, Rouen, 76000, France and
[3] INSERM U1245, University of Normandie UNIROUEN, Rouen, 76000, France.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Next Generation Sequencing (NGS) has become the go-to standard method for the detection of Single Nucleotide Variants (SNV) in tumor cells. The use of such technologies requires a PCR amplification step and a sequencing step, steps in which artifacts are introduced at very low frequencies. These artifacts are often confused with true low-frequency variants that can be found in tumor cells and cell-free DNA. The recent use of Unique Molecular Identifiers (UMI) in targeted sequencing protocols has offered a trustworthy approach to filter out artifactual variants and accurately call low frequency variants. However, the integration of UMI analysis in the variant calling process led to developing tools that are significantly slower and more memory consuming than raw-reads-based variant callers.
Results: We present UMI-VarCal, a UMI-based variant caller for targeted sequencing data with better sensitivity compared to other variant callers. Being developed with performance in mind, UMI-VarCal stands out from the crowd by being one of the few variant callers that don't rely on SAMtools to do their pileup. Instead, at its core runs an innovative homemade pileup algorithm specifically designed to treat the UMI tags in the reads. After the pileup, a Poisson statistical test is applied at every position to determine if the frequency of the variant is significantly higher than the background error noise. Finally, an analysis of UMI tags is performed, a strand bias and a homopolymer length filter are applied to achieve better accuracy. We illustrate the results obtained using UMI-VarCal through the sequencing of tumor samples and we show how UMI-VarCal is both faster and more sensitive than other publicly available solutions.
Availability: The entire pipeline is available at https://gitlab.com/vincent-sater/umi-varcal-master under MIT license.
Contact: vincent.sater@gmail.com

1 Introduction

Old traditional sequencing technologies have showed their limits and were rapidly replaced by next generation sequencing (NGS) for the detection of genomic aberrations like single nucleotide variants (SNV) and copy number variations (CNV). However, the use of such technologies requires extracted genomic DNA to be fragmented to produce DNA fragments. These fragments constitute the DNA library that has to be massively amplified in order to produce enough fragments and cover all the targeted regions. These fragments are finally sequenced by a NGS sequencer to generate reads. Nowadays, research centers rely heavily on next generation

sequencers like Illumina or Thermo Fisher as their use produces very high coverage over targeted genomic regions, therefore allowing low-frequency variants to be accurately detected. In fact, the detection of low-frequency variants is a crucial step for cancer diagnosis. It is a very active area of research as it allows to personalize the treatment according to the found mutations. Unfortunately, low-frequency variants can be very easily confused with DNA polymerase errors produced during the amplification step as well as sequencing errors produced during the sequencing step. This has led to the rise of new sequencing protocols that rely on unique molecular identifiers (UMI) to correct the technical artifacts. The UMI implementation in such protocols was shown to be very effective in many published studies (Schmitt *et al.* (2012), Kukita *et al.* (2015), Newman *et al.* (2016), Young *et al.* (2016) and Bar *et al.* (2017)). UMIs are short arbitrary oligonucleotides sequences that are attached to the library of DNA fragments by ligation prior to the amplification step. The fact that UMIs are arbitrary sequences allows for every fragment to have a unique short oligonucleotide sequence attached to it, forming a unique tag for each fragment. These UMIs or unique tags are then amplified with their respective fragments and their sequences can be figured out from the reads through sequencing.

At the moment, three UMI-based variant callers are publicly available: DeepSNVMiner Andrews *et al.* (2016), MAGERI Shugay *et al.* (2017) and smCounter2 Xu *et al.* (2019). These tools all apply the same approach that tries to correct technical artifacts by performing a majority vote within a UMI family, since theoretically, reads that have the same UMI tag should be identical. By doing that, they build a consensus read for each UMI family and then they apply a statistical method (like Beta distribution) to model background error rates at each position and apply standard filters to call final variants. In order to call variants, raw-reads-based variant callers and UMI-based variant callers use SAMtools Li *et al.* (2009) to perform the pileup step. The pileup step generates a count of insertions, deletions and substitutions at each covered position in the BAM/SAM file. The advantage of SAMtools' pileup is that it is very efficient in terms of execution time and memory consumption. This allows for raw-reads-based variant callers to be relatively fast when compared to UMI-based variant callers. On the other hand, SAMtools does not take UMI tags into account so using it in a UMI-based variant caller significantly increases execution time. Only MAGERI does not use SAMtools pileup in its pipeline. By doing that, the tool is significantly slower and more memory consuming than all the other approaches, therefore justifying why other variant callers use it.

In this article, we present UMI-VarCal, a somatic single nucleotide variant and indel caller for UMI-based targeted paired-end sequencing protocols. UMI-VarCal stands out from the crowd by being one of the few variant callers that don't rely on SAMtools to do their pileup. Instead, thanks to an innovative homemade pileup algorithm specifically designed to treat the UMI tags present in the reads, UMI-VarCal is faster than both raw-reads-based and UMI-based variant callers. To test our tool, we compare it against two of the best raw-reads-based variant callers that only need the tumor sample to call variants, SiNVICT Kockan *et al.* (2017) and outLyzer Muller *et al.* (2016) and specifically designed to detect low-frequency variants. We also demonstrate that it can be as - if not more - sensitive as other UMI-based variant callers by comparing it against DeepSNVMiner.

2 Materials and methods

2.1 Samples

The Centre Henri Becquerel in Rouen designed a targeted sequencing panel for Diffuse Large B cell Lymphoma (DLBCL) analysis. This panel is designed to identify genomic abnormalities within a list of 36 genes that are most commonly impacted in this type of lymphoma. The panel was made specifically for QIAseq chemistry in order to introduce UMI during the construction of the library. For the list of genes used in the panel and the number of targeted regions per gene, the reader can refer to the supplementary table S1. In order to test UMI-VarCal against the three variant callers DeepSNVMiner, SiNVICT and outLyzer, we randomly selected 3 samples from a very large number of patients whose DNA were sequenced at the Centre Henri Becquerel and all suffering from DLBCL. Sample 1 and sample 3 are frozen biopsies extracted from 2 different patients at the Centre Henri Becquerel while sample 2 is a DNA extracted from a cell line. The selected samples have a corresponding histopathologic review and the quality of the DNA was checked to be adequate for sequencing.

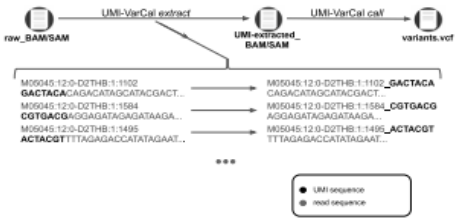


Fig. 1. Software input: UMI-VarCal can handle raw and UMI-extracted BAM or SAM files. If raw files are provided, the dedicated UMI extraction tool must be run prior to the calling tool. UMI tags are extracted from the read sequence and added to the end of the read ID. If BAM files are provided, they will be converted into SAM format. The variant calling tool can only start when the UMI-extracted SAM file is ready.

2.2 Software input

In order to run UMI-VarCal, three files are required: the paired-end SAM/BAM aligned file, the BED file containing the coordinates of the targeted genomic regions and a reference genome FASTA file with BWA index files. For ease of use, UMI-VarCal can accept BAM files as well as SAM files as input. We developed UMI-VarCal as a standalone variant caller that doesn't require any external tools to run. However, if the input is given under BAM format, SAMtools will be called in order to convert the BAM file into SAM format. Also, we integrated a UMI extraction tool in the software meaning that it can handle raw BAM/SAM files as well as UMI-extracted alignment files (Figure 1). Our tool can accept a fourth file under the PILEUP format. This file is only optional. In fact, when running UMI-VarCal on a sample, a PILEUP file is automatically produced. Giving this file to the software at execution will allow it to skip the pileup generation step (refer to section 2.3.1 for details) and load the old pileup instead. Being the step that takes most of the execution time, skipping it and loading the PILEUP file will make the user gain significant time.

2.3 Workflow

2.3.1 Pileup

The first step of the workflow is to generate the pileup. A pileup consists of the total count of match, substitution, insertion and deletion events at each position covered by the BED file. In fact, after filtering all the reads with low quality values, UMI-VarCal loops through every pair of reads and counts how many times each event was observed. Since each read is associated with a UMI tag, it means that every observed event

at each position can be tagged with the corresponding UMI. After going through all the reads, this step will generate the complete list of A, C, G, T, insertion and deletion counts as well as their corresponding UMI tags at each position and for each chromosome. Also, this is when our algorithm estimates background error noise for each position. After completing the pileup, UMI-VarCal will automatically generate a PILEUP file with all the necessary informations. This file can be used if the user wishes to launch a new analysis of the same sample but with different variant calling parameters since changing these don't affect the pileup but only the variants called. In fact, this allows the analysis to complete faster since loading the pileup is very much faster that regenerating it.

2.3.2 Estimating the background error rate

In order to distinguish between real variants and technical artifacts (DNA polymerase and sequencing errors), the background error rate must be estimated. We already know that the background error rate is not constant and can vary at different positions so we can assume that each position has a specific error rate. In order to estimate site-specific error rates, some variant callers require a matched normal sample along with the tumor sample to make the analysis, while others use many control samples to model the error noise and provide the variant calling tool with the built-in model. While the first approach is definitely the best to estimate the error rate, matched normal samples are very hard to get, especially for cell-free DNA samples. Estimating the model on a number of control samples is a good approach that is capable of filtering many technical artifacts but has the limitation of being specific to the panel sequencing protocol and therefore, cannot be used across different panels. That's why UMI-VarCal uses base quality scores at each position to estimate the corresponding base error probabilities.

Since each sequenced base is associated with a quality score, we can use it to determine the base error probability for each position by calculating the average mean quality score. Assuming that X_i represents the total number of reads n covering the position i as $X_i = \{x_i^1, x_i^2, ..., x_i^n\}$, and that Q_i represents the quality scores of the bases at the position i for each read as $Q_i = \{q_i^1, q_i^2, ..., q_i^n\}$, we can easily calculate the average quality score of the position i by

$$\overline{q}_i = \frac{\sum_{j=1}^{n} q_i^j}{n} \tag{1}$$

In fact, a quality score is a prediction of the probability of an error in base calling so the $\overline{q}_i$ that we calculated above reflects the sequencing quality or the base error probability at the position. Using the mean average qscore, we can compute the base error probability ϵ_i at each position i by

$$\epsilon_i = 10^{\frac{-\overline{q}_i}{10}} \tag{2}$$

2.3.3 Searching for candidate positions

The generated pileup contains the counts of A, C, G, T, insertions and deletions for all the positions covered in the BED file. UMI-VarCal loops through all these positions and applies at each one a Poisson test to determine if the alternative allele can be distinguished from the background error. We supposed that the presence of a variant is a rare event that could be treated as a hypothesis testing problem, where the null hypothesis (H_0) is that the alternative allele (substitutions, deletions and insertions) cannot be separated from background errors and the alternative hypothesis (H_1) is that the alternative allele can be distinguished from background errors and could actually represnt a true variant. At a position i, we define d_i as the depth at position i, ϵ_i as its base error probability and k_i as the total number of the alternative allele observations at position i. Under (H_0), k_i follows a Poisson distribution (λ_i) where λ_i is the number of errors expected to be found at a position i. We can simply calculate λ_i by

$$\lambda_i = d_i.\epsilon_i \tag{3}$$

At a position i, we can then calculate the p-value that represents the probability of observing more than λ_i errors as follows

$$p(k_i; \lambda_i) = 1 - \sum_{j=0}^{k_i} \frac{e^{-\lambda_i}.\lambda_i^j}{j!} \tag{4}$$

When we are conducting multiple hypothesis tests, we have an increased probability of false positives meaning that if we perform the same test multiple times, the chances of calling a null result as significant become higher. The false positive rate (FPR) refers to the number of false positives we expect when we perform a hypothesis test. So if we set the type 1 error probability (alpha) at 0.05, we can ensure that at worse, the percentage of false positives in all the tests we performed will be at 5%.. For example, if we test 10 000 positions and control the FPR at 0.05 (5%), on average 500 false variants (10000×0.05) will be called significant. This method poses a problem when we are conducting multiple tests as it becomes too permissive and we do not want to have such a great number of false positives. Typically, multiple comparison procedures control the false discovery rate (FDR) by trying to identify the most significant features and trying to filter out as much false positives as possible at the same time. UMI-VarCal applies the Benjamini-Hochberg procedure Benjamini and Hochberg (1995) in order to decrease the FDR, thus significantly reducing the number of total false positives. After applying the FDR correction to the p-values, we obtain the corresponding q-values. If the q-value is $\geq \alpha$, we can accept the null hypothesis and filter out the position as it means that the alternative allele observed at this position is most probably a technical artifact. Even with the FDR correction, the Poisson modeling applied to this situation maintains a relatively high sensitivity leaving us with a non negligible number of false positives. This is mainly due to the fact that the test does not take into consideration the strand bias nor the surrounding context of the variant. Therefore, in order to reduce the number of false positives, we apply three post-processing procedures as described below.

2.3.4 UMI analysis

When we apply the Poisson test to each position covered by the BED file, three scenarios are possible:

1. no alternative allele is found at the position: the position is filtered out
2. the q-value of the test is $\geq \alpha$ which means that the alternative allele is probably a technical artifact and therefore, the position is filtered out
3. the q-value of the test is $< \alpha$ which means that the alternative allele is most probably a true variant

In this last case only, a UMI analysis is applied. This step consists mainly of separating the list of all unique UMI tags found at a position into three different lists:

1. ref_umi: a list of all unique UMI tags found on the reads with the reference allele
2. alt_umi: a list of all unique UMI tags found on the reads with the alternative allele
3. $noise\_umi$: a list of all unique UMI tags found on the reads with neither the reference nor the alternative allele

Theoretically, if a variant is a true variant, it had to be found on the initial DNA fragment. So when we tag the DNA fragment with a unique UMI, we are also tagging the variant with the same unique UMI. After amplification, the DNA fragment is amplified and will produce thousands of reads, all

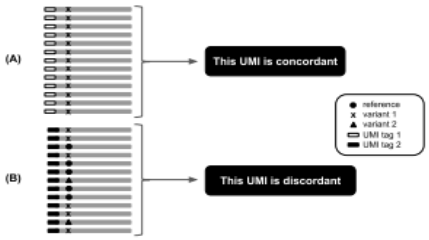


Fig. 2. The difference between a discordant UMI and a concordant UMI. (A) All the reads with the same green UMI tag present variant A: the green UMI tag is concordant. (B) The black UMI tag is found on 13 reads. Of the 13 reads, 6 present variant A, 5 present the reference allele and 2 present variant B. Since not all the reads have the same variant, we conclude that the black UMI tag is discordant.

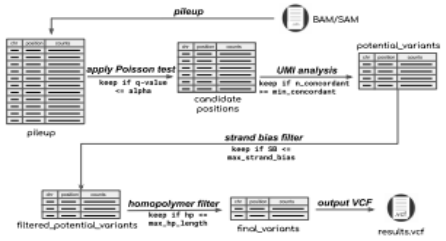


Fig. 3. UMI-VarCal workflow: UMI-Varcal starts by building a pileup from the BAM/SAM file provided. The pileup consists on A,C,G,T, insertion and deletion counts for each postilion covered by the alignment file. Since we know that not all positions contains mutations, UMI-VarCal starts looping through the pileup and applying a Poisson test at each position to filter out positions that don't seem to contain variants. To each of these candidate positions, an UMI analysis is carried out. At this step, one of 2 conditions are required in order to keep the variant. If the UMI analysis is successful, the potential variant must go through a strand bias filter to make sure that it isn't strand biased. If it passes that test, a final homopolymer region length filter is applied to make sure that the alternative allele is not due to the variant's presence in a long homopolymer region. If the alternative allele passes all these filters, it will be called and present in the final VCF file.

carrying the same UMI as well as the same alternative allele. This means that at a specific position, if the alternative allele represents a true variant, all the reads that have the same UMI tag must present the same alternative allele. If that's the case, the UMI is called concordant (Figure 2A). On the other hand, if some of the reads with the same UMI tag present the reference or a noise allele, the UMI is called discordant (Figure 2B). Each concordant UMI tag characterizes a single DNA fragment. Using the three lists ref_umi, alt_umi and $noise\_umi$, we can calculate the number of concordant and discordant UMI tags for each variant. The more concordant UMI tags a variant has, the more DNA fragments it was present on initially. UMI-VarCal uses a concordant UMI tags threshold in order to filter out variants with too little concordant UMI counts. This UMI-based filter guarantees that the variants that pass through are not technical artifacts. These variants are called potential variants as they haven't passed through all the post-processing steps yet.

2.3.5 Strand bias filtering

This is the second filter and it is only applicable for potential variants (variants that have passed the Poisson test and the UMI analysis process). It was proven by Guo *et al.* (2012a) that a high strand bias (SB) could point out to a potential high false-positive rate, especially in Illumina short-read sequencing data. In this step, our strand bias filter calculates the strand bias score for each potential variant and, with the use of a threshold, aims to filter out all strand biased variants. Guo et al., 2012 compared three different methods to calculate the strand bias score (the traditional SB score, the GATK-SB score and the SB Fisher score) and demonstrated that the traditional SB calculation and the Fisher score can capture false positives better than the GATK-SB method. In addition, the traditional method used by Guo *et al.* (2012b) to detect false positives in variants from mitochondrial DNA samples showed very good results with a threshold of 1.0. UMI-VarCal uses the traditional SB calculation method (Equation 5) and applies the threshold of 1.0 in order to filter out the highest number of false positives among potential variants without being restrictive. We define R_f and R_r as the forward and reverse strands allele counts of the major allele, and V_f and V_r as the forward and reverse strands allele counts of the alternative allele.

$$SB = \frac{\left| \frac{V_f}{R_f+V_f} - \frac{V_r}{R_r+V_r} \right|}{\frac{V_f+V_r}{R_f+R_r+V_f+V_r}} \quad (5)$$

2.3.6 Filtering variants in homopolymer regions

Both pyrosequencing and ion semiconductor sequencing have difficulties to call correctly the bases situated in long homopolymer-containing regions. The uncertainty is due to the fact that the repeating identical nucleotides have to be incorporated during the same synthesis cycle. Ivády *et al.* (2018) demonstrated that the base calling accuracy suffers greatly as the length of the homopolymer region increases (> identical 4 bases). SomaticSniper Larson *et al.* (2012) is a variant calling tool that applies a homopolymer length filter in order to remove variants that occur in long homopolymer regions as this would mean that they are most probably artifacts due to sequencing errors. UMI-VarCal uses the same filter to remove variants found in a homopolymer region with a length > 7.

2.4 Implementation

The overall workflow (Figure 3) is comprised of Python modules that are called by a main Python script. All the modules are compiled in Cython to achieve better overall performance. UMI-VarCal is available for Python version 2 and 3. UMI-VarCal doesn't rely on any external program to launch. It only requires SAMtools if the input file is provided under BAM format.The extraction tool and the variant calling tool are executed through a UNIX command line interface. All the parameters and thresholds (minimum base quality, minimum read quality, minimum mapping quality, type 1 error rate alpha, minimum number of concordant UMI tags, maximum strand bias and maximum homopolymer region length) are customizable in order to allow the user total control over his results.

2.5 Software output

By default, UMI-VarCal automatically produces three files:

1. A standard VCF file containing all the variants that passed the tests and were successfully reported. For each variant, allele frequency, alternative allele observation count, total read depth, homopolymer length, variant type and confidence are provided. A confidence level is provided for each variant and is computed based on the variant's strand bias, homopolymer region length, q-value and the

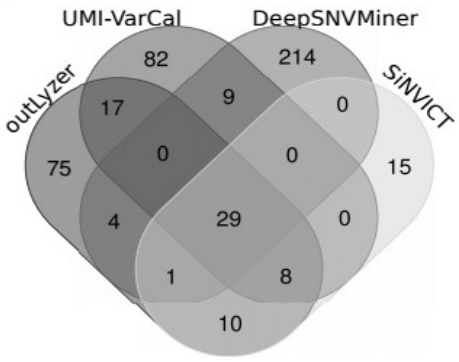


Fig. 4. Venn diagram of variants found by UMI-VarCal, DeepSNVMiner, SiNVICT and outLyzer in Sample 1

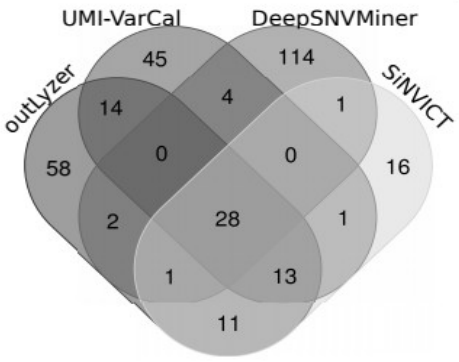


Fig. 5. Venn diagram of variants found by UMI-VarCal, DeepSNVMiner, SiNVICT and outLyzer in Sample 2

concordant/discordant ratio). Five levels are possible ranging from low to certain (low < average < high < strong < certain).

2. A VARIANTS file containing all the variants that were successfully reported. This file contains the same variants of the VCF file, in addition to detailed metrics for each variant.
3. A binary PILEUP file that corresponds to the entire pileup dumped. This file can be used to skip the pileup regeneration and load the pileup directly if the analysis was already done on the sample.

3 Results

At this moment, three UMI-based variant callers are publicly available: DeepSNVMiner, MAGERI and smCounter2. smCounter2 is relatively fast but has a theoretical detection limit of only 0.5%. MAGERI has a theoretical detection limit of 0.1% but is very slow and consumes a lot of memory (in our tests, it took 1 hour of execution time and a minimum of 200 GB of RAM to analyze one sample). Finally, DeepSNVMiner presents the advantage of having the same detection limit as MAGERI (0.1%) and is way more efficient in terms of execution time and memory consumption. To demonstrate our tool superiority over other available tools in terms of variant detection as well as performance, we compared it against DeepSNVMiner and also against two of the best raw-reads-based variant callers, outLyzer and SiNVICT that are designed specifically to detect low-frequency variants. In the following, we will compare the detection performance of the 4 variant callers on three different samples.

3.1 Variant detection comparison

3.1.1 Sample 1

In total, 464 variants were found (all the variants are detailed in Supplementary Table S2) (Figure 4). UMI-VarCal accounts for 145 variants, while DeepSNVMiner, outLyzer and SiNVICT detected 257, 144 and 63 variants respectively. Among these 145 variants, 29 are also found by all the other three tools and 63 were found by at least one other variant caller. 214 variants were only found by DeepSNVMiner: 139/214 didn't pass the Poisson test, 60/214 didn't pass the UMI analysis test, 4/214 are most probably strand biased and 1/214 is in a long homopolymer region. 75 variants were found only by outLyzer: 3/75 didn't pass the Poisson test, 64/75 didn't pass the UMI analysis test, 3/75 are most probably strand biased and 5/75 are in a long homopolymer region. 15 variants were found only by SiNVICT: 3/15 didn't pass the Poisson test, 7/15 didn't pass the UMI analysis test and 5/15 are detected in positions that are not covered by the provided BED file. 10 variants were found by both SiNVICT and outLyzer: all 10 variants are in a long homopolymer region. 4 variants were detected by both DeepSNVMiner and outLyzer: all four variants didn't pass the UMI analysis test and one of them is also most probably strand biased. 1 variant was found by DeepSNVMiner, SiNVICT and outLyzer: this variant is in a very long homopolymer region (length = 16). 82 variants were detected only by UMI-VarCal: 74/82 (90.2%) have a frequency below 1% and 28/82 (34.1%) have a frequency below 0.5%. UMI-VarCal detected 8 variants at a frequency < 0.4% but no variant was detected under the 0.3% frequency. Also, only 1/82 (1.2%) had a low level of confidence while 73/82 (89%) had at least a high confidence level.

3.1.2 Sample 2

In total, 308 variants were found (all the variants are detailed in Supplementary Table S3) (Figure 5). UMI-VarCal accounts for 105 variants, while DeepSNVMiner, outLyzer and SiNVICT detected 150, 127 and 71 variants respectively. Among these 105 variants, 28 are also found by all the other three tools and 60 were found by at least one other variant caller. 114 variants were only found by DeepSNVMiner: 63/114 didn't pass the Poisson test, 48/114 didn't pass the UMI analysis test and 3/114 are most probably strand biased. 58 variants were found only by outLyzer: 5/58 didn't pass the Poisson test, 46/58 didn't pass the UMI analysis test, 2/58 are most probably strand biased and 5/58 are in a long homopolymer region. 16 variants were found only by SiNVICT: 2/16 didn't pass the Poisson test, 7/16 didn't pass the UMI analysis test, 1/16 is in a long homopolymer region and 6/16 are detected in positions that are not covered by the provided BED file. 11 variants were found by both SiNVICT and outLyzer: 10/11 variants are in a long homopolymer region and one have 0 concordant UMI tags and therefore didn't pass the UMI analysis test. 2 variants were detected by both DeepSNVMiner and outLyzer: both of them didn't pass the UMI analysis test and one of them is also most probably strand biased. 1 variant was found by DeepSNVMiner, SiNVICT and outLyzer: this variant is in a long homopolymer region (length = 8) and has 0 concordant UMI tags. 45 variants were detected only by UMI-VarCal: 39/45 (86.7%) have a frequency below 1% and 13/45 (28.9%) have a frequency below 0.5%. UMI-VarCal detected 3 variants at a frequency

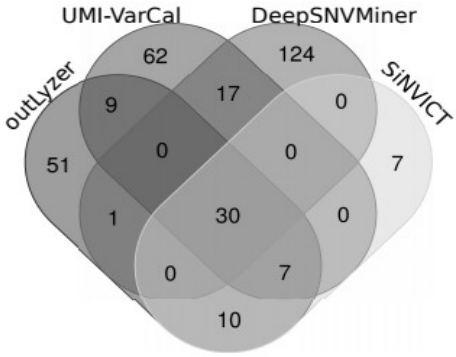


Fig. 6. Venn diagram of variants found by UMI-VarCal, DeepSNVMiner, SiNVICT and outLyzer in Sample 3

< 0.4% but no variant was detected under the 0.3% frequency. Also, none of the 45 variants had a low level of confidence while 37/45 (82.2%) had at least a high confidence level.

3.1.3 Sample 3

In total, 318 variants were found (all the variants are detailed in Supplementary Table S4) (Figure 6). UMI-VarCal accounts for 125 variants, while DeepSNVMiner, outLyzer and SiNVICT detected 172, 108 and 54 variants respectively. Among these 145 variants, 30 are also found by all the other three tools and 83 were found by at least one other variant caller. 124 variants were only found by DeepSNVMiner: 88/124 didn't pass the Poisson test and 36/124 didn't pass the UMI analysis test. 51 variants were found only by outLyzer: 45/51 didn't pass the UMI analysis test, 1/51 are most probably strand biased and 5/51 are in a long homopolymer region. 7 variants were found only by SiNVICT: 3/7 didn't pass the UMI analysis test, 1/7 is in a long homopolymer region and 4/7 are detected in positions that are not covered by the provided BED file. 10 variants were found by both SiNVICT and outLyzer: 9/10 variants are in a long homopolymer region and one is most probably strand biased. 1 variant was detected by both DeepSNVMiner and outLyzer: this variant didn't pass the UMI analysis test. 62 variants were detected only by UMI-VarCal: 51/62 (82.3%) have a frequency below 1% and 10/62 (16.1%) have a frequency below 0.5%. UMI-VarCal detected 2 variants at a frequency < 0.4% but no variant was detected under the 0.3% frequency. Also, none of the 62 variants had a low level of confidence while 55/62 (88.7%) had at least a high confidence level.

3.2 Performance comparison

In order to compare the performance of UMI-VarCal with the other three variant callers, we artificially created 5 different samples with increasing size (1, 2, 3, 5 and 10 million reads). This will allow to compare not only the performance of the tools but also to have a look at how the performance varies with sample size. All these tests are performed on a one core CPU running at 2.20 GHz. All measurements were done 3 times and the average was used for the comparison (Figure 7). To analyze 1 million reads, UMI-VarCal is the fastest with 62 seconds to complete the analysis. It is followed by SiNVICT that takes 138 seconds and outLyzer with 228 seconds. The slowest tool is DeepSNVMiner as it takes 1107 seconds to complete its

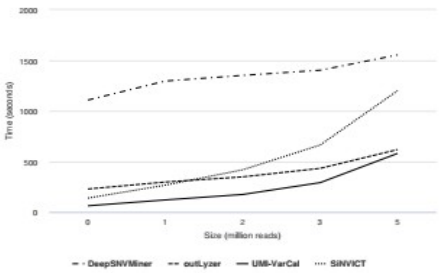


Fig. 7. Performance comparison between UMI-VarCal, DeepSNVMiner, SiNVICT and outLyzer

analysis. For the 2 million reads analysis, the ranks don't change as UMI-VarCal is still the fastest of the four tools and DeepSNVMiner the slowest. The analysis of 3 million reads is the fastest on UMI-VarCal and the slowest on DeepSNVMiner as well. However, outLyzer sets the better time versus SiNVICT as the latter's performance seeming to suffer when increasing sample size. The ranks don't change at the 5 million reads mark as UMI-VarCal sill outperforms the three other variant callers and DeepSNVMiner being the slowest. Finally, the analysis of the 10-million-read sample is the fastest again on UMI-VarCal taking only 580 seconds to complete. outLyzer is closely behind and completes the analysis 38 seconds after (618 seconds). SiNVICT maintains the third place as it takes 1200 seconds to finish its analysis. DeepSNVMiner is still last with it taking 1553 seconds to have the final results ready. We note that both UMI-VarCal and outLyzer can be executed on multiple cores which can significantly decrease running time, especially on very large samples. In terms of memory consumption, UMI-VarCal is not very demanding. Memory consumption is not only impacted by the number of the reads but also by other factors such as the amplification factor of the sample and the maximum sequencing depth. Therefore, measuring the variation of memory consumtion with sample size is not significant. In our tests on the 3 DNA samples we selected, UMI-VarCal needed approximately 3 GB of RAM.

4 Discussion

Detecting somatic mutations with low allelic frequency is a challenge but is primordial in cancer studies in order to characterize tumor heterogeneity. Many raw-reads-based variant callers are available and do a good job in detecting most variants within a sample. Some of them however need a matched normal sample in order to perform the analysis: this can be problematic as these samples are difficult to find and might not exist in some applications. Other tools like SiNVICT and outLyzer do an outstanding job actually at detecting variants with frequencies as low as 0.5% but at the cost of having a high number of false positives: it is expected as these tools don't integrate a UMI analysis and thus cannot efficiently filter out false positives. UMI-based variant callers don't have this problem since they perform a UMI analysis that allows them to filter out most false positives. MAGERI showed some very good results in the publication with a theoretical detection limit of 0.1% but suffers in terms of performance as it consumes a lot of memory and is very slow. Another UMI-based variant caller is smCounter2 that has good performance but a detection limit of only 0.5%. DeepSNVMiner is a UMI-based variant caller that presents a theoretical detection limit of 0.1% and is relatively fast, compared to other

UMI-based variant callers. It starts by generating an initial list of variants using SAMtools calmd and then selects only those that have strong UMI support. However, in our tests, it seems to generate a lot of false positives since it doesn't contain a strand bias filter nor a homopolymer region length filter.

UMI-VarCal was able to perform better than the 3 other variant callers. It could easily detect the true variants found by the others and filter out the false positives due to its multi-step post-processing filters (UMI analysis filter, strand bias filter and homopolymer length filter). In addition, it was able to detect a high number of low-frequency variants (AF $\leq$ 1%) not found by other tools, of which 85% (on average) have at least a high level of confidence. In terms of execution time, we must admit that it is somewhat unfair to compare a UMI-based variant caller such as DeepSNVMiner to raw-reads-based variant callers. In our comparison, we showed not only that our tool can easily outperform an UMI-based variant caller but can only beat one of the fastest raw-reads based variant callers, outLyzer.

5 Conclusion

Here, we present UMI-VarCal: a standalone UMI-based variant caller developed to achieve more accurate low-frequency variant detection in paired-end sequencing NGS libraries. Also, thanks to a new pileup algorithm specifically designed to integrate the UMI tags in the reads, it is able to achieve excellent performance, in terms of both execution time and memory consumption, making it one of the fastest - if not the fastest - variant callers out there. In addition of its outstanding performance, UMI-VarCal is capable of detecting a large number of variants with frequencies as low as 0.3% that were completely missed out by the other tools. Among these variants, approximately 85% have at least a high confidence level meaning that they are most likely true variants. UMI-VarCal was built to allow total control for the user over his analysis since all the filters' parameters are customizable. This makes this tool adequate and available to a large number of clinical and research applications.

Funding

This work was partly funded by the University of Rouen Normandie and Vincent Sater is funded by a PhD fellowship from the Région Normandie.

References

Andrews, T. D., Jeelall, Y., Talaulikar, D., Goodnow, C. C., and Field, M. A. (2016). DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. *PeerJ*, **4**.

Bar, D. Z., Arlt, M. F., Brazier, J. F., Norris, W. E., Campbell, S. E., Chines, P., Larrieu, D., Jackson, S. P., Collins, F. S., Glover, T. W., and Gordon, L. B. (2017). A novel somatic mutation achieves partial rescue in a child with Hutchinson-Gilford progeria syndrome. *J Med Genet*, **54**(3), 212–216.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**(1), 289–300.

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D. C., and Shyr, Y. (2012a). The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.

Guo, Y., Cai, Q., Samuels, D. C., Ye, F., Long, J., Li, C.-I., Winther, J. F., Tawn, E. J., Stovall, M., Lähteenmäki, P., Malia, N., Levy, S., Shaffer, C., Shyr, Y., Shu, X.-o., and Boice, J. D. (2012b). The use of Next Generation Sequencing Technology to Study the Effect of Radiation Therapy on Mitochondrial DNA Mutation. *Mutat Res*, **744**(2), 154–160.

Ivády, G., Madar, L., Dzsudzsák, E., Koczok, K., Kappelmayer, J., Krulisova, V., Macek, M., Horváth, A., and Balogh, I. (2018). Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *BMC Genomics*, **19**.

Kockan, C., Hach, F., Sarrafi, I., Bell, R. H., McConeghy, B., Beja, K., Haegert, A., Wyatt, A. W., Volik, S. V., Chi, K. N., Collins, C. C., and Sahinalp, S. C. (2017). SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics*, **33**(1), 26–34.

Kukita, Y., Matoba, R., Uchida, J., Hamakawa, T., Doki, Y., Imamura, F., and Kato, K. (2015). High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients. *DNA Res*, **22**(4), 269–277.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**(3), 311–317.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Muller, E., Goardon, N., Brault, B., Rousselin, A., Paimparay, G., Legros, A., Fouillet, R., Bruet, O., Tranchant, A., Domin, F., San, C., Quesnelle, C., Frebourg, T., Ricou, A., Krieger, S., Vaur, D., and Castera, L. (2016). OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. *Oncotarget*, **7**(48), 79485–79493.

Newman, A. M., Lovejoy, A. F., Klass, D. M., Kurtz, D. M., Chabon, J. J., Scherer, F., Stehr, H., Liu, C. L., Bratman, S. V., Say, C., Zhou, L., Carter, J. N., West, R. B., Sledge, G. W., Shrager, J. B., Loo, B. W., Neal, J. W., Wakelee, H. A., Diehn, M., and Alizadeh, A. A. (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol*, **34**(5), 547–555.

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., and Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A*, **109**(36), 14508–14513.

Shugay, M., Zaretsky, A. R., Shagin, D. A., Shagina, I. A., Volchenkov, I. A., Shelenkov, A. A., Lebedin, M. Y., Bagaev, D. V., Lukyanov, S., and Chudakov, D. M. (2017). MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS Comput Biol*, **13**(5).

Xu, C., Gu, X., Padmanabhan, R., Wu, Z., Peng, Q., DiCarlo, J., and Wang, Y. (2019). smCounter2: an accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers. *Bioinformatics*, **35**(8), 1299–1309.

Young, A. L., Challen, G. A., Birmann, B. M., and Druley, T. E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat Commun*, **7**.

C. Détection des CNV avec UMI : algorithme mCNA

C.1. Introduction

Il existe différentes approches pour identifier les CNV à partir de données de séquençage NGS : celles utilisant l'information portée par les paires de lectures (read-pair, RP), celles utilisant l'information sur les lectures scindées (split-read, SR) ou celles enfin se basant sur la profondeur de séquençage (read-depth, RD).

Les approches RP telles que BreakDancer [282], PEMer [283] ou Ulysses [284] utilisent l'information portée par les paires de lectures afin d'identifier des gains ou des pertes dans une région donnée. En effet, un certain nombre d'informations peuvent être extraites :

- l'orientation des lectures : si une lecture s'aligne dans un sens du génome de référence, alors en l'absence d'anomalie son homologue doit s'aligner dans le sens opposé
- la taille de l'insert entre les couples de lectures pairées est supposée suivre une distribution uni-modale dans une librairie de séquençage

Les algorithmes cherchent donc à extraire des anomalies au sein de ces couples de lectures afin d'identifier des gains de matériel (insertion) ou des pertes de matériel (délétion).

Les approches SR telles que SVseq2 [285], Gustaf [286] ou PRISM [287] utilisent elles aussi l'information portée par les couples de lecture mais cette fois-ci en cherchant à identifier les problèmes d'alignement au sein de chaque couple. Elles vont chercher à identifier, par exemple, les couples de séquences pour lesquels l'une des séquences s'alignent parfaitement sur le génome de référence tandis que l'autre ne s'alignent pas, ou sur un chromosome différent de son homologue. Ces problèmes d'alignement au sein de chaque couple peuvent donner potentiellement le point de cassure d'une insertion ou d'une délétion. Néanmoins, ces algorithmes ne sont pas adaptés à la détection de larges remaniements.

Enfin, les approches RD consistent à compter le nombre de lectures s'alignant dans une fenêtre glissante définie. Ces données de comptage sont alors normalisées de sorte à rendre possible la comparaison entre un échantillon à tester et un autre échantillon servant de référence. Une perte locale de profondeur entre échantillon et témoin sera corrélée à une perte de matériel et donc une possible délétion, tandis qu'à l'inverse une augmentation locale de la profondeur de séquençage sera corrélée à un gain de matériel. Cette stratégie est très largement répandue notamment dans le contexte de séquençage de panel de gènes au diagnostic pour lesquels une liste de régions d'intérêt est séquencée profondément. Certains

outils, comme ONCOCNV 28 , ont été spécialement développés pour l'analyse de panels de séquençage ciblés. Néanmoins, ces approches RD nécessitent d'appliquer des stratégies de normalisation souvent complexes du fait de l'amplification des librairies. L'amplification induit un biais important dans le comptage des lectures alignées et empêche une quantification directe du nombre de molécules d'ADN présentes avant amplification pour chacune des régions ciblées.

L'introduction des UMI dans la construction des librairies peut permettre de développer une nouvelle approche bioinformatique pour la quantification des séquences uniques par région. Il s'agit non plus de quantifier le nombre de lectures s'alignant sur les régions d'intérêt mais directement le nombre d'UMI. Ainsi, les données de comptage pour déterminer la présence ou non de CNV ne sont plus dépendantes de l'efficacité d'amplification propre à chaque région ciblée.

Cette approche est particulièrement intéressante pour l'analyse d'échantillons de cfDNA. En effet, les fragments d'ADN libérés dans la circulation sanguine par les cellules tumorales sont souvent fragmentés entraînant ainsi des séquences alignées plus courtes. Ce biais empêche l'utilisation des approches RD pour la détection de CNV car il entraîne des pertes de profondeur locales qui ne sont pas liées à une variation de copies mais qui sont uniquement le fait de fragments courts alignés. Le fait de compter le nombre d'UMI par position permet de rendre les données de comptage indépendantes de ce facteur de taille.

C.2. Implémentation

mCNA (molecular Copy Number Alteration) est une nouvelle approche bioinformatique permettant la détection de CNV via l'information portée par les UMI pour des panels ciblés de séquençage.

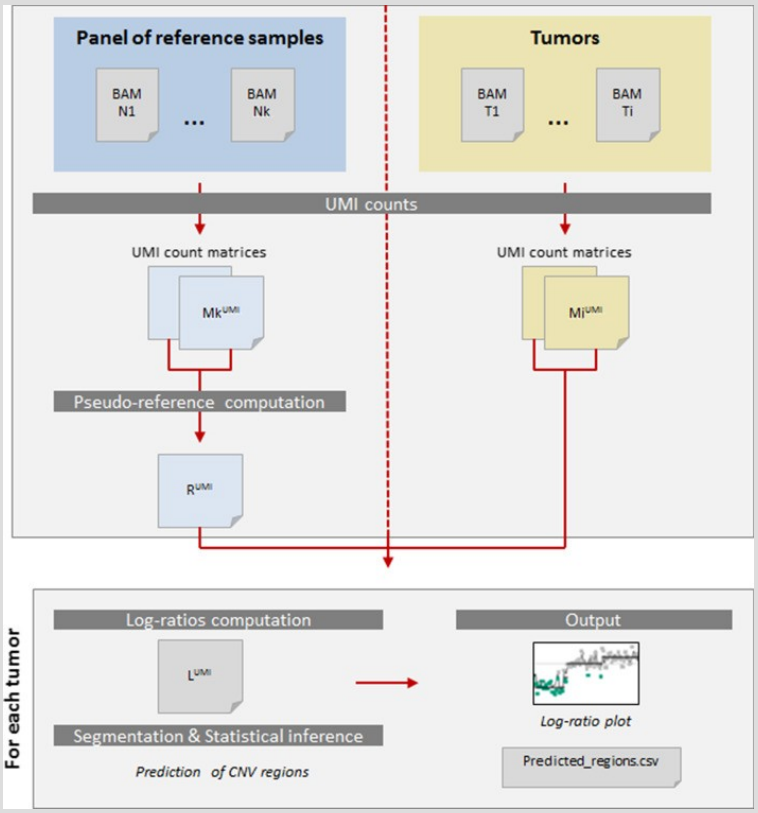


Figure 58: mCNA : fonctionnement de l'algorithme.

mCNA repose sur quatre grandes étapes : la construction des matrices de comptage en UMI, la construction d'une pseudo-référence, l'estimation des log-ratios pour chaque région ciblée et finalement l'estimation des gains ou des pertes de copies après segmentation (figure 58).

L'algorithme procède au comptage du nombre d'UMI lus par région à partir de données de séquençage alignées afin d'établir, pour chaque échantillon séquencé, une matrice de comptage en UMI. Ces matrices de comptage sont divisées respectivement par le nombre d'UMI moyen lus de chaque échantillon de sorte à permettre leur comparaison. En effet, la quantité d'ADN utilisé lors de la préparation des librairies a un impact direct sur le nombre d'UMI moyen lus et il est donc nécessaire de gommer ce biais avant les étapes ultérieures.

A partir des matrices de comptages calculées sur des échantillons de référence, une référence moyennée appelée « pseudo-référence » est créée. Ce profil témoin est obtenu en appliquant, pour chaque région ciblée, la moyenne géométrique du nombre d'UMI normalisés de chacun des échantillons contrôle. Dès cette étape, un contrôle qualité de la pseudo-référence est réalisé en calculant l'écart entre les données de comptage de chacun des échantillons de référence et la pseudo-référence. L'objectif est d'établir si la pseudo-référence est bien le reflet des échantillons contrôles et si des variations de comptage sont déjà présentes

dans des échantillons pourtant dépourvus d'anomalie. Le cas échéant, les échantillons ayant des profils discordants sont exclus tout comme les régions ayant des variabilités de comptage trop importantes. L'idée sous-jacente est de ne pas considérer des régions déjà anormalement bruitées dans les échantillons contrôles lors de l'analyse des profils tumoraux.

Finalement, les matrices de comptages des échantillons tumoraux sont comparées à la pseudo-référence via le calcul de log-ratios pour chaque région. Les profils ainsi obtenus sont segmentés par l'algorithme PSCBS [289], une implémentation de l'algorithme *Circulary Binary Segmentation* sous R, puis un test de comparaison de moyenne est réalisé afin de déterminer si la la moyenne observée des log-ratios de chaque segment est significativement différent de la valeur référence 0.

Le détail complet de toutes les étapes de l'algorithme mCNA est présent en fin de chapitre dans l'article publié dans BMC Bioinformatics.

C.3. Validation biologique

Robustesse de la quantification

Nous nous sommes tout d'abord intéressés à la qualité de l'acquisition des données de comptage en amont de l'algorithme. Nous avons démontré à partir d'échantillons contrôles que la variance observée dans les comptages UMI était significativement inférieure aux comptages utilisant le nombre de lectures (figure 59). En d'autres termes, il est plus intéressant de quantifier le nombre d'UMI par région que le nombre de lectures si l'on souhaite détecter une baisse ou une augmentation de matériel dans des échantillons.

Afin de vérifier la reproductibilité des données générées après normalisation des comptages, nous avons corrélé les données obtenues par mCNA sur la lignée cellulaire REC-1. Deux librairies ont été construites indépendamment à partir du PanLymphome et séquencées sur deux runs différents. Les log-ratios calculés sur les réplicats sont très fortement corrélés (r=0.93) alors même que les profondeurs moyennes obtenues à partir des deux librairies étaient différentes.

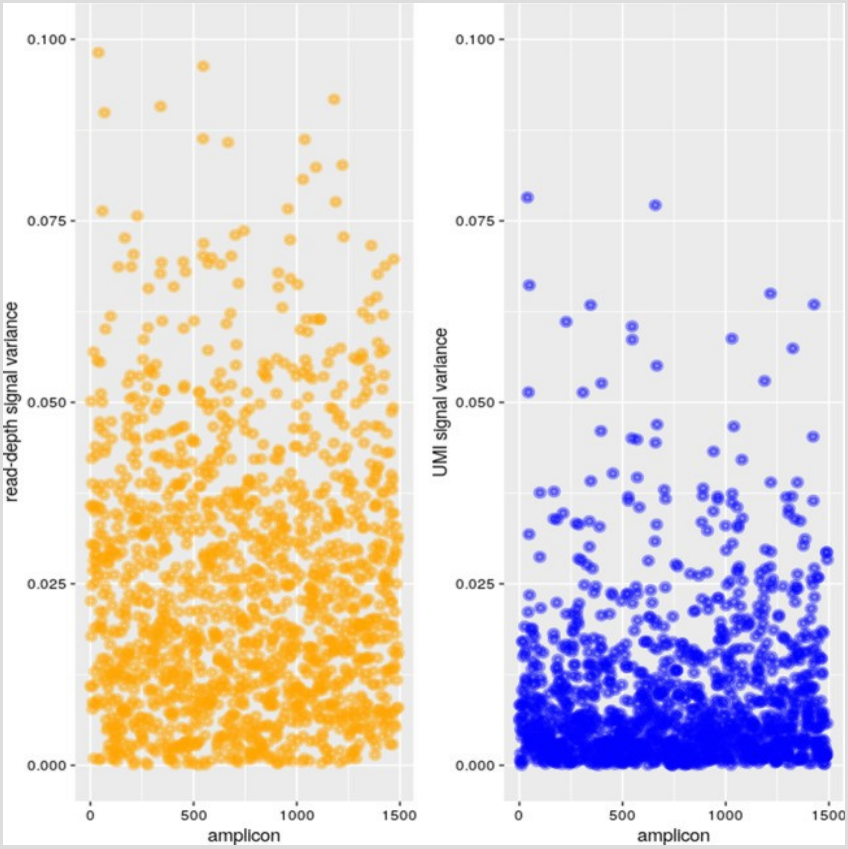


Figure 59: mCNA : variance des signaux en UMI et en nombre de lectures.

Le graphique représente la variance des comptages normalisés par amplicon basés sur le nombre de lectures alignées (à gauche) ou le nombre d'UMI (à droite).

Limite de sensibilité et spécificité

Afin d'évaluer la capacité de détection de l'approche, nous avons réalisé tout d'abord une évaluation *in silico* des performances de l'algorithme en modifiant les données de comptages en UMI dans la matrice de comptage d'un échantillon séquencé.

A partir d'un panel de séquençage Qiaseq appelé PanLymphome, couvrant 69 gènes d'intérêt (1493 amplicons), un échantillon témoin a tout d'abord été séquencé. A partir des données alignées, nous avons introduit artificiellement dans l'échantillon une amplification de *XPO1*, un gain de *IRF4*, une délétion hétérozygote de *CDKN2A* et une délétion homozygote de *CDKN2B*. Nous avons ensuite appliqué des dilutions de ces segments anormaux en faisant varier le pourcentage de cellules tumorales de notre échantillon *in silico* de sorte à évaluer les

capacités de détection de l'algorithme après les différentes phases de normalisation, de segmentation et le test de comparaison de moyenne.

Nous avons observé que les log-ratios attendus et mesurés à la fin du traitement bioinformatique des données brutes par mCNA sont très fortement corrélés sur ces données simulées (figure 60). mCNA est en mesure de détecter toutes les anomalies introduites pour des pourcentages de cellules tumorales comprises entre 10 et 100 %. A 5 %, seules les anomalies impliquant le gain ou la perte de plus d'une copie sont retrouvées. Cette observation nous indique donc une limite de sensibilité théorique comprise entre 5 et 10 %.

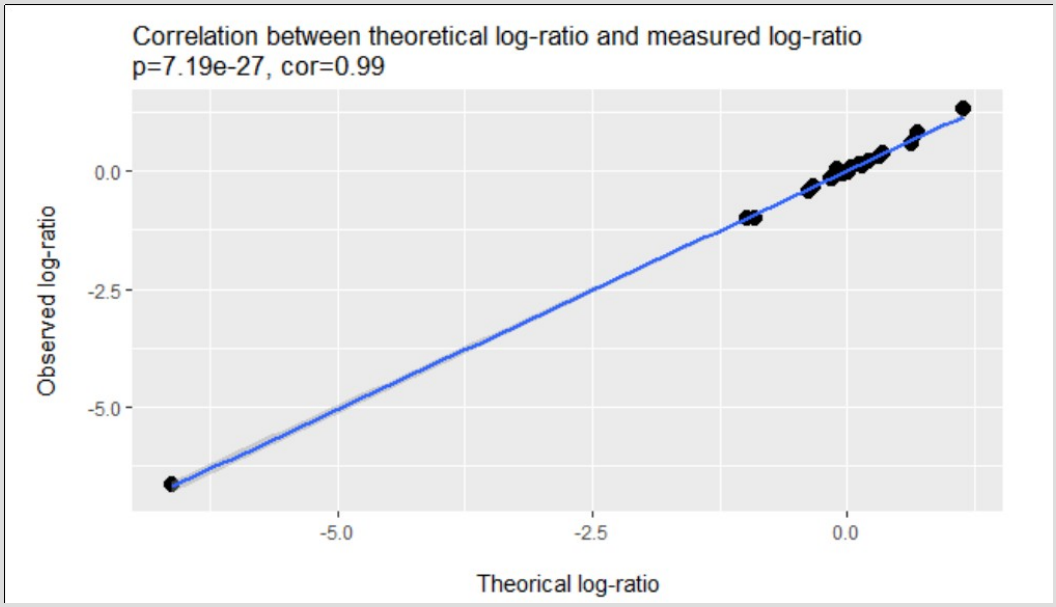


Figure 60: Corrélation entre les log-ratios attendus et calculés par mCNA sur le jeu de données simulées.

Chaque point du graphique correspond au log-ratio mesuré pour chaque segment anormal *in silico* introduit dans un échantillon de référence à des pourcentages de cellules tumorales variables (100 %, 50 %, 20 %, 10 % et 5%). Ces segments anormaux correspondent à une amplification du gène *XPO1*, un gain de *IRF4*, une délétion hétérozygote de *CDKN2A* et une délétion homozygote de *CDKN2B*.

Afin de valider ces résultats *in silico*, nous avons séquencé de façon indépendante la lignée REC-1 en duplicat afin d'établir un profil de CNV de référence pour cette lignée cellulaire. Nous trouvons tout d'abord une forte corrélation des log-ratios mesurés entre les deux réplicats ($r=0,99$, $p < 0,001$) alors que les profondeurs moyennes de séquençage ne sont pas identiques (1851X / 2217X). 30/31 segments prédits comme étant amplifiés ont été retrouvés

dans les deux réplicats tout comme 21/23 segments normaux et 17/18 segments délétés, donnant ainsi une concordance globale du profil de 94,17 %.

Cette lignée a ensuite été diluée dans de l'ADN témoin Promega dépourvu de variation de nombre de copies sur les régions séquencées. La gamme repose sur des dilutions à 50 %, 30 %, 20 %, 10 % et 5 % d'ADN de lignée REC-1. Ces différentes dilutions ont été séquencées indépendamment de sorte à déterminer le seuil à partir duquel mCNA n'est plus en mesure de retrouver les anomalies trouvées initialement dans la lignée. Les résultats, visibles sur la figure 61, montre un seuil de sensibilité limite comparable à celui déterminé à partir des échantillons *in silico* compris entre 5 et 10 % de cellules tumorales. En dessous de 10 % de contingent tumoral, seules les délétions homozygotes sont retrouvées et les gains de copies semblent plus difficiles à détecter.

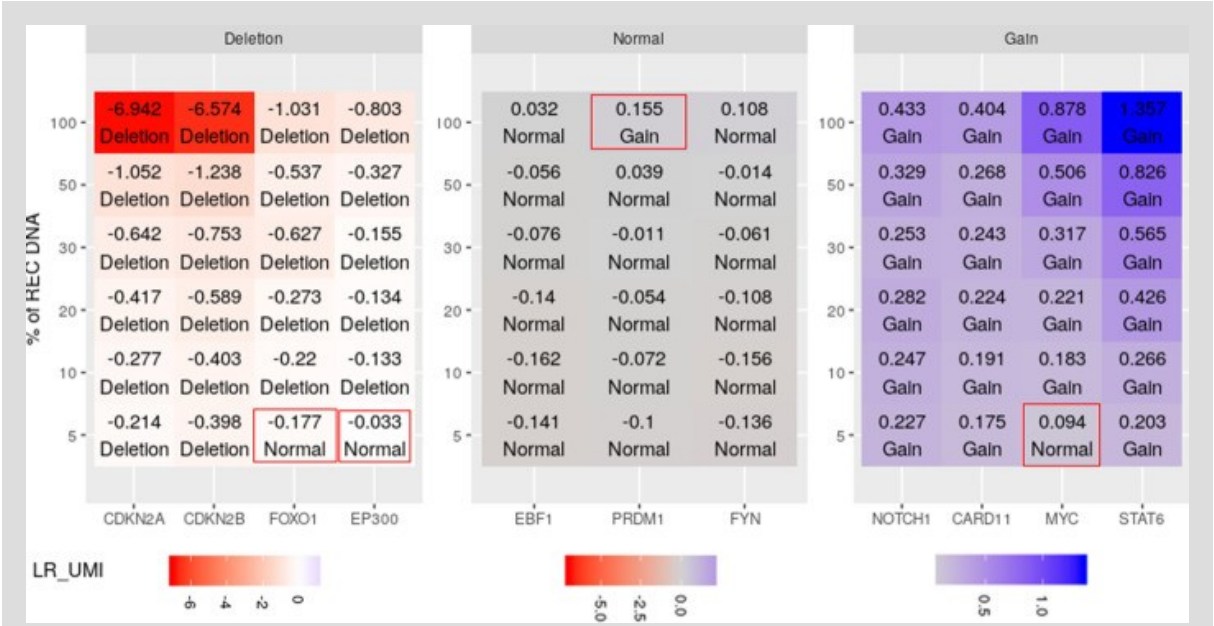


Figure 61: Résultats du traitement des dilutions de la lignée REC-1 par mCNA.

Le tableau indique la liste des anomalies de REC-1 en fonction du pourcentage de cellules tumorales. Chaque cellule rapporte le log-ratio estimé par mCNA sur un segment donné ainsi que la classe prédite (Délétion, Normal, Gain). Les discordances de classe par rapport aux résultats attendus sont contourées en rouge.

Interopérabilité entre séquenceurs

Afin d'évaluer la robustesse de la pseudo-référence calculée à partir d'échantillons contrôles, nous avons comparé les résultats obtenus pour un même set d'échantillons entre des librairies séquencées sur MiSeq et NextSeq. L'objectif est de déterminer si les comptages en

UMI varient entre les deux séquenceurs que ce soit pour les échantillons de référence ou les échantillons testés.

Un exemple de résultat de comparaison est donné en figure 62. On y retrouve les profils CNV d'un échantillon analysé sur un panel de gènes dédié à la détection des anomalies dans les LLC au diagnostic au Centre Henri Becquerel à Rouen. Ce panel cible tout ou partie des gènes *BIRC3*, *ATM*, *POU6F1*, *MDM2*, *DLEU2*, *PLCG2*, *TP53*, *BCL2*, *XPO1*, *CXCR4*, *SF3B1*, *CECR1*, *MYD88*, *FBXW7*, *SEC63*, *BRAF*, *NOTCH1* et *BTK*. Les résultats nous montrent des profils presque superposables pour un même échantillon quelque soit la machine ayant servi pour le séquençage de l'échantillon testé et/ou des échantillons contrôles.

Figure 62: Comparaison des résultats de mCNA sur un échantillon d'ADN extrait d'un patient atteint de LLC en fonction de la nature du séquenceur utilisé.

Les résultats montrent des profils superposables quelque soit la nature de la technologie de séquençage utilisée pour le séquençage de l'échantillon ou des témoins à l'origine du calcul de la pseudo-référence.

Résultats préliminaires sur des échantillons de cfDNA

La détection des CNV dans les échantillons de cfDNA est encore un vaste sujet de recherche en bioinformatique. Les échantillons de cfDNA étant plus courts que les fragments d'ADN extraits des biopsies des patients, les algorithmes classiques basés sur une

comparaison de la profondeur de séquençage par fenêtre ne sont pas adaptés. Ainsi, nous travaillons à évaluer la pertinence de l'algorithme mCNA dans ce contexte.

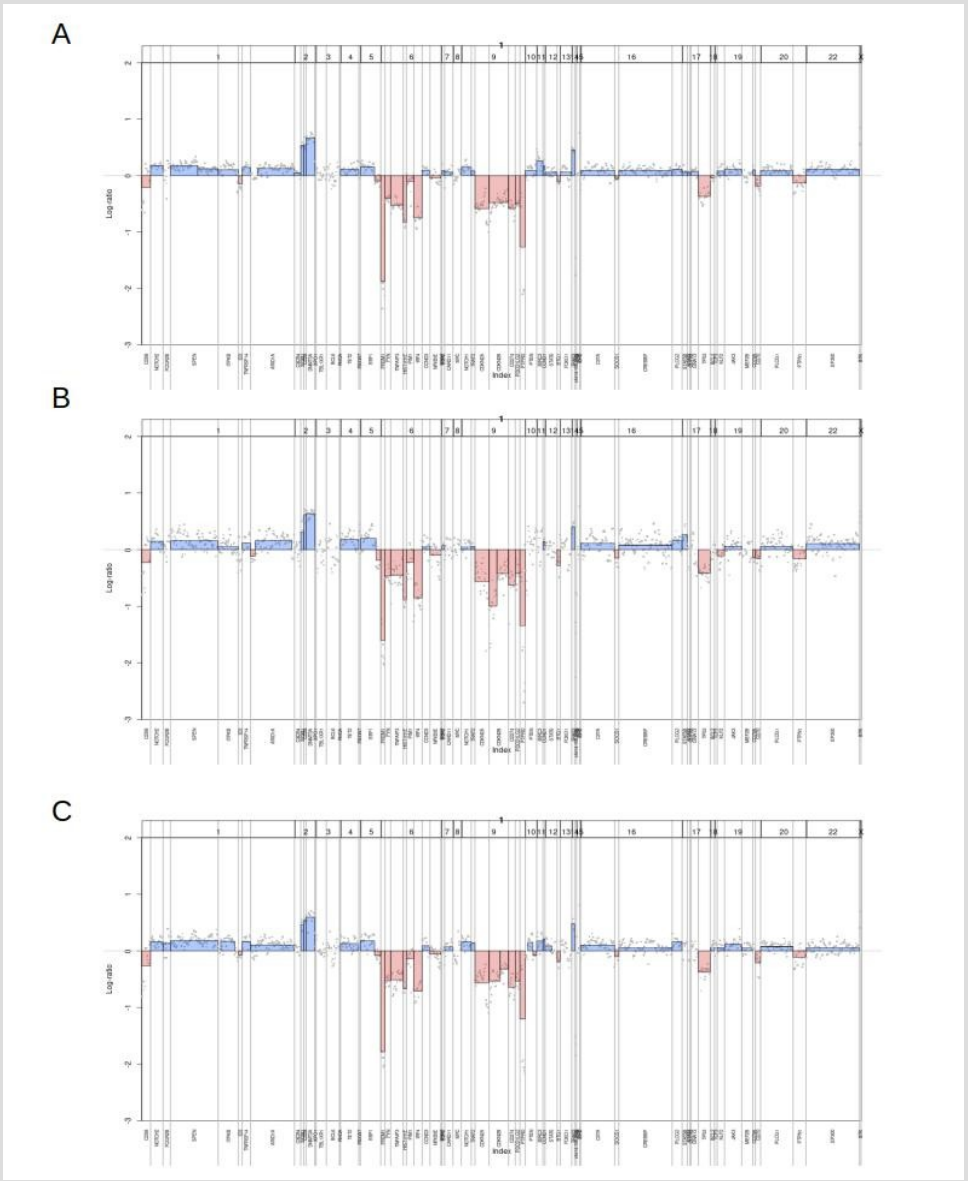


Figure 63: Profil de CNV obtenu par mCNA sur un échantillon de cfDNA au diagnostic d'un patient atteint de LDGCB séquencé en triplicat.

Les profils A, B et C correspondent à un même échantillon plasmatique au diagnostic de LDGCB séquencé sur le panel PanLymphome. Les profils montrent une concordance importante entre les différents réplicats.

Nous détaillerons dans cette section quelques résultats préliminaires concernant la détection de CNV à partir d'échantillons plasmatiques sur deux panels de séquençage précédemment décrits (Lymphopanel et Panlymphome).

Nous avons tout d'abord cherché à déterminer si les profils de CNV obtenus à partir de l'ADN extrait d'un même plasma sont reproductibles. Nous avons donc réalisé trois librairies à partir d'un même échantillon plasmatique de LDGCB pour lequel nous avions déjà détecté des mutations à une fréquence allélique moyenne de 40 % afin de nous placer dans les conditions les plus favorables pour établir un profil. Les résultats de l'analyse sont visibles sur la figure 63. Les profils des triplicats sont similaires pour cet échantillon ce qui valide le concept de détection des CNV via mCNA dans des échantillons de cfDNA. Par ailleurs, nous avons essayé d'obtenir des profils via des algorithmes classiques tels que ONCOCNV mais ceux-ci sont en échec.

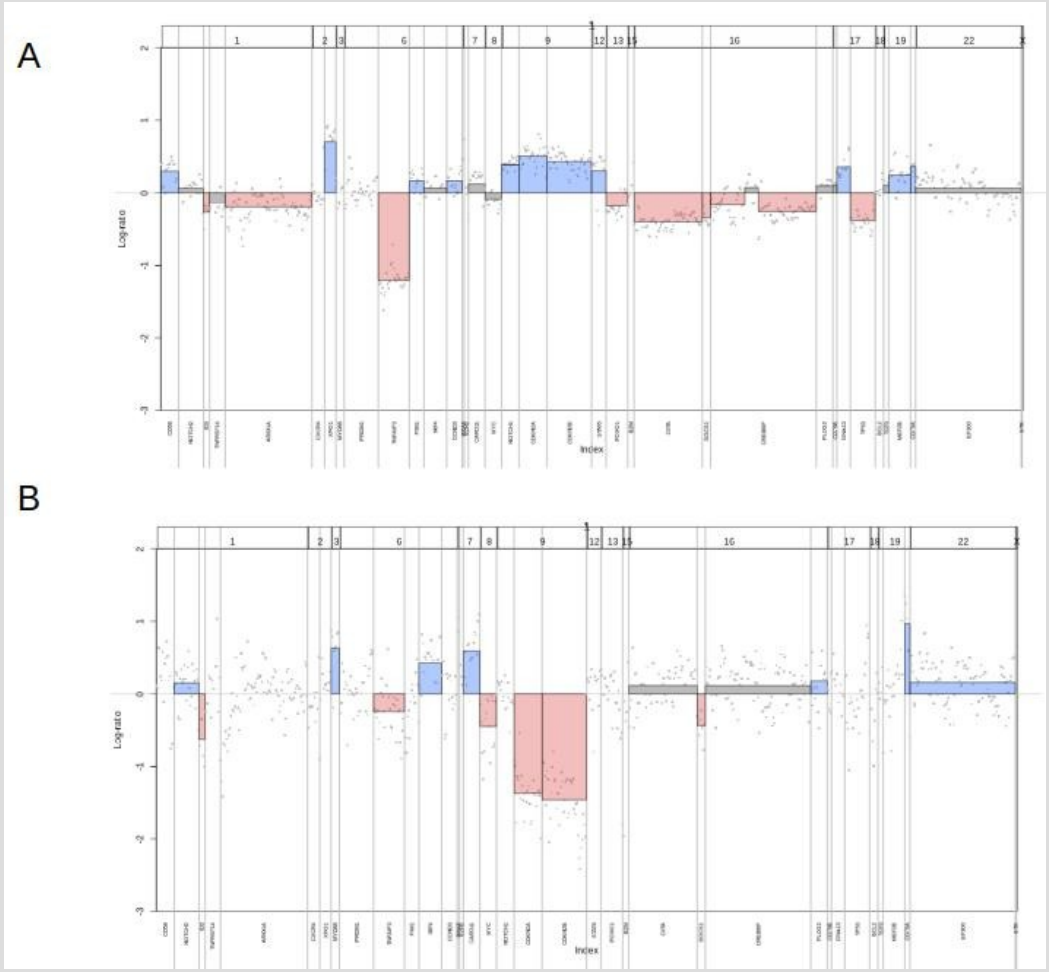


Figure 64: Profils de CNV obtenus par mCNA sur deux échantillons plasmatiques au diagnostic de patients atteints de LDGCB.

Les profils (A) et (B) ont été obtenus à partir de deux échantillons séquencés. Si la variance de la mesure des log-ratios est faible dans l'échantillon (A), on observe un bruit de mesure plus important dans l'échantillon (B) rendant complexe l'interprétation des profils.

D’autres profils de CNV à partir de plasmas ont été obtenus sur le LymphoPanel (figure 64). On y retrouve des anomalies récurrentes dans les LDGCB avec par exemple des délétions des régions 6q et de *CDKN2A/B*. Néanmoins, on observe que la variance observée des log-ratios n’est pas homogène entre les deux échantillons malgré les différentes étapes de normalisation implémentées dans mCNA.

C.4. Limitations et perspectives

Nous avons développé un nouvel algorithme de détection des CNV basé non plus sur la profondeur de séquençage mais sur le nombre d’UMI. Cette approche apporte une amélioration significative de la qualité des signaux d’acquisition en gommant les principaux biais d’amplification lors de la préparation des librairies. Les résultats sur les ADN extraits de biopsies permettent d’interpréter des profils en échec via les approches bioinformatiques classiques basées sur les comptages de lecture. L’apport de cette approche est particulièrement important sur les échantillons FFPE qui sont souvent dégradés. Nous avons pu démontrer la reproductibilité des profils et tenté d’apporter des arguments tangibles sur la limite de détection de cette nouvelle approche entre 5 et 10 %.

Les résultats de mCNA sur les échantillons de cfDNA montrent des premiers résultats prometteurs. Néanmoins, la recherche de CNV reste en pratique très complexe dans ces échantillons pauvres en contingent tumoral. On observe une variance dans les signaux de log-ratios souvent bien supérieure dans les échantillons de cfDNA que dans les biopsies sans pour autant parvenir à en identifier la cause. Nous avons pu déterminer, sur les premières cohortes de patients séquencés, une limite de sensibilité probablement plus proche de 20 % contre 5-10 % dans les biopsies. Cette limite exclut un nombre important d’échantillons de cfDNA analysables en fonction des pathologies. Néanmoins, l’approche reste intéressante dans certaines pathologies comme dans le LDGCB ou le PMBL dans lesquelles des fractions importantes de ctDNA sont retrouvées dans le plasma des patients au diagnostic.

C.5. Articles

Viailly et al. BMC Bioinformatics (2021) 22:120
https://doi.org/10.1186/s12859-021-04060-4

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers

Pierre-Julien Viailly[1,2*†], Vincent Sater[1,2,4†], Mathieu Viennot[1,2], Elodie Bohers[1,2], Nicolas Vergne[5], Caroline Berard[4], Hélène Dauchel[4], Thierry Lecroq[4], Alison Celebi[1,2,3], Philippe Ruminy[1,2], Vinciane Marchand[1,2], Marie-Delphine Lanic[1,2], Sydney Dubois[1,2], Dominique Penther[1,2], Hervé Tilly[1,2], Sylvain Mareschal[6] and Fabrice Jardin[1,2]

*Correspondence: pierre-julien.viailly@chb.unicancer.fr
†Pierre-Julien Viailly and Vincent Sater contributed equally to this work.
[1] INSERM U1245, Team Genomics and Biomarkers of Lymphoma and Solid Tumors, Normandie Univ, UNIROUEN, Rouen, France
Full list of author information is available at the end of the article

Abstract

Background: Recently, copy number variations (CNV) impacting genes involved in oncogenic pathways have attracted an increasing attention to manage disease susceptibility. CNV is one of the most important somatic aberrations in the genome of tumor cells. Oncogene activation and tumor suppressor gene inactivation are often attributed to copy number gain/amplification or deletion, respectively, in many cancer types and stages. Recent advances in next generation sequencing protocols allow for the addition of unique molecular identifiers (UMI) to each read. Each targeted DNA fragment is labeled with a unique random nucleotide sequence added to sequencing primers. UMI are especially useful for CNV detection by making each DNA molecule in a population of reads distinct.

Results: Here, we present molecular Copy Number Alteration (mCNA), a new methodology allowing the detection of copy number changes using UMI. The algorithm is composed of four main steps: the construction of UMI count matrices, the use of control samples to construct a pseudo-reference, the computation of log-ratios, the segmentation and finally the statistical inference of abnormal segmented breaks. We demonstrate the success of mCNA on a dataset of patients suffering from Diffuse Large B-cell Lymphoma and we highlight that mCNA results have a strong correlation with comparative genomic hybridization.

Conclusion: We provide mCNA, a new approach for CNV detection, freely available at https://gitlab.com/pierrejulien.viailly/mcna/ under MIT license. mCNA can significantly improve detection accuracy of CNV changes by using UMI.

Keywords: UMI, CNV calling, Next generation sequencing

Background

Recently, copy number variations (CNV) impacting genes involved in oncogenic pathways have attracted an increasing attention to manage disease susceptibility [1, 2]. CNV is one of the most important somatic aberrations in the genome of tumor cells.

Oncogene activation and tumor suppressor gene inactivation are often attributed to copy number gain/amplification or deletion, respectively, in many cancer types and stages.

CNV analysis refers to the detection of a difference in the dosage of a genomic locus containing one or more dosage-sensitive genes (zygosity). The resolution limit of conventional cytogenetics (approximately 5 Mb) has been improved by molecular cytogenetics using comparative genomic hybridization (CGH) and more recently array comparative genomic hybridization (aCGH). These technologies make it possible to detect genomic imbalances of < 100 kb, whereas more specialized array designs increase the resolution to ≤ 200 bp for specific targeted regions. Despite these performances, aCGH requires the purchase of a specific platform for data acquisition and its resolution is limited to the detection of tumoral clones that differ substantially in DNA content from a reference.

Next Generation Sequencing technologies (NGS) have rapidly supplanted traditional Sanger sequencing as the preferred methodology for the detection of actionable single nucleotide variations (SNV) in oncology. Diagnostic laboratories are now massively equipped with Illumina/Thermofisher sequencers. Massively parallel sequencing offers many advantages including high sensitivity and specificity for SNV and CNV detection within a single platform. Nevertheless, libraries must be amplified by PCR to produce a sufficient amount of signal. This amplification step introduces many biases for counting reads because the number of produced reads is no longer directly proportional to the number of initial unique targeted DNA fragments. The amplification factor of each region is unknown and depends on many parameters such as library size, GC content, region length or competition between primers overlapping the same locus while using amplicon-based libraries.

There are three main approaches to identify CNV from NGS data: read-pair (RP), split-read (SR), and read-depth (RD).

RP methods (BreakDancer [3], PEMer [4], Ulysses [5]) consist in comparing the average insert size between the sequenced read-pairs with an expected size based on a reference genome. The discordance between mapped paired-reads and the predetermined average insert size is then used to identify gain and loss of materials. Shorter/longer insert size than expected will correlate to the loss/gain of material, respectively.

SR methods evaluate CNV using paired reads where only one read of the pair has a reliable mapping quality whereas the other one partially fails to map to the reference sequence. These discrepancies within a read pair can potentially provide the precise position of insertion/deletion events. Several tools implementing SR strategies enable the detection of these breakpoints (SVseq2 [6], Gustaf [7], PRISM [8]) but they are limited to short insertions or deletions.

The RD approach consists in counting aligned reads overlapping a genomic region in a sliding window. These read counts (RC) are then compared between the sample of interest and a reference to compute CNV segmentation. A local decrease in sequencing depth will be associated with a loss of genomic material whereas its increase will be correlated to locus gain/amplification. Several tools were developed using RD-based approaches (CNVnator [9], CNV-seq [10]). This strategy seems particularly promising for the analysis of targeted sequencing experiments (TSE). TSE enables the sequencing of key genes or regions of interest to high depth (500–1000X or higher) and provides a cost-effective

strategy to identify variants at low allele frequencies. Some tools, such as ONCOCNV [11], were specially developed for the analysis of targeted amplicon-based libraries. Many biases due to the amplification step while preparing this type of library prevent the direct quantification of loci copy-number (size of the library, GC percentage, amplicon length, primer melting temperature, competition between primers...). It implies the use of normalization strategies to allow the comparison of read counts between samples.

Recent advances in NGS protocols allow for the addition of unique molecular identifiers (UMI) to each read. Each targeted DNA fragment is labeled by a unique random nucleotide sequence added to sequencing primers. UMI are especially useful for CNV detection by making each DNA molecule in a population of reads distinct. They allow the direct count of targeted DNA molecules before the library amplification by simply counting the number of unique UMI sequences per position of the alignment.

Here, we present mCNA (molecular Copy Number Alteration), a new methodology allowing the detection of copy number changes using UMI. We demonstrate the success of our algorithm on a dataset of patients diagnosed with Diffuse Large B-cell Lymphoma (DLBCL) and we highlight that mCNA results have a strong correlation with CGH. To assess the robustness and sensitivity limit of our approach, we used in silico simulation of copy number aberrations in a control sample and also sequential dilutions of REC-1 cell line.

Methods

Library construction

A Pan-lymphoma panel was designed using the QIAseq Targeted DNA Custom Panel Builder (QIAGEN) to identify alterations within important genes for lymphomagenesis. This panel targets 69 genes (hotspots, regions or whole gene) using 1493 gene specific primers. List of genes and number of GSP per gene are provided in Additional file 1: Table S1.

The QIAseq Targeted DNA chemistry introduces molecular barcodes (UMI) to enable digital sequencing and to identify PCR duplicates (Fig. 1). The molecular barcodes are short aleatory nucleotide sequences of 12 bp length added to each read before the library amplification. Statistically, this process provides 4^{12} possible indices per adapter; hence, each DNA molecule in the sample receives a unique UMI sequence.

Subjects and methods

Study design and patients

22 adult patients with de novo CD20+ Diffuse Large B-cell Lymphoma (DLBCL) or primary mediastinal B-cell lymphoma (PMBL) were selected from the prospective, multicenter, and randomized LNH-03B LYSA trials with available frozen tumor samples and adequate DNA quality. CGH was previously performed for these samples after whole-genome amplification against a Promega normal DNA pool using Agilent SurePrint G3 4 × 180 K microarrays. Briefly, arrays were scanned with Agilent Feature Extraction and processed with cghRA pipeline as previously described [12].

DNA from REC-1 cell line, established from the lymph node of a 61-year-old man with refractory B-cell lymphoma, was extracted. Dilutions at 50%, 30%, 20%, 10% and

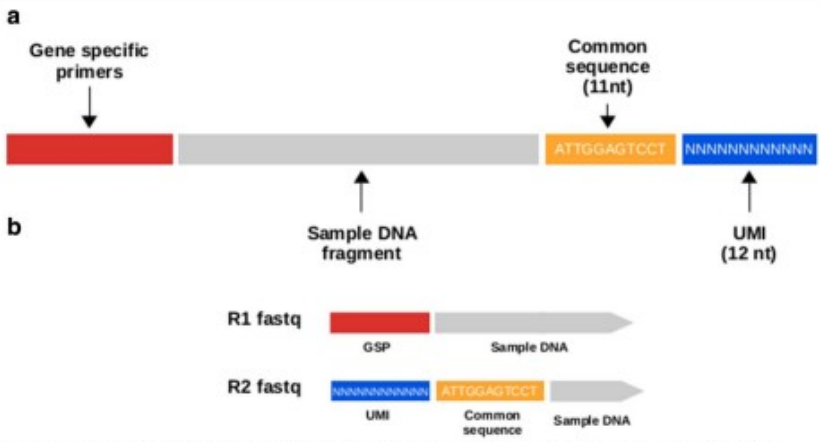


Fig. 1 Qiaseq library construction scheme. **a** Libraries were composed of a pool of gene specific primers (GSP) designed to target genomic regions of interest, a sample DNA fragment, a common sequence of 11 nucleotides used during the library construction as an universal primer and finally a random sequence of 12 nucleotides (UMI). **b** Structure of reads in the Fastq files obtained after paired-end sequencing

5% of this DNA were performed using Human Mixed Genomic DNA Promega. Human Genomic DNA comes from multiple anonymous donors.

Five blood samples of healthy individuals were collected and used as a control to construct the pseudo-reference profile.

Sample collection and sequencing

Tumor genomic DNA (gDNA) was isolated from fresh diagnostic tissue biopsies or blood. Samples were quantified using QuBit High Sensitivity dsDNA (Thermo Fisher Scientific).

gDNA samples were sequenced with the entire Pan-lymphoma panel. 30 ng of gDNA were enzymatically fragmented and end repaired, followed by ligation of the molecular barcoded adaptators (UMI). After purification, target enrichment was carried out using the set of 1493 gene specific primers. Then, enriched DNA was submitted to universal PCR with a number of cycle adapted to this number of primers. Purified libraries were quantified using QuBit High Sensitivity dsDNA.

Finally, libraries were sequenced on Illumina MiSeq (paired-end, 2 × 150 bp) following manufacturer's user manual (Illumina, CA).

Library sequencing and bioinformatics pre-processing

Briefly, gene-specific primers and common regions were trimmed from R1 and R2 fastq using an in-house program. UMI sequences were extracted from read construction using UMI-tools [13].

Reads were aligned against hg19 reference genome using BWA-mem [14] and standardized according to the GATK3 Best Practices recommendations. A detailed bioinformatics pipeline is provided in Additional file 1: Fig. S1.

mCNA algorithm

In this article, we present a new strategy to detect copy number changes for targeted panels of genes using UMI. The algorithm is composed of four main steps: the construction of UMI count matrices, the use of control samples to construct a pseudo-reference, the computation of log-ratios (LR), the segmentation and finally the statistical inference of abnormal segmented breaks (Fig. 2).

Prerequisites

mCNA algorithm requires sequencing libraries introducing one or more short aleatory sequences (Unique Molecular Identifiers, UMI) in reads construction. UMI sequences

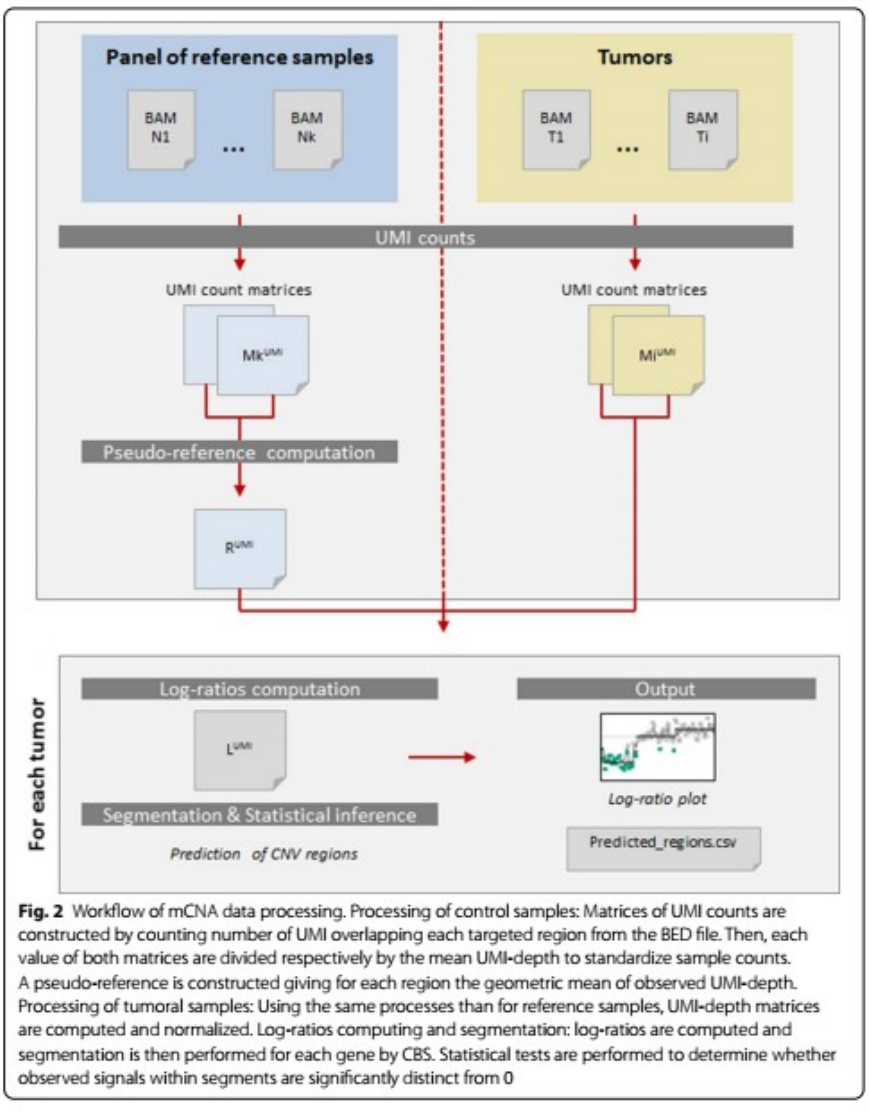


Fig. 2 Workflow of mCNA data processing. Processing of control samples: Matrices of UMI counts are constructed by counting number of UMI overlapping each targeted region from the BED file. Then, each value of both matrices are divided respectively by the mean UMI-depth to standardize sample counts. A pseudo-reference is constructed giving for each region the geometric mean of observed UMI-depth. Processing of tumoral samples: Using the same processes than for reference samples, UMI-depth matrices are computed and normalized. Log-ratios computing and segmentation: log-ratios are computed and segmentation is then performed for each gene by CBS. Statistical tests are performed to determine whether observed signals within segments are significantly distinct from 0

must be extracted from raw FASTQ files before alignment and appended to read identifiers using UMI-tools [13]. Processed reads must be aligned against a reference genome to produce BAM file. A BED file is also required, giving for each targeted region the chromosome name, the start/end positions of the locus and the gene name.

Details for complete bioinformatics processing of QIAseq Targeted DNA Panel are provided in Additional file 1: Fig. S1.

Construction of UMI-depth matrices

We define M^{UMI} as the UMI-depth matrix of one BAM file. P is the total number of targeted regions. $C_{\mathrm{p}}^{\mathrm{UMI}}$ reflects the number of unique UMI overlapping p region and U the total number of unique UMI of one sample.

Each region supplied in the BED file is scanned using *scanBam* function of Rsamtools package [15]. $C_{\mathrm{p}}^{\mathrm{UMI}}$ is computed from unique UMI sequences extracted from read names overlapping p.

Each matrix M^{UMI} is finally normalized by U to allow the comparison between samples, as shown in the Additional file 1: Fig S2.

Pseudo-reference construction

From M^{UMI} matrices of normal samples, a geometric mean of $C_{\mathrm{p}}^{\mathrm{UMI}}/U$ is computed line by line to create a vector R^{UMI} of dimensions $(1, P)$.

To automatically detect outlier samples, Root-Mean-Square Deviations (RMSD) are computed between $C_{\mathrm{p}}^{\mathrm{UMI}}/U$ and $R_{\mathrm{p}}^{\mathrm{UMI}}$ for each region p of each control sample.

Samples with at least 20% of regions with $RMSD_p > T$ are excluded from baseline construction, with T defined as:

$$T = Q3(RMSD_p) + 1.5 \times IQR(RMSD_p)$$

If at least one sample is filtered, R^{UMI} vector is updated with passing filter M^{UMI} matrices only. The same process is applied to detect outlier noisy regions. These positions are defined as sequenced regions with at least $RMSD_p > T$ in 50% of control samples.

Log-ratios and signal centering

We define the log-ratio $L_{\mathrm{p}}^{\mathrm{UMI}}$ as:

$$L_{\mathrm{p}}^{\mathrm{UMI}} = log2\left(\frac{M_{\mathrm{p}}^{\mathrm{UMI}}}{R_{\mathrm{p}}^{\mathrm{UMI}}}\right)$$

where $M_{\mathrm{p}}^{\mathrm{UMI}}$ is the UMI count of a tumor sample for the region p and $R_{\mathrm{p}}^{\mathrm{UMI}}$ is the UMI pseudo-reference vector of control samples for the region p.

A Gaussian mixture model with one to three mixture components is estimated from $L_{\mathrm{p}}^{\mathrm{UMI}}$ using *Mclust* function of R package mclust [16]. The estimated gaussian closest to $\widehat{L^{\mathrm{UMI}}} = 0$ is used to center the signal by subtracting its average from the $L_{\mathrm{p}}^{\mathrm{UMI}}$ values. Indeed, we assume that the Gaussian of our signal closest to 0 corresponds to a diploid state. This centering step could be disabled via the program's arguments.

Segmentation

Each gene is composed of n consecutive regions and we define a vector of log-ratios V_n^{UMI} used for segmentation, as:

$$V_n^{UMI} = \left\{ L_p^{UMI}; L_{p+1}^{UMI}; \cdots \right\} (p \in n)$$

mCNA uses the circular binary segmentation (CBS) method implemented in the R package PSCBS [17] to segment V_n^{UMI}.

To avoid breakpoints at outlier values, a vector of weights W is given to CBS segmentation function. W is inversely proportional to the variances of C_p^{UMI}/U observed in the control samples and defined as:

$$W_p = \frac{1}{var(M_p^{UMI}/U)} \text{(within controls)}$$

W_p are then transformed to be limited to the interval [0,1] as follows:

$$W'_p = \frac{W_p - min(W_p)}{max(W_p) - min(W_p)}$$

Finally, a Student's t-test is performed on each segmented region to test whether or not the V_n^{UMI} vector is significantly different from the reference value of 0. To avoid false positive segments, a FDR correction is applied.

Estimation of tumoral content

We define G_n^{UMI} and D_n^{UMI} the vectors V_n^{UMI} of significant amplified/deleted segments, respectively. We use D_n^{UMI} and G_n^{UMI} distributions to estimate tumor enrichment assuming that means of these distributions reflect a gain/loss of one segment copy and that log-ratios are a mixture of both tumoral and normal signals. We define as c the percentage of tumor enrichment to estimate.

Two independent estimates of c were produced: one from the significantly deleted segments and the other from those amplified. The estimation of c cannot be done in one step because log-ratios involving one gain or one loss are not symmetrical. For example, the loss of one copy of a segment in a sample containing only tumor cells will lead to a log-ratio equal to $\log_2(\frac{1}{2}) = -1$ while a gain of one copy will lead to $\log_2(\frac{3}{2}) = 0.58$.

From amplified regions, the distribution of L_p^{UMI} can be decomposed as follows:

$$\begin{aligned} L_p^{UMI} = \log_2\left(c \times \frac{3}{2} + (1-c) \times \frac{2}{2}\right) &\iff 2^{L_p^{UMI}} = \frac{1}{2}c + 1 \\ &\iff c = 2\left(2^{L_p^{UMI}} - 1\right) \end{aligned} \quad (1)$$

The mean value of the distribution of G_n^{UMI} is used in order to complete this Eq. (1) to estimate c.

The same decomposition is carried out considering the loss of a copy:

$$L_p^{UMI} = \log_2\left(c \times \frac{1}{2} + (1-c) \times \frac{2}{2}\right) \iff c = -2\left(2^{L_p^{UMI}} - 1\right) \quad (2)$$

The mean value of the distribution of D_n^{UMI} is used in order to complete this Eq. (2) to estimate c.

The algorithm output by default the mean value of this two independant estimates of c

Results

Comparison between read-depth and UMI-depth signals

To allow the comparison between read-depth and UMI-depth signals, we extracted respective counts from our reference samples for each targeted region. Theses counts were normalized respectively by the mean read-depth/the mean UMI-depth to make samples comparable. Measured variances were significantly lower when taking into account the UMI-depth and not the read-depth (p value < 2.2e−16), as shown in Fig. 3.

Construction of Pan-Lymphoma baseline

From mCNA quality control step, one control sample (CTL-22081) was excluded during pseudo-reference computation because of too high RMSD. The distribution of normalized UMI counts for this sample was clearly distinct from others as shown in the

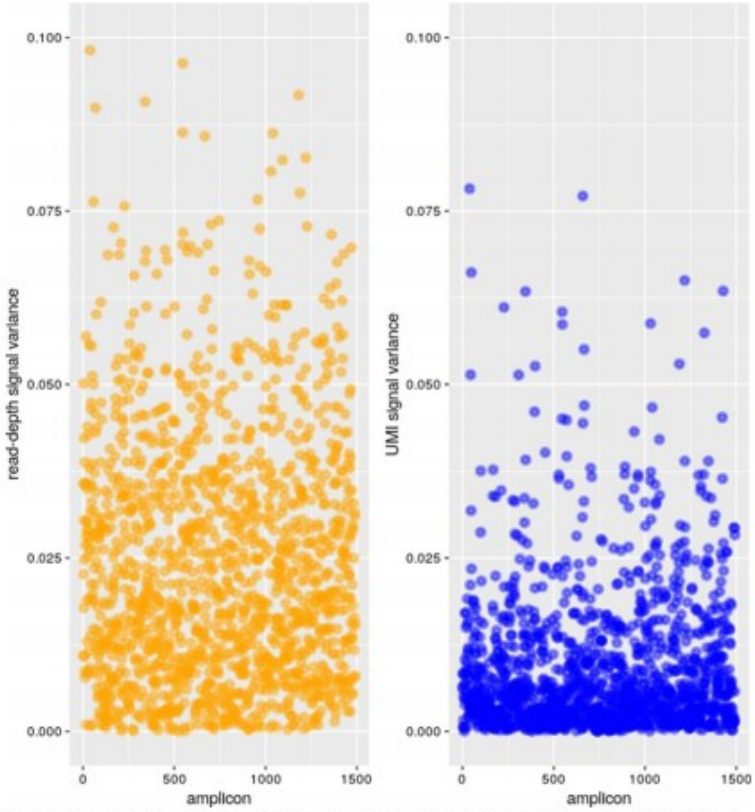


Fig. 3 Variance of measurements within reference samples. The scatter plots represent the normalized read-depth counts on the left and the normalized UMI-counts on the right within control samples. Each targeted region is represented by a point

Additional file 1: Fig. S2. mCNA also detected 31 targeted regions not passing RMSD filters which were excluded. List of outliers and their characteristics are provided in Additional file 1: Table S2.

To validate our approach, we determined the correlation between normalized UMI count matrices of control samples and the computed pseudo-reference vector for each targeted region (Fig. 4). Signals were significantly and strongly correlated (r>0.96, p < 2.2e−16), which means that the computed pseudo-reference perfectly reflects the controls.

Example of mCNA profile

For each tested sample, mCNA generates a csv file summarizing by segment the measured data and the significance of the tests. A graph is also provided representing the log-ratios by region, the segmented signal and the results of the test. An example of profile is shown in Fig. 5.

Comparison between mCNA and CGH data

In order to validate mCNA approach, we first compared CGH and NGS data (Fig. 6). We estimated log-ratios for each targeted region of the Pan-lymphoma panel using mCNA approach and then those estimated from CGH.

We observe a strong correlation between log-ratios of both technologies (r = 0.74). The majority of discrepancies are visible for $L_{\mathrm{CGH}} = 0$ which may show a lack of sensitivity of CGH due to a lack of probe coverage.

To further our comparison, we extracted all predicted mCNA segments of our 22 tumor samples. These segments were then annotated with CGH results. 114/120 (95%)

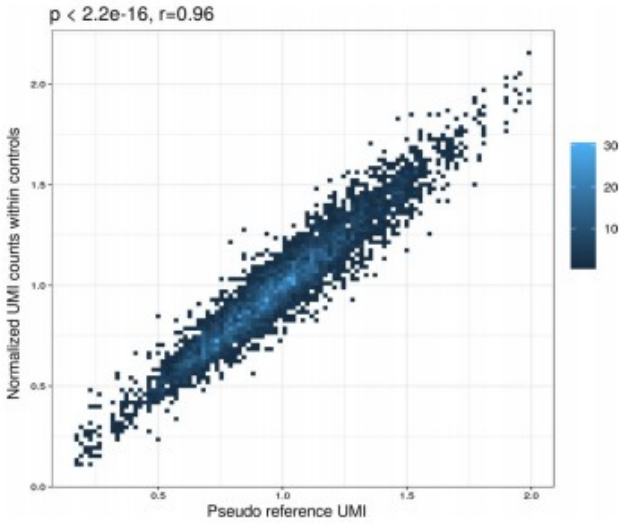


Fig. 4 Correlation between normalized UMI count matrices and matched pseudo-reference vector. The scatter plot represents the normalized counts $C_{\mathrm{p}}^{\mathrm{UMI}}/U$ of M^{UMI} matrices within control samples and the associated pseudo-reference vector R^{UMI}. Each targeted region is represented by a point. A Pearson correlation test is computed and correlation coefficient is displayed at the top of the figure

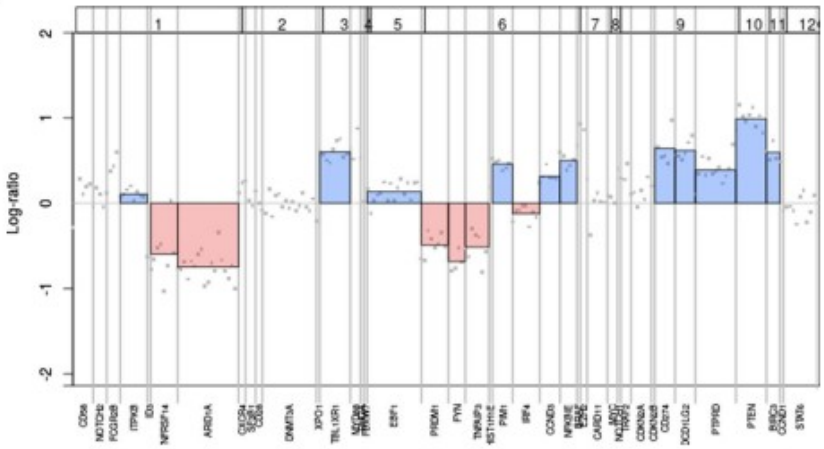


Fig. 5 Example of mCNA profile. The top of the panel represents the chromosome location. Each vertical line delimits a gene. The log-ratio plot gives estimated log-ratios for the sample and the results of CBS segmentation. Blue/red segments represent respectively significant gain/lost segments. Gene names are displayed at the bottom of the graph

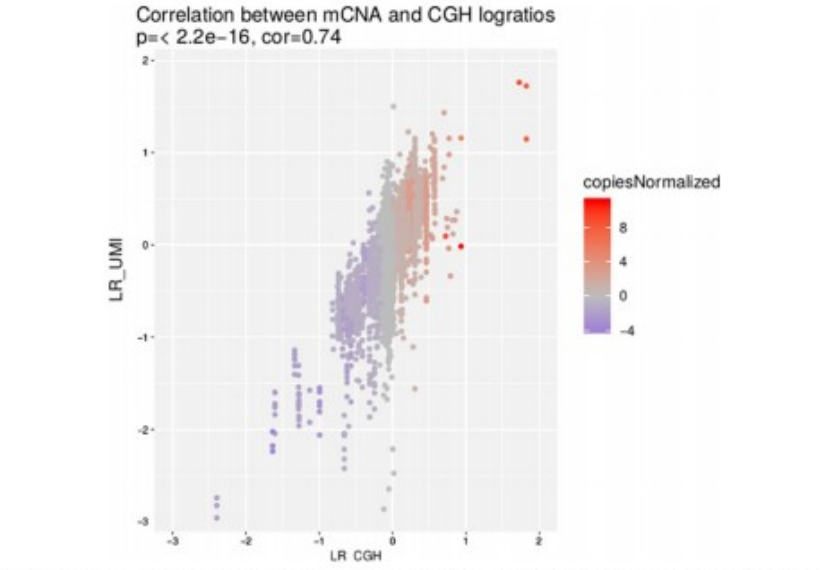


Fig. 6 Correlation between mCNA and CGH log-ratios. Each point of the plot corresponds to a region of the Pan-Lymphoma Panel annotated by CGH results and estimated by mCNA. Colors show the number of gain or loss estimated by CGH data regarding the region. Pearson correlation coefficient is displayed at the top of the plot

mCNA segments were predicted deleted by CGH and 175/221 (79%) were predicted as gain. 723/978 were predicted normal by mCNA and confirmed in CGH (74%), leading to an overall agreement between the two datasets of 83%.

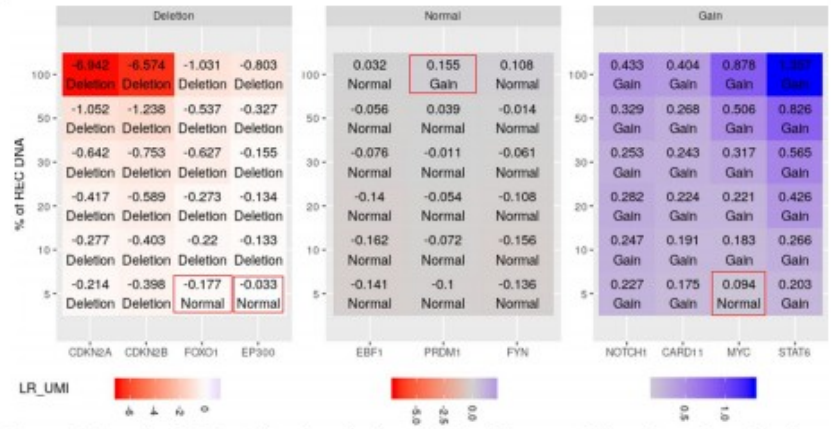


Fig. 7 mCNA results of REC-1 cell line through dilutions. Each cell contains the log values obtained for the given gene and the final status of the segment given by mCNA

Robustness and sensitivity limit

To estimate theoretical sensitivity limit of mCNA approach, we first edited M^{UMI} matrix of UMI count of one control sample (16,464) to introduce amplification of *XPO1*, gain of *IRF4*, heterozygous deletion of *CDKN2A* and homozygous deletion of *CDKN2B*. We applied an in silico dilution of these abnormal segments at 100%, 50%, 20%, 10% and 5% of tumor cells and applied mCNA to determine whether or not segments were significantly found after signal centering, segmentation and statistical test application. Results were summarized in Additional file 1: Fig. S4 and Additional file 1: Table S3. We found a strong correlation between expected and computed log-ratios ($r = 0.99$, $p = 7.19e8\text{-}27$) after signal centering and segmentation. mCNA was able to detect all in silico abnormal segments for tumor cell percentage between 10% and 100%. At 5%, only segments involving gain or loss of more than one copy were significantly found.

To confirm in silico results, REC-1 cell line was sequenced on two different runs to estimate the robustness of $L_{\mathrm{p}}^{\mathrm{UMI}}$ measurement using the Pan-Lymphoma panel. We found a strong correlation between the two replicates ($r = 0.93$, $p < 0.001$) even if the sequencing depths were not the same (1851X/2217X). Details are provided in Additional file 1: Fig. S3.

30/31 segments were predicted as gains in both replicates (96.77%), 21/23 (91.30%) as normal and 17/18 (94.44%) as deleted, thus giving an average agreement of 94.17%. Discordant predictions result from segments having a low number of targeted regions and a small log-ratio variation.

Finally, dilutions of REC-1 DNA were performed at 50%, 30%, 20%, 10% and 5%. REC-1 is a near-diploid cell line of male origin with a modal chromosome number of 45 and a polyploidy rate of 10%. Its karyotype is highly rearranged with approximately 5–6 derivative chromosomes in the karyology that have been described. Significant segments in this cell line were selected from the initial profile to evaluate the sensitivity of our approach through the different dilutions. Results seem consistent up to a threshold of 10% enrichment (Fig. 7). Above this threshold, the evaluation of tumor content

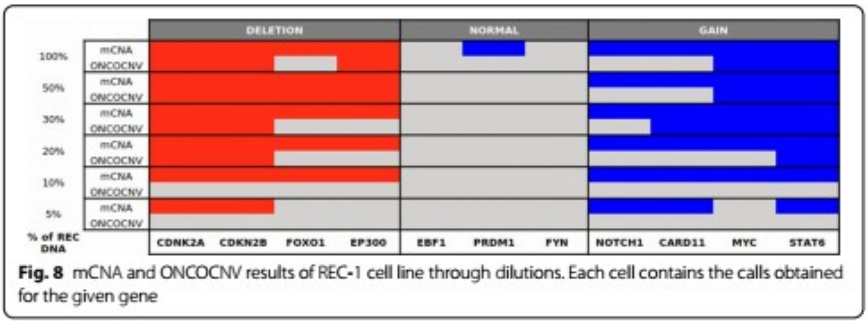


Fig. 8 mCNA and ONCOCNV results of REC-1 cell line through dilutions. Each cell contains the calls obtained for the given gene

seems consistent between expected and estimated percentage of tumor cells ($r = 0.98$) as shown in Additional file 1: Fig. S5.

Comparison to read-depth algorithm

To assess mCNA's analytical performance, we decide to compare our UMI-depth approach to the read-depth algorithm ONCOCNV [11] using REC-1 dataset. ONCOCNV was commonly used for the analysis of targeted sequencing panel of genes. It uses several normalization steps on read counts to erase library amplification biases such as library size, GC content of each region or amplicon length.

We hypothesized that the direct count of UMI could improve the limit sensitivity of ONCOCNV insofar as we have shown that the signal in UMI was less noisy than read counts. ONCOCNV results were generated for all REC-1 dilutions using the same control samples as those used to construct mCNA baseline.

As expected, mCNA achieved much higher prediction accuracies than ONCOCNV as the percentage of tumor cells decreases (Fig. 8). Here, accuracy measures the proportion of genes with correctly annotated copy number status compared to the initial REC-1 profile: normal, gain or deletion. Considering the results of the algorithms from 100 to 10% of REC-1 DNA, the overall prediction accuracy fluctuated from 0.90 to 0.27 for ONCOCNV, while it was significantly higher for mCNA : 1.0 to 0.90. Interestingly, while mCNA results look consistent at 10%, ONCOCNV fails to detect heterozygous deletions of *FOXO1* and *EP300* at 30% of tumoral cell.

Discussion

We proposed a new methodology to be used to detect copy number changes for targeted panels of genes using unique molecular identifiers. By changing the source of information from sequencing depth to UMI depth, mCNA provides a simple and robust methodology for the detection of CNV.

We demonstrated that using UMI-depth signal, and not read-depth signal, seems more robust in samples without abnormal copies. The algorithm uses a pool of reference samples to construct a pseudo-reference and includes a filtering step to automatically exclude samples and/or targeted regions with abnormal variance. We

demonstrated that this in silico baseline profile reflects the reference samples and enables the estimation of CNV changes in unpaired tumor samples.

mCNA provides a strong estimation of log-ratios which correlates to our CGH dataset of 22 DLCBL samples. As we expected, the majority of discrepancies are visible for short breaks within genes probably due to a lack of probe coverage of Agilent SurePrint G3 4x180K microarrays. To avoid overestimation of breakpoints due to outlier values, mCNA provides a vector of weights to CBS segmentation function. We also recommend the use of at least 6 non-overlapping amplicons to properly estimate the state of a targeted region.

As we expected, we failed to detect CNA for samples that were highly contaminated by normal cells (less than 10% of tumor content). In this case, the noise in measurements is higher than the expected difference between measurements in the case of one CNV event. This observed threshold of 10% was confirmed by in silico simulation and also by sequential dilution of REC-1 cell line.

Our approach is designed to be used for targeted gene panels and thus doesn't allow the combination of UMI-depth signal and B allele frequencies to improve the sensitivity of our CNV calling approach, as for analyses at the exome scale for example.

Another limitation of mCNA approach is the assumption that the majority of the signal corresponds to a diploid state. Polyploid profiles for example still remain challenging because the algorithm proceeds to center the signals. We recommend for panels targeting very frequently altered genes to deactivate this centering step.

Finally, mCNA gives the opportunity to obtain both the mutational and the copy number status at no additional cost. It helps in the interpretation of frequently altered genes, such as *TP53* for example, for which mutations are often associated with copy abnormalities.

Conclusion

In this article, we present a new strategy to detect copy number changes for targeted panels of genes using UMI. mCNA is composed of four main steps: the construction of UMI count matrices, the use of reference samples to construct a pseudo-reference, the computation of log-ratios , the segmentation and finally the statistical inference of segmented breaks.

Abbreviations

CNV: Copy number variation; UMI: Unique Molecular Identifiers; mCNA: Molecular Copy Number Alteration; LR: Log-ratio; NGS: Next Generation Sequencing; SNV: Single Nucleotide Variation; PCR: Polymerase chain reaction; RP: Read-pair; SR: Split-read; RD: Read-depth; RC: Read count; TSE: Targeted Sequencing Experiment; DLBCL: Diffuse Large B-Cell Lymphoma; PMBL: Primary Mediastinal B-cell Lymphoma; GSP: Gene Specific Primer; BED: Browser Extensible Data; BAM: Binary Alignment Map; RMSD: Root-Mean-Square Deviation; FDR: False Discovery Rate.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04060-4.

Additional file 1. Supplementary Figures S1–S5, Supplementary Tables S1–S3.

Acknowledgements

None.

Authors' contributions
PJV and VS conceived the algorithm. MV, MDL and EB performed the experiments. NV, CB, HD, TL, AC and SM contributed to the statistical design of the study. PR, VM, MDL, DP, and HT contributed to data interpretation. PJV, VS, SD and FJ contributed to the writing of the manuscript. All authors discussed the results and contributed to the final manuscript. All authors read and approved the final manuscript.

Funding
This study was funded by grants from the Centre Henri Becquerel (Rouen, France). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
mCNA is available at https://gitlab.com/pierrejulien.viailly/mcna/ under MIT license. The datasets analysed during the current study are also available in mCNA data repository.

Declarations

Ethics approval and consent to participate
Sequencing data results from patients enrolled in the prospective, multicenter, and randomized LNH-03B LYSA clinical trials. This study was performed with approval of the Ethic Committee Haute-Normandie on 2003 and written informed consent was obtained from all participants at the time of enrollment.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] INSERM U1245, Team Genomics and Biomarkers of Lymphoma and Solid Tumors, Normandie Univ, UNIROUEN, Rouen, France. [2] Centre Henri Becquerel, Rouen, France. [3] Master Bioinformatique BIM, Normandie Univ, UNIROUEN, Rouen, France. [4] LITIS EA 4108, Normandie Univ, UNIROUEN, Rouen, France. [5] LMRS UMRS 6085, Normandie Univ, UNIROUEN, Rouen, France. [6] INSERM U1052 UMR CNRS 5286, Cancer Research Center of Lyon, Lyon, France.

Received: 10 August 2020 Accepted: 2 March 2021
Published online: 12 March 2021

References

1. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1(6):62.
2. Jardin F, Jais J-P, Molina T-J, Parmentier F, Picquenot J-M, Ruminy P, Tilly H, Bastard C, Salles G-A, Feugier P, Thieblemont C, Gisselbrecht C, de Reynies A, Coiffier B, Haioun C, Leroy K. Diffuse large B-cell lymphomas with CDKN2a deletion have a distinct gene expression signature and a poor prognosis under R-CHOP treatment: a GELA study. Blood. 2010;116(7):1092–104.
3. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: identification of genomic structural variation from paired-end read mapping. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors. Current protocols in bioinformatics. Wiley; 2014. p. 15-6115611. https://doi.org/10.1002/0471250953.bi1506s45.
4. Korbel JO, Abyzov A, Mu X, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009;10(2):23.
5. Gillet-Markowska A, Richard H, Fischer G, Lafontaine I. Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. Bioinformatics. 2015;31(6):801–8.
6. Zhang J, Wang J, Wu Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinform. 2012;13 Suppl 6:6.
7. Trappe K, Emde A-K, Ehrlich H-C, Reinert K. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics (Oxford, England). 2014;30(24):3484–90.
8. Jiang Y, Wang Y, Brudno M. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics (Oxford, England). 2012;28(20):2576–83.
9. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.
10. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinform. 2009;10(1):80.
11. Boeva V, Popova T, Lienard M, Toffoli S, Kamal M, Le Tourneau C, Gentien D, Servant N, Gestraud P, Rio Frio T, Hupé, P, Barillot E, Laes J-F. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. Bioinformatics. 2014;30(24):3443–50
12. Mareschal S, Ruminy P, Alcantara M, Villenet C, Figeac M, Dubois S, Bertrand P, Bouzelfen A, Viailly P-J, Penther D, Tilly H, Bastard C, Jardin F. Application of the cghRA framework to the genomic characterization of Diffuse Large B-Cell Lymphoma. Bioinformatics (Oxford, England). 2017;33(19):2977–85.
13. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017;27(3):491–9.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754–60.

15. Morgan M, Pagès H, Obenchain V, Hayden N. Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. Bioconductor version: Release (3.10); 2019. https://bioconductor.org/packages/Rsamtools/. Accessed 2019-12-04.
16. Fraley C, Raftery AE, Scrucca L, Murphy TB, Fop M. mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation; 2019. https://CRAN.R-project.org/package=mclust. Accessed 2019-03-29.
17. Bengtsson H, Neuvial P, Seshan VE, Olshen AB, Spellman PT, Olshen RA. PSCBS: analysis of parent-specific DNA copy numbers; 2019. https://CRAN.R-project.org/package=PSCBS. Accessed 2019-12-04.

Publisher's Note

D. Simulation de données avec UMI : UMI-Gen

D.1. État de l'art

Nous avons vu dans les chapitres précédents que la détection des variants de faible fréquence est devenue un enjeu majeur pour l'analyse d'échantillons dans lesquels les fragments tumoraux sont particulièrement dilués. Des nouvelles chimies de préparation de librairie permettent l'introduction de barcodes moléculaires uniques appelés UMI visant à améliorer la sensibilité de détection de ces variants de faible fréquence et ont conduit au développement des nouveaux algorithmes précédemment cités : UMI-VarCal (chapitre III.B) et mCNA (chapitre III.C).

Afin de pouvoir comparer les outils de détection de variants objectivement, il est très difficile de se baser sur une analyse comparative des résultats des algorithmes à partir d'échantillons biologiques dilués dans la mesure où la vérité biologique est très souvent inconnue. Il est donc nécessaire de pouvoir simuler informatiquement des fichiers dans lesquels des mutations sont introduites à des fréquences alléliques connues afin de pouvoir évaluer la sensibilité et la spécificité de chaque algorithme. De nombreux programmes existent pour simuler des données de séquençage et introduire des mutations ou des variations de nombre de copies comme IntSIM [290] ou SVSR [291]. Néanmoins, il n'existe pas dans la littérature d'algorithmes de simulation intégrant des UMI dans les données générées.

Nous avons ainsi chercher à développer un nouvel algorithme de simulation de données de séquençage avec UMI appelé UMI-Gen. UMI-Gen utilise plusieurs échantillons biologiques afin d'estimer le bruit de fond de séquençage et les scores de qualité par position de l'alignement. Il permet ensuite d'ajouter des artefacts et des mutations à des fréquences alléliques souhaitées dans les lectures générées *in silico*. Afin de valider l'approche, nous avons utilisé 6 échantillons contrôles afin d'estimer le bruit de fond d'un panel de séquençage puis nous avons introduit 15 variants à différentes positions de l'alignement. Finalement, à partir de ces données générées, nous avons comparé 4 algorithmes : SiNVICT [292], OutLyzer [276], DeepSNVMiner [277] et UMI-VarCal [293].

D.2. Concepts et implémentation

UMI-Gen nécessite trois paramètres d'exécution : une liste de fichiers BAM/SAM témoins, un fichier BED contenant les coordonnées des régions alignées et le fichier FASTA du génome de référence indexé. Les fichiers contrôles sont des fichiers BAM ou SAM

d'échantillons de référence dépourvus d'anomalies somatiques. L'outil a été développé dans le but de simuler des résultats de séquençage ciblé et il est donc obligatoire de fournir une liste de coordonnées génomiques pour réaliser la simulation. Les différentes étapes de l'algorithme UMI-Gen sont visibles sur la figure 65.

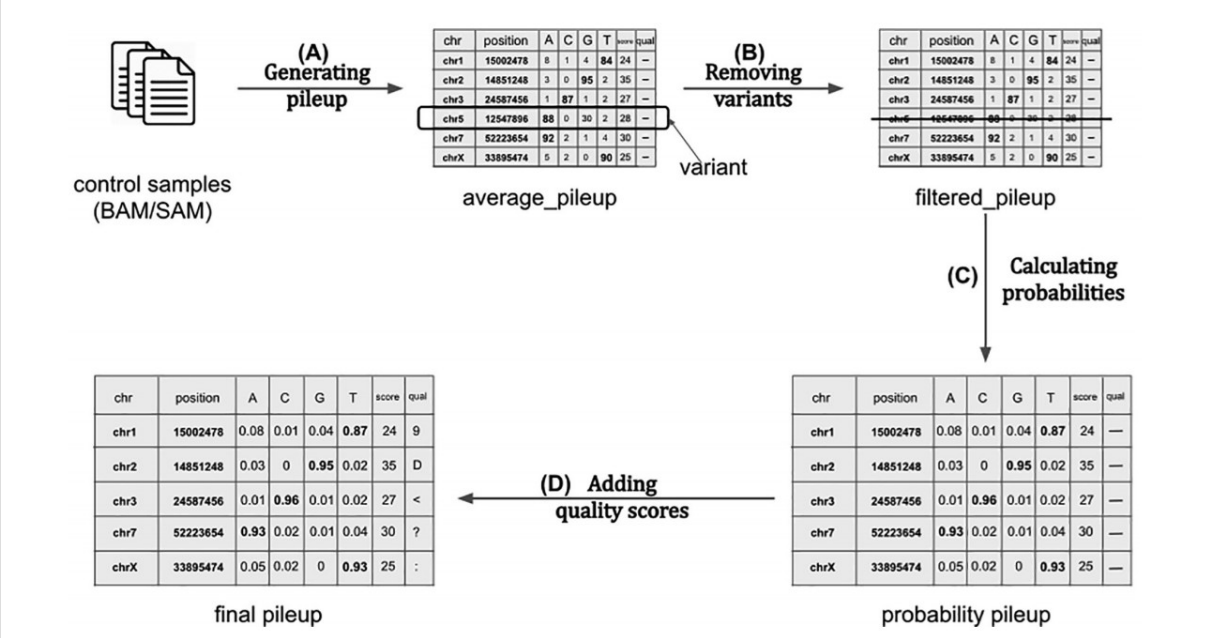


Figure 65: UMI-Gen - Création du *pileup* témoin à partir des échantillons contrôles.

L'algorithme UMI-Gen intègre différentes étapes successives afin de construire un *pileup* utilisé pour la simulation à partir des échantillons de référence. La première étape (A) consiste à réaliser un *pileup* moyen à partir des différents fichiers en entrée de l'algorithme afin d'obtenir le nombre de A, T, G, C, d'insertions et de délétions moyen observé à chaque position de l'alignement dans les régions spécifiées dans le fichier BED. Ce *pileup* est ensuite filtré (B) de sorte à éliminer les variants spécifiques de chaque témoin puis les données de comptages sont converties en fréquences. Le score de qualité moyen à chaque position est finalement ajouté à cette matrice (D).

La première étape de l'algorithme UMI-Gen vise à générer un fichier *pileup* pour chaque témoin à partir de la liste de fichiers fournie lors de l'exécution. Ces fichiers *pileup* sont ensuite moyennés. Le fichier *pileup* moyenné peut être sauvegardé de sorte à être réutilisé en argument de UMI-Gen si l'on souhaite simuler plusieurs fichiers à partir d'une même liste de témoins. Ce fichier *pileup* moyenné contient l'occurrence moyenne du nombre de bases et d'insertions/délétions retrouvée à chaque position du fichier BED. On y retrouve aussi des métriques de qualité telles que la profondeur moyenne et le score de qualité de base moyen

observés à chaque position. Afin d'éliminer les variants de chaque témoin, les fonctions de détection des variants de l'algorithme UMI-VarCall sont réutilisées sur chacun des témoins de sorte à déterminer une liste de variants à éliminer du *pileup*. On souhaite en effet que les données simulées ne reprennent pas les variants spécifiques de chaque témoin tels que les polymorphismes, mais uniquement le bruit de fond de séquençage. Après cette étape de filtration, l'estimation du bruit de fond est réalisée en convertissant les occurrences de chaque position du *pileup* en fréquence d'apparition. Cette matrice de fréquence sera la base pour créer le jeu de données simulées.

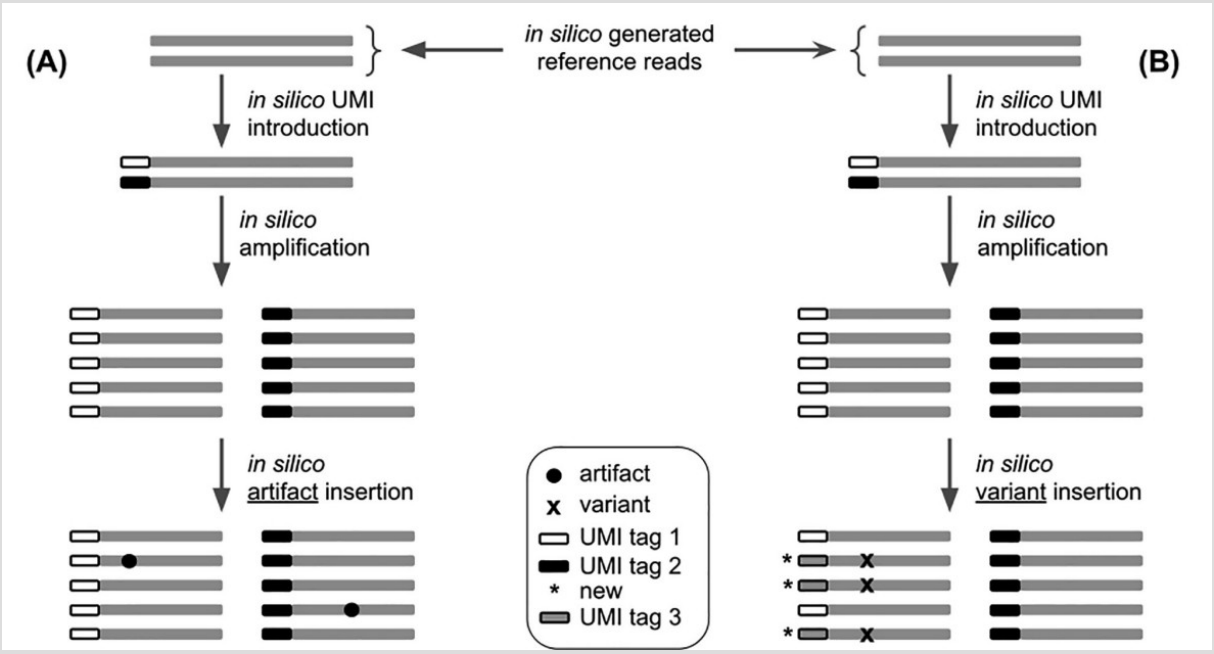


Figure 66: UMI-Gen - Introduction du bruit de fond de séquençage et des vrais positifs.

Les faux positifs sont introduits sans édition de l'UMI en (A) de sorte à reproduire les erreurs de séquence et de PCR observées dans les témoins. L'insertion des vrais positifs soumis par l'utilisateur est effectuée de sorte à générer des UMI concordants, c'est à dire des UMI pour lesquels l'ensemble des lectures provenant de chaque UMI supporte la présence du variant à la position du variant.

Selon la profondeur de séquençage et la longueur des séquences souhaitées, UMI-Gen réalise un fichier d'alignement dans lequel des lectures sont générées. Ces lectures à cette étape sont parfaitement identiques à la séquence du génome de référence. Un UMI est attaché à chacune de ces séquences selon le nombre d'UMI uniques soumis par l'utilisateur. Ces séquences comportant un UMI unique sont ensuite amplifiées jusqu'à atteindre la profondeur souhaitée à toutes les positions. Cette étape peut être réalisée automatiquement à partir des

fréquences alléliques des variants à insérer dans les données simulées si l'utilisateur n'a pas de prérequis.

Enfin, la dernière étape de l'algorithme (figure 66) vise à éditer les lectures générées. UMI-Gen édite, position par position, les lectures de sorte à respecter les fréquences observées de A, T, G, C, d'insertions et de délétions observées dans les échantillons témoins. Cette édition est réalisée aléatoirement sans tenir compte de l'UMI attaché à chaque lecture de sorte à reproduire les artefacts de PCR et d'amplification observés dans de vrais résultats de séquençage. Une fois cette étape d'édition réalisée pour insérer le bruit de fond de séquençage, la même étape est répétée pour insérer des vrais variants à des positions souhaitées. L'insertion de ces vrais positifs est réalisée en éditant cette fois-ci la séquence de l'UMI de sorte à ce que l'ensemble des lectures porteuses du même UMI intègrent ce vrai-positif.

UMI-Gen fournit à la fin de son exécution les fichiers FASTQ pairés, un fichier BAM aligné par BWA et l'index de ce fichier BAM. Des exemples d'application pour la comparaison des résultats de plusieurs algorithmes de détection de variants sont présentés dans l'article référencé en fin de chapitre.

D.3. Articles

Computational and Structural Biotechnology Journal 18 (2020) 2270–2280

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

UMI-Gen: A UMI-based read simulator for variant calling evaluation in paired-end sequencing NGS libraries

Vincent Sater [a,c,*,1], Pierre-Julien Viailly [b,c,1], Thierry Lecroq [a], Philippe Ruminy [b,c], Caroline Bérard [a], Élise Prieur-Gaston [a], Fabrice Jardin [b,c]

[a] University of Rouen Normandy UNIROUEN, LITIS EA 4108, 76000 Rouen, France
[b] Department of Pathology, Centre Henri Becquerel, 76000 Rouen, France
[c] INSERM U1245, University of Rouen Normandy UNIROUEN, 76000 Rouen, France

ARTICLE INFO

Article history:
Received 5 May 2020
Received in revised form 3 August 2020
Accepted 5 August 2020
Available online 27 August 2020

Keywords:
Sequence analysis
UMI
Simulator
Variant calling
NGS

ABSTRACT

Motivation: With Next Generation Sequencing becoming more affordable every year, NGS technologies asserted themselves as the fastest and most reliable way to detect Single Nucleotide Variants (SNV) and Copy Number Variations (CNV) in cancer patients. These technologies can be used to sequence DNA at very high depths thus allowing to detect abnormalities in tumor cells with very low frequencies. Multiple variant callers are publicly available and are usually efficient at calling out variants. However, when frequencies begin to drop under 1%, the specificity of these tools suffers greatly as true variants at very low frequencies can be easily confused with sequencing or PCR artifacts. The recent use of Unique Molecular Identifiers (UMI) in NGS experiments has offered a way to accurately separate true variants from artifacts. UMI-based variant callers are slowly replacing raw-read based variant callers as the standard method for an accurate detection of variants at very low frequencies. However, benchmarking done in the tools publication are usually realized on real biological data in which real variants are not known, making it difficult to assess their accuracy.

Results: We present UMI-Gen, a UMI-based read simulator for targeted sequencing paired-end data. UMI-Gen generates reference reads covering the targeted regions at a user customizable depth. After that, using a number of control files, it estimates the background error rate at each position and then modifies the generated reads to mimic real biological data. Finally, it will insert real variants in the reads from a list provided by the user.

Availability: The entire pipeline is available at https://gitlab.com/vincent-sater/umigen under MIT license.

1. Introduction

Nowadays, next generation sequencers such as Thermo Fisher or Illumina have become the standard go-to method for DNA sequencing. Prior to sequencing, DNA must be extracted and amplified by PCR in order to generate enough fragments to cover the wanted amplicons. After amplification, the sequencer handles the obtained fragments and generates their sequences in the form of reads. In most applications, especially ones that handle variant detection, the obtained reads must then be aligned to a reference genome in order to be used effectively. Today, cancer diagnosis is a very active area of research and one of its most important applications is the detection of Single Nucleotide Variants (SNV) in tumor cells. In fact, each cancer type has a specific profile of genetic mutations in specific genes. Therefore, establishing a precise profile of variants in a cancer patient allows to better understand the cancer evolution and customize the treatment according to the established profile.

Detecting and calling out variants in the aligned reads is done through a variant calling analysis. Generally, variant calling tools can detect mutational events such as substitutions, insertions and deletions very efficiently. However, at very low variant allele frequencies (VAFs) (under 1%), it becomes very challenging for raw-read-based variant callers to accurately call variants. In fact, PCR amplification and the sequencing step can introduce errors

* Corresponding author at: University of Rouen Normandy UNIROUEN, LITIS EA 4108, 76000 Rouen, France.
E-mail address: vincent.sater@gmail.com (V. Sater).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.csbj.2020.08.011

in the final reads. These errors are called artifacts and occur at very low VAFs which can lead to the confusion between them and true low-frequency variants. Multiple studies [1–5] have shown the effectiveness of using Unique Molecular Identifiers as a way to filter out PCR and sequencing artifacts. UMIs are short arbitrary oligonucleotide sequences that are attached to DNA fragments by ligation before the PCR amplification. By definition, the UMI tags must be random sequences so each fragment can have a unique short oligonucleotide sequence attached to it, giving each fragment a unique sequence tag. During the amplification, the UMI tags are amplified with their respective fragments. After sequencing, each UMI tag can be figured out from the reads. The idea behind using UMI tags in NGS experiments to filter out artifacts is explained in Fig. 1. In fact, if a variant is a true mutation, it means that it must have been present on the initial DNA fragment so when we tag the DNA fragment with a UMI, we are also tagging the mutation. The fragments that result from the amplification of that mutated DNA fragment must all be tagged by the same UMI tag and carry the same mutation (Fig. 1A). On the other hand, if the variant is a sequencing error, it means that the initial DNA fragment did not have the mutation in the first place and that it appeared later in the sequencing step. Therefore, during the amplification step, all the fragments resulting from the amplification of that DNA fragment should theoretically be tagged with the same UMI and should not present the mutation. The mutation will be produced later on, in the sequencing step, affecting only some reads but not all of them, thus creating discrepancies in the same UMI group (Fig. 1B).

With the growing number of variant calling tools, it has become hard to choose the right tool adapted to a certain experiment. Data simulation can play an important role for testing different tools on a dataset that we have control on, a control that we do not have on real biological data. At the moment, many short read simulators exist such as IntSIM [6] that can simulate somatic variants using HMM models trained on real sequencing genomes and SVSR [7] that is specifically designed to simulate datasets with structural variations and is compatible with multiple sequencing platforms. These tools are publicly available for researchers and allow them to test their algorithms on a simulated dataset in which variants are inserted at different frequencies and at different positions. The usage of the read simulators enable having a very accurate benchmarking of each variant calling tool ability. Surprisingly, no simulation software exists at the moment that let users generate reads with UMI tags. In this article, we present UMI-Gen, a UMI-based read simulator that can be used not only to test raw-read based variant callers but most importantly, UMI-based ones. UMI-Gen uses multiple real biological samples to estimate background error rate and base quality scores at each position. Then, it will introduce real variants in the final reads. To test our tool, we used 6 control samples and show exactly how our algorithm estimates the background error rate at each position. Then we give it a list of 15 variants at different positions and at different frequencies to introduce them in the final reads. Finally, we used 2 raw-read-based variant callers: SiNVICT[8] and OutLyzer [9] and two UMI-based variant callers: DeepSNVMiner [10] and UMI-VarCal [11] in order to compare the 4 tools performance and demonstrate that UMI-Gen correctly inserts the given variants at their respective positions and at the correct frequencies in a dataset that mimics perfectly what is seen in biological samples.

2. Materials and methods

2.1. Software input

UMI-Gen requires a minimum of three parameters at execution: a list of control BAM/SAM samples, the BED file with the coordinates of the targeted genomic regions and a reference genome FASTA file with BWA index files. In fact UMI-Gen is designed to work on targeted sequencing data only thus a BED file is always required. UMI-Gen can also accept a fourth optional file under the PILEUP format. In fact, when running UMI-Gen on control samples, a PILEUP file is automatically produced. This file contains the A, C, G and T average counts at each position for all the control samples. This file can be given to UMI-Gen at execution time and will allow the software to reload the pileup generated during the last analysis instead of regenerating it. This will allow the user to gain some significant time since the pileup generation is the most time-consuming step.

2.1.1. Control samples

Control samples are BAM/SAM files that are obtained by sequencing healthy individuals and normally should not contain any somatic variant. UMI-Gen can accept input files in BAM and SAM formats. A pileup is performed on each sample and a final average pileup is generated from the counts of all control samples.

2.1.2. Variant file

This file contains a list of the variants the user wishes to insert in the simulated reads. These are the only variants that should be reported in the variant callers VCF file during variant calling benchmarks. The variant file is a Comma Separated Values (CSV) file that contains 2 columns: the first column contains the variant ID with the HGVS nomenclature and the second column being the variant's desired frequency. UMI-Gen will then go to each position and insert these variants in order to produce final reads.

2.2. Generating the final pileup

2.2.1. Pileup

The first step of the workflow (Fig. 2) consists of generating the final pileup. For each control sample, our pileup algorithm will count the occurrences of each A, C, G and T. The counts will be stored for each position of the BED file as well as the average quality of the position and its depth. This is basically the same algorithm that is used by UMI-based variant caller UMI-VarCal that has been reintegrated in this tool for its high efficiency in treating reads with UMI tags. When all the pileups for all the control samples are ready, they will be merged in a final pileup that contains the average statistics (counts, depth and quality score) at each position based on the observations on all control samples (Fig. 2A). When the average pileup is complete and ready, it will be automatically dumped as a PILEUP file that contains all the calculated information on the set of control samples. If the user wishes to generate simulated data based on the same BED file and the same set of control samples, the dumped pileup can be used directly which allows the program to skip the pileup generation step and go directly to the variant calling step, saving the user much significant time.

2.2.2. Variant calling

Even though the control samples are theoretically variant-free, SNP and undetected mutations could still be present in the files. These potential variants must be removed so they would not be present in the final reads. To do so, we used the same variant calling method implemented in UMI-VarCal to call out potential variants and remove them from the pileup. This step will produce what we call a filtered pileup (Fig. 2B).

2.2.3. Background noise estimation

The background noise estimation step consists of calculating the frequency of observing an A/C/G/T at each position. Without the background errors, at each position the reference base should

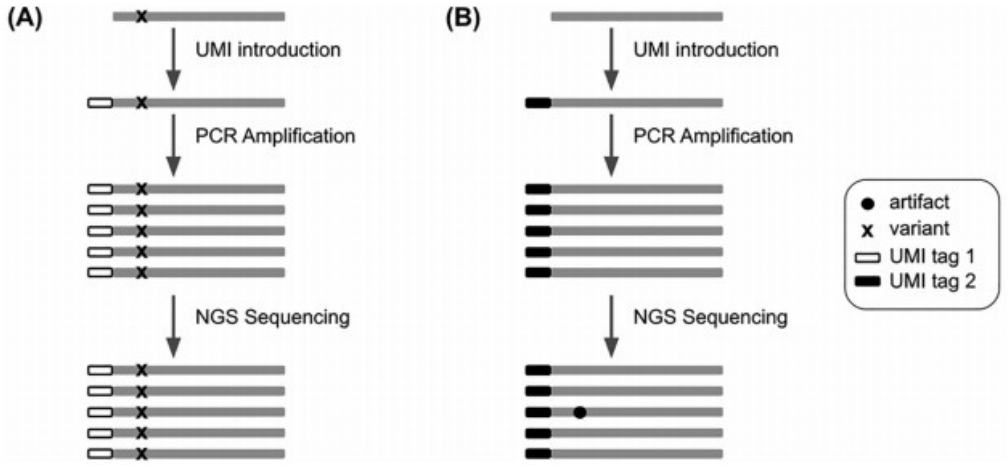


Fig. 1. The difference between a true variant and an artifact from a UMI perspective. (A) A true variant is present on the DNA fragment so when the UMI tag 1 is added, it tags the fragment and the mutation as well. After amplification, all the fragments tagged with the UMI tag 1 carry the same mutation. (B) An artifact is not present on the DNA fragment but rather appears at the steps that follow the UMI introduction. That is why not all fragments with the same UMI tag 2 carry the same artifact.

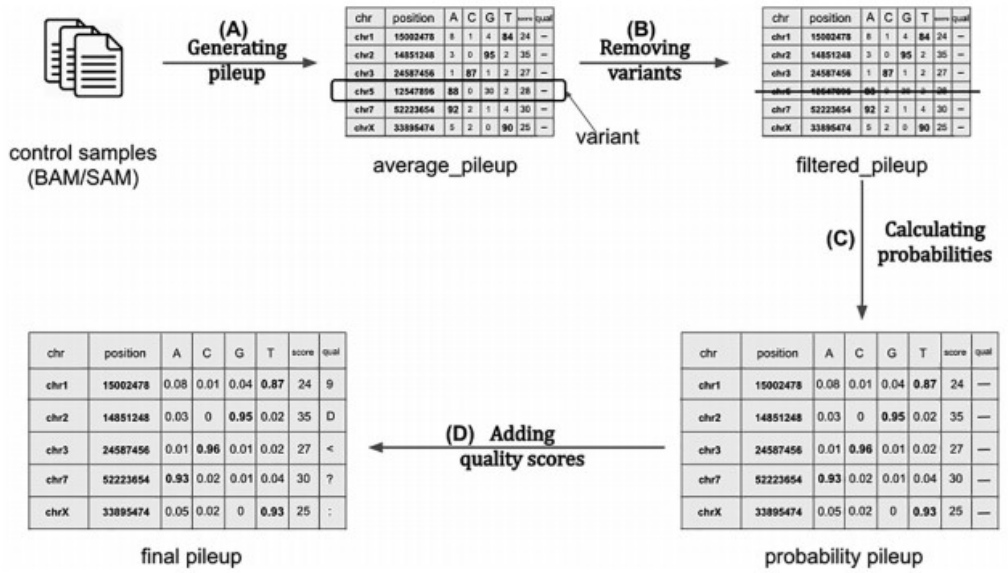


Fig. 2. Background error estimation workflow. (A) The first step runs over every position in all control samples and counts the total occurrences of every A, C, G and T. It also stores the average base quality score for each position. (B) The second step's goal is to remove any suspected variant from the pileup as our objective is to estimate background error noise only. (C) In this step, the counts are converted to probabilities by dividing them by the depth for each position. (D) The final step consists of converting the base quality score of each position to the corresponding ASCII + 33 character.

have a frequency of 1 while the remaining three bases should be at 0. The total of the four frequencies must be equal to 1. However, we know that artifacts exist in our control samples and these artifacts represent the background noise that we normally encounter in a normal NGS experiment. Since our aim is to simulate reads that are highly similar to those produced with real sequencing experiments, UMI-Gen calculates the real base frequencies from the control samples at each position. The frequencies will then be used as a probability matrix when producing the final reads. When this step is complete, a probability pileup is generated (Fig. 2C). Insertions and deletions are not considered during the background noise estimation and thus, are not present in the final pileup as their occurrence has a much lower rate (~1000 times lower) than that of substitutions) especially in second and third generation sequencers [12]. Therefore, we judge that their inclusion is not worth complicating the algorithm for.

2.2.4. Quality scores estimation

Our tool was developed on sequencing files produced by an Illumina sequencer. In the FASTQ files produced by Illumina sequencers, quality scores are encoded into a compact form, which uses only 1 byte per quality value [13,14]. The full table of encoding is available in Table S1. UMI-Gen is therefore only compatible with sequencers that use the same encoding. UMI-Gen calculates the average quality score for each position based on the qualities in all control samples and then converts the quality score to the corresponding ASCII character to be inserted in the final FASTQ file. This is the final step of the pileup generation workflow and will produce the final pileup (Fig. 2D). Moreover, UMI-Gen also models the base quality scores per position in read on the control samples and introduces the estimation in the final reads. Based on all the reads in the control samples, our tool will calculate a median base quality score for each position in the reads to produce a quality per position matrix. This matrix is then used at the end to recalibrate the quality scores according to each base's position in the read. For example, this allows UMI-Gen to mimic the loss of quality at the end of the reads when present.

2.3. Producing the reads

The main objective of UMI-Gen is to generate paired-end reads that mimic reads obtained from real life experiments. To do so, it starts exactly the way a real-life sequencing experiment starts: getting the DNA fragments. At the beginning, our tool will generate a number of initial sequences that only present the reference base at each position.

The user can explicitly specify the desired length for all the reads at execution. It should be noted that the algorithm will only create reads that will exactly align on the specified positions from the BED file so off-target amplification is not considered. Then, a UMI tag is attached to each initial sequence. Depending on the amplification factor and the desired depth chosen by the user, the algorithm will keep amplifying the initial sequences until the desired depth is reached at all positions. In fact, at this step, default values for the amplification factor and initial DNA fragments are automatically calculated in order to ensure optimal performance of the tool. We do so by analyzing the depth chosen by the user and the VAFs of the variants that he wishes to introduce. Using these numbers, we calculate the minimum number of initial DNA fragments needed for the true variant insertion. Even though this will ensure optimal performance, the user is free to change these parameters as long as they are mathematically allowed. Once we have the reference reads, the second step consists of adding the background noise (refer to Section 2.2.3) to these reads (Fig. 3A). Using the probability matrix calculated before, UMI-Gen modifies the reads at each position for them to match the calculated probabilities. These modifications are done without changing the reads' UMI so they mimic PCR and sequencing artifacts: they are false positives and should not be called by variant callers. Finally, UMI-Gen parses the variant file provided by the user in order to insert true mutations in the final reads. The algorithm will go to each position, change the probability of the variant to the corresponding frequency from the variant file. In this step, since UMI-Gen is adding a true variant, the UMI tags of the modified reads are also modified in order to produce concordant UMI tags (Fig. 3B). A concordant UMI tag is a UMI whose all reads carry the exact same mutation. Also, since UMI-Gen generates paired-end data, when adding a mutation on one read, the variant is automatically added to its mate (since we only generate paired reads that always overlap).

2.4. Software output

Once all variants are inserted, UMI-Gen will generate the two FASTQ files (R1 and R2). It will then call BWA [15] to do the alignment, a step that will produce a BAM file. SAMtools [16] is finally called to create the BAM's index file and convert the BAM into SAM. All five files are generated in the desired output directory. In addition, UMI-Gen generates a binary PILEUP file that

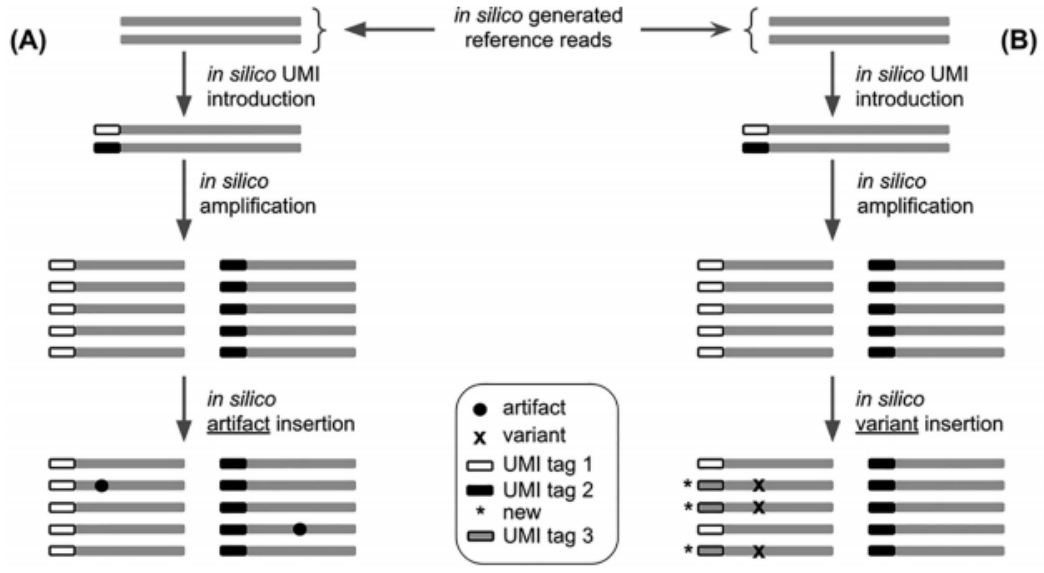


Fig. 3. The difference between adding a true variant and adding an artifact in generated reads. (A) Adding an artifact is relatively easy as all the tool has to do is to modify the base at the wanted position without touching the read's UMI tag. (B) On the other hand, in order to add a true variant, the software must change the base at the wanted position on a set of reads. Then it will create a new UMI tag (UMI tag 3) and change the UMI tag of all the affected reads to UMI tag 3.

corresponds to the dumped average pileup. This file can be used to skip the pileup regeneration and load the pileup directly if the analysis was already done on the same control samples.

2.5. Implementation

Launching UMI-Gen's workflow (Fig. 4) is handled by a main Python script that controls many Python3 modules. In order to achieve better overall performance, Cython was used to compile all Python modules. UMI-Gen requires for the tools BWA and SAMtools to be installed on the PC/server: BWA is called for the alignment step and SAMtools for converting, sorting and indexing the generated BAM files. Our tool can be executed through a UNIX/Linux command line interface. In total, UMI-Gen can accept 20 parameters at execution. Managing these parameters allows the user to have full control over his simulated data. A list of all the parameters and thresholds is available in Table S2.

3. Results

3.1. Control samples

A targeted sequencing panel was designed at the Centre Henri Becquerel in Rouen (France) to search for specific mutations in the DNA of patients suffering from Diffuse Large B cell Lymphoma (DLBCL). This panel of 76,630 bases is designed to identify genomic abnormalities within a list of 36 genes that are most commonly impacted in this type of lymphoma. The panel was specifically designed for QIAseq chemistry allowing UMI introduction in the DNA fragments during the construction of the library. A list of the genes used in the panel and their corresponding number of targeted regions is provided in the supplementary Table S3. In order to test our tool's ability to mimic and reproduce average sequencer background noise in the produced sample, we randomly selected 6 samples from a very large number of patients whose DNA were sequenced at the Centre Henri Becquerel. All six samples are liquid biopsies with circulating cell-free DNA that was checked to be adequate for sequencing. We preferred the use of liquid biopsies as these samples usually contain a high number of very low frequency variants and artifacts. Using such samples as control samples will produce simulated data with a relatively high number of artifacts. This will allow us to have an accurate estimate of the specificity of each tested variant caller.

Table 1 shows the exact counts of A, C, G, T for position 2,493,165 on chromosome 1 for each control sample. The fist control sample counts (0,11,10,874), the second sample has (0,1,7,843), the third one has (0,2,2,860), the fourth sample shows (1,6,9,965), the fifth one has (1,2,4,867) and the final one counts (3,2,2,880). As explained in Section 2.2.3, UMI-Gen will calculate

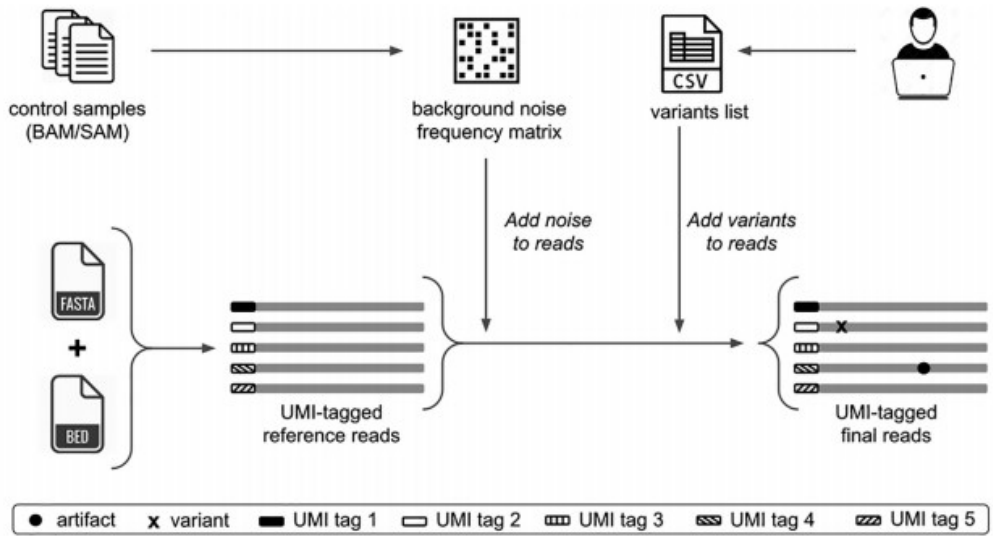


Fig. 4. UMI-Gen's workflow: Control samples are used to create a background noise frequency matrix and the user provides a CSV file with a list of the wanted variants. Using the FASTA and the BED files, UMI-Gen creates a first set of UMI-tagged reference reads. Artifacts are then inserted to mimic the sequencer's background noise. Finally, the tool uses the list provided by the user to insert variants at their exact locations.

Table 1
The A, C, G and T breakdown at position 2,493,165 of chromosome 1 for the six control samples.

Sample	A	C	G	T
Control 1	0	11	10	874
Control 2	0	1	7	843
Control 3	0	2	2	860
Control 4	0	6	9	965
Control 5	1	2	4	867
Control 6	3	2	2	880

an average count for each base and then estimate its probability. In our case and for this position, the obtained average count has 4 A, 24 C, 34 G and 5289 T with a total count of 5351 bases. To obtain the probabilities for this position on this chromosome, we simply divide each base count by the total count of the 4 bases, obtaining the final probability vector (0.0007, 0.0045, 0.0064, 0.9884). If, for example, we wanted to produce a BAM file with a depth of 3000x, this position would have 2 A, 14 C, 19 G and 2965 T. The probability matrix mentioned in Section 2.2.3 is basically the probability vectors of each position of the panel, merged together. In our test and in order to demonstrate our results, we simulated two artificial samples in which we added the calculated background error noise. The first sample or Sample 1 has an average depth of 1000× (+/− 15% at each position) and Sample 2 has an average depth of 10,000×. To make sure that the artifacts were correctly added to the reads, we used IGV (version 2.4.16) [17] to visualize the reads. Fig. 5 shows how the background error noise is properly and very accurately added at position 2,493,165 of chromosome 1 with the probabilities calculated from the 6 control samples above.

3.2. Simulated data validation

In order to validate our simulated dataset, we compared it to the control samples used to generate it. First, we compared the base quality scores distribution in the reads. Fig. 6A shows the variation of the median base quality scores with the position of base in the read for the control samples. We can clearly see that the median score is very high and very stable at the start and all along the read's length (≥34). However, a first drop in quality is noted at position 138 and a second more considerable one at position 145. In our simulated data, we chose an average length for the reads of about 110 bp so the longest read had a length of 127. We can see, in Fig. 6B, how the algorithm perfectly recreates the stability of the scores all along the simulated reads. However, since the simulated reads did not have lengths >135 bp, we do not see that little drop at the end of the simulated reads. In fact, to be sure that our quality score estimation works correctly, we simulated a drop in quality at the position 85 and wanted to see if it will be inserted in the simulated reads. Fig. 7 shows how the simulated drop in quality (38 → 34) at position 85 was perfectly reproduced in the simulated data (36 → 33). Another parameter we wanted to verify is the %GC variation between the control and the simulated data. Fig. 8 clearly shows how the median %GC of reads in the control data (Fig. 8A – 56% GC) is nearly identical to that of the simulated reads (Fig. 8B – 57% GC).

3.3. Inserted variants

Two different lists of mutations were created to go along with each simulated sample. The first list contains 11 substitution variants with frequencies that go from 0.9 (90%) to 0.01 (1%), one deletion at 1% and one insertion at 1%. This list is used to produce the simulated Sample 1 with a depth of 1000x. The second list contains 13 substitution variants with frequencies that go from 0.9 (90%) to 0.001 (0.1%), one deletion at 1% and one insertion at 1%. This list is used to produce the simulated Sample 2 with a depth of 10,000x. Two very low frequency variants (frequency <1%) were added to the second list to test the variant insertion accuracy of UMI-Gen. In fact, very low frequency variants are the hardest to detect and should be systematically used to rigorously test any variant caller. In order to verify that the wanted variants were added at the exact locations with the correct frequencies, we used IGV to visualize the reads. Fig. 9 shows the variants added in both samples and Table 2 details the exact variants that we inserted at the specific locations. Next generation sequencers have difficulties with accurately detecting variants in long homopolymer regions. Some variant callers automatically filter out variants that occur in such regions and others do not. In order to avoid any bias, we chose each variant's location carefully to make sure that it is not inserted in a homopolymer region. Fig. 10 demonstrates that our tool is capable of accurately adding variants in the final reads at the specified locations for both samples.

3.4. Variant detection

We tested the ability of four different variant callers to correctly detect the true variants added in Section 3.3 and filter out sequencing errors/artifacts added in 3.1. We used SiNVICT and OutLyzer, two raw-read-based variant callers specifically developed to detect low frequency variants and two UMI-based variant callers (DeepSNVMiner and UMI-VarCal) with a very low frequency detection threshold and that analyze UMI tags in order to produce more accurate results. The four variant callers were tested on the two artificial samples: Sample 1 that contains 13 known variants and a depth of 1000× and Sample 2 that contains 15 known variants and a depth of 10,000×.

Both samples have a total of 76,630 sequenced positions which corresponds to the size of the sequencing panel. Tables 3 and 4 detail the results of each tool for Sample 1 and 2 respectively. The total number of positives corresponds to the number of variants found in the result VCF file. The total number of negatives is then calculated by subtracting total positives from the total

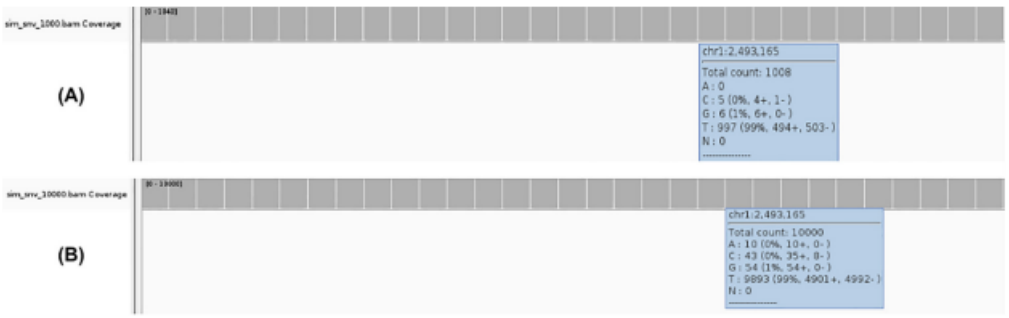


Fig. 5. The A, C, G and T breakdown at the position 2,493,165 of the chromosome 1 in the produced samples: Sample 1 with a depth of 1000× (A) and Sample 2 with the depth of 10,000× (B).

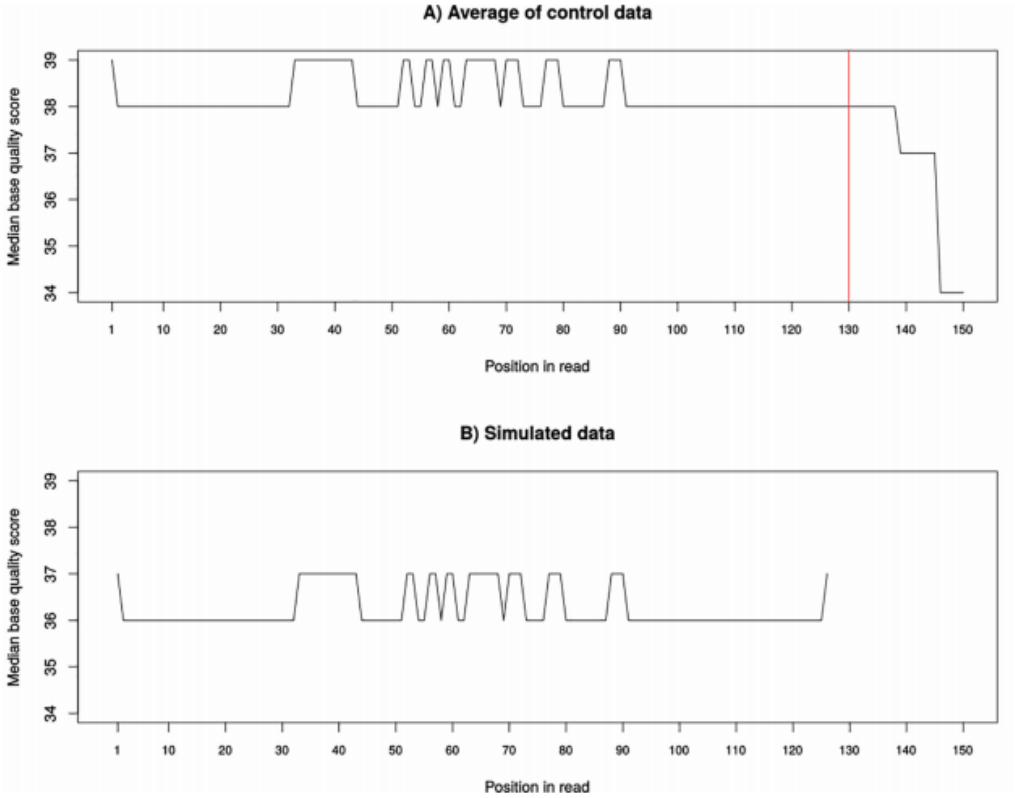


Fig. 6. The variation of the median base quality score with position in read in real samples (A) and in the simulated data (B).

number of positions (76,630). The four variant callers had comparable results between the two samples. Starting with SiNVICT, it detected 241 variants in Sample 1 and 463 in Sample 2 but with the same number of true positives. This corresponds to a sensitivity of 61.5%/53.4% which is relatively acceptable and a specificity of 99.7%/99.4% on Sample 1/2. Moving on to OutLyzer, the tool detected 109 variants in Sample 1 and three times more variants in Sample 2 (342). Unfortunately, this corresponded to one more true positive, the rest being only false positives. Outlyzer scored good sensitivities (>80%) and excellent specificities (99.9%/99.6%) on both samples.

Concerning DeepSNVMiner, the tool managed to detect all the inserted variants except the deletion in both samples. The tool scored very high scores on sensitivity (92.3%/93.4%) as well as specificity (99.95%/99.99%) for both datasets. Finally, UMI-VarCal was able to achieve a perfect score (100%) in terms of sensitivity and specificity on both samples detecting all the 13/15 variants in Sample 1/2 with no false positives for both configurations.

3.5. Performance

In order to evaluate UMI-Gen's performance, we simulated four samples with increasing depths: 500, 1000, 5000 and 10,000. For each simulated sample, execution time and memory consumption were reported. The four samples were simulated using the same six control samples. The first time we run UMI-Gen, the pileup generation step is mandatory. The pileup generation step only depends on the control samples and takes about 1.5 min per sample. The quality estimation step following the pileup is also essential and takes on average 0.5 min per sample. However, these 2 steps generate files that can be given directly to the program at the execution. This means that for the other times the user wants to simulate data using the same control samples, the pileup file and the quality matrix file can be used directly allowing to save considerable time. Table 5 details the execution time numbers and the memory needed to generate each sample. Generating the FASTQ files takes only 1.57 min for the 500× sample and uses only 1 GB of RAM. On the other side, 16.58 min are needed for a sample of 10,000× and memory consumption goes up to 5.1 GB. All these tests were performed on a computer running Linux (Ubuntu 16.04) using only one core CPU running at 2.20 GHz and equipped with 16 GB of RAM. All measurements were done three times and the average was used for the comparison. After the FASTQ generation, BWA and SAMtools are called from within the tool to generate the corresponding BAM and SAM files.

4. Discussion

Tagging DNA fragments with UMI tags have proved itself as a very reliable method to significantly reduce – if not completely remove – the number of false positives upon variant calling. A huge number of variant callers are publicly available at the moment but unfortunately, only 4 of them are specifically developed to treat UMI tags in reads. For raw-read-based variant callers, a lot of

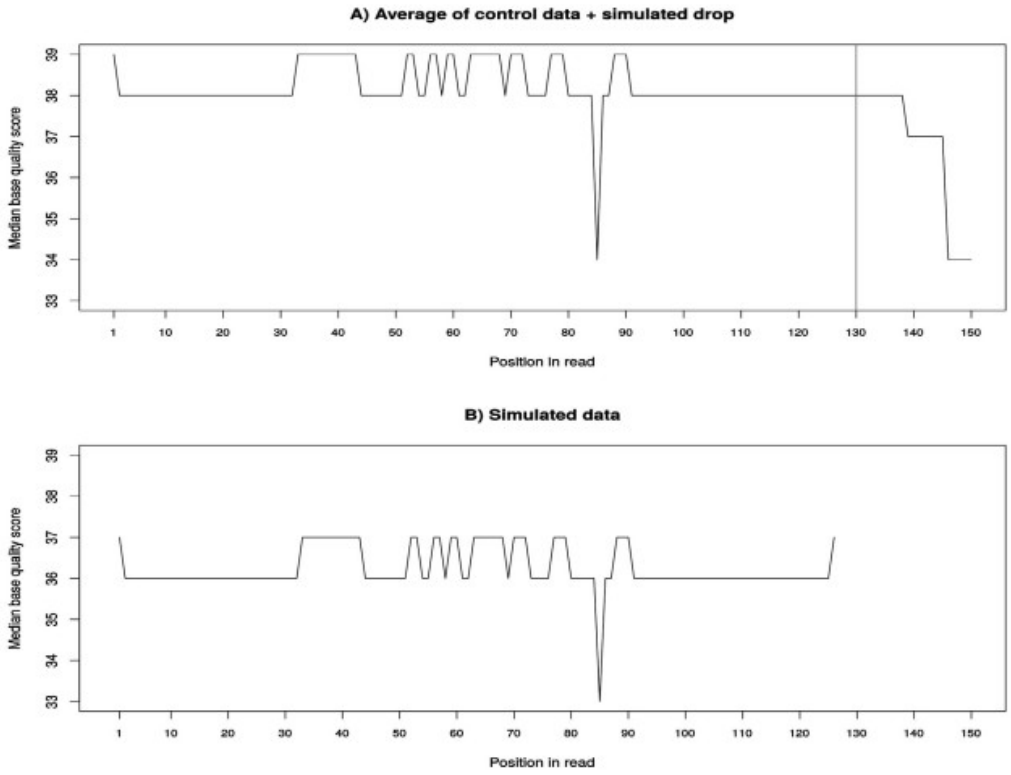


Fig. 7. The variation of the median base quality score with position in read in real samples (A) and in the simulated data (B). A simulated drop in quality was simulated in scenario A and its reproduction in the simulated dataset (B).

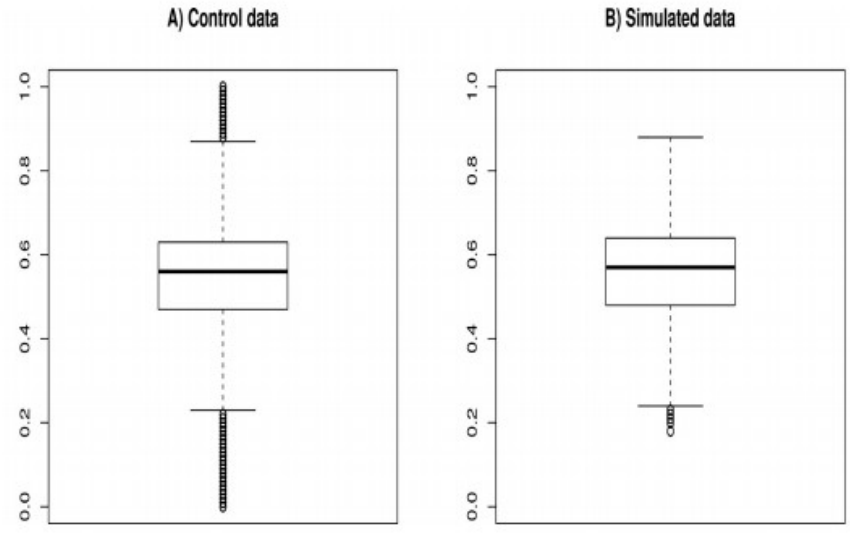
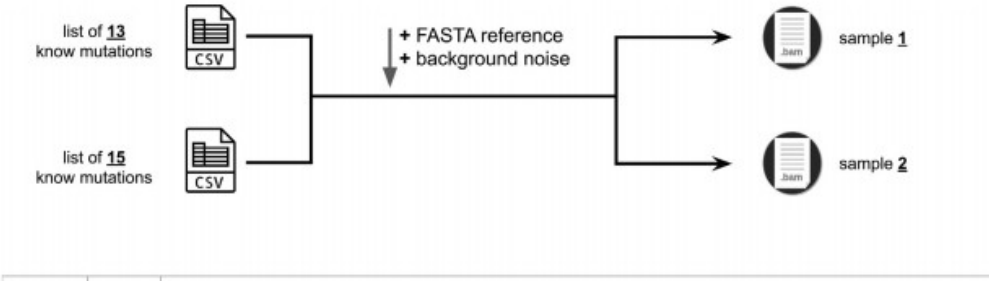


Fig 8. The repartition of the %GC in reads in the real data (A) and in the simulated data (B).



sample	depth	inserted mutations frequencies												
		0.90	0.80	0.70	0.60	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.005	0.001
sample 1	1000	1	1	1	1	1	1	1	1	3	1	1	X	X
sample 2	10 000	1	1	1	1	1	1	1	1	3	1	1	1	1

Figure 9. Along with the reference genome FASTA file and the BED file, two different lists were used, one with 13 variants and the other with 15 variants to respectively produce the artificial samples Sample 1 and Sample 2.

Table 2
Detailed list of the inserted mutations. In this test, all mutations are inserted on chromosome 1.

Position	Reference allele	Variant allele	Frequency	Sample
2,488,101	G	A	0.9	S1 & S2
2,489,200	C	A	0.8	S1 & S2
2,491,260	A	G	0.7	S1 & S2
2,493,201	T	A	0.6	S1 & S2
2,494,300	G	A	0.5	S1 & S2
23,885,600	C	A	0.4	S1 & S2
23,885,800	A	T	0.3	S1 & S2
27,022,900	C	A	0.2	S1 & S2
27,023,200	C	A	0.1	S1 & S2
27,093,001	G	A	0.05	S1 & S2
27,100,350	C	A	0.01	S1 & S2
27,106,500	G	A	0.005	S2 only
117,057,400	T	A	0.001	S2 only
120,458,000	C	CTA	0.1	S1 & S2
120,466,600	TGTC	T	0.1	S1 & S2

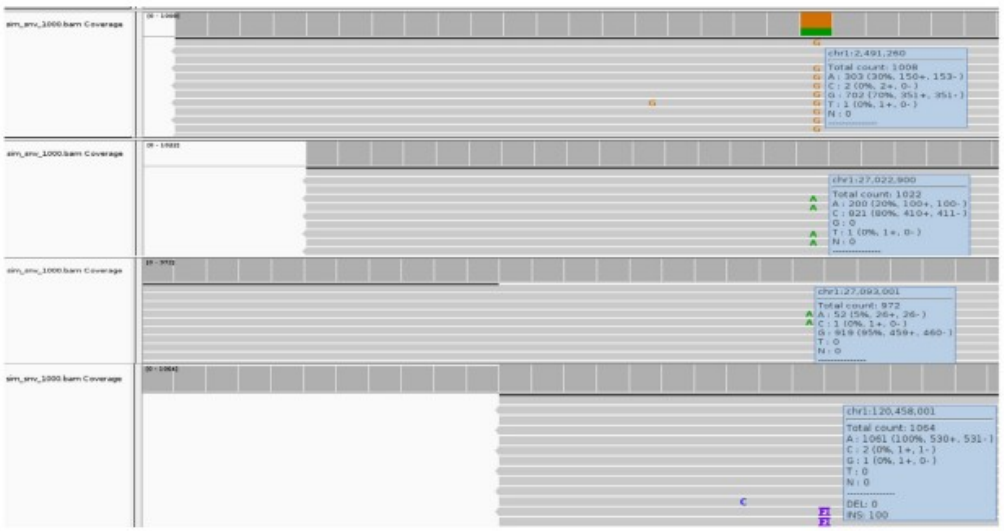


Figure 10. The inserted mutations were correctly added to the reads with their exact locations at their corresponding frequencies. Here, we see four mutations: chr1:2491260A > G at 70%, chr1:27022900C > A at 20%, chr1:120458000C > CTA at 10% and chr1:27093001G > A at 5%.

Table 3
Variant calling results on Sample 1. Four variant callers were tested: SiNVICT, OutLyzer, DeepSNVMiner and UMI-VarCal and for each tool, True Positives (TP), False Positives (FP), False Negatives (FN), sensitivity and specificity are reported.

Variant Caller	TP	FP	FN	Sensitivity (%)	Specificity (%)
SiNVICT	8	233	5	61.5	99.7
OutLyzer	11	98	2	84.6	99.9
DeepSNVMiner	12	37	1	92.3	99.95
UMI-VarCal	13	0	0	100	100

Table 4
Variant calling results on Sample 2. Four variant callers were tested: SiNVICT, OutLyzer, DeepSNVMiner and UMI-VarCal and for each tool, True Positives (TP), False Positives (FP), False Negatives (FN), sensitivity and specificity are reported.

Variant Caller	TP	FP	FN	Sensitivity (%)	Specificity (%)
SiNVICT	8	455	7	53.4	99.4
OutLyzer	12	330	3	80	99.6
DeepSNVMiner	14	2	1	93.4	99.99
UMI-VarCal	15	0	0	100	100

Table 5
Performance analysis of UMI-Gen: the variation of execution time and memory consumption with the simulated data's depth.

Sample	Data Simulation (min)	FASTQ to BAM (s)	RAM Usage (GB)
500×	1.57	8	1.0
1000×	1.87	14	1.1
5000×	6.97	52	2.6
10,000×	16.58	99	5.1

artificial read simulators exist and can satisfy everyone's needs. However, to our knowledge, no tool is publicly available to simulate artificial reads with UMI tags. Such a tool is very important as it allows developers to accurately test the specificity and the sensitivity of their variant callers on artificial reads in which real variants are known instead of testing them on biological samples whose mutational profile is completely or partially unknown.

Our main objective was to develop a UMI-based read simulator that is fast, accurate and reliable. UMI-Gen is able to estimate the background error noise of a given control dataset and then reproduce it accurately in the produced reads. Doing so, it allows to mimic the sequencer's background noise of a real sequencing experiment. We also showed that our simulator is able to accurately insert variants if provided with a list of variants with exact locations and their corresponding frequencies and produce reads that mimic ones produced in real life experiments. In our tests, we were able to insert mutations as low as 0.1% but theoretically, we can go as low as we want provided that the depth of the produced sample is accordingly increased.

Moreover, in our variant caller comparison, SiNVICT did a decent job detecting the 8 of the added variants and went as low as 5%. Impressively, we judge the performance of OutLyzer as excellent as it detected 12 of the 15 variants (Sample 2) and showed a detection threshold of 0.5% which is very respectable. However, SiNVICT and OutLyzer being raw-read-based variant callers, UMI tags were not treated in the reads and therefore, both tools produced a high percentage of false positives. On the other hand, DeepSNVMiner results were near perfect as expected from a decent UMI-based variant caller detecting all variants except one in both scenarios with only a couple of false positives. Finally, UMI-VarCal was successfully able to treat UMI tags allowing it to filter out all false positives and only call out the 13 added variants in Sample 1 and all of the 15 in Sample 2. These results demonstrate how the UMI-based variant calling approach is much more efficient and accurate than raw-read-based ones allowing to detect variants with VAFs as low as 0.1% without sacrificing specificity. It also highlights the need to the development and usage of UMI-based read simulators in order to test these new algorithms.

5. Conclusion

Here, we present UMI-Gen: a standalone UMI-based read simulator for variant calling evaluation in paired-end sequencing NGS libraries. UMI-Gen produces sequencing files (FASTQ, BAM and SAM) for an artificial sample to be used for UMI-based variant calling testing purposes. By using a set of control DNA samples, our tool is capable of accurately mimicking the background error noise of the sequencer and adding it into the reads. After that, it can insert specific mutations at specific locations and at very precise frequencies that can go as low as 0.1% (and even lower). In our tests, all added artifacts were correctly inserted in the reads, causing a high number of false positives in the raw-read-based variant callers results. Also, all inserted true variants were visualized with a genome visualization tool (IGV) and were detected by at least one of the four variant calling tools we tested. Finally, we note that UMI-Gen's filters and parameters (such as read length and UMI tag length) are customizable which gives the user total control over his produced samples. This level of customization allows the tool to be adequate for a high number of research applications.

Funding

This work was partly funded by the Université de Rouen Normandie and Vincent Sater is funded by a PhD fellowship from the Région Normandie.

CRediT authorship contribution statement

Vincent Sater: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Pierre-Julien Viailly:** Conceptualization, Methodology, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Thierry Lecroq:** Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Philippe Ruminy:** Methodology, Supervision, Project administration. **Caroline Bérard:** Methodology, Formal analysis. **Élise Prieur-Gaston:** Conceptualization, Supervision. **Fabrice Jardin:** Resources, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.08.011.

References

[1] Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 2012;109:14508–13. https://doi.org/10.1073/pnas.1208715109. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3437896/.

[2] Kukita Y, Matoba R, Uchida J, Hamakawa T, Doki Y, Imamura F, et al. High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients. DNA Res 2015;22:269–77. https://doi.org/10.1093/dnares/dsv010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4535617/.

[3] Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 2016;34:547–55. https://doi.org/10.1038/nbt.3520. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4907374/.

[4] Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat Commun 2016;7. https://doi.org/10.1038/ncomms12484. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4996934/.

[5] Bar DZ, Arlt MF, Brazier JF, Norris WE, Campbell SE, Chines P, et al. A novel somatic mutation achieves partial rescue in a child with Hutchinson-Gilford progeria syndrome. J Med Genet 2017;54:212–6. https://doi.org/10.1136/jmedgenet-2016-104295. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5384422/.

[6] Yuan X, Zhang J, Yang L. IntSIM: An integrated simulator of next-generation sequencing data. IEEE Trans Biomed Eng 2017;64:441–51. https://doi.org/10.1109/TBME.2016.2560939.

[7] Yuan X, Gao M, Bai J, Duan J. SVSR: A program to simulate structural variations and generate sequencing reads for multiple platforms. IEEE/ACM Trans Comput Biol Bioinf 2020;17:1082–91. https://doi.org/10.1109/TCBB.2018.2876527.

[8] Kockan C, Hach F, Sarrafi I, Bell RH, McConeghy B, Beja K, et al. SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. Bioinformatics 2017;33:26–34. https://doi.org/10.1093/bioinformatics/btw536.

[9] Muller E, Goardon N, Brault B, Rousselin A, Paimparay G, Legros A, et al. OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. Oncotarget 2016;7:79485–93. https://doi.org/10.18632/oncotarget.13103. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5346729/.

[10] Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ 2016;4. https://doi.org/10.7717/peerj.2074. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888318/.

[11] Sater V, Viailly P-J, Lecroq T, Prieur-Gaston E, Bohers E, Viennot M, et al. UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. Bioinformatics (Oxford, England) 2020. https://doi.org/10.1093/bioinformatics/btaa053.

[12] Schirmer M, Damore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinf 2016;17. https://doi.org/10.1186/s12859-016-0976-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4787001/.

[13] Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res 1998;8:186–94. https://doi.org/10.1101/gr.8.3.186. URL: http://genome.cshlp.org/content/8/3/186.

[14] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res 1998;8:175–85. https://doi.org/10.1101/gr.8.3.175. URL: http://genome.cshlp.org/content/8/3/175.

[15] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705234/.

[16] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/.

[17] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011;29:24–6. https://doi.org/10.1038/nbt.1754. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346182/.

IV. DISCUSSION ET PERSPECTIVES

Les objectifs de ce travail de thèse étaient le développement de nouveaux algorithmes bioinformatiques pour améliorer la détection des variations (SNV, CNV) à partir d'échantillons biologiques pauvres en contingent tumoral tels que le cfDNA. Nous avons cherché à améliorer à la fois cette détection sur le plan qualitatif, en cherchant à discriminer le bruit de fond technique du signal biologique avec plus d'efficacité, et sur le plan quantitatif via l'utilisation des UMI et des technologies de séquençage à haut-débit.

La première approche, appelée LowVarFreq, visait à éliminer les artefacts biologiques et techniques à partir de données de séquençage Ion Torrent n'intégrant pas de barcodes moléculaires. L'algorithme, en mesurant une liste d'artefacts à partir d'échantillons témoins afin de mieux les éliminer dans les échantillons testés, a permis d'interpréter les résultats de séquençage d'une série d'ADN extraits d'échantillons tumoraux et plasmatiques de lymphomes de Hodgkin. Néanmoins, si LowVarFreq est en capacité d'estimer le bruit de fond de séquençage par échantillon, il n'est pas en mesure de pouvoir le corriger afin d'accroître la sensibilité de détection en quantifiant des anomalies en dessous ce bruit de fond. De plus, LowVarFreq intègre de plusieurs algorithmes bioinformatiques de *variant calling* dont les sensibilités et spécificités sont difficilement quantifiables en l'absence d'échantillons de référence dont la liste des vraies mutations aurait été évaluée en amont. De ce fait, nous avons considéré que les vrai-positifs avaient une chance accrue d'être détectée par plusieurs algorithmes dans un même échantillon mais cette hypothèse est probablement très réductrice.

Nous nous sommes ensuite intéressés à l'introduction de barcodes moléculaires appelés UMI dans les librairies de séquençage. Ces barcodes moléculaires, qui permettent d'identifier les fragments d'ADN avant toute étape d'amplification, ont conduit au développement de deux nouveaux algorithmes : UMI-VarCal et mCNA.

UMI-VarCal est un algorithme de détection de variations ponctuelles (SNV, insertions et délétions) prenant en compte l'information portée par les UMI afin d'identifier les erreurs de séquence et de PCR. En quantifiant le nombre d'UMI uniques concordants et discordants par variation candidate, il permet d'améliorer la sensibilité et la spécificité de détection. La comparaison avec d'autres outils de la littérature a conduit à valider cette approche sur des fichiers *in silico* et sur des ADN extraits de tumeur. Néanmoins, cette approche reste limitée à la quantité d'ADN disponible pour la préparation de la librairie. En effet, pour une même

quantité de lectures séquencées, une baisse de la quantité d'ADN pour un échantillon en amont du séquençage va entraîner mécaniquement une augmentation de facteur d'amplification et de relecture des séquences de chaque UMI. Plus le facteur de relecture est important et plus la probabilité d'observer une erreur de PCR ou de séquençage dans les lectures d'un UMI unique augmente. En d'autres termes, la probabilité d'observer un UMI discordant est directement proportionnelle à ce facteur de relecture. Il est donc très important d'adapter la profondeur de séquençage d'un échantillon à la quantité de matériel disponible. UMI-VarCal est aujourd'hui utilisé au laboratoire de diagnostic du Centre Henri Becquerel à Rouen pour toutes les recherches de variants par séquençage à haut-débit. Parallèlement, l'équipe de recherche est entrain de valider l'approche sur plusieurs centaines d'échantillons de cfDNA de cohortes de lymphomes.

L'introduction des UMI a aussi permis de développer un nouvel algorithme de détection de CNV appelé mCNA. Cette approche bioinformatique vise à non plus utiliser la profondeur de séquençage afin de détecter les gains ou les pertes de matériel mais à quantifier le nombre d'UMI uniques par fenêtre glissante. Nous avons démontré la supériorité de l'approche en comparaison des algorithmes publiés dans la littérature à la fois sur des données simulées mais aussi sur des dilutions de lignées et par comparaison avec les résultats de CGH obtenus sur des ADN extraits de biopsies de DLBCL. Ces résultats nous ont permis d'estimer la sensibilité de l'approche entre 10 et 5 %. Néanmoins, les premiers résultats obtenus sur des échantillons de cfDNA sur les cohortes d'échantillons de cfDNA nous montrent que certains d'entre eux, même ayant des mutations détectées à une fréquence moyenne supérieure à ce seuil de 10 %, nous donnent des profils qui ne sont pas exploitables avec une variance importante de la mesure des log-ratios. Cela signifie qu'une part de la variance des comptages n'est pas corrigée par l'étape de normalisation dans certains échantillons de cfDNA. Nous ne sommes pour l'instant pas parvenu à établir les paramètres à l'origine de cette observation.

Enfin, nous avons développé dans le cadre de cette thèse le tout premier simulateur de données de séquençage à haut-débit intégrant des UMI appelé UMI-Gen. Ce simulateur, en permettant d'introduire des vrai-positifs dans des échantillons mimant le bruit de fond de séquençage d'échantillons témoins, permet une comparaison objective des résultats des différents outils de *variant calling* disponibles dans la littérature. Ce simulateur a permis de montrer la supériorité de l'approche UMI-VarCal pour la détection des variants de faible fréquence en terme de sensibilité et de spécificité. Les fichiers générés *in silico* vont permettre, à chaque mise à jour de nos chaînes de traitement bioinformatique, de vérifier si les

changements apportés n'induisent pas l'absence de détection de vrai-positifs. La stabilité des traitements informatiques des données de séquençage sera en effet très vraisemblablement un enjeu majeur pour les prochaines années tant sur le plan du diagnostic que pour la construction de projets de recherche multicentriques faisant intervenir du séquençage NGS.

A. Perspectives

A.1. Amélioration de la détection des anomalies dans le cfDNA pour le suivi de maladie résiduelle

La détection des anomalies de faible fréquence dans le cfDNA demeure un vaste sujet de recherche. Les perspectives d'amélioration sont nombreuses allant de l'amélioration des protocoles d'extraction de cfDNA jusqu'au traitement informatique des données, en passant par l'optimisation de la qualité de séquençage des séquenceurs de nouvelle génération.

Les approches sans UMI comme LowVarFreq ne semblent plus adaptées aujourd'hui aux applications d'analyse du ctDNA. Il est primordial d'introduire des UMI dans les protocoles de construction de librairies afin d'aider les algorithmes bioinformatiques à discriminer le signal biologique du bruit de fond de séquençage.

De nombreuses études ont limité la quantification du ctDNA à la détection des mutations somatiques sur les régions codantes des gènes. Cette approche est pertinente au diagnostic puisqu'elle permet de trouver des mutations récurrentes sur des gènes dont les fonctions biologiques sont connues et sont souvent associées à un sous-type de lymphome. Cependant, pour le suivi de maladie résiduelle, cette stratégie ne semble pas être la plus adaptée. En effet, en MRD, on souhaite suivre l'évolution de la quantité de ctDNA en cours de traitement afin de prévenir l'apparition de clones résistants au traitement. L'estimation de la concentration de ctDNA est calculée à partir des fréquences alléliques moyennes des mutations circulantes détectées. De ce fait, pour des tumeurs peu mutées, le nombre de biomarqueurs analysables pour le suivi peut devenir très restreint et donc prévenir le calcul de cette concentration. Pour des quantités faibles de ctDNA, la probabilité de détecter un événement chute et il est donc primordiale de pouvoir suivre un nombre important d'événements détectés au diagnostic.

Une nouvelle approche pour la détection de variants de phase (Phased variant, PV), appelée PhasED-Seq, semble très prometteuse [294]. Cette approche vise à non plus cibler les parties codantes des gènes mais les régions portant plusieurs mutations sur un même fragment d'ADN afin d'accroître à la fois la sensibilité et la spécificité de détection (figure 67). Ce ne sont plus cette fois-ci quelques mutations qui sont suivies en MRD mais plusieurs

combinaisons de mutations portées par des même fragments de ctDNA. Cette combinatoire est particulièrement intéressante en terme de spécificité. Là où quelques bases mutées retrouvées au suivi chez un patient sur une mutation ponctuelle présente au diagnostic est particulièrement difficile à interpréter, la présence de trois variants et plus sur un même fragment de cfDNA ne peut être expliquée par des d'artefacts de PCR ou de séquençage.

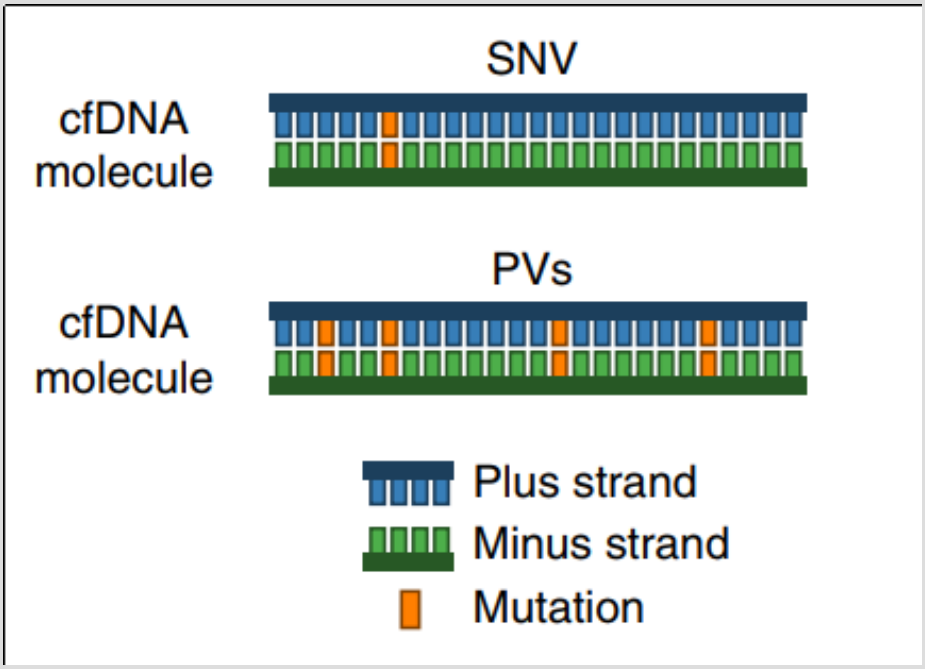


Figure 67: Représentation schématique des variants de phase.

Cette figure nous montre la représentation d'un SNV recherché dans des échantillons de cfDNA d'une part et un variant de phase (PV) d'autre part. Les PV sont caractérisés par la présence de plusieurs anomalies portées sur un même fragment de ctDNA. Adaptée de Kurtz et al. (2021) [294].

Les remaniements VDJ sont eux aussi des éléments intéressants à suivre en MRD. Nous avons vu dans ce manuscrit que la recherche de clonalité, c'est à dire l'identification du réarrangement VDJ du clone tumoral, est utilisée au diagnostic. La détection de ces réarrangements dans les échantillons de cfDNA au diagnostic et au suivi est donc une marque forte de la présence persistante du clone tumoral. Néanmoins, ces approches sont parfois difficiles à mettre en place dans la mesure où les fragments de ctDNA sont des séquences parfois trop courtes pour lire l'intégrité du CDR3 des immunoglobulines.

A.2. Harmonisation des pratiques d'analyse du ctDNA

Le développement des analyses du ctDNA pour une utilisation au diagnostic conduira nécessairement à l'harmonisation des pratiques d'extraction, de séquençage et d'analyse bioinformatique.

Une première étude clinique non-interventionnelle portée par le groupe coopérateur du LYSA (Lymphoma Study Association), appelée RT3, visait à dépeindre pour chaque patient atteint de lymphome un portrait de plusieurs caractéristiques moléculaires au diagnostic à partir de matériel biologique extrait de la tumeur en impliquant un réseau national de plateformes spécialisées. RT3 a permis une première harmonisation du rendu des résultats pour les patients avec une première centralisation des analyses bioinformatiques provenant de plusieurs centres hospitaliers.

Un nouvel essai appelé French Connect (French Cooperative Network for ctDNA in lymphoma), porté par le CALYM et piloté par le Pr MH Delfau-Larue, s'intéresse cette fois-ci à l'analyse du ctDNA avec 7 équipes multidisciplinaires (Créteil, Dijon, Lyon, Nantes, Rennes, Rouen et Toulouse). En impliquant à la fois des biologistes, des cliniciens, des pathologistes, des imageurs et des bioinformaticiens, French Connect a pour ambition d'harmoniser et d'optimiser l'analyse du ctDNA au sein des membres de l'Institut CALYM avec un panel de séquençage commun et un traitement centralisé des données de séquençage à haut-débit sur une plateforme bioinformatique commune. Les différents outils bioinformatiques développés dans le cadre de cette thèse, et plus particulièrement UMI-VarCal et mCNA, seront utilisés dans le cadre de cet essai afin d'évaluer leurs performances sur des données de vie réelle. Sur le plan clinique, un résultat concomitant de la réponse moléculaire du ctDNA et de la réponse métabolique en imagerie (PETSCAN) permettra une évaluation globale de la réponse au traitement. Ce projet sera l'occasion d'intégrer la détection des variants de phase à partir des données de séquençage ainsi que l'identification et le suivi du réarrangement VDJ du clone tumoral au diagnostic et en cours de traitement.

V. ANNEXES

A. Revue publiée (Pharmaceuticals 2021)

Suite aux différents travaux menés par l'unité INSERM U1245, une revue a été commandée à notre laboratoire par le journal *Pharmaceuticals* (IF 5.68).

Notre article, publié en 2021 [153], donne une vue d'ensemble des différentes étapes nécessaires à la détection des anomalies génétiques à partir d'échantillons de cfDNA depuis le pré-analytique jusqu'aux différents biais bioinformatiques inhérents à ce type d'approche. Il permet d'avoir une vue exhaustive des différentes technologies d'analyse du cfDNA et de remettre en perspective les différents travaux menés dans le cadre de cette thèse vis à vis des méthodes existantes.

pharmaceuticals

MDPI

Review

cfDNA Sequencing: Technological Approaches and Bioinformatic Issues

Elodie Bohers *, Pierre-Julien Viailly and Fabrice Jardin

INSERM U1245, Henri Becquerel Center, IRIB, Normandy University, 76000 Rouen, France; pierre-julien.viailly@chb.unicancer.fr (P.-J.V.); fabrice.jardin@chb.unicancer.fr (F.J.)
* Correspondence: elodie.bohers@chb.unicancer.fr

Abstract: In the era of precision medicine, it is crucial to identify molecular alterations that will guide the therapeutic management of patients. In this context, circulating tumoral DNA (ctDNA) released by the tumor in body fluids, like blood, and carrying its molecular characteristics is becoming a powerful biomarker for non-invasive detection and monitoring of cancer. Major recent technological advances, especially in terms of sequencing, have made possible its analysis, the challenge still being its reliable early detection. Different parameters, from the pre-analytical phase to the choice of sequencing technology and bioinformatic tools can influence the sensitivity of ctDNA detection.

Keywords: cell-free DNA; circulating tumoral DNA; sequencing technologies; bioinformatics

Citation: Bohers, E.; Viailly, P.-J.; Jardin, F. cfDNA Sequencing: Technological Approaches and Bioinformatic Issues. *Pharmaceuticals* **2021**, *14*, 596. https://doi.org/10.3390/ph14060596

Academic Editor: Gerald Reischl

Received: 14 May 2021
Accepted: 18 June 2021
Published: 21 June 2021

1. Introduction

Cell free circulating DNA (cfDNA) refers to DNA fragments present outside of cells in body fluids such as plasma, urine, and cerebrospinal fluid (CSF). CfDNA was first identified in 1948 from plasma of healthy individuals [1]. Afterward, studies showed that the quantity of this cfDNA in the blood was increased under pathological conditions such as auto-immune diseases [2] but also cancers [3]. In 1989, Philippe Anker and Maurice Stroun, from the University of Geneva, demonstrated that this cfDNA from cancer patients carries the characteristics of the DNA from tumoral cells [4]. Next, using the recently developed technique of PCR, David Sidransky and his team found the same mutations of *TP53* in bladder tumoral samples and urine pellets from patients [5]. Then, the research and identification of genomic anomalies specific of a cancer type in the circulating DNA, such as *NRAS* and *KRAS* mutations or *HER*-2 amplifications [6–8], started to expand, and for the first time, the term of circulating tumor DNA (ctDNA) appeared.

Since the highlighting of this circulating DNA of tumoral origin, technological developments in molecular biology, from quantitative and digital PCR to Next Generation Sequencing, turned it into a powerful liquid biopsy tool. At the era of precision medicine, it seems crucial to identify molecular alterations that will be able to guide the therapeutic management of patients. As tumors release DNA in the blood or other body fluids such as urine, this circulating tumoral DNA, containing the molecular characteristics of the tumor, can be collected with a simple body fluid sample. Since it is minimally invasive, this liquid biopsy is easily repeatable during follow up and in case of relapse. It is also of major interest in some particular cancers where a tumoral biopsy is difficult to obtain such as primary central nervous system lymphoma [9] or cancer subtypes with tissue biopsy containing very little tumoral cells such as Hodgkin lymphoma (HL) for which Reed–Sternbeg cells represent only 0.1 to 2% of the tumoral mass [10,11]. In these particular conditions and malignancies, the sequencing of ctDNA in body fluids could serve as a surrogate for a tumor biopsy. Other body fluids than blood are often used according to the localization of the tumor, such as urine for bladder cancers or cerebrospinal fluid for cerebral tumors [9,12] but blood is the body fluid most often used in studies.

In blood, average cfDNA concentration in healthy individuals can range between 0 and 100 ng/mL of plasma with an average of 30 ng/mL of plasma and is significantly higher in blood of cancer patients, varying between 0 and 1000 ng/mL, with an average of 180 ng/mL [13]. This concentration is correlated with the stage of the cancer, increasing with higher stages, and the size of the tumor. Circulating DNA of tumoral origin represents from 0.01 to more than 90% of the total cell free DNA found in blood [14]. In different types of cancers, a large scale ctDNA sequencing study has shown an association between ctDNA levels and mutational tumor burden [15]. Moreover, given the spatial heterogeneity observed in tumor tissue, ctDNA analysis can determine the complete molecular landscape of a patient's tumor and give supplementary information on drug targetable alterations and resistant variants [16]. ctDNA kinetics during follow up is correlated with prognosis, as a drastic reduction in its level after treatment is associated with better prognosis, whereas an increase usually means the evolution of drug resistant clones and an ultimate therapeutic failure [17–20].

Detection of ctDNA during MRD follow up to predict early relapse and at diagnosis in early stages of cancer continues to be a challenge, as the fraction of tumoral DNA contents in total circulating DNA may be <0.01% [21,22]. The development of sequencing technologies being more and more sensitive allows the detection of alterations present in cfDNA at very low variant allele frequencies (VAF), not only for mutational profiling at diagnosis but also for the early detection of disease recurrence and monitoring for therapy response. However, several parameters can affect the sensitivity of ctDNA detection. First, adequate handling of the blood sample, from blood collection to the quality control of the cfDNA extracted, is crucial in analysis. Next, an important step is the choice of the biomarker (s) and the sequencing technology used to detect it. Then, bioinformatic analysis, using error suppression algorithms, is the ultimate tool to discriminate the true variant from false positives.

2. Pre-Analytical Requirements

Some pre-analytical parameters can affect the sensitivity of ctDNA detection, which strongly relies on input material quantity and quality. Precautions have to be taken to limit degradation of cfDNA and contamination with genomic DNA. Ruptured blood cells were described to be a main source of cfDNA contamination, but it can be in part avoidable by improved pre-analytical processing. The nature of the anticoagulant or stabilizing agent contained in the blood collection tube, the volume of plasma, the preservation, as well as the cfDNA extraction kit, are key elements of the pre-analytical phase, conditioning the accuracy and limit of detection of the analysis [23]. The main steps for blood sample processing are represented in Figure 1.

For optimal extraction of cfDNA from plasma samples, it is recommended to use blood collected into sample collection tubes that provide efficient stabilization of plasma. Several studies have compared different blood collection tubes (BCTs), especially conventional anticoagulant EDTA tubes and as well as long-term storage BCTs from four different manufacturers (such as Streck (cfDNA BCT), Roche Diagnostics (Cell-Free DNA Collection Tube), Qiagen (PAXgene Blood ccfDNA Tube), and Norgen Biotek Corp. (cf-DNA/cf-RNA Preservative tubes)). These last BCTs are pre-coated with preservatives to prevent cell lysis and, therefore, reduce the release of RNA and DNA from hematopoietic cells. All these studies concluded that time between sampling and first centrifugation is a major point when using EDTA collection tubes, and that this time should not exceed 4 h. Whereas, the other BCTs, containing the stabilizing agent, can be stored at room temperature several days without affecting the further analytical performances (up to 14 days recommended but until 3 days for better results). However, despite the presence of stabilizing agents, the ambient temperature must be respected to avoid contamination with normal genomic DNA [24–27]. Therefore, EDTA tubes are suitable for internal analysis or monocentric studies, but if blood has to be shipped for external analysis, specialized long-term BCTs are more convenient.

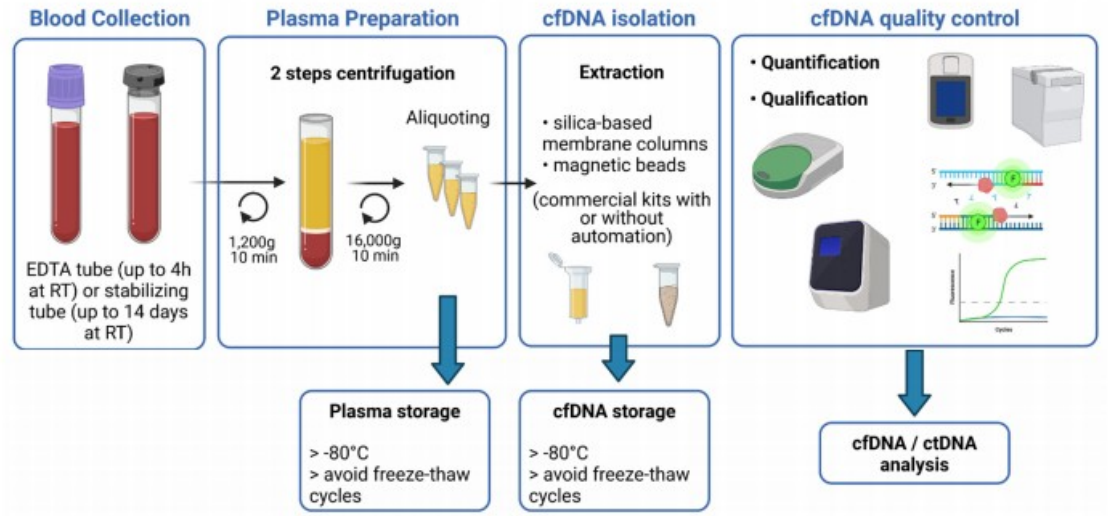


Figure 1. Schematic overview of the main steps for blood sample processing and cfDNA extraction. Blood, collected in EDTA or stabilizing tubes, goes through two rounds of centrifugation to obtain plasma samples. CfDNA is isolated from plasma using commercial kits and is quantified and qualified for further analysis.

Once sampling, blood collection tubes have to be centrifuged for plasma separation. This step can also affect cfDNA concentration and several studies set out to determine the best centrifugation protocol. The two-step centrifugation protocols turned out to be the suitable ones to prevent unwanted release of genomic DNA. Blood cells first have to be removed by slow centrifugation (1200–2000× *g* for 10 min at +4 °C or RT) in order to avoid cell lysis. Whereas afterwards, cellular debris and fragments will be removed by short-term high-speed microcentrifugation of the plasma supernatant (12,000–16,000× *g* for 10 min at +4 °C or RT), either before or after a freeze–thaw cycle [28,29]. The most crucial step is to not disturb the buffy coat while collecting the plasma after the first spin. Plasma samples should then be aliquoted to avoid repeated freeze–thaw cycles and kept at −80 °C for long-term storage.

Yield of cfDNA can also differ according to the extraction kit. Various commercial purification kits have been tested, in particular kits from Qiagen (QIAamp circulating nucleic acid kit and QIAamp min Elute ccfDNA mini kit), Promega (Maxwell RSC ccfDNA plasma kit), Applied Biosystems (Mag MAX cell-free DNA isolation kit), and Norgen Biotek (plasma/serum cell-free circulating DNA Purification midi kit). These kits work either with columns or magnetic beads and some are or can be automated [30,31]. All studies agreed to conclude that Qiagen kits, with or without automation, give the best performances. Once extracted, cfDNA should be stored at −80 °C.

Several studies emphasized the importance to perform consistent quality controls (QC) on the isolated circulating DNA. cfDNA is released through apoptosis and necrosis of normal and malignant cells and is highly fragmented [32]. Its size ranged between 20 and 220 bp with a maximum peak at 167 bp, corresponding to the length of DNA wrapped around a single nucleosome [33]. The use of fluorimetric methods is not suitable to accurately quantify cfDNA as it will not discriminate cfDNA from contaminating genomic DNA (gDNA). Contrariwise, capillary electrophoresis can measure the size of DNA fragments and give an estimation of the absolute concentration of cfDNA [34,35], but will not evaluate the presence of impurities that could inhibit downstream enzymatic reactions. QPCR-based and ddPCR methods can evaluate amplificability of cfDNA, as well as concentration and integrity, but are negatively impacted by gDNA contamination, through distorting the ratio between short and long amplicons [36]. Recently, Alcaide and

colleagues developed a promising multiplex ddPCR single-well assay, which can evaluate the quantity, quality, and fragment size distribution of cfDNA samples using low inputs and without the need of reference samples and calibration curves. This assay targets at the same time several olfactory receptor genes, representing three fragment size ranges, and a customizable control diploid locus. Unfortunately, the determination of cfDNA yields can still be affected by gDNA contamination and by copy number alterations [37].

Despite recent promising progresses, the pre-analytical process of blood samples still need standardization and further investigations to improve quality controls of the cfDNA that will be used to detect circulating DNA of tumoral origin.

3. Detection of ctDNA by Sequencing Technologies

Sequencing technologies for detection and analysis of the ctDNA range from point mutations analyses using PCR-based methods to analyses of whole genome using NGS based methods. The choice of the method employed depends on the application and the sensitivity intended (see Table 1 for comparison of some selected techniques).

Table 1. Comparison of some sequencing technologies for ctDNA detection.

Analysis Type		Technique	Sensitivity (LoD)	Targets	Applications	Advantages	Limitations
PCR based methods	qPCR	ARMS-PCR PNA-LNA Clamp PCR COLD PCR	0.01–0.1%	Hotspot mutation	Cancer detection and monitoring, targetable alterations, some assays approved for clinical use	High specificity and sensitivity, cost effective, rapid, ease of use	No multiplexing, limited to detection of known mutations
	digital PCR	ddPCR BEAMing	0.01–0.1%	Hotspot mutations, gene fusions, CNV	Cancer detection and monitoring, targetable alterations, some assays approved for clinical use	Up to 5 targets, high sensitivity and specificity, absolute quantification, single molecule analysis, cost effective, rapid, ease of use	Limited multiplexing (number of fluorescent colors), limited to detection of known mutations
PCR based methods	PCR coupled to spectrometry	SERS UltraSEEK	0.1–1%	Known mutations	Cancer detection and monitoring, targetable alterations, for research use	Multiplexing capacity	Limited to detection of known mutations
NGS based methods	targeted	Tam-Seq	2%	Known and unknown mutations, indels, CNV, chromosomal rearrangements (capture)	Cancer detection and monitoring, classification, targetable alterations, for research use	High specificity	Amplicon methods by multiplex PCR (depend on fragment size), no error correction
		eTam-Seq	0.02%			Error correction	Amplicon methods by multiplex PCR
		Safe-SeqS	0.01–0.05%			Error correction by SSCS	Amplicon methods by multiplex PCR
		Duplex sequencing	0.0001–0.1%			Error correction by DSCS	Amplicon methods by multiplex PCR
		TEC-Seq	0.05–0.1%			Error correction by SSCS, Hybrid capture method (not dependent on fragment size)	Less comprehensive than WGS or WES
		single primer extension (SPE)	0.5–1%			Amplicon methods by SPE (not dependent on fragment size), error correction by SSCS	Less comprehensive than WGS or WES
		SPE-duplex UMI	0.1–0.2%			Error correction by DSCS	Less comprehensive than WGS or WES
		CAPP-Seq	0.02%			Hybrid capture method (not dependent on fragment size)	Need large input, allelic bias (capture), stereotypical errors (hybridization step), less comprehensive than WGS or WES
		iDES eCAPP-Seq	0.00025–0.004%			Error correction by DSCS and correction of stereotypical errors	Less comprehensive than WGS or WES
		Ig-HTS	0.001%	VDJ rearrangements	Non-invasive monitoring, approved for clinical use	Very high sensitivity	Tissue biopsy needed

Table 1. *Cont.*

Analysis Type	Technique	Sensitivity (LoD)	Targets	Applications	Advantages	Limitations
Untargeted	WES	5%	Coding regions, intron-exon junctions, promoters, untranslated regions, non-coding DNA of miRNA genes	Cancer detection, monitoring of resistant clones in metastasis, for research use	Mutation discovery and signatures, detection of CNV, fusion genes, rearrangements, predicted neoantigens and Tumor Mutational Burden	Low sensitivity (increasing depth lead to high cost), need bioinformatic expertise
	WGS	5–10%	Structural variants (fragmentation pattern, genome-wide CNV, methylation profile)	Cancer localization and origin, early detection (early and late stage), for research use	Shallow sequencing, genome wide profiling, identification of cancer signatures	Expensive, variable sensitivity (low) and specificity, need bioinformatic expertise, lots of data generated

Abbreviations: PCR—polymerase chain reaction; ARMS—amplification refractory mutation system; qPCR—quantitative real-time PCR; ddPCR—droplet digital PCR; BEAMing—beads, emulsion, amplification, magnetics; SERS—surface-enhanced Raman spectroscopy; PNA/LNA—peptide nucleic acid/locked nucleic acids; NGS—next-generation sequencing; Tam-Seq—Tagged-amplicon deep sequencing; TEC—targeted error correction; CAPP-Seq—Cancer Personalized Profiling by Deep Sequencing; iDES—Integrated Digital Error Suppression; Ig-HTS—Immunoglobulin high-throughput sequencing; WES—whole exome sequencing; WGS—whole genome sequencing; LoD—Limit of Detection; CNV—Copy Number Variation; indels—insertions/deletions; SSCS—single-stranded consensus sequence; DSCS—double-stranded consensus sequence.

Targeted approaches can detect, with high sensitivity, specificity and at a fast and cost-effective rate, already known recurrent mutations. These hotspot mutations frequently occur in a specific type of tumor and can be, most of the time, targeted by a therapy. Thus, targeted approaches can be very useful for the follow up of minimal residual disease to early detect relapse or track resistant mutations. Contrariwise, untargeted approaches are less sensitive but are useful for the discovery of new DNA mutations and genome wide alterations such as copy number variations (CNV, or copy number alterations, CNA).

Several parameters in the sequencing processing can affect the sensitivity of detection. One of them, also depending in part on the pre analytical process, is to put enough genome equivalent of cfDNA in the sequencing reaction to have enough altered molecule to detect. For example, as around 3000 copies of haploid genome are present in 10 ng of DNA, approximatively 60 ng of cfDNA will be required for a sensitivity of 0.01% (one rare event in 10,000 molecules), which is often challenging, even more if we consider that more than one observation is necessary to determine a true variant. Amplification steps cannot replace low input of cfDNA because the polymerase will introduce errors, increasing the risk to have false positive variants. Another parameter that may improve sensitivity is to monitor multiple alterations simultaneously in order to increase the chances of detecting ctDNA. With a binomial simulation, Van der Pol and Mouliere showed that, in theory and at a given concentration of cfDNA, increasing the number of mutations analyzed could improve the detection of low fraction ctDNA [21]. This kind of analysis was made possible with the advent of next-generation sequencing technologies, by increasing the possibility of multiplexing.

3.1. PCR-Based Methods

PCR-based methods, such as the derivatives of qPCR and digital PCR, are fast, cost-effective, and relatively simple to carry out and analyze. They allow detection of single or few mutations at low variants allele frequency, up to 0.1% and less, with high specificity.

3.1.1. Quantitative PCR

At first, the quantitative PCR (qPCR) method, by measuring the fluorescence emitted by a labeled probe during amplification of a targeted gene, was used to estimate the concentration of cfDNA in plasma of patients with cancer [38]. Later, qPCR assays were developed to detect mutations in tumoral cfDNA and the sensitivity of detection was improved by promoting the specific amplification of the mutant allele. Among the most used techniques, we can find ARMS-PCR, PNA-LNA Clamp PCR, or COLD PCR.

ARMS-PCR (amplification-refractory mutation system) is a simple method for detecting point mutations or small deletions, in which DNA is amplified by allele specific primers. In this technique, the lack of 3′ to 5′ exonuclease proofreading activity of the Taq polymerase reduces dramatically the annealing and hence the amplification in case of mismatch at the 3′ end of the primer. The limit of detection for this technique seems very variable according to the studies published, depending on the method, the samples used to determine this threshold or the mutations themselves. Although there are some improvements of the method, the false positive rate is still high with a limit of detection around 0.5 to 1% in plasma samples [39,40]. This limit can go down to 0.015% with ARMS-plus that includes a "Wild-type blocker" and in which amplicons were shortened to 50–80 bp, prohibiting the non-specific amplification and thus increasing the detection specificity [41].

PNA-LNA (peptide nucleic acid-locked nucleic acid) Clamp PCR uses a blocking synthetic nucleic acid analog complementary to wild type sequence to favor the amplification of the mutant allele. This method is particularly used in non-small cell lung cancer (NSCLC) to detect *EGFR* mutations, especially T790M mutations in tumor resistant to EGFR-TKIs (tyrosine kinase inhibitors), where cfDNA could be an alternative to the re-biopsy. This technique shows a high sensitivity with the detection of 0.1% mutant allele and a specificity of 79%. Using smaller PCR products and by increasing the number of cycles, Watanabe and colleagues reached less than 0.1% detection rate [42]. More recently, a dual PNA clamping-mediated LNA-PNA PCR clamp (LNA-dPNA PCR clamp) assay with two PCR rounds of PNA clamping succeeded in achieving a limit of detection of 0.01% [43].

COLD PCR (co-amplification at lower denaturation temperature-PCR) is an amplification method that selectively enriches low-abundance variant alleles from a mixture of wild-type and variation-containing DNA, irrespective of mutation type and position, by exploiting the critical denaturation temperature. The use of a lower denaturation temperature results in selective denaturation of molecules containing wild-type mutant heteroduplexes, which is followed by amplification. COLD-PCR has been used to improve the reliability of a number of different assays that traditionally use conventional PCR, such as Sanger sequencing, pyrosequencing or qPCR, greatly increasing their sensitivity. Thus, this method can detect mutant allele fraction down to 0.1% [44,45].

3.1.2. Digital PCR

As an example in lymphoma, this technique has a potential clinical use in diffuse large B cell lymphoma (DLBCL), as co-occurring mutations in *MYD88* and *CD79B* can predict response to Ibrutinib treatment, thus providing a predictive molecular tool for patient and therapy selection [46]. As well, in primary central nervous system lymphoma (PCNSL), mutation *MYD88* L265P was identified by ddPCR in cerebrospinal fluid or vitreous fluid with a superior sensitivity when compared with qPCR [47,48]. Since this mutation is found in up to 85% of PCNSL cases and not in non-hematological brain tumors, this ddPCR assay may be a promising technique for minimally invasive confirmation of PCNSL diagnosis.

BEAMing (beads, emulsion, amplification, magnetics) is a highly sensitive digital PCR method that combines emulsion PCR and flow cytometry to identify and quantify specific somatic mutations present in DNA [49]. Diehl and coworkers used a BEAMing approach to detect mutations in cfDNA from patients with colorectal cancer, showing that ctDNA dynamics reflects tumor responses and progression, and that ctDNA detection after surgery represented a marker of residual disease [50]. This method, mainly used so far in solid tumors, such as colorectal [51], breast [52], and lung cancers [53], has a highly sensitive detection rate with variant allele fraction as low as 0.01%.

Although ddPCR allows for quantitative assessment of mutant frequencies in cfDNA, it is limited by the number of fluorescent probes that can be used in one assay (up to five) [54,55].

Copy number variations have also been investigating in cfDNA using ddPCR. Even if the number of targets is limited, it can be a useful tool for detecting, simply and rapidly,

some gains or losses, which are associated with poor prognosis at diagnosis or during follow-up [56,57].

DdPCR can also be suitable to detect chromosomal rearrangements, especially in hematological malignancies. Among others, assays have been developed for translocation t(11;14) deregulating the *CCND1* gene and translocation t(14;18) deregulating the *BCL2* gene, which are frequently observed in Mantle cell lymphoma (MCL) and follicular lymphoma (FL), respectively [58,59]. The sensitivity of these techniques can go down to 0.01%.

3.1.3. PCR Coupled with Mass Spectrometry

The major limitation of the previous PCR-based approaches is their very limited multiplexing ability. Mass spectrometry-based methods such as surface-enhanced Raman spectroscopy (SERS) and UltraSEEK are adaptation of the conventional PCR method with a unique advantage in multiplexing to detect ctDNA mutations at low frequency with low input amount of cfDNA and fast turnaround time.

SERS is a surface-sensitive technique that enhances Raman scattering by molecules adsorbed on rough metal surfaces or by nanostructures such as plasmonic-magnetic silica nanotubes [60]. The detection of target specific DNA is based on the use of labeled nanotags (Raman reporters) and the measurement of the Shift in the spectrum of Raman reporter that can provide information about low-frequency transitions in molecules. The status of mutations is then analyzed with SERS spectrum where unique spectral peaks demonstrated the presence of targeted mutations. Multiplex PCR/SERS identifying three hotspot mutations has been developed in melanoma and colorectal cancer with a limit of detection as few as 0.1% [61,62].

The UltraSEEK chemistry is able to interrogate multiple informative variants within a single reaction. In this method, the mutant allele is specifically targeted by a primer extension step that omits the wild type allele. Reaction products are subsequently captured to a solid support, washed and released. Eluted products are then submitted to MALDI-TOF Mass Spectrometry. The use of a 68 mutations panel on cfDNA from melanoma patients showed the same sensitivity as ddPCR [63]. In NSCLC, the limit of detection of the UltraSEEK Lung Panel, consists of 73 variants, was 0.125–1% with low input of specific tumoral cfDNA fragments beforehand measured with the LiquidIQ Panel [64]. Of note, this study showed the importance of preanalytical cfDNA quality control and input amount for the accuracy of liquid biopsy testing. The comparison between UltraSEEK and a real-time PCR test (cobas *EGFR* Mutation test v2) showed a concordance of 100% with more than 10 ng of cfDNA, whereas it fell to 73–84% when less than 8 ng were used, implying a loss of sensitivity.

Overall, these PCR-based assays are very effective tools for detecting mutations at a relatively low-cost, which make them feasible in routine clinical practices. The main limitation is the limited multiplexing ability, which restricts the possibility of targets and can lead to a greater consumption of material. Furthermore, the alterations detected must be previously known such as hotspot mutations, which is more suitable for a minimal residual disease but less as a diagnostic tool.

3.2. Targeted NGS-Based Methods

Targeted deep sequencing techniques are still limited to a certain number of regions but can cover entire genes or entire coding regions of genes. Thus, they are suitable for genes without hotspot mutations, which is often the case for loss of function mutations in tumor suppressor genes.

Targeted enrichment in library construction can be achieved by direct amplification (amplicon or multiplex PCR) or hybridization capture (hybrid capture) of the DNA regions of interest. Techniques using multiplex PCR-based methods are more dependent on the length of the fragments and may require several simultaneous reactions for target enrichment to cover a large region of a gene, consuming more DNA. Hybrid capture

some gains or losses, which are associated with poor prognosis at diagnosis or during follow-up [56,57].

DdPCR can also be suitable to detect chromosomal rearrangements, especially in hematological malignancies. Among others, assays have been developed for translocation t(11;14) deregulating the *CCND1* gene and translocation t(14;18) deregulating the *BCL2* gene, which are frequently observed in Mantle cell lymphoma (MCL) and follicular lymphoma (FL), respectively [58,59]. The sensitivity of these techniques can go down to 0.01%.

3.1.3. PCR Coupled with Mass Spectrometry

The major limitation of the previous PCR-based approaches is their very limited multiplexing ability. Mass spectrometry-based methods such as surface-enhanced Raman spectroscopy (SERS) and UltraSEEK are adaptation of the conventional PCR method with a unique advantage in multiplexing to detect ctDNA mutations at low frequency with low input amount of cfDNA and fast turnaround time.

SERS is a surface-sensitive technique that enhances Raman scattering by molecules adsorbed on rough metal surfaces or by nanostructures such as plasmonic-magnetic silica nanotubes [60]. The detection of target specific DNA is based on the use of labeled nanotags (Raman reporters) and the measurement of the Shift in the spectrum of Raman reporter that can provide information about low-frequency transitions in molecules. The status of mutations is then analyzed with SERS spectrum where unique spectral peaks demonstrated the presence of targeted mutations. Multiplex PCR/SERS identifying three hotspot mutations has been developed in melanoma and colorectal cancer with a limit of detection as few as 0.1% [61,62].

The UltraSEEK chemistry is able to interrogate multiple informative variants within a single reaction. In this method, the mutant allele is specifically targeted by a primer extension step that omits the wild type allele. Reaction products are subsequently captured to a solid support, washed and released. Eluted products are then submitted to MALDI-TOF Mass Spectrometry. The use of a 68 mutations panel on cfDNA from melanoma patients showed the same sensitivity as ddPCR [63]. In NSCLC, the limit of detection of the UltraSEEK Lung Panel, consists of 73 variants, was 0.125–1% with low input of specific tumoral cfDNA fragments beforehand measured with the LiquidIQ Panel [64]. Of note, this study showed the importance of preanalytical cfDNA quality control and input amount for the accuracy of liquid biopsy testing. The comparison between UltraSEEK and a real-time PCR test (cobas *EGFR* Mutation test v2) showed a concordance of 100% with more than 10 ng of cfDNA, whereas it fell to 73–84% when less than 8 ng were used, implying a loss of sensitivity.

Overall, these PCR-based assays are very effective tools for detecting mutations at a relatively low-cost, which make them feasible in routine clinical practices. The main limitation is the limited multiplexing ability, which restricts the possibility of targets and can lead to a greater consumption of material. Furthermore, the alterations detected must be previously known such as hotspot mutations, which is more suitable for a minimal residual disease but less as a diagnostic tool.

3.2. Targeted NGS-Based Methods

Targeted deep sequencing techniques are still limited to a certain number of regions but can cover entire genes or entire coding regions of genes. Thus, they are suitable for genes without hotspot mutations, which is often the case for loss of function mutations in tumor suppressor genes.

Targeted enrichment in library construction can be achieved by direct amplification (amplicon or multiplex PCR) or hybridization capture (hybrid capture) of the DNA regions of interest. Techniques using multiplex PCR-based methods are more dependent on the length of the fragments and may require several simultaneous reactions for target enrichment to cover a large region of a gene, consuming more DNA. Hybrid capture

methods employ custom RNA probes complementary to targeted regions and are able to detect both single nucleotide variants (SNV) and structural variants [65]. In this method of enrichment, the fragmentation of cfDNA can lead to a heterogeneous coverage across targeted exons with a lower fragment depth in the edge regions of exons, which must be taken into consideration when designing the panels for ctDNA sequencing [66].

The main issue of going down in sensitivity is the reliability of interpretation in the discrimination between the true and the false variants. Although they have high sensitivity and specificity, NGS platforms show a random error rate between 0.1 and 1.5% per base call, but library preparation protocols have been upgraded to improve the detection of rare variants [67,68]. In targeted DNA sequencing, the use of few DNA molecules combined with ultra-deep sequencing increases the risk to read several times the same molecule where polymerase errors are introduced at any step during the NGS process, leading to the inability to confidently call rare variants. One of the major recent technological advances is the use of molecular barcodes, which are random sequences introduced before any amplification step. They allow the counting of original DNA molecules instead of PCR duplicates, thereby enabling digital sequencing and resulting in unbiased and accurate mutation profiles with an increased sensitivity [69–72].

- Tagged-amplicon deep sequencing (Tam-Seq)

Tam-Seq is an amplicon method using a target enrichment array with barcoded primers to prepare the amplicon library for NGS. First, an initial targeted preamplification step is carried out, followed by a selective amplification of the regions of interest in single-plex reactions. Then, sequencing adaptors and sample-specific barcodes are attached to the amplicons in a further PCR. It was first able to detect mutations in circulating DNA with high sensitivity and specificity (>97%) at allele frequencies as low as 2% [73]. The technique has been recently improved (enhanced Tam-Seq, eTam-Seq) with a primer design strategy, allowing for amplification of highly fragmented DNA, a workflow reducing the background error rate, and a more efficient calling algorithm with better detection of SNV and indels (insertions/deletions), and also CNV [74]. This assay, using an optimal amount of DNA, detected 94% mutations at 0.25–0.33% allele fraction (AF) with a limit of detection down to 0.02% AF with high per-base specificity (99.9997%). In this study comparison of eTam-Seq with dPCR showed a good concordance between the two techniques, demonstrating the quantitative accuracy of eTAm-Seq technology for reliable detection of mutations at low allele frequency [74].

- Safe-Sequencing System (Safe-SeqS)

This amplicon method was originally described by the group of Bert Vogelstein [69]. It was the first approach using molecular barcodes in DNA sequencing, to increase sensitivity of massively parallel sequencing. In this technique, a unique identifier (UID) is assigned to each template molecule before any amplification. Thereby, PCR fragments with the same UID are considered mutant if more than 95% of them contain the identical mutation. Thus, this method allows a correction of amplification and sequencing errors and can quantify rare mutations with a sensitivity of 0.05% of allele fraction. Safe-SeqS showed high performance in detecting mutations in cfDNA from patients with solid tumors, for molecular profiling as well as real-time monitoring of minimal residual disease [75]. A recent study on three independent cohort of nonmetastatic colorectal cancer, showed a median mutant allele frequency of 0.046% with a minimum of 0.01% [76].

- Duplex sequencing

Duplex sequencing is an improvement of the Safe-SeqS technique [77,78]. In this method, a semi-degenerated double stranded unique barcoded adapter is ligated to a target double stranded DNA. After sequencing, molecules with the duplex adaptors are compared and mutations are retained only if there is a consensus between both strands. Thus, in addition to get rid of PCR and sequencing errors, the advantage of this technique is to identify artifacts due to sample alterations [79] because it can examine both strands individually and the damage to them is usually not identical (error correction by double-

stranded consensus sequence). The theoretically sensitivity of this approach to discovering mutants is one molecule among 10^7 which is much higher in accuracy than conventional next-generation sequencing methods [77,78].

Several studies, in various types of cancers, applied this method on plasma cfDNA. In combination with target enrichment using hybrid capture, this approach allowed detection of tumoral fraction at 0.1% and below with high sensitivity and specificity, providing a powerful tool for diagnosis as well as longitudinal monitoring of disease [80–82].

- Targeted error correction sequencing (TEC-Seq)

In this technique, molecular barcoding is also used to facilitate the discrimination between true mutations and false positive variants. DNA fragments are tagged each one with a different "exogenous" DNA barcode before any amplification, as for Safe-SeqS, but not only. The start and end genome mapping positions of paired-end sequenced fragments were also used as "endogenous barcodes" to distinguish between individual molecules. This combination of barcodes allows keeping track of each fragment as they are sequenced around 30,000 times [70]. This approach was applied to several type of solid cancers and demonstrated ability for early stage detection. The analytical sensitivity was 100% and 89% for detecting mutations present at 0.2% and 0.1%, respectively, using minimum thresholds of 0.05% in hot-spot positions and 0.1% at all other locations, resulting in a sensitivity of 97.4% overall, and without detection of false positives (less than one error in three million bases sequenced).

- Single primer extension (SPE) with unique molecular barcode

SPE is an amplicon-based method used by QIAGEN in their QIASeq targeted DNA panel kits. This approach uses only one gene specific primer (GSP) for amplification of each genomic region, which makes it less dependent on the size of DNA fragments than PCR using two primers and offers a uniform coverage. As for capture, the first step is a fragmentation step in which the buffer used inhibits fragmentation of the high length fragments of DNA such as contaminating gDNA. The following steps are reparation and ligation of adapters. These adapters will be used for amplification of targeted region (together with GSP) and contain the degenerated molecular barcodes (UMI, Unique Molecular Index). Moreover, given this UMI contains 12 base pairs, it allows a large number of combinations and a very little risk for redundancy [71]. Theorical sensitivity threshold of this technique is 0.5–1% with over 90% sensitivity and a very few number of false positive. Recently, improvement by using duplex UMI adapters lowered the sensitivity up to 0.1–0.2% allele fractions [83].

This technique of deep sequencing, using molecular barcodes to improve accuracy in variant detection, has been used at diagnosis in order to identify actionable genetic alteration with targeted therapies available for treatment or hotspot mutations to be tracked with ddPCR during follow up, with a detection of variant allele frequency down to 1–5% [84,85]. Further investigations are needed to find the real limit of detection of this technology, which may be below 1% as other techniques using molecular barcoding.

This approach also allowed detection of CNV. In PCR-based library construction, amplification introduces biases in further reads count because the amplification factor is dependent on many parameters such as library size, GC content, region length or competition between primers overlapping the same locus. Thus, the use of UMI via the mCNA tool allows the direct count of targeted DNA molecules before any amplification and the detection of CNV in a robust and sensitive way [86].

- Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq)

CAPP-Seq is an ultra-sensitive assay consisting of a hybrid capture-based NGS method developed for ctDNA detection. In this technique, the first important step is to query cancer databases to identify known recurrent mutations for a particular cancer type. Then, biotinylated oligonucleotide probes, named "Selector", are designed to target large segments of the concerned regions. The protocol is optimized for low DNA levels and sensitivity is increased using deep sequencing [87,88]. The sensitivity is also improved by its ability to

detect simultaneously various types of alterations: single nucleotide variants, rearrangements, insertions/deletions, and copy number alterations. It was originally described to detect and monitor lung cancer but was successfully adapted to a broad range of cancers, including different types of solid tumors as well as hematological malignancies such as DLBCL, LF, and HL [10,20,89–91].

With this method, ctDNA was detected in blood of NSCLC patients with 96% specificity for mutant allele fraction down to 0.02%. It was improved in 2016, with the use of iDES (Integrated Digital Error Suppression). This iDES-enhanced CAPP-Seq combines CAPP-Seq with duplex barcoding sequencing technology and with a computational algorithm that removes stereotypical errors associated with the CAPP-Seq hybridization step. This improved version of CAPP-Seq has shown a high sensitivity in the detection of EGFR mutations in cfDNA of NSCLC patients, with variant allele frequency as low as 0.004% with >99.99% specificity. Moreover, using duplex sequencing and covering a large number of mutations (≥200), the authors outperformed iDES and managed to detect ctDNA down to 0.00025%, with an input of only 32 ng of cfDNA [92].

- Immunoglobulin high-throughput sequencing (Ig-HTS)

This test was specifically developed for MRD in hematological malignancies. In this method, ultra-deep sequencing of genomic DNA, with a set of locus-specific multiplex PCR covering all possible rearranged IgH, IgK, and IgL receptor gene sequences, firstly identifies the tumor-specific clonotype. Then, this clonotype can be tracked as a specific fingerprint to quantify ctDNA in lymphoma disease monitoring with a sensitivity of approximatively 10-6 [93–95]. This technique presents some technical limitations, including the need of tissue biopsy to identify clonotype and difficulties to identify clonotype sequences in some lymphoma types such as DLBCL of the germinal center type and FL because of somatic hypermutation (SHM). Nevertheless, this method has shown high performance in surveillance ctDNA, after complete remission, to identify risk of recurrence before any clinical evidence of disease in most patients (with a median of 3.5 months) [93,94].

This approach was also used for MRD monitoring in DLBCL patients after CAR-T cell therapy, showing correlation with clinical and radiologic outcomes for all the patients tested [96].

3.3. Untargeted NGS-Based Methods

As mentioned previously, untargeted approaches, namely whole exome and whole genome sequencing (WES, WGS), are less sensitive than targeted approaches. The sensitivity of these techniques on cfDNA is estimated around 5–10%, as compared to less than 0.1% for a targeted sequencing approach [97], making it difficult to detect rare events, especially in situations of early detection or minimal residual disease. Moreover, these technologies are more expensive and require both very high throughput sequencing equipment and expertise to analyze the large amount of data generated, which makes its implementation in routine practices challenging. However, these approaches may be necessary for the discovery of new alterations in the context of initial profiling at diagnosis, to provide information for the use of more sensitive targeted techniques during disease monitoring. Even if they are not suitable to detect subclonal events, they may be useful, considering intra tumoral heterogeneity, to highlight new drug targets or to track drug resistance clones [98].

WES is, most of the time, limited to coding regions and splicing sites of genes but it is a good compromise for exploration of unknown mutations at a reasonable cost. It can identify both driver and passenger mutations and also can be extended to promoters, untranslated regions, and non-coding DNA of miRNA genes. Even if protein-coding genes constitute only approximately 1.5% of the human genome, they contain a great majority of the disease-causing mutations [99]. The technical feasibility of whole-exome sequencing (WES) on cfDNA has been demonstrated in various solid tumors and some hematological malignancies [98]. Low coverage and sensitivity, compared to targeted NGS technologies does not allow for the detection of rare variants but WES of cfDNA is suitable for mutational analysis of patients with advanced tumors and increased ctDNA fractions (>5% mutant

allele fraction). The first exome-wide sequencing analysis of ctDNA was performed to analyze serial plasma samples (before initiating treatment and at disease recurrence), in order to track genomic evolution and response to therapy in patients with metastatic cancer (breast, ovarian, and lung cancer) receiving systemic therapy [100]. These samples contained high percentages of ctDNA (between 5% and 55%) and the average depth of sequencing coverage ranged from 31- to 160-fold. This study showed the possibility to identify candidate genetic alterations driving treatment resistance using cfDNA analysis. These findings largely agreed with additional studies demonstrating that whole-exome sequencing of cfDNA in metastatic patients could serve as a surrogate for tumor genome analysis, considering the difficulties of doing multiple biopsies and the high ctDNA allele frequencies making WES possible [101–104].

Additionally, given intra tumoral heterogeneity, analysis comparing mutational profile between tumor and cfDNA mostly identified more mutations in cfDNA with a high prevalence of targetable genes. Beyond SNV detection, WES of cfDNA also allowed analysis of mutational signatures, copy number variations, fusion genes, rearrangements, predicted neoantigens, and tumor mutational burden [98].

Contrariwise to WES, WGS technologies is more suitable to detect ctDNA by identifying structural and non-coding variations such as genome-wide copy number aberrations, methylation profiles, and fragmentation patterns.

To override the cost and analysis time limitations caused by WGS, Heitzer and colleagues developed a shallow genome-wide sequencing approach called Plasma-Seq [105]. This method uses an Illumina MiSeq instrument, which is a benchtop high-throughput sequencing platform often available in routine laboratories. This technique does not have a sufficient sequencing resolution to identify SNV but is able to detect CNV in cfDNA at a depth of 0.1×, with a specificity >80% when ctDNA fraction is ≥10%. Recently, this approach of shallow WGS has been successfully used in cfDNA of DLBCL and HL patients to identify copy number patterns that can differentiate the two diseases at diagnosis [106]. These copy number aberrations were also correlated with clinical parameters, and longitudinal analyses showed correlation with disease status. Moreover, the sensitivity and informativity for HL was better in cfDNA than in tumor, as for mutation detection [10,11,106].

Aneuploidy has also been explored with WGS derived techniques such as Fast-SeqS (Fast Aneuploidy Screening Test-Sequencing System) and WALDO (Within Sample Aneuploidy Detection), using a single specific primer pair to amplify dispersed retrotransposon regions throughout the genome (long interspersed nuclear elements (LINEs)) [107,108]. By simulations with synthetic DNA, the bioinformatic tool WALDO showed high performance to detect individual chromosome arm gain or loss with a fraction of ctDNA >5%, and up to 1% of tumoral fraction with a sensitivity of 78%. However, due to their mechanism of detection, these techniques are limited to cancers presenting aneuploidy.

In order to detect genomic rearrangements, Leary et al. developed a technique called PARE (personalized analysis of rearranged ends), which uses WGS mate-paired analysis of the tumoral DNA to identify patient specific genomic rearrangement. This assay is highly sensitive with detection of ctDNA lower than 0.001% of total cfDNA [109]. Analyses, in breast and colorectal cancers, suggest that ctDNA concentrations at levels >0.75% could be detected in the cfDNA of patients with a sensitivity >90% and a specificity >99%, and that even a single copy of rearrangement from ctDNA can be detected without false positives [110]. In a recent study, PARE was employed to detect rearrangements in gastric tumor, which were used to design a quantitative PCR assay targeting rearranged loci for quantitative monitoring in cfDNA. Thus, the authors were able to predict relapse as the presence of postoperative ctDNA was significantly correlated with cancer recurrence within 12 months of surgery [111].

WGS, combined with artificial intelligence, can also identify genome-wide fragmentation patterns in cfDNA. Several studies in different cancer types have shown that these patterns can be used to detect ctDNA in body fluids and with very low plasma ctDNA

allele fraction). The first exome-wide sequencing analysis of ctDNA was performed to analyze serial plasma samples (before initiating treatment and at disease recurrence), in order to track genomic evolution and response to therapy in patients with metastatic cancer (breast, ovarian, and lung cancer) receiving systemic therapy [100]. These samples contained high percentages of ctDNA (between 5% and 55%) and the average depth of sequencing coverage ranged from 31- to 160-fold. This study showed the possibility to identify candidate genetic alterations driving treatment resistance using cfDNA analysis. These findings largely agreed with additional studies demonstrating that whole-exome sequencing of cfDNA in metastatic patients could serve as a surrogate for tumor genome analysis, considering the difficulties of doing multiple biopsies and the high ctDNA allele frequencies making WES possible [101–104].

Additionally, given intra tumoral heterogeneity, analysis comparing mutational profile between tumor and cfDNA mostly identified more mutations in cfDNA with a high prevalence of targetable genes. Beyond SNV detection, WES of cfDNA also allowed analysis of mutational signatures, copy number variations, fusion genes, rearrangements, predicted neoantigens, and tumor mutational burden [98].

Contrariwise to WES, WGS technologies is more suitable to detect ctDNA by identifying structural and non-coding variations such as genome-wide copy number aberrations, methylation profiles, and fragmentation patterns.

To override the cost and analysis time limitations caused by WGS, Heitzer and colleagues developed a shallow genome-wide sequencing approach called Plasma-Seq [105]. This method uses an Illumina MiSeq instrument, which is a benchtop high-throughput sequencing platform often available in routine laboratories. This technique does not have a sufficient sequencing resolution to identify SNV but is able to detect CNV in cfDNA at a depth of 0.1×, with a specificity >80% when ctDNA fraction is ≥10%. Recently, this approach of shallow WGS has been successfully used in cfDNA of DLBCL and HL patients to identify copy number patterns that can differentiate the two diseases at diagnosis [106]. These copy number aberrations were also correlated with clinical parameters, and longitudinal analyses showed correlation with disease status. Moreover, the sensitivity and informativity for HL was better in cfDNA than in tumor, as for mutation detection [10,11,106].

Aneuploidy has also been explored with WGS derived techniques such as Fast-SeqS (Fast Aneuploidy Screening Test-Sequencing System) and WALDO (Within Sample Aneuploidy Detection), using a single specific primer pair to amplify dispersed retrotransposon regions throughout the genome (long interspersed nuclear elements (LINEs)) [107,108]. By simulations with synthetic DNA, the bioinformatic tool WALDO showed high performance to detect individual chromosome arm gain or loss with a fraction of ctDNA >5%, and up to 1% of tumoral fraction with a sensitivity of 78%. However, due to their mechanism of detection, these techniques are limited to cancers presenting aneuploidy.

In order to detect genomic rearrangements, Leary et al. developed a technique called PARE (personalized analysis of rearranged ends), which uses WGS mate-paired analysis of the tumoral DNA to identify patient specific genomic rearrangement. This assay is highly sensitive with detection of ctDNA lower than 0.001% of total cfDNA [109]. Analyses, in breast and colorectal cancers, suggest that ctDNA concentrations at levels >0.75% could be detected in the cfDNA of patients with a sensitivity >90% and a specificity >99%, and that even a single copy of rearrangement from ctDNA can be detected without false positives [110]. In a recent study, PARE was employed to detect rearrangements in gastric tumor, which were used to design a quantitative PCR assay targeting rearranged loci for quantitative monitoring in cfDNA. Thus, the authors were able to predict relapse as the presence of postoperative ctDNA was significantly correlated with cancer recurrence within 12 months of surgery [111].

WGS, combined with artificial intelligence, can also identify genome-wide fragmentation patterns in cfDNA. Several studies in different cancer types have shown that these patterns can be used to detect ctDNA in body fluids and with very low plasma ctDNA

fraction [112,113]. Indeed, ctDNA fragments are generally shorter and more variable in their length than those found in controls are. Moreover, beyond this difference of size of fragments in cfDNA between healthy individuals and patients with cancer, their location in the genome can be informative of the epigenetic profile of the origin cells. Indeed, the cfDNA fragmentation landscape represents a nucleosome footprint reflecting the cell and tissue of origin, potentially enabling non-invasive diagnosis of cancer type [112]. Recently, Cristiano et al. used this approach for the early detection of ctDNA from 236 patients with various cancers and reported sensitivities ranging from 57 to 99% with a specificity of 98% [114]. This nucleosome footprinting firstly identified by WGS represents nucleosome depletion at transcription start sites of highly expressed genes and the capture of this chromatin accessibility profile was used by CAPP-Seq technology to define gene expression differences and thus determine the cell-of-origin in DLBCL subjects from cfDNA [115].

Among epigenetic alterations, aberrant DNA methylation events can also represent an ideal biomarker for detection and classification of early stage cancer, as they occur early in cancer development, sometimes before the acquisition of SNVs. Multiple liquid biopsy studies have been performed utilizing DNA methylation markers in various cancer types [21]. As whole genome bisulfite sequencing is inefficient due to low recovery and degradation of DNA after bisulfite conversion [116], high cost and limited information recovery given the low genome-wide abundance of CpGs, techniques has been developed to pre-enrich methylated DNA fragments with or without bisulfite treatment. These strategies are either very targeted, as methylation events of interest occur at known, stereotyped positions [117], or larger to identify methylation patterns, which have been shown to enable accurate determination of cell-of-origin from cfDNA and non-invasive cancer classification. For example, a technique for cell-free methylated DNA immunoprecipitation followed by high throughput sequencing (cfMe-DIP) has been developed for genome-wide methylation exploration of bisulfite-free plasma DNA, on low input cfDNA and with enough sensitivity for early detection of cancer [118]. More recently, a semi targeted assay of 595 genomic regions covering 11,787 CpG sites, named PanSeer assay, allowed the detection of five types of cancer in 88% of post-diagnosis patients with a specificity of 96% [119]. Even if the result needs confirmation, the authors also detected cancer in 95% of asymptomatic individuals who were later diagnosed, demonstrating that cancer can be non-invasively detected up to four years before diagnosis. In lymphoma, aberrant promoter methylation patterns detected in cfDNA have been shown to be an independent and significant poor prognostic factor for 5-year overall survival in DLBCL, outperforming existing clinical risk parameters an independent [120,121].

Moreover, as healthy cells also participate to epigenetic changes, it may need to be distinguished from these of cancers cells [21]. Thus, it could be of major interest to combine epigenetic analysis of the entire cfDNA pool with mutational analysis of ctDNA molecules.

4. Bioinformatical Methods

While cfDNA seems to be a promising screening tool, it still remains a real challenge for bioinformatics. While common bioinformatics strategies allow variant identification down to 2–5% allele frequency, in most cases, ctDNA accounts for a small fraction of total cfDNA since most of cfDNA is derived from non-cancer cells and especially blood cells. ctDNA fraction can be lower than 0.1%, leading to the detection of somatic mutations at the same level as the sequencing noise. It implies the use of in silico strategies to distinguish true positive variant calls from sequencing noise.

It has been reported from healthy controls that under an allele fraction of 0.02%, more than 50% of sequenced genomic positions had sequencing artifacts [92]. These errors are particularly due to library preparation, the error rates of NGS technologies, and the physical characteristic of the cfDNA fragments.

In addition, there are many tools and therefore many bioinformatics parameters that need to be optimized when analyzing cfDNA samples. While major progress has been made in the harmonization of tumor analyzes with the GATK4 Best Practices Workflows [122],

there is not yet an international consensus for bioinformatic cfDNA analysis and research in this area remains very active.

4.1. Adapter Contamination

The quality of cfDNA analysis is particularly impacted by adapter contamination. cfDNA fragments could be shorter than usual which may result in the sequencing of adapters due to too many sequencing cycles compared to their lengths. Consequently, these reads could be either unmappable to the reference genome or could have a lower alignment score. These alignment scores are considered by a large number of bioinformatic tools and could finally affect the results of variant caller algorithms. Adapter contamination could be found both in 5′ and in 3′ of sequenced reads.

Many softwares were developed to find and trim adapters, like Cutadapt [123], Tag-Cleaner [124], Trim Galore [125], or Trimmomatic [126]. In general, these algorithms also integrate the trimming of low quality nucleotides and the extraction of molecular barcodes.

4.2. Library Biases and Molecular Barcoding

The amplification of the libraries by PCR includes many biases for counting mutated reads because the number of aligned reads is no longer directly proportional to the number of initial unique targeted DNA fragments. The amplification factor of each region is unknown and depends on many parameters such as library size, GC content, or fragment length. This bias is particularly present for samples with low DNA concentration at extraction. Recent advances in library preparation allow the addition of Unique Molecular Identifiers (UMI) to each read. UMI are especially useful to correct library amplification biases by making each DNA molecule in a population of reads distinct.

There are two main bioinformatic approaches to use UMI for cfDNA analysis.

The first one consists in grouping PCR duplicates prior to any downstream analysis by merging sequences harboring the same UMI tag. To perform this task, the most popular tool is UMI-tools [127]. The advantage of this approach is that it allows the use of classic bioinformatic pipelines after deduplicating the reads. It erases amplification biases due to cfDNA characteristics. However, it no longer provides access to essential information such as the amplification factor of each UMI or the discordant mutation calls of reads having the same UMI.

More recent approaches consist in using new bioinformatic algorithms for variant and CNV calling which are able to take into account the information carried by the UMIs after alignment, i.e., at the end of data processing.

For example, the UMI-VarCal algorithm [128] tries to quantify the number of concordant and discordant UMIs for each candidate variant during the variant calling process. Concordant UMIs were defined as number of unique UMIs for which all the reads carrying these UMIs validate the presence of the variant. Conversely, discordant UMIs quantify the number of abnormal substitutions like sequencing or PCR errors. Another example of barcode-aware variant caller is SmCounter [71]. SmCounter uses a barcode level allele probability and UMI counts to reject candidate mutations lacking enough barcodes with good read evidence. These approaches make it possible to distinguish true mutations at low frequency from sequencing noise and is particularly useful for cfDNA analysis.

Many biases due to the amplification step while preparing sequencing libraries prevent the direct quantification of loci copy-number [129]. cfDNA fragments are often shorter than DNA extracted from tissue and make it impossible to use conventional approaches for the detection of CNV such as read-depth algorithms. Recent approaches, like mCNA [86], use the UMI counts instead of read counts to improve high-resolution copy number variation of genes.

4.3. Bioinformatics Processing

There is not yet an international consensus for bioinformatic cfDNA analysis pipeline. The bioinformatics tools and parameters must be adapted to the nature of the sequenced

samples (quantity of DNA, quality of extraction, integrity of the cfDNA, etc.), to the kits used to prepare libraries, to the presence of UMI in library construction or not, and finally to the sequencing depth. In addition, sequencing biases are often sample specific which requires an objective assessment of sequencing noise at sample level.

However, this variability, specific to each sample, is not incompatible with an objective evaluation of the performance of bioinformatic algorithms. Some first tools, like UMI-Gen [130], allow to create in silico alignment datasets to evaluate the performance of variant calling and filtration tools. UMI-Gen is a UMI-based read simulator, which reproduces targeted sequencing paired-end alignment files (BAM) by estimating sequencing noise from a set of reference BAM files. It is particularly useful for evaluating the performance of variant calling tools because it allows to vary many parameters (sequencing depth, number of initial UMI, etc.) and to insert variants at frequencies of interest during the simulation. It thus makes it possible to optimize bioinformatic pipelines according to the targeted panels or the sequencing technology.

5. Conclusions

Many studies have demonstrated that analysis of ctDNA, as a liquid biopsy, is a powerful tool for non-invasive genotyping across various cancer types, in solid tumors, as well as in hematological malignancies. Investigations have shown the possibility to use ctDNA both at diagnosis, for prognosis or targeted therapies, and during longitudinal monitoring of the disease, as a dynamic biomarker of tumor burden during treatment and to detect relapse after treatment. Moreover, liquid biopsy could be a surrogate for tissue biopsy in some particular cases of tumors not accessible for surgery or spread of tumoral mass and metastasis, given the intra tumoral heterogeneity.

Nowadays, the main issue in ctDNA analysis is to go down in sensitivity without generating false positive results, especially for early detection of cancer at diagnosis and relapse. Due to its short fragment length, low quantity, and small fraction in cfDNA [131], reliable detection of ctDNA can still be a major technical challenge. That is why the most suitable ctDNA assay for a specific application has to be chosen according to its analytical performance characteristics [66]. However, recent optimizations in techniques, from standardization of preanalytical processing to the development of high sensitive sequencing technologies with the help of bioinformatical algorithms for error correction, has the potential to detect ctDNA at the molecular level with a great accuracy.

The next step in the near future could be the integration of ctDNA detection assays into prospective multicentric studies and clinical trials to establish its true clinical utility.

Author Contributions: Writing—original draft preparation, E.B. and P.-J.V.; writing—review and editing, E.B., P.-J.V. and F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Figure 1 was created by E.B. with BioRender.com (21 June 2021, Agreement number: LU22M8WBV5).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mandel, P.; Metais, P. Nuclear acids in human blood plasma. *C. R. Seances Soc. Biol. Fil.* **1948**, *142*, 241–243.
2. Koffler, D.; Agnello, V.; Winchester, R.; Kunkel, H.G. The Occurrence of Single-Stranded DNA in the Serum of Patients with Systemic Lupus Erythematosus and Other Diseases. *J. Clin. Investig.* **1973**, *52*, 198–204. [CrossRef]
3. Leon, S.A.; Shapiro, B.; Sklaroff, D.M.; Yaros, M.J. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res.* **1977**, *37*, 646–650. [PubMed]

4. Stroun, M.; Anker, P.; Maurice, P.; Lyautey, J.; Lederrey, C.; Beljanski, M. Neoplastic Characteristics of the DNA Found in the Plasma of Cancer Patients. *Oncology* **1989**, *46*, 318–322. [CrossRef]
5. Sidransky, D.; Von Eschenbach, A.; Tsai, Y.C.; Jones, P.; Summerhayes, I.; Marshall, F.; Paul, M.; Green, P.; Hamilton, S.R.; Frost, P.; et al. Identification of p53 gene mutations in bladder cancers and urine samples. *Science* **1991**, *252*, 706–709. [CrossRef] [PubMed]
6. Vasioukhin, V.; Anker, P.; Maurice, P.; Lyautey, J.; Lederrey, C.; Stroun, M. Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.* **1994**, *86*, 774–779. [CrossRef]
7. Anker, P.; Lefort, F.; Vasioukhin, V.; Lyautey, J.; Lederrey, C.; Chen, X.Q.; Stroun, M.; Mulcahy, H.E.; Farthing, M.J. K-ras mutations are found in DNA extracted from the plasma of patients with colorectal cancer. *Gastroenterology* **1997**, *112*, 1114–1120. [CrossRef]
8. Austrup, F.; Uciechowski, P.; Eder, C.; Böckmann, B.; Suchy, B.; Driesel, G.; Jäckel, S.; Kusiak, I.; Grill, H.J.; Giesing, M. Prognostic value of genomic alterations in minimal residual cancer cells purified from the blood of breast cancer patients. *Br. J. Cancer* **2000**, *83*, 1664–1673. [CrossRef] [PubMed]
9. Bobillo, S.; Crespo, M.; Escudero, L.; Mayor, R.; Raheja, P.; Carpio, C.; Rubio-Perez, C.; Tazón-Vega, B.; Palacio, C.; Carabia, J.; et al. Cell free circulating tumor DNA in cerebrospinal fluid detects and monitors central nervous system involvement of B-cell lymphomas. *Haematologica* **2020**, *106*, 513–521. [CrossRef]
10. Spina, V.; Bruscaggin, A.; Cuccaro, A.; Martini, M.; Di Trani, M.; Forestieri, G.; Manzoni, M.; Condoluci, A.; Arribas, A.; Terzi-Di-Bergamo, L.; et al. Circulating tumor DNA reveals genetics, clonal evolution, and residual disease in classical Hodgkin lymphoma. *Blood* **2018**, *131*, 2413–2425. [CrossRef]
11. Camus, V.; Viennot, M.; LeQuesne, J.; Viailly, P.-J.; Bohers, E.; Bessi, L.; Marcq, B.; Etancelin, P.; Dubois, S.; Picquenot, J.-M.; et al. Targeted genotyping of circulating tumor DNA for classical Hodgkin lymphoma monitoring: A prospective study. *Haematologica* **2020**, *106*, 154–162. [CrossRef]
12. Satyal, U.; Srivastava, A.; Abbosh, P.H. Urine Biopsy—Liquid Gold for Molecular Detection and Surveillance of Bladder Cancer. *Front. Oncol.* **2019**, *9*, 1266. [CrossRef]
13. Esposito, A.; Criscitiello, C.; Trapani, D.; Curigliano, G. The Emerging Role of "Liquid Biopsies," Circulating Tumor Cells, and Circulating Cell-Free Tumor DNA in Lung Cancer Diagnosis and Identification of Resistance Mutations. *Curr. Oncol. Rep.* **2017**, *19*, 1. [CrossRef]
14. Elazezy, M.; Joosse, S.A. Techniques of Using Circulating Tumor DNA as a Liquid Biopsy Component in Cancer Management. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 370–378. [CrossRef]
15. Zill, O.A.; Banks, K.; Fairclough, S.R.; Mortimer, S.A.; Vowles, J.V.; Mokhtari, R.; Gandara, D.R.; Mack, P.C.; Odegaard, J.I.; Nagy, R.J.; et al. The Landscape of Actionable Genomic Alterations in Cell-Free Circulating Tumor DNA from 21,807 Advanced Cancer Patients. *Clin. Cancer Res.* **2018**, *24*, 3528–3538. [CrossRef]
16. Stanta, G.; Bonin, S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. *Front. Med.* **2018**, *5*, 85. [CrossRef]
17. Scherer, F.; Kurtz, D.M.; Newman, A.M.; Craig, M.A.; Stehr, H.; Zhou, L.; Glover, C.; Kohrt, H.; Levy, R.; Diehn, M.; et al. Noninvasive Detection of Ibrutinib Resistance in Non-Hodgkin Lymphoma Using Cell-Free DNA. *Blood* **2016**, *128*, 1752. [CrossRef]
18. Bohers, E.; Viailly, P.-J.; Becker, S.; Marchand, V.; Ruminy, P.; Maingonnat, C.; Bertrand, P.; Etancelin, P.; Picquenot, J.-M.; Camus, V.; et al. Non-invasive monitoring of diffuse large B-cell lymphoma by cell-free DNA high-throughput targeted sequencing: Analysis of a prospective cohort. *Blood Cancer J.* **2018**, *8*, 1–13. [CrossRef]
19. Thompson, J.R.; Menon, S.P. Liquid Biopsies and Cancer Immunotherapy. *Cancer J.* **2018**, *24*, 78–83. [CrossRef]
20. Kurtz, D.M.; Scherer, F.; Jin, M.C.; Soo, J.; Craig, A.F.; Esfahani, M.S.; Chabon, J.J.; Stehr, H.; Liu, C.L.; Tibshirani, R.; et al. Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol.* **2018**, *36*, 2845–2853. [CrossRef]
21. Van der Pol, Y.; Mouliere, F. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer Cell* **2019**, *36*, 350–368. [CrossRef]
22. Cheng, F.; Su, L.; Qian, C. Circulating tumor DNA: A promising biomarker in the liquid biopsy of cancer. *Oncotarget* **2016**, *7*, 48832–48841. [CrossRef]
23. Meddeb, R.; Pisareva, E.; Thierry, A.R. Guidelines for the Preanalytical Conditions for Analyzing Circulating Cell-Free DNA. *Clin. Chem.* **2019**, *65*, 623–633. [CrossRef] [PubMed]
24. Diaz, I.M.; Nocon, A.; Mehnert, D.H.; Fredebohm, J.; Diehl, F.; Holtrup, F. Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing. *PLoS ONE* **2016**, *11*, e0166354. [CrossRef]
25. Alidousty, C.; Brandes, D.; Heydt, C.; Wagener, S.; Wittersheim, M.; Schäfer, S.C.; Holz, B.; Merkelbach-Bruse, S.; Büttner, R.; Fassunke, J.; et al. Comparison of Blood Collection Tubes from Three Different Manufacturers for the Collection of Cell-Free DNA for Liquid Biopsy Mutation Testing. *J. Mol. Diagn.* **2017**, *19*, 801–804. [CrossRef]
26. Gahlawat, A.W.; Lenhardt, J.; Witte, T.; Keitel, D.; Kaufhold, A.; Maass, K.K.; Pajtler, K.W.; Sohn, C.; Schott, S. Evaluation of Storage Tubes for Combined Analysis of Circulating Nucleic Acids in Liquid Biopsies. *Int. J. Mol. Sci.* **2019**, *20*, 704. [CrossRef]
27. Zhao, Y.; Li, Y.; Chen, P.; Li, S.; Luo, J.; Xia, H. Performance comparison of blood collection tubes as liquid biopsy storage system for minimizing cfDNA contamination from genomic DNA. *J. Clin. Lab. Anal.* **2019**, *33*, e22670. [CrossRef] [PubMed]
28. El Messaoudi, S.; Rolet, F.; Mouliere, F.; Thierry, A.R. Circulating cell free DNA: Preanalytical considerations. *Clin. Chim. Acta* **2013**, *424*, 222–230. [CrossRef]

4. Stroun, M.; Anker, P.; Maurice, P.; Lyautey, J.; Lederrey, C.; Beljanski, M. Neoplastic Characteristics of the DNA Found in the Plasma of Cancer Patients. *Oncology* **1989**, *46*, 318–322. [CrossRef]
5. Sidransky, D.; Von Eschenbach, A.; Tsai, Y.C.; Jones, P.; Summerhayes, I.; Marshall, F.; Paul, M.; Green, P.; Hamilton, S.R.; Frost, P.; et al. Identification of p53 gene mutations in bladder cancers and urine samples. *Science* **1991**, *252*, 706–709. [CrossRef] [PubMed]
6. Vasioukhin, V.; Anker, P.; Maurice, P.; Lyautey, J.; Lederrey, C.; Stroun, M. Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.* **1994**, *86*, 774–779. [CrossRef]
7. Anker, P.; Lefort, F.; Vasioukhin, V.; Lyautey, J.; Lederrey, C.; Chen, X.Q.; Stroun, M.; Mulcahy, H.E.; Farthing, M.J. K-ras mutations are found in DNA extracted from the plasma of patients with colorectal cancer. *Gastroenterology* **1997**, *112*, 1114–1120. [CrossRef]
8. Austrup, F.; Uciechowski, P.; Eder, C.; Böckmann, B.; Suchy, B.; Driesel, G.; Jäckel, S.; Kusiak, I.; Grill, H.J.; Giesing, M. Prognostic value of genomic alterations in minimal residual cancer cells purified from the blood of breast cancer patients. *Br. J. Cancer* **2000**, *83*, 1664–1673. [CrossRef] [PubMed]
9. Bobillo, S.; Crespo, M.; Escudero, L.; Mayor, R.; Raheja, P.; Carpio, C.; Rubio-Perez, C.; Tazón-Vega, B.; Palacio, C.; Carabia, J.; et al. Cell free circulating tumor DNA in cerebrospinal fluid detects and monitors central nervous system involvement of B-cell lymphomas. *Haematologica* **2020**, *106*, 513–521. [CrossRef]
10. Spina, V.; Bruscaggin, A.; Cuccaro, A.; Martini, M.; Di Trani, M.; Forestieri, G.; Manzoni, M.; Condoluci, A.; Arribas, A.; Terzi-Di-Bergamo, L.; et al. Circulating tumor DNA reveals genetics, clonal evolution, and residual disease in classical Hodgkin lymphoma. *Blood* **2018**, *131*, 2413–2425. [CrossRef]
11. Camus, V.; Viennot, M.; LeQuesne, J.; Viailly, P.-J.; Bohers, E.; Bessi, L.; Marcq, B.; Etancelin, P.; Dubois, S.; Picquenot, J.-M.; et al. Targeted genotyping of circulating tumor DNA for classical Hodgkin lymphoma monitoring: A prospective study. *Haematologica* **2020**, *106*, 154–162. [CrossRef]
12. Satyal, U.; Srivastava, A.; Abbosh, P.H. Urine Biopsy—Liquid Gold for Molecular Detection and Surveillance of Bladder Cancer. *Front. Oncol.* **2019**, *9*, 1266. [CrossRef]
13. Esposito, A.; Criscitiello, C.; Trapani, D.; Curigliano, G. The Emerging Role of "Liquid Biopsies," Circulating Tumor Cells, and Circulating Cell-Free Tumor DNA in Lung Cancer Diagnosis and Identification of Resistance Mutations. *Curr. Oncol. Rep.* **2017**, *19*, 1. [CrossRef]
14. Elazezy, M.; Joosse, S.A. Techniques of Using Circulating Tumor DNA as a Liquid Biopsy Component in Cancer Management. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 370–378. [CrossRef]
15. Zill, O.A.; Banks, K.; Fairclough, S.R.; Mortimer, S.A.; Vowles, J.V.; Mokhtari, R.; Gandara, D.R.; Mack, P.C.; Odegaard, J.I.; Nagy, R.J.; et al. The Landscape of Actionable Genomic Alterations in Cell-Free Circulating Tumor DNA from 21,807 Advanced Cancer Patients. *Clin. Cancer Res.* **2018**, *24*, 3528–3538. [CrossRef]
16. Stanta, G.; Bonin, S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. *Front. Med.* **2018**, *5*, 85. [CrossRef]
17. Scherer, F.; Kurtz, D.M.; Newman, A.M.; Craig, M.A.; Stehr, H.; Zhou, L.; Glover, C.; Kohrt, H.; Levy, R.; Diehn, M.; et al. Noninvasive Detection of Ibrutinib Resistance in Non-Hodgkin Lymphoma Using Cell-Free DNA. *Blood* **2016**, *128*, 1752. [CrossRef]
18. Bohers, E.; Viailly, P.-J.; Becker, S.; Marchand, V.; Ruminy, P.; Maingonnat, C.; Bertrand, P.; Etancelin, P.; Picquenot, J.-M.; Camus, V.; et al. Non-invasive monitoring of diffuse large B-cell lymphoma by cell-free DNA high-throughput targeted sequencing: Analysis of a prospective cohort. *Blood Cancer J.* **2018**, *8*, 1–13. [CrossRef]
19. Thompson, J.R.; Menon, S.P. Liquid Biopsies and Cancer Immunotherapy. *Cancer J.* **2018**, *24*, 78–83. [CrossRef]
20. Kurtz, D.M.; Scherer, F.; Jin, M.C.; Soo, J.; Craig, A.F.; Esfahani, M.S.; Chabon, J.J.; Stehr, H.; Liu, C.L.; Tibshirani, R.; et al. Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol.* **2018**, *36*, 2845–2853. [CrossRef]
21. Van der Pol, Y.; Mouliere, F. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer Cell* **2019**, *36*, 350–368. [CrossRef]
22. Cheng, F.; Su, L.; Qian, C. Circulating tumor DNA: A promising biomarker in the liquid biopsy of cancer. *Oncotarget* **2016**, *7*, 48832–48841. [CrossRef]
23. Meddeb, R.; Pisareva, E.; Thierry, A.R. Guidelines for the Preanalytical Conditions for Analyzing Circulating Cell-Free DNA. *Clin. Chem.* **2019**, *65*, 623–633. [CrossRef] [PubMed]
24. Diaz, I.M.; Nocon, A.; Mehnert, D.H.; Fredebohm, J.; Diehl, F.; Holtrup, F. Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing. *PLoS ONE* **2016**, *11*, e0166354. [CrossRef]
25. Alidousty, C.; Brandes, D.; Heydt, C.; Wagener, S.; Wittersheim, M.; Schäfer, S.C.; Holz, B.; Merkelbach-Bruse, S.; Büttner, R.; Fassunke, J.; et al. Comparison of Blood Collection Tubes from Three Different Manufacturers for the Collection of Cell-Free DNA for Liquid Biopsy Mutation Testing. *J. Mol. Diagn.* **2017**, *19*, 801–804. [CrossRef]
26. Gahlawat, A.W.; Lenhardt, J.; Witte, T.; Keitel, D.; Kaufhold, A.; Maass, K.K.; Pajtler, K.W.; Sohn, C.; Schott, S. Evaluation of Storage Tubes for Combined Analysis of Circulating Nucleic Acids in Liquid Biopsies. *Int. J. Mol. Sci.* **2019**, *20*, 704. [CrossRef]
27. Zhao, Y.; Li, Y.; Chen, P.; Li, S.; Luo, J.; Xia, H. Performance comparison of blood collection tubes as liquid biopsy storage system for minimizing cfDNA contamination from genomic DNA. *J. Clin. Lab. Anal.* **2019**, *33*, e22670. [CrossRef] [PubMed]
28. El Messaoudi, S.; Rolet, F.; Mouliere, F.; Thierry, A.R. Circulating cell free DNA: Preanalytical considerations. *Clin. Chim. Acta* **2013**, *424*, 222–230. [CrossRef]

29. Sorber, L.; Zwaenepoel, K.; Jacobs, J.; De Winne, K.; Goethals, S.; Reclusa, P.; Van Casteren, K.; Augustus, E.; Lardon, F.; Roeyen, G.; et al. Circulating Cell-Free DNA and RNA Analysis as Liquid Biopsy: Optimal Centrifugation Protocol. *Cancers* **2019**, *11*, 458. [CrossRef]
30. Diefenbach, R.J.; Lee, J.; Kefford, R.; Rizos, H. Evaluation of commercial kits for purification of circulating free DNA. *Cancer Genet.* **2018**, *228–229*, 21–27. [CrossRef] [PubMed]
31. Van Dessel, L.F.; Vitale, S.R.; Helmijr, J.C.A.; Wilting, S.M.; Vlugt-Daane, M.; Hoop, E.O.-D.; Sleijfer, S.; Martens, J.W.M.; Jansen, M.P.H.M.; Lolkema, M.P.; et al. High-throughput isolation of circulating tumor DNA: A comparison of automated platforms. *Mol. Oncol.* **2018**, *13*, 392–402. [CrossRef]
32. Jahr, S.; Hentze, H.; Englisch, S.; Hardt, D.; Fackelmayer, F.O.; Hesch, R.D.; Knippers, R. DNA fragments in the blood plasma of cancer patients: Quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* **2001**, *61*, 1659–1665.
33. Henikoff, S.; Church, G.M. Simultaneous Discovery of Cell-Free DNA and the Nucleosome Ladder. *Genetics* **2018**, *209*, 27–29. [CrossRef]
34. Lapin, M.; Oltedal, S.; Tjensvoll, K.; Buhl, T.; Smaaland, R.; Garresori, H.; Javle, M.; Glenjen, N.I.; Abelseth, B.K.; Gilje, B.; et al. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. *J. Transl. Med.* **2018**, *16*, 1–10. [CrossRef] [PubMed]
35. Nikolaev, S.; Lemmens, L.; Koessler, T.; Blouin, J.-L.; Nouspikel, T. Circulating tumoral DNA: Preanalytical validation and quality control in a diagnostic laboratory. *Anal. Biochem.* **2018**, *542*, 34–39. [CrossRef] [PubMed]
36. Devonshire, A.S.; Whale, A.S.; Gutteridge, A.; Jones, G.; Cowen, S.; Foy, C.A.; Huggett, J.F. Towards standardisation of cell-free DNA measurement in plasma: Controls for extraction efficiency, fragment size bias and quantification. *Anal. Bioanal. Chem.* **2014**, *406*, 6499–6512. [CrossRef]
37. Alcaide, M.; Cheung, M.; Hillman, J.; Rassekh, S.R.; Deyell, R.; Batist, G.; Karsan, A.; Wyatt, A.W.; Johnson, N.; Scott, D.W.; et al. Evaluating the quantity, quality and size distribution of cell-free DNA by multiplex droplet digital PCR. *Sci. Rep.* **2020**, *10*, 1–10. [CrossRef] [PubMed]
38. Hohaus, S.; Giachelia, M.; Massini, G.; Mansueto, G.; Vannata, B.; Bozzoli, V.; Criscuolo, M.; D'Alò, F.; Martini, M.; Larocca, L.M.; et al. Cell-free circulating DNA in Hodgkin's and non-Hodgkin's lymphomas. *Ann. Oncol.* **2009**, *20*, 1408–1413. [CrossRef]
39. Aung, K.L.; Donald, E.; Ellison, G.; Bujac, S.; Fletcher, L.; Cantarini, M.; Brady, G.; Orr, M.; Clack, G.; Ranson, M.; et al. Analytical Validation of BRAF Mutation Testing from Circulating Free DNA Using the Amplification Refractory Mutation Testing System. *J. Mol. Diagn.* **2014**, *16*, 343–349. [CrossRef]
40. Siggillino, A.; Ulivi, P.; Pasini, L.; Reda, M.S.; Chiadini, E.; Tofanetti, F.R.; Baglivo, S.; Metro, G.; Crinó, L.; Delmonte, A.; et al. Detection of *EGFR* Mutations in Plasma Cell-Free Tumor DNA of TKI-Treated Advanced-NSCLC Patients by Three Methodologies: Scorpion-ARMS, PNAClamp, and Digital PCR. *Diagnostics* **2020**, *10*, 1062. [CrossRef]
41. Zhang, X.; Chang, N.; Yang, G.; Zhang, Y.; Ye, M.; Cao, J.; Xiong, J.; Han, Z.; Wu, S.; Shang, L.; et al. A comparison of ARMS-Plus and droplet digital PCR for detecting EGFR activating mutations in plasma. *Oncotarget* **2017**, *8*, 112014–112023. [CrossRef] [PubMed]
42. Watanabe, K.; Fukuhara, T.; Tsukita, Y.; Morita, M.; Suzuki, A.; Tanaka, N.; Terasaki, H.; Nukiwa, T.; Maemondo, M. EGFR Mutation Analysis of Circulating Tumor DNA Using an Improved PNA-LNA PCR Clamp Method. *Can. Respir. J.* **2016**, *2016*, 1–7. [CrossRef]
43. Zhang, S.; Chen, Z.; Huang, C.; Ding, C.; Li, C.; Chen, J.; Zhao, J.; Miao, L. Ultrasensitive and quantitative detection ofEGFRmutations in plasma samples from patients with non-small-cell lung cancer using a dual PNA clamping-mediated LNA-PNA PCR clamp. *Analyst* **2019**, *144*, 1718–1724. [CrossRef]
44. Milbury, C.A.; Li, J.; Liu, P.; Makrigiorgos, G.M. COLD-PCR: Improving the sensitivity of molecular diagnostics assays. *Expert Rev. Mol. Diagn.* **2011**, *11*, 159–169. [CrossRef]
45. Galbiati, S.; Damin, F.; Burgio, V.; Brisci, A.; Soriani, N.; Belcastro, B.; Di Resta, C.; Gianni, L.; Chiari, M.; Ronzoni, M.; et al. Evaluation of three advanced methodologies, COLD-PCR, microarray and ddPCR, for identifying the mutational status by liquid biopsies in metastatic colorectal cancer patients. *Clin. Chim. Acta* **2019**, *489*, 136–143. [CrossRef]
46. Wilson, W.H.; Young, R.M.; Schmitz, R.; Yang, Y.; Pittaluga, S.; Wright, G.; Lih, C.-J.; Williams, P.M.; Shaffer, A.L.; Gerecitano, J.; et al. Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. *Nat. Med.* **2015**, *21*, 922–926. [CrossRef]
47. Hiemcke-Jiwa, L.S.; Minnema, M.C.; Loon, J.H.R.-V.; Jiwa, N.M.; De Boer, M.; Leguit, R.J.; De Weger, R.A.; Huibers, M.M. The use of droplet digital PCR in liquid biopsies: A highly sensitive technique for MYD88 p.(L265P) detection in cerebrospinal fluid. *Hematol. Oncol.* **2018**, *36*, 429–435. [CrossRef]
48. Chen, K.; Ma, Y.; Ding, T.; Zhang, X.; Chen, B.; Guan, M. Effectiveness of digital PCR for MYD88L265P detection in vitreous fluid for primary central nervous system lymphoma diagnosis. *Exp. Ther. Med.* **2020**, *20*, 301–308. [CrossRef]
49. Dressman, D.; Yan, H.; Traverso, G.; Kinzler, K.W.; Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 8817–8822. [CrossRef]
50. Diehl, F.; Schmidt, K.; Choti, M.A.; Romans, K.; Goodman, S.; Li, M.; Thornton, K.; Agrawal, N.; Sokoll, L.; Szabo, S.A.; et al. Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **2008**, *14*, 985–990. [CrossRef]
51. Holdhoff, M.; Schmidt, K.; Donehower, R.; Diaz, L.A. Analysis of Circulating Tumor DNA to Confirm Somatic KRAS Mutations. *J. Natl. Cancer Inst.* **2009**, *101*, 1284–1285. [CrossRef]

52. O'Leary, B.; Hrebien, S.; Beaney, M.; Fribbens, C.; Garcia-Murillas, I.; Jiang, J.; Li, Y.; Bartlett, C.H.; Andre, F.; Loibl, S.; et al. Comparison of BEAMing and Droplet Digital PCR for Circulating Tumor DNA Analysis. *Clin. Chem.* **2019**, *65*, 1405–1413. [CrossRef]
53. Garcia, J.; Gauthier, A.; Lescuyer, G.; Barthelemy, D.; Geiguer, F.; Balandier, J.; Edelstein, D.L.; Jones, F.S.; Holtrup, F.; Duruisseau, M.; et al. Routine Molecular Screening of Patients with Advanced Non-SmallCell Lung Cancer in Circulating Cell-Free DNA at Diagnosis and During Progression Using OncoBEAMTM EGFR V2 and NGS Technologies. *Mol. Diagn. Ther.* **2021**, *25*, 239–250. [CrossRef]
54. Butler, T.; Spellman, P.T.; Gray, J. Circulating-tumor DNA as an early detection and diagnostic tool. *Curr. Opin. Genet. Dev.* **2017**, *42*, 14–21. [CrossRef] [PubMed]
55. Milbury, C.A.; Zhong, Q.; Lin, J.; Williams, M.; Olson, J.; Link, D.R.; Hutchison, B. Determining lower limits of detection of digital PCR assays for cancer-related gene mutations. *Biomol. Detect. Quantif.* **2014**, *1*, 8–22. [CrossRef]
56. Shoda, K.; Ichikawa, D.; Fujita, Y.; Masuda, K.; Hiramoto, H.; Hamada, J.; Arita, T.; Konishi, H.; Komatsu, S.; Shiozaki, A.; et al. Monitoring the HER2 copy number status in circulating tumor DNA by droplet digital PCR in patients with gastric cancer. *Gastric Cancer* **2016**, *20*, 126–135. [CrossRef]
57. Lee, K.S.; Nam, S.K.; Seo, S.H.; Park, K.U.; Oh, H.-K.; Kim, D.-W.; Kang, S.-B.; Kim, W.H.; Lee, H.S. Digital polymerase chain reaction for detecting c-MYC copy number gain in tissue and cell-free plasma samples of colorectal cancer patients. *Sci. Rep.* **2019**, *9*, 1–9. [CrossRef]
58. Delfau-Larue, M.-H.; Van Der Gucht, A.; Dupuis, J.; Jais, J.-P.; Nel, I.; Beldi-Ferchiou, A.; Hamdane, S.; Benmaad, I.; Laboure, G.; Verret, B.; et al. Total metabolic tumor volume, circulating tumor cells, cell-free DNA: Distinct prognostic value in follicular lymphoma. *Blood Adv.* **2018**, *2*, 807–816. [CrossRef]
59. Pott, C.; Brüggemann, M.; Ritgen, M.; Van Der Velden, V.H.J.; Van Dongen, J.J.M.; Kneba, M. MRD Detection in B-Cell Non-Hodgkin Lymphomas Using Ig Gene Rearrangements and Chromosomal Translocations as Targets for Real-Time Quantitative PCR. *Meth. Mol. Biol.* **2019**, 199–228. [CrossRef]
60. Pyrak, E.; Krajczewski, J.; Kowalik, A.; Kudelski, A.; Jaworska, A. Surface Enhanced Raman Spectroscopy for DNA Biosensors—How Far Are We? *Molecular* **2019**, *24*, 4423. [CrossRef] [PubMed]
61. Wee, E.J.; Wang, Y.; Tsao, S.C.-H.; Trau, M. Simple, Sensitive and Accurate Multiplex Detection of Clinically Important Melanoma DNA Mutations in Circulating Tumour DNA with SERS Nanotags. *Theranostics* **2016**, *6*, 1506–1513. [CrossRef] [PubMed]
62. Lyu, N.; Rajendran, V.K.; Diefenbach, R.; Charles, K.; Clarke, S.J.; Engel, A.; Rizos, H.; Molloy, M.P.; Wang, Y. Sydney 1000 Colorectal Cancer Study Investigators* Multiplex detection of ctDNA mutations in plasma of colorectal cancer patients by PCR/SERS assay. *Nanotheranostics* **2020**, *4*, 224–232. [CrossRef]
63. Gray, E.S.; Witkowski, T.; Pereira, M.; Calapre, L.; Herron, K.; Irwin, D.; Chapman, B.; Khattak, M.A.; Raleigh, J.; Hatzimihalis, A.; et al. Genomic Analysis of Circulating Tumor DNA Using a Melanoma-Specific UltraSEEK Oncogene Panel. *J. Mol. Diagn.* **2019**, *21*, 418–426. [CrossRef] [PubMed]
64. Lamy, P.-J.; Van Der Leest, P.; Lozano, N.; Becht, C.; Duboeuf, F.; Groen, H.J.M.; Hilgers, W.; Pourel, N.; Rifaela, N.; Schuuring, E.; et al. Mass Spectrometry as a Highly Sensitive Method for Specific Circulating Tumor DNA Analysis in NSCLC: A Comparison Study. *Cancers* **2020**, *12*, 3002. [CrossRef] [PubMed]
65. Mamanova, L.; Coffey, A.J.; Scott, C.E.; Kozarewa, I.; Turner, E.; Kumar, A.; Howard, E.; Shendure, J.; Turner, D.J. Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **2010**, *7*, 111–118. [CrossRef]
66. Deveson, I.W.; Gong, B.; Lai, K.; LoCoco, J.S.; Richmond, T.A.; Schageman, J.; Zhang, Z.; Novoradovskaya, N.; Willey, J.C.; Jones, W.; et al. Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology. *Nat. Biotechnol.* **2021**, 1–14. [CrossRef]
67. Glenn, T.C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **2011**, *11*, 759–769. [CrossRef]
68. Loman, N.J.; Misra, R.V.; Dallman, T.J.; Constantinidou, C.; Gharbia, S.E.; Wain, J.; Pallen, M.J. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **2012**, *30*, 434–439. [CrossRef]
69. Kinde, I.; Wu, J.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9530–9535. [CrossRef]
70. Phallen, J.; Sausen, M.; Adleff, V.; Leal, A.; Hruban, C.; White, J.; Anagnostou, V.; Fiksel, J.; Cristiano, S.; Papp, E.; et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **2017**, *9*, eaan2415. [CrossRef]
71. Xu, C.; Ranjbar, M.R.N.; Wu, Z.; Dicarlo, J.; Wang, Y. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. *BMC Genom.* **2017**, *18*, 5. [CrossRef]
72. Salk, J.J.; Schmitt, M.W.; Loeb, L.A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* **2018**, *19*, 269–285. [CrossRef]
73. Forshew, T.; Murtaza, M.; Parkinson, C.; Gale, D.; Tsui, D.W.Y.; Kaper, F.; Dawson, S.-J.; Piskorz, A.M.; Jimenez-Linan, M.; Bentley, D.; et al. Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. *Sci. Transl. Med.* **2012**, *4*, 136ra68. [CrossRef]
74. Gale, D.; Lawson, A.R.J.; Howarth, K.; Madi, M.; Durham, B.; Smalley, S.; Calaway, J.; Blais, S.; Jones, G.; Clark, J.; et al. Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free DNA. *PLoS ONE* **2018**, *13*, e0194630. [CrossRef] [PubMed]

75. Fostira, F.; Oikonomopoulou, P.; Kladi, A.; Edelstein, D.; Stieler, K.; Heim, D.; Gkotzamanidou, M.; Anastasiou, M.; Kotsantis, I.; Kavourakis, G.; et al. Blood-based testing of mutations in patients with head and neck squamous cell carcinoma (HNSCC) using highly sensitive SafeSEQ technology. *Ann. Oncol.* **2019**, *30*, v469. [CrossRef]
76. Tie, J.; Cohen, J.D.; Lo, S.N.; Wang, Y.; Li, L.; Christie, M.; Lee, M.; Wong, R.; Kosmider, S.; Skinner, I.; et al. Prognostic significance of postsurgery circulating tumor DNA in nonmetastatic colorectal cancer: Individual patient pooled analysis of three cohort studies. *Int. J. Cancer* **2021**, *148*, 1014–1026. [CrossRef] [PubMed]
77. Schmitt, M.W.; Kennedy, S.R.; Salk, J.J.; Fox, E.; Hiatt, J.B.; Loeb, L.A. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 14508–14513. [CrossRef] [PubMed]
78. Kennedy, S.R.; Schmitt, M.W.; Fox, E.J.; Kohrn, B.; Salk, J.J.; Ahn, E.H.; Prindle, M.J.; Kuong, K.J.; Shen, J.-C.; Risques, R.-A.; et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **2014**, *9*, 2586–2606. [CrossRef]
79. Costello, M.; Pugh, T.J.; Fennell, T.J.; Stewart, C.; Lichtenstein, L.; Meldrim, J.C.; Fostel, J.L.; Friedrich, D.C.; Perrin, D.; Dionne, D.; et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **2013**, *41*, e67. [CrossRef]
80. Alcaide, M.; Yu, S.; Davidson, J.; Albuquerque, M.; Bushell, K.; Fornika, D.; Arthur, S.; Grande, B.M.; McNamara, S.; du Tertre, M.C.; et al. Targeted error-suppressed quantification of circulating tumor DNA using semi-degenerate barcoded adapters and biotinylated baits. *Sci. Rep.* **2017**, *7*, 10574. [CrossRef]
81. Ren, Y.; Zhang, Y.; Wang, D.; Liu, F.; Fu, Y.; Xiang, S.; Su, L.; Li, J.; Dai, H.; Huang, B. SinoDuplex: An Improved Duplex Sequencing Approach to Detect Low-frequency Variants in Plasma cfDNA Samples. *Genom. Proteom. Bioinform.* **2020**, *18*, 81–90. [CrossRef] [PubMed]
82. Mallampati, S.; Zalles, S.; Duose, D.Y.; Hu, P.C.; Medeiros, L.J.; Wistuba, I.I.; Kopetz, S.; Luthra, R. Development and Application of Duplex Sequencing Strategy for Cell-Free DNA–Based Longitudinal Monitoring of Stage IV Colorectal Cancer. *J. Mol. Diagn.* **2019**, *21*, 994–1009. [CrossRef] [PubMed]
83. Peng, Q.; Xu, C.; Kim, D.; Lewis, M.; Dicarlo, J.; Wang, Y. Targeted Single Primer Enrichment Sequencing with Single End Duplex-UMI. *Sci. Rep.* **2019**, *9*, 4810. [CrossRef] [PubMed]
84. Sim, W.C.; Loh, C.H.; Toh, G.L.-X.; Lim, C.W.; Chopra, A.; Chang, A.Y.C.; Goh, L.L. Non-invasive detection of actionable mutations in advanced non-small-cell lung cancer using targeted sequencing of circulating tumor DNA. *Lung Cancer* **2018**, *124*, 154–159. [CrossRef]
85. Zou, D.; Day, R.; Cocadiz, J.A.; Parackal, S.; Mitchell, W.; Black, M.A.; Lawrence, B.; Fitzgerald, S.; Print, C.; Jackson, C.; et al. Circulating tumor DNA is a sensitive marker for routine monitoring of treatment response in advanced colorectal cancer. *Carcinogenesis* **2020**, *41*, 1507–1517. [CrossRef]
86. Viailly, P.-J.; Sater, V.; Viennot, M.; Bohers, E.; Vergne, N.; Berard, C.; Dauchel, H.; Lecroq, T.; Celebi, A.; Ruminy, P.; et al. Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers. *BMC Bioinform.* **2021**, *22*, 1–15. [CrossRef]
87. Newman, A.; Bratman, S.V.; To, J.; Wynne, J.F.; Eclov, N.C.W.; Modlin, L.A.; Liu, C.L.; Neal, J.W.; Wakelee, H.A.; Merritt, R.E.; et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **2014**, *20*, 548–554. [CrossRef]
88. Bratman, S.V.; Newman, A.; Alizadeh, A.A.; Diehn, M. Potential clinical utility of ultrasensitive circulating tumor DNA detection with CAPP-Seq. *Expert Rev. Mol. Diagn.* **2015**, *15*, 715–719. [CrossRef]
89. Klass, D.; Newman, A.; Lovejoy, A.; Zhou, L.; Stehr, H.; Xu, T.; He, J.; Komaki, R.; Liao, Z.; Maru, D.; et al. Analysis of Circulating Tumor DNA in Esophageal Carcinoma Patients Treated with Chemoradiation Therapy. *Int. J. Radiat. Oncol.* **2015**, *93*, S104–S105. [CrossRef]
90. Scherer, F.; Kurtz, D.M.; Newman, A.; Stehr, H.; Craig, A.F.M.; Esfahani, M.S.; Lovejoy, A.F.; Chabon, J.J.; Klass, D.M.; Liu, C.L.; et al. Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. *Sci. Transl. Med.* **2016**, *8*, 364ra155. [CrossRef]
91. Dudley, J.C.; Schroers-Martin, J.; Lazzareschi, D.V.; Shi, W.Y.; Chen, S.B.; Esfahani, M.S.; Trivedi, D.; Chabon, J.J.; Chaudhuri, A.A.; Stehr, H.; et al. Detection and Surveillance of Bladder Cancer Using Urine Tumor DNA. *Cancer Discov.* **2019**, *9*, 500–509. [CrossRef]
92. Newman, A.; Lovejoy, A.F.; Klass, D.M.; Kurtz, D.M.; Chabon, J.J.; Scherer, F.; Stehr, H.; Liu, C.L.; Bratman, S.V.; Say, C.; et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **2016**, *34*, 547–555. [CrossRef]
93. Roschewski, M.; Dunleavy, K.; Pittaluga, S.; Moorhead, M.; Pepin, F.; Kong, K.; Shovlin, M.; Jaffe, E.S.; Staudt, L.M.; Lai, C.; et al. Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: A correlative biomarker study. *Lancet Oncol.* **2015**, *16*, 541–549. [CrossRef]
94. Kurtz, D.M.; Green, M.R.; Bratman, S.V.; Scherer, F.; Liu, C.L.; Kunder, C.A.; Takahashi, K.; Glover, C.; Keane, C.; Kihira, S.; et al. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood* **2015**, *125*, 3679–3687. [CrossRef]
95. Sarkozy, C.; Huet, S.; Carlton, V.E.; Fabiani, B.; Delmer, A.; Jardin, F.; Delfau-Larue, M.-H.; Hacini, M.; Ribrag, V.; Guidez, S.; et al. The prognostic value of clonal heterogeneity and quantitative assessment of plasma circulating clonal IG-VDJ sequences at diagnosis in patients with follicular lymphoma. *Oncotarget* **2017**, *8*, 8765–8774. [CrossRef]

96. Hossain, N.M.; Dahiya, S.; Le, R.; Abramian, A.M.; Kong, K.A.; Muffly, L.S.; Miklos, D.B. Circulating tumor DNA assessment in patients with diffuse large B-cell lymphoma following CAR T-cell therapy. *Leuk. Lymphoma* **2019**, *60*, 503–506. [CrossRef] [PubMed]
97. Corcoran, R.B.; Chabner, B.A. Application of Cell-free DNA Analysis to Cancer Treatment. *N. Engl. J. Med.* **2018**, *379*, 1754–1765. [CrossRef] [PubMed]
98. Bos, M.K.; Angus, L.; Nasserinejad, K.; Jager, A.; Jansen, M.P.; Martens, J.W.; Sleijfer, S. Whole exome sequencing of cell-free DNA—A systematic review and Bayesian individual patient data meta-analysis. *Cancer Treat. Rev.* **2020**, *83*, 101951. [CrossRef]
99. Choi, M.; Scholl, U.I.; Ji, W.; Liu, T.; Tikhonova, I.R.; Zumbo, P.; Nayir, A.; Bakkaloğlu, A.; Özen, S.; Sanjad, S.; et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19096–19101. [CrossRef]
100. Murtaza, M.; Dawson, S.-J.; Tsui, D.W.Y.; Gale, D.; Forshew, T.; Piskorz, A.M.; Parkinson, C.; Chin, S.-F.; Kingsbury, Z.; Wong, A.S.C.; et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nat. Cell Biol.* **2013**, *497*, 108–112. [CrossRef] [PubMed]
101. Heidary, M.; Auer, M.; Ulz, P.; Heitzer, E.; Petru, E.; Gasch, C.; Riethdorf, S.; Mauermann, O.; Lafer, I.; Pristauz, G.; et al. The dynamic range of circulating tumor DNA in metastatic breast cancer. *Breast Cancer Res.* **2014**, *16*, 1–10. [CrossRef]
102. Murtaza, M.; Dawson, S.; Pogrebniak, K.; Rueda, O.M.; Provenzano, E.; Grant, J.; Chin, S.-F.; Tsui, D.W.Y.; Marass, F.; Gale, D.; et al. Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.* **2015**, *6*, 8760. [CrossRef]
103. Lebofsky, R.; Decraene, C.; Bernard, V.; Kamal, M.; Blin, A.; Leroy, Q.; Frio, T.R.; Pierron, G.; Callens, C.; Bieche, I.; et al. Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. *Mol. Oncol.* **2015**, *9*, 783–790. [CrossRef]
104. Chicard, M.; Colmet-Daage, L.; Clement, N.; Danzon, A.; Bohec, M.; Bernard, V.; Baulande, S.; Bellini, A.; Eve, L.; Pierron, G.; et al. Whole-Exome Sequencing of Cell-Free DNA Reveals Temporo-spatial Heterogeneity and Identifies Treatment-Resistant Clones in Neuroblastoma. *Clin. Cancer Res.* **2018**, *24*, 939–949. [CrossRef]
105. Heitzer, E.; Ulz, P.; Belic, J.; Gutschi, S.; Quehenberger, F.; Fischereder, K.; Benezeder, T.; Auer, M.; Pischler, C.; Mannweiler, S.; et al. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Med.* **2013**, *5*, 30. [CrossRef] [PubMed]
106. Raman, L.; Van Der Linden, M.; De Vriendt, C.; Broeck, B.V.D.; Muylle, K.; Deeren, D.; Dedeurwaerdere, F.; Verbeke, S.; Dendooven, A.; De Grove, K.; et al. Shallow-depth sequencing of cell-free DNA for Hodgkin and diffuse large B-cell lymphoma (differential) diagnosis: A standardized approach with underappreciated potential. *Haematologica* **2020**. [CrossRef]
107. Kinde, I.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B. FAST-SeqS: A Simple and Efficient Method for the Detection of Aneuploidy by Massively Parallel Sequencing. *PLoS ONE* **2012**, *7*, e41162. [CrossRef] [PubMed]
108. Douville, C.; Springer, S.; Kinde, I.; Cohen, J.D.; Hruban, R.H.; Lennon, A.M.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B.; Karchin, R. Detection of aneuploidy in patients with cancer through amplification of long interspersed nucleotide elements (LINEs). *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 1871–1876. [CrossRef] [PubMed]
109. Leary, R.J.; Kinde, I.; Diehl, F.; Schmidt, K.; Clouser, C.; Duncan, C.; Antipova, A.; Lee, C.; McKernan, K.; De La Vega, F.M.; et al. Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. *Sci. Transl. Med.* **2010**, *2*, 20ra14. [CrossRef]
110. Leary, R.J.; Sausen, M.; Kinde, I.; Papadopoulos, N.; Carpten, J.D.; Craig, D.; O'Shaughnessy, J.; Kinzler, K.W.; Parmigiani, G.; Vogelstein, B.; et al. Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing. *Sci. Transl. Med.* **2012**, *4*, 162ra154. [CrossRef] [PubMed]
111. Kim, Y.-W.; Song, Y.; Kim, H.-S.; Sim, H.W.; Poojan, S.; Eom, B.W.; Kook, M.-C.; Joo, J.; Hong, K.-M.; Kim, Y.-H. Monitoring circulating tumor DNA by analyzing personalized cancer-specific rearrangements to detect recurrence in gastric cancer. *Exp. Mol. Med.* **2019**, *51*, 1–10. [CrossRef]
112. Snyder, M.W.; Kircher, M.; Hill, A.J.; Daza, R.M.; Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **2016**, *164*, 57–68. [CrossRef] [PubMed]
113. Mouliere, F.; Chandrananda, D.; Piskorz, A.M.; Moore, E.K.; Morris, J.; Ahlborn, L.B.; Mair, R.; Goranova, T.; Marass, F.; Heider, K.; et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **2018**, *10*, eaat4921. [CrossRef]
114. Cristiano, S.; Leal, A.; Phallen, J.; Fiksel, J.; Adleff, V.; Bruhm, D.C.; Jensen, S.Ø.; Medina, J.E.; Hruban, C.; White, J.R.; et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nat. Cell Biol.* **2019**, *570*, 385–389. [CrossRef]
115. Mehrmohamadi, M.; Esfahani, M.S.; Soo, J.; Scherer, F.; Schroers-Martin, J.G.; Chen, B.; Kurtz, D.M.; Hamilton, E.; Liu, C.L.; Diehn, M.; et al. Distinct Chromatin Accessibility Profiles of Lymphoma Subtypes Revealed By Targeted Cell Free DNA Profiling. *Blood* **2018**, *132*, 672. [CrossRef]
116. Ørntoft, M.-B.W.; Jensen, S.Ø.; Hansen, T.B.; Bramsen, J.B.; Andersen, C.L. Comparative analysis of 12 different kits for bisulfite conversion of circulating cell-free DNA. *Epigenetics* **2017**, *12*, 626–636. [CrossRef]
117. Cirillo, M.; Craig, A.F.; Borchmann, S.; Kurtz, D.M. Liquid biopsy in lymphoma: Molecular methods and clinical applications. *Cancer Treat. Rev.* **2020**, *91*, 102106. [CrossRef] [PubMed]

118. Shen, S.Y.; Singhania, R.; Fehringer, G.; Chakravarthy, A.; Roehrl, M.H.A.; Chadwick, D.; Zuzarte, P.C.; Borgida, A.; Wang, T.T.; Li, T.; et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nat. Cell Biol.* **2018**, *563*, 579–583. [CrossRef]
119. Chen, X.; Gole, J.; Gore, A.; He, Q.; Lu, M.; Min, J.; Yuan, Z.; Yang, X.; Jiang, Y.; Zhang, T.; et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.* **2020**, *11*, 1–10. [CrossRef] [PubMed]
120. Kristensen, L.S.; Hansen, J.W.; Kristensen, S.S.; Tholstrup, D.; Harsløf, L.B.S.; Pedersen, O.B.; Brown, P.D.N.; Grønbæk, K. Aberrant methylation of cell-free circulating DNA in plasma predicts poor outcome in diffuse large B cell lymphoma. *Clin. Epigenet.* **2016**, *8*, 1–11. [CrossRef]
121. Wedge, E.; Hansen, J.W.; Garde, C.; Asmar, F.; Tholstrup, D.; Kristensen, S.S.; Munch-Petersen, H.D.; Ralfkiaer, E.; Brown, P.; Grønbaek, K.; et al. Global hypomethylation is an independent prognostic factor in diffuse large B cell lymphoma. *Am. J. Hematol.* **2017**, *92*, 689–694. [CrossRef] [PubMed]
122. Van der Auwera, G.A.; O'Connor, B.D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st ed.; O'Reilly Media: Cambridge, MA, USA, 2020.
123. Kechin, A.; Boyarskikh, U.; Kel, A.; Filipenko, M. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J. Comput. Biol.* **2017**, *24*, 1138–1143. [CrossRef]
124. Schmieder, R.; Lim, Y.W.; Rohwer, F.; Edwards, R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinform.* **2010**, *11*, 341. [CrossRef]
125. Babraham Bioinformatics—Trim Galore! Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 11 May 2021).
126. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
127. Smith, T.; Heger, A.; Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **2017**, *27*, 491–499. [CrossRef]
128. Sater, V.; Viailly, P.-J.; Lecroq, T.; Prieur-Gaston, É.; Bohers, É.; Viennot, M.; Ruminy, P.; Dauchel, H.; Vera, P.; Jardin, F. UMI-VarCal: A new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. *Bioinform. Oxf. Engl.* **2020**, *36*, 2718–2724. [CrossRef]
129. Boeva, V.; Popova, T.; Lienard, M.; Toffoli, S.; Kamal, M.; Le Tourneau, C.; Gentien, D.; Servant, N.; Gestraud, P.; Frio, T.R.; et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinform. Oxf. Engl.* **2014**, *30*, 3443–3450. [CrossRef] [PubMed]
130. Sater, V.; Viailly, P.-J.; Lecroq, T.; Ruminy, P.; Bérard, C.; Prieur-Gaston, É.; Jardin, F. UMI-Gen: A UMI-based read simulator for variant calling evaluation in paired-end sequencing NGS libraries. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2270–2280. [CrossRef]
131. Volckmar, A.-L.; Sültmann, H.; Riediger, A.; Fioretos, T.; Schirmacher, P.; Endris, V.; Stenzinger, A.; Dietz, S. A field guide for cancer diagnostics using cell-free DNA: From principles to practice and clinical applications. *Genes Chromosom. Cancer* **2018**, *57*, 123–139. [CrossRef] [PubMed]

VI. BIBLIOGRAPHIE

[1] D. Luo, T. Zhou, Y. Tao, Y. Feng, X. Shen, and S. Mei, “Exposure to organochlorine pesticides and non-Hodgkin lymphoma: a meta-analysis of observational studies,” *Sci. Rep.*, vol. 6, p. 25768, May 2016, doi: 10.1038/srep25768.

[2] N. L. Harris *et al.*, “A revised European-American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group,” *Blood*, vol. 84, no. 5, pp. 1361–1392, Sep. 1994.

[3] J. J. M. van Dongen *et al.*, “Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936,” *Leukemia*, vol. 17, no. 12, pp. 2257–2317, Dec. 2003, doi: 10.1038/sj.leu.2403202.

[4] F. Hamadeh, S. P. MacNamara, N. S. Aguilera, S. H. Swerdlow, and J. R. Cook, “MYD88 L265P mutation analysis helps define nodal lymphoplasmacytic lymphoma,” *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc*, vol. 28, no. 4, pp. 564–574, Apr. 2015, doi: 10.1038/modpathol.2014.120.

[5] S. Dubois *et al.*, “Biological and Clinical Relevance of Associated Genomic Alterations in MYD88 L265P and non-L265P-Mutated Diffuse Large B-Cell Lymphoma: Analysis of 361 Cases,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 23, no. 9, pp. 2232–2244, May 2017, doi: 10.1158/1078-0432.CCR-16-1922.

[6] E. Tiacci *et al.*, “Simple genetic diagnosis of hairy cell leukemia by sensitive detection of the BRAF-V600E mutation,” *Blood*, vol. 119, no. 1, pp. 192–195, Jan. 2012, doi: 10.1182/blood-2011-08-371179.

[7] E. Tiacci *et al.*, “BRAF mutations in hairy-cell leukemia,” *N. Engl. J. Med.*, vol. 364, no. 24, pp. 2305–2315, Jun. 2011, doi: 10.1056/NEJMoa1014209.

[8] F. Jardin *et al.*, “Diffuse large B-cell lymphomas with CDKN2A deletion have a distinct gene expression signature and a poor prognosis under R-CHOP treatment: a GELA study,” *Blood*, vol. 116, no. 7, pp. 1092–1104, Aug. 2010, doi: 10.1182/blood-2009-10-247122.

[9] V. Bobée *et al.*, “Determination of Molecular Subtypes of Diffuse Large B-Cell Lymphoma Using a Reverse Transcriptase Multiplex Ligation-Dependent Probe Amplification Classifier: A CALYM Study,” *J. Mol. Diagn. JMD*, vol. 19, no. 6, pp. 892–904, Nov. 2017, doi: 10.1016/j.jmoldx.2017.07.007.

[10] F. Drieux *et al.*, “Defining signatures of peripheral T-cell lymphoma with a targeted 20-marker gene expression profiling assay,” *Haematologica*, vol. 105, no. 6, pp. 1582–1592, Jun. 2020, doi: 10.3324/haematol.2019.226647.

[11] V. Bobée *et al.*, “Combining gene expression profiling and machine learning to diagnose B-cell non-Hodgkin lymphoma,” *Blood Cancer J.*, vol. 10, no. 5, p. 59, May 2020, doi: 10.1038/s41408-020-0322-5.

[12] S. Huet *et al.*, “A gene-expression profiling score for prediction of outcome in patients with follicular lymphoma: a retrospective training and validation analysis in three international cohorts,” *Lancet Oncol.*, vol. 19, no. 4, pp. 549–561, Apr. 2018, doi: 10.1016/S1470-2045(18)30102-5.

[13] S. SH *et al.*, *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. Accessed: Jul. 06, 2021. [Online]. Available: https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-Haematopoietic-And-Lymphoid-Tissues-2017

[14] I. S. Lossos and R. D. Gascoyne, “Transformation of follicular lymphoma,” *Best Pract. Res. Clin. Haematol.*, vol. 24, no. 2, pp. 147–163, Jun. 2011, doi: 10.1016/j.beha.2011.02.006.

[15] R. L. M. C. Agbay, S. Loghavi, L. J. Medeiros, and J. D. Khoury, “High-grade Transformation of Low-grade B-cell Lymphoma: Pathology and Molecular Pathogenesis,” *Am. J. Surg. Pathol.*, vol. 40, no. 1, pp. e1-16, Jan. 2016, doi: 10.1097/PAS.0000000000000561.

[16] L. H. Sehn *et al.*, “Introduction of combined CHOP plus rituximab therapy dramatically improved outcome of diffuse large B-cell lymphoma in British Columbia,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 23, no. 22, pp. 5027–5033, Aug. 2005, doi: 10.1200/JCO.2005.09.137.

[17] B. Coiffier *et al.*, “Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the Groupe d’Etudes des Lymphomes de l’Adulte,” *Blood*, vol. 116, no. 12, pp. 2040–2045, Sep. 2010, doi: 10.1182/blood-2010-03-276246.

[18] L. H. Sehn and R. D. Gascoyne, “Diffuse large B-cell lymphoma: optimizing outcome in the context of clinical and biologic heterogeneity,” *Blood*, vol.

125, no. 1, pp. 22–32, Jan. 2015, doi: 10.1182/blood-2014-05-577189.
[19] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray,” *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995, doi: 10.1126/science.270.5235.467.
[20] J. DeRisi *et al.*, “Use of a cDNA microarray to analyse gene expression patterns in human cancer,” *Nat. Genet.*, vol. 14, no. 4, pp. 457–460, Dec. 1996, doi: 10.1038/ng1296-457.
[21] A. Alizadeh *et al.*, “The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes,” *Cold Spring Harb. Symp. Quant. Biol.*, vol. 64, pp. 71–78, 1999, doi: 10.1101/sqb.1999.64.71.
[22] A. A. Alizadeh *et al.*, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, no. 6769, pp. 503–511, Feb. 2000, doi: 10.1038/35000501.
[23] A. Rosenwald *et al.*, “The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma,” *N. Engl. J. Med.*, vol. 346, no. 25, pp. 1937–1947, Jun. 2002, doi: 10.1056/NEJMoa012914.
[24] G. Wright, B. Tan, A. Rosenwald, E. H. Hurt, A. Wiestner, and L. M. Staudt, “A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 17, pp. 9991–9996, Aug. 2003, doi: 10.1073/pnas.1732008100.
[25] G. Lenz *et al.*, “Stromal gene signatures in large-B-cell lymphomas,” *N. Engl. J. Med.*, vol. 359, no. 22, pp. 2313–2323, Nov. 2008, doi: 10.1056/NEJMoa0802885.
[26] R. Coutinho *et al.*, “Poor concordance among nine immunohistochemistry classifiers of cell-of-origin for diffuse large B-cell lymphoma: implications for therapeutic strategies,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 19, no. 24, pp. 6686–6695, Dec. 2013, doi: 10.1158/1078-0432.CCR-13-1482.
[27] S. Mareschal *et al.*, “Accurate Classification of Germinal Center B-Cell-Like/Activated B-Cell-Like Diffuse Large B-Cell Lymphoma Using a Simple and Rapid Reverse Transcriptase-Multiplex Ligation-Dependent Probe Amplification Assay: A CALYM Study,” *J. Mol. Diagn. JMD*, pp. S1525-1578(15)00046-X, Apr. 2015, doi: 10.1016/j.jmoldx.2015.01.007.
[28] D. W. Scott *et al.*, “Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue,” *Blood*, vol. 123, no. 8, pp. 1214–1217, Feb. 2014, doi: 10.1182/blood-2013-11-536433.
[29] G. Lenz and L. M. Staudt, “Aggressive lymphomas,” *N. Engl. J. Med.*, vol. 362, no. 15, pp. 1417–1429, Apr. 2010, doi: 10.1056/NEJMra0807082.
[30] D. W. Scott *et al.*, “Prognostic Significance of Diffuse Large B-Cell Lymphoma Cell of Origin Determined by Digital Gene Expression in Formalin-Fixed Paraffin-Embedded Tissue Biopsies,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 33, no. 26, pp. 2848–2856, Sep. 2015, doi: 10.1200/JCO.2014.60.2383.
[31] K. Dunleavy *et al.*, “Differential efficacy of bortezomib plus chemotherapy within molecular subtypes of diffuse large B-cell lymphoma,” *Blood*, vol. 113, no. 24, pp. 6069–6076, Jun. 2009, doi: 10.1182/blood-2009-01-199679.
[32] L. A. Mathews Griner *et al.*, “High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 6, pp. 2349–2354, Feb. 2014, doi: 10.1073/pnas.1311846111.
[33] G. S. Nowakowski *et al.*, “Lenalidomide can be safely combined with R-CHOP (R2CHOP) in the initial chemotherapy for aggressive B-cell lymphomas: phase I study,” *Leukemia*, vol. 25, no. 12, pp. 1877–1881, Dec. 2011, doi: 10.1038/leu.2011.165.
[34] G. S. Nowakowski *et al.*, “Lenalidomide combined with R-CHOP overcomes negative prognostic impact of non-germinal center B-cell phenotype in newly diagnosed diffuse large B-Cell lymphoma: a phase II study,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 33, no. 3, pp. 251–257, Jan. 2015, doi: 10.1200/JCO.2014.55.5714.
[35] F. Offner *et al.*, “Frontline rituximab, cyclophosphamide, doxorubicin, and prednisone with bortezomib (VR-CAP) or vincristine (R-CHOP) for non-GCB DLBCL,” *Blood*, vol. 126, no. 16, pp. 1893–1901, Oct. 2015, doi: 10.1182/blood-2015-03-632430.
[36] W. H. Wilson *et al.*, “Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma,” *Nat. Med.*, vol. 21, no. 8, pp. 922–926, Aug. 2015, doi: 10.1038/nm.3884.
[37] Y. Yang *et al.*, “Exploiting synthetic lethality for the therapy of ABC diffuse large B cell

lymphoma," *Cancer Cell*, vol. 21, no. 6, pp. 723–737, Jun. 2012, doi: 10.1016/j.ccr.2012.05.024.
[38] C. G. Mullighan, "Genome sequencing of lymphoid malignancies," *Blood*, vol. 122, no. 24, pp. 3899–3907, Dec. 2013, doi: 10.1182/blood-2013-08-460311.
[39] S. P. Treon *et al.*, "MYD88 L265P somatic mutation in Waldenström's macroglobulinemia," *N. Engl. J. Med.*, vol. 367, no. 9, pp. 826–833, Aug. 2012, doi: 10.1056/NEJMoa1200710.
[40] R. D. Morin *et al.*, "Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin," *Nat. Genet.*, vol. 42, no. 2, pp. 181–185, Feb. 2010, doi: 10.1038/ng.518.
[41] L. Pasqualucci *et al.*, "Inactivating mutations of acetyltransferase genes in B-cell lymphoma," *Nature*, vol. 471, no. 7337, pp. 189–195, Mar. 2011, doi: 10.1038/nature09730.
[42] V. N. Ngo *et al.*, "Oncogenically active MYD88 mutations in human lymphoma," *Nature*, vol. 470, no. 7332, pp. 115–119, Feb. 2011, doi: 10.1038/nature09671.
[43] R. D. Morin *et al.*, "Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma," *Nature*, vol. 476, no. 7360, pp. 298–303, Jul. 2011, doi: 10.1038/nature10351.
[44] J. G. Lohr *et al.*, "Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 10, pp. 3879–3884, Mar. 2012, doi: 10.1073/pnas.1121343109.
[45] A. Gonzalez-Aguilar *et al.*, "Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas," *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 18, no. 19, pp. 5203–5211, Oct. 2012, doi: 10.1158/1078-0432.CCR-12-0845.
[46] J. Zhang *et al.*, "Genetic heterogeneity of diffuse large B-cell lymphoma," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 4, pp. 1398–1403, Jan. 2013, doi: 10.1073/pnas.1205299110.
[47] S. Mareschal *et al.*, "Whole exome sequencing of relapsed/refractory patients expands the repertoire of somatic mutations in diffuse large B-cell lymphoma," *Genes. Chromosomes Cancer*, vol. 55, no. 3, pp. 251–267, Mar. 2016, doi: 10.1002/gcc.22328.
[48] E. Cerami *et al.*, "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer Discov.*, vol. 2, no. 5, pp. 401–404, May 2012, doi: 10.1158/2159-8290.CD-12-0095.
[49] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signal.*, vol. 6, no. 269, p. pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.
[50] S. Dubois *et al.*, "Next-Generation Sequencing in Diffuse Large B-Cell Lymphoma Highlights Molecular Divergence and Therapeutic Opportunities: a LYSA Study," *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 22, no. 12, pp. 2919–2928, Jun. 2016, doi: 10.1158/1078-0432.CCR-15-2305.
[51] R. E. Davis *et al.*, "Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma," *Nature*, vol. 463, no. 7277, pp. 88–92, Jan. 2010, doi: 10.1038/nature08638.
[52] G. Lenz *et al.*, "Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 36, pp. 13520–13525, Sep. 2008, doi: 10.1073/pnas.0804295105.
[53] S. Monti *et al.*, "Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma," *Cancer Cell*, vol. 22, no. 3, pp. 359–372, Sep. 2012, doi: 10.1016/j.ccr.2012.07.014.
[54] L. Pasqualucci and R. Dalla-Favera, "SnapShot: diffuse large B cell lymphoma," *Cancer Cell*, vol. 25, no. 1, pp. 132-132.e1, Jan. 2014, doi: 10.1016/j.ccr.2013.12.012.
[55] L. Pasqualucci *et al.*, "Analysis of the coding genome of diffuse large B-cell lymphoma," *Nat. Genet.*, vol. 43, no. 9, pp. 830–837, Jul. 2011, doi: 10.1038/ng.892.
[56] M. Scandurra *et al.*, "Genomic lesions associated with a different clinical outcome in diffuse large B-Cell lymphoma treated with R-CHOP-21," *Br. J. Haematol.*, vol. 151, no. 3, pp. 221–231, Nov. 2010, doi: 10.1111/j.1365-2141.2010.08326.x.
[57] R. Scholtysik *et al.*, "Characterization of genomic imbalances in diffuse large B-cell lymphoma by detailed SNP-chip analysis," *Int. J. Cancer*, vol. 136, no. 5, pp. 1033–1042, Mar. 2015, doi: 10.1002/ijc.29072.
[58] E. Sebastián *et al.*, "High-resolution copy number analysis of paired normal-tumor samples from diffuse large B cell lymphoma," *Ann. Hematol.*, vol. 95, no. 2, pp. 253–262, Jan. 2016, doi: 10.1007/s00277-015-2552-3.
[59] S. Mareschal *et al.*, "Application of the cghRA framework to the genomic characterization of Diffuse Large B-Cell Lymphoma," *Bioinforma. Oxf. Engl.*, vol. 33, no. 19, pp. 2977–2985, Oct. 2017, doi: 10.1093/bioinformatics/btx309.

[60] T. Akasaka *et al.*, "Nonimmunoglobulin (non-Ig)/BCL6 gene fusion in diffuse large B-cell lymphoma results in worse prognosis than Ig/BCL6," *Blood*, vol. 96, no. 8, pp. 2907–2909, Oct. 2000.

[61] S. L. Barrans *et al.*, "Rearrangement of the BCL6 locus at 3q27 is an independent poor prognostic factor in nodal diffuse large B-cell lymphoma," *Br. J. Haematol.*, vol. 117, no. 2, pp. 322–332, May 2002, doi: 10.1046/j.1365-2141.2002.03435.x.

[62] J. Iqbal *et al.*, "Distinctive patterns of BCL6 molecular alterations and their functional consequences in different subgroups of diffuse large B-cell lymphoma," *Leukemia*, vol. 21, no. 11, pp. 2332–2343, Nov. 2007, doi: 10.1038/sj.leu.2404856.

[63] I. S. Lossos *et al.*, "Expression of a single gene, BCL-6, strongly predicts survival in patients with diffuse large B-cell lymphoma," *Blood*, vol. 98, no. 4, pp. 945–951, Aug. 2001, doi: 10.1182/blood.v98.4.945.

[64] K. Offit *et al.*, "Rearrangement of the bcl-6 gene as a prognostic marker in diffuse large-cell lymphoma," *N. Engl. J. Med.*, vol. 331, no. 2, pp. 74–80, Jul. 1994, doi: 10.1056/NEJM199407143310202.

[65] Q. Ye *et al.*, "Prognostic impact of concurrent MYC and BCL6 rearrangements and expression in de novo diffuse large B-cell lymphoma," *Oncotarget*, vol. 7, no. 3, pp. 2401–2416, Jan. 2016, doi: 10.18632/oncotarget.6262.

[66] S. M. Aukema *et al.*, "Double-hit B-cell lymphomas," *Blood*, vol. 117, no. 8, pp. 2319–2331, Feb. 2011, doi: 10.1182/blood-2010-09-297879.

[67] C. Copie-Bergman *et al.*, "MYC-IG rearrangements are negative predictors of survival in DLBCL patients treated with immunochemotherapy: a GELA/LYSA study," *Blood*, vol. 126, no. 22, pp. 2466–2474, Nov. 2015, doi: 10.1182/blood-2015-05-647602.

[68] A. M. Staiger *et al.*, "Clinical Impact of the Cell-of-Origin Classification and the MYC/ BCL2 Dual Expresser Status in Diffuse Large B-Cell Lymphoma Treated Within Prospective Clinical Trials of the German High-Grade Non-Hodgkin's Lymphoma Study Group," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 35, no. 22, pp. 2515–2526, Aug. 2017, doi: 10.1200/JCO.2016.70.3660.

[69] A. Younes *et al.*, "Randomized Phase III Trial of Ibrutinib and Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone in Non-Germinal Center B-Cell Diffuse Large B-Cell Lymphoma," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 37, no. 15, pp. 1285–1295, May 2019, doi: 10.1200/JCO.18.02403.

[70] A. Davies *et al.*, "Gene-expression profiling of bortezomib added to standard chemoimmunotherapy for diffuse large B-cell lymphoma (REMoDL-B): an open-label, randomised, phase 3 trial," *Lancet Oncol.*, vol. 20, no. 5, pp. 649–662, May 2019, doi: 10.1016/S1470-2045(18)30935-5.

[71] R. D. Morin *et al.*, "Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing," *Blood*, vol. 122, no. 7, pp. 1256–1265, Aug. 2013, doi: 10.1182/blood-2013-02-483727.

[72] C. Steidl and R. D. Gascoyne, "The molecular pathogenesis of primary mediastinal large B-cell lymphoma," *Blood*, vol. 118, no. 10, pp. 2659–2669, Sep. 2011, doi: 10.1182/blood-2011-05-326538.

[73] B. Chapuy *et al.*, "Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes," *Nat. Med.*, vol. 24, no. 5, pp. 679–690, May 2018, doi: 10.1038/s41591-018-0016-8.

[74] R. Schmitz *et al.*, "Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma," *N. Engl. J. Med.*, vol. 378, no. 15, pp. 1396–1407, Apr. 2018, doi: 10.1056/NEJMoa1801445.

[75] A. Reddy *et al.*, "Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma," *Cell*, vol. 171, no. 2, pp. 481-494.e15, Oct. 2017, doi: 10.1016/j.cell.2017.09.027.

[76] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007, doi: 10.1126/science.1136800.

[77] M. S. Lawrence *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, Jul. 2013, doi: 10.1038/nature12213.

[78] S. Dubois *et al.*, "Refining diffuse large B-cell lymphoma subgroups using integrated analysis of molecular profiles," *EBioMedicine*, vol. 48, pp. 58–69, Oct. 2019, doi: 10.1016/j.ebiom.2019.09.034.

[79] "A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. The Non-Hodgkin's Lymphoma Classification Project," *Blood*, vol. 89, no. 11, pp. 3909–3918, Jun. 1997.

[80] D. Cazals-Hatem *et al.*, "Primary mediastinal large B-cell lymphoma. A clinicopathologic study of 141 cases compared with 916 nonmediastinal large B-cell lymphomas, a GELA ('Groupe d'Etude des

Lymphomes de l'Adulte') study," *Am. J. Surg. Pathol.*, vol. 20, no. 7, pp. 877–888, Jul. 1996, doi: 10.1097/00000478-199607000-00012.

[81] P. L. Zinzani *et al.*, "Practice guidelines for the management of extranodal non-Hodgkin's lymphomas of adult non-immunodeficient patients. Part I: primary lung and mediastinal lymphomas. A project of the Italian Society of Hematology, the Italian Society of Experimental Hematology and the Italian Group for Bone Marrow Transplantation," *Haematologica*, vol. 93, no. 9, pp. 1364–1371, Sep. 2008, doi: 10.3324/haematol.12742.

[82] G. Chen, A. P. C. Yim, L. Ma, P. Gaulard, and J. K. C. Chan, "Primary pulmonary large B-cell lymphoma--mediastinal type?," *Histopathology*, vol. 58, no. 2, pp. 324–326, Jan. 2011, doi: 10.1111/j.1365-2559.2011.03745.x.

[83] A. Rosenwald *et al.*, "Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma," *J. Exp. Med.*, vol. 198, no. 6, pp. 851–862, Sep. 2003, doi: 10.1084/jem.20031074.

[84] K. J. Savage *et al.*, "The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma," *Blood*, vol. 102, no. 12, pp. 3871–3879, Dec. 2003, doi: 10.1182/blood-2003-06-1841.

[85] J. Yuan *et al.*, "Identification of Primary Mediastinal Large B-cell Lymphoma at Nonmediastinal Sites by Gene Expression Profiling," *Am. J. Surg. Pathol.*, vol. 39, no. 10, pp. 1322–1330, Oct. 2015, doi: 10.1097/PAS.0000000000000473.

[86] F. Menestrina *et al.*, "Mediastinal large-cell lymphoma of B-type, with sclerosis: histopathological and immunohistochemical study of eight cases," *Histopathology*, vol. 10, no. 6, pp. 589–600, Jun. 1986, doi: 10.1111/j.1365-2559.1986.tb02512.x.

[87] M. Paulli *et al.*, "Mediastinal B-cell lymphoma: a study of its histomorphologic spectrum based on 109 cases," *Hum. Pathol.*, vol. 30, no. 2, pp. 178–187, Feb. 1999, doi: 10.1016/s0046-8177(99)90273-3.

[88] A. Traverse-Glehen *et al.*, "Mediastinal gray zone lymphoma: the missing link between classic Hodgkin's lymphoma and mediastinal large B-cell lymphoma," *Am. J. Surg. Pathol.*, vol. 29, no. 11, pp. 1411–1421, Nov. 2005, doi: 10.1097/01.pas.0000180856.74572.73.

[89] W. H. Wilson *et al.*, "A prospective study of mediastinal gray-zone lymphoma," *Blood*, vol. 124, no. 10, pp. 1563–1569, Sep. 2014, doi: 10.1182/blood-2014-03-564906.

[90] J. P. Higgins and R. A. Warnke, "CD30 expression is common in mediastinal large B-cell lymphoma," *Am. J. Clin. Pathol.*, vol. 112, no. 2, pp. 241–247, Aug. 1999, doi: 10.1093/ajcp/112.2.241.

[91] M. Calaminici, K. Piper, A. M. Lee, and A. J. Norton, "CD23 expression in mediastinal large B-cell lymphomas," *Histopathology*, vol. 45, no. 6, pp. 619–624, Dec. 2004, doi: 10.1111/j.1365-2559.2004.01969.x.

[92] C. Copie-Bergman *et al.*, "The MAL gene is expressed in primary mediastinal large B-cell lymphoma," *Blood*, vol. 94, no. 10, pp. 3567–3575, Nov. 1999.

[93] C. Copie-Bergman *et al.*, "MAL expression in lymphoid cells: further evidence for MAL as a distinct molecular marker of primary mediastinal large B-cell lymphomas," *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc*, vol. 15, no. 11, pp. 1172–1180, Nov. 2002, doi: 10.1097/01.MP.0000032534.81894.B3.

[94] F. Leithäuser, M. Bäuerle, M. Q. Huynh, and P. Möller, "Isotype-switched immunoglobulin genes with a high load of somatic hypermutation and lack of ongoing mutational activity are prevalent in mediastinal B-cell lymphoma," *Blood*, vol. 98, no. 9, pp. 2762–2770, Nov. 2001, doi: 10.1182/blood.v98.9.2762.

[95] P. Tsang, E. Cesarman, A. Chadburn, Y. F. Liu, and D. M. Knowles, "Molecular characterization of primary mediastinal B cell lymphoma," *Am. J. Pathol.*, vol. 148, no. 6, pp. 2017–2025, Jun. 1996.

[96] A. Mottok *et al.*, "Genomic Alterations in CIITA Are Frequent in Primary Mediastinal Large B Cell Lymphoma and Are Associated with Diminished MHC Class II Expression," *Cell Rep.*, vol. 13, no. 7, pp. 1418–1431, Nov. 2015, doi: 10.1016/j.celrep.2015.10.008.

[97] C. Steidl *et al.*, "MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers," *Nature*, vol. 471, no. 7338, pp. 377–381, Mar. 2011, doi: 10.1038/nature09754.

[98] M. Bentz *et al.*, "Gain of chromosome arm 9p is characteristic of primary mediastinal B-cell lymphoma (MBL): comprehensive molecular cytogenetic analysis and presentation of a novel MBL cell line," *Genes. Chromosomes Cancer*, vol. 30, no. 4, pp. 393–401, Apr. 2001, doi: 10.1002/1098-2264(2001)9999:9999<::aid-gcc1105>3.0.co;2-i.

[99] M. Shi *et al.*, "Expression of programmed cell death 1 ligand 2 (PD-L2) is a distinguishing feature of primary mediastinal (thymic) large B-cell lymphoma and associated with PDCD1LG2 copy gain," *Am. J. Surg. Pathol.*, vol. 38, no. 12, pp. 1715–1723, Dec. 2014, doi: 10.1097/PAS.0000000000000297.

[100] F. Feuerhake *et al.*, "NFkappaB activity, function, and target-gene signatures in primary mediastinal large B-cell lymphoma and diffuse large B-cell lymphoma subtypes," *Blood*, vol. 106, no. 4, pp. 1392–1399, Aug. 2005, doi: 10.1182/blood-2004-12-4901.

[101] R. Schmitz *et al.*, "TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma," *J. Exp. Med.*, vol. 206, no. 5, pp. 981–989, May 2009, doi: 10.1084/jem.20090528.

[102] F. Jardin *et al.*, "Recurrent mutations of the exportin 1 gene (XPO1) and their impact on selective inhibitor of nuclear export compounds sensitivity in primary mediastinal B-cell lymphoma," *Am. J. Hematol.*, vol. 91, no. 9, pp. 923–930, Sep. 2016, doi: 10.1002/ajh.24451.

[103] M. Janz *et al.*, "Classical Hodgkin lymphoma is characterized by high constitutive expression of activating transcription factor 3 (ATF3), which promotes viability of Hodgkin/Reed-Sternberg cells," *Blood*, vol. 107, no. 6, pp. 2536–2539, Mar. 2006, doi: 10.1182/blood-2005-07-2694.

[104] H. Stein *et al.*, "Down-regulation of BOB.1/OBF.1 and Oct2 in classical Hodgkin disease but not in lymphocyte predominant Hodgkin disease correlates with immunoglobulin transcription," *Blood*, vol. 97, no. 2, pp. 496–501, Jan. 2001, doi: 10.1182/blood.v97.2.496.

[105] T. Marafioti, M. Pozzobon, M.-L. Hansmann, G. Delsol, S. A. Pileri, and D. Y. Mason, "Expression of intracellular signaling molecules in classical and lymphocyte predominance Hodgkin disease," *Blood*, vol. 103, no. 1, pp. 188–193, Jan. 2004, doi: 10.1182/blood-2003-05-1487.

[106] A. Ehlers, E. Oker, S. Bentink, D. Lenze, H. Stein, and M. Hummel, "Histone acetylation and DNA demethylation of B cells result in a Hodgkin-like phenotype," *Leukemia*, vol. 22, no. 4, pp. 835–841, Apr. 2008, doi: 10.1038/leu.2008.12.

[107] R. Küppers *et al.*, "Identification of Hodgkin and Reed-Sternberg cell-specific genes by gene expression profiling," *J. Clin. Invest.*, vol. 111, no. 4, pp. 529–537, Feb. 2003, doi: 10.1172/JCI16624.

[108] S. Mathas *et al.*, "Intrinsic inhibition of transcription factor E2A by HLH proteins ABF-1 and Id2 mediates reprogramming of neoplastic B cells in Hodgkin lymphoma," *Nat. Immunol.*, vol. 7, no. 2, pp. 207–215, Feb. 2006, doi: 10.1038/ni1285.

[109] I. Schwering *et al.*, "Loss of the B-lineage-specific gene expression program in Hodgkin and Reed-Sternberg cells of Hodgkin lymphoma," *Blood*, vol. 101, no. 4, pp. 1505–1512, Feb. 2003, doi: 10.1182/blood-2002-03-0839.

[110] M. Hinz *et al.*, "Nuclear factor kappaB-dependent gene expression profiling of Hodgkin's disease tumor cells, pathogenetic significance, and link to constitutive signal transducer and activator of transcription 5a activity," *J. Exp. Med.*, vol. 196, no. 5, pp. 605–617, Sep. 2002, doi: 10.1084/jem.20020062.

[111] M. Hinz, P. Löser, S. Mathas, D. Krappmann, B. Dörken, and C. Scheidereit, "Constitutive NF-kappaB maintains high expression of a characteristic gene network, including CD40, CD86, and a set of antiapoptotic genes in Hodgkin/Reed-Sternberg cells," *Blood*, vol. 97, no. 9, pp. 2798–2807, May 2001, doi: 10.1182/blood.v97.9.2798.

[112] U. E. Höpken *et al.*, "Up-regulation of the chemokine receptor CCR7 in classical but not in lymphocyte-predominant Hodgkin disease correlates with distinct dissemination of neoplastic cells in lymphoid organs," *Blood*, vol. 99, no. 4, pp. 1109–1116, Feb. 2002, doi: 10.1182/blood.v99.4.1109.

[113] B. F. Skinnider *et al.*, "Signal transducer and activator of transcription 6 is frequently activated in Hodgkin and Reed-Sternberg cells of Hodgkin lymphoma," *Blood*, vol. 99, no. 2, pp. 618–626, Jan. 2002, doi: 10.1182/blood.v99.2.618.

[114] N. Siddiqui, B. Ayub, F. Badar, and A. Zaidi, "Hodgkin's lymphoma in Pakistan: a clinico-epidemiological study of 658 cases at a cancer center in Lahore," *Asian Pac. J. Cancer Prev. APJCP*, vol. 7, no. 4, pp. 651–655, Dec. 2006.

[115] X. G. Zhou *et al.*, "Epstein-Barr virus (EBV) in Chinese pediatric Hodgkin disease: Hodgkin disease in young children is an EBV-related lymphoma," *Cancer*, vol. 92, no. 6, pp. 1621–1631, Sep. 2001, doi: 10.1002/1097-0142(20010915)92:6<1621::aid-cncr1488>3.0.co;2-p.

[116] E. Tiacci *et al.*, "Pervasive mutations of JAK-STAT pathway genes in classical Hodgkin lymphoma," *Blood*, vol. 131, no. 22, pp. 2454–2465, May 2018, doi: 10.1182/blood-2017-11-814913.

[117] J. Reichel *et al.*, "Flow sorting and exome sequencing reveal the oncogenome of primary

Hodgkin and Reed-Sternberg cells," *Blood*, vol. 125, no. 7, pp. 1061–1072, Feb. 2015, doi: 10.1182/blood-2014-11-610436.
[118] V. Spina *et al.*, "Circulating tumor DNA reveals genetics, clonal evolution, and residual disease in classical Hodgkin lymphoma," *Blood*, vol. 131, no. 22, pp. 2413–2425, May 2018, doi: 10.1182/blood-2017-11-812073.
[119] L. Mansouri *et al.*, "Frequent NFKBIE deletions are associated with poor outcome in primary mediastinal B-cell lymphoma," *Blood*, vol. 128, no. 23, pp. 2666–2670, Dec. 2016, doi: 10.1182/blood-2016-03-704528.
[120] K. Van Roosbroeck *et al.*, "JAK2 rearrangements, including the novel SEC31A-JAK2 fusion, are recurrent in classical Hodgkin lymphoma," *Blood*, vol. 117, no. 15, pp. 4056–4064, Apr. 2011, doi: 10.1182/blood-2010-06-291310.
[121] A. Mottok *et al.*, "Inactivating SOCS1 mutations are caused by aberrant somatic hypermutation and restricted to a subset of B-cell lymphoma entities," *Blood*, vol. 114, no. 20, pp. 4503–4506, Nov. 2009, doi: 10.1182/blood-2009-06-225839.
[122] M. A. Weniger *et al.*, "Mutations of the tumor suppressor gene SOCS-1 in classical Hodgkin lymphoma are frequent and associated with nuclear phospho-STAT5 accumulation," *Oncogene*, vol. 25, no. 18, pp. 2679–2684, Apr. 2006, doi: 10.1038/sj.onc.1209151.
[123] J. Gunawardana *et al.*, "Recurrent somatic mutations of PTPN1 in primary mediastinal B cell lymphoma and Hodgkin lymphoma," *Nat. Genet.*, vol. 46, no. 4, pp. 329–335, Apr. 2014, doi: 10.1038/ng.2900.
[124] M. Kleppe *et al.*, "Mutation analysis of the tyrosine phosphatase PTPN2 in Hodgkin's lymphoma and T-cell non-Hodgkin's lymphoma," *Haematologica*, vol. 96, no. 11, pp. 1723–1727, Nov. 2011, doi: 10.3324/haematol.2011.041921.
[125] V. Camus *et al.*, "Detection and prognostic value of recurrent exportin 1 mutations in tumor and cell-free circulating DNA of patients with classical Hodgkin lymphoma," *Haematologica*, vol. 101, no. 9, pp. 1094–1101, Sep. 2016, doi: 10.3324/haematol.2016.145102.
[126] M. S. Lim *et al.*, "T-cell/histiocyte-rich large B-cell lymphoma: a heterogeneous entity with derivation from germinal center B cells," *Am. J. Surg. Pathol.*, vol. 26, no. 11, pp. 1458–1466, Nov. 2002, doi: 10.1097/00000478-200211000-00008.
[127] V. Diehl *et al.*, "Clinical presentation, course, and prognostic factors in lymphocyte-predominant Hodgkin's disease and lymphocyte-rich classical Hodgkin's disease: report from the European Task Force on Lymphoma Project on Lymphocyte-Predominant Hodgkin's Disease," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 17, no. 3, pp. 776–783, Mar. 1999, doi: 10.1200/JCO.1999.17.3.776.
[128] B. Pellegrino *et al.*, "Lymphocyte-predominant Hodgkin's lymphoma in children: therapeutic abstention after initial lymph node resection--a Study of the French Society of Pediatric Oncology," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 21, no. 15, pp. 2948–2952, Aug. 2003, doi: 10.1200/JCO.2003.01.079.
[129] F. B. Coles, R. W. Cartun, and W. T. Pastuszak, "Hodgkin's disease, lymphocyte-predominant type: immunoreactivity with B-cell antibodies," *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc*, vol. 1, no. 4, pp. 274–278, Jul. 1988.
[130] G. S. Pinkus and J. W. Said, "Hodgkin's disease, lymphocyte predominance type, nodular--a distinct entity? Unique staining profile for L&H variants of Reed-Sternberg cells defined by monoclonal antibodies to leukocyte common antigen, granulocyte-specific antigen, and B-cell-specific antigen," *Am. J. Pathol.*, vol. 118, no. 1, pp. 1–6, Jan. 1985.
[131] G. S. Pinkus and J. W. Said, "Hodgkin's disease, lymphocyte predominance type, nodular--further evidence for a B cell derivation. L & H variants of Reed-Sternberg cells express L26, a pan B cell marker," *Am. J. Pathol.*, vol. 133, no. 2, pp. 211–217, Nov. 1988.
[132] V. Brune *et al.*, "Origin and pathogenesis of nodular lymphocyte-predominant Hodgkin lymphoma as revealed by global gene expression analysis," *J. Exp. Med.*, vol. 205, no. 10, pp. 2251–2268, Sep. 2008, doi: 10.1084/jem.20080809.
[133] A. R. Huppmann *et al.*, "EBV may be expressed in the LP cells of nodular lymphocyte-predominant Hodgkin lymphoma (NLPHL) in both children and adults," *Am. J. Surg. Pathol.*, vol. 38, no. 3, pp. 316–324, Mar. 2014, doi: 10.1097/PAS.0000000000000107.
[134] T. Ohno, J. A. Stribley, G. Wu, S. H. Hinrichs, D. D. Weisenburger, and W. C. Chan, "Clonality in nodular lymphocyte-predominant Hodgkin's disease," *N. Engl. J. Med.*, vol. 337, no. 7, pp. 459–465, Aug. 1997, doi: 10.1056/NEJM199708143370704.
[135] T. Marafioti *et al.*, "Origin of nodular lymphocyte-predominant Hodgkin's disease from a clonal expansion of highly mutated germinal-center B cells," *N. Engl. J. Med.*, vol. 337, no. 7, pp. 453–458, Aug. 1997, doi: 10.1056/NEJM199708143370703.

[136] C. Atayar *et al.*, “BCL6 alternative breakpoint region break and homozygous deletion of 17q24 in the nodular lymphocyte predominance type of Hodgkin’s lymphoma-derived cell line DEV,” *Hum. Pathol.*, vol. 37, no. 6, pp. 675–683, Jun. 2006, doi: 10.1016/j.humpath.2006.01.018.
[137] C. Renné, J. I. Martín-Subero, M.-L. Hansmann, and R. Siebert, “Molecular cytogenetic analyses of immunoglobulin loci in nodular lymphocyte predominant Hodgkin’s lymphoma reveal a recurrent IGH-BCL6 juxtaposition,” *J. Mol. Diagn. JMD*, vol. 7, no. 3, pp. 352–356, Aug. 2005, doi: 10.1016/S1525-1578(10)60564-8.
[138] I. Wlodarska, M. Stul, C. De Wolf-Peeters, and A. Hagemeijer, “Heterogeneity of BCL6 rearrangements in nodular lymphocyte predominant Hodgkin’s lymphoma,” *Haematologica*, vol. 89, no. 8, pp. 965–972, Aug. 2004.
[139] A. Liso *et al.*, “Aberrant somatic hypermutation in tumor cells of nodular-lymphocyte-predominant and classic Hodgkin lymphoma,” *Blood*, vol. 108, no. 3, pp. 1013–1020, Aug. 2006, doi: 10.1182/blood-2005-10-3949.
[140] S. Hartmann *et al.*, “Highly recurrent mutations of SGK1, DUSP2 and JUNB in nodular lymphocyte predominant Hodgkin lymphoma,” *Leukemia*, vol. 30, no. 4, pp. 844–853, Apr. 2016, doi: 10.1038/leu.2015.328.
[141] O. T. Avery, C. M. Macleod, and M. McCarty, “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III,” *J. Exp. Med.*, vol. 79, no. 2, pp. 137–158, Feb. 1944, doi: 10.1084/jem.79.2.137.
[142] J. D. Watson and F. H. Crick, “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid,” *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 1953, doi: 10.1038/171737a0.
[143] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.
[144] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E. Hood, “The synthesis of oligonucleotides containing an aliphatic amino group at the 5’ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis,” *Nucleic Acids Res.*, vol. 13, no. 7, pp. 2399–2412, Apr. 1985, doi: 10.1093/nar/13.7.2399.
[145] W. Ansorge, B. S. Sproat, J. Stegemann, and C. Schwager, “A non-radioactive automated method for DNA sequence determination,” *J. Biochem. Biophys. Methods*, vol. 13, no. 6, pp. 315–323, Dec. 1986, doi: 10.1016/0165-022x(86)90038-2.
[146] W. Ansorge, B. Sproat, J. Stegemann, C. Schwager, and M. Zenke, “Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis,” *Nucleic Acids Res.*, vol. 15, no. 11, pp. 4593–4602, Jun. 1987, doi: 10.1093/nar/15.11.4593.
[147] J. M. Prober *et al.*, “A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides,” *Science*, vol. 238, no. 4825, pp. 336–341, Oct. 1987, doi: 10.1126/science.2443975.
[148] H. Swerdlow and R. Gesteland, “Capillary gel electrophoresis for rapid, high resolution DNA sequencing,” *Nucleic Acids Res.*, vol. 18, no. 6, pp. 1415–1419, Mar. 1990, doi: 10.1093/nar/18.6.1415.
[149] J. A. Luckey *et al.*, “High speed DNA sequencing by capillary electrophoresis,” *Nucleic Acids Res.*, vol. 18, no. 15, pp. 4417–4421, Aug. 1990, doi: 10.1093/nar/18.15.4417.
[150] S. Bobillo *et al.*, “Cell free circulating tumor DNA in cerebrospinal fluid detects and monitors central nervous system involvement of B-cell lymphomas,” *Haematologica*, vol. 106, no. 2, pp. 513–521, Feb. 2021, doi: 10.3324/haematol.2019.241208.
[151] V. Camus *et al.*, “Targeted genotyping of circulating tumor DNA for classical Hodgkin lymphoma monitoring: a prospective study,” *Haematologica*, vol. 106, no. 1, pp. 154–162, Jan. 2021, doi: 10.3324/haematol.2019.237719.
[152] U. Satyal, A. Srivastava, and P. H. Abbosh, “Urine Biopsy-Liquid Gold for Molecular Detection and Surveillance of Bladder Cancer,” *Front. Oncol.*, vol. 9, p. 1266, 2019, doi: 10.3389/fonc.2019.01266.
[153] E. Bohers, P.-J. Viailly, and F. Jardin, “cfDNA Sequencing: Technological Approaches and Bioinformatic Issues,” *Pharm. Basel Switz.*, vol. 14, no. 6, p. 596, Jun. 2021, doi: 10.3390/ph14060596.
[154] P. Mandel and P. Metais, “Nuclear Acids In Human Blood Plasma,” *C. R. Seances Soc. Biol. Fil.*, vol. 142, no. 3–4, pp. 241–243, Feb. 1948.
[155] D. Koffler, V. Agnello, R. Winchester, and H. G. Kunkel, “The occurrence of single-stranded DNA in the serum of patients with systemic lupus

erythematosus and other diseases," *J. Clin. Invest.*, vol. 52, no. 1, pp. 198–204, Jan. 1973, doi: 10.1172/JCI107165.

[156] S. A. Leon, B. Shapiro, D. M. Sklaroff, and M. J. Yaros, "Free DNA in the serum of cancer patients and the effect of therapy," *Cancer Res.*, vol. 37, no. 3, pp. 646–650, Mar. 1977.

[157] M. Stroun, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Beljanski, "Neoplastic characteristics of the DNA found in the plasma of cancer patients," *Oncology*, vol. 46, no. 5, pp. 318–322, 1989, doi: 10.1159/000226740.

[158] D. Sidransky *et al.*, "Identification of p53 gene mutations in bladder cancers and urine samples," *Science*, vol. 252, no. 5006, pp. 706–709, May 1991, doi: 10.1126/science.2024123.

[159] V. Vasioukhin, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Stroun, "Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia," *Br. J. Haematol.*, vol. 86, no. 4, pp. 774–779, Apr. 1994, doi: 10.1111/j.1365-2141.1994.tb04828.x.

[160] P. Anker *et al.*, "K-ras mutations are found in DNA extracted from the plasma of patients with colorectal cancer," *Gastroenterology*, vol. 112, no. 4, pp. 1114–1120, Apr. 1997, doi: 10.1016/s0016-5085(97)70121-5.

[161] F. Austrup *et al.*, "Prognostic value of genomic alterations in minimal residual cancer cells purified from the blood of breast cancer patients," *Br. J. Cancer*, vol. 83, no. 12, pp. 1664–1673, Dec. 2000, doi: 10.1054/bjoc.2000.1501.

[162] R. H. Dennin, "DNA of free and complexed origin in human plasma: concentration and length distribution," *Klin. Wochenschr.*, vol. 57, no. 9, pp. 451–456, May 1979, doi: 10.1007/BF01477498.

[163] S. Jahr *et al.*, "DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells," *Cancer Res.*, vol. 61, no. 4, pp. 1659–1665, Feb. 2001.

[164] M. Ivanov, A. Baranova, T. Butler, P. Spellman, and V. Mileyko, "Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation," *BMC Genomics*, vol. 16 Suppl 13, p. S1, 2015, doi: 10.1186/1471-2164-16-S13-S1.

[165] F. Diehl *et al.*, "Detection and quantification of mutations in the plasma of patients with colorectal tumors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 45, pp. 16368–16373, Nov. 2005, doi: 10.1073/pnas.0507904102.

[166] Y. M. D. Lo *et al.*, "Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus," *Sci. Transl. Med.*, vol. 2, no. 61, p. 61ra91, Dec. 2010, doi: 10.1126/scitranslmed.3001720.

[167] A. H. Wyllie, R. G. Morris, A. L. Smith, and D. Dunlop, "Chromatin cleavage in apoptosis: association with condensed chromatin morphology and dependence on macromolecular synthesis," *J. Pathol.*, vol. 142, no. 1, pp. 67–77, Jan. 1984, doi: 10.1002/path.1711420112.

[168] T. Beiter, A. Fragasso, J. Hudemann, A. M. Niess, and P. Simon, "Short-term treadmill running as a model for studying cell-free DNA kinetics in vivo," *Clin. Chem.*, vol. 57, no. 4, pp. 633–636, Apr. 2011, doi: 10.1373/clinchem.2010.158030.

[169] Y. M. Lo *et al.*, "Quantitative abnormalities of fetal DNA in maternal serum in preeclampsia," *Clin. Chem.*, vol. 45, no. 2, pp. 184–188, Feb. 1999.

[170] Y. M. Lo, J. Zhang, T. N. Leung, T. K. Lau, A. M. Chang, and N. M. Hjelm, "Rapid clearance of fetal DNA from maternal plasma," *Am. J. Hum. Genet.*, vol. 64, no. 1, pp. 218–224, Jan. 1999, doi: 10.1086/302205.

[171] S. C. Y. Yu *et al.*, "High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing," *Clin. Chem.*, vol. 59, no. 8, pp. 1228–1237, Aug. 2013, doi: 10.1373/clinchem.2013.203679.

[172] P. Burnham *et al.*, "Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma," *Sci. Rep.*, vol. 6, p. 27859, Jun. 2016, doi: 10.1038/srep27859.

[173] M. B. Giacona, G. C. Ruben, K. A. Iczkowski, T. B. Roos, D. M. Porter, and G. D. Sorenson, "Cell-free DNA in human blood plasma: length measurements in patients with pancreatic cancer and healthy controls," *Pancreas*, vol. 17, no. 1, pp. 89–97, Jul. 1998, doi: 10.1097/00006676-199807000-00012.

[174] M. Fleischhacker and B. Schmidt, "Circulating nucleic acids (CNAs) and cancer--a survey," *Biochim. Biophys. Acta*, vol. 1775, no. 1, pp. 181–232, Jan. 2007, doi: 10.1016/j.bbcan.2006.10.001.

[175] M. Elazezy and S. A. Joosse, "Techniques of using circulating tumor DNA as a liquid biopsy component in cancer management," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 370–378, 2018, doi: 10.1016/j.csbj.2018.10.002.

[176] F. Scherer *et al.*, "Noninvasive Detection of Ibrutinib Resistance in Non-Hodgkin Lymphoma Using Cell-Free DNA," *Blood*, vol. 128, no. 22, pp. 1752–1752, Dec. 2016, doi: 10.1182/blood.V128.22.1752.1752.

[177] E. Bohers *et al.*, “Non-invasive monitoring of diffuse large B-cell lymphoma by cell-free DNA high-throughput targeted sequencing: analysis of a prospective cohort,” *Blood Cancer J.*, vol. 8, no. 8, p. 74, Aug. 2018, doi: 10.1038/s41408-018-0111-6.

[178] J. R. Thompson and S. P. Menon, “Liquid Biopsies and Cancer Immunotherapy,” *Cancer J. Sudbury Mass*, vol. 24, no. 2, pp. 78–83, Apr. 2018, doi: 10.1097/PPO.0000000000000307.

[179] D. M. Kurtz *et al.*, “Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 36, no. 28, pp. 2845–2853, Oct. 2018, doi: 10.1200/JCO.2018.78.5246.

[180] Y. van der Pol and F. Mouliere, “Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA,” *Cancer Cell*, vol. 36, no. 4, pp. 350–368, Oct. 2019, doi: 10.1016/j.ccell.2019.09.003.

[181] F. Cheng, L. Su, and C. Qian, “Circulating tumor DNA: a promising biomarker in the liquid biopsy of cancer,” *Oncotarget*, vol. 7, no. 30, pp. 48832–48841, Jul. 2016, doi: 10.18632/oncotarget.9453.

[182] I. Medina Diaz, A. Nocon, D. H. Mehnert, J. Fredebohm, F. Diehl, and F. Holtrup, “Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing,” *PloS One*, vol. 11, no. 11, p. e0166354, 2016, doi: 10.1371/journal.pone.0166354.

[183] C. Alidousty *et al.*, “Comparison of Blood Collection Tubes from Three Different Manufacturers for the Collection of Cell-Free DNA for Liquid Biopsy Mutation Testing,” *J. Mol. Diagn. JMD*, vol. 19, no. 5, pp. 801–804, Sep. 2017, doi: 10.1016/j.jmoldx.2017.06.004.

[184] A. Ward Gahlawat *et al.*, “Evaluation of Storage Tubes for Combined Analysis of Circulating Nucleic Acids in Liquid Biopsies,” *Int. J. Mol. Sci.*, vol. 20, no. 3, p. E704, Feb. 2019, doi: 10.3390/ijms20030704.

[185] Y. Zhao, Y. Li, P. Chen, S. Li, J. Luo, and H. Xia, “Performance comparison of blood collection tubes as liquid biopsy storage system for minimizing cfDNA contamination from genomic DNA,” *J. Clin. Lab. Anal.*, vol. 33, no. 2, p. e22670, Feb. 2019, doi: 10.1002/jcla.22670.

[186] S. El Messaoudi, F. Rolet, F. Mouliere, and A. R. Thierry, “Circulating cell free DNA: Preanalytical considerations,” *Clin. Chim. Acta Int. J. Clin. Chem.*, vol. 424, pp. 222–230, Sep. 2013, doi: 10.1016/j.cca.2013.05.022.

[187] L. Sorber *et al.*, “Circulating Cell-Free DNA and RNA Analysis as Liquid Biopsy: Optimal Centrifugation Protocol,” *Cancers*, vol. 11, no. 4, p. E458, Mar. 2019, doi: 10.3390/cancers11040458.

[188] R. J. Diefenbach, J. H. Lee, R. F. Kefford, and H. Rizos, “Evaluation of commercial kits for purification of circulating free DNA,” *Cancer Genet.*, vol. 228–229, pp. 21–27, Dec. 2018, doi: 10.1016/j.cancergen.2018.08.005.

[189] L. F. van Dessel *et al.*, “High-throughput isolation of circulating tumor DNA: a comparison of automated platforms,” *Mol. Oncol.*, vol. 13, no. 2, pp. 392–402, Feb. 2019, doi: 10.1002/1878-0261.12415.

[190] S. Henikoff and G. M. Church, “Simultaneous Discovery of Cell-Free DNA and the Nucleosome Ladder,” *Genetics*, vol. 209, no. 1, pp. 27–29, May 2018, doi: 10.1534/genetics.118.300775.

[191] M. Lapin *et al.*, “Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer,” *J. Transl. Med.*, vol. 16, no. 1, p. 300, Nov. 2018, doi: 10.1186/s12967-018-1677-2.

[192] S. Nikolaev, L. Lemmens, T. Koessler, J.-L. Blouin, and T. Nouspikel, “Circulating tumoral DNA: Preanalytical validation and quality control in a diagnostic laboratory,” *Anal. Biochem.*, vol. 542, pp. 34–39, Feb. 2018, doi: 10.1016/j.ab.2017.11.004.

[193] A. S. Devonshire *et al.*, “Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification,” *Anal. Bioanal. Chem.*, vol. 406, no. 26, pp. 6499–6512, Oct. 2014, doi: 10.1007/s00216-014-7835-3.

[194] B. Vogelstein and K. W. Kinzler, “Digital PCR,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 16, pp. 9236–9241, Aug. 1999, doi: 10.1073/pnas.96.16.9236.

[195] K. Chen, Y. Ma, T. Ding, X. Zhang, B. Chen, and M. Guan, “Effectiveness of digital PCR for MYD88L265P detection in vitreous fluid for primary central nervous system lymphoma diagnosis,” *Exp. Ther. Med.*, vol. 20, no. 1, pp. 301–308, Jul. 2020, doi: 10.3892/etm.2020.8695.

[196] L. S. Hiemcke-Jiwa *et al.*, “The use of droplet digital PCR in liquid biopsies: A highly sensitive technique for MYD88 p.(L265P) detection in cerebrospinal fluid,” *Hematol. Oncol.*, vol. 36, no. 2, pp. 429–435, Apr. 2018, doi: 10.1002/hon.2489.

[197] T. M. Butler, P. T. Spellman, and J. Gray, “Circulating-tumor DNA as an early detection and diagnostic tool,” *Curr. Opin. Genet. Dev.*, vol. 42,

pp. 14–21, Feb. 2017, doi: 10.1016/j.gde.2016.12.003.
[198] C. A. Milbury *et al.*, “Determining lower limits of detection of digital PCR assays for cancer-related gene mutations,” *Biomol. Detect. Quantif.*, vol. 1, no. 1, pp. 8–22, Sep. 2014, doi: 10.1016/j.bdq.2014.08.001.
[199] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein, “Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 15, pp. 8817–8822, Jul. 2003, doi: 10.1073/pnas.1133470100.
[200] F. Diehl *et al.*, “Circulating mutant DNA to assess tumor dynamics,” *Nat. Med.*, vol. 14, no. 9, pp. 985–990, Sep. 2008, doi: 10.1038/nm.1789.
[201] M. Holdhoff, K. Schmidt, R. Donehower, and L. A. Diaz, “Analysis of circulating tumor DNA to confirm somatic KRAS mutations,” *J. Natl. Cancer Inst.*, vol. 101, no. 18, pp. 1284–1285, Sep. 2009, doi: 10.1093/jnci/djp240.
[202] B. O’Leary *et al.*, “Comparison of BEAMing and Droplet Digital PCR for Circulating Tumor DNA Analysis,” *Clin. Chem.*, vol. 65, no. 11, pp. 1405–1413, Nov. 2019, doi: 10.1373/clinchem.2019.305805.
[203] J. Garcia *et al.*, “Routine Molecular Screening of Patients with Advanced Non-SmallCell Lung Cancer in Circulating Cell-Free DNA at Diagnosis and During Progression Using OncoBEAMTM EGFR V2 and NGS Technologies,” *Mol. Diagn. Ther.*, vol. 25, no. 2, pp. 239–250, Mar. 2021, doi: 10.1007/s40291-021-00515-9.
[204] K. Shoda *et al.*, “Monitoring the HER2 copy number status in circulating tumor DNA by droplet digital PCR in patients with gastric cancer,” *Gastric Cancer Off. J. Int. Gastric Cancer Assoc. Jpn. Gastric Cancer Assoc.*, vol. 20, no. 1, pp. 126–135, Jan. 2017, doi: 10.1007/s10120-016-0599-z.
[205] K. S. Lee *et al.*, “Digital polymerase chain reaction for detecting c-MYC copy number gain in tissue and cell-free plasma samples of colorectal cancer patients,” *Sci. Rep.*, vol. 9, no. 1, p. 1611, Feb. 2019, doi: 10.1038/s41598-018-38415-4.
[206] M.-H. Delfau-Larue *et al.*, “Total metabolic tumor volume, circulating tumor cells, cell-free DNA: distinct prognostic value in follicular lymphoma,” *Blood Adv.*, vol. 2, no. 7, pp. 807–816, Apr. 2018, doi: 10.1182/bloodadvances.2017015164.
[207] C. Pott, M. Brüggemann, M. Ritgen, V. H. J. van der Velden, J. J. M. van Dongen, and M. Kneba, “MRD Detection in B-Cell Non-Hodgkin Lymphomas Using Ig Gene Rearrangements and Chromosomal Translocations as Targets for Real-Time Quantitative PCR,” *Methods Mol. Biol. Clifton NJ*, vol. 1956, pp. 199–228, 2019, doi: 10.1007/978-1-4939-9151-8_9.
[208] T. C. Glenn, “Field guide to next-generation DNA sequencers,” *Mol. Ecol. Resour.*, vol. 11, no. 5, pp. 759–769, Sep. 2011, doi: 10.1111/j.1755-0998.2011.03024.x.
[209] N. J. Loman *et al.*, “Performance comparison of benchtop high-throughput sequencing platforms,” *Nat. Biotechnol.*, vol. 30, no. 5, pp. 434–439, May 2012, doi: 10.1038/nbt.2198.
[210] J. J. Salk, M. W. Schmitt, and L. A. Loeb, “Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations,” *Nat. Rev. Genet.*, vol. 19, no. 5, pp. 269–285, May 2018, doi: 10.1038/nrg.2017.117.
[211] I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein, “Detection and quantification of rare mutations with massively parallel sequencing,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 23, pp. 9530–9535, Jun. 2011, doi: 10.1073/pnas.1105422108.
[212] J. Phallen *et al.*, “Direct detection of early-stage cancers using circulating tumor DNA,” *Sci. Transl. Med.*, vol. 9, no. 403, p. eaan2415, Aug. 2017, doi: 10.1126/scitranslmed.aan2415.
[213] T. Forshew *et al.*, “Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA,” *Sci. Transl. Med.*, vol. 4, no. 136, p. 136ra68, May 2012, doi: 10.1126/scitranslmed.3003726.
[214] D. Gale *et al.*, “Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free DNA,” *PloS One*, vol. 13, no. 3, p. e0194630, 2018, doi: 10.1371/journal.pone.0194630.
[215] F. Fostira *et al.*, “1158P - Blood-based testing of mutations in patients with head and neck squamous cell carcinoma (HNSCC) using highly sensitive SafeSEQ technology,” *Ann. Oncol.*, vol. 30, p. v469, Oct. 2019, doi: 10.1093/annonc/mdz252.050.
[216] J. Tie *et al.*, “Prognostic significance of postsurgery circulating tumor DNA in nonmetastatic colorectal cancer: Individual patient pooled analysis of three cohort studies,” *Int. J. Cancer*, vol. 148, no. 4, pp. 1014–1026, Feb. 2021, doi: 10.1002/ijc.33312.
[217] C. Xu, M. R. Nezami Ranjbar, Z. Wu, J. DiCarlo, and Y. Wang, “Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller,”

BMC Genomics, vol. 18, no. 1, p. 5, Jan. 2017, doi: 10.1186/s12864-016-3425-4.

[218] W. C. Sim *et al.*, "Non-invasive detection of actionable mutations in advanced non-small-cell lung cancer using targeted sequencing of circulating tumor DNA," *Lung Cancer Amst. Neth.*, vol. 124, pp. 154–159, Oct. 2018, doi: 10.1016/j.lungcan.2018.08.007.

[219] D. Zou *et al.*, "Circulating tumor DNA is a sensitive marker for routine monitoring of treatment response in advanced colorectal cancer," *Carcinogenesis*, vol. 41, no. 11, pp. 1507–1517, Nov. 2020, doi: 10.1093/carcin/bgaa102.

[220] P.-J. Viailly *et al.*, "Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers," *BMC Bioinformatics*, vol. 22, no. 1, p. 120, Mar. 2021, doi: 10.1186/s12859-021-04060-4.

[221] S. V. Bratman, A. M. Newman, A. A. Alizadeh, and M. Diehn, "Potential clinical utility of ultrasensitive circulating tumor DNA detection with CAPP-Seq," *Expert Rev. Mol. Diagn.*, vol. 15, no. 6, pp. 715–719, Jun. 2015, doi: 10.1586/14737159.2015.1019476.

[222] A. M. Newman *et al.*, "An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage," *Nat. Med.*, vol. 20, no. 5, pp. 548–554, May 2014, doi: 10.1038/nm.3519.

[223] D. Klass *et al.*, "Analysis of Circulating Tumor DNA in Esophageal Carcinoma Patients Treated With Chemoradiation Therapy," *Int. J. Radiat. Oncol.*, vol. 93, no. 3, Supplement, pp. S104–S105, Nov. 2015, doi: 10.1016/j.ijrobp.2015.07.251.

[224] F. Scherer *et al.*, "Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA," *Sci. Transl. Med.*, vol. 8, no. 364, p. 364ra155, Nov. 2016, doi: 10.1126/scitranslmed.aai8545.

[225] J. C. Dudley *et al.*, "Detection and Surveillance of Bladder Cancer Using Urine Tumor DNA," *Cancer Discov.*, vol. 9, no. 4, pp. 500–509, Apr. 2019, doi: 10.1158/2159-8290.CD-18-0825.

[226] E. Bohers *et al.*, "Somatic mutations of cell-free circulating DNA detected by next-generation sequencing reflect the genetic changes in both germinal center B-cell-like and activated B-cell-like diffuse large B-cell lymphomas at the time of diagnosis," *Haematologica*, vol. 100, no. 7, pp. e280-284, Jul. 2015, doi: 10.3324/haematol.2015.123612.

[227] M. Fontanilles *et al.*, "Non-invasive detection of somatic mutations using next-generation sequencing in primary central nervous system lymphoma," *Oncotarget*, vol. 8, no. 29, pp. 48157–48168, Jul. 2017, doi: 10.18632/oncotarget.18325.

[228] V. Camus *et al.*, "Digital PCR for quantification of recurrent and potentially actionable somatic mutations in circulating free DNA from patients with diffuse large B-cell lymphoma," *Leuk. Lymphoma*, vol. 57, no. 9, pp. 2171–2179, Sep. 2016, doi: 10.3109/10428194.2016.1139703.

[229] P. Vandenberghe *et al.*, "Non-invasive detection of genomic imbalances in Hodgkin/Reed-Sternberg cells in early and advanced stage Hodgkin's lymphoma by sequencing of circulating cell-free DNA: a technical proof-of-principle study," *Lancet Haematol.*, vol. 2, no. 2, pp. e55-65, Feb. 2015, doi: 10.1016/S2352-3026(14)00039-8.

[230] L. Bessi *et al.*, "Somatic mutations of cell-free circulating DNA detected by targeted next-generation sequencing and digital droplet PCR in classical Hodgkin lymphoma," *Leuk. Lymphoma*, vol. 60, no. 2, pp. 498–502, Feb. 2019, doi: 10.1080/10428194.2018.1492123.

[231] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, May 2016, doi: 10.1038/nrg.2016.49.

[232] E. R. Mardis, "Next-generation sequencing platforms," *Annu. Rev. Anal. Chem. Palo Alto Calif*, vol. 6, pp. 287–303, 2013, doi: 10.1146/annurev-anchem-062012-092628.

[233] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using phred. I. Accuracy assessment," *Genome Res.*, vol. 8, no. 3, pp. 175–185, Mar. 1998, doi: 10.1101/gr.8.3.175.

[234] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Res.*, vol. 8, no. 3, pp. 186–194, Mar. 1998.

[235] T. Kivioja *et al.*, "Counting absolute numbers of molecules using unique molecular identifiers," *Nat. Methods*, vol. 9, no. 1, pp. 72–74, Nov. 2011, doi: 10.1038/nmeth.1778.

[236] T. Smith, A. Heger, and I. Sudbery, "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome Res.*, vol. 27, no. 3, pp. 491–499, Mar. 2017, doi: 10.1101/gr.209601.116.

[237] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *ArXiv13033997 Q-Bio*, May 2013, Accessed: Jul. 30, 2021. [Online]. Available: http://arxiv.org/abs/1303.3997

[238] G. A. Van der Auwera *et al.*, “From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline,” *Curr. Protoc. Bioinforma.*, vol. 43, p. 11.10.1-11.10.33, 2013, doi: 10.1002/0471250953.bi1110s43.

[239] B. Merriman, Ion Torrent R&D Team, and J. M. Rothberg, “Progress in ion torrent semiconductor chip based sequencing,” *Electrophoresis*, vol. 33, no. 23, pp. 3397–3417, Dec. 2012, doi: 10.1002/elps.201200424.

[240] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, Art. no. 1, May 2011, doi: 10.14806/ej.17.1.200.

[241] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinforma. Oxf. Engl.*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[242] C. Del Fabbro, S. Scalabrin, M. Morgante, and F. M. Giorgi, “An extensive evaluation of read trimming effects on Illumina NGS data analysis,” *PloS One*, vol. 8, no. 12, p. e85024, 2013, doi: 10.1371/journal.pone.0085024.

[243] A. D. Smith, Z. Xuan, and M. Q. Zhang, “Using quality scores and longer reads improves accuracy of Solexa read mapping,” *BMC Bioinformatics*, vol. 9, p. 128, Feb. 2008, doi: 10.1186/1471-2105-9-128.

[244] H. Li, J. Ruan, and R. Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores,” *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, Nov. 2008, doi: 10.1101/gr.078212.108.

[245] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li, “ZOOM! Zillions of oligos mapped,” *Bioinforma. Oxf. Engl.*, vol. 24, no. 21, pp. 2431–2437, Nov. 2008, doi: 10.1093/bioinformatics/btn416.

[246] H. Jiang and W. H. Wong, “SeqMap: mapping massive amount of oligonucleotides to the genome,” *Bioinforma. Oxf. Engl.*, vol. 24, no. 20, pp. 2395–2396, Oct. 2008, doi: 10.1093/bioinformatics/btn429.

[247] R. Li, Y. Li, K. Kristiansen, and J. Wang, “SOAP: short oligonucleotide alignment program,” *Bioinforma. Oxf. Engl.*, vol. 24, no. 5, pp. 713–714, Mar. 2008, doi: 10.1093/bioinformatics/btn025.

[248] D. Campagna *et al.*, “PASS: a program to align short sequences,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 7, pp. 967–968, Apr. 2009, doi: 10.1093/bioinformatics/btp087.

[249] Y. J. Kim *et al.*, “ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 11, pp. 1424–1425, Jun. 2009, doi: 10.1093/bioinformatics/btp178.

[250] M. Burrows and D. J. Wheeler, “A block-sorting lossless data compression algorithm,” p. 24.

[251] B. Langmead, “Aligning short sequencing reads with Bowtie,” *Curr. Protoc. Bioinforma.*, vol. Chapter 11, p. Unit 11.7, Dec. 2010, doi: 10.1002/0471250953.bi1107s32.

[252] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[253] M. A. DePristo *et al.*, “A framework for variation discovery and genotyping using next-generation DNA sequencing data,” *Nat. Genet.*, vol. 43, no. 5, pp. 491–498, May 2011, doi: 10.1038/ng.806.

[254] “Genomics in the Cloud [Book].” https://www.oreilly.com/library/view/genomics-in-the/9781491975183/ (accessed Jul. 30, 2021).

[255] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010, doi: 10.1093/nar/gkq603.

[256] W. McLaren *et al.*, “The Ensembl Variant Effect Predictor,” *Genome Biol.*, vol. 17, no. 1, p. 122, Jun. 2016, doi: 10.1186/s13059-016-0974-4.

[257] 1000 Genomes Project Consortium *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.

[258] K. J. Karczewski *et al.*, “The ExAC browser: displaying reference data information from over 60 000 exomes,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D840–D845, Jan. 2017, doi: 10.1093/nar/gkw971.

[259] A. J. Iafrate *et al.*, “Detection of large-scale variation in the human genome,” *Nat. Genet.*, vol. 36, no. 9, pp. 949–951, Sep. 2004, doi: 10.1038/ng1416.

[260] S. T. Sherry, M. Ward, and K. Sirotkin, “dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation,” *Genome Res.*, vol. 9, no. 8, pp. 677–679, Aug. 1999.

[261] S. A. Forbes *et al.*, “The Catalogue of Somatic Mutations in Cancer (COSMIC),” *Curr. Protoc. Hum. Genet.*, vol. Chapter 10, p. Unit 10.11, Apr. 2008, doi: 10.1002/0471142905.hg1011s57.

[262] M. J. Landrum *et al.*, “ClinVar: improving access to variant interpretations and supporting evidence,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062–D1067, Jan. 2018, doi: 10.1093/nar/gkx1153.

[263] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, “Identifying a

High Fraction of the Human Genome to be under Selective Constraint Using GERP++," *PLOS Comput. Biol.*, vol. 6, no. 12, p. e1001025, Dec. 2010, doi: 10.1371/journal.pcbi.1001025.

[264] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, Jan. 2010, doi: 10.1101/gr.097857.109.

[265] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003, doi: 10.1093/nar/gkg509.

[266] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010, doi: 10.1038/nmeth0410-248.

[267] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, Jan. 2019, doi: 10.1093/nar/gky1016.

[268] D. Quang, Y. Chen, and X. Xie, "DANN: a deep learning approach for annotating the pathogenicity of genetic variants," *Bioinformatics*, vol. 31, no. 5, pp. 761–763, Mar. 2015, doi: 10.1093/bioinformatics/btu703.

[269] S. Hwang, E. Kim, I. Lee, and E. M. Marcotte, "Systematic comparison of variant calling pipelines using gold standard personal exome variants," *Sci. Rep.*, vol. 5, no. 1, p. 17875, Dec. 2015, doi: 10.1038/srep17875.

[270] A. Supernat, O. V. Vidarsson, V. M. Steen, and T. Stokowy, "Comparison of three variant callers for human whole genome sequencing," *Sci. Rep.*, vol. 8, no. 1, p. 17851, Dec. 2018, doi: 10.1038/s41598-018-36177-7.

[271] H. M. Schilbert, A. Rempel, and B. Pucker, "Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data," *Plants*, vol. 9, no. 4, p. 439, Apr. 2020, doi: 10.3390/plants9040439.

[272] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *ArXiv12073907 Q-Bio*, Jul. 2012, Accessed: Aug. 30, 2021. [Online]. Available: http://arxiv.org/abs/1207.3907

[273] A. Cornish and C. Guda, "A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference," *BioMed Res. Int.*, vol. 2015, p. 456479, 2015, doi: 10.1155/2015/456479.

[274] R. Poplin *et al.*, "Scaling accurate genetic variant discovery to tens of thousands of samples," Jul. 2018. doi: 10.1101/201178.

[275] A. Wilm *et al.*, "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets," *Nucleic Acids Res.*, vol. 40, no. 22, pp. 11189–11201, Dec. 2012, doi: 10.1093/nar/gks918.

[276] E. Muller *et al.*, "OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice," *Oncotarget*, vol. 7, no. 48, pp. 79485–79493, Nov. 2016, doi: 10.18632/oncotarget.13103.

[277] T. D. Andrews, Y. Jeelall, D. Talaulikar, C. C. Goodnow, and M. A. Field, "DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations," *PeerJ*, vol. 4, p. e2074, 2016, doi: 10.7717/peerj.2074.

[278] M. Shugay *et al.*, "MAGERI: Computational pipeline for molecular-barcoded targeted resequencing," *PLoS Comput. Biol.*, vol. 13, no. 5, p. e1005480, May 2017, doi: 10.1371/journal.pcbi.1005480.

[279] C. Xu *et al.*, "smCounter2: an accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers," *Bioinforma. Oxf. Engl.*, vol. 35, no. 8, pp. 1299–1309, Apr. 2019, doi: 10.1093/bioinformatics/bty790.

[280] Y. Guo, J. Li, C.-I. Li, J. Long, D. C. Samuels, and Y. Shyr, "The effect of strand bias in Illumina short-read sequencing data," *BMC Genomics*, vol. 13, p. 666, Nov. 2012, doi: 10.1186/1471-2164-13-666.

[281] G. Ivády *et al.*, "Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system," *BMC Genomics*, vol. 19, no. 1, p. 158, Feb. 2018, doi: 10.1186/s12864-018-4544-x.

[282] X. Fan, T. E. Abbott, D. Larson, and K. Chen, "BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping," *Curr. Protoc. Bioinforma.*, vol. 45, p. 15.6.1–11, 2014, doi: 10.1002/0471250953.bi1506s45.

[283] J. O. Korbel *et al.*, "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome Biol.*, vol. 10, no. 2, p. R23, Feb. 2009, doi: 10.1186/gb-2009-10-2-r23.

[284] A. Gillet-Markowska, H. Richard, G. Fischer, and I. Lafontaine, "Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries," *Bioinforma. Oxf. Engl.*, vol.

31, no. 6, pp. 801–808, Mar. 2015, doi: 10.1093/bioinformatics/btu730.
[285] J. Zhang, J. Wang, and Y. Wu, “An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data,” *BMC Bioinformatics*, vol. 13 Suppl 6, p. S6, Apr. 2012, doi: 10.1186/1471-2105-13-S6-S6.
[286] K. Trappe, A.-K. Emde, H.-C. Ehrlich, and K. Reinert, “Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone,” *Bioinforma. Oxf. Engl.*, vol. 30, no. 24, pp. 3484–3490, Dec. 2014, doi: 10.1093/bioinformatics/btu431.
[287] Y. Jiang, Y. Wang, and M. Brudno, “PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants,” *Bioinforma. Oxf. Engl.*, vol. 28, no. 20, pp. 2576–2583, Oct. 2012, doi: 10.1093/bioinformatics/bts484.
[288] V. Boeva *et al.*, “Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data,” *Bioinforma. Oxf. Engl.*, vol. 30, no. 24, pp. 3443–3450, Dec. 2014, doi: 10.1093/bioinformatics/btu436.
[289] H. Bengtsson, P. Neuvial, V. E. Seshan, A. B. Olshen, P. T. Spellman, and R. A. Olshen, *PSCBS: Analysis of Parent-Specific DNA Copy Numbers*. 2019. Accessed: Apr. 13, 2021. [Online]. Available: https://CRAN.R-project.org/package=PSCBS
[290] X. Yuan, J. Zhang, and L. Yang, “IntSIM: An Integrated Simulator of Next-Generation Sequencing Data,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 441–451, Feb. 2017, doi: 10.1109/TBME.2016.2560939.
[291] X. Yuan, M. Gao, J. Bai, and J. Duan, “SVSR: A Program to Simulate Structural Variations and Generate Sequencing Reads for Multiple Platforms,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, no. 3, pp. 1082–1091, Jun. 2020, doi: 10.1109/TCBB.2018.2876527.
[292] C. Kockan *et al.*, “SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA,” *Bioinforma. Oxf. Engl.*, vol. 33, no. 1, pp. 26–34, Jan. 2017, doi: 10.1093/bioinformatics/btw536.
[293] V. Sater *et al.*, “UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries,” *Bioinforma. Oxf. Engl.*, vol. 36, no. 9, pp. 2718–2724, May 2020, doi: 10.1093/bioinformatics/btaa053.
[294] D. M. Kurtz *et al.*, “Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA,” *Nat. Biotechnol.*, Jul. 2021, doi: 10.1038/s41587-021-00981-w.

RÉSUMÉ

Développement d'outils bioinformatiques pour l'analyse de l'ADN tumoral libre circulant des lymphomes

En France, le lymphome est le 6e cancer le plus fréquent avec chaque année environ 15 000 nouveaux cas diagnostiqués et près de 4 500 décès. Derrière cette maladie se cache en réalité une très grande hétérogénéité tant sur le plan clinique que phénotypique. Le développement des approches d'immunohistochimie, de cytogénétique et l'avènement récent des séquenceurs de nouvelle génération permettent une caractérisation toujours plus précise de cette maladie via la quantification de plusieurs biomarqueurs à partir de la tumeur. L'intégration de ces différentes sources de données a permis une meilleure classification des lymphomes aujourd'hui scindés en plusieurs dizaines d'entités distinctes.

Le concept de biopsie liquide, qui regroupe un ensemble d'examens réalisés à partir de fluides biologiques tels que le plasma, est devenu un enjeu majeur de ces dernières années. La biopsie liquide, en permettant une détection non invasive des biomarqueurs issus de la tumeur à différents temps de la prise en charge du patient, permet de suivre l'évolution de la maladie et pourrait permettre à plus ou moins moyen terme de proposer aux patients le bon diagnostic, le bon traitement et au bon moment de la maladie via le développement des thérapies ciblées.

Les travaux de ce mémoire visent à présenter les différents développements bioinformatiques menés afin de mieux caractériser les biopsies liquides par séquençage à haut-débit. Différents algorithmes, intégrant ou non des barcodes moléculaires, seront détaillés et associés à des exemples d'application en conditions réelles. Un état de l'art antérieur au développement des nouveaux outils sera présenté et leurs limites seront discutées.

Mots-clés Lymphome, Biopsie Liquide, Séquençage de nouvelle génération