Industrialisation des procédures d'analyses de données de séquençage pan-génomiques constitutionnelles

Quentin Testard

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Biologie du développement - Oncogenèse

Arrêté ministériel : 25 mai 2016

Présentée par

Quentin TESTARD

Thèse dirigée par **Julien Thevenon**, Université Grenoble Alpes,

co-encadrée par **Jean-François Taly**, Eurofins Biomnis
et **Laure Raymond**, Eurofins Biomnis.

Préparée au sein du **Laboratoire IAB : Epigenetics, Environment, Cell Plasticity, Cancer (UGA / Inserm U1209 / CNRS UMR 5309)** et de **l'École Doctorale Chimie et Sciences du Vivant**.

Industrialisation des procédures d'analyses de données de séquençage pan-génomiques constitutionnelles.

Industrialization of constitutional pan-genomic sequencing data analysis procedures.

Thèse soutenue publiquement le **10 décembre 2021**,
devant le jury composé de :

Monsieur Julien Thevenon
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Université Grenoble Alpes / CHU Grenoble Alpes, Directeur de thèse
Madame Marie De Tayrac
PROFESSEURE DES UNIVERSITES - PRATICIEN HOSPITALIER, Université de Rennes 1 / CHU Rennes, Rapporteure
Madame Christel Thauvin-Robinet
PROFESSEURE DES UNIVERSITES - PRATICIEN HOSPITALIER, Université de Bourgogne / CHU Dijon Bourgogne, Rapporteure
Monsieur Nicolas Chatron
MAITRE DE CONFERENCE DES UNIVERSITES - PRATICIEN HOSPITALIER, Université Claude Bernard Lyon / HCL Lyon, Examinateur
Madame Véronique Geoffroy
INGENIEURE DE RECHERCHE, Université de Strasbourg, Examinateure
Monsieur Pierre Ray
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER, Université Grenoble Alpes / CHU Grenoble Alpes, Président du jury

Remerciements

Si quelqu'un m'avait dit un jour, lorsque j'ai choisi de commencer mes études supérieures par un DUT de technicien de laboratoire, que je finirais par faire un doctorat, ma réponse aurait très probablement été : "Absolument aucune chance, impossible". Clairvoyant et peu têtue que j'étais (ça n'a pas changé et je dirai même que ça a empiré), j'étais résolu à faire des études courtes. Pourtant, le DUT a entraîné une inscription en troisième année de licence qui elle-même a motivé une inscription en première année de Master de Biologie.

Puis un jour, le questionnement : "Vais-je trouver un travail avec un Master de Biologie ? Est-ce que je me vois faire de la recherche théorique en Biologie toute ma vie ? Certainement pas". Ce que je voyais autour de moi lors de mes différents stages m'intéressait intellectuellement, mais ne me plaisait pas en tant qu'activité professionnelle. Le manque d'impact concret ou de reconnaissance dans une carrière de recherche académique en Biologie ? Allez savoir.

Les hasards de la vie (car la vie, *je dirais que c'est d'abord des rencontres*), ont fait que c'est à ce moment-là que je me suis intéressé à la Bioinformatique. J'ai donc décidé de changer de voie et de m'inscrire en Master de Bioinformatique. Après la découverte d'une science aux ramifications insoupçonnées, je n'avais absolument aucune idée de ce dans quoi je m'engageais. Puis la fin du Master arriva et mon avis n'avait pas vraiment changé, "Une thèse académique ? Au grand jamais !".

Pourtant, c'est parfois la rencontre de deux personnes qui peuvent bouleverser le destin d'une troisième. Celle d'un médecin, à la recherche d'un profil atypique et d'une dirigeante de Master, avec un étudiant étrange qui refuse de s'engager dans la recherche académique. Mais au moins pour une fois, j'avais vu juste j'ai bien fini par faire une thèse, mais elle n'était ni académique, ni théorique.

Gravitant dans le domaine hospitalier et plus largement dans le domaine de la santé depuis le début de mes études, cette thèse m'a confirmé que ce qui me plaît, c'est de savoir que mon travail est utile. Non pas que la recherche théorique soit inutile, sans elle, je ne serais rien. Il y a pourtant un lien bien plus direct et concret entre mon travail et son application, que si je travaillais dans un institut de recherche en tant que chercheur. En tout cas, c'est comme cela que je vois les choses.

Alors, je commencerais par remercier le Dr Anna-Sophie Fiston-Lavier, pour m'avoir encouragé en tant qu'étudiant, mais également pour avoir su glisser mon nom (le bon, je présume) à la bonne personne, au bon moment. J'aimerais également remercier le Pr Julien Thevenon, le Dr Laure Raymond et le Dr Jean-François Taly, de tout d'abord avoir cru en moi alors que j'étais à 16 000 km de la France. De m'avoir fait découvrir le monde de la génétique et de la bioinformatique clinique. D'avoir pu faire preuve d'écoute et de soutien dans les moments les plus difficiles et les plus délicats. Je tenais à vous remercier pour ces bientôt 4 ans.

J'aimerais ensuite remercier mes collègues de travail et plus particulièrement la Team Bio IT d'Eurofins Biomnis, Aurore, Nicolas et Fanny. Merci pour votre aide et votre temps. Une pensée également amicale pour Francis qui soutiendra toute sa vie une équipe honteuse, sans histoire, valant le PIB du Malawi, pour faire 1-1 contre le FC Bruges. J'espère que quand tu liras ces mots, Messi aura mis au moins un but en Ligue 1. Je remercie également l'équipe technique, les garçons aux mains en or, Mohamed et Pascal.

Je remercie Quentin, Virginie et Kévin, compagnons bioinformaticiens du Pavillon Chissé, merci pour votre aide, pour les randonnées, les pique-niques, les soirées jeux, les bières et les viennoiseries ainsi que les gâteaux ! Et oui, il s'agirait de ne pas laisser vos ordinateurs sans surveillance.

Je remercie également Émilien, pour l'aide que tu as pu m'apporter notamment lors du projet HPC-CHU, ainsi que pour ta curiosité et ton intérêt envers la Bioinformatique.

Je remercie Bruno, pour ton aide et ta bienveillance quant à mon utilisation des *clusters* CIMENT.

Merci à Florent pour ton aide, les cafés et tes formations.

À mes amis,
Aux compagnons de galère, Quentin, Valentin et Anna. Savoir que l'on n'est pas seul dans les moments de doute est important. Il fut assez surprenant de remarquer que broyer du noir en votre compagnie pouvait apporter du réconfort, en tout cas pour moi. Mention spéciale tout de même pour Valentin, qui doit supporter la vision d'horreur de mes one-liners Bash, de mon code spaghetti, mais aussi l'odeur de mes spaghettis, entre autres, dans la vraie vie.

Au sang de la vigne, car c'est le sobriquet du moment, Hugo, Xavier, Rémy, Catalina, Nicolas pour les tranches de rigolade, les commentaires de société frustrants, les pics acerbes (sans animosité aucune, la plupart du temps en tout cas) et les parties de jeu. Tout ça grâce au sacro-saint Telegram. C'est-y pas génial la technologie ? J'ai hâte qu'on se retrouve tous ensemble à Montpellier.

Aux poulettes, Laura, Gonché et Marianne, j'espère que l'on va se voir tous ensemble très rapidement, sinon je vais être 13NRV. Maintenant que tout le monde est un docteur poulette, on va pouvoir se déhancher de fou sur la piste de danse, tout en buvant du pastis, en tout cas pour ma part. Si la Bioinformatique ça capote, on arrête tout et on monte un restaurant, brunch à base de mozzarella uniquement. Je vous fais un topo du POC ASAP sur ma Lenovo Yoga Tablet 2. Deal ? Mention spéciale à Marianne, car ton soutien dans les moments les plus difficiles a beaucoup compté.

Aux Martégaux, Fabien, Romain et Léopold, j'espère pouvoir vous voir plus souvent, mais il va falloir arrêter de croire que je vais venir vivre à Marseille. Hein Léopold ?

Aux Bretons, Zap, Nathan, Fabien, Annaïg, Mathilde, Motig. Sans déconner, quelle idée d'aller TOUS vous fourrer là-bas, c'est à l'autre bout de la France. Déjà qu'on ne se voyait pas assez souvent, vous, vous allez vous enterrer à TROU pauMÉ. Je vous déteste, surtout toi Annaïg, avec un prénom comme ça, on sait tout de suite que c'est ta faute. Vous auriez pu aller en Islande, j'aurais mis autant de temps pour venir vous voir, mais au moins ça ne sentirait pas le lisier et le vote LFI. Mais non, ça va boire du cidre et se tremper le cul dans la fontaine de Brest. Je sais que vous ne m'aimez pas, vous allez même à la Baraka sans moi. Tout ça parce que j'ai jamais été disponible pendant neuf ans et d'un coup d'un seul ça va habiter dans une région

où les gens peuvent se déboîter congénitalement la hanche. En plus pour venir vous voir faut prendre des deux fois deux voies ou des quatre voies je sais pas quoi, super les potes ... Je vous aime, j'arrive, faites-moi une place.

À Damien, si toi aussi tu pouvais aller te perdre en Bretagne ça m'évitera des aller-retours, merci.

À Zonzon et Alex, allez l'OM.

À Jérémie et Audrey, je ne vous promets pas de vous inviter à vivre à la maison, mais j'espère vous voir bientôt.

À mes parents, à ma famille et à ma belle-famille, qui doivent se demander pourquoi j'ai bien pu choisir une pareille carrière. Veuillez m'excuser d'avance pour toutes les choses que j'ai promis de faire après la thèse, car je doute de les tenir. Merci d'être aussi patient avec moi et de comprendre les sacrifices que j'ai pu faire.

À ma moitié, à ma compagne, Cynthia, sans qui je n'aurai jamais pu réussir. Tu es le plus grand soutien que je puisse espérer, même si tu ne t'en rends pas compte. Je t'aime.

À mes poulets ... Miaou ?

Table des matières

Table des figures

Liste des tableaux

Liste des snippets

Introduction

Le nombre de maladies rares est estimé à entre 5000 et 8000[1] pathologies distinctes. Elles sont individuellement rares puisque par définition elles affectent moins de 1 individu sur 2000 dans la population générale, mais leur nombre les rend collectivement fréquentes. En effet, entre 3 à 6% de la population mondiale est affectée par une maladie rare, soit théoriquement 3 à 4 millions de Français ou 25 millions d'Européens.

En raison du nombre limité de patients atteints de certaines maladies et du manque de connaissances à leur sujet, le diagnostic pour les patients est souvent retardé. Cela peut entraîner des complications et souffrances irréversibles pour les patients, leurs familles et amis. C'est ce qui est appelé, l'odyssée diagnostique[2]. Il est estimé que 70% des maladies rares d'origine génétique se manifestent dès l'enfance. Les maladies rares sont une des causes les plus fréquentes de mortalité infantile puisque 30% des enfants atteints meurent avant l'âge de 5 ans[2][3].

Plus de 80% des maladies rares auraient une origine génétique, en grande majorité monogénique[4]. L'importance de disposer d'un diagnostic précis pour personnaliser la prise en charge explique la recrudescence d'analyses de biologie moléculaire (analyse à l'échelle du génome, d'un gène ou d'un nucléotide). La démocratisation des technologies de séquençage à haut débit depuis 2005 a permis l'acquisition massive de données génomiques, que ce soit d'individus sains, ou de patients[5]. Pourtant, établir un diagnostic de maladie rare avec les technologies de biologie moléculaire actuelles reste difficile.

L'application à échelle industrielle des analyses de séquençage d'exome et de génome associée avec l'augmentation constante des connaissances médicales représente un espoir concret pour apporter un diagnostic à la majorité des patients concernés par une maladie rare suspecte d'être génétique.

L'objectif de cette thèse a été tout d'abord de mettre en place un environnement opérationnel d'analyse de données de séquençage pan-génomiques constitutionnelles selon les standards industriels et les bonnes pratiques en vigueur au sein des laboratoires de biologie médicale Eurofins Biomnis et du CHU Grenoble-Alpes. Une fois ceci fait, les données ont ensuite été traitées à l'aide de méthodologies novatrices afin de bénéficier d'informations supplémentaires sur des variations potentiellement pathogènes, notamment les variations de nombre de copies, qui étaient jusqu'alors inexploitées et ainsi d'améliorer le rendement diagnostic des analyses.

Depuis le transfert vers les équipes, le travail de cette thèse est utilisé avec succès dans la routine diagnostique du laboratoire Eurofins Biomnis et du CHU Grenoble-Alpes.

Mes travaux ont également été utilisés dans des projets de recherche d'équipes de l'Institute for Advanced Bioscience (IAB) et sont présentés dans ce manuscrit. Ils comprennent : la détection de variations somatiques de petite taille dans un modèle *in vitro* de développement de cancer lié au microenvironnement cellulaire, la détection de variations structurales à l'aide de la technologie de séquençage récente Oxford Nanopore Technologies et la comparaison de méthodologies de priorisations de variations génétiques à l'aide de descriptions cliniques basées sur les termes HPO.

État de l'art

1.1 Généralité sur les acides nucléiques

1.1.1 L'ADN

L'acide désoxyribonucléique plus connu sous son acronyme ADN est le support de l'information génétique chez l'ensemble des organismes vivants ainsi que chez certains virus. La molécule d'ADN est un polymère, c'est-à-dire une molécule composée de monomères (unité de répétition) enchaînés un grand nombre de fois, formant ainsi une macromolécule. Chacun de ces monomères est également appelé nucléotide. Au nombre de quatre, ils sont formés d'un groupement phosphate, lié à un pentose, un monosaccharide à cinq atomes de carbone, le désoxyribose, lui-même lié à une base nucléique qui diffère par son groupement azoté. Sont présentes d'une part les purines, l'Adénine (A) et la Guanine (G) et d'autre part les pyrimidines, la Cytosine (C) et la Thymine (T). Des propriétés physico-chimiques des nucléotides dépendent la fameuse structure en double hélice de l'ADN, dont la découverte, longtemps attribuée à Watson et Crick pour leurs travaux publiés en 1953[6], n'aurait pu être possible sans les travaux préliminaires de Rosalind Franklin[7].

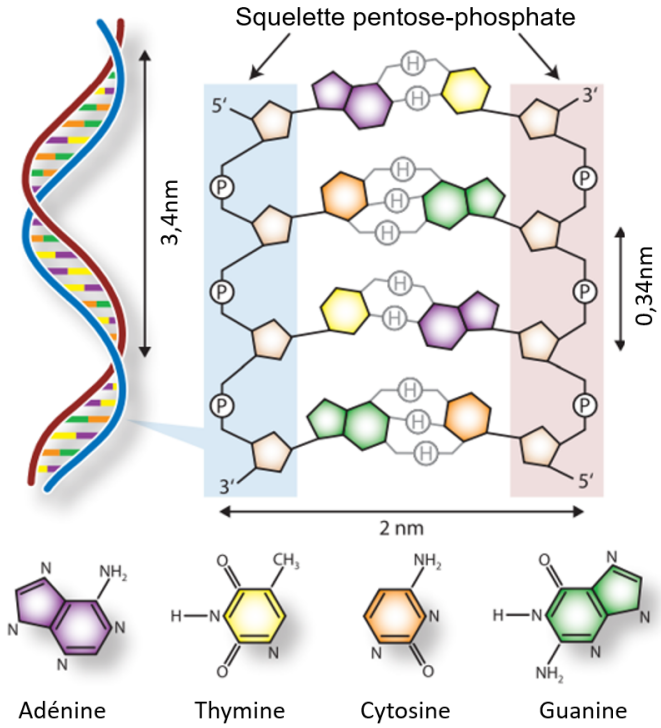


FIGURE 1.1 – Représentation de la structure en double hélice de l'ADN conséquence des liaisons physico-chimiques entre nucléotides.
Adapté d'après *Gauthier, Michel. (2007). Simulation of polymer translocation through small channels : A molecular dynamics study and a new Monte Carlo approach.*

Les nucléotides sont assemblés par paires complémentaires. Les molécules de phosphate et de pentose des nucléotides sont liées *via* des liaisons phosphodiesters et forment un squelette pentose-phosphate sur lequel se rattachent les différentes bases azotées (à l'extérieur du squelette). Ces dernières sont reliées entre elles *via* des liaisons hydrogènes entre bases puriniques et pyrimidiques selon les couples Adénine-Thymine (deux ponts hydrogénés) et Guanine-Cytosine (trois ponts hydrogénés), voir *Figure 1.1*.

L'ADN est constitué de deux molécules (brins) complémentaires (bicaténaire) et antiparallèles, chaque brin étant le complément inverse de l'autre. Par convention, les brins d'ADN sont représentés en fonction de l'orientation de leur sucre initial et terminal de la gauche vers la droite dans le sens 5' vers 3'. Le côté 5' désignant l'extrémité du brin terminant par un groupement phosphate et le 3' par un groupement hydroxyle.

L'ADN nucléaire est stocké à l'intérieur du noyau de chaque cellule sous la forme de chromosomes. Ces derniers sont constitués d'une longue molécule d'ADN qui va s'enrouler de manière successive afin de se densifier et se compresser. La molécule d'ADN va tout d'abord s'enrouler autour de protéines appelées histones afin de former des nucléosomes (146 à 147 nucléotides sont nécessaires pour former un nucléosome). Les nucléosomes ainsi enchainés vont eux même s'organiser en boucle afin de se condenser une fois de plus, formant ainsi de la chromatine, unité structurale du chromosome, voir *Figure 1.2*.

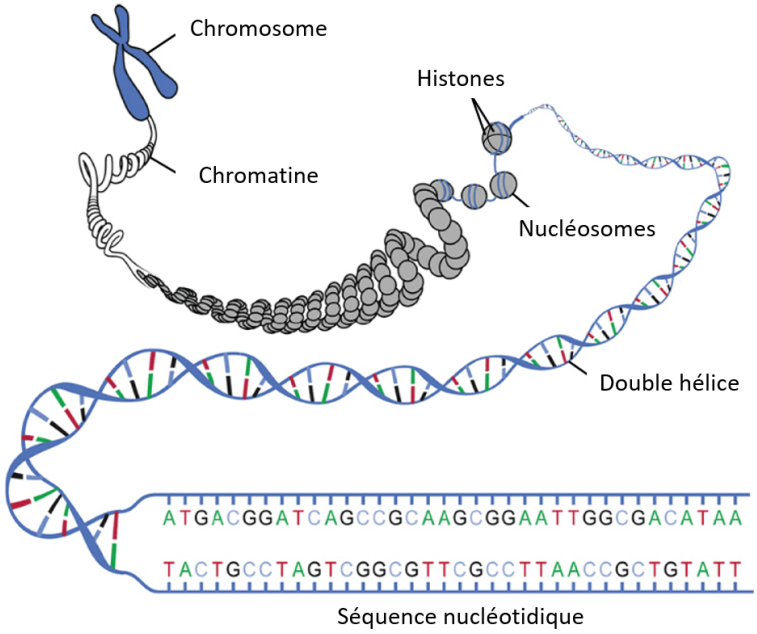


FIGURE 1.2 – Représentation schématique de la structure de l'ADN bicaténaire au chromosome. Adapté d'après *DeSaix, P., Betts, J. G., Johnson, E., Johnson, J. E., Korol, O., Kruse, D. H., ... Young, K. A. (2013). Anatomy Physiology : openStax.*

Le génome humain est d'une complexité importante, non élucidée à ce jour. Certaines régions sont connues pour être relativement difficiles à étudier, notamment les régions situées au niveau des centromères ainsi qu'à leurs extrémités (télomères). Ces zones sont flanquées par de l'hétérochromatine, un type de chromatine plus dense que l'euchromatine qui constitue le reste des chromosomes, au niveau des régions péricentromérique et subtélomériques. Ces zones difficiles marquent la délimitation entre les séquences spécifiques des chromosomes avec les zones riches en séquences et éléments répétés (répétitions en tandem, ADN satellites, transposons ...). Ces dernières sont d'intérêt médical, car associées à de nombreuses pathologies, mais aussi, car elles ont un rôle dans la stabilité des chromosomes. Pourtant, leur résolution moléculaire à l'échelle individuelle reste inaccessible. Leurs séquences peuvent être absentes des génomes de référence. Ainsi, les bases (A, T, G et C) sont remplacées par des longues répétitions de la lettre "N", qui signifie que la composition de la séquence est inconnue. Les régions répétées, plus ou moins résolues moléculairement constituent plus de la moitié de notre génome, voir partie sur la composition du génome.

1.1.2 L'ARN

Chez les eucaryotes (du grec *eu* "bien" et *karuon* "noyau", relatif à tous les organismes à noyau, regroupant vulgairement les animaux, les champignons, les plantes et les protozoaires), l'ADN nucléaire est stocké dans le noyau des cellules et n'en sort jamais. Pourtant, les protéines sont produites dans le cytoplasme des cellules. Il existe une étape permettant la jonction entre les deux et celle-ci se repose notamment sur la production d'un ARN messager ou ARNm.

L'ARN ou acide ribonucléique est au même titre que l'ADN un acide nucléique que l'on retrouve chez la plupart des êtres vivants, mais aussi quelques virus. Il diffère pourtant de l'ADN en plusieurs points. L'ARN est tout d'abord un acide nucléique monocaténaire formé d'un enchaînement de nucléotides. Contrairement à l'ADN, le désoxyribose est absent en la faveur d'un autre sucre, le ribose, beaucoup moins stable. De plus, l'Uracile (U) aux propriétés similaires à la Thymine (T) remplace cette dernière. Cela lui permet de se lier de la même manière à l'Adénine (A).

Dans le cadre de la production de protéines, l'ARN a pour rôle d'être le support temporaire de l'information génétique, du noyau où réside l'ADN, vers le cytoplasme, où demeure la machinerie de synthèse protéique. C'est un type d'ARN particulier qui est produit, l'ARN messager ou ARNm, lors de ce qu'on appelle la transcription, représentée *Figure 1.3*. À partir d'une portion d'un brin matrice d'ADN (contenant un ou plusieurs gènes), un complexe enzymatique, les ARN polymérases, va produire le brin complémentaire d'ARNm en substituant les thymines par des uraciles. Ce brin d'ARNm est appelé transcrit primaire. Il va ensuite être maturé, par l'ajout d'une coiffe en 5' et par l'ajout d'une queue poly(A) en 3'. Il est par la suite épissé, c'est-à-dire que les introns du transcrit vont être excisés pour notamment ne laisser que des séquences codantes et régulatrices du matériel génétique initial. Le transcrit est alors mature, il peut passer du noyau vers le cytoplasme guidé par la coiffe et la queue poly(A) où il sera traduit en protéine.

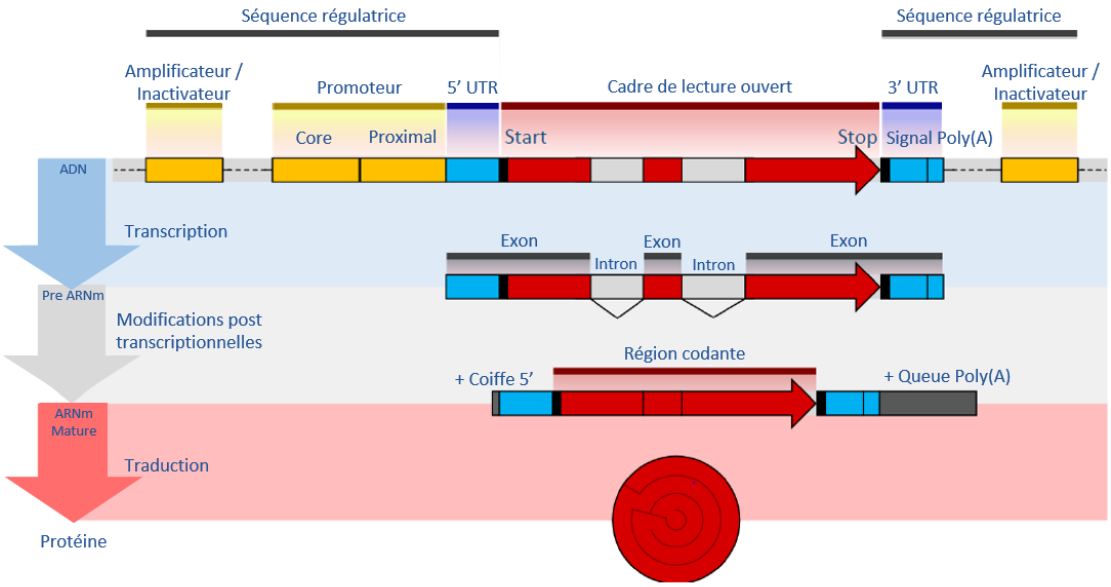


FIGURE 1.3 – Représentation schématique des étapes de la transcription d'un brin d'ADN en ARNm jusqu'à la traduction de l'ARNm en protéine.
Adapté d'après *Shafee, Thomas, and Rohan Lowe. "Eukaryotic and prokaryotic gene structure." WikiJournal of Medicine 4.1 (2017) : 2.*

Tous les transcrits produits au cours de la transcription ne sont pas traduits en protéines. La base de données GENCODE[8] version 38 recense 237 012 transcrits différents chez l'Homme dont seulement 86 757 ont une partie ou la totalité de leurs séquences qui sont ensuite traduites en protéines. Le reste des ARN et transcrits non codants (150 255) ont des formes et des utilités très diverses qui pour certaines restent encore à découvrir à ce jour. Certains ARN permettent par exemple, les mécanismes de traduction des transcrits vers la protéine, l'épissage alternatif, la maturation des ARNm ou encore la régulation de l'expression de certains gènes.

Les initiatives RefSeq[9] (NCBI) et ENSEMBL[10] (EMBL-EBI) produisent les séquences de l'ensemble des transcrits connus et notamment celles des transcrits de références (notés NM_ pour RefSeq ou ENST_ pour ENSEMBL). Les transcrits sont prédits à l'aide de pipelines d'annotations automatisés utilisant un assemblage de génome de référence comme base. Les lacunes ou erreurs de l'assemblage de référence peuvent produire des erreurs dans la prédiction des transcrits. Les transcrits de référence sont eux soumis à une curation manuelle, mais peuvent néanmoins se révéler faux. La liste des transcrits est actualisée périodiquement. L'ensemble des transcrits de référence représente la cible médicale d'intérêt lors de l'appel de variations génétiques dans le cadre du séquençage par capture d'exome.

1.1.3 Le gène

Chez les eucaryotes, le gène est l'unité fonctionnelle de l'ADN supportant l'information génétique qui va être transcrite en ARNm afin de produire une ou plusieurs protéines. Comme représenté *Figure 1.3*, un gène est constitué d'exons, qui représentent la partie codante (traduite en protéine *via* la transcription puis la traduction) de l'ADN, mais également d'introns et de régions régulatrices. Ces dernières sont constituées des promoteurs et des UTR (*UnTranslated Regions* pour régions non traduites). Les promoteurs et les UTR sont des séquences nucléotidiques particulières en 5' et 3' du ou des gènes auxquelles elles sont rattachées et contiennent des séquences régulatrices de l'expression des gènes. Ces séquences ont notamment pour rôle de recruter les complexes protéiques nécessaires à l'initiation de la transcription. Ils contrôlent le moment, le lieu et la quantité d'ARNm produit, selon le besoin.

Les introns sont des séquences non codantes qui ont pour rôle de permettre par combinatoire la sélection de la totalité ou d'une partie seulement des exons lors de l'épissage afin de produire plusieurs ARNm matures différents. Plusieurs protéines ou versions d'une protéine peuvent donc être produites à partir d'un même transcrit primaire, c'est l'épissage alternatif. Un gène est donc un ensemble de séquences codantes, non codantes et de séquences régulatrices. Une modification au sein d'un gène pourrait entraîner diverses altérations qui auraient un impact sur la structure de l'ANRm et donc sur la structure ou la fonction de la protéine finale, augmentant ainsi le risque de pathologie, voir partie sur les variations génomiques. En revanche, toutes les variations de la séquence nucléotidique d'un gène ne sont pas automatiquement pathogènes. Les gènes existent en de multiples versions au sein de la population qui n'entraînent pas pour la plupart de changement majeur dans la protéine pour laquelle ils codent (mutations silencieuses). Ces différentes versions de gènes sont appelées allèles. Chaque chromosome issu de la même paire est le support d'un allèle. Les paires d'allèles peuvent être identiques (homozygotes) ou non (hétérozygotes).

Le séquençage d'exome se limite spécifiquement aux séquences codantes uniquement, soit à l'ensemble des exons, bien que certains kits de capture puissent également comporter les UTR. En revanche, le séquençage de génome lui comprend l'entièreté des séquences génétiques de l'individu étudié.

1.1.4 Le génome

Le génome issu de l'ADN nucléaire représente l'ensemble du matériel génétique chez un organisme donné. La taille du génome varie en fonction des organismes et des espèces. Chez l'être humain, sa taille est estimée à environ 3,2 x 10^9 nucléotides (bases) distribués sur 23 paires de chromosomes. Selon les connaissances actuelles, le génome humain nucléaire comprendrait environ 20000 gènes pour produire 80000 protéines différentes[11]. Parmi l'ensemble de ces plus de trois milliards de bases qui constituent le génome humain, seul un à deux pour cent du génome total (exons) code effectivement pour des protéines. L'ensemble des zones géniques (exons, introns et séquences régulatrices appartenant aux gènes) représentent seulement un quart du génome total, voir *Figure 1.4*.

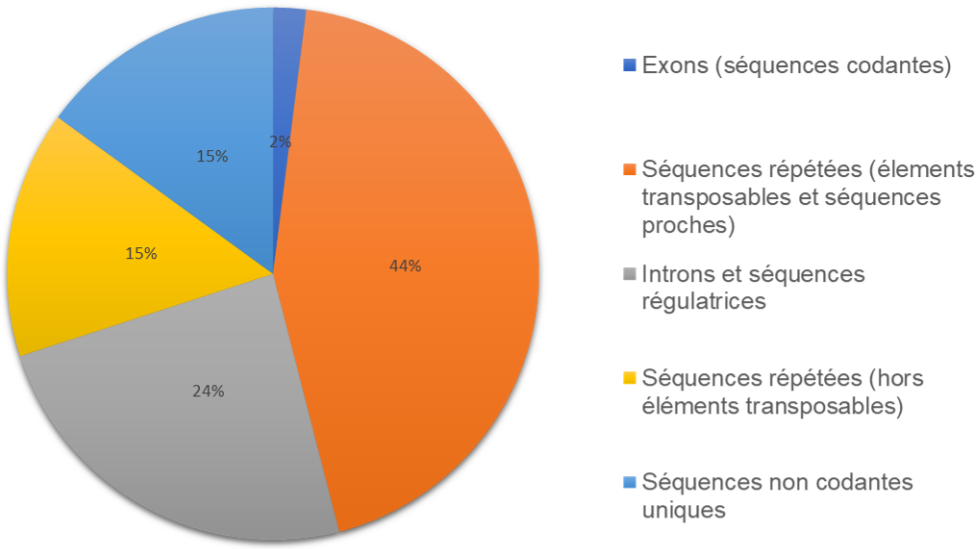


FIGURE 1.4 – Composition du génome humain.
Adapté d'après *Diagram of components of the genome as estimated in 2014, NHS National Genetics and Genomics Education Centre.*

Le reste du génome est occupé par différents types de séquences. Tout d'abord, plus de 50% du génome humain est constitué de séquences répétées. Elles sont de plusieurs types. Les plus abondantes sont les éléments transposables ou transposons, des séquences capables de se déplacer (excision puis insertion) de manière autonome dans le génome grâce à un mécanisme connu sous le nom de transposition. Leur mobilité au sein du génome leur confère une part importante dans les mécanismes évolutifs et adaptatifs du vivant. En effet, dans les cas où les mécanismes de transposition auraient lieu dans des zones codantes ou régulatrices, si l'insertion confère un caractère présentant un avantage évolutif, celle-ci sera probablement sélectionnée et perdurera dans la descendance.

Les séquences répétées non liées aux éléments transposables sont de deux types, les microsatellites qui sont des séquences relativement courtes (d'un à quatre nucléotides) dont le motif se répète jusqu'à une centaine de fois et les duplications segmentales qui sont des séquences quasi identiques de plus de 1 kb, retrouvées en plusieurs endroits, tout au long du génome.

Bien que le génome nucléaire soit supposé avoir été séquencé en entier, en réalité, n'a été séquencé que ce qui le pouvait. Les technologies de séquençage utilisées par le GRC (Genome Reference Consortium) jusqu'à la version GRCh38 ne peuvent séquencer efficacement certaines parties des chromosomes, comme

les centromères et les télomères du fait de leur nature hautement répétitive. De nouveaux projets visent à fermer certains de ces lacunes (*gaps*) à l'aide de nouvelles technologies de séquençage[12] et de produire un assemblage du génome humain véritablement complet sans *gaps* dans la séquence de référence. C'est notamment le but du consortium Telomere-to-Telomere (T2T) qui a déjà produit un assemblage complet du chromosome X[13] et qui devrait participer à la prochaine version du génome de référence. Cette version corrigée à l'aide de données issues de technologies de séquençage de troisième génération est attendue comme l'amélioration la plus importante apportée au génome humain de référence depuis sa publication initiale[14].

1.2 Génome et ADN

1.2.1 Le Projet Génome Humain

En 1990 est lancé le Projet Génome Humain (HGP pour Human Genome Project) dont le but est de séquencer pour la première fois l'entièreté du génome humain afin d'en assembler une version de haute qualité. La connaissance de l'époque se limitait à seulement quelques gènes ou portions du génome humain. Le terme séquençage désigne les différentes technologies permettant de déterminer l'enchaînement des nucléotides (A, C, T, G) d'un fragment ou d'une molécule d'ADN, voir partie sur le séquençage haut débit.

La première ébauche de génome a été publiée en 2001[15] par le HGP et a été reçue par la communauté comme une des avancées majeures de la science moderne. Une version définitive a ensuite été publiée en 2003 (NCBI Build 34 ou hg16) corrigeant nombre des erreurs et des manquements de la première version. Le projet prévu pour une durée de 15 ans et pour un montant de trois milliards de dollars en a duré seulement 13 pour un coût estimé à 2,7 milliards de dollars.

Le génome produit par le HGP marque le véritable début de la génomique. Il a permis de poser les prémices de la cartographie du génome humain dans son ensemble et a été à l'origine du développement de nouvelles technologies à visée pan-génomique (développement de sondes pour puces à ADN, séquençage d'exomes ...).

Grâce aux progrès technologiques, avec notamment la diversification des technologies et la baisse des prix de séquençage, de nombreux projets de séquençage massifs ont vu le jour avec comme objectif l'annotation fonctionnelle du génome produit par le HGP, comme le projet ENCODE[16] ou le 1000 Genomes Project (1KGP)[17].

Le génome produit par le HGP a également pu servir par la suite de guide ou de référence afin de faciliter l'assemblage des génomes de références suivants plus complets ou corrigés par d'autres consortiums, notamment le consortium de référence sur le génome humain.

Les diverses estimations quant à la taille ou les éléments qui composent le génome nous proviennent de projets qui ont eu pour objet la production de génomes de référence humains. Un génome de référence ou assemblage de référence est une séquence nucléique censée représenter l'ensemble des nucléotides d'un individu chimérique idéalisé qui n'existe pas. La séquence est produite à partir des données de plusieurs individus de l'organisme donné et a pour but de servir de canevas lors de la recherche de variations chez un individu à partir de la référence.

1.2.2 Le consortium de référence sur le génome humain

Le consortium de référence sur le génome humain (GRC pour Genome Reference Consortium), fondé en 2007, est un consortium international constitué d'une vingtaine d'experts en séquençage, assemblage de génome, ainsi qu'en bioinformatique, ayant pour but de produire une séquence de référence du génome humain au plus proche de la réalité[18].

Plusieurs versions du génome de référence ont été publiées par le GRC depuis la publication initiale du HGP. Aujourd'hui encore, deux versions coexistent dans les laboratoires, GRCh37 (hg19) ou GRCh38 (hg38) publiées respectivement en 2009 et 2013. Chacune de ces versions à reçu des correctifs ou mises à jour mineures corrigeant ou améliorant en partie leur séquence, GRCh37.p13 publié en 2013 et GRCh38.p13 publié en 2019. Chaque nouvelle séquence de référence est supposée plus complète et plus proche de la réalité biologique que la précédente.

Si GRCh37 n'a pas été complètement abandonné, c'est d'une part, par habitude des laboratoires de travailler avec cette version du génome de référence. Beaucoup de leurs outils sont seulement disponibles dans cette version du génome. La mettre à jour nécessiterait de, soit payer une mise à jour dans le cas d'outils commerciaux, soit des efforts de recherche et développement parfois longs et coûteux dans le cas d'outils développés en interne. Pendant longtemps, le frein majeur qui a empêché la mise à jour de la version du génome de référence dans certains processus était l'absence pour GRCh38 de certaines bases de métadonnées d'annotation disponibles pour GRCh37. Les initiatives les maintenant étant pour la plupart communautaires, il y a un décalage logique entre la publication d'une nouvelle référence et celles des bases de données qui lui sont associées. Aujourd'hui, la totalité des bases de données nécessaires pour le diagnostic génomique est disponible sous GRCh38 et a même pour la plupart été enrichie par rapport à la version précédente de l'assemblage de référence. Pourtant, il faut encore souvent composer avec ces deux versions de la référence humaine au jour le jour en tant que bioinformaticien clinique.

Le fait que l'assemblage GRCh38 soit plus complet n'est pas dû qu'à sa date de publication plus récente[19], voir *Table 1.1* . La N50, définie comme la longueur de la séquence du *contig* (séquence génomique continue et ordonnée générée par l'assemblage), le plus petit à 50% de la longueur totale du génome, est une des métriques les plus importantes relatives à la qualité d'un assemblage. Cela traduit le fait que le génome de référence version 38 comporte moins de *gaps* que la version précédente et que les *contigs* de cette référence soient plus longs.

	GRCh37.p13	GRCh38.p13
Nb de base total	3,23 Gb	3,27 Gb
Nb de base total (sans N)	2,99 Gb	3,11 Gb
N50	46 Mb	67 Mb
Loci alternatifs	9	261

TABLE 1.1 – Caractéristiques des assemblages de référence GRCh37.p13 et GRCh38.p13.
Adapté d'après *https://www.ncbi.nlm.nih.gov/grc/human/data.*

De plus, bien que ce ne soit pas une spécificité du génome de référence GRCh38, celui-ci possède des *contigs* alternatifs en un nombre supérieur aux assemblages de référence précédents voir *Table 1.2*. Ceux-ci représentent des portions de l'assemblage hautement variable dans la population (notamment détectées grâce au 1KGP) sur lesquels sont présents des gènes en différentes versions (*loci*). Aligner et faire l'appel de variants de ces régions permet de réduire le nombre de variants faux positifs détectés au sein de ces gènes[19].

Type	**Nom**	**Nombre**	**Taille (bp)**
Chromosomes canoniques	chr1-22, chrX, chrY, chrM	25	3 088 286 401
Patches	Variables	187	62 830 845
Contigs alternatifs	Variables (*_alt)	261	109 535 387
Non mappés	Variables (*_random)	42	6 978 808
Non placés	Variables (chrUn_*)	127	4 485 509
Séquence indéterminée	N	*	161 368 351

TABLE 1.2 – Description de la constitution de l'assemblage de référence du génome GRCh38.p13 par GENCODE 32.
Numéro d'accession GCA_000001405.28. Adapté d'après *https://www.gencodegenes.org/.*

Enfin, il est conseillé par de nombreux acteurs, dont les équipes du *Broad Institute*, institut de référence mondiale du domaine de la génomique, de rajouter des séquences qui ne sont pas humaines dans la séquence de référence utilisée pour conduire les analyses bioinformatiques. Celles-ci, appelées leurres (*decoy*), ont pour rôle de capter les lectures qui n'ont aucun intérêt dans le cadre diagnostic lors de l'étape d'alignement contre une référence. La séquence la plus plébiscitée est celle du génome du virus d'Epstein-Barr (EBV) présent dans les lignées de cellules immortalisées, mais aussi dans 90% des cellules de la population mondiale.

La plupart des analyses se concentrent sur l'étude du génome nucléaire, il est pourtant également possible d'étudier l'ADN mitochondrial humain qui répond à différentes règles par rapport au génome nucléaire. Tout d'abord, il est issu des mitochondries, organites cytoplasmiques présents dans la plupart des cellules eucaryotes et intervenant dans la respiration cellulaire, processus catabolique permettant les transferts d'énergie cellulaire. Le génome mitochondrial est uniquement hérité de la mère et non des deux parents et est hétéroplasmique, c'est-à-dire que cohabite plus d'une version du génome au sein d'un même individu. Le génome mitochondrial est étudié par un consortium qui lui est dédié, le MSeqDR[20] responsable des bases de données curées par les experts MITOMAP[21] et HmtDB22[22].

1.2.3 Spectre de variation du génome humain

1.2.3.1 Les variations mononucléotidiques

Les polymorphismes d'un seul nucléotide (SNP pour *Single-Nucleotide Polymorphism*) peuvent être des substitutions, insertions ou délétions par rapport à une séquence de référence. Un SNP est par définition fréquent dans la population et n'induit pas forcément de pathologies, ce sont d'abord les principaux acteurs de la diversité génétique. Ne sont considérées comme SNP que les variations présentes à plus de 5% dans la population. En dessous de ce seuil, ces variations sont caractérisées de variants d'un seul nucléotide (*Single-Nucleotide* Variant, SNV). Les variations d'un seul nucléotide (mononucléotidiques) sont les variations les plus fréquentes au sein du génome humain, on en compte environ 4 à 5 millions par individu, soit une variation mononucléotidique toutes les 1000 paires de bases environ. La plupart de ces variations mononucléotidiques sont des variations héritées des parents, voir *Table 1.4*.

Les variations *de novo* sont plus rares que les variations héritées, leur fréquence est estimée à 2 x 10^{-8} par paire de bases, par génération[23]. Lorsqu'une variation est présente dans l'ensemble des cellules d'un individu, mais que cette variation n'était pas présente chez ses parents, alors, elle est dite *de novo* (DNM). Les DNM apparaissent en général dans l'un des gamètes des deux parents ou dans le zygote (cellule fécondée), selon différents mécanismes propres aux différents types de variations pouvant apparaître.

Les variations ponctuelles sont extrêmement divers parmi la population générale, 675 millions de variations distinctes sont décomptées dans la version 153 (août 2019) de la base de données dbSNP[24], censée contenir l'ensemble des SNP détectés chez l'Homme à ce jour.

Les SNP ou SNV peuvent avoir différentes conséquences selon s'ils prennent place dans une région codante, intronique, intergénique ou régulatrice et selon leur impact, variation silencieuse (pas de changement d'acide aminé dans la protéine), faux-sens (changement de l'acide aminé) ou non-sens (apparition d'un codon-stop et production d'un ARNm tronqué). La plupart des variations mononucléotidiques et des petites insertions et délétions sont des polymorphismes, car par définition fréquents dans la population. Plus de 90% de ces variations sont présentes à plus de 5% dans la population générale et 5% des variations ont une fréquence comprise entre 1 et 5%, voir *Table 1.3*.

Sur le séquençage de génome de cet individu ont été retrouvées un total de **5195673 variations**. Les filtres suivants ont été appliqués :

Filtres appliqués	Compte
Variations à plus de 5% dans GnomAD *	4297261
Variations entre 1% et 5% dans GnomAD *	216317
Variations à plus de 10% dans la cohorte *	505174
Variations intergéniques profondes (à plus de 2kb d'un gène) *	69835
Variations avec une lecture alternative	7596
Variations avec une balance allélique <10%	13036
Total des variations non considérées	**5109219**

* Les variations pathogènes ou probablement pathogènes telles que référencées dans ClinVar sont conservées sur ces étapes.

TABLE 1.3 – Exemples de filtres de fréquence appliqués à la détection de variations de petite taille sur des données de séquençage WGS
Données internes.

Classe de variation	Sous-classe / autre terme	Taille	Nombre par WGS Illumina	Nombre *de novo* estimé	Puce à ADN	WGS short-read	WGS long-read
SNV	**Mutation ponctuelle, substitution**	**1 pb**	**$3,5 \times 10^6$**	**44 - 82**	**XX**	**XXX**	**XX**
Indel	**Insertion, délétion, indel complexes**	**1-49 pb**	**$4,5 \times 10^5$**	**2,9 - 9**	**XX**	**XX**	**XXX**
SV		**> 50 pb**		**0,19**			
CNV	Délétion		5 000		X	XX	XXX
	Duplication		1 000		X	XX	XXX
	Multiplication (3+ copies)		450		X	XX	XX
Insertion	Nouvelle, modèle ou répétition	> 50pb	1500		-	X	XXX
Réarrangements équilibrés	Inversion	> 50pb	40		-	XX	XXX
	Translocation réciproque	Interchromosomal	0,001		-	XX	XXX
Réarrangements génomiques complexes	SV complexes, chromothripsis	> 1 mb	0,01		-	XX	XXX
CNV extrêmement large	Aneuploïdie, anomalies chromosomiques	> 1 mb	0,01		XXX	XXX	XXX
Insertion de rétrogène	Rétroduplication, rétrocopie	Partie codantes d'un gène	10		-	XX	XXX
Éléments transposables	SINE, LINE, SVA	0,3 – 7 kb	2 000		-	X	XXX
Variations de répétitions en tandem				**Inconnu**			
Répétitions en tandem courtes (STR)	Microsatellites, répétitions de séquences simples	Unité de répétition de 1 à 6 pb	1×10^5		-	X	XXX
Répétitions en tandem de nombre variable (VNTR)	Minisatellites	Unité de répétition de 7 à 49 pb	Inconnu		-	X	XX
Répétitions centromériques et hétérochromatiques	ADN satellite (α, β, 1-3)	Variable	Inconnu		-	-	X

TABLE 1.4 – Caractéristiques, nombre moyen et potentiel de détection en fonction de la technologie utilisée des différentes classes de variations génétiques chez un individu.
Adapté d'après *Lappalainen, T., Scott, A. J., Brandt, M., Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. Cell, 177(1), 70-84. et Eichler, E. E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. New England Journal of Medicine, 381(1), 64-74. et Acuna-Hidalgo, R., Veltman, J. A., Hoischen, A. (2016). New insights into the generation and role of de novo variations in health and disease. Genome biology, 17(1), 1-19.*

1.2.3.2 Les indels

Les indels sont des insertions ou des délétions d'une taille comprise entre 2 et 50pb. Les indels peuvent respecter la phase du cadre de lecture ou non. En effet, celles qui ne sont pas d'une taille multiple de trois induisent une variation avec décalage du cadre de lecture du transcrit, aboutissant le plus souvent à la production d'une protéine tronquée et dégradée par la machinerie cellulaire. Bien qu'en nombre moins important que les SNP, les indels affectent environ la même proportion génomique et participent également à la diversité génétique, voir *Table 1.4*.

1.2.3.3 Les SV

Les variations de structure ou variants structuraux (SV pour *Structural Variation*), sont des variations génétiques de grandes tailles[25] (supérieures à 50pb) résultantes de réarrangements ou de cassures chromosomiques suivies d'une étape de réparation des extrémités cassées entraînant un nouvel arrangement des gènes emportés par le réarrangement chromosomique[26].

Ces anomalies peuvent être équilibrées, c'est-à-dire qu'il n'y a ni addition, ni perte de matériel génétique. L'individu touché est dans la plupart du temps de phénotype sain ou "porteur sain". Une pathologie peut être associée si le point de cassure impact un gène ou une région génomique fonctionnelle. La variation présente un risque pour la descendance dans le cas où le SV représenterait un point de cassure favorable à la survenue d'un déséquilibre chromosomique. Dans la population, la majorité des cas de phénotypes anormaux sont issus de réarrangements *de novo* rares dans la population générale. Les différents types de SV connus aujourd'hui sont répertoriés dans la base de données DbVar[27].

Les anomalies équilibrées regroupent plusieurs types d'altérations. Tout d'abord les inversions de matériel génétique, c'est-à-dire qu'une séquence nucléotidique continue est inversée dans la même position. Les translocations intrachromosomiques, lorsque les régions concernées proviennent du même chromosome et interchromosomiques, lorsque les régions concernées proviennent de chromosomes différents.

À l'inverse les anomalies peuvent être déséquilibrées, c'est-à-dire qu'il y a perte, ou gain de matériel génétique. Les variants structuraux déséquilibrés regroupent, les insertions de nouvelles séquences qui n'ont aucune correspondance avec le génome de référence ainsi que les insertions ou délétions d'éléments mobiles grâce à leur propriété de transposition. Elles rassemblent également les duplications en tandem. Une duplication en tandem est constituée de 2 régions adjacentes identiques, qui peuvent être répétées un nombre variable de fois. Une répétition de séquence en tandem est étendue ou contractée par rapport à une référence. Enfin, parmi les anomalies déséquilibrées, on retrouve les duplications segmentales, des séquences de plus de 1 kb présentes plus d'une fois dans le génome, dont les copies sont identiques à plus de 90%.

Les variations de nombre de copies (CNV pour *Copy Number Variants*) représentent une sous-classe d'anomalies de structure, plus fréquente que les précédentes et sont détaillées ci-dessous.

Le domaine d'étude des variants structuraux est un espace de recherche extrêmement actif. L'accès plus large au séquençage de génome entier et aux technologies de séquençage de troisième génération permet la classification de variants qui étaient difficiles à étudier auparavant, voir *Figure 1.5*. La méthodologie de classification de l'ensemble des types de variants de structure est disponible sur le site de la base de données DbVar (https://www.ncbi.nlm.nih.gov/dbvar/content/overview/).

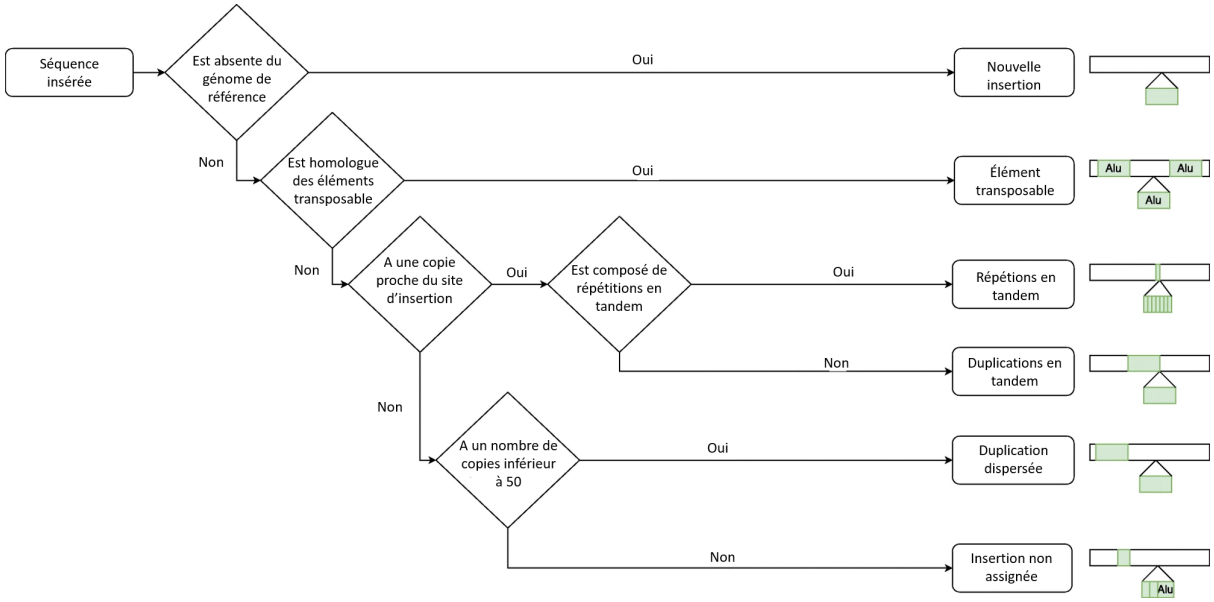


FIGURE 1.5 – Méthodologie de classification des insertions en fonction de la nature de leur séquence. Adapté d'après *Delage, W. J., Thevenon, J., Lemaitre, C. (2020). Towards a better understanding of the low recall of insertion variants with short-read based variant callers. BMC genomics, 21(1), 1-17.*

1.2.3.4 Les CNV

Les CNV sont des variants structuraux déséquilibrés impliquant que le nombre de copies d'un ensemble de gènes particulier varie d'un individu à l'autre par rapport à une référence, voir *Figure 1.6* . Ils sont classés en deux catégories, les gains (duplications) et les pertes (délétions), de matériel génétique. Les CNV sont de tailles diverses, bien qu'ils soient définis comme des variations d'une taille minimale de 1 kb, la limite de détection minimale opérationnelle est de l'ordre de 50 pb[23]. Dans les cas les plus extrêmes, les CNV peuvent être des aberrations chromosomiques, c'est-à-dire, la délétion ou la duplication de chromosomes entiers aboutissant à un nombre incorrect de chromosomes (aneuploïdies).

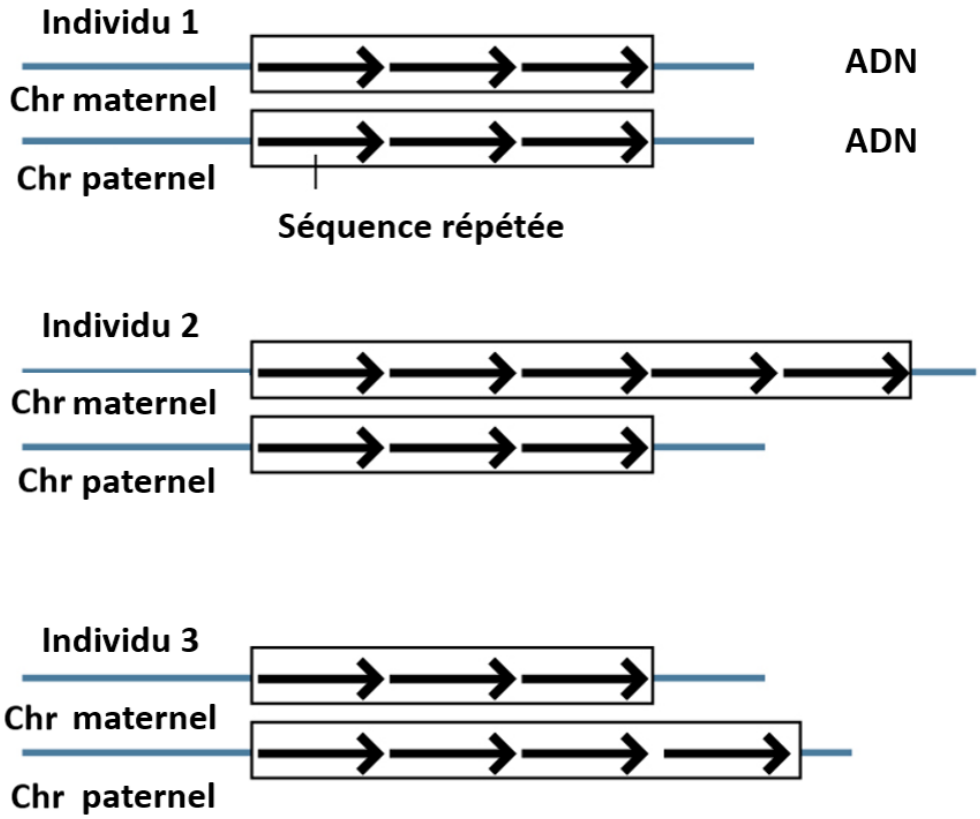


FIGURE 1.6 – Illustration de variations de nombre de copies.
Adapté d'après https://www.genome.gov/genetics-glossary/Copy-Number-Variation.

Les anomalies de grandes tailles telles que les aneuploïdies et certains CNV (> 100kb) sont rares, environ une naissance pour 42 présente un CNV de grande taille[23]. Néanmoins, du fait de leur taille, ils affectent en moyenne une part plus importante du génome d'un individu que les SNV et les indels combinés, voir *Figure 1.7*. La relation inverse entre la taille et la fréquence des variations s'explique par l'impact sur le pronostic d'un individu que de telles anomalies peuvent provoquer. De plus, plus la taille du CNV augmente, plus la probabilité qu'il n'ait pas été hérité, mais qu'il soit *de novo* augmente[23].

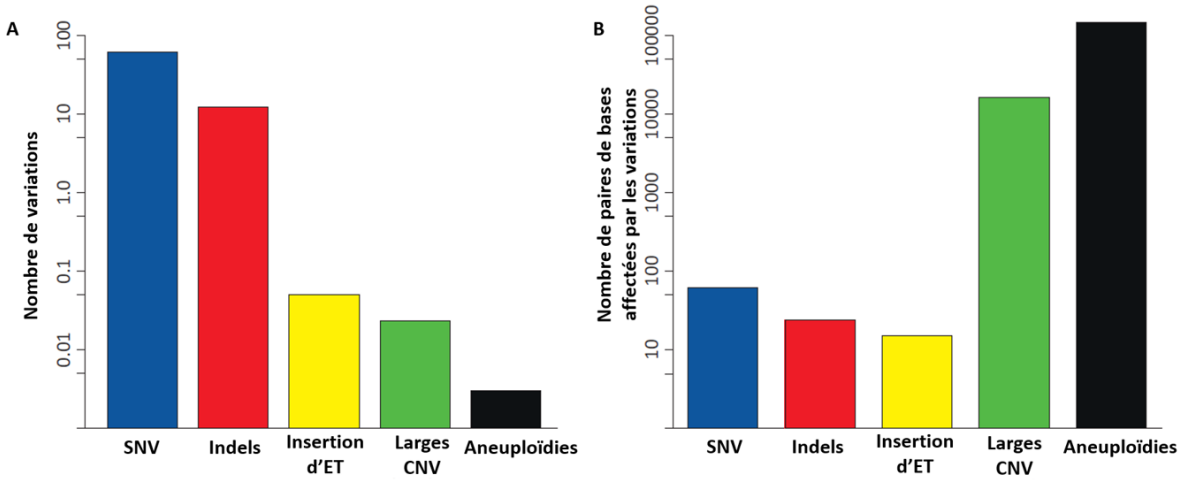


FIGURE 1.7 – Comparaison de la fréquence et de l'ampleur des différentes formes de variation génétique.
A : Nombre moyen de variations de chaque type par individu. B : Nombre moyen de bases affecté par chaque type de variation en moyenne chez un individu. ET : éléments transposables.
Adapté d'après *Campbell, C. D., Eichler, E. E. (2013). Properties and rates of germline mutations in humans. Trends in Genetics, 29(10), 575-584.*

Au même titre que les autres types de SV, les CNV ne sont pas distribués de manière uniforme sur le génome. Certaines zones, comme les régions subtélomériques et péricentriques sont enrichies en CNV. De plus, certaines régions favorisent l'apparition de CNV *de novo*, notamment les régions flanquées par des duplications segmentales[28].

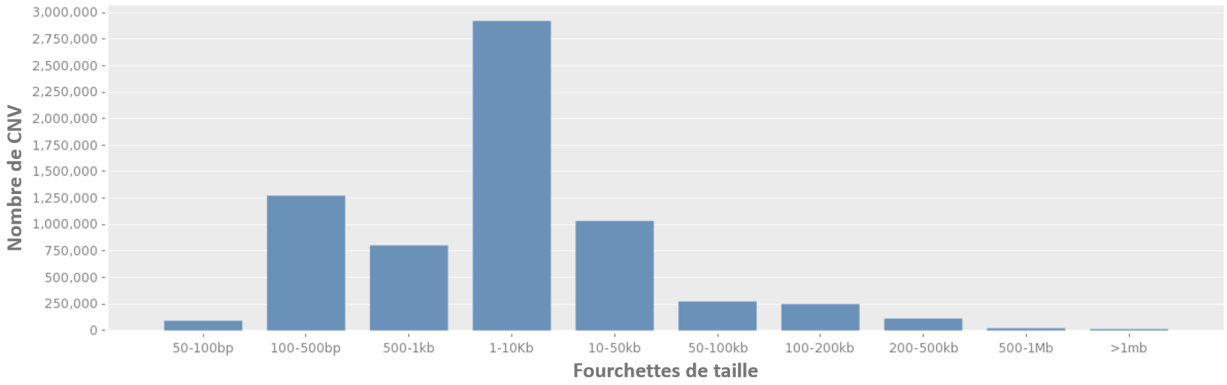


FIGURE 1.8 – Distribution de taille des CNV de la base de données DGV version 106.
Adapté d'après *http://dgv.tcag.ca/v106/app/statistics.*

Même si les CNV de petite taille ($<$ 100 kb) n'entrainent pas pour la plupart de pathologies, ils n'en restent pas moins d'intérêt clinique. Bien qu'ils soient plus difficiles à détecter que leurs homologues de grande taille, car uniquement détectables de manière globale depuis l'arrivée des technologies de Séquençage à Haut Débit (HTS pour *High Throughput Sequencing*), ils représentent la majorité des CNV détectées, voir *Figure 1.8*. La base de données DGV[29] (*Database of Genomic Variants*) regroupe à la manière de dbSNP[24] les variations de plus de 50 paires de base détectées. Les CNV compris entre 1 et 10 kb y sont les plus représentés. Ces données peuvent représenter la réalité mais aussi les biais et limitations technologiques.

En effet, des biais dans leur détection existent, les CNV les plus larges sont bien évidemment les plus simples à détecter, mais également, les technologies HTS détectent plus de délétions que d'insertions à l'inverse des technologies de cartographie génétiques. En moyenne sont détectées 2 délétions pour une duplication avec les technologies HTS[28].

1.3 Détecter les variations du génome humain

Bien avant le séquençage d'ADN à haut débit pour le diagnostic des maladies rares, le diagnostic d'un patient reposait sur une approche clinique combinée avec des examens de génétique chromosomique, puis moléculaire. Cela englobe les techniques de descriptions et classifications phénotypiques de syndromes déjà connus, les analyses de constantes biologiques, les analyses de biologie moléculaire ciblées (type MLPA) et de cytogénétique (type FISH ou caryotype) ainsi que le séquençage Sanger d'un gène candidat. L'association de ces méthodes a un taux de diagnostic d'environ 30%[30].

Les innovations technologiques majeures qui ont permis de faire transitionner la génétique clinique vers la génomique clinique sont les technologies d'Analyse Chromosomique sur Puce à ADN (ACPA) apparues dans les années 90 et les technologies de séquençage d'ADN à haut débit par synthèse apparues à partir de l'année 2005. Ces deux technologies complémentaires restent actuellement les piliers de la médecine génomique de routine dans le cadre du diagnostic de maladie rare[2][31]. Une sélection de différentes méthodes pour la détection de variations est représentée *Figure 1.9*.

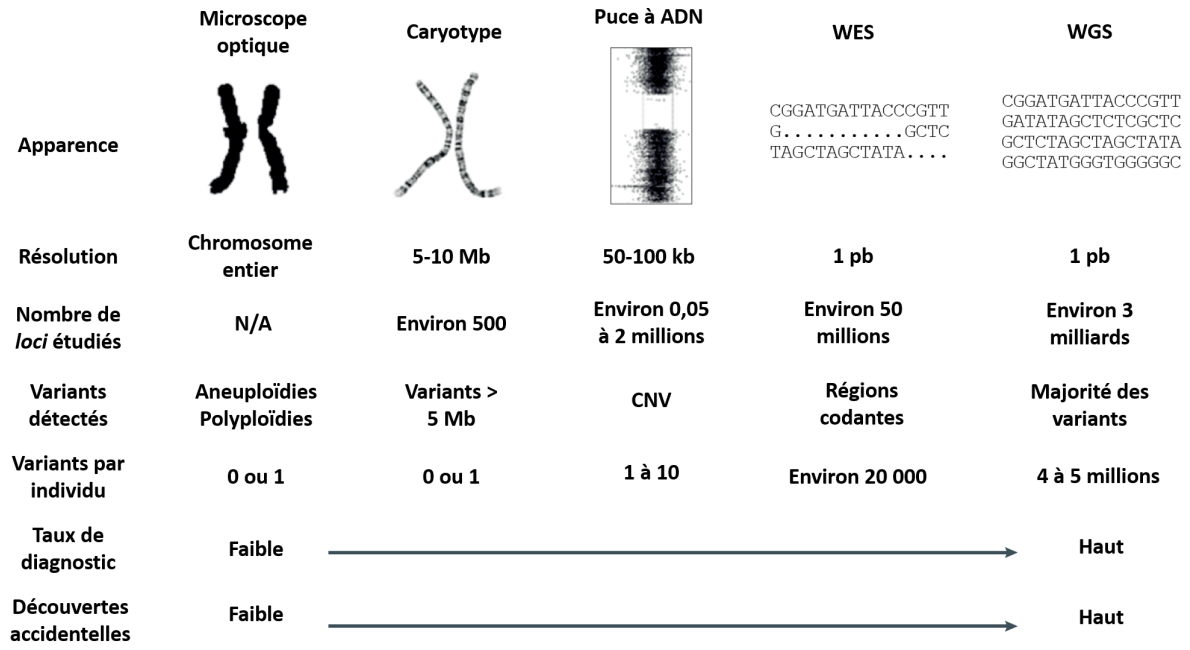


FIGURE 1.9 – Stratégies de test et hétérogénéité clinique pour la détection de variations génomiques. Adapté d'après *Wright, C. F., FitzPatrick, D. R., Firth, H. V. (2018). Paediatric genomics : diagnosing rare disease in children. Nature Reviews Genetics, 19(5), 253-268.*

1.3.1 Technologies de cartographie génétique

1.3.1.1 Analyse Chromosomique sur Puce à ADN

À partir des années 2000, la démocratisation des technologies d'Analyse Chromosomique sur Puce à ADN (ACPA) a permis de rechercher des anomalies chromosomiques déséquilibrées (notamment les CNV) à une résolution plus faible que celle du caryotype (100 kb au lieu de 5 mb) de manière pan-génomique (sur l'ensemble du génome)[32][33]. Elles sont des deux types principaux, les puces d'hybridation génomique comparative[34] (*CGH array* ou a-CGH pour *array Comparative Genomic Hybridization*) et les puces à SNP[35] (*SNP array*). Ces deux technologies sont supposées équivalentes à conception équivalente en ce qui concerne les performances et la fiabilité de détection de CNV même si leur principe n'est pas exactement le même. C'est pourquoi elles sont utilisées comme technologies de référence pour la détection de CNV par la quasi-totalité des laboratoires.

Une puce à ADN comprend un grand nombre de fragments d'ADN déposés sur une surface solide telle que du verre, du silicium ou du plastique[36]. Ces fragments d'ADN spécifiques d'une taille de 25 à 80 paires de bases sont appelés sondes. Le design des sondes sur la puce est à la discrétion du fabricant, c'est-à-dire que les sondes sont choisies pour recouvrir des régions stratégiques dans le cadre du diagnostic, notamment les régions codantes, mais pas seulement. Le nombre varie en fonction du type de puce. Plus il y a de sondes, plus la puce est résolutive et plus elle est chère. Les sondes sont déposées sur la puce sous la forme de taches (*spots*) sur lesquelles sera mesurée l'hybridation comparative. Ces deux technologies sont automatisables et permettent notamment l'étude des CNV à haut débit.

1.3.1.2 La puce d'hybridation génomique comparative

Le principe de la *CGH array* est basé sur celui de l'hybridation comparative de l'ADN. Une quantité d'ADN du patient à tester est marquée à l'aide d'un fluorochrome vert et la même quantité d'ADN de référence est marquée à l'aide d'un fluorochrome rouge. Le tout est mélangé puis déposé sous la forme de spots déposés de manière ordonnée sur la puce, voir *Figure 1.10*.

Après hybridation comparative des ADN, les résultats sont lus à l'aide d'un scanner, en général vendu par le fournisseur de la puce. Il a pour fonction d'enregistrer et convertir les signaux lumineux en valeur d'intensité de fluorescence pour l'ensemble des sondes de la puce. Ces valeurs sont ensuite analysées par un logiciel propriétaire, en général, basé sur un modèle de Markov caché (HMM pour *Hidden Markov Model*). Le logiciel détermine le différentiel de nombre de copies (gains ou pertes) des différents spots en fonction de la valeur d'intensité de fluorescence de l'ADN contrôle et de la valeur des sondes environnantes.

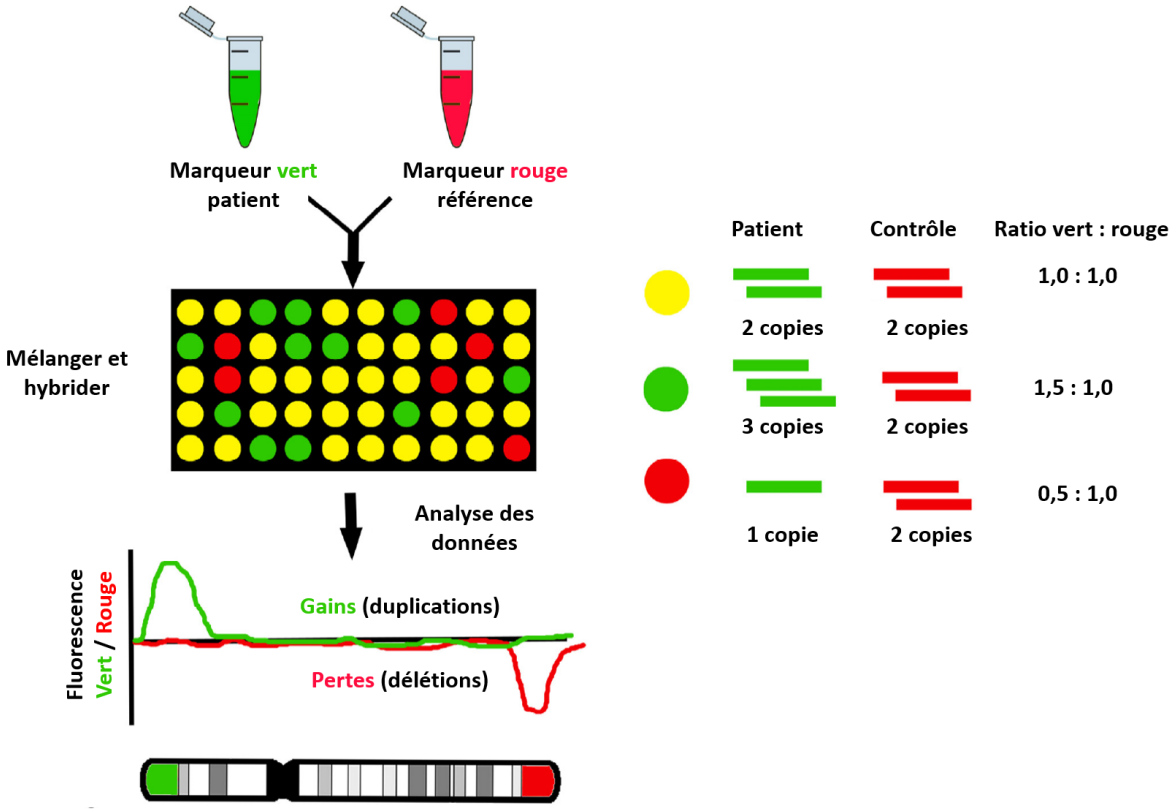


FIGURE 1.10 – Principe des puces d'hybridation génomique comparative ou *CGH-array*.
Adapté d'après *Karampetsou, E., Morrogh, D., Chitty, L. (2014). Microarray Technology for the Diagnosis of fetal chromosomal aberrations : which platform should we use ?. Journal of clinical medicine, 3(2), 663-678.*

L'intensité de fluorescence des différents ADN est généralement représentée de manière longitudinale le long des chromosomes sous la forme d'un graphique d'intensité en logarithme base 2, appelé Log Ratio, 0 représentant l'égalité parfaite entre l'ADN patient et contrôle, voir *Figure 1.10*. Une perte est en général suspectée en dessous d'un ratio de 0,80 et un gain suspecté au-delà de 1,25. Certains logiciels peuvent identifier certaines zones conflictuelles qui nécessitent une intervention humaine pour valider la présence ou non d'un CNV.

1.3.1.3 La puce à SNP

La *SNP-array* contrairement à la *CGH-array* repose sur une hybridation des fragments à une sonde complémentaire sans compétition, voir *Figure 1.11* . Grâce à la carte des SNP dont l'établissement a été initié par le HGP, les constructeurs peuvent concevoir des sondes contenant les 2 allèles du SNP pour un *locus* donné. L'ADN est hybridé et comme précédemment, grâce à un scanner, les intensités de fluorescence vont permettre de génotyper (déterminer les allèles) de l'individu pour un ensemble déterminé de *loci*.

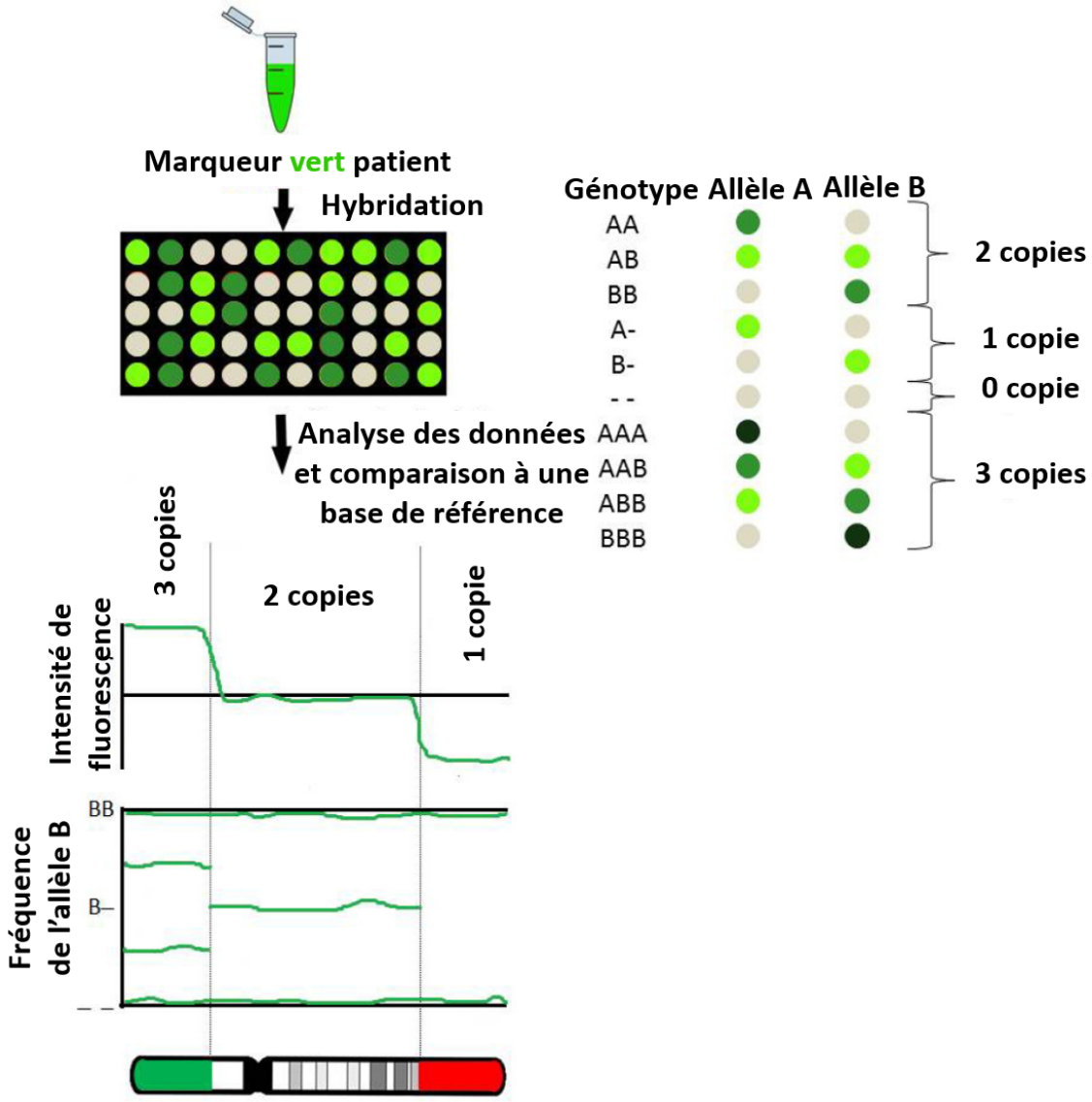


FIGURE 1.11 – Principe des *SNP-array*.
Adapté d'après*Karampetsou, E., Morrogh, D., Chitty, L. (2014). Microarray Technology for the Diagnosis of fetal chromosomal aberrations : which platform should we use ?. Journal of clinical medicine, 3(2), 663-678.*

Ces puces ont initialement été développées à des fins d'haplotypage, c'est-à-dire la détermination d'haplotypes, des groupes d'allèles de différents *loci* situés sur un même chromosome et transmis ensemble. Toutefois, elles peuvent être utilisées pour déterminer le nombre de copies des différents *loci* étudiés et ainsi détecter les CNV, voir *Figure 1.12*. Elles permettent également la détection de perte d'hétérozygotie (LOH pour *Loss Of Heterozygosity*) et notamment les cas de disomie uniparentale avec l'étude des parents pour en déterminer l'origine.

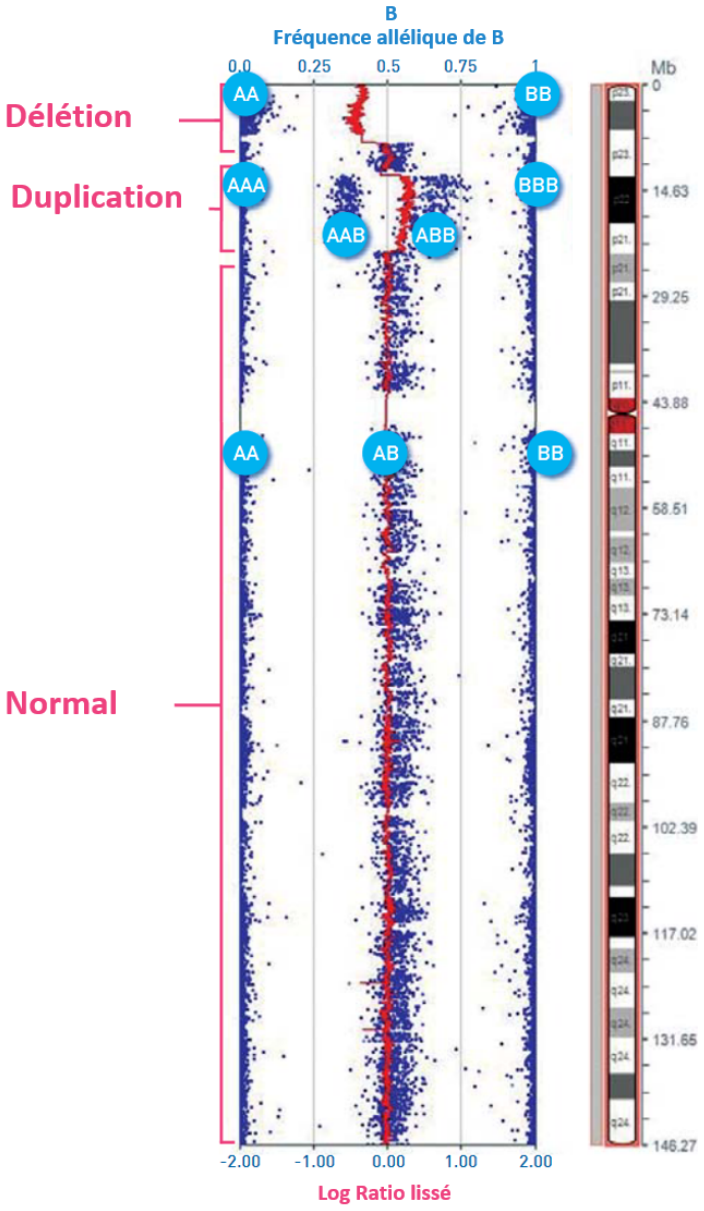


FIGURE 1.12 – Représentation des résultats d'une puce *HumanCytoSNP-12 BeadChip* (Illumina). Présentation de modifications génomiques complexes (une duplication et une délétion) pour un allèle A et un allèle B, sur le chromosome 8p. [arr8p23.3.p23.1(221411-6914226)x1,8p23.1.p21.3(12583059-22995348)x3]. Adapté d'après *Genome-Wide analysis by SNP-array*.

1.3.1.4 Bionano Genomics

La technologie développée par Bionano Genomics est une technologie de cartographie optique du génome qualifiée de *Whole Genome Imaging*, permettant la production de cartographie génétique ou *Whole Genome Map*[37]. Le principe de la technologie Saphyr vendu par Bionano Genomics repose sur la représentation graphique (par fluorescence) des sites de restriction d'enzymes au niveau du génome, permettant l'élaboration d'une carte de sites de restriction de manière semi-automatique. La comparaison de cette carte à d'autres cartes de référence ou d'échantillons proches permet la détection précise de variants de structure chromosomique.

La première difficulté dans l'utilisation de cette technologie réside dans le fait qu'elle nécessite une grande quantité d'ADN de très haut poids moléculaire (200 à 400 kb) qui n'est pas toujours disponible, voir *Figure 1.13*. Cet ADN est ensuite digéré par des enzymes de restriction modifiées qui créent une coupure simple brin au niveau d'un site de reconnaissance spécifique afin d'exciser un nucléotide et de le remplacer par un nucléotide analogue marqué par fluorescence. De nouvelles chimies permettent même l'utilisation d'enzymes transférant seulement le marqueur fluorescent sans la modification du nucléotide en une seule étape[38]. L'ADN marqué est alors transféré dans des puces propriétaires, jouant le rôle de *flowcells*.

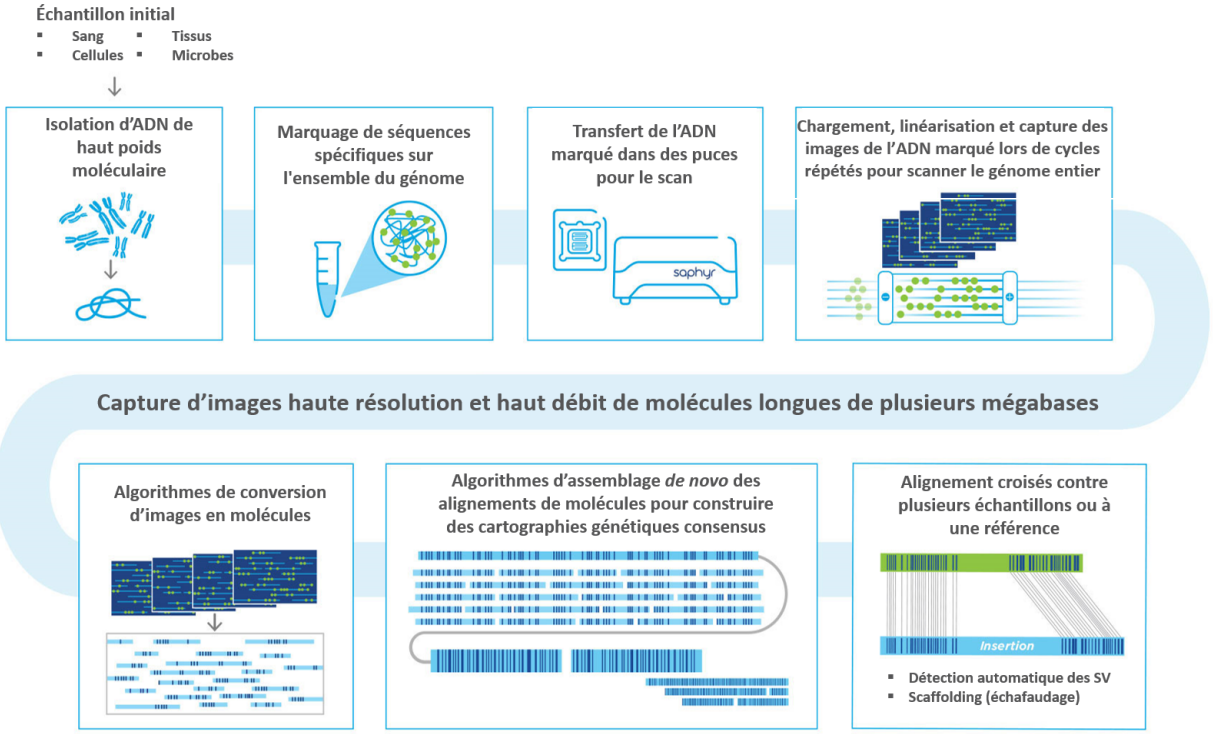


FIGURE 1.13 – Représentation schématique du *workflow* Saphyr de la technologie Bionano Genomics. Adapté d'après *Bionano Platform Technology*.

L'ADN ainsi transféré va ensuite être forcé à migrer vers des centaines de milliers de nanocanaux de 100 à 200 nm de diamètre. L'ADN en solution libre adopte une position tridimensionnelle dite de pelote aléatoire, voir *Figure 1.14*. L'ADN est par la suite démêlé dans la puce par l'action de structures solides, allant de l'échelle microscopique à l'échelle nanoscopique, jusqu'au sein des nanocanaux. Les molécules y sont alors linéarisées.

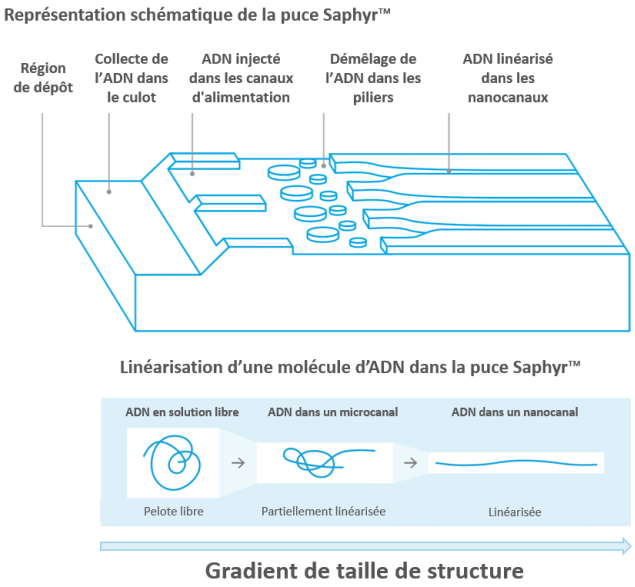


FIGURE 1.14 – Représentation schématique d'une puce Saphyr et de son rôle dans la linéarisation des molécules d'ADN.
Adapté d'après *https://bionanogenomics.com/ technology/platform -technology/.*

Une fois dans les nanocanaux, les molécules sont immobilisées et le scanner capture des images haute définition des molécules marquées. Les images des molécules dépassant la résolution du capteur (250 kb de long) sont assemblées afin de bénéficier du maximum d'information disponible par molécule. Une fois les molécules complètement capturées, les nanocanaux sont lavés et le processus se répète. Plus de 25 gigabases d'ADN par *flowcell* peuvent être capturées par heure.

Une fois l'ensemble des données brutes d'imageries capturées, les données sont converties en profils de restriction numériques par molécule, voir *Figure 1.13*. Un logiciel propriétaire permet l'analyse des données et leur assemblage *de novo* en une carte génétique. Cette carte peut être utilisée pour plusieurs types d'analyses dont la détection de variants structurants par rapport à une carte de référence.

La technologie de Bionano Genomics complète le marché des technologies de séquençage haut débit. Certaines de ses applications, notamment l'échafaudage (*scaffolding*) hybride afin d'obtenir un assemblage de génome, nécessitent en plus des données de cartographie optique des données de séquençage haut débit. Il est possible que cette technologie représente le futur des technologies pour la détection des variations génomiques de grande taille. Pour le moment, elle est assez peu répandue (seulement deux plateformes en France la proposent) et toujours assez onéreuse, à la manière de certaines technologies de séquençage de troisième génération.

1.3.2 Obtention et qualification des acides nucléiques

Préalablement aux étapes de préparation de la librairie de séquençage, il est nécessaire d'obtenir du matériel génétique de bonne qualité. L'ADN à séquencer est extrait d'un tissu biologique. La plupart du temps, dans le cadre de séquençages d'ADN pour le diagnostic de maladies rares, l'ADN est issu de sang recueilli sur EDTA (acide éthylènediaminetétraacétique, un anticoagulant permettant la préservation des cellules contenant l'ADN). Cette étape est appelée extraction de l'ADN. Les procédés d'extraction d'ADN, bien que multiples, manuels ou automatisés, reposent néanmoins sur des étapes communes.

Tout d'abord, les cellules contenues dans les tissus sont lysées (leur membrane est détruite) par l'ajout d'un agent chaotrope (qui détruit la structure spatiale des macromolécules, dont celles des protéines membranaires) ou par le biais d'une force mécanique (centrifugation, pompe à vide), voir les deux. Ensuite, par l'utilisation de divers solvants, l'ADN est solubilisé, récolté, précipité puis lavé afin d'être purifié et séparé des autres constituants du lysat cellulaire. Dans le cadre des extractions d'ADN en phase solide, la solution de cellules lysées passe dans une colonne contenant divers éluant et agent adsorbant liquides ou solides (comme la silice) ayant une affinité particulière pour certains des composants afin de récolter les acides nucléiques séparés et purifiés.

Afin de contrôler que l'extraction s'est déroulée correctement et que les acides nucléiques sont de qualité suffisante pour le séquençage, ils doivent être qualifiés. Encore une fois, de nombreuses techniques manuelles ou plus ou moins automatisées existent, mais leur principe technique reste le même.

La pureté de l'ADN extrait est déterminée à l'aide de la mesure de ratio d'absorbance par spectrophotométrie UV. Le ratio 260/280 nm permet de détecter une possible contamination des acides nucléiques par des protéines. La valeur attendue pour de l'ADN double brin pur varie entre 1,8 et 2,0. Le ratio 260/230 permet de détecter une possible contamination par des composés tels que l'EDTA, le phénol ou certains sucres. La présence de certains contaminants peut avoir un impact important sur certaines étapes inhérentes à la préparation de librairie de séquençage, comme la PCR, ou le séquençage lui-même, c'est pourquoi il est crucial de pouvoir détecter toute contamination.

La quantification de la concentration des acides nucléiques peut-être dérivée depuis les ratios d'absorbance calculés, mais les méthodes de quantification par fluorescence sont aujourd'hui préférées. Elles permettent le dosage de l'ADN double brin *via* l'interaction entre un fluorophore et les molécules d'ADN sous leur forme native. Le taux de fixation du fluorophore étant directement lié à la quantité d'ADN, l'intensité de fluorescence émise permet de déterminer la quantité d'ADN double brin en solution. De plus, cette technique permet une meilleure séparation de la quantification de la concentration d'ARN et d'ADN qui réagissent à des longueurs d'onde proches, ce qui peut fausser les résultats obtenus par spectrophotométrie.

Enfin, il est parfois nécessaire d'évaluer l'intégrité de l'ADN extrait par une méthode permettant d'estimer la distribution des tailles des fragments extraits. Cela est d'autant plus important lorsqu'un ADN de haut poids moléculaire est recherché ou en cas de suspicion d'échec technique. La migration sur gel d'agarose automatisée représente la méthodologie la plus répandue au sein des laboratoires et permet de déterminer la taille des fragments sur une plage plus ou moins étendue en fonction de la technique.

Une fois, l'ensemble de ces contrôles qualité passés avec succès, l'échantillon est qualifié pour être utilisé comme matrice pour la préparation de librairies de séquençage ou diverses techniques de biologie moléculaire. L'échantillon peut également être congelé afin d'être conservé en prévision de futures analyses.

1.3.3 Technologies de première génération

1.3.3.1 Le séquençage Sanger

Comme indiqué précédemment, le terme séquençage désigne toutes les technologies permettant de déterminer l'enchaînement d'acides aminés d'un fragment ou d'une molécule d'ADN à l'aide d'un automate, le séquenceur. Sont qualifiées de technologies de séquençage de première génération les méthodes apparues à la fin des années 70. Celle qui s'est imposée, car plus efficace et moins dangereuse que ses alternatives est la méthode mise au point par Frederick Sanger[39]. C'est cette technologie qui a permis le séquençage du premier génome de référence par le HGP.

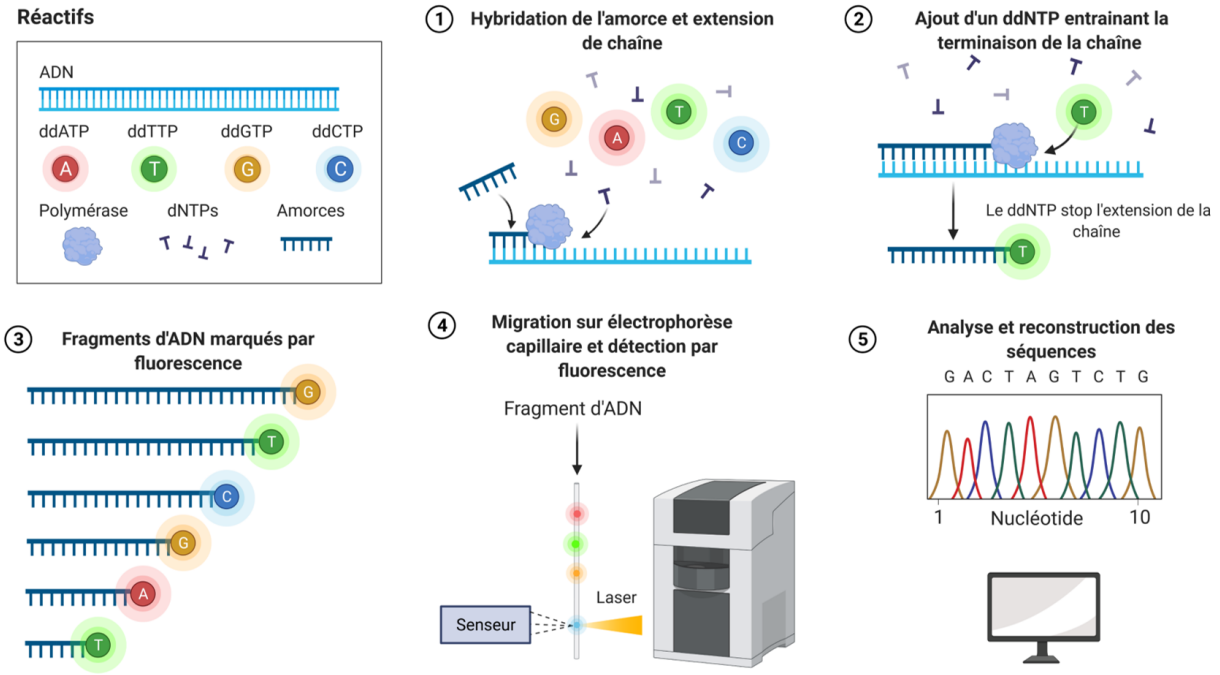


FIGURE 1.15 – Principes et étapes du séquençage Sanger.
Adapté d'après *Sanger Sequencing, BioRender.com (2021).*

Le principe de la méthode de séquençage Sanger, visible *Figure 1.15*, repose sur le principe de la réplication de l'ADN. Le séquençage se déroule dans un milieu contrôlé notamment riche en désoxyribonucléosides quelconques (dNTP), c'est-à-dire un mélange de nucléotides précurseurs de l'ADN et en didésoxyribonucléotide quelconque (ddNTP), un mélange de désoxyribonucléosides dont un groupement hydroxyle est absent.

Premièrement, la structure double brin de l'ADN est dénaturée par chauffage afin d'obtenir un simple brin de la molécule d'ADN à séquencer. Ensuite, une amorce spécifiquement conçue s'hybride à l'endroit souhaité du simple brin. L'ADN polymérase réplique ensuite le simple brin en incorporant les nucléotides présents dans le milieu réactif jusqu'à incorporer un ddNTP. L'absence du groupement hydroxyle des ddNTP empêche la polymérase de continuer à incorporer de nouveaux nucléotides au sein du nouveau brin. La réaction de polymérisation se termine, le dernier nucléotide de la molécule possède un fluorochrome.

Au sein des systèmes modernes de séquençage Sanger, chaque ddNTP est relié à un fluorochrome spécifique de la base azotée qui lui est associée et dont les longueurs d'onde d'excitation ainsi que d'émission sont différentes. Les fragments de différentes tailles obtenus précédemment subissent une migration par électrophorèse capillaire. Les fragments migrent dans le gel ou le tampon de migration et passent devant un laser excitant ainsi les flurochromes du ddNTP dernièrement incorporé. Grâce à un système de captation d'image, les longueurs d'onde émises par les fragments sont enregistrées puis converties en bases correspondantes (*basecalling*).

Le séquençage Sanger moderne est considéré comme du séquençage massivement parallèle (MPS pour *Massively Parallel Sequencing* à ne pas confondre avec le HTS décrit par la suite) car il peut séquencer entre 16 et 384 échantillons dans des capillaires différents (environ 300kb en 3h). Malgré le développement du séquençage à haut débit, le séquençage Sanger reste utilisé comme technologie de référence, du fait de son très faible taux d'erreur et de sa très forte disponibilité dans les laboratoires.

1.3.4 Technologies de seconde génération

1.3.4.1 Généralités sur le séquençage de seconde génération

Le séquençage d'ADN à haut débit (HTS pour *High-Throughput Sequencing*), séquençage de seconde génération, séquençage *short-reads* ou encore parfois qualifié du terme galvaudé de séquençage nouvelle génération (NGS), commercialisé à partir de 2005 est arrivé dans les laboratoires lors de la décennie suivante. Il a depuis pénétré la quasi-totalité des laboratoires de biologie médicale et les hôpitaux, car son adoption massive a entraîné une baisse concomitante du coût de séquençage. En effet, entre 2001 et 2021, le coût d'un séquençage de génome a été divisé par 100 000, dépassant même les prédictions de la loi de Moore qui prédit le doublement des performances de calcul informatique tous les deux ans, voir *Figure 1.16*. Le terme HTS désigne aujourd'hui presque exclusivement le séquençage basé sur la technologie Illumina, *leader* sur le marché.

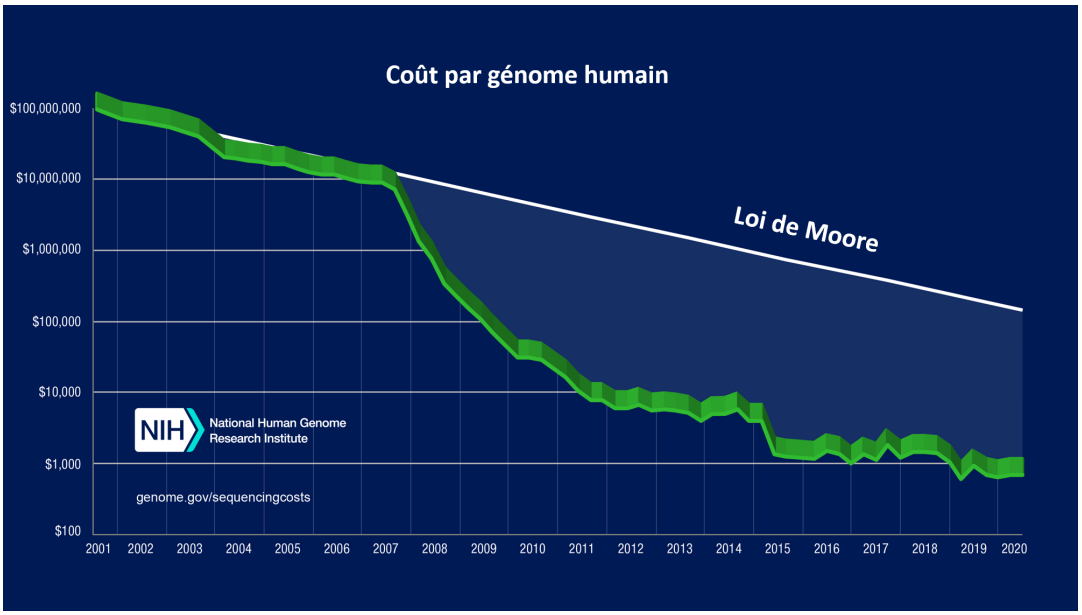


FIGURE 1.16 – Diminution du coût de séquençage humain de 2001 à 2020.
Adapté d'après *NIH DNA Sequencing Costs Data.*

1.3.4.2 Le séquençage Illumina

La première étape commune à tout séquençage Illumina est ce qui est appelé la préparation de la librairie de séquençage. L'ADN extrait est tout d'abord fragmenté, le plus souvent par fragmentation mécanique, enzymatique ou ultrasonication, voir *Figure 1.17*. Des adaptateurs spécifiques sont ensuite ajoutés aux extrémités 5' et 3' des fragments nouvellement produits par une enzyme l'ADN ligase. La librairie est prête est peut-être utilisée sur le champ ou congelée pour un séquençage ultérieur.

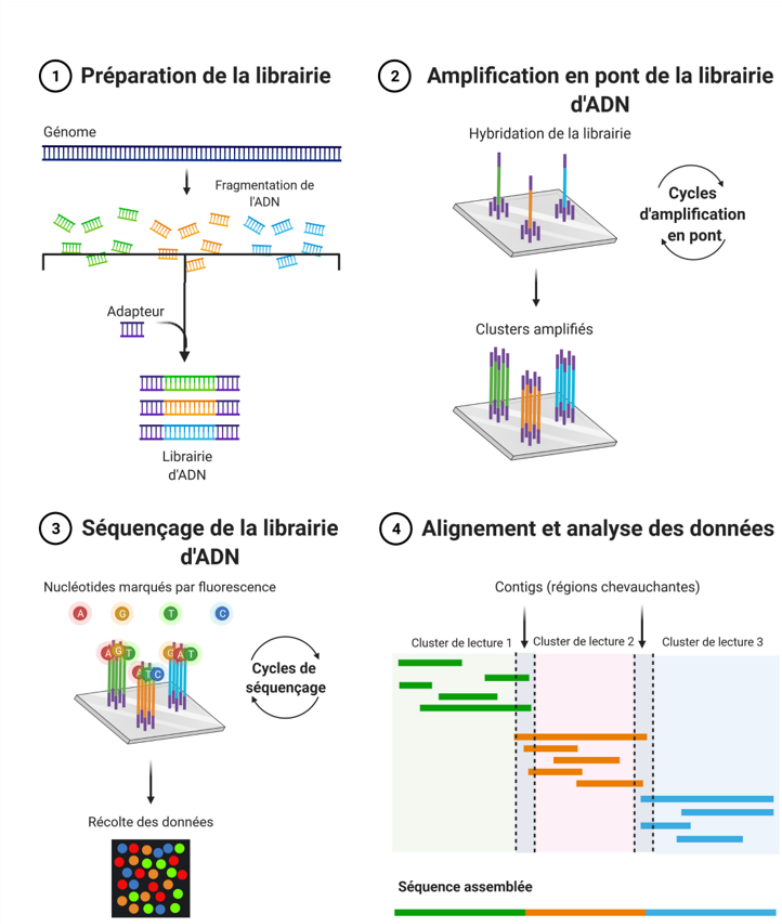


FIGURE 1.17 – Principes et étapes du séquençage Illumina.
Adapté d'après *Next Generation Sequencing (Illumina), BioRender.com (2021).*

La librairie est ensuite chargée sur une *flowcell*, un support de verre contenant des micropuits (le nombre varie en fonction du modèle de séquenceur) où se déroulent les étapes suivantes d'amplification et de séquençage. Les adaptateurs liés aux fragments d'ADN permettent la liaison de ces derniers à des oligonucléotides complémentaires (d'autres adaptateurs) disposés dans les micropuits de la *flowcell*. Les fragments sont alors dénaturés par chauffage, puis une extrémité du simple brin s'hybride à la *flowcell*. Le brin original est répliqué par une ADN polymérase puis est éliminé par lavage. Les simples brins complémentaires hybridés à la *flowcell* subissent plusieurs cycles d'amplification par PCR en pont (*bridge PCR amplification*). Lors de chaque cycle, la deuxième extrémité libre du fragment s'hybride à la *flowcell*. La formation en pont permet la réplication et la création d'un nouveau brin complémentaire. À la fin des cycles d'amplification, il en résulte des *clusters* clonaux de notre fragment originel.

S'ensuit l'étape de séquençage à proprement parler même si le terme est en général utilisé pour désigner toutes les étapes se déroulant au sein d'un séquenceur. Les simples brins des *clusters* vont être répliqués. D'une manière similaire au séquençage Sanger, des dNTP couplés à des fluorochromes vont être incorporés au brin complémentaire en cours de la synthèse. Après chaque incorporation (cycle), le *cluster* est excité par un laser. Celui-ci répond par un flash lumineux d'une longueur d'onde correspondante au nucléotide intégré. Le nombre de cycles détermine la longueur des lectures (*reads*), la retranscription sous la forme de fichiers textuels de l'enchaînement des nucléotides d'un fragment issu d'un *cluster* à l'aide de l'alphabet A, C, T, G. Le nombre de cycles est déterminé en fonction de la méthodologie de préparation des librairies et de la longueur estimée des fragments séquencés. Les fichiers obtenus à l'issue du séquençage sont des fichiers au format BCL (*Binary Base Call*) qui sont des fichiers représentant les intensités lumineuses des différents *clusters* excités par les fluorochromes. Pour obtenir les fichiers des *reads* textuels au format FASTQ, il est nécessaire de passer par une étape d'appel de bases (*basecalling*).

La technologie de séquençage Illumina est extrêmement fiable avec en moyenne 0.1% d'erreurs à la base. La limitation principale de la technologie est la taille de ses *reads*. En effet, les *reads* produits avec cette technologie sont assez courts (en général entre 75 et 250pb), c'est pourquoi elle est souvent qualifiée de séquençage *short-reads*. Cette faible taille de fragments induit de moins bonnes performances notamment pour caractériser les zones répétées ou les variations de structure du génome. De plus, l'amplification par PCR des régions répétées et des régions de faible complexité est moins efficace que dans le reste du génome[40].

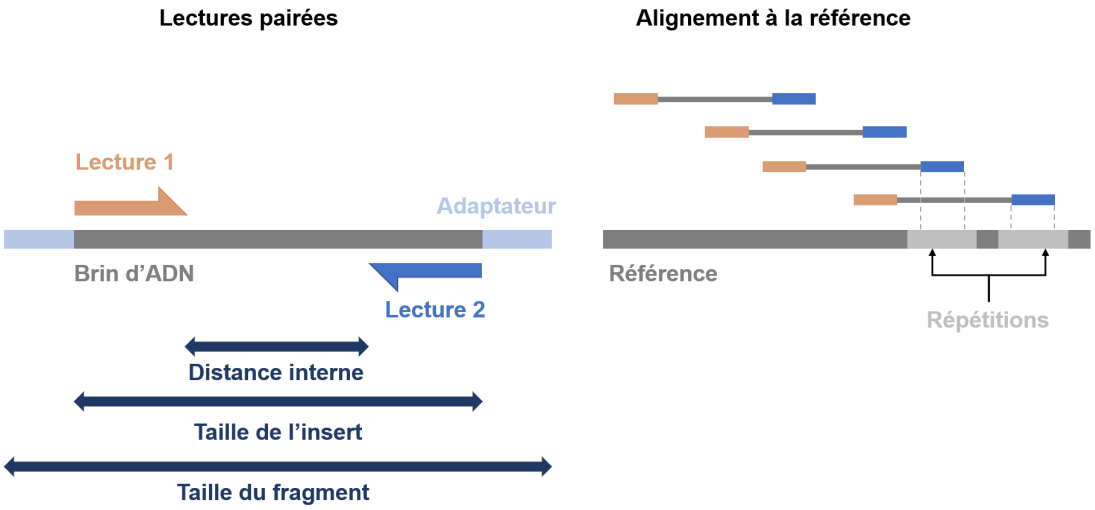


FIGURE 1.18 – Principe du séquençage par paire et de l'alignement des lectures pairées à la référence.

Une avancée majeure de la technologie Illumina pour résoudre en partie ces problèmes a été la mise au point des méthodologies de séquençage en paires ou par lectures pairées (PE pour *paired-end*). Grâce à des adaptateurs spécifiques, les fragments d'ADN issus des *clusters* sont lus en sens et antisens lors du séquençage, voir *Figure 1.18*. Ainsi, la position de la paire de lectures permet de déterminer plus précisément les positions des lectures sur le génome de référence lors de l'alignement, notamment dans le cas de zones répétées. Les distributions de tailles d'inserts (*insert size*) et les distances internes (*inner distance*) sont également des indicateurs extrêmement importants afin de déterminer les paires discordantes. En effet, dans le cas où l'une des lectures d'une paire tomberait dans un SV, selon le type, la taille de l'insert ou la distance interne de la paire aurait une valeur aberrante par rapport à la distribution normale et indiquerait donc la présence potentielle d'une variation de structure.

Comme indiqué précédemment, le séquençage Illumina est capable de produire une quantité phénoménale de données par expérience (*run*). La quantité de données produites et le nombre d'échantillons séquencés en même temps varient en fonction de ce qui est séquencé, comment et avec quel séquenceur.

Illumina a sur le marché une gamme de séquenceurs extrêmement étendue qui permet de couvrir la quasi-totalité des besoins en ce qui concerne le séquençage. Les séquenceurs se distinguent principalement par le rendement en termes de séquences produites, le prix à l'achat et le prix de réalisation d'une librairie. Tous les séquenceurs ne sont pas indiqués pour séquencer un génome humain et peuvent être destinés pour d'autres analyses, voir *Figure 1.19*.

En général, plus un séquenceur est cher à l'achat, plus il est capable de produire beaucoup de données en un seul *run*. Encore faut-il pouvoir, assumer l'investissement initial, pouvoir recruter assez d'échantillons afin de remplir le séquenceur, ainsi qu'utiliser le séquenceur à son plein potentiel pour faire baisser les coûts de séquençage. Les coûts de séquençage ne sont que la face émergée de l'iceberg en ce qui concerne le coût réel d'une analyse de génome. Plus les données séquencées sont importantes, plus les coûts de gestion et de stockage des données augmentent ainsi que les coûts et temps relatifs à l'analyse des données. Ces raisons, parmi d'autres, sont certaines des motivations qui ont poussées de nombreux laboratoires dont le CHU de Grenoble et Eurofins Biomnis à préférer le séquençage d'exome plutôt que le séquençage de génome entier en tant qu'activité de routine diagnostic.

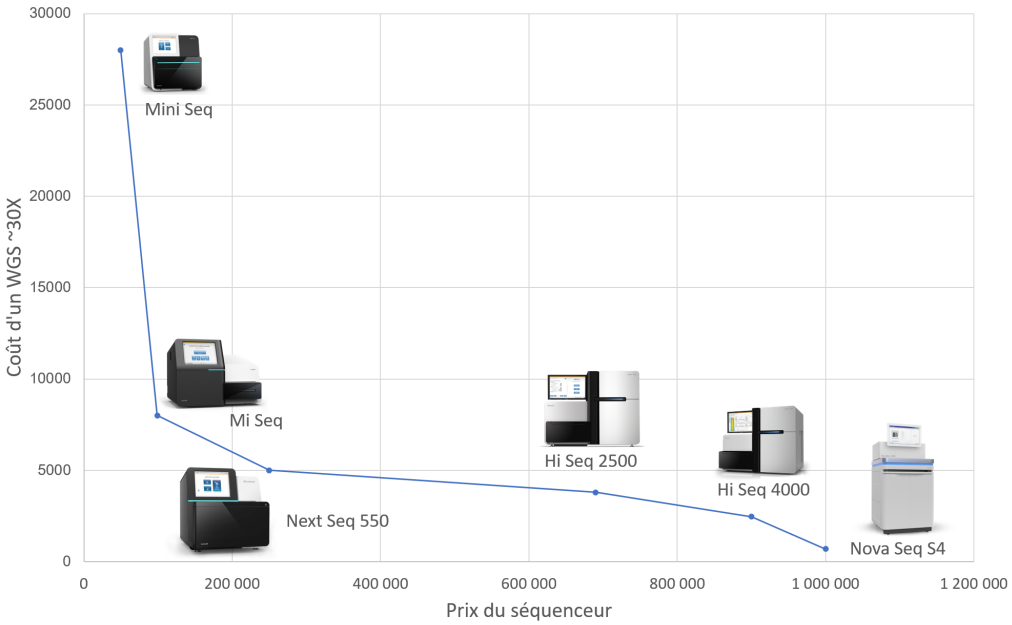


FIGURE 1.19 – Prix d'un séquençage de WGS humain à environ 30X en fonction du prix du séquenceur. D'après les données de *coût de séquençage, maintenues par Albert Viella.*

Aujourd'hui le séquençage de génome entier (WGS pour *Whole Genome Sequencing*) Illumina est utilisé pour la détection de SNV, d'indels et de SV supposés à l'origine du phénotype clinique du patient séquencé. Le WGS Illumina est encore assez peu répandu dans les laboratoires français hospitaliers. Néanmoins, le plan France Médecin Génomique 2025 (PFMG 2025) avec la mise en place de laboratoires de biologie médicale de séquençage à haut débit de génome vise à démocratiser l'utilisation du WGS dans le cadre du diagnostic de maladies rares en France.

1.3.4.3 Le séquençage par capture d'exome

Le séquençage d'exome (WES pour *Whole Exome Sequencing*) est une méthode de séquençage de l'ADN ciblée. En théorie, seules les parties codantes du génome humain sont ciblées ou capturées lors d'une étape de la préparation de librairie, puis séquencées comme une librairie Illumina classique. Le WES est extrêmement populaire, car il possède un très bon rapport de taux de diagnostic sur investissement[41], l'exome étant le support de 85% des variations supposées pathogènes selon les connaissances médicales actuelles[42]. En pratique, le terme de capture désigne l'ensemble des étapes nécessaires pour séparer puis récupérer les fragments d'ADN issus des exons et les autres. De nombreux kits de capture commerciaux existent avec diverses compositions et stratégies. Lors de cette thèse, deux kits de capture ont été utilisés, tout d'abord le kit *Roche MedExome*, remplacé par le plus performant *Twist Bioscience Human Core Exome*.

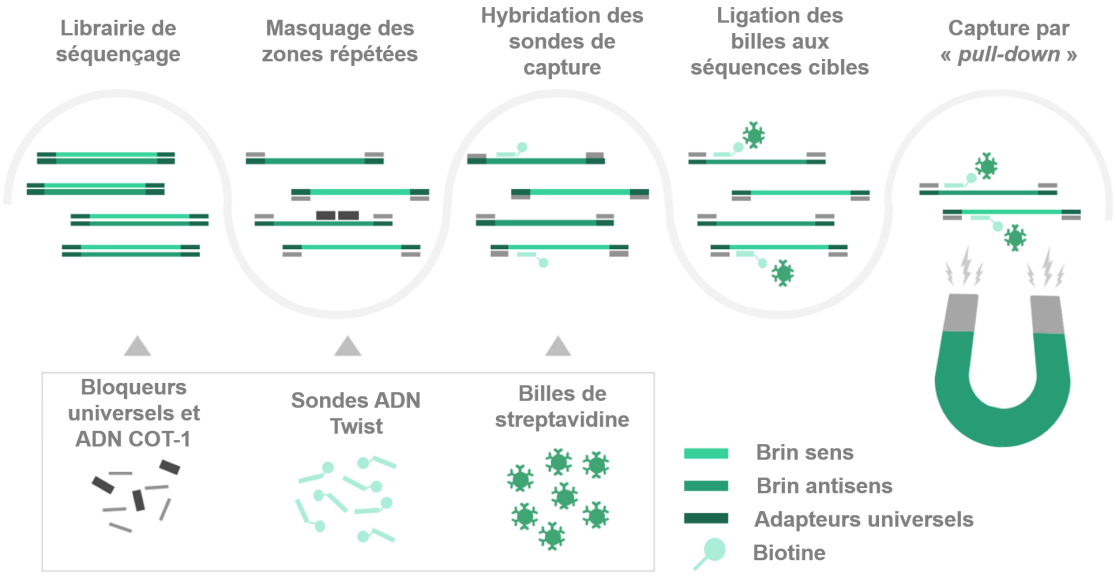


FIGURE 1.20 – Principe du séquençage par capture *Twist Bioscience Human Core Exome*.
Adapté d'après *Universal Blockers*.

Le principe général de la capture est l'introduction d'oligonucléotides (sondes) complémentaires aux régions codantes et spécifiques à chaque constructeur. Dans le cas du kit *Twist Bioscience*, cette étape est précédée d'un ajout de séquences bloqueuses, notamment de l'ADN Cot-1, voir *Figure 1.20*. Ces oligonucléotides s'hybrident avec les fragments d'ADN issus de régions répétées non spécifiques, les masquant, pour éviter qu'ils ne soient capturés. Ensuite, les sondes sont ajoutées au mélange. Les sondes sont des séquences doubles brin d'une longueur comprise entre 50 et 120 paires de bases. Le mélange de fragments d'ADN et de sondes est ensuite chauffé. Ainsi, les sondes se dénaturent et forment de simples brins. En abaissant la température, les sondes tentent de retrouver leur forme stable d'ADN double brin et par complémentarité, s'hybrident aux fragments d'ADN ciblés par la capture.

Les sondes possèdent une molécule de biotine à une de leurs extrémités. La biotine est une molécule qui est utilisée en raison de sa très haute affinité de fixation avec la protéine appelée streptavidine, fixation qui est l'une des plus fortes interactions connues en biologie. Des billes magnétiques enrobées de streptavidine sont alors ajoutées au mélange. Une fois que les billes enrobées sont fermement liées aux sondes biotinylées, un aimant est utilisé pour séparer l'ADN de l'exome du reste. La solution est ensuite lavée pour éliminer les fragments non souhaités. Ces différentes opérations constituent l'étape de *pull-down*, les fragments issus de la capture sont alors prêts pour le séquençage.

Au même titre que le WGS, le WES permet de détecter les SNV et les indels issus des parties codantes supposées à l'origine du phénotype clinique du patient séquencé. Le WES ne permet pas de détecter tous les types de SV, seuls les CNV impliquant les exons sont détectables à l'heure actuelle.

1.3.5 Technologies de troisième génération

Comme indiqué précédemment, la limite principale de la technologie Illumina est la taille de ses *reads* qui s'avère limitante pour la caractérisation de certaines régions du génome humain. Plusieurs technologies ont été développées pour pallier ce défaut, notamment les lectures liées (*linked-reads*) de 10X Genomics ou les lectures longues (*long-reads*) d'Oxford Nanopore Technologies et Pacific Bioscience. Ce sont surtout ces dernières qui nous intéressent dans le cadre de cette thèse.

Les caractéristiques principales des techniques de séquençage de longues lectures sont, le séquençage d'une seule molécule d'ADN (*Single Molecule Sequencing*) et les tailles des lectures qu'elles produisent. Les technologies *long-reads* comme leur nom l'indique produisent donc des lectures plus grandes que les technologies *short-reads*. La taille moyenne des lectures produites est au moins supérieure à 1 kb (même si la moyenne de taille des *reads* générés par les deux technologies abordées par la suite est en général au minimum centrée autour de 5 kb). Deuxièmement, ces technologies ont un temps de séquençage réduit qui peut varier entre quelques heures ou jours (voire minutes pour certaines technologies *real-time* en fonction de l'organisme étudié). De plus, ces techniques ne nécessitent pas d'amplification de l'ADN. L'amplification de l'ADN est une étape pouvant induire de nombreuses erreurs de réplication et est donc un biais en moins pour les TGS[43].

1.3.5.1 Oxford Nanopore Technologies

Le séquençage à l'aide de nanopores est une idée datant des années 90[44]. Diverses idées et améliorations ont permis la commercialisation du dispositif MinION au grand public par Oxford Nanopore Technologies (ONT) en 2015. Le MinION se présente sous la forme d'un petit boîtier, pouvant se brancher en USB 3 à un ordinateur. Il est composé de deux parties distinctes, la *flowcell*, consommable, qui embarque la microélectronique et la microfluidique nécessaire au séquençage et le boîtier qui permet d'alimenter et de transférer les données, voir *Figure 1.21*. Après chargement des échantillons et le séquençage initié, ce dernier est modulable à l'aide d'une interface utilisateur disponible sur l'ordinateur.

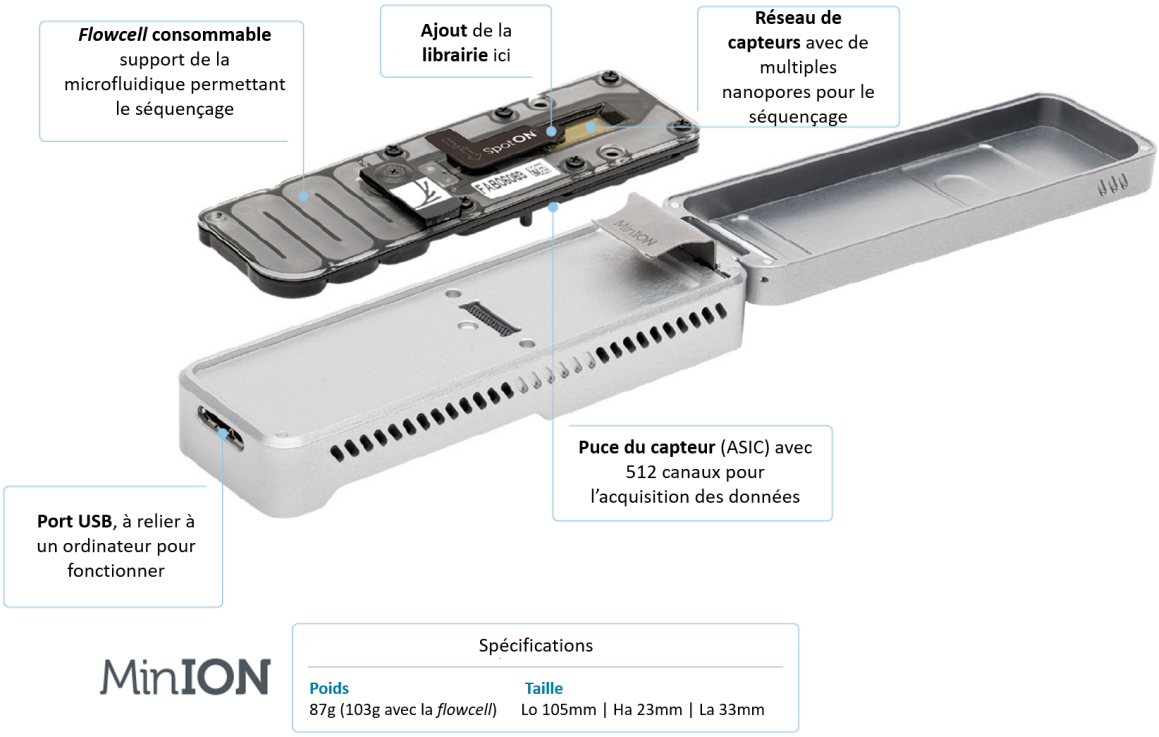


FIGURE 1.21 – Représentation d'un séquenceur MinION.
Adapté d'après *MinION Brochure.*

De la même manière que son concurrent direct, ONT a diversifié son offre de séquenceurs pour couvrir le plus d'applications possible, du séquençage de bactérie et de virus jusqu'au séquençage du génome humain en passant par le séquençage de données de capture, voir *Figure 1.22*.

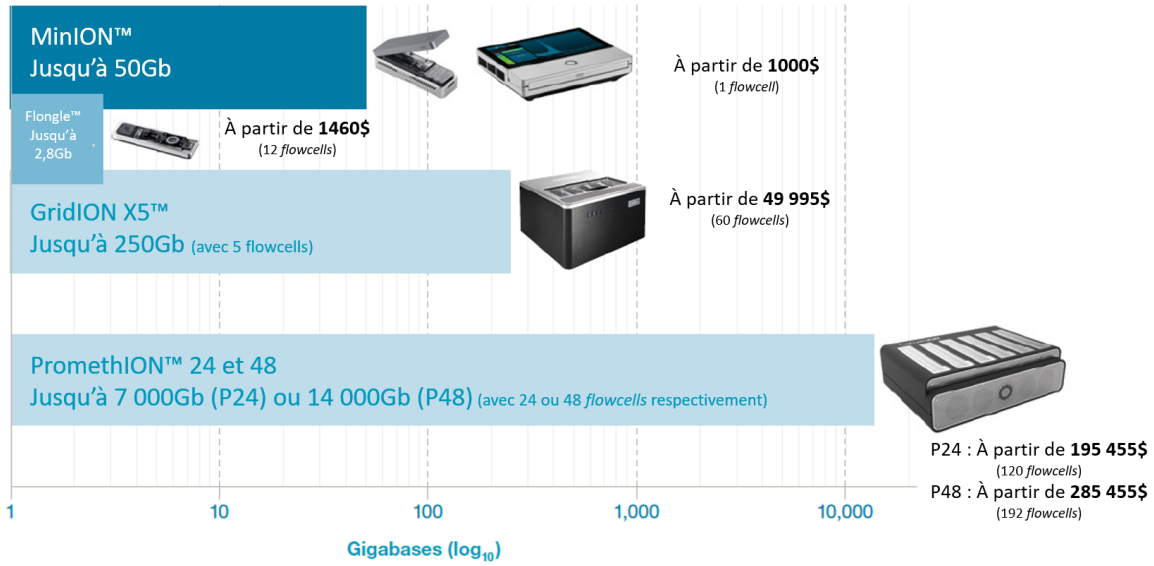


FIGURE 1.22 – Gamme de séquenceurs Oxford Nanopore Technologie prix et rendements associés. Adapté d'après *MinION Brochure.*

Le Flongle est destiné aux séquençages qui ne nécessitent pas énormément de production de données comme le séquençage d'amplicons ou de petits organismes (virus, bactéries). Le MinION et le GridION X5 se partagent les mêmes *flowcells*, la seule différence étant que le GRIDION est capable de séquencer jusqu'à 5 *flowcells* MinION en parallèle. Le PromethION est le plus gros séquenceur de la gamme d'ONT et utilise des *flowcells* différentes possédant un plus grand nombre de pores. Il a été conçu pour le séquençage de génomes de taille importante (plantes, mammifères) et est actuellement utilisé pour du diagnostic clinique par une poignée de pionniers à travers le monde. Le prix d'un séquençage humain 30X avec un PromethION est d'environ 5000 $[45].

Le principe du séquençage par nanopore est le même pour tous les séquenceurs d'ONT, seul change le type et le nombre de pores. Le séquençage nanopore peut être qualifié de *strand sequencing* ou séquençage d'un seul brin, sans amplification au préalable. Le principe de cette technologie repose sur un complexe protéique spécifique, le nanopore. Un nanopore est constitué d'un polymère de protéines ancré dans une membrane électro résistante également constituée de polymères synthétiques[46]. Le cœur de ce complexe protéique forme un long tube creux de quelques nanomètres de diamètre qui traverse de part en part la membrane synthétique. Les *flowcells* MinION disposent de 512 canaux permettant le séquençage, chacun hébergeant 4 nanopore ce qui représente le séquençage hypothétique de 2048 brins en parallèle. Malheureusement, en réalité, seul environ 60% de ces pores seront correctement fonctionnels lors de l'expérimentation. Ils sont détectés lors de tests de contrôle de la qualité de la *flowcell* en amont du séquençage.

Un potentiel électrique est appliqué à travers la membrane ce qui a pour conséquence de créer un courant électrique à travers l'ouverture du nanopore. Les molécules passant à travers le nanopore causent un différentiel de potentiel électrique. En théorie, en mesurant cette perturbation qui est caractéristique des molécules, celle-ci peut être identifiée, voir *Figure 1.23*.

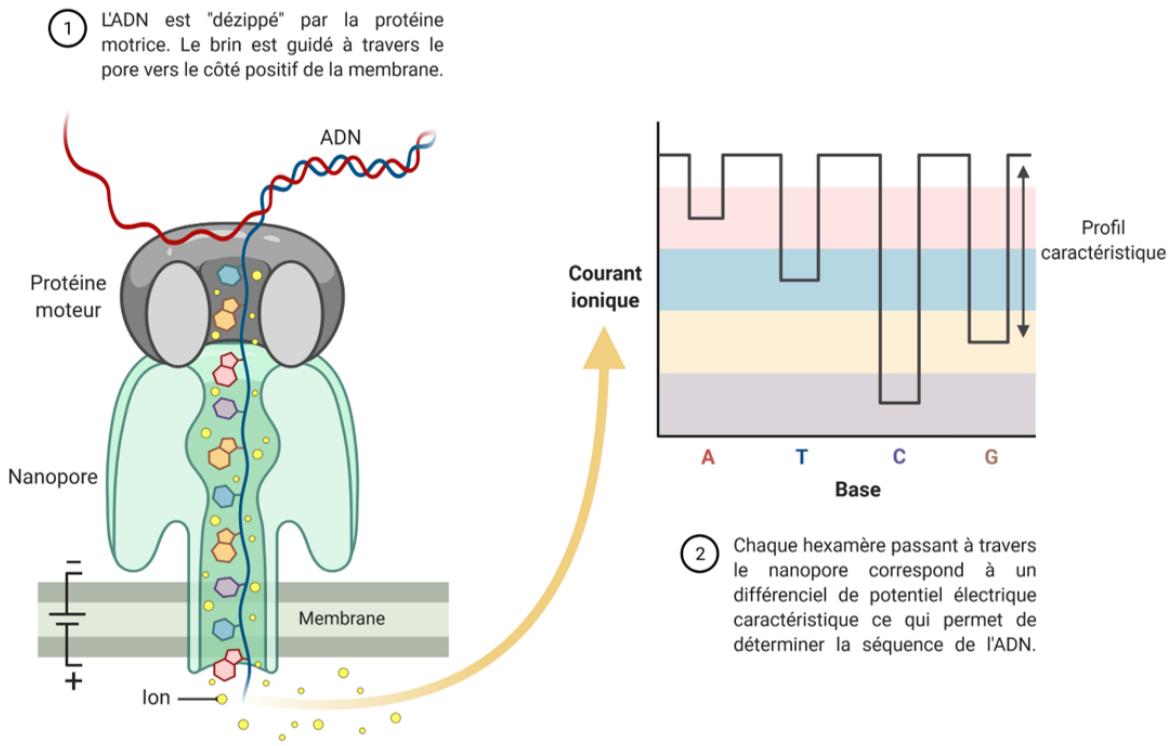


FIGURE 1.23 – Schéma d'un nanopore et d'un *squiggle plot*, représentation schématique la variation du courant électrique au sein de la membrane.
Adapté d'après *Nanopore Sequencing, BioRender.com (2021).*

La préparation d'une librairie de séquençage nanopore consiste à tout d'abord fragmenter l'ADN extrait en longs fragments. Cette étape n'est pas obligatoire, ONT vend sa technologie comme pouvant séquencer en théorie des brins entiers. Mais dans la pratique, de trop longs brins auront tendance à boucher les nanopores et diminuer le rendement du séquençage. En général, pour un séquençage de génome humain, les *reads* sont fragmentés afin d'obtenir une longueur entre 10 et 15 kb. La diminution de la taille des lectures a pour effet d'augmenter le rendement du séquençage, jusqu'à une certaine limite. Au contraire, plus les *reads* à séquencer sont longs, moins le séquençage produira de données[12].

La préparation de librairie consiste en la ligation d'une ADN polymérase modifiée aux brins à séquencer et à l'ajout d'un adaptateur "en Y" qui aura pour rôle de permettre le couplage de l'ADN polymérase au nanopore. Le rôle de la polymérase ou protéine moteur est de « dézipper » le brin bicaténaire au même titre que le ferait une Fermeture Éclair et de guider le brin à travers le pore, à une vitesse de 400 bases par seconde. Lors du passage du simple brin, les modifications caractéristiques du potentiel électrique engendré par le passage des différentes combinaisons de bases sont enregistrées par la puce ASIC du dispositif et sont stockées sous la forme de valeurs électriques. Ces valeurs peuvent être représentées sous la forme d'un graphique d'intensité appelée *squiggle plot* voir *Figure 1.23*. Ces valeurs d'intensité seront par la suite traduites en fichiers de séquence de base au format FASTQ lors du *basecalling*, grâce à des modèles d'apprentissage

(*machine-learning* ou *deep-learning*). Les outils et donc les modèles de *basecalling* à utiliser dépendent du type de chimie et du type de données séquencées. Lorsque le séquençage de la molécule est terminé, la protéine motrice se détache et le pore est disponible pour séquencer un nouveau fragment.

En réalité, la variation du potentiel électrique qui est détectée par les dispositifs senseurs du puits n'est pas dépendante d'une seule base, mais de plusieurs. Le nombre varie en fonction de la chimie du kit, de l'outil et la version du *basecaller* (logiciel de *basecalling*) utilisé. Pour le kit version *R9.4.1*, la détermination des bases peut soit s'effectuer sur des pentamères ou des hexamères. Il y a donc 45 (1024) ou 46 (4096) possibilités. Il existe un biais préexistant très important dû au principe de détection de penta ou hexamères. Lors de l'étape de *basecalling*, lorsque les signaux électriques sont convertis en séquences, les portions homo-kmerique (ou homopolymèriques) telles que "AAAAA..." ou "TTTTT..." sont sous-représentées en longueur, du fait de la non-variation du potentiel électrique. Afin de pallier ce problème, ONT a développé un nouveau type de pore disponible sur les chimies ultérieures à la version R10, voir *Figure 1.24*.

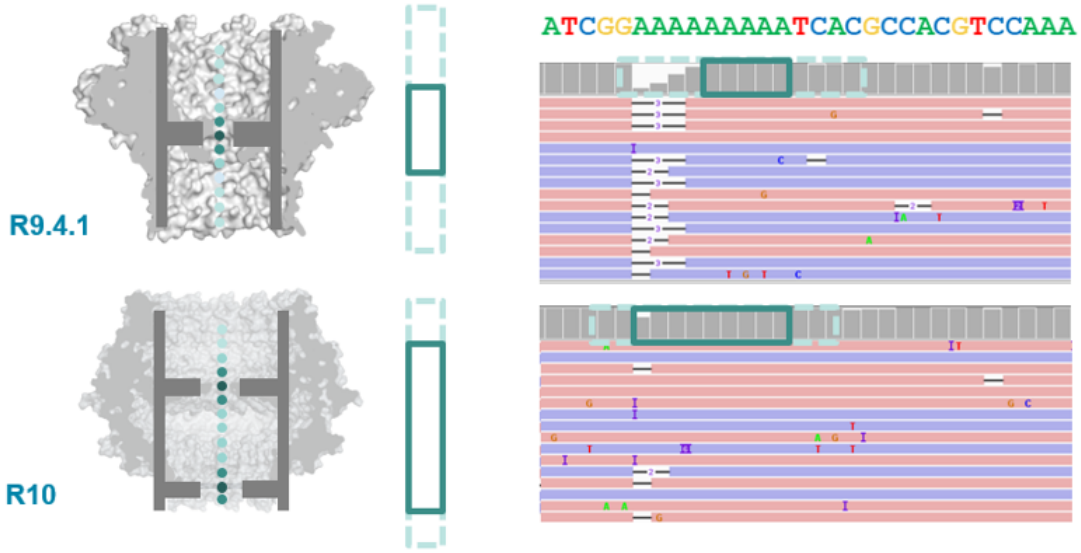


FIGURE 1.24 – Représentation d'un pore de kit *R9.4.1* et d'un pore *R10* et ultérieure et leur impact lors du séquençage d'une région homopolymérique.
Adapté d'après *R10.3: the newest nanopore for high accuracy nanopore sequencing.*

Les pores ultérieurs à la version de chimie de kit R10, la chimie *R10.3* étant la chimie la plus avancée à l'heure actuelle, ont deux têtes de lecture contrairement aux pores des kits *R9.4.1*. Cela a pour avantage permettre une meilleure résolution des régions homopolymériques et d'améliorer la précision des données de séquençage nanopore, au coût d'un rendement moins élevé par rapport à la chimie *R9.4.1*. Les usages de ces deux versions de chimie de séquençage sont différents, c'est pourquoi ONT maintient et tente d'améliorer les deux versions. L'idéal serait une chimie dont les pores disposent de deux têtes de lecture tout en ayant un rendement aussi important qu'avec une.

La technologie ONT de par la taille de ses *reads* a une meilleure sensibilité de détection de SV même à basse couverture (environ 5X) dans des zones répétées par rapport au séquençage Illumina[47]. Néanmoins, la précision à la base de la technologie ONT est plus faible que celle de la technologie Illumina[12], il est assez difficile de l'utiliser pour la détection de SNV pour du génome humain à l'aide de *flowcells* MinION malgré sa capacité de détection de SV et d'identification de longues répétitions. Cela est pourtant possible à l'aide de séquençage PromethION, car il permet de séquencer plus profondément et d'ainsi corriger les erreurs de

séquençage, mais son usage n'est pas encore très répandu. Pour l'instant la technologie ONT est assez peu utilisée en séquençage humain de routine dans le monde.

Une *flowcell* de MinION ou de GridION permet, lorsqu'exploitée à son plein potentiel, de séquencer un génome humain à environ 10X pour un coût estimé à 1200 euros par échantillon (données du laboratoire).

1.3.5.2 Pacific Bioscience

La technologie de Pacific Bioscience (ou PacBio) SMRT (*Single-Molecule Real-Time*), est une méthode de séquençage direct de l'ADN qui ne nécessite pas d'amplification au même titre qu'ONT. Introduite sur le marché en 2010, la technologie SMRT se base sur la réplication de l'ADN par l'ADN polymérase de la même manière que le séquençage Sanger ou Illumina[48].

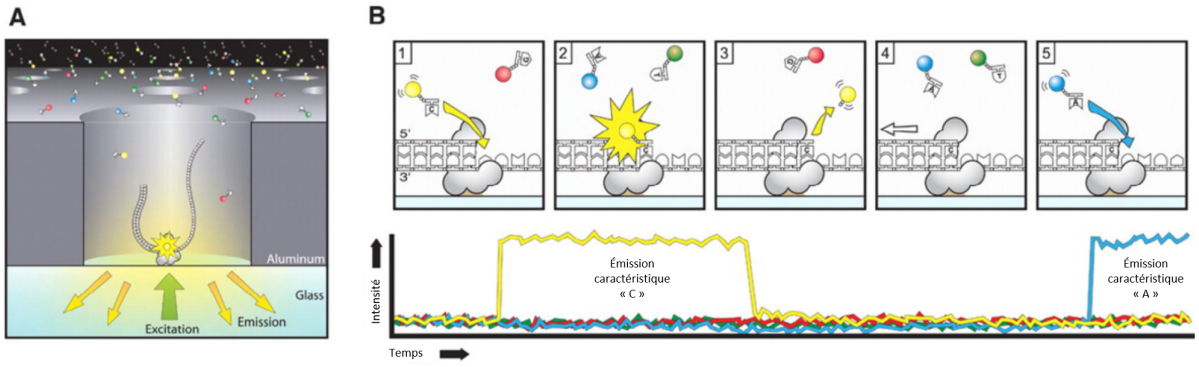


FIGURE 1.25 – Principe du séquençage PacBio SMRT.
Adapté d'après *Rhoads, A., Au, K. F. (2015). PacBio sequencing and its applications. Genomics, proteomics bioinformatics, 13(5), 278-289.*

Les nucléotides qui vont être incorporés dans le milieu sont couplés à des fluorochromes. Chaque base possède un fluorochrome d'une certaine longueur d'onde spécifique à son type (A, C, T ou G) capable d'être détecté et différencié. Le fluorochrome est lié au phosphate terminal du nucléotide (contrairement aux technologies précédentes où le fluorophore est directement sur la base). Lorsque le nucléotide est incorporé dans le brin complémentaire nouvellement synthétisé par la polymérase, ce fluorophore est clivé. Le fluorophore est excité, un flash de lumière qui sera détecté par le séquenceur est émis et le brin d'ADN est alors totalement naturel, voir *Figure 1.25*.

Les brins d'ADN à séquencer sont déposés sur des cellules SMRT (*SMRT-cell*), éponyme à la technique. Ces cellules contiennent 150 000 microstructures appelées détecteurs ZMW (*Zero Mode Waveguides*) qui jouent le rôle de chambres de visualisation. Elles sont le support du séquençage et ressemblent dans le principe aux *flowcells*. Entre 35 000 et 75 000 seront effectives et produiront des *reads*. Chaque ZMW contient une polymérase qui est accrochée au support en verre. Afin d'optimiser la technique, au préalable, lors de la préparation des échantillons des ZMW, les brins d'ADN sont mis sous la forme *SMRTbell* ou *SMRTbell template*. C'est-à-dire, un ADN circulaire monocaténaire fermé qui est créé par la ligature d'adaptateurs en épingle à cheveux aux deux extrémités d'une molécule d'ADN bicaténaire, visible *Figure 1.26*. Cette forme garantit la linéarité et la stabilité de la molécule, permettant à la polymérase de synthétiser le brin complémentaire nécessaire au séquençage. Cette synthèse commence au niveau d'un des adaptateurs en épingle à cheveux.

Les ZMW, en aluminium, d'environ 70 nm de largeur, permettent l'observation du clivage des fluorophores par la polymérase lors de l'ajout du nucléotide. Grâce à leur faible volume (20 zeptolitres ou 20 x 10^{-21} litres, le plus petit volume possible pour la visualisation d'un phénomène lumineux) les signaux sont concentrés et détectés sans bruit de fond par un système optique à travers le support de verre de la cellule. Les ZMW actifs permettent la parallélisation massive du processus. Les différents signaux émis lors du séquençage d'un *SMRTbell* d'un ZMW donné sont alors enregistrés par l'appareil et décodés en nos séquences génomiques ou en nos longues lectures lors du *basecalling*.

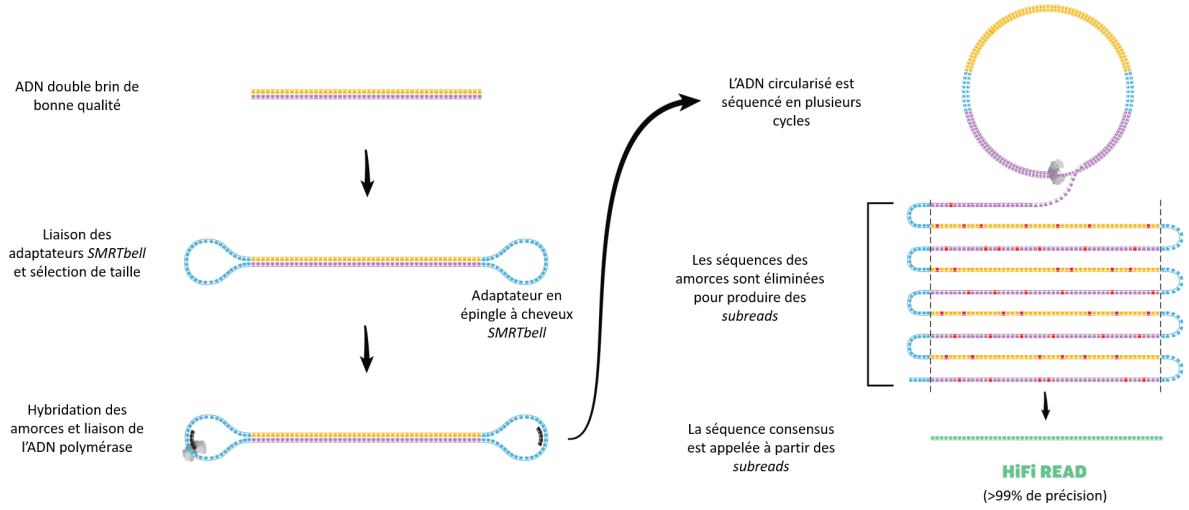


FIGURE 1.26 – Librairie de séquençage SMRT et production de lectures HiFi.
Adapté d'après *CCS docs*

Les polymérases ont une durée de vie limitée. Comme la molécule répliquée est circulaire, une fois le tour effectué, la polymérase peut recommencer. Chaque tour est appelé un cycle ou *pass*. Il en résulte un *read* circulaire associant des copies des deux brins d'ADN bout à bout, c'est le consensus circulaire (ou CCS pour *Circular Consensus Sequencing*). Celui-ci pourra être coupé en *subreads* (sous-ensemble d'un *read* issu d'un *read* parent) dans lesquels seront éliminées les séquences des adaptateurs. Une séquence consensus extrêmement précise va être déterminée à partir des *subreads* produits, elle est appelée lecture HiFi pour *High Fidelity*[49]. Le nombre de *pass* sur l'ensemble des séquences équivaut à la profondeur de séquençage.

La durée de vie de la polymérase et donc le nombre de *pass* par molécule dépendent de deux choses. Premièrement, la taille de la séquence, car logiquement, à durée de vie égale, plus la séquence est courte, plus l'enzyme fera de tours. Deuxièmement, les kits et réactifs de séquençage et notamment les tampons dans lesquels la polymérase synthétise le brin complémentaire. Depuis la mise au point du séquençage SMRT, Pacific Bioscience améliore la qualité et la formule de ses réactifs permettant d'obtenir des lectures de plus en plus longues. Grâce à ces nouveaux kits, la longueur moyenne des *reads* produits par les séquenceurs PacBio de la gamme Sequel II (les seuls encore maintenus) oscille entre 5 000 et 20 000 pb.

Actuellement, la PacBio HiFi est la technologie de séquençage de troisième génération avec la plus haute précision à la base[45][49] (99% de fiabilité médiane), c'est-à-dire qu'elle ne produit que très peu d'erreurs de séquençage. En revanche, c'est également la technologie la plus onéreuse pour ce qui est de séquencer un génome humain, voir *Table 1.5*. Son fort coût l'empêche pour l'instant d'être adoptée dans une majorité de laboratoires. Son utilisation reste pour l'instant réservée à des activités de recherche dans de très grands centres internationaux. Elle n'a donc pas pour l'instant vocation à être utilisée en diagnostic de routine, mais est utilisée notamment par l'assemblage de génome de référence (humain entre autres). Elle a été utilisée par le consortium T2T pour différencier des copies de répétitions ou des haplotypes subtilement divergents[14].

Consommable / étape de préparation de librairie	**Quantité**	**Prix**
Contrôle qualité de l'échantillon à séquencer	1	8,44 $
Préparation de la librairie SMRTbell	1	700,51 $
Préparation supplémentaire	1	265,16 $
Fragmentation et sélection de fragment taille BluePippin	1	363,49 $
Flowcell Sequel II 8M SMRT séquençage long	3	10 315,11 $
Analyse de données CCS	3	984,78 $
Total	-	12 713,49 $

TABLE 1.5 – Prix d'un séquençage de génome humain entre 25 et 30X avec un séquenceur Sequel II issu de la technologie PacBio HiFi.
Adapté d'après *HiFi Data collection (25-30X coverage of a human-sized genome), University of Washington PacBio Sequencing Services.*

1.3.5.3 Technologie de séquençage *linked reads* Chromium

Les technologies de séquençage de lectures liées (*linked reads*) font partie des technologies de séquençage de troisième génération, mais ne sont pas des technologies de séquençage à lecture longue à proprement parler. L'objectif initial de la technologie Chromium 10X Genomics (https://www.10xgenomics.com/) est de permettre aux laboratoires possédant seulement des séquenceurs Illumina d'effectuer du WGS comportant une information supplémentaire de *barcoding* ajouté lors de la préparation de librairie. Les lectures liées fournissent des informations à longue distance absentes des approches *short-reads* classiques. Tout le principe de la technologie est basé sur le fait que toutes les lectures qui partagent un même code-barres peuvent être regroupées comme provenant d'une seule longue molécule d'entrée. Ainsi, il est possible de lier les lectures entre elles et donc de lier différents *loci* éloignés séquencés comme provenant du même brin. Grâce à cette information supplémentaire, il est en théorie plus simple d'effectuer des opérations telles que le phasage des haplotypes et la détection de variation de structure.

La *barcoding* des *reads* est effectué à l'aide d'un automate, spécifique à la technologie, appelé *Chromium controller*, visible *Figure 1.27*. Celui-ci est un dispositif microfluidique de préparation de librairies compatibles avec les séquenceurs Illumina. Le dispositif Chromium a pour rôle de mélanger des billes de gel fonctionnalisées contenant des codes-barres uniques avec des enzymes et une quantité limitée d'ADN génomique de haut poids moléculaire. Ces composants sont ensuite encapsulés dans de l'huile pour produire un GEM (*Gel-bead in EMulsion*), voir*Figure 1.27*.

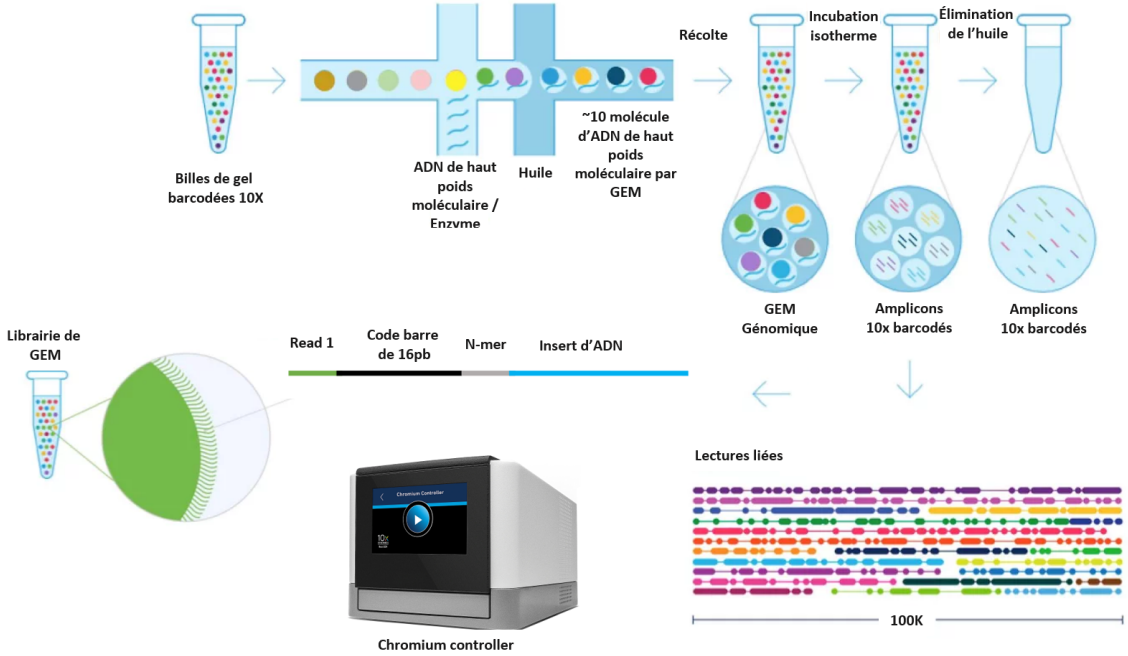


FIGURE 1.27 – Principe des lectures liées Chromium 10x Genomics.
GEM : Billes de gel dans l'émulsion (*Gel Bead-in-Emulsion*).
Adapté d'après *A Technical Note: An Introduction to Linked-Read Technology for a More Comprehensive Genome and Exome Analysis*

Une préparation standard de librairie Chromium contient jusqu'à 4 millions de codes-barres uniques et ceux-ci sont partagés par environ 1,4 million de GEM. Chaque GEM contient en moyenne 10 molécules

d'ADN. Les codes-barres 10x d'une longueur de 16 paires de bases sont conçus de manière à s'hybrider après la première lecture de la paire, puis d'être suivi d'un hexamère aléatoire, puis par l'insert d'ADN.

Les fichiers obtenus par le séquençage de librairies Chromium doivent être démultiplexées et *basecall* par un outil propriétaire appelé Long Ranger(https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger). Ce dernier est un outil propriétaire "boîte noire" qui permet le démultiplexage des données obtenues, ainsi que diverses opérations comme l'alignement des données contre un génome de référence ou l'appel de variation à partir de fichier d'alignement.

Même si le postulat de départ de la technologie Chromium 10X Genomics *linked reads* était plutôt intelligent, la technologie n'a malheureusement eu que peu de succès dans la communauté scientifique. Il y a plusieurs raisons à cela, tout d'abord, le fait que la technologie de barre code soit propriétaire et que le code de Long Ranger ne soit pas open source a très fortement impacté l'engouement de la communauté bioinformatique dans le développement d'outils pour cette technologie. Des données produites avec cette technologie sont ainsi difficilement exploitables, hormis avec les outils fournis par 10X Genomics. Deuxièmement, bien que supposée plus performante que le séquençage *short-reads* simple, le peu d'informations disponibles et le peu de données disponibles publiques produites à partir de cette technologie limitent sa compréhension et son utilisation. La technologie Chromium 10X Genomics *linked reads* a été abandonnée par son fabricant le 30 juin 2020[50], bien que l'entreprise reste active dans le domaine du séquençage *single cell*, mais aussi de la protéomique et de la transcriptomique spatiale.

1.4 Approches Bioinformatiques pour traiter les données issues de séquençage Illumina

1.4.1 *Basecalling* et démultiplexage

Le *basecalling* ou appel des bases est l'étape qui permet de traduire les données acquises lors d'un séquençage, en nucléotides pour l'ensemble des lectures, la séquence n'étant "lue" qu'indirectement par le séquenceur[51]. Les méthodologies diffèrent en fonction des technologies de séquençage.

Les séquenceurs Illumina enregistrent des images au format BCL de chacun des cycles d'incorporation des nucléotides fluorescents et de l'émission du signal lumineux pour chaque *cluster*. Les signaux émis au niveau des *clusters* sont caractérisés par le moment de leur émission, l'intensité de leur fluorescence et le bruit de la mesure. La plupart des outils de *basecalling* reposent sur des algorithmes de HMM. Ces derniers ont pour rôle de déterminer une matrice de transition indiquant le nombre et le type de bases incorporées au sein du *cluster* à chaque cycle[51]. En temps normal seule une base est incorporée, mais il est possible qu'aucune (*phasing*) ou que deux bases (*pre-phasing*) soient incorporées. Chaque base appelée l'est avec un score de qualité associé représentant la fiabilité de l'appel.

Le *basecalling* des données issues de séquençage Illumina représente communément la première étape de nombre de pipelines bioinformatiques. L'outil consensus utilisé par la communauté, bcl2fastq, *wrapper* (script ou outil permettant l'exécution) de CASAVA[52], permet également le démultiplexage des données, c'est-à-dire, l'assignation des lectures à l'échantillon correspondant. Lors de la préparation de la librairie de séquençage pour un échantillon, des séquences index spécifiques à l'échantillon sont hybridées à chaque lecture. Les différentes librairies de patients sont ensuite mélangées en quantités supposées égales afin d'obtenir une librairie multiplexée, ce qui permet de séquencer plusieurs individus ou échantillons en même temps. Afin de démultiplexer correctement lors du *basecalling*, un document (*samplesheet*) récapitulant les paramètres du séquençage (nombre de cycles, tolérance à l'erreur pour l'assignation des bases) et les couples de séquences index, échantillon sont fournis à bcl2fastq. Lors du *basecalling*, l'outil reconnaît aux extrémités des *reads* les index et ainsi assigne chaque lecture propre à son patient en y rognant les séquences des adaptateurs et des index.

Aujourd'hui, la majorité des données de séquençage produites avec un séquençage Illumina nécessitent un *basecalling* indépendant de l'exécution du séquenceur sur des serveurs informatiques. Néanmoins, les séquenceurs les plus récents Illumina embarquent des serveurs DRAGEN permettant d'effectuer le *basecalling* en même temps que le séquençage sans intervention de l'utilisateur. Le principe du *basecalling* pour les données issues de séquençage ONT a déjà été décrit précédemment. Le *basecalling* des données PacBio est directement fait par le séquenceur pour chaque *SMRT cell*. Ces dernières ne permettent pas de multiplexage d'échantillons au sein d'une même *flowcell*.

Le résultat du *basecalling* et du démultiplexage est un fichier au format FASTQ, format usuel de départ de nombre de pipelines bioinformatiques, regroupant une part ou la totalité des reads produites lors d'un *run*. Un *read* est composé de trois parties principales, l'en tête (*header*) qui commence toujours par le caractère "@", la séquence d'acides nucléiques et la séquence de qualité de l'appel de chaque base, voir *Snippet 1.1*. Le format FASTQ n'a aucune spécification officielle, il existe donc des variantes en fonction des générations de séquenceur et des constructeurs.

Snippet 1.1 – *Read* issu d'un fichier FASTQ pairé obtenu avec la version 1.8 de CASAVA.

```
1 @NB501194:1048:H3CLJBGXG:1:11101:9245:1056  2:N:0:TTAACGCAGA+GGCGGATCAA
2 GCATCATGTGGTCTCAGCGTGATACATCACTTCGCAAAAGTGGAGTAGGCAACATATTCATTAAAAATCTGG
3 +
4 AAAAAEEEEEEEEEEEEEEEEE/EEEEEAAEAE<EEEEEEEEEE//EAEEE/EEAEEEEE/EEAEEAE/E
```

NB501194	Nom unique du séquenceur
1048	Identifiant du projet (run)
H3CLJBGXG	Identifiant de la flowcell
1	Numéro de la piste (*lane*) de la flowcell
11101	Numéro de la zone (tile) au sein de la *lane*
9245	Coordonnée x du *cluster* sur la *tile*
1056	Coordonnée y du *cluster* sur la *tile*
2	Numéro du membre de la paire de lectures, 1 ou 2 (paired-end)
N	Passage du filtre, Y (yes) indique un read de mauvaise qualité, N (no)
0	0 lorsqu'aucun des bit contrôles n'est activé, sinon c'est un int
TTAACGCAGA+ GGCGGATCAA	Index (étiquette nucléotidique ou tag) de la séquence pour le démultiplexage

TABLE 1.6 – Différents champs composants l'en tête d'un *read* obtenu avec la version 1.8 de CASAVA.

L'en-tête est composé de plusieurs informations qui composent le nom du *read*, voir *Table 6*. La ligne suivant l'entête est celle comportant la séquence de la lecture. Enfin la séquence de qualité représentant le score de qualité de l'appel de chacune des bases est séparée du reste par le symbole "+". La séquence de qualité est encodée en symboles de l'alphabet ASCII décalés de 33 (car les 32 premiers caractères représentent des caractères de contrôle), le 33^e caractère signifiant alors une qualité de 0. Ce score de qualité est nommé score *Phred* ou *Phred+33*. D'autres encodages du score *Phred*, notamment avec un alphabet réduit existent pour diminuer la taille des fichiers FASTQ.

1.4.2 Alignements sur génome de référence

Le séquençage nous permet d'obtenir des millions de lectures fragmentées à partir du matériel génétique de base utilisé pour la préparation de librairie, mais pas de déterminer de quelle partie les lectures proviennent. Cette étape est souvent comparée à la résolution d'un puzzle dont les lectures en seraient les pièces et le génome le modèle à reconstruire, leurs motifs représentant les caractéristiques génomiques des lectures et du génome lui-même.

C'est grâce à des outils et algorithmes bioinformatiques qu'il est possible d'exploiter le potentiel des fichiers format FASTQ produit lors du *basecalling*. Deux analyses principales existent à partir de ces fichiers, l'assemblage et l'alignement (*mapping*). L'alignement constitue généralement la première étape d'analyse de données bioinformatiques lorsqu'un génome de référence est disponible, car bien plus rapide et moins coûteux en ressources informatiques que l'assemblage. De sa bonne exécution sont conditionnés les résultats des étapes suivantes. L'alignement consiste à retrouver la position propre de chaque lecture par rapport à un génome de référence, voir *Figure 1.28*. Les algorithmes d'alignement diffèrent en fonction des caractéristiques des *reads* à aligner.

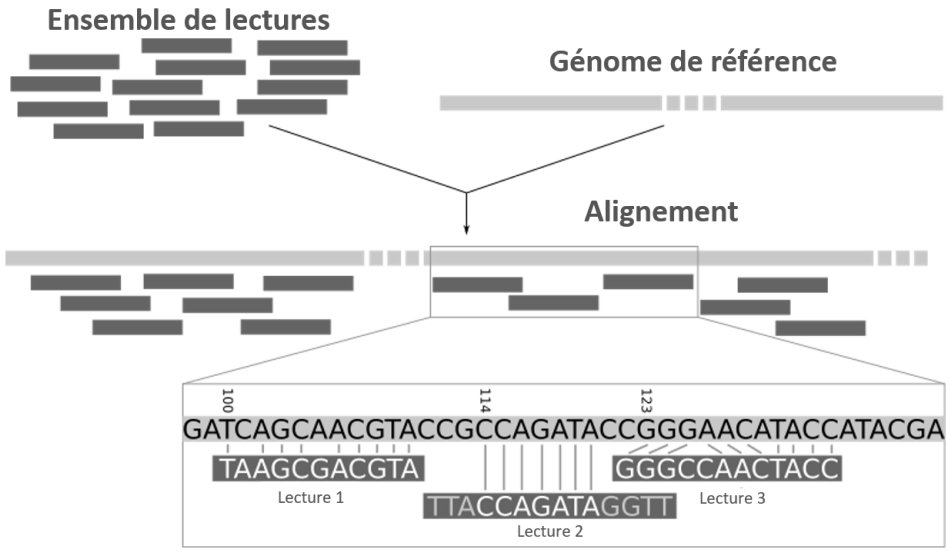


FIGURE 1.28 – Illustration du processus d'alignement ou *mapping*.
Adapté d'après *Galaxy Project mapping tutorial*.

La plupart des outils d'alignement de lectures courtes fonctionnent selon le principe du *seed and extend* (chaînage et extension), voir *Figure 1.29*. Les graines (*seed*) correspondent à des mots (enchaînement de lettres) ou k-mer de taille variable (défini par l'utilisateur, par défaut ou de taille arbitraire selon l'outil) qui vont être recherchées au sein du génome de référence au format FASTA et de ses index. Lorsqu'il y a correspondance, la graine est alors étendue en fonction d'une matrice de score, le but étant de maximiser le score tout en étendant la graine jusqu'à un maximum de la taille du *read*, voir *Figure 1.29*. L'extension de la graine peut autoriser des erreurs dans le cas où cela contribuerait à maximiser le score.

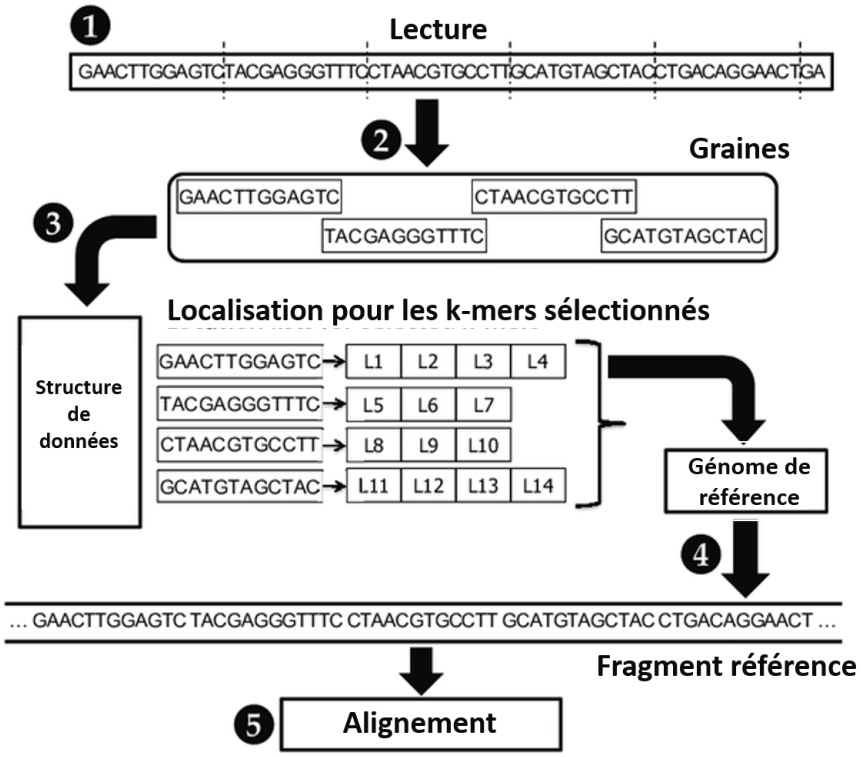


FIGURE 1.29 – Principe du *seed and extend* pour l'alignement de lectures contre un génome de référence. Adapté d'après *Kim, J. S., Cali, D. S., Xin, H., Lee, D., Ghose, S., Alser, M., ... Mutlu, O. (2018). GRIM-Filter : Fast seed location filtering in DNA read mapping using processing-in-memory technologies. BMC genomics, 19(2), 23-40.*

Les outils développés pour les lectures courtes peuvent être classifiés en deux types principaux, selon leur stratégie pour trouver et étendre les graines. Les premiers comme MAQ[53] utilisent des tables de hachage et des arbres de préfixe/suffixe pour indexer les lectures et permettre de retrouver la position propre des lectures sur le génome[54]. Les seconds, comme BWA[55] implémentent les arbres de préfixe/suffixe en utilisant un FM-Index basé sur la Transformée de Burrows–Wheeler (BWT), une structure d'indexation permettant d'indexer le génome et ainsi permettre un alignement rapide à partir de séquences courtes, lorsqu'elles ne divergent pas trop du génome de référence.

Pour les lectures issues de séquençage génomique Illumina récent, le consensus s'est fait autour de l'outil BWA-MEM[56], évolution de l'algorithme initial de BWA qui implémente le principe de *supermaximal exact matches* (SMEMs) permettant de trouver les graines les plus longues possibles pour chacun des *reads*. Avec le développement des technologies *long-reads* c'est posé la question du développement d'outils

et d'algorithmes pour être adaptés à leurs profils. Les algorithmes développés précédemment ne peuvent pas directement être appliqués sur des données de troisième génération. Les outils de *mapping short-reads* effectuent pour la plupart un alignement sans ou avec peu d'erreurs (*mismatch*) autorisées (en général au maximum une seule), du fait de leur haute précision à la base. Les données de troisième génération et notamment ONT présentent un taux d'erreur à la base assez élevé, car les indels artéfacts y sont fréquents.

L'outil qui fait consensus pour l'alignement des données ONT est Minimap2[57], outil successeur de BWA-MEM. Minimap2 intègre le même principe de *seed-chain-align*. Il indexe les graines de la référence avec une table de hachage. Ces graines de longueur fixe sont moins efficaces en termes de spécificité par rapport aux graines de longueur variable en théorie, mais peuvent être calculées beaucoup plus efficacement en pratique et sont beaucoup plus adaptées à des *reads* qui contiennent du bruit.

N° de colonne	Nom	Description
1	QNAME	Nom du *read* ou de la paire
2	FLAG	Drapeau décrivant l'alignement
3	RNAME	Nom de la séquence de référence
4	POS	Position sur la référence du début de l'alignement
5	MAPQ	Score de qualité de l'alignement
6	CIGAR	Chaîne de caractère CIGAR
7	MRNM	Nom du second *read* de la paire
8	MPOS	Position sur la référence du début de l'alignement du second *read* de la paire
9	ISIZE	Longueur inférée de la taille de l'insert
10	SEQ	Séquence de la référence à la position du *read*
11	QUAL	Score qualité (*Phred*) du *read*

TABLE 1.7 – Descriptions des différents champs d'un fichier au format SAM.
smallAdapté d'après, *Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.*

Que les données d'entrées soient issues de séquençage *short* ou *long-reads*, l'alignement produit des fichiers tabulés au format SAM[58] (*Sequence Alignment Map*), voir *Table 1.7*. Ce format récapitule pour chaque *read* aligné, la position et la nature de l'alignement, notamment grâce à la chaîne de caractère CIGAR (*Concise Idiosyncratic Gapped Alignment Report*). Ces fichiers sont facilement manipulables notamment grâce à l'outil SAMtools[58] basé sur la librairie HTSlib[59]. Généralement, les fichiers SAM sont compressés en fichiers au format BAM (*Binary Alignment Map*), puis indexés. Cela permet de stocker l'information des fichiers SAM sans perte d'information en économisant de l'espace, mais aussi de pouvoir les requêter rapidement grâce à un index. La plupart des outils les plus populaires pour la manipulation de fichiers BAM et SAM sont basés sur cette même librairie HTSlib, spécialisée dans la lecture et l'écriture de données HTS[59]. D'autres formats compatibles avec cette librairie existent pour le stockage de données d'alignement, notamment le format CRAM[60].

1.4.3 Appel de variations

L'appel de variations ou *variant calling* est l'étape qui consiste à déterminer les différences entre l'alignement pour un échantillon donné et le génome de référence utilisé. Toutes les variations détectées ne sont pas des variants d'intérêt pathologique, mais pour la plupart des polymorphismes. Certaines erreurs de séquençage, inhérentes à la technologie, peuvent également être détectées comme des variations. Ces erreurs peuvent être dues aux cycles PCR nécessaires à la préparation des librairies, aux étapes d'amplification en amont du séquençage (comme pour la technologie Illumina), aux propriétés intrinsèques des génomes (régions répétées) ou aux erreurs de *basecalling*. Les algorithmes et outils de détection de variant (*variant callers*), diffèrent en fonction du type de variations recherché et de la méthode utilisée.

Pour l'appel de variations, deux grands types de méthodes existent. Tout d'abord, les méthodes probabilistes utilisent des modèles statistiques (bayésien la plupart du temps) pour estimer la probabilité de chaque génotype possible (homozygote référence, hétérozygote et homozygote variant) à chaque position en fonction des variations recherchées. Les génotypes sont estimés en prenant en compte les différents biais pouvant introduire du bruit dans les données à partir d'informations telles que la couverture, la qualité et le nombre de *reads* variants ou références à une position donnée. C'est la stratégie de l'outil consensus pour le SNV *calling* constitutif, GATK[61] Haplotypecaller[62] développé par les équipes du *Broad Institute* et dont le principe est visible Figure *Figure 1.30*.

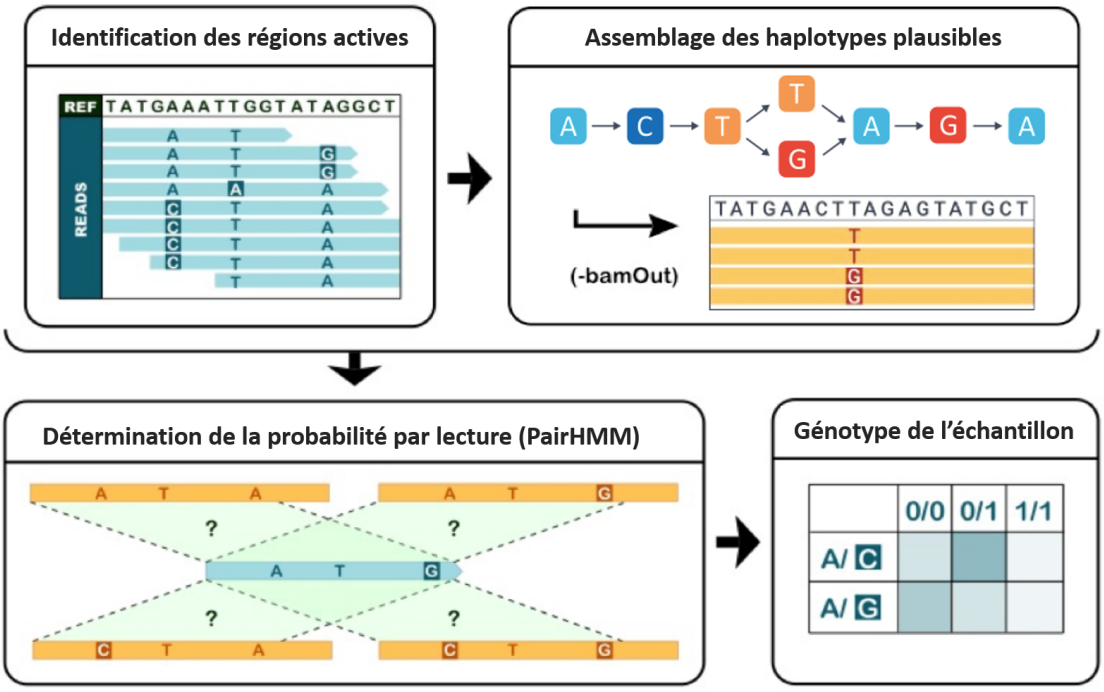


FIGURE 1.30 – Principe du *variant calling* et de la méthode de l'outil GATK Haplotypecaller.
Adapté d'après *Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., ... Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv, 201178.*

Les méthodes heuristiques, elles, utilisent des seuils déterminés arbitrairement ou empiriquement pour détecter ou non les variants. Ces méthodes sont peu utilisées par les méthodes de SNP *calling*, en revanche elles le sont par les outils de SV *calling* comme l'outil Sniffles[63]. Ces derniers exploitent les informations de score de qualité, couverture ou encore fréquence allélique, mais également d'autres signatures des lectures alignées propres aux variants structuraux, voir *Figure 1.31*. Les algorithmes heuristiques sont en général dépréciés dans les applications diagnostiques où la reproductibilité du résultat est recherchée au-delà de la performance de l'appel.

En général, les CNV sont détectés grâce à l'information de la profondeur de couverture (DoC pour *Depth of Coverage*) à une position donnée. En excluant les biais de captures et de séquençage, les chutes et augmentations soudaines maintenues de la couverture sont souvent dues à des variations de nombre de copies. Un *read* aligné est fendu (SR pour *Split-Reads*), lorsqu'une partie est alignée sur une région génomique et que l'autre est alignée ailleurs ou n'est pas alignée. La présence de *split-reads* évoque la présence d'un événement génomique à la position où les lectures s'alignent. Enfin, dans le cas du séquençage pairé, lorsqu'une paire est discordante (RP pour discordant *Read-Pair*), c'est-à-dire que seul un *read* de la paire est aligné ou que les deux sont alignés, mais que la valeur de la distance de l'insert n'est pas contenue dans la distribution normale de la librairie, alors, il y a suspicion de variation structurale.

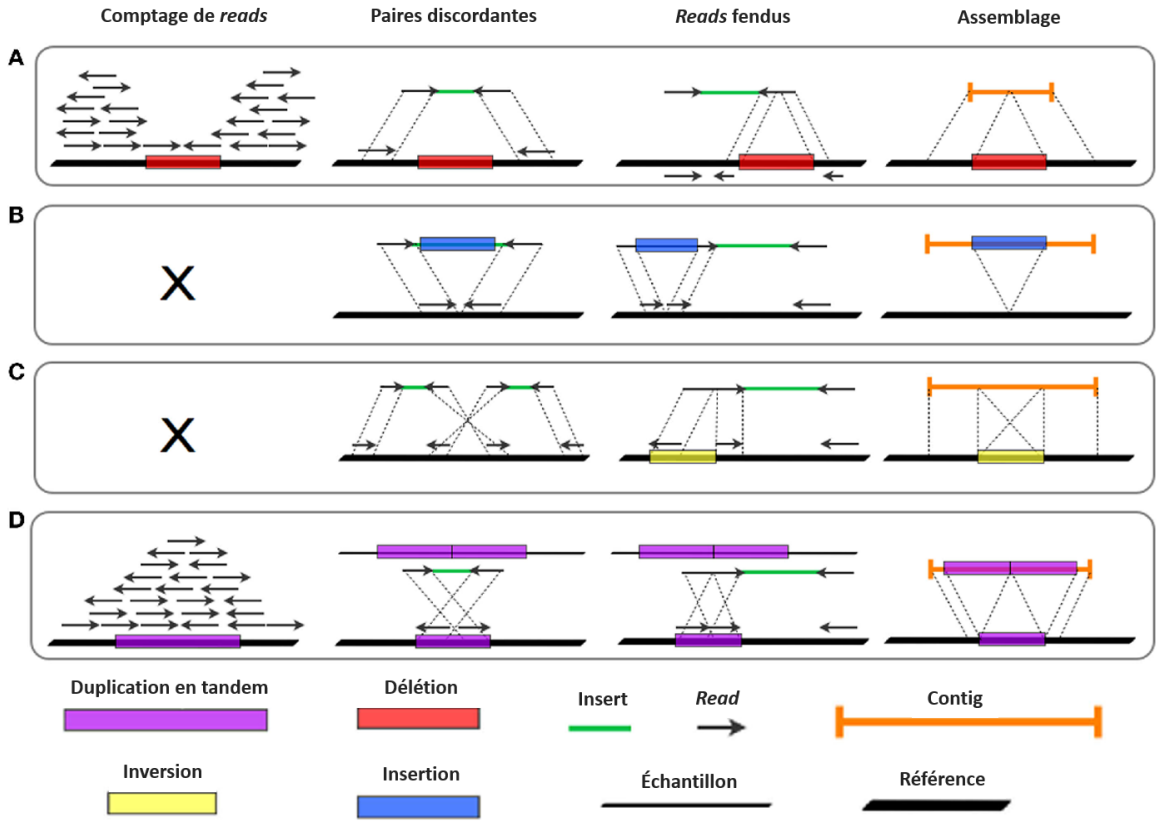


FIGURE 1.31 – Méthodologie de détection des variations de structure, comptage de lectures, paires discordantes, lectures fendues et assemblage.
Adapté d'après *Tattini, L., D'Aurizio, R., Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. Frontiers in bioengineering and biotechnology, 3, 92.*

Peu importe l'outil utilisé, dans la plupart des cas, le fichier obtenu après l'appel des variants est au format VCF[64] (*Variant Calling Format*). Ce format de fichier correspond à un standard international de description de variations génétiques. Le format VCF est un format de fichier tabulé récapitulant à chaque ligne un variant, son type, son génotype et diverses informations de qualité, voir *Table 1.8*. C'est ce format de fichier qui est le support de l'annotation.

N° de colonne	Nom	Description
1	CHROM	Nom du chromosome sur lequel le variant est détecté
2	POS	Position de la variation sur la séquence
3	ID	Identifiant de la variation (identifiant dbSNP24 si connue)
4	REF	Séquence de référence sur laquelle la variation est détectée
5	ALT	Liste des allèles alternatifs à la position
6	QUAL	Score de qualité associé à l'inférence des allèles donnés.
7	FILTER	Drapeau indiquant quel set de filtration ont été appliqués
8	INFO	Liste de champs clefs valeurs décrivant la variation
9	FORMAT	Liste de champs décrivant la variation (génotype, balance allélique, profondeur allélique...)
+	SAMPLE(S)	Si plus d'un échantillon, redite des valeurs du champ format pour l'échantillon donné

TABLE 1.8 – Descriptions des différents champs d'un fichier au format VCF.
Adapté d'après, *Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.*

1.4.4 Annotation, filtration et priorisation des variations

L'appel de variations détecte un nombre énorme de variations entre l'échantillon et la référence, environ 30 000 pour un WES et 4 à 5 millions pour un WGS, soit environ 1 SNP toutes les 1 kb[65][66]. Discriminer les variations cliniquement pertinentes du reste consiste à établir un lien entre le phénotype du patient et l'information médicale disponible sur les variations détectées. Pour cela, il faut rajouter de la méta-information à ces dernières, les annotations, ce qui permettra l'établissement de critères de filtration pour l'interprétation. Les annotations employées correspondent aux standards de la littérature et des pratiques internationales pour la classification des variations génétiques proposées par l'ACMG et relayées en français par le réseau national NGS-diag[67].

Suite à la filtration, les variations sont priorisées (ordonnées) en fonction de la concordance entre phénotypes clinique et la connaissance médicale des variations détectées pour faciliter l'interprétation[68], voir principe de la priorisation *Figure 1.32*. Les connaissances cliniques sur les maladies rares sont contenues dans des bases de données gérées par des experts, dans la littérature scientifique évaluée par des pairs et accessibles *via* le biais de plateformes communautaires de partage d'informations entre praticiens experts.

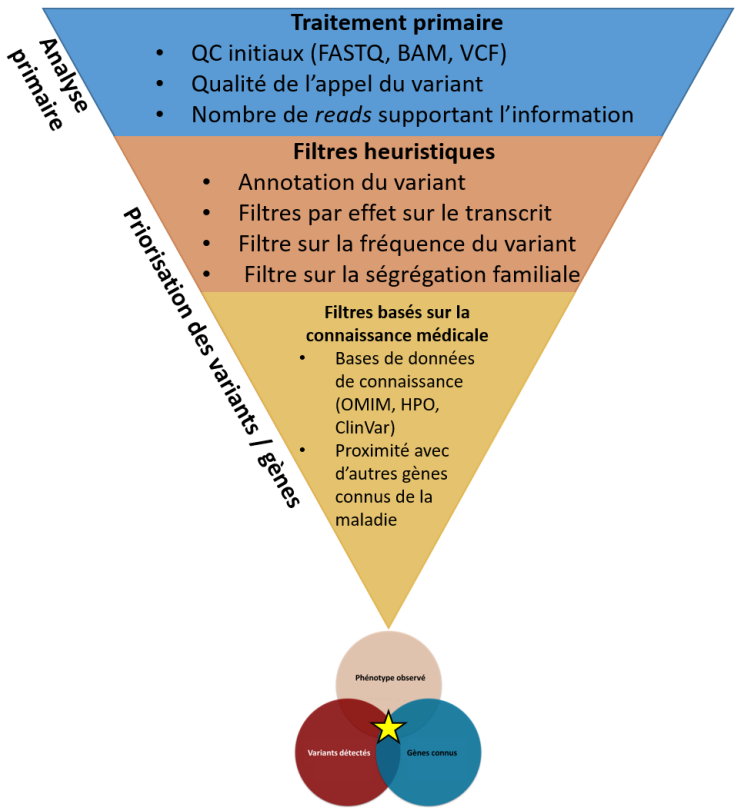


FIGURE 1.32 – Principe de la filtration et priorisation des variations pour obtenir une variation candidate pour le diagnostic.

Un variant peut être annoté à plusieurs niveaux[69]. Au niveau du variant lui-même :

— Les informations propres à l'appel de variation, comme le génotype, la profondeur de couvertures des différents allèles ainsi que la qualité de l'appel de variant.

— La fréquence des variations dans la population générale. Pour cela, plusieurs bases de données sont disponibles, celle du 1KGP[17], celle de l'*Exome Aggregation Consortium*[65] (ExaC) qui regroupe les données de 60 706 exomes ainsi que celle de la *Genome Aggregation Database*[70] (gnomAD) qui regroupe les données de 125 748 exomes et 15 708 génomes (version *2.1*). Le contexte médical de cette thèse étant les maladies rares, ces bases de données permettent d'éliminer les variations présentes à plus de 1% de la population, hormis quelques exceptions qui sont relativement circonscrites.

— L'impact médical de la variation sur le transcrit ou la protéine sous-jacente. Les variations sont classées en fonction de la gravité de son effet estimé ou validé par des pairs. C'est notamment le cas des bases de données ClinVar[71] et *Matchmaker Exchange*[72] qui regroupent des informations soumises et validées par la communauté scientifique sur les variations génomiques et leur relation avec certains phénotypes cliniques.

Des outils standards permettent d'annoter les variations génétiques détectées, tels que SnpEff[73], SnpSift[74] et le *Variant Effect Predictor*[75](VEP) du projet Ensembl[10]. En plus des catégories d'annotations mentionnées ci-dessus, ces logiciels permettent d'ajouter des annotations fonctionnelles aux variants dé-

tectés et d'ainsi prédire leur effet sur le transcrit ou la protéine sous-jacente tels que les scores Polyphen2[76], CADD[77] ou la pLI[65].

Au niveau du ou des gènes touchés par la variation. Lorsque le gène a été décrit comme relié à un phénotype clinique, lorsqu'il est altéré, il est primordial de prioriser les variations qui le touchent. Plusieurs bases de données reliant l'information génique au phénotype clinique existent. Tout d'abord, la base de données OMIM[78] (*Online Mendelian Inheritance in Man*) récapitule des phénotypes de maladies pour lesquels les bases moléculaires sont connues. Cela inclut, les troubles et traits mendéliens monogéniques, les susceptibilités au cancer et aux maladies complexes (par exemple, BRCA1 et la susceptibilité de cancer du sein et de l'ovaire familial ou CFH et la dégénérescence maculaire), les variations qui conduisent à des valeurs anormales, mais bénignes des tests de laboratoire et des groupes sanguins (par exemple, le déficit en lactate déshydrogénase B et le système de groupes sanguins ABO) et certaines maladies génétiques somatiques cellulaires (par exemple, GNAS et le syndrome de McCune-Albright et IDH1 et le glioblastome multiforme). De la base de données OMIM a été dérivé le projet d'Ontologie du Phénotype Humain[79] (HPO pour *Human Phenotype Ontology*). Comme son nom l'indique, la base de données HPO est une ontologie qui relie les différents termes cliniques, plus ou moins spécifiques en fonction de leur position dans l'arbre, à différents gènes causaux. HPO contient actuellement plus de 13 000 termes et plus de 50 000 annotations entre phénotype et diverses maladies héréditaires.

La priorisation des variants qui découle de l'annotation a pour but de croiser les informations phénotypiques, avec celles des variations détectées et de la connaissance médicale à propos de cette dernière. Lorsqu'une variation détectée concorde, alors elle est une variation candidate pour le diagnostic et devra être étudiée attentivement par l'interpréteur, voir *Figure 1.32*.

Le nombre de variations détectées par WGS avant priorisation en fonction de leur type et de la région génomique où elles ont lieu est visible *Table 1.9*. Environ 10 000 variations silencieuses et 10 000 variations non-synonymes sont attendues dans les zones codantes de l'exome. À cela s'ajoutent environ 10 000 variations au niveau des jonctions exons-introns qui ne sont pas représentées *Table 1.9*. Avant priorisation, un exome contient donc environ 30 000 variations de petite taille.

Table 2. SNPs Identified through Whole-Genome Sequencing of DNA from the Proband.*

SNP Type	No. of SNPs
Nongene	2,255,102
Gene	1,165,204
Intron	1,064,655
Promoter	60,075
3′ UTR	16,350
5′ UTR	3,517
Splice regulatory site	2,089
Splice site	112
Synonymous	9,337
Stop→stop	17
Nonsynonymous	9,069
Stop→gain	121
Stop→loss	27
Total	3,420,306

TABLE 1.9 – Type de variants de petite taille détectés par WGS en fonction de la région génomique. Adapté d'après, *Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L., ... Gibbs, R. A. (2010). Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. New England Journal of Medicine, 362(13), 1181-1191.*

À partir de ces 30 000 variations exoniques, 500 demeurent après l'application de filtres sur la fréquence ainsi que sur les conséquences estimées. Après filtration sur la connaissance médicale, il ne reste alors plus qu'environ 80 variations. Parmi celles-ci, seules 30 variations sont en lien possible avec le phénotype patient. Enfin, seulement 0 à 3 variations suivent les lois de transmission mendéliennes, voir *Table 1.10*.

Table 2. Variant Categories in Clinical Reports of Whole-Exome Sequencing.*

Category	No. of Variants†
Focused report	
Deleterious mutation related to the disease phenotype	0–2
VUS related to the disease phenotype	4–9
Medically actionable mutation‡	0 or 1
Autosomal recessive carrier status§	0 or 1
Pharmacogenetic variant¶	0–4
Expanded report	
Deleterious mutation unrelated to the disease phenotype	1–3
VUS unrelated to the disease phenotype‖	17–41
Truncating mutation in genes with no known association with disease	17–25
Not included in report	
VUS unrelated to the disease phenotype in which only one mutant allele was identified in a gene associated with a recessive disorder	26–64
VUS in gene with no known association with disease	300–600

TABLE 1.10 – Nombre de variants par type détectés par WES.
Adapté d'après, *Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., ... Eng, C. M. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. New England Journal of Medicine, 369(16), 1502-1511.*

1.5 Industrialisation des procédés d'analyses bioinformatiques

Cette thèse prend place dans un contexte de partenariat privé-public entre un institut de recherche, un hôpital universitaire et un laboratoire de biologie médicale privé. Il a été convenu dès le début qu'il allait falloir suivre certains standards et bonnes pratiques de développement industriels. De plus, nous sommes dans une ère où être capable d'assurer la reproductibilité des analyses, qu'elles soient bioinformatiques ou non, est une exigence. Cette reproductibilité est assurée en Biologie par l'automatisation grandissante des analyses et de la mise en place de procédures à suivre dans le cadre des différentes activités de production. En Bioinformatique, cette exigence de reproductibilité peut être assurée par plusieurs moyens :

— Un meilleur contrôle des versions des outils et des bases de données associées exécutées lors des analyses.

— Un meilleur contrôle et traçabilité de l'exécution des analyses.

— Un cadre de développement facilitant la modification et la traçabilité de la modification du code.

— La mise en place de procédures, automatiques ou non pour le lancement et le bon déroulement des analyses.

Ces divers moyens s'appuient sur différents outils qui ont été utilisés tout au long de cette thèse, notamment, les logiciels de conteneurisation, les gestionnaires de *workflow*, les ordonnanceurs ou *jobs scheduler* et les logiciels de gestion de versions. L'ensemble de ces outils sont aujourd'hui très utilisés par la communauté scientifique pour développer des pipelines bioinformatiques, un pipeline étant *un groupe de logiciels exécutés en série de telle façon que la sortie d'un logiciel sert d'entrée pour le suivant*[80].

1.5.1 Les logiciels de conteneurisation informatique

Les logiciels de conteneurisation informatique sont des environnements dédiés, légers et portables qui permettent l'exécution d'une application et de ses librairies sur un système hôte. Les conteneurs rappellent le principe des machines virtuelles. La virtualisation est bien présente dans les deux cas, mais ce ne sont pas les mêmes composants qui sont virtualisés, voir *Figure 1.33*.

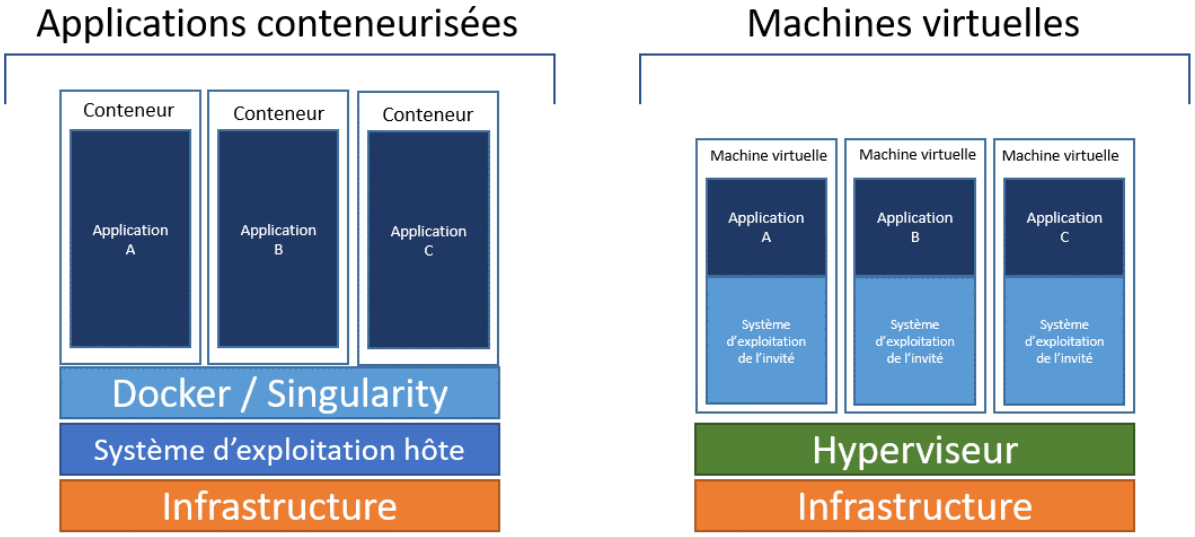


FIGURE 1.33 – Principe de la conteneurisation des applications, comparaison avec des applications embarquées dans les machines virtuelles.

Les machines virtuelles disposent de leur propre système d'exploitation en plus d'embarquer les applications et librairies nécessaires à leur bon fonctionnement. Les conteneurs utilisent le système d'exploitation de la machine hôte *via* le logiciel de conteneurisation et ne contiennent donc que des applications et librairies. Les conteneurs sont alors beaucoup plus légers, portables et facilement déployables. Ils ne nécessitent que l'installation d'un logiciel de conteneurisation sur la machine hôte afin d'être opérationnels. Les pipelines bioinformatiques à base de conteneurs nécessitent une maintenance moindre que les machines virtuelles. Si un processus utilisant plusieurs applications venait à être modifié, si jamais les applications sont contenues de manière unitaire (atomique), alors seul un conteneur du processus entier sera reconstruit. Dans le cas d'une machine virtuelle, c'est dans son entier qu'elle devra être mise à jour.

Les conteneurs sont construits à l'aide de "recettes", des fichiers textuels au format spécifique à l'outil utilisé. Ces recettes décrivent l'ensemble des étapes pour la construction du conteneur, voir *Snippet 1.2*. Une recette s'appuie sur une image de base de plus ou moins bas niveau. Dans l'exemple représenté, l'image de base contient le gestionnaire de paquet Conda préinstallé. Conda est utilisé dans ce cas-ci pour installer les outils nécessaires pour l'alignement de fichiers FASTQ contre un fichier de référence.

Snippet 1.2 – Exemple d'un fichier de recette Docker (*Dockerfile*).

```
1  ################## BASE IMAGE ######################
2  FROM continuumio/miniconda
3
4  ################## METADATA ########################
5
6  MAINTAINER  Quentin Testard <Quentin.Testard@biomnis.com>
7  LABEL description="Docker image for executing the Lygrexome pipeline bcl2fastq module"
8  LABEL tags="Genomics"
9
10 ################## CONDA TOOLS #####################
11
12 RUN conda update -n base conda \
13   conda config --add channels conda-forge \
14   conda config --add channels bioconda \
15   conda install -y bwa=v0.7.17 samtools=v1.13 gatk4=v4.1.4.1
```

Les conteneurs sont utilisés en calcul scientifique pour plusieurs raisons. Ils facilitent l'utilisation de nombreux outils au sein d'un même environnement, notamment des outils utilisant des versions de librairies incompatibles entre elles. Cela permet également de pouvoir en théorie reproduire à l'identique une analyse faite il y a plusieurs années, si les conteneurs ont été conservés. Attention, reconstruire un conteneur avec la même recette à plusieurs mois ou années d'intervalle ne garantit pas la construction de l'exact même conteneur, c'est pourquoi, on aura tendance à archiver ceux construits et utilisés dans les diverses versions de notre pipeline. Enfin, ils rendent le déploiement de pipelines utilisant des conteneurs beaucoup plus évidents, car les conteneurs sont interopérables sur une autre infrastructure à partir du moment où celle-ci possède une installation du même logiciel de conteneurisation dans une version compatible.

Les deux outils de conteneurisation les plus répandus sont Docker et Singularity. Docker est l'outil de conteneurisation le plus utilisé au monde, mais celui-ci demande des permissions utilisateur de trop haut niveau qui ne peuvent pas être permises sur la plupart des *clusters* de calcul partagés, car trop lourd à administrer. L'alternative est le logiciel Singularity qui peut transformer des conteneurs Docker en conteneur Singularity ou tout simplement en construire à partir de son propre modèle de recettes et qui ne nécessite pas autant d'administration que l'outil Docker. C'est pourquoi il est assez répandu sur les infrastructures de calcul universitaires[81].

1.5.2 Les gestionnaires de *workflow*

Historiquement, les diverses étapes qui constituaient un pipeline bioinformatique étaient contenues dans des scripts, le plus souvent en langage Bash et permettaient de lancer des outils en lignes de commandes ou d'autres scripts. Cette approche bien que simple à mettre en place n'est absolument pas robuste ni facilement maintenable à long terme. C'est pourquoi des outils spécifiques ont été développés pour tenir le rôle de "squelette" des pipelines et permettre d'organiser diverses tâches dans un ordre prédéfini en faisant

transiter les sorties des différentes étapes vers les suivantes. Ces outils sont nommés gestionnaires de flux de travail ou gestionnaires de *workflow*. De nombreux gestionnaires de *workflow* existent, certains assez généralistes et développés par de grands groupes informatiques comme luigi (https://github.com/spotify/luigi) développé par Spotify ou des efforts plus communautaires et spécialisés avec notamment Nextflow[82] et Snakemake[83]. Ces deux gestionnaires de *workflow* ont spécialement été développés pour la Bioinformatique mais peuvent très bien être utilisés de manière plus générale.

Nextflow et Snakemake sont à peu près équivalents dans leurs possibilités et utilisations. Le premier est néanmoins basé sur le langage Groovy, un dérivé du Java et le second sur Python. La plus grande différence entre ces outils est la dynamique de leur communauté. En effet, sous l'impulsion du consortium nf-core[84], ayant pour but de mettre à disposition du public des pipelines Nextflow pour des tâches bioinformatiques génériques, comme l'analyse de données de HTS pour la détection de SNV[85], ou la discrimination de souches du virus SARS-CoV-2 (Viralrecon), la popularité de Nextflow semble avoir dépassé celle de Snakemake, bien qu'aucune donnée chiffrée n'existe. Nous allons nous concentrer sur Nextflow qui a été l'outil utilisé durant cette thèse.

Nextflow permet d'enchaîner des tâches, processus ou *process* qui sont décrits au sein d'un script, voir *Snippet 1.3*. Dans l'exemple ci-joint, le *process* exécute l'outil de qualité FastQC[86] à partir de données FASTQ pairées fournies au script *via* des valeurs définies dans un fichier de configuration. L'outil produit comme résultat une archive ZIP et un fichier HTML qui seront fournis, ainsi que le dossier de sortie à un *process* ultérieur.

Snippet 1.3 – Exemple d'un *process* Nextflow, ici l'outil de mesures de qualité FastQC.

```
process fastqc {
        errorStrategy 'retry'
        maxRetries 3
        tag "$pair_id"
        publishDir "${params.resultDir}/metrics/fastqc", mode: 'copy'

        input:
        set val(pair_id), file(reads) from fastqc_files_ch

        output:
        file("*_fastqc.{zip,html}") into fastqc_results
        val("${params.resultDir}/metrics/fastqc") into fastqc_rep

        script :
        """
        fastqc -t ${task.cpus} $reads
        """
    }
```

Nextflow produit plusieurs fichiers qui permettent le suivi de l'exécution de l'analyse, dont un rapport au format HTML, traçant les différents *process* en détaillant leur exécution (temps, consommation CPU, consommation RAM, lecture/écriture...). Nextflow permet également de faire du "retour sur point". Si un *process* échoue, le pipeline s'arrête. L'utilisateur peut alors identifier la source du problème et relancer l'analyse de l'endroit où elle s'était arrêtée. Nextflow gère de manière transparente les conteneurs Singularity et Docker, permettant ainsi d'assurer la reproductibilité des analyses. Enfin, Nextflow est compatible avec plusieurs ordonnanceurs, ce qui permet de l'utiliser sur des infrastructures de calcul distribué.

1.5.3 Les ordonnanceurs ou *jobs scheduler*

Les calculs bioinformatiques pouvant être assez exigeants en matière de puissance nécessaire pour mener l'analyse à bien, il est assez courant d'utiliser des grappes de serveurs ou *cluster* de calcul. Les *clusters* également appelés supercalculateurs, sont des regroupements d'ordinateurs indépendants, appelés nœuds ou *nodes*, organisés en grappes de manière à pouvoir faire du calcul massivement parallèle (le même calcul distribué sur plusieurs nœuds) ou massivement séquentiel (plusieurs occurrences d'un même calcul sur plusieurs nœuds). Un *cluster* est en général constitué d'un ou plusieurs nœuds maîtres ou *master* qui ont le rôle de distribuer les différentes tâches qui leur sont soumises par les utilisateurs aux restes des nœuds dits esclaves ou *slave*. La constitution d'un *cluster* a plusieurs avantages. Elle permet de pouvoir proposer plus de disponibilités de temps de calcul aux utilisateurs, de mutualiser les coûts des machines, des infrastructures, de main-d'œuvre et de maintenance, ainsi que la plupart du temps, proposer une infrastructure de calcul homogène (tous les nœuds ou de grands groupes de nœuds ayant les mêmes caractéristiques).

Afin d'organiser les nœuds en *cluster*, il est nécessaire d'utiliser un outil à cette fin, un ordonnanceur. Outre le fait que d'organiser le *cluster*, la plupart des ordonnanceurs permettent entre autres, d'optimiser les ressources allouées aux différents utilisateurs, de gérer le lancement des tâches ou jobs, de fournir une interface de suivi des performances du *cluster* en temps réel, de gérer les permissions et les quotas de calcul entre utilisateurs et de garantir la disponibilité des nœuds pour les tâches réservées, tout ça, de manière automatique.

L'intégration de nombreux ordonnanceurs dans le gestionnaire de *workflow* Nextflow permet l'utilisation de pipelines standardisés sur les différents *clusters* de calcul dont l'intégralité possède des ordonnanceurs, plus ou moins populaires.

1.5.4 Les logiciels de gestion de version

L'informatique et la science en général étant de plus en plus collaborative et ouverte, il est nécessaire d'avoir un répertoire compilant l'ensemble du code et permettant la traçabilité de la chronologie des modifications du code par les différents contributeurs. Les logiciels de versions décentralisés comme le populaire Git (https://git-scm.com/) permettent cela en créant un dépôt central où seront hébergés les fichiers. Ces dépôts sont hébergés sur des sites spécialisés dans l'hébergement de projets de développement logiciel comme GitHub (https://github.com/) ou GitLab (https://gitlab.com). Une fois les dépôts clonés sur sa machine locale, il est possible pour les différents utilisateurs de travailler de manière désynchronisée, sans être forcément connecté à internet.

Lors d'une évolution, c'est-à-dire une modification du code, une nouvelle version est créée par l'outil de versionnage. Ces différentes versions, d'utilisateurs différents par exemple, peuvent être sur des branches différentes pour ensuite être réunies sur une branche principale lors des *merge*. Les fichiers ainsi que toutes leurs différentes versions au cours du temps sont conservés dans le dépôt.

1.6 État de l'art sur l'exhaustivité de la détection des variations génomique par les technologies de séquençage

Indépendamment des différents types de variations ou des différents types de technologies utilisées, la mise en place d'une nouvelle analyse nécessite de déterminer si le test permet de réaliser la mesure d'intérêt avec exactitude et fiabilité. C'est le principe de validité analytique. En général, cela passe par une caractérisation de la sensibilité, de la spécificité, de la reproductibilité, de la robustesse et de la satisfaction des contrôles qualité de l'analyse[87]. Il n'est pas possible de détecter toutes les variations dans le champ d'application de la technique, il suffit d'en connaître les limites et de détecter au minimum toutes les variations qui l'étaient par la précédente technologie de référence utilisée. S'il n'existe pas d'autre technique à ce jour, il faut démontrer que les performances analytiques d'une technologie permettent son utilisation dans un cadre de diagnostic clinique en les caractérisant[88]. Dans le cadre d'analyses de diagnostic clinique, la validité clinique plutôt que la validité analytique peut suffire, c'est-à-dire la propension du test à prédire avec précision et fiabilité le phénotype clinique d'intérêt.

Les analyses issues de technologies de seconde ou de troisième génération de séquençage ont pour visée de remplacer dans les laboratoires des tests pan-génomiques comme la puce à ADN ou des techniques ciblées comme la FISH ou la PCR *long-range*.

Les méthodologies préférées pour la validation analytique ou clinique sont la comparaison de variation de sets de vérités publiés par des institutions de référence sur des organismes *gold standards* (HG002, NA12878 ...) ou la vérification de variations pathologiques connues par méthodes orthogonales.

1.6.1 Échantillons de référence

En complément d'un génome de référence, l'existence d'échantillons d'étalonnage de méthodes est un prérequis aux applications diagnostiques. Il est d'une importance majeure de mesurer les performances de nos processus de production de données de séquençage ainsi que de nos pipelines d'analyse bioinformatiques. Pour cela, diverses initiatives ont vu le jour. Elles ont pour but de créer des ensembles (*sets*) de variations de référence standards (*gold standards*) auxquels sont comparées les variations détectées issues de séquençages d'individus de référence. L'initiative la plus connue est celle du consortium Genome in a Bottle (GIAB) hébergée par le NIST (*National Institute of Standards and Technology*) et le JIMB (*Joint Initiative for Metrology in Biology*).

Le GIAB a produit plusieurs ensembles de variations de référence à partir d'individus différents. Les plus utilisées sont les sets de petites variations (SNV, SNP et Indels) issus de plusieurs individus du 1KGP, dont l'individu NA12878 (HG001), du trio Juif Ashkénaze, HG002, HG003, HG004, dont les différents types de variations sont visibles *Table 1.11*, ainsi que du trio Chinois Han, NA24631 (HG005), NA24694 et NA24695[89]. Ces individus ont été séquencés de manière intensive, des centaines de fois, par plusieurs centres de séquençage de référence. Les données ont ensuite été compilées par le GIAB pour en publier un set de référence complet pour chaque individu. Le GIAB a également publié des coordonnées de zones dites de haute confiance qui sont en général celles sur lesquelles sont effectués les points de repère et comparaison (*benchmark*) bioinformatiques en vue de la validation ou de la qualification des processus d'analyse[90]. Ces zones ont été déterminées en intégrant les données de plusieurs technologies et représentent environ 80% des bases totales du génome humain[91]. Les appels de variants des régions de haute confiance ont tendance à inclure un sous-ensemble de variants de régions qui sont plus faciles à détecter.

Echantillons	NA12878	HG002	HG003	HG004
HC SNV	3102724	3271601	3271601	3102724
Petits indels	16667	15874	15874	16667
Insertions	293598	268387	268387	293598
Délétions	296423	287319	287319	296423
CNV >50 kb	-	5[92]	-	-

TABLE 1.11 – Différents types de variations caractérisées par le GIAB chez NA12878 et le trio Ashkénaze.
HC : haute confiance.
Adapté d'après, *https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/.*

Des sets de références de SV ont également été publiés par le GIAB[92], comme le trio Yoruba, NA19238, NA19239, NA19240, le trio Portoricain HG00731, HG00732, HG00733 et le trio Chinois Han. Ces sets de référence ne sont malheureusement pas encore aussi aboutis que ceux des petites variations.

Des initiatives indépendantes ont également vu le jour pour mettre à disposition des sets de références ou des données de séquençage de ces mêmes individus notamment séquencés avec des technologies de troisième génération[47][93].

1.6.2 SNV et indels

Plusieurs jeux *gold standard* ont été publiés au cours du temps pour caractériser les performances analytiques de la détection des SNP et indels des technologies HTS[89][92][94][95][96]. La quasi-totalité des sets de variations de petite taille ont été produits avec diverses générations des technologies Illumina, depuis plus de dix ans, ce qui peut entraîner l'apparition d'artéfacts entre technologies, voire au sein de plusieurs générations d'une même technologie.

La validation analytique d'un processus de détection de SNV comprend généralement la préparation de la librairie à partir d'un ou plusieurs ADN de référence, le séquençage et l'analyse bioinformatique. Une fois la liste de variations détectées par le pipeline bioinformatique obtenue, celle-ci est comparée à l'ensemble de variations de vérité édité par l'organisme de référence avec des outils comme Hap.py[97]. Cette méthodologie de validation permet de caractériser les performances analytiques (rappel, précision) sur ces organismes. En revanche, ces performances ne sont pas généralisables sur l'ensemble des échantillons qui seront traités par le pipeline bioinformatique, beaucoup des outils d'appel de variants étant calibrés sur ces organismes pour obtenir de bons résultats lors des *benchmarks*. De plus, la majorité de ces *benchmarks* sont effectués sur des

zones dites de haute confiance qui représentent un sous-ensemble de variation et de zones plus simples à explorer lors d'un appel de variant. Les zones de haute confiance représentent environ 80% du génome total et les performances attendues sur ces zones sont de l'ordre de 99% de rappel pour les SNP. En revanche, les scores attendus sur les zones à l'extérieur de ces zones de haute confiance sont de l'ordre de 76.5%[97].

Les performances de détection de variation de petite taille sont hautement influencées par le contexte génomique dans lequel elles se trouvent, par leur taille, mais également par la préparation de la librairie et notamment la présence ou pas d'une étape d'amplification lors de la préparation de la librairie, voir *Table 1.12*.

Contexte génomique	Type	Sensibilité (sans PCR)	Sensibilité (avec PCR)	Précision (sans PCR)	Précision (avec PCR)
Tout	SNV	98.4	98.4	86.0	86.0
	Indel	85.9	97.1	59.0	56.3
Pas dans des homopolymères ou STR	SNV	98.6	98.5	87.7	87.8
	Indel	98.3	98.4	75.4	75.5
Dans des homopolymères ou STR	SNV	95.6	97.2	61.5	60.7
	Indel	78.2	96.4	50.3	48.4
Pourcentage en GC >85%	SNV	84.7	94.4	50.4	49.2
	Indel	73.2	97.3	27.9	27.0
Toutes les bases dans des STR de dinucléotide	Indel	45.3	80.9	22.3	27.6
Toutes les bases dans des STR de dinucléotide AT	Indel	12.3	28.1	4.3	6.8

TABLE 1.12 – Rappel et précision stratifiés par contexte génomique et type de variations.
Données issues de séquençages WGS Illumina avec et sans étape PCR, comparés aux régions de haute confiance du GIAB v. 3.3.2 (STR : *Short Tandem Repeat*, Courtes répétitions en tandem).
Adapté d'après, *Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., Francisco, M., Moore, B. L., ... Zook, J. M. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nature biotechnology, 37(5), 555-560.*

Il est considéré pour le séquençage HTS Illumina, que les SNP sont détectés avec précision à partir du moment où les zones où ils se trouvent sont assez couvertes (environ 20X), à l'exception des zones problématiques pour la technologie (zones répétées ou pseudogènes). C'est en général le cas pour le WGS mais est plus délicat pour le WES à cause des biais de capture et d'amplification[98], voir *Table 1.12*. Si certaines zones sont trop couvertes, alors de faux positifs peuvent également apparaître diminuant la précision de la méthode[99]. Les indels eux sont plus ou moins détectables facilement en fonction de leur taille, leur complexité, mais aussi de la taille des *reads* séquencés.

Il n'existe pas de méthode orthogonale permettant la détection de SNP et indels pan-génomique autre que les technologies de HTS. Il est préconisé en plus du *benchmark* contre les organismes de références de valider cliniquement des résultats obtenus par des résultats antérieurs également obtenus par HTS (s'il y en a), à l'aide de cas cliniques. En effet, les variations des sets de référence étant des polymorphismes (car représentant la quasi-totalité des variations d'organismes de référence supposés sains), cette méthodologie n'a aucune pertinence en ce qui concerne la validité clinique.

1.6.3 SV et CNV

Les sets de variation de références de SV et CNV existent, mais sont moins nombreux et moins éprouvés que les ensembles de variations publiés sur les variations de petite taille[92][93][100]. Les sets de références de variations structurales se reposant seulement sur des données de séquençage *short-read* n'ont pas permis de quantifier de manière exhaustive la variabilité des variations structurales dans les organismes de référence. C'est avec l'arrivée récente de sets de vérité agrégeant des données issues de technologies de séquençage de seconde et troisième génération ainsi que de cartographie optique que des données de référence plus riches ont été produites[93]. Les performances analytiques de la détection de SV et de CNV sont, de plus, beaucoup plus dépendantes de la technologie de séquençage et d'appel de variation utilisée, chacune ayant des qualités et des faiblesses[92]. Enfin, la variabilité des SV chez la population générale entre les ethnies semble être également beaucoup plus importante que pour les SNP[17][93].

Pour l'instant, l'utilisation de ces sets de référence avec des outils de comparaison comme Truvari (https://github.com/spiralgenetics/truvari, solution utilisée par le GIAB) ou SURVIVOR[101] est possible, mais les résultats, même en effectuant seulement sur les régions de tiers 1 (l'équivalent des régions de haute confiance) est d'au mieux 85% si l'on ne sacrifie pas la précision, que ce soit avec des techniques short ou *long-reads*[92]. De plus, les sets de référence de variations entre eux ne sont pas forcément concordants. Les nombres de SV détectés varient entre individus de 10 000 à 15 000[92][93] pour les plus conservatifs, jusqu'à 25 000 pour les individus les moins séquencés et avec les techniques les plus exploratoires[102]. Les divers consortiums chargés de l'établissement de ces sets de validation travaillent sur la diminution de taux de détection de faux positifs (FDR).

De la même manière que précédemment, vu qu'il n'existe pas de méthodologie de détection pan-génomique pour tous les types de SV, les validations de méthode de détection de SV se font à partir de cas clinique. En revanche, il existe des méthodes orthogonales de référence pan-génomique pour la détection de CNV, les puces à ADN et de grands SV, le caryotype. Même si leur résolution est inférieure à celle des techniques de HTS, la validation peut s'effectuer seulement sur des variations connues validées par une technologie de référence orthogonale.

1.7 Impact des variations génétiques en maladie humaine

1.7.1 Variation en population générale

On compte une variation toutes les 1000 bases en moyenne, soit 4 à 5 millions par individu, mais la distribution de ces variations n'est pas régulière et leur impact n'est pas non plus le même. Certaines zones, en majorité intergénique, regroupent un nombre important de variations et d'autres zones contenant certains gènes, subissent une pression de sélection très importante ce qui a pour effet d'y sélectionner négativement les variations. Ces zones doivent donc représenter un intérêt critique pour la survie de l'individu jusqu'à être en âge de procréer pour subir une telle pression[65].

Une métrique, la pLI, la probabilité d'être intolérant à la perte de fonction a été proposée et mesurée à partir des données du consortium ExaC[65] comportant plus de 60 706 exomes. Une grande quantité de données ainsi qu'une bonne représentation des différentes ethnies sont nécessaires afin d'augmenter la puissance de détection des variations les plus rares dans la population générale.

3 230 gènes ont une pLI supérieure ou égale à 0,9, signifiant qu'ils sont intolérants à la perte de fonction et qu'ils subissent une pression de sélection négative. À l'inverse, 10 374 gènes ont une pLI inférieure ou égale à 0,1, signifiant qu'ils sont tolérants à la perte de fonction. La pLI est corrélé à de nombreux facteurs, mais notamment avec le nombre de partenaires d'interaction physique d'un produit génique. Les mesures de pLI élevées se sont révélées concordantes avec pratiquement tous les gènes de maladies humaines haploinsuffisants graves connus. De plus, malgré leur pLI élevée, 72% des gènes intolérants à la perte de fonction n'ont pas encore été attribués à un phénotype de maladie humaine[65].

De par leur effet délétère, les variants aux pronostics les plus graves sont également les moins fréquents, hormis quelques exceptions bien documentées dans certaines ethnies. Les données du consortium ExaC[65] mesurent en moyenne 85 variants hétérozygotes et 35 variants homozygotes ayant un effet sur la protéine sous-jacente dont 2 qui leur sont uniques (singletons) et parmi lesquels 0,14 ayant une pLI supérieure ou égale à 0,9, par individu (fréquence à moins de 1% dans la population générale).

D'autres mesures, comme les scores Polyphen2[76] ou CADD[77] (*Combined Annotation-Dependent Depletion*), sont utilisées pour prédire le caractère probablement délétère des variations. Le score pLI a pour but d'être remplacé par la LOEUF[70], issue du consortium GnomAD, successeur du consortium ExaC. La LOEUF est la probabilité de perte de fonction observée/attendue (pLoF) de la fraction supérieure (supérieur à 90%) de l'intervalle de confiance. Elle permet de corriger certains biais de la pLI, notamment sur la taille du gène étudié, mais est encore peu utilisée par la communauté médicale.

Les données du consortium GnomAD constituées (125 748 exomes et 15 708 genomes pour la version *2.1* ainsi que 76 156 génome et aucun exome pour la version *3.1*) augmentent encore la puissance de détection des variations rares par rapport aux données du consortium ExaC. Avec les méthodes actuelles, il a été estimé que le jeu de données permet de détecter les SV présents jusqu'à 0.004% dans la population étudiée supposée représentative de la population générale. Il a été estimé que 3,8 % de la population générale est porteuse d'au moins un SV autosomique rare de très grande taille ($>$ 1 Mb), dont environ la moitié (45,2 %) serait équilibrée ou complexe[103]. De plus, 0,32 % des échantillons étaient porteurs d'un SV très rare (fréquence $<$ 0,1 %) entraînant pour un gène une pLoF pour laquelle les découvertes fortuites seraient cliniquement pertinentes et dont près de la moitié (0,13 % de tous les échantillons) répondrait aux critères de diagnostic comme étant pathogène ou probablement pathogène selon les recommandations de l'*American College of Medical Genetics*[104].

Les prochains projets de séquençage massif dans les années à venir permettront sans doute de détecter un nombre de variants potentiellement actionnable du même ordre de grandeur que celui du nombre de maladies rares dans la population générale.

1.7.2 Modes de transmission génétique de maladies mendéliennes

Les différents mécanismes d'héritabilité mendéliens des maladies rares ont d'abord été décrits par Victor McKusick[105] pour décrire les modalités de transmission de maladie qui auraient pour origine une variation pathologique au sein d'un allèle de l'un ou des deux parents. Celui-ci décrit les grands types de modes de transmission des maladies suivant les lois mendéliennes de l'hérédité, voir *Figure 1.34*. Ils sont de quatre types, autosomique dominant (AD), autosomique récessif (AR), dominant lié à l'X, récessif lié à l'X.

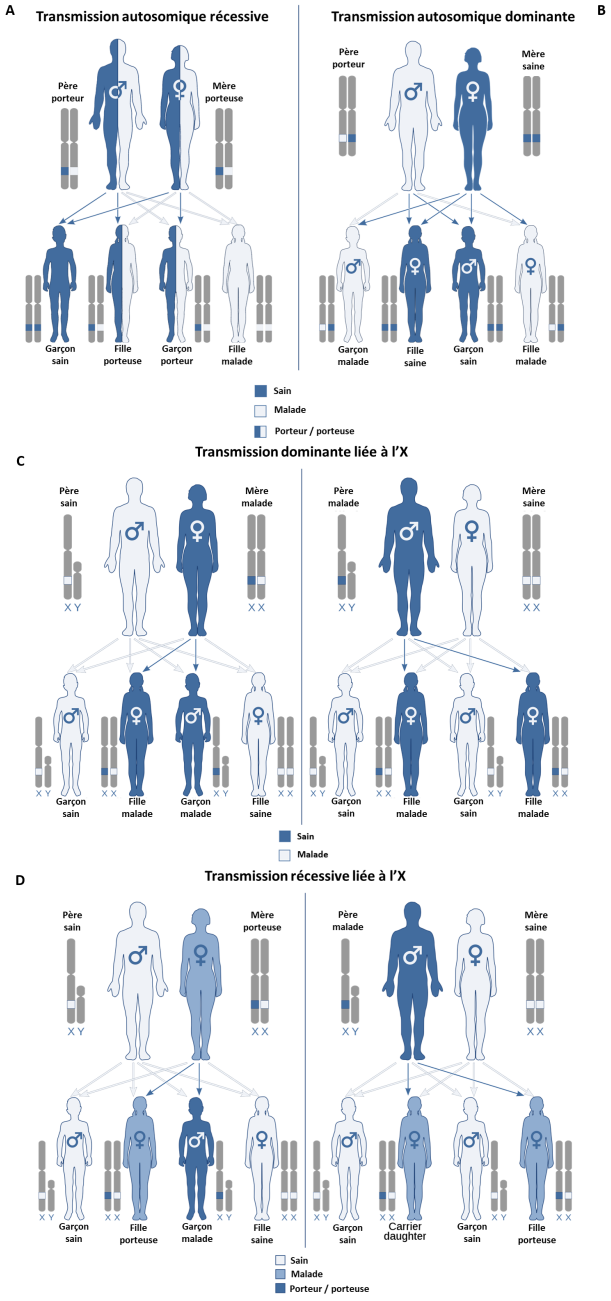


FIGURE 1.34 – Mode de transmission des pathologies régies par les lois de l'hérédité mendélienne. Adapté d'après *Wikipédia transmission mendélienne.*

Une pathologie est transmise selon le mode de transmission autosomique dominante lorsqu'un allèle muté dit "morbide" hérité d'un parent est présent sur un gène d'un chromosome autosome à l'état hétérozygote, c'est-à-dire en association avec un allèle sain (non muté) et qu'il s'exprime préférentiellement face à ce dernier. À l'inverse, la pathologie est dite autosomique récessive lorsque la présence deux allèles mutés est nécessaire pour entraîner la maladie. Les allèles mutés peuvent avoir la même variation (homozygote) ou deux variations différentes induisant une pathologie (hétérozygote composite). S'il n'y a pas de variation alors l'individu est homozygote référence.

Contrairement aux pathologies à transmission autosomique, les pathologies liées à l'X sont dépendantes du phénotype sexuel. Dans le cas de pathologies dominantes, les hommes avec un seul allèle muté (hémizygotes) et les femmes hétérozygotes sont malades. Certaines pathologies liées à l'X avec un mode de transmission dominant entraînent une mort *in utero* pour les hommes. Dans les pathologies récessives, les hommes porteurs sont également malades et les femmes hétérozygotes sont porteuses sain de la maladie.

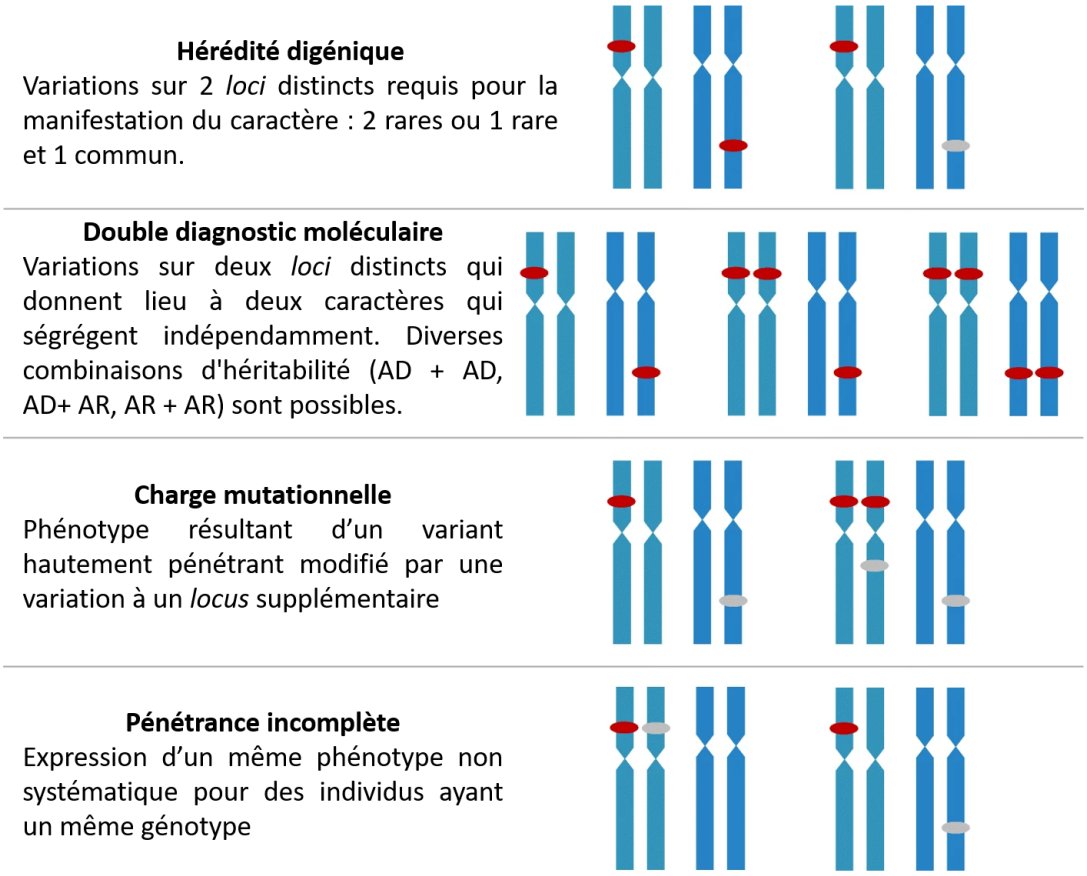


FIGURE 1.35 – Modalités de transmission complexes de certaines maladies mendéliennes.
Adapté d'après *Posey, J. E. (2019). Genome sequencing and implications for rare disorders. Orphanet journal of rare diseases, 14(1), 1-10.*

Certaines maladies suivent des modalités de transmission plus complexes que les modes d'hérédité mendéliens standard décrits initialement[106]. Elles peuvent être caractérisées par une hérédité digénique, des diagnostics moléculaires doubles, une charge variationnelle modifiée par la pénétrance des variations et une hérédité composée de variants rares et communs, voir *Figure 1.35*. La pénétrance est la proportion

d'individus qui, porteurs d'un génotype donné, vont exprimer le phénotype correspondant à ce génotype. La pénétrance d'une variation est complète lorsque tous les individus du génotype auquel elle est liée expriment le phénotype associé. Dans le cas contraire, elle est incomplète. La pénétrance peut varier en fonction de l'âge, de facteurs environnementaux et de facteurs épigénétiques. Les mécanismes exacts de la pénétrance ne sont pas encore connus, mais la plupart des études actuelles montrent des cas de "court-circuitage" de la voie mutée par un épissage alternatif[107], un différentiel de méthylation de la variation délétère menant à une réduction de sa transcription[108] ou une adaptation par la surtranscription de gènes compensatoires[109]. La pénétrance ne doit pas être confondue avec les différents degrés d'expressivité clinique du génotype aboutissant à un phénotype de diverses intensités. L'expressivité décrit la variabilité individuelle et non la variabilité statistique parmi un ensemble de génotypes.

Les individus peuvent également présenter du mosaïcisme constitutif, c'est-à-dire, la coexistence de deux ou plusieurs populations cellulaires de génotypes différents issues d'erreurs lors de la division de l'œuf fécondé. Les variations mosaïques peuvent aller de modifications d'un seul nucléotide aux altérations chromosomiques à grande échelle. Selon le pourcentage de cellules atteintes, la maladie génétique s'exprime plus ou moins intensément et sera plus ou moins facile à détecter.

Aujourd'hui le cadre d'étude dans lequel se placent le CHU de Grenoble, le laboratoire Eurofins-Biomnis ainsi que la quasi-totalité des acteurs de la génomique germinale française est celui de la recherche de variants responsables de maladies rares monogéniques, suivant les lois de l'hérédité mendélienne. Pourtant, comme vue précédemment, l'association d'une seule variation avec un phénotype particulier n'est pas toujours vraie. De plus, certaines variations ne suivent pas *stricto sensu* les lois mendéliennes d'hérédité décrites initialement, telles que les variations chromosomiques, les variations de l'ADN mitochondriales et les modifications épigénétiques. Il existe donc des limites inhérentes à cette stratégie qui sont néanmoins connues par l'ensemble des praticiens prescripteurs d'analyses génétiques.

1.7.3 Type de variation et rendement en diagnostic

Le profil type de la variation détectée en diagnostic de maladie mendélienne est une variation rare soit *de novo* (avec un modèle d'hérédité dominant), soit biallélique. Les variations bialléliques sont responsables de maladies récessives et peuvent être homozygotes dans des familles consanguines[110], voir *Table 1.13*. Cela peut s'expliquer par le fait que ces variations soient les mieux décrites dans la littérature scientifique et donc les plus faciles à identifier.

Mode d'hérédité	Nombre de cas
Autosomique dominant (n = 280)	280
De novo	208
Gène soumis à empreinte	6
Mosaïcisme	3
Mosaïcisme chez un parent	2
Autosomique récessif (n = 180)	180
SNV hétérozygote composites	104
SNV hétérozygote composites et CNV	4
Variants homozygotes (trio)	59
Variants homozygotes (singleton)	9
Variants homozygotes dus à une disomie uniparentale	5
Liés à l'X (n = 65)	65
De novo	40
Mosaïcisme	2
Désordres mitochondriaux (n = 1)	1
De novo	1
Double diagnostic (n = 23)	23
Autosomique dominant + autosomique dominant	7
Autosomique dominant + autosomique récessif	8
Autosomique dominant + lié à l'X	4
Autosomique récessif + autosomique récessif	3
Autosomique récessif + lié à l'X	1

TABLE 1.13 – Variations génétiques détectées par WES.
Étude de recherche sur 2000 patients, principalement pédiatriques ou embryonnaires, avec des troubles divers, mais majoritairement neurologiques. Taux de diagnostic de 25,2% (504/2000).
Adapté d'après, *Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., ... Eng, C. M. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. Jama, 312(18), 1870-1879.*

Le nombre de diagnostics et le type de variations identifiées varient également en fonction de la technique utilisée pour l'examen, mais aussi du type de maladie recherchée, voir *Figure 1.36* et *Figure 1.37*. En effet, comme vu précédemment, chaque technologie actuellement utilisée en clinique a ses propres forces et faiblesses. Jusqu'à aujourd'hui, à l'aube de l'utilisation du WGS en France, la plupart des diagnostics se font grâce à la combinaison de techniques de cytogénétique, de puces à ADN et de séquençages HTS. Il est donc tout à fait possible que la prévalence de certains types de variations en clinique soit sous-estimée de par la difficulté de les détecter avec ces techniques.

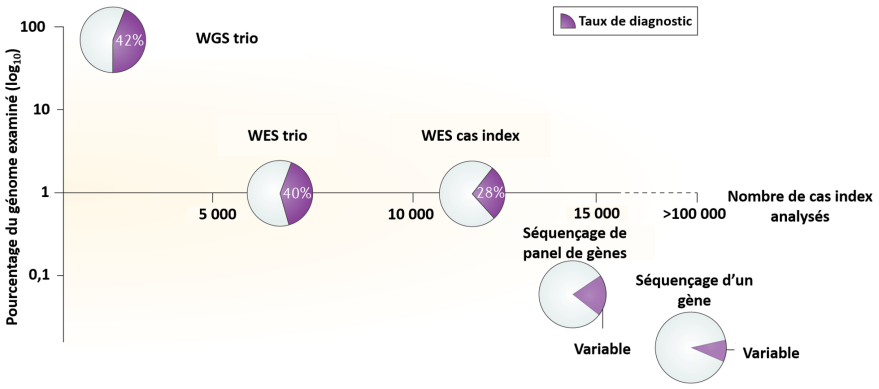


FIGURE 1.36 – Taux de diagnostic des différentes stratégies de séquençage HTS.
Adapté d'après *Wright, C. F., FitzPatrick, D. R., Firth, H. V. (2018). Paediatric genomics : diagnosing rare disease in children. Nature Reviews Genetics, 19(5), 253-268.*

De plus, la génétique clinique n'a encore que peu pénétré certaines spécialités de la médecine, ce qui fait que l'impact de la génétique dans certaines de ces indications est peut-être sous-estimé, voir *Figure 1.37*.

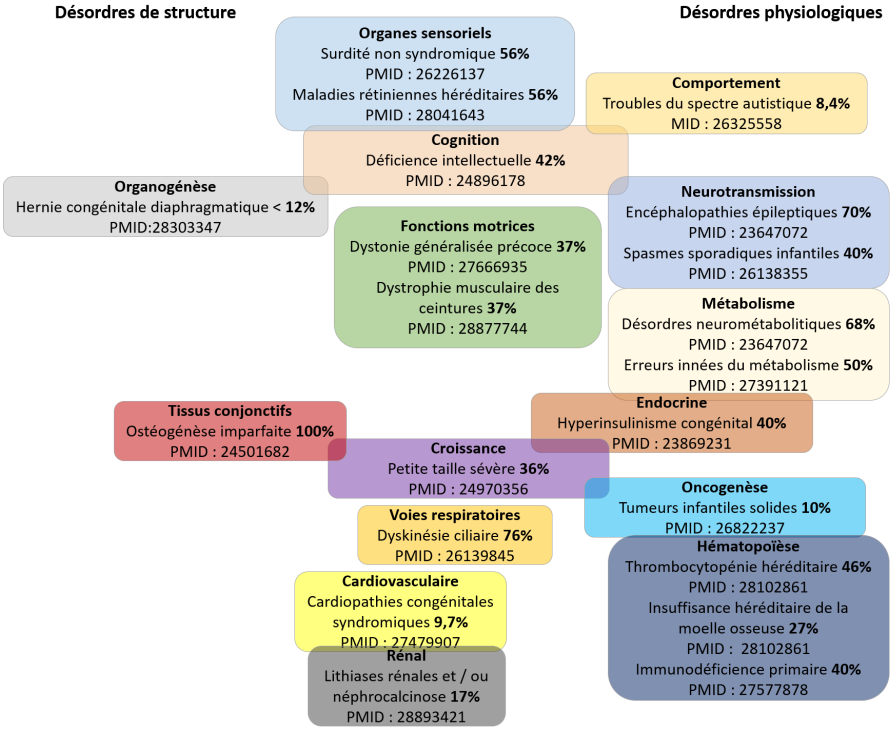


FIGURE 1.37 – Taux de diagnostic estimé d'études de séquençage HTS sur différentes indications médicales.
Adapté d'après *Wright, C. F., FitzPatrick, D. R., Firth, H. V. (2018). Paediatric genomics : diagnosing rare disease in children. Nature Reviews Genetics, 19(5), 253-268.*

1.8 Limites des approches actuelles et perspectives

1.8.1 Limites du séquençage d'exome entier *short-reads*

Comme vu précédemment, même si le séquençage d'exome entier est une analyse avec un excellent rapport coût sur diagnostic, il n'en est pas moins limité. Tout d'abord, même s'il est globalement accepté que les régions exoniques soient le support de la majorité des variations supposées pathogènes dans le cadre des maladies rares[42], elles ne représentent qu'une petite fraction du génome entier. Ces estimations ont été formulées conformément aux connaissances de leur temps et pourront se révéler inexactes dans quelques années. C'est en tout cas ce que tendent à montrer les données issues d'outils comme SpliceAI[111] et des scores dérivés comme le CADD-Splice[112], permettant de prédire les effets des variants sur l'épissage, effets qui sont actuellement sous-estimés. En accumulant de la donnée issue de plusieurs technologies de séquençage et de techniques de détection orthogonales, il est possible de s'approcher de la photographie complète des variations génomiques pathologiques. Pourtant, un grand nombre de mécanismes moléculaires aboutissant à certaines pathologies restent pour le moment inconnus, de même que la fonction de certains gènes ou certaines protéines. Il reste donc encore de la marge en ce qui concerne le diagnostic des maladies liées à une variation en régions codantes.

La technique du WES est également limitée à cause de sa conception. En effet, un séquençage d'exome comporte en général, fragmentation (donc petite taille de *reads*), capture, enrichissement et amplification. La petite taille des *reads* a un impact sur la capacité de résolution des petits indels ainsi que des zones faiblement complexes. La couverture des zones capturées d'un exomes est plus ou moins inconsistante en fonction de certains facteurs comme, la qualité du kit de capture, du matériel génétique de départ (contamination, dégradation ...) et du manipulateur, ce sont les biais techniques. De plus, certaines décisions lors de la préparation de la librairie comme le choix d'une fragmentation enzymatique beaucoup plus dépendantes des pourcentages en GC de la séquence, plutôt qu'une fragmentation mécanique, peuvent avoir un impact sur la qualité de la librairie et donc du séquençage[113].

En moyenne, un exome de bonne qualité a au moins 95 % de ses bases couvertes séquencées par au moins 20 lectures. Les zones faiblement complexes comme les zones riches en GC et homopolymériques sont moins couvertes que le reste de l'exome, ce qui rend ces zones difficiles à étudier. Ce biais provient en partie d'une hybridation moins efficace des sondes sur ces zones spécifiquement[40]. De plus, la petite taille relative des *reads* séquencés en WES ne permet pas non plus la reconstruction de zones et de variations complexes.

1.8.2 Limites du séquençage de génome entier *short-reads*

Le WGS incarne la révolution de la médecine personnalisée. Cet examen a toutefois des limitations. Chaque technologie de séquençage produit des erreurs qui lui sont caractéristiques. Dans le cas de la technologie Illumina, les erreurs proviennent de la fiabilité de détection par le capteur optique et la chimie utilisée. Ces erreurs, bien qu'assez peu fréquentes (0,1 % en moyenne), sont enrichies par certains motifs, comme les répétitions simples de GC[113].

De la même manière que le WES, le WGS comporte des étapes prônes aux biais, encore une fois, la fragmentation et l'amplification (s'il y en a). Cependant, dans la plupart des cas, le WGS couvre globalement mieux les régions exoniques qu'un WES, même de bonne qualité. De plus, le WGS permet l'étude de régions qui ne sont pas capturées en WES, notamment les introns, les UTR et les régions intergéniques, mais aussi les régions mal capturées en WES. Le WGS permet également la détection de plus de types de SV, notamment dans les zones non détectées en WES. L'apport supplémentaire de ces zones permet au WGS un taux de diagnostic supplémentaire par rapport au WES, dans une mesure qui n'est pas encore définitive, car tous les

types de variations détectables par le WGS ne sont pas encore pleinement exploités[114][115][116]. Le WGS *short-reads* ne permet toujours pas la caractérisation fine des zones répétées et de faible complexité, mais représente une évolution conséquente du WES pour un prix plus élevé. Peu de laboratoires aujourd'hui en France peuvent se permettre d'acquérir des séquenceurs type NovaSeq permettant le séquençage WGS au plus bas prix. C'est le rôle d'initiatives comme le Plan France Médecine Génomique 2025 d'acquérir l'infrastructure nécessaire pour la génération et le traitement de données WGS *short-reads* au niveau national et faire entrer le WGS dans le parcours de soin français.

1.8.3 Limites du séquençage de génome entier *long-reads*

Les technologies de séquençage de troisième génération *long-reads* compensent certaines des faiblesses du WGS *short-reads*, mais en comportent d'autres. La complexité du *basecalling* de la technologie ONT fait que la précision à la base de la technologie est bien moins élevée que celles des technologies de seconde génération. De ce fait, les zones homopolymériques favorisent l'apparition de petits indels, la technologie ne permettant pas de déterminer la longueur exacte de ces zones.

En revanche, cette technologie peut s'avérer utile en théorie pour la détection de variations de structure de petite taille dans des zones assez mal caractérisées par des techniques de séquençage basées sur des lectures courtes. Cependant, ces variations n'étant détectables que depuis récemment exclusivement avec ces technologies relativement jeunes, la connaissance médicale de ces régions émerge progressivement. La technologie PacBio, bien que plus précise que la technologie ONT est beaucoup plus chère et bien moins démocratisée dans les laboratoires. Si son prix restait un frein à son adoption massive, il faudrait que certaines découvertes majeures seulement possibles avec cette technologie poussent les laboratoires à s'équiper. Hormis quelques pionniers, ces technologies n'ont pas encore été largement adoptées par de grands organismes dédiés à la génomique pour le diagnostic humain.

1.9 Contexte du travail de thèse

1.9.1 Au niveau national et international

De nombreux pays ont mis en place ces dernières années des capacités d'acquisition et d'analyse de données génomiques pour le diagnostic. Certains pays comme l'Estonie, la Slovénie ou la Hollande ont déjà intégré depuis quelques années la médecine génomique de précision à leur parcours de Santé. D'autres pays comme les États-Unis (2014), la Chine et le Royaume-Uni (2012) en tête de proue avec le projet 100 000 génomes pour un coût de 300 millions de livres, s'appuient sur des plans d'investissement étatiques pour le développement de cette nouvelle filière de santé. Depuis les années 2010 avec la démocratisation du séquençage haut débit, la France accuse un retard envers ces pays qui disposent d'infrastructures de séquençage haut débit nationales.

La France dispose tout de même d'infrastructures de séquençage publiques ou privées à disposition de la recherche ainsi que du diagnostic. Elle peut, de plus, compter sur une communauté de professionnels de santé particulièrement dynamique en ce qui concerne les maladies génétiques. Mais des investissements locaux épars ne peuvent rivaliser avec des structures étrangères organisées de manière industrielle. Celles-ci captent souvent, en sous-traitance, les financements de recherche français plutôt que les structures existantes françaises de par la différence de compétitivité. De plus, les données génomiques de patients français peuvent attiser la convoitise de nombreux états, mais aussi de nombreuses sociétés privées spécialisées dans le domaine de la donnée comme Google, Apple, Facebook, Amazon, Microsoft, Samsung (GAFAMS), mais aussi les sociétés spécialisées dans le séquençage comme Illumina. Il est donc critique pour des raisons de souveraineté nationale de garder le contrôle des données de santé des Français, d'apprendre à les traiter, les analyser et en exploiter leur plein potentiel. Tout cela permettra ainsi d'éviter de voir émerger un tourisme médical génomique vers certains de nos pays voisins, créant ainsi une inégalité des citoyens face au soin.

C'est pourquoi la France a initié en 2015 le PFMG 2025[117] (Plan France Médecine Génomique 2025), un plan d'investissement pour créer une structure nationale composée de plusieurs plateformes de séquençage et d'une infrastructure informatique nationale afin de préparer l'intégration de la médecine génomique dans le système de santé français à l'horizon 2025. Le plan français a des objectifs similaires aux différents plans nationaux mondiaux, à savoir, rattraper le retard à l'international *via* la constitution d'une base de données représentative de la population française, l'augmentation du taux de diagnostic des maladies rares afin de diminuer l'errance diagnostique et les examens coûteux inutiles, ainsi qu'une meilleure caractérisation des cancers et des tumeurs afin d'adapter au mieux les traitements et améliorer le pronostic patient. L'organisation en une structure nationale permet l'acquisition du matériel pour la génération de séquences ainsi que l'infrastructure informatique pour l'analyse de données génomiques, mais aussi de développer la filière de la médecine génomique française en favorisant les partenariats publics, universitaires et industriels pour la recherche. Actuellement, deux plateformes françaises ont vu le jour, la plateforme parisienne SeqOIA recevant le prélèvement des patients franciliens ainsi que de la région Nord et la plateforme AURAGEN basée en région Auvergne-Rhône Alpes recevant les prélèvements de la région Sud et des DOM-TOM.

1.9.2 Contexte du travail du doctorant

Cette thèse a été menée *via* une Convention Industrielle de Formation par la REcherche (CIFRE) ayant pour but de rapprocher le monde universitaire du monde industriel.

Les partenariats publics, privés ayant été favorisés en vue du lancement du PMFG 2025, ma thèse a été à l'origine d'un rapprochement entre l'Université de Grenoble-Alpes *via* l'*Institute for Advanced Bioscience* (IAB), le CHU de Grenoble ainsi que le laboratoire de biologie médicale privé Eurofins Biomnis. Bien que cette thèse n'ait pas directement pris place au sein de la plateforme AURAGEN, le pipeline d'analyse de données d'exome constitutionnelles développé qui en a résulté a été la V0 des codes utilisés par la plateforme pour les analyses germinales. L'objectif de cette thèse est dans une moindre mesure le même que celui du PMFG 2025, à savoir rattraper le retard à l'international sur les analyses de données de WES dans un premier temps, puis sur des données pan-génomiques tout court.

Dans le cadre du partenariat privé, public, la partie universitaire profitait de la puissance de séquençage de l'industriel avec plus de 3000 données d'exomes produites dont plus de 300 exomes de patients grenoblois et l'industriel profitait de la connaissance académique sur l'analyse et le métier de généticien concernant la filtration des variations génomiques détectées. De plus, la plus grande partie des développements et des calculs de cette thèse ont pu être menés à bien grâce à l'infrastructure de calcul HPC (*High-Performance Computing*) CIMENT de l'UMS GRICAD de Grenoble (https://gricad.univ-grenoble-alpes.fr/), donnant accès au doctorant à plusieurs de *clusters* de calculs pour un total de plusieurs milliers de cœurs disponibles.

Cette thèse a été à l'inititive du développement de la Bioinformatique au sein des deux parties prenantes. Initialement intégré au sein de la BU de Génétique d'Eurofins Biomnis en tant que second bioinformaticien (après le Dr Jean-François Taly), une division bioinformatique spécifique y a été créé, l'équipe BioIT. Elle comprend aujourd'hui six personnes (m'incluant). De la même manière, j'ai été le premier bioinformaticien localisé au CHU de Grenoble. Initialement dans les locaux du service de Génétique et Procréation, celui-ci est devenu par la suite service de Génétique, Génomique et Procréation. J'ai ainsi pu présenter plus en détail ce qu'était la Bioinformatique au personnel médical, ainsi qu'aux membres de la Direction des Services Numériques de l'hôpital de Grenoble. Enfin, je suis localisé au sein des locaux AURAGEN Grenoble (même si ne faisant pas parti du projet), portant le nombre de bioinformaticiens à quatre (m'incluant).

Résultats

Développement d'un pipeline industriel d'analyse de DNAseq Illlumina

2.1 Contexte de développement du pipeline industriel d'analyse de *DNAseq* Illlumina

2.1.1 Motivation

Antérieurement au début de la thèse, chacune des deux parties possédait déjà un pipeline d'analyse de données d'exome constitutionnelles, mais chacun avec des perspectives d'améliorations qui leur étaient propres, voir *Table 2.14*.

	Grexome (J.Thevenon)	Biexome (JF. Taly)	Lygrexome (Q. Testard + JF. Taly + J. Thevenon)
Gestionnaire de *workflow*		Nextflow	Nextflow
Conteneurs		Singularity	Singularity
Démarrage automatique à partir de la production de données de séquençage		Watcher + Nextflow	Watcher + Nextflow
Version du génome	hg38	hg19	hg19 et hg38
Alignement	BWA MEM + Samtools	BWA MEM + Samtools	BWA MEM + Samtools
Production de VCF standardisés	GATK 3.8 + Picard Tools	GATK 3.6 + Picard Tools	GATK4
Annotation des VCF	SnpSift + SnpEff	SnpSift + SnpEff	SnpSift + SnpEff
Filtration et curation des variants	Scripts Python		Scripts Python
Production d'un rapport de variants minimal	Scripts Python		Scripts Python
Appel de CNV	XHMM		GATK4 + scripts Python
Mesure de métriques de qualité	GATK 3.8 + Samtools + scripts Python	GATK 3.6 + Samtools + identito-vigilance	GATK4 + Samtools + scripts Python + MultiQC

Absent du pipeline
Présent dans le pipeline
Absent du pipeline pour à cause de facteurs extérieurs

TABLE 2.14 – Description des différents pipelines existant au démarrage de la thèse et du pipeline développé.

Le pipeline de Grenoble, développé lorsque le Pr Julien Thévenon était membre de la FHU-TRANSLAD à Dijon, était constitué d'un *wrapper* en Bash, qui lorsqu'il était exécuté appelait d'autres scripts Bash exécutant soit des outils bioinformatiques, soit des scripts Python. Les résultats produits par ce pipeline étaient biologiquement très pertinents, mais son organisation complexe empêchait toute tentative de modification et de mise à jour. Il était requis de reprendre toute sa base de code Bash et Python.

De l'autre côté, à Eurofins Biomnis, le pipeline développé par le Dr Jean-François Taly reposait sur une base solide pour le développement de pipelines bioinformatiques au standard industriel (avec la combinaison des outils Singularity et Nextflow) mais dont les résultats biologiques se limitaient au suivi strict des bonnes pratiques GATK[61].

La motivation initiale au développement d'un nouveau pipeline était donc la constitution d'un pipeline combinant les forces des deux déjà existants tout en permettant l'intégration de ce pipeline dans les infrastructures grenobloises et lyonnaises ainsi qu'en mettant à jour et en intégrant certaines des dernières nouveautés apparues pour l'analyse de données d'exome ces dernières années, notamment la version 4 de l'outil GATK. La résultante de ce développement est le pipeline Lygrexome (composé de "Ly" pour Lyon, "gre" pour Grenoble et "exome").

La seule régression du nouveau pipeline par rapport aux pipelines originaux est l'absence d'un système de surveillance (*watcher*) permettant le démarrage automatique de l'analyse sans intervention humaine. Il n'a pas été intégré au nouveau pipeline, car celui-ci, sur les différentes infrastructures sur lesquelles il est déployé, n'a pas directement accès à des données produites par des séquenceurs. Ces données sont transférées manuellement par les divers acteurs réalisant le séquençage des données qui sont analysées par le pipeline. Aucune procédure n'a pour l'instant été mise en place avec ces différents acteurs pour le transfert standardisé des données et des fichiers annexes permettant le démarrage automatique du pipeline.

Bien que ce pipeline ait été conçu comme un pipeline d'analyse de données d'exome, il peut tout à fait être utilisé avec des données de WGS Illumina en entrée. Le résultat se limitera alors aux variations des régions exoniques, mais seront en général de meilleures qualités que des données de WES. Le pipeline a été utilisé plusieurs fois de la sorte au cours du doctorat.

2.1.2 Contraintes de l'environnement de développement

La quasi-totalité du développement et des analyses effectuées lors des deux premières années du doctorat l'ont été sur les *clusters* Luke et Dahu de l'infrastructure CIMENT opérés par l'UMS GRICAD. En 2018, les *clusters* ne possédaient ni logiciel de conteneurisation, ni ordonnanceur compatible avec la version de Nextflow de l'époque.

Premièrement, il a fallu installer et tester Singularity sur l'infrastructure CIMENT. Cela a été fait en partenariat avec Bruno Bzeznik, ingénieur de recherche à l'UMS GRICAD. Dans un premier temps, Singularity a été installé sur le *cluster* Luke. Malheureusement, celui-ci étant constitué de machines dépareillées installées sous différentes versions du noyau Linux, Singularity avait certains comportements inopportuns. Singularity a alors été installé sur le *cluster* Dahu qui venait à peine d'être mis en place et qui était encore en phase de rodage. Celui-ci étant composé de machines identiques sous une version du noyau linux récente, l'utilisation du logiciel s'est révélée stable.

Ensuite, il a fallu faire en sorte que Nextflow puisse soumettre des tâches à l'ordonnanceur du *cluster* Dahu. Ce dernier est orchestré par le logiciel OAR (https://oar.imag.fr), développé à Grenoble depuis les années 2000 au sein de l'IMAG, institut hébergeant les infrastructures CIMENT. OAR a été développé afin de pouvoir bénéficier de certaines fonctionnalités qui n'étaient pas disponibles dans les ordonnanceurs de l'époque. OAR n'étant pas supporté par Nextflow, il a donc fallu l'intégrer en modifiant le code source

Nextflow. Le repository Github de Nextflow (https://github.com/nextflow-io/nextflow) a été fork et Bruno Bzeznik ainsi que moi-même, sommes partis de l'intégration native de SGE, un ordonnanceur très populaire des années 2000 à 2010 dont est dérivé OAR. La modification fut un succès mitigé, car bien qu'elle permît de lancer des jobs OAR avec Nextflow, elle ne permettait pas d'utiliser certaines fonctionnalités de Nextflow, l'intégration était minimale.

De plus, entre le moment du fork et la finalisation de l'intégration, la version live de Nextflow subit une évolution majeure qui rendit l'intégration caduque. Il a donc été décidé de ne pas mettre d'efforts supplémentaires dans le développement de l'intégration d'OAR au sein de Nextflow et de rester avec cette version de 2018, tant que cela était possible.

En 2020, le Dr Maxime Vallée, ingénieur hospitalier bioinformaticien aux Hospices Civils de Lyon (HCL) fort d'une expérience dans le développement de programmes d'interface avec des ordonnanceurs a repris cette intégration pour la finaliser *via* une pull request sur le dépôt officiel Github. Depuis le support d'OAR est intégré nativement et est maintenu par Paolo Di Tommaso, développeur principal de Nextflow.

2.1.3 Méthodologie de développement

2.1.3.1 Conteneurisation

Comme indiqué précédemment, la composante industrielle du pipeline développé au cours de cette thèse repose sur deux outils principaux, le gestionnaire de *workflow* Nextflow et le logiciel de conteneurisation Singularity.

Les outils utilisés au sein de ce pipeline sont donc tous sans exception conteneurisés. La plupart des conteneurs Singularity utilisés ont été créés à partir de conteneurs Docker. Il y a plusieurs raisons à cela. Premièrement, au début du développement du pipeline, en 2018, les infrastructures communautaires de Singularity étaient assez peu développées. En effet, la communauté Docker est assez active grâce à l'existence du *Docker Hub* (https://hub.docker.com/), un catalogue de conteneurs public dont l'équivalent Singularity n'est que très récent. Nombre d'outils usuels en bioinformatique ont maintenant pour habitude de fournir un *Dockerfile* (fichier de recette Docker), soit sur le dépôt de leur outil, soit sur le *Docker Hub* directement. *Docker Hub* permet donc à tout un chacun de déposer ses *Dockerfiles* et ses conteneurs dans ses différentes versions sur un espace dédié sécurisé. Même si la décision devait être prise aujourd'hui avec les améliorations apportées aux espaces communautaires Singularity, le choix pourrait toujours être celui de Docker, tant la communauté est plus active. La seule inconnue qui subsiste quant à Docker est son modèle économique, bien que gratuit depuis le début, de plus en plus de services auparavant gratuits deviennent payant. Actuellement, très peu d'outils en libre accès proposent des recettes Singularity ou des conteneurs déjà construits sur le *Singularity Hub*.

La plupart des conteneurs utilisés au sein du pipeline sont donc basés sur un conteneur Docker. Les conteneurs sont créés localement à l'aide de *Docker Engine* sur les systèmes Linux ou *Docker Desktop* sur les systèmes Windows et Mac. Les conteneurs reposent sur une image de base, éditée par le consortium Continuum (https://www.continuum.io) responsable des gestionnaires de paquet anaconda et miniconda où ce dernier est préinstallé. Miniconda est assez populaire en bioinformatique, notamment grâce à son canal Bioconda (https://bioconda.github.io/) regroupant de nombreux outils bioinformatiques à jour et permettant leur installation de manière simplifiée. Un outil en particulier, GATK 4 nécessite d'être installé avec le gestionnaire de paquet miniconda avec un environnement particulier fourni à l'installation (notamment pour le module d'appel de CNV germinaux). Les paquets sont alors simplement installés dans le conteneur à l'aide d'une commande d'installation miniconda. Un exemple d'un *Dockerfile* produisant un conteneur minimal pour l'alignement de *reads* contre une référence à l'aide de miniconda est visible *Snippet 1.2*. Le

reste des outils qui ne sont pas dans les canaux miniconda, ont pour la plupart un *Dockerfile* fournit par leur développeur et sont construits à partir de celui-ci.

Une fois le conteneur Docker construit et envoyé sur *Docker Hub*, il est converti en conteneur Singularity local *via* une commande. Les conteneurs créés ne sont pas tous "atomiques", c'est-à-dire qu'ils peuvent contenir plus d'un outil. Cela est impératif pour les *process* Nextflow qui nécessitent plus d'un seul outil, un *process* ne pouvant utiliser qu'un seul conteneur distinct. Les outils qui doivent être utilisés ensemble dans divers *process* sont donc présents dans un seul même conteneur.

2.1.3.2 Organisation du code

Le code du pipeline est divisé en plusieurs fichiers. Tout d'abord, un fichier principal (*main*) regroupant l'essentiel du code exécuté (actuellement environ 1500 lignes de code). Ce fichier est subdivisé en plusieurs modules exécutables plus ou moins indépendamment les uns des autres. À l'origine, les différents modules devaient être dans des fichiers séparés, mais cela pouvait créer des comportements inattendus dans le cas de lancement asynchrone de tâches pour les différents processus Nextflow. Les différents modules du code peuvent être lancés à la suite ou de manière indépendante, un à la fois.

Le code comporte également plusieurs fichiers de configuration, avec des profils différents en fonction de l'infrastructure sur laquelle le code est exécuté. Ces profils concernent particulièrement le nombre de tâches et de cœurs utilisables en fonction de la taille de l'infrastructure, mais aussi l'emplacement des divers fichiers et bases de données de référence.

Le code est hébergé sur un dépôt GitLab (https://gitlab.com/chu_grenoble/cifre_biomnis/Lygrexome). Le dépôt contient le script principal, les scripts de configuration, les scripts Python ou Bash utilisés dans le pipeline et un README décrivant les fonctionnalités et paramètres du pipeline ainsi qu'une procédure pour lancer l'analyse. Les fichiers et bases de données de référence sont beaucoup trop volumineux (plus de 300 Go) pour être versionnés.

2.2 Modules du pipeline

Comme indiqué précédemment, le pipeline Lygrexome est divisé en plusieurs modules, chacun ayant une des grandes fonctions décrites dans les bonnes pratiques GATK, éditées par le *Broad Institute*, institut de référence mondial pour la génomique, voir *Figure 2.38*.

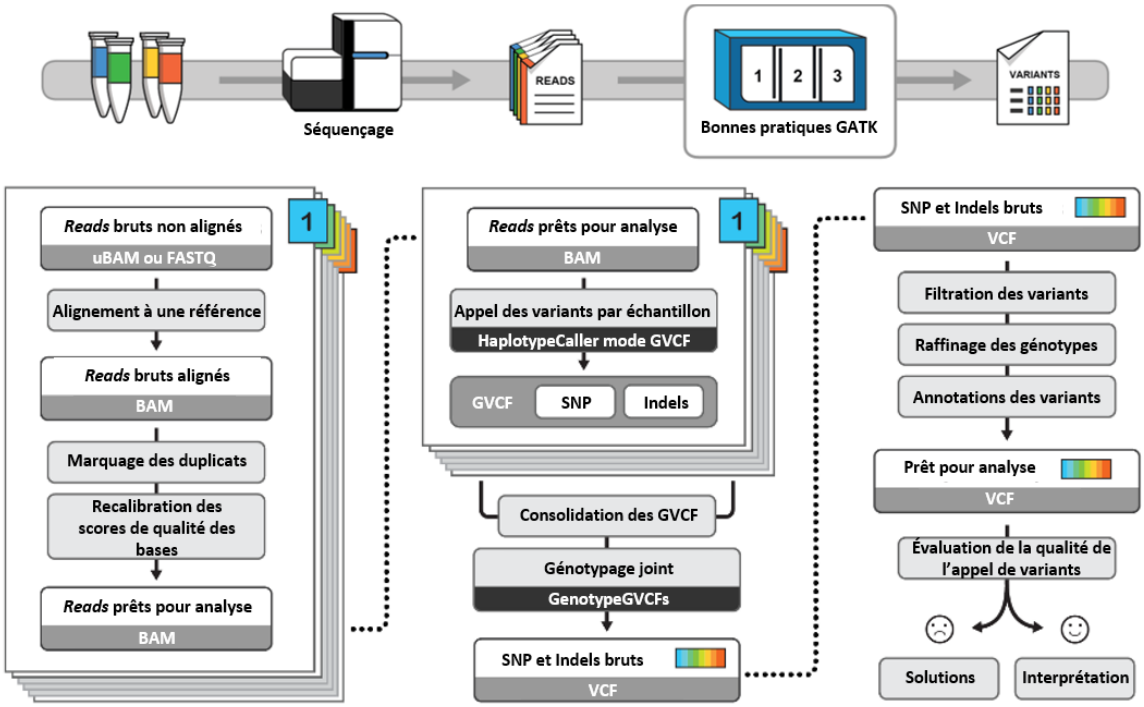


FIGURE 2.38 – Bonnes pratiques GATK pour la détection de variants germinaux de petite taille. Adapté d'après *Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high-confidence variant calls : the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 43(1), 11-10*

Sont à noter entre autres, la production d'un fichier d'alignement BAM pour l'appel de variants à partir de fichier de lectures FASTQ et d'un génome de référence format FASTA, la production d'un VCF génotypé à partir du fichier BAM et l'annotation et la curation des variants du fichier VCF. GATK[61] (*Genome Analysis ToolKit*) représente l'outil de référence de manipulation de fichiers au format BAM, VCF et bien d'autres formats, obtenus par séquençage *short-reads* Illumina et permet d'effectuer la quasi-totalité des étapes décrites dans les bonnes pratiques. C'est pourquoi l'influence du *Broad Institute* et des membres de ses équipes est très importante dans la communauté bioinformatique spécialisée en traitement de données HTS, car elle se traduit dans les outils qui y sont développés.

Lors du lancement d'au minimum un module, le pipeline effectue des vérifications quant à la présence de certains des fichiers nécessaires pour la bonne exécution de l'outil, comme la présence du génome spécifié, de ses index ou d'autres fichiers d'entrée. S'ils sont absents, alors un message d'erreur est affiché ou les fichiers sont calculés lorsque c'est possible (notamment pour les index du génome).

Le choix des modules se fait dans la commande servant à exécuter le pipeline. Cette commande permet de spécifier de nombreux paramètres comme la version du génome à utiliser, le kit de capture utilisé et bien d'autres qui sont décrites dans le README de l'outil, voir *Snippet 2.4*.

Snippet 2.4 – Exemple de commande pour lancer l'ensemble des modules du pipeline Lygrexome.

```
run_name=test_grexome ; ./nextflow -c Lygrexome.nf_config run Lygrexome.nf --fastq2bam 0
--bam2gvcf 0 --vcf2annotate 0 --metrics 0 --cnv 0 --sampleID ${run_name} --genomeVers hg38
--genomeID hg38.fasta --kit Twist --queue servoz --cnvmodel 1730
-with-report SAMPLE/${run_name}/run_reports/${run_name}_report.html
--with-trace SAMPLE/${run_name}/run_reports/${run_name}_trace.txt
-with-timeline SAMPLE/${run_name}/run_reports/${run_name}_timeline.html
-with-dag SAMPLE/${run_name}/run_reports/${run_name}_dag.png -resume
```

Certains modules nécessitant par conception des résultats produits lors d'étapes précédentes (le module produisant le VCF nécessitant un fichier BAM) peuvent être lancés si une exécution précédente du module nécessaire a eu lieu.

2.2.1 Module *fastq2bam*

2.2.1.1 Préambule

Le module fastq2bam a pour objet de produire des fichiers au format BAM prêts pour l'appel de variations à partir de données FASTQ et d'un génome de référence format FASTA indexé, voir *Figure 2.39*. Le module se base sur 3 outils principaux, BWA-MEM[56] *commit 0.7.17-r1188*, SAMtools[55] version *1.7* et GATK version *4.1.4.1*[61].

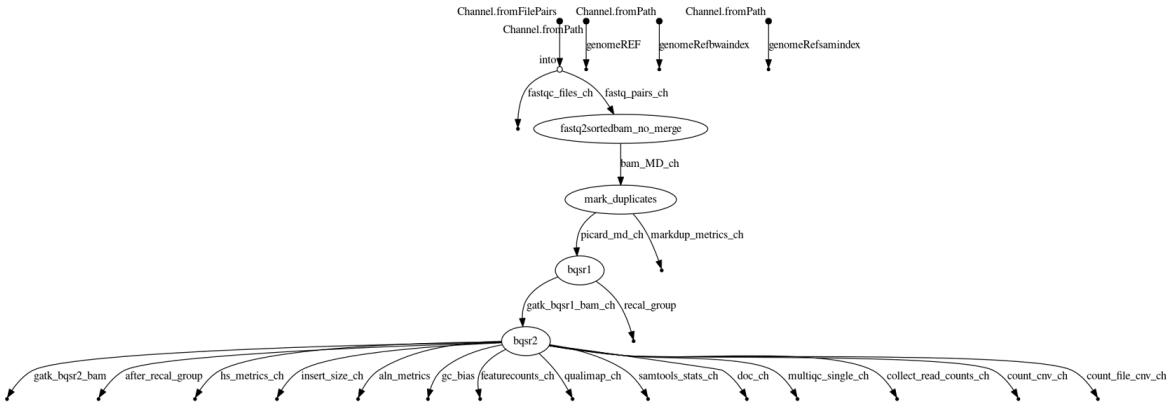


FIGURE 2.39 – Graphe orienté acyclique du module *bcl2fastq* du pipeline Lygrexome à partir de données FASTQ séquencées sur une seule *lane*.

2.2.1.2 Alignement

Premièrement, lors de l'étape *fastq2sortedbam_no_merge*, les fichiers FASTQ sont alignés contre le génome de référence spécifié dans la ligne de commande. L'alignement est effectué à l'aide de l'outil BWA-MEM au format SAM vers la sortie standard. La commande d'alignement est enchaînée ("pipée") dans l'outil SAMtools afin de ne pas produire de fichier SAM, mais directement des fichiers BAM triés moins volumineux.

2.2.1.3 Marquage des duplicats

Ensuite, lors de l'étape *mark_duplicates*, les fichiers BAM produits ont leur duplicats de PCR (étape inhérente au séquençage WES Illumina) marqués à l'aide de l'outil MarkDuplicates de la suite PicardTools (http://broadinstitute.github.io/picard/) inclue dans la version 4 de GATK. Le marquage de duplicats permet aux outils d'appel de variants d'ignorer certains *reads* au cours de leur analyse.

2.2.1.4 Recalibration des scores de qualité des bases

Enfin, les scores de qualité des bases sont recalibrés à l'aide des outils BaseRecalibrator, ApplyBQSR et AnalyzeCovariates de l'outil GATK. Cette étape est nécessaire afin de détecter les erreurs systématiques commises par les séquenceurs lorsqu'ils estiment la précision de chaque appel de base et de les corriger. La recalibration des bases se fait en plusieurs étapes lors des étapes *bqsr1* et *bqsr2* contrairement à la version 3 de GATK.

Les données FASTQ en entrée du module peuvent être compressées au format GZIP. Le pipeline peut également accepter en entrée des données séquencées sur plusieurs *lanes* d'un séquenceur, grâce à un paramètre le spécifiant le nombre de *lanes*, dans la ligne de commande servant à exécuter Lygrexome. Dans ce cas-ci, les différents FASTQ d'un même patient sont alignés de manière indépendante au sein de la tâche *fastq2sortedbam_pre_merge*, puis combinés (*merge*) en un seul fichier BAM lors de l'étape *merge_bam*, tout en conservant les informations des *lanes* dans l'entête. Le fichier d'alignement obtenu est ensuite traité comme n'importe quel autre BAM (marquage des duplicats puis recalibration des scores de qualité des bases).

2.2.2 Module *bam2gvcf*

2.2.2.1 Préambule

Le module *bam2gvcf* a pour objet de produire un VCF génotypé pour chacun des cas index (patient atteint) fournis au module, voir *Figure 2.40*. En effet, le pipeline peut analyser des cas index seuls (singletons), des duos ou des trios (en général, le cas index est un enfant et un ou deux apparentés). Pour cela, la structure familiale doit être fournie au module. Celle-ci est décrite dans un fichier appelé *batch.ped* au format "PED like", visible *Snippet 2.5*, est nécessaire au bon déroulement de l'analyse. Le module se base principalement sur l'outil GATK version *4.1.4.1*, ainsi que sur des scripts Bash et Python *2.7*.

Snippet 2.5 – Exemple d'un fichier *batch.ped* nécessaire à la description de la structure familiale des individus analysés par le module *bam2gvcf*.

```
Family_ID   Individual_ID   Paternal_ID Maternal_ID Sex Phenotype
FAM21-0441  21A1658_S1  21A1671_S2  0   0   3
FAM21-0498  21A1817_S3  21A1777_S5  21A1778_S4  0   3
FAM21-0531  21A1792_S6  21A1791_S8  21A1790_S7  0   3
FAM0001 21A1689_S9  0   0   0   3
FAM0002 21A1783_S10 0   0   0   3
FAM0003 21A1789_S11 0   0   0   3
FAM0004 20A0865_S12 0   0   0   3
```

Dans cet exemple, sont fournis au module, un duo, deux trios et quatre singletons.

Le format PED (*pedigree*), est un format utilisé par certains outils pour décrire la structure familiale d'un ensemble de données de séquençage. Dans le cas du pipeline, seules 4 colonnes du fichier *batch.ped* ont leur utilité,

les colonnes *Individual_ID* (identifiant du cas index), *Paternal_ID* (identifiant du premier apparenté), *Maternal_ID* (identifiant du deuxième apparenté), et la colonne *Phenotype* (si la valeur est de 3, alors l'analyse est lancée sur le cas index). Les autres ont été ajoutées en prévision de futures améliorations du pipeline, notamment la colonne *Family_ID* qui pourrait être utilisée dans le suivi des analyses et la colonne *Sex* qui permettrait des analyses d'identito-vigilance.

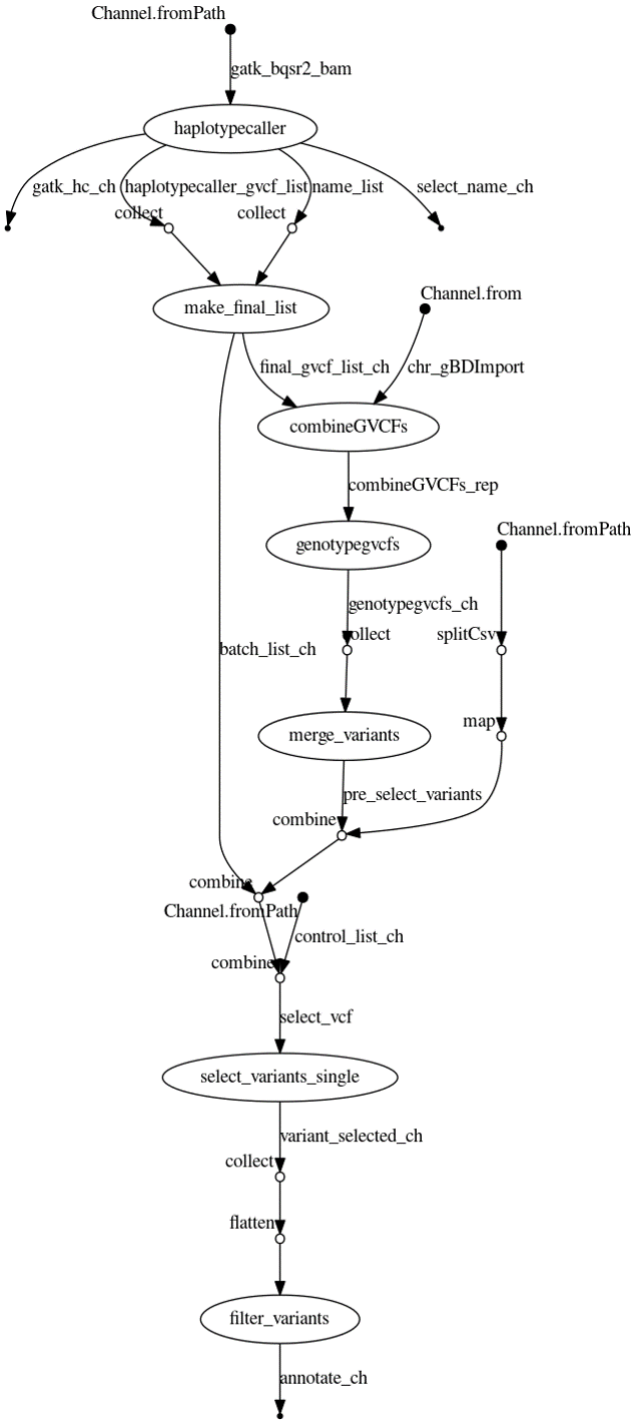


FIGURE 2.40 – Graphe orienté acyclique du module *bam2gvcf* du pipeline Lygrexome.

2.2.2.2 Appel de variants de petite taille

La première étape du module *bam2gvcf* est l'utilisation de l'outil HaplotypeCaller[62] de la suite GATK, en mode GVCF (*Genomic VCF*) pour le SNP *calling* lors de l'étape *haplotypecaller*. L'algorithme de l'outil pour la détection de SNP est schématisé *Figure 1.30*. L'appel de variant est effectué à partir de chacun des fichiers BAM produits précédemment et d'une cible (*target*) décrivant les différentes zones où doit être effectué le SNP *calling* est également requise.

C'est dans le cadre de ce module que les analyses sont dirigées strictement vers les exons. En effet, la cible d'appel est un fichier BED comportant les régions exoniques identifiées par la base de données RefSeq[9] étendue bilatéralement (*padding*) de 20 paires de bases. L'appel de variation de petite taille est exclusif aux intervalles décrits dans le fichier BED.

Les GVCF ne doivent pas être confondus avec les VCF génotypés qui seront décrits plus tard. Un GVCF est un type de fichier suivant la spécification VCF[64] mais qui apporte en plus quelques informations supplémentaires. Il existe deux types de GVCF produits par HaplotypeCaller, ceux produits avec l'option *-ERC BP_RESOLUTION* sont des fichiers VCF avec une ligne par position génomique de la cible d'appel, que le site soit variant ou référence. Les fichiers produits avec l'option *-ERC GVCF*, ce qui est le cas au sein de ce module, sont des fichiers VCF comportant une ligne par site variant et une ligne par bloc de sites références contigus. Ces blocs peuvent être fragmentés dans le cas où une variation trop importante de la qualité du génotypage (GQ) est observée. Les plages de variations tolérées sont présentes dans l'entête du fichier.

2.2.2.3 Génotypage joint

L'étape suivante consiste au génotypage, c'est-à-dire, la détermination de l'ensemble des allèles des différents fichiers GVCF (*batch*) obtenus précédemment. Le génotypage joint, qui est une étape recommandée par les bonnes pratiques GATK[61], consiste en l'utilisation de fichiers GVCF de contrôle afin d'augmenter la puissance de détection du génotype des différents sites des échantillons du *batch*. La finalité est la production de fichiers VCF génotypés regroupant les différents échantillons décrits dans le fichier PED (d'un à trois selon la structure familiale). Génotyper des données d'intérêt contre une cohorte a plusieurs avantages :

— Augmenter le nombre de données sur lequel est effectué le génotypage permet de distinguer plus facilement les sites référence homozygotes des sites avec des données manquantes. Dans le cadre du WES, cela permet de détecter s'il y a un problème de couverture au niveau du *batch* ou si c'est un problème récurrent du procédé de capture.

— Conférer une plus grande sensibilité pour la détection de variants à basse fréquence. En combinant les informations de tous les échantillons, le génotypage joint permet de récupérer certains génotypes sur les sites faiblement couverts pour un ou plusieurs échantillons donnés, mais où d'autres échantillons du *batch* et de la cohorte combinée ont une même variation.

— Conférer une plus grande capacité à filtrer les faux positifs. Les stratégies de filtration de variations à partir de leur occurrence peuvent se baser sur l'information de la fréquence du génotype à une position donnée dans le *batch* ainsi que dans la cohorte.

La combinaison de l'ensemble des GVCF du *batch* avec ceux de la cohorte est effectuée à l'aide de l'outil CombineGVCFs lors de l'étape éponyme. Idéalement la cohorte utilisée sera la plus proche possible des fichiers du *batch*, c'est-à-dire, produite avec le même type de séquenceur ou le même kit de capture, par exemple. La cohorte de normalisation comporte donc un nombre d'échantillons variable en fonction du mode utilisé du pipeline qui est notamment dépendant du kit (*Twist Human Core Exome* ou *Roche MedExome*) et la version du génome (hg19 ou hg38). La cohorte est organisée de manière que tous les GVCF contrôles soient déjà combinés entre eux et divisés par chromosome. Il y a donc 24 fichiers GVCF de cohorte et 24 tâches de combinaison de GVCF lancées à chaque exécution de module, peu importe le nombre d'échantillons du *batch*.

Ensuite, les 24 fichiers GVCF scindés par chromosome obtenus par combinaison sont génotypés à l'aide de l'outil GenotypeGVCFs lors de l'étape éponyme. 24 VCF génotypés sont alors produits contenant l'ensemble des échantillons de la cohorte et du *batch*. Ces 24 fichiers sont alors combinés en un seul VCF génotypé lors de l'étape *merge_variants*.

Les cas index et leurs apparentés (s'il y en a), décrits dans le fichier PED sont extrait du VCF génotypé combiné par un script Python afin d'obtenir un VCF par structure familiale lors de l'étape *select_variants_single*. Enfin, l'étape *filter_variants* à l'aide de l'outil VariantFiltration permet d'éliminer certaines variations de mauvaise qualité ou mal formées.

2.2.3 Module *vcf2annotate*

2.2.3.1 Préambule

Le module *vcf2annotate* a pour objet la production d'un rapport de petites variations annotées (SNV et indels) minimal contenant les variants cliniquement pertinents pour l'interprétation à partir d'un VCF génotypé, voir *Figure 2.41*.

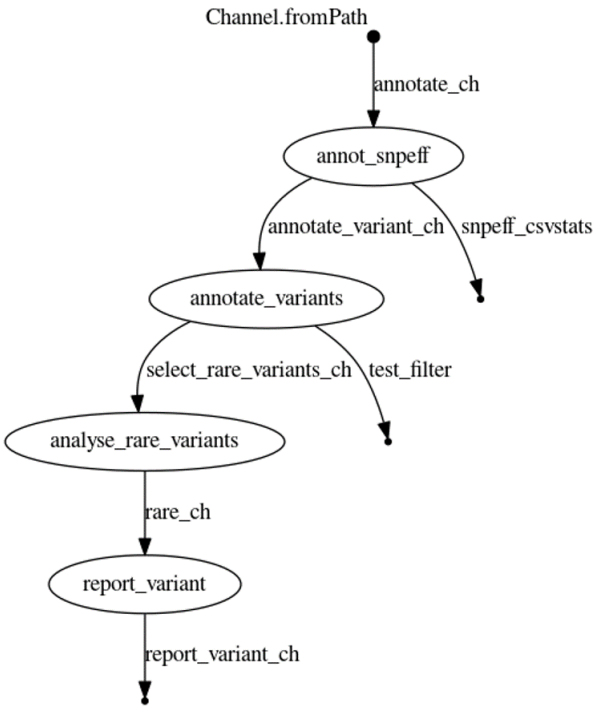


FIGURE 2.41 – Graphe orienté acyclique du module *vcf2annotate* du pipeline Lygrexome.

2.2.3.2 Annotations

La première étape afin d'obtenir un rapport minimal de variations pour le clinicien est le rajout de métadonnées (annotations) sur les différentes variations afin de pouvoir appliquer des critères de filtration et ne garder que celles qui cliniquement pertinentes.

L'annotation est faite en deux temps, premièrement, les VCF génotypés sont annoté à l'aide des outils SnpEff[73] et SnpSift[74] lors de l'étape *annot_ snpeff*. Les bases de données utilisées pour l'annotation avec ces deux outils sont indiquées *Table 2.15*.

Base de donnée	Lien d'accès	Information médicale contenue
dbSNP[24]	https ://www.ncbi.nlm.nih.gov/snp/	Liste de variations connues
Kaviar[118]	http ://db.systemsbiology.net/kaviar/	Fréquence dans la population de SNV et indels connus
ExaC[65]	https ://gnomad.broadinstitute.org/	Fréquence dans la population de SNV et indels connus
GnomAD[70]		Fréquence dans la population de SNV, indels et SV connus
dbNSFP[119]	https ://sites.google.com/site/jpopgen/dbNSFP	Prédictions et annotations fonctionnelles

TABLE 2.15 – Bases de données et caractéristiques de leurs annotations appliquées par les outils SnpEff et SnpSift.

Ensuite, certaines annotations supplémentaires sont ajoutées aux fichiers nouvellement annotés par un script Python lors de l'étape *annotate_ variants*. Les bases de données utilisées pour l'annotation avec ce script sont indiquées *Table 2.16*. L'ensemble des annotations sont ajoutées dans le fichier VCF dans l'entête et dans le champ INFO comme indiqué dans la spécification du format VCF[64].

Base de donnée	Lien d'accès	Information médicale contenue
ClinVar[71]	https ://www.ncbi.nlm.nih.gov/clinvar/	Signification clinique de variations de tout type
COSMIC[120]	https ://cancer.sanger.ac.uk/cosmic	Catalogue de variations somatiques impliquées dans le développement de cancers
OMIM[78]	https ://www.omim.org/	Catalogue de liens entre phénotypes de maladies rares et bases moléculaires connus
ACMG[121]	https ://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/	Liste de gène d'intérêt en cas de découverte secondaire
ExaC[65]	https ://gnomad.broadinstitute.org/	Liste de scores de prédiction du consortium ExaC

TABLE 2.16 – Bases de données et caractéristiques de leurs annotations appliquées par le script *annotate_ variants.py*.

2.2.3.3 Filtration des variations

L'étape suivante est la filtration des variations sur des critères établis préalablement se basant sur les annotations précédemment ajoutées. La filtration des variations est effectuée lors de l'étape *analyse_rare_variants* et aboutit à la production d'un fichier VCF filtré. Les arguments fournis au script *analyse_rare_variants.py* sont décrits *Table 2.17*. Ce script ainsi que le script *annotate_variants.py* se basent sur une précédente version développée à Dijon par le Dr Émilie Tisserant et par Yannis Duffourd et ont été remaniés pour être compatible avec la version actuelle du pipeline et de la spécification VCF *4.2*.

Argument	Critère de filtration
–pass	Sélections des variants indiqués comme "*PASS*" dans le champ *FILTER* (passent les critères de qualité)
–denovo	Drapeau pour les variants *de novo* (oui, non)
–min_dp 5	Profondeur minimale de couverture de la variation
–min_alt_dp 2	Nombre minimum de lectures soutenant l'allèle alternatif
–min_alt_frac 0.1	Fraction minimale d'allèles alternatifs
–max_samples 3	Nombre maximum d'échantillons du *batch* et de la cohorte de contrôle comportant la même variation
–nsssi	Sélections des variants non-synonymes (NS), site accepteur ou site donneur de l'épissage (SS), Indels codant (I)
–freq_exac 0.01	Fréquence maximale tolérée dans la population de la variation selon ExaC
–filter_sites dbSNP	Sélections des variants renseignés dans dbSNP[24]
–recessive	Drapeau pour les variants récessif (oui, non)
–freq_gnomadex 0.01	Fréquence maximale tolérée dans la population de la variation selon GnomAD[65] Exome
–freq_gnomadge 0.01	Fréquence maximale tolérée dans la population de la variation selon GnomAD[65] Génome

TABLE 2.17 – Critères appliqués pour la filtration de variations de fichiers VCF annotés pour le diagnostic de maladies rares du script *analyse_rare_variants.py*.

2.2.3.4 Production du rapport

Dernière étape du module *vcf2annotate*, la production d'un rapport tabulé pour l'interprétation par le clinicien lors de l'étape *report_variant* par le script *report_vcf.py*. Les rapports sont ensuite le support de l'interprétation à l'aide de logiciels tableur comme Microsoft Excel ou Open Office Calc.

2.2.4 Module *metrics* (contrôle qualité)

2.2.4.1 Préambule

Le module *metrics* a pour objectif de calculer et compiler de nombreuses données de qualité mesurables (métriques) qui aideront le clinicien à statuer si le séquençage est d'assez bonne qualité pour rendre un résultat sur l'analyse. Lorsqu'une variation candidate est retrouvée pour un patient donné, la vérification de la qualité des données est importante, mais pas cruciale, car le résultat sera vérifié par une technique orthogonale de référence dans la quasi-majorité des cas. En revanche, lorsque rien n'a été trouvé, il faut s'assurer que les données étaient d'assez bonne qualité pour statuer que l'analyse s'est déroulée dans des conditions permettant de rendre un résultat négatif. Les métriques sont calculées à partir de différents types de fichiers utilisés ou produits par le pipeline (FASTQ, BAM, VCF ...) en fonction de ce qui est mesuré. Enfin, de nombreuses données de qualité peuvent être intégrées dans l'outil MultiQC[122].

2.2.4.2 Calcul des métriques

De nombreuses métriques sont calculées au sein de ce module, celles-ci sont récapitulées *Table 2.18*.

Outil	Type de fichier en entrée	Fonction
FastQC[86]	FASTQ	Contrôle qualité des données brutes FASTQ (score de qualité des bases, [61] contenu en adaptateur, duplicats ...)
GATK CollectInsert SizeMetrics	BAM	Détermination de la distribution de la taille d'insert de la librairie
GATK[61] CollectHsMetrics	BAM	Métriques relatives à l'efficacité de la capture
GATK CollectAlignment SummaryMetrics	BAM	Métriques relatives à la qualité des alignements des *reads*
GATK[61] CollectGcBiasMetrics	BAM	Métriques relatives aux biais de GC
GATK[61] DepthOfCoverage	BAM	Métriques relatives à la profondeur de couverture
annotate_sample_interval_summary_with_genes.py	DepthOfCoverage	Métriques relatives à la profondeur de couverture par intervalle de la cible / gène
SAMtools[58] idxstats	BAM	Métriques relatives à la qualité des alignements des *reads*
Samtools[58] flagstat	BAM	Statistiques sur les drapeaux des fichiers BAM relatifs à la qualité des alignements des *reads*
Qualimap[123] bamqc	BAM	Métriques relatives à la profondeur de couverture
featureCounts[124]	BAM	Caractéristiques génomiques des *reads* alignés

TABLE 2.18 – Outils de calcul de métriques de qualité utilisés dans le module *metrics*.

2.2.4.3 Rapport qualité

La quasi-totalité des outils exécutés dans le module metrics peut avoir leurs résultats d'intégrés dans l'outil MultiQC[122]. De plus, certains fichiers calculés lors d'étapes précédentes (rapport de recalibration des bases de GATK, données d'annotation fonctionnelle de SnpEff) sont également intégrables dans MultiQC. Ce dernier permet la création d'un rapport au format HTML, dynamique, configurable, organisé en plusieurs volets pour chaque outil, voir *Figure 2.42*. Les résultats des différents outils sont représentés sous forme de graphiques ou de tableaux dynamiques, pour la plupart. Le module *metrics* exécute l'outil MultiQC plusieurs fois, une fois par échantillon et une fois pour l'ensemble du *batch*. Lorsque le *batch* est trop important en taille, certains graphiques perdent leurs propriétés dynamiques au profit d'images fixes.

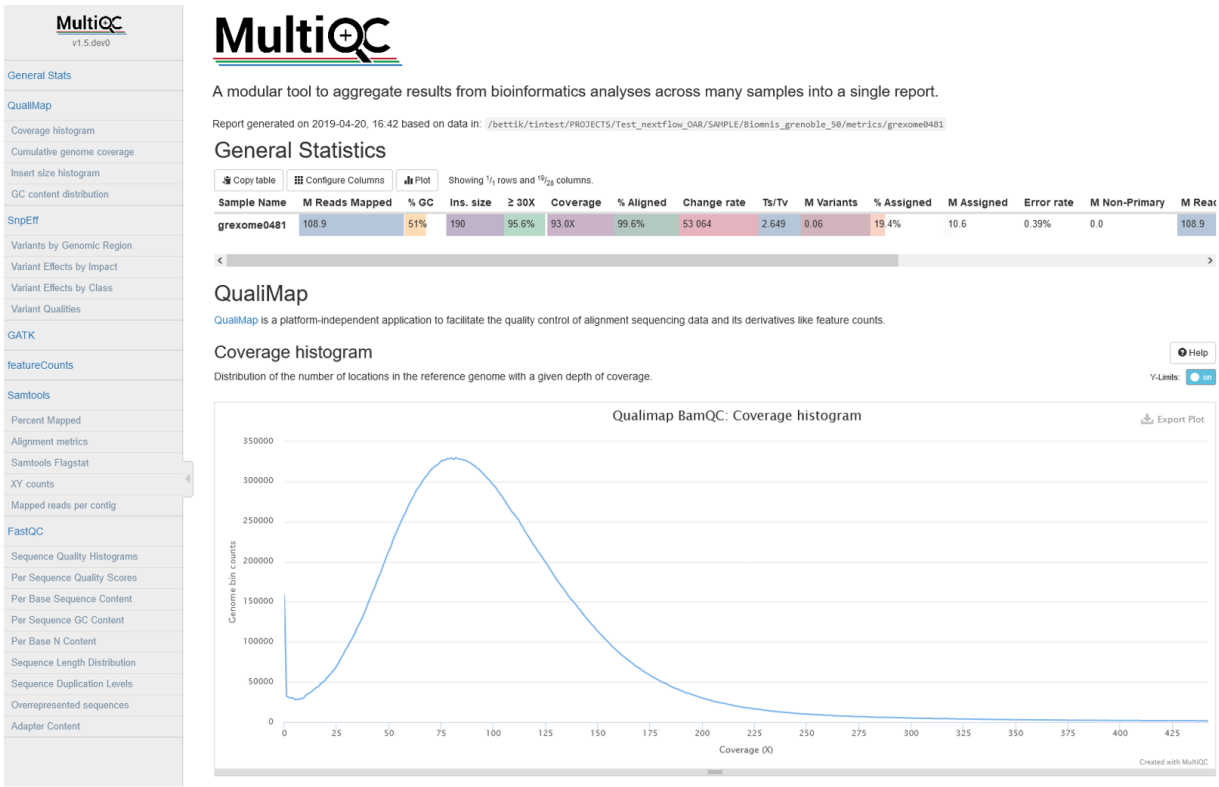


FIGURE 2.42 – Rapport MultiQC représentant les statistiques générales d'un échantillon et les données de l'outil Qualimap.

Le rapport MultiQC ne dispense pas totalement la consultation de certains fichiers textuels calculés dans le module, mais permet un suivi rapide et graphique sur de nombreuses métriques importantes.

2.3 Validation du pipeline sur données de WES

2.3.1 Validation sur données de référence

La validation des processus sur des données de référence est une part très importante de la qualification de tout outil bioinformatique[97]. Dans le cadre de la validation du pipeline bioinformatique, une analyse de non-régression du pipeline déjà en place chez Eurofins utilisant l'outil GATK[61] en version 3 contre le pipeline Lygrexome utilisant l'outil GATK en version 4, a été effectuée. Le test de non-régression a porté sur le *benchmarking* de détection de SNV et d'indels sur les régions de haute confiance du GIAB version 3.3.2 (qui était la version la plus récente à l'époque des tests) de l'individu HG002, en version hg19 du génome. Ce *benchmark* a été effectué à l'aide de l'outil Hap.py[97] avec les fichiers produits par les deux pipelines et le set de référence précédemment cité, édité par le GIAB[89]. La version du génome hg19, bien que considérée comme obsolète, est la seule version compatible du pipeline GATK3 et est encore largement utilisée dans les laboratoires de biologie médicale. À titre indicatif ont été indiqués les résultats obtenus pour la condition GATK 4 et la version hg38 du génome pour le même ensemble de données de référence. Il est difficile de comparer les valeurs obtenues pour les deux différentes versions du génome, car il est impossible de savoir si les différences peuvent être imputées aux différences entre les deux versions du génome de référence ou à la méthodologie bioinformatique. Les résultats de cette étude sont visibles *Table 2.19*.

	Type	Filtre	Vérité Total	Vérité VP	Vérité FN	Requête Total	Requête FP	FP génotype	Rappel	Précision
GATK3 hg19	INDEL	ALL	1 483	1 284	199	1 527	242	72	0.865	0.841
	SNP	ALL	23 377	23 298	79	23 606	311	10	0.996	0.986
GATK4 hg19	INDEL	ALL	1 483	1 277	206	1 976	696	65	0.861	0.647
	SNP	ALL	23 377	23 300	77	23 791	494	8	0.996	0.979
GATK4 hg38	INDEL	ALL	1 402	1 059	343	1 953	243	32	0.755	0.814
	SNP	ALL	23 131	22 988	143	31 452	706	10	0.993	0.970

TABLE 2.19 – Récapitulatif du test de non-régression.
Test de non régression entre le pipeline Biexome basé sur GATK 3 version hg19 ainsi que le pipeline Lygrexome basé sur GATK 4 version hg19 et hg38, effectués avec Hap.py. VP : vrai positif, FN : faux négatif, FP : faux positif, FP génotype : variant appelé à une position existante dans le set de vérité, mais avec le mauvais génotype.

Les valeurs les plus importantes de cette table sont les taux de rappel et de précision, dont la méthode de calcul est schématisée *Figure 2.43*. Même si la précision est un paramètre important qui augmente notamment lorsque le nombre de faux positifs est minimisé, c'est le rappel qui est la statistique cruciale. En effet, elle représente le nombre de variations cliniquement pertinentes retrouvées avec succès. Les pipelines de détection de variations auront toujours pour but de maximiser leurs performances en rappel, quitte à sacrifier un peu de précision s'il le faut.

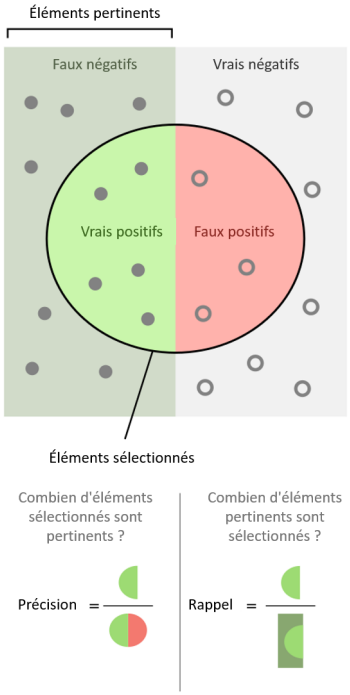


FIGURE 2.43 – Méthode de calcul de la précision et du rappel à partir de données de détection de variations face à un ensemble de données de référence.
Adapté d'après, Wikipédia F1-Score.

Le set de référence version hg19 de l'individu HG002 comporte 23 377 SNP et 1 483 indels issues des zones de haute confiance à retrouver. Les taux de rappel, que ce soit pour le pipeline utilisant la version 3 de l'outil GATK ou la version 4 sont extrêmement similaires et sont au niveau de l'attendu (>95%). Il est communément admis que le taux de rappel des indels soit plus faible que celui des SNP, de par leur plus grande difficulté à être détectés correctement avec la technologie de séquençage Illumina. La seule différence notable entre les deux conditions est l'apparition d'un nombre d'indels faux positif très élevé pour GATK4 (696) hg19 par rapport à GATK3 (242). Il est possible qu'une partie de ces faux positifs n'en soit pas en réalité, mais plutôt des données manquantes du set de vérité. En effet, la version du set de référence pour cette étude étant assez ancienne (2017), celui-ci a été constitué de données appelées, entre autres, à l'aide de l'outil HaplotypeCaller de l'outil GATK version 3. L'algorithme de celui-ci s'étant perfectionné dans sa dernière itération, il est tout à fait possible que l'outil dans sa version la plus récente détecte des variations qu'il ne détectait pas auparavant. Ainsi ces indels nouvellement détectés seraient absents de l'ensemble de variations de vérité pour l'individu HG002. L'important est que le taux de rappel entre les deux conditions est resté stable, malgré la baisse en précision.

Le fait que la version 3 de GATK ne soit plus maintenue au profit de la version 4 et que les mesures de rappel soient stables entre les deux conditions permet de conclure que le test de non-régression est un succès malgré la baisse de précision au niveau de la détection des indels pour la condition GATK4.

En revanche, il est intéressant de se pencher sur les zones aveugles des deux pipelines, à savoir les SNP et indels faux négatifs. D'un nombre relativement équivalent entre les deux conditions, nous avons déterminé que la plupart des faux négatifs étaient communs aux deux conditions, voir *Figure 2.44*. Les données d'entrée des deux pipelines sont des FASTQ produits par le protocole de génération de données d'Eurofins Biomnis utilisant le kit *Twist Human Core Exome*. L'explication la plus probable est que ces variations soient dans les zones aveugles du kit de capture (peu ou pas hybridées par les sondes de capture). Il est également possible que ces variations, soit, n'existent pas ou qu'elles ne soient pas détectables par la stratégie algorithmique de l'outil HaplotypeCaller.

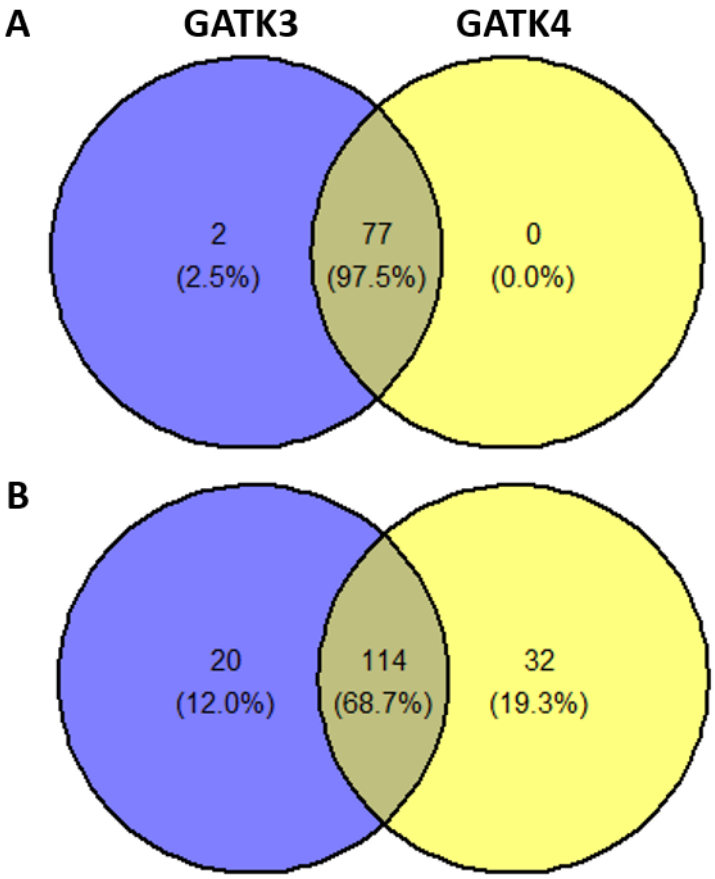


FIGURE 2.44 – Variants faux négatifs communs entre les conditions GATK3 et GATK4 hg19.
A : SNP, B : Indels.

Il est intéressant de noter que le nombre de faux négatifs est assez constant pour les conditions en hg19 par rapport à la condition hg38, où le nombre de faux négatifs de la seconde est en hausse par rapport à la première. Encore une fois ces disparités peuvent être dues au contenu du set de vérité et mériteraient d'être actualisées avec une version plus récente des données de référence.

2.3.2 Validation sur données cliniques

Ce pipeline a été utilisé depuis l'année 2019 pour traiter diverses cohortes de patients produites par les différentes parties de ma thèse. Eurofins Biomnis a utilisé ce pipeline sur plus de 3 512 exomes dont 2 507 cas Index. Le recrutement des patients séquencés chez Biomnis est très divers. La plupart des exomes qui y sont séquencés proviennent de cliniciens français qui prescrivent un exome après avoir déjà fait quelques analyses de routine en interne (biochimie, caryotype, ACPA, panels ...). De plus, les indications des patients qui sont adressés varient, mais une grande partie de ces exomes proviennent de patients avec des déficiences intellectuelles et des maladies rénales. Le taux de diagnostic sur cette cohorte est de l'ordre de 25 %.

Le reste des exomes reçus chez Eurofins-Biomnis proviennent de prescripteurs étrangers et n'ont en général pas eu d'analyse génomique au préalable. Les analyses révèlent parfois des diagnostics comme des aneuploïdies ou de très grands CNV qui auraient pu être détectés par des technologies comme l'ACPA ou le caryotype. Pourtant, ces anomalies sont détectées avec succès par l'exome, ce qui prouve que la stratégie de "l'*exome-first*" est envisageable. Le taux de diagnostic sur cette cohorte est d'environ 50 %.

Nombre d'exome	300+
Nombre de cas index séquencés	258
Exomes rendus négatifs	93
Exomes rendus positifs	108
Échecs techniques	1
En cours d'interprétation	38
En cours de validation	18

TABLE 2.20 – Nombre d'exomes du Service de Génétique, Génomique et Procréation analysés par le pipeline Lygrexome durant ma thèse.

Une grande campagne de réanalyse des données d'exomes du Service de Génétique, Génomique et Procréation de l'Hôpital Couple Enfant de Grenoble, produits pour la majorité par Eurofins Biomnis et l'IBP de Grenoble est en cours avec mon pipeline. Les résultats obtenus à ce jour sont visibles *Table 2.20*.

2.4 Déploiement, utilisation et prise en mains par les utilisateurs du pipeline

2.4.1 Déploiement sur l'infrastructure Eurofins Biomnis

Durant l'été 2020, lorsque le pipeline est entré dans une phase où son développement était quasi finalisé, une passation du code s'est organisée avec Aurore Perdriau, ingénieure en Bioinformatique et Jean-François Taly, chef de la division Bioinformatique clinique (BioIT), tous deux employés par Eurofins Biomnis. Le code et les bases de données nécessaires à son bon fonctionnement ont été envoyés, archivés et déposés sur l'infrastructure industrielle.

L'infrastructure de calcul informatique et de production de données de séquençage d'Eurofins Biomnis étant très particulière, il a été décidé que le pipeline développé au cours de cette thèse et le pipeline déployé sur l'infrastructure industrielle bifurquent, bien qu'étant toujours très proches. Le déploiement s'est fait avec succès assez rapidement, grâce à Nextflow intégrant l'ordonnanceur SLURM[125] installé sur l'infrastructure Biomnis et aux conteneurs Singularity[81] déjà utilisés par le pipeline précédemment utilisé en routine. De plus, deux modules déjà en place dans ce dernier, le module combinant démultiplexage et *basecalling*, ainsi que le module d'identito-vigilance, ont été intégrés dans le pipeline nouvellement appelé Cifrexome.

Depuis, les modifications d'intérêts apportés au pipeline Lygrexome sont intégrées en miroir par Aurore Perdriau et Fanny Ponce, également ingénieure en Bioinformatique.

2.4.2 Infrastructure HPC-CHU

2.4.2.1 Mise en place de l'infrastructure HPC-CHU

Dans le cadre de ma thèse, j'ai été sollicité ainsi que Valentin Klein, ingénieur en Bioinformatique de la Plateforme AURAGEN, par Émilien Beaussart, directeur du secteur RS4 (réseau et sécurité) au sein de la Direction des Services Numériques (DSN) du CHU Grenoble Alpes, afin organiser un *cluster* de calcul à partir de machines décommissionnées. Ont été mis à notre disposition :

— Huit serveurs équipés de processeurs Intel Xeon X5560, 4 cœurs, 4 threads, cadencés à 2,80-3,20 GHz et 128 Go de RAM, afin d'être installés comme nœuds esclaves.

— Un serveur équipé d'un processeur Intel Xeon E5-2450, 8 cœurs, 16 threads cadencé à 2,10-2,90 GHz et 192 Go de RAM, afin d'être installé comme nœud maître.

— Un espace de stockage NFS de 20 To non redondé pour le stockage des données.

L'hébergement, la connexion des serveurs ainsi que l'installation des systèmes des exploitations a été pris en charge par Nicolas Lecertisseur, ingénieur hospitalier informaticien à la DSN. Valentin Klein et moi-même nous sommes occupés d'organiser ces différents serveurs en un *cluster*. Pour cela, a été installé l'ordonnanceur SLURM[125], le gestionnaire de conteneurs Singularity[81] et certains paquets usuels au bon fonctionnement d'un serveur. Nous avons également connecté les différents serveurs à l'espace de stockage et fait en sorte que les espaces */home* des utilisateurs soient hébergés sur l'espace de stockage NFS et accessibles à tous les nœuds. De plus, nous avons documenté la procédure d'installation sur un dépôt git hébergé par le CHU (http://git-bioinfo/) et rendu le déploiement d'hypothétiques futurs nœuds semi-automatisés en utilisant une librairie Python spécialisée dans le déploiement de serveurs appelée Fabric (https://www.fabfile.org/).

Une fois mis en service, ce *cluster* a permis l'analyse de plusieurs *runs* de données de séquençage d'exome. Familièrement nommé "Brouette" car "Ça roule, mais il faut le porter à bout de bras", celui-ci a servi comme preuve de concept de la faisabilité d'un *cluster* hospitalier pour la bioinformatique et plus largement pour le traitement informatique de données médicales. Cela a permis le déblocage de fonds par la DNS afin d'acheter plusieurs machines plus récentes et taillées pour le calcul bioinformatique pour les années à venir. Ont été mis à notre disposition :

— Trois serveurs équipés de deux processeurs AMD EPYC 7742 chacun, 64 cœurs, 128 threads cadencés à 2,25-3,40 GHz et 1 To de RAM, afin d'être installés comme nœuds esclaves.

— Un GPU de calcul scientifique GPU PNY Tesla V100 32 Go, payé avec la dotation de ma thèse, couplé à l'un des nœuds esclave.

— Un espace de stockage NFS supplémentaire non redondé de 43 To.

L'organisation du *cluster* HPC-CHU est visible *Figure 2.45*. Les trois nouveaux serveurs ont été déployés en utilisant les scripts *Fabric* développés lors de la précédente intégration de nœuds. Les deux ensembles de serveurs ont été assignés à deux queues de calculs différentes gérés par SLURM. En effet, l'énorme différence de puissance de calcul entre les deux sous-ensembles de serveurs induit la nécessité de pouvoir choisir l'un ou l'autre au moment de lancer calcul. Les serveurs les plus anciens ont alors été placés dans la queue "Brouette" et les plus récents dans la queue "Servoz", dans une seule même entité nommée HPC-CHU.

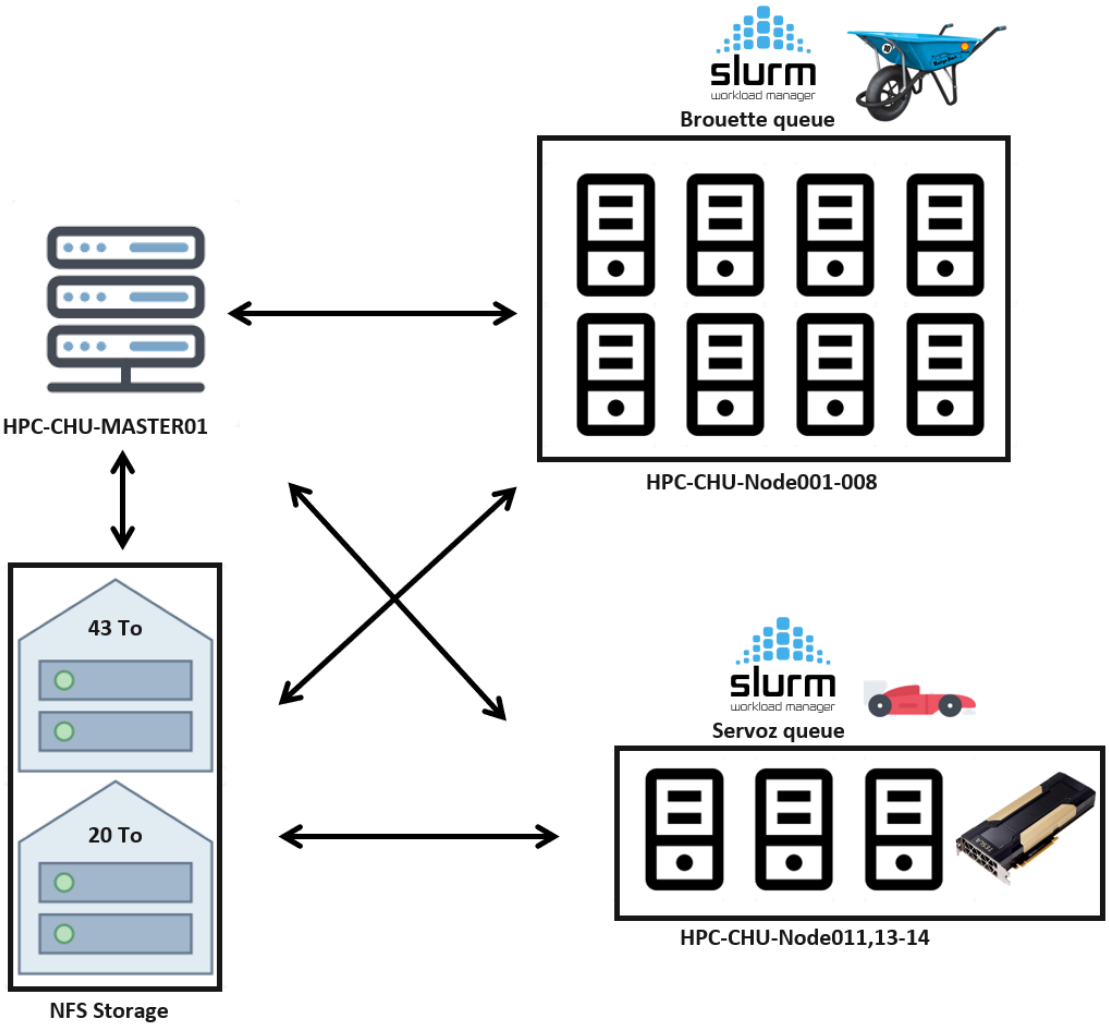


FIGURE 2.45 – Organisation du *cluster* HPC-CHU et des queues Brouette et Servoz.

2.4.2.2 Test de l'infrastructure HPC-CHU

Depuis sa mise en place, l'infrastructure HPC-CHU a été utilisée par de nombreux utilisateurs, même si j'en reste le principal, voir *Figure 2.46*. La plupart des calculs effectués sur cette infrastructure sont relatifs à de la manipulation et de l'analyse de données de séquençage. En revanche, ces données ne sont pas exclusivement humaines. Une activité d'analyse de données de séquençage de SARS-CoV 2 en routine a notamment commencé depuis le début de l'année 2021.

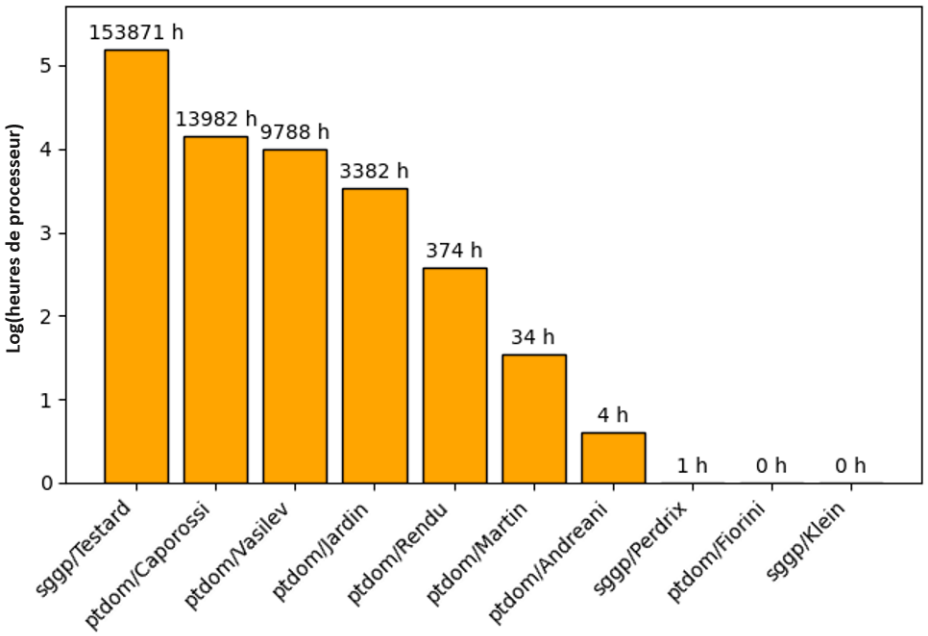


FIGURE 2.46 – Utilisation du *cluster* HPC-CHU (heures de CPU) du 01/01/21 jusqu'au 03/09/21.

Un test de performance a été effectué sur certaines des infrastructures sur lesquelles le pipeline a été déployé. Le test consistait en l'analyse de données d'un *run* de neuf exomes avec tous les modules du pipeline d'exécutés, voir *Table 2.21*. C'est l'infrastructure Servoz qui a été la plus rapide en temps réel, mais pas en temps CPU. Cela peut être expliqué par le fait que les CPU de l'infrastructure Dahu ont une horloge de cœur turbo plus élevée que celles des CPU de l'infrastructure Servoz, mais aussi, car le *cluster* Dahu est partagé et que certains processus aient pu être mis en attente le temps d'en avoir une place de disponible, bien que le pipeline ait été lancé en heure creuse pour éviter cela au maximum. Les résultats obtenus par la queue Servoz étaient tout de même très satisfaisants, car au niveau d'un *cluster* HPC régional. Brouette est l'infrastructure la plus lente comme attendue.

Queue	Temps réel	Heures de CPU
Dahu	7h 49min	359
Brouette	16h 5mn	431
Servoz	6h 12	433

TABLE 2.21 – Temps d'exécution du pipeline sur différentes infrastructures lors de l'analyse d'un *run* test.

2.4.2.3 Déploiement sur l'infrastructure HPC-CHU

Lors de la mise en place de Brouette, un des premiers tests de l'infrastructure à été le déploiement du pipeline Lygrexome et l'analyse d'un *run* de test. Le déploiement s'est fait extrêmement simplement, de la même manière que sur l'infrastructure Eurofins Biomnis. Le code a été cloné sur HPC-CHU depuis le dépôt GitLab (https://gitlab.com/chu_grenoble/cifre_biomnis/Lygrexome) où il était stocké et les fichiers et bases de données de référence ont été archivés, compressés, envoyés, depuis le *cluster* Dahu, puis décompressés sur HPC-CHU. Après transfert, décompression des références nécessaires au bon fonctionnement du pipeline ainsi que quelques modifications dans le fichier de configuration du pipeline, celui-ci a été fonctionnel sur la nouvelle infrastructure en moins d'une heure grâce à l'installation de Singularity et la compatibilité par défaut de Nextflow avec SLURM. Depuis, des profils de configuration en fonction de l'infrastructure sur laquelle le pipeline est exécuté (Dahu, Brouette, Servoz) ont été ajoutés au pipeline.

2.4.2.4 Utilisation du pipeline sur l'infrastructure HPC-CHU

Depuis le milieu de l'année 2020, une activité de séquençage d'exome a été mise en place conjointement entre l'Unité Transversale de Production de Biologie Moléculaire (UTP-BM) de l'Institut de Biologie et Pathologie (IBP) de Grenoble, qui produit les séquences, le Service de Génétique, Génomique et Procréation de l'Hopital Couple Enfant ainsi que de l'unité Biochimie Génétique et Moléculaire (BGM) du CHU de Grenoble. Les patients recrutés par l'HCE sont en général des situations de diagnostic prénatal référées par le Centre Pluridisciplinaire de diagnostic prénatal. La prescription est effectuée en trio dans ce cas. Les patients recrutés par la BGM sont d'indications plus diverses et majoritairement des prescriptions de recours pour le diagnostic de maladies neuromusculaires, ou d'infertilité sévère.

Cette activité correspond à douze exomes séquencés tous les mois à analyser au plus vite. En général, les données sont disponibles le lundi et doivent être analysées pour le mercredi ou jeudi suivant afin de laisser le temps aux médecins pour leur revue de cas (*staff*) qui a lieu les vendredis, voir *Table 2.22*. À ce jour, plus de 175 exomes produits par l'UTP-BM ont été analysés par mon pipeline.

	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Préparation de librairie							
Séquençage							
Transfert des données							
Analyse bioinformatique							
Interprétation et rendu							

TABLE 2.22 – Diagramme de Gantt du déroulement sur deux semaines consécutives d'un séquençage d'exome jusqu'à son rendu en passant par son analyse sur le *cluster* HPC-CHU.

Depuis juin 2021, le rôle de lancer le pipeline sur les données d'exome et de transmettre les données aux différents participants de ce projet a été transféré à l'ingénieur hospitalier bioinformaticien Ivaylo Vasilev, membre de l'UTP-BM. La passation de l'activité, c'est fait très simplement. Le code a été cloné depuis le dépôt git vers le répertoire */home* d'Ivaylo, puis un accès vers les fichiers et les bases de données de référence a été ouvert à l'aide de liens symboliques. Il ne restait plus qu'à modifier quelques valeurs du fichier de configuration et l'analyse était exécutable par Ivaylo. Nous échangeons régulièrement sur les modifications et les avancées qui sont régulièrement appliquées au pipeline et ses dépendances.

2.5 Discussion

Le pipeline Lygrexome a été développé, maintenu et enrichi en fonctionnalités tout au long de cette thèse. Il a permis l'analyse ou la réanalyse périodique de plusieurs milliers d'exomes cliniques ou de recherche, de plusieurs centres et a été le support de l'activité d'interprétation de plusieurs équipes de clinicien. Le pipeline a permis d'atteindre un taux de diagnostic au moins équivalent à l'attendu, de rattraper le retard organisationnel du CHUGA et de poser des bases solides à une routine diagnostique. En cela, le développement de cet outil est un succès. Malgré tout, même si le pipeline développé au cours de cette thèse est parfaitement fonctionnel et qu'il remplit nombre des fonctions attendues par les deux parties constituantes de ce doctorat, il pourrait tout de même être amélioré en plusieurs points.

Tout d'abord, bien que le lancement du pipeline soit semi-automatisé, il n'en demeure pas moins qu'il reste des étapes manuelles nécessaires pour son exécution. Ces étapes sont bien entendu des risques d'erreurs supplémentaires lors du lancement des analyses. Il est déjà arrivé que le pipeline ne soit pas lancé avec les bons arguments et que soit, l'analyse échoue, soit les résultats produits soient de mauvaise qualité. Cela n'est pas acceptable dans le cadre d'une analyse de routine clinique, qui plus est, parfois urgente dans le cas des analyses de données de WES prénatal. Pour remédier à cela, il faudrait développer un système d'interfaçage entre les systèmes de gestion de laboratoire (SGL) du CHU de Grenoble, de l'UTP-BM ou d'Eurofins Biomnis en lien avec les équipes concernées. Ce travail pour les SGL grenoblois avait été envisagé et concrétisé par le recrutement d'une étudiante alternante en Bioinformatique, mais malheureusement, le travail n'a pas pu être mené à son terme. De plus, il était également prévu de développer une connexion entre le pipeline et la base de données regroupant les phénotypes cliniques des patients grenoblois organisés à l'aide de l'outil PhenoTips[126]. Ce travail avait également été assigné à l'étudiante en alternance et n'a donc pas pu être mené à bien.

Suite à l'automatisation et l'extraction des phénotypes cliniques des patients, il était prévu le développement et l'intégration d'un module de priorisation des variations à l'aide des informations phénotypiques sous la forme de termes de l'ontologie HPO[79]. L'outil Phrank[127] avait été sélectionné pour cette tâche, mais son intégration n'a jamais eu lieu.

Lors du déploiement de l'outil sur les différentes infrastructures sur lequel il a été installé, le code a été très facilement exporté grâce au dépôt versionné de la plateforme GitLab, mais certaines de ses dépendances doivent être transférées *via* l'outil *scp* ou méthode équivalente. Le pipeline est donc plus transportable que portable. Plusieurs types de fichiers subsistent parmi ces dépendances. Tout d'abord, les bases de données et de connaissance de référence médicale. Certaines d'entre elles sont directement téléchargeables en licence libre et en libre accès sur le site internet dont elles dépendent, d'autres sont payantes ou nécessitent une licence spécifique pour être utilisées et enfin, certaines ont été créés de toutes pièces à partir de données analytiques et sont spécifiques aux données traitées, comme les cohortes de normalisation pour le SNP *calling* (séquenceurs, kits). Toutes ne sont donc pas téléchargeables ou constructibles automatiquement, néanmoins il serait utile de mettre au point une procédure pour le téléchargement et la construction des dépendances du pipeline, même si celle-ci ne peut pas être automatisable en intégralité. De plus, les références du pipeline mériteraient d'être mises à jour, car la dernière date de l'année 2020.

De plus, la plateforme GitLab permet la mise en place de tests d'intégration, c'est-à-dire, l'exécution de tests qui permettent la vérification des performances et de la fiabilité du code après sa modification. Bien que des tests périodiques comme les tests de non-régression soient effectués de manière ponctuelle lors de modifications majeures du pipeline, une méthodologie de tests automatiques, déclenchés après certains types de modifications permettrait de valider le bon fonctionnement de l'outil au cours de son cycle de vie.

Enfin, bien que les résultats produits par ce pipeline soient pertinents biologiquement et de bonne qualité, le rapport final pour l'interprétation n'est pas très facile à appréhender pour les utilisateurs (*user-friendly*). Une des pistes principales pour rendre l'interprétation plus agréable et ainsi plus rapide par les cliniciens est l'intégration des résultats d'appel de variations dans des interfaces permettant la filtration des variations sur les champs annotés des fichiers VCF, comme l'outil Cutevariant[128].

Malgré tout cela, ce pipeline a été choisi comme la solution utilisée par les différents centres grenoblois analysant les exomes produits par la plateforme de l'UTP-BM à la préférence des solutions déjà en place, unifiant ainsi les pratiques d'analyses des données de la communauté médicale grenobloise.

Analyse de données WGS tumorales

3.1 Motivation

Au cours de cette thèse, mon aide a été sollicitée par Sophie Park et Mathieu Meunier, hématologues au CHU de Grenoble. L'objectif de cette collaboration était l'analyse de données de modèles de cancer *in vitro* obtenues par différentes techniques de biologie cellulaire et moléculaire dont différentes technologies omiques. L'analyse conduite lors de cette thèse avait pour objet l'exploitation de données somatiques pairées obtenus par WGS Illumina et WGS lectures liées (*linked reads*) Chromium 10X Genomics (https://www.10xgenomics.com/products/linked-reads).

Cette étude avait plusieurs objectifs. Premièrement, de me familiariser avec les analyses bioinformatiques génomiques usuelles en oncogénétique et en cancérologie. Deuxièmement, de me faire découvrir la technologie Chromium 10X Genomics. Enfin, de me présenter un cas d'usage, de quel pouvaient être les besoins en Bioinformatique à l'hôpital dans le cadre de projets de recherche. Tout cela s'intégrait parfaitement dans l'objectif de cette thèse.

3.2 Contexte de l'étude

Des cellules souches précurseures hématopoïétiques immatures (CD34+) de trois donneurs sains ont été exposées à des vésicules extracellulaires de petite taille contenant des exosomes sains ou précancéreux issus de cellules stromales mésenchymateuses saines ou dérivées de syndrome myélodysplasique pendant 7 jours. Cette étude avait pour objectif de mieux comprendre la physiopathologie des syndromes myélodysplasiques et notamment l'effet du microenvironnement sur leur développement *via* l'étude de l'effet des exosomes précancéreux sur ces cellules entraînant une hématopoïèse défectueuse.

Pour cela deux types de séquençages ont été effectués. Tout d'abord le WGS Illumina de trois patients selon les conditions, cellules exposées à des exosomes sains ou précancéreux et le WGS Chromium 10X Genomics selon les mêmes conditions. L'objectif était de trouver des variations signatures de syndromes myélodysplasiques communes aux trois patients. Les approches par séquençage ne constituaient qu'une partie des investigations réalisées sur ces modèles cellulaires.

3.3 Matériel et méthode

Les données utilisées au cours de cette analyse ont été séquencées en 2 x 150 paires de bases, *short-reads paired-end* à l'aide du kit de préparation de librairies sans PCR d'ADN TruSeq d'Illumina, conformément aux instructions du fabricant. Après normalisation et contrôle de qualité, les bibliothèques qualifiées ont été séquencées sur un HiSeqX5 par la plateforme de séquençage du Centre National de Recherche en Génomique Humaine (CNRGH) d'Évry-Courcouronnes, centre de référence national en génomique. Les fichiers ont été récupérés depuis le FTP du CNRGH au format FASTQ.

Les données issues de librairies Chromium ont été transférées après avoir été démultiplexées par l'outil Long Ranger version *2.2.2* et de la commande *mkfastq*. Les données ont ensuite été alignées contre le génome hg38 *contigs* canoniques et l'appel de SNV, de SV, ainsi que le phasage des haplotypes ont été obtenus grâce à la commande *wgs*.

Les données issues de séquençage Illumina ont été traitées par le module *fastq2bam* du pipeline d'analyse de données DNA-seq Illumina pour leur alignement contre le génome version hg38 *contigs* canoniques. Les contrôles qualité sur ces données ont été effectués à l'aide des outils usuels décrits précédemment. L'ensemble des données patients séquencées disposaient d'au moins 30X de couverture chacune.

La détection de variations de petite taille somatiques a été effectuée sur les données Illumina et les données Chromium alignées à l'aide de l'outil GATK[61] version 4.0.9.0 et notamment l'outil Mutect2[129]. Pour cela, le *Broad Institute* conseille une marche à suivre pour obtenir des appels de SNV et d'indels de bonne qualité à partir de données somatiques pairées, visible *Figure 3.47*.

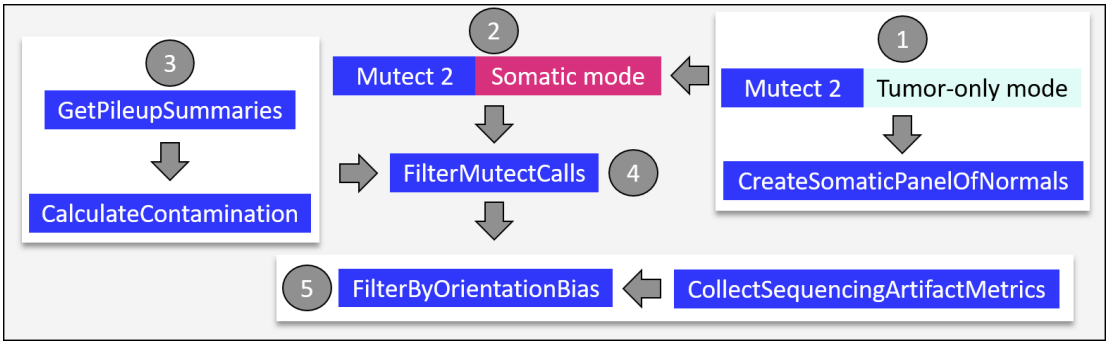


FIGURE 3.47 – Étapes et ordre à suivre pour la détection de variations de petite taille à partir de données somatiques.
Adapté d'après, *Call somatic mutations using GATK4 Mutect2.*

La première étape pour l'appel de variation de petite taille somatique est la constitution d'un panel de personnes normales (*Panel of Normal* ou PoN). Un PoN est un type de ressource utilisé dans l'analyse des variants somatiques. Il est constitué d'échantillons normaux, c'est-à-dire obtenus à partir de tissus sains supposés ne pas avoir subi d'altérations somatiques. Le PoN a pour rôle la caractérisation des artéfacts techniques récurrents afin de pouvoir les éliminer lors de l'appel de variation et ainsi améliorer la précision de l'analyse. Dans notre cas, nous avions à notre disposition 53 données WGS Illumina séquencées par le CNRGH similaires aux données utilisées au cours de cette étude et de treize fichiers seulement issus de WGS Chromium afin de constituer nos PoN. Deux PoN différents ont été construits à l'aide de l'outil Mutect2 en mode tumeur seule, en discriminant par la technologie de séquençage.

Ensuite, l'appel des variations a été effectué à l'aide de l'outil Mutect2 en mode somatique, c'est-à-dire à partir de données alignées somatiques et supposées saines issues de chaque patient (paire normale, tumeur). En plus des fichiers d'alignement, l'outil nécessite le PoN, ainsi qu'un fichier de référence de fréquences des variations germinales dans la population. Dans le cadre de cette étude, nous avons utilisé le fichier VCF du projet GnomAD[70], fournit dans le *bundle* GATK, contenant les données de plus de 200 000 exomes et 16 000 génomes, de concert avec un paramètre de Mutect2 permettant d'exclure toutes les variations présentes dans les données patient présentes à une fréquence supérieure à 2,5 x 10^{-6}.

De manière concomitante, les outils GetPileupSummaries et CalculateContamination de GATK4 sont utilisées pour estimer la contamination inter-échantillon. Il est nécessaire de tenir compte du taux de contamination estimé par cet outil lors de la filtration des variations sur leur taux de fréquence allélique alternatif.

Les variations obtenues à l'aide de l'outil Mutect2 sont filtrées à l'aide de l'outil FilterMutectCalls. Quatorze filtres sur différents critères sont appliqués aux fichiers de variations (qualité, fréquence allélique, nombre de *reads* supportant l'information, présence dans le *panel of normal* ...). Les variants qui passent l'ensemble de ces seuils de filtration ont la mention *PASS* dans le champ *FILTER*. Enfin, une étape optionnelle a été réalisée, l'utilisation de l'outil FilterByOrientationBiais afin de filtrer certains types d'artéfacts caractéristiques détectés au préalable par l'outil CollectSequencingArtifactMetrics.

L'annotation des fichiers VCF produits a été effectuée avec les outils SnpEff[73] et SnpSift[74] afin d'ajouter les méta-informations de la base de données COSMIC[120] notamment.

3.4 Résultats

3.4.1 Données Chromium 10x Genomics

Les données Chromium produites avec l'outil Long Ranger *wgs* ont été soumises à interprétation pour les biologistes, que ce soit pour les variations de petite taille ou les variations de structure. Ces données n'ont pas permis d'amener un résultat significatif et n'ont finalement pas été associées à la publication. Les données produites par l'outil Long Ranger *align* ont été soumises aux mêmes étapes de détection de variations somatiques de petite taille à l'aide de Mutect 2. Les résultats se sont trouvés être beaucoup plus bruités par rapport aux données WGS Illumina seules, sans doute à cause d'un PoN trop peu fourni. Aucun résultat issu de cette technologie n'a été intégré à l'article.

3.4.2 Appel de variation sur données de WGS Illumina

Les données WGS Illumina ont été traitées comme indiqué selon les bonnes pratiques GATK pour l'appel de variation de petite taille somatiques, depuis les fichiers FASTQ jusqu'aux fichiers VCF filtrés. Cela a permis de produire 3 fichiers VCF par patient contenant SNV et Indels. Les nombre de variations totales et filtrées sont visibles *Table 3.23*.

	Patient 1	Patient 2	Patient 3
Nombre de variants totaux	89476	87791	93503
Nombre de variants filtrés	86680	87055	92686
Nombre de variants PASS	796	736	817

TABLE 3.23 – Nombre de variations issues des fichiers VCF des 3 patients traités au cours de cette analyse. Certains variants peuvent avoir été filtrés pour plusieurs raisons.

Seul environ 1% des variations des fichiers VCF des trois patients traités passent avec succès l'ensemble des filtres de qualité (*PASS*). Les variations restantes sont pour une majorité (86.3% de moyenne entre les trois patients) des SNV, 9.5% des indels et 4,1% de polymorphismes multinucléotidiques (lorsque la séquence de référence et la séquence alternative sont supérieures à une paire de bases et qu'elles sont de même longueur), voir *Figure 3.48*. Les variations détectées sont majoritairement des transitions (64,18%) plutôt que des transversions ayant lieu majoritairement dans des régions non codantes. Des variants faux-sens ont été identifiés dans les régions codantes de quatre gènes (EEF1A1, WDR66, KIR3DL3 et ANKRD1) parmi les données des trois patients.

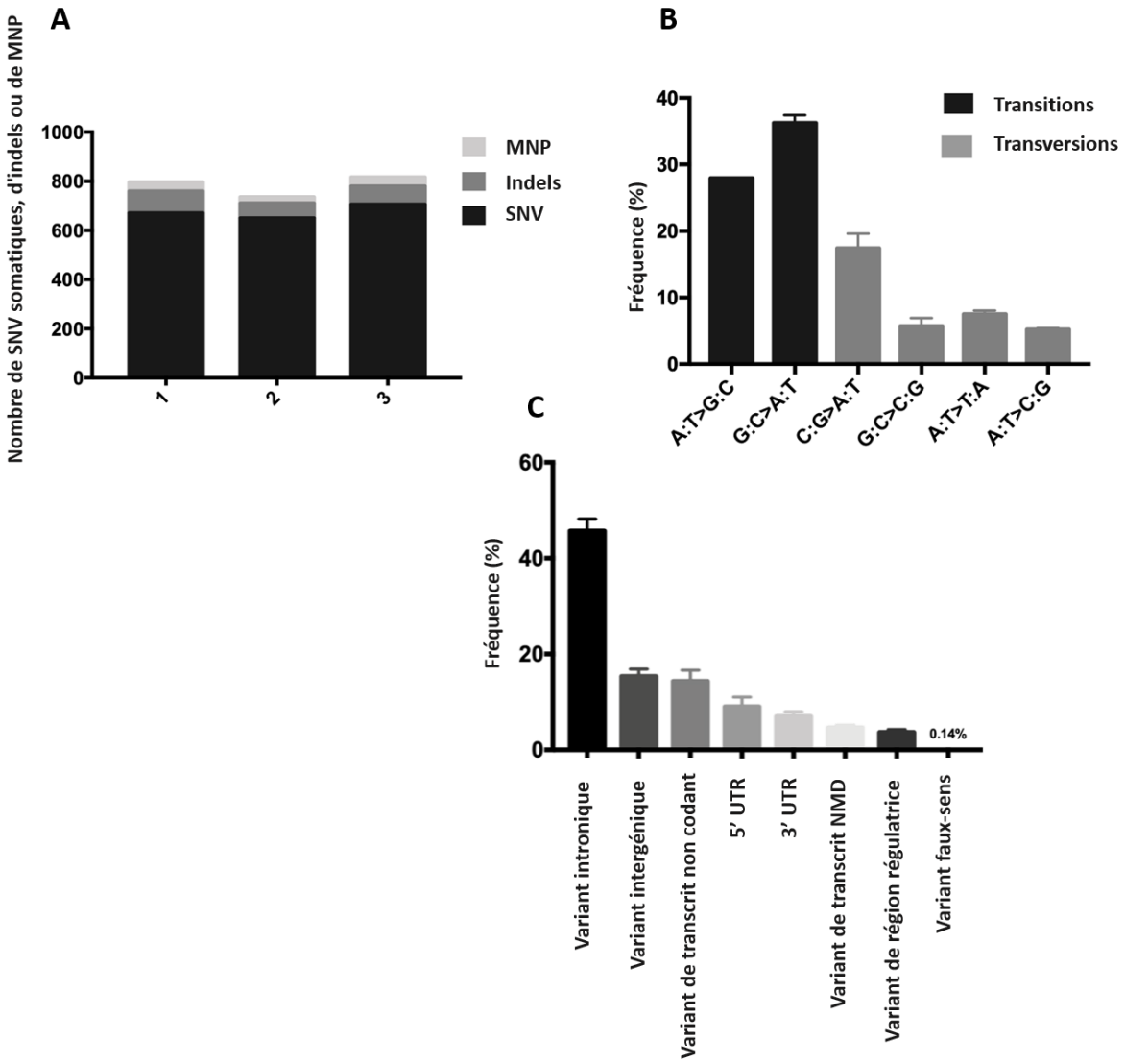


FIGURE 3.48 – Caractéristiques des SNP détectés par Mutect 2.
A : Nombre de SNV, d'indels et de polymorphismes multi-nucléotidiques dans chaque échantillon. B : Distribution de la fréquence des substitutions de base pour les trois échantillons, transversions en noir et transitions en gris. C : Distribution de la fréquence de la répartition des variants somatiques entre les régions génomiques pour les trois patients.

Les données de variations ont également été utilisées pour déterminer les signatures mutationnelles, décrites par la base de données COSMIC[120], induite par l'exposition à des exosomes précancéreux, à l'aide de l'application web MuSiCa[130], voir *Figure 3.49*. Les signatures mutationnelles détectées parmi les trois patients étaient assez comparables. Celles détectées l'étaient avec la plus grande fiabilité pour les signatures 1 et 5, qui sont également les plus communes parmi l'ensemble des types de cancer. Quant au reste, aucune signature particulière ne s'est détachée.

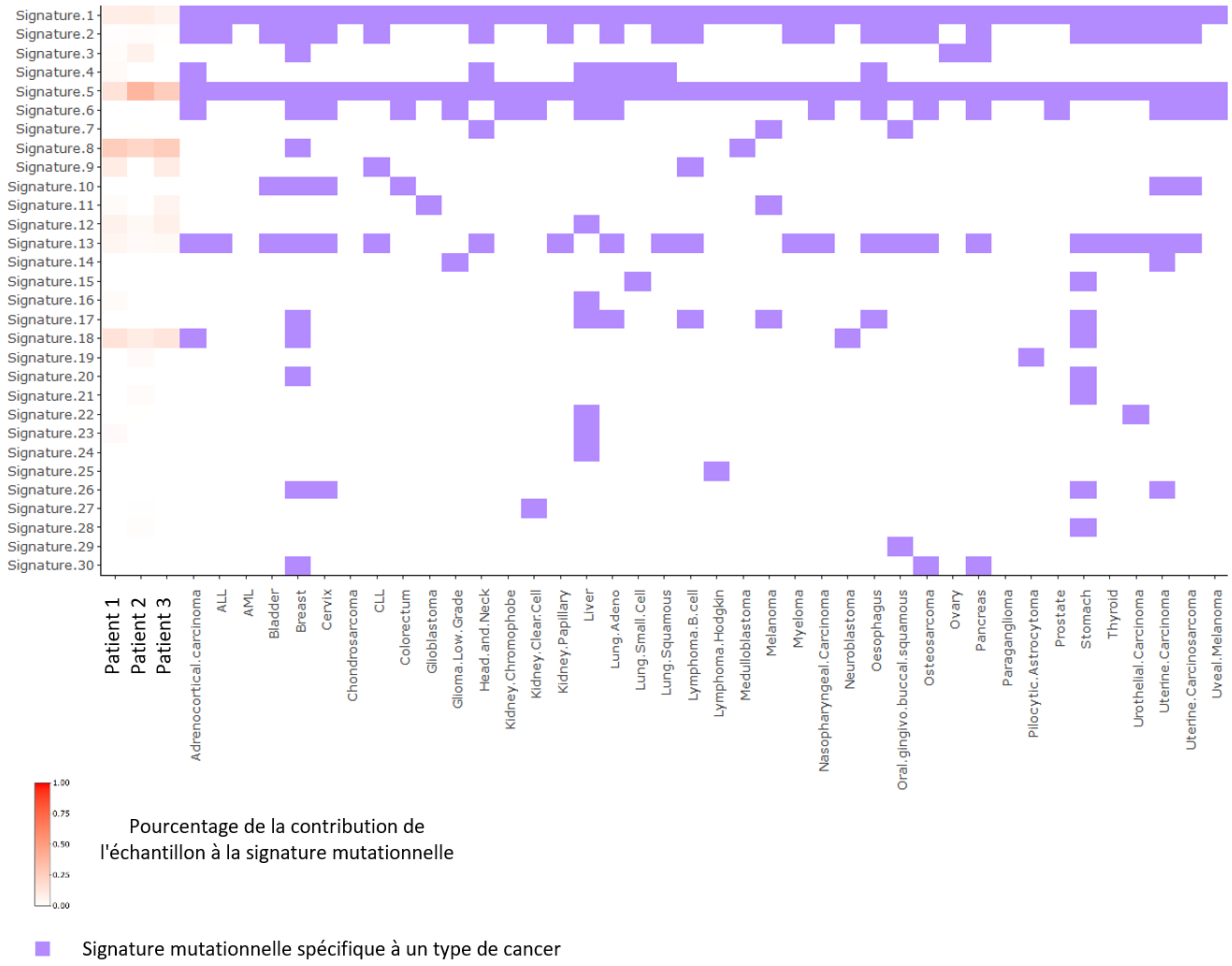


FIGURE 3.49 – Signatures variationnelles des trois patients étudiés obtenues à l'aide de l'outil MuSiCa.

Ensuite, nous avons cherché à déterminer si certains gènes avaient été enrichis en variations au cours de cette étude. Pour cela, ont été observés les différents gènes, présentant au moins une variation en communs entre les trois patients, voir *Figure 3.50*. Aucune variation commune n'a été retrouvée chez les trois patients, en revanche, 23 gènes différents communs à au moins deux des patients ont été retrouvés mutés, dont deux communs aux trois conditions. En revanche, lorsque des critères de filtration stricts sont appliqués comme la filtration sur la VAF (Fréquence des allèles du variant), ce nombre diminue fortement. La VAF est un critère de choix pour la filtration de variants somatiques, car il est attendu que l'apparition de variation somatique soit reliée à l'apparition d'une ou plusieurs populations clonales cancéreuses.

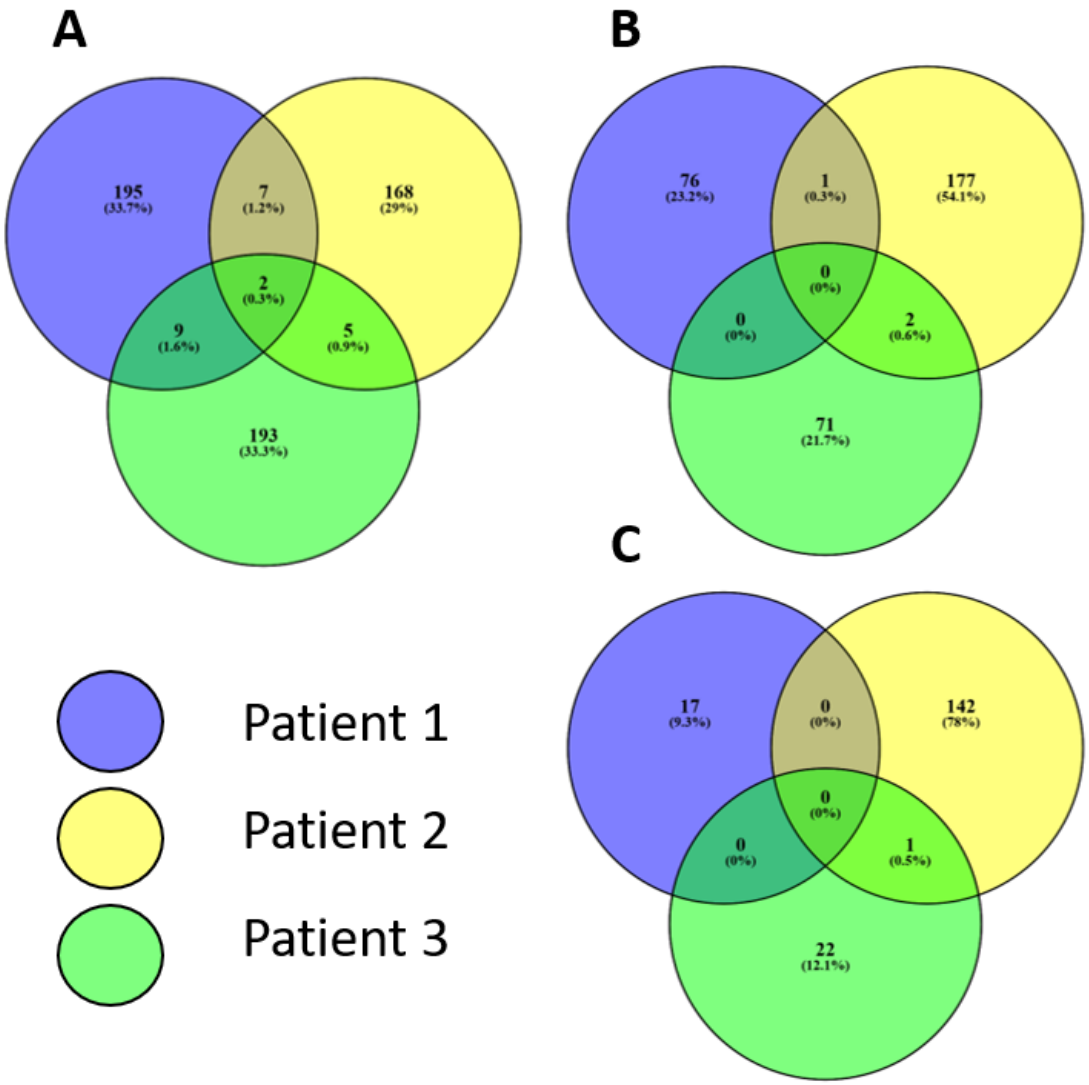


FIGURE 3.50 – Diagramme de Venn produit à l'aide de l'outil Venny représentant les gènes en commun possédant des variations PASS détectées parmi les trois patients.
A : toutes les variations PASS, B : seulement celles avec une VAF $>$ 5 %, C : seulement celles avec une VAF $>$ 10 %.

Enfin, les 23 gènes supportant une variation chez au moins deux des conditions ont été recherchés dans l'ontologie de la base de données DAVID[131] spécialisée dans les mutations somatiques, voir *Figure 3.51*. L'analyse de l'ontologie fonctionnelle liée à ces gènes a révélé des fonctions dans des processus biologiques l'épissage alternatif, la production de phosphoprotéines, de protéines du cytoplasme et du système endomembranaire.

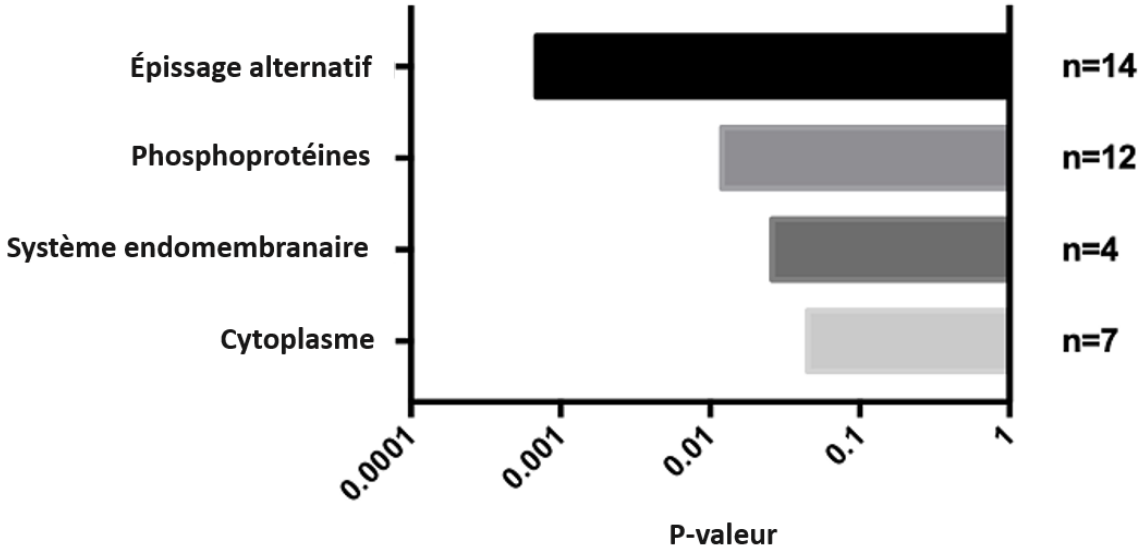


FIGURE 3.51 – Fonction supposée des 23 gènes soumis à l'ontologie de la base de données DAVID.

3.5 Discussion

Planifié initialement à la fin de l'année 2018, ce travail a pu être réalisé sur plusieurs mois jusqu'au début de l'année 2019. Cette collaboration, pour l'analyse de données HTS somatiques, a mis en relation un doctorant et son directeur de thèse spécialisés en analyse de données HTS germinales avec une équipe de biologiste spécialisée en cancérologie peu rompue à l'analyse de données HTS. Bien que les analyses réalisées au cours de cette expérimentation étaient assez éloignées de la zone de confort de chacun, elles se sont néanmoins concrétisées par une publication en 2020. Ce travail représente à ce jour ma seule expérience en traitement de données somatiques.

La tentative d'exploitation des données Chromium a été une véritable déception. Au moment de la conception du plan expérimental de cette analyse, en 2017, la technologie pouvait encore être considérée comme une technologie en expansion, potentiellement prometteuse. Le manque d'information et de données sur la technologie nous a pourtant été préjudiciable. Nous aurions peut-être pu exploiter les données si nous avions eu en notre possession plus de données de référence fournies par le CNRGH afin de pouvoir constituer un PoN plus important et ainsi produire des données moins bruitées. Néanmoins, même en changeant le PoN par celui du 1KGP[17], les données sont restées inexploitables, potentiellement à cause des artefacts techniques n'étant pas éliminés par un PoN produit sur une plateforme différente de celle qui a servi à générer les données de cette étude.

Deuxièmement, l'obligation de passer par la boîte noire Long Ranger a été un véritable cauchemar technique. En effet, l'outil étant peu ou passablement optimisé, ses temps d'exécution sont extrêmement longs. Une des spécificités du *cluster* CIMENT est une limite de calcul fixée à 48 heures maximum par tâche. La commande *wgs* de Long Ranger enchaînant les différentes étapes de l'alignement jusqu'à l'appel des variants sans qu'aucune interaction soit possible, il m'a été impossible de réaliser les calculs nécessaires jusqu'aux vacances de Noël 2018, lorsqu'une dérogation temporaire m'a été accordée, mais également période où le *cluster* Dahu a connu sa plus grande période d'instabilité depuis sa mise en place jusqu'à nos jours.

C'est pourquoi nous avons pris la décision d'exploiter uniquement les données provenant de la technologie WGS Illumina. Les résultats obtenus en plus des autres expérimentations conduites pour la publication de cet article ont permis de conclure le microenvironnement cellulaire pourrait être à l'origine de l'apparition de syndrome myélodysplasique en induisant un stress génotoxique par des vésicules extracellulaires assurant la communication intercellulaire. L'analyse WGS seule que nous avons conduite n'aurait pas pu permettre une telle hypothèse.

3.6 Article

Leukemia
https://doi.org/10.1038/s41375-020-0738-8

LETTER

Myelodysplastic syndrome

Extracellular vesicles from myelodysplastic mesenchymal stromal cells induce DNA damage and mutagenesis of hematopoietic stem cells through miRNA transfer

Mathieu Meunier[1,2] · Audrey Guttin[3] · Sarah Ancelet[2] · David Laurin[2] · Johanna Zannoni[1,2] · Christine Lefebvre[4] · Sylvie Tondeur[4] · Virginie Persoons[5] · Mylène Pezet[2] · Karin Pernet-Gallay[6] · Florent Chuffart[2] · Sophie Rousseaux[2] · Quentin Testard[2,7,8] · Julien Thevenon[2,7] · Claire Jouzier[1,2] · Jean-François Deleuze[9] · Karine Laulagnier[3] · Rémy Sadoul[3] · Christine Chatellard[3] · Pierre Hainaut[2] · Benoît Polack[1] · Jean-Yves Cahn[1] · Jean-Paul Issartel[3] · Sophie Park[1,2]

Received: 8 September 2019 / Revised: 15 December 2019 / Accepted: 30 January 2020

To the Editor:

Physiopathology of myelodysplastic syndrome (MDS) remains poorly understood and the role of the microenvironment is increasingly highlighted. Recent studies in mouse models demonstrate that abnormalities of mesenchymal stromal cells (MSC) contribute to the physiopathology of MDS. In particular, genetic deletion of *dicer1*, a gene encoding a type III RNase essential for the genesis of miRNA, in murine MSC-derived osteoprogenitors led to a pathological microenvironment generating myelodysplastic features in hematopoietic progenitors and ultimately leading to acute myeloid leukemia [1]. In human, there is an increased susceptibility to senescence of the MDS mesenchymal stem cells and defects in the support properties of the growth of hematopoietic stem cells (HSC) [2]. These observations establish a causal relationship between deregulation of the hematopoietic niche and MDS pathogenesis. However, so far only few studies have addressed the mechanisms by microenvironmental MSC and HSC exchange signals that may interfere with miRNA processing, specifically in the human MDS microenvironment.

Low risk MDS stroma supporting a defectuous hematopoiesis of normal HSPC was confirmed in our coculture model (Supplementary Data, Fig. 1). We focused therefore on the intrinsic properties of MDS-derived MSC and compared them to normal MSC (patient and healthy donors characteristics are described in Supplementary Tables 1, 2).

Presented in part at the 35th French Society of Hematology Annual Meeting (SFH), April 1–3 2015, Paris; at the 11th Annual Meeting of Groupe Francophone des Myélodysplasies (GFM), May 26-27 2016, Grenoble; 38th French Society of Hematology Annual Meeting, March 28–30 2018, Paris; and at the 12th Annual Meeting of Groupe Francophone des Myélodysplasies, May 30–31 2018, Tours, France.

Supplementary information The online version of this article (https://doi.org/10.1038/s41375-020-0738-8) contains supplementary material, which is available to authorized users.

✉ Mathieu Meunier, Md, PhD
mmeunier2@chu-grenoble.fr

✉ Sophie Park, MD, PhD
spark@chu-grenoble.fr

[1] Department of Hematology, CHU Grenoble Alpes, Grenoble, France

[2] CNRS UMR 5309, INSERM, U1209, Université Grenoble Alpes, Institute for Advanced Bioscience, 38700 Grenoble, France

[3] Grenoble Institut des Neurosciences, INSERM U836, Grenoble Alpes University, Grenoble, France

[4] Laboratoire de Génétique des hémopathies, CHU Grenoble Alpes, Grenoble, France

[5] Unité de Thérapie et d'Ingénierie Cellulaire, EFS Auvergne Rhône Alpes et UF Génétique moléculaire et maladie héréditaire et oncologie, CHUGA, Grenoble, France

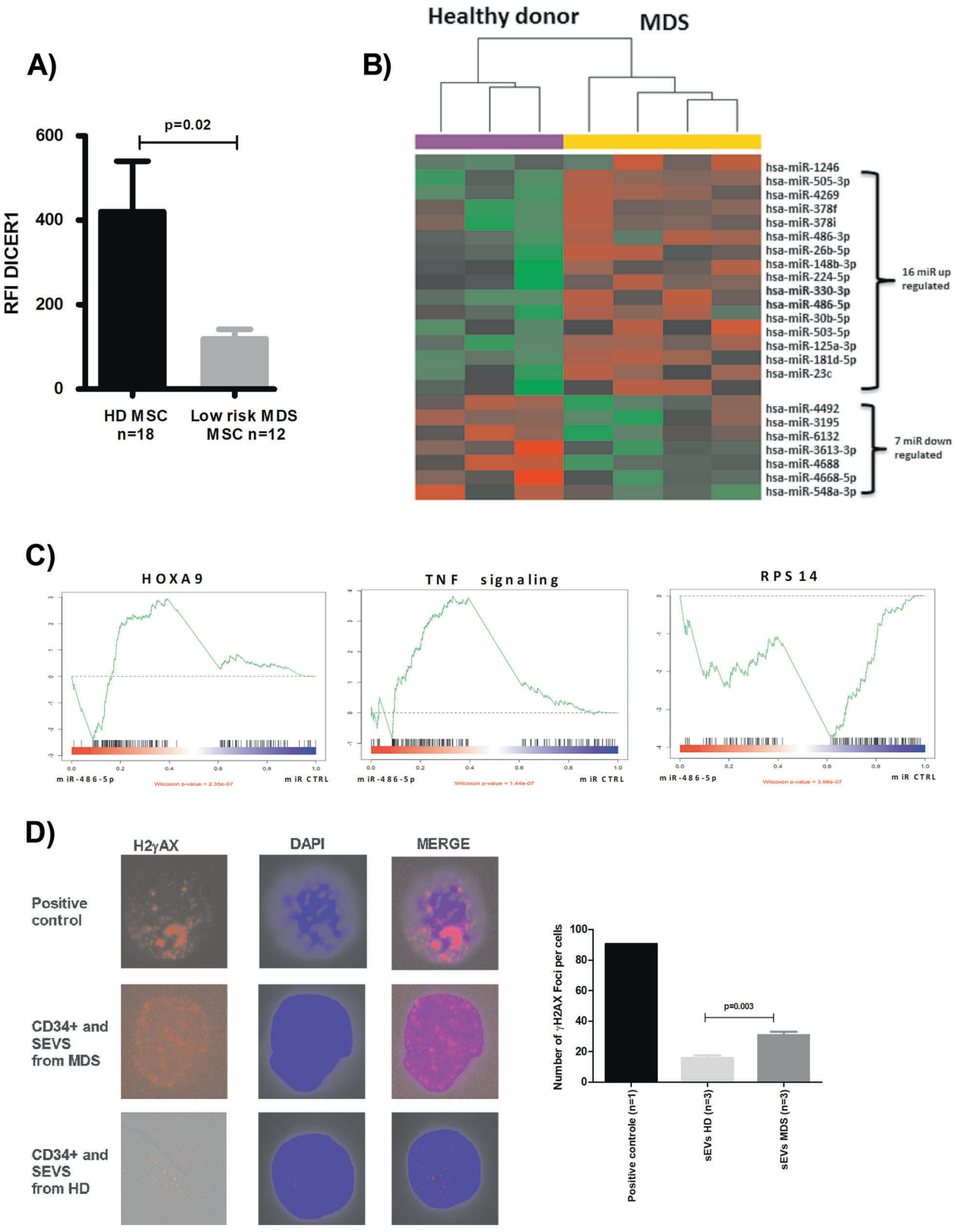
[6] Grenoble Institut des Neurosciences, Plateforme de Microscopie électronique, Grenoble Alpes University, Grenoble, France

[7] Département de Génétique et Procréation, CHU Grenoble Alpes, Université Grenoble Alpes, Grenoble, France

[8] Bioinformatics Unit, Eurofins Biomnis, Lyon, France

[9] Centre National de Génotypage, Institut de Génomique, Evry and Centre d'Etude du Polymorphisme Humain, 2 rue Gaston Crémieux, 91000 Paris, France

Published online: 12 February 2020



Thus, we investigated DICER1 expression on MSC from MDS-patient and compared them to healthy donors (HD) cells. We performed flow cytometry on fresh unseparated bone marrow from 18 HD and 12 low risk MDS patients according to IPSS score (Supplementary Data Fig. 2a). We observed a DICER1 down regulation in low risk MDS

◀ **Fig. 1 DICER1 expression in healthy donor and MDS mesenchymal stromal cells. a** DICER1 expression a directly analyzed on total bone marrow MSC harvested from HD or MDS patients by flow cytometry. Results are expressed in term of relative fluorescence intensity (RFI) of DICER1 fluorescence/condition without staining. **b** Heat map of miRNA with at least a 2-fold reduced or increased mean expression in MDS versus HD MSCs. Sixteen miRNAs were upregulated and 7 seven were down regulated. **c** RT-qPCR of miR-486-5p in 5 MDS and 5 HD MSCs. Results are given as $2^{-\Delta Ct}$ (Delta Ct: $\Delta Ct = Ct$ miRNA – mean Ct (RNU6B/miR-26a-3p). **d** Visualization of MDS-derived MSC sEVs incorporation into HD HSPC by confocal microscopy. HSC were CD34-FITC labeled, nucleus were Dapi labeled and sEVs were Vybrant DiD labeled; visualization was on a Zeiss Dynascope LSM710 NLO, 633/1.4 oil-immersion objective, at room temperature. **e** RTqPCR of miR-486-5p in sEVs from MDS and HD MSC. Results are given as $2^{-\Delta Ct}$ (Delta Ct: $\Delta Ct = Ct$ miRNA – mean Ct (miR-200/miR-101). **f** RT-qPCR of miR-486-5p in HSPC from healthy donors incubated with sEVs produced by MDS MSCs and HD MSCs. Results are given using the $2^{-\Delta\Delta Ct}$ formula. **g** Gene Sets Enrichment Analysis (GSEA) done on the RNA sequencing data from HSPC cells overexpressing miR-486-5p. HD healthy donor, MDS myelodysplastic syndrome, MSC mesenchymal stromal cell, miRNA, microRNA ns, no significant, RFI relative fluorescence intensity, RT-qPCR reverse transcription quantitative polymerase chain reaction.

patients compared to HD ($p = 0.02$, Fig. 1a). Since MSC expansion in vitro requires multiple passages, we confirmed that the down regulation observed directly in fresh bone marrow ex vivo still persisted after MSC expansion in vitro by flow cytometry and RTqPCR (Supplementary Data Fig. 2b–d).

Because DICER1 has a pivotal role in miRNA processing [3], we hypothesized that MSC from low risk MDS could present, as a consequence, a deregulated miRNA profile. Thus, the miRNA profile from four low risk MDS patients and three HD MSC were investigated using Affymetrix GeneChip miRNA 4.0 Array. We observed 16 miRNA upregulated and seven downregulated miRNA in MDS-derived MSC when compared to HD cells (Fig. 1b).

Interestingly, among the 16 upregulated MDS-derived MSC miRNA, 6 miR (according to different complementary databases based on prediction or literature data mining like Targetscan, miRTarbase, and microT-CDS) have DICER1 for potential target and potentially and partially explain the DICER1 downregulation in low risk patients MDS-derived MSC. Overexpression of those miR was confirmed by RT-qPCR in another validating cohort of MSC (5 HD and 5 low risk MDS) and only miR-26b-5p and miR148b-3p were significantly overexpressed ($p = 0.03$ and $p = 0.04$ respectively, Supplementary Data Fig. 3a). We confirmed the hypothesis of negative modulation of DICER1 by the miR-26b-5p and miR-148b-3p, using stable cells line overexpressing those 2 miRNA and DICER1 luciferase reporter plasmid (Supplementary Data, Fig. 3b–d).

Within the upregulated miRNA identified in MDS-derived MSC not involving DICER1 target, we outlined miR-486-5p as a potential interesting modulation actor. Indeed, miR-486-5p is implicated in leukemogenesis as it is over-expressed in leukemic cells of Down syndrome patients [4] and plays a role in MSC senescence [5]. We confirmed miR-486-5p overexpression by RT-qPCR in our validating cohort of MSC ($p = 0.007$, Supplementary Data Fig. 4a). A possible way of cell-to-cell communication in the hematopoietic niche is extracellular vesicles. When vesicles are isolated without specific attention to their size, the smaller fraction (<200 nm) is called small extracellular vesicles (sEVs) [6]. sEVs contain exosomes which are the most described vesicles containing protein and RNA species and especially miRNA. We have isolated sEVs from supernatant of expanded MSC from both healthy donors and low risk MDS patients. This fraction of sEVs was positive for exosome confirmed by western blot, transmission electron microscopy assay and NTA profile (Supplementary Data, Fig. 5a–c).

Then, we confirmed that sEVs could be incorporated into hematopoietic stem progenitor cells using confocal microscopy (Supplementary Data, Fig. 5d) and demonstrated by RT-qPCR that sEVs from MDS-derived MSC are enriched in miR-486-5p ($p = 0.04$, (Supplementary Data, Fig. 4b). We measured miR-486-5p expression level in $CD34^+$ hematopoietic stem progenitor cells (HSPC) from HD after co-incubation with MSC derived sEVs from MDS or from control HD. miR-486-5p was upregulated in HSPC incubated with MDS MSC derived sEVs ($p = 0.047$) compared to the cells incubated with HD MSC derived sEVs (Supplementary Data, Fig. 4c). This result highly suggests that sEVs from MDS-derived MSC carry miR-486-5p as a cargo to the HSPC.

To gain insight the signaling pathways which could be involved in HSPC overexpressing miR-486-5p, we transfected this miR in healthy donor HSPC ($n = 3$, supplementary data, Fig. 6a) and performed RNA sequencing of this transfected HSPC. Our results showed that 49 genes were significantly upregulated whereas 14 were downregulated (Supplementary Data, Fig. 6c) (The full list of genes is given in Supplementary Data, Table 3). We carried out a Gene Sets Enrichment Analysis (GSEA) on the collections of gene sets made available by the Broad Institute. We observed an enrichment of genes implicated in TNF signaling pathway (Fig. 1c) and other immune pathways like INFα or INFγ and TGFβ (Supplementary Data, Fig. 6d). Moreover, we also observed enrichment of the HOXA9 gene signature and on the contrary a defective RPS14 signature (Fig. 1c).

In order to investigate the putative functional effects induced by MDS MSC-derived sEVs, we incubated directly those vesicles with HD $CD34^+$ HSPC, to bypass cell to cell contacts and secreted cytokines by the microenvironment. Then, after 48 h of incubation with

sEVS, we observed more apoptotic HSPC cells with sEVs from MDS-derived MSC than with sEVs from HD MSC ($p = 0.04$) (Supplementary Data, Fig. 7a). There was no difference of the percentage of cycling cells with sEVs from MDS-derived MSC or from HD ($p = 0.45$) nor differentiation of the $CD34^+$ HSPC assessed by the loss of $CD34^+$ marker ($p = 0.24$) nor on the clonogenicity (Supplementary Data, Fig. 7b–d).

As oxidative stress due to reactive oxygen species (ROS) is involved in MDS development [7], we assessed ROS levels with the fluorescent DHE probe the mitochondria-targeted superoxide-reacting MitoSOX probe in the HSPCs incubated with sEVs derived from MDS or HD MSC. The DHE and MitoSOX ratio of specific fluorescence (RFI) was increased by an average of 40% ($p = 0.046$) and average increase of 212% ($p = 0.031$) respectively in HSPC exposed to MDS sEVs compared to sEVs from HD (Supplementary Data, Fig. 7e). It is well known that excess of ROS is deleterious for cells and induce DNA damage. We assessed by immuno-fluorescence confocal microscopy the number of γH2AX foci in HSPC treated with sEVs from MDS MSC or HD at Day 7 of incubation. We observed that sEVs from MDS MSC induced a higher number of γH2AX foci 30.80 (±2.33) versus 15.80 (±1.77) per cell for HD MSC derived sEVs ($p = 0.008$) (Fig. 1d), suggesting more DNA damages mediated by EVs from MDS MSC. Of note, when we observed the presence of γH2AX foci initially at Day 2 of the culture, the DNA damages were not present (data not shown).

Lastly, we analyzed by whole-genome sequencing (WGS) $CD34^+$ HSPC from HD incubated 7 days with sEVs from either HD MSC ($n = 3$) or MDS-derived MSC ($n = 3$). We observed that $CD34^+$ HSPC incubated with sEVs from MDS derived MSC carried 783 (range 736–817) somatic variants in comparison with those incubated with sEVs from HD MSC (Supplementary Data, Fig. 8a). Those somatic variants are single-nucleotide variants (SNVs) in 86.3% (range 84.3–88.3) of the cases, short insertions and deletions in 9.5% (range 8.3–11.2) and finally multi-nucleotide polymorphisms (MNPs) in 4.1% (range 3.4–4.5). SNVs were transversions for 34.82% and transitions for 64.18% of the cases. Those variants were mostly in non-coding regions (Supplementary Data, Fig. 8b, c). Missense variants have been identified in coding regions in four genes (*EEF1A1, WDR66, KIR3DL3*, and *ANKRD1*). We have then compared our data with previously observed mutational signatures from COSMIC (Catalogue of Somatic Mutations in Cancer) database and we found similarities with four known mutational signatures: signature 1, 5, 8, and 18 (Supplementary, Data Fig. 8d). Signatures 1 and 5 are found in all cancer types and most cancer samples. Signature 8 is found in breast and medulloblastoma, whereas signature 18 is seen in neuroblastoma, breast and stomach carcinomas. Then, we compared mutations induced by the sEVs between the three samples to determine if they were identical or if these mutations were stochastically distributed. We observed no overlapping somatic variant between the three samples analyzed but when we focused on the genes targeted by these somatic mutations, we found 23 overlapping genes (lists of genes are given in Supplementary Data, Table 4) between the three samples. Gene ontology analysis of these genes revealed functions in enriched biological process like alternative splicing, phosphoproteins, cytoplasm proteins, endomembrane system (Supplementary Data Fig. 8e, f). Alternative splicing was the most enriched biological pathway affected with 14 out of 23 genes mutated involved in RNA alternative splicing.

Together those findings support the view that medullar microenvironment could drive MDS initiation by inducing genotoxic stress by extracellular vesicles mediating cell-to-cell communication.

Acknowledgements MM has received a financial grant from Université Grenoble Alpes (Programme AGIR 2014) and from Bourse "Espoir Sandrine". The bioinformatics and biostatistics analyses of RNA sequencing presented in this paper were performed by the Epimed platform in IAB (UGA - INSERM 1209 – CNRS UMR 5309). We thank also Cecile Martin for providing us healthy donor bone marrows. We thank Jean-Marc Moulis for the luminescence measurement facilities.

Author Contributions MM, AG, CJ, SA, DL, JZ, CJ and VP performed experiments; FC and SR performed the bioinformatics and statistical analyses of the RNAseq data; KL, CC and RM performed the nanosight analyses; KPG performed electronic microscopic pictures; JFD performed the WGS; QT, ST, CL and JT performed the bioinformatics analyses of WGS; ML performed the confocal microscopy pictures; AG and JPI analyzed the transcriptomics results; MM and SP designed the research, analyzed results and wrote the paper. PH, BP and JYC reviewed the paper.

Compliance with ethical standards

Conflict of interest The authors declare no relevant conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Raaijmakers MH, Mukherjee S, Guo S, Zhang S, Kobayashi T, Schoonmaker JA, et al. Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. Nature. 2010;464:852–7.
2. Geyh S, Oz S, Cadeddu RP, Frobel J, Bruckner B, Kundgen A, et al. Insufficient stromal support in MDS results from molecular and functional deficits of mesenchymal stromal cells. Leukemia. 2013;27:1841–51.
3. Foulkes WD, Priest JR, Duchaine TF. DICER1: mutations, microRNAs and mechanisms. Nat Rev Cancer. 2014;14:662–72.

4. Shaham L, Vendramini E, Ge Y, Goren Y, Birger Y, Tijssen MR, et al. MicroRNA-486-5p is an erythroid oncomiR of the myeloid leukemias of Down syndrome. Blood. 2015;125:1292–301.
5. Kim YJ, Hwang SH, Lee SY, Shin KK, Cho HH, Bae YC, et al. miR-486-5p induces replicative senescence of human adipose tissue-derived mesenchymal stem cells and its expression is controlled by high glucose. Stem Cells Dev. 2012;21:1749–60.
6. Tkach M, Thery C. Communication by extracellular vesicles: where we are and where we need to go. Cell. 2016;164:1226–32.
7. Meunier M, Ancelet S, Lefebvre C, Arnaud J, Garrel C, Pezet M, et al. Reactive oxygen species levels control NF-kappaB activation by low dose deferasirox in erythroid progenitors of low risk myelodysplastic syndromes. Oncotarget. 2017; 8:105510–24.

Évaluation de la détection de variants structuraux par séquençage Oxford Nanopore Technologies dans un contexte de routine diagnostique

4.1 Motivation

Bien que le WGS *short-reads* commence à peine à se démocratiser en routine diagnostique, le domaine de recherche des technologies de séquençage est extrêmement actif, notamment en ce qui concerne les technologies de troisième génération. Grâce à l'amélioration constante de ces technologies et des outils bioinformatiques associés à l'analyse des données qui en sont issues depuis ces dernières années, elles ont atteint un niveau de précision et de rendement qui permettrait en théorie de les utiliser à des fins cliniques en routine. De plus, la supériorité du séquençage à longues lectures pour la détection de SV comparé aux techniques existantes a été démontrée par des études exploratoires[92] et des études menées par des consortiums comme le GIAB[93].

Les analyses actuellement en place en routine diagnostic en laboratoires de Biologie Médicale ne permettent que de détecter un nombre limité de type de SV de manière pan-génomique (non ciblée) chez les patients, à savoir, les CNV de grande taille. La technologie Oxford Nanopore (ONT) permet, en théorie, de détecter le plus grand nombre de types de SV différents parmi les technologies de séquençage génomique. C'est pourquoi à partir de l'automne 2019, Eurofins Biomnis s'est équipé d'un séquenceur GridION ONT. L'objectif de cette étude était de déterminer si un séquençage à relative basse couverture (5-10X) sur *flowcell* MinION permettait la détection de variants structuraux de grande taille et si cette nouvelle analyse pouvait s'intégrer dans le cadre des analyses de routine clinique du laboratoire.

Préalablement au séquençage d'échantillons patients et de référence par le laboratoire, un *benchmarking* de différents outils bioinformatiques contre des sets de variation de référence à partir de données de séquençage ONT publiques a été conduit. Les résultats du *benchmark* sur les données PromethION ont une valeur indicative, car il n'est pas garanti que les résultats obtenus sur de telles données peuvent être transposés sur des données produites par le séquenceur MinION. En revanche, ce *benchmark* a permis de mettre en place la méthodologie qui a ensuite été répliquée sur des données produites par le laboratoire Biomnis à partir d'ADN issus de lignées immortalisées d'individus *gold standard* sur séquenceur MinION. Enfin, plusieurs patients avec des variants de structure connus détectés par le laboratoire ont été séquencés afin d'évaluer s'il était possible de retrouver ces variations avec la technologie ONT.

Ce travail a été présenté en session simultanée lors des Assises de Génétiques de Tours en 2020.

4.2 Matériel et méthodes

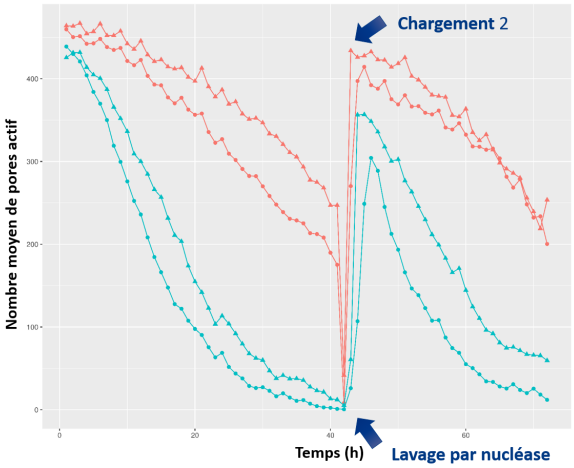
Plusieurs échantillons dont treize données patients et un échantillon de référence (HG002) ont été séquencés au cours de plusieurs séries de séquençages différentes sur un séquenceur GridION d'ONT. L'ADN des patients a été extrait à partir de sang EDTA à l'aide du *Qiagen MidiKit* ou à l'aide du *Gentra Puregene Blood Kit*. La qualification des ADN extraits et commerciaux a été réalisée à l'aide de l'automate d'électrophorèse capillaire *Fragment Analyser*. La fragmentation a été effectuée à l'aide de *Covaris G-tube*. Une opération de purification et sélection des fragments a été entreprise pour certains des échantillons à l'aide de kits *Nanobind* de l'entreprise *Circulomics*. Enfin, la préparation de la librairie a été effectuée à l'aide du kit de ligation *SQK-LSK109* et chargée sur des *flowcell* MinION *R9.4.1*. Les données ont été basecall avec l'outil Guppy (https://nanoporetech.com/community) version *3.2.6+afc8e14*.

L'alignement des données obtenues par séquençage a été effectué à l'aide des outils Minimap2[57] *commit 2.17-r941*, LAST[132] version *1021* et NGMLR[63] version *0.2.7* contre le génome hg38 sans les *contigs* alternatifs. Les variants structuraux ont été appelés à l'aide des *variant callers* pbsv (https://github.com/PacificBiosciences/pbsv) *commit v2.2.2-2-gf1f52e1*, SVIM[133] version *1.2.0*, Sniffles[63] version *1.0.11*, ainsi que le logiciel de détection de répétition NCRF[134] version *1.01.00* et le logiciel de détection d'aneuploïdies AneuFinder[135] version *1.14.0*. Enfin, certains des variants des patients de références ont été contrôlés à l'aide de l'outil de visualisation IGV[136]. La plupart des calculs ont été menées sur le *cluster* Dahu de CIMENT de manière parallélisée à l'aide de GNU Parallel[137] version *20191122*.

4.3 Optimisation du rendement de séquençage

Même si la détection de SV par alignement des lectures contre un génome de référence puis par appel de variant ne nécessite pas autant de profondeur de couverture ni des lectures aussi longues que pour un assemblage, il est nécessaire d'essayer d'exploiter la technologie au maximum de ses capacités. Les seuils qui avaient été établis pour déterminer si cette technologie pouvait ou non être intégrable dans la routine diagnostic, était la capacité de générer environ 33 Gb par *flowcell* (équivalent à environ 10X de profondeur de couverture sur le génome humain), tout en ayant une longueur de fragment la plus longue possible. C'est pourquoi ont été testés différents protocoles de laboratoire humide afin de pouvoir choisir celui qui offrirait le meilleur rapport longueur de lectures sur rendement du séquençage.

La relation inverse entre taille des lectures moyenne et rendement moyen de la librairie de la technologie ONT est connue. Elle a notamment été observée lors du développement du protocole de séquençage *ultra long-reads* afin de produire des données qui serviront à un nouvel assemblage de référence[12]. Cette relation a été vérifiée en testant le rendement du séquençage Nanopore entre des librairies dont la seule différence dans la préparation a été la fragmentation ou non de l'ADN, voir *Figure 4.52*.



Échantillon	Randement	N50
076-484-43	18Gb	18Kb
076-482-03	13Gb	26Kb
076-484-43 Fragmenté 15Kb	39Gb (>2x)	9Kb
076-482-03 Fragmenté 15Kb	33Gb (>2x)	10KB

FIGURE 4.52 – Nombre de pores actifs au cours du séquençage corrélé au rendement du séquençage et à la N50 des lectures produites.

Deux échantillons ont été séquencés avec ou sans fragmentation de l'ADN. Nous avons observé que l'ADN fragmenté à une taille moyenne de 15 kb permettait un meilleur rendement de séquençage. De plus, la valeur de N50 est plus élevée avec un ADN fragmenté. La N50 est la longueur de la lecture la plus courte dans le groupe des lectures les plus longues qui représentent ensemble 50% des nucléotides de l'ensemble des séquences.

Au cours du séquençage, les fragments d'ADN peuvent bloquer les pores de la *flowcell*. Périodiquement, le logiciel de contrôle du séquençage peut inverser le courant appliqué aux pores supposés bloqués afin de faire ressortir le fragment d'ADN. Si la tentative de déblocage échoue, alors le pore est bloqué définitivement. Lors du lavage à l'aide d'un kit contenant une solution de nucléase, certains pores peuvent être débloqués. La procédure de déblocage et le lavage sont plus efficaces avec les fragments courts plutôt qu'avec les longs. C'est pourquoi le nombre moyen de pores actifs est plus élevé lors des séquençages de fragments courts avant le lavage et que le lavage permet de récupérer plus de pores après le chargement de la seconde librairie, augmentant ainsi le rendement moyen du séquençage.

La fragmentation ou non des fragments d'ADN préalablement à la préparation de librairie n'est qu'un des paramètres pouvant avoir un effet sur le rendement du séquençage. C'est pourquoi divers protocoles ont été testés au cours du séquençage des quatorze échantillons utilisés dans cette étude, voir *Figure 4.53*.

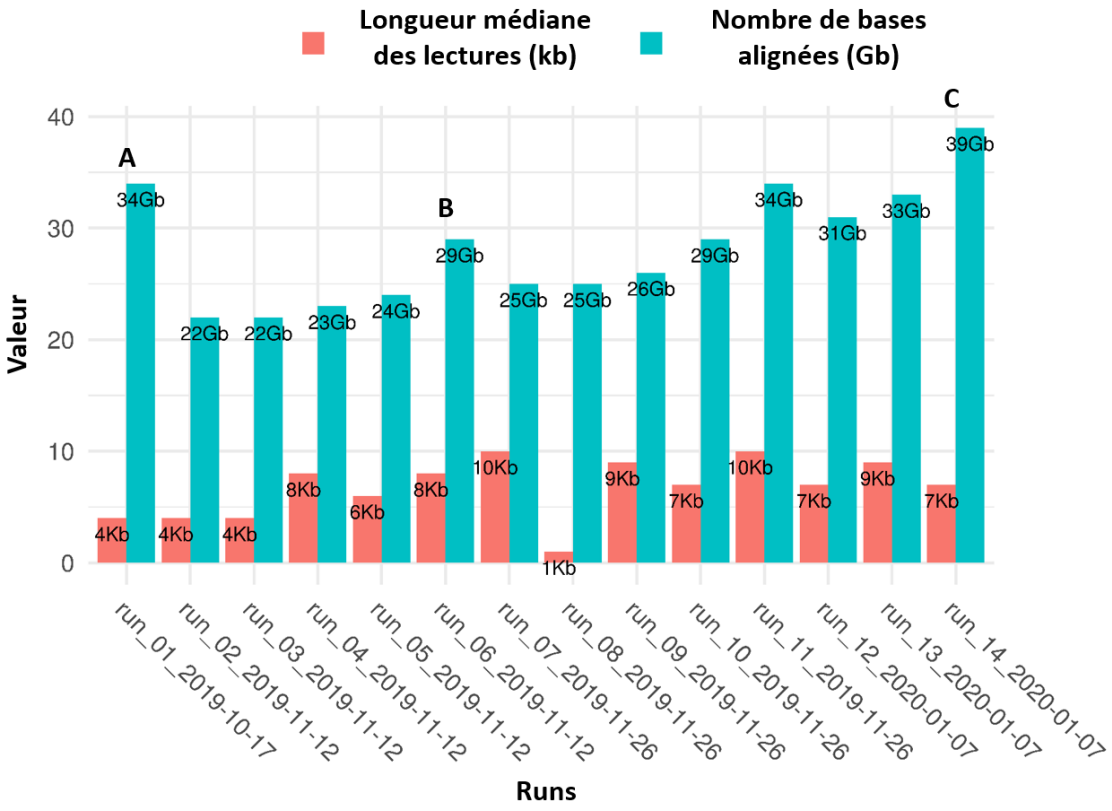


FIGURE 4.53 – Rendement et longueur médiane de quatorze échantillons séquencés sur GridION par Eurofins Biomnis.

Pour le *run*, noté A, le protocole utilisé a été le suivant, extraction de l'ADN à l'aide du *Qiagen MidiKit*, fragmentation à 15 kb *via Covaris G-tube* et préparation de deux librairies distinctes par ligation. Le séquençage a été mené sur 72 h, avec le chargement d'une première librairie à T0, puis un lavage des pores de la *flowcell* à l'aide du kit de lavage, puis le chargement de la seconde librairie à T36 h. Bien que le rendement de ce séquençage ait été assez important, la N50 du *run* était plus faible qu'attendu, très probablement à cause d'un nombre de fragments de petite taille trop important.

Pour les *runs* du 12 novembre, incluant le *run* noté B, le même protocole a été appliqué par rapport à précédemment, excepté une étape supplémentaire de purification et sélection des fragments à l'aide du kit *Circulomics*. L'effet de ce kit peut être observé *Figure 4.54*. Les fragments de petite taille qui étaient présents après l'extraction sont éliminés. Grâce à cela, la N50 des différents *runs* était plus élevée, mais l'effet sur le rendement de séquençage était trop important.

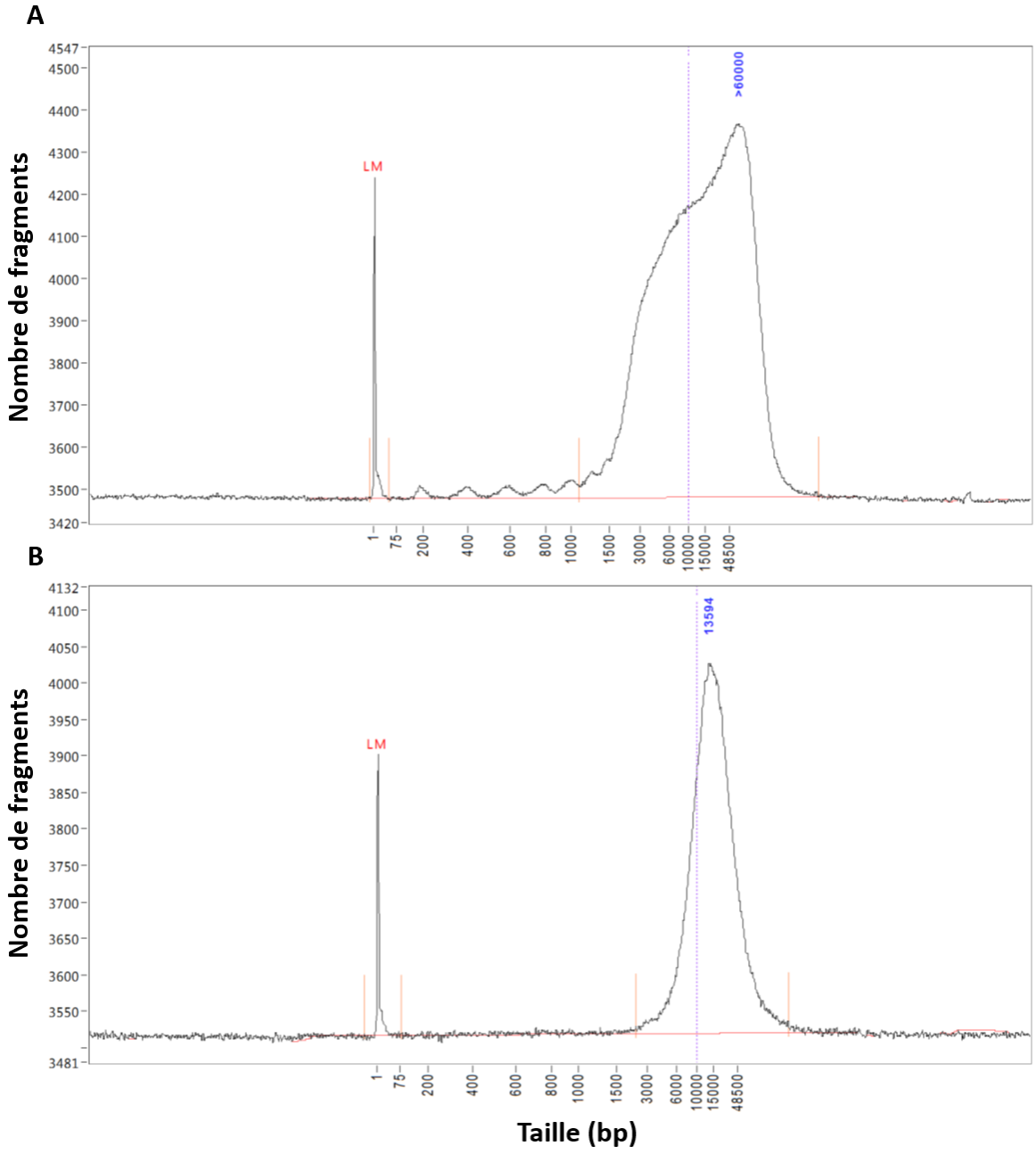


FIGURE 4.54 – Distribution de la taille des fragments mesurée par l'automate *Fragment Analyser*. A : Après extraction. B : Après fragmentation et sélection des fragments. LM : Marqueur léger de migration sur gel permettant de calculer la distribution de la taille des fragments.

Enfin, pour les *runs* des 26 novembre et 7 janvier, incluant le *run* noté C, une seule librairie avait été préparée par *run*, sans sélection de fragments. En revanche, l'extraction avait été réalisée à l'aide du kit *Gentra Puregene* supposé produire moins de petits fragments, sans trop consommer d'ADN par rapport au kit *Circulomics*. L'augmentation du rendement avec une taille de fragment plus qu'honorable avec seulement une librairie avait été mise sur le compte de plusieurs facteurs. Tout d'abord, à un ADN d'excellente qualité (extrait dans les semaines précédant le séquençage, contrairement à la plupart des autres échantillons congelés). Mais également, à un protocole optimisé, ne produisant que peu de petits fragments et conservant une quantité d'ADN importante. Enfin, à la montée en expertise de l'équipe technique composée du Dr Francis Rousseau, Mohamed Taoudi et Pascal Mouty. Les derniers séquençages effectués avec la technologie rencontraient pour la plupart les critères qui avaient été définis, soit un rendement de plus de 33 Gb, pour une longueur de fragment autour de 10 kb.

4.4 *Benchmarking* de détection de variants de structure sur données issues de séquenceur PromethION

4.4.1 Données de référence

De la même manière que la validation des résultats de SNV *calling* par le pipeline Lygrexome, un *benchmarking* des performances concernant la détection de SV par différents couples de logiciels d'alignement et de variant callers a été conduit. Les données de référence utilisées au cours de cette étude sont celles de l'individu HG002 publiées par le GIAB[92]. Pareillement aux données de références pour la détection de SNV du GIAB, les fichiers de SV de l'individu HG002 possèdent des SV de haute confiance appelés tiers 1, disponibles ici : http://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/. L'étude s'est concentrée sur ces derniers. À cette époque, le set de référence n'étant qu'en version hg19, la totalité des analyses ont donc été effectuées sur les *contigs* canoniques du génome version 19.

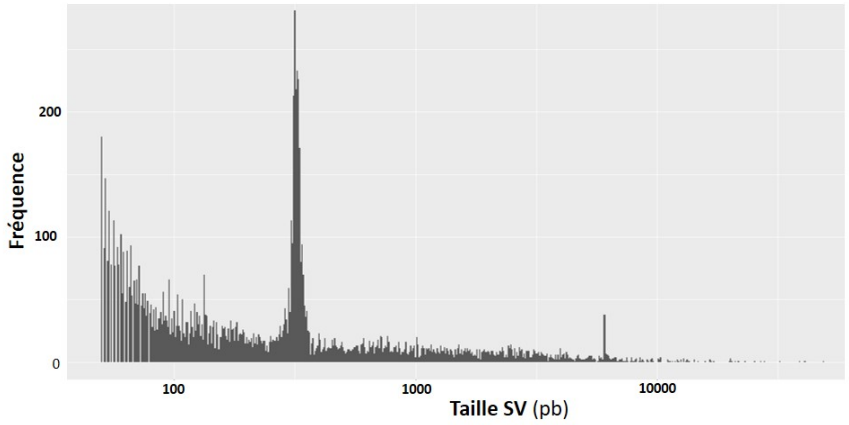


FIGURE 4.55 – Distribution de la longueur des variants de haute confiance présents dans l'ensemble de référence de vérité de l'individu HG002 publié par le GIAB.

L'individu HG002 étant un individu sain, l'ensemble de référence de haute confiance est majoritairement composé de variants de petite taille et ne possède que peu de réarrangements de grande taille (53 d'une taille contenue entre 10 et 50 kb) et encore moins de très grande taille (5 d'une taille supérieure à 50 kb, dont deux supérieures à 100 kb), avec la plus grande variation d'une taille de 114 kb. La distribution des tailles et des types des variants de haute confiance présents dans l'ensemble de référence de vérité de l'individu HG002 sont visibles *Figure 4.55* et *Table 4.24*.

A

Type de SV	Nb SV
SIMPLEDEL	2895
SIMPLEINS	3321
DUP	1772
SUBSDEL	232
SUBSINS	250
CONTRAC	1176

B

Taille des SV (pb)	Nb SV
50 à 99	2955
100 à 299	2349
300 à 999	3057
1000 à 5 000	1019
5 000 à 10 000	208
10 000 à 50 000	53
50 000+	5
Total	9646
Médiane	232
Moyenne	706

TABLE 4.24 – Caractéristiques des variations contenues dans l'ensemble de référence de vérité de l'individu HG002 publié par le GIAB.
A : type. B : taille.

Les types de variations décrits Table *Table 4.24* A, sont définis par les auteurs de l'étude qui a permis la publication de cet ensemble de référence[92] de la manière suivante :

— SIMPLEDEL : Suppression d'au moins une séquence unique.

— SIMPLEINS : Insertion d'au moins une séquence unique.

— CONTRAC : Contraction ou suppression d'une séquence entièrement similaire à la séquence restante.

— DUP : Duplication ou insertion d'une séquence entièrement similaire à une séquence préexistante.

— INV : Inversion.

— SUBSINS : Insertion d'une nouvelle séquence avec modification d'une séquence préexistante.

— SUBSDEL : Suppression d'une séquence avec modification d'une partie de la séquence restante.

Les pertes de matériel (délétions) peuvent englober les trois types de variations décrites par le GIAB que sont, les SIMPLEDEL, CONTRAC et SUBSDEL. Les gains de matériel (insertions) peuvent englober les trois types de variations décrites par le GIAB que sont, les SIMPLEINS, SUBSINS et DUP. Aucune inversion ne fait partie des variants de haute confiance dans ce set de donnée.

4.4.2 Données de séquençage

Les données de séquençage utilisées au cours de ce *benchmark* sont issues d'une publication du consortium GIAB, dont l'objectif était le séquençage, puis l'assemblage de onze génomes, dont celui de l'individu de référence HG002[138]. La préparation de librairies *ultra long-reads* a été effectuée à l'aide du kit de ligation *SQK-LSK109* et le séquençage a été réalisé sur un séquenceur PromethION avec des *flowcell* de type *FLO-PRO002*. Trois séquençages ont été effectués, à environ 8, 16 et 21X de profondeur de séquençage. Les fichiers FASTQ dont l'appel des bases a été effectué avec guppy version *3.2.5* sont disponibles au téléchargement ici : https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/. Après téléchargement, nous avons combiné les fichiers FASTQ pour obtenir une profondeur moyenne de 45X, puis sous-échantillonés les lectures avec l'outil Seqtk (https://github.com/lh3/seqtk) *commit 1.3-r106* à environ 30, 20, 15, 10, 5 et 1X, 3 fois pour chaque condition.

La décision a été prise de ne pas appliquer de filtres de qualité ou de taille sur les lectures malgré l'enrichissement du fichier FASTQ 45X en *reads* de petite taille (longueur de lecture médiane de 450 pb), car notre objectif était d'avoir un taux de rappel maximal et donc de garder toute l'information possible, voir *Figure 4.56*. Malgré cela, les données de séquençage comportent également un grand nombre de *reads* de très grande taille (longueur moyenne de lecture de 8,1 kb et N50 de 48 kb).

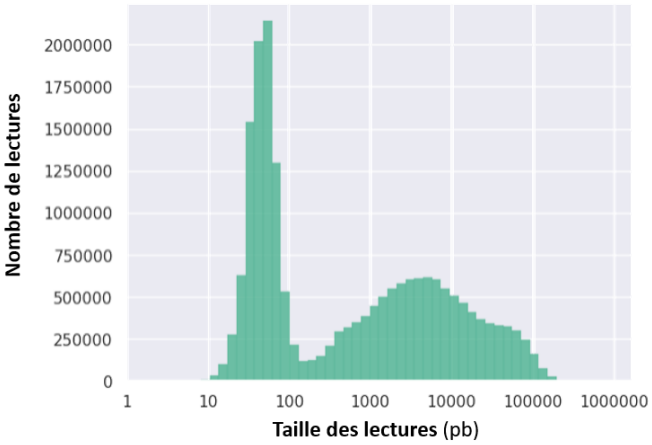


FIGURE 4.56 – Distribution de la taille des lectures du fichier FASTQ 45X du GIAB. Produit à l'aide de NanoPlot[139].

Bien que ces données aient un profil différent de celles produites par le séquençage sur *flowcell* MinION, elles représentent un point de comparaison intéressant, car elles sont en théorie ce qui se fait de mieux avec la technologie ONT.

4.4.3 Outils utilisés pour l'alignement

L'alignement des différents fichiers FASTQ sous-échantillonnés a été effectué à l'aide de deux outils selon trois conditions. Le premier outil à avoir été utilisé est l'aligneur Minimap2[57] qui fait office de référence pour l'alignement de longues lectures (au même titre que BWA-MEM[55] pour l'alignement de lectures courtes). L'outil Minimap2 a été utilisé selon deux modes de configuration différents. Premièrement, avec l'ensemble de paramètres prédéfinis pour l'alignement de données issues de la technologie ONT avec la commande *minimap2 -x map-ont*, condition qui sera par la suite nommée *minimap2_ont*. La deuxième condition consiste en un ensemble de paramètres spécifiques permettant l'utilisation de l'outil pbsv à partir des données d'alignement produites de cette manière, voir *Snippet 4.6*. Ces paramètres ne font pas seulement que changer le format des fichiers d'alignement les rendant compatibles avec pbsv, ils changent également le comportement de l'outil en diminuant les pénalités d'extension de *gaps* et ainsi favorisent des alignements plus longs[47]. Cette condition sera par la suite nommée *minimap2_pbsv*.

Snippet 4.6 – Paramètres de l'outil Minimap2 pour la condition *minimap2_pbsv*.

```
1 minimap2 -a --MD --eqx -L -O 5,56 -E 4,1 -B 5 --secondary=no -z 400,50 -r 2k -Y
```

Le second outil pour l'alignement des lectures contre un génome de référence est NGMLR[63]. Il a été sélectionné, car il est l'outil d'alignement conseillé par l'auteur de l'outil de *variant calling* Sniffles[63] lui-même utilisé par le GIAB. À noter que l'auteur de Sniffles est également l'auteur de NGMLR et fait partie du consortium GIAB. L'outil a été utilisé avec les paramètres par défaut.

4.4.4 Outils utilisés pour l'appel de variant

Comme indiqué précédemment, nous avons utilisé trois outils de *variant calling* différents. Tout d'abord, Sniffles, dont seul le paramètre *-s [int]* a varié, conformément aux recommandations de l'auteur[63]. Ce dernier représente le nombre minimum de lectures qui soutiennent un SV. Sachant que la couverture des données que nous allions produire allait être relativement réduite, nous avons défini ce paramètre à deux (condition nommée *sniffles_s2*) afin de pouvoir détecter les SV hétérozygotes même peu couverts.

Le second outil utilisé est pbsv, le *variant caller* de l'entreprise Pacific Bioscience qui peut tout à fait être utilisé sur des données ONT. Pbsv a été utilisé avec les paramètres par défaut et n'a pu produire des données qu'à partir des fichiers d'alignement de la condition *minimap_pbsv*, car seule à comporter les informations nécessaires à son bon fonctionnement.

Enfin, le dernier *variant caller* utilisé est l'outil SVIM[133], lancé avec les paramètres par défaut, conformément aux recommandations de l'auteur[133].

4.4.5 Méthodologie de comparaison

Les fichiers VCF produits par les différents outils d'appel de variants ont été comparés au fichier VCF de référence à l'aide de l'outil Truvari (https://github.com/spiralgenetics/truvari) en utilisant les mêmes paramètres que ceux utilisés par le GIAB[92], voir *Snippet 4.7*. L'outil produit grâce au paramètre *–giabreport* un résumé comportant par exemple, le nombre de faux positifs, de faux négatifs ou encore le taux de rappel et de précision de la comparaison. Les paramètres de Truvari utilisés pour la comparaison sont assez permissifs, ils sont décrits *Table 4.25*.

Snippet 4.7 – Exemple de commande de l'outil Truvari pour la comparaison d'un fichier VCF produit par un couple *aligner*, *variant caller* avec le fichier VCF de référence du GIAB.

```
truvari –b HG002_SVs_Tier1_v0.6.vcf –c query.vcf –o query –r 2000 ––sizemax 1000000 ––pctsim 0 ––passonly ––includebed HG002_SVs_Tier1_v0.6.bed ––giabreport
```

Paramètres	Signification
-b HG002_SVs_Tier1_v0.6.vcf.gz	Fichier VCF de référence
-c query.vcf.gz	Fichier VCF requête
-r 2000	Distance maximale de l'emplacement de référence
–sizemax 1000000	Taille maximale du variant à prendre en compte pour la comparaison
–pctsim 0	Pourcentage minimum de similarité de séquence entre les allèles. Défini à 0 pour ignorer.
–passonly	Ne considérer que les variants avec le champ *FILTER* = *PASS*
–includebed HG002_SVs_Tier1_v0.6.bed	Fichier BED des régions de tiers 1 pour inclure seulement des appels chevauchants
–giabreport	Création d'un rapport présentant des statistiques sur la comparaison à partir des données du GIAB

TABLE 4.25 – Signification des paramètres utilisés pour l'outil Truvari.

4.4.6 Résultats

Les valeurs moyennes de rappel et de précision issues de la comparaison des résultats des différents couples d'outil d'alignement et d'appel de variations sur les trois répétitions des données sous échantillonnées de 1 à 45X sont présentées *Figure 4.58*.

Contrairement aux trois autres outils d'appel de variant, pour pbsv, plus la profondeur de séquençage augmente, plus le rappel et la précision augmentent. Au sein des autres conditions, l'augmentation de la profondeur de séquençage augmente le taux de rappel, mais fait diminuer significativement la précision, car de nombreux faux positifs sont rapportés.

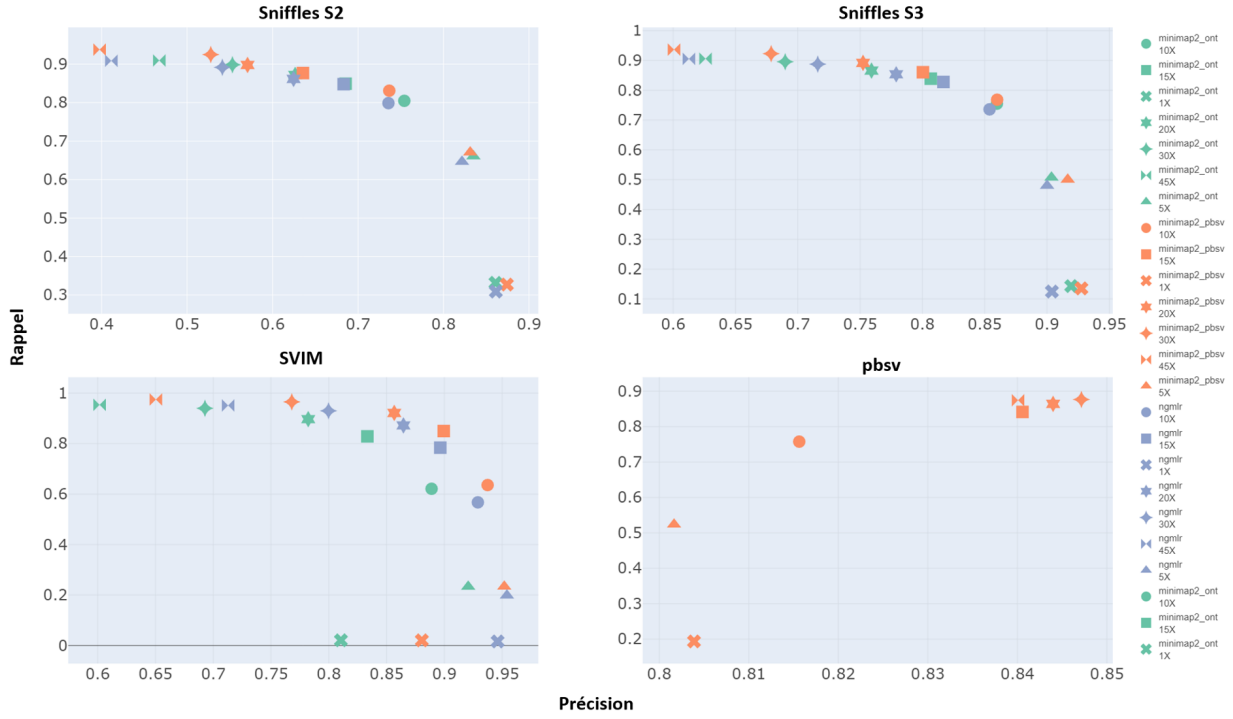


FIGURE 4.57 – Moyenne des taux de rappel et précision des trois sous échantillonnages. Appels de variant effectués avec les couples *minimap_ont*, *minimap_pbs*, NGML et *sniffles_s2*, *sniffles_s3*, pbsv et SVIM, de 1 à 45X

Bien que le couple *minimap_pbsv*, pbsv, présente le meilleur rapport rappel sur précision à hautes profondeurs (15 à 45X), ce n'est pas l'outil qui présente le taux de rappel le plus élevé. C'est la condition *minimap_pbsv* avec le *variant caller* SVIM qui se rapproche le plus proche des 100 % de *recall* (97,5 %) avec une précision relativement honorable (65 %). La condition *sniffles_s3* présente un rappel quasiment équivalent à haute profondeur avec la condition *sniffles_s2*, avec néanmoins une meilleure précision.

En revanche, pour les conditions 5 et 10X de profondeur de séquençage, qui représentent les profondeurs qui nous intéressent le plus, car profondeur estimée d'un séquençage MinION, c'est le couple *minimap_pbsv*, *sniffles_s2* qui présente le meilleur taux de rappel parmi toutes les conditions, voir *Table 4.26*.

A

	Pbsv	Svim			Sniffles S2		
	Minimap2_pbsv	Minimap2_pbsv	Minimap2_ont	ngmlr	Minimap2_pbsv	Minimap2_ont	ngmlr
SIMPLEDEL	0,60	0,28	0,28	0,34	0,75	**0,77**	0,75
SIMPLEINS	0,53	0,22	0,22	0,16	**0,70**	0,67	0,64
DUP	0,33	0,18	0,18	0,16	**0,50**	0,45	0,44
SUBSDEL	0,56	0,19	0,19	0,16	0,77	0,79	**0,80**
SUBSINS	0,52	0,22	0,22	0,14	0,67	**0,71**	0,69
CONTRAC	0,59	0,25	0,25	0,25	0,63	0,63	**0,65**
TOTAL	0,52	0,22	0,22	0,19	**0,67**	0,66	0,64

B

	Pbsv	Svim			Sniffles S2		
	Minimap2_pbsv	Minimap2_pbsv	Minimap2_ont	ngmlr	Minimap2_pbsv	Minimap2_ont	ngmlr
SIMPLEDEL	0,85	0,68	0,68	0,67	0,92	**0,93**	0,92
SIMPLEINS	0,86	0,62	0,60	0,47	**0,88**	0,84	0,81
DUP	0,48	0,56	0,51	0,51	**0,60**	0,54	0,54
SUBSDEL	0,83	0,63	0,66	0,59	**0,96**	0,94	0,95
SUBSINS	0,84	0,59	0,62	0,46	**0,93**	0,92	**0,93**
CONTRAC	**0,83**	0,68	0,67	0,68	0,77	0,75	0,78
TOTAL	0,76	0,63	0,62	0,56	**0,83**	0,81	0,79

TABLE 4.26 – Moyenne des taux de rappel, des trois sous échantillonnages des de la comparaison des différents types de variations.
Produit avec les couples d'outil *minimap_ont*, *minimap_pbsv*, NGMLR et *sniffles_s2*, *sniffles_s3*, pbsv et SVIM, à 5X (A) et 10X (B).

Parmi tous les types de SV détectés présents dans le set de variation, ce sont les duplications et dans une moindre mesure les contractions qui sont les plus difficiles à détecter. Dans les deux cas, ces SV se traduisent par le gain ou la perte d'une portion de séquence strictement identique à une séquence préexistante. Ce motif est particulièrement difficile à détecter à faible couverture, ce qui explique les mauvaises performances pour ces conditions particulières. Pour les autres types de variations, leur taux de rappel semble plutôt équivalent à une profondeur donnée. Entre 5 et 10X, c'est le couple *minimap_pbsv* et *sniffles_s2* qui présentent les meilleures performances de rappel. *Sniffles_s2* présente les meilleurs taux de rappel pour tous les types de SV, sauf pour les contractions qui sont détectées le plus efficacement par pbsv.

À noter que, que ce soit utilisé de concert avec l'outil SVIM ou l'outil Sniffles, l'outil d'alignement NGMLR présente systématiquement les taux de rappels les plus faibles parmi les conditions.

4.5 *Benchmarking* de détection de variants de structure sur données issues de séquenceur MinION

4.5.1 Données de séquençage

Les données de séquençage utilisées au cours du réplicat du *benchmarking* précédent proviennent de deux *runs*, séquencés à 3,5 et 4,5X environ combinés en un seul fichier FASTQ à environ 8X. La distribution de la taille des lectures du fichier FASTQ combiné est visible *Figure 4.58*. Bien que les lectures de ce jeu de données soient moins longues en moyenne que celles du GIAB (N50 à 10,4 kb et taille de moyenne de lecture à 6,1 kb), il n'y a en revanche pas d'enrichissement en petites lectures (taille médiane des lectures à 4,1 kb) grâce à la fragmentation à 15 kb, puis la sélection des fragments opérée avec le kit *Circulomics*.

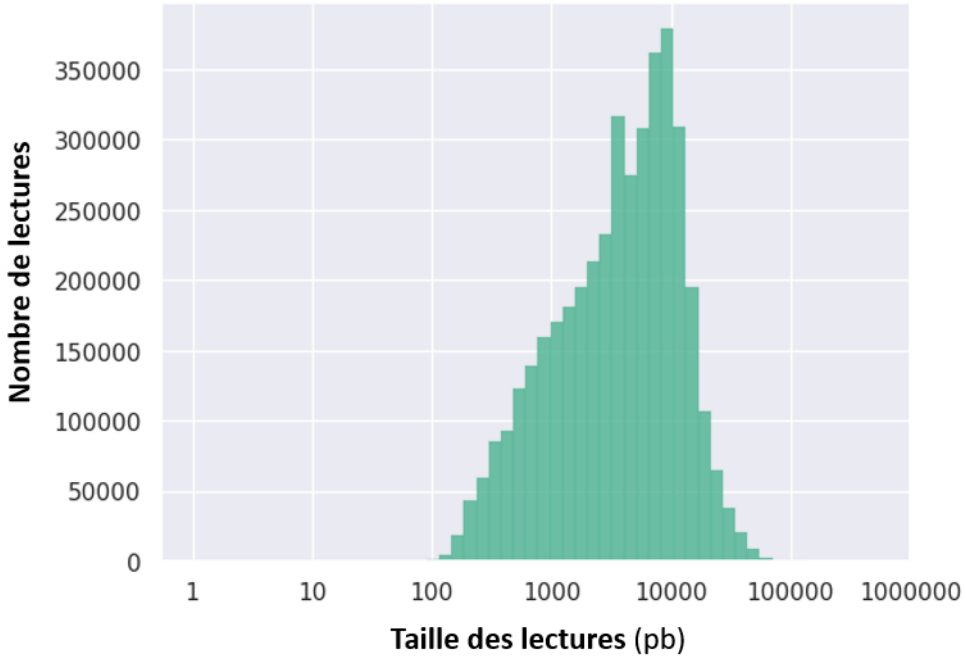


FIGURE 4.58 – Distribution de la taille des lectures du fichier FASTQ 8X produit par Eurofins Biomnis après transformation logarithmique.
Produit à l'aide de NanoPlot[139].

4.5.2 Résultats

Les taux de rappel de l'appel de variants à partir des données de séquençage de l'individu HG002 8X Eurofins Biomnis par rapport aux conditions 5 et 10X du *benchmark* des données du GIAB sont visibles *Figure 4.59*.

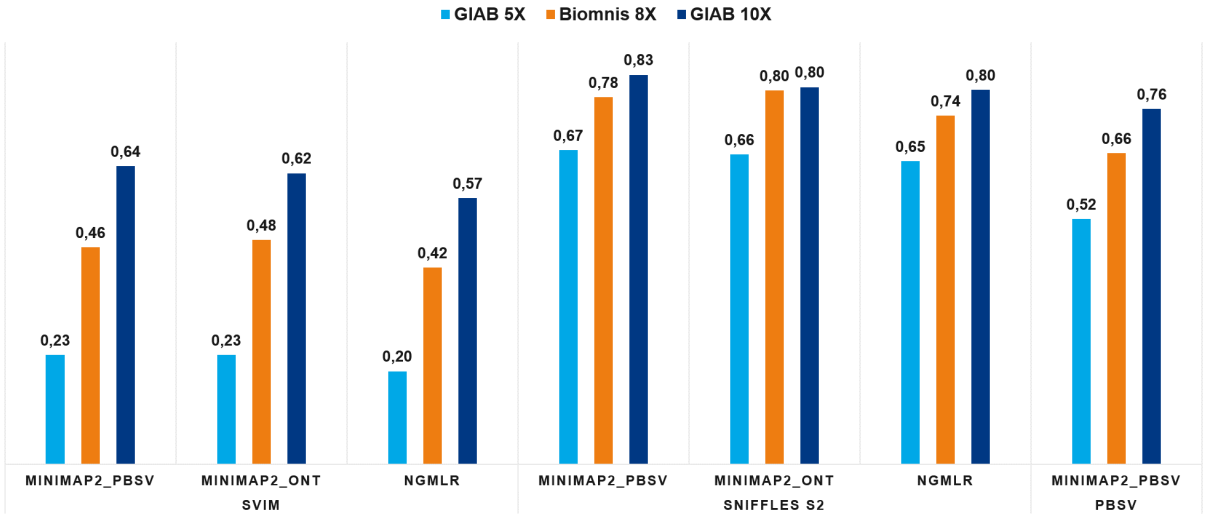


FIGURE 4.59 – Taux de rappel des données Biomnis 8X contre les données GIAB 5 et 10X.

Tout d'abord, peu importe l'outil d'alignement utilisé, l'outil SVIM présente les taux de rappel les plus bas à 5, 8 et 10X. À noter que le taux de rappel pour les conditions SVIM à 8X sont contenus entre les valeurs des conditions 5 et 10X, comme attendu.

La valeur de rappel la plus importante pour la condition 8X est obtenue avec le couple *minimap_ont*, *sniffles_s2* qui atteint les mêmes performances à 8X sur les données Biomnis qu'à 10X sur les données GIAB. C'est pourtant parmi les trois conditions de profondeur de séquençage, le couple *minimap_pbsv*, *sniffles_s2* qui donne les meilleurs résultats de rappel sur les données 10X. Cette différence s'explique peut-être par la différence de profil des données PromethION et MinION utilisées lors de ce benchmark.

	Pbsv	Svim			Sniffles S2		
	Minimap2_pbsv	Minimap2_pbsv	Minimap2_ont	ngmlr	Minimap2_pbsv	Minimap2_ont	ngmlr
SIMPLEDEL	0,73	0,49	0,52	0,52	0,86	**0,89**	0,88
SIMPLEINS	0,69	0,44	0,46	0,31	0,80	**0,82**	0,72
DUP	0,45	0,46	0,44	0,42	**0,65**	0,64	0,56
SUBSDEL	0,65	0,34	0,42	0,35	0,78	**0,88**	0,86
SUBSINS	0,70	0,35	0,39	0,29	0,74	**0,82**	0,72
CONTRAC	**0,76**	0,52	0,52	0,51	0,75	0,73	0,74
TOTAL	0,66	0,46	0,47	0,41	0,78	**0,79**	0,74

TABLE 4.27 – Taux de rappel des différents types de SV de la comparaison avec les données Eurofins Biomnis 8X.
Produit avec les couples d'outil *minimap_ont*, *minimap_pbsv*, NGMLR et *sniffles_s2*, *sniffles_s3*, pbsv et SVIM.

Les taux de rappels des différents types de SV de la comparaison avec les données Eurofins Biomnis 8X sont visibles *Table 4.27*. Comme précédemment, les duplications sont le type de SV détecté le moins efficacement. En revanche, alors que la plupart des meilleurs taux de rappels avaient été détectés avec le couple *minimap2_pbsv*, *sniffles_s2* sur les données PromethION, c'est le couple *minimap2_ont*, *sniffles_s2* qui performe le mieux sur les

données MinION. Comme précédemment, les contractions sont détectées le plus efficacement par l'outil pbsv. Le couple d'outils *minimap2_ont*, *sniffles_s2* étant celui qui présente les meilleurs taux de rappel sur les données issues de séquençage MinION, c'est celui qui a été choisi pour être utilisé lors de l'étape suivante de détection de SV références issus de données patients.

4.6 Détection de variants structuraux connus dans le cadre de séquençage de routine clinique Oxford Nanopore Technologies

4.6.1 Séquençage et alignement de données issues patients provenant de la routine diagnostique présentant des SV de référence

Treize patients issus de la routine diagnostic du laboratoire Eurofins Biomnis présentant des variations de structures de types et de taille divers, validées par au moins une technique de référence, ont été séquencés sur flowcell MinION. La description des différents SV de référence et des données séquencées est visible *Table 4.28*.

N°	Variants	Diagnostic	Profondeur moyenne	Taille médiane lectures	N50
1	Inversion	46,XY,inv(12)(p13.1q12).ish	9X	3.7kb	6.9kb
2	Expansions nucléotidiques	X-fragile; exon 1 de FMR1: >200 "GCC"	9X	7.5kb	10.1kb
3	Translocation réciproque	46,XY,t(18,22)(q21.1;q13.3).ish	11X	7.0Kb	9.2kb
4		46,XX,t(3,17)(p21.31;q25.3).ish	6X	3.6kb	7.6kb
5		46,XX,t(7;20)(q22;q13.3).ish	6X	4.1kb	7.4kb
6		46,XX,t(4;8)(p13;q22).ish	9X	9.8kb	12.4kb
7	Aneuploïdie	47,XXY.nuc	7X	5.7kb	8.6kb
8	Délétion	CGH => 15:23652850-28544359;DEL; length=**4.9Mbp**	6X	7.5kb	8.9kb
9	Duplication	CGH => 22q11.1q11.21(16114244_18844632)x4; length=**2,8Mb**	8X	8.3kb	10.2kb
10		CGH => 1q21.1q21.2(145388137_147820342)x3;lenght=**2,4Mb**	8X	7.1kb	11,2kb
	Translocations impliquant un centromère ou un télomère				
11	Centromère	46,XY,t(11;17)(p10;q10).ish	7X	8.6kb	10.4kb
12	Télomère	46,XX,t(1;5)(q44;q23.2).ish	9X	8.7kb	10.7kb
13	Robertsonienne	45,XX,der(13;14)(q10;q10)nuc	7X	1.3kb	6.7kb

TABLE 4.28 – Description des données des treize SV validés par méthode de référence orthogonale issues de patients de la routine clinique du laboratoire Eurofins Biomnis, séquencés par MinION.

La totalité des patients présentant ces SV ont été séquencés sur flowcell MinION et présentent des profondeurs moyennes comprises entre 6 et 11X ainsi que des N50 comprises entre 6 et 13 kb. L'ensemble des données d'alignement ont été produites à l'aide de l'outil Minimap2[57] et des paramètres prédéfinis pour les données issues de la technologie Oxford Nanopore.

Certains de ces SV de référence présentent des variations très différentes de celles testées dans les *benchmarks* précédents. En effet, les variations du set de référence ne dépassent pas 114 kb et ne contiennent pas tous les types de SV rencontrés en routine clinique. Il a donc fallu intégrer de nouveaux outils spécialisés dans la détection d'autres types d'anomalies, notamment de variant de très grande taille.

4.6.2 Outils supplémentaires utilisés pour la détection de types de réarrangements spécifiques

Trois outils supplémentaires d'alignement et, ou, d'appel de variation ont été utilisés pour détecter les SV de nos treize patients. Le premier est l'outil LAST[132], qui est un *aligner*, assez ancien dans sa conception et son fonctionnement, mais aussi un *variant caller*. À l'aide de commandes visibles *Snippet 4.8*, LAST permet de détecter les *split-reads* et ainsi de détecter de possibles réarrangements structuraux, au format MAF. Ce dernier, au même titre que le format VCF est un format tabulé qui contient pour chaque variant une ligne informant des caractéristiques de ce dernier. LAST est un outil extrêmement lent en comparaison avec les outils les plus consensuels du domaine, mais également plus sensible.

Snippet 4.8 – Exemple d'un alignement et d'un appel des lectures scindées avec l'outil LAST.

```
1 lastdb -uNEAR db genome.fasta
2 last-train -Q1 db q.fastq > train.out
3 lastal -p train.out db q.fastq | last-split > out.maf
```

1 : Indexation du génome. 2 : détermination des paramètres des séquences. 3 : alignement et appel de variations

Le second outil, NCRF[134] pour *Noise-Cancelling Repeat Finder*, est un logiciel spécialisé dans la détection de motifs particuliers insérés dans des répétitions en tandem issues de données de séquençage *long-reads* bruitées. Cette stratégie est tout à fait adaptée à la détection de maladies dites à triplets, caractérisées par un nombre pathologique d'expansions nucléotidiques spécifiques d'une taille d'un multiple de trois. Le nombre de répétitions étant cruciale pour la détermination du pronostic pour le patient et ces expansions pouvant aller jusqu'à plusieurs centaines de bases, l'utilisation de longues lectures est tout indiquée, bien que des techniques de références de biologie moléculaire existent déjà.

Le dernier outil, AneuFinder[135] est un outil de détection de CNV de grande taille, d'aneuploïdies et de leurs points de cassure à partir de données de séquençage. Il permet également la mesure de statistiques au niveau du caryotype et la mise en forme des données en des représentations type CGH ou caryotype.

La totalité des anomalies recherchées a été vérifiée à l'aide de l'outil de visualisation IGV[136] à partir des fichiers d'alignement.

4.6.3 Résultats

Les résultats issus de la détection des SV chez les treize données patients présentant un CNV validé par au moins une technique de référence orthogonale sont représentés *Table 4.29*.

N°	Variants	Diagnostic	Résultats	Profondeur moyenne	Taille médiane lectures	N50
1	Inversion	46,XY,inv(12)(p13.1q12).ish	sniffle_s2	9X	3.7kb	6.9kb
2	Expansions nucléotidiques	X-fragile; exon 1 de FMR1: >200 "GCC"	NCRF	9X	7.5kb	10.1kb
3	Translocation réciproque	46,XY,t(18,22)(q21.1;q13.3).ish	sniffle_s2	11X	7.0Kb	9.2kb
4		46,XX,t(3,17)(p21.31;q25.3).ish	sniffle_s2	6X	3.6kb	7.6kb
5		46,XX,t(7;20)(q22;q13.3).ish	Last	6X	4.1kb	7.4kb
6		46,XX,t(4;8)(p13;q22).ish	sniffle_s2	9X	9.8kb	12.4kb
7	Aneuploïdie	47,XXY.nuc	AneuFinder	7X	5.7kb	8.6kb
8	Délétion	CGH => 15:23652850-28544359;DEL; length=**4.9Mbp**	AneuFinder	6X	7.5kb	8.9kb
9	Duplication	CGH => 22q11.1q11.21(16114244_18844632)x4; length=**2,8Mb**	AneuFinder	8X	8.3kb	10.2kb
10		CGH => 1q21.1q21.2(145388137_147820342)x3;lenght=**2,4Mb**	no	8X	7.1kb	11,2kb
	Translocations impliquant un centromère ou un télomère					
11	Centromère	46,XY,t(11;17)(p10;q10).ish	no	7X	8.6kb	10.4kb
12	Télomère	46,XX,t(1;5)(q44;q23.2).ish	no	9X	8.7kb	10.7kb
13	Robertsonienne	45,XX,der(13;14)(q10;q10)nuc	no	7X	1.3kb	6.7kb

TABLE 4.29 – Résultats du SV *calling* sur les données des treize SV validés par méthode de référence orthogonale issues de patients de la routine clinique du laboratoire Eurofins Biomnis, séquencés par MinION.

Les SV ont d'abord été divisés en deux catégories, les translocations impliquant un centromère ou un télomère et les autres. Les premières ne sont pas détectables en l'état en utilisant des stratégies d'alignement sur un génome de référence, puis d'appel de variant, car ces zones sont absentes du génome de référence.

Pour les autres, 90 % des variations séquencées ont été détectées. La duplication notée numéro 10 est la seule qui n'a été détectée par aucune de nos stratégies. Néanmoins, la zone présentant le SV validé a été observée avec l'outil de visualisation IGV[136] et une augmentation sensible de la couverture à cet endroit a été observée. Malheureusement, cet échantillon étant assez peu séquencé en moyenne et l'anomalie à détecter étant une duplication (par essence difficile à séquencer à basse profondeur, comme vu précédemment), il peut être supposé qu'un séquençage plus profond aurait permis de détecter le SV. Un autre séquençage à 6X avec un tirage plus favorable de séquençage des régions d'intérêt aurait peut-être pu permettre la détection du SV, mais la seule manière de minimiser l'impact du hasard sur la détection est de maximiser la profondeur moyenne de séquençage.

En ce qui concerne les autres anomalies, les SV notés 1, 3, 4 et 6 ont été détectés à l'aide du couple *minimap_ont*, *sniffles_s2*, ce qui signifie que le SV de référence connu a été détecté avec succès et est présent dans le fichier VCF de l'appel de variant. La translocation réciproque notée 5 n'a pas été détectée par le couple *minimap_ont*, *sniffles_s2*, mais l'a été par LAST et sa commande *last-split*. LAST étant un outil plus sensible que le couple précédent, cela peut expliquer pourquoi il a été en mesure de le détecter. Encore une fois, l'échantillon étant assez peu couvert en moyenne, il est probable qu'avec des données mieux couvertes, il aurait été détecté par l'association de *minimap_ont* et *sniffles_s2*.

L'expansion nucléotidique notée 2 a été détectée avec succès grâce à l'outil NCRF. Comme expliqué précédemment, NCRF est un outil excellent dans la détection de répétitions de petits motifs particuliers. Le Syndrome de l'X Fragile est caractérisé par la répétition de triplets CCG dans la séquence de l'exon 1 du gène FMR1. Ici, l'outil a permis de détecter que le nombre de répétitions de triplets était supérieur à 200, ce seuil représentant la mutation dite complète de l'exon, entraînant l'inactivation du gène FMR1 et le pire pronostic possible pour ce syndrome.

Enfin, les CNV de grande taille ou aneuploïdies notées de 7 à 9 ont été détectées avec succès à l'aide de l'outil AneuFinder[135].

4.7 Discussion

L'utilisation de la technologie ONT et plus particulièrement du séquençage MinION permet donc la détection de SV présents dans les zones inclus dans le génome de référence utilisé pour l'alignement. Les données issues d'ONT permettent de détecter au sein d'une même technologie de nombreux types d'anomalies différents qui sont actuellement détectés en routine par plusieurs technologies de référence, comme la *CGH-array*, le caryotype ou encore les tests moléculaires ciblés comme la FISH ou la MLPA. Néanmoins, les technologies ONT ne peuvent pas encore remplacer toutes ces technologies, car certains types d'anomalies, comme les translocations impliquant un centromère ou un télomère ou les duplications sont difficiles voir impossibles à détecter en l'état actuel de la technologie. Même si certains de ces SV peuvent également être détectés par séquençage *short-reads*, les technologies *long-reads* permettent d'étudier certaines zones qui étaient auparavant inaccessibles avec les lectures courtes[92][93].

Certaines de ces limitations pourront être corrigées par l'amélioration des génomes de référence disponibles pour la communauté, notamment avec des références contenant certaines des zones qui en sont actuellement absentes, comme les télomères et les centromères[13]. D'autres de ces limitations sont corrigibles en séquençant plus profondément[47]. L'évolution de la technologie permettra sans doute dans quelques années d'avoir un rendement plus important par *flowcell*, ce qui est déjà le cas aujourd'hui. Il y a quelques années, il était inenvisageable de séquencer un échantillon humain à 10X en une seule *flowcell* MinION. Pour surmonter ces limitations, il est également possible de passer au modèle de séquenceur PromethION permettant le séquençage d'un échantillon humain à environ 30 à 40X sur une seule *flowcell* avec un protocole de séquençage *long-reads* classique. Séquencer à de telles profondeurs permet en plus de la détection des SV, l'appel de variation de petite taille. Mais le séquenceur PromethION étant assez récent et les technologies Oxford Nanopore moins robustes que le séquençage WGS Illumina, cela représente des freins encore trop importants pour leur démocratisation en routine clinique diagnostique au sein des laboratoires.

N° échantillon	Nombre variants	Nombre variants spécifiques
1	23467	1483
2	25320	2316
3	22347	1491
4	24438	4473
5	20990	1281
6	19732	1523
7	19738	1550
8	19456	1647
9	22034	1825
10	20387	1664
11	21916	1405
12	21234	1395
13	21550	1896
14	19498	1768
15	23746	2816
Moyenne	**21723**	**1902**

TABLE 4.30 – Nombre de variants totaux et uniques partagés entre quinze différents échantillons. Appel effectué avec le couple *minimap2_ont*, *sniffles_s2*

En revanche, bien qu'un séquençage MinION à basse couverture (5-10X) puisse permettre la détection de SV, il est difficile d'utiliser cette technologie en première intention. En effet, le séquençage, puis l'appel de variation produisent énormément de variants, dont une grande partie est composée de polymorphismes, mais aussi de faux positifs, voir *Table 4.30*.

Le nombre de SV détectés par individu produit par le couple *minimap2_ont*, *sniffles_s2* est trop important pour espérer détecter des SV non connus de prime abord. Ce nombre a potentiellement été maximisé, car nous avons choisi de favoriser le taux de rappel au détriment de la précision. Malgré cela, nous pouvons nous permettre de filtrer les variants récurrents, car nous nous concentrons uniquement sur les variants rares. Lorsque ce travail a été conduit, de la fin de l'année 2019 à mi 2020, la version 3.1 de GnomAD[103], qui représente la base de données de référence de fréquence pour les SV, n'avait pas encore été publiée (octobre 2020). Si ce travail devait être poursuivi afin de permettre la filtration de SV détectés à partir de données de séquençage faiblement couvertes, la stratégie choisie serait dans un premier temps la filtration sur la fréquence des variants dans la population générale à partir d'une cohorte interne et de bases de données telles que GnomAD SV. Si jamais cela venait à ne pas être suffisant, le choix d'un autre couple outil avec un meilleur rapport rappel sur précision, comme l'association *minimap_pbsv*, pbsv pourrait être envisagée. Enfin, l'utilisation de bases de données curées par les experts associant SV et impact clinique, telle que ClinVar[71], pourrait également être envisagée.

Le séquençage ONT représente donc une des futures possibilités pour la détection de SV de manière pan-génomique dans un cadre de routine clinique. Bien qu'encore trop immature pour une intégration dans la routine actuelle, les évolutions futures de la technologie, ainsi que les outils et supports pour détecter, filtrer et trier les SV à partir des données de séquençage ONT, permettront un jour peut-être, son arrivée massive en diagnostic humain.

Priorisation phénotypique de données d'exomes cas index pour le diagnostic de maladies rares

5.1 Motivation

Au cours de l'année 2018, j'ai été sollicité pour participer à un projet réunissant plusieurs acteurs dont le CHU Grenoble Alpes, le CHU Dijon Bourgogne, le CHU de Rennes, le CHU de Brest, les HCL de Lyon, le laboratoire Eurofins Biomnis et l'entreprise SeqOne.

Bien que la détection de variations reste un enjeu, elle est néanmoins de mieux en mieux maitrisée. Cela induit, de manière concomitante avec l'augmentation du nombre d'examens génomiques réalisés, la production de volumes de variations à interpréter de plus en plus importants. La priorisation des variations permet d'adresser certains des problèmes posés par l'interprétation d'examen génomiques, en réduisant et ordonnant le nombre de variations d'intérêt à étudier, notamment lorsque les stratégies de diagnostic tendent à maximiser le rappel et donc augmenter le nombre de variants à étudier.

La comparaison a été effectuée sur un ensemble de données d'exomes récolté parmi les différents centres. Chacun de ces exomes était accompagné de termes HPO décrivant le phénotype du patient ainsi que d'un ou plusieurs gènes supposés causaux du phénotype patient. Les données de variations utilisées dans cette étude ne portaient donc seulement que sur les variations de petite taille exoniques.

5.2 Jeu de données d'exomes phénotypés

Les exomes ont été produits et interprétés de manières propres aux différents centres. Les fichiers ont été transférés dans différents formats ce qui a nécessité une étape d'unification pour leur exploitation. Seules les informations de l'ID du patient, du centre prescripteur, du ou des gènes causaux, des termes HPO associés au patient ainsi que le fichier VCF récapitulant les variations ont été utilisés de cette étude.

Le jeu de données d'exome sur lequel ont été testés les différents outils de priorisation phénotypique est composé de 322 patients, dont 6 avec deux gènes potentiellement causaux, soit 328 diagnostics différents. La distribution du nombre de termes HPO par centre est visible *Figure 5.60*.

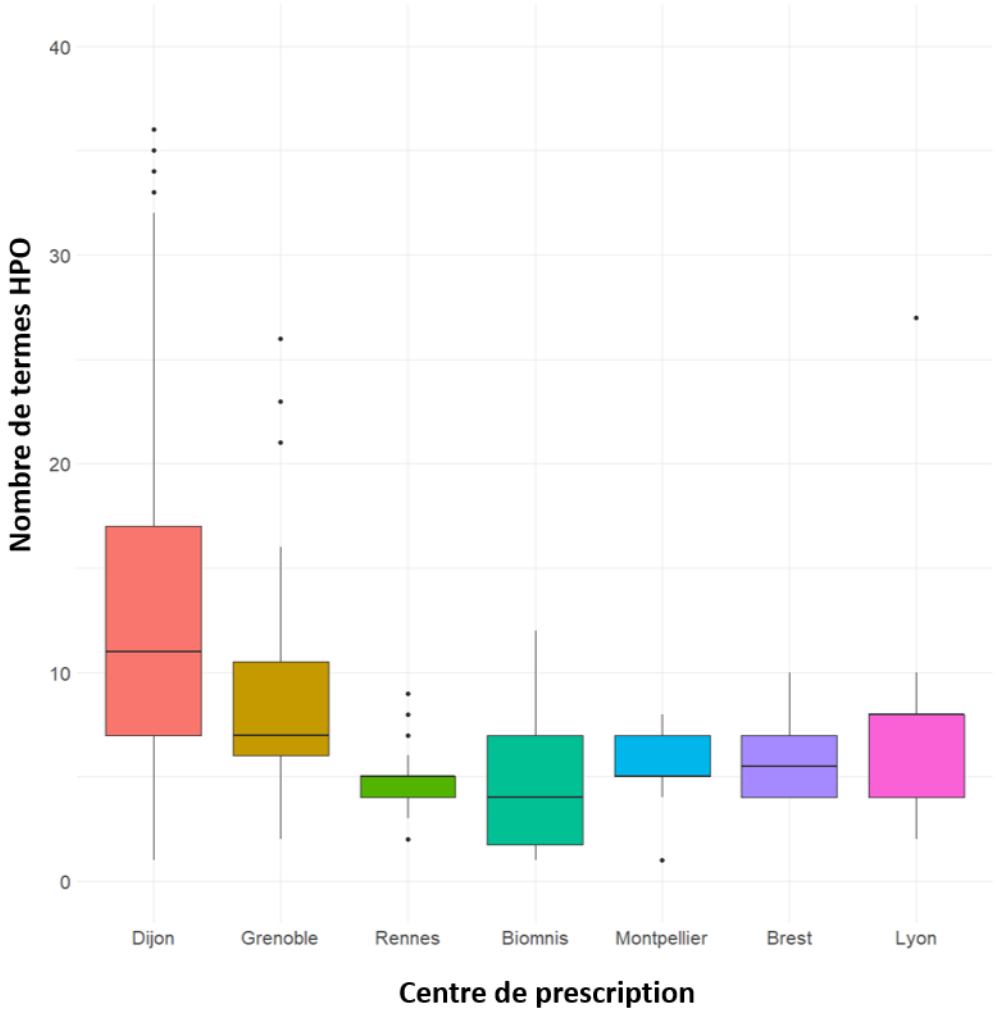


FIGURE 5.60 – Distribution du nombre de termes HPO par patient en fonction du centre de prescription.

Plusieurs patients ne possèdent qu'un seul terme HPO, alors que le nombre maximal de termes HPO renseigné pour un patient est de 71, puis 36. Les patients sont donc phénotypés de manière assez inégale entre centres et même au sein d'un même centre. Cela est probablement dû à une variabilité de méthodologie de phénotypage entre les cliniciens, aucune consigne particulière n'ayant été donnée, hormis de renseigner au moins un terme. Il peut être supposé que la manière de phénotyper de l'ensemble des centres ayant fourni un nombre assez important de patients de cette cohorte représente les pratiques de phénotypage de routine de ces centres au moment de l'étude. Il est communément admis qu'un phénotypage de bonne qualité contient cinq termes HPO, les plus précis possibles (en fonction de la position des termes dans l'ontologie)[140]. Trop de termes peu informatifs peuvent être aussi contreproductif que pas assez.

5.3 Outils de priorisation phénotypique

Plusieurs stratégies de priorisation existent. Celle qui a été retenue dans cette étude est la priorisation phénotypique. Elle consiste en la liaison entre certains gènes décrits comme impliqués en maladie humaine et les phénotypes provoqués par l'altération de ces gènes. À partir des informations des gènes pour lesquels des variations ont été détectées ainsi que de la description phénotypique de l'individu, les outils de priorisation vont mettre en valeur certains gènes susceptibles d'être à l'origine du phénotype pathologique.

Trois outils ont été comparés au sein de cette étude. Les outils académiques Exomiser[141] et AMELIE[142], ainsi que l'outil de priorisation développé par l'entreprise SeqOne (https://seqone.com/).

5.3.1 Exomiser

Exomiser est un outil de priorisation phénotypique basé sur des filtres de fréquence et d'impact estimé des variations, combinés à des modèles d'association préexistants de phénotypes humains et animaux. Les différentes étapes de l'outil Exomiser sont schématisées *Figure 5.61*. Exomiser nécessite en entrée, les termes HPO, le ou les fichiers VCF et un fichier PED dans le cas où serait fourni à l'outil plus d'un individu, comme pour l'analyse d'un trio par exemple.

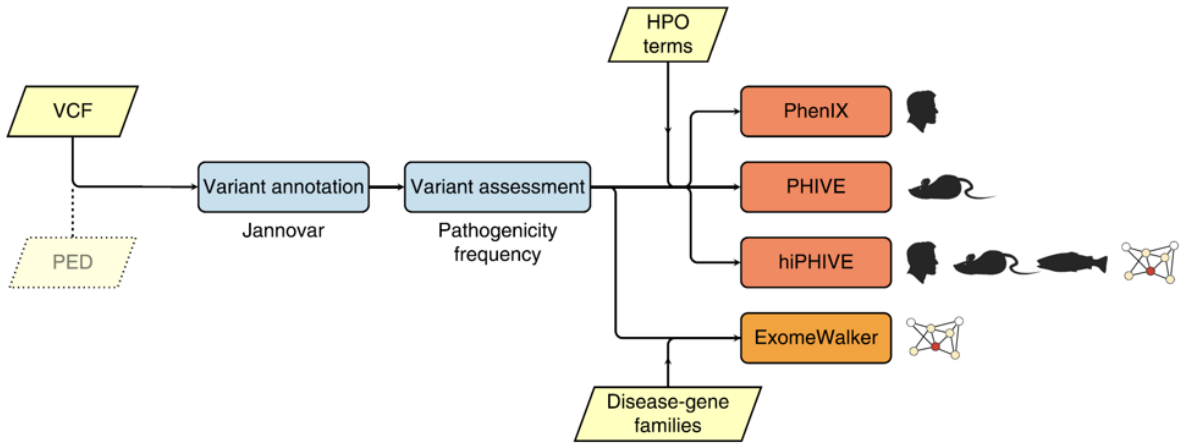


FIGURE 5.61 – Schématisation des différentes étapes de l'outil Exomiser..
D'après, *Smedley, D., Jacobsen, J. O., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., ... Robinson, P. N. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature protocols, 10(12), 2004-2015.*

Les fichiers VCF sont tout d'abord annotés puis filtrés en fonction d'un score par variant lié à la fréquence de ce dernier dans les données du 1KGP et d'un score par gène relatif à sa pathogénicité estimée si altéré.

Deux modèles différents d'association gène phénotype sont encore maintenus au sein de l'outil Exomiser, *PheniX* et *hiPHIVE*. Le modèle *PheniX* se base sur des données de comparaison de phénotypes avec des données cliniques préexistantes et est à télécharger séparément de l'outil. Les variants sont priorisés à l'aide d'un score basé sur la pathogénicité et les similarités sémantiques de phénotypes de patient en lien avec des termes HPO de maladie dont l'étiologie a été clarifiée. Le point faible de ce modèle est donc qu'il ne permet d'attribuer un score qu'aux variants et aux gènes qui ont déjà été mis en relation avec une ou plusieurs pathologies. Plusieurs versions du modèle *PheniX* étaient disponibles au moment de l'étude (**18***02*, **18***05*, **année**, *mois*).

Le modèle *hiPHIVE* (*human interactome PHIVE*) est basé sur des comparaisons de phénotypes humains et inter-espèces (souris, poisson-zèbre). Contrairement au modèle *PheniX*, le modèle *hiPHIVE* peut donner des résultats

pour des associations gène, phénotype qui n'ont pas encore été mis en évidence chez l'humain. Cette stratégie suppose que s'il existe un modèle animal pour le gène contenant la mutation associée à la maladie, il est probable qu'il présente une similarité phénotypique avec les phénotypes cliniques humains.

Les données murines proviennent de la MGD[143] (*Mouse Genome Database*) et de l'IMPC[144] (*International Mouse Phenotyping Consortium*). La couverture des gènes codant pour des protéines humaines par des données phénotypiques de souris était de 33 % en 2017 pour 7000 maladies rares[145]. Les données du poisson-zèbre proviennent de la base de données ZFIN (*Zebrafish Model Organism database*) utilisant l'ontologie anatomique du poisson-zèbre[146], l'ontologie des gènes du poisson-zèbre[147] ainsi que l'ontologie des traits du poisson-zèbre[148]. Ces deux ensembles de données sont croisés avec les bases OMIM[105] et Orphanet[149] pour trouver de potentielles associations gène, pathologie entre l'humain et l'animal. Plusieurs versions du modèle *hiPHIVE* étaient disponibles au moment de l'étude (1711, 1802, 1805).

5.3.2 AMELIE

L'outil AMELIE[150] (Automatic Mendelian Literature Evaluation) est un outil de priorisation des variants à partir d'une base de connaissance littéraire. Le principe de construction de la base de connaissance est visible *Figure 5.62*.

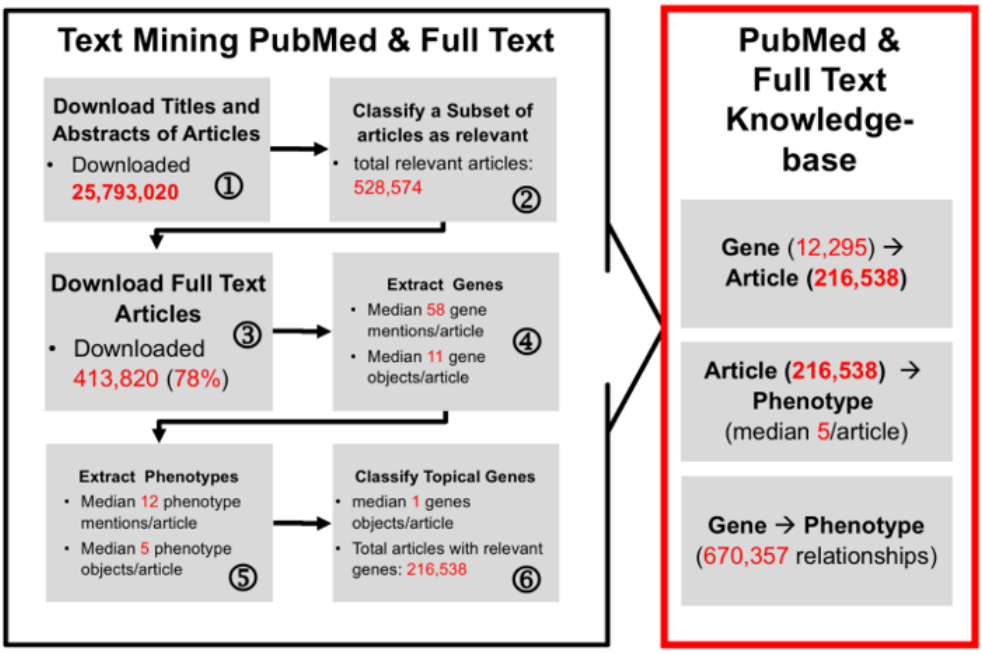


FIGURE 5.62 – Méthodologie de construction de la base de connaissance de l'outil AMELIE. D'après, *Birgmeier, J., Haeussler, M., Deisseroth, C. A., Jagadeesh, K. A., Ratner, A. J., Guturu, H., ... Bejerano, G. (2017). AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. BioRxiv, 171322.*

L'intégralité de la base de données bibliographique relative aux sciences biologiques et biomédicales, MEDLINE, a été téléchargée *via* son moteur de recherche PubMed (https://pubmed.ncbi.nlm.nih.gov/). Un algorithme de fouille de données a été entraîné sur l'ensemble des articles téléchargés afin de détecter les phénotypes et les gènes dans le corps de texte. Un deuxième classifier a été entraîné afin de corréler les phénotypes et les gènes entre eux, voir *Figure 5.63*.

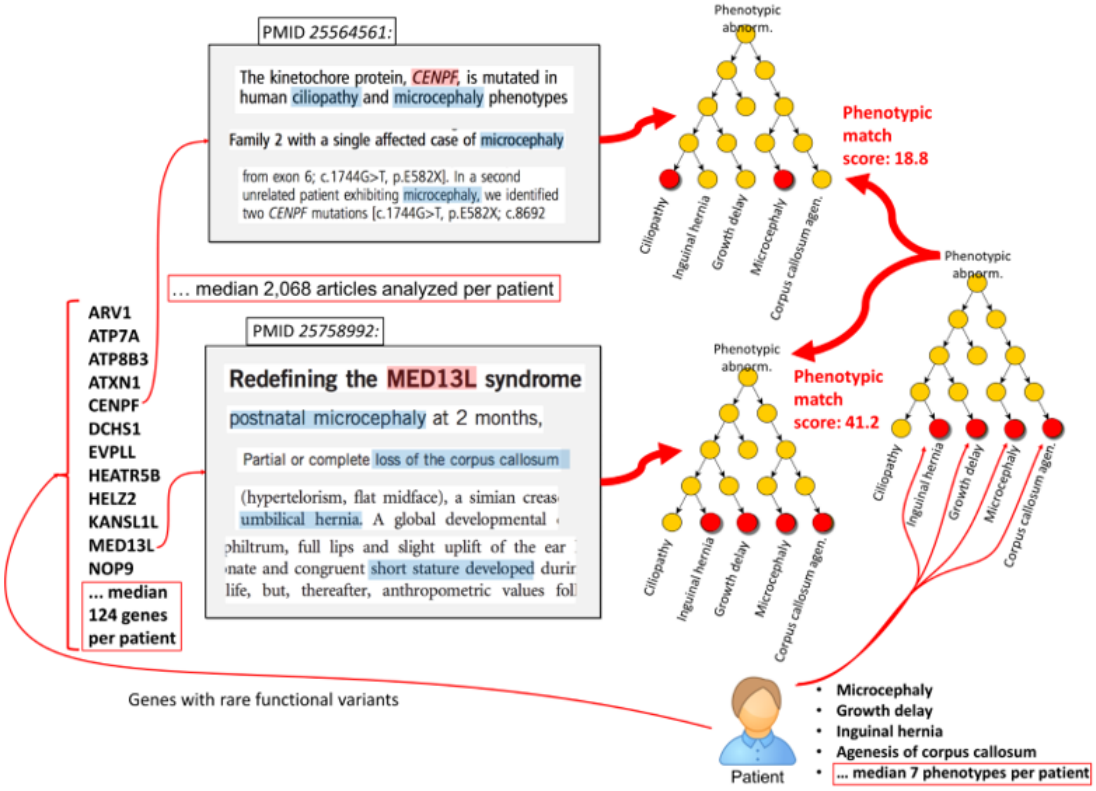


FIGURE 5.63 – Méthodologie d'extraction et de classification des relations gènes, phénotype par l'outil AMELIE.
D'après, *Birgmeier, J., Haeussler, M., Deisseroth, C. A., Jagadeesh, K. A., Ratner, A. J., Guturu, H., ... Bejerano, G. (2017). AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. BioRxiv, 171322.*

L'outil AMELIE est requêtable *via* un portail web ou une API et nécessite une liste de termes HPO ainsi qu'un fichier VCF pour la priorisation des variations.

5.3.3 SeqOne Scout

L'entreprise SeqOne a développé l'outil de priorisation SeqOne Scout et l'a intégré à l'interface d'interprétation. Les cliniciens peuvent ainsi sélectionner des filtres qui prioriseront les variations restantes.

Bien que le fonctionnement de l'outil exact ne soit pas connu, celui-ci adopte une stratégie similaire à l'outil AMELIE. SeqOne Scout est basé sur une base de connaissance littéraire issue de MEDLINE, mais pas seulement. Des bases de données comme ClinVar[71] et OMIM[105] sont intégrées à l'algorithme de l'outil pour donner du poids à certains variants déjà renseignés comme pathologique.

L'algorithme a été décliné en quatre versions, plus ou moins élaborées, pour ce test, SeqOne Brut, SeqOne Basic, SeqOne Scout. Chacun de ces trois modèles existe en version sans ClinVar. Cette distinction est importante, car certains des diagnostics utilisés dans cette étude avaient été soumis dans la base de données ClinVar, créant ainsi un

biais de détection favorisant les conditions avec ClinVar. Les outils SeqOne Brut et SeqOne Basic se basent seulement sur des filtres de fréquence et de pathogénicité en utilisant les associations gènes phénotypes des termes HPO et de la base de données OMIM.

5.4 Résultats

Les résultats de priorisation des différents outils sur le jeu de données d'exome sont représentés *Figure 5.64*. Les différentes positions de chacun des diagnostics pour chaque outil et ses modèles associés ont été comptabilisées.

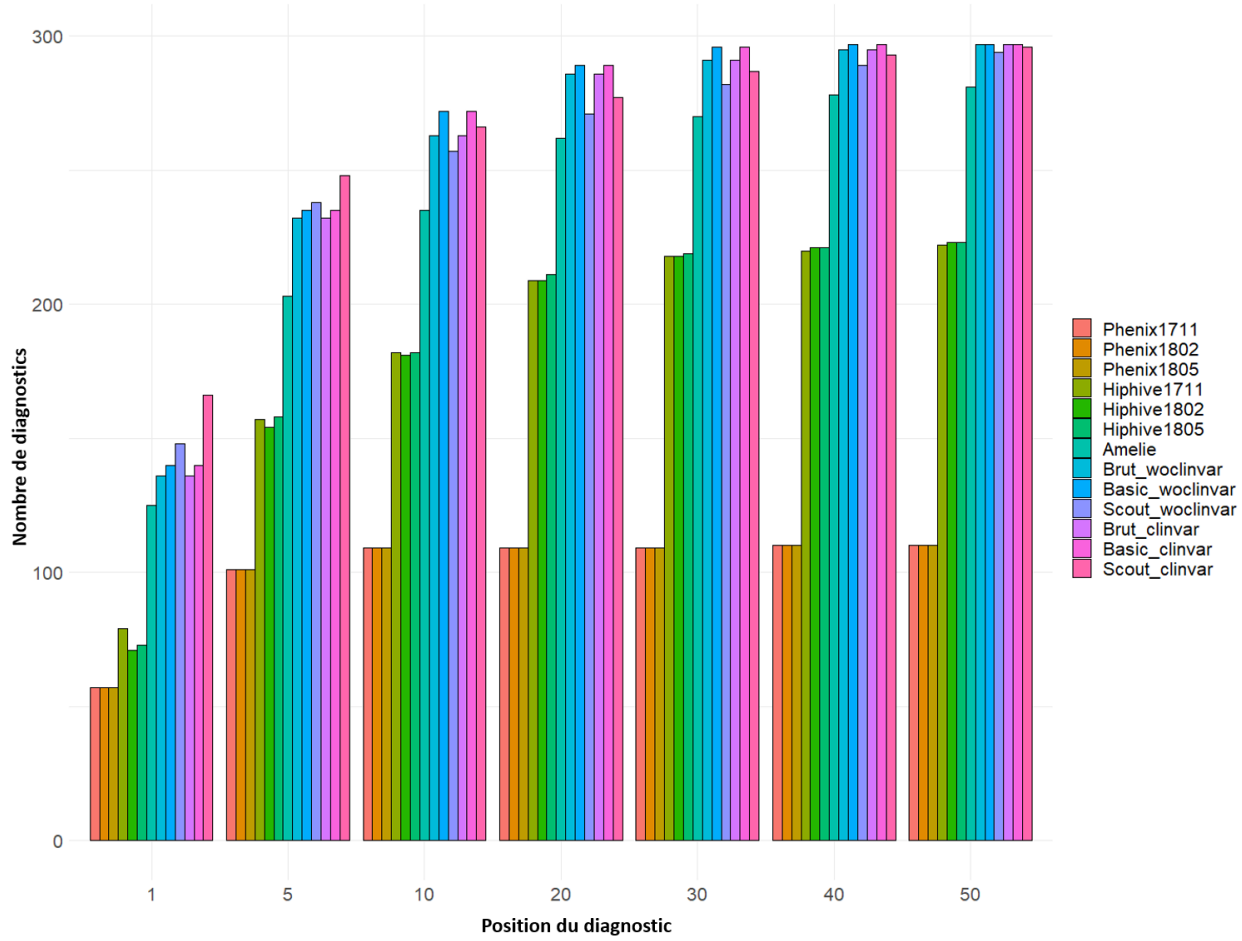


FIGURE 5.64 – Nombre de diagnostics à la position X ou inférieur en fonction de l'outil et du modèle utilisé. wo_clinvar : sans ClinVar.

Tout d'abord, le modèle *PheniX* de l'outil Exomiser est celui qui présente les résultats les plus faibles, ce qui était attendu. En effet, l'outil atteint un plateau extrêmement rapidement, à partir de la condition "position 10 et moins", aucun nouveau diagnostic n'est apporté, peu importe la version du modèle utilisée. Les 109 diagnostics retrouvés par le modèle *PheniX* peuvent être considérés comme les plus simples à détecter, car ce modèle étant le plus limité. En ce qui concerne le modèle *hiPHIVE*, celui-ci présente également un plateau plus haut que le modèle précédent à partir de la condition "position 20 et moins". En revanche, le nombre de diagnostics augmente tout de même, de manière limitée, de la position 20 à la position 50 (211 à 223). Les différences entre les résultats des versions du modèle *hiPHIVE* sont quasi-nulles. L'outil Exomiser est le moins performant des trois.

Les résultats des outils AMELIE et SeqOne sans ClinVar sont très proches pour les diagnostics à la première position. Ensuite, l'écart se creuse jusqu'à la position 50. Les conditions Scout avec ClinVar, suivi par la condition sans ClinVar, dominent jusqu'aux 10 premières places. Elles sont ensuite rattrapées par la version Basic de l'outil, puis la version Brut à partir de la position 20, pour atteindre un plateau à la position 50. À la position 50, l'outil Scout présente le plus grand nombre de diagnostics (297/328), puis l'outil AMELIE (281) puis l'outil Exomiser (223).

Certains des diagnostics sont retrouvés au-delà de la position 50, mais certains ne le sont jamais, voir *Figure 5.65*. Sans surprises, l'outil le moins performant est Exomiser et son modèle *PheniX* suivi du modèle *hiPHIVE*. L'outil AMELIE arrive ensuite avec un nombre de diagnostics manqué relativement honorable. Enfin, sans surprise parmi les outils SeqOne, SeqOne Brut est celui qui manque le moins de diagnostics, car étant la version la plus sensible de l'outil. Certains de ces diagnostics étant retrouvés à plus de la position 200, il faut néanmoins tempérer ce résultat. Les outils de priorisation ayant pour but de faciliter et de réduire le temps d'interprétation, si le diagnostic est au-delà de la position 50, l'intérêt de ces stratégies est alors limité. L'idéal serait de pouvoir classer le diagnostic dans les 5 à 10 premières positions, à ce titre, l'outil SeqOne Scout est le plus performant suivi de très près par l'outil AMELIE.

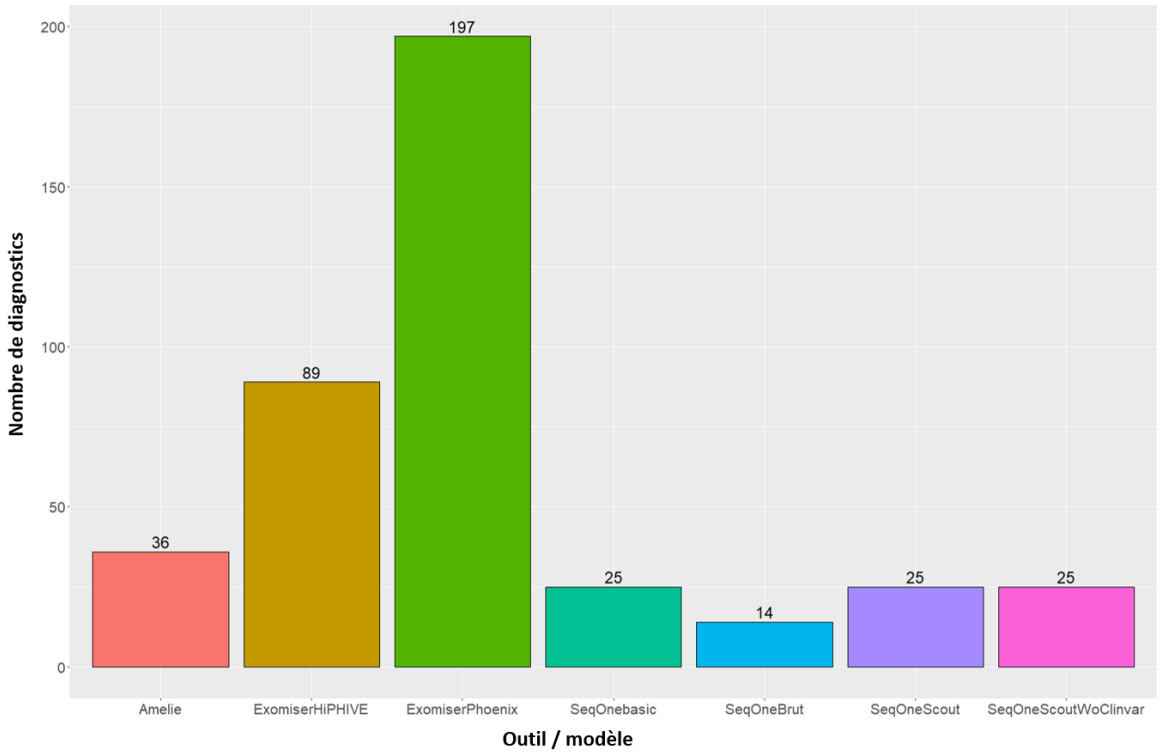


FIGURE 5.65 – Nombre de diagnostics qui ne sont pas retrouvés par les outils et leurs différents modèles.

5.5 Discussion

Les résultats de priorisations obtenus sur notre cohorte avec les différents outils sont supérieurs à ce que nous attendions. En effet, l'outil AMELIE sur les données issues de sa publication retrouve le diagnostic dans les dix premières positions dans 45 % des cas[150]. Au sein de notre cohorte, l'outil AMELIE a retrouvé le bon diagnostic dans les dix premières positions 235 fois, soit dans 71 % des cas. Parmi les 328 diagnostics de la cohorte, 131 sont retrouvés avec succès par le modèle *PheniX* de l'outil Exomiser. Il peut être supposé qu'un tiers des diagnostics de cette étude soient évidents, car détecté par l'ensemble des modèles. De plus, depuis 2018, date de cette étude, chacun de ces outils a reçu des améliorations, de nouvelles versions et modèles pour l'outil sont disponibles pour Exomiser[151], une version 2 de l'outil AMELIE[152], ainsi qu'une amélioration continue de l'algorithme de SeqOne Scout. Les résultats obtenus avec ces outils pourraient donc être encore meilleurs aujourd'hui.

Cette étude n'a été qu'une preuve de concept que la priorisation de variation à partir de données phénotypiques représentait l'étape suivante dans la filtration et l'aide à l'interprétation pour le clinicien. Elle a également permis de montrer que les meilleures solutions académiques et industrielles sont très proches en performances. Malheureusement, l'outil AMELIE n'est utilisable que *via* un site web ou une API. Il existe un outil de priorisation de variation exécutable localement développé par les équipes en charge d'AMELIE, Phrank[127]. Malgré le fait qu'il soit plus performant que le seul autre outil local testé, Exomiser, il reste moins performant qu'AMELIE[150]. C'est pourquoi c'est l'outil Phrank qui avait été choisi pour être intégré au pipeline d'analyse de données d'exome.

Le jeu de données d'exomes phénotypés et interprétés récolté au cours de cette étude est utilisé comme jeu de validation pour le développement de l'outil PhenoGenius par le Dr Kévin Yauy. PhenoGenius est un outil de suggestion de diagnostic génétique basé sur la description clinique du patient ainsi qu'un outil de suggestion d'investigations moléculaires, ciblée ou pan-génomique, obtenu par détection d'enrichissement statistique de symptômes par exploration des termes HPO couramment associés dans la littérature aux termes renseignés pour un patient donné.

La priorisation des variations représente un champ de développement particulièrement dynamique visant à résoudre certains des problèmes posés par la génération de plus en plus importante de données génomiques. La priorisation des variations du rapport final d'interprétation de l'outil SeqOne ainsi que l'interface *user friendly* qui lui est associée sont parmi les principaux arguments de charme qui explique une pénétration aussi rapide de l'outil dans les différents laboratoires français publics et privés. À nous de nous en inspirer par la suite pour le développement de notre outil d'analyse de données d'exome.

Modèles d'apprentissage pour pipeline d'appel de CNV GATK4 sur données WES constitutionnelles

6.1 Motivation

Un des objectifs majeurs de cette thèse est l'exploration des données de séquençage d'exome constitutionnelles pour la détection de variations potentiellement pathogènes. La détection de CNV sur exome est une analyse sous exploitée par de nombreux laboratoires alors qu'elle est techniquement réalisable. Elle présente un intérêt, car capable de détecter des anomalies de taille inférieure au seuil de détection des techniques de référence actuellement implantées dans les laboratoires (caryotype et l'ACPA). La mise en place d'une telle technique représente donc un apport diagnostic potentiel qui n'a pas d'autre coût que son développement et les coûts associés à l'analyse bioinformatique. Ces coûts sont négligeables par rapport à ceux d'une seconde analyse génomique et fait de l'*exome-first*, une stratégie intéressante en terme d'organisation de l'offre de soin en ce qui concerne le diagnostic de maladies génétiques.

Un des deux pipelines qui a servi de modèle au développement de Lygrexome possédait un modèle d'appel de CNV basé sur le l'outil XHMM[153]. La fonctionnalité de détection de CNV devait donc être portée sur le nouveau pipeline. Le choix d'outil s'est donc tout naturellement porté sur XHMM celui-ci produisant des résultats de bonne facture. Lorsque le *Broad Institute* a publié la version 4 de son outil GATK, embarquant un module d'appel de CNV constitutifs (pour faire la distinction avec un module de détection de CNV somatique), la décision a été prise de tester l'outil. En effet, de nombreux outils publiés par le *Broad Institute* font aujourd'hui consensus. Prendre de l'avance sur un de leurs outils nous permet de nous positionner en meneur si celui-ci s'il venait à être considéré comme un outil de référence.

À sa sortie, l'outil était encore en bêta, instable et peu documenté. Il a fallu de nombreux mois d'essais et d'optimisation pour enfin pouvoir produire les premiers modèles d'apprentissage optimisés pour la détection de CNV germinaux à partir de données d'exome.

Ces travaux ont fait l'objet d'une présentation à l'ACLF 2021 et ont été soumis aux Assises de Génétique 2022. Ils font également l'objet d'une prépublication en tant que co-premier auteur sur la validation de la méthode d'appel de CNV et de son apport en diagnostic de routine face aux technologies de référence.

6.2 Article en prépublication

Supplementary Tables : https://www.medrxiv.org/content/medrxiv/early/2021/10/16/2021.10.14.21264732/DC2/embed/media-2.xlsx 

Exome sequencing as a first-tier test for copy number variant detection : retrospective evaluation and prospective screening in 2418 cases.

Quentin Testard[1,2,3]* Xavier Vanhoye[1]*, Kevin Yauy[2,13], Marie-Emmanuelle Naud[1], Gaelle Vieville[3], Francis Rousseau[1], Benjamin Dauriat[4], Valentine Marquet[4], Sylvie Bourthoumieu[4], David Genevieve[5,6], Vincent Gatinois[6], Constance Wells[6], Marjolaine Willems[6], Christine Coubes[6], Lucile Pinson[6], Rodolphe Dard[7], Aude Tessier[7], Bérénice Hervé[7], François Vialard[7], Ines Harzallah[8], Renaud Touraine[8], Benjamin Cogné[9], Wallid Deb[9], Thomas Besnard[9], OIivier Pichon[9], Béatrice Laudier[10], Laurent Mesnard[11], Alice Doreille[11], Tiffany Busa[12], Chantal Missirian[12], Véronique Satre[3,14], Charles Coutton[3,14], Tristan Celse[3], Radu Harbuz[3], Laure Raymond[1], Jean-François Taly[1#] , Julien Thevenon[2,3#]

[1] Service de Génétique, Eurofins Biomnis, Lyon, France

[2] CNRS UMR 5309, INSERM, U1209, Université Grenoble Alpes, Institute for Advanced Biosciences, Grenoble, France

[3] Service de Génétique et Procréation, CHU Grenoble Alpes, Grenoble, France

[4] Service de Cytogénétique, Génétique Médicale et Biologie de la Reproduction, CHU de Limoges, Limoges, France

[5] Université Montpellier, Unité INSERM U1183, Montpellier, France

[6] Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, CHU Montpellier, Montpellier, France

[7] Département de Génétique, CHI Poissy-Saint-Germain en Laye, Poissy, France

[8] Service de génétique clinique, chromosomique et moléculaire, CHU de Saint-Étienne, Saint-Étienne, France

[9] Service de Génétique Médicale, CHU de Nantes, Nantes, France

[10] Laboratoire d'Immunologie et Neurogénétique Expérimentales et Moléculaires INEM UMR7355, CHR d'Orléans, Orléans, France

[11] Sorbonne Université, Urgences Néphrologiques et Transplantation Rénale, AP HP, Hôpital Tenon, Paris, France

[12] Département de génétique médicale, AP HM, Hôpital de la Timone Enfant, Marseille, France

[13] SeqOne Genomics, Montpellier, France

[14]Équipe "Génétique, Epigénétique et Thérapies de l'Infertilité", IAB, INSERM 1209, CNRS UMR 5309, Université Grenoble Alpes, France

* These authors contributed equally

Co-corresponding authors Julien THEVENON, jthevenon@chu-grenoble.fr and Jean-François TALY, Jean-Francois.TALY@biomnis.com

ABSTRACT

Purpose: Despite exome (ES) or genome sequencing (GS) availability, chromosomal microarray (CMA) remains the first-line diagnostic tests in most rare disorders diagnostic work-up, looking for Copy-number variations (CNV), with a diagnostic yield of 10-20%. The question of the equivalence of CMA and ES in CNV calling is an organisational and economic question, especially when ordering a GS after a negative CMA and/or ES.

Methods: This work measures the equivalence between CMA and GATK4 exome sequencing depth of coverage method in detecting coding CNV on a retrospective cohort of 615 unrelated individuals. A prospective detection of ES CNV on a cohort of 1803 unrelated individuals was performed.

Results: On the retrospective validation cohort every CNV was accurately detected (64/64 events). In the prospective cohort, 32 diagnostics were performed among the 1803 individuals with CNVs ranging from 704bp to aneuploidy. An incidental finding was reported. The overall increase in diagnostic yield was of 1.7%, varying from 1.2% in individuals with multiple congenital anomalies to 1.9% in individuals with chronic kidney failure.

Conclusions: Combining SNV and CNV detection increases the suitability of exome sequencing as a first-tier diagnostic test for suspected rare mendelian disorders. Before considering the prescription of a GS after a negative ES, a careful reanalysis with updated CNV calling and SNV annotation should be considered.

INTRODUCTION

Copy Number Variants (CNV) represent the imbalance of the genomic material compared to the reference genome, resulting in an increase or decrease in genomic material. CNVs vary in size, although they are defined as variants with a minimum size of 1 kb[1]. Adoption of Chromosomal Microarray Analysis[2] (CMA) techniques have proven invaluable in discovering pathogenic CNVs in a wide variety of diseases, especially for diagnosing multiple congenital anomalies (MCA). In routine practice, a diagnostic yield of ~15% is reached for patients with intellectual disability disorder or MCA, and can be attributed to large CNVs (> 100 kb)[3]. Despite the rapid adoption of next generation sequencing, standard chromosomal analysis and CMA remain the first-tier tests for most rare disorders diagnostic work-up [4,5].

In practice, the average resolution of CMA technologies implemented in laboratories is about 50 kb[3]. In theory, Genome Sequencing (GS) CNV calling is the golden path for CNV calling. However, exome sequencing is notably widespread and more affordable, thus an accurate CNV calling should be advised on existing data before ordering an additional diagnostic test.

Although ES has intrinsic limitations, common problems are shared by GS and ES in calling CNV such as extreme GC contents or low complexity regions. In GS, algorithms strategies of type Depth of Coverage (DoC), Split Read , Discordant Pairs and Assembly[6] can be used, whereas ES CNV calling tools can only use DoC. ES specifically encounters additional limitations regarding the targeted enrichment (known as capture bias), leading to non-uniform read depths impacting the reproducibility and robustness of CNV calling tools[7]. The ratio of read count between a test and a reference is usually preferred to a single-sample analysis, which could lead to many false positive[8].

Numerous tools such as XHMM[9], CODEX[10], CANOES[11], CoNIFER[12] or ExomeDepth[8] were developed when germline ES started to be democratized, none of them has really imposed itself as the reference tool. In January 2018 the Broad Institute released the fourth version of its GATK[13] tool (GATK4) including several tools forming a CNV detection module[14]. This module is based on the principle of constructing a learning model from a cohort of patients DoC data that can be further reused.

This study presents an analytical validation framework for a clinical routine of GATK4 gCNV calling on ES data supported by a retrospective benchmark on 615 unrelated index cases with previously acquired CNVs. Results include the prospective screening for CNV in 1803 unrelated individuals with no previous CMA.

PATIENTS AND METHODS

Individuals gathering

Patients were ascertained in the diagnostic routine of the Eurofins Biomnis Laboratory (Lyon, France). The referring clinical centers included Nantes, Lyon, Montpellier, Paris (Tenon), Grenoble, Besançon, Saint Etienne, Limoges, Poissy, Marseille, Orléans and international laboratories (details provided in *Supplementary Material Table 1*). Patients provided written consent. A total of 2418 individuals were included in the work. Overall, 615 had CMA, MLPA or NGS-based data available as tabulated files and were used as the analytical retrospective validation cohort. Files formats were normalized during this study. For the remaining 1803 individuals, no question was asked regarding previously available CMA results, and are further referred as the prospective screening cohort.

ES capture sequencing

For all the 2418 probands, ES libraries were generated using standard procedures (*Supplementary Materials*) for 3 different capture protocols for sequencing Roche Medexome kit (n= 447), Twist Bioscience Human Comprehensive Exome kit + RefSeq + UTR spike (n= 988), Twist Bioscience Human Comprehensive Exome kit + RefSeq spike (n= 983). Libraries were sequenced on Illumina NextSeq 500 sequencers in paired-end mode (2 x 76bp).

ES analysis for CNV calling

Exome Sequencing data was mapped against the hg38 genome, following the Broad Institute GATK best practice guidelines [15]. CNV calls were performed with the GATK4 CNV calling module. Fine-tuning of ES learning model creation was performed according to parameters provided by the Broad Institute teams (shown in Supplementary Material). It was therefore decided to divide the calling target into 4 bins with the GATK IntervalListTools in order to run four instances of the GermlineCNVCaller in parallel on our computing infrastructure. The full methodology of model building is available in the *Supplementary Material*. Each VCFs were then annotated with AnnotSV[16] version 2.5.1 to add crucial metadata for interpretation by the clinician. The output files by AnnotSV were processed by an in-house Python script to keep only the annotations of interest, but also to add the occurrence cohort counts of each CNV.

The diagnostic target represented 41 935 379 bp, defined by the merging of UCSC RefSeq and RefSeq Curated[17] intervals, with 5'-3' padding of 20bp. This diagnostic target included 21450 genes with 198188 exonic intervals.

For all samples, CNV were analyzed at the same time as SNV analysis. SNV interpretation was done following ACMG recommendations[18] . CNVs were prioritized based on their frequency in our cohort, and in DGV[19] ; the inclusion of an OMIM Morbid gene ; the quality metrics of the CNV and the inheritance of the CNV. Recurrent CNV were specifically analyzed according to gene content and recurrent CNV list of the French AChroPuce consortium (https://acpa-achropuce.com/).

Analytical retrospective validation cohort

Biological results from 615 individuals with previously identified clinically relevant CNV were gathered and compared to CNV detection by ES. To ensure comparable results across detection techniques, only coding CNV were compared. Overall, 72 CNVs were considered as clinically relevant. 64 CNV were used for comparison, either classified as VUS, likely pathogenic or pathogenic. Frequent polymorphisms and technical artefacts may be confusing and were excluded from the analysis. The 64 CNVs included 30 loss, 31 gain (including a XXY phenotype) and 3 VUS with a copy number of 2 (chromosome X), with sizes ranging from an intragenic single exon deletion to large anomalies including aneuploidy (summarized in *Supplementary material table 2*).

Prospective screening cohort

Prospective cohort included 2418 individuals. CNVs were called only on ES data. Each CNV larger than 1 Mb was individually interpreted. Regarding smaller CNV, filtering was performed (i) on the quality score QA > 20 and QS > 20; (ii) overlapping or impacting a gene referenced in the OMIM database with suspected or demonstrated dosage sensitivity (pLI >0.9); (iii) autosomal dominant inheritance for heterozygous CNV inheritance. Every homozygous and hemizygous CNV were considered. Each filtered CNV was interpreted and classified. Downstream CNV validations were performed by the referring centers using standard procedures.

RESULTS

Statistical description of CNV calls

Across capture kits, the distribution of the CNVs larger than 50kb number was varying from an average of 5-10 events. The median number of CNVs smaller than 50kb varied from 31-36 across capture kits (*Figure 1A*). The median number of CNV encompassing an OMIM morbid gene was comparable across capture kits. For morbid CNVs, their distribution is comparable between the 3 models, with a median of 4 (< 50kb) or 1 (> 50kb) (*Figure 1B*).

Finally, detected CNVs were intersected with the DGV database. Intervals were considered comparable when at least 80% of reciprocal overlap was observed. A median of 75,56%, 77,78% and 75,76% (Roche, Twist, Twist+UTR) of detected CNVs were referenced in the DGV database (*Figure 1C*). A median of one CNV overlapping an OMIM morbid gene and absent from the DGV database was observed (*Figure 1D*).

Defining the model size for ES-CNV calling

From the Twist model data set (n = 1154), several models were built of different sizes and random data (50, 100, 150, 200, 300, 600 samples), with three subsamples for each size condition. Then, from the Twist data set, 154 samples were randomly selected and were used as a fixed cohort. Iteratively, CNVs were called on those samples against the previously constructed models (*Figure 2*).

The lower the number of samples used to build the model, the higher the average number of CNVs per patient and vice versa (*Figure 2*). In addition, the smaller the models, the more variable are the distributions between the subsamples. Among the 1154 samples, and independently from the calling model, 23 individuals were continuously leading to high numbers of CNV calls (> 200).

Isolating outliers of the ES-CNV calling pipeline

Among the whole cohort (2418 samples), 2275 individuals had fewer than 200 events. 143 samples were leading to an excess of CNV calls across capture kits and calling models. The distribution of CNV counts is represented by *Supplementary material Figure 4*. These 143 outlier samples were excluded from the interpretation and further analysis. Among the 143 samples, 66 were concentrated in seven sequencing runs with technical issues; 67 samples were DNA received from collaborators (60 DNA extracted from blood and 7 DNA extracted from tissues); 10 were blood samples received by the laboratory.

Defining recurrent uncallable regions

ES CNV calling was unable to quantify the copy number ratio for a significant portion of the diagnostic target, 10 and 11% for Twist capture kits and 8.76% for the Roche kit. Focusing on the 3593 genes of the OMIM morbidmap identified 32 genes totally uncallable for coding CNV (*AHDC1, AMER1, BBS12, CHAMP1, CRYAA, CSF2RA, DOLK, FLRT3, FZD2, GP1BA, HPS6, IRF2BPL, IRS4, KCNA1, KCNA4, KCNA5, MAGEL2, MKRN3, MYORG, PIGW, POMGNT2, RAG2, SAMD9, SAMD9L, SLC18A3, SLITRK1, SLITRK6, THBD, TRIM32, UBQLN2, ZNF469, MARCH2*).

Across capture kits and each calling model, an average of 410 genes are partially represented and CNV calling might be impacted (*Supplementary material Figure 5*).

Analytical retrospective validation cohort

Overall, 615 samples were available. Twenty-five (4.0%) samples were excluded from the analysis because they were classified as outliers. Among the 72 selected CNVs, 8 were excluded because they were localised in intergenic regions or in a previously defined uncallable region (*Supplementary material Figure 6*). For the 590 remaining samples, the 64 CNVs were accurately detected and genotyped (*Supplementary material table 2*). No additional large and rare CNV was reported.

Prospective screening cohort

Among the 1803 individuals, 32 CNV and 2 aneuploidies were diagnosed. Among the 615 individuals with MCA, 20 diagnoses were performed. Among the 631 individuals with chronic kidney failure, 12 diagnoses were performed (*Supplementary material table 3*). Regarding the 22 pathogenic or likely pathogenic CNV larger than 50 kB, ES was the first genetic investigation.

Patient 4 was presenting with chronic kidney failure and kidney cysts in adulthood, revealed an intragenic deletion of *COL4A3* at heterozygous state. BAM viewing emphasized breakpoints in exon 9 (*Figure 3*). Breakpoints were verified using Sanger sequencing, allowing characterization of the variation : NC_000002.12:g.227248049_227251231del ; NM_000091.4:c.469-394_609+29del. Small pathogenic or likely pathogenic CNV in genes of recessive inheritance, associated with a pathogenic or likely pathogenic SNV on the other allele for 2 patients were detected. Patient 5, presenting with dilated cardiomyopathy and facial

dysmorphism, carried NM_006663.3(*PPP1R13L*):c.1871_1872del ; p.(Arg624Profs*119), maternally inherited, and intragenic duplication of *PPP1R13L* (duplication of exon 2 to exon 7, of 13). Patient 6 presented with growth delay, facial dysmorphia, delayed psychomotor development, hyperextensibility, cortical atrophy, thin corpus callosum and hypomyelination. ES detected a deletion of the whole *PYCR2* gene, maternally inherited and a hemizygous point variant paternally inherited : NM013328.3(*PYCR2*):c.751C>T ; p.(Arg251Cys).

DISCUSSION

This study assessed the analytical validity of gCNV calling in an ES routine based on a 615 individuals retrospective validation cohort and demonstrated the positive impact on ES diagnostic yield through the screening of 1803 individuals. In this first-tier ES routine, CNV calling identified 2 aneuploidy, 22 large CNV and 10 small CNV.

The 64 CNV gathered from the retrospective validation cohort were accurately detected and genotyped by the ES procedure. Previous study had demonstrated the equivalence of ES against CMA [20]. Another published cohort included 147 samples with 102 CNV, and they performed comparison between aCGH CNV detection and CANOES CNV detection[11]. The recall was 87.2% (89/102). They suggested that the missed CNV by ES might be secondary to the capture design or size of the event with only 1 or 2 targets [11]. Our retrospective validation dataset included very small events such as hemizygous deletion of one exon in *DMD* gene or gain of one exon in *IL1RAPL1* (respectively for individuals 2 and 1, *Supplementary material table 2)*, which were accurately identified. These observations suggest that this work may add an important validation of the procedure for a clinical ES routine.

In the validation cohort, the exhaustive detection of CNV may be secondary to the preliminary definition of predictive limitations of the procedure. These limitations included the definition of uncallable regions, and the prediction of aberrant and noisy samples. This study did not aim at deciphering the underlying causes for these limitations.

Analysis of outliers of CNV-ES detection reveal that our workflow is robust and suitable for routine diagnosis, with 4% of failed samples (143/1804). Most of those

outlier samples could be explained by pre-analytical or analytical issues. Only 10 blood samples (among 1698) were classified as “outliers”. This failure rate of 0.6% is acceptable and comparable or below those of CMA in our practice. To further investigate those outliers, we analyzed CNV calls for outliers of the validation cohort : all medically relevant CNV were properly called, with high quality metrics. Those data suggest that CNV calling is possible for samples initially classified as outliers, but require intensive filtration and interpretation, to distinguish authentic CNV and background noise.

Tools to model coverage distributions across exons are widespread in the clinical bioinformatics community. On the other hand, the possibility of being able to build a learning model, and then to reuse it later on, seems genuinely new. The performances of the CNV calling models are certainly correlated to the number of data items that were used to build them. However, two models built with the same number of data and different sequencing depths will have different results. It is therefore more likely that the efficiency of the model is correlated to the cumulative sequencing depth of the data that compose it as well as their homogeneity across individuals. With the current sequencing data generation processes in our lab, if we ever had to reconstruct a model, the number of samples required would most likely be around 300.

GS has been proven to be more efficient for diagnosis than ES, both for SNV and CNV[21–24]. Indeed, in addition to being able to detect exonic, intronic and intergenic SNVs and indels, GS can more accurately detect exonic, intronic and intergenic structural variants. Unlike ES, the production of GS data does not require prior amplification or capture steps. This limits the variability of depth between exons, and virtually extinct the uncallable regions. Nevertheless, even if the set of

uncaptured zones represents about 4 mb or 10% compared to the defined medical target in this study. However, only 0.9% of morbid genes have their entire sequence in the blind areas of our pipeline. Copy number variations in these genes will not be detected. However, large CNVs encompassing such genes might be detected.

Careful examination of the data generated by the pipeline allowed identification of causing-disease CNV for 35 patients. Among these 35 positive results, 8 individuals had a negative CMA before ES prescription. In the neurodevelopmental disorder cohort, the added diagnosis range is 1,2% (20/1787). This percentage is relatively low compared with the yield of >10% reported for genomic microarrays. This can easily be explained by the fact that the vast majority of patients with a neurodevelopmental disorder were previously screened negative for CNV microarray analysis, resulting in a depletion of pathogenic CNVs in this patient group. Clinically relevant CNVs were observed only in patients who had previously been screened on a (low-resolution) microarray platform or in patients who did not receive microarray-based CNV profiling. This percentage is consistent with previous studies analyzing exome based CNV calling within ID cohorts (1.3%[25] ; 1.6%[26]). Among individuals with chronic kidney failure, the diagnosis yield reaches 1,9 % (12/631). Only few data highlights the implication of CNV in renal disease. Previous studies demonstrated an added diagnosis range of 3.6% (2 of 56 patients)[25] with CNV detection.

Of note, using an exome-wide CNV detection pipeline raises new incidental findings. We identified a deletion of 6 exons of LDLR (responsible for familial hypercholesterolemia [OMIM:# 143890)] for a patient referred for neurodevelopmental disorders.

The commitment to make ES a frontline analysis is not new[27]. On one hand, it has already been shown that ES can be much more efficient than traditional methods in terms of diagnostic rates as well as cost-effectiveness[28]. On the other hand, ES has already shown its superiority against some routine genetic analyses such as gene panels and single gene testing[24,29]. The ability to bundle the detection of exonic SNVs, Indels and CNVs make the ES strategy an extremely competitive and efficient first-tier analysis. In this cohort, two diagnoses were performed by combining CNV and SNV calling (0.11%, 2/1803). This observation is consistent with data from a large study of 12000 individuals combining CMA and ES for the identification of 17 diagnoses (0.11%)[30]. Despite limitations, thousands of exomes will be produced in the coming years for the diagnosis of rare disorders. A careful and updated analysis will enhance the diagnostic yield of the tests and will participate in reducing the diagnostic odyssey of patients with undiagnosed disorders.

This study highlights the technical validity and the clinical utility of exome-based CNV screening. Incorporation of CNV analysis in exome sequencing data-analysis pipelines increases the diagnostic yield of exome sequencing by up to 1,9%. Of importance, this increase in diagnostic yield is obtained without any additional direct laboratory costs. Combining SNV and CNV detection increases the suitability of exome sequencing as a first-tier diagnostic test for many, if not most, suspected genetic disorders. Before considering the prescription of a GS after a negatif ES, a careful reanalysis with updated CNV calling and SNV annotation should be considered.

REFERENCES

1. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. *Trends Genet*. 2013;29(10):575-584.

2. Boone PM, Bacino CA, Shaw CA, et al. Detection of clinically relevant exonic copy-number changes by array CGH. *Hum Mutat*. 12/2010;31(12):1326-1342.

3. Miller DT, Adam MP, Aradhya S, et al. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *The American Journal of Human Genetics*. 2010;86(5):749-764. doi:10.1016/j.ajhg.2010.04.006

4. Manning M, Hudgins L. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genet Med*. 11/2010;12(11):742-745.

5. Sagoo GS, Mohammed S, Barton G, et al. Cost Effectiveness of Using Array-CGH for Diagnosing Learning Disability. *Appl Health Econ Health Policy*. 8/2015;13(4):421-432.

6. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. *Front Genet*. 2015;06. doi:10.3389/fgene.2015.00138

7. Hong CS, Singh LN, Mullikin JC, Biesecker LG. Assessing the reproducibility of exome copy number variations predictions. *Genome Med*. 12/2016;8(1):82.

8. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. :8.

9. Fromer M. Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. :11.

10. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*. 2015;43(6):e39-e39.

11. Quenez O, Cassinari K, Coutant S, et al. Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. *Eur J Hum Genet*. 2021;29(1):99-109.

12. Krumm N, Sudmant PH, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res*. 2012;22(8):1525.

13. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.

14. Babadi M, Lee SK, Smirnov AN. GATK gCNV: accurate germline copy-number variant discovery from sequencing read-depth data. :4.

15. Auwera GA, Carneiro MO, Hartl C, et al. From FastQ Data to High□Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics*. 10/2013;43(1). doi:10.1002/0471250953.bi1110s43

16. Geoffroy V, Herenger Y, Kress A, et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*. 2018;34(20):3572-3574.

17. Pruitt KD. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2004;33(Database issue):D501-D504.

18. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.

19. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986.

20. Rajagopalan R, Murrell JR, Luo M, Conlin LK. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med*. 2020;12(1):14.

21. Gross AM, Ajay SS, Rajan V, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med*. 5/2019;21(5):1121-1130.

22. Ellingford JM, Campbell C, Barton S, et al. Validation of copy number variation analysis for next-generation sequencing diagnostics. *Eur J Hum Genet*. 6/2017;25(6):719-724.

23. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*. 2015;112(17):5473-5478.

24. Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med*. 4/2018;20(4):435-443.

25. Pfundt R, del Rosario M, Vissers LEL, et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine*. 2017;19(6):667-675. doi:10.1038/gim.2016.163

26. Marchuk DS, Crooks K, Strande N, et al. Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS One*. 2018;13(12):e0209185.

27. Melbourne Genomics Health Alliance, Stark Z, Tan TY, et al. A prospective

evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet Med*. 11/2016;18(11):1090-1096.

28. Yeung A, Tan NB, Tan TY, et al. A cost-effectiveness analysis of genomic sequencing in a prospective versus historical cohort of complex pediatric patients. *Genet Med*. 12/2020;22(12):1986-1993.

29. Sun Y, Ruivenkamp CAL, Hoffer MJV, et al. Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Hum Mutat*. 06/2015;36(6):648-655.

30. Yuan B, Wang L, Liu P, et al. CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels. *Genet Med*. 2020;22(10):1633-1641.

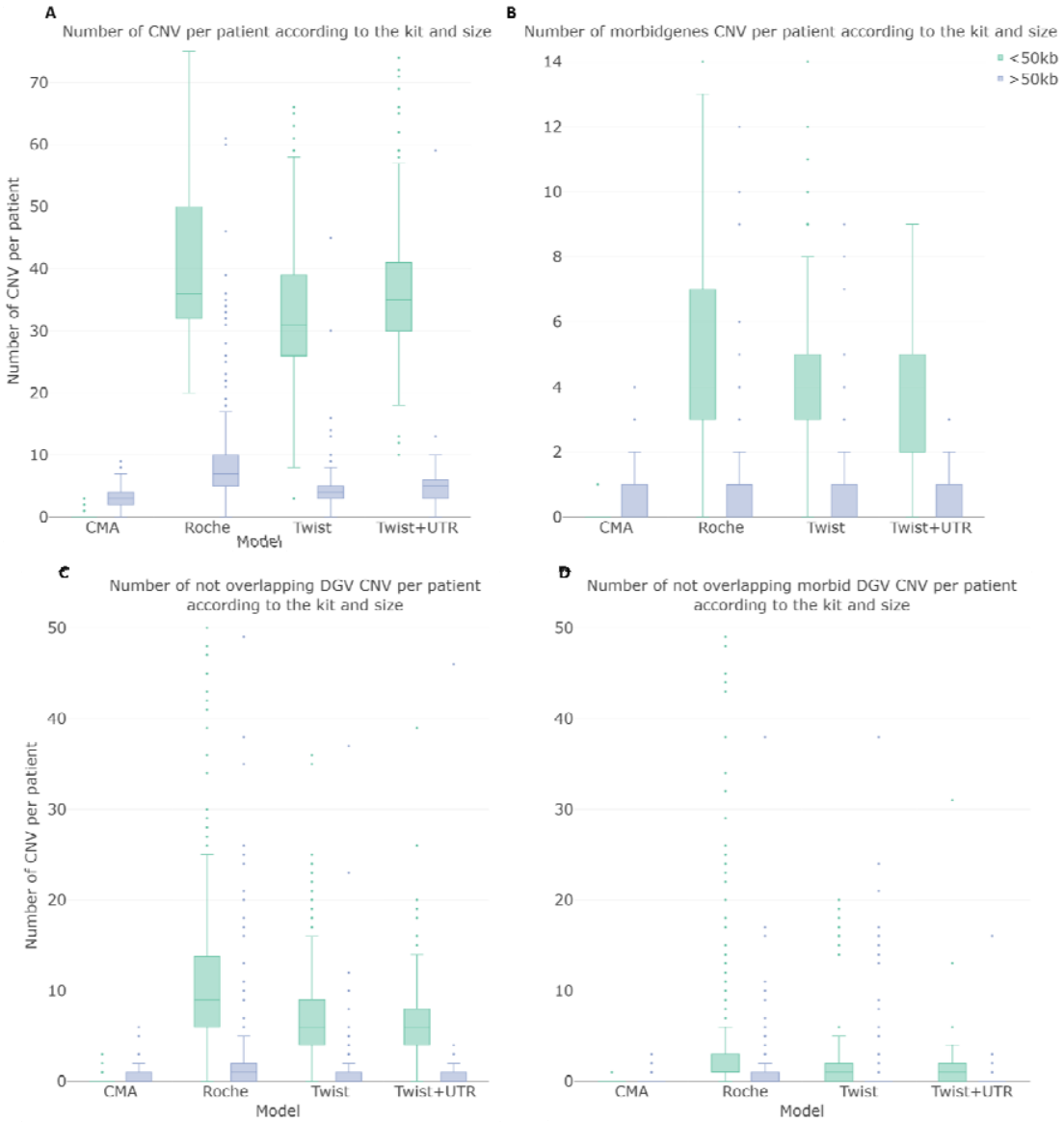


Figure 1. Distribution of the number of called CNV. *(A) The total number of CNV, (B) the number of CNV containing at least one morbid gene, (C) the number of CNV not present in DGV, (D) the number of CNV containing at least one morbid gene not present in DGV, per patient according to the CNV size and the model used compared to CMA data . CMA (n=300), Roche (n=511), Twist (n=1154), Twist UTR (n=383).*

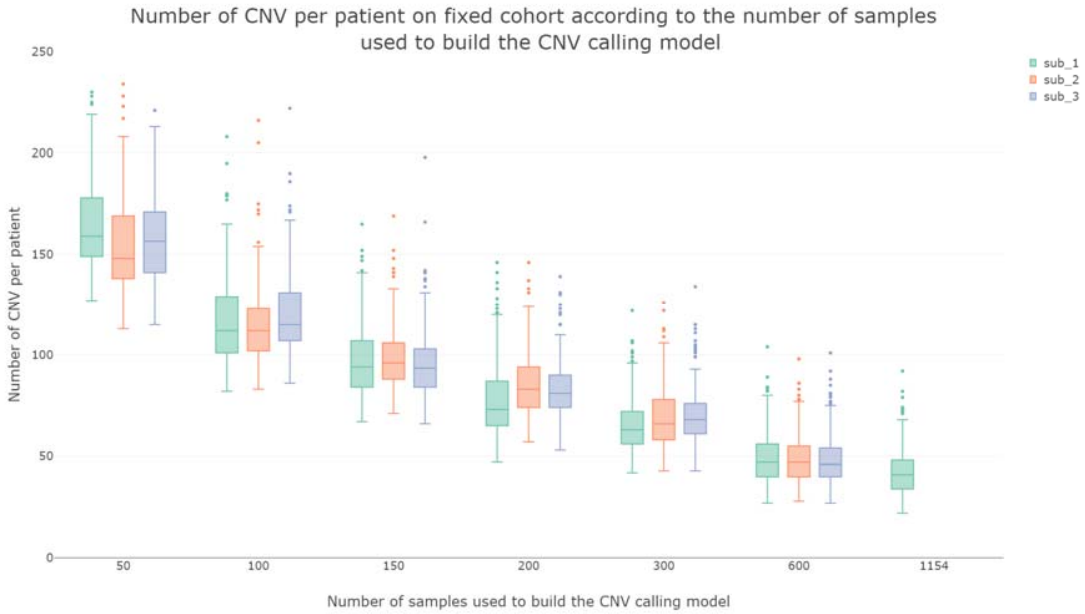


Figure 2. Distribution of the number of CNVs per patient in the cohort of 154 fixed patients according to model size and subsampling. 3 sub-samples (sub 1-3) of built CNV calling models consisting of 50 to 600 samples sequenced with the Twist Human Core Exome kit. CNV reused to call CNV 154 randomly selected samples (the same samples for every model) compared to the results of the Twist model consisting of 1154 samples on these 154 samples.

Figure 3. BAM visualisation of Intragenic heterozygous deletion of COL4A3 exon 9. *Reads colored, oriented and sorted by insert size with IGV software.*

Supplementary Materials

Exome sequencing as a first-tier test for copy number variant detection : retrospective evaluation and prospective screening in 2418 cases.

Quentin Testard[1,2,3], Xavier Vanhoye[1], Kevin Yauy [2,13], Marie-Emmanuelle Naud[1], Gaelle Vieville[3], Benjamin Dauriat[4], Valentine Marquet[4], Sylvie Bourthoumieu[4], David Genevieve[5,6], Vincent Gatinois[6], Constance Wells[6], Marjolaine Willems[6], Christine Coubes[6], Lucile Pinson[6], Rodolphe Dard[7], Aude Tessier[7], Bérénice Hervé[7], François Vialard[7], Ines Harzallah[8], Renaud Touraine[8], Benjamin Cogné[9], OIivier Pichon[9], Wallid Deb[9], Thomas Besnard[9], Béatrice Laudier[10], Laurent Mesnard[11], Alice Doreille[11], Tiffany Busa[12], Chantal Missirian[12], Véronique Satre[3], Charles Coutton[3], Tristan Celse[3], Radu Harbuz[3], Laure Raymond[1], Jean-François Taly[1] , Julien Thevenon[2,3].

[1] Service de Génétique, Eurofins Biomnis, Lyon, France

[2] CNRS UMR 5309, INSERM, U1209, Université Grenoble Alpes, Institute for Advanced Biosciences, Grenoble, France

[3] Service de Génétique et Procréation, CHU Grenoble Alpes, Grenoble, France

[4] Service de Cytogénétique, Génétique Médicale et Biologie de la Reproduction, CHU de Limoges, Limoges, France

[5] Université Montpellier, Unité INSERM U1183, Montpellier, France

[6] Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, CHU Montpellier, Montpellier, France

[7] Département de Génétique, CHI Poissy-Saint-Germain en Laye, Poissy, France

[8] Service de génétique clinique, chromosomique et moléculaire, CHU de Saint-Étienne, Saint-Étienne, France

[9] Service de Génétique Médicale, CHU de Nantes, Nantes, France

[10] Laboratoire d'Immunologie et Neurogénétique Expérimentales et Moléculaires INEM UMR7355, CHR d'Orléans, Orléans, France

[11] Sorbonne Université, Urgences Néphrologiques et Transplantation Rénale, AP HP, Hôpital Tenon, Paris, France

[12] Département de génétique médicale, AP HM, Hôpital de la Timone Enfant, Marseille, France

[13] SeqOne Genomics, Montpellier, France

Correspondence to :
Pr Julien THEVENON
jthevenon@chu-grenoble.fr

Prescribing center	Number of patient reference data provided
Nantes	349
Montpellier	79
Saint-Étienne	59
Limoges	55
Poissy	50
Marseille	11
Besançon	2
Orléans	2
Tenon	1
International Laboratories	7
Total	615

Supplementary Table 1. Origin of reference CMA data from the validation cohort.

DNA preparation for ES sequencing

The genomic DNA is extracted from blood samples using the QIAsymphony® DSP DNA Mini Kit on a QIAsymphony instrument following the recommendation of QiaGen.

QC DNA and normalisation

The concentration of the genomic DNA is established using Optical density at 260 nm with a Spectrophotometer (DropSense 16).

After normalisation, 50 ng of genomic DNA are engaged in a library preparation step. The generation of the captured pooled libraries are prepared using the Human Comprehensive Exome kit delivered by Twist Bioscience following their recommendation.

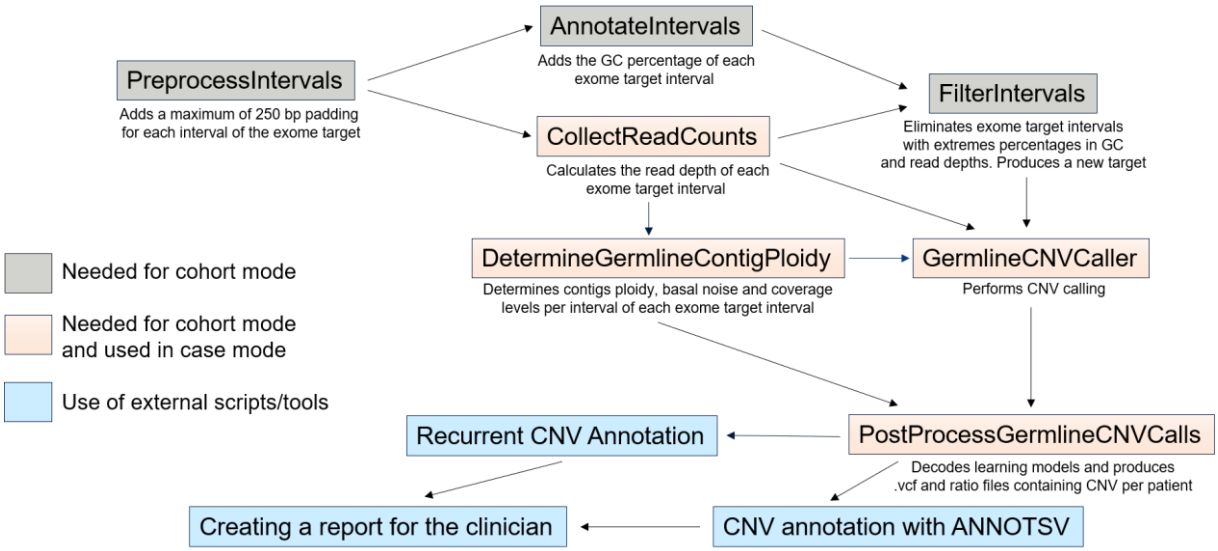
ES capture libraries preparation

It consists of a first step that combines an enzymatic fragmentation, repair ends and dA tailing. The dA-tailed DNA fragments are then ligated to universal adapters and purify on magnetic beads. These purified fragments are amplified by PCR (7 cycles) using Universal Dual Index Primers (Truseq compatible) and the KAPA Hifi HotStart enzyme. After cycling the amplicons are purified on magnetic beads. Each library sample is quantified using the Thermo Fisher Scientific Qubit dsDNA Broad Range Quantitation Assay on Qubit (Thermo fisher Scientific) and the size range is validated on a migration on a Fragment Analyzer (Agilent) using the NGS Fragment analysis kit.

CNV model building and reusing

ES pre-CNV calling target processing

The GATK4 tools from the gCNV module used in our ES CNV calling models creation are shown in *Supp Fig 1*. Model creation was done by launching scripts and command lines, some steps were parallelized with the GNU Parallel[32] tool. Some steps must have been taken beforehand to obtain the final CNV calling model target from our medical reference target, RefSeq[17], in Picard[33] INTERVAL_LIST format. The PreprocessIntervals tool was applied with a padding parameter, expanding to a maximum of 250 base pair (interval expansion stopped when it was likely to overlap with another interval) each interval of RefSeq. Then, a depth of coverage counts of the expanded target intervals of all patient BAM files included in the cohort was performed with the CollectReadCounts tool. Meanwhile, the percentage in GC of the extended target was calculated with the AnnotateIntervals tool. The annotated and extended targets with the cohort count dataset were provided to the FilterInterval tool, which filtered out intervals with extreme GC content, but also over or under-captured intervals in the count dataset. The final filtered target was used to recalculate all the patient coverage data of the cohort.



Supplementary Figure 1. gCNV calling workflow. Gray rectangles represent the preliminary steps in obtaining a suitable calling target for the pipeline, with areas which are not suitable for calling masked. Those steps are performed only once during the model creation. Orange rectangles represent the steps that create learning models in *COHORT* mode and reuse them in *CASE* mode. Yellow rectangles represent scripts or external tools to GATK4 used to add metadata to the produced *VCF* files. Only tools in the orange and blue rectangles are present in our production pipeline.

ES learning model creation and CNV calling

Patient count data and the filtered calling target were given in input to the DetermineGermline- ContigPloidy tool in *COHORT* mode. A DetermineGermlineContigPloidy learning model and call files were created from the cohort. Call files summarized the ploidy of every chromosome for each of the cohort patient. *COHORT* mode produced models and calls whereas *CASE* mode only produced calls by reusing a model created by a previous *COHORT* mode execution.

Patient count data, filtered calling target and DetermineGermlineContigPloidy model were provided as input to the GermlineCNVCaller tool. As above, it created a

model from the cohort patient dataset and call files. The PostProcessGermlineCNVcalls tool converted the models and calls produced by the DetermineGermlineContigPloidy and GermlineCNVCaller into ratio files (with the estimated number of copies per interval for each patient), interval VCF files (reporting all intervals or bin from the filtered target), but also segment VCF files (merging contiguous intervals supposed to belong to the same original CNV) which were the files used afterwards.

Learning model built and used for CNV calling

Currently, 3 models are used in diagnostic routine, the Roche model built with data from the Roche protocol (n = 511), the Twist model built with data from the Twist Bioscience Human Core Exome kit + RefSeq spike protocol (n = 1154) and the Twist UTR model built with data from the Twist Bioscience Human Core Exome kit + RefSeq spike + UTR spike protocol (n = 383). Two additional models were used as a reference to study the performance of the models currently in production, but were not used for diagnosis. The Twist Legacy model, populated with the entire Twist protocol data set at the time of its creation (n = 432) and the Twist All model, containing the entire Twist and Twist UTR data set plus demultiplexing duplicates (n = 1730).

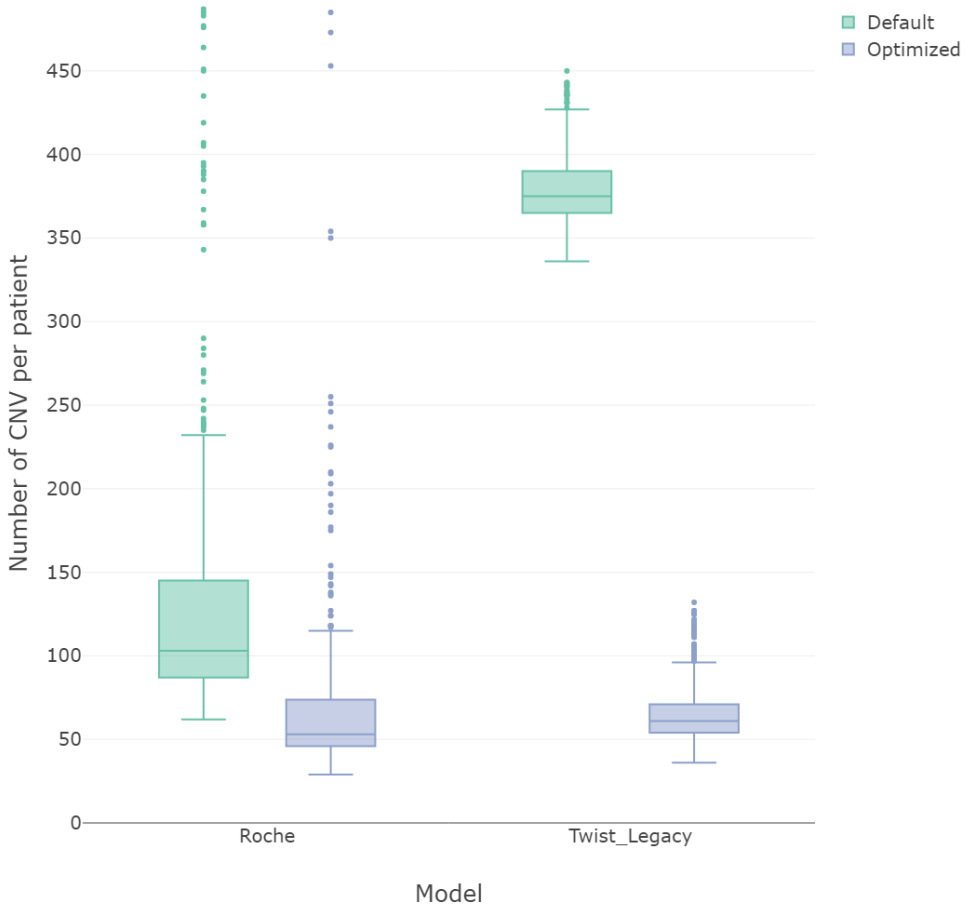
CNV calling ES learning model re-use

The reuse of the different machine learning models from the DetermineGermlineContigPloidy and GermlineCNVCaller tools was done within the same Nextflow pipeline that we used to produce the BAM files. Reusing models rather than building them was a much less demanding processor in terms of resources and computing time. The sequence of steps in the pipeline was the same as when the model was created, but these steps are performed within an automated production

framework. The only difference between the two execution modes is that the CNV count of the VCF files produced in a model reuse is added to the CNV count of the cohort used to build the model, in order to have a more accurate recurrent CNV count.

CNV Fine tuning model creation parameters before CNV calling results

During our various preliminary tests with the gCNV calling pipeline of GATK4, we realized that the CNV calling results could be significantly improved, depending on the data, by changing some of the default parameters. We discussed with the Broad Institute team in charge of the pipeline development and we determined a list of parameters allegedly "optimized" for our exome sequencing data. We tested the impact of these parameters against the default parameters with the datasets we had at the time, *i.e.,* a Roche model consisting of 511 samples (still the model used for the Roche data, as we have not produced any new data using this wet lab protocol since then) and a Twist Legacy model consisting of 432 samples (samples that are fully included in the Twist model with 1154 samples constructed *a posteriori* with all the data produced with this wet lab protocol to date), results are shown in *Supp Fig 2*.



Supplementary Figure 2. Number of CNV in VCF file per patient (only CNV genotype different to 0) according to the set of parameters and the model used for the model construction (Roche n=511, Twist_Legacy n=432).

For both kits optimized parameters have a lower number of CNVs distribution per patient in regards to default parameters, which is expected for a more effective learning model and thus a more effective CNV calling model. Nevertheless, the mean or median number of CNVs per patient for the default condition is very different between the two kits, whereas using the optimized settings brings this number to about the same level. It is difficult to comment on CNV calling model differences, as the two datasets were produced with a different wet lab protocol, so they cannot be compared.

We have nevertheless tried to determine what could be the cause of these differences in profile. We first thought of an effect of patient recruitment bias that would lead to over-recruitment of poor-quality samples with too many CNVs. However, it is the Roche dataset that presents the most data with an atypical number of CNVs per patient, as it is the first exome data that we have produced ES data and on which we have learned the most. This does not correspond with a very high number of CNVs for the default condition of the Twist Legacy model compared to the Roche model. We compared many metrics between these two datasets and finally found that the real difference would be in the sequencing depth.

Optimized parameters for DetermingermlineContigPloidy tool

--sample-psi-scale 0.001

--global-psi-scale 0.05

--mean-bias-standard-deviation 1.0

Optimized parameters for GermlineCNVCaller

--adamax-beta-2 0.97

--caller-update-convergence-threshold 1e-06

--class-coherence-length 100000.0

--cnv-coherence-length 100000.0

--convergence-snr-trigger-threshold 0.2

--copy-number-posterior-expectation-mode EXACT

--depth-correction-tau 100.0

--interval-psi-scale 0.002

--learning-rate 0.03

--log-emission-sampling-median-rel-error 0.001

--log-emission-sampling-rounds 20

--log-mean-bias-standard-deviation 10.0

--max-advi-iter-first-epoch 5000

--max-advi-iter-subsequent-epochs 200

--max-bias-factors 16

--max-calling-iters 20

--max-training-epochs 20

--min-training-epochs 5

--num-thermal-advi-iters 4000

--p-active 0.01

--p-alt 0.001

--sample-psi-scale 1e-07

Effect of binned target on CNV calling results

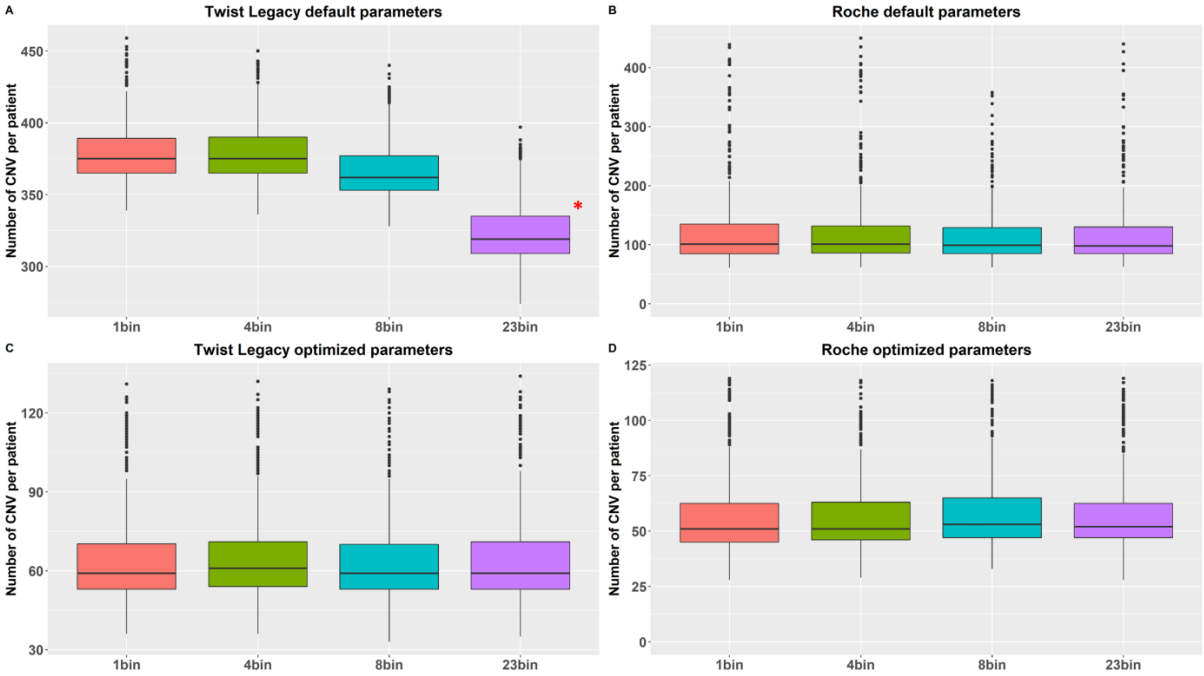
The creation of the learning model and more particularly the model of the GermlineCNVCaller step was a particularly long step and consumed computational resources, it was decided to divide the calling target to be able to parallelize their creation. We therefore divided our calling target into 4, 8 and 23 bins. For the first two conditions, we used GATK's IntervalListTools, which divides files in INTERVAL_LIST format into roughly equal parts. For the 23 bins condition, we made one bin per autosomal chromosome and one bin for the two gonosomal chromosomes.

We therefore divided our calling target into 4, 8 and 23 bins. For the first two conditions, we used GATK's IntervalListTools, which divides files in INTERVAL_LIST format into roughly equal parts. For the 23 bins condition, we made one bin per autosomal chromosome and one bin for the two gonosomal chromosomes.

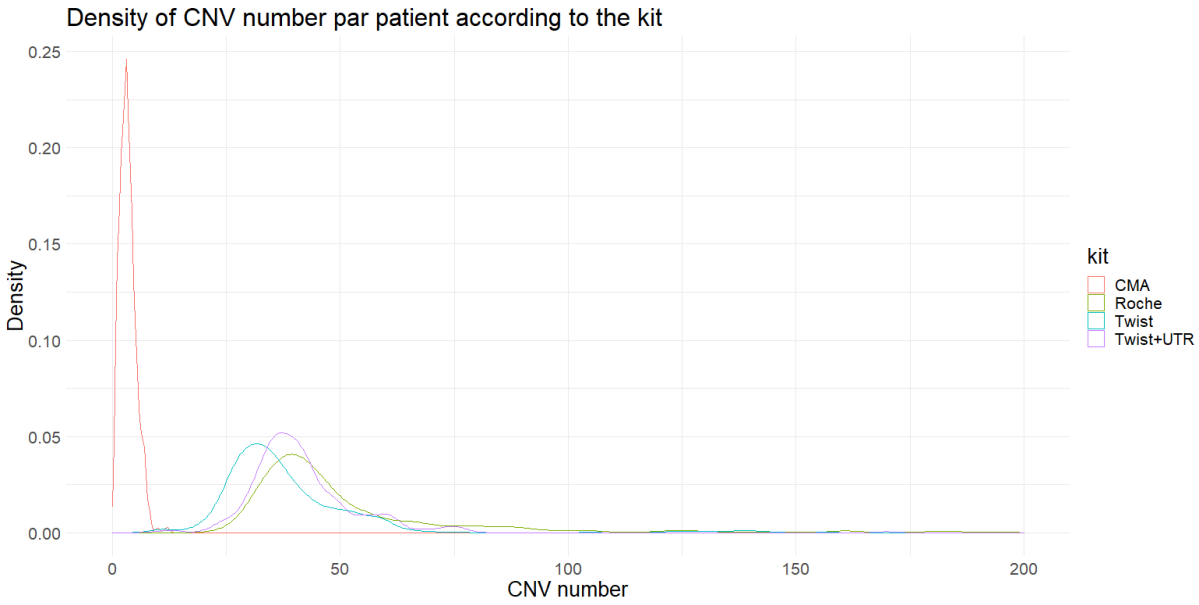
To determine whether the number of bins had any effect on the outcome of CNV calling, we decided to recreate models by reusing the data from the Roche (n = 511) and Twist Legacy (n = 432) models and the different binned targets previously created. The results of the different models are shown in *Supp Fig 3*.

A slight decrease in the number of unexplained CNVs per patient is visible for the Twist Legacy, default, “8 bin”, condition. A clear decrease in the number of CNV per patient is visible for the condition Twist Legacy, default, “23 bin”, due to the fact that the chr2 bin could never be completed with GATK version 4.1.4.1 due to a bug known by the developers. No significant difference in distribution attributable to the number of bins is visible.

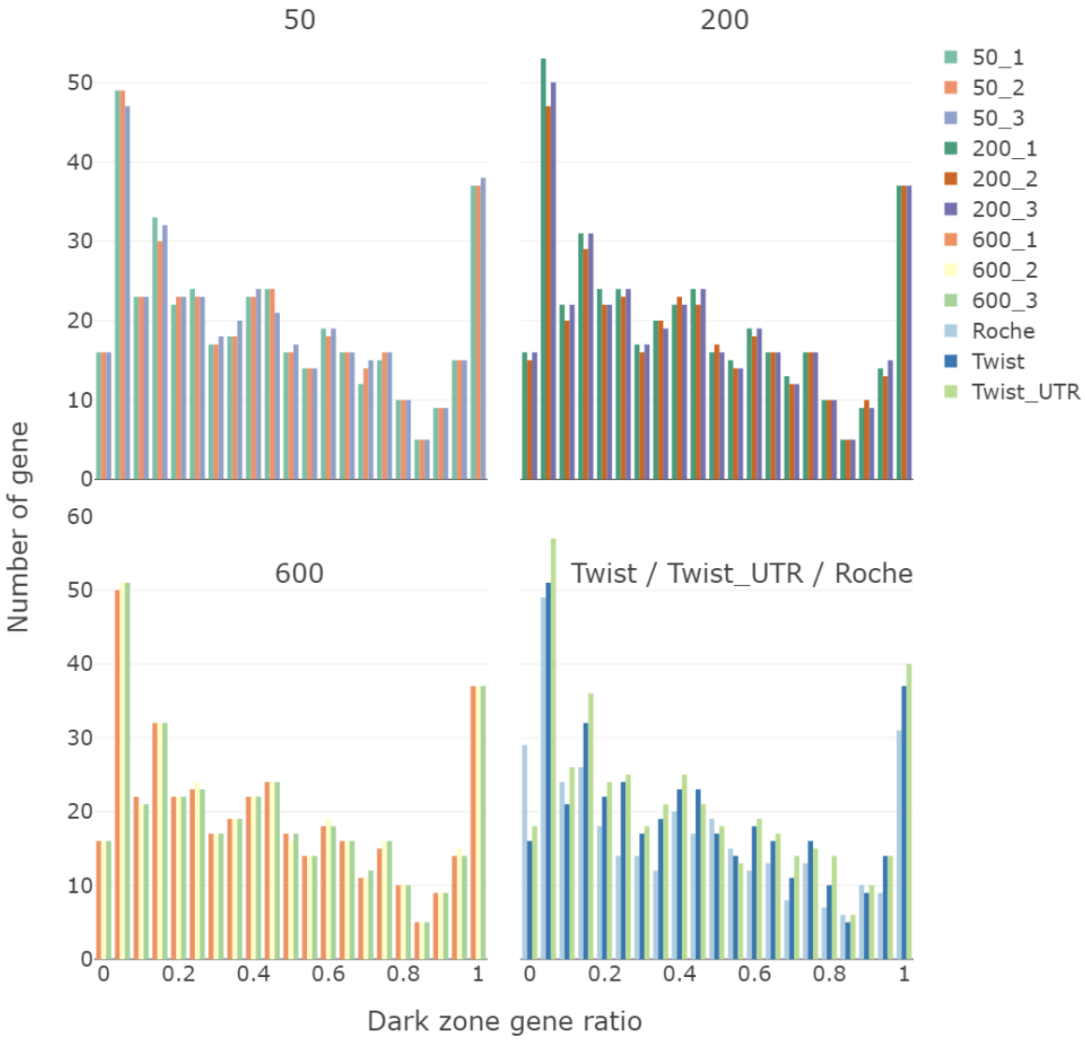
We have chosen for reasons of practicality related to the size of our computing infrastructure to create our models by dividing them into 4 bins.



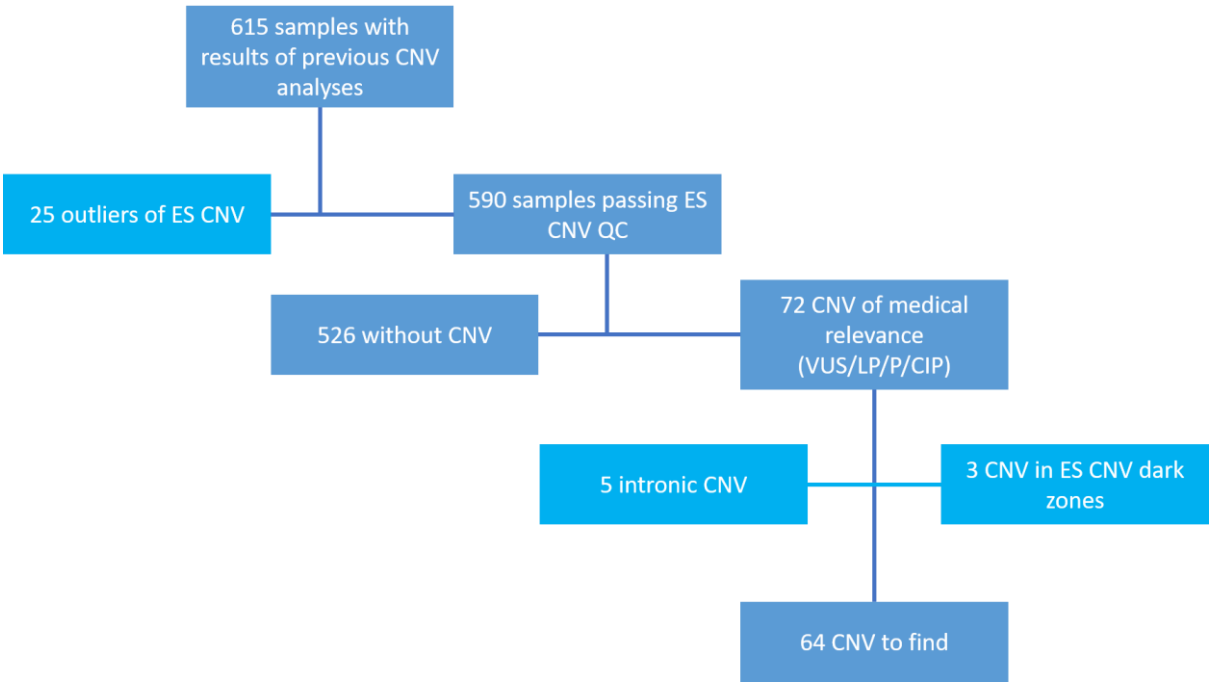
Supplementary Figure 3. Distribution of the CNV number per patient according to the parameter set, the kit used and the number of bins.



Supplementary Figure 4. Density of CNV counts per patient according to the kit / CNV model. *CMA (n=300), Roche (n=511), Twist (n=1154), Twist UTR (n=383).*



Supplementary Figure 5. Proportion of morbid genes coding sequence in CNV calling pipeline models black area of. Each color represents the number and proportion of genes in the black zone per model. Results shown from populated models of 50, 200, and 600 samples sequenced with the Twist Human Core exome kit, as well as results from Twist (n =1154), Roche (n=511), and Twist+UTR (n=383) models. 1 : 100% of the coding sequence is in the black zone.



Supplementary Figure 6. Methodology for determining the validation cohort CNVs from pre-existing CMA data. Light blue : CNVs eliminated from the validation cohort. VUS : Variant of Uncertain Significance, LP : Likely Pathogenic, P : Pathogenic, CIP : CNV with incomplete penetrance.

6.3 Discussion

L'appel de CNV sur donnés HTS n'est pas nouveau[153], néanmoins, l'appel de CNV sur données d'exome est plus complexe que sur les données de génome entier[154]. L'amélioration, la spécification et la diversification des stratégies de détection des outils d'appel de CNV sur données d'exome, ainsi que l'agrégation massive de données, ont pu permettre le développement de méthodologies d'appel de CNV sur exome performantes[155].

Ce travail s'inscrit donc dans la continuité de l'effort collectif du domaine. En revanche, il se distingue par plusieurs aspects. Tout d'abord, il se concentre sur une méthodologie novatrice, l'outil GATK4, développé par le *Broad Institute*, basé sur la construction de modèles d'apprentissage d'appel de CNV. La plupart des outils d'appel de CNV modernes adaptés aux données d'exome construisent un modèle de référence contre lequel sera effectué l'appel de CNV, afin d'éliminer le bruit induit par le séquençage par capture. En revanche, peu d'entre eux sont capables de construire puis réutiliser ces modèles, la plupart nécessitent le lancement des nouveaux échantillons avec le *batch* de contrôle, généralement constitué de l'ensemble des échantillons produits précédemment. De plus, la méthodologie de construction de modèles d'apprentissage à partir d'agrégation de données patients présentent des performances corrélées au nombre d'échantillons utilisé. Cette stratégie est particulièrement intéressante pour un laboratoire de Biologie Médicale qui produit beaucoup de données par an, mais plus généralement dans une époque où le volume de données produit est de plus en plus important. En revanche, cette stratégie n'est pas forcément adaptée à n'importe quel utilisateur et n'importe quel cas d'usage. La récolte de centaines voir milliers d'échantillons homogènes n'est à la portée que de certains laboratoires et industriels.

La méthodologie se distingue également par ses performances plus élevées que les stratégies publiées dans la littérature[155][156]. En effet, 100% des variants de la cohorte de validation, composée de CNV de diverses tailles et nombre de copies, censée représenter la routine clinique usuelle, ont été retrouvés par notre modèle actuellement en routine. Celui-ci a nécessité pour sa construction, l'agrégation de 1154 données patients homogènes (séquencées selon le même protocole paillasse) ainsi que près d'un mois de calcul sur la quasi-totalité de l'infrastructure Servoz. La puissance nécessaire de calcul pour la création de ces modèles représente également un frein pour l'adoption de cette stratégie. Pourtant, l'utilisation de modèles plus petits, avec des critères de filtration stringents en plus d'une base de données de récurrence des CNV permet l'élimination de la plupart des CNV faux positifs et artéfactuels. Néanmoins, plus on a de patients, moins le nombre de CNV à filtrer est important et plus notre base de données de récurrence est efficace. La volonté permanente d'augmentation du nombre de patients dans le modèle est donc positive.

Outre la nécessité d'un nombre de données important, il l'est important d'avoir une vision à plus long terme pour l'appel de CNV sur données sur exome qu'habituellement. Comme indiqué précédemment, les stratégies de construction de modèle d'apprentissage sont extrêmement sensibles aux biais techniques de production des données, beaucoup plus que l'appel de SNV. Il faudra alors anticiper les changements de protocole paillasse sur son activité de CNV *calling*, quitte à devoir attendre d'avoir engrangé plusieurs centaines d'échantillons pour la production d'un modèle qui ne produira pas de CNV faux positifs en un nombre aberrant.

De plus, l'appel de CNV sur exome présente les mêmes limites inhérentes à l'ensemble des méthodologies de séquençage par capture. Même les kits les plus performants de nos jours produisent des zones mal capturées et ainsi mal séquencées. Ces zones sont favorables à l'apparition de CNV artéfactuels dans les résultats d'appel de CNV, il a donc été choisi de les masquer. Ce sont les zones noires. Elles sont liées aux kits et non à la méthodologie en elle-même. Un changement de protocole de génération de données et donc de kit produit donc des données dont les zones noires sont différentes. En théorie, la plupart de ces zones sont communes, car la mauvaise hybridation des sondes repose sur la composition biologique de ces séquences mal capturées.

Les résultats produits sont assez bons pour pouvoir imaginer le remplacement partiel des technologies de puce à ADN pour la détection de CNV dans le cadre de la routine diagnostic maladie rare en première intention. Depuis des années, la stratégie de l'*exome-first* est plébiscitée par de nombreux acteurs du domaine[2]. L'amélioration des performances de CNV *calling* sur données d'exome est un argument complémentaire pour envisager une réorganisation de l'offre de soin autour du diagnostic de maladies rares.

Cette méthodologie a été intégrée dans le module *cnv* du pipeline Lygrexome et permet la réutilisation de modèles et des bases de données de récurrence associées construits précédemment. Ce module est utilisé en routine depuis plus d'un an au sein du laboratoire Eurofins Biomnis et de la communauté maladie rare grenobloise.

Conclusion et perspectives

Cette thèse avait pour but l'exploitation de données de séquençage pan-génomiques constitutionnelles par la mise en place de processus d'analyses industrielles. Plus globalement, elle avait également pour objectif de me former au métier de bioinformaticien clinique. À ses différentes composantes et exigences qui sont celles du métier d'aujourd'hui, ainsi qu'aux compétences transversales qui sont requises. De la Génétique, à la Biologie Moléculaire, dans le cadre de l'aide à l'interprétation, en passant par la Bioinformatique, pour l'analyse de données germinales ainsi que somatiques, sans oublier l'Informatique, les Systèmes et les Réseaux, pour la mise en place d'une structure informatique d'analyse, le tour d'horizon aura été assez complet. D'aucuns pourraient trouver que je me suis trop éparpillé, mais la réalité tient aussi du contexte de cette thèse. Pionnière en termes de recherche en Bioinformatique au CHU de Grenoble ainsi qu'à Eurofins Biomnis, il aura fallu construire les prémices d'une unité de Bioinformatique clinique au sein des deux structures.

Cela s'est traduit par, la mise en place d'un pipeline d'analyse de données d'exome constitutionnelles aux standards industriels, suivants les bonnes pratiques du domaine, de la mise en place d'une infrastructure informatique pour le traitement de données bioinformatique au sein du CHU de Grenoble. Mais également, a la mise en place d'un module d'appel de CNV issue d'une méthodologie de détection novatrice. À ces travaux, se sont rajoutés plusieurs projets de recherche en collaboration. Notamment, une analyse de données somatiques issue d'un modèle *in vitro* de cancer, une étude sur la performance d'outils de priorisation phénotypique, ainsi qu'une étude sur la détection de SV à partir de la technologie de séquençage novatrice *long-reads* ONT.

Certains de ces travaux mériteraient d'être approfondis, par exemple, la question de l'utilisation de séquençage ONT en routine. Bien que le séquençage à faible profondeur (5-10X), permette la caractérisation de SV connus, il ne permet pas, en l'état, la détection de SV pathologiques de petite taille sans *a priori*. Par la filtration de SV récurrents issus de bases de données de fréquence et de cohorte de normalisation, la plus grande partie des SV polymorphiques ou artéfactuels devrait être éliminée. L'utilisation de méthodes d'autocorrection des lectures pourrait également permettre la diminution du nombre de faux positifs issu de l'appel des SV. Néanmoins, ces méthodes sont encore extrêmement demandeuses en termes de ressource et de temps de calcul pour un résultat, certes meilleur, mais encore loin de la fiabilité à la base de la technologie Illumina[157]. Une autre possibilité serait l'utilisation de méthodes de correction ou d'assemblage hybride (petites et longues lectures) afin de marier le meilleur des deux mondes. Peu de données existent sur l'utilisation de ces deux stratégies sur des données humaines à basse couverture. Le coût d'un WGS Illumina et d'un WGS ONT à 10X étant plus élevé que le coût d'un exome ou que celui d'un WGS Illumina 30X sur une plateforme type PFMG 2025, il faudra déterminer dans quels cas cette association pourrait être utile.

En ce qui concerne le pipeline d'analyse d'exome, de nombreux ajouts et améliorations ont été envisagés durant la thèse et pourraient être entrepris si volonté de maintenir et améliorer cet outil il y a. L'ajout de tests d'intégration, la mise à jour de certains outils et bases de données, la refonte du module d'annotation ainsi que que de la stratégie de filtration des variations, l'interconnexion de l'outil avec le SGL et la base de données phénotypiques du CHUGA, le départ automatique des analyses, l'ajout d'un module de priorisation des variations, ainsi que le choix et la mise en place d'une solution de visualisation et de trie des variants des rapports d'interprétation (SNV et CNV) dans une interface plus *user friendly* qu'un fichier tabulé.

La mise en place du *cluster* HPC-CHU permettrait également de tester certains outils d'apprentissage machine utilisant le calcul GPU. Il serait intéressant de tester des outils tels que DeepVariant[158], même sur des données d'exome, pour constater si l'apprentissage machine permettrait de détecter des variations supposées difficiles. Les

profils de variations ciblés sont les variants mosaïques, les variants de séquences pseudogéniques ou les variants provenant de zones mal capturées.

Enfin, d'autres variations restent encore à détecter à partir des données de WES. Par exemple, le génome mitochondrial est capturé de manière aspécifique par les kits de capture d'exome, notamment grâce à son homologie avec certaines portions ciblées. L'ajout d'un module de détection de variations mitochondriales au pipeline d'analyse de données d'exome est envisagé.

Bibliographie

[1] Melissa Haendel, Nicole Vasilevsky, Deepak Unni, Cristian Bologa, Nomi Harris, Heidi Rehm, Ada Hamosh, Gareth Baynam, Tudor Groza, Julie McMurry, Hugh Dawkins, Ana Rath, Courtney Thaxon, Giovanni Bocci, Marcin P Joachimiak, Sebastian Köhler, Peter N Robinson, Chris Mungall, and Tudor I Oprea. How many rare diseases are there ? *Nat. Rev. Drug Discov.*, 19(2) :77–78, November 2019.

[2] J Thevenon, Y Duffourd, A Masurel-Paulet, M Lefebvre, F Feillet, S El Chehadeh-Djebbar, J St-Onge, A Steinmetz, F Huet, M Chouchane, V Darmency-Stamboul, P Callier, C Thauvin-Robinet, L Faivre, and J B Rivière. Diagnostic odyssey in severe neurodevelopmental disorders : toward clinical whole-exome sequencing as a first-line diagnostic test. *Clin. Genet.*, 89(6) :700–707, June 2016.

[3] Monica H Wojcik, Talia S Schwartz, Inbar Yamin, Heather L Edward, Casie A Genetti, Meghan C Towne, and Pankaj B Agrawal. Genetic disorders and mortality in infancy and early childhood : delayed diagnoses and missed opportunities. *Genetics in Medicine*, 20(11) :1396–1404, 2018.

[4] M J Field and T F Boat. Rare diseases and orphan products : Accelerating research and development. 2010.

[5] C Gilissen, A Hoischen, H G Brunner, and J A Veltman. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.*, 20(5), May 2012.

[6] J D Watson and F H C Crick. Molecular structure of nucleic acids : A structure for deoxyribose nucleic acid. *Nature*, 171(4356) :737–738, April 1953.

[7] R E Franklin and R G Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356) :740–741, April 1953.

[8] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1) :D766–D773, October 2018.

[9] K D Pruitt. NCBI reference sequence (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 33(Database issue) :D501–D504, December 2004.

[10] T Hubbard, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, T Down, R Durbin, E Eyras, J Gilbert, M Hammond, L Huminiecki, A Kasprzyk, H Lehvaslaiho, P Lijnzaad, C Melsopp, E Mongin, R Pettett, M Pocock, S Potter, A Rust, E Schmidt, S Searle, G Slater, J Smith, W Spooner, A Stabenau, J Stalker, E Stupka, A Ureta-Vidal, I Vastrik, and M Clamp. The ensembl genome database project. *Nucleic Acids Res.*, 30(1) :38–41, January 2002.

[11] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles

Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J Hubbard. GENCODE : the reference human genome annotation for the ENCODE project. *Genome Res.*, 22(9) :1760–1774, September 2012.

[12] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, 36(4) :338–345, April 2018.

[13] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Amy B Wilfert, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823) :79–84, September 2020.

[14] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J Hoyt, Mark Diekhans, Glennis A Logsdon, Michael Alonge, Stylianos E Antonarakis, Matthew Borchers, Gerard G Bouffard, Shelise Y Brooks, Gina V Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G de Lima, Philip C Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T Fiddes, Giulio Formenti, Robert S Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G S Grady, Tina A Graves-Lindsay, Ira M Hall, Nancy F Hansen, Gabrielle A Hartley, Marina Haukness, Kerstin Howe, Michael W Hunkapiller, Chirag Jain, Miten Jain, Erich D Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V Maduro, Tobias Marschall, Ann M McCartney, Jennifer McDaniel, Danny E Miller, James C Mullikin, Eugene W Myers, Nathan D Olson, Benedict Paten, Paul Peluso, Pavel A Pevzner, David Porubsky, Tamara Potapova, Evgeny I Rogaev, Jeffrey A Rosenfeld, Steven L Salzberg, Valerie A Schneider, Fritz J Sedlazeck, Kishwar Shafin, Colin J Shew, Alaina Shumate, Yumi Sims, Arian F A Smit, Daniela C Soto, Ivan Sović, Jessica M Storer, Aaron Streets, Beth A Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P Walenz, Aaron Wenger, Jonathan M D Wood, Chunlin Xiao, Stephanie M Yan, Alice C Young, Samantha Zarate, Urvashi Surti, Rajiv C McCoy, Megan Y Dennis, Ivan A Alexandrov, Jennifer L Gerton, Rachel J O'Neill, Winston Timp, Justin M Zook, Michael C Schatz, Evan E Eichler, Karen H Miga, and Adam M Phillippy. The complete sequence of a human genome. May 2021.

[15] International Human Genome Sequencing Consortium and International Human Genome Sequencing Consortium. Erratum : Initial sequencing and analysis of the human genome, 2001.

[16] Kate R Rosenbloom, Timothy R Dreszer, Michael Pheasant, Galt P Barber, Laurence R Meyer, Andy Pohl, Brian J Raney, Ting Wang, Angie S Hinrichs, Ann S Zweig, Pauline A Fujita, Katrina Learned, Brooke Rhead, Kayla E Smith, Robert M Kuhn, Donna Karolchik, David Haussler, and W James Kent. ENCODE whole-genome data in the UCSC genome browser. *Nucleic Acids Res.*, 38(Database issue) :D620–5, January 2010.

[17] A global reference for human genetic variation. *Nature*, 526(7571) :68–74, September 2015.

[18] Deanna M Church, Valerie A Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M McLaren, Graham R S Ritchie, Derek Albracht, Milinn Kremitzki, Susan Rock, Holland Kotkiewicz, Colin Kremitzki, Aye Wollam, Lee Trani, Lucinda Fulton, Robert Fulton, Lucy Matthews, Siobhan Whitehead, Will Chow, James Torrance, Matthew Dunn, Glenn Harden, Glen Threadgold, Jonathan Wood, Joanna Collins, Paul Heath, Guy Griffiths, Sarah Pelan, Darren Grafham,

Evan E Eichler, George Weinstock, Elaine R Mardis, Richard K Wilson, Kerstin Howe, Paul Flicek, and Tim Hubbard. Modernizing reference genome assemblies. *PLoS Biol.*, 9(7), July 2011.

[19] V A Schneider, T Graves-Lindsay, K Howe, N Bouk, H C Chen, P A Kitts, T D Murphy, K D Pruitt, F Thibaud-Nissen, D Albracht, R S Fulton, M Kremitzki, V Magrini, C Markovic, S McGrath, K M Steinberg, K Auger, W Chow, J Collins, G Harden, T Hubbard, S Pelan, J T Simpson, G Threadgold, J Torrance, J M Wood, L Clarke, S Koren, M Boitano, P Peluso, H Li, C S Chin, A M Phillippy, R Durbin, R K Wilson, P Flicek, E E Eichler, and D M Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, 27(5), May 2017.

[20] M J Falk, L Shen, M Gonzalez, J Leipzig, M T Lott, A P Stassen, M A Diroma, D Navarro-Gomez, P Yeske, R Bai, R G Boles, V Brilhante, Ral, J T DaRe, R Shelton, S F Terry, Z Zhang, W C Copeland, M van Oven, H Prokisch, D C Wallace, M Attimonelli, D Krotoski, S Zuchner, and X Gai. Mitochondrial disease sequence data resource (MSeqDR) : a global grass-roots consortium to facilitate deposition, curation, annotation, and integrated analysis of genomic data for the mitochondrial disease clinical and research communities. *Mol. Genet. Metab.*, 114(3), March 2015.

[21] Andreas M Kogelnik, Marie T Lott, Michael D Brown, Shamkant B Navathe, and Douglas C Wallace. MITOMAP : A human mitochondrial genome database. *Nucleic Acids Res.*, 24(1) :177–179, January 1996.

[22] Francesco Rubino, Roberta Piredda, Francesco Maria Calabrese, Domenico Simone, Martin Lang, Claudia Calabrese, Vittoria Petruzzella, Mila Tommaseo-Ponzetta, Giuseppe Gasparre, and Marcella Attimonelli. HmtDB, a genomic resource for mitochondrion-based human variability studies. *Nucleic Acids Res.*, 40(Database issue) :D1150, January 2012.

[23] Catarina D Campbell and Evan E Eichler. Properties and rates of germline mutations in humans. *Trends Genet.*, 29(10) :575–584, October 2013.

[24] S T Sherry, M H Ward, M Kholodov, J Baker, L Phan, E M Smigielski, and K Sirotkin. dbSNP : the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1) :308–311, January 2001.

[25] J L Freeman, G H Perry, L Feuk, R Redon, S A McCarroll, D M Altshuler, H Aburatani, K W Jones, C Tyler-Smith, M E Hurles, N P Carter, S W Scherer, and C Lee. Copy number variation : new insights in genome diversity. *Genome Res.*, 16(8), August 2006.

[26] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping, 2011.

[27] I Lappalainen, J Lopez, L Skipper, T Hefferon, J D Spalding, J Garner, C Chen, M Maguire, M Corbett, G Zhou, J Paschall, V Ananiev, P Flicek, and D M Church. DbVar and DGVa : public archives for genomic structural variation. *Nucleic Acids Res.*, 41(Database issue), January 2013.

[28] Mehdi Zarrei, Jeffrey R MacDonald, Daniele Merico, and Stephen W Scherer. A copy number variation map of the human genome. *Nat. Rev. Genet.*, 16(3) :172–183, February 2015.

[29] Jeffrey R MacDonald, Robert Ziman, Ryan K C Yuen, Lars Feuk, and Stephen W Scherer. The database of genomic variants : a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, 42(D1) :D986–D992, October 2013.

[30] Maladies rares, le modèle français. http://www.academie-medecine.fr/wp-content/uploads/2016/04/rapport-global-31-03-2016.pdf. Accessed : 2021-7-6.

[31] Joo Wook Ahn, Susan Bint, Anne Bergbaum, Kathy Mann, Richard P Hall, and Caroline Mackie Ogilvie. Array CGH as a first line diagnostic test in place of karyotyping for postnatal referrals - results from four years' clinical application for over 8,700 patients. *Mol. Cytogenet.*, 6(1) :1–6, April 2013.

[32] Daniel Pinkel, Richard Segraves, Damir Sudar, Steven Clark, Ian Poole, David Kowbel, Colin Collins, Wen-Lin Kuo, Chira Chen, Ye Zhai, Shanaz H Dairkee, Britt-Marie Ljung, Joe W Gray, and Donna G Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays, 1998.

[33] Stephen Scherer. Faculty opinions recommendation of diagnostic genome profiling in mental retardation, 2005.

[34] S Solinas-Toldo, S Lampel, S Stilgenbauer, J Nickolenko, A Benner, H Döhner, T Cremer, and P Lichter. Matrix-based comparative genomic hybridization : biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20(4), December 1997.

[35] D G Wang, J B Fan, C J Siao, A Berno, P Young, R Sapolsky, G Ghandour, N Perkins, E Winchester, J Spencer, L Kruglyak, L Stein, L Hsie, T Topaloglou, E Hubbell, E Robinson, M Mittmann, M S Morris, N Shen, D Kilburn, J Rioux, C Nusbaum, S Rozen, T J Hudson, R Lipshutz, M Chee, and E S Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366), May 1998.

[36] Brynn Levy and Rachel D Burnside. Are all chromosome microarrays the same ? what clinicians need to know. *Prenat. Diagn.*, 39(3) :157–164, February 2019.

[37] Robert Riehn, Manchun Lu, Yan-Mei Wang, Shuang Fang Lim, Edward C Cox, and Robert H Austin. Restriction mapping in nanofluidic devices. *Proc. Natl. Acad. Sci. U. S. A.*, 102(29) :10012–10016, July 2005.

[38] Giulio Formenti, Matteo Chiara, Lucy Poveda, Kees-Jan Francoijs, Andrea Bonisoli-Alquati, Luca Canova, Luca Gianfranceschi, David Stephen Horner, and Nicola Saino. SMRT long reads and direct label and stain optical maps allow the generation of a high-quality genome assembly for the european barn swallow (hirundo rustica rustica). *Gigascience*, 8(1), November 2018.

[39] F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, J C Fiddes, C A Hutchison, P M Slocombe, and M Smith. Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, 265(5596) :687–695, February 1977.

[40] Michael J Clark, Rui Chen, Hugo Y K Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte, and Michael Snyder. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, 29(10) :908–914, September 2011.

[41] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery, 2011.

[42] J Majewski, J Schwartzentruber, E Lalonde, A Montpetit, and N Jabado. What can exome sequencing do for you ? *J. Med. Genet.*, 48(9), September 2011.

[43] Hayan Lee, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W Richard McCombie, and Michael C Schatz. Third-generation sequencing and the future of genomics. April 2016.

[44] J J Kasianowicz, E Brandin, D Branton, and D W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24), November 1996.

[45] Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. Long-read human genome sequencing and its applications, 2020.

[46] H Lu, F Giordano, and Z Ning. Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics*, 14(5), October 2016.

[47] Wouter De Coster, Peter De Rijk, Arne De Roeck, Tim De Pooter, Svenn D'Hert, Mojca Strazisar, Kristel Sleegers, and Christine Van Broeckhoven. Structural variants identified by oxford nanopore PromethION sequencing of the human genome. *Genome Res.*, 29(7) :1178–1187, July 2019.

[48] A Rhoads and K F Au. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, 13(5), October 2015.

[49] A M Wenger, P Peluso, W J Rowell, P C Chang, R J Hall, G T Concepcion, J Ebler, A Fungtammasan, A Kolesnikov, N D Olson, A Töpfer, M Alonge, M Mahmoud, Y Qian, C S Chin, A M Phillippy, M C Schatz, G Myers, M A DePristo, J Ruan, T Marschall, F J Sedlazeck, J M Zook, H Li, S Koren, A Carroll, D R Rank, and M W Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10), October 2019.

[50] Linked-Reads genomics - 10x genomics. https://www.10xgenomics.com/products/linked-reads. Accessed : 2021-9-11.

[51] Christophe Dessimoz Christian Ledergerber. Base-calling for next-generation sequencing platforms. *Brief. Bioinform.*, 12(5) :489, September 2011.

[52] bcl2fastq conversion software. https://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html. Accessed : 2021-8-25.

[53] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11) :1851–1858, November 2008.

[54] K Reinert, B Langmead, D Weese, and D J Evers. Alignment of Next-Generation sequencing reads. *Annu. Rev. Genomics Hum. Genet.*, 16, 2015.

[55] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14) :1754–1760, May 2009.

[56] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.

[57] Heng Li. Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18) :3094–3100, May 2018.

[58] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16) :2078–2079, June 2009.

[59] James K Bonfield, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies. HTSlib : C library for reading/writing high-throughput sequencing data. *Gigascience*, 10(2), February 2021.

[60] Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, 21(5) :734–740, May 2011.

[61] Geraldine A Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V Garimella, David Altshuler, Stacey Gabriel, and Mark A DePristo. From FastQ data to High-Confidence variant calls : The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, 43(1), 2013.

[62] Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples.

[63] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, 15(6) :461–468, April 2018.

[64] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15) :2156, August 2011.

[65] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, and Daniel G MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616) :285–291, August 2016.

[66] Claudia Gonzaga-Jauregui, James R Lupski, and Richard A Gibbs. Human genome sequencing in health and disease, 2012.

[67] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L Rehm. Standards and guidelines for the

interpretation of sequence variants : a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.*, 17(5) :405–423, March 2015.

[68] Mahjoubeh Jalali Sefid Dashti and Junaid Gamieldien. A practical guide to filtering and prioritizing genetic variants. *Biotechniques*, 62(1) :18–30, January 2017.

[69] D G MacArthur, T A Manolio, D P Dimmock, H L Rehm, J Shendure, G R Abecasis, D R Adams, R B Altman, S E Antonarakis, E A Ashley, J C Barrett, L G Biesecker, D F Conrad, G M Cooper, N J Cox, M J Daly, M B Gerstein, D B Goldstein, J N Hirschhorn, S M Leal, L A Pennacchio, J A Stamatoyannopoulos, S R Sunyaev, D Valle, B F Voight, W Winckler, and C Gunter. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497) :469–476, April 2014.

[70] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, Benjamin M Neale, Mark J Daly, and Daniel G MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809) :434–443, May 2020.

[71] M J Landrum, J M Lee, M Benson, G R Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, W Jang, K Karapetyan, K Katz, C Liu, Z Maddipatla, A Malheiro, K McDaniel, M Ovetsky, G Riley, G Zhou, J B Holmes, B L Kattman, and D R Maglott. ClinVar : improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, 46(D1), January 2018.

[72] Anthony A Philippakis, Danielle R Azzariti, Sergi Beltran, Anthony J Brookes, Catherine A Brownstein, Michael Brudno, Han G Brunner, Orion J Buske, Knox Carey, Cassie Doll, Sergiu Dumitriu, Stephanie O M Dyke, Johan T den Dunnen, Helen V Firth, Richard A Gibbs, Marta Girdea, Michael Gonzalez, Melissa A Haendel, Ada Hamosh, Ingrid A Holm, Lijia Huang, Matthew E Hurles, Ben Hutton, Joel B Krier, Andriy Misyura, Christopher J Mungall, Justin Paschall, Benedict Paten, Peter N Robinson, François Schiettecatte, Nara L Sobreira, Ganesh J Swaminathan, Peter E Taschner, Sharon F Terry, Nicole L Washington, Stephan Züchner, Kym M Boycott, and Heidi L Rehm. The matchmaker exchange : a platform for rare disease gene discovery. *Hum. Mutat.*, 36(10) :915–921, October 2015.

[73] P Cingolani, A Platts, Wang le L, M Coon, T Nguyen, L Wang, S J Land, X Lu, and D M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff : SNPs in the genome of drosophila melanogaster strain w1118 ; iso-2 ; iso-3. *Fly*, 6(2), 2012.

[74] P Cingolani, V M Patel, M Coon, T Nguyen, S J Land, D M Ruden, and X Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.*, 3, March 2012.

[75] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biol.*, 17(1) :1–14, June 2016.

[76] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, 0 7 :Unit7.20, January 2013.

[77] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD : predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, 47(D1) :D886–D894, October 2018.

[78] Joanna Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. McKusick's online mendelian inheritance in man (OMIM®). *Nucleic Acids Res.*, 37(Database issue) :D793, January 2009.

[79] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology : A tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, 83(5) :610, November 2008.

[80] Akira. Introduction aux pipelines - blog bioinformatique communautaire scientifique. https://bioinfo-fr.net/introduction-aux-pipelines, October 2012. Accessed : 2021-8-26.

[81] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity : Scientific containers for mobility of compute. *PLoS One*, 12(5) :e0177459, May 2017.

[82] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35(4) :316–319, April 2017.

[83] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine, 2018.

[84] Philip A Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.*, 38(3) :276–278, February 2020.

[85] Maxime Garcia, Szilveszter Juhos, Malin Larsson, Pall I Olason, Marcel Martin, Jesper Eisfeldt, Sebastian DiLorenzo, Johanna Sandgren, Teresita Díaz De Ståhl, Philip Ewels, Valtteri Wirta, Monica Nistér, Max Käller, and Björn Nystedt. Sarek : A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res.*, 9 :63, January 2020.

[86] Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed : 2021-9-6.

[87] Test compagnon associé à une thérapie ciblée : définitions et méthode d'évaluation. https://www.has-sante.fr/upload/docs/application/pdf/2014-04/guide_meth_court_test_cpagnon_vd.pdf. Accessed : 2021-8-16.

[88] Christian R Marshall, Shimul Chowdhury, Ryan J Taft, Mathew S Lebo, Jillian G Buchan, Steven M Harrison, Ross Rowsey, Eric W Klee, Pengfei Liu, Elizabeth A Worthey, Vaidehi Jobanputra, David Dimmock, Hutton M Kearney, David Bick, Shashikant Kulkarni, Stacie L Taylor, John W Belmont, Dimitri J Stavropoulos, and Niall J Lennon. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. *npj Genomic Medicine*, 5(1) :1–12, October 2020.

[89] Justin M Zook, Jennifer McDaniel, Nathan D Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A Irvine, Len Trigg, Rebecca Truty, Cory Y McLean, Francisco M De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.*, 37(5) :561–566, April 2019.

[90] Peter Krusche, Len Trigg, Paul C Boutros, Christopher E Mason, Francisco M De La Vega, Benjamin L Moore, Mar Gonzalez-Porta, Michael A Eberle, Zivana Tezak, Samir Lababidi, Rebecca Truty, George Asimenos, Birgit Funke, Mark Fleharty, Marc Salit, Justin M Zook, and the Global Alliance for Genomics and Health Benchmarking Team. Best practices for benchmarking germline small variant calls in human genomes.

[91] Rachel L Goldfeder, James R Priest, Justin M Zook, Megan E Grove, Daryl Waggott, Matthew T Wheeler, Marc Salit, and Euan A Ashley. Medical implications of technical accuracy in genome sequencing. *Genome Med.*, 8(1) :1–12, March 2016.

[92] Justin M Zook, Nancy F Hansen, Nathan D Olson, Lesley Chapman, James C Mullikin, Chunlin Xiao, Stephen Sherry, Sergey Koren, Adam M Phillippy, Paul C Boutros, Sayed Mohammad E Sahraeian, Vincent Huang, Alexandre Rouette, Noah Alexander, Christopher E Mason, Iman Hajirasouliha, Camir Ricketts, Joyce Lee, Rick Tearle, Ian T Fiddes, Alvaro Martinez Barrio, Jeremiah Wala, Andrew Carroll, Noushin Ghaffari, Oscar L Rodriguez, Ali Bashir, Shaun Jackman, John J Farrell, Aaron M Wenger, Can Alkan, Arda Soylev, Michael C Schatz, Shilpa Garg, George Church, Tobias Marschall, Ken Chen, Xian Fan, Adam C English, Jeffrey A Rosenfeld, Weichen Zhou, Ryan E Mills, Jay M Sage, Jennifer R Davis, Michael D Kaiser, John S Oliver, Anthony P Catalano, Mark J P Chaisson, Noah Spies, Fritz J Sedlazeck, and Marc Salit. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.*, 38(11) :1347–1355, June 2020.

[93] Mark J P Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, Xian Fan, Jia Wen, Robert E Handsaker, Susan

Fairley, Zev N Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M Wenger, Alex R Hastie, Danny Antaki, Thomas Anantharaman, Peter A Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T Chuang, Christine C Lambert, Deanna M Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David U Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Ernest T Lam, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M Munson, Fabio C P Navarro, Bradley J Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy W C Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C J Spierings, Alistair Ward, Annemarie E Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B Gerstein, Pui-Yan Kwok, Peter M Lansdorp, Gabor T Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E Devine, Michael E Talkowski, Ryan E Mills, Tobias Marschall, Jan O Korbel, Evan E Eichler, and Charles Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, 10(1) :1–16, April 2019.

[94] Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.*, 32(3) :246–251, March 2014.

[95] Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, Elizabeth Henaff, Alexa B R McIntyre, Dhruva Chandramohan, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M Truty, Christopher C Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Chunlin Xiao, Jonathan Trow, Stephen T Sherry, Alexander W Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace X Y Zheng, Michael Schnall-Levin, Heather S Ordonez, Patrice A Mudivarti, Kristina Giorda, Ying Sheng, Karoline Bjarnesdatter Rypdal, and Marc Salit. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3(1) :1–26, June 2016.

[96] Michael A Eberle, Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, Semyon Kruglyak, Elliott H Margulies, Gil McVean, and David R Bentley. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, 27(1) :157–164, January 2017.

[97] Peter Krusche, Len Trigg, Paul C Boutros, Christopher E Mason, Francisco M De La Vega, Benjamin L Moore, Mar Gonzalez-Porta, Michael A Eberle, Zivana Tezak, Samir Lababidi, Rebecca Truty, George Asimenos, Birgit Funke, Mark Fleharty, Brad A Chapman, Marc Salit, Justin M Zook, and Global Alliance for Genomics and Health Benchmarking Team. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.*, 37(5) :555–560, May 2019.

[98] Janine Meienberg, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. Clinical sequencing : is WGS the better WES ? *Hum. Genet.*, 135(3) :359–362, March 2016.

[99] H Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20), October 2014.

[100] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K Konkel, Ankit Malhotra, Adrian M Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J P Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y K Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A Gibbs, Gabor Marth, Christopher E Mason, Androniki Menelaou, Donna M Muzny, Bradley J Nelson, Amina Noor, Nicholas F Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E Schadt, Mallory Romanovitch,

Andreas Schlattl, Robert Sebra, Andrey A Shabalin, Andreas Untergasser, Jerilyn A Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A Batzer, Steven A McCarroll, Ryan E Mills, Mark B Gerstein, Ali Bashir, Oliver Stegle, Scott E Devine, Charles Lee, Evan E Eichler, and Jan O Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571) :75–81, September 2015.

[101] Daniel C Jeffares, Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J Sedlazeck. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.*, 8 :14061, January 2017.

[102] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, Wesley C Warren, Vincent Magrini, Sean D McGrath, Yang I Li, Richard K Wilson, and Evan E Eichler. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3) :663–675.e19, January 2019.

[103] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, Nicholas A Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher W Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Caroline Cusick, Lauren Margolin, Kent D Taylor, Henry J Lin, Stephen S Rich, Wendy S Post, Yii-Der Ida Chen, Jerome I Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M Neale, Sekar Kathiresan, Mark J Daly, Eric Banks, Daniel G MacArthur, and Michael E Talkowski. A structural variation reference for medical and population genetics. *Nature*, 581(7809) :444–451, May 2020.

[104] Robert C Green, Jonathan S Berg, Wayne W Grody, Sarah S Kalia, Bruce R Korf, Christa L Martin, Amy L McGuire, Robert L Nussbaum, Julianne M O'Daniel, Kelly E Ormond, Heidi L Rehm, Michael S Watson, Marc S Williams, and Leslie G Biesecker. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing, 2013.

[105] Victor A McKusick. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, 80(4) :588–604, 2007.

[106] Jennifer E Posey. Genome sequencing and implications for rare disorders, 2019.

[107] L Zinman, H N Liu, C Sato, Y Wakutani, A F Marvelle, D Moreno, K E Morrison, K L Mohlke, J Bilbao, J Robertson, and E Rogaeva. A mechanism for low penetrance in an ALS family with a novel SOD1 deletion. *Neurology*, 72(13) :1153–1159, March 2009.

[108] Evidence that paternal expression of the ϵ-Sarcoglycan gene accounts for reduced penetrance in Myoclonus-Dystonia. *Am. J. Hum. Genet.*, 71(6) :1303–1311, December 2002.

[109] Zhipeng Ma, Peipei Zhu, Hui Shi, Liwei Guo, Qinghe Zhang, Yanan Chen, Shuming Chen, Zhe Zhang, Jinrong Peng, and Jun Chen. PTC-bearing mRNA elicits a genetic compensation response via upf3a and COMPASS components. *Nature*, 568(7751) :259–263, April 2019.

[110] Kaitlin E Samocha, Elise B Robinson, Stephan J Sanders, Christine Stevens, Aniko Sabo, Lauren M McGrath, Jack A Kosmicki, Karola Rehnström, Swapan Mallick, Andrew Kirby, Dennis P Wall, Daniel G MacArthur, Stacey B Gabriel, Mark DePristo, Shaun M Purcell, Aarno Palotie, Eric Boerwinkle, Joseph D Buxbaum, Edwin H Cook, Jr, Richard A Gibbs, Gerard D Schellenberg, James S Sutcliffe, Bernie Devlin, Kathryn Roeder, Benjamin M Neale, and Mark J Daly. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, 46(9) :944–950, September 2014.

[111] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, Eric D Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3) :535–548.e24, January 2019.

[112] Philipp Rentzsch, Max Schubach, Jay Shendure, and Martin Kircher. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores, 2021.

[113] Xiaotu Ma, Ying Shao, Liqing Tian, Diane A Flasch, Heather L Mulder, Michael N Edmonson, Yu Liu, Xiang Chen, Scott Newman, Joy Nakitandwe, Yongjin Li, Benshang Li, Shuhong Shen, Zhaoming Wang, Sheila Shurtleff, Leslie L Robison, Shawn Levy, John Easton, and Jinghui Zhang. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.*, 20, 2019.

[114] Aziz Belkadi, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.*, 112(17) :5473–5478, April 2015.

[115] C Gilissen, J Y Hehir-Kwa, D T Thung, M van de Vorst, B W van Bon, M H Willemsen, M Kwint, I M Janssen, A Hoischen, A Schenck, R Leach, R Klein, R Tearle, T Bo, R Pfundt, H G Yntema, B B de Vries, T Kleefstra, H G Brunner, L E Vissers, and J A Veltman. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509), July 2014.

[116] E E Palmer, R Sachdev, R Macintosh, U S Melo, S Mundlos, S Righetti, T Kandula, A E Minoche, C Puttick, V Gayevskiy, L Hesson, S Idrisoglu, C Shoubridge, M H N Thai, R L Davis, A P Drew, H Sampaio, P I Andrews, J Lawson, M Cardamone, D Mowat, A Colley, S Kummerfeld, M E Dinger, M J Cowley, T Roscioli, A Bye, and E Kirk. Diagnostic yield of whole genome sequencing after nondiagnostic exome sequencing or gene panel in developmental and epileptic encephalopathies. *Neurology*, 96(13), March 2021.

[117] France médecine génomique 2025. https://www.inserm.fr/wp-content/uploads/2017-11/aviesan-planfrancemedecinegenomique-2025.pdf. Accessed : 2021-8-30.

[118] Gustavo Glusman, Juan Caballero, Denise E Mauldin, Leroy Hood, and Jared C Roach. Kaviar : an accessible system for testing SNV novelty. *Bioinformatics*, 27(22) :3216–3217, September 2011.

[119] Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu. dbNSFP v4 : a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.*, 12(1) :1–8, December 2020.

[120] S A Forbes, G Bhamra, S Bamford, E Dawson, C Kok, J Clements, A Menzies, J W Teague, P A Futreal, and M R Stratton. The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.*, CHAPTER :Unit, April 2008.

[121] David T Miller, Kristy Lee, Wendy K Chung, Adam S Gordon, Gail E Herman, Teri E Klein, Douglas R Stewart, Laura M Amendola, Kathy Adelman, Sherri J Bale, Michael H Gollob, Steven M Harrison, Ray E Hershberger, Kent McKelvey, C Sue Richards, Christopher N Vlangos, Michael S Watson, and Christa Lese Martin. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing : a policy statement of the american college of medical genetics and genomics (ACMG). *Genet. Med.*, 23(8) :1381–1390, May 2021.

[122] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC : summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19) :3047–3048, June 2016.

[123] Konstantin Okonechnikov, Ana Conesa, and Fernando García-Alcalde. Qualimap 2 : advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2) :292, January 2016.

[124] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts : an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7) :923–930, November 2013.

[125] Andy B Yoo, Morris A Jette, and Mark Grondona. SLURM : Simple linux utility for resource management. In *Job Scheduling Strategies for Parallel Processing*, pages 44–60. Springer, Berlin, Heidelberg, June 2003.

[126] M Girdea, S Dumitriu, M Fiume, S Bowdin, K M Boycott, S Chénier, D Chitayat, H Faghfoury, M S Meyn, P N Ray, J So, D J Stavropoulos, and M Brudno. PhenoTips : patient phenotyping software for clinical and research use. *Hum. Mutat.*, 34(8), August 2013.

[127] Karthik A Jagadeesh, Johannes Birgmeier, Harendra Guturu, Cole A Deisseroth, Aaron M Wenger, Jonathan A Bernstein, and Gill Bejerano. Phrank measures phenotype sets similarity to greatly improve mendelian diagnostic disease prioritization. *Genet. Med.*, 21(2) :464–470, July 2018.

[128] Sacha Schutz, Tristan Montier, and Emmanuelle Genin. Cutevariant : a GUI-based desktop application to explore genetics variations. February 2021.

[129] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. Calling somatic SNVs and indels with mutect2.

[130] M Díaz-Gay, M Vila-Casadesús, S Franch-Expósito, E Hernández-Illán, J J Lozano, and S Castellví-Bel. Mutational signatures in cancer (MuSiCa) : a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics*, 19(1), June 2018.

[131] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. The DAVID gene functional classification tool : a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, 8(9) :R183, 2007.

[132] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome Res.*, 21(3) :487–493, March 2011.

[133] David Heller and Martin Vingron. SVIM : structural variant identification using mapped long reads. *Bioinformatics*, 35(17) :2907–2915, January 2019.

[134] R S Harris, M Cechova, and K D Makova. Noise-cancelling repeat finder : uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics*, 35(22), November 2019.

[135] Bjorn Bakker, Aaron Taudt, Mirjam E Belderbos, David Porubsky, Diana C J Spierings, Tristan V de Jong, Nancy Halsema, Hinke G Kazemier, Karina Hoekstra-Wakker, Allan Bradley, Eveline S J M de Bont, Anke van den Berg, Victor Guryev, Peter M Lansdorp, Maria Colomé-Tatché, and Floris Foijer. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.*, 17(1) :1–15, May 2016.

[136] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nat. Biotechnol.*, 29(1) :24, January 2011.

[137] Ole Tange. *GNU Parallel 2018*. April 2018.

[138] Kishwar Shafin, Trevor Pesout, Ryan Lorig-Roach, Marina Haukness, Hugh E Olsen, Colleen Bosworth, Joel Armstrong, Kristof Tigyi, Nicholas Maurer, Sergey Koren, Fritz J Sedlazeck, Tobias Marschall, Simon Mayes, Vania Costa, Justin M Zook, Kelvin J Liu, Duncan Kilburn, Melanie Sorensen, Katy M Munson, Mitchell R Vollger, Jean Monlong, Erik Garrison, Evan E Eichler, Sofie Salama, David Haussler, Richard E Green, Mark Akeson, Adam Phillippy, Karen H Miga, Paolo Carnevali, Miten Jain, and Benedict Paten. Nanopore sequencing and the shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.*, 38(9) :1044–1053, September 2020.

[139] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack : visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15) :2666–2669, March 2018.

[140] Kristin D Kernohan, Taila Hartley, Najmeh Alirezaie, Care4Rare Canada Consortium, Peter N Robinson, David A Dyment, and Kym M Boycott. Evaluation of exome filtering techniques for the analysis of clinically relevant genes. *Hum. Mutat.*, 39(2) :197–201, February 2018.

[141] Damian Smedley, Julius O B Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J Buske, Nicole L Washington, William P Bone, Melissa A Haendel, and Peter N Robinson. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat. Protoc.*, 10(12) :2004–2015, December 2015.

[142] Johannes Birgmeier, Maximilian Haeussler, Cole A Deisseroth, Karthik A Jagadeesh, Alexander J Ratner, Harendra Guturu, Aaron M Wenger, Peter D Stenson, David N Cooper, Christopher Ré, Jonathan A Bernstein, and Gill Bejerano. AMELIE accelerates mendelian patient diagnosis directly from the primary literature.

[143] Judith A Blake, Carol J Bult, James A Kadin, Joel E Richardson, Janan T Eppig, and Mouse Genome Database Group. The mouse genome database (MGD) : premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.*, 39(Database issue) :D842–8, January 2011.

[144] Gautier Koscielny, Gagarine Yaikhom, Vivek Iyer, Terrence F Meehan, Hugh Morgan, Julian Atienza-Herrero, Andrew Blake, Chao-Kung Chen, Richard Easty, Armida Di Fenza, Tanja Fiegel, Mark Grifiths, Alan Horne, Natasha A Karp, Natalja Kurbatova, Jeremy C Mason, Peter Matthews, Darren J Oakley, Asfand Qazi, Jack Regnart, Ahmad Retha, Luis A Santos, Duncan J Sneddon, Jonathan Warren, Henrik Westerberg, Robert J

Wilson, David G Melvin, Damian Smedley, Steve D M Brown, Paul Flicek, William C Skarnes, Ann-Marie Mallon, and Helen Parkinson. The international mouse phenotyping consortium web portal, a unified point of access for knockout mice and related phenotyping data, 2014.

[145] Human diseases. https://www.mousephenotype.org/human-diseases/, November 2018. Accessed : 2021-9-23.

[146] Ceri E Van Slyke, Yvonne M Bradford, Monte Westerfield, and Melissa A Haendel. The zebrafish anatomy and stage ontologies : representing the anatomy and development of danio rerio. *J. Biomed. Semantics*, 5(1) :12, February 2014.

[147] Gene Ontology Consortium. Gene ontology consortium : going forward. *Nucleic Acids Res.*, 43(Database issue) :D1049–56, January 2015.

[148] Georgios V Gkoutos, Chris Mungall, Sandra Dolken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian Kohler, and Peter N Robinson. Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2009 :7069–7072, 2009.

[149] S S Weinreich, R Mangon, J J Sikkens, M E Teeuw, and M C Cornel. [orphanet : a european database for rare diseases]. *Ned. Tijdschr. Geneeskd.*, 152(9) :518–519, March 2008.

[150] Johannes Birgmeier, Maximilian Haeussler, Cole A Deisseroth, Ethan H Steinberg, Karthik A Jagadeesh, Alexander J Ratner, Harendra Guturu, Aaron M Wenger, Mark E Diekhans, Peter D Stenson, David N Cooper, Christopher Ré, Alan H Beggs, Jonathan A Bernstein, and Gill Bejerano. AMELIE speeds mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.*, 12(544), May 2020.

[151] Valentina Cipriani, Nikolas Pontikos, Gavin Arno, Panagiotis I Sergouniotis, Eva Lenassi, Penpitcha Thawong, Daniel Danis, Michel Michaelides, Andrew R Webster, Anthony T Moore, Peter N Robinson, Julius O B Jacobsen, and Damian Smedley. An improved Phenotype-Driven tool for rare mendelian variant prioritization : Benchmarking exomiser on real patient Whole-Exome data. *Genes*, 11(4), April 2020.

[152] Johannes Birgmeier, Maximilian Haeussler, Cole A Deisseroth, Ethan H Steinberg, Karthik A Jagadeesh, Alexander J Ratner, Harendra Guturu, Aaron M Wenger, Mark E Diekhans, Peter D Stenson, David N Cooper, Christopher Ré, The Manton Center, Alan H Beggs, Jonathan A Bernstein, and Gill Bejerano. AMELIE 2 speeds up mendelian diagnosis by matching patient phenotype & genotype to primary literature. November 2019.

[153] Shaun M Purcell Menachem Fromer. Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet.*, 81 :7.23.1, 2014.

[154] Celine S Hong, Larry N Singh, James C Mullikin, and Leslie G Biesecker. Assessing the reproducibility of exome copy number variations predictions. *Genome Med.*, 8(1) :82, 2016.

[155] Olivier Quenez, Kevin Cassinari, Sophie Coutant, Francois Lecoquierre, Kilan Le Guennec, Stéphane Rousseau, Anne-Claire Richard, Stéphanie Vasseur, Emilie Bouvignies, Jacqueline Bou, et al. Detection of copy-number variations from ngs data using read depth information : a diagnostic performance evaluation. *European Journal of Human Genetics*, 29(1) :99–109, 2021.

[156] Ramakrishnan Rajagopalan, Jill R Murrell, Minjie Luo, and Laura K Conlin. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome medicine*, 12(1) :1–11, 2020.

[157] Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, and Arnaud Lefebvre. CONSENT : Scalable long read self-correction and assembly polishing with multiple sequence alignment. April 2020.

[158] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean, and Mark A DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, 36(10) :983–987, September 2018.

Résumé

Le nombre de maladies rares est estimé à entre 5000 et 8000 pathologies distinctes. Elles sont individuellement rares puisque par définition elles affectent moins de 1 individu sur 2000 dans la population générale, mais leur nombre les rend collectivement fréquentes. En raison du nombre limité de patients atteints de certaines maladies et du manque de connaissances à leur sujet, le diagnostic pour les patients est souvent retardé.

Plus de 80 % des maladies rares auraient une origine génétique, en grande majorité monogénique. La démocratisation des technologies de séquençage à haut débit depuis 2005 a permis l'acquisition massive de données génomiques que ce soit d'individus sains ou de patients. Pourtant, établir un diagnostic de maladie rare avec les technologies de biologie moléculaire actuelle reste difficile.

L'application à échelle industrielle des analyses de séquençage d'exome et de génome associée avec l'augmentation constante des connaissances médicales représente un espoir concret pour apporter un diagnostic à la majorité des patients concernés par une maladie rare suspecte d'être génétique.

Cette thèse s'est particulièrement intéressée à la mise en place de processus d'analyse de données de séquençage pan-génomiques constitutionnelles selon les standards industriels et les bonnes pratiques en vigueur dans le domaine. Puis, à la détection de variations de nombre de copies à partir de données de séquençage d'exomes, analyse souvent inexplorée par de nombreux laboratoires.

Enfin, cette thèse aura permis le développement et l'initiation de collaborations encore actives à ce jour. Cela s'est concrétisé par une étude ayant comme objet la détection de variations la détection somatique de petite taille dans un modèle *in vitro* de développement de cancer lié au microenvironnement cellulaire, la détection de variations structurales à l'aide de la technologie de séquençage novatrice Oxford Nanopore et la comparaison de méthodologies de priorisations de variations génétiques à l'aide de descriptions cliniques basées sur les termes HPO.

Mots clefs : Séquençage Haut-Débit, Analyse de données, Industrialisation, Bioinformatique.

Abstract

The number of rare diseases is assessed between 5000 and 8000 distinct pathologies. Individually, these diseases are rare, but together they represent a major health problem at the population level. Because of the limited number of patients with certain diseases and the lack of knowledge about them, diagnosis for patients is often delayed.

More than 80 % of rare diseases have a genetic origin, the vast majority of which are monogenic. The democratization of high-throughput sequencing technologies since 2005 has allowed the massive acquisition of genomic data from both healthy individuals and patients. However, establishing a diagnosis of a rare disease with current molecular biology technologies remains difficult.

The industrial-scale application of exome and genome sequencing analyses associated with the constant increase in medical knowledge represents a concrete hope to bring a diagnosis to the majority of patients concerned by a rare disease suspected to be genetic.

This thesis was particularly interested in the implementation of processes for constitutional genome-wide sequencing data analysis following industrial standards and bioinformatics good practices. Then, in the detection of copy number variations from exome sequencing data, but under-explored by many laboratories.

Finally, this thesis has allowed the development and the initiation of collaborations still active today. This has been concretized by a study on detection of small somatic variations in an in vitro model of cancer development linked to the cellular microenvironment, the detection of structural variations using the innovative Oxford Nanopore sequencing technology and the comparison of prioritization methodologies for genetic variations using clinical descriptions based on HPO terms.

Keywords : High Throughput Sequencing, Data analysis, Industrialization, Bioinformatics.