Sélection d'un pipeline bioinformatique pour la détection des variations du nombre de copies dans le cadre des prédispositions aux cancers

Robin Delacroix

▶ **To cite this version:**

Robin Delacroix. Sélection d'un pipeline bioinformatique pour la détection des variations du nombre de copies dans le cadre des prédispositions aux cancers. Sciences pharmaceutiques. 2020. dumas-02959761

HAL Id: dumas-02959761

https://dumas.ccsd.cnrs.fr/dumas-02959761

Submitted on 7 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MÉMOIRE
DU DIPLÔME D'ÉTUDES SPÉCIALISÉES
DE BIOLOGIE MÉDICALE

Soutenu le 24/09/2020
Par Mr Delacroix Robin
Né le 23 juin 1992 à Digne-les-Bains

Conformément aux dispositions du décret n°2003-76
du 23 janvier 2003, tenant lieu de

THÈSE

POUR LE DIPLÔME D'ÉTAT DE DOCTEUR EN PHARMACIE

----oOo----

Sélection d'un pipeline bioinformatique pour la détection des variations du nombre de copies dans le cadre des prédispositions aux cancers.

----oOo----

JURY :

Président : **Pr Catherine BADENS**

Membres : **Pr Hagay SOBOL**
Dr Audrey REMENIERAS
M Quentin DA COSTA

Membre invité : **Dr David SALGADO**

Faculté de Pharmacie
Aix•Marseille Université

27 Boulevard Jean Moulin – 13385 MARSEILLE Cedex 05
Tel. : 04 91 83 55 00 – Fax : 04 91 80 26 12

ADMINISTRATION :

Doyen :	Mme Françoise DIGNAT-GEORGE
Vice-Doyens :	M. Jean-Paul BORG, M. François DEVRED, M. Pascal RATHELOT
Chargés de Mission :	Mme Pascale BARBIER, M. David BERGE-LEFRANC, Mme Manon CARRE, Mme Caroline DUCROS, Mme Frédérique GRIMALDI, M. Guillaume HACHE
Conseiller du Doyen :	M. Patrice VANELLE
Doyens honoraires :	M. Patrice VANELLE, M. Pierre TIMON-DAVID,
Professeurs émérites :	M. José SAMPOL, M. Athanassios ILIADIS, M. Henri PORTUGAL, M. Philippe CHARPIOT
Professeurs honoraires :	M. Guy BALANSARD, M. Yves BARRA, Mme Claudette BRIAND, M. Jacques CATALIN, Mme Andrée CREMIEUX, M. Aimé CREVAT, M. Gérard DUMENIL, M. Alain DURAND, Mme Danielle GARÇON, M. Maurice JALFRE, M. Joseph JOACHIM, M. Maurice LANZA, M. José MALDONADO, M. Patrick REGLI, M. Jean-Claude SARI
Chef des Services Administratifs :	Mme Florence GAUREL
Chef de Cabinet :	Mme Aurélie BELENGUER
Responsable de la Scolarité :	Mme Nathalie BESNARD

DEPARTEMENT BIO-INGENIERIE PHARMACEUTIQUE
Responsable : Professeur Philippe PICCERELLE

PROFESSEURS

BIOPHYSIQUE	M. Vincent PEYROT M. Hervé KOVACIC
GENIE GENETIQUE ET BIOINGENIERIE	M. Christophe DUBOIS
PHARMACIE GALENIQUE, PHARMACOTECHNIE INDUSTRIELLE, BIOPHARMACIE ET COSMETOLOGIE	M. Philippe PICCERELLE

MAITRES DE CONFERENCES

BIOPHYSIQUE	M. Robert GILLI Mme Odile RIMET-GASPARINI Mme Pascale BARBIER M. François DEVRED Mme Manon CARRE M. Gilles BREUZARD Mme Alessandra PAGANO
GENIE GENETIQUE ET BIOTECHNOLOGIE	M. Eric SEREE-PACHA Mme Véronique REY-BOURGAREL
PHARMACIE GALENIQUE, PHARMACOTECHNIE INDUSTRIELLE, BIOPHARMACIE ET COSMETOLOGIE	M. Pascal PRINDERRE M. Emmanuel CAUTURE Mme Véronique ANDRIEU Mme Marie-Pierre SAVELLI
BIO-INGENIERIE PHARMACEUTIQUE ET BIOTHERAPIES PHARMACO ECONOMIE, E-SANTE	M. Jérémy MAGALON Mme Carole SIANI

ENSEIGNANTS CONTRACTUELS

ANGLAIS	Mme Angélique GOODWIN

DEPARTEMENT BIOLOGIE PHARMACEUTIQUE

Responsable : Professeur Françoise DIGNAT-GEORGE

PROFESSEURS

BIOLOGIE CELLULAIRE	M. Jean-Paul BORG
HEMATOLOGIE ET IMMUNOLOGIE	Mme Françoise DIGNAT-GEORGE Mme Laurence CAMOIN-JAU Mme Florence SABATIER-MALATERRE Mme Nathalie BARDIN
MICROBIOLOGIE	M. Jean-Marc ROLAIN M. Philippe COLSON
PARASITOLOGIE ET MYCOLOGIE MEDICALE, HYGIENE ET ZOOLOGIE	Mme Nadine AZAS-KREDER

MAITRES DE CONFERENCES

BIOCHIMIE FONDAMENTALE, MOLECULAIRE ET CLINIQUE	M. Thierry AUGIER M. Edouard LAMY Mme Alexandrine BERTAUD Mme Claire CERINI Mme Edwige TELLIER M. Stéphane POITEVIN
HEMATOLOGIE ET IMMUNOLOGIE	Mme Aurélie LEROYER M. Romaric LACROIX Mme Sylvie COINTE
MICROBIOLOGIE	Mme Michèle LAGET Mme Anne DAVIN-REGLI Mme Véronique ROUX M. Fadi BITTAR Mme Isabelle PAGNIER Mme Sophie EDOUARD M. Seydina Mouhamadou DIENE
PARASITOLOGIE ET MYCOLOGIE MEDICALE, HYGIENE ET ZOOLOGIE	Mme Carole DI GIORGIO M. Aurélien DUMETRE Mme Magali CASANOVA Mme Anita COHEN
BIOLOGIE CELLULAIRE	Mme Anne-Catherine LOUHMEAU

ATER

BIOCHIMIE FONDAMENTALE, MOLECULAIRE ET CLINIQUE	Mme Anne-Claire DUCHEZ
BIOLOGIE CELLULAIRE ET MOLECULAIRE	Mme Alexandra WALTON

A.H.U.

HEMATOLOGIE ET IMMUNOLOGIE	Mme Mélanie VELIER

DEPARTEMENT CHIMIE PHARMACEUTIQUE

Responsable : Professeur Patrice VANELLE

PROFESSEURS

CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION	Mme Catherine BADENS
CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES	M. David BERGE-LEFRANC
CHIMIE MINERALE ET STRUCTURALE – CHIMIE THERAPEUTIQUE	M. Pascal RATHELOT M. Maxime CROZET
CHIMIE ORGANIQUE PHARMACEUTIQUE	M. Patrice VANELLE M. Thierry TERME
PHARMACOGNOSIE, ETHNOPHARMACOGNOSIE	Mme Evelyne OLLIVIER

MAITRES DE CONFERENCES

BOTANIQUE ET CRYPTOGAMIE, BIOLOGIE CELLULAIRE	Mme Anne FAVEL Mme Joëlle MOULIN-TRAFFORT
CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION	Mme Catherine DEFOORT M. Alain NICOLAY Mme Estelle WOLFF Mme Elise LOMBARD Mme Camille DESGROUAS M. Charles DESMARCHELIER
CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES	M. Pierre REBOUILLON
CHIMIE THERAPEUTIQUE	Mme Sandrine ALIBERT Mme Caroline DUCROS M. Marc MONTANA Mme Manon ROCHE Mme Fanny MATHIAS
CHIMIE ORGANIQUE PHARMACEUTIQUE HYDROLOGIE	M. Armand GELLIS M. Christophe CURTI Mme Julie BROGGI M. Nicolas PRIMAS M. Cédric SPITZ M. Sébastien REDON
PHARMACOGNOSIE, ETHNOPHARMACOLOGIE	M. Riad ELIAS Mme Valérie MAHIOU-LEDDET Mme Sok Siya BUN Mme Béatrice BAGHDIKIAN

MAITRES DE CONFERENCE ASSOCIES A TEMPS PARTIEL (M.A.S.T.)

CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION	Mme Anne-Marie PENET-LOREC
CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES	M. Cyril PUJOL
DROIT ET ECONOMIE DE LA PHARMACIE	M. Marc LAMBERT
GESTION PHARMACEUTIQUE, PHARMACOECONOMIE ET ETHIQUE PHARMACEUTIQUE OFFICINALE, DROIT ET COMMUNICATION PHARMACEUTIQUES A L'OFFICINE ET GESTION DE LA PHARMAFAC	Mme Félicia FERRERA

A.H.U.

CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION	M. Mathieu CERINO

ATER

CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES	M. Duje BURIC

DEPARTEMENT MEDICAMENT ET SECURITE SANITAIRE
Responsable : Professeur Benjamin GUILLET

PROFESSEURS

PHARMACIE CLINIQUE	M. Stéphane HONORÉ
PHARMACODYNAMIE	M. Benjamin GUILLET
TOXICOLOGIE ET PHARMACOCINETIQUE	M. Bruno LACARELLE Mme Frédérique GRIMALDI M. Joseph CICCOLINI

MAITRES DE CONFERENCES

PHARMACODYNAMIE	M. Guillaume HACHE Mme Ahlem BOUHLEL M. Philippe GARRIGUE
PHYSIOLOGIE	Mme Sylviane LORTET Mme Emmanuelle MANOS-SAMPOL
TOXICOLOGIE ET PHARMACOCINETIQUE	Mme Raphaëlle FANCIULLINO Mme Florence GATTACECCA
TOXICOLOGIE GENERALE ET PHARMACIE CLINIQUE	M. Pierre-Henri VILLARD Mme Caroline SOLAS-CHESNEAU Mme Marie-Anne ESTEVE

A.H.U.

PHYSIOLOGIE / PHARMACOLOGIE	Mme Anaïs MOYON
PHARMACIE CLINIQUE	M. Florian CORREARD

ATER.

TOXICOLOGIE ET PHARMACOCINETIQUE	Mme Anne RODALLEC

CHARGES D'ENSEIGNEMENT A LA FACULTE

Mme Valérie AMIRAT-COMBRALIER, Pharmacien-Praticien hospitalier
M. Pierre BERTAULT-PERES, Pharmacien-Praticien hospitalier
Mme Marie-Hélène BERTOCCHIO, Pharmacien-Praticien hospitalier
Mme Martine BUES-CHARBIT, Pharmacien-Praticien hospitalier
M. Nicolas COSTE, Pharmacien-Praticien hospitalier
Mme Sophie GENSOLLEN, Pharmacien-Praticien hospitalier
M. Sylvain GONNET, Pharmacien titulaire
Mme Florence LEANDRO, Pharmacien adjoint
M. Stéphane PICHON, Pharmacien titulaire
M. Patrick REGGIO, Pharmacien conseil, DRSM de l'Assurance Maladie
Mme Clémence TABELE, Pharmacien-Praticien attaché
Mme TONNEAU-PFUG, Pharmacien adjoint
M. Badr Eddine TEHHANI, Pharmacien – Praticien hospitalier
M. Joël VELLOZZI, Expert-Comptable

Mise à jour le 23 janvier 2020

Remerciements

À **Madame la Professeure Catherine BADENS**, qui me fait l'honneur de présider ce jury de thèse. Soyez assuré de ma sincère reconnaissance et de mon profond respect.

À **Madame la Docteur Audrey REMENIERAS,** qui m'a fait l'honneur de diriger ce travail et d'encadrer ma thèse. Un grand merci pour ta confiance, pour tes conseils et pour les heures consacrées à relire et corriger ce travail.

À **Monsieur le Professeur Hagay SOBOL,** qui me fait l'honneur de siéger à ce jury de thèse. Je vous remercie d'avoir accepté de juger ce travail et vous prie de trouver ici l'expression de mon respect et de ma sincère reconnaissance.

À **Monsieur Quentin DA COSTA.** Pour m'avoir fait l'honneur de participer à ce jury, sois assuré de ma profonde gratitude et de mon estime. Je te remercie de l'attention que tu as bien voulu porter à mon travail.

À **Monsieur le Docteur David SALGADO,** merci d'avoir accepté de siéger à ce jury de thèse, et d'avoir participé avec bienveillance à ma formation au cours de l'année passée au sein du laboratoire INSERM U1251**.** J'ai vraiment apprécié de travailler avec toi. Sois certain que tu m'as énormément appris.

À Madame la Docteur Violaine BOURDON, un grand merci pour ta bonne humeur chaque jour. J'ai vraiment apprécié de travailler avec toi.

À Monsieur le Docteur Tetsuro NOGUCHI, merci beaucoup pour tout ce que tu m'as apporté cette année. Tu débordes de connaissances et j'ai adoré discuter de sujets et d'autres avec toi chaque jour. Tu as l'esprit très ouvert et c'est très appréciable.

À Madame la Docteur Samira FEKAIRI, merci chère voisine de m'avoir supporté tout au long de cette année. J'ai vraiment apprécié ton honnêteté et ta bonne humeur.

À toute l'équipe de bioinformatique du laboratoire INSERM U1251, BEROUD Christophe, SALGADO David, DESVIGNES Jean-Pierre, GUIEN Céline, GARIBAL Marc, GRATTEPANCHE Coralie, CORCUFF Mélanie. Cette année avec vous était tout à fait singulière. J'y ai découvert une autre ambiance de travail, à la fois sérieuse et familière. J'ai apprécié de travailler et de converser avec chacun d'entre vous. Vous m'avez tous apporté quelque chose. Je vous souhaite le meilleur.

Aux personnels du laboratoire d'oncogénétique moléculaire de l'IPC : Merci pour votre accueil, votre bonne humeur et votre gentillesse. Je vous suis reconnaissant d'avoir participé à ma formation.

Table des matières

Introduction **1**

Partie 1 : Généralités **3**

1. Les variations structurales génomiques 3
2. Les variations du nombre de copies 6
3. Mécanismes à l'origine des SV : 7
4. Détection des SV, méthodes de référence : 10
   - 4.1. Techniques de cytogénétique : 10
   - 4.2. Techniques de CGH-array : 12
   - 4.3. Techniques de biologie moléculaire - PCR 14
5. Séquençage haut débit et bioinformatique 16
   - 5.1. Séquençage haut débit 16
   - 5.2. Technologies NGS à lectures courtes ou « *short reads* » : 16
   - 5.3. Technologies NGS à lectures longues ou « long reads » : 22
     - 5.3.1. Les technologies SMRT : 23
       - 5.3.1.1. Pacific Biosciences « PacBio » : 23
       - 5.3.1.2. Oxford Nanopore Technologies : 24
     - 5.3.1. Les technologies synthétiques : 26
   - 5.4. Bioinformatique 28
     - 5.4.1. Pipeline Bioinformatique : 28
     - 5.4.2. Trimming qualité et filtrage : 29
     - 5.4.3. Étape d'alignement globale : 30
     - 5.4.4. Gestion des duplicats de PCR : 32
     - 5.4.5. Le réalignement des indels (réalignement local) : 33
     - 5.4.6. Le recalibrage des scores qualité : 34
     - 5.4.7. L'identification des variants (*Variant calling*) 35

5.4.8. Le Génome humain de référence : 35
6. Détection des CNV par séquençage haut débit (NGS) : 36
6.1. Approche « Read-Pairs » ou « Paired-end mapping » 37
6.2. Approche *Split-read* 40
6.3. Approche par analyse de la profondeur de lecture 42
6.4. Approche par assemblage *de novo* 44
Partie 2 : Logiciels disponibles et sélectionnés 45
1. ExomeDepth : 46
2. Delly : 49
3. DECoN : 51
4. Smoove/Lumpy : 52
5. Biomedical Genomics Workbench : 53
Partie 3 – Matériel et Méthode 55
1. Échantillons utilisés : 55
2. Composition du panel de gènes 56
3. CNV du set de référence 57
4. Gestion des pseudogènes : 60
Partie 4 : Résultats 61
1. Sélection et utilisation des logiciels : 61
1.1. Expérimentation des 5 logiciels sélectionnés 61
1.2. Etude des cas discordants entre les logiciels : 67
1.2.1. *BRCA2* - Délétion de l'exon 7 : 68
1.2.2. *MSH2* délétion des exons 8 à 16 et *BRCA1* délétion des exons 8 à 13 68
1.2.3. *PMS2* - Duplication des exons 11 et 12 70
1.2.4. *MSH6* - Délétion des exons 5 et 6 70
1.3. Optimisation de l'utilisation de ExomeDepth 71
1.3.1. Optimisation d'ExomeDepth concernant *PMS2* et *PM2CL* : 73

1.3.2. Limites analytiques : duplication des exons 11 et 12 de *PMS2* : 76
1.3.3. Phase 2 de l'optimisation d'ExomeDepth : 77
1.3.4. Automatisation de l'exécution logicielle 79
Discussion 80
Conclusion 82
Index des figures 85
Index des tableaux 90
Annexe 1 : Format FASTQ 91
1. Description 91
2. Structure du format FASTQ 91
Annexe 2 : Formats SAM et BAM 92
1. Description 92
2. Structure du format SAM 93
Annexe 3 : Format BED 94
1. Description : 94
2. Structure du format BED 94
Annexe 4 : Format FASTA 95
1. Description : 95
2. Structure du format FASTA : 95
Annexe 5 : Format VCF 96
1. Description : 96
2. Structure du format VCF : 97
Annexe 6 : Procédure 98
Références bibliographiques 103

Liste des abréviations

A :	Adénine
ACPA :	Analyse Chromosomique sur Puce à ADN
ADN :	Acide Désoxyribonucléique
ADNdb :	ADN double-brin
ADNsb :	ADN simple-brin
ARN :	Acide Ribonucléique
ASCII :	American Standard Code for Information Interchange
ATP :	Adénosine Triphosphate
BAC :	Bacterial Artificial Chromosome
BGW :	Biomedical Genomics Workbench
C :	Cytosine
CGH-array :	Comparative Genomic Hybridization array
CNP :	Copy-Number Polymorphism
CNV :	Copy-Number Variation
ddNTP :	di-désoxyriboNucléotide TriphosPhate
dNTP :	désoxyriboNucléoside TriphosPhate
DoC :	Depth Of Coverage
ED :	ExomeDepth
EDTA :	ÉthylèneDiamineTétraAcétique
FISH :	Fluorescence In Situ Hybridization
G :	Guanine
HBOC :	Hereditary Breast and Ovarian Cancer syndrome
Indels :	*Contraction de* Insertion-Deletion
Kb :	Kilobases
LCRs :	Low Copy Repeats
LRS :	Long-Read Sequencing
MAPH :	Multiplex Amplifiable Probe Hybridization
Mb :	Mégabases
MEI :	Mobile Element Insertion
MLPA :	Multiplex Ligation-dependent Probe Amplification
NAHR :	Non-Allelic Homologous Recombination
NGS :	Next-Generation Sequencing

NHEJ :	Non-Homologous End Joining
ONT :	Oxford Nanopore Technologies
PAC :	P1 Artificial Chromosome
PacBio :	Pacific Biosciences
pb :	Paire de bases
PCR :	Polymerase Chain Reaction
PEM :	Paired-End Mapping
QMPSF :	Quantitative Multiplex PCR of Short Fluorescent Fragments
qPCR :	quantitative Polymerase Chain Reaction
RP :	Read-Pair
SBL :	Sequencing by Ligation (*séquençage par ligation*)
SBS :	Sequencing By Synthesis (*séquençage par synthèse*)
SINE :	Short Interspersed Nuclear Element
SMRT :	Single Molecule Real Time Technology
SNP :	Single Nucleotide Polymorphism
SNV :	Single-Nucleotide Variant
SR :	Split-Read
SV :	Structural Variant *ou* Structural Variation
T :	Thymine
tg-NGS :	targeted-NGS
WES :	Whole Exome sequencing
WGS :	Whole Genome Sequencing
ZMW :	Zero-Mode Waveguides

« L'Université n'entend donner aucune approbation, ni improbation aux opinions émises dans les thèses. Ces opinions doivent être considérées comme propres à leurs auteurs. »

Introduction

Les Variations du Nombre de Copies (CNV) sont des altérations structurales chromosomiques qui consistent en une anomalie du nombre de copies d'un fragment du génome, en gain ou en perte. Leur taille est comprise entre 50b et 5Mb (1). Les CNV sont une composante importante de la diversité génétique (2–6), mais sont également à l'origine d'une proportion importante des pathologies humaines (7,8). Leur impact réel reste cependant méconnu et leur implication est probablement sous-estimée, car ils ne sont pas détectables par les techniques Sanger classiques. Les CNV d'une taille supérieure à 1Mb sont identifiés par les techniques de cytogénétique, ceux d'une taille inférieure par les techniques d'Analyse Chromosomique par Puce à ADN (ACPA) dont la résolution peut descendre jusqu'à quelques kilobases. Ce sont actuellement les techniques de référence pour la détection des CNV (9–11). Les délétions de tailles inférieures (moins d'un kilobase), mais trop étendues pour être détectées par les techniques Sanger (quelques dizaines de bases) peuvent être recherchées par d'autres techniques de biologie moléculaire comme l'amplification multiplex de sondes dépendantes d'une ligation ou Multiplex Ligation-Dependent Probe Amplification (MLPA), qui sont cependant des techniques coûteuses et contraignantes.

Il subsiste donc un fossé analytique concernant la détection des CNV dont la taille est comprise entre 50b et quelques kilobases, en découle une méconnaissance de leur impact biologique et de leur fréquence réelle.

Depuis une quinzaine d'années, les techniques de séquençage dites « de nouvelle génération » ont permis la généralisation des analyses génétiques. Les techniques de séquençage de nouvelle génération ou Next-Generation Sequencing (NGS) permettent une détermination rapide et précise de la séquence nucléotidique de l'ADN. À partir des données générées, et après analyse bioinformatique, ces techniques mettent en évidence les variations d'une seule base et les petites inversions/délétions, mais pourraient permettre également la détection des CNV (12–14). La détection des CNV dans les données de NGS nécessite cependant des processus bioinformatiques spécifiques surajoutés aux pipelines classiques.

Dans le cadre de la mise en place dans notre laboratoire d'un second pipeline de détection des CNV, quelques-uns des logiciels les plus fréquemment cités dans la littérature ont été testés. La performance de 5 d'entre eux sur nos données a été évaluée et leurs atouts et limites ont été discutés. Notre sélection a également dû tenir compte de la facilité d'intégration dans le pipeline

bioinformatique en place dans notre laboratoire, de leur facilité d'utilisation ainsi que de leur exigence en ressources humaines et économiques. Notre panel de routine contenant plusieurs gènes possédant une grande homologie avec des pseudogènes, nous avons ensuite procédé à des adaptations pour une plus grande efficacité analytique sur ces régions.

Ce travail peut ainsi servir de base à la mise en place d'un pipeline d'analyse des CNV à partir de données de NGS pour d'autres laboratoires utilisant des techniques de NGS par panel de gènes. Nous proposerons également quelques adaptations pour une plus grande efficacité sur les régions présentant des pseudogènes, sources d'erreurs et de faux positifs. Un des objectifs principaux de ce travail est donc de concevoir un processus applicable en routine et prenant en compte des limitations matérielles, humaines et économiques.

Partie 1 : Généralités

1. Les variations structurales génomiques

Lors de l'analyse les génomes humains, on ne peut que constater la similarité des séquences génomiques entre les différents individus autour du globe (15–17). Il est estimé que deux humains pris au hasard ont un génome similaire à 99,9% en séquence nucléotidique. C'est donc dans une fraction infime du génome que se situent les variations génétiques conduisant à la diversité phénotypique observable entre les individus et à la prédisposition aux maladies. Le spectre des variations génétiques chez l'humain s'étend de la simple paire de bases nucléiques à de grands réarrangements chromosomiques. Il est aujourd'hui admis que les génomes des membres d'une même espèce diffèrent davantage les uns des autres en raison de variations structurelles que de différences entre paires de bases. Ces Variations de Structure (SV) génomiques contribuent de manière considérable à l'hétérogénéité globale du génome humain. (3,18,19)

Les SV génomiques sont historiquement définies comme un ensemble d'altérations génomiques de taille supérieure à 1kb (20). Cette définition inclut les SV équilibrés (inversions et translocations), et les variations déséquilibrées (insertions/duplications et délétions) qui peuvent être pathologiques ou bénignes. Les SV sont donc des variations chromosomiques de différents types qui peuvent être quantitatives (délétions, duplications et insertions) ou qualitatives (translocations et inversions), par leur localisation et leur orientation (*Figure 1)*.

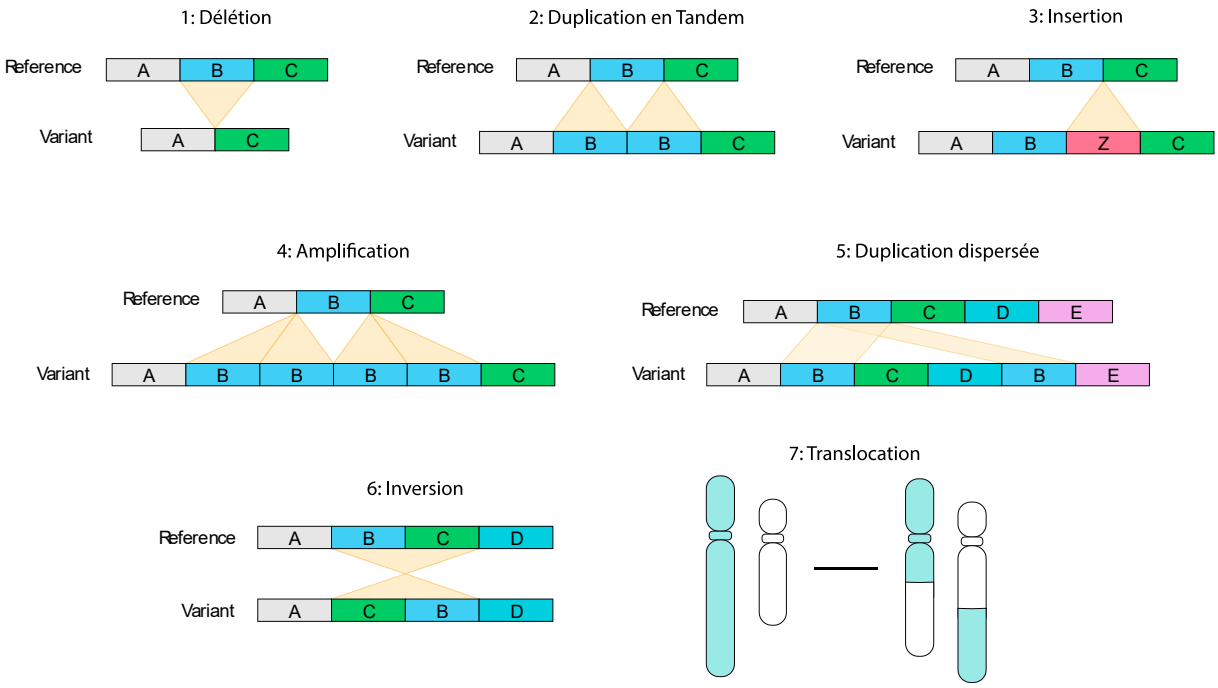


Figure 1 : Schématisation des différents types de variations structurelles. 1 - Délétion, 2 - Duplication en tandem (ou répétition en tandem) ; 3 - Insertion : 4 - Amplification ; 5 - Duplication dispersée (ou répétition dispersée) ; 6 - Inversion ; 7 - Translocation ;

Le choix historique du seuil de taille (> 1 kb) avait pour but de combler le vide descriptif existant entre les petites séquences répétées (comme les séquences répétées en tandem) et celles détectées à l'échelle d'un caryotype. Plus récemment la définition des SV a été révisée et l'on considère les SV comme étant des altérations de taille supérieure « à la taille d'un exon » ou encore une altération de taille supérieure à 50pb (5). C'est donc cette limite qui différentie aujourd'hui les SV des indels. Cette modification a été rendue nécessaire par la mise en place progressive en routine du séquençage génome entier (21) et illustre les changements globaux que subit la génétique avec l'avènement du séquençage haut débit.

En parallèle des SV, une catégorie est à prendre à part, ce sont les éléments mobiles ou Mobile Element Insertion (MEI) aussi appelés éléments transposables ou simplement transposons. Leur classification au sein des SV n'est pas bien établie et selon les sources ils sont inclus ou non. Les éléments transposables sont l'une des principales sources d'instabilité génomique par divers mécanismes complexes. Parmi eux, les séquences Alu sont les éléments les plus fréquents, on estime qu'elles composeraient environ 10% du génome humain répartis sur l'ensemble des chromosomes. Les séquences Alu sont des rétrotransposons non autonomes, elles font partie de la famille des petits éléments nucléaires intercalés ou SINE (*short interspersed nuclear element*)

avec les séquences MIR et MIR3. Les séquences Alu sont impliquées en pathologie humaine (22) comme le résume le tableau 1.

Gene	Position	Subfamily	Mechanism	Disease
ACE	Chr 17	*AluY*a5	Insertion	Alzheimer's disease
ALMS1	Chr 2	*AluY*a5	Insertion	Alström syndrome
BMPR2	Chr 2	*AluY* *AluS*	ARMD_NAHR NHEJ	Pulmonary arterial hypertension
CDSN	Chr 6	*AluS*	ARMD_NHEJ	Peeling skin disease
COL4A5	Chr X	*AluY*	Insertion	Alport syndrome
FA	Chr X	*AluY*	ARMD_NAHR	Fanconi anemia
GBA1	Chr 1	*AluS*x	ARMD_NAHR	Gaucher disease
GGA	Chr 17	*AluS*	ARMD_NAHR	Pomp disease
GLA	Chr X	*Alu*	Insertion mediated deletion	Fabry disease
MUTYH	Chr 1	*AluY*b8	Insertion	Breast cancer/gastric cancer
PMP22	Chr 17	*AluY*/*AluS*c	ARMD_NAHR	Charcot-Marie-Tooth disease
SOX10	Chr 22	*AluS*	FoSTes/MMBIR	Waardenburg syndrome type 4
SPAST	Chr 2 Chr 2	*AluY*/*AluS* *AluY*	FoSTes/MMBIR	Hereditary spastic paraplegia
SPG11	Chr 15 Chr 15	*AluY*/*AluS* *AluS*	ARMD_NAHR	Spastic paraplegias
STK11	Chr 19	*AluY*	ARMD_NAHR	Peutz-Jeghers syndrome

Tableau 1: Plusieurs exemples de séquences ALU impliquées en pathologie humaine. ***Source : Kim, S., Cho, C.-S., Han, K., & Lee, J. (2016). Structural Variation of Alu Element and Human Disease. Genomics & Informatics, 14(3), 70. doi:10.5808/gi.2016.14.3.70 (22).***

Comme les SV, les transposons imposent des défis bioinformatiques spécifiques (23). Ils peuvent provoquer des réarrangements de grande taille, leur localisation et leur nombre de copies sont variables. Ils sont ainsi à l'origine de problèmes majeurs lors de l'étape d'alignement comme l'illustre la figure 2 ci-dessous.

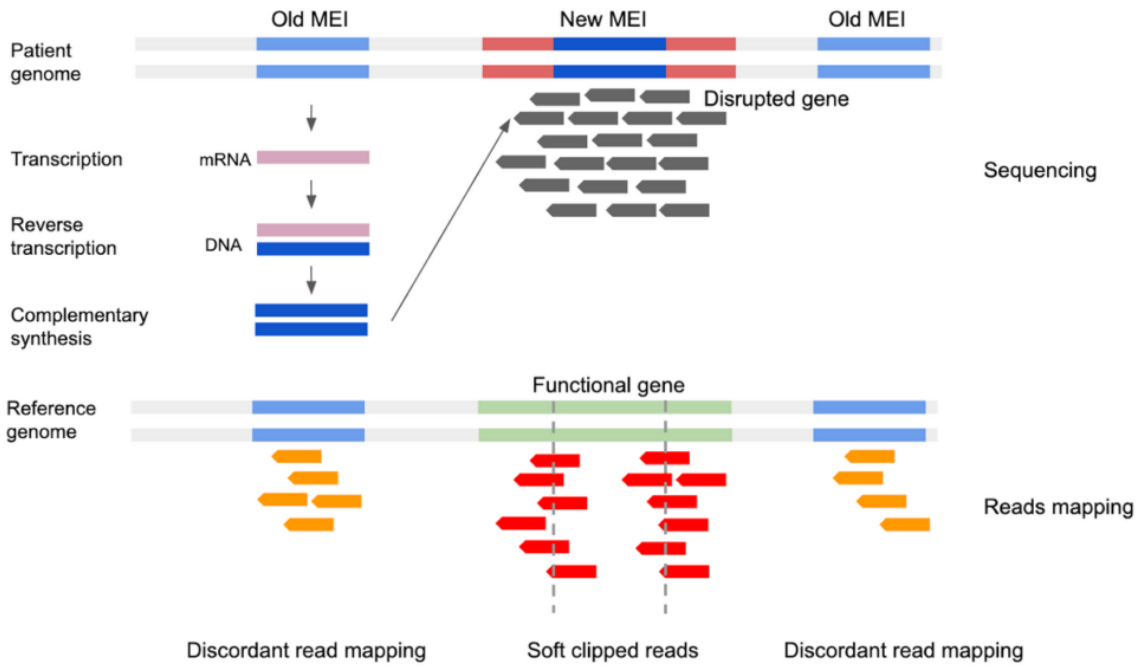


Figure 2: Difficultés d'alignement des lectures en lien avec la présence de transposons, la présence des transposons en plusieurs exemplaires, et de localisation variable, entraine des erreurs lors de l'application des logiciels d'alignement. *Source de la figure : Mobile element insertion (MEI) detection for NGS based clinical diagnostics – SeqOne (24)*

Les transposons sont des éléments impliqués en pathologie humaine et leur détection est un challenge auquel les bio-informaticiens et biologistes se heurtent chaque jour. Leurs moyens de détection ne sont pas standardisés et assurer leur détection en routine dans notre laboratoire n'est pas l'objectif de cette thèse qui porte sur les CNV. Cependant la recherche et l'interprétation de ces séquences Alu comme des variations structurelles équilibrées (inversions) seront probablement l'étape suivante dans l'évolution du pipeline bioinformatique de notre laboratoire.

2. Les variations du nombre de copies

Les CNV sont un sous-ensemble des SV. Ils sont définis comme un segment d'ADN de taille supérieure à 50pb dont le nombre de copies est différent par rapport au génome de référence (3,6). Ce sont donc des SV déséquilibrés.

Les CNV peuvent correspondre à un gain d'ADN (duplication et amplifications) ou à une perte (délétion) par rapport à un génome de référence. Les CNV fréquents, présents chez plus de 1 % de la population, sont appelés Copy Number Polymorphism (CNP). Un CNV peut contenir des

gènes, parties de gènes et/ou leurs régions de régulation. Ils peuvent également n'être constitués que de séquences non codantes (20,25,26).

Avec les avancées et la démocratisation des techniques d'Hybridation Génomique Comparative (CGH-array), l'importance des CNV en pathologie humaine a récemment été mise en lumière (27–32). Ils sont responsables d'un large spectre de maladies génétiques et l'étendue de leur implication en pathologie humaine est largement méconnue (33).

Bien qu'il ne soit pas nécessaire de connaitre les détails des mécanismes moléculaires qui sont à l'origine des SV pour assimiler le contenu de cette thèse, on peut s'y intéresser car ils sont une bonne illustration de la complexité des rouages moléculaires agissant sur le génome humain.

3. Mécanismes à l'origine des SV :

Trois principaux mécanismes conduisent à l'apparition de SV. De manière globale elles sont la conséquence de cassures chromosomiques suivies par un ou plusieurs recollements anormaux.

Le premier mécanisme est appelé recombinaison homologue non allélique (*non-allelic homologous recombination* ou NAHR). Il survient pendant la méiose ou la mitose et nécessite deux répétitions segmentaires (*low copy repeats ou LCR*) ou duplicons (34). En raison de leur haut degré de similarité de séquence, les copies non alléliques de répétitions segmentaires peuvent parfois être alignées en méiose ou en mitose à la place des copies aux positions alléliques habituelles. Ce phénomène est appelé mésappariement et peut entrainer des remaniements chromosomiques dans les cellules filles (Figure 3).

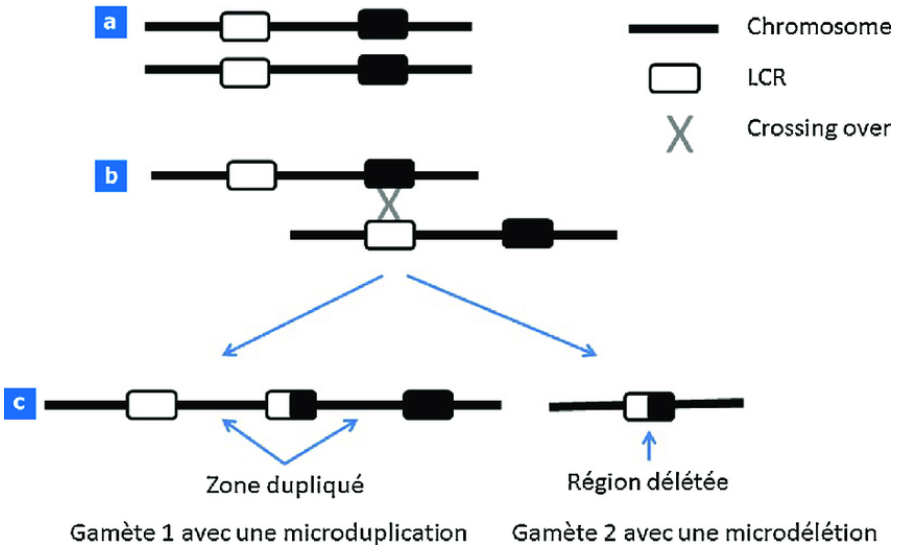


Figure 3 : Illustration du mécanisme de recombinaison allélique non homologue (NAHR) interchromatidien : a- Les rectangles représentent les régions hautement répétées ou LCR. Les rectangles blancs et noirs correspondent à deux LCR distincts qui possèdent un haut degré de similarité ; b – Entre les deux LCR peut se produire un mésappariement lors de la méiose ; c - Quand un échange (*crossing over*) non allélique se produit, on observe la création d'un gamète avec une microdélétion et un gamète avec une microduplication. *Source : Poisson A, Nicolas A, Sanlaville D, Cochat P, Leersnyder HD, Rigard C, et al. Le syndrome de Smith-Magenis, une association unique de troubles du comportement et du cycle veille/sommeil.*

Le second mécanisme est appelé jonction d'extrémités non homologues ou NHEJ (*Non-Homologous End Joining*). Il s'agit d'un un outil moléculaire de réparation de l'ADN ayant pour objectif la réparation de cassures double brin. C'est un mécanisme dit non conservatif, car il ne restaure pas la séquence initiale de l'ADN, il assure seulement la continuité d'un ADN endommagé par une cassure double brin. Cette réparation peut ainsi conduire à une modification de l'information génétique et généralement à une délétion (Figure 4).

C'est un phénomène décrit chez tous les organismes, des bactéries jusqu'aux mammifères. Il est couramment utilisé pour réparer les cassures physiologiques (ex. : recombinaisons des régions VDJ dans le système immunitaire) ou pathologiques induites par les radiations ionisantes ou les espèces réactives de l'oxygène.

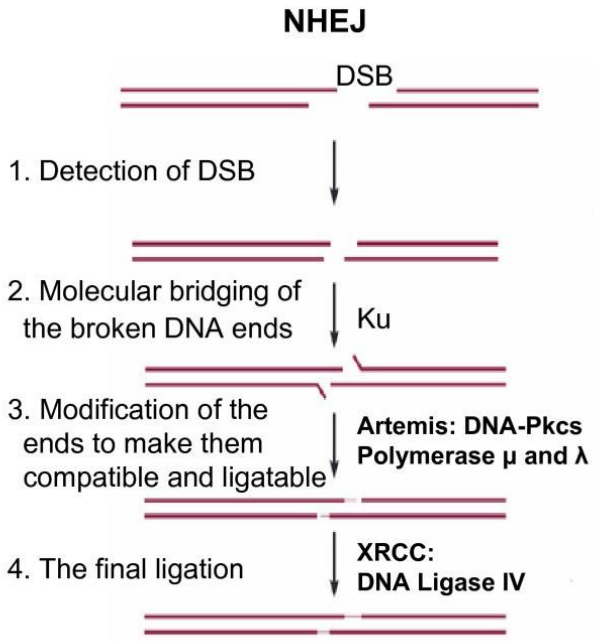


Figure 4: Illustration des 4 étapes des NHEJ : 1- identification d'une cassure double-brins ; 2- Création de ponts moléculaires entre les deux terminaisons cassées ; 3- Légère modification des terminaisons pour en augmenter la compatibilité (digestion enzymatique) ; 4- Ligation des deux brins. *Source : Gu, Wenli & Zhang, Feng & Lupski, James. (2008). Mechanisms for human genomic rearrangements. PathoGenetics. 1. 4. 10.1186/1755-8417-1-4.*

Le dernier mécanisme que nous citerons sans toutefois le détailler est désigné par l'acronyme FoSTeS (*Fork Stalling and Template Switching)*. Il n'est pas en lien avec une cassure double brin et entraine la formation de réarrangements non récurrents complexes.

Ces 3 mécanismes sont parfois groupés en 2 catégories : les recombinaisons homologues (NAHR) et non homologues (NHEJ et FoSTeS).

Il est probable que tous les phénomènes biologiques conduisant à la création et à la transmission des SV ne soient pas encore découverts. La génétique est un domaine très évolutif qui implique le développement constant de nouvelles techniques d'analyse s'appuyant sur les technologies disponibles. C'est pourquoi nous ferons une brève description des techniques de référence pour l'étude des SV.

4. Détection des SV, méthodes de référence :

4.1. Techniques de cytogénétique :

La détection des anomalies chromosomiques a commencé avec la possibilité d'observer les différents chromosomes en microscopie dans les années 1950-1960 (Figure 5). Deux grands types de mutations ont alors été distinguées : les anomalies de nombres (aneuploïdies et polyploïdies) et les anomalies de structure (équilibrées et déséquilibrées).

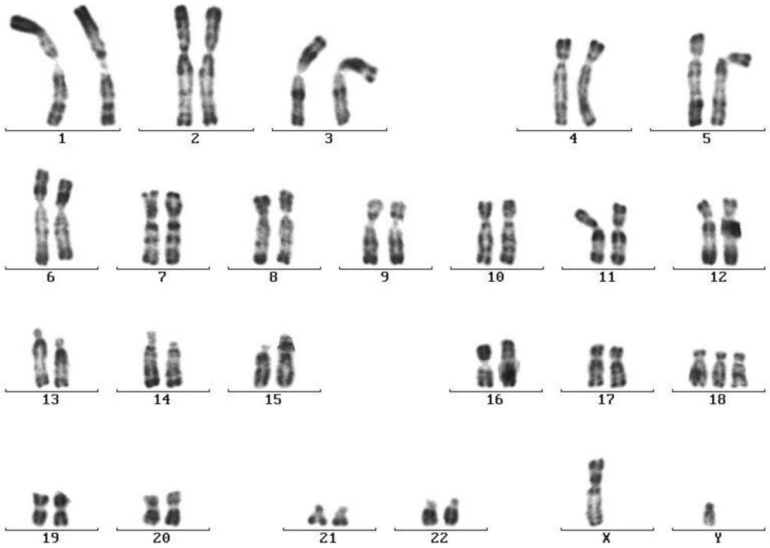


Figure 5: Résultat d'un caryotype humain masculin normal après tri des chromosomes en paires et de 1 à 22 (+X et Y). ***Source : Laboratoire de cytogénétique du Centre Hospitalier Universitaire de REIMS – 04/12/1999.***

Au fil du temps, les avancées en matière d'optique et de « banding » des chromosomes ont permis une lente amélioration des techniques de cytogénétique. De l'observation d'anomalies à l'échelle du chromosome entier (aneuploïdies, grands réarrangements, chromosomes acrocentriques) jusqu'à l'observation d'anomalies de plus petite taille : translocations, duplications et délétions de taille modeste (>3Mb) (6) (Figure 6).

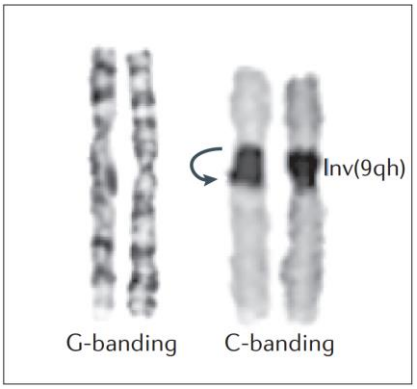


Figure 6: Le Giemsa banding (G-banding) utilise un colorant chimique qui révèle des bandes sombres sur les chromosomes métaphasiques. Ces méthodes permettent de détecter des anomalies structurales de grande taille (> 3Mb), on peut voir ici une inv(9qh). ***Source : Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97. doi :10.1038/nrg1767.***

En plus d'une sensibilité limitée aux anomalies de plusieurs mégabases, le caryotype est une technique longue et coûteuse. Elle n'est pas automatisable et nécessite l'intervention de l'œil du cytogénéticien ce qui en fait une technique dépendante de la performance du lecteur et soumise à interprétation.

Dans le même temps, le perfectionnement des techniques de Fluorescence In Situ Hybridization (FISH) a permis un gain de sensibilité et la possibilité d'observer au microscope des anomalies plus petites. Ces techniques consistent en l'utilisation de sondes, marquées par un fluorochrome. Elles permettent, après hybridation avec l'ADN fixé sur une lame de verre, d'observer des colorations au microscope à fluorescence, et ainsi d'observer les variations structurelles (Figure 7).

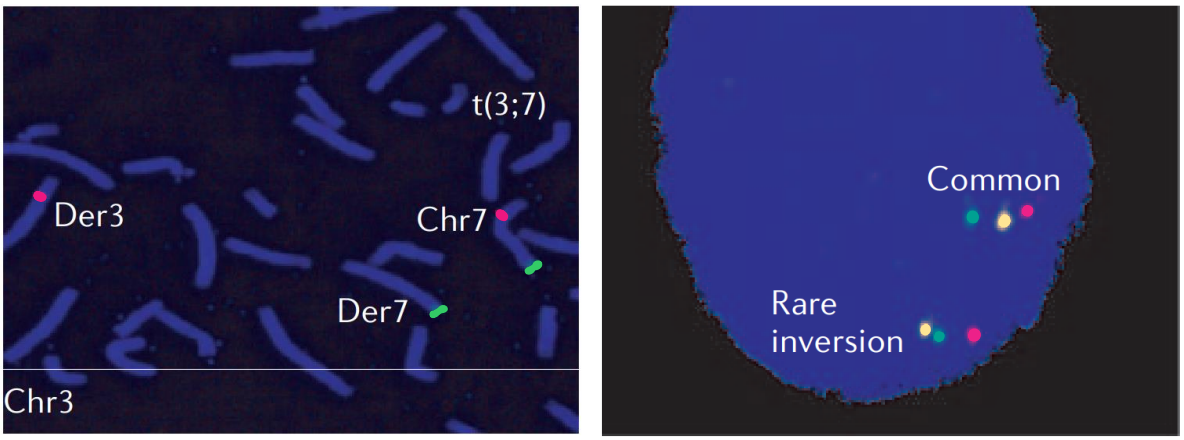


Figure 7: Sur l'image de gauche, une FISH sur chromosomes en métaphase sur laquelle on observe une translocation t(3;7) cryptique, un exemple d'anomalie structurelle équilibrée. Sur l'image de droite, une FISH sur noyaux en interphase sur laquelle on observe une micro-inversion 7p22 de 700 kb. C'est un exemple de SV bénigne (polymorphisme).

Les techniques de FISH permettent de détecter des remaniements chromosomiques de moins de 2Mb (avec une taille des sondes comprise entre 100 et 200kb). Cependant ce sont des techniques nécessitant des sondes spécifiques de la région à analyser ce qui implique une connaissance de l'anomalie recherchée. Ce n'est pas le cas des techniques de peinture chromosomiques, « *spectral karyotyping* » qui permettent de détecter des SV sans ciblage, mais qui sont très coûteuses. Elles nécessitent une étape de culture cellulaire et leur interprétation est difficile (Figure 8).

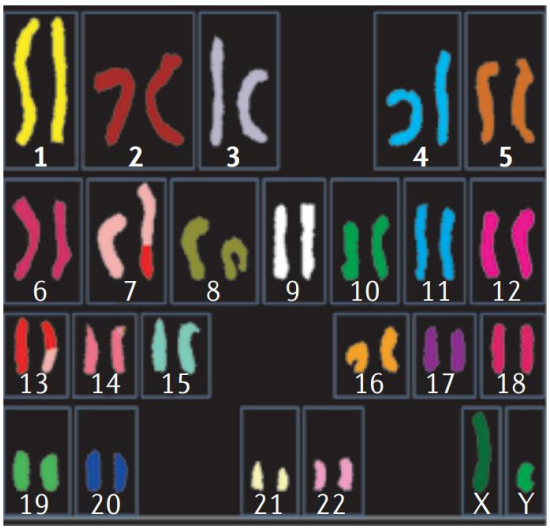


Figure 8: La peinture chromosomique (spectral karyotyping) est utile pour détecter les réarrangements interchromosomiques. Ici une translocation t(7;13). ***Source : Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97. doi:10.1038/nrg1767***

Pendant de nombreuses années la cytogénétique a été la référence absolue pour la détection et la caractérisation des anomalies de structure. Elle possède cependant certaines limites, dont la principale est sa résolution limitée. C'est pourquoi l'hybridation génomique comparative, technique développée dans les années 1990, a rapidement trouvé sa place dans le diagnostic de routine des SV.

4.2. Techniques de CGH-array :

Actuellement les techniques de référence pour la détection des SV génomiques déséquilibrées sont des techniques basées sur la CGH. Les plus utilisées sont les analyses chromosomiques par puce à ADN (ACPA ou CGH-*array*) (25,37–39). Elles sont classées parmi les techniques de cytogénétique moléculaire avec les techniques de FISH (*cf. Partie 1 - Chapitre 4.1*) et appartiennent au domaine de la cytogénétique.

L'ACPA consiste à réaliser une hybridation compétitive entre un patient et un témoin contrôle. C'est la taille des fragments d'ADN de la puce qui conditionne la définition de la technique. Dans le cas des puces à haute densité plus d'un million de « spots » contenant des fragments d'ADN différents peuvent être rassemblés sur une même lame de verre. Cela permet une couverture pangénomique et une résolution de l'ordre de quelques kilobases, on parle parfois de caryotype moléculaire (40,41).

La figure 9 ci-dessous illustre le résultat d'une puce à ADN, chaque cercle/hexagone représente un « spot » de la puce. Sur cette dernière sont fixées des milliers de copies d'un même fragment d'ADN, de position et de séquence connues.

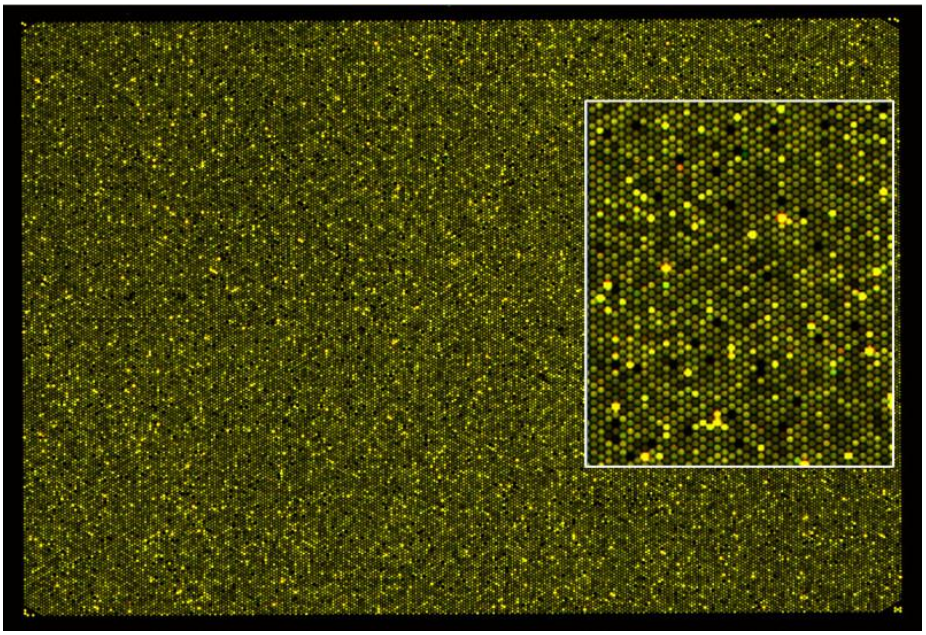


Figure 9 : Illustration des résultats d'une puce à ADN : le rectangle blanc représente un agrandissement. ***Source : Ahn JW, Coldwell M, Bint S, Mackie Ogilvie C. Array comparative genomic hybridization (array CGH) for detection of genomic copy number variants. J Vis Exp. 2015*.**

Les résultats d'une technique de CGH-array sont donc une multitude de rapports de fluorescence liés à des fragments d'ADN de séquences connues.

Ainsi, dans les techniques de CGH-array comme en biologie moléculaire, les ressources bioinformatiques sont indispensables, tant pour l'analyse des données (puces haute-densité) que pour leur interprétation (bases de données). Les progrès bioinformatiques sont intimement liés à l'évolution des puces ADN. Des puces avec une résolution toujours plus fine impliquent des moyens bioinformatiques toujours plus puissants.

Il n'existe pas de règles strictes pour le traitement bioinformatique des données issues de CGH-array et chaque logiciel commercialisé est différent, on notera toutefois que plusieurs étapes sont généralement partagées :

- Analyse des images.
- Normalisation spatiale (42) et/ou interéchantillon.
- Traitement des doublons.
- Identification des régions avec une perte ou un gain de matériel et des points de cassures (43,44).

Tout au long de l'analyse, plusieurs paramètres informatiques de contrôle qualité sont extraits et interprétés.

Une fois l'ensemble de ces étapes réalisées, on obtient une valeur théorique de rapport de signaux de fluorescence qui correspond à la quantité d'ADN pour chaque segment chez le patient par rapport à celle de l'échantillon de référence. Les puces à ADN sont donc des méthodes rapides, sensibles et semi-automatisables qui ont rapidement trouvé leur place en routine pour la détection des SV déséquilibrés. Leur résolution est intermédiaire entre la cytogénétique classique et la biologie moléculaire. Cependant ce sont des techniques qui ont un coût élevé, qui nécessitent l'achat d'un appareillage spécifique et un excellent niveau d'expertise.

4.3. Techniques de biologie moléculaire - PCR

En génétique biologique les techniques sont classiquement séparées entre les domaines de la cytogénétique (*cf. Partie 1- Chapitres 4.1 et 4.2*) et de la biologie moléculaire. Avec l'évolution des méthodes, la frontière entre ces deux domaines se fait de plus en plus mince. Les techniques de PCR décrites ci-après sont des techniques de biologie moléculaire.

Pour le criblage de régions ciblées du génome, les tests les plus performants sont principalement basés sur la Réaction en Chaîne par Polymérase ou Polymerase Chain Reaction (PCR). Ils ne sont pas adaptés au multiplexage (étude de plusieurs régions simultanées) c'est pourquoi plusieurs méthodes dérivées ont été développées. La plus populaire est la MLPA (49), illustrée dans la figure 10 ci-dessous.

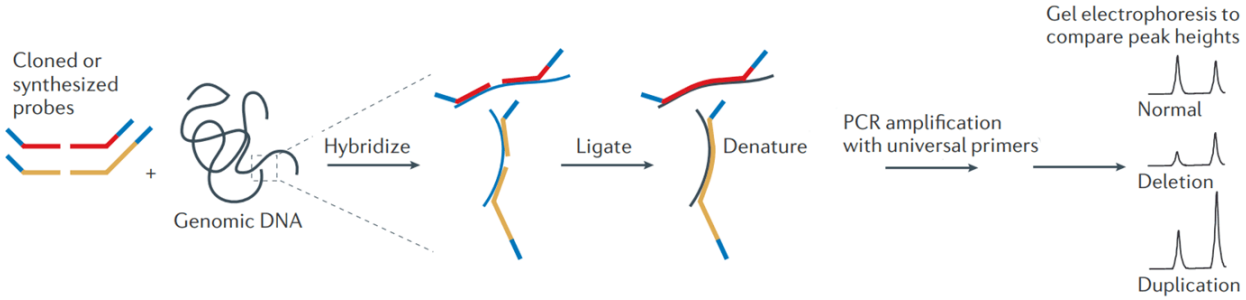


Figure 10: Illustration d'une méthode de PCR multiplexe pour la détection des CNV, Technique Multiplex Ligation-dependent Probe Amplification (MLPA). ***Source : Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97. doi:10.1038/nrg1767***

S'achève ici la description des techniques « classiques » de détection des SV. Elles sont toujours les techniques de référence. Le tableau 2 récapitule leurs limites analytiques.

Méthode	Détecte les translocations ?	Détecte les inversions ?	Détecte les CNV > 50kb	Détecte les CNV de 1-50kb	Détecte les CNV < 1kb
Technique sans ciblage					
Caryotype	Oui (> 3Mb)	Oui (>3Mb)	Oui (>3Mb)	Non	Non
CGH-array (nucléotide)	Non	Non	Oui	Oui	Non
Technique ciblée					
MLPA	Non	Non	Oui	Oui	Oui
FISH	Oui	Oui	Oui	Oui	Oui

Tableau 2: Tableau récapitulatif des limites analytiques des techniques de cytogénétique et des techniques basées sur la PCR-quantitative citées.

L'excellent recul dont les biologistes bénéficient sur ces techniques est un atout crucial pour un rendu biologique qui peut avoir d'importantes conséquences cliniques. Cependant, en parallèle de l'utilisation et du perfectionnement des techniques de cytogénétique, les techniques de séquençage ont amorcé une révolution scientifique d'envergure.

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides. Ces techniques permettent de détecter des anomalies à l'échelle de l'acide nucléique et de quelques acides nucléiques. Elles sont utilisées en routine dans presque tous les domaines médicaux et de recherche. Le séquençage de l'ADN est un domaine très actif dans lequel les révolutions technologiques s'enchaînent frénétiquement depuis la conclusion de l'Human Genome Project le 14 avril 2003. Aujourd'hui les approches scientifiques permettant le séquençage de l'ADN

sont très diverses et de nouvelles apparaissent régulièrement. Cependant concernant la détection des SV, les techniques de séquençage ne sont pas suffisamment performantes pour remplacer complètement les techniques de cytogénétique et de PCR.

5. Séquençage haut débit et bioinformatique

5.1. Séquençage haut débit

Les techniques de séquençage à haut débit, aussi appelées techniques de séquençage de nouvelle génération, font partie du domaine de la biologie moléculaire. Ces techniques permettent le séquençage de l'ADN en grands volumes, en un temps très court et à un coût moins important qu'avec la méthode manuelle de Sanger. Le séquençage est réalisé au sein d'automates dont le fonctionnement repose sur différentes technologies brevetées que l'on peut séparer en deux groupes : les technologies à lectures courtes (*short reads*) et les technologies à lecture longue (*long reads*).

Les technologies à lectures courtes sont de loin les plus répandues. Elles permettent la synthèse de lectures d'une taille comprise entre 35pb et 700pb. Les technologies à lectures longues outrepassent cette limite de taille mais sont considérablement plus coûteuses. Leur débit est cependant bien inférieur, ce qui limite leur utilisation.

5.2. Technologies NGS à lectures courtes ou « *short reads* » :

Au sein des technologies NGS short reads, on distingue deux écoles principales, celles basées sur le séquençage par ligation (SBL), et celles basées sur le séquençage par synthèse (SBS).

Ces deux approches prennent en entrée des fragments d'ADN de taille homogène qui présentent à leurs extrémités de courtes séquences nucléotidiques appelées « étiquettes » (*tags*).

Ces tags ont plusieurs fonctions essentielles, ils permettent :

- La conservation de l'information de l'origine du fragment tout au long du séquençage, alors que plusieurs échantillons sont analysés simultanément, on parle de multiplexage. Le tag assurant cette fonction est appelé index ou code-barre.
    - Les codes-barres sont de petits oligonucléides uniques pour chaque patient au sein d'une même expérience de séquençage.
- La fixation du fragment sur un support solide (billes ou *flow-cell* selon la technologie). Le tag assurant cette fonction est appelé adaptateur, il est différent pour chaque plateforme de séquençage.

- L'initialisation de la réaction de séquençage.

On notera que parfois les index sont considérés comme faisant pleinement partie des adaptateurs. La figure 11 ci-dessous est une illustration des adaptateurs (technologie à lectures courtes Illumina).

Figure 11: Composition des adaptateurs Illumina (modèle *paired-end*). SP : Amorce (SP+P5/P7 = 95pb) - P5 : Adaptateur commun - P7 : Adaptateur commun contenant l'index – Index : Fragment spécifique de chaque patient (6pb) – Insert : Fragment d'intérêt. Source : Jacques V. L'intégration du séquençage de nouvelle génération dans le diagnostic médical : application aux leucémies aiguës myéloïdes et syndromes myélodysplasiques (51).

L'étape de départ commune et qui permet la création de ces fragments et est appelée « préparation des libraires ». Différentes approches existent, nous citerons la préparation de librairies par amplification et la préparation de librairies par fragmentation. La Figure 12 ci-dessous illustre ces deux approches de préparation des librairies.

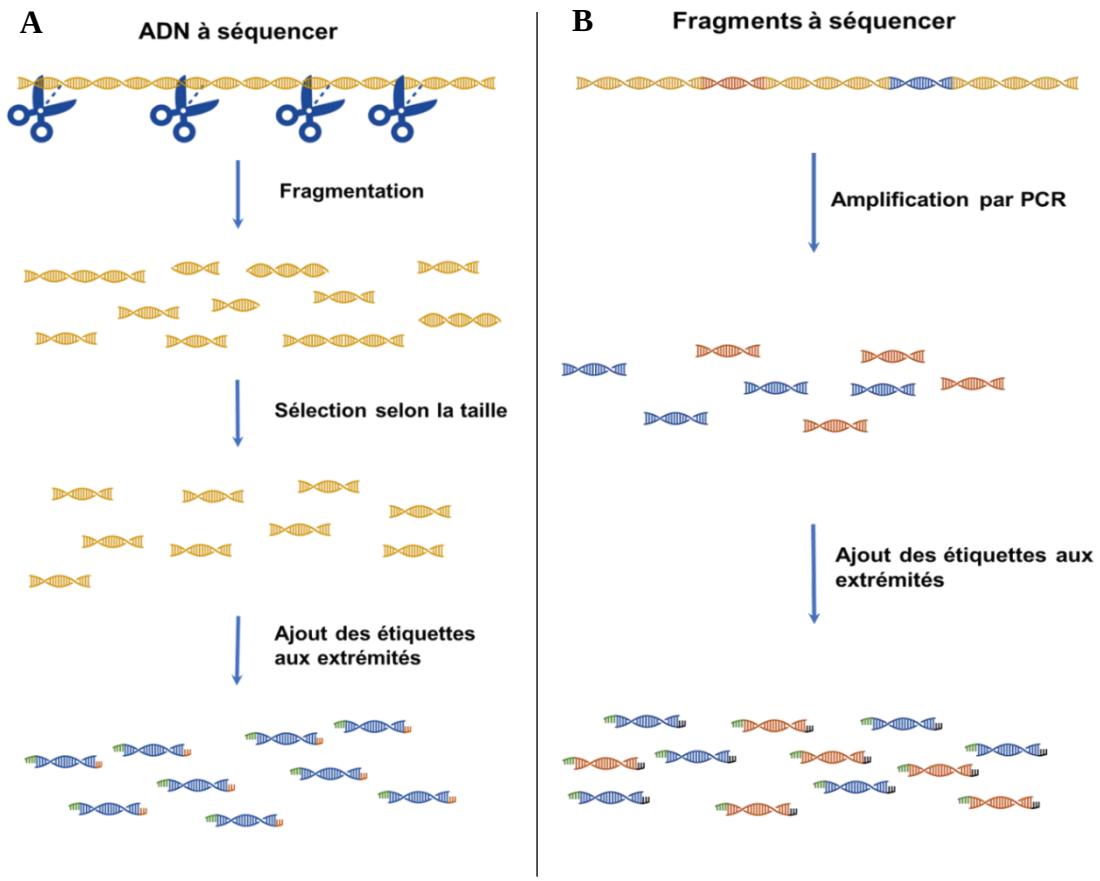


Figure 12: A- La préparation de librairies par fragmentation : le matériel génétique est découpé en fragments de différentes tailles par un procédé qui peut être mécanique, chimique ou enzymatique. Une sélection des fragments par la taille est réalisée. La dernière étape consiste en l'ajout d'étiquettes aux extrémités. B- La préparation de librairies par amplification : des régions d'intérêt sont amplifiées par PCR grâce à des amorces spécifiques. Des étiquettes sont ajoutées aux extrémités des amplicons par ligation ou par PCR. *Source de la figure : https://www.biomnigene.fr/.*

La préparation des librairies par fragmentation nécessite quelques précisions. Son principe est qu'après fragmentation de l'ADN de départ il est improbable que deux fragments identiques aient été produits. Chaque fragment généré est considéré comme unique. À la suite de cette fragmentation et après ligation aux amorces, une étape de PCR courte (moins de 6 cycles) est effectuée. Cette étape crée intentionnellement plusieurs clones de chaque fragment d'ADN. Elle est nécessaire pour obtenir une sensibilité suffisante pour l'étape de séquençage. Il sera donc inévitable de séquencer plusieurs clones d'un même fragment d'ADN, on parle de duplicats de PCR *(cf. Partie 1 -Chapitre 5.4.4).*

Une fois la préparation des librairies terminée, la seconde étape du séquençage est également une amplification par PCR, généralement sur surface solide, soit au sein d'une émulsion (PCR par émulsion) à la surface de billes individuellement emprisonnées dans des gouttelettes (Figure 13), soit à la surface d'une *flowcell* (« Bridge » ou « Cluster » PCR) (Figure 14).

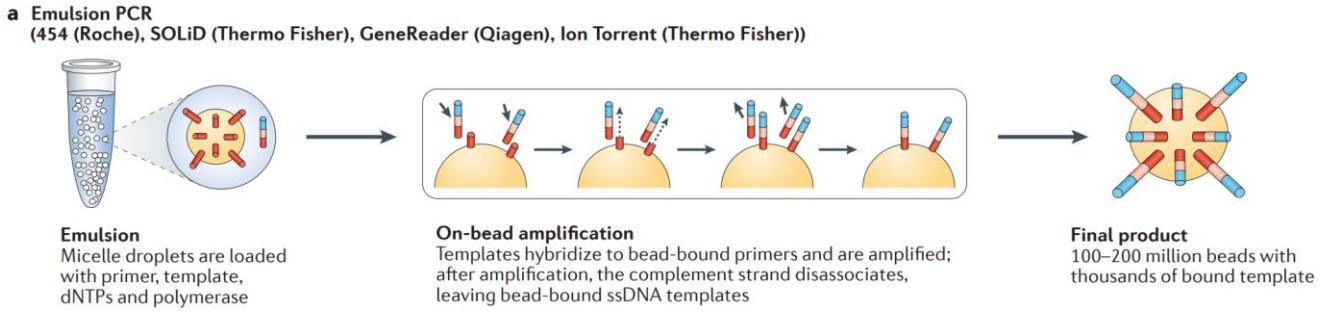


Figure 13 : Illustration de la technologie de PCR par d'émulsion. L'étape de PCR est effectuée sur une bille, à l'intérieur d'une micelle, recouvrant chaque bille de milliers de copies de la même séquence d'ADN. Source : ***Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).***

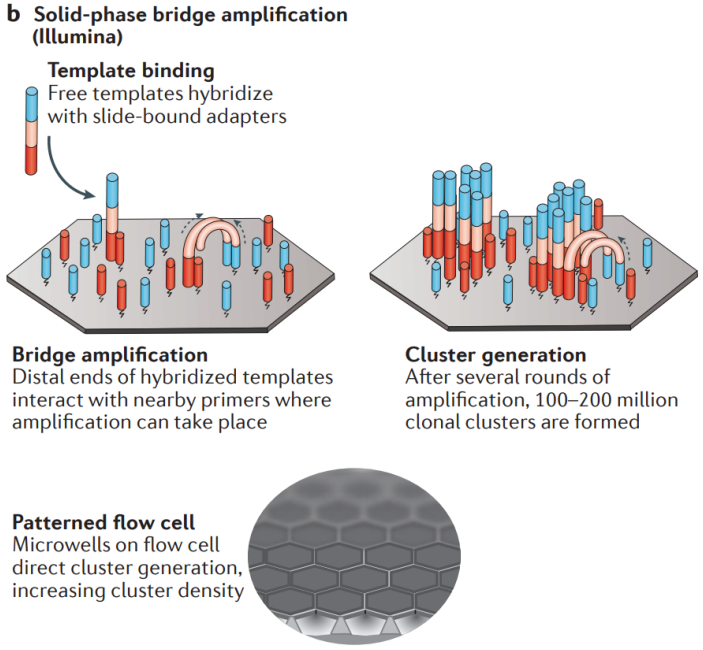


Figure 14 : Illustration de la technologie « bridge PCR » d'Illumina. ***Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).***

Cette étape d'amplification permet donc d'obtenir des milliers de copies d'un fragment de librairie sur une surface très restreinte. Le signal généré par ce « *cluster* » de fragments d'ADN identiques pourra ainsi être significativement distingué du bruit de fond.

La dernière étape est en réalité la seule étape de séquençage au sens strict du terme. Elle consiste à identifier la séquence nucléotidique par la détection et l'interprétation d'un signal dont la nature dépend de la technologie utilisée :

- Utilisation de nucléotides fluorescents : Illumina, SOLiD (ThermoFisher).
- Libération d'adénosine triphosphate : 454 Pyroséquençage (Roche).
- Détection de variations de protons H^+ (Proton Thermofisher)

La distinction entre le séquençage par ligation et le séquençage par synthèse est liée à la technologie utilisée lors de cette étape. Le séquençage par synthèse regroupe un ensemble de technologies qui emploient une ADN-polymérase lors de l'étape finale de séquençage, par opposition aux technologies de ligation qui n'en utilisent pas.

Le séquençage paired-end : Un point nécessite tout particulièrement d'être détaillé. Il s'agit du principe et des implications bioinformatiques du séquençage *paired-end*.

Le séquençage « *paired-end* » est l'évolution du séquençage « *single-end* ». La quasi-totalité des technologies à lectures courtes est utilisée en paired-end (illumina / Ion-torrent). La différence entre *single-end* et *paired-end* ne se situe pas dans une modification des automates ou de la réaction de séquençage mais dans la préparation des librairies et le choix des amorces.

Le paired-end consiste à séquencer les lectures par paires. Dans une paire les deux lectures sont alors séparées par une distance connue (taille de l'insert). On distingue deux distances pour une paire de lectures (Figure 15).

- La distance « *inner* » est la distance qui sépare les deux lectures, la taille des lectures est exclue. Cette taille correspond à la taille de l'insert.
- La distance « *outer* » inclut la taille des lectures dans le calcul de la distance.

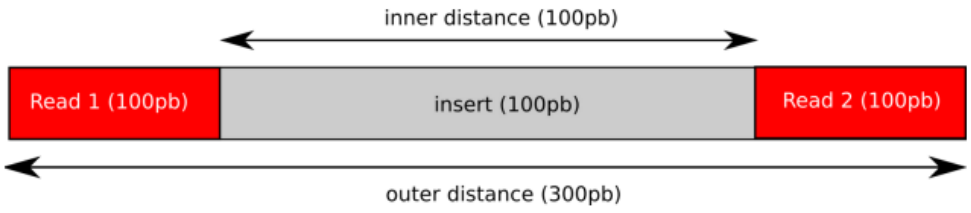


Figure 15: Illustration des distances calculables au sein d'une paire de lectures paired-end. Source : *Ségolène CABOCHE, Gaël EVEN – Paired-End versus mate-pair v1.0 17.02.2012 – Ressource en ligne - http://www.biorigami.com/*

Les deux lectures peuvent être non chevauchantes et séparées par une distance fixe, ou chevauchantes quand la taille de la totalité de la librairie est inférieure à la somme des tailles des lectures (*forward* et *reverse*) séquencées.

Chaque fragment d'ADN est donc séquencé à partir des deux extrémités et la distance entre les deux lectures d'une paire est connue. Cette conformation permet de gagner grandement en précision quand il s'agit d'aligner ensuite ces lectures sur la référence au sein de régions répétées (*cf. Partie 1 - Chapitre 5.4.3*). Ce paramètre est au cœur du principe de détection des CNV par l'approche *Paired-End Mapping* (*cf. Partie 1 - Chapitre 6.1*). La figure 16 ci-dessous est une illustration du séquençage paired-end.

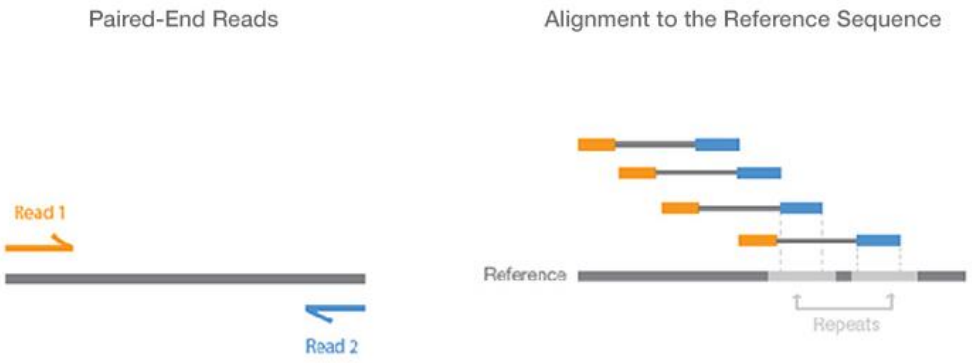


Figure 16: Illustration du principe du séquençage Paired-End tel qu'appliqué dans les technologies Illumina. *Source : « Single-read sequencing - Understand the key differences between these sequencing read types », Ressource en ligne : https://emea.illumina.com/.*

Attention toutefois à une confusion courante dans les termes utilisés, le « paired-end » ne doit pas être confondu avec le « mate-pair ». Le mate-pair est un autre processus de préparation des librairies permettant une circularisation des fragments d'ADN qui génère des paires de lectures avec un insert très long (2 à 20kpb), *a contrario* dans le paired-end les deux lectures d'une paire sont peu éloignées (moins de 600pb).

Le séquençage à lecture courte est un moyen fiable et très efficace de séquencer l'ADN. Ce sont des technologies relativement anciennes et répandues sur lesquelles nous avons un recul suffisant. Cependant le point commun et la principale limite de l'ensemble des technologies à lectures courtes est qu'elles nécessitent une étape d'amplification qui expose au biais de PCR.

5.3. Technologies NGS à lectures longues ou « long reads » :

Actuellement la majorité des laboratoires de biologie moléculaire de routine sont équipés avec des automates Illumina (technologie dominante) basés sur une technologie à lectures courtes sur laquelle nous avons un excellent recul. Cependant les technologies à lectures longues, plus récentes, tendent à se perfectionner et pourraient devenir la référence dans les années à venir.

À l'opposé des lectures courtes, les technologies de séquençage à lectures longues (*long-read sequencing* ou LRS) permettent d'obtenir des lectures de plusieurs kilobases successives (> 1000 pb). Elles peuvent ainsi couvrir des régions complexes ou répétitives avec une seule lecture continue. Cela permet d'éliminer les ambiguïtés dans les positions ou la taille des éléments génomiques. Elles sont également utiles pour les études transcriptomiques en couvrant des transcrits d'ARN messagers entiers et de manière continue, permettant l'étude de la connectivité précise des exons et de discerner les isoformes des gènes.

À ce jour, il existe deux principales approches technologiques à lecture longue : les approches de séquençage en temps réel de molécules uniques (*real-time long-read sequencing*) abrégées SMRT (*Single Molecule Real-time Technology*) également parfois appelées technologies LRS « vraies » et les approches synthétiques (*synthetic long-read sequencing*).

Les technologies SMRT diffèrent des approches à lecture courte en ce qu'elles ne reposent pas sur une population clonale de fragments d'ADN amplifiés pour la génération d'un signal détectable. Elles permettent de séquencer sans préamplifier. De plus elles ne nécessitent pas un cycle de réaction pour chaque ajout de désoxyriboNucléotide TriPhosphate (dNTP).

Les deux technologies de SMRT les plus remarquables sont celles développées respectivement par Pacific Biosciences (PacBio) (52) et Oxford Nanopore Technologies (ONT) (Figures 17 et 18). A l'opposé, les approches synthétiques s'appuient sur les technologies à lecture courte pour reconstruire *in silico* des lectures longues, elles ne génèrent donc pas de « véritables » lectures longues ; il s'agit plutôt d'une modification de la composition de la librairie, qui en utilisant un grand nombre de codes-barres, permet l'assemblage artificiel de fragments plus grands. Les deux technologies dominant les approches synthétiques sont celles commercialisées par 10X Genomics et Illumina.

5.3.1. Les technologies SMRT :

5.3.1.1. Pacific Biosciences « PacBio » :

Quand les technologies NGS à lecture courte fixent l'ADN et permettent à la polymérase de se mouvoir le long du brin, la technologie PacBio guide un fragment d'ADN simple brin au travers d'un puits jusqu'à la polymérase fixée. Ces puits, de quelques picolitres, ont un fond transparent et sont appelés « *zero-mode waveguides* » ZMW, ils sont plusieurs milliers répartis à la surface d'une *flow-cell*.

Cette technologie impose une étape de préparation des librairies bien spécifique dont l'objectif est d'obtenir des fragments d'ADN « en cloche ». La figure 17 illustre les étapes du séquençage par la technologie PacBio. On remarquera que, contrairement aux lectures courtes, le phénomène est ici mesuré en temps réel, sans avoir à procéder à des étapes de rinçage par solvant entre chaque incorporation de dNTP.

Grâce à l'utilisation des ADN « en cloche », cette technologie permet le séquençage d'un modèle d'ADN en boucle (circulaire). Cela permet le séquençage multiple de chaque brin d'ADN à chaque rotation complète autour de la polymérase. Bien qu'il soit difficile de séquencer plusieurs fois les segments d'ADN de plus de 3kb, ces passages répétés permettent la génération d'une lecture consensuelle, connue sous le nom de « séquence consensuelle circulaire » et de réduire les erreurs stochastiques.

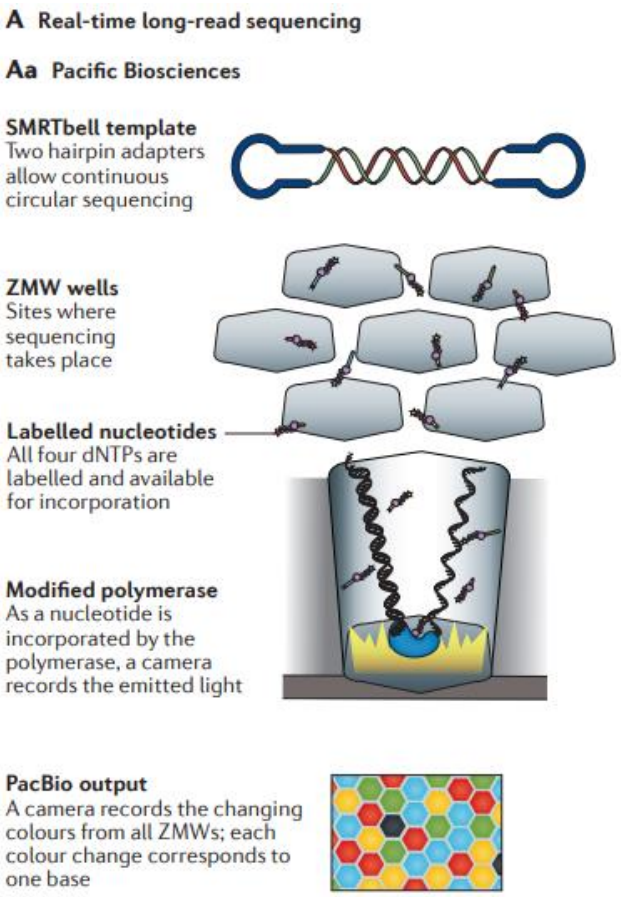


Figure 17: Illustration de la technologie commercialisée par Pacific Biosciences (PacBio). ***Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).***

5.3.1.2. Oxford Nanopore Technologies :

Cette seconde technologie à lecture longue est certainement celle qui se différencie le plus de toutes les autres technologies de séquençage. Singulièrement, les séquenceurs de nanopores ne sont pas basés sur l'incorporation de nucléotides ou sur une hybridation. Quand les autres plateformes sont basées sur la détection d'un signal secondaire (émission de signaux lumineux, détection de variation de pH), les séquenceurs de nanopores analysent de manière directe la composition en acides nucléiques d'une molécule d'ADN natif. Ce séquençage direct est effectué en forçant le passage de l'ADN à travers un pore protéique disposé sur une membrane synthétique et au travers duquel un courant électrique est mesuré (Figure 18).

Une variation de tension électrique est détectée lors de la translocation de l'ADN sous l'action d'une protéine motrice. Le traçage temporel des charges appelé « espace de gribouillis »

(*squiggle space)*, est caractéristique de la séquence nucléotidique qui doit alors être interprétée comme un k-mer.

En bioinformatique, un k-mer est une séquence biologique de longueur k elle-même contenue dans une séquence biologique. Par exemple la séquence « AGAT » possède 4 monomères (A, G, A, et T), trois dimères (2-mer) AG, GA et AT, deux trimères (3-mer) AGA et GAT et enfin un quadrimère (4-mer) AGAT.

Lorsque les autres plateformes ont 1 à 4 signaux à analyser, les instruments nanopore en ont plus de 1000 différents (un pour chaque k-mer possible), plus encore si l'on tient compte des bases modifiées que l'on peut retrouver sur l'ADN natif.

Le brin natif analysé étant double brin, la réaction de séquençage a lieu de telle façon que le brin sens puis le brin antisens sont séquencés l'un à la suite de l'autre, ce qui permet de générer une séquence consensus appelée séquence "2D".

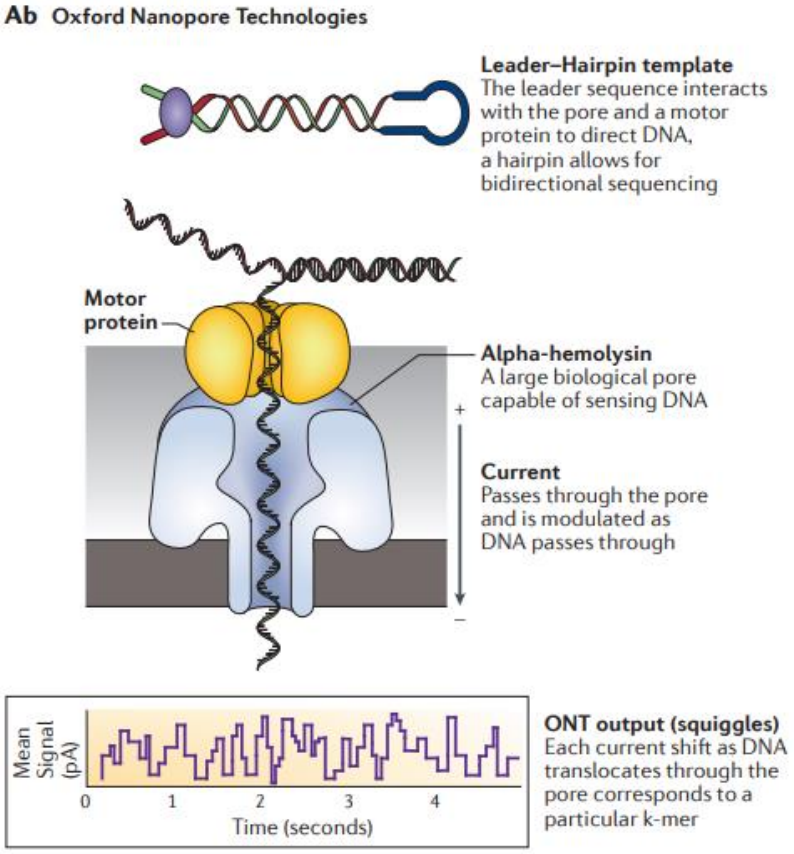


Figure 18: Illustration de la technologie commercialisée par Oxford Nanopore Technologies (ONT). Source : *Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).*

5.3.1. Les technologies synthétiques :

Contrairement aux véritables plateformes de séquençage, les technologies à lectures longues synthétiques reposent sur un système de codes-barres. Elles permettent à partir de l'association d'une multitude de fragments séquencés sur des séquenceurs à lectures courtes, de recréer *in silico* des lectures longues.

Dans les approches synthétiques, de grands fragments d'ADN sont répartis dans des puits de microtitrage (ou dans une émulsion) de façon à obtenir quelques milliers de fragments différents par puits, on appelle parfois ces puits les « containers ». Dans chaque puits, ces segments d'ADN sont à nouveau fragmentés et associés à des codes-barres uniques. S'en suit une étape de séquençage sur des plateformes à lecture courte classique. L'assemblage *in silico* de grands fragments est permis grâce aux informations transportées par les codes-barres.

Après le séquençage, les lectures sont démultiplexées à l'aide des codes-barres. Une étape appelée « sous-assemblage » permet de reconstituer synthétiquement des lectures de plusieurs kilobases, qui correspondent aux fragments d'ADN issus de la première fragmentation. Ces technologies sont basées sur une utilisation différente des amorces avec un système plus complet de codes-barres et un traitement bioinformatique spécifique. Le traitement informatique des données issues de ce type de séquençage est donc complexe et les constructeurs préservent une certaine confidentialité sur leur fonctionnement.

Il existe trois principaux systèmes permettant de générer des séquences longues synthétiques : la plateforme commercialisée par Illumina (Figure 19-Ba), le système 10X Genomics (Figure 19-Bb) et un système commercialisé par BionanoGenomics.

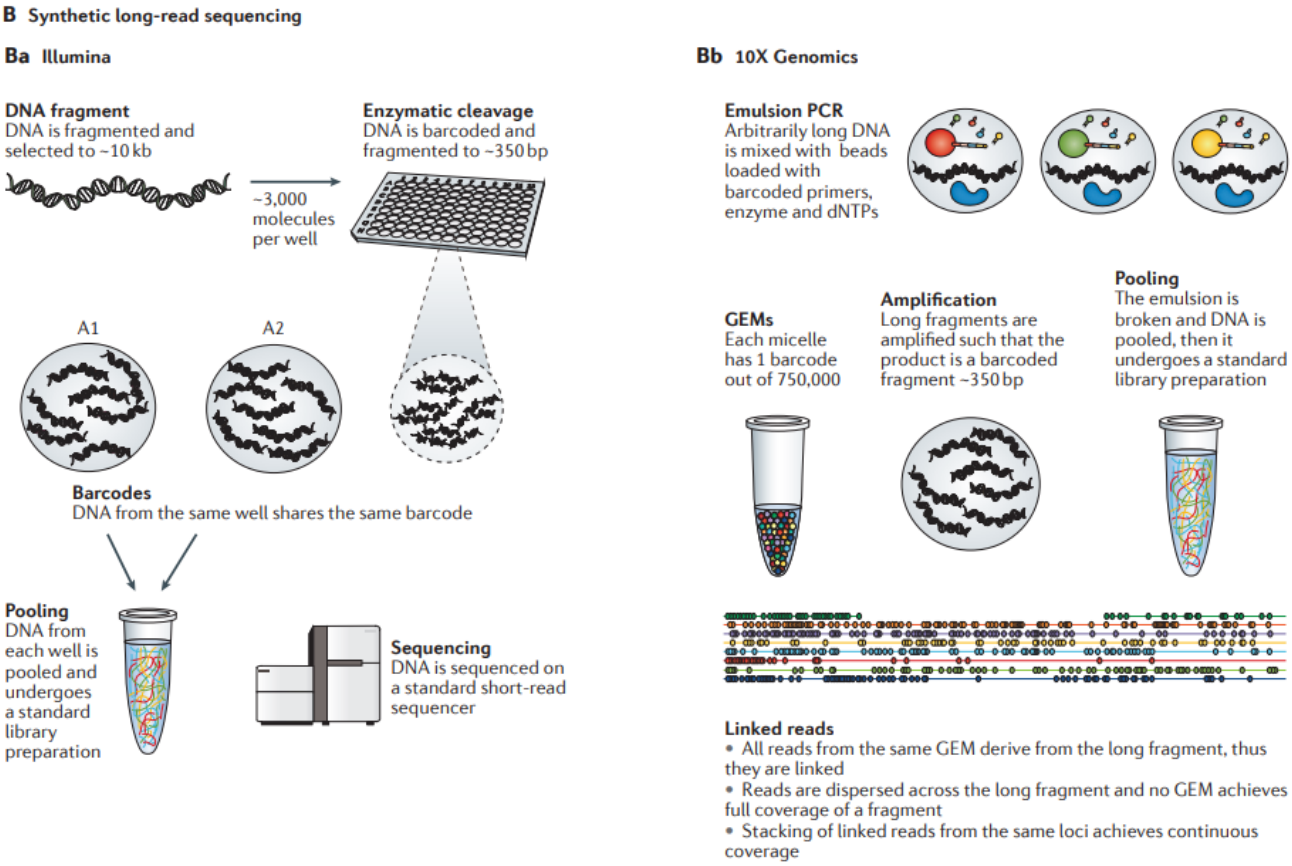


Figure 19: Illustration des technologies synthétiques. Ba : Illumina ; Bb : Séquençage en émulsion de 10X Genomics. ***Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).***

Chacune de ces technologies présente des avantages et des inconvénients en fonction du taux d'erreur acceptable, du budget financier, de la capacité en traitement bioinformatique ou encore de la qualité de l'ADN à séquencer. À ce jour les technologies à lectures courtes, et tout particulièrement le séquençage par synthèse d'Illumina, sont les technologies dominantes. Elles ont l'avantage d'être reproductibles et d'être les moins coûteuses du marché. Les techniques décrites précédemment sont à la pointe de la technologie, et ce dans un milieu qui connait une évolution très rapide. D'années en années les automates sont plus puissants, plus précis et plus rapides. A chaque augmentation de la quantité de données générées, la puissance nécessaire au traitement informatique de ces données augmente également.Cette complexification des données demande un investissement humain considérable. L'expertise des bio-informaticiens est nécessaire pour rendre les données exploitables. Les résultats de ces analyses doivent être parfaitement fiables pour aboutir à des décisions médicales. Pour bien comprendre l'enjeu de cette thèse et les processus bioinformatiques utilisés dans l'analyse des CNV, il semble nécessaire de rappeler les notions de base de la bioinformatique pour le séquençage haut débit. Le traitement bioinformatique débute dès la fin de l'étape de séquençage.

5.4. Bioinformatique

Les signaux analysés lors de la dernière étape du séquençage permettent la détermination de l'ordre des bases nucléiques qui constituent l'ADN (*base calling*). Les séquences nucléotidiques générées sont ainsi appelées "reads", ou « lectures », dans lesquelles chaque base nucléique A (Adénine), C (Cytosine), T (Thymine) ou G (Guanine) est associée à un score de qualité (score Phred). Ce dernier représente la confiance accordée à la base et est encodé en ASCII pour tenir sur un seul caractère.

L'ASCII (*American Standard Code for Information Interchange*) est une norme informatique de codage de caractères. Elle définit 128 codes à 7 bits avec 95 caractères imprimables : les chiffres arabes de 0 à 9, les lettres minuscules et capitales de A à Z et plusieurs symboles mathématiques et de ponctuation. Les caractères de numéro 0 à 31 et le 127 ne sont pas affichables car ils correspondent à des commandes de contrôle de terminal informatique et le numéro 32 correspond à un « espace ». C'est pourquoi le score Phred le plus bas (0) est codé avec le caractère de la table ASCII le plus petit, affichable et visible (33) qui correspond au symbole « ! ».

La quantité de lectures différentes générées pour chaque expérience est variable selon la technologie. Il y en a environ un million avec l'automate 454 de Roche et une dizaine de milliards avec les automates Illumina. L'acquisition et le traitement de cette importante quantité de données nécessitent donc le développement d'outils informatiques spécialisés et performants, car c'est à la fin du séquençage que la chimie fait place à la bioinformatique.

5.4.1. Pipeline Bioinformatique :

Appliquée au séquençage haut débit, la bioinformatique est organisée en « pipeline ». Un pipeline correspond à un enchaînement prédéterminé de processus informatiques ayant chacun un rôle dans le traitement des données brutes afin de les rendre interprétables par l'Homme. La figure 20 ci-dessous illustre les principales étapes d'un pipeline de bioinformatique appliqué au NGS (53).

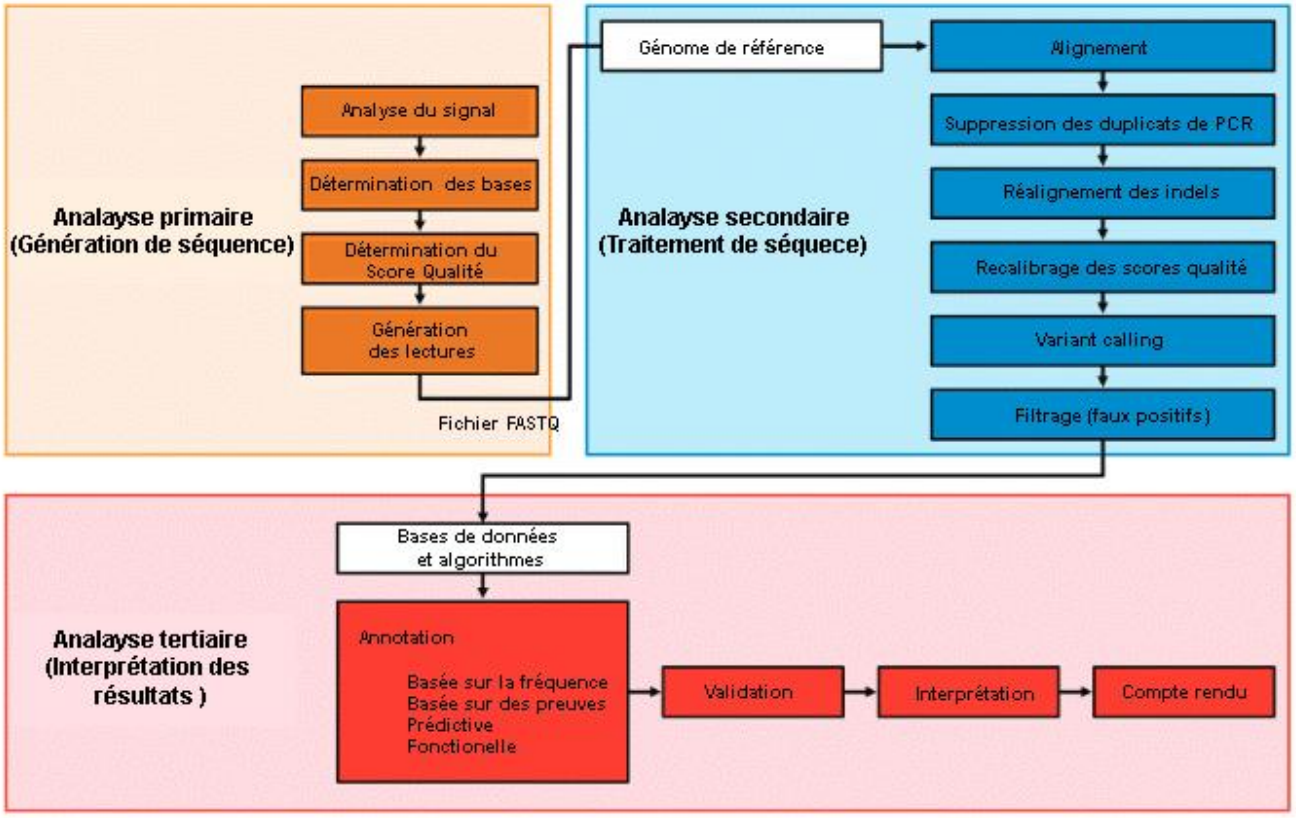


Figure 20: Principales étapes d'un pipeline bioinformatique utilisé pour l'analyse des données de séquençage à haut débit. ***Source : Grzych G. Evaluation of in silico prediction tools and interest of functional tests in the interpretation of identified variants by next-generation sequencing in human genetics. 2018.***

Il n'est pas nécessaire de parfaitement connaitre l'ensemble des étapes du pipeline pour comprendre le travail rapporté dans cette thèse. Cependant une brève description des étapes principales permettra d'insister sur l'impact majeur qu'a chaque élément du pipeline sur le résultat final.

5.4.2. Trimming qualité et filtrage :

On appelle ces étapes « *pre-processing* », elles sont utilisées pour exclure les lectures de mauvaise qualité qui auraient pu apparaitre lors du séquençage. (54)

Le *trimming* est un outil bioinformatique permettant de supprimer les régions de mauvaise qualité dans les lectures tout en essayant d'en conserver la longueur maximale. Comme pour chaque étape du pipeline, une mauvaise maitrise de ce processus peut entrainer des faux positifs (conservation de segments de mauvaise qualité comportant des artefacts analytiques) ou négatifs (suppression de régions dans lesquelles se trouverait une mutation).

Le filtrage (*filtering*) est la seconde méthode de correction qualité des lectures. Il est basé sur l'hypothèse que la majorité des erreurs sont rares et aléatoires, et qu'elles peuvent donc être éliminées si l'on ne se réfère qu'au modèle de lectures les plus fréquemment rencontrés. L'utilisation de modèles statistiques classiques (Markov caché) ou spécifiques (k-mer) permet cette étape de correction qui, une fois encore, doit être parfaitement maitrisée pour assurer un rendu clinique final (54,55).

5.4.3. Étape d'alignement globale :

A la sortie de l'automate de séquençage et des étapes de *pre-processing*, l'ensemble des lectures (qui contiennent les séquences et le score qualité associé à chaque base) sont sauvegardées dans un fichier dit « FASTQ » (*cf. Annexe 1 : Format FASTQ*). La longueur de ces lectures dépend de la technologie utilisée (à lecture longue ou courte). C'est à partir du fichier FASTQ qu'est réalisée l'étape d'alignement globale, elle consiste à essayer de localiser sur le génome humain la position de chacune des lectures.

Deux méthodes d'alignement existent :

1. L'assemblage *de novo* : Les fragments d'ADN qui se chevauchent permettent de reconstruire une séquence génomique continue appelée *contig*. L'assemblage des différents *contigs* entre eux permet d'obtenir une séquence plus grande appelée *scaffold*. Un *scaffold* contient un ou plusieurs trous (*gaps*). Ces gaps peuvent être comblés partiellement par l'utilisation de technologies à lectures longues, ou laissés tels quels dans un génome provisoire. Il est également possible de réaliser une étape de « *gap-filling* » pour accentuer le séquençage sur ces régions en particulier (Figure 21). Cette technique est très coûteuse en temps de calcul et est principalement employée pour reconstruire des génomes non connus pour lesquels il n'y a pas de références. Elle n'est pas utilisée en routine pour les données issues de l'humain (56).

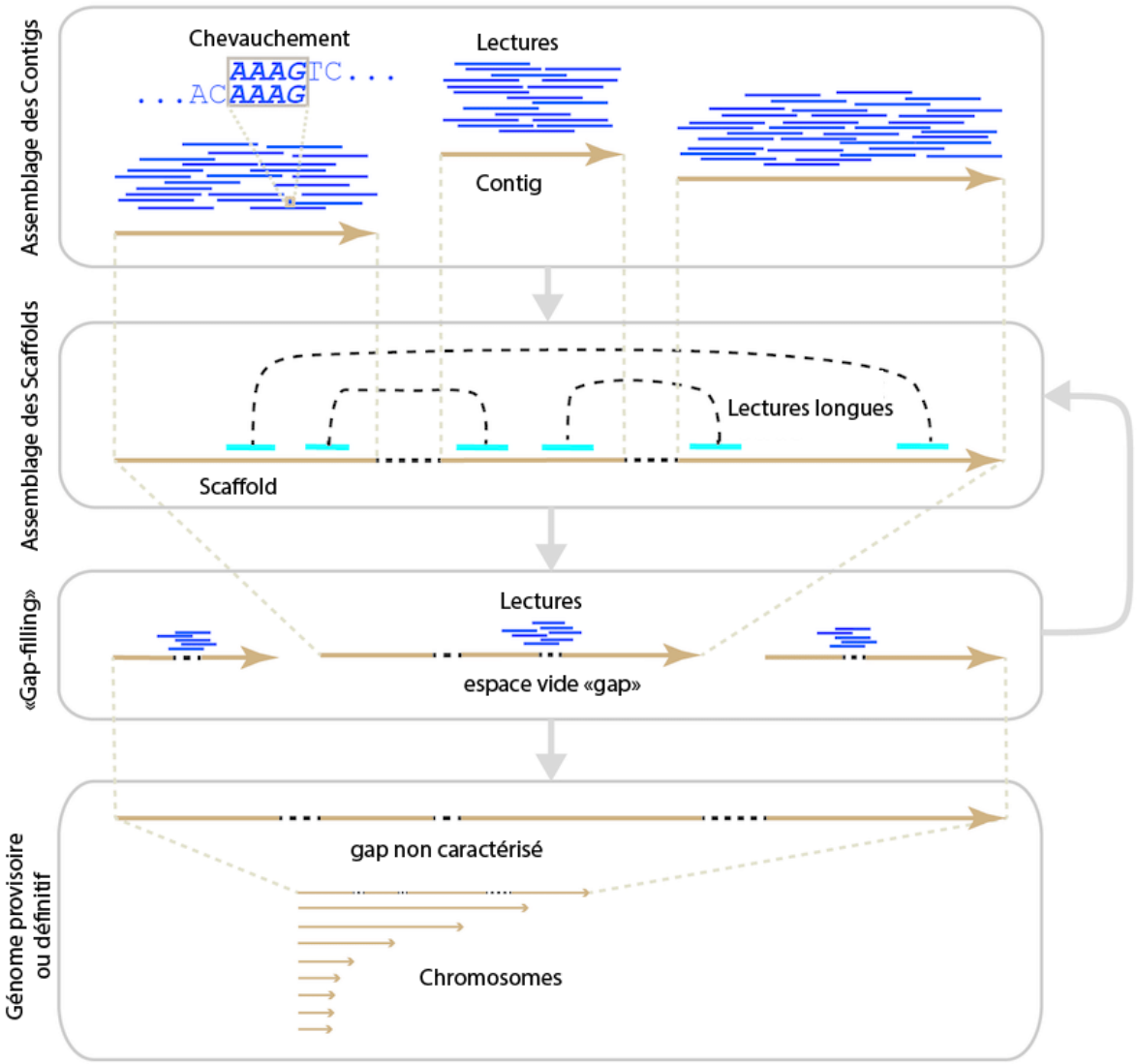


Figure 21: Illustration de l'assemblage *de novo* d'un génome entier. Grâce au chevauchement existant entre les lectures, on peut assembler des contigs, les espaces vides (*gaps*) peuvent ensuite être comblés grâce aux lectures longues. Il est enfin possible de procéder à une ou plusieurs étape(s) de « *gap-filling* » pour analyser les derniers espaces vides. Il est ainsi possible de reconstruire des chromosomes et un génome, qui peut être soit complet, soit provisoire avec la persistance de certains *gaps*. *Schéma extrait et traduit depuis la publication : Sohn, jang-il & Nam, Jin-Wu. (2016). The present and future of de novo whole-genome assembly. Briefings in Bioinformatics. 19. bbw096. 10.1093/bib/bbw096.*

2. L'alignement à un génome de référence : il consiste à faire correspondre chaque lecture à une région du génome par complémentarité, en se référant à une version déterminée du génome humain. Les calculs sont moins complexes que pour l'assemblage *de novo*, mais cette opération nécessite de disposer d'un génome de référence. Les logiciels qui procèdent à l'alignement utilisent généralement la transformée de Burrows Wheeler qui a inspiré le nom du logiciel d'alignement le plus connu : BWA (57).

Cette première étape d'alignement est appelée alignement global car il utilise l'ensemble des éléments de chacune des séquences. Quand il y a peu de différence entre les longueurs des séquences, des insertions sont créées dans la séquence la plus courte pour réussir à aligner les deux séquences d'une extrémité à l'autre. A l'inverse quand il y a beaucoup de différence entre les longueurs des séquences, chaque position d'une séquence longue est considérée comme étant un point de départ d'alignement avec une séquence courte.

Les aligneurs globaux utilisent des algorithmes de type II de Collins et Coulson aussi appelés « algorithmes de meilleure localisation ».

Cependant comme tout outil d'analyse, ces logiciels présentent un certain taux d'erreurs et sont moins efficaces quand les lectures à aligner s'éloignent de la référence (évènements complexes) ou si elles contiennent des séquences répétées. C'est pourquoi d'autres algorithmes d'alignements dits « locaux », basés sur la localisation des similarités, sont utilisés pour effectuer un alignement plus précis sur chaque région et détecter plus efficacement de petites délétions/insertions *(cf. Partie 1 - Chapitre 5.4.5).*

A la sortie de l'étape d'alignement, le fichier contenant les associations entre chaque lecture et ses coordonnées génomiques peut être soit un fichier « SAM » (pour *Sequence Alignment Map*), soit un fichier binaire compressé « BAM » (pour *Binary Alignment Map*) (58), ou encore un fichier binaire ultra-compressé « CRAM » (59). Les différences entre ces formats sont détaillées dans l'annexe 2 : Format SAM et BAM.

À partir du BAM, plusieurs étapes bioinformatiques sont encore nécessaires pour obtenir des données exploitables par l'Homme, à commencer par le traitement des duplicats de PCR.

5.4.4. Gestion des duplicats de PCR :

Le marquage des duplicats de PCR est une étape conseillée dans la grande majorité des pipelines bioinformatiques, on notera que leur suppression n'est pas toujours obligatoire.

Les duplicats de PCR peuvent être définis comme des lectures qui résultent du séquençage de deux copies ou plus du même fragment d'ADN ; ce sont des lectures séquencées indépendamment qui sont cependant exactement identiques. Il faut savoir qu'en l'absence de phénomènes parasites (clonaux ou optiques), deux lectures sont toujours différentes. Deux fragments d'ADN parfaitement identiques ne peuvent être que des clones car il est statistiquement improbable que de tels fragments soient générés lors de l'étape fragmentation.

La génération des duplicats de PCR se produit donc lorsque plusieurs clones d'un même fragment d'ADN (produits lors de la phase d'amplification de la préparation de la librairie) se lient au support solide (*flow-cell* ou bille) avant la seconde phase du séquençage (phase d'amplification). Chacun est alors amplifié et séquencé. Les lectures obtenues sont alors strictement identiques. Il est ainsi logique que l'augmentation du nombre de cycles de PCR lors de la préparation de la librairie augmente la quantité de duplicats de PCR.

Il existe aussi un phénomène de duplicats « optique » qui survient lorsqu'un cluster unique est identifié par erreur par l'automate comme étant deux clusters distincts, c'est une autre explication possible à la présence de duplicats dans les résultats.

Les duplicats de PCR sont responsables de deux phénomènes parasites :

- Ils peuvent contenir une ou plusieurs mutations erronées introduites pendant l'amplification PCR, produisant des faux positifs.
- Ils peuvent introduire un biais de brin en augmentant artificiellement l'occurrence de l'allèle séquencé en duplicata par rapport à l'autre allèle.

La gestion des duplicats de PCR est un problème inhérent aux techniques de séquençage haut débit, c'est un obstacle fréquent, mais dont l'impact réel est très dépendant de la quantité de duplicats générés. Dans les faits, si la technique est bien maitrisée, l'impact des duplicats de PCR reste faible (60).

5.4.5. Le réalignement des indels (réalignement local) :

L'étape familièrement appelée « réalignement des indels » est une étape d'alignement local qui a pour objectif de mettre en évidence des insertions/duplications ou des délétions qui auraient été ratées par la première étape d'alignement global dont le rôle était de localiser la position d'une lecture entière sur le génome.

L'objectif des alignements locaux est de détecter, sans préjuger de la longueur, les régions les plus similaires entre deux séquences. L'alignement local compare donc une partie de chacune des séquences et non leur totalité (contrairement aux alignements globaux). Comme le traitement informatique d'une délétion à l'intérieur d'une séquence est considéré comme une insertion dans la séquence lui faisant face, on utilise le terme « d'indel » (INsertion-DELetion).

L'étape de réalignement des indels est nécessaire pour compléter le travail des aligneurs globaux qui présentent des difficultés à aligner à la perfection des lectures de petite taille présentant des évènements complexes (plus d'un changement d'un seul nucléotide) et particulièrement lors d'insertions-délétions. Il n'est pas nécessaire de décrire l'importance de la détection des indels, ces derniers ayant généralement un impact majeur sur la fonction des gènes dans lesquels ils apparaissent. Cela est d'autant plus vrai lorsqu'ils sont localisés à l'intérieur d'un exon et entrainent un décalage du cadre de lecture (*frameshift*). C'est pourquoi l'étape de réalignement local des indels est une étape cruciale du pipeline.

Les algorithmes utilisés par les aligneurs locaux différents de ceux des alignements globaux sur le principe mathématique. Les deux sont basés sur des calculs matriciels mais se distinguent à la fois par les matrices utilisées et la pondération des espaces (*gaps*). Les principes d'alignement et les matrices ne sont pas détaillés dans ce document.

5.4.6. Le recalibrage des scores qualité :

Comme précédemment décrit, le séquençage génère pour chaque base un score qualité associé. Il quantifie la probabilité d'erreur lors de l'identification de la base entre A, C, T et G. Ce score qualité est présent dès le départ des processus bioinformatiques et sert de référence tout au long de l'analyse. Cependant il comporte lui-même des erreurs lors de son évaluation (61).

L'étape de recalibrage des scores qualité est une étape standard dans nombre de pipelines bioinformatiques. Elle est basée sur le principe selon lequel certaines erreurs de séquençage sont déjà connues et reproductibles. Les logiciels de recalibrage s'appuient donc sur des bases de données de variants, ainsi que sur la position de la base au sein de la lecture (62). Quand un variant présent est connu, le recalibrage modifie le score qualité de la base, passant du score observé au score empirique (recalibré).

Le recalibrage des scores qualité a pour principal intérêt de diminuer le nombre de variants détecté lors de l'étape d'identification des variants, en éliminant les faux positifs (63).

5.4.7. L'identification des variants (*Variant calling*)

L’étape d’identification des variants est une étape automatisée de mise en évidence des variants à partir des données ayant préalablement subi toutes les étapes décrites précédemment. Ce n’est pas l’étape mathématiquement ou informatiquement la plus complexe, mais sa performance dépend grandement de l’ensemble du pipeline bioinformatique, de la qualité de l’échantillon de départ, de la technique de séquençage et des paramètres choisis pour filtrer les résultats.

Lorsque se termine le "variant calling", les résultats sont disposés au sein d’un fichier texte appelé VCF pour *Variant Call Format* et détaillé dans l’annexe 5. Ce fichier contient l’ensemble des variants du sujet séquencé et les informations les concernant. C’est à partir de ce dernier que le travail d’interprétation médicale débutera.

L’étape d’identification des variants (*variant calling*) est généralement la dernière étape des processus bioinformatiques. Elle se concentre sur la détection des SNV et des petites indels (<50 pb) qui ont longtemps été les cibles principales des logiciels de variant calling. En effet les SNV sont les évènements les mieux connus et les plus décrits, ils sont plutôt bien détectés par les logiciels classiques des pipelines bioinformatiques en NGS. C’est ainsi que dès lors que la détection des SNV et des indels a été maitrisée, et grâce à l’évolution des technologies de séquençage, les bio-informaticiens ont entrepris de développer des logiciels spécifiquement dédiés à la détection des SV dans les données de NGS.

5.4.8. Le Génome humain de référence :

Le génome humain de référence est, comme son nom l’indique, un génome de référence et donc une base de données numérique de séquences d'acides nucléiques assemblées comme un exemple représentatif de l'ensemble des gènes dans un organisme individuel idéalisé d'une espèce, dans notre cas l’*Homo sapiens.*

Ce génome virtuel est assemblé à partir des données de séquençage d'un certain nombre d’individus, il ne représente donc pas avec précision l'ensemble des variations d'un organisme pris à l’échelle individuelle, mais fournit une mosaïque haploïde de différentes séquences d'ADN issu de plusieurs « donneurs ». Dans un univers idyllique, le génome de référence devrait représenter fidèlement le génome de l’ensemble de l’humanité. En réalité c’est un objectif extrêmement difficile en raison de la diversité retrouvée dans certaines parties du génome entre les populations les plus éloignées les unes des autres. Cette limitation de la

diversité contenue dans le génome de référence est aggravée par la sélection historique des individus inclus, principalement européens et nord-américains. Il est également important de comprendre que ce n'est pas le génome « le plus commun » ni un génome « sain ».

Depuis la publication initiale du génome humain, plusieurs « versions » du génome de référence se sont enchaînées. Chaque version est de meilleure qualité grâce aux évolutions technologiques et tente de rectifier le défaut d'inclusion des populations initialement sous-représentées. Cette sous-représentation entraine des implications réelles, y compris en ce qui concerne les résultats cliniques. Notre capacité à interpréter une variation dans la séquence du génome d'un individu dépend directement de notre capacité à distinguer ce qui est polymorphique de ce qui est pathologique.

Le génome de référence est donc une entité virtuelle « évolutive ». Il est modifié au travers de versions similairement à un programme informatique. Il y a des « mises à jour » (*releases*) majeures (ex. : GRCh37 vers GRCh38) et mineures (ex. : GRCh38.p13 vers GRCh38.p14), des « *patches* » lui sont appliqués. Ces patchs peuvent être nouveaux (*Novel patches*) et apportent de nouvelles informations ou correctifs (*Fix patches*). Actuellement la dernière version majeure du génome humain est la version GRCh38 (ou HG38). Cependant la version GRCh39 était prévue pour l'année 2020.

Le génome de référence est un outil indispensable pour toute analyse en biologie moléculaire humaine, nécessaire pour l'étape d'alignement, la caractérisation des variants et dans notre cas les logiciels de détection des CNV.

6. Détection des CNV par séquençage haut débit (NGS) :

Comme détaillé précédemment, les automates de NGS génèrent des millions de lectures indépendantes, qui après traitement bioinformatique, permettent d'obtenir une liste des variants dont la taille est inférieure à quelques dizaines de bases. La détection de ces petits SNV et indels est globalement maitrisée et de nouvelles approches analytiques recherchant les SV ont alors été mises au point.

Réussir à détecter les SV dans les données de NGS avec une fiabilité égale à celle de la détection des SNV est un objectif ambitieux. Ainsi de nombreuses équipes se sont emparées du sujet et plusieurs centaines de logiciels différents sont aujourd'hui disponibles avec une qualité variable. Ils sont généralement adaptés pour le séquençage génome entier (WGS) ou exome entier (WES). Seule une minorité a été testée et/ou adaptée à un usage sur un panel de gènes

(tg-NGS), pourtant la technique la plus utilisée dans les laboratoires de génétique moléculaire de routine à ce jour.

Malgré leur nombre, ces logiciels dont l'objectif est de détecter les anomalies de structure dans les données issues du NGS ne sont basés que sur 4 approches principales : le « Paired-end mapping » (PEM), l'analyse de la profondeur de lecture « Depth Of Coverage » (DoC), l'approche « Split-read » (SR) et l'approche par assemblage *de novo*.

6.1. Approche « Read-Pairs » ou « Paired-end mapping »

L'approche « Paired-End mapping » (PEM), aussi appelée approche « Read-Pair » (RP), est la plus ancienne approche développée pour détecter les SV dans les données de NGS (20,64,65). Elle est basée sur la détection de mésappariement des paires de lectures, la variation de la taille des inserts (66–68) et l'orientation des lectures.

Cette approche n'est envisageable que dans le cas d'un séquençage « paired-end ». A chaque lecture correspond un partenaire. Ensemble elles permettent le séquençage d'un fragment d'ADN depuis chaque extrémité. Pour chaque paire de lectures, la distance entre une lecture (R1) et son partenaire (R2) est connue sous le nom de « taille d'insert ». Dans le séquençage par paires (*paired-end sequencing*), les fragments séquencés avec la même librairie sont supposés présenter une distribution bien spécifique de la taille de l'insert (Figure 22). Ainsi, une discordance entre la taille moyenne observée des inserts et celle supposée, fais suspecter la présence d'un SV. Les méthodes de RP comparent donc, pour chaque paire de lectures, la taille moyenne de l'insert avec sa taille estimée.

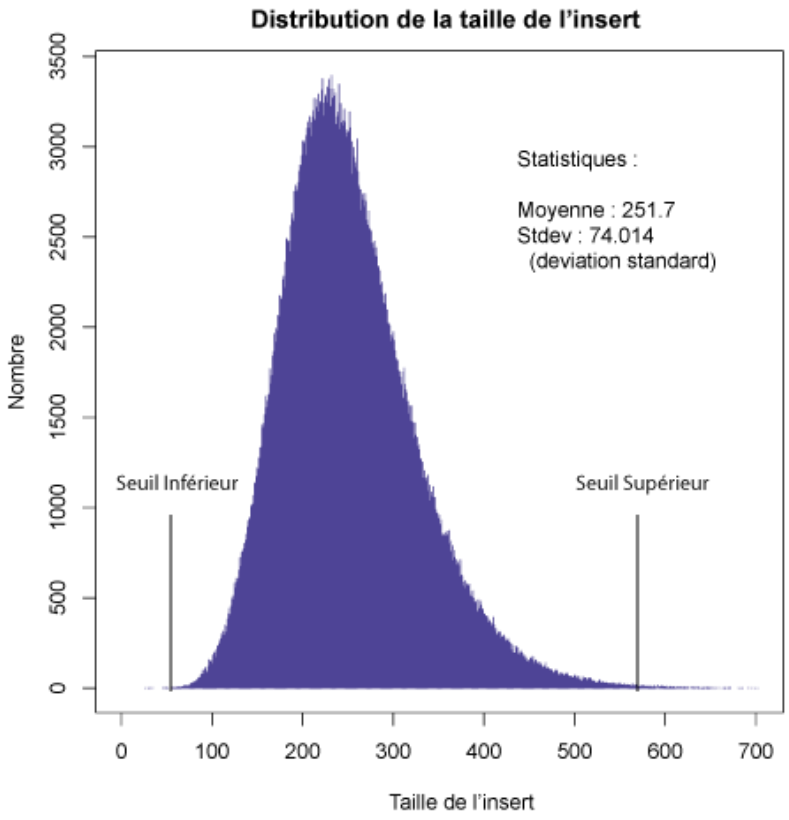


Figure 22: Distribution de la taille de l'insert (en pb). L'approche PEM impose la fixation d'un seuil inférieur et d'un seuil supérieur. Une taille comprise à l'extérieur de ces seuils indique la présence d'une SV. Statistiques extraites d'un fichier BAM représentatif du groupe de notre étude.

Au sein de l'approche PEM, deux logiques se distinguent quant à l'estimation de la taille de l'insert, elle peut être calculée soit à partir d'un génome de référence (approche basée sur une référence), soit à partir d'une valeur préfixée (l'approche par regroupement ou « clustering »).

L'approche par regroupement utilise une distance prédéfinie pour identifier les lectures discordantes, tandis que l'approche basée sur une référence utilise un test de probabilité pour distinguer une distance inhabituelle entre les paires de lectures par rapport à cette distance dans un génome de référence.

Les paires de lectures dont la taille de l'insert est celle attendue, qui sont correctement alignées sur le génome de référence, et dont l'orientation est correcte, sont appelées paires concordantes. Dans tous les autres cas les paires sont dites discordantes et suggèrent la présence d'un SV.

Les méthodes PEM permettent d'identifier efficacement les CNV (insertions, délétions, insertions d'éléments mobiles, duplications en tandem), mais également les SV équilibrés

comme les inversions. Elles permettent une détermination relativement précise des points de cassure.

La figure 23 illustre les différentes configurations que peut prendre une paire de lecture discordante en fonction du type de SV rencontré. Chaque type de SV peut être détecté en fonction de la taille de l'insert et de l'orientation de chacune des lectures d'une ou plusieurs paires.

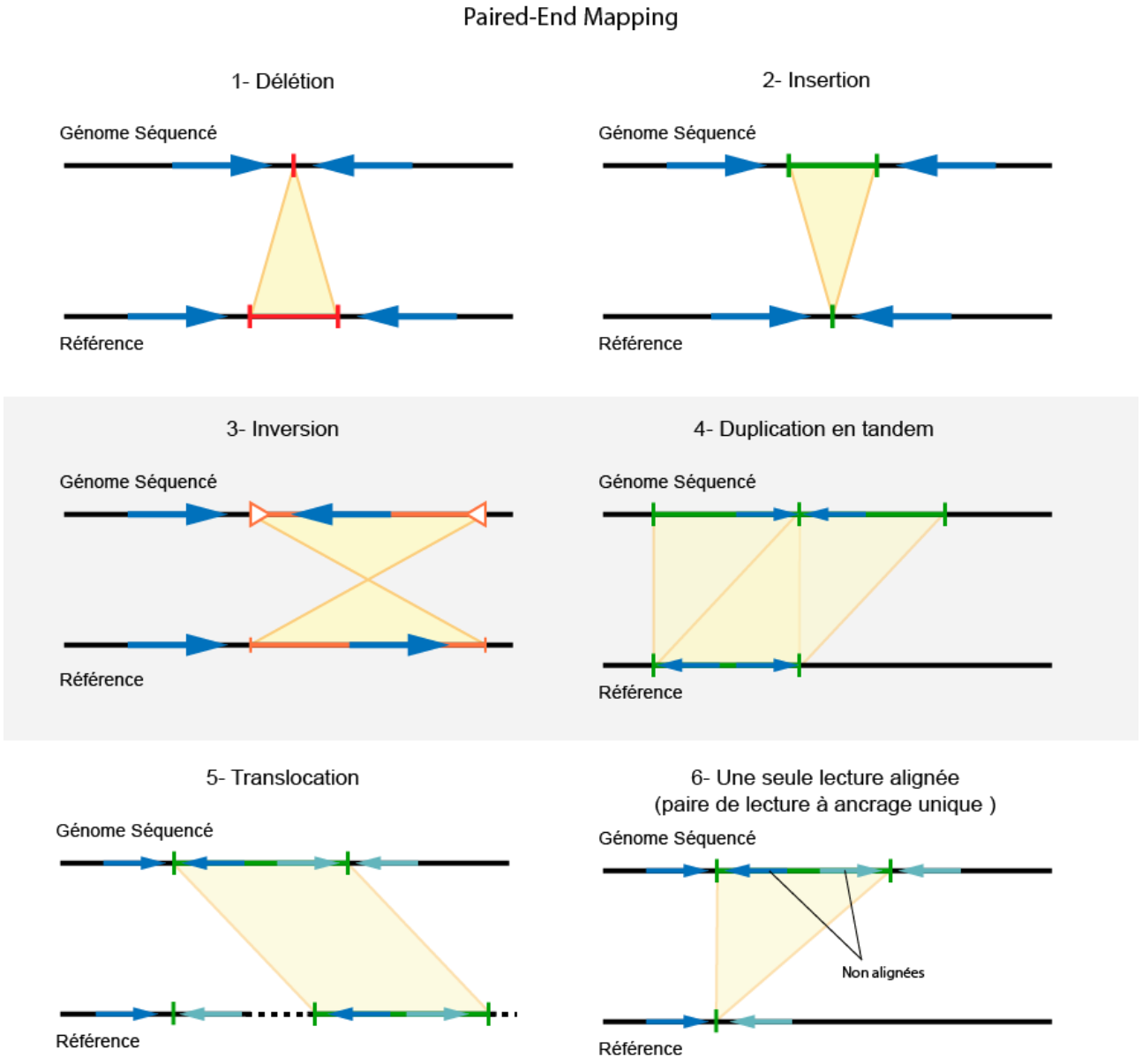


Figure 23 : : Illustration des différentes configurations que peuvent prendre les paires de lectures dans les différents types de SV. 1- Délétion : Quand une paire de lectures encadre la région délétée, la taille de l'insert après alignement sera nettement supérieure à celle attendue. 2- Insertion : Quand une paire de lectures encadre la région insérée, la taille de l'insert après alignement sera nettement inférieure à celle attendue. 3- Inversion : Une inversion peut être détectée quand une des lectures d'une paire se trouve sur la région inversée, l'orientation de cette lecture sera alors incorrecte. 4- Duplication en tandem : Quand les deux lectures d'une paire encadrent le point central d'une duplication en tandem, on retrouve après alignement ces deux lectures dans une orientation correcte, mais avec une position inversée l'une par rapport à l'autre.

5- Translocation : Pour détecter une translocation, il faut que deux paires de lectures encadrent chacune un point de cassure. Dans ce cas, après alignement, on retrouve chaque lecture dans une orientation correcte, mais avec des positions les unes par rapport aux autres bouleversées. Si la translocation est interchromosomique, les deux extrémités de chaque paire se retrouvent alignées sur un chromosome différent. 6- Paires de lectures à ancrage unique : Quand, pour deux paires de lectures, une seule lecture est correctement alignée, il est possible que ces dernières encadrent en réalité un fragment d'ADN dont la séquence n'existe pas sur le génome de référence. ***Figure traduite et inspirée de la publication Xi, Ruibin & Kim, Tae-Min & Park, Peter. (2010). Detecting structural variations in the human genome using next generation sequencing. Briefings in functional genomics. 9. 405-15. 10.1093/bfgp/elq025.*** **(69)**

Cependant l'approche PEM présente plusieurs limites :

- Elle ne permet pas de détecter les insertions/duplications de taille supérieure à la taille moyenne des inserts dans la librairie utilisée (70).
- Elle ne permet pas la détection des petits CNV, en raison de la difficulté à distinguer de petites variations dans la taille de l'insert (70).
- Elles ne sont pas applicables à la détection des CNV dans les régions de faible complexité avec duplication segmentaire (66).
- Les approches PEM ne sont pas applicables en cas de séquençage *single-end*.

6.2. Approche *Split-read*

Comme l'approche PEM, l'approche « *Split-Read* » n'est utilisable qu'en cas de séquençage *Paired-End*. Elle est basée sur l'analyse des paires de lectures dont seulement une a été correctement alignée, ces paires sont appelées des paires de lectures à ancrage unique.

La théorie sur laquelle repose cette approche est la suivante : en cas de SV il peut être impossible d'aligner correctement la lecture sur la référence. Les lectures non alignées peuvent donc contenir une information indirecte sur la présence d'un SV. Cependant pour conserver l'information concernant la localisation d'une lecture non alignée, il faut que la seconde lecture de la paire soit alignée.

L'analyse split-read consiste ainsi à utiliser toutes les paires dont une lecture n'a pas été alignée pour détecter des SV dans la région dans laquelle cette lecture aurait dû être alignée. Pour cela la lecture de la paire qui a été correctement alignée est utilisée comme une ancre, et la lecture non alignée est fragmentée avant d'être de nouveau alignée localement. La figure 24 ci-dessous illustre les différentes configurations possibles dans l'approche SR.

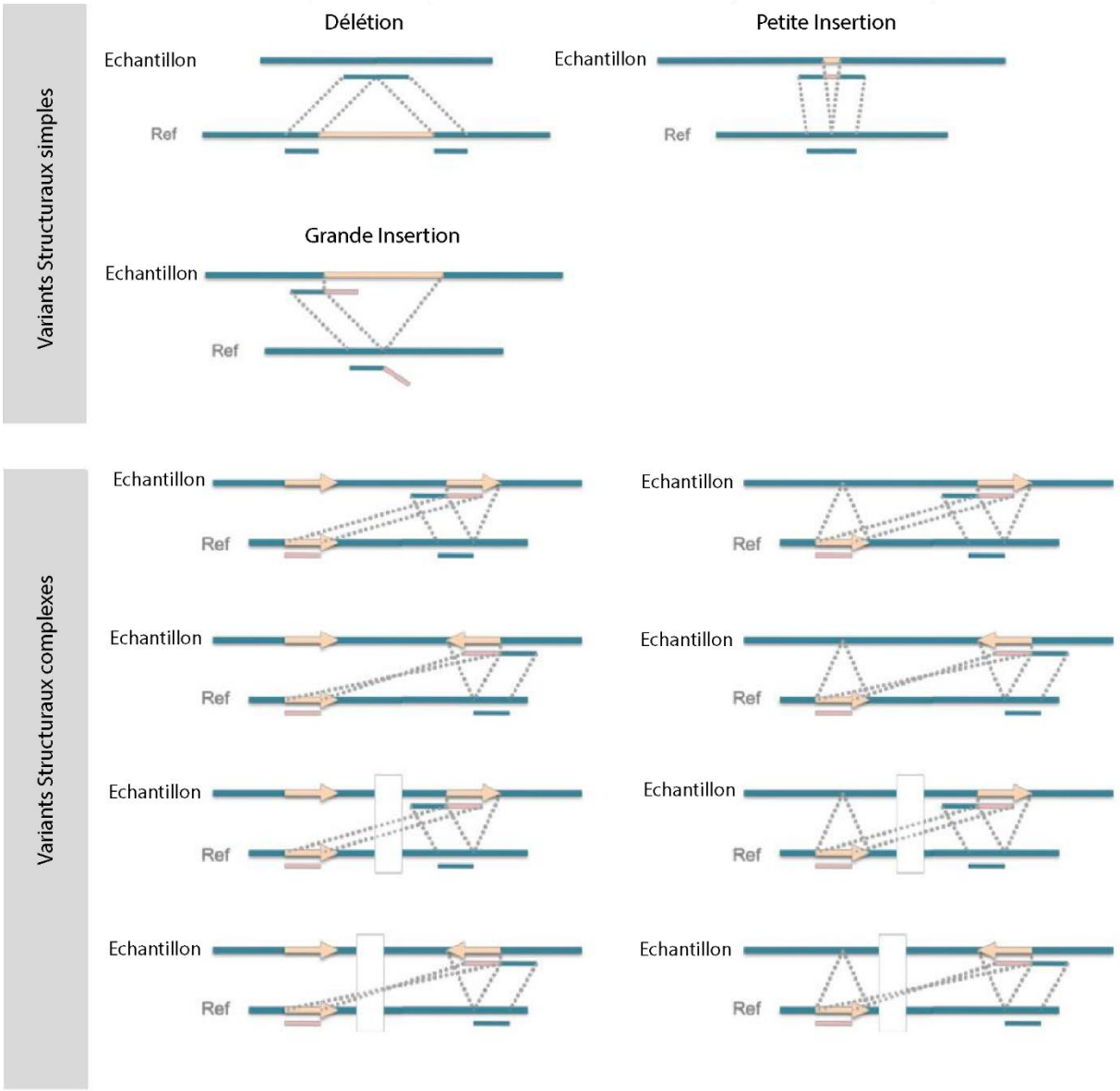


Figure 24: Illustration de l'analyse basée sur l'approche SR. Les SV peuvent être détectées lorsque des lectures recouvrent le point de cassure. Cette approche permet une détection directe des délétions et duplications/insertions. Les réarrangements plus complexes sont identifiés par une réanalyse globale en présence de plusieurs évènements de type délétions/insertions. Traduit de la publication *Identification of genomic indels and structural variations using split-read* (71).

6.3. Approche par analyse de la profondeur de lecture

L'approche par analyse de la profondeur de lecture est basée sur l'hypothèse qu'il existe une corrélation entre la profondeur de séquençage d'une région génomique et le nombre de copies de cette région dans l'ADN initial (72).

Ces approches détectent les CNV en comparant le nombre de lectures observées dans une fenêtre chromosomique prédéfinie avec le nombre estimé dans cette même région par un modèle statistique comme illustré dans la figure 25. Selon la performance des logiciels, différentes étapes de normalisation permettent d'éliminer de potentiels biais, généralement dus à la teneur en GC et aux régions répétées (73,74).

Il existe trois grands types de logiciels de DoC, ceux analysant un échantillon unique, ceux utilisant un couple échantillon testé – échantillon référence et ceux utilisant un ensemble d'échantillons de référence.

Les logiciels utilisant un échantillon unique, à défaut de pouvoir comparer à une référence, travaillent à partir des valeurs absolues du nombre de lectures. Pour les deux autres types de logiciels, les références servent de témoin négatif et permettent une analyse à partir de valeurs de profondeur relatives (75). L'utilisation de références a pour principal intérêt de réduire l'impact de la variabilité interéchantillon en lien avec les disparités d'efficacité de capture entre les exons (57-59). Cette seconde catégorie de logiciels représente la grande majorité des logiciels de DoC et semble posséder les meilleures performances.

L'approche DoC apporte une notion quantitative qui est absente dans les approches PEM et SR. Quand la SR et la PEM permettent de détecter la position potentielle des CNV elles ne permettent pas de quantifier cette variation du nombre de copies (délétion hétérozygote, ou homozygote, duplication ou amplification). En outre, l'approche DoC est plus efficace pour détecter les CNV de grande taille pour lesquels le PEM et le SR sont inefficaces (76). Elle est particulièrement bien adaptée aux méthodes de séquençage discontinu (tg-NGS et WES).

Cependant elle est moins performante dans les régions riches en séquences répétées et dans toute région qui pose des problèmes lors de l'alignement, comme celles possédant des pseudogènes (77). Les pseudogènes sont des gènes apparus par duplication au cours de l'évolution puis ayant, après divers évènements génétiques, perdu la capacité d'être transcrit ou traduit. Ils ne sont donc pas exprimés, mais leur homologie structurale avec certains gènes parfaitement fonctionnels entraine des difficultés lors du séquençage ou de l'alignement.

Compte tenu de l'état actuel du NGS en diagnostic, avec une utilisation très majoritaire des analyses par panels de gène (tg-NGS) et des exomes (WES), l'approche DoC est la plus efficace pour détecter des CNV (78).

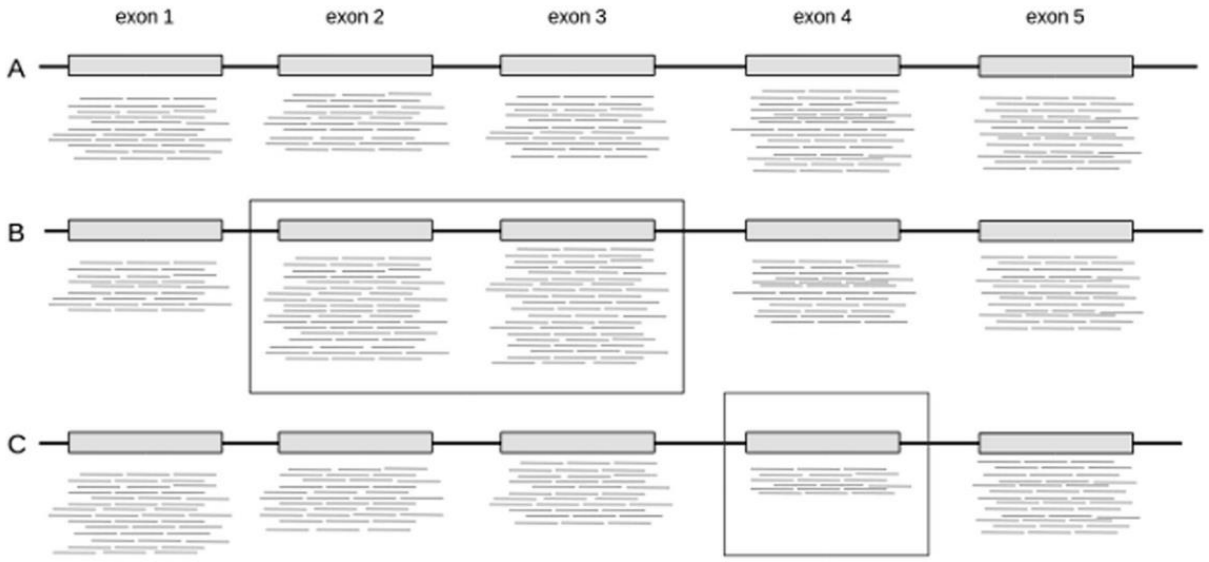


Figure 25 : Illustration de l'approche par profondeur de lecture (DoC) – A : Absence de CNV ; B : Duplication de deux exons (2 et 3) ; C : Délétion complète de l'exon 4. La profondeur de lecture est déterminée par comptage des lectures au sein de régions prédéfinies non chevauchantes (par exemple ici un exon). La profondeur observée dans cette fenêtre est ensuite comparée à une valeur théorique issue d'une ou plusieurs références. *Source : 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 (2020).* (79).

On remarquera que la traduction mot à mot de « *Depth Of Coverage* » serait profondeur de couverture et non de lectures. Cependant la définition bioinformatique de « couverture » ne permet pas cette association. Comme l'illustre la figure 26, la couverture correspond à une proportion de bases séquencées couvertes par une certaine quantité de lectures, par exemple la couverture à 30X est le pourcentage de bases couvertes par au moins 30 lectures. Pour correctement étudier une région en laboratoire de diagnostic, la couverture doit être de 100%. Quant à la profondeur, elle correspond au nombre de lectures chevauchant une base séquencée en particulier, ou à une moyenne du nombre de lectures si elle est évaluée sur plusieurs bases.

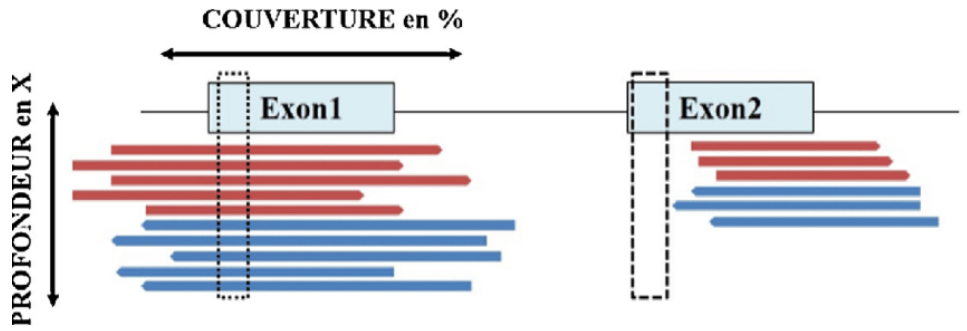


Figure 26: Illustration des termes « Couverture » et « Profondeur » de séquençage. Source : Lacoste, C. & Fabre, Alexandre & Pécheux, C. & Lévy, Nicolas & Krahn, Martin & Malzac, P. & Bonello-Palot, N. & Badens, Catherine & Bourgeois, Patrice. (2017). Le séquençage d'ADN à haut débit en pratique clinique. Archives de Pédiatrie. 24. 10.1016/j.arcped.2017.01.008 (80).

La valeur de couverture est dépendante d'une valeur préfixée de profondeur. L'association de mots « Profondeur de couverture » n'est donc pas adaptée et « *Depth of Coverage* » doit être traduite en « Profondeur de lecture » ou « Profondeur de séquençage ».

6.4. Approche par assemblage *de novo*

L'approche par assemblage de novo est similaire à l'étape d'alignement par assemblage *de novo* détaillée dans la Partie 1 Chapitre 5.4.3. À cela près qu'après avoir constitué la séquence à partir des lectures, une étape de comparaison avec une référence est effectuée. On peut ainsi mettre en évidence de potentiel SV (81–84).

L'avantage de l'approche par assemblage *de novo* est qu'elle permet de détecter des SV plus petites. Cependant elle fonctionne mal dans les régions riches en séquences répétées et nécessite une profondeur de séquençage élevée. C'est une approche qui n'est applicable qu'en cas d'utilisation de technologie de séquençage à lectures longues (y compris synthétiques).

Partie 2 : Logiciels disponibles et sélectionnés

L'objectif de ce travail est la mise à disposition du laboratoire d'un second pipeline bioinformatique pour la détection des CNV. Ce second pipeline permet d'assurer un contrôle qualité du pipeline actuellement utilisé (Sophia Genetics).

Dans ce but nous devions :

- sélectionner un ou plusieurs logiciels de détection des CNV,
- les tester sur un panel d'échantillons pour les évaluer.

Après la sélection d'un ou plusieurs logiciels, l'objectif est d'adapter leur utilisation, puis organiser leur inclusion dans le pipeline de routine.

Les critères qui ont pesé sur la sélection du/des logiciels étaient nombreux. Il fallait que l'analyse des CNV soit au moins aussi efficiente que celle proposée par le pipeline bioinformatique de Sophia Genetics. Mais au-delà de la performance scientifique, il nous fallait également prendre en compte des limites en lien avec une activité de routine :

- Limitations économiques : les dépenses budgétaires sont contrôlées.
- Limitation en ressources humaines : notre laboratoire dispose pour la bioinformatique de routine d'un bio-informaticien. Ce dernier est déjà employé aux activités préexistantes. La création d'une nouvelle charge de travail se doit d'être organisée pour éviter une surcharge du personnel. Cette limitation en ressources humaines a orienté nos choix vers un logiciel devant être accessible et rapidement maitrisable.

La négation de critères humains et économiques serait contre-productive pour permettre la concrétisation à court terme de ce projet.

La première étape consiste à sélectionner plusieurs logiciels puis à les tester. Pour cela nous nous sommes basés sur la littérature scientifique disponible sur PubMed. Il existe de très nombreux logiciels de détection des CNV qui diffèrent sur :

- leur approche scientifique,
- le langage informatique utilisé,
- la performance,
- l'équipe à l'origine du développement,
- les types de séquençages (panel de gènes, exome entier, génome entier),
- la technologie de séquençage utilisée (lectures longues ou lectures courtes),

- le type de prélèvement (somatique ou constitutionnel),
- la taille des CNV recherchés,
- la performance sur les chromosomes sexuels.

Le nombre de logiciels disponible ne cesse d'augmenter et il n'est donc pas ici question de réaliser un inventaire des logiciels disponibles. En correspondance avec nos objectifs nous avons sélectionné et testé 5 logiciels de détection des CNV :

1. ExomeDepth
2. Delly
3. DECoN
4. Smoove
5. BGW

Ces choix sont issus d'une volonté de diversification des approches théoriques utilisées et chacun de ces logiciels a présenté des avantages et des inconvénients.

1. ExomeDepth :

ExomeDepth est un standalone software (qui fonctionne hors-ligne, ne nécessitant pas obligatoirement une connexion internet pour fonctionner) – distribué sous la forme d'un package R (V > 3.4).

Publié dans la revue Bioinformatics en 2012 (85) le logiciel ExomeDepth est basé que l'approche DoC et compare la profondeur de lecture de chaque exon à la profondeur de lecture du même exon sur un set de références. Ce panel de référence doit être issu de la même technique de séquençage et de la même analyse bioinformatique que l'échantillon testé. Il est constitué d'une sélection parmi les autres échantillons lancés conjointement dans l'analyse ExomeDepth. Ainsi chaque échantillon analysé lors de la recherche de CNV l'est comparativement à ses pairs.

ExomeDepth a été développé pour s'affranchir des limites des logiciels basés sur une distribution binomiale, cette dernière étant trop sensible à la variabilité inter-échantillon. Pour cela il utilise une distribution bêta binomiale qui tient compte de la dispersion dans le ratio du nombre de lectures.

ExomeDepth peut être utilisé sur des données issues de l'analyse d'un panel de gènes, dès lors qu'il contient suffisamment d'exons. Il est adapté à la recherche de CNV rares car ExomeDepth suppose que les CNV recherchés dans l'échantillon ne sont pas présents dans le set de référence.

En conséquence, l'analyse de trios doit être faite séparément. Les développeurs précisent également que le ratio optimal set-référence:échantillon-test pour détecter un CNV est de 10:1.

La constitution du set de référence est une étape clé de la performance de l'analyse. La première étape réalisée par ExomeDepth consiste à ordonner l'ensemble des échantillons par corrélation avec l'échantillon testé, à la suite de quoi le set de référence est constitué par ajout séquentiel ; à chaque itération, les modèles mathématiques sont ajustés pour estimer la probabilité de détection d'une délétion hétérozygote d'un seul exon. La constitution du set de référence s'arrête une fois que cette probabilité cesse d'augmenter. Les concepteurs d'ExomeDepth ont constaté que la taille optimale du set de référence était d'environ 10 échantillons (quand il s'agit d'exomes).

Pour chaque exon, le modèle bêta binomial génère une probabilité pour trois scénarios distincts, la présence d'une délétion, d'une addition, ou aucun CNV, pour combiner la probabilité entre chaque exon successif, ExomeDepth utilise alors un modèle de Markov caché, aussi appelé Chaîne de Markov caché. Il s'agit d'un modèle statistique composé d'états et de transitions. Une transition matérialisant la possibilité de passer d'un état à un autre. Dans notre application chaque étape du Modèle Markov caché est un exon et les états sont un nombre de copies normal *(cn = 2)* et la présence d'un CNV *(cn = 1 ou 3).* Le modèle de Markov caché à un double objectif : fusionner les CNV entre plusieurs exons et déterminer la probabilité d'observer un CNV en fonction de l'état de l'exon précédent.

Dans les résultats, ExomeDepth fournit un facteur de Bayes (Statistique bayésienne) ou « rapport de vraisemblance ». Ce facteur de Bayes transforme le rapport des probabilités *a priori* en rapport de probabilités *a posteriori* et pondère la probabilité qu'un CNV détecté par ExomeDepth soit réel (vrai positif). Il n'est pas renseigné de seuil.

Un autre avantage notable du logiciel ExomeDepth est la possibilité de tenir compte du GC-content mais nécessite de fournir au logiciel un fichier au format FASTA du génome de référence. L'influence du contenu en GC dans la variation interindividuelle de la profondeur est un obstacle à la détection des CNV par approche DoC. En perturbant les étapes de capture et de séquençage, une forte teneur en GC entraine un bruit de fond. L'inclusion de ce facteur dans l'analyse de régression permet une réduction limitée mais significative de ce bruit de fond.

Plusieurs paramètres clés de ce logiciel sont à maitriser :

- Ne sont inclus dans l'analyse que les *lectures paires (paired reads)* cohérentes, soit deux lectures situées à moins de 1000 paires de bases l'une de l'autre, dans la bonne

orientation et avec un score de qualité Phred suffisant (par défaut >= 20). Ce critère peut être modifié par l'utilisateur (*paramètre min.mapq dans la fonction getBamCounts()*).

- Les exons espacés de moins de 50 paires de bases sont fusionnés en un seul.

Un paramètre peut être modifié lors de l'analyse : la « *transition.probability* » dans la fonction *CallCNV()*. Il est fixé par défaut à $10^{-4.}$ Il s'agit de la valeur donnée au modèle Markov caché pour le passage d'un état normal à un état délété ou dupliqué, en faisant diminuer ce paramètre on augmente la sensibilité et en l'augmentant on augmente la spécificité.

Il faut savoir lors de l'utilisation d'ExomeDepth :

- Les duplications sont plus difficiles à détecter que les délétions.
- Plus un CNV est long, plus il est détectable.
- Les CNV homozygotes sont plus faciles à détecter que les hétérozygotes.
- La profondeur minimum pour mettre en évidence un CNV sur un exon est de 30x (sur l'échantillon testé), ce n'est pas la profondeur d'efficacité optimale, mais la profondeur minimale pour espérer un résultat de qualité.
- Le logiciel prend en compte le GC-content.
- Les trios doivent être séparés pour une recherche de CNV par ExomeDepth.

Les avantages d'ExomeDepth :

- Simple d'utilisation avec quelques connaissances en R.
- Un manuel (vignette R) complet et bien renseigné.
- Logiciel rapide : analyse complète d'un run de 48 échantillons en un temps compris entre 7 et 8mins.
- Grand retour d'expérience : 2012-2020 => 8 ans d'expérience.
- Toujours actualisé, la dernière mise à jour date de 2020-01-09.

Les limites d'ExomeDepth :

- ExomeDepth peut ne pas détecter des CNV fréquents (polymorphismes) s'ils sont présents dans le set de référence.
- Limites : Inadapté à la détection des CNV sur les chromosomes sexuels.

Version utilisée : ExomeDepth 1.1.15

2. Delly :

Initialement publié en septembre 2012 dans le journal scientifique « Bioinformatics » (86), Delly est un logiciel de détection des CNV basé sur les approches PEM et SR. C'est un logiciel sur lequel la communauté de bio-informaticiens possède un bon recul. C'est un gratuiciel *open source* écrit en langage C++ dont la description est très détaillée.

Analyse PEM de Delly :

Pour chaque échantillon, Delly analyse l'orientation par défaut des paires de lectures et la distribution de la taille de l'insert (médiane et déviation standard). Grâce à ces paramètres, il identifie l'ensemble des paires discordantes (orientation anormale ou taille de l'insert anormale). Une paire est jugée discordante lorsqu'elle diffère de la médiane de plus de 3 déviations standards. Ce paramètre peut être modifié par l'utilisateur.

Toutes les paires discordantes sont alignées sur chaque chromosome permettant la détection des translocations. Pour chaque anomalie de structure identifiée, Delly effectue une analyse pour y faire correspondre chaque paire discordante.

Chaque type de SV est analysé séparément, ce qui permet l'identification des SV complexes qui seraient un ensemble de plusieurs SV différents (*cf. Partie 4 - Chapitre 1.1*).

Chaque type de SV modifiant la taille de l'insert de manière caractéristique :

- Les délétions et insertions sont recherchées dans les valeurs aberrantes à l'extrémité de la distribution de la taille de l'insert, mais dont l'orientation est correcte.
- Les inversions sont détectées comme des paires dont l'orientation est anormale, mais avec une taille d'insert correcte.
- Les duplications en tandem sont détectées comme des paires pour lesquelles l'ordre relatif entre la première et la seconde lecture a changé, mais ayant une orientation correcte.
- Les translocations sont détectées comme des paires dont les deux lectures sont alignées sur des chromosomes différents. Quatre types de translocations sont différenciés.

Analyse Split-Read Delly :

Dans l'analyse split-read de Delly, les paires distinguées lors de l'analyse PEM précédemment décrite sont interprétées comme des fragments génomiques contenant potentiellement différents points de cassure. Ils subissent une étape d'alignement très fine (à l'échelle du nucléotide) afin

de mettre en évidence des micro-homologies et des micro-insertions. L'analyse procède en plusieurs étapes, comprenant :

1. la recherche de candidats parmi les lectures,
2. l'extraction des SV de référence,
3. l'indexation et le comptage des k-mers,
4. la détection du point de cassure le plus probable,
5. le calcul d'un consensus à partir les différentes lectures,
6. l'alignement de cette séquence consensus par rapport aux SV de référence.

Pour chaque SV potentielle, Delly recherche localement ce qui est appelé des « lectures appariées à ancrage unique ». Une paire de lectures à ancrage unique est une paire dans laquelle l'une des lectures est correctement alignée alors que l'autre non. La lecture non alignée est donc un candidat pour l'analyse split-read. Pour obtenir une spécificité maximale, Delly enregistre et applique ensuite pour chaque lecture non alignée une analyse de direction (forward ou reverse), qui peut être déduite de sa lecture paire correctement alignée et de l'orientation par défaut de la librairie. Cette orientation par défaut est également utilisée pour déduire si un variant structurel est attendu en amont ou en aval de la lecture alignée.

L'ensemble des étapes qui suivent ainsi que le raisonnement global est similaire à celui de la définition des analyses split-read telles que décrites dans la Partie 1 Chapitre 6.2 de cette thèse.

Caractéristiques techniques :

- En entrée, pour chaque patient, le logiciel Delly nécessite un fichier BAM trié, indexé et dans lequel les duplicats de PCR sont marqués.
- Un fichier FASTA du génome de référence indexé est nécessaire pour effectuer l'analyse SR. *La concordance entre le génome de référence donné en entrée et celui utilisé pour l'alignement du BAM doit être parfaite sans quoi l'analyse retourne une erreur.*
- La sortie est un fichier au format BCF avec un index au format CSI (une extension classique pour un fichier d'index).
- Delly est disponible sur *docker* et sur *Bioconda*.
- Delly peut être utilisé avec les technologies à lectures longues.

Version de Delly utilisée : Delly2 V0.8.3

3. DECoN :

DECoN est un logiciel de détection des CNV par approche DoC décrit comme « développé spécifiquement pour les panels de gènes ». Il a été publié en 2016 dans la revue médicale Wellcome Open Research (87).

DECoN est en réalité une version modifiée d'ExomeDepth V1.0.0 qui comprend deux modifications. La première est qu'il permet de détecter les variantes affectant le premier exon sur un chromosome, tel que défini dans le fichier BED. La seconde est que les probabilités de transition entre les différents états (normal, délété, dupliqué) lors de l'application du modèle Markov caché ont été modifiées pour dépendre de la distance entre les exons. De cette manière les exons adjacents dans la liste des régions ciblées sont traités indépendamment s'ils sont situés loin l'un de l'autre sur le chromosome. Ces deux modifications ont cependant été intégrées dans ExomeDepth à partir de la version v.1.1.0.

DECoN comprend également plusieurs modifications visant à améliorer et à élargir l'utilisation d'ExomeDepth. ExomeDepth est un progiciel R qui nécessite donc qu'un utilisateur R averti sélectionne, spécifie et exécute les fonctions appropriées dans l'ordre correct pour générer un résultat. Il nécessite également un certain nombre de dépendances, qui peuvent elles-mêmes avoir des versions différentes selon l'installation locale de R par l'utilisateur.

DECoN optimise, standardise et automatise l'analyse des CNV par ExomeDepth. Cette approche serait une garantie supplémentaire pour convenir aux laboratoires de routine, DECoN a été modifié pour ne pas être impacté par la mise à jour des dépendances. L'ensemble des opérations d'analyse sont les mêmes qu'ExomeDepth.

DECoN nécessite en entrée les fichiers BAM, un fichier BED et un fichier FASTA de référence.

La lecture de la publication initiale de DECoN laisse penser qu'il s'agit d'une version améliorée d'ExomeDepth, ce qui était peut-être le cas en 2016 lors de sa parution. Depuis, les équipes d'ExomeDepth ont comblé les quelques lacunes que DECoN était supposé corriger. Concernant les actualisations, la dernière mise à jour de DECoN date de mai 2018, alors que la dernière version d'ExomeDepth date de janvier 2020. Cette absence d'actualisation de l'outil fait redouter une absence de développement actif.

4. Smoove/Lumpy :

Smoove est un gratuiciel open source de détection des CNV optimisé pour les technologies à lectures courtes. Il est basé sur une combinaison des approches SR et PEM. Le développement a également été orienté afin d'améliorer la spécificité, défaut important des logiciels basés sur ces approches, en écartant de nombreux signaux parasites courants.

Smoove utilise plusieurs autres logiciels open source dont Lumpy, au cœur de l'analyse des CNV. Lumpy est un logiciel de détection des CNV free-access et open source (C et C++) basé sur une combinaison de l'approche SR et PEM qui a été publié en 2014 dans la revue scientifique Genome Biology (88).

Smoove permet de paralléliser l'utilisation de *lumpy_filter* qui extrait les lectures discordantes et identifie les candidats au split-read. Ces deux types de lectures étant le point d'entrée du logiciel Lumpy qui réalise la recherche de CNV. Il permet également d'automatiser l'utilisation de Lumpy_filter pour supprimer les régions à profondeur anormalement importante, les régions parasites et les lectures dites « singleton » (pour lesquelles la lecture paire a été supprimée par l'un des filtres précédents). Ces étapes préliminaires automatisées permettent de réduire considérablement la durée et la puissance de calcul nécessaires à l'analyse par Lumpy.

Smoove transmet également en temps réel les données issues de l'analyse par Lumpy directement au logiciel *svtyper* qui réalise le génotypage, parallèlement à son exécution. Une étape finale permet de trier, compresser et indexer le fichier VCF de sortie.

Concernant les dépendances, Smoove exige :

- Lumpy et Lumpy_filter
- Samtools
- gsort
- bgzip+tabix

Et facultativement :

- svtyper
- svtools
- mosdepth
- bcftools
- duphold

Smoove/Lumpy est un logiciel décrit comme « optimisé pour les lectures courtes », il semblait donc être un bon choix pour tester les approches SR et PEM sur les données issues de notre panel. Cependant l'optimisation est faire pour les technologies à « lectures courtes » de type lectures longues synthétiques (PacBio), ce qui explique les résultats médiocres obtenus sur nos analyses.

5. Biomedical Genomics Workbench :

BGW (Biomedical Genomics Workbench) est une ressource privée de la société QIAGEN qui est un fournisseur d'échantillons et de technologies d'analyse pour le diagnostic moléculaire, les tests appliqués, la recherche universitaire et pharmaceutique. C'est une *holding* néerlandaise cotée en bourse et présente dans une trentaine de pays. Son siège opérationnel principal est situé à Hilden en Allemagne.

BGW et son outil d'analyse des CNV se démarquent des autres logiciels sélectionnés pour cette étude car il n'est pas ni un gratuiciel (freeware) ni *open source*.

L'algorithme de détection des CNV de BGW est conçu pour détecter les variations du nombre de copies à partir des données de séquençage qu'il s'agisse de panels de gènes ou d'exomes entiers. Cet algorithme, dont le code n'est pas consultable, est basé sur un logiciel open source appelé CONTRA. Ce dernier a été publié en 2012 dans la revue scientifique Bioinformatics (89) et s'inspire de la publication de Niu et Zhang parue en 2012 dans la revue Annals of Applied Statistics (90). Il est basé sur l'approche DoC et procède comme précédemment décrit (cf. Partie 1 chapitre 6.3) mais sans communiquer sur les méthodes statistiques utilisées.

Pour utiliser cet outil de détection des CNV, il faut au minimum un échantillon à tester et un échantillon de contrôle qui ne doit pas contenir le CNV à détecter. Il est conseillé que l'échantillon de contrôle partage autant de paramètres expérimentaux (par exemple, le sexe ou la méthode d'enrichissement et de séquençage) que possible avec l'échantillon testé.

BGW fournit tout de même une description partielle du fonctionnement de son logiciel qui effectue l'analyse des CNV en plusieurs étapes :

1. Les profondeurs de lecture sont analysées pour chacun des échantillons.
2. Une valeur de profondeur de référence est générée à partir des échantillons de contrôle.
3. Une analyse de profondeur de chaque chromosome est effectuée sur l'échantillon testé. Tous les chromosomes présentant des couvertures anormalement élevées ou faibles sont identifiés et ils ne seront pas utilisés pour établir les modèles statistiques de référence.

4. Les profondeurs des échantillons sont alors normalisées.
5. Chaque chromosome est segmenté en régions ayant un rapport de couverture similaire.
6. Un modèle statistique analyse alors la profondeur de ces régions.
7. Les CNV au niveau régional sont identifiés en utilisant le modèle de l'étape précédente.

Le logiciel de détection des CNV de BGW possède le lourd désavantage de ne pas être open source ni gratuit. Nous disposons au laboratoire de la version 12.0.3 de BGW.

Partie 3 – Matériel et Méthode

1. Échantillons utilisés :

Les logiciels d'analyse des CNV sélectionnés ont été testés sur un set d'échantillons ayant pour origine les activités de routine du laboratoire entre l'année 2017 et janvier 2020.

Ce set se compose ainsi :

26 « runs » NGS indépendants, chacun constitués de :

- 46 échantillons de patients
- 1 échantillon contrôle négatif (H_2O)
- 1 patient contrôle

Le total est de 1248 échantillons, dont 26 H_2O et 26 patients contrôles.

Chacun de ces runs a été choisi car il contient au moins un CNV confirmé par MLPA. Le tableau 4 récapitule les 30 CNV connus présents dans le set d'échantillons. Pour chaque échantillon, l'ADN est extrait à partir de sang périphérique (en tube additionné d'EDTA ou recueilli sur du buvard) puis séquencé sur un automate MiSeq (Illumina) en séquençage paired-end avec une librairie conçue en coopération avec la société Sophia Genetics.

À partir d'une série d'échantillons séquencés, les étapes classiques de traitement bioinformatique (*cf. Partie 1 - Chapitre 5.4.1*) et l'appel des variants sont effectués parallèlement sur deux pipelines indépendants. Le premier pipeline bioinformatique est celui commercialisé par la société Sophia Genetics. Son fonctionnement détaillé est soumis à des restrictions de propriétés intellectuelles. Le second pipeline est un pipeline développé au sein du laboratoire, basé sur BGW (Biomedical Genomics Workbench) et qui a été organisé par les bio-informaticiens du service. Il accomplit les mêmes tâches que celui de Sophia Genetics.

Ces deux pipelines parallèles assurent une sécurité et sont un gage de qualité. A la fin de leur fonctionnement, ils génèrent tous deux les fichiers FASTQ et BAM pour chaque échantillon, ainsi qu'un fichier VCF contenant les variants.

Une dernière étape réalise une comparaison entre les variants mis en évidence par chacun des pipelines. Elle est réalisée par un script python développé par notre équipe de bio-informaticiens pour générer un fichier Excel (Microsoft).

2. Composition du panel de gènes

Le laboratoire utilise en routine un kit appelé « HBOC ». Il correspond à l'indication « Syndrome de cancers du sein et de l'ovaire héréditaire ». Il inclut 27 gènes, dont seuls les 13 suivants font l'objet d'une analyse et d'un rendu de résultat en routine :

1. *BRCA1* (NM_007294)
2. *BRCA2* (NM_000059)
3. *CDH1* (NM_004360)
4. *EPCAM* (NM_002354)
5. *MLH1* (NM_000249)
6. *MSH2* (NM_000251)
7. *MSH6* (NM_000179)
8. *PALB2* (NM_024675)
9. *PMS2* (NM_000535)
10. *PTEN* (NM_000314)
11. *RAD51C* (NM_058216)
12. *RAD51D* (NM_002878)
13. *TP53* (NM_000546)

Le tableau 3 ci-après reprend les différents gènes inclus dans le panel « HBOC » utilisé en routine au laboratoire.

N°	Nom usuel	Nom complet	Ref NCBI (transcrit)	Ref Ensembl (transcrit)	Chromosome
1	*ABRAXAS1*	abraxas 1, *BRCA1* A complex subunit	NM_139076	ENST00000321945	4
2	*APC*	*APC* regulator of WNT signaling pathway	NM_000038	ENST00000257430	5
3	*ATM*	*ATM* serine/threonine kinase	NM_000051	ENST00000278616	11
4	*BARD1*	*BRCA1* associated RING domain 1	NM_000465	ENST00000260947	2
5	*BRCA1*	*BRCA1* DNA repair associated	NM_007294	ENST00000357654	17
6	*BRCA2*	*BRCA2* DNA repair associated	NM_000059	ENST00000544455	13
7	*BRIP1*	*BRCA1* interacting protein C-terminal helicase 1	NM_032043	ENST00000259008	17
8	*CDH1*	cadherin 1	NM_004360	ENST00000261769	16
9	*CHEK2*	checkpoint kinase 2	NM_007194	ENST00000328354	22

10	*EPCAM*	epithelial cell adhesion molecule	NM_002354	ENST00000263735	2
11	*MLH1*	mutL homolog 1	NM_000249	ENST00000231790	3
12	*MRE11*	MRE11 homolog, double strand break repair nuclease	NM_005591	ENST00000323929	11
13	*MSH2*	mutS homolog 2	NM_000251	ENST00000233146	2
14	*MSH6*	mutS homolog 6	NM_000179	ENST00000234420	2
15	*MUTYH*	mutY DNA glycosylase	NM_001048171	ENST00000372115	1
16	*NBN*	nibrin	NM_002485	ENST00000265433	8
17	*PALB2*	partner and localizer of *BRCA2*	NM_024675	ENST00000261584	16
18	*PIK3CA*	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	NM_006218	ENST00000263967	3
19	*PMS2*	*PMS1* homolog 2, mismatch repair system component	NM_000535	ENST00000265849	7
20	*PMS2CL*	*PMS2* C-terminal like pseudogene	Non codant : NR_002217	Non codant : ENSG00000187953	7
21	*PTEN*	phosphatase and tensin homolog	NM_000314	ENST00000371953	10
22	*RAD50*	*RAD50* double strand break repair protein	NM_005732	ENST00000378823	5
23	*RAD51C*	*RAD51* paralog C	NM_058216	ENST00000337432	17
24	*RAD51D*	*RAD51* paralog D	NM_002878	ENST00000345365	17
25	*STK11*	serine/threonine kinase 11	NM_000455	ENST00000326873	19
26	*TP53*	tumor protein p53	NM_000546	ENST00000269305	17
27	*XRCC2*	X-ray repair cross complementing 2	NM_005431	ENST00000359321	7

Tableau 3: Gènes séquencés dans le Panel HBOC. En bleu : les gènes faisant l'objet d'une analyse et d'un rendu de résultat en routine.

3. CNV du set de référence

Dans notre set d'échantillons, 30 CNV différents devaient être détectés. Ces différents CNV ne proviennent pas de la même famille à une exception près (numéros 3 et 4 du tableau 4).

Tous ont été vérifiés par MLPA et ont été retrouvés lors de l'analyse bioinformatique du pipeline de Sophia Genetics.

N°	Gène	Nomenclature (Transcrit)	Nomenclature (Protéine)	Description	Longueur supposée (pb)
1	*BRCA2*	c.(516+1_517-1)_(631+1_632-1)del	p.(Gly173Serfs*19)	Délétion de l'exon 7	164 à 3209
2	*MSH2*	c.(1386+1_1387-1)_(1661+1_1662-1)del	p.(Val463Glnfs*7)	Délétion des exons 9 et 10	3827 à 24703
3	*BRCA1*	c.(4484+1_4485-1)_(4986+1_4987-1)del	p.?	Délétion des exons 15 et 16	3643 à 10391
4	*BRCA1*	c.(4484+1_4485-1)_(4986+1_4987-1)del	p.?	Délétion des exons 15 et 16	3643 à 10391
5	*BRCA1*	c.(134+1_135-1)_(547+1_548-1)del	p.?	Délétion des exons 5 à 8	6808 à 9628
6	*MSH2*	c.(1276+1_1277-1)_(*1_?)del	p.?	Délétion des exons 8 à 16	37451 à ?
7	*MSH2*	c.(366+1_367-1)_(645-1_646-1)del	p.(Ala123_Gln215del)	Délétion de l'exon 3	328 à 3574
8	*PALB2*	c.(3113+1_3114-1)_(3201+1_3202-1)del	p.(Asn1039Glyfs*7)	Délétion de l'exon 11	137 à 13398
9	*MSH2*	c.(942+1_943-1)_(1076+1_1077-1)del	p.(Gly315Ilefs*29)	Délétion de l'exon 6	183 à 15372
10	*PALB2*	c.(3113+1_3114-1)_(3201+1_3202-1)del	p.(Asn1039Glyfs*7)	Délétion de l'exon 11	137 à 13398
11	*PALB2*	c.(1684+1_1685-1)_(2748+1_2749-1)del	p.(Gly562Valfs*6)	Délétion des exons 5 à 7	4283 à 10816
12	*PALB2*	c.(3113+1_3114-1)_(3201+1_3202-1)del	p.(Asn1039Glyfs*7)	Délétion de l'exon 11	137 à 13398
13	*EPCAM*	c.(858+1_859-1)_(*1_?)del	p.?	Délétions des exons 8 et 9	1497 à ?
14	*BRCA1*	c.(5332+1_5333-1)_(5406+1_5407-1)del	p.(Asp1778Glyfs*27)	Délétion de l'exon 22	123 à 3380
15	*PMS2*	c.(1144+1_1145-1)_(2174+1_2175-1)dup	p.?	Duplication des exons 11 et 12	< 6306
16	*RAD51C*	c.(705+1_706-1)_(837+1_838-1)dup	p.?	Duplication de l'exon 5	181 à 17465
17	*BRCA1*	c.(441-1_442-1)_(4357+1_4358-1)del	p.?	Délétion des exons 8 à 13	20596 à 25047
18	*MSH6*	c.(3172+1_3173-1)_(3556+1_3557-1)del	p.(Asp1058_Ser1185del)	Délétion des exons 5 et 6	1657 à 5145
19	*BRCA2*	c.(-?_-227)_(316+1_317-1)dup	p.?	Duplication du 5'UTR à l'exon 3	2914 à ?
20	*BRCA1*	c.(5332+1_5333-1)_(5406+1_5407-1)del	p.(Asp1778Glyfs*27)	Délétion de l'exon 22	123 à 3380
21	*BRCA1*	c.(5406+1_5407-1)_(5592+1_?)del	p.?	Délétion des exons 23 et 24	2075 à ?
22	*APC*	c.(?_-1)_(*1_?)del	p.?	Délétion complète du gène	136745 à ?
23	*BRCA2*	c.(8331+1_8332-1)_(8632+1_8633-1)del	p.?	Délétion des exons 19 et 20	748 à 8014

24	*BRCA1*	c.(4484+1_4485-1)_(4675+1_4676-1)del	p.?	Délétion de l'exon 15	3643 à 9967
25	*PMS2*	c.(903+1_904-1)_(1148+1_1145-1)del	p.(Cys303Thrfs*2)	Délétion des exons 9 et 10	2307 à 7962
26	*BRCA1*	c.(-20+1_-19-1)_(80+1_81-1)del	p.?	Délétion de l'exon 2	129 à 9521
27	*BRCA1*	c.(5277+1_5278-1)_(5332+1_5333-1)del	p.(Phe1761Asnfs*14)	Délétion de l'exon 21	104 à 3389
28	*PALB2*	c.(2586+1_2587-1)_(3201+1_3202-1)del	p.(Asn863_Met1067del)	Délétion des exons 7 à 11	12443 à 21240
29	*BRCA1*	c.(441-1_442-1)_(4357+1_4358-1)del	p.?	Délétion des exons 8 à 13	20596 à 25047
30	*CDH1*	c.(387+1_388-1)_(1711+1_1712-1)dup	p.?	Duplication des exons 4 à 11	11051 à 20156

Tableau 4: CNV présents dans notre panel d'échantillons. La longueur supposée des CNV a été calculée en additionnant la longueur des introns bordant chaque point de cassure estimé. Quand le CNV se termine en amont de la région 5'UTR ou en aval du 3'UTR la taille n'est pas calculable et cette incertitude est représentée par un point d'interrogation.

4. Gestion des pseudogènes :

Notre panel de gènes inclut plusieurs gènes qui possèdent des régions analogues dans des pseudogènes : *PMS2*, *PIK3CA*, *PTEN*, *CHEK2*. Seul *PTEN* et *PMS2* voient leurs analyses inscrites sur les comptes-rendus de routine.

Le haut degré d'homologie qui peut exister entre ces gènes et d'autres régions du génome entraine des difficultés d'alignement. La répartition des lectures entre deux régions homologues peut se faire selon différents principes qui dépendent de la programmation du panel. Il est possible de les répartir au hasard, équitablement, ou encore selon une proportion précise choisie.

Cette répartition influence grandement la performance de l'analyse des CNV car l'étape d'alignement est cruciale pour le fonctionnement de ces logiciels.

Partie 4 : Résultats

1. Sélection et utilisation des logiciels :

1.1. Expérimentation des 5 logiciels sélectionnés

Les 5 logiciels sélectionnés ont été testés en vue d'évaluer leur correspondance avec les objectifs préfixés. Ces derniers doivent pouvoir détecter un maximum de CNV de notre set de référence tout en minimisant le nombre de faux positifs.

Ont également été évalués :

- leur facilité d'utilisation,
- leur accessibilité,
- leur fiabilité d'utilisation,
- le temps d'analyse,
- le temps nécessaire à l'interprétation des résultats.

Le tableau 5 ci-dessous reprend les résultats de l'analyse des 5 logiciels vis-à-vis des 30 CNV de notre set de référence.

N°	Gène	Description	ExomeDepth	Delly	Smoove	DECoN	BGW
1	*BRCA2*	Délétion de l'exon 7	O	N	O	O	O
2	*MSH2*	Délétion des exons 9 et 10	O	N	N	O	O
3	*BRCA1*	Délétion des exons 15 et 16	O	N	N	O	O
4	*BRCA1*	Délétion des exons 15 et 16	O	N	N	O	O
5	*BRCA1*	Délétion des exons 5 à 8	O	N	N	O	O
6	*MSH2*	Délétion des exons 8 à 16	O	O/N	O/N	O	O
7	*MSH2*	Délétion de l'exon 3	O	N	N	O	O
8	*PALB2*	Délétion de l'exon 11	O	N	N	O	O
9	*MSH2*	Délétion de l'exon 6	O	N	N	O	O
10	*PALB2*	Délétion de l'exon 11	O	N	N	O	O
11	*PALB2*	Délétion des exons 5 à 7	O	N	N	O	O

12	*PALB2*	Délétion de l'exon 11	O	N	N	O	O
13	*EPCAM*	Délétions des exons 8 et 9	O	N	N	O	O
14	*BRCA1*	Délétion de l'exon 22	O	N	N	O	O
15	*PMS2*	Duplication des exons 11 et 12	N	O	O	N	N
16	*RAD51C*	Duplication de l'exon 5	O	N	N	O	O
17	*BRCA1*	Délétion des exons 8 à 13	O	O/N	O/N	O	O
18	*MSH6*	Délétion des exons 5 et 6	O	O	O	O	O
19	*BRCA2*	Duplication du 5'UTR à l'exon 3	O	N	N	O	O
20	*BRCA1*	Délétion de l'exon 22	O	N	N	O	O
21	*BRCA1*	Délétion des exons 23 et 24	O	N	N	O	O
22	*APC*	Délétion complète du gène	O	N	N	O	O
23	*BRCA2*	Délétion des exons 19 et 20	O	N	N	O	O
24	*BRCA1*	Délétion de l'exon 15	O	N	N	O	O
25	*PMS2*	Délétion des exons 9 et 10	O	N	N	O	N
26	*BRCA1*	Délétion de l'exon 2	O	N	N	O	O
27	*BRCA1*	Délétion de l'exon 21	O	N	N	O	O
28	*PALB2*	Délétion des exons 7 à 11	O	N	N	O	O
29	*BRCA1*	Délétion des exons 8 à 13	O	N	N	O	O
30	*CDH1*	Duplication des exons 4 à 11	O	N	N	O	O

Tableau 5: Capacité des différents logiciels testés à détecter correctement les 30 CNV de référence. O : CNV correctement détecté par le logiciel ; N : CNV non détecté par le logiciel. La codification O/N est utilisée quand un CNV est bien détecté par un logiciel, mais pour lequel la taille rapportée n'est pas compatible avec la réalité (*cf. Partie 4 - Chapitre 1.2*).

La première observation qui peut être faite à partir de ces données est qu'il existe une forte hétérogénéité dans la capacité des différents logiciels à détecter nos CNV. Pour l'ensemble de cette phase de test les logiciels ont été conservés dans leur paramétrage par défaut. Nous précisons tout de même qu'à la suite de l'obtention de ces résultats, de nombreux essais faisant varier différents paramètres actionnables ont été effectués sans obtenir de bénéfices significatifs.

Les logiciels ExomeDepth et DECoN ont démontré les mêmes performances, ce qui n'est pas surprenant car DECoN fonctionne sur la base de l'utilisation du code d'ExomeDepth. Cependant DECoN semble posséder une moins bonne spécificité avec une moyenne de 25 CNV rapportés par patients, contre une vingtaine avec ExomeDepth. En termes de rapidité, ces deux logiciels procèdent à l'analyse d'un patient en 10,5 secondes en moyenne (8min et 30s environ pour un run contenant 48 échantillons).

Les deux logiciels basés sur les approches SR et PEM ont montré une efficacité limitée, seuls 2 CNV ont été correctement détectés par Delly et 3 par Smoove. Deux CNV supplémentaires ont correctement été localisés sur une région rapportée comme délétée, mais la taille de ces dernières paraissait erronée *(cf. Partie 4 - Chapitre 1.2).* Les logiciels basés sur les approches PEM et SR sont à la fois plus difficiles à prendre en main que les logiciels basés sur l'approche DoC mais aussi plus longs pour effectuer l'analyse avec une moyenne de 5 à 10 minutes par échantillon. Les résultats générés sont également plus difficiles à interpréter.

Les logiciels basés sur l'approche DoC sont reconnus pour ne pas rapporter les points de cassures précis. Les CNV sont alors considérés comme commençant au début de la région présentant une variation significative de profondeur, jusqu'à la région de retour à la normale. Dans ces logiciels (DoC), les CNV rapportés sont de 2 natures différentes : délétion (DEL) ou duplication (DUP). Les résultats sont faciles à interpréter, de la forme « CHROMOSOME:DEBUT-FIN ».

A contrario les logiciels de SR et PEM obtiennent des résultats plus précis mais plus complexes qui peuvent être de 5 natures :

- DEL : Délétion
- DUP : Duplication
- INV : Inversion
- BND : Points de cassure
- INS : Insertion

Les DEL, DUP, et INV sont relativement faciles à aborder. Ils sont présentés sous la forme « CHROMOSOME \t DÉBUT \t LONGUEUR » ou « CHROMOSOME:DEBUT-FIN ». « \t » étant la représentation graphique du caractère informatique « tabulation horizontale ».

Les BND nécessitent une expertise plus poussée. Ils sont représentés conformément aux spécifications du format VCF et les explications ci-après sont issues du document « The Variant Call Format (VCF) Version 4.3 Specification ».

Contrairement aux autres annotations de CNV, le BND (pour « *breakend* ») n'est pas réellement un « type » de réarrangement mais l'annotation d'un point de cassure unique. Les BND sont utilisés pour annoter des réarrangements complexes avec plusieurs points de cassure. Ils sont utilisés lorsque sont détectés plusieurs réarrangements adjacents. Chaque réarrangement possédant 2 points de cassures. Pour chacun des deux points de cassure de chaque réarrangement, le logiciel produit une ligne de type BND sur le fichier VCF.

Dans l'annotation d'un BND, les champs « REF » et « POS » (*cf. annexe n°5 : Format VCF*) sont les mêmes que pour tout variant, la séquence nucléotidique du champ « REF » étant représentée par la lettre « s » ci-après.

Contrairement à l'ensemble des annotations de variants, dans le cas des BND le champ « ALT » ne contient pas uniquement la séquence nucléotidique remplaçant « s », mais aussi un ensemble d'informations spatiales.

Le contenu de ce champ « ALT » est composé de trois parties :

1. La chaîne de caractère « t » qui remplace la séquence présente dans le champ REF (« s »). La chaîne « t » peut être une version étendue de « s » s'il y a insertion de bases.
2. La position « p » du point de cassure, indiquée par une chaîne de caractère de la forme « CHR : POS ».
3. La direction dans laquelle la séquence se poursuit à partir de « p », cette information étant contenue dans l'orientation des crochets entourant « p ».

Pour créer le champ « ALT », ces 3 éléments sont combinés de 4 différentes manières. Dans chacun des 4 cas, il est supposé que « s » soit remplacé par « t », et qu'ensuite un réarrangement commençant à la position « p » soit joint à « t ».

Les différents cas possibles sont alors les suivants :

REF	ALT	Signification
s	t[p[	La région en aval de la position « p » est localisée après « t »
s	]p]t	La région en amont de la position « p » est localisée avant « t »
s	[p[t	La région (*complément inverse*) en aval de la position « p » est localisée avant « t ».
s	t]p]	La région (*complément inverse*) en amont de la position « p » est localisée après « t ».

La figure 27 ci-dessous illustre les 4 cas possibles ci-dessus en prenant pour exemple un point de cassure sur le chromosome 1 :

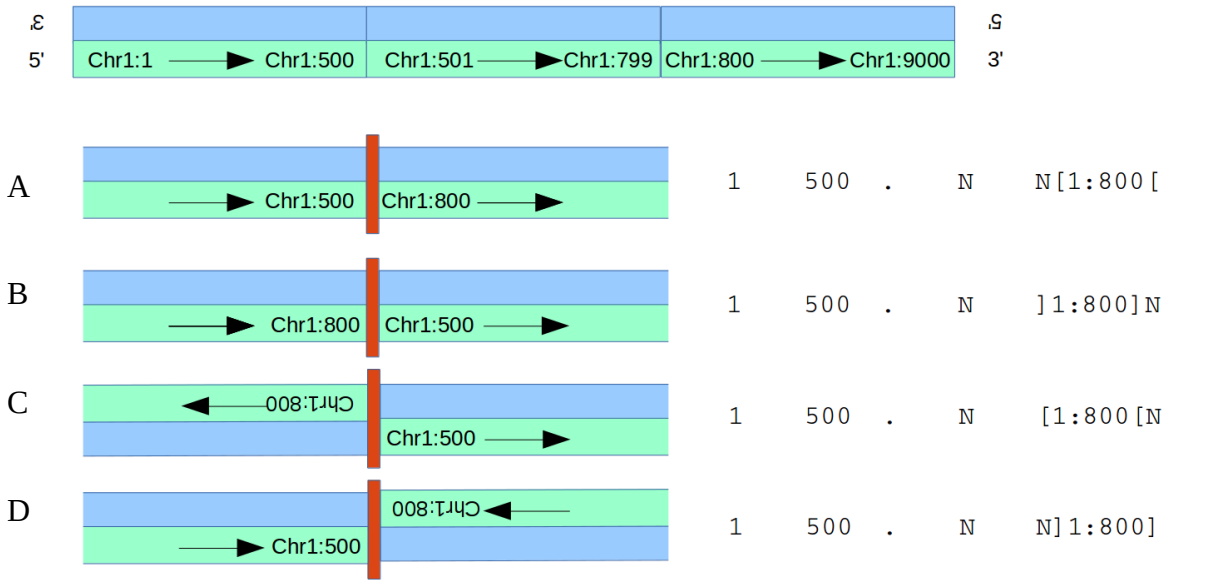


Figure 27: Illustration des différentes annotations d'un point de cassure en fonction des différentes possibilités de réarrangements. <u>Cas A</u> : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant à partir de ce point correspond à celle se trouvant en aval de la position « p » (position 800 du chromosome 1). <u>Cas B</u> : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant avant ce point correspond à celle se trouvant en amont de la position « p » (position 800 du chromosome 1). <u>Cas C</u> : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant avant ce point correspond à celle se trouvant en aval de la position « p » (position 800 du chromosome 1), il s'agit en réalité de la séquence complément inverse. <u>Cas D</u> : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant à partir de ce point correspond à celle se trouvant en amont de la position « p » (position 800 du chromosome 1), il s'agit de la séquence complément inverse. *Source : Merging Structural Variant Calls from Different Callers - simpsonlab.github.io*

L'annotation de réarrangement par BND est plus complexe mais plus complète. Elle est utilisable pour les réarrangements simples (DUP, DEL), par exemple la représentation A de la figure 27 correspond à une délétion de 299pb commençant en position 500 du chromosome 1.

Elle serait annotée dans le VCF par « 1 500 . N <DEL> ... END=799; »

Second exemple, la représentation D de la figure 27 correspond à une inversion et serait annotée dans le VCF par « 1 500 . N <INV> ... END=800; »

Chaque point de cassure peut être annoté à partir de chacun des deux côtés. La règle est généralement d'annoter à partir de la position la plus petite. Le tableau ci-dessous illustre les deux annotations possibles pour chacun des cas illustrés dans la figure 27.

Annotation à partir de la position 500	Annotation à partir de la position 800
1 500 . N N[1:800[	1 800 . N]1:500]N
1 500 . N]1:800]N	1 800 . N N[1:500[
1 500 . N [1:800[N	1 800 . N [1:500[N
1 500 . N N]1:800]	1 800 . N N]1:500]

Les SV de nature BND sont utilisés quand le logiciel ne trouve qu'un seul point de cassure, ou un réarrangement de complexité supérieure à DEL, DUP, INV. Ils compliquent cependant grandement les résultats. Dans notre set d'échantillons le nombre de BND rapportés par patients a été très variable et fortement lié à la qualité du séquençage. Il n'était pas rare que le résultat d'un échantillon contienne une vingtaine de résultats de BND.

Le logiciel proposé par BGW a montré des performances tout à fait satisfaisantes, mais son utilisation s'est révélée complexe. En effet ce logiciel est proposé sous la forme d'un module dans le cadre du logiciel de pipeline bioinformatique de BGW. Il s'agit d'un système représenté sous forme de « blocs » graphiques, assemblables via des voies d'entrée et de sortie. Ce logiciel s'utilise sans écriture de code et sans lignes de commandes, le but étant de rendre la bioinformatique accessible au plus grand nombre. De ce fait, le logiciel est peu accessible. Les interactions avec l'utilisateur sont plus limitées qu'avec l'emploi d'ordres directs via ligne de commande. Il est également impossible de réaliser une automatisation rapide (par exemple via un fragment de code linux). Quant à ses performances elles sont comparables à celles retrouvées par les autres logiciels basés sur l'approche DoC. La position des points de cassure est déterminée selon le même principe. Les annotations des points de cassure sont donc identiques entre ces 3 logiciels.

Une seule discordance survient entre les 3 logiciels de DoC. BGW ne retrouve pas le CNV consistant en une délétion de 2 exons sur *PMS2*. Plus exactement il rapporte une délétion de 3 exons. Le choix des références utilisées, associé à la variabilité des profondeurs sur *PMS2*, est à l'origine de cette différence ponctuelle.

Ces arguments sont plutôt en défaveur de l'utilisation du module de BGW en routine dans notre laboratoire. En effet les résultats sont similaires à ceux de ExomeDepth et ces deux logiciels sont basés sur la même approche, ils ne sont donc pas complémentaires.

En réponse à nos critères de sélection et dans le respect de nos objectifs, il a été pris la décision d'utiliser ExomeDepth comme principal logiciel de détection des CNV. En effet il s'est révélé être efficace, très rapide et facile à prendre en main, il est écrit dans un langage informatique des plus simple (R). Il est également facile à installer, open source, gratuit, et régulièrement mis à jour.

Bien que le choix du logiciel ait été en faveur d'ExomeDepth, il est intéressant de pousser les investigations sur les quelques cas dans lesquels les différents logiciels rapportaient des résultats discordants.

1.2. Etude des cas discordants entre les logiciels :

Parmi l'ensemble des résultats, 5 cas étaient particulièrement intéressants et méritent d'être détaillés, il s'agit des CNV n°1, 6, 15, 17 et 18 du tableau 5. Ils sont repris dans le tableau ci-dessous.

N°	**Gène**	**Description**	**ExomeDepth**	**Delly**	**Smoove**	**DECoN**	**BGW**
1	*BRCA2*	Délétion de l'exon 7	O	N	O	O	O
6	*MSH2*	Délétion des exons 8 à 16	O	O/N	O/N	O	O
15	*PMS2*	Duplication des exons 11 et 12	N	O	O	N	N
17	*BRCA1*	Délétion des exons 8 à 13	O	O/N	O/N	O	O
18	*MSH6*	Délétion des exons 5 et 6	O	O	O	O	O

1.2.1. *BRCA2* - Délétion de l'exon 7 :

Résultats des différents logiciels :

Logiciel	Départ (base)	Fin (base)	Longueur (pb)
ExomeDepth / DECoN / BGW	32 900 611	32 900 775	164
Delly	NA	NA	NA
Smoove	32 900 635	32 901 212	577

Il s'agit ici d'un cas dans lequel un logiciel basé sur les approches SR-PEM nous a permis d'identifier un des CNV du set de référence. La délétion rapportée ici est cohérente, elle est précise, plus encore que par l'approche DoC. Cependant, seul Smoove retrouve ce CNV. Cette discordance entre Smoove et Delly pourrait avoir pour origine une différence dans la sélection des lectures utilisées pour l'analyse SR.

Les points de cassures rapportés par Smoove et les logiciels de DoC sont différents. La taille de la région délétée est ainsi également différente. ExomeDepth et DECoN retrouvent une délétion de l'ensemble de l'exon 7 (164pb). Les points de cassure identifiés par Smoove sont en léger décalage, la région délétée est plus longue (577 pb). Cependant la différence observée entre les résultats des logiciels de DoC et de Smoove n'a pas d'impact sur la pathogénicité de cette délétion.

1.2.2. *MSH2* délétion des exons 8 à 16 et *BRCA1* délétion des exons 8 à 13

Tableau des résultats concernant la délétion des exons 8 à 16 de *MSH2* :

Logiciel	Départ	Fin	Longueur (pb)
ExomeDepth / DECoN / BGW	47 672 662	47 710 113	37451
Delly	47 596 699	107 133 406	59 536 707
Smoove	47 596 699	107 133 405	59 536 706

Dans ce cas-ci, nous avons pu constater que les délétions rapportées par les deux logiciels basés sur l'approche SR et PEM sont identiques (à une base près). Pour chacun de ces deux logiciels,

bien que le point de cassure de départ soit cohérent, le point de cassure terminal ne semble pas correct. Sur cette région (de la base 47 596 699 à la base 107 133 406 du chromosome 2) se trouve en effet le gène *MSH2* mais également le gène *MSH6*. *MSH6* s'étend sur le chromosome 2 de la base 47 922 669 à la base 48 037 240. Il est séquencé dans notre panel de gène dans les mêmes conditions que *MSH2*.

Le CNV rapporté par les logiciels Smoove et Delly implique une délétion complète de *MSH2* associé à une délétion complète de *MSH6*, or les résultats des autres logiciels, l'analyse de Sophia Genetics et la MLPA ne sont pas en faveur de cette délétion complète du gène *MSH6*. Nous pouvons en déduire que bien qu'une délétion des exons 8 à 16 de *MSH2* soit rapportée, et bien que le point de cassure initial soit théoriquement possible, la délétion rapportée est fausse.

Tableau des résultats concernant la délétion des exons 8 à 13 de *BRCA1* :

Logiciel	**Départ**	**Fin**	**Longueur (pb)**
ExomeDepth / DECoN / BGW	41 231 326	41 251 922	20 596
Delly	41 246 558	64 901 262	23 654 704
Smoove	41 196 419	73 319 847	32 123 428

Ce cas est similaire à celui détaillé ci-dessus, ici également les points de cassure initiaux rapportés par les logiciels basés sur l'approche SR et PEM sont cohérents. Cependant une fois encore la longueur de la région rapportée comme délétée n'est pas en rapport avec la réalité. Selon un raisonnement identique au cas précédent, sur le chromosome 17 et en aval du gène *BRCA1* se trouve le gène *RAD51C*, s'étendant de la base 56 769 934 à la base 56 811 703. Il est également séquencé dans les mêmes conditions. Or aucune délétion n'a été retrouvée par technique MLPA, par Sophia Genetics ou par les autres logiciels. Ce raisonnement est également applicable au gène *BRIP1* qui s'étend de la base 59 760 631 à la base 59 938 925 de ce même chromosome 17.

1.2.3. *PMS2* - Duplication des exons 11 et 12

Tableau des résultats concernant la duplication des exons 11 et 12 de *PMS2* :

Logiciel	Départ	Fin	Longueur (pb)
ExomeDepth / DECoN / BGW	NA	NA	NA
Delly	6 020 605	6 027 304	6 699
Smoove	6 020 604	6 027 268	6 664

Il s'agit ici du seul cas dans lequel les logiciels basés sur les approches SR-PEM ont montré une performance supérieure à ceux basés sur l'approche DoC. La position théorique du point de cassure initial de ce CNV se trouve entre la base 6 027 252 et 6 029 430 (intron 10-11), le second point de cassure se trouve entre la base 6 018 328 et 6 022 454 (intron 12-13).

Les résultats des logiciels basés sur l'approche SR et PEM sont donc parfaitement cohérents. Ils apportent dans ce cas une information qui est manquante dans les résultats des logiciels basés sur l'approche DoC et semblent s'affranchir des problèmes en lien avec le pseudogène *PMS2CL*.

1.2.4. *MSH6* - Délétion des exons 5 et 6

Tableau des résultats concernant la délétion des exons 5 et 6 de *MSH6* :

Logiciel	Départ	Fin	Longueur (pb)
ExomeDepth / DECoN / BGW	48 030 534	48 032 191	1 657
Delly	48 028 788	48 032 571	3 785
Smoove	48 028 786	48 032 570	3 784

Ce cas illustre le seul exemple dans lequel tous les logiciels ont correctement identifié le CNV recherché, il s'agit d'une délétion des exons 5 et 6 du gène *MSH6*.

La position du premier point de cassure est donc supposée se trouver entre la base 48 028 295 et la base 48 030 558 (intron 4-5) et le second point de cassure entre la base 48 032 167 et la base 48 032 756 (intron 6-7).

Les résultats sont cohérents. On pourrait se fier aux logiciels basés sur les approches SR-PEM réputés meilleurs dans l'estimation de la position précise des points de cassures.

1.3. Optimisation de l'utilisation de ExomeDepth

L'objectif suivant de ce travail consistait donc à optimiser l'utilisation d'ExomeDepth spécifiquement pour notre laboratoire. En effet bien qu'ayant une bonne sensibilité (tous les CNV détectés à l'exception d'un), la spécificité de ce logiciel restait un facteur négatif. Au total, pour l'ensemble des échantillons de tous les « runs » évalués et en ne considérant que les gènes de routine (et *PMS2CL*), ExomeDepth rapportait 671 CNV. La figure 30 représente la répartition de ces 671 CNV sur les différents gènes de routine. Nous retrouvons avec ExomeDepth en moyenne 0,538 CNV par échantillon analysés et 25,8 CNV par « run ». On remarquera qu'en moyenne 20,2 échantillons par run présentaient au moins un CNV. Le tableau ci-dessous reprend ces valeurs statistiques.

Nombre total de CNV rapportés	671
Nombre moyen de CNV par « run »	25,8
Nombre moyen d'échantillons présentant au moins un CNV par « run »	20,2
Nombre moyen de CNV par échantillon pour l'ensemble des échantillons testés.	0,538

Tableau rapportant les statistiques des résultats d'ExomeDepth avant la phase d'optimisation.

La charge de travail supplémentaire pour l'évaluation de ces nombreux « faux positifs » serait considérable, de même que les des surcoûts si l'on envisageait de réaliser une technique MLPA pour chacun des 25,8 CNV rapportés dans un run. Nous avons donc analysé une nouvelle fois les résultats issus de l'évaluation des logiciels. Nous nous sommes concentrés sur ceux d'ExomeDepth en particulier, et avons recherché des leviers d'action afin d'augmenter la spécificité.

Quand nous analysons les résultats, nous observons que *PMS2* et *PMS2CL* sont responsables de respectivement 40% et 52% des positifs totaux. Ces valeurs sont illustrées dans les figures 28 et 29 ci-dessous.

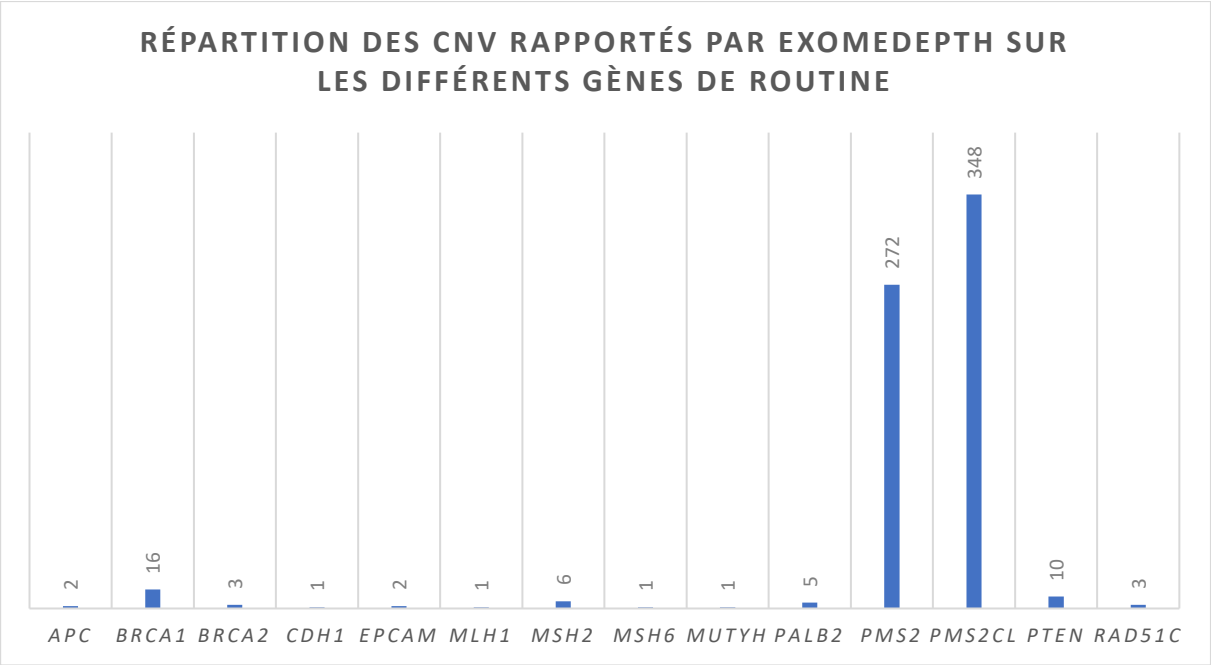


Figure 28: Illustration graphique de la répartition des 671 CNV rapportés par ExomeDepth lors de l'analyse du set d'échantillons sur les différents gènes de routine du laboratoire.

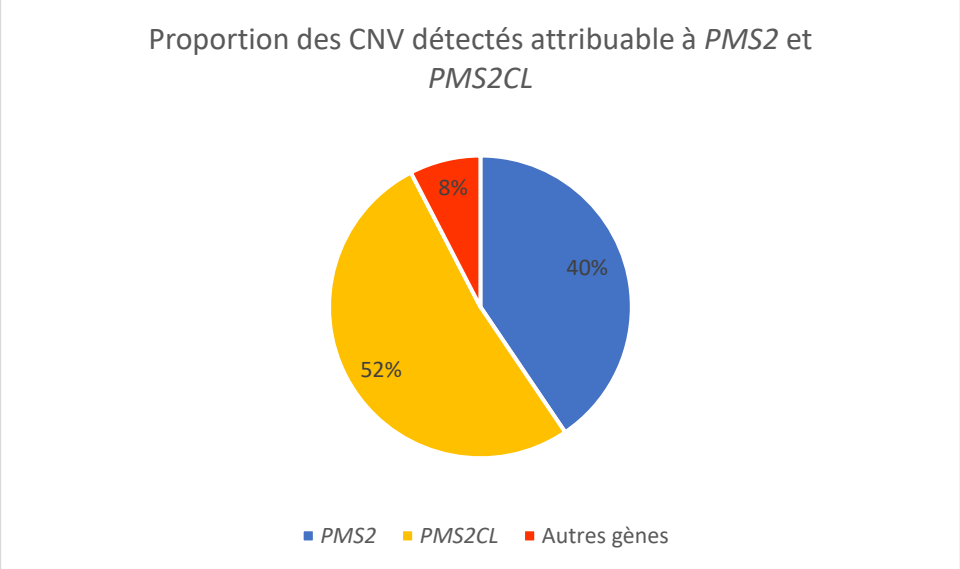


Figure 29: Proportion des CNV détectés localisés sur le gène *PMS2* et le pseudogène *PMS2CL*.

En filtrant le gène *PMS2* et le pseudogène *PMS2CL*, ExomeDepth rapportait 51 CNV pour l'ensemble des patients. Ces 51 CNV comprennent 23 supposés faux positifs et la totalité des 28 CNV à détecter. La première étape du travail d'optimisation consiste donc à améliorer la spécificité sur le gène *PMS2* et le pseudogène *PMS2CL*.

1.3.1. Optimisation d'ExomeDepth concernant *PMS2* et *PM2CL* :

Il nous faut décrire les spécificités génétiques imposés par *PSM2* et son pseudogène *PSM2CL*. Comme l'illustre la figure 30, les exons 9 et 11 à 15 de *PMS2* ont une grande homologie de séquence avec le pseudogène nommé *PMS2CL*.

Pour cette raison, l'étape d'alignement des lectures est particulièrement complexe et peut être source d'erreurs. Les régions possédant une grande homologie de séquence représentent une difficulté pour tous les logiciels de détection des CNV, ainsi que pour les autres techniques comme la MLPA.

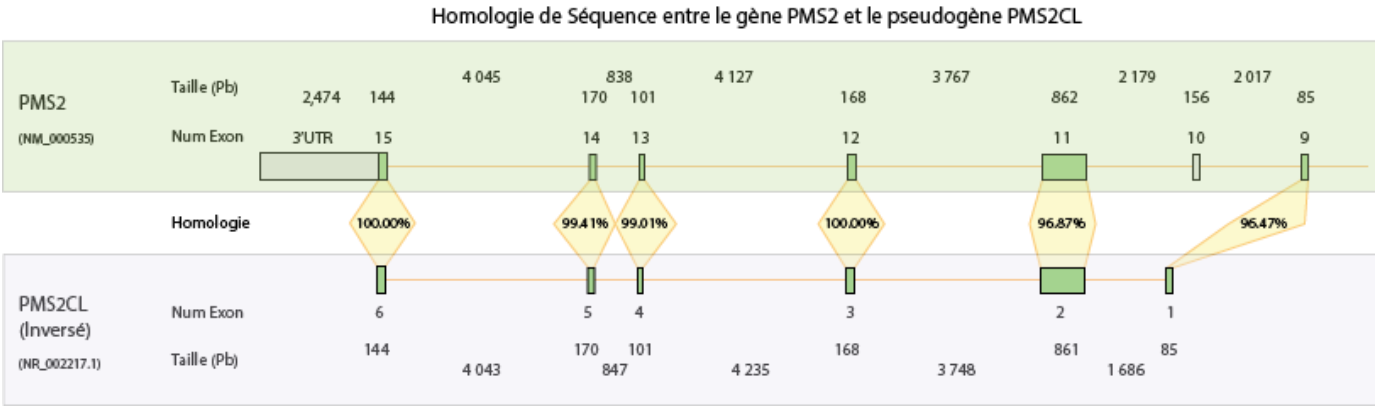


Figure 30: Illustration des exons ayant une homologie de séquence entre le gène *PSM2* et le pseudogène *PMS2CL*. Pour des raisons d'affichage, le gène *PMS2CL* est représenté en position inversée.

Quand deux régions sont homologues, voir identiques comme les exons 15 de *PMS2* et 6 de *PMS2CL*, le logiciel qui procède à l'alignement ne peut pas déterminer si la lecture provient du séquençage de l'une ou de l'autre de ces régions. C'est une limite connue des logiciels d'alignement.

Différentes solutions existent pour y parer :

- une répartition équitable entre les deux régions,
- une répartition selon un ratio prédéterminé,
- la recherche de SNP spécifiques de l'une des deux régions afin de guider la répartition.

Lorsque l'on considère des variations à l'échelle d'une seule base ce phénomène génère principalement des faux positifs. Quand une variation bénigne est localisée sur *PSM2CL* mais que les lectures issues de son séquençage sont alignées sur *PSM2*, un faux positif est alors généré. En ce qui concerne la recherche de CNV, l'approche DoC est faussée par un alignement non optimal.

En cas de répartition équitable des lectures, si on considère une région délétée à l'état hétérozygote (50% de la profondeur attendue) possédant une forte homologie avec une région non délétée (100% de la profondeur attendue). Les lectures étant réparties sur les 2 régions de manière équitable, la profondeur observée sera de 75% sur chacune des deux régions. Cela peut conduire à des faux positifs ou des faux négatifs.

Quand la répartition est faite selon un certain ratio prédéterminé ou un procédé plus complexe, la profondeur observée est imprévisible. C'est ce que nous observons lors de l'utilisation d'ExomeDepth. Dans une grande proportion des cas, une délétion rapportée sur *PMS2* est associée chez le même patient à une duplication sur la région homologue de *PMS2CL*. Peu de solutions sont proposées dans la littérature mais quelques-unes ont tout de même été trouvées. Dans son étude parue en 2017, Kerkhof propose de résoudre ce problème en pratiquant une méthode de normalisation à 4 allèles (91). À savoir de procéder à l'addition du nombre de lectures entre les régions homologues de *PMS2* et *PMS2CL* **avant** l'emploi de l'algorithme de normalisation de ExomeDepth, puis de raisonner non plus selon un modèle à 2 allèles, mais à 4 allèles.

Les délétions sont donc définies par une profondeur normalisée moyenne inférieure ou égale à 0,8 (3/4 allèles), tandis que les duplications sont définies par un ratio supérieur ou égal à 1,2 (5/4 allèles). Cette méthode permettrait d'identifier la présence d'un CNV sur *PMS2* et *PMS2CL* tout en réduisant fortement le nombre de faux positifs. Ensuite, la localisation précise des CNV rapportés entre le gène *PMS2* et son pseudogène *PMS2CL* sera faire par vérification MLPA qui, nous le rappelons, est systématique dans notre laboratoire en cas de suspicion de CNV.

Nous avons donc modifié légèrement le code de lancement d'ExomeDepth pour permettre l'addition de la profondeur sur les régions homologues de *PMS2* et *PMS2CL*.

Après l'addition de cette étape d'optimisation, ExomeDepth possède une meilleure spécificité. Concernant l'ensemble des échantillons de notre set, et en ne considérant que les gènes de routine, ED rapporte 56 CNV, soit 2,15 CNV par run en moyenne. Chaque run présente en moyenne 2,05 patients avec au moins un CNV. Chaque patient présentant en moyenne 0,045 CNV. Le tableau ci-dessous rapporte ces statistiques.

	Avant optimisation	Après optimisation
Nombre total de CNV rapportés	671	56
Nombre moyen de CNV par « run »	25,8	2,15
Nombre moyen de patients présentant au moins un CNV par « run »	20,2	2,05
Nombre moyen de CNV par patient pour l'ensemble des échantillons testés.	0,538	0,045

Tableau rapportant les statistiques des résultats d'ExomeDepth avant et après phase d'optimisation.

Sur les 30 CNV de notre set de référence, 29 sont toujours détectés correctement par l'analyse. Le CNV manquant est toujours absent des résultats. La figure 31 représente la répartition de ces 56 CNV sur les différents gènes de routine après la première phase d'optimisation.

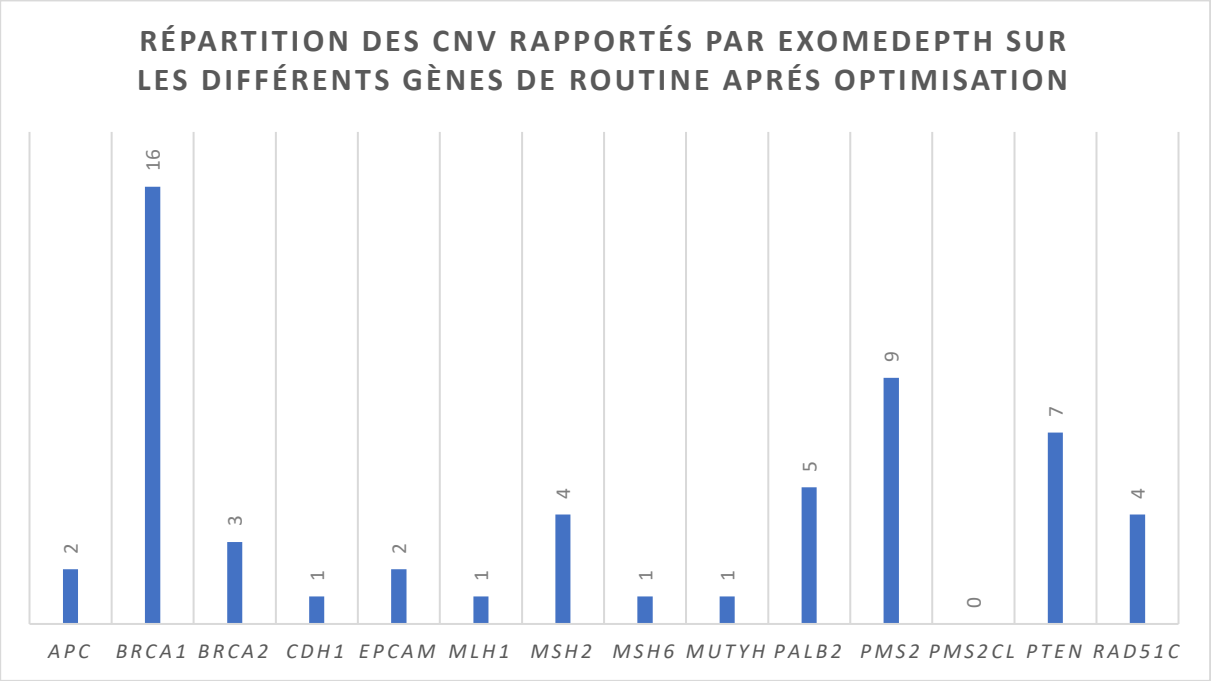


Figure 31: Illustration graphique de la répartition sur les différents gènes de routine du laboratoire des 64 CNV rapportés par ExomeDepth lors de l'analyse du set d'échantillons.

Nous remarquons une diminution très importante du nombre de CNV rapportés sur *PMS2* et *PMS2CL* mais également des changements sur des gènes qui ne sont pas directement concernés par les modifications effectuées dans le code d'ExomeDepth. En effet, si nous nous intéressons aux CNV détectés sur *PTEN* nous pouvons observer 3 CNV qui ne sont plus rapportés. A

l'inverse, sur le gène *RAD51C*, un nouveau CNV a été rapporté. Sur le gène *MSH2*, 2 CNV sont également absents dans la nouvelle analyse.

En effet la modification effectuée sur PMS2 et PMS2CL a entrainé un changement dans le choix des références par ExomeDepth lors de son analyse. Après notre modification les 10 échantillons les plus proches de l'échantillon analysés n'étaient plus les mêmes.

1.3.2. Limites analytiques : duplication des exons 11 et 12 de *PMS2* :

Un seul CNV de notre panel de référence reste indétectable avec le logiciel choisi malgré notre phase d'optimisation. Il s'agit d'une duplication des exons 11 et 12 du gène *PMS2*.

Les figures 32 et 33 ci-dessous représentent le ratio rapporté par ExomeDepth entre la profondeur observée et attendue sur chaque exon de *PMS2*. Notre optimisation a permis de recentrer les valeurs autour d'un équilibre (Ratio égal à 1). Cependant les ratios de profondeur des exons 11 et 12, bien que supérieurs à ceux des autres exons analysés, ne sont pas suffisamment élevés pour dépasser la valeur de 1,25 (5/4 copies) et être détectés comme significativement plus élevés.

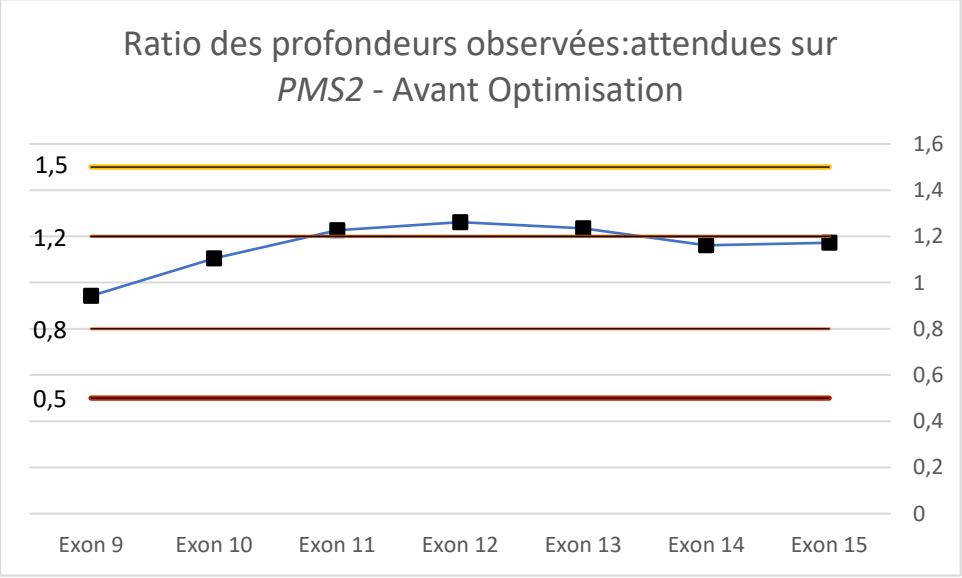


Figure 32: Ratio des profondeurs observées et attendues sur chacun des exons de *PMS2* présentant une homologie de structure avec *PMS2CL*. Ces données sont celles obtenues avant la phase d'optimisation.

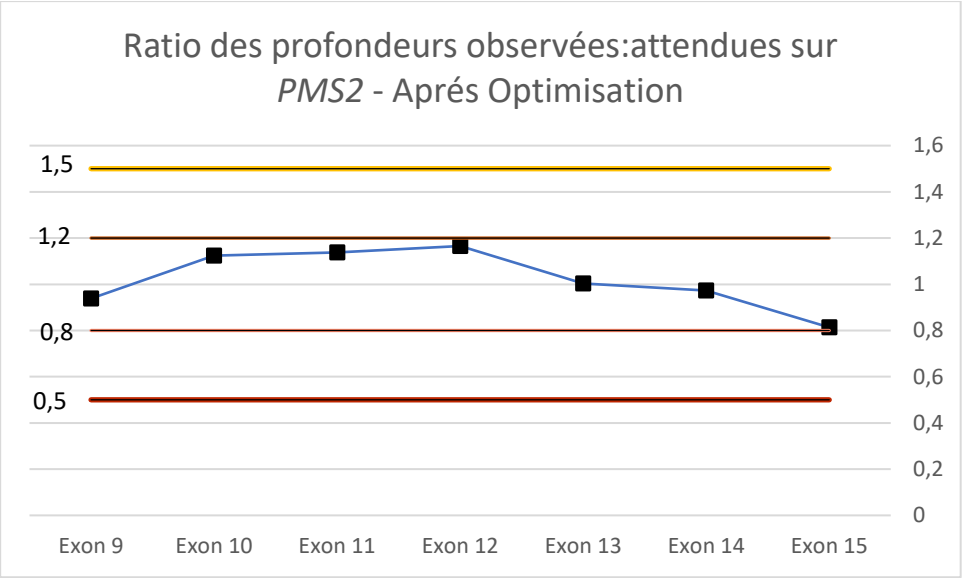


Figure 33: Ratio des profondeurs observées et attendues sur chacun des exons de *PMS2* présentant une homologie de structure avec *PMS2CL*. Ces données sont celles obtenues après la phase d'optimisation.

Ce CNV représente une double difficulté pour notre logiciel. La première est sa localisation sur le gène *PMS2*, sur une région présentant une forte homologie avec le pseudogène *PMS2CL*. La seconde vient du fait qu'il s'agisse d'une duplication, qui sont plus difficiles à détecter que les délétions. La difficulté de détection des duplications par les logiciels basés sur l'approche DoC est en lien avec la difficulté à détecter le passage d'un nombre de copie N à N+1. Ceci est d'autant plus vrai que N est élevé (ici N est égal à 4, car nous avons additionné les deux copies présentent sur le gène *PMS2* et son pseudogène *PMS2CL*).

Ce résultat met en évidence la limite de notre analyse, la détection d'un évènement sur le gène *PMS2* nécessite des attentions particulières. Quant à la détection des duplications, bien qu'elles soient décrites comme plus difficiles à mettre en évidence, ce cas est le seul dans lequel ExomeDepth s'est révélé insuffisant.

1.3.3. Phase 2 de l'optimisation d'ExomeDepth :

Après avoir réduit considérablement le nombre de CNV rapportés, nous avons étudié les supposés faux positifs restants. Pour cela un tri a été effectué par le nombre d'occurrences dans l'ensemble des runs et au sein de chaque run. En effet les CNV recherchés dans le domaine de l'oncogénétique moléculaire sont des CNV très rares. Nous précisons que les trios, ou les membres d'une même famille, n'ont pas été analysés conjointement.

Le CNV plus fréquemment rencontré correspond alors à une délétion de l'Exon 2 sur le gène *PTEN* avec un nombre d'occurrence de 5. Vient ensuite, avec un nombre d'occurrence total de 3 la délétion des exons 15 et 16 de *BRCA1*. Ces 3 cas sont cependant tous de vrai positifs. La délétion de l'exon 11 du gène *PALB2* possède également un nombre d'occurrence égal à 3. Ce sont également tous des vrais positifs. Aucun autre CNV n'avait un nombre d'occurrence supérieur à 2. Le tableau ci-dessous reprend ces informations.

CNV	**Nombre d'occurrence dans l'ensemble du set d'échantillons.**
Délétion de l'exon 2 de PTEN	5
Délétion des exons 15 et 16 de BRCA1	3
Délétion de l'exon 11 du gène PALB2	3

Tableau rapportant le nombre d'occurrence des CNV les plus fréquemment rapportés par ExomeDepth après phase d'optimisation.

Nous avons donc entrepris une réflexion autour de cette délétion récurrente survenant sur l'exon 2 de *PTEN*. *PTEN* possède un pseudogène nommé *PTENP1*. Cependant l'exon 2 n'est pas un exon qui possède une forte homologie de séquence avec une région de ce pseudogène. Cette théorie ne suffit donc pas à expliquer les résultats observés. Le score de Bayes moyen pour ces 5 cas était de 8,554 avec un maximum à 15,8. Le logiciel ExomeDepth rapporte ce CNV comme une délétion sur le chromosome 10 de la base 89653757 à la base 89653891.

PTEN exon 2 n'a donc pas d'équivalent sur le pseudogène et ce CNV pourrait être le fait d'un polymorphisme. Un polymorphisme consistant en une délétion de 899pb et se terminant en -58 de la première base de l'exon 2 de *PTEN* est en effet rapporté dans la littérature. Dans une étude parue en 2013 Sandell retrouve ce polymorphisme chez environ 4% des patients de sa cohorte (92). Cette délétion perturberait la liaison des sondes de MLPA et la liaison des sondes de PCR utilisées dans son étude. Cela entraine l'apparition d'une fausse délétion de l'exon 2 de *PTEN*.

Nous avons donc procédé à l'évaluation des logiciels préalablement sélectionnés. Puis avons ensuite étudié en détail le fonctionnement d'ExomeDepth pour l'adapter à notre activité de routine. Les résultats obtenus sont satisfaisants et concordants avec nos objectifs initiaux.

En effet ce logiciel est fiable et rapide. L'étape suivante consiste donc à l'intégrer au mieux dans le pipeline préexistant. Pour cela nous l'avons automatisé. Nous avons ensuite écrit et validé les procédures qualité nécessaires.

1.3.4. Automatisation de l'exécution logicielle

Le script développé pour automatiser l'utilisation de ExomeDepth est constitué de 2 parties. La première est un script linux simple qui liste l'ensemble des fichiers format « .BAM » d'un répertoire (ou un ensemble de répertoires). Cette première partie de script est exécutée par une simple ligne de commande. Cette liste sert d'entrée au logiciel ExomeDepth. La seconde partie est écrite en R-script et est automatiquement exécutée par le premier script pour chaque liste de fichier « .BAM ». Elle procède aux différentes étapes nécessaires à l'analyse par ExomeDepth. Enfin une troisième phase procède à la modification de la structure des résultats pour générer un fichier tableau (format .csv ou Excel) de l'ensemble du run (ExomeDepth générant quant à lui un fichier résultat par échantillon).

Notre idée de départ était de combiner plusieurs logiciels pour associer leurs points forts et augmenter la sensibilité et spécificité de l'analyse des CNV dans notre laboratoire, pour cela nous voulions utiliser plusieurs approches théoriques différentes de détection des CNV. Cependant à la suite de notre expérimentation nous concluons que, dans le cas de notre laboratoire, avec l'utilisation d'un séquençage par panel de gènes avec une technologie à lecture courte, une seule des approches théoriques est réellement applicable : l'approche DoC. Bien que notre idée de départ se soit révélée inapplicable, l'utilisation d'un seul logiciel basé sur l'approche DoC permet déjà d'approcher une qualité optimale d'analyse.

L'utilisation de logiciels basés sur les approches SR et PEM (Smoove et Delly dans notre cas) était une idée intéressante, mais qui s'est révélée à la fois difficile à mettre en place, et comportant de nombreuses limites. Cette expérimentation nous a permis de changer notre point de vue initial et de privilégier une analyse réalisée par un logiciel unique optimisé spécifiquement pour notre laboratoire.

Discussion

L'idée à l'origine de ce travail est de permettre à notre laboratoire de bénéficier d'un second pipeline bioinformatique pour la détection des CNV dans les données de NGS. Ce qui permet un contrôle qualité des résultats générés par la société Sophia Genetics actuellement en charge de l'analyse bioinformatique de l'ensemble de nos données issues de NGS.

Le nouveau procédé d'analyse devait être :

- fiable,
- rapide,
- intégrables à notre pipeline bioinformatique,
- compréhensible,
- sans entrainer de surcoûts démesurés ni nécessiter trop de temps en personnel technique, bio-informaticien et biologiste.

Notre travail avait pour objectif une application directe dans un laboratoire de routine, ce qui explique la considération de contraintes supplémentaires que ne rencontrent pas les laboratoires de recherche. Pour la réalisation de ce projet nous avons pu bénéficier du travail et de l'expérience de nombreux développeurs qui ont créé, et mis à disposition de tous, des logiciels à la fois performants, open source, et gratuits. Nous tenons à rappeler que ce travail n'aurait pas été réalisable sans leur participation indirecte.

La mise à disposition d'une grande quantité de logiciels spécialisés dans la détection des CNV nous a permis de procéder à la sélection et l'évaluation de 5 d'entre eux. Ces 5 logiciels ont été sélectionnés vis-à-vis de leur correspondance avec nos objectifs et nos capacités matérielles. Nous avons ensuite procédé à une évaluation nous permettant de juger de la performance de ces logiciels vis-à-vis de nos données et de critères qui étaient propres à notre mode de fonctionnement. Nous précisons cependant qu'il ne s'agissait en aucun cas d'une comparaison de la performance de ces logiciels entre eux, ni d'évaluer leur sensibilité et spécificité.

Les résultats de cette évaluation nous ont mené à modifier notre idée initiale, l'association de plusieurs logiciels basés sur différentes approches n'étant pas à la hauteur de nos attentes. Une telle association logicielle aurait entrainé une complexification considérable des processus et de l'interprétation des résultats, sans bénéfice sur la performance analytique.

Nous avons donc opté pour l'utilisation d'un logiciel unique, basé sur l'approche DoC qui est la plus adaptée à notre mode de séquençage en panel de gènes. S'en est suivit une phase d'optimisation de l'usage de ce logiciel, en vue d'une correspondance maximale avec nos objectifs. Nous avons donc validé des changements prudents dans le code informatique après de nombreuses recherches dans la littérature. Ces changements nous ont permis de réduire considérablement le nombre de faux positifs relatifs la présence de gènes liés à des pseudogènes dans notre panel.

En conclusion nous disposons d'un logiciel de détection des CNV dans les données de NGS performant et dont nous connaissons les limites. L'intégration de celui-ci dans notre pipeline bioinformatique s'est faite sans difficultés et son utilisation en routine a pu être initiée.

Ce travail a également mis en lumière l'expertise des bio-informaticiens, piliers des analyses NGS. En effet chaque tâche informatique nécessite une évaluation experte car chaque changement sur un élément du pipeline se répercute sur l'ensemble du process. Ce travail ne rapporte pas l'ensemble des difficultés rencontrées, nombreux ont été les « bugs », entrainants un arrêt des analyses, des résultats erronés, ou l'impossibilité d'utiliser certains logiciels. Un bio-informaticien maitrisant les logiciels utilisés est un moyen de générer des résultats fiables. L'expertise biologique reste évidemment un gage de qualité. La parfaite maitrise des gènes contenus dans un panel est primordiale. La présence ou non de pseudogènes, de polymorphismes ou de séquence ALU peuvent être des sources d'interférences analytiques. L'expertise du biologiste permet de guider l'analyse bioinformatique.

Nous retiendrons donc en marge de l'objectif principal de ce travail que la génétique à l'ère du séquençage haut débit nécessite une collaboration étroite entre deux domaines d'expertise, les bio-informaticiens et les biologistes spécialisés. Ces deux domaines d'expertise sont étroitement liés et interdépendants.

Conclusion

Ce travail permet de démontrer la faisabilité de la mise en place d'outils bioinformatiques de détection des CNV à partir de données de NGS. L'importante quantité de logiciels disponibles sur le sujet est une chance par l'accès à une grande palette de différents programmes mais renforce également la difficulté à faire un choix.

En effet ces dernières années le nombre de logiciels développés a augmenté de manière importante. Il existe plusieurs centaines de logiciels sur le sujet. Malgré leur objectif partagé, ces différents programmes diffèrent sur tout un ensemble de critères, de la performance à la rapidité en passant par le langage informatique utilisé ou encore la documentation disponible. Cette grande diversité complique la démarche de choix éclairé, d'autant plus que la communication concernant les performances analytiques est souvent orientée dans les publications initiales. Plusieurs études ont tenté de réaliser une comparaison désintéressée entre plusieurs logiciels mais il est important de remarquer qu'il serait impossible de réaliser un tel type d'étude en incluant la totalité des logiciels disponibles.

Si ces logiciels présentent une grande variabilité de performance c'est parce qu'ils ont été développés pour correspondre à une technologie de séquençage particulière, dans un contexte spécifique (constitutionnel, somatique) ou encore à des échantillons particuliers (génomes bactériens, eucaryotes non humains, végétaux, ou viraux). Mais dans chaque cas, ils reposent sur l'une des 4 approches théoriques qui ont chacune leurs avantages et limites :

- L'approche PEM est capable d'identifier presque tous les types de SV, mais sa précision est dépendante de la taille de l'insert définie par la librairie utilisée (93).
- L'approche Split-Read permet de détecter avec précision les points de cassure, mais dont la taille des SV détectables est limitée par la longueur des lectures (71).
- La méthode par assemblage de novo qui consiste à générer une longue séquence à partir de lectures courtes puis à la comparer au génome de référence est applicable principalement avec les technologies à lectures longues et est limitée quand il s'agit de détecter les duplications ou les répétitions (5).
- L'approche DoC est la plus performante pour détecter les délétions et duplications mais ne peut détecter que les CNV parmi l'ensemble des SV. Elle ne permet pas d'identifier avec précision les points de cassure (94).

L'idée de combiner plusieurs approches semble donc logique. Cependant dans certains cas l'association de différentes approches n'entraine pas de bénéfice conséquent. Il reste donc important de juger spécifiquement chaque situation et de fixer ses objectifs avant le choix de son logiciel ou de son association de logiciels. Il faut en connaitre les points forts, mais plus encore les limites. Il faut également maitriser l'ensemble des étapes bioinformatiques qui mènent à partir d'un ensemble de lectures à la génération de résultats d'une recherche de SV.

Le choix d'un logiciel n'est que la première étape de la mise en place d'un pipeline bioinformatique de détection des CNV. La qualité des résultats doit être assurée. Il est donc nécessaire de maitriser à la perfection à la fois le logiciel utilisé, la technologie de séquençage mais aussi la biologie des gènes soumis à l'analyse. L'exemple le plus évident est celui de la séparation des membres d'une même famille lors de l'analyse des CNV. Cette connaissance des limites informatiques et biologiques ainsi que des évènements à l'origine de faux positifs permet des adaptations. Ces dernières nécessitent alors à leur tour d'être validées.

C'est cette complexité importante, associée à un domaine technologique très évolutif, qui est probablement la cause de l'absence de recommandations officielles nationales ou internationales sur le sujet à ce jour. Plusieurs groupes de travail se sont d'ores et déjà emparés du sujet. La communauté « *Human Copy Number Variation Community* » de l'organisation intergouvernementale européenne ELIXIR travaille actuellement à définir un pipeline optimal de détection des CNV. Il est donc probable que les travaux de ces groupes d'experts permettent une standardisation des méthodes de détection des CNV dans les données de NGS au cours des prochaines années. Il ne faut cependant pas écarter la possibilité qu'il faille attendre que le cycle de vie technologique du séquençage haut débit atteigne un palier avant de pouvoir produire des recommandations qui resteraient applicables dans la durée. L'engouement pour les technologies de séquençage est important depuis plusieurs années et certaines technologies prometteuses sont encore en phase de maturation. On peut aisément imaginer une nouvelle révolution technologique dans un futur proche.

Il faut également noter l'absence d'échantillons de référence disponibles pour l'étude des SV. Cette lacune est un frein pour la standardisation des logiciels et la rédaction de recommandations. Sans référence, il est difficile de réaliser une comparaison de performance entre différents logiciels. Aujourd'hui la multiplication des données tend à combler ce vide. Certains échantillons issus du 1000GP ont été séquencés par de multiples méthodes, par différentes technologies et ont bénéficié d'une étude approfondie des SV notamment grâce aux puces à ADN de très haute résolution.

La mise en place d'un outil d'analyse des CNV n'est en réalité qu'une première étape vers l'objectif de détection de l'ensemble des évènements d'une taille supérieure à quelques nucléotides. La recherche et l'évaluation de la pathogénicité d'évènements plus complexes comme des inversions, l'insertion de séquences ALU, ou d'autres petits éléments nucléaires intercalés reste un défi pour les biologistes moléculaires à l'avenir. La détection de ces évènements de grande taille ainsi que des évènements introniques impose aux biologistes moléculaires de maitriser les bases de la bioinformatique et de travailler de pair avec les bio-informaticiens pour relever les défis de demain.

Index des figures

Figure 1 : Schématisation des différents types de variations structurelles. 1 - Délétion, 2 - Duplication en tandem (ou répétition en tandem) ; 3 - Insertion : 4 - Amplification ; 5 - Duplication dispersée (ou répétition dispersée) ; 6 - Inversion ; 7 - Translocation ; 4
Figure 2: Difficultés d'alignement des lectures en lien avec la présence de transposons, la présence des transposons en plusieurs exemplaires, et de localisation variable, entraine des erreurs lors de l'application des logiciels d'alignement. *Source de la figure : Mobile element insertion (MEI) detection for NGS based clinical diagnostics – SeqOne (24)* 6
Figure 3 : Illustration du mécanisme de recombinaison allélique non homologue (NAHR) interchromatidien : a- Les rectangles représentent les régions hautement répétées ou LCR. Les rectangles blancs et noirs correspondent à deux LCR distincts qui possèdent un haut degré de similarité ; b – Entre les deux LCR peut se produire un mésappariement lors de la méiose ; c - Quand un échange (*crossing over*) non allélique se produit, on observe la création d'un gamète avec une microdélétion et un gamète avec une microduplication. *Source : Poisson A, Nicolas A, Sanlaville D, Cochat P, Leersnyder HD, Rigard C, et al. Le syndrome de Smith-Magenis, une association unique de troubles du comportement et du cycle veille/sommeil.* 8
Figure 4: Illustration des 4 étapes des NHEJ : 1- identification d'une cassure double-brins ; 2- Création de ponts moléculaires entre les deux terminaisons cassées ; 3- Légère modification des terminaisons pour en augmenter la compatibilité (digestion enzymatique) ; 4- Ligation des deux brins. *Source : Gu, Wenli & Zhang, Feng & Lupski, James. (2008). Mechanisms for human genomic rearrangements. PathoGenetics. 1. 4. 10.1186/1755-8417-1-4.* 9
Figure 5: Résultat d'un caryotype humain masculin normal après tri des chromosomes en paires et de 1 à 22 (+X et Y). *Source : Laboratoire de cytogénétique du Centre Hospitalier Universitaire de REIMS – 04/12/1999.* 10
Figure 6: Le Giemsa banding (G-banding) utilise un colorant chimique qui révèle des bandes sombres sur les chromosomes métaphasiques. Ces méthodes permettent de détecter des anomalies structurales de grande taille (> 3Mb), on peut voir ici une inv(9qh). *Source : Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97. doi :10.1038/nrg1767.* 11
Figure 7: Sur l'image de gauche, une FISH sur chromosomes en métaphase sur laquelle on observe une translocation t(3;7) cryptique, un exemple d'anomalie structurelle équilibrée. Sur l'image de droite, une FISH sur noyaux en interphase sur laquelle on observe une micro-inversion 7p22 de 700 kb. C'est un exemple de SV bénigne (polymorphisme). 11
Figure 8: La peinture chromosomique (spectral karyotyping) est utile pour détecter les réarrangements interchromosomiques. Ici une translocation t(7;13). *Source : Feuk, L., Carson, A. R., & Scherer, S. W.*

(2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97. doi:10.1038/nrg1767..... 12
Figure 9 : Illustration des résultats d'une puce à ADN : le rectangle blanc représente un agrandissement. *Source : Ahn JW, Coldwell M, Bint S, Mackie Ogilvie C. Array comparative genomic hybridization (array CGH) for detection of genomic copy number variants. J Vis Exp. 2015*. 13
Figure 10: Illustration d'une méthode de PCR multiplexe pour la détection des CNV, Technique Multiplex Ligation-dependent Probe Amplification (MLPA). *Source : Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85–97. doi:10.1038/nrg1767*..... 15
Figure 11: Composition des adaptateurs Illumina (modèle *paired-end*). SP : Amorce (SP+P5/P7 = 95pb) - P5 : Adaptateur commun - P7 : Adaptateur commun contenant l'index – Index : Fragment spécifique de chaque patient (6pb) – Insert : Fragment d'intérêt. Source : Jacques V. L'intégration du séquençage de nouvelle génération dans le diagnostic médical : application aux leucémies aiguës myéloïdes et syndromes myélodysplasiques (51)..... 17
Figure 12: A- La préparation de librairies par fragmentation : le matériel génétique est découpé en fragments de différentes tailles par un procédé qui peut être mécanique, chimique ou enzymatique. Une sélection des fragments par la taille est réalisée. La dernière étape consiste en l'ajout d'étiquettes aux extrémités. B- La préparation de librairies par amplification : des régions d'intérêt sont amplifiées par PCR grâce à des amorces spécifiques. Des étiquettes sont ajoutées aux extrémités des amplicons par ligation ou par PCR. *Source de la figure : https://www.biomnigene.fr/.*..... 18
Figure 13 : Illustration de la technologie de PCR par d'émulsion. L'étape de PCR est effectuée sur une bille, à l'intérieur d'une micelle, recouvrant chaque bille de milliers de copies de la même séquence d'ADN. Source : *Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).*..... 19
Figure 14 : Illustration de la technologie « bridge PCR » d'Illumina. *Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).*..... 19
Figure 15: Illustration des distances calculables au sein d'une paire de lectures paired-end. Source : *Ségolène CABOCHE, Gaël EVEN – Paired-End versus mate-pair v1.0 17.02.2012 – Ressource en ligne - http://www.biorigami.com/* 20
Figure 16: Illustration du principe du séquençage Paired-End tel qu'appliqué dans les technologies Illumina. *Source :* « *Single-read sequencing - Understand the key differences between these sequencing read types* », *Ressource en ligne : https://emea.illumina.com/.* 21
Figure 17: Illustration de la technologie commercialisée par Pacific Biosciences (PacBio). *Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).*..... 24

Figure 18: Illustration de la technologie commercialisée par Oxford Nanopore Technologies (ONT). Source : *Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).*..... 25
Figure 19: Illustration des technologies synthétiques. Ba : Illumina ; Bb : Séquençage en émulsion de 10X Genomics. *Source : Goodwin, S., McPherson, J. & McCombie, W. Coming of age : ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).*..... 27
Figure 20: Principales étapes d'un pipeline bioinformatique utilisé pour l'analyse des données de séquençage à haut débit. *Source : Grzych G. Evaluation of in silico prediction tools and interest of functional tests in the interpretation of identified variants by next-generation sequencing in human genetics. 2018.*..... 29
Figure 21: Illustration de l'assemblage *de novo* d'un génome entier. Grâce au chevauchement existant entre les lectures, on peut assembler des contigs, les espaces vides (*gaps*) peuvent ensuite être comblés grâce aux lectures longues. Il est enfin possible de procéder à une ou plusieurs étape(s) de « *gap-filling* » pour analyser les derniers espaces vides. Il est ainsi possible de reconstruire des chromosomes et un génome, qui peut être soit complet, soit provisoire avec la persistance de certains *gaps*. *Schéma extrait et traduit depuis la publication : Sohn, jang-il & Nam, Jin-Wu. (2016). The present and future of de novo whole-genome assembly. Briefings in Bioinformatics. 19. bbw096. 10.1093/bib/bbw096.* 31
Figure 22: Distribution de la taille de l'insert (en pb). L'approche PEM impose la fixation d'un seuil inférieur et d'un seuil supérieur. Une taille comprise à l'extérieur de ces seuils indique la présence d'une SV. Statistiques extraites d'un fichier BAM représentatif du groupe de notre étude..... 38
Figure 23 : : Illustration des différentes configurations que peuvent prendre les paires de lectures dans les différents types de SV. 1- Délétion : Quand une paire de lectures encadre la région délétée, la taille de l'insert après alignement sera nettement supérieure à celle attendue. 2- Insertion : Quand une paire de lectures encadre la région insérée, la taille de l'insert après alignement sera nettement inférieure à celle attendue. 3- Inversion : Une inversion peut être détectée quand une des lectures d'une paire se trouve sur la région inversée, l'orientation de cette lecture sera alors incorrecte. 4- Duplication en tandem : Quand les deux lectures d'une paire encadrent le point central d'une duplication en tandem, on retrouve après alignement ces deux lectures dans une orientation correcte, mais avec une position inversée l'une par rapport à l'autre. 5- Translocation : Pour détecter une translocation, il faut que deux paires de lectures encadrent chacune un point de cassure. Dans ce cas, après alignement, on retrouve chaque lecture dans une orientation correcte, mais avec des positions les unes par rapport aux autres bouleversées. Si la translocation est interchromosomique, les deux extrémités de chaque paire se retrouvent alignées sur un chromosome différent. 6- Paires de lectures à ancrage unique : Quand, pour deux paires de lectures, une seule lecture est correctement alignée, il est possible que ces dernières encadrent en réalité un fragment d'ADN dont la séquence n'existe pas sur le génome de référence. *Figure traduite et inspirée de la publication Xi, Ruibin & Kim, Tae-Min & Park, Peter. (2010). Detecting*

structural variations in the human genome using next generation sequencing. Briefings in functional genomics. 9. 405-15. 10.1093/bfgp/elq025. (69)..... 39
Figure 24: Illustration de l'analyse basée sur l'approche SR. Les SV peuvent être détectées lorsque des lectures recouvrent le point de cassure. Cette approche permet une détection directe des délétions et duplications/insertions. Les réarrangements plus complexes sont identifiés par une réanalyse globale en présence de plusieurs évènements de type délétions/insertions. Traduit de la publication *Identification of genomic indels and structural variations using split-read* (71). 41
Figure 25 : Illustration de l'approche par profondeur de lecture (DoC) – A : Absence de CNV ; B : Duplication de deux exons (2 et 3) ; C : Délétion complète de l'exon 4. La profondeur de lecture est déterminée par comptage des lectures au sein de régions prédéfinies non chevauchantes (par exemple ici un exon). La profondeur observée dans cette fenêtre est ensuite comparée à une valeur théorique issue d'une ou plusieurs références. *Source : 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 (2020).* (79)..... 43
Figure 26: Illustration des termes « Couverture » et « Profondeur » de séquençage. Source : Lacoste, C. & Fabre, Alexandre & Pécheux, C. & Lévy, Nicolas & Krahn, Martin & Malzac, P. & Bonello-Palot, N. & Badens, Catherine & Bourgeois, Patrice. (2017). Le séquençage d'ADN à haut débit en pratique clinique. Archives de Pédiatrie. 24. 10.1016/j.arcped.2017.01.008 (80). 44
Figure 27: Illustration des différentes annotations d'un point de cassure en fonction des différentes possibilités de réarrangements. Cas A : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant à partir de ce point correspond à celle se trouvant en aval de la position « p » (position 800 du chromosome 1). Cas B : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant avant ce point correspond à celle se trouvant en amont de la position « p » (position 800 du chromosome 1). Cas C : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant avant ce point correspond à celle se trouvant en aval de la position « p » (position 800 du chromosome 1), il s'agit en réalité de la séquence complément inverse. Cas D : En position 500 du chromosome 1, aucune base nucléique n'est remplacée, la séquence s'étendant à partir de ce point correspond à celle se trouvant en amont de la position « p » (position 800 du chromosome 1), il s'agit de la séquence complément inverse. *Source : Merging Structural Variant Calls from Different Callers - simpsonlab.github.io* 65
Figure 28: Illustration graphique de la répartition des 671 CNV rapportés par ExomeDepth lors de l'analyse du set d'échantillons sur les différents gènes de routine du laboratoire..... 72
Figure 29: Proportion des CNV détectés localisés sur le gène *PMS2* et le pseudogène *PMS2CL*..... 72
Figure 30: Illustration des exons ayant une homologie de séquence entre le gène *PSM2* et le pseudogène *PMS2CL*. Pour des raisons d'affichage, le gène *PMS2CL* est représenté en position inversée. 73
Figure 31: Illustration graphique de la répartition sur les différents gènes de routine du laboratoire des 64 CNV rapportés par ExomeDepth lors de l'analyse du set d'échantillons. 75

Figure 32: Ratio des profondeurs observées et attendues sur chacun des exons de *PMS2* présentant une homologie de structure avec *PMS2CL*. Ces données sont celles obtenues avant la phase d'optimisation. 76
Figure 33: Ratio des profondeurs observées et attendues sur chacun des exons de *PMS2* présentant une homologie de structure avec *PMS2CL*. Ces données sont celles obtenues après la phase d'optimisation. 77

Index des tableaux

Tableau 1: Plusieurs exemples de séquences ALU impliquées en pathologie humaine. *Source : Kim, S., Cho, C.-S., Han, K., & Lee, J. (2016). Structural Variation of Alu Element and Human Disease. Genomics & Informatics, 14(3), 70. doi:10.5808/gi.2016.14.3.70 (22).* 5
Tableau 2: Tableau récapitulatif des limites analytiques des techniques de cytogénétique et des techniques basées sur la PCR-quantitative citées. 15
Tableau 3: Gènes séquencés dans le Panel HBOC. En bleu : les gènes faisant l'objet d'une analyse et d'un rendu de résultat en routine. 57
Tableau 4: CNV présents dans notre panel d'échantillons. La longueur supposée des CNV a été calculée en additionnant la longueur des introns bordant chaque point de cassure estimé. Quand le CNV se termine en amont de la région 5'UTR ou en aval du 3'UTR la taille n'est pas calculable et cette incertitude est représentée par un point d'interrogation. 59
Tableau 5: Capacité des différents logiciels testés à détecter correctement les 30 CNV de référence. O : CNV correctement détecté par le logiciel ; N : CNV non détecté par le logiciel. La codification O/N est utilisée quand un CNV est bien détecté par un logiciel, mais pour lequel la taille rapportée n'est pas compatible avec la réalité (*cf. Partie 4 - Chapitre 1.2*). 62

Annexe 1 : Format FASTQ

1. Description

Le format FASTQ est un format de fichier texte (type MIME : text/plain) qui permet le stockage concomitant de séquences d'acides nucléiques avec leurs scores de qualité associés.

Ce format a été développé par Jim Mullikin au Wellcome Trust Sanger Institute (fin du 20^{e} siècle) afin de lier à un fichier de séquence au format FASTA des données de qualité. Avec le séquençage haut débit, ce format est devenu le standard pour le stockage des données de séquenceurs. (95). Ce format possède également l'avantage d'être hautement compressible.

2. Structure du format FASTQ

Dans un fichier FASTQ, chaque séquence utilise 4 lignes (séparation : '\n').

La première ligne commence par le caractère arobase "@" suivie de l'identifiant de la séquence et éventuellement d'une description. La seconde ligne correspond à la séquence nucléique brute. La troisième ligne commence par le caractère plus "+" (parfois suivi par l'identifiant de la séquence et de sa description). La quatrième et dernière ligne contient les scores de qualité associés à chacune des bases de la séquence. En toute logique les lignes 2 et 4 font toujours exactement le même nombre de caractères, car chaque base nucléique et son score de qualité sont chacune codés avec un unique caractère ASCII.

Illustration du format FASTQ

Annexe 1 - Figure 1 – Structure des données au format FASTQ.

Annexe 2 : Formats SAM et BAM

1. Description

Le format « SAM » de cartographie d'alignement de séquence (*Sequence Alignment Map*) est un format texte tabulé utilisé pour stocker des séquences biologiques alignées sur une séquence de référence. Le format SAM a été développé par Heng Li et Bob Handsaker et *al* en 2009 (58). Il est aujourd'hui largement utilisé pour stocker les séquences nucléotidiques générées par les technologies de séquençage haut débit.

Le format SAM est un format hautement compressible et indexable, il est adapté aux lectures courtes ou longues produites par les différentes plateformes de séquençage et est utilisé pour conserver des données alignées sur une référence.

Architecturalement, il est composé d'un « header » en amont des « données d'alignement ».

Les lignes de l'en-tête commencent par le caractère arobase '@' suivit par l'un des codes à deux lettres parmi : HD (*header line*), SQ (*Reference sequence dictionary*), RG (*Read group*), PG (*Program*) et CO (*One-line text comment*). L'en-tête contient toutes les informations nécessaires sur l'origine du fichier, l'ordre dans lequel les données d'alignement sont triées, etc.

Chaque ligne contenant les données d'alignement est composée d'au moins 11 colonnes tabulées (séparées par le caractère '\t') qui contiennent l'essentiel des données sur l'alignement et des colonnes supplémentaires facultatives pouvant contenir diverses autres informations (Figure 1).

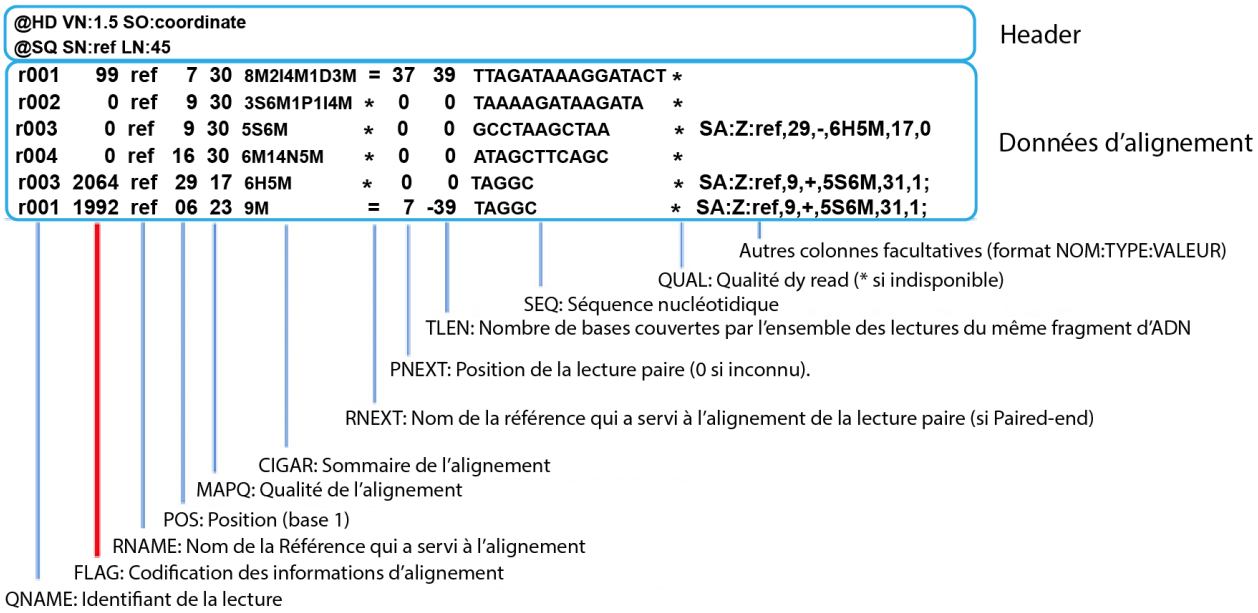
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Annexe 2 - Figure 1 – Les 11 colonnes obligatoires des données d'alignements du format SAM. String : Chaîne de caractère ; Int : Nombre intégral

La version binaire compressée du SAM s'appelle le format BAM (*Binary Alignment Map)*. La version ultra compressée est appelée CRAM.

2. Structure du format SAM

Illustration du format SAM



Annexe 2 - Figure 2 – Structure des données au format SAM.

Le format BAM est binaire et n'est pas illustré.

Annexe 3 : Format BED

1. Description :

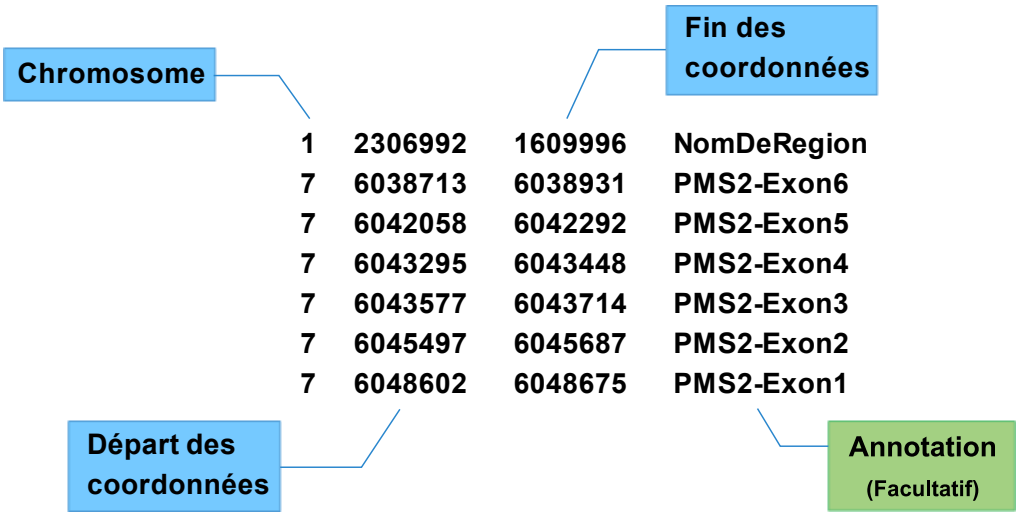
Le format « BED » (*Browser Extensible Data*) est un format de fichier texte tabulé utilisé pour stocker des régions génomiques sous forme de coordonnées avec leurs annotations associées. Il a été développé au cours du Projet Génome humain puis est rapidement devenu un standard en bioinformatique sans pour autant posséder de spécifications officielles.

Le format BED permet la manipulation de coordonnées en lieu et place de séquences nucléotidiques. Sa simplicité lui permet une accessibilité très importante autant par les outils de traitement de texte que par les langages de script (Python, Golang, Perl) même s'il existe des outils plus spécialisés (*ex. BEDTools*).

Architecturalement le format « BED » est en général constitué de 3 à 4 colonnes tabulées qui contiennent respectivement : La référence du chromosome, le point de départ des coordonnées, la fin des coordonnées, et facultativement une annotation associée à cette région. Le format BED n'étant pas verrouillé il est possible de créer des colonnes facultatives contenant tout type d'informations. De la même façon, il peut y avoir ou ne pas y avoir une ligne de « header » qui indique l'intitulé de chaque colonne.

2. Structure du format BED

Illustration du format BED



Annexe 3 - Figure 1 – Structure des données au format BED.

Annexe 4 : Format FASTA

1. Description :

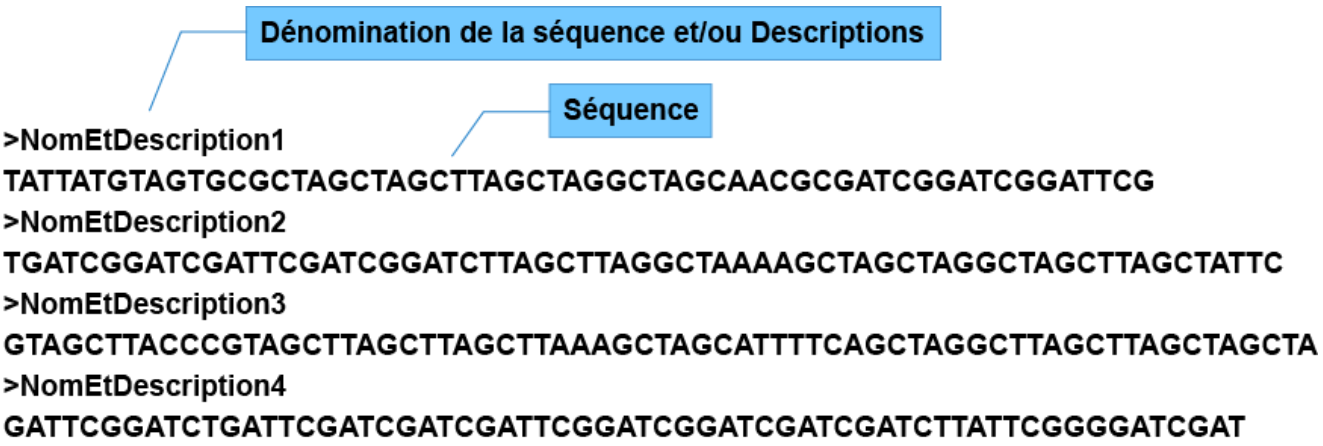
Le format FASTA (ou format Pearson) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique. Chaque séquence est représentée par une suite de lettres correspondant à chaque acide nucléique ou acide aminé et peut être précédée par une dénomination et/ou des descriptions. Ce format est originellement issu de la suite de programmes FASTA puis est devenu un standard en bioinformatique.

C'est un format texte dans lequel deux lignes sont utilisées pour chaque séquence : la première qui contient le nom ou une description commence toujours par un chevron supérieur « > », la ligne suivante contient la séquence en caractère.

La simplicité du format FASTA permet la lecture et la modification des séquences, que ce soit via l'utilisation d'outils de traitement de texte ou des langages script. Ce format présente également l'avantage d'être indexable. Les génomes de référence sont, par exemple, contenus dans des fichiers FASTA.

Un fichier au format FASTA est conventionnellement signalé par une extension « .fasta » ou « .fa ». Les index, quand ils sont créés, sont généralement au format « .fai ».

2. Structure du format FASTA :



Annexe 4 - Figure 1 – Structure des données au format FASTA.

Annexe 5 : Format VCF

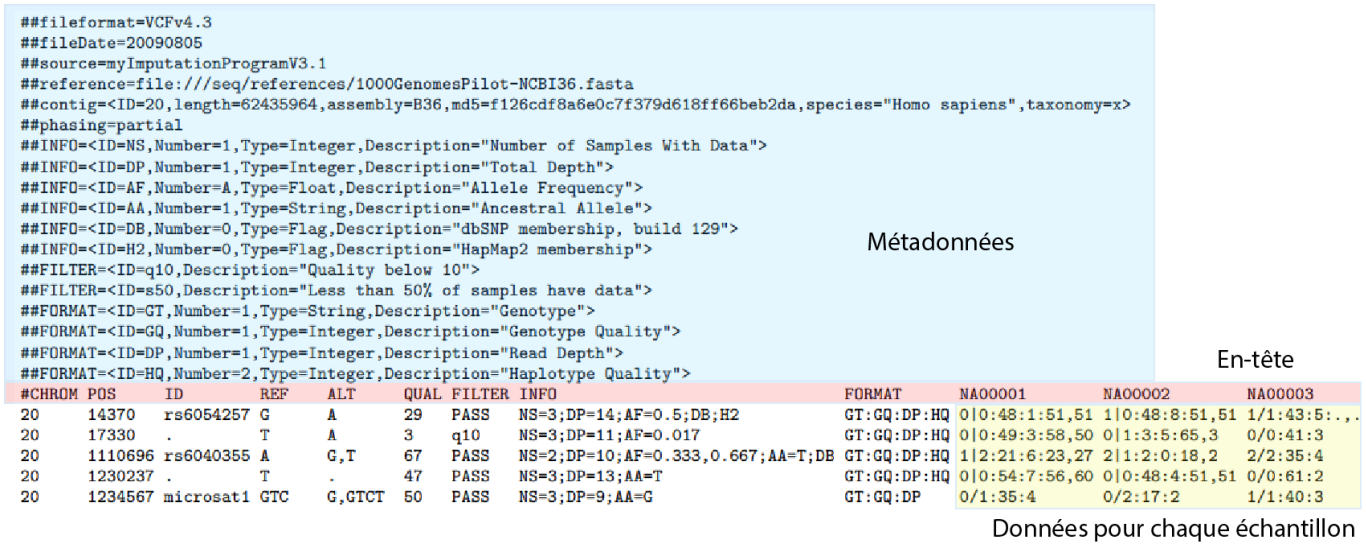
1. Description :

Parfois grossièrement traduit en « format variante d'appel » ou « format d'appel de variants », le format VCF (pour *Variant Call Format*) est un format de fichier texte utilisé en bioinformatique pour stocker les variants retrouvé par les techniques de séquençage. C'est le principal format des documents de travail des biologistes. Ce format contient successivement plusieurs lignes de métadonnées (débutant par les symboles « ## »), une ligne d'en-tête (débutant par le symbole « # »), et de multiples lignes de données qui contiennent chacune des informations sur une unique position dans le génome. Le format a également la capacité de contenir les données de génotype de chaque échantillon pour chaque position, lorsque le document correspond à l'analyse de plusieurs échantillons (champs de texte séparés par des tabulations). On notera que les champs de longueur zéro ne sont pas autorisés, et que les champs qui devraient être vides sont remplacés par un point (".").

Ce format a été développé pour s'affranchir des limites des formats préexistants, tels que le *General feature format* (GFF). En effet ces derniers stockaient l'ensemble des informations génétiques issues d'un séquençage, or une grande partie est redondante entre les différents génomes. En utilisant le format vcf, seuls les variants font l'objet d'une écriture, impliquant que sur l'ensemble des autres positions l'échantillon correspond à ce que l'on retrouverait sur le génome de référence.

La norme est actuellement dans la version 4.3 publiée le 25 juin 2020 et disponible dans le dépôt github de *samtools* : https://samtools.github.io/hts-specs/VCFv4.3.pdf.

2. Structure du format VCF :



Annexe 5 - Figure 1 – Structure des données au format VCF.

Annexe 6 : Procédure

	ANALYSE DES CNV PAR EXOMEDEPTH A PARTIR DE PANEL CONSTITUTIONNEL	**Création :** 03/09/2020 **Diffusion :** **XX/XX/2020**
	SPR-GDOC-OGM.MOP-XXXXXXXXXX	Pages : 1 /XX Version : V 1.0

IDENTIFICATION DU DOCUMENT

Domaine	Fonction	Caractéristique
SUPPORT	**GESTION DOCUMENTAIRE ET ARCHIVAGE**	**ONCOGÉNÉTIQUE**

Type	Date de péremption
MODE OPÉRATOIRE	**22/05/2021**

Mots clefs : ONCOGENETIQUE MOLECULAIRE; OCGM; LOGM; BIOINFORMATIQUE ; CNV ; SV ; EXOMEDEPTH ; NGS; PANEL; GENETIQUE CONSTITUTIONNELLE ; DBC; DEPARTEMENT DE BIOLOGIE DU CANCER; COFRAC ISO 15189; LABORATOIRES; OECI

Référentiels externes

HAS r2011:REF 21:Crit b-Démarche qualité en laboratoire de biologie médicale
ISO 15189 version 2012:5 - Exigences techniques:5.5 - Processus analytiques
OECI

DIFFUSION DU DOCUMENT

Générale

GROUPE DE TRAVAIL

NOM – PRENOM	FONCTION	NOM – PRENOM	FONCTION
DELACROIX Robin	Interne Biologie Médicale		
DA COSTA Quentin	Bioinformaticien		
REMENIERAS Audrey	Pharmacien Biologiste		
BIDAUT Ghislain	Coordinateur Bioinformaticien		

LEGITIMITE DU DOCUMENT

REDACTION	VERIFICATION	APPROBATION
Nom : DELACROIX Robin	**Nom : BOUYSSIE Martine**	**Nom : SOBOL Hagay**
Fonction : Interne Biologie Médicale	**Fonction : RAQ**	**Fonction : Responsable Oncogénétique**
Date : 03/09/2020 **Visa :**	**Date : 19/04/2018** **Visa :**	**Date : 23/05/2018** **Visa :**

Ce document ne doit être ni modifié ni reproduit
La version imprimée de ce document n'est valable que **le 16 septembre 2020** (passé cette date, elle doit être détruite)

SOMMAIRE

1.0 OBJET **3**
2.0 MOYENS **3**
2.1 Equipement **3**
2.2 Prérequis **3**
3.0 MISE EN ŒUVRE **3**
3.1 Commande **3**
3.2 Conditions de réalisation d'une analyse des CNVs **4**
3.3 Utilisation **4**
4.0 DOCUMENTS ASSOCIES **5**
4.1 Documents et Formulaires liés **5**
4.2 Autres documents **5**
4.3 Annexes **5**

ETAT DES MODIFICATIONS

N° de Version	Date	Objet de la modification
1.0	03/09/2020	Création initiale du document au sein de la GED

Ce document ne doit être ni modifié ni reproduit
La version imprimée de ce document n'est valable que **le 16 septembre 2020** (passé cette date, elle doit être détruite)

1.0 OBJET

A partir d'une série d'échantillons séquencés puis analysés parallèlement par deux pipelines indépendants : analyse des CNVs par le logiciel ExomeDepth à partir des fichiers « BAM ». ExomeDepth est un logiciel gratuit et opensource distribué sous forme de module « R ». ExomeDepth réalise la recherche des CNV par analyse de la profondeur de lecture (approche Depth-of-Coverage).

ExomeDepth est utilisé en routine par les bioinformaticiens à la suite des analyses de séquençage effectuées sur Sophia DDM (Sophia Genetics) en génétique constitutionnelle. Il prend en entrée :

- Une liste des chemins informatiques vers chacun des fichiers BAMS de l'ensemble du run (pour chaque échantillon, les autres échantillons du run servent à constituer un set de référence. Un échantillon ne peut pas être analysé seul).
- Le fichier BED correspondant au panel utilisé.
- Un fichier FASTA du génome de référence pour la prise en compte du contenu en GC.

Le set de référence est constitué des échantillons les plus proches de l'échantillon analysé en termes de répartition de la profondeur, en conséquence un CNV présent dans le set de référence peut entrainer des faux négatifs.

IMPORTANT : LES CNVs fréquents ne peuvent pas être recherchés avec ExomeDepth. Les membres d'une même famille et principalement **les trios doivent être séparés lors de cette analyse**.

2.0 MOYENS

2.1 Equipement

- Système informatique avec une distribution Linux ou autre systèmes d'exploitation Unix.
- Logiciel R installé (version ≥ 3.4.0).
- Logiciel Rscript
- L'ensemble des dépendances R de ExomeDepth installées, ainsi que les dépendances tierces.

2.2 Prérequis

- Accès au serveur O:\.
- Console possédant les commandes Bash standards.

3.0 MISE EN ŒUVRE

3.1 Commande

Note : l'ensemble des commandes ci-dessous peuvent être contenues dans un simple fichier script shell « .sh »

Génération de la liste des chemins vers les fichiers « BAM » d'entrée, localisés dans le même répertoire :

Dans une console linux: `$ readlink -f *.bam > Run_Bams_List.txt`

Ce document ne doit être ni modifié ni reproduit
La version imprimée de ce document n'est valable que **le 16 septembre 2020** (passé cette date, elle doit être détruite)

Lancement d'ExomeDepth pour l'analyse de tous les échantillons d'un run :

*ExomeDepth est un logiciel de détection des CNVs qui nécessite **pour chaque étape** des instructions en ligne de commande dans une console R. Le process a **donc été automatisé** pour ne nécessiter **qu'une seule exécution pour un run entier**. **L'utilisation se fait alors via l'exécution d'une commande Rscript dans une console Linux.***

Dans une console linux:

```
$ Rscript CHEMIN_VERS_FICHIER_RSCRIPT -s -b CHEMIN_VERS_Run_Bams_List.txt
 -o DOSSIER_DE_SORTIE -p CHEMIN_FICHIER.bed  -v  >  FICHIER_LOG
```

Avec :

- Rscript : exécute le logiciel Rscript.
- CHEMIN_VERS_FICHIER_RSCRIPT : Chemin absolu ou relatif vers le fichier Rscript contenant l'ensemble des commandes R de ExomeDepth (actuellement nommé « *runExomeDepthWithGC.r* »).
- –s : si le flag s est présent *(=true)* l'analyse va procéder à l'addition des profondeurs des régions homologues entre PMS2 et PMS2CL, dans le cas contraire, l'analyse sera faite sans cette optimisation.
  - **IMPORTANT** : L'optimisation effectuée sur PMS2 est faite par addition entre les régions homologues de PMS2 et PMS2CL, cette étape d'addition est dépendante du BED utilisé, pour chaque changement de fichier BED il faudra adapter le fichier Rscript pour additionner les bonnes lignes.
- -b CHEMIN_VERS_Run_Bams_List.txt: le chemin vers le fichier contenant la liste **des** fichiers « .BAM » du run.
- -o DOSSIER_DE_SORTIE : Chemin vers le dossier dans lequel seront créés les fichiers de résultats de chaque patient de l'analyse.
- -p CHEMIN_FICHIER.bed : Chemin vers le fichier BED d'entrée correspondant aux régions analysées.
  - **IMPORTANT :** Le nombre de colonnes du fichier BED doit entrainer un changement dans le code du fichier Rscript. La présence ou l'absence de la 4eme colonne avec la description des régions dans le fichier BED doit être connu.
- –v pour « *verbose* », sert à augmenter le nombre d'informations inscrites par ExomeDepth dans le fichier log.
- FICHIER_LOG: chemin vers le fichier log. Ce fichier Log correspond à l'analyse du **run entier**.

NB : Le chemin vers le fichier FASTA servant à la prise en compte de la teneur en GC de ExomeDepth est intégré dans le code Rscript, pour le modifier il faut modifier le fichier Rscript « runExomeDepthWithGC ».

3.2 Conditions de réalisation d'une analyse des CNVs

Pour chaque run de NGS, l'analyse des CNVs est faite par les bioinformaticiens en routine. Les résultats sont analysés par les biologistes.

3.3 Utilisation

- **Se renseigner sur la présence de membres d'une même famille et/ou de trios dans le run. Si oui, les séparer.**
- Exécuter les deux commandes décrites précédemment. Une fois pour chaque Run analysé.
- Attendre la fin de l'exécution de l'analyse, chaque échantillon du run est analysé (environ 10min pour un run entier de 48 échantillons). *Il n'est pas nécessaire d'exclure le prélèvement H2O.*
- Contrôler par lecture du fichier log le bon déroulement de l'analyse.

Ce document ne doit être ni modifié ni reproduit
La version imprimée de ce document n'est valable que **le 16 septembre 2020** (passé cette date, elle doit être détruite)

4.0 DOCUMENTS ASSOCIES

4.1 Documents et Formulaires liés

Documents liés
- SPR-GDOC-OGM.MOP-BGW_202009
- SPR-GDOC-LOG.MOP-17-0023
- SPR-GDOC-LOG.MOP-17-0024

Formulaires liés
SPR-GDOC-LAB.FOR-12-0037

4.2 Autres documents

4.3 Annexes

Ce document ne doit être ni modifié ni reproduit
La version imprimée de ce document n'est valable que **le 16 septembre 2020** (passé cette date, elle doit être détruite)

Références bibliographiques

1. Copy number variation: New insights in genome diversity [Internet]. [cité 15 janv 2020]. Disponible sur: https://genome.cshlp.org/content/16/8/949.full

2. A copy number variation map of the human genome | Nature Reviews Genetics [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.nature.com/articles/nrg3871

3. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 23 nov 2006;444(7118):444-54.

4. A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome - ScienceDirect [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.sciencedirect.com/science/article/pii/S0002929707609240

5. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. mai 2011;12(5):363-76.

6. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics. févr 2006;7(2):85-97.

7. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis - ScienceDirect [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.sciencedirect.com/science/article/pii/S0888754308002115

8. Copy Number Variation in Human Health, Disease, and Evolution | Annual Review of Genomics and Human Genetics [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.annualreviews.org/doi/10.1146/annurev.genom.9.081307.164217

9. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation [Internet]. [cité 15 janv 2020]. Disponible sur: https://genome.cshlp.org/content/13/10/2291

10. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. - PubMed - NCBI [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pubmed/9771718

11. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. - PubMed - NCBI [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pubmed/10471496

12. Sequencing studies in human genetics: design and interpretation | Nature Reviews Genetics [Internet]. [cité 15 janv 2020]. Disponible sur: https://www.nature.com/articles/nrg3455

13. New Era of Genetic Testing and Its Impact on Research and Clinical Care | Clinical Chemistry | Oxford Academic [Internet]. [cité 15 janv 2020]. Disponible sur: https://academic.oup.com/clinchem/article/58/6/1070/5620868

14. Koboldt DC, Larson DE, Chen K, Ding L, Wilson RK. Massively Parallel Sequencing Approaches for Characterization of Structural Variation. Methods Mol Biol. 2012;838:369-84.

15. Initial sequencing and analysis of the human genome | Nature [Internet]. [cité 21 févr 2020]. Disponible sur: https://www.nature.com/articles/35057062

16. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 16 févr 2001;291(5507):1304-51.

17. Finishing the euchromatic sequence of the human genome | Nature [Internet]. [cité 21 févr 2020]. Disponible sur: https://www.nature.com/articles/nature03001

18. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. avr 2010;464(7289):704-12.

19. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 11 2015;349(6253):aab3761.

20. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet. juill 2005;37(7):727-32.

21. A map of human genome variation from population scale sequencing. Nature. 28 oct 2010;467(7319):1061-73.

22. Kim S, Cho C-S, Han K, Lee J. Structural Variation of Alu Element and Human Disease. Genomics Inform. sept 2016;14(3):70-7.

23. Ewing AD. Transposable element detection from whole genome sequence data. Mobile DNA. 29 déc 2015;6(1):24.

24. Mobile element insertion (MEI) detection for NGS based clinical diagnostics – SeqOne [Internet]. [cité 5 juill 2020]. Disponible sur: https://seq.one/2020/05/14/mobile-element-insertion-mei-detection-for-ngs-based-clinical-diagnostics/

25. Large-Scale Copy Number Polymorphism in the Human Genome | Science [Internet]. [cité 22 févr 2020]. Disponible sur: https://science.sciencemag.org/content/305/5683/525.long

26. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental Duplications and Copy-Number Variation in the Human Genome. Am J Hum Genet. juill 2005;77(1):78-88.

27. Sebat J. Major changes in our DNA lead to major changes in our thinking. Nature Genetics. 27 juin 2007;39:S3-5.

28. Ommen G-JB van. Frequency of new copy number variation in humans. Nat Genet. avr 2005;37(4):333-4.

29. Human Copy Number Variation and Complex Genetic Disease [Internet]. [cité 22 févr 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6662611/

30. Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. Proc Biol Sci. 7 nov 2010;277(1698):3213-21.

31. Stankiewicz P, Lupski JR. Structural Variation in the Human Genome and its Role in Disease. Annual Review of Medicine. 2010;61(1):437-55.

32. Iskow RC, Gokcumen O, Lee C. Exploring the role of copy number variants in human adaptation. Trends Genet. juin 2012;28(6):245-57.

33. Saitou M, Gokcumen O. An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health. J Mol Evol. 1 janv 2020;88(1):104-19.

34. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends in Genetics. 1 févr 2002;18(2):74-82.

35. Lieber MR. The Mechanism of Human Nonhomologous DNA End Joining. J Biol Chem. 1 avr 2008;283(1):1-5.

36. Lee JA, Carvalho CMB, Lupski JR. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. Cell. 28 déc 2007;131(7):1235-47.

37. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. sept 2004;36(9):949-51.

38. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances - Solinas-Toldo - 1997 - Genes, Chromosomes and Cancer - Wiley Online Library [Internet]. [cité 3 mars 2020]. Disponible sur: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2264%28199712%2920%3A4%3C399%3A%3AAID-GCC12%3E3.0.CO%3B2-I?sid=nlm%3Apubmed

39. Brennan C, Zhang Y, Leo C, Feng B, Cauwels C, Aguirre AJ, et al. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. Cancer Res. 15 juill 2004;64(14):4744-8.

40. Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis. - PubMed - NCBI [Internet]. [cité 3 mars 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pubmed/15750031

41. Sanlaville D, Lapierre J-M, Turleau C, Coquin A, Borck G, Colleaux L, et al. Molecular karyotyping in human constitutional cytogenetics. Eur J Med Genet. sept 2005;48(3):214-31.

42. Neuvial P, Hupé P, Brito I, Liva S, Manié É, Brennetot C, et al. Spatial normalization of array-CGH data. BMC Bioinformatics. 22 mai 2006;7:264.

43. Hupé P, Stransky N, Thiery J-P, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics. 12 déc 2004;20(18):3413-22.

44. Polzehl J, Spokoiny V. Local likelihood modeling by adaptive weights smoothing∗. In 2002.

45. Bièche I, Olivi M, Champème MH, Vidaud D, Lidereau R, Vidaud M. Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer. Int J Cancer. 23 nov 1998;78(5):661-6.

46. Ponchel F, Toomes C, Bransfield K, Leong FT, Douglas SH, Field SL, et al. Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. BMC Biotechnol. 13 oct 2003;3:18.

47. Armour JAL, Sismani C, Patsalis PC, Cross G. Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res. 15 janv 2000;28(2):605-9.

48. Hollox EJ, Akrami SM, Armour JAL. DNA copy number analysis by MAPH: molecular diagnostic applications. Expert Rev Mol Diagn. juill 2002;2(4):370-8.

49. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 15 juin 2002;30(12):e57.

50. Charbonnier F, Raux G, Wang Q, Drouot N, Cordier F, Limacher JM, et al. Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. Cancer Res. 1 juin 2000;60(11):2760-3.

51. Jacques V. L'intégration du séquençage de nouvelle génération dans le diagnostic médical: application aux leucémies aiguës myéloïdes et syndromes myélodysplasiques. :140.

52. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2 janv 2009;323(5910):133-8.

53. Grzych G. Evaluation of in silico prediction tools and interest of functional tests in the interpretation of identified variants by next-generation sequencing in human genetics. 2018.

54. Encyclopedia of Bioinformatics and Computational Biology - 1st Edition [Internet]. [cité 6 juill 2020]. Disponible sur: https://www.elsevier.com/books/encyclopedia-of-bioinformatics-and-computational-biology/ranganathan/978-0-12-811414-8

55. Akogwu I, Wang N, Zhang C, Gong P. A comparative study of k-spectrum-based error correction methods for next-generation sequencing data analysis. Human Genomics. 25 juill 2016;10(2):20.

56. Sohn jang-il, Nam J-W. The present and future of de novo whole-genome assembly. Briefings in Bioinformatics. 14 oct 2016;19:bbw096.

57. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 15 juill 2009;25(14):1754-60.

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 15 août 2009;25(16):2078-9.

59. Efficient storage of high throughput DNA sequencing data using reference-based compression [Internet]. [cité 6 mars 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083090/

60. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 25 juill 2016;17(7):239.

61. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res. 6 janv 2009;19(6):1124-32.

62. Quality scores and SNP detection in sequencing-by-synthesis systems [Internet]. [cité 20 avr 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2336812/

63. A framework for variation discovery and genotyping using next-generation DNA sequencing data [Internet]. [cité 20 avr 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/

64. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science. 19 oct 2007;318(5849):420-6.

65. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. mai 2008;453(7191):56-64.

66. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives | BMC Bioinformatics | Full Text [Internet]. [cité 15 janv 2020]. Disponible sur: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S11-S1

67. Hayes M, Pyon YS, Li J. A Model-Based Clustering Method for Genomic Structural Variant Prediction and Genotyping Using Paired-End Sequencing Data. PLoS One [Internet]. 27 déc 2012 [cité 9 mars 2020];7(12). Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531386/

68. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Front Genet [Internet]. 13 avr 2015 [cité 9 mars 2020];6. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4394692/

69. Xi R, Kim T-M, Park P. Detecting structural variations in the human genome using next generation sequencing. Briefings in functional genomics. 1 déc 2010;9:405-15.

70. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods. 15 oct 2009;6(11s):S13-20.

71. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, et al. Identification of genomic indels and structural variations using split-read. BMC Genomics. 25 juill 2011;12:375.

72. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 1 nov 2012;28(21):2711-8.

73. Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics. 15 janv 2011;27(2):268-9.

74. Janevski A, Varadan V, Kamalakaran S, Banerjee N, Dimitrova N. Effective normalization for copy number variation detection from whole genome sequencing. BMC Genomics. 26 oct 2012;13(6):S16.

75. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 13 sept 2013;14(11):S1.

76. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. sept 2009;19(9):1586-92.

77. Macé A, Kutalik Z, Valsesia A. Copy Number Variation. In: Evangelou E, éditeur. Genetic Epidemiology: Methods and Protocols [Internet]. New York, NY: Springer New York; 2018 [cité 3 sept 2019]. p. 231-58. (Methods in Molecular Biology). Disponible sur: https://doi.org/10.1007/978-1-4939-7868-7_14

78. Yao R, Yu T, Qing Y, Wang J, Shen Y. Evaluation of copy number variant detection from panel-based next-generation sequencing data. Molecular Genetics & Genomic Medicine. 2019;7(1):e00513.

79. Quenez O, Cassinari K, Coutant S, Lecoquierre F, Le Guennec K, Rousseau S, et al. Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. European Journal of Human Genetics. 26 juin 2020;1-11.

80. Lacoste C, Fabre A, Pécheux C, Lévy N, Krahn M, Malzac P, et al. Le séquençage d'ADN à haut débit en pratique clinique. Archives de Pédiatrie. 1 févr 2017;24.

81. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. ABySS: A parallel assembler for short read sequence data. Genome Res. juin 2009;19(6):1117-23.

82. De novo assembly of human genomes with massively parallel short read sequencing [Internet]. [cité 21 août 2019]. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813482/

83. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 8 janv 2012;44(2):226-32.

84. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. mars 2012;22(3):549-56.

85. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 1 nov 2012;28(21):2747-54.

86. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 15 sept 2012;28(18):i333-9.

87. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res [Internet]. 25 nov 2016 [cité 19 juin 2020];1. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409526/

88. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology. 26 juin 2014;15(6):R84.

89. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 15 mai 2012;28(10):1307-13.

90. Niu YS, Zhang H. THE SCREENING AND RANKING ALGORITHM TO DETECT DNA COPY NUMBER VARIATIONS. Ann Appl Stat. sept 2012;6(3):1306-26.

91. Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, et al. Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. The Journal of Molecular Diagnostics. 1 nov 2017;19(6):905-20.

92. Sandell S, Schuit RJL, Bunyan DJ. An intronic polymorphic deletion in the PTEN gene: implications for molecular diagnostic testing. Br J Cancer. 5 févr 2013;108(2):438-41.

93. Le Scouarnec S, Gribble SM. Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. Heredity. janv 2012;108(1):75-85.

94. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics. oct 2009;41(10):1061-7.

95. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. avr 2010;38(6):1767-71.

SERMENT DE GALIEN

Je jure, en présence de mes maîtres de la Faculté, des conseillers de l'Ordre des pharmaciens et de mes condisciples :

- **D'honorer ceux qui m'ont instruit dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle à leur enseignement.**
- **D'exercer, dans l'intérêt de la santé publique, ma profession avec conscience et de respecter non seulement la législation en vigueur, mais aussi les règles de l'honneur, de la probité et du désintéressement.**
- **De ne jamais oublier ma responsabilité et mes devoirs envers le malade et sa dignité humaine, de respecter le secret professionnel.**
- **En aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les mœurs et favoriser des actes criminels.**

Que les hommes m'accordent leur estime si je suis fidèle à mes promesses.

Que je sois couvert d'opprobre, méprisé de mes confrères, si j'y manque.