Développement de méthodes bioinformatiques d'analyses combinatoires du patrimoine génétique de la tumeur et de son hôte : intérêt en diagnostic moléculaire et recherche transversale en cancérologie

Nicolas Soirat

Normandie Université

UNIVERSITÉ CAEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité SCIENCES DE LA VIE ET DE LA SANTE

Préparée au sein de l'Université de Caen Normandie

Développement de méthodes bioinformatiques d'analyses combinatoires du patrimoine génétique de la tumeur et de son hôte : intérêt en diagnostic moléculaire et recherche transversale en cancérologie.

Présentée et soutenue par
NICOLAS SOIRAT

Thèse soutenue le 20/12/2023 devant le jury composé de		
M. ALEXANDRE HARLE	Professeur des universités PraticienHosp, Université de Lorraine	Rapporteur du jury
M. JEAN MULLER	Maître de conférences HDR, UNIVERSITE STRASBOURG	Rapporteur du jury
M. LAURENT CASTERA	Praticien hospitalier, 14 BACLESSE CAEN	Membre du jury
M. NICOLAS SEVENET	Professeur des universités, Université de Bordeaux	Membre du jury
M. FLORIAN CLATOT	Professeur des universités PraticienHosp, Université de Rouen Normandie	Président du jury
MME SOPHIE KRIEGER	Maître de conférences HDR, Université de Caen Normandie	Directeur de thèse

Thèse dirigée par SOPHIE KRIEGER (CANCER AND BRAIN GENOMICS)

ED 497 NBISE
Normande de Biologie Intégrative, Santé, Environnement

Brain Cancer Normandy genomics

Remerciements

La rédaction de cette thèse est à ce jour, le plus grand accomplissement de ma vie. L'aboutissement de ces trois années, riches en émotions, n'a été possible que grâce à l'aide de nombreuses personnes. Je tenais à leur dédier cette partie du manuscrit afin de les remercier en bonne et due forme.

Tout d'abord, je voudrais remercier Nicolas Philippe, pour la confiance qu'il a placée en moi lorsqu'il m'a accueilli à SeqOne dans le cadre de mon stage en 2019. Merci pour tout le soutien jusqu'à la fin de la préparation de mon doctorat et de m'avoir permis de découvrir la vie dans cette belle entreprise qu'est SeqOne. Un grand merci à Dominique Vaur pour avoir permis ce projet collaboratif et m'avoir accueilli au sein de son équipe au Centre François Baclesse. Pouvoir travailler au sein d'une équipe dynamique et passionnée dans le milieu hospitalier aura été réellement enrichissant pour moi. Merci à vous deux pour toutes vos remarques pertinentes et pour m'avoir accompagnés dans cette aventure.

Je n'aurai pas pu aboutir à la conclusion de ce travail sans une direction de thèse et des encadrants incroyables. Merci à Sophie Krieger, Laurent Castera, Denis Bertrand et Anne-Laure Bougé pour avoir été mes "parents de la science", j'espère avoir pu en être un minimum digne. Merci pour toujours m'avoir stimulé scientifiquement, merci pour votre soutien dans les moments difficiles et d'avoir pu me supporter dans les moments où j'étais morose. Votre disponibilité m'a permis de toujours pouvoir discuter avec vous, que ce soit pour faire avancer ce projet ou pour des échanges plus légers mais tellement importants. Vos quatre personnalités bien que différentes auront réellement été complémentaires pour moi et j'ai essayé tout au long de ce travail, de m'inspirer de votre vision de la science, pour me développer en tant que scientifique. Merci Sophie pour ta bienveillance, ton expertise clinique et les appels téléphoniques sur nos phobies administratives. Merci Laurent pour ton franc-parler, ta rigueur (j'espère ne pas t'avoir fait arracher trop de cheveux sur la rédaction),

tes idées et nos discussions sur la tech ! Merci à Denis pour toujours m'avoir poussé au bout de mes réflexions, m'avoir accompagné dans mes rushs (je crois que j'y ai même pris goût) et ta bonne humeur contagieuse qui m'a permis de traverser bien des moments. Enfin merci à toi Anne-Laure, pour avoir été mon encadrante en stage et durant le début de ma thèse. J'ai essayé de m'inspirer de ton organisation et ta rigueur, tu sais que je partais de loin. Grâce à cela, j'ai pu démarrer ma thèse sans m'éparpiller (enfin je crois ?).

Merci aux deux équipes dans lesquelles j'ai eu la chance de travailler. Merci à Thibaut, Imène et Alexandre de l'équipe bioinfo du Centre qui m'ont aidé à m'intégrer et m'ont accompagné durant mes 6 mois à Caen. Merci à Arnaud et Julien pour avoir participé à la bonne ambiance de ce bureau de folie. Je remercie Raphaël et Camille pour m'avoir fait découvrir le monde de l'épissage alternatif de l'ARN qui m'était jusqu'alors inconnu !

Merci à l'équipe bioinfo de SeqOne : Mélanie, Raphaël, Céline G., Marie, Valentin, Céline E. et Sacha. Même si à la fin je gravitais autour de vous, je vous suis reconnaissant pour tout ce que vous m'avez apporté durant ces 3 ans, dans une ambiance presque familiale au travers de discussions sérieuses autour de la bioinfo. Je n'oublie pas les instants mangas, jeux vidéos, politiques et bien d'autres sur Discord, le Vendredi après-midi ! Merci à Jiri et Nicolas pour nos discussions sur l'IA et d'avoir pris le temps de m'introduire à ce magnifique monde, j'espère qu'un jour nous pourrons porter un projet ensemble dans ce domaine (j'y crois). Votre passion pour ce domaine est belle à voir ! Bien sûr je n'oublie pas l'ensemble de SeqOne et toutes les personnes avec qui j'ai pu travailler de loin. Vous êtes tous des gens formidables qui ont fait que je n'ai pas vu mes 3 ans passer. Merci pour toutes ces discussions tech, les afterworks, une bienveillance et une ambiance au quotidien presque irréaliste.

Je tenais à remercier Kévin Yauy qui, après avoir été mon maître de stage, est devenu mon collègue doctorant à SeqOne. J'ai pu découvrir la génétique médicale grâce à toi, suite à notre non-compréhension du langage R en workshop, à l'époque où j'étais encore étudiant. Ta positivité et ta vision de la génétique m'ont toujours inspiré. J'espère que l'on se reverra en congrès (avec Jiri bien sûr !) et de manière moins formelle autour d'une bière ! Tu as été un de mes premiers modèles en tant que bioinformaticien en santé et j'espère avoir été à la hauteur !

Je dédie aussi cette thèse à des personnes ayant participé à sa réussite au travers du soutien émotionnel qu'ils ont été pour moi. Je pense dans un premier temps à ma deuxième famille, les Penda, ce groupe d'amis qui me suit pour certains depuis la crèche. J'ai une énorme chance de vous avoir dans ma vie, merci pour toutes

ces soirées, pour les confessions, nos sessions de jeu et surtout, de m'avoir supporté pendant ces 3 ans, dans mes bons comme mauvais moments. Je suis fier de notre petit groupe (même si on est 25) et je suis rassuré de tous vous avoir à mes côtés. Merci au PVN, on a commencé ensemble en biologie, on se retrouve encore des années plus tard comme si le temps n'avait pas bougé ! Enfin merci à notre groupe Telegram bioinfo, Quentin (aka Qt), Quentin (aka Qd), Anna, Xavier, Valentin et Hugo. De collègues de promo à de véritables amis merci de m'avoir encouragé et rassuré dans cette grande aventure qu'est la thèse. Merci pour nos discussions parfois endiablées, votre humour, vos conseils et tous les moments partagés ensemble depuis le master (et il y en a eu) !

Je terminerai ces remerciements par les personnes qui m'accompagnent depuis plus de 28 ans, Papa, Maman je vous aime et je vous remercie de m'avoir poussé et soutenu dans tout ce que j'ai entrepris. Merci Papa pour m'avoir appris à être curieux et à m'intéresser au monde qui m'entoure. Merci Maman pour m'avoir appris le sérieux dans le travail et pour m'avoir empêché de me reposer sur mes acquis. Merci Thibault, mon petit frère, pour tout ce qu'on a partagé, ta complicité, ton humour, ta maturité, ta rigueur et pour nos débats interminables, tes avis comptent beaucoup pour moi, je n'aurais pu espérer un petit frère aussi fiable ! Bien sûr je remercie toute ma famille pour tous les moments passés ensemble, les repas et surtout votre unicité qui est un véritable confort au quotidien. C'est inestimable pour moi et je ne sais pas où je serai sans cela. A tous mes grand-parents, oncles, tantes, cousins et cousines, vous vous reconnaîtrez, merci à vous, du fond du coeur.

Papou, Mamou, je vous avais promis, qu'un jour, je serai docteur. Au final je ne serai pas docteur en médecine, j'ai trouvé une autre vocation entre temps et je pense que je suis mieux là qu'à cheval comme on dit.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

BRCA1 *BReast CAncer 1*. 5, 6, 12, 15, 52, 59, 65, 112, 148, 149

BRCA2 *BReast CAncer 2*. 5, 6, 12, 15, 52, 59

ACMG *American College of Medical Genetics and Genomics*. 12

ADN Acide désoxyribonucléique. ix, 1, 2, 5, 10, 21, 24–26, 29–31, 33, 35, 36, 38–40, 42, 44, 47, 48, 52, 57, 59–62, 148, 151, 171

ADNc Acide désoxyribonucléique Complémentaire. 39, 42

AMM Autorisation de Mise sur le Marché. 13, 61, 146

API *Application Programming Interface*. 56, 65

ARN Acide Ribonucléique. 10, 23, 25, 28, 34, 39, 40, 42, 45, 50–52, 57, 59, 61, 62, 147, 148, 150, 154, 171

ARNm Acide Ribonucléique messager. 25–28, 40, 60, 61, 111, 113, 147, 171

ASCO/AMP *American Society of Clinical Oncology/Association for Molecular Pathology*. 13, 14, 53

BAM *Binary Alignment Map*. 46–48

BLOSUM *BLOcks SUbstitution Matrix*. 20

BP *Base Pairs*. 20, 21

BWA *Burrows-Wheeler Aligner*. 44

BWA-MEM *Buroows-Wheeler Aligner-Maximal Exact Matches*. 44

CADD *Combined Annotation Dependent Depletion*. 50, 51

CCS *Circular Consensus Sequencing*. 37, 38

CGC *Cancer Gene Census*. 14, 15

CIViC *Clinical Interpretation of Variants in Cancer*. 11, 52, 53, 56, 62, 65, 146, 147

CKB *Clinical Knowledgebase*. 11, 53, 56, 146, 147

CLR *Continuous Long Read*. 37, 38

CNP *Copy Number Polymorphism*. 24

CNV *Copy Number Variation*. 22, 24, 51, 53, 60–62, 65, 147, 171

ComPerMed *Personalised Medicine Commission*. 12

COSMIC *Catalogue of Somatic Mutations in Cancer*. 11, 13, 15, 52, 53, 62, 65, 147

dbSNP *Single Nucleotide Polymorphism Database*. 49, 51

ddNTP Didéoxynuclotide Tri Phosphate. 29, 30

DGV *Database of Genomic Variants*. 52

DNA Deoxyribonucleic Acid. 171

EMEA *European Medicines Evaluation Agency*. 13

ESHG *Europen Society of Human Genetics*. 154

ESMO *European Society for Medical Oncology*. 13

EVE *Evolutionary model of Variant Effect*. 49

FDA *Food and Drug Administration*. 13, 66, 146

FFPE *Formalin-Fixed Paraffin-Embedded*. 54

FLAIR *Full-Length Alternative Isoform analysis of RNA*. 58

FMI® Foundation Medicine Incorporation®. 66, 146

FrOG *French OncoGenetics*. 6, 52

GATK *Genome Analysis Toolkit*. 53

Gb Gigabase. 18

GGC Groupe Génétique et Cancer. 5, 6, 10, 52

gnomAD *Genome Aggregation Database*. 51

GRC *Genome Reference Consortium*. 17–19

GRCh37 *Human genome assemblies by the Genome Reference Consortium, build 37*. ix, 18, 19

GRCh38 *Human genome assemblies by the Genome Reference Consortium, build 38*. ix, 18, 19

HBOC Hereditary Breast and Ovarian Cancer. 171

HBOC *Hereditary Breast and Ovarian Cancer*. ix, 5–7, 10, 25, 52, 112, 148, 171

HER2+ *Human Epidermal growth factor Receptor 2 positive*. 17

HGP *Human Genome Project*. 18

HiFi *High Fidelity*. 37, 59

HRD *Homologous Recombination Deficiency*. 151

HTS *High Throughput Sequencing*. 30

ICGC *International Cancer Genome Consortium*. 15

IDF Ile De France. 8, 9

INCa Institut National du Cancer. 7–9

INDEL Insertion-délétion. 20, 23, 24, 53, 60, 61, 149

IntOGen *Integrative Onco Genomics*. 15

JOBIM Journées Ouvertes en Biologie, Informatique et Mathématiques. 154

kb kilobase. 35, 37

LMC Leucémie Myéloïde Chronique. 16

LoRID *Long Read Isoform Discovery*. 62, 112–114, 147–154, 171

Mb Megabase. 18

MEI *Mobile Element Insertion*. 21

MM *MolecularMatch*. 53, 56, 65, 146, 147

MMR *Missmatch Repair*. 151

mRNA messenger Ribonucleic Acid. 171

MSI *Microsatellite Instability*. 151

NGS *Next-Generation Sequencing*. 30, 40, 60, 62

NMD *Non-sens Mediated mRNA Decay*. 16, 25

OMIM *Online Mendelian Inheritance in Man*. 52

OncoKB *Oncology Knowledge Base*. 11, 15, 53, 146, 147
ONT Oxford Nanopore Technologies. 35, 38, 46, 149

PacBio Pacific Biosciences. 35–38, 59
PCR *Polymerase Chain Reaction*. 31, 41, 54
pRNPn petites ribonucléoprotéines nucléaires. 25

RCP Réunion de Concertation Pluridisciplinaire. 11, 147, 154
RefSeq *Reference Sequence*. 52
REVEL *Rare Exome Variant Ensemble Learner*. 50
RNA Ribonucleic Acid. 171

SAM *Sequence Alignment Map*. ix, 46
SFMPP Société Française de Médecine Prédictive et Personnalisée. 154
SIFT *Sorting Intolerant From Tolerant*. 49
SMRT *Single Molecule Real-Time*. 36
SNP *Single Nucleotide Polymorphism*. 19, 51
SNV *Single-Nucleotide Variant*. 19, 47, 53, 60–62, 65, 147, 171
SOSTAR iSofOrmS annoTAtoR. 112, 113, 150, 152, 171
SPiP *Splicing Prediction Pipeline*. 50
STAR *Spliced Transcripts Alignement to a Reference*. 45, 56
SV *Structural Variant*. 21, 22, 24, 35, 40, 46, 52, 53, 61

T2T *Telomere-To-Telomere*. 19, 35
TAD *Topologically Associated Domain*. 60
TCGA *The Cancer Genome Atlas*. 2, 15
TMB *Tumor Mutation Burden*. 41, 151
TPM *Transcripts Per kilobase Million*. 150
TSG *Tumor Suppressor Gene*. 15, 16, 25, 65
TSO *TruSight Oncology*. 151, 152

UMI *Unique Molecular Identifier*. 31, 32
uORF *upstream Open Reading Frame*. 50

UTR *Untranslated Region*. 149

VCF *Variant Calling Format*. ix, 48, 49
VEP *Variant Effect Predictor*. 50
VUS *Variant of Unknown Significance*. 12

WES *Whole Exome Sequencing*. 11, 40–42
WGS *Whole Genome Sequencing*. 40–42

ZMW *Zero-Mode Waveguide*. 36

1. Introduction

1.1 Avant propos : Génétique et génomique des cancers

La découverte en 1953 de l'Acide désoxyribonucléique (ADN) par Watson et Crick [1] faisant suite aux travaux de Franklin et Gosling [2], molécule déterministe de chaque espèce vivante, a permis le développement exponentiel de la génétique et de la génomique au cours du XX^{e} siècle. Le XXI^{e} siècle est marqué par la résolution du génome humain [3] et par l'étude des génomes à très haut débit avec l'accélération des capacités de séquençage fruit du développement des nanotechnologies et de l'évolution des capacités de calcul et de stockage en informatique [4]. La génétique tend à expliquer les mécanismes héritables des caractéristiques d'un individu à sa descendance, c'est-à-dire la transmission d'un trait génétique, mais aussi les mécanismes *de novo* amenant à de nouvelles caractéristiques acquises par le biais de néo-mutations par exemple. La génomique se focalise sur l'étude des mécanismes de fonctionnement des génomes.

Les variations transmises ou acquises du génome d'un individu à l'autre rend chaque individu unique et permet d'expliquer les différences entre eux mais aussi l'apparition de certaines maladies. Parmi elles, le développement d'un cancer chez un individu fait appel à l'accumulation d'anomalies acquises au cours d'un processus oncogénique dans une cellule et ses dérivées, initiant une prolifération anormale et anarchique des cellules. Dans certains cas, une anomalie du génome est héritée ou acquise précocement au cours du développement embryonnaire, il s'agit d'anomalies constitutionnelles. Certaines anomalies constitutionnelles sont associées à un surrisque de cancer au cours de la vie d'un individu.

Les avancées majeures en diagnostic moléculaire ont fait évoluer le traitement

du cancer. Ainsi, la notion de médecine de précision est apparue, avec entre autres, l'hypothèse que l'étude des variations de l'ADN du génome d'un individu et/ou de sa tumeur permettrait de trouver une cause à l'apparition du cancer et une solution thérapeutique [5]. D'un point de vue médical, aujourd'hui, il est primordial de proposer des solutions techniques afin de faciliter l'étude et l'interprétation des génomes, afin d'orienter chaque patient vers une prise en charge thérapeutique adaptée, dépendante de son patrimoine génétique hérité et acquis. Les objectifs de cette thèse s'articulent, (i) d'une part, dans le cadre général du traitement des cancers par une thérapeutique ciblée, basée sur l'identification de traitements efficaces orientés par l'analyse d'un profil mutationnel tumoral d'un patient, (ii) d'autre part, dans le cadre de la génétique constitutionnelle des cancers, spécifiquement dans le contexte de la recherche d'évènements génétiques complexes, non mis en évidence par des techniques innovantes.

1.2 La génétique des cancers

Il est possible de distinguer deux axes en génétique des cancers [6, 7] : la génétique tumorale et constitutionnelle.

1.2.1 La génétique tumorale

Un cancer se développe au détriment d'une cellule somatique qui a accumulé une série de mutations au cours du développement d'un individu. Cette cellule acquiert des capacités de prolifération en dehors de tout contrôle physiologique

A partir de 2008 [8], le programme *The Cancer Genome Atlas* (TCGA) se développe dans le but d'améliorer la compréhension du paysage mutationnel des tumeurs [9]. TCGA représente plus de 2.5 petabytes de données issues de 11 000 patients et 33 types de tumeurs différentes (`https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history`). De nos jours, ce programme a permis (i) d'améliorer la connaissance des évènements mutationnels impliqués dans le développement tumoral (ii) d'affiner la classification des sous-types tumoraux (iii) d'identifier les mutations actionnables, c'est à dire les mutations qui peuvent être ciblées par des thérapies afin de ralentir la prolifération de la tumeur.

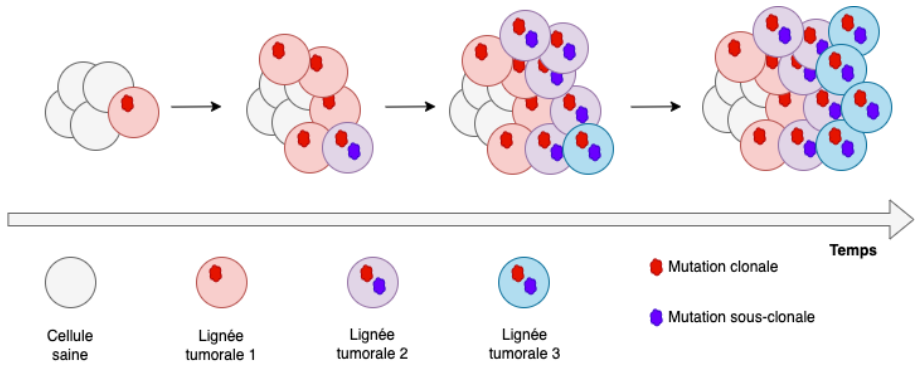


Figure 1.1 – Evolution clonale des cellules cancéreuses.
Adapté d'après les travaux de Nowell [10].

Lors d'un développement tumoral, les cellules ayant accumulé des mutations spécifiques (dites *drivers*), vont se multiplier, si bien que de nouvelles lignées vont apparaître, formant ainsi des sous-clones tumoraux pouvant eux-mêmes acquérir de nouvelles mutations. Chaque lignée clonale va alors disposer de profils mutationnels différents par rapport aux cellules cancéreuses initiales. Les mutations peuvent alors être clonales, auquel cas elles sont partagées par l'ensemble des cellules de l'échantillon tumoral (lors d'une biopsie par exemple) ou sous-clonales si elles sont partagées uniquement par une partie des cellules tumorales d'un prélèvement (voir Figure 1.1). Il existe donc une hétérogénéité tumorale, complétant la notion de clonalité décrite depuis 1976 par Nowell [10].

Les différents clones tumoraux sont soumis à une pression de sélection grâce à la grande hétérogénéité tumorale, cette pression est en faveur de clones accélérant leur prolifération [11]. L'hétérogénéité tumorale rend complexe l'identification des mutations responsables de l'oncogenèse et les mutations réellement actionnables. Aussi, il faut s'assurer que les événements observés au sein de la tumeur sont des vrais positifs et nous reviendrons plus en détail sur les moyens d'affirmer la présence d'une mutation dans le paragraphe 1.6.

De plus, la pression de sélection des lignées tumorales peut induire des phénomènes de résistance à une thérapeutique ciblée, par l'apparition d'une nouvelle mutation conférant une résistance, voire éventuellement, la disparition de la cible thérapeutique par des mécanismes de réversion [12]. Les enjeux majeurs lors du diagnostic génétique tumoral sont donc :

— De détecter des mutations actionnables dans un paysage génétique hétérogène.
— D'identifier l'efficacité d'une thérapie ciblée ou d'une résistance à un traite-

ment dans le contexte de la pathologie d'un type tumoral spécifique.

1.2.2 Génétique constitutionnelle des tumeurs

La génétique constitutionnelle des tumeurs, ou oncogénétique, cherche à identifier les caractères transmis ou transmissibles responsables d'un surrisque de cancer au cours de la vie d'un individu. Ces derniers ont pour support des mutations transmises par les cellules de la lignée germinale (voir Figure 1.2), l'ensemble des cellules de l'individu est porteur de la mutation. Les mutations peuvent également être acquises précocement lors du développement embryonnaire. L'individu est donc porteur d'une mutation en mosaïque dans ce dernier cas. Si les gamètes sont porteurs de la mutation, alors l'individu pourra transmettre cette dernière à sa descendance.

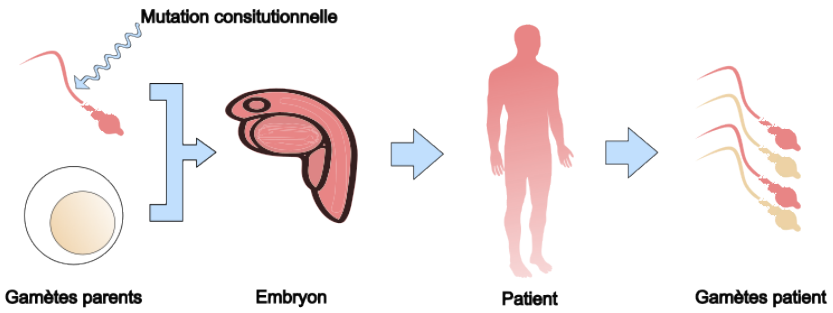


Figure 1.2 – Représentation d'une mutation constitutionnelle.
Mutations transmises par les gamètes des parents de l'individu : touchant les lignées germinales, elles sont transmises à la descendance.

L'identification de mutations, dites variants pathogènes, dans les gènes de prédisposition au cancer, chez un individu atteint (cas index d'une famille), relève d'un intérêt clinique majeur pour le patient mais aussi ses apparentés. En effet, cela permet une meilleure prise en charge clinique voire une orientation thérapeutique de celui-ci et d'élargir le conseil génétique à sa famille, en proposant des tests prédictifs basés sur l'identification de variants pathogènes, en cause dans la famille. Il est alors possible d'établir des recommandations de surveillance de l'apparition d'un cancer, voire d'orienter un membre de la famille vers des mesures chirurgicales prophylactiques. Ces tests génétiques permettent, quand ils sont négatifs, d'exclure les membres de la famille d'une surveillance médicale intensifiée. Quand ils sont positifs, ils permettent

d'orienter la vigilance chez les membres porteurs de variants pathogènes sur les gènes de prédisposition au cancer et de proposer des prises en charge adaptées.

Environ 5% à 10% des cancers sont associés à une prédisposition [13]. Il existe plus d'une dizaine de syndromes associés à des spectres tumoraux spécifiques touchant l'adulte et/ou l'enfant. La prédisposition la plus fréquente induit un surrisque aux cancers du sein et/ou de l'ovaire chez la femme, il s'agit du syndrome *Hereditary Breast and Ovarian Cancer* (HBOC). Il est recensé 61 214 nouveaux cas de cancer du sein en France en 2023 avec une progression des cas par an d'en moyenne 1.1% [14] et avec 50.3% des femmes dépistées [15]. En France, bien que moins fréquents, 5348 cancers de l'ovaire sont estimés pour 2023 [15]. Il est donc primordial parmi ces derniers d'identifier les cas prédisposés.

Initialement, le gène *BReast CAncer 1* (*BRCA1*) et le gène *BReast CAncer 2* (*BRCA2*) ont été identifiés dans ce syndrome [16, 17]. Depuis, d'autres gènes ont été démontrés comme impliqués dans ce syndrome HBOC. Le Groupe Génétique et Cancer (GGC), recommande l'étude de 13 gènes pouvant être responsables de cancers du sein ou de l'ovaire [18] : *BRCA1*, *BRCA2*, *PALB2*, *TP53*, *CDH1*, *PTEN*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, *MSH6*, *PMS2* et *EPCAM*. Cette liste de gènes n'est pas fixe et évolue avec les connaissances. En dehors de la France, d'autres panels HBOC incluent des gènes comme *ATM*, *CHEK2* et *BRIP1*, dont les récentes études ont montré une augmentation modérée de risque de cancer du sein [19, 20]. Ce syndrome est causé par la transmission d'un variant pathogène à partir d'un allèle dans un modèle répondant à une transmission mendélienne autosomique dominante [21]. En effet, ces gènes sont majoritairement des gènes suppresseurs de tumeurs impliqués dans la réparation de l'ADN (voir Table 1.1).

Gene	Nom complet	Cancer Gene Role
BRCA1	*familial breast/ovarian cancer gene 1*	TSG
BRCA2	*familial breast/ovarian cancer gene 1*	TSG
PALB2	*partner and localizer of BRCA2*	TSG
TP53	*tumor protein p53*	oncogene ; TSG
CDH1	*cadherin 1 ; type 1 ; E-cadherin (epithelial) (ECAD)*	TSG
PTEN	*phosphatase and tensin homolog gene*	TSG
RAD51C	*RAD51 paralog C*	Non renseigné
RAD51D	*RAD51 paralog D*	Non renseigné
MLH1	*E.coli MutL homolog gene*	TSG
MSH2	*mutS homolog 2 (E. coli)*	TSG
MSH6	*mutS homolog 6 (E. coli)*	TSG
PMS2	*postmeiotic segregation increased 2 (S. cerevisiae)*	TSG
EPCAM	*epithelial cell adhesion molecule*	Non renseigné
ATM	*ataxia telangiectasia mutated*	TSG
CHEK2	*CHK2 checkpoint homolog (S. pombe)*	TSG
BRIP1	*BRCA1 interacting protein C-terminal helicase 1*	TSG

Table 1.1 – Rôles des différents gènes à risque pour le syndrome HBOC
A partir des informations de *COSMIC Cancer Gene Census*

Le diagnostic moléculaire des prédispositions génétiques est en pleine expansion ; cela est étroitement lié à l'augmentation du nombre de consultations en oncogénétique en France et des avancées technologiques dans les laboratoires d'oncogénétique. 55 584 consultations pour le syndrome HBOC ont été recensées en 2020. En 2021, ce sont 1108 variants causaux recensés sur *BRCA1* et 1344 sur *BRCA2* qui étaient référencés sur la base nationale historique du groupe GGC, *French OncoGenetics* (FrOG) [22]. Il existe néanmoins d'autres syndromes prédisposant au développement de cancers [23], dont les plus fréquents en dehors de HBOC, sont cités dans la Table 1.2.

Syndrome	Acronyme	Prévalence	Transmission	Gène(s) impliqué(s)
Hereditary leiomyomatosis				
renal cancer cell syndrome	HLRCC	<1 :500	AD	FH
Lynch syndrom	LS	1 :279	AD	MLH1, MSH2, MSH6, PMS2, EPCAM
Neurofibromatosis	NF1, NF2	1 :2600 (NF1), 1 :60,000 (NF2)	AD	NF1, NF2
Li–Fraumeni syndrom	LF	1 :3500	AD	TP53
Tuberous Sclerosis Complex	TSC	1 :5800	AD	TSC1, TSC2
Familial adenomatous polyposis	FAP	1 :8000	AD	APC
Multiple endocrine neoplasia type 1	MEN 1	1 :10,000	AD	MEN1
Von Hippel–Lindau syndrome	VHLS	1 :36,000	AD	VHL
Multiple endocrine neoplasia type 2A	MEN2A	1 :44,000	AD	RET
Schwannomatosis	SCHW	1 :70,000	AD	SMARCB1, LZTR1

Table 1.2 – Syndromes les plus fréquents en terme de prévalence. Les syndromes liés à l'HBOC ont été retirés. AD : Autosomique Dominant.
Adapté d'après Garutti et al. 2023 [23]

1.3 Les problématiques en génétique médicale

1.3.1 Paysage de l'oncogénétique en France

Avec le développement des techniques de séquençage, de nombreux centres ont développé une expertise de l'oncogénétique en France. Selon l'Institut National du Cancer (INCa), en 2020, la France possède 146 sites de consultations en oncogénétique collaborant avec 26 laboratoires académiques réalisant les tests génétiques [24]. Ces sites sont répartis dans 101 villes pour un total de 17 programmes régionaux ou interrégionaux de suivi en oncogénétique (voir Figure 1.3). Ces programmes mettent en place un suivi individualisé de patients prédisposés aux cancers. Ce suivi s'effectue au travers de compétences multidisciplinaires entre établissements internes ou externes au programme, au sein d'une même région ou non. A titre d'exemple, dans le cadre du syndrome HBOC, ce suivi permet de proposer en fonction des risques calculés dans chaque famille, une surveillance mammaire par imagerie (échographie, mammographie, IRM) adaptée et de proposer des chirurgies préventives ovariennes voire mammaires.

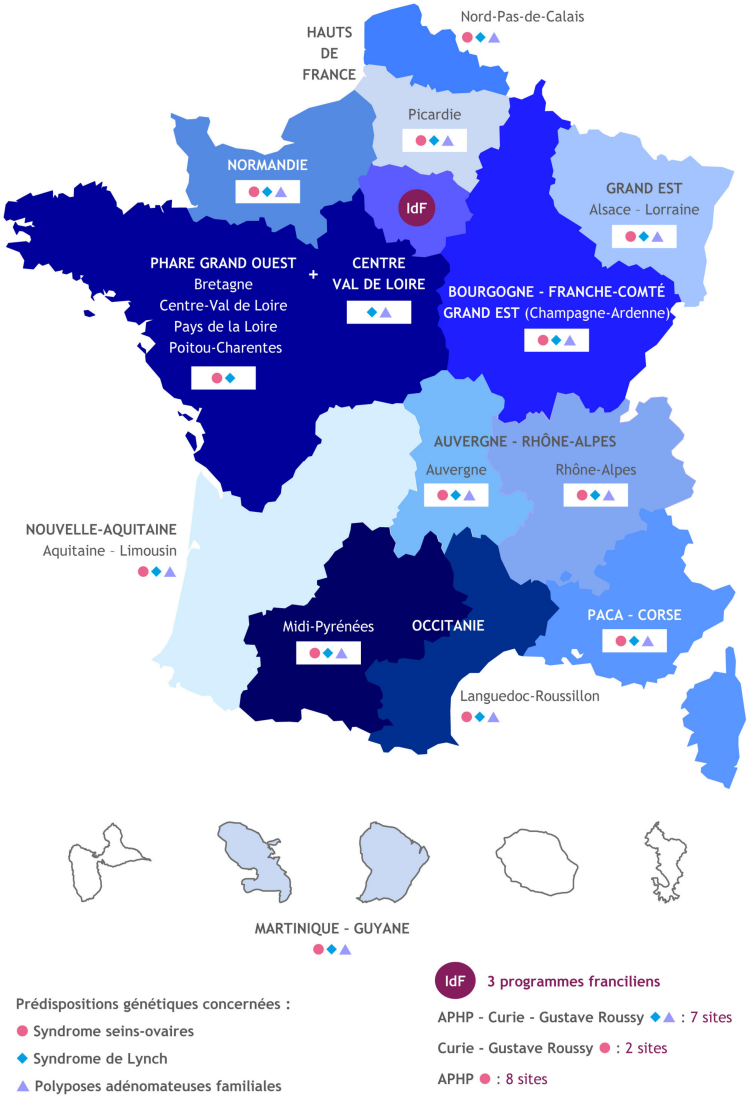


Figure 1.3 – Carte de France des 17 programmes de suivi en oncogénétique
Présentation des 17 programmes régionaux ou interrégionaux de France en oncogénétique. Trois programmes sont inclus en Ile De France (IDF).
Repris de L'INCa, 2018

1.3.2 Paysage des plateformes de génétique moléculaire des cancers

En parallèle des consultations en oncogénétique, les plateformes de génétique moléculaire des cancers se sont développées en France. Toujours selon l'INCa, 28 plateformes sont recensées sur l'ensemble du territoire français (voir Figure 1.4).

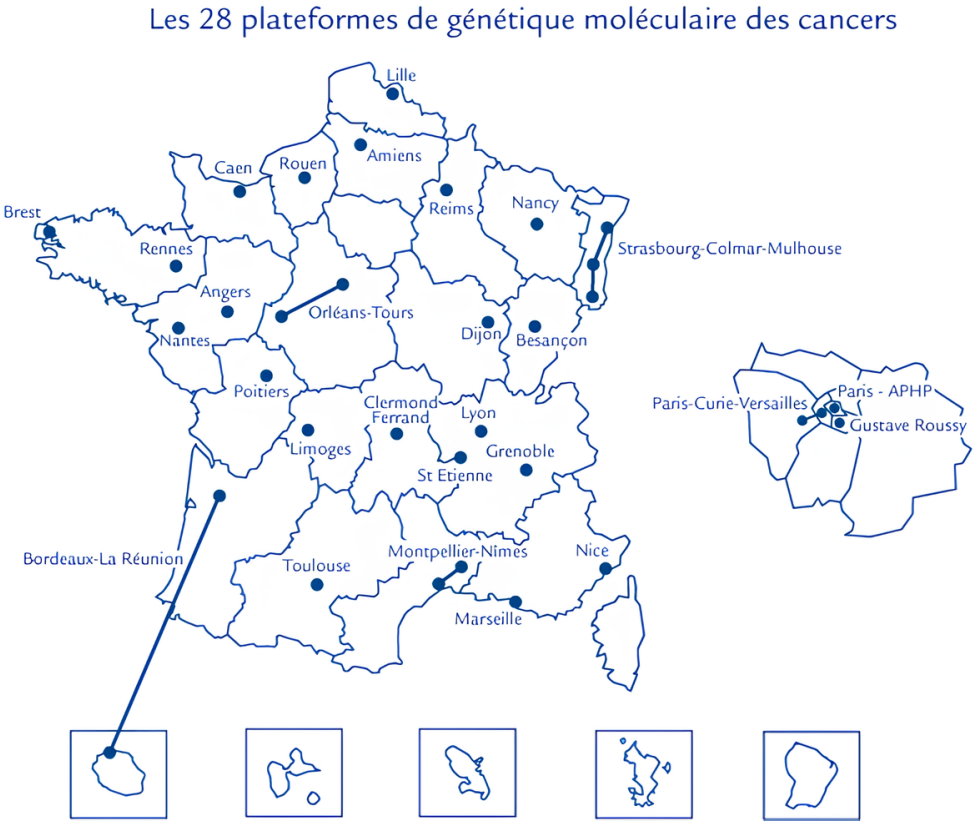


Figure 1.4 – Carte de France des 28 plateformes de génétique moléculaire
Répartition des 28 plateformes de génétique moléculaire en France. Trois plateformes sont situées IDF (Paris-Curie-Versailles, Paris-APHP et Gustave Roussy).
Repris de L'INCa

Ces différents centres réalisent des tests moléculaires ayant plusieurs objectifs :

— Décider de l'accès à une thérapie ciblée après étude du profil moléculaire du patient

— Participer au diagnostic en complémentarité de critères cliniques, morphologiques et/ou biologiques
— Suivre l'évolution du cancer
— Guider la stratégie thérapeutique du patient

Les 28 plateformes ont réalisé, en 2020, plus de 346 000 tests ayant permis l'accès aux thérapies ciblées à 142 000 patients.

1.3.3 Problématiques en génétique consitutionnelle

A titre d'exemple, le diagnostic moléculaire du syndrome HBOC fait appel au séquençage d'un panel de gènes défini au regard des recommandations du GGC. Néanmoins, ce panel de gènes *via* l'identification d'une anomalie moléculaire appelée variant pathogène, permet d'expliquer seulement 10 à 20% des histoires familiales de cancer du sein et/ou de l'ovaire [25, 26]. Les 90% restant représentent l'hérédité dite manquante. Elle peut s'expliquer, à ce jour, (i) par l'existence d'autres gènes que les gènes du panel du GGC) non identifiés mais impliqués dans ce syndrome (ii) par la présence des variants, dont la signification est inconnue, dans les gènes de ce panel, et qui sont identifiés dans 20% des familles, de véridiques variants pathogènes [22, 27] (iii) par des événements non détectables dans ces même gènes, par les techniques actuelles mises en oeuvre dans les laboratoires de diagnostic à ce jour. Concernant ce dernier point, le développement de technologies de séquençage en longs fragments (*long-reads*, pourrait permettre d'explorer des événements génétiques complexes du génome ou indirectement en étudiant les isoformes aberrantes du transcriptome [28], jusqu'alors difficilement accessibles par du séquençage de jonctions en fragments courts (*short-reads*), de l'ADN ou ARN. La résolution d'une partie de l'hérédité manquante du syndrome HBOC à l'aide de ces nouvelles techniques de séquençage, permettant l'accès à un transcriptome en haute résolution, est un objectif de cette thèse, .

1.3.4 Problématiques en génétique des tumeurs

La réalisation d'un profil génétique tumoral à partir des données issues de l'étude du génome tumoral, permet la recherche de thérapies ciblées. Ceci permet d'identifier les potentielles cibles moléculaires visées par des thérapies spécifiques, qui prédisent la sensibilité à ces traitements. Cela permet également d'identifier des anomalies moléculaires, à l'origine d'une résistance à ces mêmes traitements. Ces anomalies sont appelées des cibles ou variants actionnables. L'analyse de ce profil moléculaire demande le développement d'outils capables de colliger les résultats d'analyses des

différents types de variants et des outils fournissant un rapport à jour des thérapies associées à ces variants actionnables.

L'implémentation de Réunion de Concertation Pluridisciplinaire (RCP) (aussi appelée *tumor board* chez les anglo-saxons) dans le paysage médical permet de confronter les avis de plusieurs praticiens de santé de spécialités différentes concernant la prise en charge d'un patient. Les RCP ont été rendues obligatoires par le Plan Cancer 2003-2007 [29], "Tous les nouveaux patients atteints de cancer bénéficieront d'une concertation pluridisciplinaire et d'un programme personnalisé de soins". Dans le contexte d'une RCP dite alors moléculaire, les biologistes médicaux interprètent les nombreux variants issus du séquençage de tumeurs et mettent en avant ceux pouvant orienter les décisions thérapeutiques à apporter aux patients. De nos jours, l'identification de ces cibles pour la RCP est réalisée à partir de larges panels de gènes, du séquençage d'exome complet ou *Whole Exome Sequencing* (WES) ou de génome complet, celui-ci demeurant coûteux. L'interprétation des variants comme cibles potentiellement actionnables nécessite une maîtrise de la bibliographie et des bases de données spécialisées comme *Oncology Knowledge Base* (OncoKB) [30], *Clinical Interpretation of Variants in Cancer* (CIViC) [31], *Catalogue of Somatic Mutations in Cancer* (COSMIC) [32] ou *Clinical Knowledgebase* (CKB) [33].

Ainsi les différentes structures chargées de l'analyse du profil des variants du patient doivent faire face à la gestion d'un grand volume de données. Il est donc nécessaire, afin d'améliorer la prise en charge clinique, de mettre en place des outils d'analyse fiables et rapides. Ils doivent permettre d'aider le biologiste à sélectionner une dizaine de variants potentiellement actionnables, parmi les quelques dizaines de milliers de variants que l'on peut mettre en évidence dans un large panel de gènes. Enfin, ces outils d'analyses doivent fournir de manière optimale et automatisée une liste de thérapies ciblées en fonction de ces variants. Le développement d'un tel outil de sélection des variants et des thérapies associées, permettant l'accès à une médecine personnalisée pertinente, est aussi un objectif de cette thèse en génétique des tumeurs.

1.3.5 Classification au service de l'interprétation

Lors du séquençage de larges panels de gènes que ce soit en génétique constitutionnelle ou tumorale, plusieurs dizaines de milliers de variants sont identifiés chez un patient. Il est nécessaire de pouvoir fournir un sens biologique à ces variants, notamment au niveau de leur impact sur un phénotype, pour pouvoir les interpréter.

C'est pourquoi de nombreuses classifications ont été proposées afin d'ordonner les variants dans différentes catégories, en fonction de leur pathogénicité.

1.3.5.1 Les classifications en génétique constitutionnelle

La classification du *American College of Medical Genetics and Genomics* (ACMG) propose un modèle de classification largement utilisé de nos jours. D'autres modèles de classification existent [34] et d'autres modèles sont spécifiques pour les gènes *BRCA1* et *BRCA2* [35]. La classification ACMG fournit une liste de bonnes pratiques à suivre pour interpréter les variants afin de les classer sur une échelle de 5 classes allant de la classe 1, variant bénin, à la classe 5, variant pathogène [36]. Cette classification est composée de nombreux critères basés sur :

— La fréquence des mutations dans la population
— Les prédictions sur les conséquences sur la protéine du gène
— Des études fonctionnelles
— L'utilisation de bases de données et de connaissances

La combinaison des différents critères permet de conclure sur la pathogénicité du variant définie selon 5 termes : bénin (*benign*), probablement bénin (*likely benign*), variant de signification inconnue ou *Variant of Unknown Significance* (VUS), probablement pathogène (*likely pathogenic*) et pathogène (*pathogenic*). La classification ACMG propose un modèle de catégorisation des arguments de classification, qui alimente un arbre décisionnel. Au cours des dernières années, des mises à jour sont faites sur la classification et certains biais sont pointés notamment au niveau de la réinterprétation de certains variants en fonction du phénotype de l'individu [37, 38].

1.3.5.2 Classification somatique

En génétique somatique des tumeurs, les classifications peuvent se regrouper en deux catégories.

La première vise à catégoriser les variants en fonction de leur impact sur le développement du cancer chez un individu. La classification réalisée par *Personalised Medicine Commission* (ComPerMed) en est l'exemple. A l'image de la logique de la classification ACMG, la classification ComPerMed prédit la pathogénicité du variant sur une échelle entre le statut bénin et le statut pathogénique [39]. Tous les variants fréquents en population générale sont classés comme bénins par exemple. Les variants pathogènes sont définis à l'aide d'une liste tenue à jour par ComPerMed. Enfin, les variants de signification inconnue (ou *Variant of Unknown Significance* (VUS)) ou

catégorisés probablement pathogènes, sont évalués en fonction de la nature du variant et du rôle du gène dans l'oncogenèse, de sa localisation sur le locus, en fonction des informations présentes dans des bases de données externes (ClinVar, COSMIC) recensant les connaissances fonctionnelles du variant.

La deuxième catégorie concerne les classifications qui mettent en avant l'actionnabilité des variants. Plusieurs propositions de classification se sont développées dans diverses régions du monde. La classification *American Society of Clinical Oncology/Association for Molecular Pathology* (ASCO/AMP) [40] a été proposée aux Etats-Unis et la classification de l'*European Society for Medical Oncology* (ESMO) [41] en Europe. La classification ASCO/AMP fournit des indications pour tout type de tumeurs, la classification ESMO est spécialisée dans les variants de tumeurs solides. Ces classifications prennent en compte les thérapies approuvées par la *Food and Drug Administration* (FDA) pour les Etats-Unis, l'équivalent de l'*European Medicines Evaluation Agency* (EMEA) et de l'Autorisation de Mise sur le Marché (AMM) en France. Les variants associés à une thérapie autorisée (ou approuvée) sont attribués à un rang élevé dans la classification. Cette dernière est découpée en 4 groupes appelés tiers. Le tiers 4 correspond aux variants sans importance clinique et le tiers I correspond à ceux avec la plus grande importance clinique (Figure 1.5). Ces variants de tiers I peuvent orienter vers une thérapie spécifique, qui fait consensus dans la communauté clinique, en s'appuyant sur la littérature scientifique de ces variants. Les variants pouvant faire recommander des thérapies à partir d'arguments indirects, comme une action prouvée, mais sur d'autres types de cancer que celui du patient, à partir de résultats d'essais cliniques ou précliniques, sont classés dans le tiers II. Le tiers 3 regroupe des variants avec une importance clinique inconnue (voir exemple de la classification ASCO/AMP au travers de la Figure 1.5).

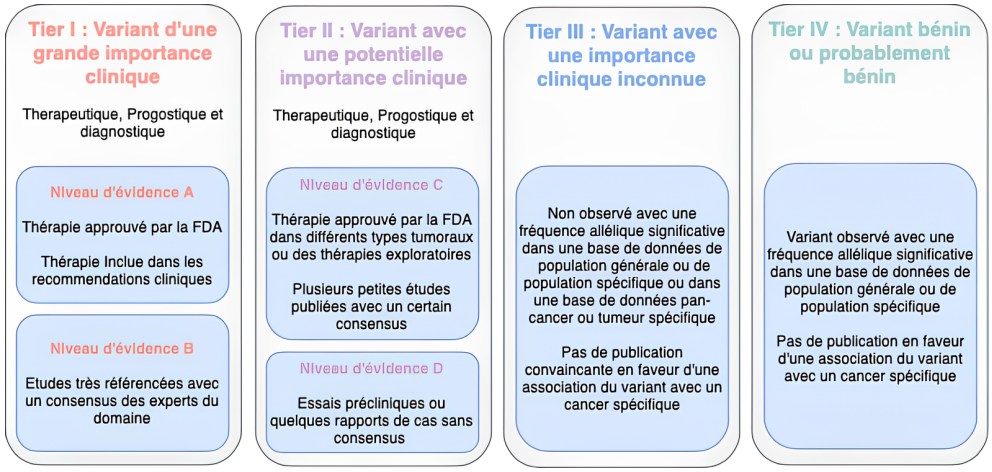


Figure 1.5 – Description de la classification ASCO/AMP
La classification ASCO/AMP se divise en 4 tiers catégorisant les variants somatiques en fonction de leur importance clinique. Plus le tiers est élevé plus il tend vers un variant supporté par des thérapies autorisées en routine clinique. Les tiers I et II se décomposent chacun en 2 sous-tiers différents : IA, IB, IIC et IID. Ces 4 sous tiers comportent des variants pouvant être utiles pour une action thérapeutique, diagnostique ou prognostique.
Adapté d'après Li et al. 2017

1.4 Pathologies moléculaires

Les classifications précédentes sous-entendent qu'il est nécessaire de faire la preuve de la causalité ou de la pathogénicité d'un variant. Un variant causal en oncogénétique est un variant associé à un risque accru de développer un cancer. La causalité d'un variant tumoral dans le développement d'une tumeur est évaluée au travers de deux aspects (i) la caractérisation du rôle du gène porteur de la mutation dans l'oncogenèse spécifique de l'organe (ii) l'identification de la mutation et de ses conséquences sur les gènes.

1.4.1 Les gènes impliqués dans l'oncogenèse

Avec l'avènement des nouvelles techniques de séquençage, le paysage et la caractérisation des variants pathogènes se précisent. Depuis la création de la première liste de gènes impliqués dans l'oncogenèse, par le *Cancer Gene Census* (CGC) en 2004 [42], de nombreuses analyses et initiatives ont eu pour but de caractériser le rôle

des gènes dans l'oncogenèse. Parmi elles on peut citer les travaux de Vogelstein en 2013[43] où 140 gènes ont été caractérisés et associés à des voies de signalisation. Des études basées sur les analyses de TCGA [44] ou du consortium international *International Cancer Genome Consortium* (ICGC) [45] ont permis de définir de nouveaux gènes et voies de signalisation impliqués dans les cancers. Plus récemment la caractérisation du rôle des gènes dans l'oncogenèse a progressé et alimente les analyses avec les prédicteurs *in silico* de l'impact des gènes dans l'oncogenèse [46, 47] et les bases de connaissances [9, 48]. L'information des rôles des gènes impliqués dans l'oncogenèse est recensée dans des bases de données comme OncoKB, CGC de COSMIC [32] ou *Integrative Onco Genomics* (IntOGen) [49]. Les gènes responsables de l'oncogenèse sont regroupés en deux types : les *Tumor Suppressor Gene* (TSG) et les oncogènes.

1.4.1.1 Les gènes suppresseurs de tumeur

L'hypothèse de Knudson a proposé en 1971 un modèle mathématique décrivant deux formes de rétinoblastomes [50] :

— Une forme bilatérale de rétinoblastome survenant dans les premiers mois de vie, au sein de cas familiaux.
— Une forme unilatérale de rétinoblastome diagnostiquée plus tardivement dans les cas sporadiques.

Selon le modèle de Knudson, deux mutations au sein d'une cellule rétinienne provoqueraient le développement d'un rétinoblastome. Dans la forme bilatérale, si une mutation constitutionnelle est déjà présente (premier *hit*), la probabilité de développer un rétinoblastome dépend alors de l'apparition d'une mutation somatique (deuxième *hit*) sur le deuxième gène cible. Donc un patient héritant d'une mutation constitutionnelle a un risque élevé de développer de manière précoce un rétinoblastome. En 1973, Comings complète le modèle de Knudson [51] supposant que deux mutations sur deux allèles d'un même gène provoqueraient son inactivation et donc le développement d'un rétinoblastome. L'analyse d'un grand nombre de rétinoblastomes a mis en évidence une région de délétion commune en 13q14 et en 1993, Toguchida caractérise la séquence complète du gène *RB1* [52], responsable du développement du rétinoblastome. La caractérisation du gène *RB1* confirme l'hypothèse de Knudson et ouvre la voie au diagnostic moléculaire des prédispositions aux cancers. C'est ainsi la première caractérisation de l'hypothèse du *two-hits* ainsi que des TSG permettant, par la suite, l'identification de nouveaux TSG comme *BRCA1*, *BRCA2* ou même *TP53* et son implication dans le syndrome Li-Fraumeni [53].

Les variants pathogènes dans les TSG sont généralement des mutations inactivatrices, c'est-à-dire qui vont provoquer la traduction d'une protéine non fonctionnelle

ou la perte du messager du gène sous l'action du *Non-sens Mediated mRNA Decay* (NMD) [54] ou des anomalies de la transcription d'épissage. L'inactivation d'un TSG favorise l'oncogenèse étant donné que les TSG régulent négativement la prolifération cellulaire (voir Figure 1.6) [55].

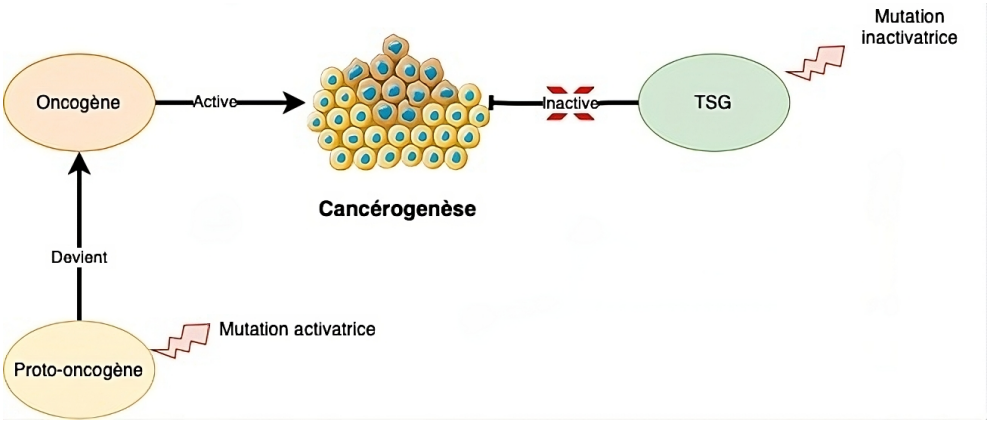


Figure 1.6 – Rôle des oncogènes et des gènes suppresseurs de tumeurs.
Un proto-oncogène sera appelé oncogène s'il subit une mutation généralement activatrice, responsable de l'augmentation de l'oncogenèse. Un gène suppresseur de tumeurs réduit l'oncogenèse. Si une mutation délétère a lieu sur ce gène, il perd son activité inhibitrice et donc favorise l'oncogenèse.

1.4.1.2 Les oncogènes

Les proto-oncogènes sont des gènes qui codent pour des protéines stimulant la division cellulaire, l'inhibition de la différenciation cellulaire et la régulation de la mort cellulaire [56]. L'apparition d'un variant pathogène sur un proto-oncogène, induit généralement la traduction de protéines fonctionnelles dont l'activité est modifiée ou insensible aux mécanismes d'inhibitions physiologiques. Le gène muté est alors défini comme oncogène. La distinction entre proto-oncogène et oncogène, qui est le terme courant, est rarement faite [57]. Les conséquences peuvent être une augmentation de la division cellulaire, une diminution de la différenciation cellulaire et enfin l'inhibition des mécanismes de mort cellulaire (voir Figure 1.6).

Dans le cadre de leurs recherches sur la Leucémie Myéloïde Chronique (LMC), dès 1960, Nowell et Hungerford découvrent une anomalie chromosomique connue sous le

nom de chromosome Philadelphie [58]. Cette découverte a été la première anomalie génétique directement liée à un cancer, corroborant l'hypothèse de l'implication d'anomalies génétiques dans les cancers. En 1973, Rowley définit l'anomalie chromosomique par une translocation réciproque t(9 ;22)(q34 ;q11) [59]. Des études transversales d'un virus induisant la leucémie chez la souris ont permis d'identifier le gène causal *v-ABL* alors retrouvé plus tard chez l'homme sous le nom *ABL1* [60, 61]. En 1984, les travaux de Groffen et son équipe mettent en évidence une région présentant des points de cassures d'une translocation génomique chez 17 patients [62]. Située au niveau du chromosome 22, ils nommèrent cette région *breakpoint cluster region* menant à la caractérisation du gène *BCR* [62]. La translocation t(9 ;22)(q34 ;q11) est alors définie comme la fusion *BCR-ABL1.* Ce gène chimérique a pour conséquence de déréguler l'activité thyrosine kinase d'*ABL1* entraînant alors (i) une augmentation de la croissance cellulaire (ii) une indépendance aux facteurs de croissance (iii) une inhibition du processus d'apoptose.

Certains oncogènes sont à présent des cibles thérapeutiques [43, 9]. Par exemple, dans les cancers métastatiques du sein, 20% sont classifiés *Human Epidermal growth factor Receptor 2 positive* (HER2+), une forme agressive de cancer du sein de mauvais prognostic [63]. Le développement de l'Herceptine (Trastuzumab) a permis de grandement améliorer le traitement de ces types de cancers [64, 63], avec des taux de survie de 90% sur les phases précoces. D'autres oncogènes comme *EGFR* dans les cancers bronchiques non à petites cellules (associé à des thérapies comme Gefitinib ou Erlotinib) [65, 66], *RET* dans les cancers de la thyroïde (associé à des thérapies comme Selpercatinib) [67, 68] ou *ERBB2* dans les cancers colorectaux métastatiques (associé à des résistances à des thérapies comme Cetuximab et Panitumumab) [69, 70] sont aussi des gènes dont les mutations pathogènes activatrices amplifient l'activité kinase des protéines codées par ces gènes.

1.4.2 Les différents types de mutations et leurs conséquences

Dans cette partie, nous introduirons la notion de génome de référence, les différents types de variants déduits de cette référence ainsi que leurs conséquences sur le génome humain et le développement tumoral.

1.4.2.1 Les génomes de référence

Le *Genome Reference Consortium* (GRC) propose en 2009 un assemblage de l'ensemble du génome humain[3]. L'objectif de cette démarche étant de proposer une séquence de référence commune aux analyses génomiques, afin que les analyses en génétique se basent sur la même référence, dans l'objectif de détecter les différences

entre le génome de la référence et de l'individu. Historiquement, le *Human Genome Project* (HGP), à l'origine de l'assemblage du premier génome humain, a proposé 18 versions (jusqu'à hg18 pour *human genome* version 18) avant que le GRC poursuive les travaux d'assemblage du génome humain. Le GRC a d'abord proposé une première version appelée hg19 ou *Human genome assemblies by the Genome Reference Consortium, build 37* (GRCh37) ayant subi plusieurs améliorations mineures jusqu'en 2013 pour aboutir à une version appelée GRCh37.p13. En 2016, le GRC publie une amélioration majeure de l'assemblage GRCh37, avec la version hg18 (GRCh38) [71].

Cette nouvelle version majeure propose des correctifs d'erreurs d'assemblage. En s'appuyant sur la table 1.3, la dernière version du GRCh38 couvre 0.07 Gigabase (Gb) de plus que la dernière version du GRCh37 pour une meilleure identification des bases avec 0.15 Gb résolues par rapport au GRCh37. Cela s'explique notamment par l'assemblage de meilleure qualité du GRCh38 supporté par plus de *contigs* (séquences continues et ordonnées résultantes de l'assemblage) possédant une valeur de qualité, appelée N50, élevée. Dans un ensemble de *contigs*, la valeur N50 est définie comme la longueur de séquence du *contig* le plus court à 50% de la longueur totale de l'assemblage. Avec une N50 élevée GRCh38 définit des *contigs* plus longs, permettant de réduire la taille de zones non couvertes, appelées *gaps*, entre *contigs*. Enfin le GRCh38 dispose de 252 *loci* alternatifs de plus que le GRCh37 aussi représentés par des *contigs* alternatifs. Les *loci* alternatifs correspondent à une représentation alternative d'une séquence d'un *locus* et leur taille est inférieure à 1 Megabase (Mb) [72]. Ces *loci* alternatifs représentent des séquences variables dans la population et permettent de limiter le nombre de faux positifs lors d'étapes effectuant la correspondance entre la séquence d'un individu et celle d'une référence (assemblage par exemple).

	GRCh38.p.14	GRCh37.p.13
Date de sortie	03/02/2022	28/06/2013
Nb total de base	3.3 Gb	3.23 Gb
Nb total de base sans N	3.14 Gb	2.99 Gb
Loci alternatifs	261	9
N50	67.79 Mb	46.40 Mb

Table 1.3 – Métriques autour de l'assemblage des génomes de référence GRCh38 et GRCh37
Selon le *Genome Reference Consortium*

Actuellement la dernière version datant de Février 2022 est la GRCh38.p14. En effet, comme vu dans la table 1.3, le nombre total de bases résolues et le nombre de bases non résolues (N) est différent entre les deux versions. Ainsi les positions génomiques des gènes peuvent être différentes d'une version à l'autre. Les étapes d'annotations dépendent de ces positions pour enrichir la connaissance sur un variant. Cette annotation est une étape majeure dans l'interprétation que l'on fait d'un variant. Ainsi certains laboratoires d'analyses, disposant d'outils calibrés pour l'interprétation de variant sur GRCh37, n'ont pas encore réalisé de migration vers la nouvelle référence GRCh38 en raison de la lourdeur des étapes de validation associées.

De nouvelles initiatives voient le jour pour affiner l'assemblage du génome humain afin de tendre vers une réduction des erreurs et la diminution du nombre de *gaps*. C'est notamment le cas pour le consortium *Telomere-To-Telomere* (T2T) proposant en 2022 une version améliorée de la référence du GRC avec une meilleure résolution des régions hétérochromatiques du génome humain par rapport au GRCh38 [73]. En 2023 le consortium complète l'assemblage du T2T en résolvant l'assemblage du chromosome Y, jusqu'alors assemblé avec de nombreux *gaps* en raison de la présence de multiples régions répétées et d'insertions microsatellitaires [74].

1.4.2.2 Les types de variants

Les différents types de variant que l'on peut détecter dans un génome se répartissent en 3 grandes catégories :

1.4.2.2.1 Variation d'un nucléotide (mononucléotidique)

Les variations mononucléotidiques regroupent les substitutions, délétions ou insertions d'un nucléotide par rapport à un nucléotide d'une référence. Ces variations sont aussi appelées *Single-Nucleotide Variant* (SNV). Quand un SNV est présent dans au moins 1% de la population on parle alors de *Single Nucleotide Polymorphism* (SNP).

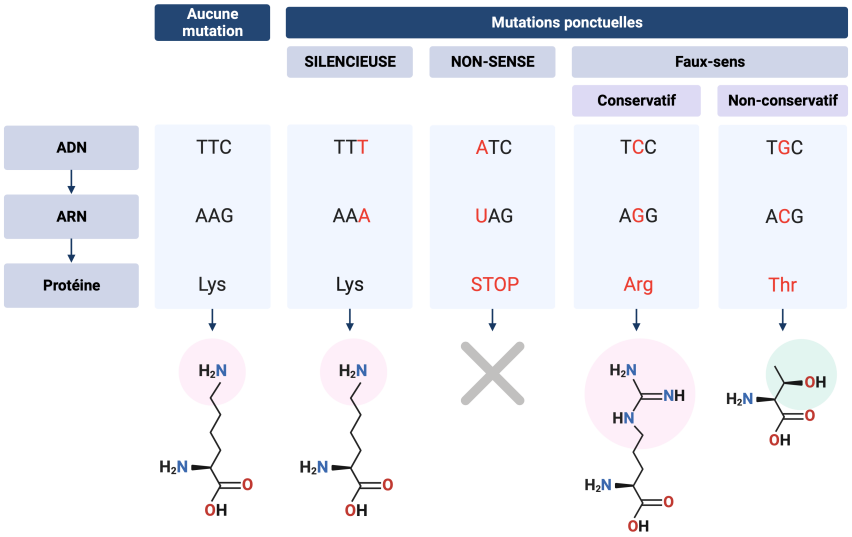


Figure 1.7 – Les types de variants mononucléotidiques
Adapté d'après *Point Mutation, BioRender*

Les substitutions des mononucléotides sont décrites selon 3 types (voir Figure 1.7) :

— Les mutations silencieuses ou isosémantiques : ce sont des mutations qui n'engendrent pas de changement sur le codon de par la redondance du code génétique
— Les mutations non-sens : substitutions qui provoquent l'apparition d'un codon stop prématuré
— Les mutations faux-sens : substitutions provoquant un changement d'acide aminé par un autre acide aminé de composition chimique similaire (on dira que le faux-sens est conservatif) ou différente (on dira que le faux-sens est non-conservatif). La similarité des acides aminés peut être évaluée par des matrices de distances comme la matrice *BLOcks SUbstitution Matrix* (BLOSUM) [75]

1.4.2.2.2 Les INDEL

Une Insertion-délétion (INDEL) est une mutation impliquant une insertion et/ou une délétion d'une séquence de nucléotide de moins de 50 paires de bases ou *Base Pairs* (BP) [76]. Par convention, au-delà de cette taille, les INDEL sont considérées comme variants structuraux. Les INDEL sont les deuxièmes types de variants les

plus fréquents, représentant 1% des mutations du génome [77].

1.4.2.2.3 Les variants structuraux

Les variants structuraux appelés *Structural Variant* (SV) sont des altérations génomiques d'au moins 50 BP [76]. Ils regroupent plusieurs catégories d'événements comme les délétions, les duplications, les insertions, les translocations et les inversions [78]. Un SV peut être composé de plusieurs de ces événements [79], ils seront alors définis comme complexes. Leur incidence est beaucoup plus importante au niveau des régions télomériques [80]. Les SV peuvent impliquer l'insertion d'éléments mobiles ou *Mobile Element Insertion* (MEI). Les MEI sont des séquences d'Acide désoxyribonucléique (ADN) pouvant s'insérer en nouvelle copie ailleurs dans le génome. Chez l'humain, seulement trois familles de MEI ont conservé leur capacité d'insertion : les LIN1, Alu et SVA [81, 82].

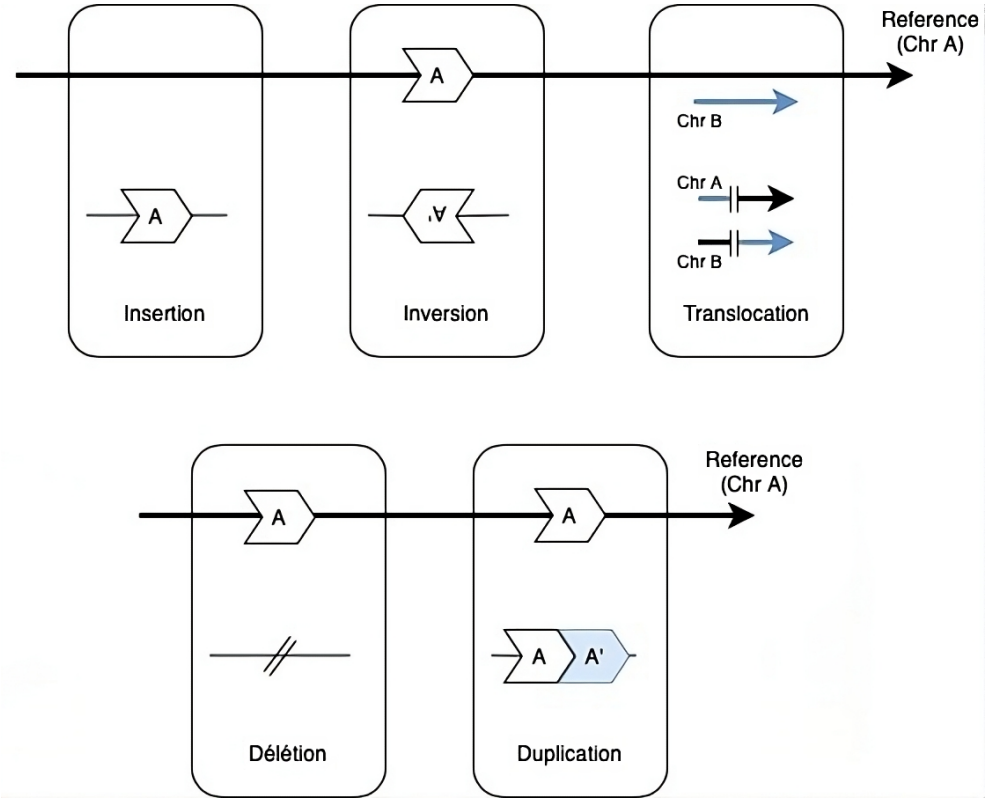


Figure 1.8 – Les types de variants structuraux
Cinq catégories de variants structuraux sont décrits. La translocation décrite représente une translocation inter-chromosomique.
Adapté d'après GATK, variant structural

Les différentes catégories de SV peuvent être regroupées en deux grandes familles d'anomalies [83] :

— Les anomalies équilibrées : concernent les anomalies qui ne vont pas induire de perte ou un gain du matériel génétique. Les inversions se rangent dans cette catégorie d'anomalies (voir Figure 1.8).

— Les anomalies déséquilibrées : concernent les anomalies qui vont induire une perte ou un gain du matériel génétique. On inclut dans cette catégorie les duplications et les délétions (voir Figure 1.8). Les SV déséquilibrés sont souvent associés au terme CNV.

Aussi, les translocations peuvent être intrachromosomiques lorsque les deux régions impliquées dans la translocation se situent sur le même chromosome. Une

translocation interchromosomique se définit lorsque les deux régions impliquées dans la translocation se situent sur deux chromosomes différents. Une translocation peut être équilibrée ou non.

1.4.2.3 Les conséquences des variants sur la transcription et la traduction d'un gène

L'ensemble des variants présentés précédemment peut avoir des conséquences différentes en fonction de leur position dans un gène et de leur nature. Les variants non-sens et faux-sens ont un impact direct sur la protéine résultante, pouvant provoquer une troncation de la protéine avec l'apparition d'un codon stop prématuré ou modification de la séquence peptidique (voir Figure 1.9). Ces variants, mais également les variants isosémantiques et les INDEL, peuvent également impacter les sites d'épissage aboutissant à des épissages aberrants [84]. Des variants dans les régions promotrices ou séquences régulatrices, tels que l'épissage (étape de maturation de l'ARN qui sera développée dans le paragraphe "Les anomalies liées à l'épissage" de la section 1.4.2.4.2), peuvent aussi entraîner des anomalies de l'expression [85, 86]. C'est pourquoi il est important d'analyser l'impact d'un variant à la fois au niveau de l'ARN et de la protéine.

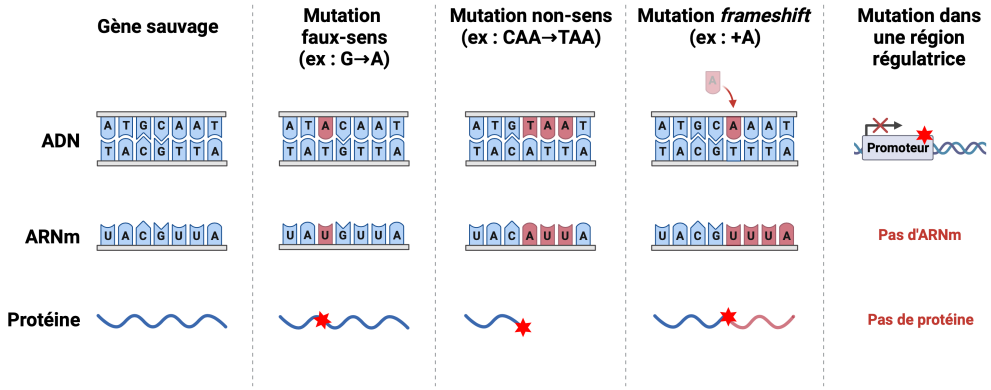


Figure 1.9 – Effets communs des mutations sur une protéine
Adapté d'après Effects of Common Mutations, BioRender

Les INDEL aboutissent à un décalage du cadre de lecture, aussi appelé *frameshift*, à l'origine d'un codon stop prématuré ou dans des cas plus rares à un allongement de la protéine (*stop loss*). Les INDEL peuvent être présentes en phase du cadre de lecture (*in-frame*). Ces variants impliquent l'ajout ou la suppression d'acides aminés pouvant altérer la fonction de la protéine, son interaction avec d'autres protéines et sa stabilité.

Parmi les SV on retrouve plusieurs conséquences :

(i) Les *Copy Number Variation* (CNV) peuvent concerner une délétion ou duplication (gain) partielle ou complète d'un gène[87] (voir Figure 1.8, cas des délétions et duplications). Cependant certains CNV sont non pathogènes en fonction de la position de la région concernée et des gènes impactés [88]. Enfin, certains peuvent être fréquents en population générale ($> 1\%$), correspondant à des polymorphismes, on parle alors de *Copy Number Polymorphism* (CNP).

(ii) Les fusions de gènes impliquent des réarrangements chromosomiques pouvant être responsables d'un échange entre gènes d'une séquence codante ou régulatrice de l'ADN. Ces mécanismes de réarrangements, généralement retrouvés dans les tumeurs, peuvent être initiés par des insertions, des duplications, des délétions ou des translocations (voir Figure 1.10). Si le point de cassure (*breakpoint*) génomique est dans le cadre de lecture (*in-frame*), la néo-protéine formée de la fusion des deux gènes, a une forte probabilité d'avoir un effet oncogénique [89]. Si un point de cassure génomique est en dehors du cadre de lecture (*out-of-frame*), la néo-protéine sera aberrante, ayant pour conséquence d'être tronquée et donc non fonctionnelle.

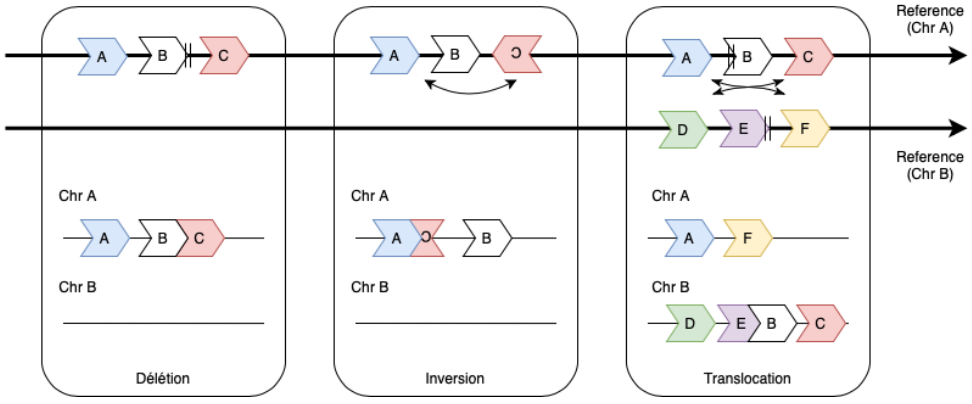


Figure 1.10 – Les principaux types de fusions de gènes
Adapté d'après Le *Broad Institute*

1.4.2.4 Impact sur la fonctionnalité du gène

1.4.2.4.1 Modification d'acide aminé, gain et pertes de fonction

La conséquence d'une mutation est dépendante, en partie, du rôle du gène dans lequel elle apparaît. Comme décrites dans la section 1.4.1, les mutations ayant pour conséquence une perte de fonction auront un impact probablement pathogénique sur les TSG (exemple des gènes du syndrome HBOC). L'apparition d'un codon stop prématuré peut aboutir à la synthèse d'une protéine tronquée. Cela peut, à l'image des *frameshift*, impacter des gènes comme les TSG qui perdront leur fonction en faveur de l'oncogenèse. Néanmoins il est possible que les transcrits présentant un codon de terminaison précoce soient pris en charge par le *Non-sens Mediated mRNA Decay* (NMD). C'est un complexe de protection du matériel génétique qui a pour but de détruire les transcrits anormaux. Cette destruction peut engendrer une diminution de l'expression du transcrit du gène muté et ainsi diminuer l'expression des TSG [54]. A l'inverse, les mutations dites activatrices auront un impact probablement pathogénique, comme par exemple un gain sur un oncogène. La fusion *BCR-ABL1*, l'amplification de *ERBB2* et des mutations sur le gène *BRAF* (mutation V600E) [90] ou sur le gène *EGFR* (mutation T790) [91], sont des exemples de mutations activatrices impliquées dans l'oncogenèse.

1.4.2.4.2 Les anomalies liées à l'épissage

L'épissage est un mécanisme permettant la formation d'un Acide Ribonucléique messager (ARNm) mature à partir d'un pré-ARN (voir Figure 1.11). Ce processus se décompose en plusieurs étapes. L'épissage est amorcé par la reconnaissance des sites d'épissage. Ces derniers correspondent à un couple de deux acides nucléiques : un site donneur GU à l'extrémité 5' de l'intron et un site accepteur AG à l'extrémité 3' de l'intron. Les sites d'épissage sont reconnus par des complexes protéiques appelés petites ribonucléoprotéines nucléaires (pRNPn) qui forment le spliceosome après s'être liés aux sites d'épissage. Il est composé de 5 différents pRNPn : U1, U2, U4, U5 et U6 ainsi que de protéines régulatrices. Les ARN U6 et U2 en se liant aux séquences conservées en 5' de l'intron vont former la boucle lasso, essentielle à l'alignement des sites d'épissage et du retrait des introns sur l'ARN pré-messager. Le spliceosome, en catalysant la coupure des extrémités 5' de l'intron en la reliant à l'extrémité 3' de l'exon suivant de l'ARN pré-messager, aboutit à la formation d'un ARNm mature.

Une molécule d'ADN peut, après les étapes de transcription et d'épissage, donner différents ARNm. Ce procédé est appelé épissage alternatif et permet d'obtenir des

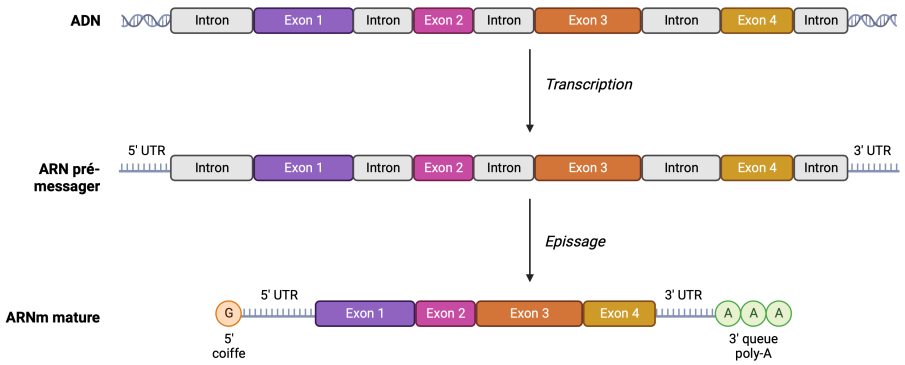


Figure 1.11 – Transcription et épissage de l'ADN à l'ARNm
Adapté de RNA processing in eukaryotes, BioRender

combinaisons d'exons différentes sur chaque isoforme d'ARNm. Mis en évidence en 1977 au travers des travaux de Chow, Gelinas, Broker et Roberts [92], il permet d'augmenter pour un génome donné, la diversité des protéines traduites. On dénote 5 mécanismes impliqués dans l'épissage alternatif (voir Figure 1.12) :

— Saut d'exon
— Site d'épissage alternatif en 5'
— Site d'épissage alternatif en 3'
— Rétention d'intron
— Exons mutuellement exclusifs

Le saut d'exon (ou *exon skipping*) se produit entre exons cassettes, c'est-à-dire des exons bordés par des introns en 5' et 3'. Le site d'épissage en 5' de l'exon est omis lors de l'épissage, entraînant la jonction avec les exons adjacents. La production d'une protéine tronquée ou la modification de la séquence protéique peut être une conséquence de ce type d'épissage alternatif.

Des sites d'épissages alternatifs en 5' ou 3' peuvent être engagés lors de l'épissage. Situés aux extrémités 5' ou 3' de l'intron en 5' de l'exon, ils peuvent conduire à l'inclusion ou l'exclusion d'un exon spécifique. La modification de la séquence protéique ou l'apparition d'un codon stop prématuré peut être une conséquence de ce type

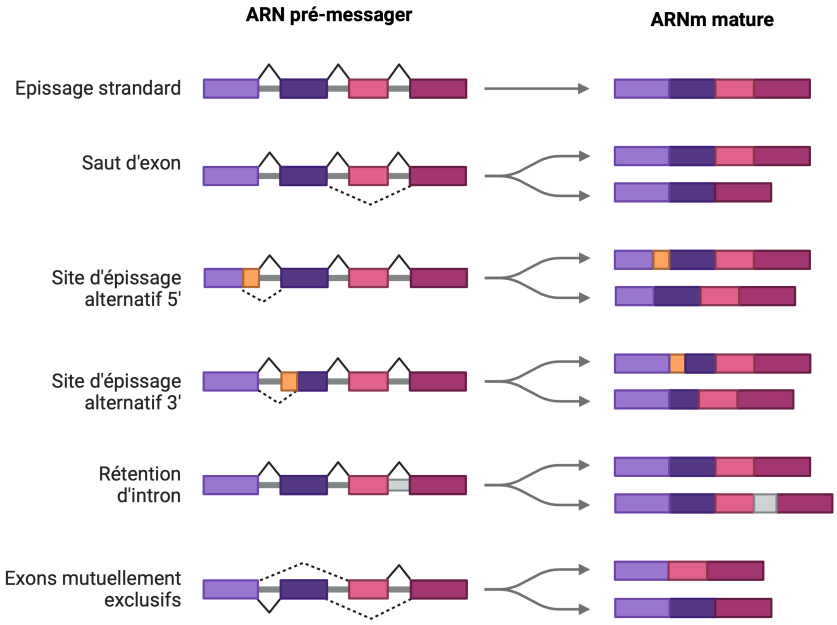


Figure 1.12 – Les différents types d'épissages alternatifs
Adapté de mRNA splicing types, BioRender

d'épissage.

La rétention d'intron est définie par la conservation d'un intron dans l'ARNm mature au lieu d'être éliminé lors de l'étape d'épissage. L'apparition d'un codon stop prématuré est fréquente dans ce cas.

Enfin les exons mutuellement exclusifs correspondent à un type d'épissage où deux exons ou plus sont mutuellement exclusifs, signifiant que seul l'un des exons est inclus dans l'ARNm mature. L'inclusion d'un exon exclut automatiquement les autres exons mutuellement exclusifs. Ce type d'épissage permet de générer plusieurs isoformes d'ARNm avec des séquences protéiques très distinctes.

Des éléments régulateurs peuvent renforcer ou affaiblir les interactions entre les sites d'épissage et leurs facteurs de régulation, modulant ainsi l'épissage alternatif. Appelés *splicing enhancers* et *splicing silencers* ils peuvent favoriser l'utilisation ou l'inhibition de certains sites d'épissage influençant la composition des exons de l'ARNm mature. L'épissage alternatif peut réguler l'expression génique en modifiant la quantité d'ARNm produit ou en influençant la stabilité de l'ARNm. L'épissage

alternatif joue un rôle dans l'adaptation à différents états physiologiques ou environnementaux, en faveur de la survie des cellules [93].

La régulation de l'épissage et les isoformes d'ARNm jouent un rôle essentiel dans la régulation de l'expression génique et dans la diversité des protéines traduites. Depuis plusieurs années, les interactions entre les anomalies de l'épissage alternatif et les cancers sont de plus en plus décrites. L'épissage alternatif peut favoriser la prolifération de la tumeur voire inhiber l'apoptose (mort cellulaire) des cellules tumorales [94]. Des mutations génomiques peuvent influencer l'épissage alternatif. Elles peuvent altérer les séquences des sites d'épissage (donneur en 5' et récepteur en 3') perturbant la liaison des facteurs d'épissage. De plus, elles peuvent affecter les séquences régulatrices et les gènes codant les facteurs d'épissage. Ces mutations entraînent alors un déséquilibre de la composition des ARNm et des protéines produites, ayant un impact direct sur la fonction cellulaire et donc le développement de maladies, y compris le cancer [95, 96]. Les anomalies génomiques (mutations du génome) peuvent donc être responsables d'anomalies du transcriptome (événements aberrants de l'ARN) à l'origine de maladies (développement tumoral par exemple).

En conséquence, les isoformes alternatives et les variants d'épissage sont considérés comme des biomarqueurs pertinents en diagnostic des cancers [97, 98, 99]. De plus l'analyse parallèle du génome et du transcriptome de l'individu s'est montrée efficace pour comprendre l'hérédité manquante des cancers, mettant en exergue de nouveaux types de mutations pathogènes héréditaires [100, 101]. Cependant, à ce jour, l'ensemble des profils d'isoformes des gènes impliqués dans la prédisposition aux cancers reste à caractériser. De même, l'utilisation de techniques accessibles *in vitro* et *in silico* pour les caractériser et les annoter, reste un défi à relever afin de progresser dans la connaissance de cette hérédité manquante.

1.5 Les séquençages du génome et du transcriptome

Afin de pouvoir détecter ces différentes mutations, depuis 2003 avec le séquençage complet du premier génome humain par le *Human Genome Project* (voir Figure 1.13 pour plus de détails sur la chronologie), les techniques de séquençage se sont développées en perfectionnant les capacités de séquençage. De nos jours, plusieurs terabases peuvent être séquencées avec un coût de séquençage en baisse, une meilleure précision de séquençage (réduction du taux d'erreur) et avec une amélioration des protocoles de séquençage, augmentant la qualité des séquençages.

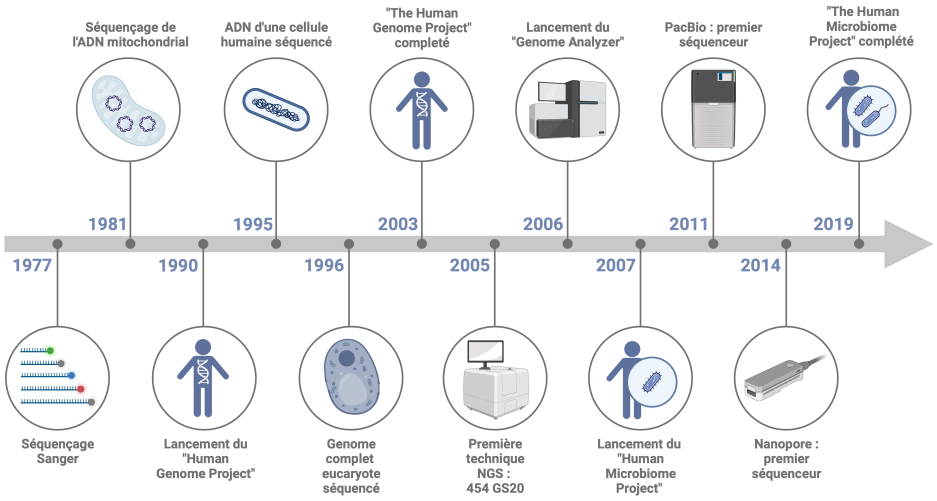


Figure 1.13 – Histoire du séquençage de l'ADN
Adapté de The History of DNA Sequencing, BioRender

1.5.1 Séquençage Sanger

A la suite de ces travaux en 1977 sur une méthode de séquençage de l'ADN [102], Frederick Sanger reçoit un deuxième prix Nobel de chimie en 1980. Les méthodes modernes de Sanger utilisent des Didéoxynuclotide Tri Phosphate (ddNTP) fluorescents permettant, à l'aide de l'utilisation d'un laser lors de la migration sur gel, d'utiliser les différentes fréquences émises par l'interaction du laser sur les fluorochromes pour définir le nucléotide (voir Figure 1.14).

Le séquençage Sanger a été une première révolution dans le monde de la génomique. Le premier génome humain complet a été séquencé notamment à l'aide de cette technique. Quelques limitations inhérentes à la méthode de ce type de séquençage existent :

— La taille des fragments d'ADN : même s'il est possible d'atteindre 1000 bases, la longueur des séquences reste un facteur limitant pour la résolution de longues régions ADN ou l'analyse de régions répétées.
— Le coût et le débit : le séquençage Sanger nécessite des réactifs coûteux et n'est pas compatible avec l'analyse de nombreux échantillons.
— Limite de détection : les variants à faible fraction allélique sont difficilement interprétables, tels que les variants en mosaïque ou les variants tumoraux,

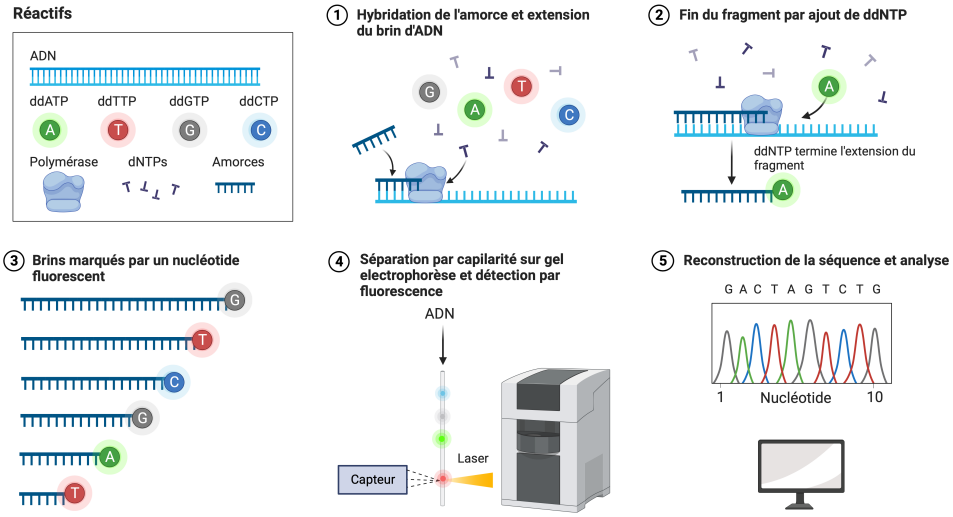


Figure 1.14 – Le Séquençage Sanger
Le protocole présenté ici inclut l'utilisation de fluorochromes liés au ddNTP.
Adapté de Sanger Sequencing, BioRender

dont la fraction allélique peut être diluée par la présence de tissu sain, dans l'échantillon analysé.

— Qualité : la qualité du séquençage Sanger se dégrade après 700 bases et peut se trouver impactée au début du fragment.

Le séquençage Sanger reste encore très utilisé notamment en clinique car très accessible et utile en technique orthogonale pour confirmer des variants identifiés en NGS ou pour mettre en place un test présymptomatique rapide et peu coûteux.

1.5.2 Séquençage haut-débit : *short-read*

Aujourd'hui de nouvelles méthodes de séquençage ont été développées et font entrer les techniques de séquençage dans l'ère du séquençage haut-débit ou *High Throughput Sequencing* (HTS). Aussi appelées *Next-Generation Sequencing* (NGS), ces nouvelles techniques reposent sur le séquençage en simultané de plusieurs millions de fragments d'ADN. De nos jours, les NGS sont associées aux séquenceurs de la société Illumina, devenue leader sur le marché.

1.5.2.1 Principes du séquençage Illumina

Le séquençage Illumina est un séquençage par synthèse d'ADN. L'ADN est fragmenté par fragmentation enzymatique ou par sonification. Des adaptateurs sont ligués aux extrémités des fragments (voir Figure 1.17 (1) et Figure 1.15). Ils ont 3 grandes fonctions :

— La génération d'amplifiats via les régions P5 et P7 interagissant avec les oligonucléotides présents à la surface des *flow cells* (*vide infra*)
— La reconnaissance des échantillons par l'ajout d'un couple d'index nucléotidiques choisi pour être unique au cours du séquençage de plusieurs échantillons. Cela permet le multiplexage (index 1 et 2), donc le séquençage de différents échantillons simultanément, ceux-ci ayant un identifiant moléculaire unique.
— La fixation d'amorces de séquençage (PS Lecture 1 et 2) des lectures 1 et 2.

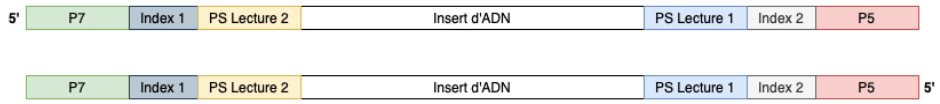


Figure 1.15 – Composition des adaptateurs Illumina
Adapté d'Illumina

Un adaptateur peut être couplé à un identifiant moléculaire unique ou *Unique Molecular Identifier* (UMI) permettant d'associer chaque molécule d'ADN initiale, à un identifiant. L'utilisation des UMI permet entre autres de limiter les biais induits par les erreurs de *Polymerase Chain Reaction* (PCR), en réalisant un consensus des fragments séquencés appelés *reads* disposant d'un même identifiant. De plus des variants avec une faible fraction allélique peuvent être mieux détectés. Un variant présent sur l'ensemble des *reads*, dérivés de la même molécule d'ADN, n'est probablement pas un artéfact de séquençage alors présent sur une faible proportion des *reads* partageant le même UMI [103, 104] (voir Figure 1.16).

Les différents fragments sont fixés sur une *flow cell*, surface en plaque de verre séparée en canaux où des oligonucléotides sont répartis dans des micropuits. Les fragments se fixent alors sur les oligonucléotides, la phase d'amplification débute. Chaque fragment bascule et s'hybride au *primer* d'un oligonucléotide à proximité. L'hybridation des fragments avec l'oligonucléotide complémentaire leur donne une structure en "pont" caractéristique de la chimie de la *flow cell*. Le fragment d'origine est retiré de la *flow cell* avec une étape de lavage, après sa réplication par l'ADN polymerase. Les étapes précédentes se répètent, amplifiant ainsi les fragments répartis

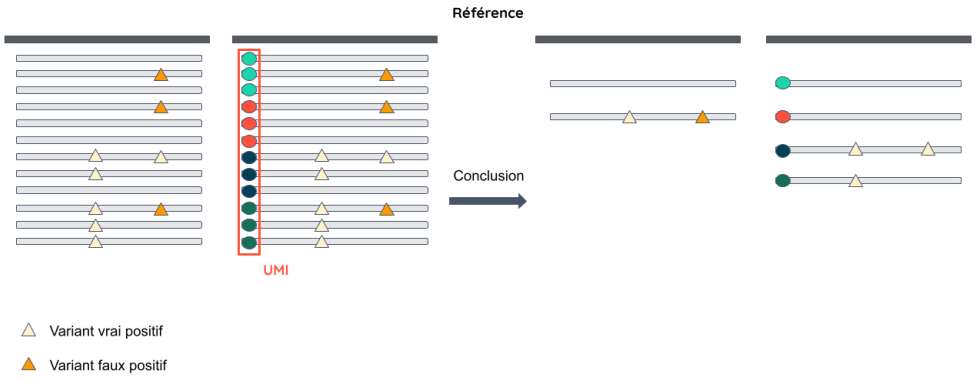


Figure 1.16 – Avantage de l'utilisation des UMI
Détection de variants supportés par des *reads* avec et sans l'utilisation des UMI. Les méthodes de consensus des *reads* marqués par des UMI permettent de retirer les variants faux positifs.

en *clusters* ; ils sont alors considérés comme amplifiats, donc clonaux (voir Figure 1.17 (2)).

Les nucléotides synthétisés sont marqués par des fluorochromes et chaque synthèse de nucléotide (ou cycle) émettra un signal lumineux après excitation par laser. Le nombre de cycles déterminera la taille de la lecture (ou *read*) et définit à terme la séquence nucléotidique : c'est l'étape de séquençage proprement dite. Les images générées par la fluorescence sont traitées par des logiciels spécialisés (inclus avec les séquenceurs Illumina). Ils convertissent les différentes fréquences émises par les fluorochromes en lettres correspondant aux nucléotides (A, T, C, G), c'est le *base calling* (voir Figure 1.17 (3)). Des méthodes de démultiplexage peuvent s'enchaîner au *base calling* afin de regrouper les *reads* de chaque patient entre eux, à condition que les index aient été rajoutés lors de la préparation des échantillons. Les suites du traitement bioinformatique seront explorées dans la section 1.6.

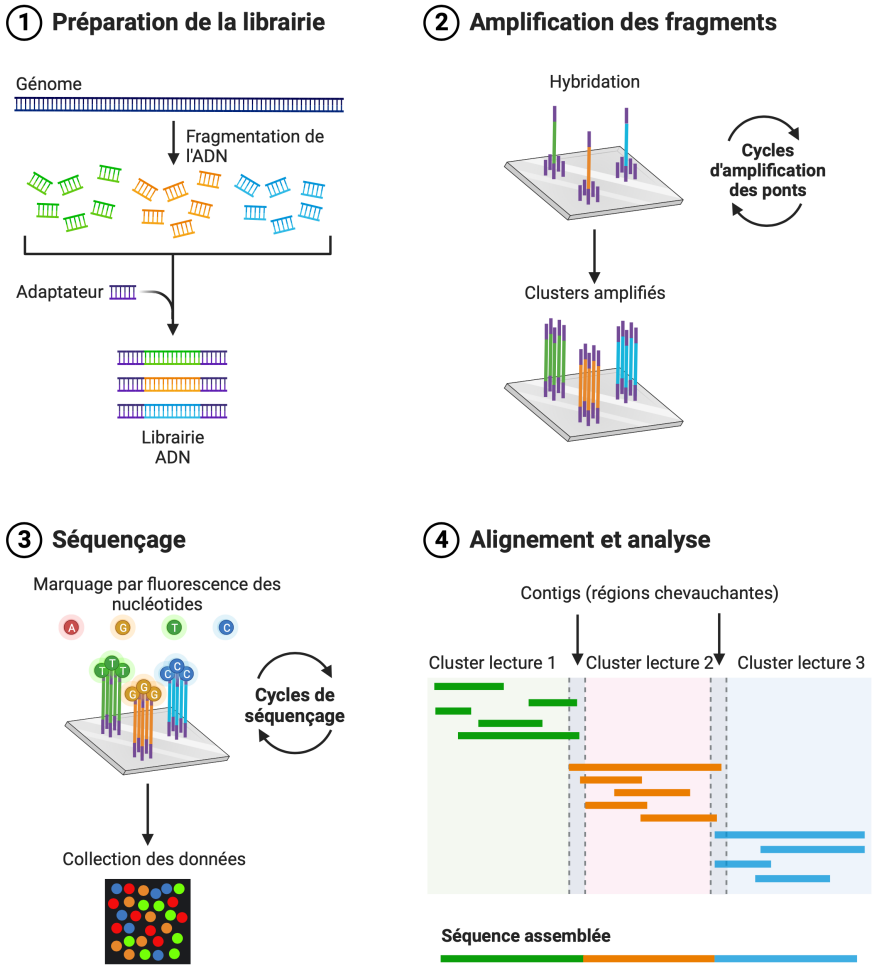


Figure 1.17 – Séquençage Illumina
Séquençage Illumina d'une librairie ADN.
Adapté de Next Generation Sequencing, BioRender

1.5.2.2 Intérêts et limites du séquençage Illumina

Le séquençage Illumina est connu pour sa rapidité (de quelques heures à 48h), sa précision (pouvant atteindre un taux d'erreur de 0.1%) et sa capacité à produire de

grandes quantités de données (les NovaSeq pouvant générer des milliards de *reads*). Les avancées technologiques faites sur les séquenceurs et leur capacité de séquençage grandissante ont permis, en un peu plus de 10 ans, de diviser le prix de séquençage d'un génome humain par 100 000 (voir Figure 1.18). L'utilisation de ce type de séquençage s'est donc largement démocratisée notamment dans le domaine de la génomique.

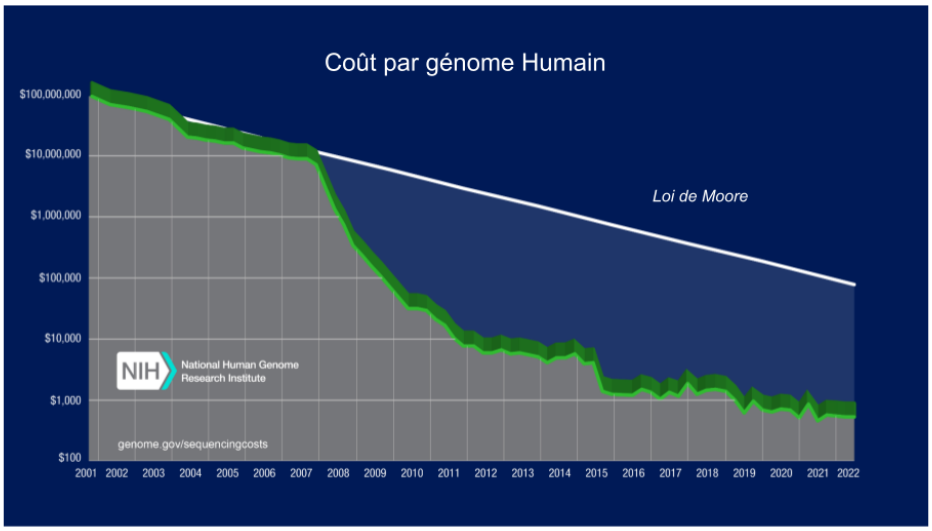


Figure 1.18 – Coût du séquençage d'un génome humain
Durant presque 7 ans, le prix du séquençage d'un génome humain suivait la loi de Moore avant de totalement s'effondrer, invalidant les prédictions de la loi.
Selon le National Human Genome Research Institue

Il reste néanmoins des limitations inhérentes aux séquençages *short-reads*. La principale limitation résulte de la taille des *reads*. Généralement longs de 150 bases, l'alignement des *reads* peut s'avérer difficile notamment dans des régions pseudo-gènes ou des régions contenant des séquences répétées de faible complexité. Pour le séquençage de l'ARN, la taille des fragments ne permet pas l'exploration complète de la structure des isoformes [105] avec des captures de petits fragments du transcrit, nécessitant des assemblages ultérieurs afin de reconstruire le transcrit complet.

1.5.3 Séquençage haut-débit : *long read*

L'impulsion de Pacific Biosciences (PacBio) en 2011 a permis le développement d'un nouveau type de séquençage : le séquençage *long-read*. Appelé séquençage de troisième génération, l'utilisation de *reads* de grandes tailles de 1 à 50 kilobase (kb) a pour but d'outrepasser les limitations du *short-read* et de résoudre les problèmes de caractérisation fine des SV ou de détection des variants présents dans des régions répétées de l'ADN. L'emploi de *reads* de grandes tailles a notamment permis de résoudre une partie de la complexité des régions répétées du génome de référence proposé par le consortium T2T [73, 74].

Les *long-reads* ont aussi permis la détection et le recensement de nouveaux SV, un *read* pouvant couvrir l'intégralité d'un SV [106]. L'avènement des technologies *long-reads* a permis l'identification de nouveaux transcrits, qui, à l'instar des SV, peuvent être résolus intégralement par un *read*. Cette nouvelle résolution permet, entre autres, de découvrir de nouveaux transcrits issus de l'épissage alternatif, d'estimer des expressions avec une abondance de transcrits plus importante, et donc ainsi découvrir de nouvelles fonctions [107].

Le séquençage de troisième génération est disponible chez deux fournisseurs : Oxford Nanopore Technologies (ONT) et PacBio.

1.5.3.1 Séquençage Pacbio

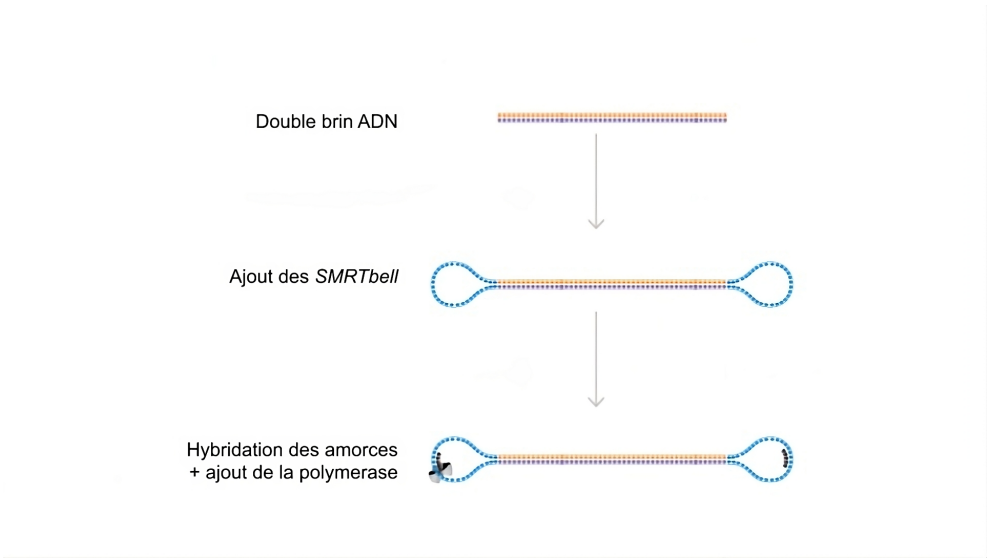


Figure 1.19 – Présentation de la librairie SMRT*bell*
Le principe de la librairie repose sur la transformation du fragment linéaire en fragment circulaire *via* l'hybridation de deux adaptateurs en tête d'épingle.
Adapté selon Wenger *et al.*, 2019

PacBio est le premier fournisseur de séquenceur à commercialiser une méthode se basant sur l'utilisation de *long-read*. Le séquençage PacBio est également connu sous le nom de séquençage d'une molécule en temps réel (ou *Single Molecule Real-Time* (SMRT) *sequencing*). Son intérêt repose dans la détection en temps réel de l'incorporation des nucléotides via leur fluorescence.

Comme toutes les méthodes de séquençage, la première étape consiste à préparer le matériel génétique à séquencer. PacBio fournit une librairie appelée SMRT *bell* qui hybride des adaptateurs en forme de tête d'épingle sur chaque extrémité du fragment d'ADN. Une polymerase est ensuite liée à un adaptateur (voir Figure 1.19). Cette conformation en ADN circulaire consolide le fragment à séquencer et jouera un rôle clé pour les étapes à venir.

Chaque fragment hybridé avec la SMRT librairie est déposé dans la SMRT *cell* contenant des millions de puits appelés *Zero-Mode Waveguide* (ZMW). Chaque ZMW ne contient qu'un seul fragment. Dans ces puits, la polymerase présente sur les fragments va inclure des nucléotides libres et marqués par fluorescence. Chaque incor-

poration va alors émettre un signal lumineux qui est mesuré en temps réel. Les différences de longueur d'ondes associées à chaque fluorochrome de nucléotides permettent d'obtenir la séquence du fragment (voir Figure 1.20).

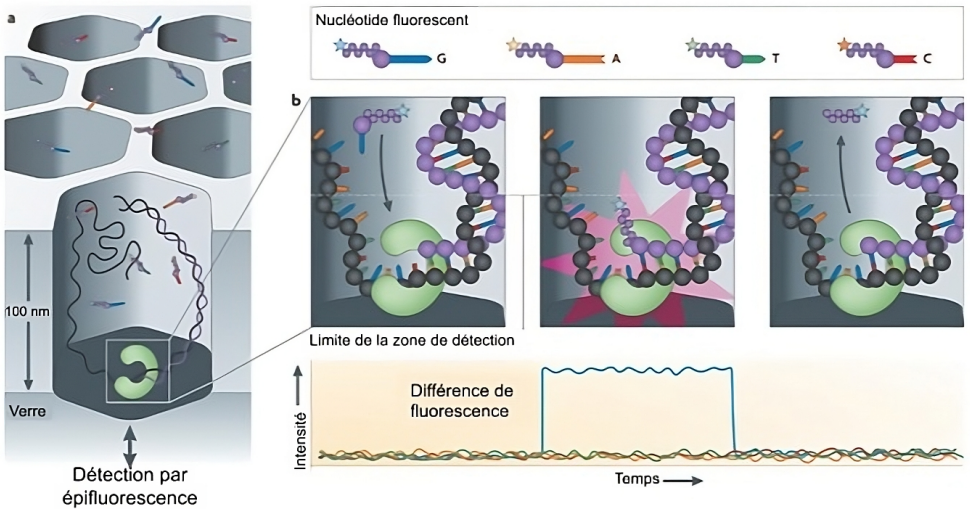


Figure 1.20 – Séquençage PacBio
Adapté selon RNA-Seq blog

Deux modes de séquençage sont disponibles avec les séquenceurs PacBio (voir Figure 1.21). Le mode *Circular Consensus Sequencing* (CCS) repose sur un séquençage en plusieurs cycles sur le fragment circulaire. Chaque cycle génère des *subreads* qui sont alors utilisés pour générer une séquence consensus aussi appelée *High Fidelity* (HiFi) *read*. Ce mode propose un avantage majeur : un taux d'erreur possible à 0.1%, similaire au séquençage *short-reads*[108]. Le deuxième mode nommé *Continuous Long Read* (CLR) réalise une seule passe sur chaque molécule. Sans consensus le taux d'erreur est plus élevé (10%) qu'avec le mode CCS. Néanmoins la taille d'insert du fragment initial est généralement plus grande (de 25 kb à 175 kb, face à 10 à 20 kb pour le mode CCS), favorisant un séquençage de *read* de plus grande taille (selon PacBio, la moitié des *reads* totaux pourrait atteindre plus de 50 kb en fonction des kits). Ainsi, le mode CCS est plus propice pour confirmer la composition d'une séquence ou des expériences avec un taux d'erreur faible, sur le séquençage des nucléotides. Le mode CLR peut par exemple être le mode de prédilection pour de l'assemblage de génome, favorisé par la grande taille des *reads* [109].

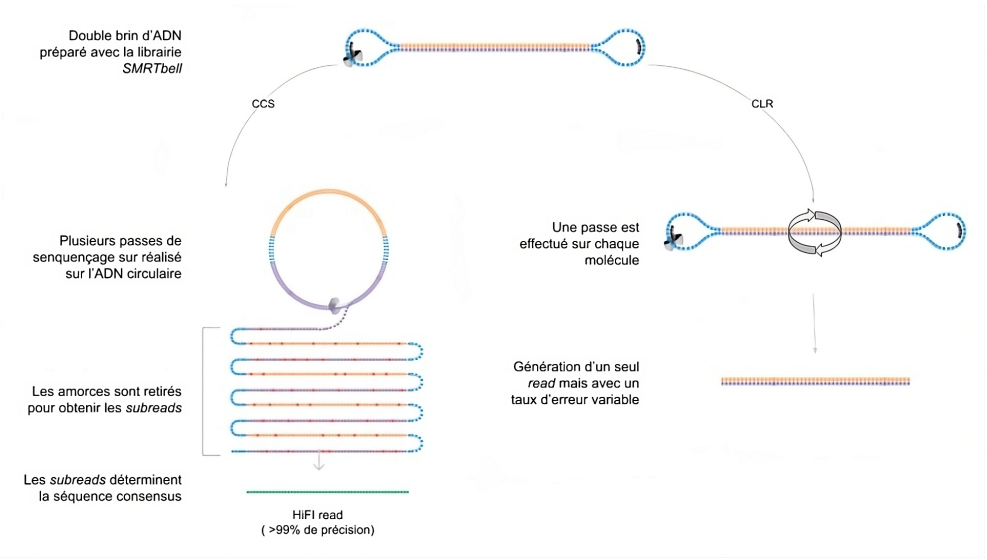


Figure 1.21 – Présentation des différents modes du séquençage PacBio
Deux modes principaux : *Circular Consensus Sequencing* (CCS) et *Continuous Long Read* (CLR).
Adapté selon Wenger *et al.*, 2019

1.5.3.2 Séquençage Oxford Nanopore Technology

Oxford Nanopore Technologies (ONT) a commercialisé son premier séquenceur, le MinION, en 2015. Une des particularités de ce séquenceur réside dans sa taille (tenant dans une paume de main), sa portabilité puisqu'il peut se connecter à n'importe quel ordinateur, ainsi que la vitesse de séquençage. Par exemple, le boîtier du MinION mesure 105 mm L x 23 mm l x 33 mm H et contient une *flow cell* sur laquelle le séquençage *long-read* s'effectue. La méthode de séquençage est commune aux différents séquenceurs ONT. Différents adaptateurs sont fixés aux séquences extraites du matériel biologique.

Au cours des années, ONT a développé des kits différents améliorant la précision de séquençage [110]. Le point commun à ces différents kits est la présence d'un adaptateur sur lequel est présente une protéine motrice qui guide et déroule les molécules d'ADN au travers d'un pore (appelé Nanopore). Le Nanopore est fixé sur une membrane electro-résistante. Un courant électrique constant est appliqué le long du pore. Au passage de la molécule du côté négatif (*cis*) vers le côté positif du pore (*trans*), l'ouverture de ce dernier va provoquer un différentiel de potentiel électrique

au travers du courant ionique, qui varie en fonction de la base nucléotidique. La variation de ce potentiel permet de déterminer l'ordre des nucléotides composant la molécule d'ADN (voir Figure 1.22) [109, 110].

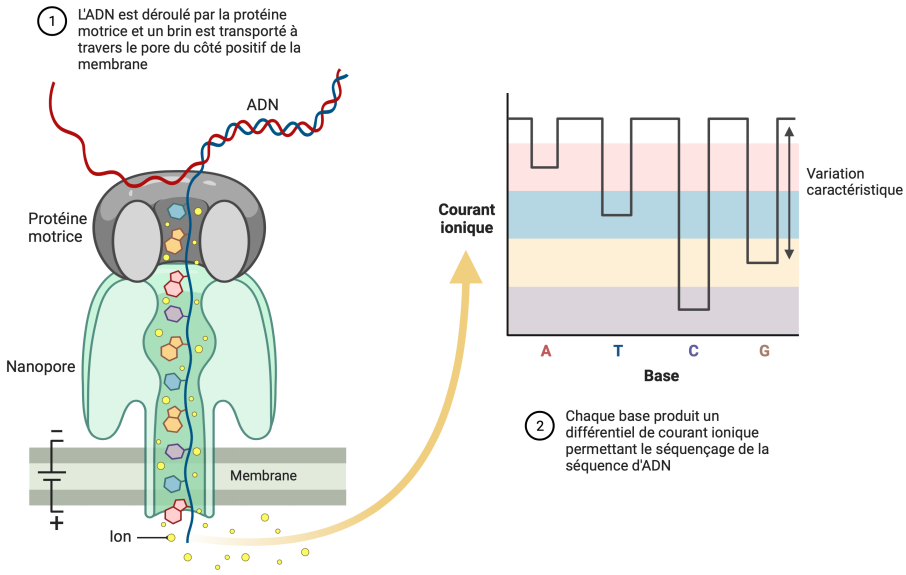


Figure 1.22 – Séquençage Nanopore
(1) Représentation du passage de la molécule d'ADN dans le Nanopore. (2) Conversion du différentiel de potentiel électrique en nucléotides (ou base) : étape de *base calling*.
Adapté de Nanopore Sequencing, BioRender

Dans le cas du MinION, 512 canaux composent la *flow cell*, contenant chacun 4 nanopores. Même si une partie des *flow cells* n'est pas utilisée, 2048 molécules d'ADN peuvent être hypothétiquement séquencées en parallèle. A la différence du séquençage *short-read* d'Illumina, la molécule d'ADN n'est pas fragmentée. L'intérêt du séquençage par Nanopore réside dans le séquençage en temps réel de la molécule complète d'ADN. Concernant l'étude de l'ARN, il peut séquencer de l'ADNc ou directement de l'ARN. Cependant, dans ce dernier cas, il nécessite de grandes quantités d'ARN pour avoir un nombre de *reads* suffisant. C'est pourquoi le séquençage direct de l'ARN en *long-read* est plus adapté à l'étude de modifications de l'ARN, telle que

la methylation des adénines [111].

Le séquençage complet de la molécule d'ADN ou d'ARN n'est pas garantie, la moyenne de taille des fragments reste variable [112, 113], pouvant limiter la description d'un SV ou d'une isoforme non supportée dans sa taille totale par un *read*. Ainsi, même si le séquençage Nanopore a permis une meilleure résolution des SV, de par la taille des fragments générés, il reste moins précis que le séquençage Illumina sur les mutations ponctuelles, avec un nombre élevé d'erreurs, principalement des insertions et des délétions [114]. L'analyse des données doit donc se faire en connaissance de cause afin d'éviter les erreurs d'interprétation.

L'émergence de nouveau procédé comme l'*adaptative sampling* a permis de contrôler l'enrichissement des séquences pendant le séquençage (via l'interface *ReadUntil*). Il permet d'exclure du pore des molécules d'ADN n'appartenant pas à une région d'intérêt. Cette décision d'inclure ou d'exclure un fragment peut se faire après la lecture et l'alignement en temps réel des premiers cent nucléotides d'une molécule d'ADN.

Cette technique semble moins adaptée à l'étude de l'ARN dont les fragments (transcrits pleines longueurs) sont plus courts que l'ADN génomique et dont certains isoformes sont sous-représentées [115]. Néanmoins, par exemple, l'exploration des longs ARN non codants a pu être réalisée par séquençage *long-read* après une capture ciblée [116, 117, 118]. Le développement de l'analyse d'un transcriptome ciblé par séquençage *long-reads* pourrait permettre la détection d'isoformes complexes ou à bas bruit.

La mise en œuvre d'une telle méthode de séquençage ciblé par une approche *long-read* de l'ARNm, couplée à un pipeline bioinformatique permettant de quantifier et annoter ces transcrits, est un objectif de cette thèse.

1.5.4 Différents protocoles de séquençage

Il existe différentes manières d'utiliser le *Next-Generation Sequencing* (NGS), soit par le ciblage d'un panel de gènes ou de l'ensemble des régions codantes d'un génome, aussi appelé *Whole Exome Sequencing* (WES), soit par le séquençage du génome complet aussi appelé *Whole Genome Sequencing* (WGS). Enfin d'autres protocoles de séquençages permettent d'obtenir les séquences du transcriptome à partir de molécules d'ARN. L'obtention du transcriptome peut être complet (on parle aussi de *Whole Transcriptome*) ou partiel, dans ce cas on parlera de transcriptomes ciblés. Chaque protocole suit une étape de préparation de librairie ADN et ARN ayant pour buts principaux d'apporter les différentes amorces et adaptateurs nécessaires à la réaction de séquençage. Ces préparations sont aussi composées d'étapes d'enri-

chissement comme les approches par capture de régions d'intérêts ou par *Polymerase Chain Reaction* (PCR) (approche amplicon).

Lors de la capture de régions d'intérêt, les sondes utilisées sont liées à une molécule de biotine, permettant la sélection des séquences d'intérêt par un système magnétique. Les sondes peuvent être chevauchantes, c'est à dire qu'une même région d'intérêt peut être capturée par plusieurs sondes, dans le but d'homogénéiser la profondeur de couverture des régions d'intérêt.

Au cours de l'enrichissement de type amplicon, des étapes de PCR multiplex vont être effectuées sur chaque échantillon sur chaque région d'intérêt délimitée par des amorces. Les approches par amplicon nécessitent moins de matériel génétique de départ que les approches par capture. Néanmoins, différents rendements des cycles de PCR rendent la profondeur de couverture des régions hétérogènes.

Le choix d'un protocole de séquençage s'effectue selon plusieurs critères :
— L'objectif de l'analyse
— Les moyens et infrastructures pour l'analyse
— Les coûts budgétaires de chaque technique
— Le temps corrélé à la cadence d'analyse du laboratoire

1.5.4.1 Le séquençage d'un panel de gènes

Le séquençage d'un panel de gènes permet le séquençage d'un nombre de gènes préalablement sélectionnés, dans le cadre d'une indication diagnostique précise. Un test génétique, visant à confirmer ou non une prédisposition à un cancer, peut se faire à l'aide d'un petit panel (10 à 30 gènes) [119]. A l'inverse, dans le cadre de la recherche de mutations actionnables dans une tumeur, plusieurs centaines de gènes sont nécessaires. L'intérêt principal de l'approche est d'être la méthode la moins coûteuse par rapport au WES ou WGS, et de faciliter les interprétations en limitant les découvertes incidentelles. Les approches en panel permettent d'atteindre une profondeur de couverture (nombre de *reads* uniques supportant un nucléotide) pouvant aller à plus de 300X. Une profondeur de couverture élevée permet d'interroger des variants avec des fractions alléliques faibles, reflet de l'hétérogénéité tumorale ou de son impureté, et dans les cas d'une analyse constitutionnelle, de détecter des mosaïques. Augmenter la profondeur de couverture permet d'augmenter la précision de l'analyse. Les panels les plus larges peuvent être la base de la détermination de signatures mutationnelles, comme la charge tumorale (*Tumor Mutation Burden* (TMB)), nécessitant d'avoir un profil mutationnel large [120].

1.5.4.2 Le séquençage complet de l'exome

Le *Whole Exome Sequencing* (WES) permet l'analyse de l'ensemble de la séquence d'ADN codante de tous les gènes d'un individu, appelé l'exome. A la différence du panel, l'exome permet d'avoir une meilleure capacité de fouille génomique, sans être limité par le nombre de gènes. Néanmoins, l'exome est limité aux séquences codantes et ne permet pas l'exploration de variants introniques. Il reste plus coûteux qu'un panel de gènes en raison de la quantité de gènes séquencés. La profondeur de couverture des régions séquencées peut variant entre 80 et 150X. Le séquençage d'exome est moins adapté à la détection de variants présents avec une faible fraction allélique en raison de cette profondeur de couverture plus faible. L'interprétation des séquençages d'exome est plus complexe que celle d'un panel de gènes en raison du grand nombre de variants identifiés dans de nombreux gènes.

1.5.4.3 Le séquençage du génome complet

Le *Whole Genome Sequencing* (WGS) permet de couvrir les parties codantes et non-codantes du génome. Cette méthode de séquençage permet la caractérisation plus fine de variants structuraux, étant donné que les points de cassure sont souvent intragéniques ou intergéniques. Le WGS permet la détection de mutations introniques profondes, non détectées par les autres méthodes. Cela peut s'avérer utile notamment pour observer des variants introniques, loin des séquences codantes, souvent qualifiés de cryptiques et pouvant impacter l'épissage du messager du gène, agissant sur les *splicing enhancers*, *splicing silencer*, sites donneurs, sites accepteurs ou même l'identification d'insertion d'éléments mobiles [121, 122]. Néanmoins, la profondeur des analyses WGS se situent entre 30X et 40X pour la majorité, avec pour conséquence une diminution de la spécificité et ne permettant pas la détection de variants avec une faible fraction allélique.

1.5.4.4 Le séquençage du transcriptome complet

Le séquençage de l'ARN s'effectue à l'aide d'une étape de *reverse transcriptase* où l'ARN est converti en Acide désoxyribonucléique Complémentaire (ADNc). Le transcriptome complet similaire au WGS permet l'estimation globale du niveau d'expression de l'ensemble des gènes, de leurs exons et des jonctions d'épissage. Cette méthode permet l'analyse différentielle de l'expression des gènes ou la découverte de transcrits de fusions de gènes [123].

1.5.4.5 Le séquenage de transcrits ciblés

Le séquençage de transcrits ciblés est une technique de séquençage permettant d'enrichir le nombre d'isoformes séquencées par gène ciblé. A la différence du transcriptome complet, plus homogène dans le séquençage, cette technique est moins adaptée pour l'élaboration d'un profil d'expression. Néanmoins l'augmentation de la profondeur de couverture du séquençage sur les régions ciblées permettrait de détecter des transcrits alternatifs faiblement exprimés, mais très nombreux dans les cellules. En génétique constitutionnelle, cette approche pourrait permettre de détecter des transcrits anormaux faiblement représentés, voire des signatures associées à un phénotype particulier [124]. En génétique somatique, cette méthode reste un choix d'intérêt diagnostic pour la détection des transcrits de fusion [125] et permet d'optimiser les procédures de laboratoires diagnostiques.

1.6 La bioinformatique en clinique des cancers

1.6.1 La bioinformatique

Un pipeline bioinformatique est défini comme une succession d'étapes, réalisées par plusieurs outils, pour enchaîner le traitement de la donnée. Il existe de nombreux pipelines bioinformatiques, mais la plupart respectent un ordre commun d'étapes essentielles à tout pipeline. On considère que le *base calling* et le démultiplexage sont réalisés en amont, ces étapes étant directement réalisées par certains séquenceurs selon les fournisseurs.

1.6.1.1 Généralité autour du FASTQ

Après les étapes de *base calling* et de démultiplexage, l'ensemble des *reads* de chaque patient est regroupé dans un fichier FASTQ. Un FASTQ est un fichier linéaire (donc non tabulé) découpé en plusieurs blocs de 4 lignes. Un bloc représente un *read* et les 4 lignes qui composent le bloc correspondent à l'identifiant du *read*, sa séquence, une séquence de description (facultative) et enfin un score de qualité des bases appelées lors du *base calling* (voir Table 1.4).

L'identifiant de la séquence est formaté selon le séquenceur qui a permis d'obtenir le *read*. Dans le format d'Illumina, chaque élément de l'identifiant est séparé par " :". Ainsi les éléments composant l'identifiant donnent des informations sur le type de séquenceur utilisé, l'identité de la *flowcell*, le numéro du *run*, le *cluster* où a été séquencé le *read* ainsi que d'éventuels paramètres utilisés lors du séquençage. La

Ligne	Description
1	Identifiant de la séquence (commence par un @)
2	Séquence du *read*
3	Ligne facultative, commence par un "+" et peut être suivi de l'identifiant de la séquence
4	Qualité de la séquence, un symbole ASCII correspond à la qualité d'une base

Table 1.4 – Description des lignes d'un fichier FASTQ

Table 1.5 décrit les composants sur un exemple d'identifiant suivant :

```
@<séquenceur>:<numéro du run>:<ID de la flowcell>:<ligne>:<tile>:<x_pos>:<y_pos>:<UMI>:<read>:<filtré>:<numéro contrôle>:<index>
```

Elément	Type	Description
@	@	Chaque identifiant comment par un @
<séquenceur>	Symboles autorisés : a-z, A-Z, 0-9, _	ID du séquenceur
<numéro du run>	Nombre	Numéro du *run* du séquenceur
<ID de la flowcell>	Symboles autorisés : a-z, A-Z, 0-9	
<ligne>	Nombre	Numéro de la ligne
<tile>	Nombre	Numéro du *tile*
<x_pos>	Nombre	Coordonnée X du cluster
<y_pos>	Nombre	Coordonnée Y du cluster
<UMI>	Symboles autorisés : A/T/C/G/N	Présent si mode UMI. Les UMI des *reads* 1 et 2 sont séparés par un "+"
<read>	Nombre	2 si *reads* pairés, 1 sinon
<filtré>	Y ou N	Y si le *read* est filtré, N sinon
<numéro contrôle>	Nombre	0 si aucun bit contrôle. Sinon c'est un nombre pair (toujours 0 sur un HiSeq X et NextSeq)
<index>	Symboles autorisés : A/T/C/G/N	Index du *read*

Table 1.5 – Format d'un identifiant de *read* Illumina dans un FASTQ

1.6.1.2 Alignement

L'alignement est l'étape qui va permettre de reconstruire le génome de l'individu. Chaque séquence d'un *read* contenue dans un fichier FASTQ est comparée à un génome de référence afin de déterminer quelles sont les coordonnées génomiques du *read* (voir Figure 1.23). L'alignement permet de déterminer la couverture ou la profondeur, une métrique importante de l'alignement. Elle détermine le nombre de fois qu'un nucléotide est présent sur un *read* à une position donnée. La profondeur de couverture est donnée en divisant la somme des couvertures à chaque position par le nombre total de positions séquencées. L'alignement est généralement la première étape d'un pipeline bioinformatique.

En santé humaine, l'outil d'alignement de *reads* de référence, en bioinformatique pour l'alignement des données de séquençage de l'ADN, est *Burrows-Wheeler Aligner* (BWA) [126]. Cet outil a été amélioré pour augmenter ses performances et le temps d'exécution[40]. Ce nouvel outil d'alignement, appelé *Buroows-Wheeler Aligner-Maximal Exact Matches* (BWA-MEM), améliore l'algorithme initial de BWA

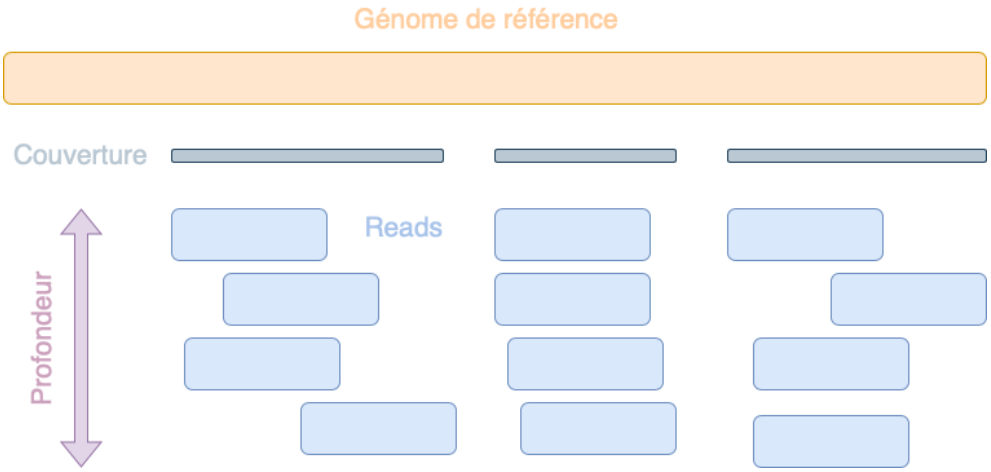


Figure 1.23 – Représentation d'un alignement de *reads*
Lorsque des *reads* sont alignés à un génome de référence, on les nomme *reads* alignés.

reposant sur un principe d'extension de graines (ou *seed*). Une graine correspond à un k-mer c'est-à-dire une séquence de taille fixe de nucléotide. Lorsqu'une graine s'aligne à une partie du génome de référence, la taille de la graine augmente (le maximum étant la taille du *read*) et chaque nouveau nucléotide est évalué d'après un score d'erreur (ou *missmatch*). Le score favorise toujours la plus grande extension de graines possible tout en ayant le moins d'erreur. Ainsi, la graine peut être étendue avec des erreurs si l'algorithme considère que c'est le meilleur compromis taille/erreur.

Des aligneurs spécifiques sont développés pour l'analyse de données ARN. Par exemple, pour le séquençage de l'ARN (RNA-Seq), l'aligneur de référence en *short-reads* est *Spliced Transcripts Alignement to a Reference* (STAR) [127]. Il repose sur un algorithme également basé sur le principe d'extension de graines. En adaptant les scores de chaque graine notamment pour détecter une jonction d'épissage et en réalisant un deuxième alignement du *read* suivant le point d'épissage, STAR propose un algorithme robuste. Ainsi, un *read* peut être décomposé en deux graines qui vont correspondre à une partie d'une séquence jointe au site donneur d'épissage et une partie d'une séquence jointe au site accepteur d'épissage (donc sur deux exons différents). STAR annote le *read* comme contenant une jonction. La détection des jonctions d'épissage lors du processus d'alignement de données ARN permet aux outils intervenant après l'alignement, de déduire les différents motifs d'épissage qui composent un transcrit.

Enfin que ce soit pour la recherche de SV ou pour l'alignement de données issues de séquençage ONT, minimap2 se profile comme l'outil de référence [128]. Son algorithme est également basé sur l'extension de graines.

Chaque outil d'alignement génère un fichier au format *Sequence Alignment Map* (SAM). Un fichier SAM contient toutes les informations déduites de l'alignement par l'aligneur. Cela inclut les positions génomiques de chaque *read*, des informations sur les outils ou références utilisées ainsi que des métriques de qualité. C'est un fichier tabulé contenant 11 colonnes (voir Table 1.6, selon la documentation officielle du fichier SAM).

Colonne	Champs	Type	Description
1	QNAME	Mot	Identifiant du *read*
2	FLAG	Entier	Drapeau en valeur bit décrivant l'alignement
3	RNAME	Mot	Nom de la séquence de référence
4	POS	Entier	Position de la première base du *read* aligné
5	MAPQ	Entier	Qualité de l'alignement
6	CIGAR	Mot	Séquence CIGAR
7	RNEXT	Mot	Identifiant du *read* pairé
8	PNEXT	Entier	Position du *read* pairé
9	TLEN	Entier	Distance absolue entre les *reads* d'une paire
10	SEQ	Mot	Séquence du *read*
11	QUAL	Mot	Score qualité du *read*

Table 1.6 – Format d'un fichier SAM

La format SAM dispose de librairies, disponibles dans plusieurs langages de programmation, basées sur la librairie initiale HTSlib fournie avec la suite de lignes de commandes Samtools [129]. Il permet notamment d'indexer ou même de compresser les fichiers SAM afin de réduire leur taille pour faciliter le stockage et accélérer la lecture. Un fichier SAM compressé se nommera alors fichier *Binary Alignment Map* (BAM).

1.6.1.3 Appel des variants

L'étape de l'appel des variants (ou *variant calling*) suit l'étape d'alignement. L'objectif de cette étape est de chercher les variations dans la séquences des *reads* alignés par rapport à la séquence de référence. C'est une étape charnière qui fournit à terme une liste de mutations observées chez un individu.

Concernant l'appel des SNV et petites délétions ou insertions, il n'y a pas réellement de consensus. Chaque outil faisant l'appel de variants, appelé *caller*, est adapté à certains types de mutations. Ainsi l'utilisation de plusieurs callers est nécessaire pour détecter différents types de variants. Ces derniers reposent, entre autres, sur la qualité des bases, données par le séquenceur, et fournie dans les fichiers d'alignement (BAM) afin d'estimer la présence d'une mutation. La prise en compte de la qualité est notamment utilisée dans les approches bayésiennes, où une qualité élevée tend à conclure sur une forte probabilité de la présence du variant.

Nom	Méthode	Application(s)
DeepVariant	*Machine learning*	Constitutionnel
FreeBayes	Analyse d'haplotype	Constitutionnel et somatique
HaplotypeCaller	Analyse d'haplotype	Constitutionnel
Octopus	Analyse d'haplotype	Constitutionnel et somatique
Strelka2	Fréquence allélique	Constitutionnel et somatique pairé
Mutect2	Analyse d'haplotype	Somatique
VarScan2	Seuil heuristique	Constitutionnel et somatique
VarDict	Seuil heuristique	Somatique
Platypus	Analyse d'haplotype	Constitutionnel et somatique

Table 1.7 – Principales méthodes d'appel de variants. Une application constitutionnelle signifie que le *variant caller* est adapté à des données sans hétérogénéité tumorale. Une application somatique signifie que le *variant caller* peut être utilisé avec l'ADN de la tumeur seule. Somatique pairé signifie que le *variant caller* est uniquement adapté pour l'appel des variants sur de l'ADN tumoral couplé avec l'ADN du tissu sain.

Les différents *variant callers* se distinguent par trois grandes approches : une approche heuristique, une approche bayésienne et une approche par *machine learning* (voir Table 1.7) [130, 131].

Les méthodes heuristiques sont parmi les premières méthodes utilisées pour l'appel des variants. Intégrées dans des outils comme VarScan2 [132] ou VarDict [133], ces méthodes reposent sur l'utilisation de seuils sur des paramètres d'un variant (fréquence du variant, nombre de reads supportant le variant). Ces seuils peuvent être couplés à des méthodes statistiques comme la *p-value* afin de déterminer des distributions de variants hétérogènes dans une région, suggérant des variants somatiques.

Les méthodes basées sur les probabilités bayésiennes concernent de nombreux outils comme Platypus [134], Mutect2 [135], Strelka2 [136], Octopus [137], HaplotypeCaller [138] ou Freebayes [139]. Parmi les approches bayésiennes, il est possible

d'estimer la probabilité qu'un variant d'une fréquence donnée sachant un génotype, soit présent. Des approches basées sur l'étude de la fréquence allélique entre population sont employées en général pour l'appel des variants de tumeurs pairées avec du tissu sain. Le *caller* pourra comparer les alignements de l'ADN tumoral mais aussi de l'ADN de tissu sain. Si une mutation est à la fois retrouvée au sein de la tumeur et du tissu sain, il y a une forte probabilité pour que la mutation ne soit pas d'origine somatique. Cette méthode, bien que beaucoup plus précise dans la détection de variant somatique, nécessite de doubler le séquençage car une tumeur doit être pairée avec du tissu sain. Le budget nécessaire pour l'analyse peut être un facteur limitant pour l'utilisation de ce mode. Une autre approche bayésienne est basée sur l'analyse de l'haplotype, un groupe d'allèles de différents gènes situés sur un même chromosome et habituellement transmis ensemble. Ces approches réalisent un assemblage local des *reads* afin de générer différentes possibilités d'haplotypes. Les *reads* sont alors alignés de nouveau sur chaque haplotype, la probabilité qu'un haplotype soit vrai est estimée par le nombre de *reads* supportant l'haplotype ainsi que leur qualité.

De nouvelles approches basées sur du *machine learning* sont intégrées dans des outils comme DeepVariant [140]. Le *machine learning* repose sur l'utilisation d'un modèle mathématique calibré à partir de données de références. Ces dernières contiennent des variables prédictives ainsi qu'une variable cible qui détermine la validité de la prédiction. Le modèle va donc mettre en corrélation les différentes variables prédictives afin d'aboutir à la même prédiction, donnée par la variable cible. Par exemple, DeepVariant entraîne un réseau neuronal convolutif à l'aide de fichiers BAM obtenus après l'alignement. DeepVariant considère chaque alignement comme une image qu'il va transformer en un groupe de vecteurs afin de pouvoir y inférer des logiques mathématiques. Lors de l'entraînement chaque image correspondant à des alignements de *reads* est associée à un génotype (variable cible) afin de pouvoir entraîner le modèle à prédire les génotypes de chaque région génomique. Les méthodes basées sur le *machine learning* sont dépendantes du modèle. Cela signifie que des différences trop prononcées entre jeux de données (régions trop hétérogènes, peu de données pour l'entraînement, trop de données pour l'entraînement) peuvent mener à des prédictions fausses. Cette problématique du *machine learning* nécessite au préalable de connaître les données utilisées et d'évaluer les biais possibles pouvant influencer les prédictions.

Enfin, l'ensemble des résultats de l'étape de l'appel des variants est donné dans un fichier appelé *Variant Calling Format* (VCF)[141]. C'est un fichier tabulé en 8 colonnes contenant différentes informations sur le variant (voir Table 1.8).

Colonne	Nom	Description
1	#CHROM	Chromosome contenant la mutation
2	POS	Position de la mutation sur la référence
3	ID	Identifiant de la mutation
4	REF	Nucléotide(s) de la référence
5	ALT	Nucléotide(s) du variant
6	QUAL	Score de qualité
7	FILTER	Détermine si le variant a passé les filtres du caller
8	INFO	Champs contenant des annotations sur le variant

Table 1.8 – Format d'un fichier VCF

1.6.1.4 Les prédicteurs de l'impact fonctionnel d'un variant

Après l'étape de l'appel des variants, une dernière étape consiste à fournir de l'information sur les variants détectés, c'est l'étape d'annotation. En dehors des variants entraînant une perte de fonction du gène, la majorité des variants nécessitent des outils *in-silico* capables de déduire l'impact potentiel du variant sur le gène. Afin d'augmenter la compréhension, la caractérisation et l'interprétation de ces conséquences, des prédicteurs ont été développés, par exemple :

— *MutationTaster* [142] utilise la base de données dbSNP afin de déterminer l'aspect polymorphique du variant. Si le variant n'est pas considéré comme polymorphique, l'outil évalue l'impact du variant sur la protéine au travers de différents critères comme une modification de la taille de la protéine, impactant la séquence Kozak (essentielle à l'initiation de la traduction) ou l'évaluation de la fonctionnalité de la protéine suite au changement d'acide aminé.

— *Sorting Intolerant From Tolerant* (SIFT) définit le caractère pathogène d'un variant [143] en fonction de l'homologie des séquences peptides au sein d'une famille de protéines partagées entre plusieurs espèces. Si une famille de protéines ne contient qu'un seul acide aminé ou des acides aminés d'une même composition chimique (hydrophobe, polarité...) et que la mutation modifie la conservation (interespèce) de la séquence peptidique, SIFT conclura à un effet pathogène.

— *Evolutionary model of Variant Effect* (EVE) [144] est un prédicteur basé sur du *deep learning*. Le modèle de l'outil est entraîné sur des séquences peptidiques de protéines présentes dans plus de 140 000 espèces. Le modèle prend en compte l'évolution des gènes et leur impact sur la vie de l'individu en cas de mutation. Ainsi EVE conclut sur l'impact d'un variant en fonction de la mo-

dification de la protéine qu'il va engendrer et de la viabilité de la protéine par homologie, déduite d'un modèle interespèce. Les gènes inclus dans le modèle sont des gènes associés à des maladies.

— SpliceAI [145] est un prédicteur spécialisé dans l'évaluation de l'impact d'un variant sur l'épissage. L'outil est basé sur un modèle d'apprentissage (*machine learning*) entraîné sur des sites d'épissage présents dans les séquences d'ARN pré-messager de la base GENCODE [146]. Le modèle est entraîné sur 10 000 nucléotides de la séquence flanquante pour prédire la fonction d'épissage de chaque position dans le transcrit de l'ARN pré-messager. Ainsi le modèle permet de prendre en compte la séquence d'ARN environnante pour évaluer l'impact probable sur l'épissage.

— *Splicing Prediction Pipeline* (SPiP) [147] est également un prédicteur spécialisé dans l'évaluation de l'impact d'un variant sur l'épissage. L'outil est basé sur un modèle d'apprentissage de type *random forest*. Il a été élaboré à l'aide d'un jeu de 4616 variants répartis sur 227 gènes. Il permet de prédire la probabilité d'impact sur l'épissage d'un variant au niveau de l'ensemble des motifs d'épissage (5'/3', point de branchement, éléments régulateurs ESR). Il est complémentaire de l'outil SpliceAI. Ce dernier est plus performant pour prédire si un variant permet la création d'un site 5'/3' intronique profond, tandis que SPiP est plus performant pour les autres prédictions.

Des prédicteurs comme *Variant Effect Predictor* (VEP) [148], *Rare Exome Variant Ensemble Learner* (REVEL) [149] ou *Combined Annotation Dependent Depletion* (CADD) [150, 151] intègrent différents prédicteurs afin de produire des métascores, prenant en compte les scores individuels de chaque prédicteur, en fonction de différentes interactions qu'un variant peut avoir sur l'ARN et sur la protéine.

VEP, par exemple, peut prédire 41 conséquences possibles pour un variant. Néanmoins parmi ces 41 conséquences, seulement 10 sont considérées comme ayant un impact élevé [152]. Parmi elles, on retrouve :

— L'ablation ou l'élongation du transcrit, pouvant avoir comme conséquence la traduction d'une protéine tronquée ou d'une protéine anormalement exprimée.

— La modification des codons *start* et *stop*. Dans le premier cas, un codon *start* anormal et non fonctionnel peut aboutir en l'absence de la traduction de la protéine ou même, si la mutation provoque un décalage du cadre de lecture, à la production d'une protéine tronquée. Bien que VEP ne prenne pas en compte les variants dans les *upstream Open Reading Frame* (uORF), une modification de la séquence d'un uORF, régulatrice de la traduction, peut induire des différentiels d'expression de la protéine traduite. Enfin une mutation dans le

codon *stop* peut altérer la taille du transcrit, aboutissant généralement à une élongation du transcrit.
— Des conséquences impactant les sites d'épissage, dont les sites canoniques donneurs ou accepteurs, modifiant l'épissage de l'ARN, peuvent aboutir à des isoformes alternatives aberrantes. Des outils tels que l'extension de CADD, CADD-Splice [153] ou SpliceAI permettent d'affiner l'interprétation de l'impact de variants d'épissage. Ces prédicteurs évaluent l'impact d'un variant sur différents motifs d'épissage. Ces derniers proposent alors un métascore afin d'uniformiser les différents scores intégrés, tous basés sur du *machine learning*. L'évaluation de l'impact de variants sur l'épissage peut permettre de s'orienter vers l'hypothèse d'un épissage alternatif aberrant.
— Des conséquences sur le décalage du cadre de lecture (ou *frameshift*), pouvant aboutir à l'apparition d'un codon stop prématuré ou de la modification de la séquence peptidique.

D'autres prédicteurs d'impact se développent aussi sur d'autres types de mutations comme pour les CNV avec ClassifyCNV [154]. Ce dernier se base sur des caractéristiques fonctionnelles du CNV, comme le nombre de promoteurs chevauchant le CNV, le nombre de gènes codant chevauchant le CNV, le chevauchement avec des gènes haploinsuffisants (se dit d'un gène dont un allèle sain ne suffit pas à produire une quantité suffisante de protéine saine en raison de la présence d'un allèle pathologique) et la fréquence du CNV dans la population [155].

1.6.1.5 Les bases de données

L'annotation d'un variant peut aussi s'effectuer à partir de la connaissance que la communauté scientifique a rassemblée dans diverses bases de données, afin d'annoter les variants détectés. La qualité de l'annotation dépend de la qualité des sources utilisées. La diversité des sources d'annotation permet d'enrichir la connaissance autour d'un variant, sur différents aspects.

Des bases de données comme *Genome Aggregation Database* (gnomAD) [156] recensent la fréquence de mutations observées dans plus de 120 000 séquençages d'exomes et plus de 15 000 séquençages de génomes d'ethnies différentes. Son utilisation en génétique humaine est cruciale car cette base permet, entre autres, d'identifier les variants fréquents dans les populations et donc ayant de fortes probabilités d'être constitutionnels et bénins. D'autres bases généralistes agrègent de la connaissance sur des variants mononucléotidiques. On peut citer par exemple, dbSNP [157] qui permet d'acquérir de la connaissance sur les SNP et leur présence dans la population

générale, avec cependant un biais, car d'authentiques variants pathogènes figurent dans cette base. La base ClinVar [158] met en relation des phénotypes cliniques et des variants soumis par la communauté scientifique, permettant de donner un argument en faveur de leur pathogénicité.

Des bases spécialisées sur des gènes très étudiés en oncologie, peuvent permettre d'affiner l'interprétation des variants sur ces gènes. Par exemple *BRCA Exchange* [159] est spécialisée pour les gènes *BRCA1* et *BRCA2*. La *TP53 Database* [160, 161] est spécialisée dans le gène *TP53* pour le syndrome de Li et Fraumenie ou encore la base *French OncoGenetics* (FrOG) du GGC spécialisée dans le recueil des variants impliqués en oncogénétiques, en particulier dans les cancers héréditaires du syndrome HBOC et du syndrome de Lynch. Ces bases de données recensent la connaissance réunie par de nombreux experts sur des gènes, dont des informations sur la pathogénicité du variant. Elles permettent d'améliorer l'annotation de certains variants sur ces gènes, absents des bases de données généralistes.

Des outils comme AnnotSV [162] permettent d'agréger la connaissance issue de plusieurs bases de données comme :

— *Reference Sequence* (RefSeq) [163], base contenant de nombreuses séquences nucléotidiques provenant d'un matériel génétique ADN ou ARN. Chaque séquence est annotée et curée et les séquences ADN sont liées à des transcrits ARN dits de référence.
— *Online Mendelian Inheritance in Man* (OMIM) [164], base faisant le lien entre une anomalie génétique et un phénotype associé. Elle permet d'évaluer les risques qu'une mutation soit connue au sein d'une pathologie.
— *Database of Genomic Variants* (DGV) [165], base contenant des SV soumis par la communauté scientifique dans le cadre de la recherche clinique ou du diagnostic.

AnnotSV est un exemple des possibilités de l'étape d'annotation. En plus des bases de données, il prend en compte d'autres paramètres comme par exemple l'haploinsuffisance d'un gène ou la présence de mutations au sein du SV ou la présence de séquences répétées autour des points de cassure du SV. De plus, dbVar [166], agrège la connaissance (preuve clinique, région génomique, description de l'évènement attendu) autour des SV connus [166].

Plusieurs bases sont spécialisées dans la détermination de l'implication d'un variant dans un cancer, comme par exemple : *Clinical Interpretation of Variants in Cancer* (CIViC) [31] ou *Catalogue of Somatic Mutations in Cancer* (COSMIC) [167]. Ces bases permettent de trouver des sources d'annotation des variants tels que (i)

des citations sur les études fonctionnelles sur le variant (ii) les cancers où sont retrouvés majoritairement ces variants (iii) des informations sur les voies de signalisations associées aux variants (iv) des informations sur des thérapies associées.

COSMIC fournit une base appelée COSMIC Fusion cataloguant des fusions observées en cancerologie. La connaissance grandissante autour des thérapies ciblées a permis le développement de bases de données en lien avec les thérapies ciblées comme : *MolecularMatch* (MM) [168], *Oncology Knowledge Base* (OncoKB)[30], *Clinical Knowledgebase* (CKB)[33] ou des bases comme COSMIC et CIViC. Ces deux dernières ont ajouté des notions d'actionnabilité à leurs bases de données. L'ensemble de ces bases fournit des associations entre des thérapies et des cibles moléculaires dites actionnables. Ces cibles peuvent être des variants de types différents (SNV, SV, CNV, fusion de gènes) et leurs associations sont supportées par un score d'évidence clinique. Ce score d'évidence décrit si la thérapie soumise au variant est prédictive d'une sensibilité ou d'une résistance à une thérapie en fonction d'un type de cancer. Chaque évidence est généralement associée à une classification (basée sur les recommandations de l'ASCO/AMP) définissant sa pertinence en fonction du variant associé et du type de cancer du patient.

1.6.2 Les challenges de l'analyse de l'ADN

En partie grâce aux recommandations de GATK [138] et à des positions de consensus dans la communauté scientifique, les pipelines bioinformatiques d'analyse de données génomes du cancer présentent de bonnes performances [169]. Ils permettent d'identifier les SNV, INDEL, CNV et transcrits de fusions.

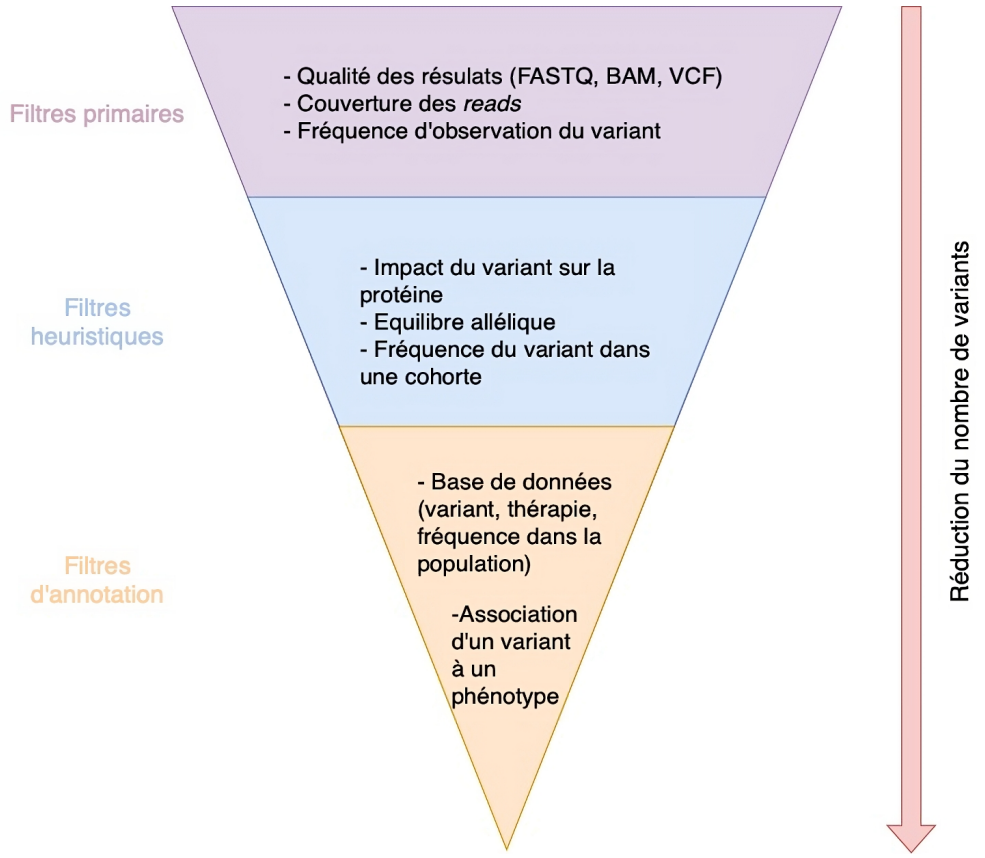


Figure 1.24 – Etape principale du filtrage de variants
Les règles de filtrations sont établies selon trois grandes catégories (i) des filtres primaires basés sur la qualité des variants (ii) des filtres heuristiques basés notamment sur la fonctionnalité du variant (iii) des filtres sur l'annotation provenant de bases de données ou de la littérature et d'informations phénotypiques.

Il est alors nécessaire d'établir, après l'annotation, des règles de filtrage afin de ne conserver que les variants hypothétiquement impactant dans le développement de la tumeur. Ces filtres visent, dans un premier temps, essentiellement à réduire les faux positifs, que ce soit des erreurs de séquençage induites par les étapes d'amplification par PCR, la conservation des tumeurs par FFPE et les différentes erreurs des outils du pipeline. Ainsi, les règles de filtration se basent généralement sur trois critères principaux (i) la qualité des variants, (ii) leur fonctionnalité et les conséquences sur le transcrit ou la protéine (iii) leurs annotations (voir Figure 1.24).

Enfin, le nombre de thérapies, mais surtout d'essais cliniques ne fait qu'augmenter

d'année en année. Selon ClinicalTrials.gov plus de 464 000 essais cliniques (dont 185 588 impliquant une cible moléculaire) seront recensés à la fin de l'année 2023, soit presque 30 000 de plus que l'année 2022 (voir Figure 1.25).

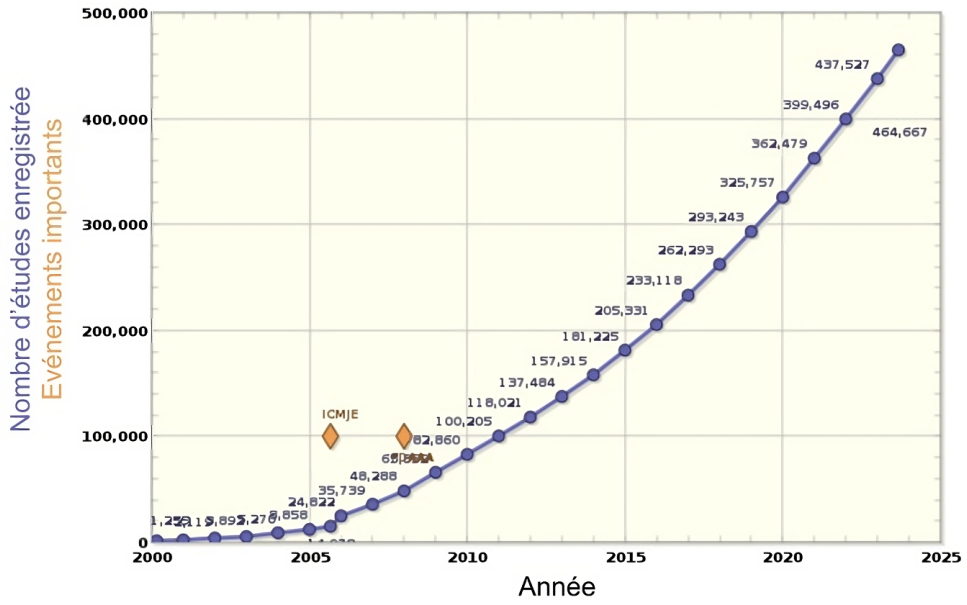


Figure 1.25 – Evolution du nombre d'études cliniques depuis 2000
Les derniers chiffres sont en date du 1er Septembre 2023.
Adapté d'après ClinicalTrials.gov

Comme décrit précédemment, de nombreuses bases de données et d'outils existent pour l'annotation de variants afin d'évaluer leur impact, l'association du gène à une pathologie connue ou la connaissance clinique d'un variant actionnable. Toutefois, la diversité de l'information engendre des problématiques auxquelles il faut faire face et l'interrogation de ces bases de données doit se faire de manière logique.

Les thérapies ciblées ne sont pas forcément associées à un variant précis, mais à un résultat moléculaire décrivant une anomalie plus "générique" pouvant être présentée, par exemple, comme une perte ou un gain de fonction du gène. Il est alors nécessaire de pouvoir comparer un variant obtenu par séquençage par rapport à ces résultats moléculaires présents dans les bases de données. Cette analyse doit être réalisée dans un premier temps en interrogeant le variant exact dans ces bases, puis élargir la recherche en comparant ce variant avec les résultats moléculaires obtenus. Ces résultats peuvent être décrits soit au niveau d'un codon, d'un exon ou d'un gène,

quand ces informations sont disponibles dans ces bases. Cette hétérogénéité des bases de données s'applique également aux bases thérapeutiques [170], nécessitant alors de trouver des solutions afin de combler le manque d'information d'une base en la couplant à l'interrogation simultanée d'autres bases. La redondance de l'information dans l'ensemble des bases de données doit également être prise en compte.

L'utilisation de ces nombreuses bases de données est un défi pour la veille médicale concernant les thérapies ciblées, puisqu'elle nécessite de maintenir les sources de données à jour. La mise à jour des bases de données nécessite des infrastructures capables de garantir la mise en place des nouvelles versions de ces bases, contenant les dernières annotations disponibles. Des bases comme *MolecularMatch* (MM), CIViC ou même CKB mettent à disposition une interface de programme aussi appelée *Application Programming Interface* (API). Les API sont des moyens d'interaction entre logiciels. Elles permettent notamment de pouvoir obtenir en temps réel les informations d'une base, sans les télécharger au préalable, au travers de requêtes. Cela permet de garantir que les informations reçues soient toujours celles les plus à jour dans la base. Néanmoins l'utilisation d'API peut être un challenge pour les serveurs de l'émetteur et du récepteur de la requête, spécialement sur les gros volumes de données. Cela explique, entre autres (en plus du temps de développement de l'API), pourquoi toutes les sources de données n'en disposent pas, ce qui est une contrainte pour leur intégration et leur maintenance.

Par conséquent, il est nécessaire de pouvoir agréger les différentes sources d'annotations afin d'en tirer une information unique. Certaines sources d'informations, de par leur redondance, ne sont pas utiles à coupler ou doivent alors nécessiter un pré-traitement de l'information. Dans l'exemple d'un outil permettant d'associer des thérapies ciblées à des variants, la quantité d'informations issue de ces nombreuses sources d'information doit être ordonnée et une logique de traitement solide doit être développée. Ainsi, la compilation des sources est un prérequis à la mise en œuvre d'un outil d'aide à l'interprétation, afin de hiérarchiser correctement les associations entre les variants et les thérapies. La mise en œuvre d'un tel outil est un objectif de cette thèse.

1.6.3 Les challenges de l'analyse de l'ARN

Dans l'analyse de données issues du transcriptome de tumeur, la détection des transcrits de fusions de gènes est bien maîtrisée et réalisée, notamment à l'aide de plusieurs outils comme Arriba [171] ou STARFusion [172] basés sur l'interprétation des alignements des *reads*. D'autres méthodes comme JAFFA [173] proposent une

détection des fusions après un assemblage des transcrits. La détection de ces événements est à présent bien maîtrisée [125].

Dans le domaine des prédispositions au cancer, il est probable que certains événements génomiques à l'origine d'une prédisposition au cancer ne soient pas identifiés à partir de l'ADN génomique. Cependant la répercussion de ces événements pourrait être détectée par des anomalies d'expression des gènes altérés. Ces anomalies d'expression peuvent être quantitatives ou qualitatives. Cela peut se traduire soit par une perte d'expression d'un allèle, soit par la production d'une isoforme anormale, soit par la modification d'un profil d'expression des isoformes des messagers du gène.

En conséquence l'étude du transcriptome global, ou préférentiellement du transcriptome ciblé afin d'obtenir une forte profondeur de séquençage permettant une définition de l'ensemble des isoformes à haute résolution, aurait un intérêt dans le diagnostic moléculaire des prédispositions au cancer. De plus le séquençage en *long-reads* semble être une technologie prometteuse pour la description des transcrits d'un gène [109, 100, 101, 174]. En effet, les techniques de séquençage *short-read* limitent la résolution de ces événements, parfois complexes et de grandes tailles (type insertion ALU), qui composent l'hérédité manquante pouvant expliquer des cas de cancer héréditaire. Il y a donc un réel besoin à comprendre les variations du patron d'épissage de ces gènes, dans l'objectif de détecter des altérations des isoformes de l'ARN reflétant possiblement un processus pathogène.

Une des solutions pourrait résider dans l'utilisation de techniques de séquençage de troisième génération. En effet les *long-reads* peuvent atteindre des tailles couvrant l'ensemble d'un transcrit appelé *full-length*. Ainsi si on considère qu'un *read* correspond à un transcrit, même si des erreurs surviennent dans la séquence du *read* l'interprétation du mécanisme d'épissage ayant abouti au transcrit observé est possible.

En *long-read*, les *reads* de grandes tailles permettent de couvrir totalement ou presque totalement la taille d'un transcrit. Cette particularité fait qu'à la différence de l'assemblage en *short-read* nécessitant des constructions successives de *contigs* et *scafolds* pour aboutir à un transcrit complet, le *long-read* permet d'obtenir directement ce niveau de précision. Néanmoins, en *long-read*, les techniques bioinformatiques d'assemblage des transcrits semblent également utiles pour définir les transcrits pleine longueur.

L'assemblage peut être guidé avec un génome de référence. Si l'assemblage est fait sans guide, il est alors considéré comme *de-novo*, les algorithmes déduiront alors les meilleurs assemblages possibles de reads pour obtenir une séquence finale. Ce choix

du meilleur assemblage possible dépend de l'algorithme utilisé par les assembleurs. Des outils comme *Full-Length Alternative Isoform analysis of RNA* (FLAIR), StringTie ou Iso-Seq essayent de dresser un profil d'isoformes alternatives (aberrantes ou physiologiques) présentes chez un patient.

1.6.3.1 StringTie

StringTie [175, 176] intègre un algorithme qui repose sur l'élaboration d'un consensus de *reads*, à partir d'un assemblage local ou *de-novo*. L'algorithme définit des fenêtres de *reads* à chaque locus du gène. Chaque *read* d'une même fenêtre est alors intégré dans un graphe d'épissage alternatif. Chaque noeud du graphe correspond à une région du génome ininterrompue par une jonction d'épissage. Les *reads* ayant un profil d'épissage identique seront situés sur le même chemin du graphe. Dès lors, les *reads* qui couvrent le chemin le plus abondant sont retirés et associés à l'isoforme construite par le graphe. L'algorithme est itératif et recommence à construire le graphe avec les *reads* présents dans la même fenêtre, moins les *reads* formant la première isoforme. Le graphe évalue, de nouveau, la conformation majoritaire des *reads* restants, en déduit une deuxième isoforme et recommence jusqu'à ce qu'il n'y ait plus de *reads* ou que la couverture en *reads* du chemin majoritaire ne soit pas assez importante (2.5 *reads* par base, par défaut).

1.6.3.2 FLAIR

FLAIR [177] est un pipeline complet proposant l'alignement des *reads* en utilisant minimap2. Il est basé sur la correction de sites d'épissage contenant des erreurs d'alignement, en utilisant des annotations du génome ou les annotations des jonctions d'épissage issues du *short-read*. Enfin un module permettra de combiner les isoformes. L'algorithme de combinaison s'initie par l'assemblage. Les sites de début et de fin de transcription sont déterminés par la densité des coordonnées de début et de fin des *reads*. Par défaut FLAIR crée une fenêtre de 100 nucléotides à la fin des sites et sélectionne le site le plus représenté dans chaque fenêtre (seuil modifiable). Ainsi la première isoforme est assemblée (première passe), FLAIR réalise un deuxième alignement par minimap2 conservant l'isoforme si elle est supportée par au moins trois *reads* avec une qualité d'alignement (MAPQ) supérieure ou égale à 1. Le processus se répète alors afin d'assembler de nouvelles isoformes en fonction de l'abondance des *reads* alignés à cette dernière.

1.6.3.3 Iso-Seq

PacBio propose une méthode propre à son séquençage de *reads* HiFi : Iso-Seq [178]. Les propriétés du séquençage HiFi dont la haute précision (99%) et la taille des *reads* pouvant couvrir la taille totale d'un transcrit, permet à Iso-Seq de ne pas effectuer d'étape d'assemblage, permettant la réalisation de séquences consensus déjà opérée lors du séquençage HiFi. En effet, dans le pipeline Iso-Seq, un *read* correspond à un transcript. Les différentes étapes d'assemblage ou de corrections ne sont pas nécessaires grâce au faible taux d'erreur du séquençage PacBio. Iso-Seq propose alors un *clustering* des *reads* HiFi, générant alors un *read* consensus pour chaque *cluster*.

Même si des méthodes bioinformatiques existent pour détecter des isoformes alternatives de l'ARN, il reste alors le problème de l'interprétation de ces données. En effet, rien que sur la base Ensembl, il existe plus de 300 transcrits des gènes *BRCA1* et *BRCA2*. En plus de l'identification des isoformes, il faut être capable de séparer les isoformes physiologiques des isoformes aberrantes. L'annotation des isoformes est donc cruciale afin de permettre de comprendre tous les mécanismes d'épissage observés dans l'isoforme. Ainsi à l'image des challenges de l'analyse de l'ADN, les défis de l'annotation sont aussi présents pour les analyses ARN afin de (i) pouvoir observer des épissages complexes de l'ARN (ii) mettre en évidence des isoformes expliquant l'hérédité manquante (iii) filtrer le nombre d'isoformes (iv) faciliter l'interprétation par le biologiste.

Le deuxième objectif de cette thèse est le développement d'un pipeline adapté aux données *long-reads* obtenues par un séquençage ciblé de l'ARN. Ce pipeline doit être un outil aidant à la recherche et la résolution de transcrits aberrants pouvant être des biomarqueurs importants pour des cancers du sein et de l'ovaire, expliquant une partie de l'hérédité manquante. En complément de la détection, nous souhaitons proposer un module d'annotation afin de décrire finement les différentes jonctions d'épissage supportées par un isoforme. Ce module permettra de filtrer des isoformes, de les annoter et donc d'aider l'utilisateur dans sa démarche d'interprétation.

2. Objectifs de la thèse

Le séquençage de l'humain est passé de l'état de projet à une réalité entraînant de nouveaux axes de recherche et de développement pour le diagnostic moléculaire des prédispositions aux cancers et de la caractérisation génétique des tumeurs, dans le cadre de l'orientation thérapeutique, du pronostic et de la réponse au traitement. Les approches de NGS reposent aujourd'hui sur des technologies de séquençage massivement parallèle de petits fragments (*short-read*) ou de longs fragments en temps réel (*long-read*) et ont grandement favorisé la compréhension des caractères constitutionnels génétiques prédisposant aux cancers et aux mécanismes génétiques acquis aboutissant au développement tumoral. La génétique médicale a profité de ces nouvelles technologies en proposant de nouvelles méthodes pour le diagnostic des prédispositions aux cancers ou la recherche de thérapies ciblées.

Néanmoins, une grande partie des situations évocatrices d'une prédisposition aux cancers reste aujourd'hui inexpliquée. Jusqu'à présent le diagnostic moléculaire de ces prédispositions repose sur l'exploration par NGS, de panels de quelques dizaines de gènes, mais elle ne résout que 10 à 20% de ces situations [179]. L'exploration du génome a pu mettre en évidence de nombreux facteurs génétiques de forte [180, 181, 182] et de faible pénétrance [182, 183, 184] mais certaines situations de part leur précocité au diagnostic, la multiplicité des cancers individuels ou dans une famille laissent penser que persistent des facteurs génétiques très fortement pénétrants et rares, restant à caractériser. Ces événements sont peu accessibles par des techniques conventionnelles de séquençage en *short-read* puisque qu'ils sont très probablement complexes et parfois intergéniques (dits profonds). Ils peuvent impliquer par exemple des *Topologically Associated Domain* (TAD) [185], des régions introniques ou exoniques, et sont composés de SNV, d'Insertion-délétion (INDEL) dont les insertions de séquences répétées du génome, de CNV ou de variants structuraux complexes. Ainsi, l'ensemble de ces événements semble aujourd'hui accessibles sur l'ADN génomique ou l'ARNm par les technologies de séquençage en combinant les approches en *short-read*

et *long-read*. En effet, il est possible que ces situations de prédisposition puissent être plus facilement détectées au niveau de l'ARN des patients en modifiant les isoformes alternatives d'ARNm (expression ou signature qualitative des isoformes). La combinatoire du séquençage en *short-read* (expression des ARNm et quantification des jonctions) et en *long-read* (accès à la phase de l'ARNm pleine longueur) permettrait d'accéder également à ces événements. Quelle que soit la méthode de séquençage, le séquençage d'ADN ou d'ARN après enrichissement d'un panel de gènes candidats ou la réalisation de panels *in silico* à partir d'exome/génome ou transcriptome semble être une méthode permettant une exploitation rapide, standardisable et compatible avec la prise en charge clinique pouvant offrir deux niveaux de lecture. Un premier à la recherche des événements les plus connus ou communs et un second plus exploratoire pour les événements plus rares ou complexes.

Aussi, l'avènement des thérapies ciblées a largement changé les approches dans le traitement des cancers. L'Imatinib est la première molécule de thérapie ciblée à avoir été utilisée. Cet inhibiteur spécifique de tyrosine kinase est l'étendard de cette nouvelle classe thérapeutique. En ciblant la protéine chimérique BCR-ABL, l'Imatinib a profondément changé le pronostic des patients atteints de leucémie myéloïde chronique [186]. Aujourd'hui plusieurs dizaines de thérapies ciblées ont une AMM pour des tumeurs solides ou hématopoïétiques. Ce faible chiffre est à opposer aux nombreuses molécules développées et au faible pourcentage des nombreuses molécules anti-cancéreuses qui dépassent le cadre des essais cliniques. Il est avancé pour expliquer ces nombreux échecs que dans une majorité des cas, les populations d'études ne sont pas correctement sélectionnées [187, 188]. L'absence de biomarqueurs (variants ou signatures moléculaires caractéristiques d'un cancer, actionnables ou non) connus pour une majorité de molécules mises en développement clinique, ne permet pas de faire bénéficier les patients des traitements les plus pertinents [189]. Une systématisation de l'établissement de profils de variants tumoraux pourrait aider à la mise en place de thérapeutiques personnalisées et permettre de découvrir de nouveaux marqueurs. La recherche et l'identification de l'ensemble de ces variants *drivers* probablement « actionnables », c'est-à-dire que leur présence ou leur absence peut orienter la prise charge du patient, au sein d'une tumeur semble être le prérequis à l'établissement d'un traitement personnalisé. Comme développé précédemment, cette approche est rendue possible par le développement du séquençage massivement parallèle qui permet le séquençage de génomes entiers ou de centaines de gènes par une approche ciblée. Les événements à rechercher sont composés de SNV, INDEL et SV (gènes de fusion, CNV)

Pour établir ces profils de variants au cours du séquençage d'une tumeur, il peut

être utile d'introduire un processus bioinformatique, capable de discriminer les variants actionnables des autres. La mise en œuvre d'un profil moléculaire doit prendre en compte les contraintes des laboratoires diagnostiques en termes de capacité de production, d'analyse et de traitement mais également de coût et de temps de réalisation qui doivent rester compatibles avec la prise en charge d'un nombre croissant de patients pouvant bénéficier de ces approches. Les approches en panels de gènes séquencés possiblement en *short-read* après enrichissement par une technologie de capture de l'ADN ou l'ARN du patient ou de sa tumeur semblent répondre à ce cahier des charges. Ainsi, il est communément admis, aujourd'hui, que le séquençage de 400 à 500 gènes impliqués dans le cancer et une cinquantaine de transcrits de fusion, est suffisant pour établir un profil moléculaire cliniquement pertinent et couvrant l'ensemble des variants actionnables nécessaires à l'utilisation des thérapeutiques ciblées autorisées ou en développement. L'analyse bioinformatique doit être capable de mettre en avant les biomarqueurs à usage diagnostique, pronostique, théranostique, pour un patient parmi l'ensemble des variants génétiques identifiés par NGS. Cette notion de priorisation des variants est une notion clé, afin d'orienter le patient vers de meilleures solutions thérapeutiques. De nombreuses sources compilant les connaissances déjà existantes, comme les bases de données OncoKB, CIViC, clinicaltrials.gov, COSMIC existent et peuvent servir à l'annotation des variants. Il est donc nécessaire de pouvoir proposer les meilleurs outils, afin de les regrouper dans des pipelines efficaces pour proposer une solution qui pourrait s'appliquer à une routine clinique. Le nombre d'informations devient vite exponentiel et doit rester cliniquement pertinent et utilisable.

Le travail de cette thèse a permis le développement de deux outils cherchant à proposer des solutions pour dépasser les contraintes énoncées précédemment en génétique de la tumeur et en génétique constitutionnelle. Afin d'aider dans l'interprétation des données issues du séquençage de tumeurs, nous avons développé DrugOrder, un outil adapté aux larges panels de gènes.

DrugOrder permet l'association de SNV, CNV et fusions de gènes à des thérapies ciblées, afin de détecter les variants actionnables issus du profil moléculaire du patient. DrugOrder contient un *framework* de hiérarchisation des associations de drogues (thérapies) et de variants afin d'ordonner les associations les plus pertinentes pour le patient.

Dans le contexte de la génétique constitutionnelle, afin de résoudre les problèmes d'identification complète d'isoformes alternatives de l'ARN, les techniques de séquençage en *long-reads* ont été mises à profit. *Long Read Isoform Discovery* (LoRID) est un pipeline de détection et d'annotation des isoformes alternatives de l'ARN. LoRID

inclut un outil indépendant d'annotation des isoformes fournissant des informations sur l'expression de l'isoforme ainsi qu'une nomenclature descriptive des différents épissages alternatifs qui la compose.

Ces pipelines doivent être sensibles et spécifiques, s'intégrant dans une démarche qualité forte, les rendant compatibles avec une industrialisation et une utilisation dans un cadre de diagnostic.

3. Profil moléculaire de la tumeur et hiérarchisation des thérapies

3.1 Contexte du développement de DrugOrder

Le séquençage de larges panels de gènes pour identifier des biomarqueurs tumoraux menant à un traitement optimisé ou à un essai clinique devient une pratique courante dans le parcours de soins des patients atteints de cancer. Une des limites de ces analyses est la difficulté pour les biologistes d'effectuer une hiérarchisation standardisée et actualisée des variants actionnables associés à un médicament pertinent, en tenant compte de l'histologie de la tumeur du patient. En effet, le nombre de variants pouvant bénéficier de ces analyses, croit d'année en année et requiert l'emploi d'outils automatisés ayant pour objectif de fournir une aide dans l'interprétation des variants, présents dans une tumeur pour trouver de potentielles cibles thérapeutiques. Ces outils doivent annoter plusieurs dizaines de milliers de variants présents après le séquençage des larges panels. L'annotation nécessite l'utilisation de bases de données variées (impact des variants ou gènes dans les cancers, bases thérapeutiques, impact fonctionnel des variants, bases cliniques), qui agrègent de plus en plus d'informations au cours des années. De plus, il est nécessaire de mettre en relation ces différentes bases de données afin de conclure sur l'intérêt thérapeutique d'un variant.

Il apparaît alors un besoin de mettre en place des outils automatisés afin de faciliter le traitement des données de séquençage d'une tumeur. Ces outils doivent disposer d'algorithmes décisionnels qui permettent de créer des liens logiques entre les bases de données ; de filtrer et ordonner les variants pour rendre une liste réduite ; d'aboutir à des conclusions aidant le biologiste dans ses interprétations cliniques. De plus, ces outils doivent gérer ou être en lien avec d'autres gestionnaires de versions de bases

de données, afin de garantir l'accès à des informations actualisées. Pour répondre à ce besoin nous avons développé une méthode automatisée nommée DrugOrder, permettant de hiérarchiser les variants associés à un médicament, basée sur 19 critères de classification, qui hiérarchisent les associations en fonction des preuves cliniques, de la maladie du patient, de l'impact du variant et de la similarité du variant avec des variants cités dans les bases de données de référence.

DrugOrder a été développé en Python et se décompose en trois étapes :

— En entrée, DrugOrder utilise les fichiers de sortie de différents pipelines spécialisés dans l'appel de variants de différents types (SNV, CNV et fusions de gènes). Les annotations des différents variants sont produites par les pipelines respectifs, DrugOrder conserve les annotations dans cette première lecture des fichiers afin de les utiliser dans l'étape de hiérarchisation.

— S'ensuit l'identification des associations thérapies-variants à l'aide de la base de données *MolecularMatch* (MM), qui propose une API afin de pouvoir effectuer des requêtes à jour, sur les thérapies disponibles pour chaque variant. L'API de MM propose un moteur de recherche capable d'élargir les recherches pour un variant n'ayant pas de thérapies directement associées à ce dernier. MM propose donc des thérapies à l'échelle du variant, du codon, de l'exon ou du gène. De plus, le moteur de MM permet de requêter des associations directement pour un type de variant. Nous avons utilisé cette option pour rechercher les associations de perte de fonction sur les TSG. A titre d'exemple, en considérant le variant *frameshift* E482Rfs*36 sur le gène *BRCA1*, plutôt que de rechercher directement les associations liées directement à ce variant *frameshift* (au travers d'une requête "*BRCA1* E482Rfs*36"), il est possible de rechercher directement les associations liées à une perte de fonction du gène *BRCA1* (au travers d'une requête "*BRCA1* Loss").

— La hiérarchisation des variants s'effectue sur l'ensemble des associations. Le *framework* de classification reprend les annotations de la première étape ainsi que les informations issues de MM afin de classer les variants. Bien que nous ayons utilisé pour DrugOrder différentes bases de données comme COSMIC, CIViC ou encore MM, le *framework* n'est pas dépendant de ces bases de données et se veut compatible avec différentes bases, tant qu'elles exposent les annotations requises pour la hiérarchisation (cf. Supplementary Table S2 du papier associé). La publication de DrugOrder se concentre principalement sur la description des différentes composantes de la hiérachisation effectuée par DrugOrder.

Les performances de DrugOrder ont d'abord été évaluées sur 371 tumeurs simu-

lées contenant au moins un variant actionnable. Toutes les associations impliquant une thérapie approuvée par la FDA ont été identifiées et 94% des variants identifiés ont été classés en première position. DrugOrder a également été évalué en analysant 15 tumeurs séquencées préalablement analysées par Foundation Medicine Incorporation® (FMI®) et reséquencées après la mise au point d'un test à façon comprenant la capture d'un panel de 639 gènes de cancer et un panel ciblant 57 gènes impliqués dans des transcrits de fusion. Les analyses au laboratoire avec DrugOrder ont correctement identifié les variants associés à des thérapies précédemment détectées par FMI®. Enfin, 62 autres tumeurs ont été séquencées au laboratoire puis analysées par un biologiste sans l'aide de DrugOrder. DrugOrder a permis d'identifier tous les variants actionnables associés aux médicaments approuvés par la FDA et a classé en première position 88% des variants identifiés par le biologiste. DrugOrder a identifié également des cibles non reportées par le biologiste mais pouvant faire l'objet de discussion clinico-biologique.

DrugOrder présente de bonnes performances et permet d'aider les biologistes à mettre les variants les plus importants cliniquement au premier plan, ce qui représente un gain de temps considérable à l'analyse.

3.2 DrugOrder : une méthode automatique pour prioriser les associations thérapies-variants au travers de l'utilisation d'un large panel de gènes

Le présent travail est en cours de soumission auprès de l'*European Journal of Cancer*. Toutes les figures, tables et données supplémentaires sont disponibles dans le répertoire GitHub suivant : `https://github.com/Nedss/these-data` dans le répertoire DrugOrder.

DrugOrder: an automated method for ranking drug-variant associations using large panel of cancer genes

Nicolas Soirat[1,2,3,4], Sophie Krieger[2,3,4], Nicolas Hamadouche[2], Nicolas Goardon[2], Mélanie Broutin[1], Raphaël Lanos[1], Jiri Ruzicka[1], Sacha Beaumeunier[1], Anne-Laure Bougé[1], Nicolas Philippe[1], Dominique Vaur[2,3], Denis Bertrand[a,1], Laurent Castéra[* a,2,3]

1. SeqOne, Montpellier, France
2. Laboratoire de biologie et de génétique du cancer, Centre François Baclesse, 14000 Caen, France
3. Inserm U1245 Cancer and Brain Genomics, Normandie, France; Univ, UNIROUEN, Rouen, France
4. Unicaen, Caen, France

[a] This Author contributed equally to this study

* Corresponding author

l.castera@baclesse.unicancer.fr

Centre François Baclesse

Département de BioPathologie, Laboratoire de biologie et de génétique du cancer

Inserm U1245 - Cancer and Brain Genomics

3, avenue du général Harris

14076 Caen

+33 2 31 45 51 54

Abstract

Large gene panel sequencing to pinpoint tumor biomarkers, which can lead to optimized treatments or clinical trials, is becoming a standard practice in the optimal care pathways for cancer patients. However, one challenge is the difficulty for biologists to standardize and keep up-to-date the tumoral knowledge linked to drugs, while considering the tumor histology.

We developed a new method called DrugOrder that prioritizes actionable variants (*i.e* leading to optimed treatment), taking into account the patient's tumor mutational landscape and tumor histology. DrugOrder prioritizes associations based on clinical evidence, patient's disease, variant impact, and variant similarity with known targetable variants.

We first tested the performance of DrugOrder's on 371 simulated tumors, each containing a single actionable variant. All associations involving an FDA-approved drug were identified, and 94% (283/301) of the identified variants were ranked in the top position. To further evaluate DrugOrder's effectiveness on realèlife samples, we re-sequenced 15 DNA/RNA tumors. These were initially sequenced and analyzed by Foundation Medicine Incorporation (FMI)®. We used a lab panel of 639 cancer genes and 57 gene fusion transcripts. DrugOrder successfully identified associations previously detected by FMI®. Lastly, we sequenced and analyzed 62 tumors, with variants previously interpreted by laboratory pathologists. DrugOrder identified all targetable variants associated with FDA-approved drugs and ranked 88% (37/42) of identified variants in the top position.

DrugOrder should enable biologists to more easily report relevant therapeutic choices to oncologists. This makes it possible to use large scale panels in the context of precision medicine.

Keywords

Cancer, Computational Biology, Sequence Analysis, Pharmacogenomic Variants, Classification, Precision Medicine

Introduction

Over the past decade, the advent of Next-Generation Sequencing (NGS) has paved the way for a comprehensive understanding of tumor variants (1–3). This characterization of tumors has spotlighted targetable oncogenic variants that can predict a treatment's effectiveness or resistance. These variants are called targetable variants, *i.e* they guide therapeutic management in the context of precision medicine. The care and support, but also survival rates of cancer patients are improving annually, partly due to the ability to prescribe drugs based on relevant tumoral genetic markers identified (4–7). This approach could benefit from whole genome tumor sequencing, exome or from the sequencing of known cancer genes at the DNA level or at the RNA level for targeted transcripts (1,8) to identify Single Nucleotide Variants (SNVs), Copy Number Variation (CNVs) and gene fusions. Nowadays, the cost-effective and medically efficient method is to sequence gene panels that include both tumor DNA and RNA (9). Sequencing a large panel of genes may allow the detection of actionable variants that are either approved by the Food and Drug Administration (FDA) or usable in clinical trials as prognosis molecular biomarkers or targetable variants in the field of precision medicine.

The sequencing of extensive gene panels often uncovers a multitude of variants, most of which have uncertain significance. Therefore, characterizing these variants is crucial in identifying potential targetable variants. Interpreting different types of variants necessitates the use of numerous databases, including generalist databases like dbNSFP (10–13), cancer variant databases like COSMIC or ClinVar (14–16) and therapeutic databases like DrugBank (17), OncoKB (18), CIViC (19) or COSMIC Actionability (20,21). These databases help identify targetable variants and their corresponding targeted therapies (drugs).

Frameworks proposed by AMP/ASCO/CAP (22) or SoVaD (23) aid in harmonizing and standardizing the ranking of drugs associated with tumor variants. These frameworks require features such as automation for up-to-date bibliography, characterization of the variant's impact on the gene, characterization of the gene itself, and characterization of the biological pathway affected by the drug.

Moreover, any method used to rank variants and their associated drugs should consider the type of cancer, as a targetable variant may be validated for a specific cancer type but may be debatable in another.

Given the complexity of classifying variants derived from sequencing hundreds of genes in a tumor and the huge number of targeted therapies, an automated informatics method is essential. This method should be capable of identifying and ranking numerous pairs of molecular targets and their associated drugs (referred to here as drug-variant associations).

The number of approved therapies, therapies recommended in guidelines, and clinical trials is vast and continually evolving with new drugs and molecular targets (24,25). Therefore, molecular biologists/pathologists require precise methods to highlight targetable variants and prioritize them. Several tools have been developed to address this issue, automating clinical recommendations for certain types of variants like SNV or CNV and specific cancer patient cohorts (26,27). These tools offer prioritization methods (28–30), but they are not thoroughly described for all variant types, such as fusions (31,32).

In this context, we have developed a drug-variant association prioritization framework, named DrugOrder, to assist medical biologists or pathologists. DrugOrder considers SNV, CNV, and gene fusion, providing a score for each identified drug-variant association. The score is determined by the predicted

functional impact of the variant, the clinical classification (ASCO/AMP), the patient's disease, and the relevance of the drug-variant association. The scoring system is based on a comprehensive variant annotation, drawing conclusions from different generalist databases and cancer-related databases such as dbNSFP, COSMIC, and MolecularMatch (33). DrugOrder has been integrated into a bioinformatics pipeline specifically designed for analyzing tumor sequencing using custom large DNA and RNA gene panels.

To assess the effectiveness of our prioritization, we tested DrugOrder on variants from 371 simulated tumors. We then validated DrugOrder on 62 tumors, sequenced using a large DNA gene panel, for which we had access to both laboratory pathologists' clinical conclusions and the clinical conclusions from another external genetic test for 15 of these cases (34). We compared these analyses with the results from DrugOrder to evaluate the tool's ability to draw conclusions similar to those of a pathologist.

Materials and Methods

Tumor samples

We obtained 62 paraffin-embedded tumor samples from the Anatomic Pathology Laboratory of the Comprehensive Cancer Center (CCC) François Baclesse (CFB) in Caen, France (Supplementary Table S1). In compliance with French regulations, patients were informed about the research conducted on their tissue specimens and did not object. Fifteen of these tumors were sequenced and analyzed, with targetable variants listed in a clinical report provided by Fondation Medicine Incorporation (FMI) ®.

Sample preparation and next-generation-sequencing (in-house test)

The mutational status (SNV, indels, CNV and translocations) of the original tumor samples was determined by targeted re-sequencing of a panel of 638 genes implicated in cancer for DNA and a panel of 57 genes implicated in protein fusion for RNA (Supplementary Data S1). For RNA sequencing, gene fusion probes were selected to target one partner of the gene fusion, regardless of the second partner. DNA and RNA from samples were extracted using the truXTRAC FFPE total NA Plus Kit - Magnetic Bead (COVARIS), following the manufacturer's instructions. DNA and RNA library sample preparations and targeted enrichment were performed using the SureSelect XT HS2 DNA Reagent Kit (Agilent) and SureSelect XT HS2 RNA Reagent Kit (Agilent), respectively, as per the manufacturer's instructions.

Paired-end sequencing at 2x100 pb was conducted on a Nextseq500 sequencer (Illumina).

Bioinformatic pipeline

For DNA alignment, we used BWA (v.0.7.15-r1142-dirty) (35) with the -M option. Non-aligned reads were realigned using Minimap2 (v2.24) (36). For SNV, variant calling was performed by two callers: Freebayes (v1.3.6) (37) and Mutect2, from the GATK package (v4.1.4.1) (38,39).

CNVPanel, a proprietary tool from SeqOne, was used for CNV calling. This tool is based on average coverage by regions normalized on the patient cohorts. Only amplifications impacting ≥ 90% of the gene were considered as true CNV events.

UMI data was processed using the standard FGBio UMI (v0.8.1) pipeline (40), which contains 3 submodules: GroupReadByUMI with "adjacency" as the strategy option, CallMolecularConsensus, and FilterConsensusRead. UMI processed reads were aligned according to the process described previously.

For RNA, UMI data was processed as for DNA and aligned using STAR (v.2.7.3a) (41). RNA fusion was detected using Arriba (v1.2.0) (42). Fusion events supported by at least 2 discordant mates and split reads were conserved. A complete description of tool options is provided in Supplementary Table S2.

Annotation pipeline

Each gene of the panel was annotated for the cancer gene role either as oncogene or tumor suppressor genes (TSG) according to a consensus based on different databases: COSMIC Gene Census (21), Vogelstein (43), Cancermine (44), ONGene (45), TCGA (1), TSGDatabase (46) and IntOGen (47). For genes without annotation from these databases, JaxCKB (48) and OncoKB (18) databases were used, as we do not have full access. Role was deducted according to the absolute majority. If the role could not be deduced, it was annotated as oncogene/TSG (Supplementary Data S1, Supplementary Figure S1).

SNVs were first annotated with VEP using the RefSeq database version (v104) (49). Variant frequency was determined with the maximum value among all populations of gnomAD (v2.1) (50). Variants were annotated with the COSMIC Mutation Census (CMC v92) tier and using ClinVar (v2021-10) pathogenicity tier (pathogenic, unknown significance, benign). CIViC (December 2017 version) was also used to evaluate the pathogenicity of a variant in cancer. To determine the impact of the variant over proteins, we used the following predictor scores from dbNSFP (4.1a): Sift, MutationTaster, fathmm-MKL, LRT, MetaLR, MetaSVM, PROVEAN, FATHMM, DANN and MutationAssessor.

Arriba is used for gene fusions detection, annotating transcripts with ENSEMBL (51) and predicting their effects(out-of-frame, in-frame). Protein domain annotations were performed using Pfam (v28.0) (52). As RNA pipeline was built for Ensembl annotation, for coherence with DNA annotations, Ensembl annotations for RNA pipelines were converted in RefSeq annotation using a conversion table from NCBI (Supplementary Data S3). We annotated fusions using COSMIC Fusion (v91). A transcript fusion identified in a tumor is associated with a fusion from COSMIC fusion (20) with the closest breakpoint genomic position on the same transcripts.

Drug-variant associations were extracted from the MolecularMatch database (commercial access API version v5). Complete description of the annotations required by DrugOrder are listed in the Supplementary Table S3.

Protocol for simulated tumor generation

The results of the sequencing of 638 targeted genes corresponding to the SureSelect targeted enrichment were simulated to create a dataset of variants from 371 tumors. Each simulated tumor was defined to include 20 SNV per megabase

corresponding to 19 passenger variants and 1 targetable variant. Passenger variants were randomly selected from coding variants in the COSMIC Mutation Census database. A targetable variant was defined as a variant associated with at least one drug conferring a therapeutic option. Targetable SNVs, with valid HGVSp, were selected in the CIViC database (December 2017 version, see Supplementary Data S2 for the whole list). For each targetable variant the clinical evidence was extracted from CIViC and evidence levels were given as follows: level A = validated association; level B = a clinical evidence; level C = a case study and level D = a preclinical data. For variants with multiple clinical evidences for a same disease, only the assertion with the highest level was used. Variants supported by an E evidence level (indirect clinical evidence) were discarded as they only represented 5 variants. The tumor type of the simulated tumor was defined by the tumor type where the targetable variant was described in CIViC.

Manual interpretation of sequenced tumor

A pathologist from the CFB performed the analysis of 62 tumors sequenced using the laboratory panel. The interpretation of the data was carried out on the SeqOne platform, following variant filtering (Supplementary Table S4), but without access to the DrugOrder scoring (see *infra*). The targetability of the variant was manually evaluated using the OncoKB database.

DrugOrder scoring

DrugOrder is a tool designed for identification and prioritization of couples of targeted therapy and targeted variant called drug-variant association. The method is constructed around a four-part framework that evaluates the significance of a drug-variant association. Each part of DrugOrder assigns a tier that provides : (i) a

clinical evidence classification score (Figure 1a), (ii) a similarity score between the patient's disease and the drug's tumor target found in the association (Figure 1b), (iii) a score that measures the similarity between the tumor variant and a drug-variant association in a database (Figure 1c) and (iv) a pathogenicity level score of the variant at the gene level (Figure 1d, e, f). Each part contributes a score that corresponds to a tier symbolized by a letter for the part and a number for the score's weight (e.g. A1, D4, S2 Figure 1). The values for tier scores have been determined empirically.

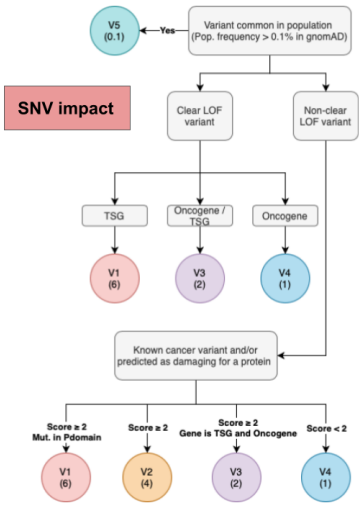
a.

Clinical evidence score		
Tier	**Description**	**Score**
A1	FDA approved therapy	10
A2	Powered studies backed by expert concensus	8
A3	FDA-approved therapy for different tumor type	6
A4	Preclinical studies or lacking expert concensus	4
A5	Variant of unknown or benign clinical significance	2

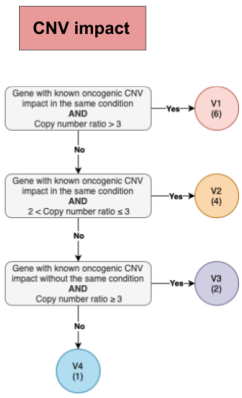
b.

Disease similarity score		
Tier	**Description**	**Score**
D1	Assertion found in the same cancer type than the patient	100
D2	Assertion found in the same organ than the patient	70
D3	Assertion found in the same hematopoietic or solid tumor than the patient	25
D4	Assertion found has no disease similarity with the patient	1

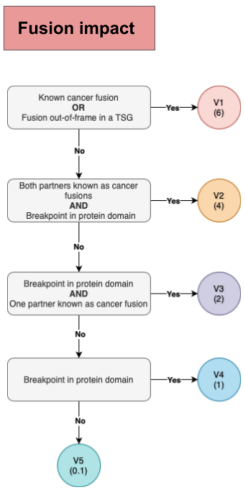
d.



e.



f.



c.

Variant similarity score											
	SNV				**CNV**		**Fusion**				
Tier	S1	S2	S3	S4	S1	S2	S1	S2	S3	S4	S5
Description	Assertion on same variant than the patient **OR** LoF in a TSG	Assertion on same codon than the patient	Assertion on same exon than the patient	Assertion on same gene than the patient	Assertion on the same amplified gene than the patient	Assertion on same gene than the patient	Assertion in the same exons and genes than the patient **OR** LoF in a TSG	Assertion in the same genes than the patient **AND** Assertion in the same exon than the patient for one partner	Assertion in the same genes than the patient	Fusion assertion in only one gene of the patient fusion	Assertion on same genes than the patient
Score	4	3	2	1	4	3	4	3	2	1	0.1

Figure 1: Description of DrugOrder prioritization framework parts.

(a) Table showcasing the Clinical Evidence Score. **(b)** Table demonstrating the Disease Similarity Score. **(c)** The Variant Similarity Score. The scoring for SNV, CNV, and fusion tiers are individually described, as unique features are utilized for their

respective scoring. **(d, e, f)** correspond to the SNV, CNV, and fusion impact tiers, respectively. We classify a known mutation as one that is documented in cancer databases. The colored circles symbolize each tier, with their values indicated in parentheses.

i) Clinical evidence classification score:

This part of the framework is structured around five tiers aligning with ASCO/AMP guidelines: FDA-approved therapies or those recommended in professional guidelines (A1); powered studies backed by expert consensus (A2); FDA-approved for different tumor types (A3); preclinical reports or lacking expert consensus (A4); variant of unknown or benign clinical significance (A5). Each tier is assigned a score between 10 for A1 and 2 for A5 (Figure 1a).

ii) Disease similarity score

DrugOrder determines a score by assessing the resemblance between the disease specified in the assertion and the patient's disease. It classifies both the patient's disease and the assertion disease based on the following criteria: identical disease (D1); disease affecting the same organ (D2); both patient's tumor and the assertion disease are either hematopoietic or solid tumor (D3); no similarity in disease detected (D4). The top tier, D1, is assigned a score of 100, while the lowest, D4, receives a score of 1 (Figure 1b).

iii) Variant similarity score

DrugOrder assesses the similarity between each drug-variant assertions and the patient's variants, based on their type (Figure 1c).

For SNVs, DrugOrder uses a four-tier system: perfect match (S1); same codon (S2); same exon (S3); same gene (S4). A perfect match (S1) means the patient and the

assertion share the same amino acid change. If the patient has a clear loss of function (LoF) variant (frameshift, nonsense, or AG/GT splice sites) in a TSG, S1 applies to all assertions with a LoF in that TSG. The same codon (S2) refers to a different amino acid change at the same position (e.g., KRAS G12D and KRAS G12V). Tiers S3 and S4 apply when the patient's SNV shares the same exon or gene, respectively, with the assertion's SNV.

For Copy Number Variants (CNVs), DrugOrder uses a two-tier system. An assertion is considered a perfect match (S1) if the amplified gene in the patient and the assertion are identical. The same gene (S2) tier applies to an assertion that describes a targetable gene identical to the patient's CNV gene but without targeted therapy dedicated for amplification for this gene.

For gene fusions, DrugOrder uses a five-tier system: perfect match (S1); partner single exon matching (S2); partner gene matching (S3); one partner matching (S4); gene matching (S5). A perfect match (S1) is when the assertion has the same partners and the breakpoints are in the same exons as the patient's. If the patient has an out-of-frame fusion in a TSG, S1 applies to all assertions with a LoF in that TSG. Tiers S2 and apply when the assertion and the patient's fusion share the same gene partners and one breakpoint of the tumor fusion is in the same exon as the assertion’s fusion. Tier S3 is given when the assertion and the patient’s fusion have the same gene partners and no breakpoint of tumor fusion is in the same exon as the assertion’s fusion. One partner matching (S4) applies when at least one partner of the patient's fusion is part of the assertion's fusion. Gene match (S5) applies to all other cases.

Each tier is associated with a score ranging from 0.1 to 4, with intermediate values depending on the variant types (Figure 1c).

iv) Variant pathogenicity score

DrugOrder assesses the pathogenicity level for all types of patient variants (Figure 1d, e, f).

For Single Nucleotide Variants (SNVs), we have adapted ComPerMed scoring (Supplementary Table S5) and method (53) (Figure 1d) with the following definition: Pathogenic (V1), Likely Pathogenic (V2), Variant of Uncertain Significance (VUS) High (V3), VUS Low (V4), and Likely Benign or Benign (V5). The Pathogenic tier (V1) is characterized by clear Loss of Function (LoF) variants in Tumor Suppressor Genes (TSGs). For non-clear LoF variants with a ComPerMed score of 2 or more in an oncogene's protein domain, pathogenic tier (V1) is given. The Likely Pathogenic (V2) tier includes non-clear LoF variants with a ComPerMed score of 2 or more. The VUS High (V3) tier is associated with clear LoF variants or non-clear LoF variants with a score of 2 or more, in a gene that has both TSG and oncogene roles. The VUS Low tier (V4) is assigned to clear LoF in oncogene or non-clear LoF variant with a ComPerMed score of less than 2. The Likely Benign (V5) tier corresponds to variants with a minor allele frequency of more than 0.1% in at least one population in gnomAD.

For Copy Number Variations (CNVs), we have chosen the COSMIC Cancer Gene Census to determine the potential oncogenic impact of the CNV. The impact evaluation of CNVs is divided into four tiers (Figure 1f): High CNV Impact (V1), Likely High CNV Impact (V2), Moderate CNV Impact (V3), and Unknown CNV Impact (V4). Tier V1 corresponds to a CNV with a copy number ratio of 3 or more on a gene with a known oncogenic CNV impact in the same disease as the patient. Similarly, tier V2 corresponds to a copy number ratio of 2 to less than 3. Genes with a copy number ratio of 3 or more with a known oncogenic CNV impact on a different disease of the

patient are defined as Moderate Impact (V3). In other cases, the CNV impact is considered as Moderate (V4).

For gene fusions, DrugOrder's impact prioritization steps are based on a comparison of known cancer fusions from the COSMIC Fusion database. The prioritization is divided into 5 tiers (Figure 1g): High Impact Fusion (V1), Likely High Impact Fusion (V2), Moderate Impact Fusion (V3), Unknown Impact Fusion (V4), and Low Impact Fusion (V5). A gene fusion is considered to have a high impact (V1) if a known cancer fusion has the same partners. This tier also applies to out-of-frame fusions in TSGs. Partners involved on 2 different known cancer fusions are considered to have a likely high impact (V2). Tier V3 is given if there is a breakpoint over the protein domain and only one partner is involved as a known cancer fusion. Tier V4 is given when no partner is involved as known cancer fusion. Other fusions are considered to have a low impact (V5). Each tier is associated with a score from 6 to 0.1 (Figure 1g).

v) DrugOrder score calculation

The DrugOrder score (DOscore) is calculated for each tier using the formula:

$$DOscore = \frac{A_{score} * D_{score} * V_{score} * S_{score}}{Normalization\ factor}$$

This normalization factor is the maximum possible value for each tier, used to standardize the DOscore for every variant type. The DrugOrder score calculation is dependent on the role of the cancer gene. For genes that are considered to have dual roles, DrugOrder calculates a score for variants using both Oncogene and TSG logic, assigning the highest score. The variant DOscores are determined by taking the highest associated assertion DOscore. Once the scores are sorted, the position

in the sorted list determines the DrugOrder rank (for instance, position 1 in the sorted list corresponds to rank 1).

Results

Evaluation of DrugOrder ranking on simulated tumors

We initially assessed the DrugOrder on a simulated group of 371 tumors, each containing 19 passenger variants and 1 targetable variant (Figure 2a). The variants were categorized based on their highest CIViC clinical evidence, resulting in 6 variants with a validated association (level A), 60 variants with clinical evidence (level B), 151 with case study evidence (level C), and 154 with preclinical evidence (level D) (Figure 2b).

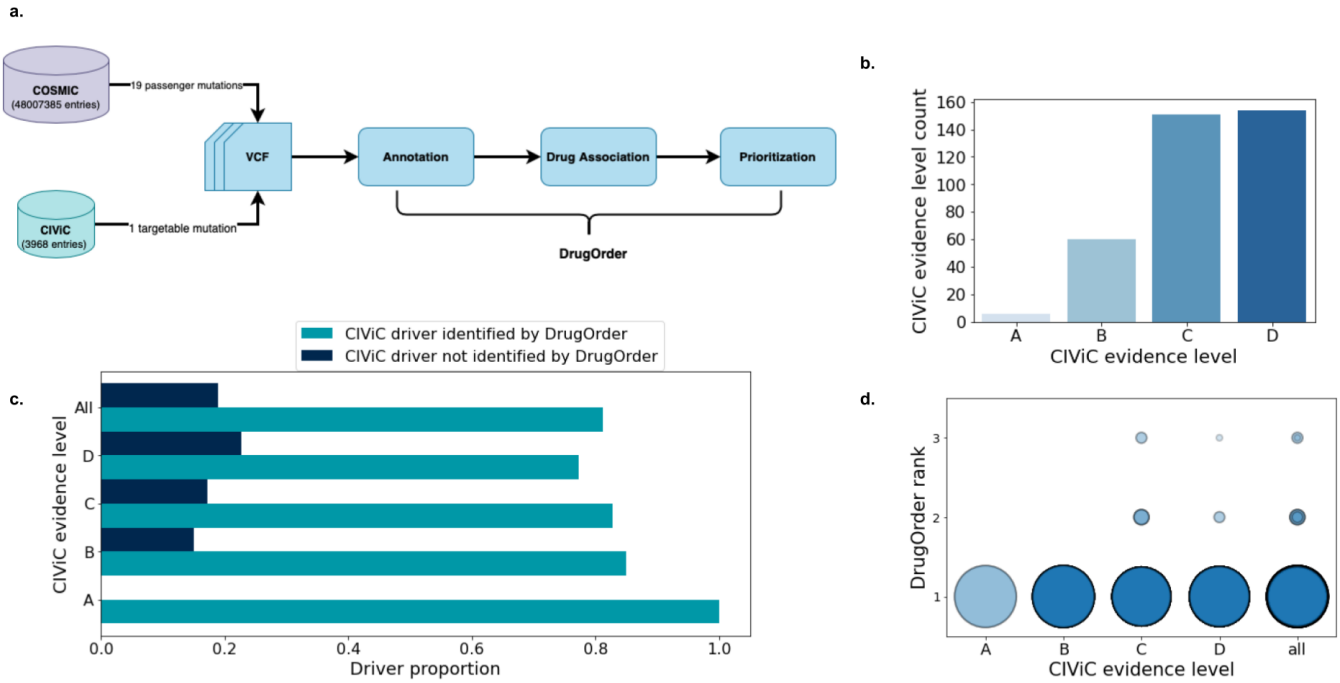


Figure 2: Evaluation of simulated tumors

(a) Protocol for simulated tumor generation. **(b)** Targetable Variants from simulated tumors based on CIViC Evidence Levels. **(c)** Proportion of CIViC targetable variants detected (or not) by DrugOrder, categorized by their respective evidence levels. The

"All" category includes variants from all CIViC evidence levels. **(d)** DrugOrder ranking of targetable variants based on their clinical evidence. The color gradient illustrates the proportion of variants within each CIViC evidence level.

DrugOrder successfully annotated all level A variants. Interestingly, 18% (70/371) of targetable variants that DrugOrder missed were evenly distributed across the lower evidence levels (Figure 2c). As for the unidentified level B variants, we observed that they were either associated with another level or absent in the OncoKB database (Supplementary Table S6). This suggests potential inconsistencies between databases when evidence levels are lower.

DrugOrder assigned a rank 1 to all targetable variants with level A or B (Figure 2d). More than 90% (226/244) of level C or D targetable variants were ranked first by DrugOrder (Figure 2d). This demonstrates that DrugOrder effectively identifies and prioritizes the most pertinent variants in these simulated tumors.

Unfortunately, we could not compare the similarity of each drug recommended by CIViC and DrugOrder. As Wagner et al. (54) pointed out, our comparison of drug name similarity on 9 variants for MM, FMI®, OncoKB, and CIViC (Supplementary Figure S2) revealed significant heterogeneity in the drugs reported in different databases. Hence, for all our experiments, we solely compared results based on the scale of variant targetability. This implies that the highest DrugOrder score for a drug, associated with the targetable variant, is attributed to the variant itself.

Relevance of DrugOrder analysis compared with drug-variant associations characterized by Foundation Medicine Inc. (FMI)®

We conducted sequencing on 15 bladder tumors using both CFB's custom in-house test and FMI®'s sequencing and analysis procedure (Figure 3a and Supplementary Data S4). FMI® reported a median of one targetable variant per patient, totaling 13 SNVs and 4 CNVs across the cohort. Our in-house tests, however, reported a median of 117 high-quality variants per patient that passed the filter (Supplementary Table S7). Post DrugOrder, an average of 6.5 variants per tumor were linked to a drug (Figure 3b).

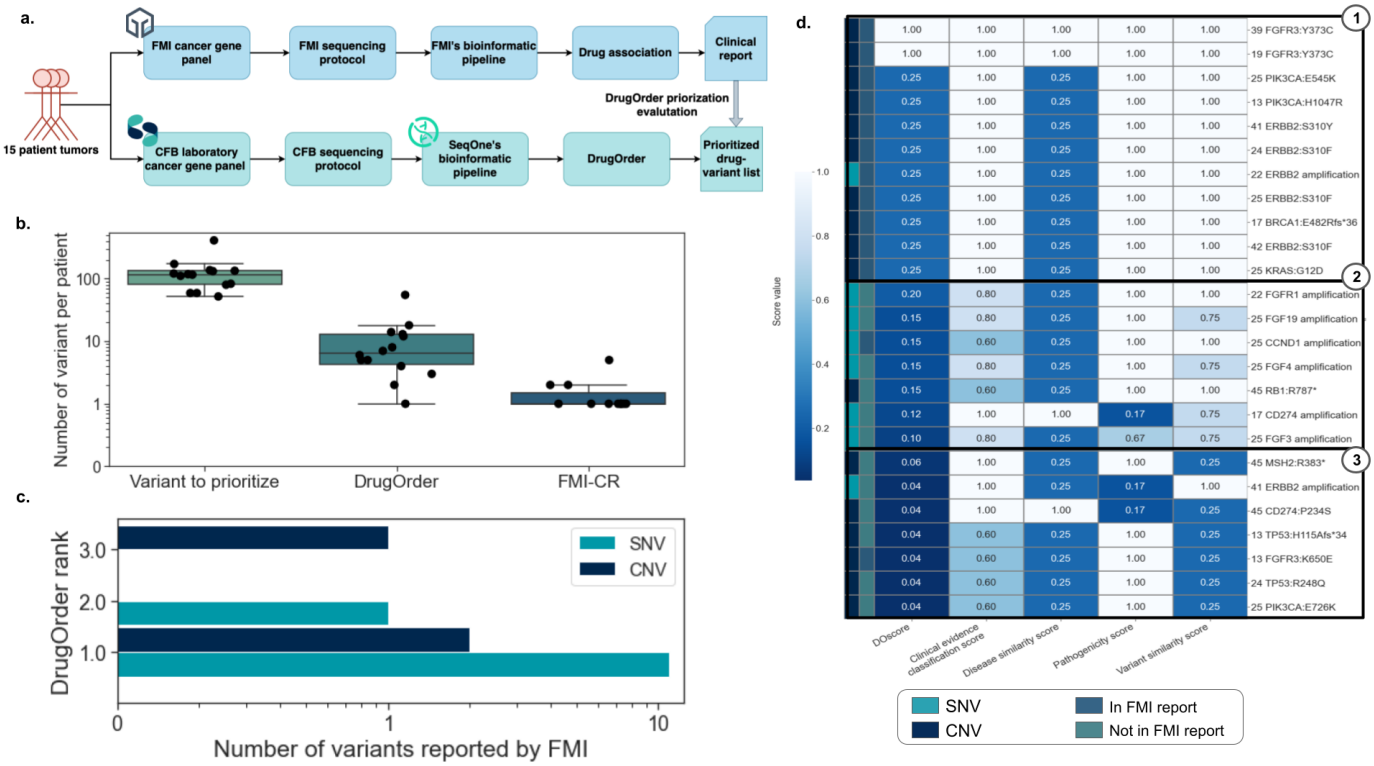


Figure 3: Comparison with FMI® characterization

(a) This illustrates the comparison protocol between FMI® and DrugOrder analysis. **(b)** The boxplot depicts the number of variants per patient, both before and after the

prioritization by DrugOrder, and within the FMI® clinical reports (FMI-CR). **(c)** The DrugOrder ranking of targetable variants as identified by FMI® is shown. The blue bar signifies SNV, while the sea-blue bar represents CNV. **(d)** A heat map of the top 20 DOscored variants is presented. The first column on the left indicates the variant type, with SNV in blue and CNV in sea-blue. The second column on the left shows whether the variant is included in the FMI® report or not. The heat map columns represent the DOscore and the framework parts. DOscore is normalized based on the highest sample DOscore, while framework part tier scores are normalized based on the best framework part tier score. The heat map is divided into three distinct groups based on the sample DOscore.

In detail, out of the 13 SNVs mentioned in the FMI® reports, 11 were ranked first by DrugOrder. One variant was ranked second in a sample because another variant also reported by FMI® took the top spot (Figure 3c and sample 13 in Supplementary Data S4). A frameshift variant in *TSC1*, identified by our bioinformatics pipeline and included in the FMI® report, was not ranked by DrugOrder as it was not associated with any drug in Molecular Match. However, it was present in OncoKB with an FDA evidence level of 1 for oncogenic mutations.

Of the 4 targetable CNVs in the FMI® reports, 3 were ranked by DrugOrder. Two took the top spot, and one was ranked third in a sample due to another variant in the FMI® report already occupying the first position (Figure 3c and sample 25 in Supplementary Data S4). A *RAF1* amplification, identified by our bioinformatics pipeline and included in the FMI® report, was not ranked by DrugOrder. This amplification was not associated with any drug in Molecular Match but was present in OncoKB with an FDA evidence level of 3.

We further examined the DrugOrder predictions by analyzing their DOscore and tier score (Figure 3d and Supplementary Data S4). We categorized these into three groups based on their DOscore values. As anticipated, variants found in FMI® clinical reports made up Group 1 (DOscore ≥ 0.2) as they had the highest DOscore. Group 2 (DOscore between 0.1 and 0.2) included one variant reported by FMI® and variants identified as targetable exclusively by DrugOrder. Among these, DrugOrder identified a *FGF19* and *FGF4* amplification associated with Lucitanib (55), supported by 1B clinical evidence, in the same patient. Group 3 (DOscore < 0.1) contained targetable variants with a DOscore less than 0.1, indicating lower confidence in variant targetability. Despite this, Group 3 still contained some variants with 1A clinical evidence, but the associated drugs are not specifically recommended for a particular variant in the database, resulting in a low variant similarity score (0.25).
In conclusion, our analysis protocol has proven effective in identifying and ranking top variants mentioned in the FMI® report.

Comparison of DrugOrder with a biologist analysis from a laboratory cohort of cancer patients

We assessed DrugOrder's performance using a cohort of the whole 62 patient tumors, sequenced by our in-house test (Figure 4a), and corroborated with internal pathologist clinical reports (Supplementary Data S5). The raw variants underwent filtering steps (Supplementary Table S4), which reduced the variant count to a median of 100 per patient. Filtering steps are adapted for larger cohorts (***e.g*** variant frequency). DrugOrder identified at least one targetable variant per patient, with a median of 10 variants per patient. Conversely, pathologist interpretation revealed a

median of one variant per sample, with 19 patients having no detectable targetable variants (Figure 4b).

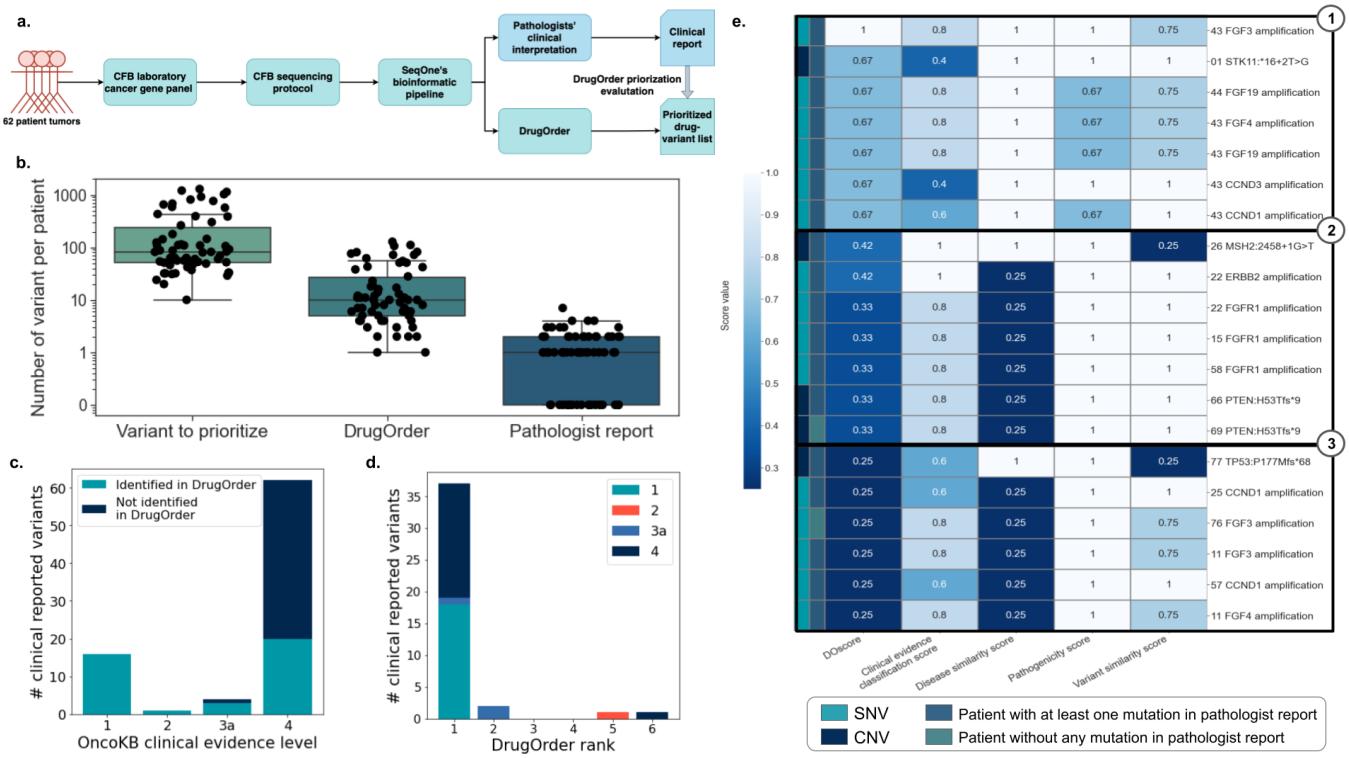


Figure 4: Comparison CFB clinical conclusions

(a) Protocol for comparing pathologist clinical reports: The CFB laboratory sequenced 62 tumor samples from cancer patients, with bioinformatics analysis conducted using SeqOne's pipelines. Clinical reports were compiled by laboratory pathologists, while DrugOrder generated a prioritized list of actionable variants. **(b)** Boxplot illustrating the number of variants per patient before and after DrugOrder prioritization, as well as in the pathologist clinical reports. **(c)** Variants either identified or missed by DrugOrder, categorized by their OncoKB clinical evidence level. Variants identified by DrugOrder are depicted by blue bars, while those missed are shown in sea-blue bars. **(d)** DrugOrder's ranking for CFB-CR variants, based on the OncoKB clinical evidence level. **(e)** Heatmap of the top 20 DOscored variants not

included in the pathologist clinical reports. The first column on the left indicates the variant type, with SNV in blue and CNV in sea-blue. The second column on the left shows whether the sample has one or no mutations in the pathologist clinical report. The heatmap columns represent the DOscore and the framework parts. DOscore is normalized according to the highest sample DOscore, while framework part tier scores are normalized based on the best framework part tier score. The heatmap is divided into three distinct groups based on the sample DOscore.

DrugOrder successfully identified all variants with an OncoKB level 1, as indicated in the pathologist clinical reports, and 49.4% of the variants detailed in the report (Figure 4c). Variants undetected by DrugOrder were predominantly (97.7%) of an OncoKB level 4, which have low clinical evidence. Moreover, 81.25% of variants reported by the pathologists were ranked in the top positions. All OncoKB level A variants were among these (Figure 4d).

As anticipated, all variants in the pathologist report were ranked at the top. Therefore, we examined the top 20 variants exclusively identified by DrugOrder (Figure 4e). Group 1 (DOscore ≥ 0.60) consists of level D ASCO/AMP variants (according to MolecularMatch), including *FGF* gene family amplifications, linked to preclinical use of Lucitanib (55) for the same disease as the patient. Group 2 (DOscore between 0.3 and 0.6) includes variants associated with a drug with a lower clinical evidence level score and recommended for a different disease. However, we noted a *PTEN* frameshift in a patient where no variants were manually found. Even though the drug is in an early clinical stage (56), its association with a drug is supported by OncoKB (FDA level 3) and is an intriguing candidate for further

investigation. Group 3 (DOscore < 0.3) comprises *FGF* amplifications, but the disease similarity decreased the rank of these variants.

In conclusion, we observed the top 20 variants (highest DOscore) in patients without manually reported variants (Supplementary Figure S3). We also detected a frameshift variant in *PTEN* and a *FGF3* amplification, but the disease similarity is low. The remaining variants in this heat map displayed low DOscores with low pathogenicity (average of VUS), low similarity score (gene scale), and disease therapy that does not match the patient's disease. These findings suggest that DrugOrder may reveal additional drug-variant associations outside standard care, potentially leading to therapy or clinical trial inclusion that might have been overlooked by manual analysis.

Discussion

We have developed a flexible framework, DrugOrder, for prioritizing drug recommendations based on a patient's genomic landscape. When tested on various datasets, DrugOrder's performance matched that of human interpretations of variants. Our evaluation method involved using a simulated dataset and laboratory tumors sequenced by our in-house test and the FMI® tests. The simulated dataset demonstrated DrugOrder's ability to effectively prioritize targetable variants amidst mutational background noise. Furthermore, DrugOrder proved its capacity to accurately prioritize targetable variants in comparison to those mentioned in pathologist reports.

In our study, we focused on the main variant types as drug targets: CNV, SNV, and gene fusion. However, there are potential targetable variant types, such as RNA aberrant splicing or loss of gene copy (or partially). They could be integrated into the framework in the near future. We investigate inclusion of Microsatellite Instability (MSI), Tumor Mutation Burden (TMB) as mutational signatures in the DrugOrder prioritization framework.

In France, the Institut National du Cancer (INCa) estimates that 80,000 molecular genetic tests are prescribed each year. With the increasing number of patients in Cancer centers, laboratories require efficient tools to support diagnostic workflows and boost productivity. DrugOrder highlights targetable mutations, offering two distinct analysis dimensions. Firstly, it simplifies the identification of FDA-approved drugs. Secondly, it opens the door to alternative treatments when first-line treatments have been ineffective, thereby increasing the chances of patient inclusion in clinical trials or compassionate care.

Our experiment comparing CIViC, FMI®, and OncoKB results revealed discrepancies due to missing assertions in databases (e.g. RAF1 amplification) or assertions with varying levels of clinical evidence (e.g. simulated tumor analysis). This underscores the need for multiple databases to enhance the detection and prioritization of drug-variant associations. We have designed our framework to be adaptable, allowing for the integration of several databases in the future such as Agency for Research on Cancer TP53 database (57) and BRCA exchange (58); therapeutic databases like CIViC, OncoKB (18), JaxCKB (48). Efforts to query cancer gene-specific databases and therapeutic databases should further enhance DrugOrder's performance. However, integrating a large number of databases necessitates the development of an up-to-date automated prioritization method, requiring robust technical infrastructure and tools for updating databases to the latest versions.

Looking ahead, we know that in cases of first-line treatment failure, a patient's profile can be re-analyzed. In such instances, DrugOrder could identified drugs available in clinical trials. Our future framework upgrades aim to identify the most suitable clinical trials for patients, prioritizing trials based on their inclusion and exclusion criteria, given the patient's genomic landscape and clinical information. There are strong lines of evidence that genomic profiling in cancer will enhance therapeutic possibilities for patients and increase participation in clinical trials (59). DrugOrder could be a valuable tool for physicians in selecting therapeutic options, particularly for therapies not yet approved, leading to discussions during tumor board meetings.

Acknowledgments

We would like to thank Raphael Leman for his review of this manuscript.

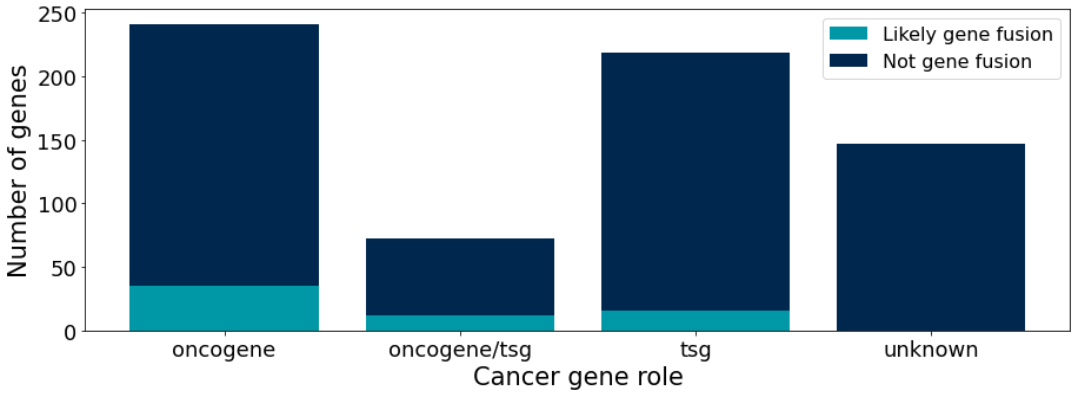
Declaration of interests

N. Hamadouche, N. Goardon, D. Vaur, S. Krieger and L. Castera declare that they have no competing interests. N.Soirat is employed by SeqOne Genomics for the time period October 2020 to present in the context of a public-private PhD project (CIFRE fellowship #2020/0103) partnership between INSERM and SeqOne Genomics. D.Bertrand, A.L. Bougé, M. Broutin, R. Lanos, J. Ruzicka, S. Beaumeunier and N. Philippe are employed by SeqOne.

Declaration of Generative AI and AI-assisted technologies in the writing process

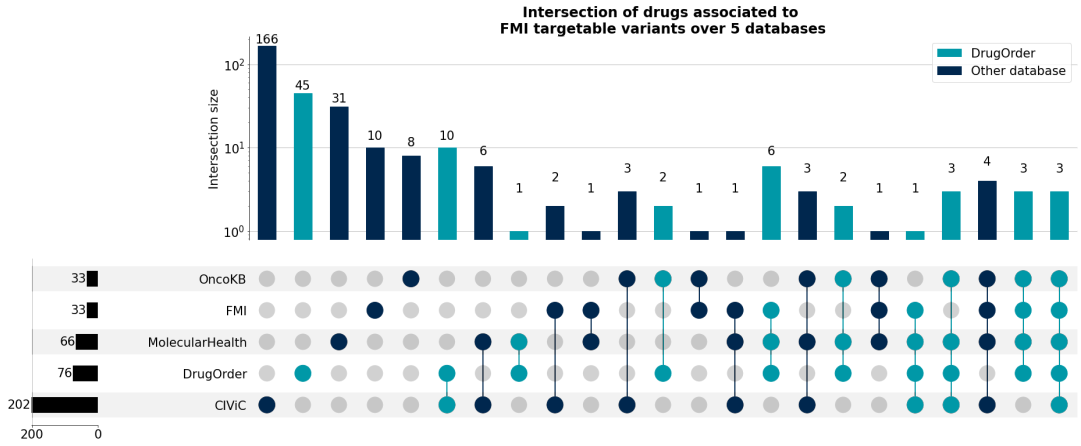
Statement: During the preparation of this work the author(s) used ChatGPT 4 in order to improve readability and language of this manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Supplementary Figures



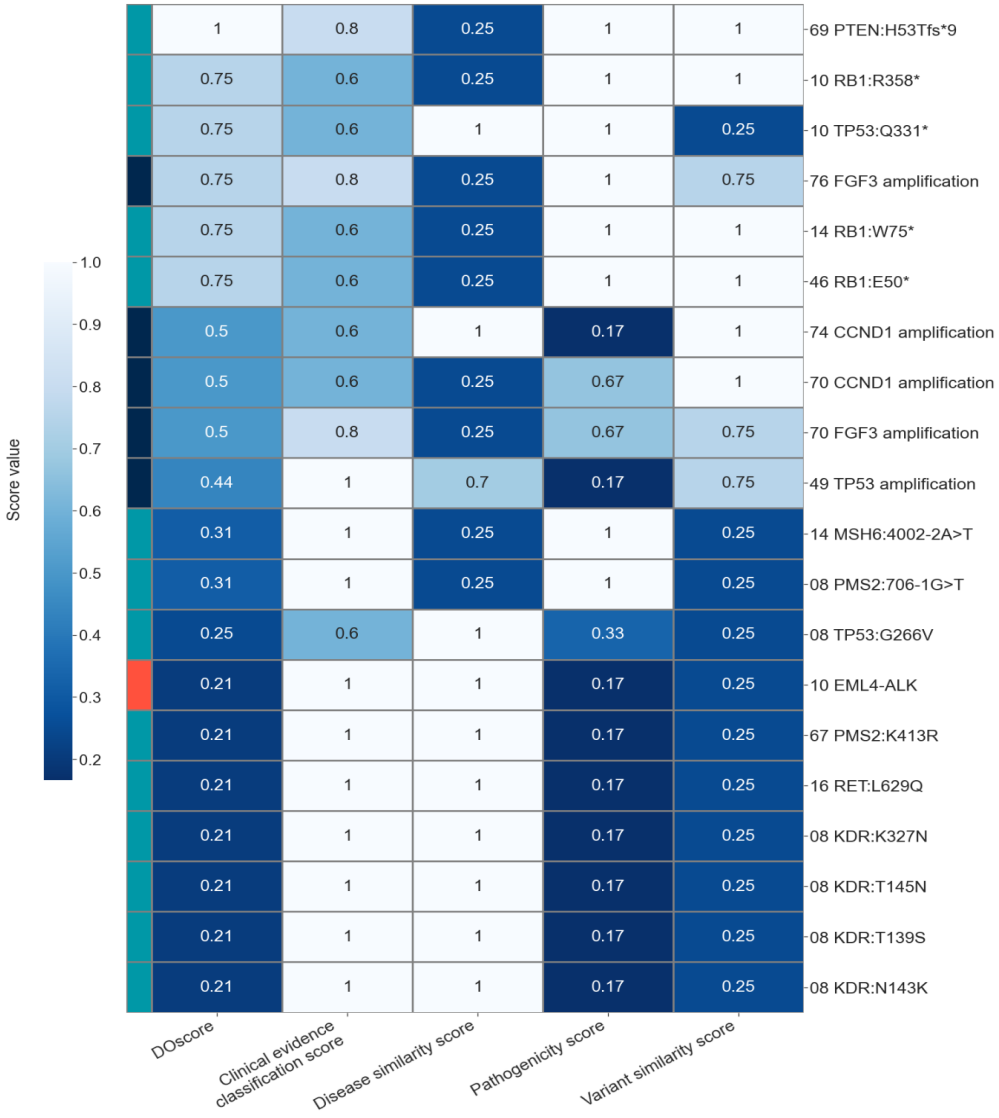
Supplementary Figure S1: Cancer gene role distribution among genes of laboratory panel

We utilized resources such as COSMIC Cancer Gene Census, Vogelstein, Cancermine, ONGene, TCGA, TSGDatabase, JaxCKB, and OncoKB for the annotations of TSG and Oncogenes. It's important to note that a single gene can fall into multiple categories. For the identification of genes implicated in oncogenic fusions, we specifically referred to the annotations provided by the COSMIC Cancer Gene Census. Unknown status is given for genes without any annotation in these databases.



Supplementary Figure S2: Therapeutic database heterogeneity

This figure illustrates the diversity within our therapeutic database, as depicted by an upset plot. The bar chart on the lower left shows the total count of drugs per database. The lower section of the plot represents a specific intersection of databases. The upper section displays the count of drugs that intersect with the database selected in the lower section. Blue bars signify each intersection that involves DrugOrder, while the sea-blue bars represent intersections that exclude DrugOrder.



Supplementary Figure S3: Heatmap of the 20 best DOscored variants on patient without conclusions in clinical reports

The first column on the left illustrates the variant type, with SNV in blue, CNV in sea-blue, and fusion in red. The columns within the heatmap represent the DOscore and its corresponding tiers. DOscore is normalized based on the highest sample DOscore, while framework part tier scores are normalized based on the best framework part tier score. The heatmap is divided into three distinct clusters, each determined by the sample DOscore.

Supplementary Tables

Supplementary Table S1: Description of organs and FMI® supported analysis for the 62 tumors samples

Sample	Supported with FMI® analysis	Organ
RCP01	No	Lung
RCP02	No	Lung
RCP03	No	Lung
RCP04	No	Lung
RCP05	No	Lung
RCP06	No	Lung
RCP07	No	Lung
RCP08	No	Lung
RCP09	No	Lung
RCP10	No	Lung
RCP11	No	Bladder
RCP12	Yes	Bladder
RCP13	Yes	Bladder
RCP14	No	Bladder
RCP15	No	Bladder
RCP16	No	Lung
RCP17	Yes	Bladder
RCP19	Yes	Bladder
RCP20	No	Bladder
RCP21	Yes	Bladder
RCP22	Yes	Bladder
RCP23	Yes	Bladder
RCP24	Yes	Bladder
RCP25	Yes	Bladder
RCP26	No	Endometrium
RCP36	No	Lung
RCP37	No	Lung
RCP38	No	Salivary Glands
RCP39	Yes	Bladder
RCP40	No	Salivary Glands
RCP41	Yes	Bladder
RCP42	Yes	Bladder
RCP43	No	Breast
RCP44	No	Breast
RCP45	Yes	Bladder

RCP46	Yes	Bladder
RCP47	Yes	Bladder
RCP48	No	Thyroid
RCP49	No	Thyroid
RCP50	No	Lung
RCP51	No	Lung
RCP52	No	Lung
RCP53	No	Brain
RCP54	No	Bladder
RCP56	No	Bladder
RCP57	No	Bladder
RCP58	No	Bladder
RCP59	No	Duodenum
RCP60	No	Kidney
RCP61	No	Skin
RCP66	No	Ovary
RCP67	No	Endometrium
RCP68	No	Ovary
RCP69	No	Ovary
RCP70	No	Ovary
RCP71	No	Breast
RCP72	No	Ovary
RCP73	No	Ovary
RCP74	No	Breast
RCP75	No	Ovary
RCP76	No	Ovary
RCP77	No	Breast

Supplementary Table S2: Command line options of each tools used for the sequencing data calling

Tool	Option
GroupReadByUMI	adjacency strategy
CallMolecularConsensus	--cc 3
FilterConsensusRead	-M 1 -N 40
BWA	-M
Freebayes	-F 2 -C 2 --genotype-qualities --pooled-continuous --pooled-discrete
Mutect2	--max-reads-per-alignment-start 0 --max-mnp-distance 3 --pcr-snv-qual 70 --pcr-indel-qual 70
STAR	--twopassMode Basic --alignSplicedMateMapLminOverLmate 0.5 --peOverlapNbasesMin 10 --outFilterMultimapNmax 100 --outFilterMismatchNmax 33 --seedSearchStartLmax 12 --alignSJoverhangMin 15 --outFilterMatchNminOverLread 0 --outFilterScoreMinOverLread 0.3 --alignSJDBoverhangMin 1 --outFilterType BySJout --chimSegmentMin 10 --chimOutType WithinBAM SoftClip --chimJunctionOverhangMin 10 --chimScoreDropMax 30 --chimScoreJunctionNonGTAG 0 --chimScoreSeparation 1 --alignSJstitchMismatchNmax 5 -1 5 5 --chimSegmentReadGapMax 3 --chimMultimapNmax 50
Arriba	-u with blacklist from original github repository
Minimap2	-ax sr

Supplementary Table S3: List of features required for drug prioritization

Database or tool used	Description	Feature
VEP	Effect prediction and description of variants	Variant impact
		Variant coordinates (coordinates, exons, codons…)
Pfam	Protein domain database	Protein domain annotation
COSMIC Fusion	Database of gene fusions involved in cancer	Fusion involvement in cancer
Cancer Gene Census (COSMIC)	Annotation of genes involved in cancer	CNV cancer involvement
Laboratory gene panel annotation (COSMIC Gene Census, Vogelstein, Cancermine, ONGene, TCGA, TSGDatabase, JaxCKB and OncoKB, see Supplementary Figure 4)	Genes of the panel annotated depending on their cancer role	Consensus of Gene Role (oncogene / TSG)
RefSeq	Gene and transcript database	Transcript annotation
		Last exon annotation
Ensembl	Gene and transcript database	Transcript annotation
Custom disease ontology	Cancer ontology	Type of cancer (disease name)
		Name of the organ
		Liquid tumor or solid tumor
ComPerMed CPV list	List of pathogenic variants according to ComPerMed publication	Pathogenic variant annotation
gnomAD	Population database	Average of all population frequency
ClinVar	Clinical database	ClinVar description: (Likely) pathogenic, unknown, (likely) benign
CIViC	Cancer variant database	CIViC description: (Likely) pathogenic, unknown, (likely) benign

Sift	Variant predictor	Impact of the mutation over protein
MutationTaster		
fathmm-MKL		
LRT		
MetaLR		
MetaSVM		
PROVEAN		
FATHMM		
DANN		
MutationAssessor		
MolecularMatch	Therapeutic database	Drug name
		Clinical evidence tier
		Disease name
		Variant coordinates (gene, coordinates, exon and codon)
CIViC actionability	Therapeutic database	Drug name
		Clinical evidence tier
		Disease name
		Variant coordinates (gene, coordinates, exon and codon)

Supplementary Table S4: Filters used for CFB cohort variant interpretation

Filter	Value
Coverage	≥ 30x
Read alternate observation	≥ 2
Quality phred score	≥ Medium
Allele frequency	≥ 4%
Variant frequency in the cohort	≤ 50%
Variant effect (VEP)	missense, intronic, splice polypyrimidine, splice region, 5 prime UTR, stop gained, frameshift, inframe deletion, splice donor region, splice acceptor, start lost, splice donor, stop lost, inframe insertion, stop retained, coding sequence
Relative population frequency	≤ 10^{-3}

Supplementary Table S5: Adapted ComPerMed score table for non-clear LoF mutations

Parameter	Score +2	Score +1	Score +0.5
Mutation tier from COSMIC Mutation Census	1, 2, 3	Other tier	/
In-silico variant damaging prediction tools*	/	/	⅔ of the prediction as damaging
ClinVar pathogenic annotation	/	/	Pathogenic / Likely Pathogenic
Present in CIViC	/	/	Yes

*In silico variant damaging prediction tools used : MutationTaster, fathmm-MKL, LRT, MetaLR, MetaSVM, PROVEAN, FATHMM, DANN, MutationAssessor

Supplementary Table S6: CIViC clinical evidence (B) not identified by DrugOrder

Sample	Variant_Name	HGVS	Drug(s)	CIViC evidence level	Disease	OncoKB level
tumor56	1_H3-3A_226064434_A_T	K28M	Akt/ERK Inhibitor ONC201	B	glioma	None
tumor57	11_GSTP1_67352689_A_G	I105V	FOLFOX Regimen	B	carcinoma of colon	None
tumor99	20_GNAS_57428713_T_C	393T>C	Cisplatin Fluorouracil	B	neoplasm of esophagus	None
tumor196	19_ERCC2_45854919_T_G	K751Q	Paclitaxel Carboplatin	B	non-small cell lung cancer	3A*
tumor239	1_NRAS_115258747_C_T	G12D	Cetuximab	B	carcinoma of colon	R1**
tumor282	19_ERCC2_45854919_T_G	K751Q	Cisplatin	B	neoplasm of bone	3A
tumor302	20_GNAS_57428713_T_C	393T>C	Erlotinib Gefitinib	B	non-small cell lung cancer	None
tumor318	7_ABCB1_87160618_A_T	S893T	Paclitaxel	B	neoplasm of ovary	None
tumor356	7_ABCB1_87138645_A_G	I1145I	Carboplatin Cisplatin	B	non-small cell lung cancer	None

***** Compelling clinical evidence supports the biomarker as being predictive of response to a drug in this indication but neither biomarker and drug are standard of care

****** Standard of care biomarker predictive of resistance to an FDA-approved drug in this indication

Supplementary Table S7: Filters used for FMI tumor interpretation

Filter	Value
Coverage	≥ 10x
VAF	≥ 5%
Allele strand ration	≥ 10/90
Population frequency	≤ 0.01%

References

1. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. Cell Press; 2018;173:371-385.e18.
2. García-Nieto PE, Morrison AJ, Fraser HB. The somatic mutation landscape of the human body. Genome Biol. 2019;20:298.
3. Lin P-C, Yeh Y-M, Hsu H-P, Chan R-H, Lin B-W, Chen P-C, et al. Comprehensively Exploring the Mutational Landscape and Patterns of Genomic Evolution in Hypermutated Cancers. Cancers. 2021;13:4317.
4. Cowppli-Bony A, Uhry Z, Remontet L, Voirin N, Guizard A-V, Trétarre B, et al. Survival of solid cancer patients in France, 1989–2013: a population-based study. Eur J Cancer Prev. 2017;26:461–8.
5. Ross JS, Gay LM. Comprehensive genomic sequencing and the molecular profiles of clinically advanced breast cancer. Pathology (Phila). Elsevier B.V.; 2017;49:120–32.
6. Wagener-Ryczek S, Merkelbach-Bruse S, Siemanowski J. Biomarkers for Homologous Recombination Deficiency in Cancer. J Pers Med. 2021;11:612.
7. Amato M, Franco R, Facchini G, Addeo R, Ciardiello F, Berretta M, et al. Microsatellite Instability: From the Implementation of the Detection to a Prognostic and Predictive Role in Cancers. Int J Mol Sci. Multidisciplinary Digital Publishing Institute; 2022;23:8726.
8. Haynes BC, Blidner RA, Cardwell RD, Zeigler R, Gokul S, Thibert JR, et al. An Integrated Next-Generation Sequencing System for Analyzing DNA Mutations, Gene Fusions, and RNA Expression in Lung Cancer. Transl Oncol. 2019;12:836–45.
9. Christofyllakis K, Bittenbring JT, Thurner L, Ahlgrimm M, Stilgenbauer S, Bewarder M, et al. Cost-effectiveness of precision cancer medicine-current challenges in the use of next generation sequencing for comprehensive tumour genomic profiling and the role of clinical utility frameworks (Review). Mol Clin Oncol. Spandidos Publications; 2022;16:1–4.
10. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32:894–9.
11. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Hum Mutat. 2013;34:E2393–402.
12. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37:235–41.
13. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12:103.
14. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.
15. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980-985.
16. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862-868.
17. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–82.
18. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. Wolters Kluwer; 2017;1–16.

19. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. Nature Publishing Group; 2017;49:170–4.
20. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47:D941–7.
21. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. Nature Publishing Group; 2018;18:696–705.
22. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. J Mol Diagn JMD. 2017;19:4–23.
23. Koeppel F, Muller E, Harlé A, Guien C, Sujobert P, Trabelsi Grati O, et al. Standardisation of pathogenicity classification for somatic alterations in solid tumours and haematologic malignancies. Eur J Cancer. 2021;159:1–15.
24. Trends, Charts, and Maps - ClinicalTrials.gov [Internet]. [cited 2022 Dec 1]. Available from: https://clinicaltrials.gov/ct2/resources/trends
25. Mullard A. 2021 FDA approvals. Nat Rev Drug Discov. 2022;21:83–8.
26. Rieke DT, Lamping M, Schuh M, Tourneau CL, Basté N, Burkard ME, et al. Comparison of Treatment Recommendations by Molecular Tumor Boards Worldwide. https://doi.org/101200/PO1800098. American Society of Clinical Oncology; 2018;1–14.
27. Langenberg KPS, Meister MT, Bakhuizen JJ, Boer JM, Eijkelenburg NKA van, Hulleman E, et al. Implementation of paediatric precision oncology into clinical practice: The Individualized Therapies for Children with cancer program 'iTHER.' Eur J Cancer. Elsevier; 2022;175:311–25.
28. Boichard A, Richard SB, Kurzrock R. The Crossroads of Precision Medicine and Therapeutic Decision-Making: Use of an Analytical Computational Platform to Predict Response to Cancer Treatments. Cancers [Internet]. Multidisciplinary Digital Publishing Institute (MDPI); 2020;12. Available from: /pmc/articles/PMC7017109/
29. Tamborero D, Dienstmann R, Rachid MH, Boekel J, Lopez-Fernandez A, Jonsson M, et al. The Molecular Tumor Board Portal supports clinical decisions and automated reporting for precision oncology. Nat Cancer. Nature Publishing Group; 2022;3:251–61.
30. Pishvaian MJ, Blais EM, Joseph Bender R, Rao S, Boca SM, Chung V, et al. A virtual molecular tumor board to improve efficiency and scalability of delivering precision oncology to physicians and their patients. JAMIA Open. Oxford Academic; 2019;2:505–15.
31. Gaonkar KS, Marini F, Rathi KS, Jain P, Zhu Y, Chimicles NA, et al. annoFuse: an R Package to annotate, prioritize, and interactively explore putative oncogenic RNA fusions. BMC Bioinformatics. 2020;21:577.
32. Paciello G, Ficarra E. FuGePrior: A novel gene fusion prioritization algorithm based on accurate fusion structure analysis in cancer RNA-seq samples. BMC Bioinformatics. 2017;18:58.
33. MolecularMatch database [Internet]. [cited 2022 Dec 1]. Available from: https://www.molecularmatch.com/
34. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31:1023–31.
35. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
36. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
37. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv; 2012 [cited 2023 Oct 4]. Available from: http://arxiv.org/abs/1207.3907
38. Geraldine A Van der Auwera BDO. Genomics in the Cloud [Internet]. O'Reilly Media, Inc; 2020 [cited 2022 Dec 7]. Available from: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

39. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2 [Internet]. bioRxiv; 2019 [cited 2023 Sep 18]. page 861054. Available from: https://www.biorxiv.org/content/10.1101/861054v1
40. fgbio [Internet]. Fulcrum Genomics; 2022 [cited 2022 Dec 7]. Available from: https://github.com/fulcrumgenomics/fgbio
41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
42. Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021;31:448–60.
43. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339:1546–58.
44. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods. Nature Publishing Group; 2019;16:505–7.
45. Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. J Genet Genomics. 2017;44:119–21.
46. Zhao M, Sun J, Zhao Z. TSGene: a web resource for tumor suppressor genes. Nucleic Acids Res. 2013;41:D970–6.
47. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. Nature Publishing Group; 2020;20:555–72.
48. Patterson SE, Liu R, Statz CM, Durkin D, Lakshminarayana A, Mockus SM. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genomics. 2016;10:4.
49. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
50. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes [Internet]. bioRxiv; 2022 [cited 2023 Sep 4]. page 2022.03.20.485034. Available from: https://www.biorxiv.org/content/10.1101/2022.03.20.485034v2
51. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51:D933–41.
52. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9.
53. Froyen G, Mercier ML, Lierman E, Vandepoele K, Nollet F, Boone E, et al. Standardization of somatic variant classifications in solid and haematological tumours by a two-level approach of biological and clinical classes: An initiative of the belgian compermed expert panel. Cancers. 2019;11.
54. Wagner AH, Walsh B, Mayfield G, Tamborero D, Sonkin D, Krysiak K, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet. Nature Publishing Group; 2020;52:448–57.
55. Liao M, Zhou J, Wride K, Lepley D, Cameron T, Sale M, et al. Population Pharmacokinetic Modeling of Lucitanib in Patients with Advanced Cancer. Eur J Drug Metab Pharmacokinet. 2022;47:711–23.
56. Choudhury AD, Higano CS, de Bono JS, Cook N, Rathkopf DE, Wisinski KB, et al. A Phase I Study Investigating AZD8186, a Potent and Selective Inhibitor of PI3Kβ/δ, in Patients with Advanced Solid Tumors. Clin Cancer Res. 2022;28:2257–69.
57. Hernandez-Boussard T, Rodriguez-Tome P, Montesano R, Hainaut P. IARC p53 mutation database: a relational database to compile and analyze p53 mutations in human tumors and cell lines. International Agency for Research on Cancer. Hum Mutat. 1999;14:1–8.
58. Cline MS, Liao RG, Parsons MT, Paten B, Alquaddoomi F, Antoniou A, et al. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2.

PLoS Genet. 2018;14:e1007752.
59. Huey RW, Shah AT, Reddi HV, Dasari P, Topham JT, Hwang H, et al. Feasibility and value of genomic profiling in cancer of unknown primary: real-world evidence from prospective profiling study. J Natl Cancer Inst. 2023;115:994–7.

4. Détection et annotation d'isoformes aberrantes à l'aide du séquençage de troisième génération

4.1 Contexte du développement de LoRID

La résolution de la structure complète des isoformes des transcrits des gènes de prédisposition aux cancers chez un patient pourrait devenir un élément du diagnostic moléculaire. Les variants génomiques peuvent avoir un effet sur l'épissage alternatif de l'ARNm. Les études du transcriptome ont permis de corréler la présence des variations génomiques avec l'apparition de transcrits alternatifs aberrants, impliqués dans des cas de prédispositions au cancer [190, 191, 192]. La résolution de l'impact fonctionnel de ces variants génomiques a permis de placer l'étude des isoformes de l'ARNm dans l'interprétation des cas cliniques [193, 194, 195]. Ainsi, l'étude du transcriptome pourrait permettre l'identification de nouveaux événements, induits par des anomalies génomiques cryptiques transmissibles. L'étude quantitative et qualitative du profil d'expression d'un individu peut donc mettre en évidence des isoformes aberrantes, pouvant expliquer des cas jusqu'alors non résolus, résolvant une partie de l'hérédité manquante.

La représentation complète des transcrits alternatifs de l'ARNm nécessite un assemblage (ou reconstruction) des différentes structures du transcrit. Les techniques de séquençage en *short-reads* ont montré leur capacité à représenter partiellement la structure d'un transcrit [196]. L'émergence des techniques de séquençage de troisième génération a permis d'utiliser des *long-reads* dont la taille correspond à l'entièreté d'un transcrit, le transcrit est alors dit *full-length*. L'assemblage des transcrits se fait

alors à partir de *reads* pouvant décrire l'ensemble des épissages qui aboutissent à une conformation de transcrit.

L'annotation de transcrits reste une étape majeure, afin d'identifier des transcrits physiologiques, aberrants ou inconnus. Les bases de données comme Ensembl [197], RefSeq [163] ou Gencode [146] regroupent les transcrits physiologiques alternatifs des gènes. L'utilisation de ces bases permet l'annotation des transcrits et de choisir un transcrit de référence. Des valeurs d'expression au sein de la cellule permettent d'estimer leur proportion par rapport à un transcrit majoritairement attendu dans le tissu. Les transcrits absents de ces bases nécessitent également d'être annotés afin de décrire leur différence vis-à-vis d'un transcrit de référence. Toutefois, l'étape d'annotation nécessite au préalable de pouvoir identifier l'ensemble des transcrits d'un même gène. Cette identification est rendue complexe pour les gènes faiblement exprimés. Notre travail, présenté sous forme de la publication suivante, décrit un protocole d'enrichissement des messagers d'un panel de gènes ainsi que le développement d'un outil appelé *Long Read Isoform Discovery* (LoRID). L'enrichissement des messagers permet une forte profondeur de séquençage sur les différentes régions d'intérêt d'un panel. Ainsi, la détection de messagers faiblement exprimés est rendue possible, permettant de compléter le profil transcriptionnel d'un patient ainsi que l'interprétation de transcrits minoritaires.

Aussi, établir des profils d'expression pourrait permettre d'identifier un transcrit aberrant inconnu des bases de référence ou une modification des expressions relatives entre les transcrits qui serait spécifique d'une prédisposition aux cancers.

LoRID est un pipeline bioinformatique d'identification et d'annotation des transcrits alternatifs. Il dispose de deux modules, le premier réalisant l'assemblage des transcrits à l'aide de StringTie, le second annotant les transcrits assemblés à l'aide d'un outil développé pour cette étude, iSofOrmS annoTAtoR (SOSTAR). SOSTAR propose notamment une nomenclature décrivant l'ensemble des épissages produits au sein de l'isoforme, produisant alors une description complète de l'événement, en comparaison avec un transcrit de référence.

L'évaluation des performances de LoRID a été réalisée au moyen de 8 contrôles dont 4 positifs au syndrome et 4 contrôles négatifs. Les 4 contrôles positifs correspondent à des variants identifiés chez des patients présentant un syndrome HBOC, impactant l'épissage. LoRID a également permis d'identifier un cas d'hérédité non expliquée, jusqu'alors au niveau moléculaire, au travers de l'identification d'un SVA retrotransposon au sein du gène *BRCA1*.

Dans le cadre de cette thèse, en tant que second auteur, mon travail s'est concentré sur la veille bibliographique et technique des méthodes existantes dans l'identification et la quantification de transcrits de l'ARNm. Mon travail a permis la mise en place du pipeline complet, en Snakemake, ainsi que la conteneurisation du pipeline au travers de Docker et Singularity, complétant la veille scientifique, après avoir choisi les composantes du pipeline. Le module SOSTAR a été développé par le premier auteur. Enfin, j'ai mené le traitement des données, à l'exception de la bioanalyse menée par le premier auteur et participé à l'interprétation des résultats. Ce travail a été le fruit de nombreuses itérations décidées conjointement avec le premier auteur, permettant l'amélioration des performances de LoRID et l'optimisation du pipeline.

4.2 Cartographie fine de la diversité des isoformes d'ARN à l'aide d'un protocole innovant de séquençage ciblé de l'ARN en *long-reads* et d'un nouveau pipeline bioinformatique dédié

Le présent travail est en cours de soumission auprès de l'*American Journal of Human Genetics*. Toutes les figures, tables et données supplémentaires sont disponibles dans le répertoire GitHub suivant : `https://github.com/Nedss/these-data` dans le répertoire LoRID.

Fine mapping of RNA isoform diversity using an innovative targeted long-read RNA sequencing protocol with novel dedicated bioinformatics pipeline

Camille Aucouturier,[1,2,3] Nicolas Soirat,[2,3,4] Laurent Castéra,[1,2] Denis Bertrand,[4] Alexandre Atkinson,[1] Thibaut Lavolé,[1] Nicolas Goardon,[1,2] Céline Quesnelle,[1] Julien Levilly,[1] Sosthène Barbachou,[1] Angelina Legros,[1] Agathe Ricou,[1,2] Flavie Boulouard,[1,2] Dominique Vaur,[1,2] Sophie Krieger,[1,2,3†] Raphael Leman[1,2†*]

1. Laboratoire de biologie et de génétique du cancer, Centre François Baclesse, Caen, 14000, France
2. Inserm U1245, Cancer and Brain Genomics, FHU G4 Genomics, Normandie University, Rouen, 76183, France
3. Normandie Univ, UNICAEN, Caen, 14000, France
4. SeqOne Genomics, Montpellier, 34000, France

†These authors contributed equally

*Correspondence: r.leman@baclesse.unicancer.fr

Abstract

Solving the structure of mRNA transcripts is a major challenge for both research and molecular diagnostic purposes. The emergence of third generation sequencing can fulfill this challenge. However, genes with low expression levels are difficult to study with the whole transcriptome sequencing approach. The use of RT-PCR long range is then required, but can lead to potential PCR bias. To fix these technical limitations, we propose a novel method to capture transcripts of a gene panel using a targeted enrichment approach suitable for Pacific Biosciences and Oxford Nanopore Technologies platforms. We designed a set of probes to capture transcripts of a panel of genes involved in hereditary breast and ovarian cancer syndrome. We present LoRID (Long Read Isoform

Discovery), a versatile pipeline to assemble and annotate isoforms from long read sequencing using a new tool called SOSTAR (iSofOrmS annoTAtoR). The reliability of the data was verified on a collection of negative and positive RNA control samples derived from lymphoblastoid cell lines on *BRCA1* and *BRCA2* genes. Using the LoRID pipeline, a case of unexplained inheritance in a family with a history of breast and ovarian cancer was solved by identifying an SVA retrotransposon in the *BRCA1* gene.

Introduction

Splicing of pre-mRNAs is a major source of transcript diversity. Also known as alternative splicing, this mechanism concerns at least 90% of multi-cassette genes (1). Three main mechanisms lead to this transcripts diversity: the exon skipping, the use of alternative splice sites and the complete intron retention (Figure S1). As an example for the human gene *KCNMA1*, more than 500 transcripts were described (2,3). The knowledge of this diversity is a crucial step to understand and explore physiological and pathological processes (4–6). Moreover, a variety of genomic variations could affect RNA splicing. Indeed, 3.8% of genomic variations from Exome Aggregation Consortium (ExAC) had an impact on splicing (7). The study of RNA transcripts also resolves the functional impact of genomic variation in the context of inheritance disease (8–10). In addition, alternative splicing could interfere with the clinical interpretation of genomic variations (11–13).

Over the past decade, several new methods based on short read sequencing (SRS) partially dealt with these challenges. High-throughput RNA sequencing (14), in particular targeted RNA sequencing (Davy *et al.*, 2017) and high-throughput minigene splicing assay (7,16) have been developed and widely used. Despite the advantage of these technologies, they are limited in exploring the complete structure of transcript isoforms (17). Then, the development of long read sequencing (LRS), also known as third generation sequencing, recently proven its interest in describing the structure of isoforms (18). The two main technologies to perform this sequencing are provided by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). Both platforms are based on single molecule sequencing but differ in nucleotide detection. Briefly, ONT sequencing platform uses

variation of ion torrent in pores within an impermeable membrane, while PacBio platform synthetizes the complementary DNA strand with a fluorescent-labeled nucleotide.

RNA long read sequencing by ONT or PacBio platforms was performed either on direct RNA/mRNA (19) or total cDNA. However, this approach does not provide a comprehensive isoforms collection, especially for genes with low expression levels. The second known approach used RT-PCR long range amplicons (20,21) but limits targeting to one or a few genes with the potential addition of a major PCR bias. Recent methods using hybridization RNA capture combined with long read sequencing were developed (22–24).

Therefore, we described a new protocol of target enrichment by probes for RNA long read sequencing compatible with both ONT and PacBio sequencing platforms (Figure 1A). In this study, we designed probes to capture transcripts from a panel of genes involved in hereditary breast and ovarian cancer (HBOC) syndrome. Indeed, genes involved in this syndrome were well known, notably *BRCA1* (NM_007294, OMIM#113705) and *BRCA2* (NM_000059, OMIM#600185). In addition, we developed and evaluated a dedicated bioinformatics pipeline named LoRID for “Long Read Isoform Discovery”, to assemble isoforms from ONT sequencing. LoRID provides options for annotating isoforms regardless of the assembly method used (Figure 1B). We validated our protocol by sequencing negative controls from healthy donors, and positive controls from patients carrying *BRCA1* spliceogenic variants.

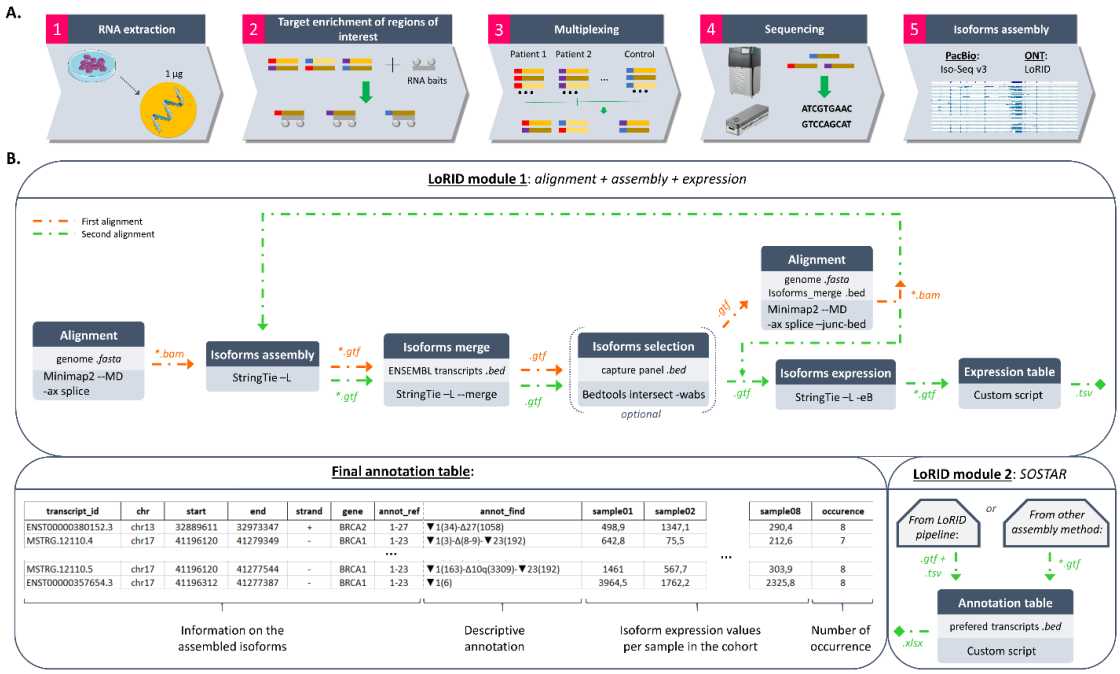


Figure 1: Targeted long read RNA sequencing workflow. A. Overview of the sequencing protocol from cell lines to isoform assembly B. Description of the LoRID pipeline.

Material and Methods

Sample collection

Lymphoblastoid cell lines from 8 patients having signed consent were established for this work. Four samples were negative controls, *i.e.* healthy donors, without any molecular abnormalities. These healthy donor's samples were collected during the clinical protocol CASOHAR (*CAncer du Sein et/ou de l'Ovaire Héréditaire – ARN*, N°ID-RCB 2015-A00598-41). Two patients were positive controls: one patient carries a duplication of *BRCA1* exon 7, and the other an intronic genomic deletion in *BRCA1* intron 14. Several criteria were defined to select highly predisposed families for long read RNA sequencing to investigate unexplained hereditary causes (Table S1). We selected a family with cancer according to these criteria to explore the unexplained inheritance.

Probes design

A custom panel of 28 genes (Table S2) was designed using Integrated DNA Technologies (IDT) xGen Lockdown probes. Transcript structures and sequences described in the RefSeq database (25) were used for the design. In the first step, probes positions were automatically set with a tailing from 0.1x to 1x by 0.1 step according to transcriptomic sequence of the gene panel (script available on request). Probes overlapping between two contiguous exons were removed to reduce the risk of over selection of a particular isoform. As for some gene of our panel had short transcript sequence, we arbitrarily set a minimal number of 5 probes per gene. Thus, in the second step, probes for genes with low probes number according to this threshold were added. These new probes were from previous designs with a higher rate of tailing. Then, the GC percent and the probe sequence specificity to the target were checked according to the IDT recommendations. Finally, a set of 367 probes was used to capture our panel, with for example ten probes for *BRCA1* gene (Figure S2).

Library preparation

Specifically, for PacBio platform, the AMPure beads (Beckman, #A63881) required a washing according to the following protocol. From 500 µL of AMPure beads, the supernatant was saved after beads precipitation on magnetic rack. The beads were washed 5 times with 1 mL of nuclease free water plus one more wash with 1 mL of Qiagen elution buffer (Qiagen, #19086). Then, they were resuspended with the saved supernatant.

RNAs were extracted from cells using the RNeasy Plus Mini Kit (Qiagen, #74134) according to the manufacturer's instructions. All samples had a RNA Integrity Number (RIN) above 9.

First strand cDNA synthesis was carried out with the SMARTer PCR cDNA Synthesis Kit (Clontech #634925) according to the manufacturer's protocol, from 1 µg of total RNA. The resulting first strand product was diluted to 150 µL of H_2O and used for large scale PCR.

For each sample, sixteen PCR reactions were performed using the PrimeSTAR GXL DNA Polymerase (Clontech #R050A). Thermal cycling conditions were 98°C for 30 seconds, followed by 12 cycles of 98°C for 10 seconds, 65°C for 15 seconds, and 68°C for 10 minutes, and a final extension of 68°C for 5

minutes. Amplicons were separated in 2 fractions to perform purification. Fraction 1 was purified twice using 1X washed AMPure beads and fraction 2 once using 0.4X washed AMPure beads. The two fractions were quantified using the 2200 TapeStation system (Agilent) and then pooled to obtain an equimolar pool of 1 µg of DNA.

The libraries were hybridized according to the IDT manufacturer's recommendations using the specific blockers (5' AAG CAG TGG TAT CAA CGC AGA GTA C 3') and (5' TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT / 3InvdT/3') at 1 mM.

Two PCR reactions were performed using Takara LA Taq DNA Polymerase Hot-Start version (Clontech, #RR042A) for each sample. Thermal cycling conditions were 95°C for 2 minutes, followed by 14 cycles of 95°C for 20 seconds, and 68°C for 10 minutes, and a final extension of 72°C for 10 minutes. Amplicons were then purified using 1X washed AMPure beads.

After DNA repair, barcodes were ligated using the SMRTbell Library Construction and Sequencing kits (PacBio) in accordance with the manufacturer's protocol (PN 101-070-200 Version 05 [November 2017]). SMRTbell libraries were loaded on Sequel II to get the subreads sequencing.

End repair and dA tailing were performed using the NEBNext Ultra™ End Repair/dA-Tailing module (New England BioLabs #E7546S). Barcoding was performed with the SQK-LSK109 ligation sequencing kit (ONT) in accordance with Nanopore community protocol. Samples were washed twice with Short Fragment Buffer (ONT) before a final elution in 15 µL of elution buffer (ONT). Samples were quantified using Qubit® Fluorometer (ThermoFisher Scientific).

ONT sequencing was performed on the MinIon Mk1B device and PacBio sequencing on a Sequel II device. Prepared libraries were loaded on MinION flow cell (R9.4.1) according to instructions. MinKNOW software (v21.06.0) was used for running the MinION sequencer during 64 hours. Additional flush buffer was added when the number of pores being used decreased.

Targeted RNA-seq short read

Splicing junctions detected by LRS were checked using targeted RNA-seq short read from the same RNA extraction. The four negative control samples and the two probands were sequenced on the same gene panel using this approach according to the protocol described in (Davy *et al.*, 2017). The data was analyzed by SpliceLauncher tool (27). Only the splicing junctions supported by at least 10 reads were retained for further analyses.

Bioinformatics analysis

PacBio Iso-Seq

Highly accurate long reads (HiFi reads) were generated from PacBio sequencing using the Iso-Seq v3 pipeline (https://github.com/PacificBiosciences/IsoSeq) with default parameters. HiFi reads were aligned against version 37 of human reference genome assembly (GRCh37). Only high quality isoforms were considered for analyses.

ONT LoRID

ONT basecalling and data demultiplexing were performed using Guppy tool (v5.0.11) with default options. These data were then processed by the LoRID pipeline. LoRID is available on GitHub (https://github.com/LBGC-CFB/LoRID), and was divided into two modules that can be run separately.

The first module aligned ONT data, assembled isoforms and computed isoforms expression. Alignment to the human reference genome assembly hg19 (GRCh37) was first performed using Minimap2 (28) (v2.24) in the splice mode with the options "-MD --ax splice". Isoforms were assembled using StringTie tool (29) (v2.0). From the bam files, StringTie was first used in the assembly mode to produce a GTF file of all potential isoforms per sample. These files were then merged using the StringTie "merge mode" to build a reference of all potential isoforms detected in the patient cohort. During this merge step, isoforms were annotated using ENSEMBL annotation. Using bedtools intersect (v2.30.0), this merge file was then filtered on the 28-gene panel with "-S" option to force strand specificity. A second round of alignment was then performed using Minimap2 with the same

options as before, adding the "--junc-bed" option with the isoforms from the merge file produced earlier. Using the new alignment files, isoforms were assembled as described above up to the new merge file. Finally, isoforms expression metrics per patient were calculated using the StringTie expression mode. Expression metrics obtained were coverage, Fragments Per Kilobase Million (FPKM) and Transcripts Per Million (TPM).

The second module named “SOSTAR” (iSofOrmS annoTAtoR) provided a descriptive and comprehensive annotation of each isoform structure. This module is compatible with any long read technology and any assembly method. Isoforms were described relative to reference transcripts (Table S2) by an annotation including only the alternative splicing events. The alternative splicing events were annotated according to Lopez and coll (13). This nomenclature uses “Δ” to refer to a skipping of a reference exon, “▼” an inclusion of a reference intron, “p” a shift of an acceptor site and “q” a shift of a donor site. For partial skipping or inclusion, the number of skipped or retained nucleotides is indicated between brackets (Figure S1). Relative positions are indicated between square brackets. “Exo” refers to an exonization of an intronic sequence, and “int” to an intronization of an exonic sequence with the relative intron or exon number in front. SOSTAR used systematic numbering for exon naming. A final table of all annotated isoforms was generated. This table allowed different levels of filtering, including isoform ID, isoform coordinates, gene name, reference annotation, descriptive annotation, expression values per patient and number of occurrences in the patient cohort.

Data Analysis

The aligned data were visualized with Integrative Genomics Viewer (IGV) software (30). Depth coverage of the 8 patients was calculated using featureCounts tool (v2.0.6) for the 28 captured genes with “-L” option to count the long reads. This tool was used with a reference human genome assembly version 37 (GRCh37) file downloaded from GENCODE (https://www.gencodegenes.org/human/release_19.html) and filtered on the 28 captured genes.

Percentages of on and off target were calculated using featureCount results. On target was calculated by the number of aligned reads assigned to the 28 genes out of the total number of aligned reads. Using the expression module of StringTie tool, expression values per isoform were calculated on the isoforms assembled by LoRID pipeline. In the patient cohort, the splicing junction expressions were calculated using formula (1) for LRS or formula (2) for SRS:

$$Expression_{i^{th}junction} = \sum_{j=0}^{n_{sample}} \left(\sum_{k=0}^{n_{isoforms}} read\_count_{ijk} \right) \quad (1)$$

$$Expression_{i^{th}junction} = \sum_{j=0}^{n_{sample}} \left(read\ count_{ij} \right) \quad (2)$$

Results

Raw sequencing data

Our library preparation generated 40.16 ng/µL of library, sufficient to perform both PacBio and ONT sequencing. Average length of library fragments were 2 kb with a range length from a few hundreds nucleotides (nt) to over 50,000 nt. From PacBio platform a total of 9 Gb of data were generated, representing 5M Circular Consensus Sequencing (CCS) reads. ONT platform generated 27 Gb of 10M reads.

An example of IsoSeq result for *RAD51C* gene (NM_058216, OMIM#602774) revealed the wide variety of transcript isoforms, compared to transcripts described in RefSeq or Ensembl databases (Figure 2A). Read count for the 28 genes was correlated between the two sequencing platforms with a correlation coefficient of 0.8419 (Figure 2B). Targeting enrichment was less efficient for *STK11* and *XRCC3* genes, with an average read count of 49 reads [PacBio: 29 reads; ONT: 68 reads] and 317 reads [PacBio: 184 reads; ONT: 450 reads] respectively. On the gene panel, average reads count per gene was 1,648 reads for PacBio and 29,624 reads for ONT per sample. For both technologies, *PTEN* and

NBN represented the most covered genes with an average read count of 106,570 reads [PacBio: 6,217 reads; ONT: 206,923 reads] and 42,669 reads [PacBio: 4,546 reads; ONT: 80,791 reads], respectively.

On target rate estimated from read count was similar between samples (Figure 2C) for both platforms. While PacBio sequencing reached an average on target of 55%, ONT sequencing achieved 70% of on target.

IsoSeq and LoRID pipelines assembled a total of 138,956 and 1,170 isoforms respectively in selected genes. Isoforms length distribution was investigated between these two pipelines (Figure 2D). IsoSeq-assembled isoforms averaged 2kb in length, while LoRID supported the longest isoforms. Indeed, isoforms detected by IsoSeq were not longer than 8,000 nt instead of 12,000 nt for LoRID. The limit of 12,000 nt represented the longest full length transcript of our panel for *ATM* gene (12,915 nt).

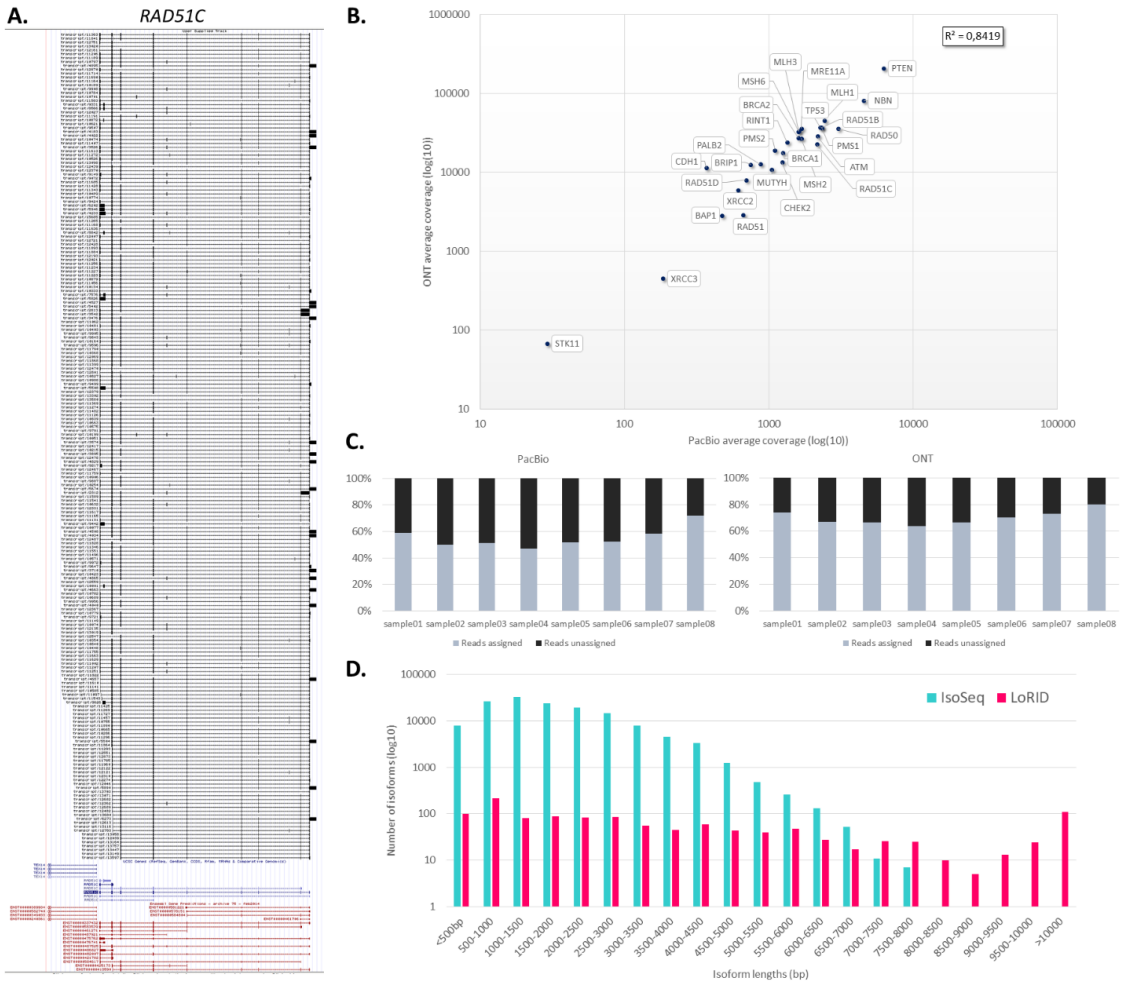


Figure 2: Overview of the results generated by targeted long read sequencing. A. RAD51C isoforms assembled by LRS (black) compared to isoforms described in Ensembl (red) and Refseq (blue) databases B. Average coverage calculated for the 28 genes in the patient cohort. Values are plotted on a logarithmic scale C. Percentage of on and off target for the 28 genes in the patient cohort D. Distribution of isoform lengths assembled by the long read pipelines.

Alternative splicing of BRCA1 and BRCA2 genes

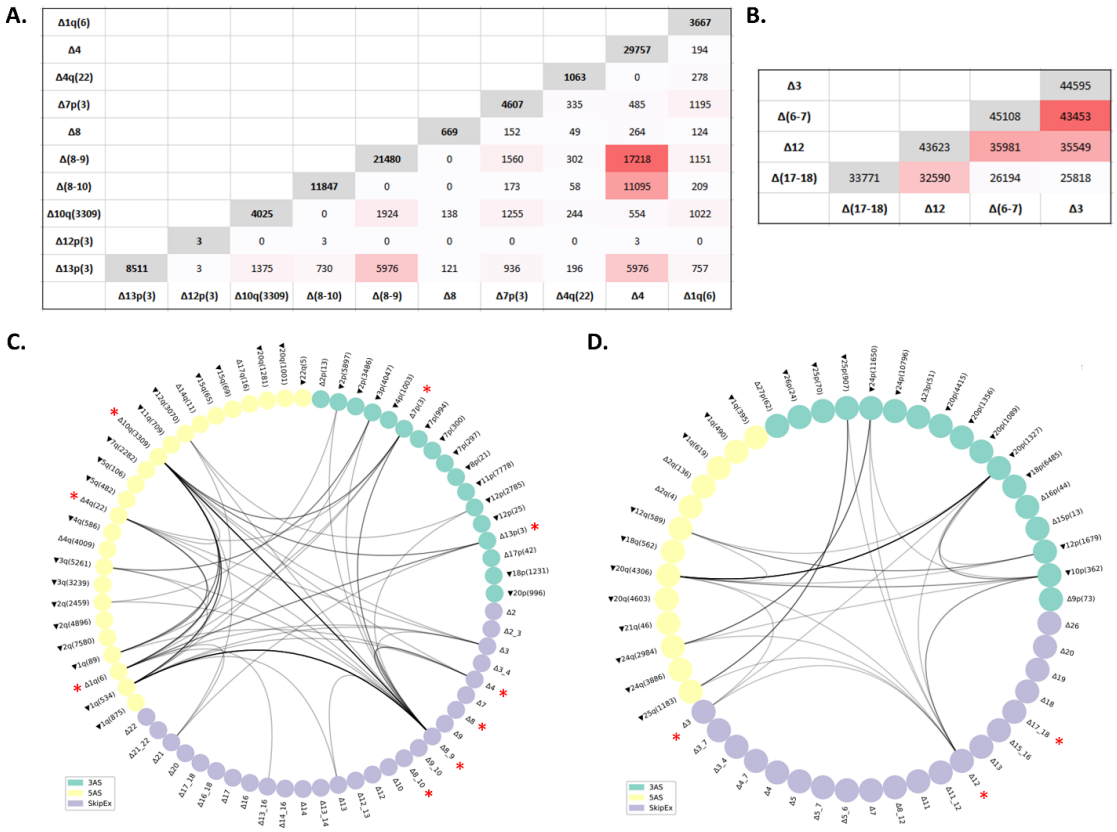
Alternatives splicing events of *BRCA1* (NM_007294.4) and *BRCA2 (*NM_000059.3*)* genes were explored on the aligned long read to validate the data obtained by our protocol.

Previously, Colombo and coll (31) described 10 predominant *BRCA1* splicing events. The in-frame events: Δ1q(6), Δ4, Δ7p(3), Δ(8-9), Δ(8-10), Δ10q(3309), Δ12p(3), Δ13p(3) and the out-of-frame events: Δ4q(22), Δ8. All of these *BRCA1* events were found in our data. The major alternative splicing event found was the Δ4. Except the Δ12p(3) which is missed, the two out-of-frame splicing events were the least represented of all splicing events. Our results showed that these events co-occur (Figure 3A). Interestingly, all pairwise combinations of these splicing events are possible, except when the 2 events are physically incompatible *i.e.* Δ(8-10) and Δ10q(3309). Pairwise combinations of Δ4 and Δ(8-9) represented the majority of all possible pairwise. These splicing event combinations were in-frame.

For *BRCA2* gene, four predominant alternative splicing events were described by Fackenthal and coll (32): the Δ3, Δ(6-7), Δ12 in-frame events and the Δ(17-18) out-of-frame event. As observed in *BRCA1*, all pairwise combinations were found in the long reads (Figure 3B). The Δ(6-7) splicing event was mostly found, and combinations of Δ3 and Δ(6-7) represented the major pairwise combinations. This combination was out-of-frame.

Among the alternatives junctions detected by SRS, the five most frequent combinations within *BRCA1* gene [▼1q(534)/Δ(8_9); Δ(8_9)/Δ10q(3309); Δ1q(6)/Δ10q(3309); ▼1q(534)/Δ10q(3309); Δ7p(3)/Δ(8_9)] represented 29.69% (76/256) of total junction combinations. In addition, other major event were found: ▼1q(89) (Figure 3C). This event was mainly observed with the alternative junctions previously described by Colombo and coll (31). Several pseudo-exons were also identified in intron 3 (▼3p(4047) and ▼3q(5261) corresponding to an exon of 116 nt) and in intron 12 (▼12p(2785) and ▼12q(3070) corresponding to an exon of 66 nt).

For *BRCA2* gene, ▼20p(1327), ▼20q(4306); ▼25p(907), ▼25q(1183) and ▼24p(11650), ▼24q(2984), junctions were over represented (Figure 3D). These junctions corresponded to pseudo-exons creation in intron 20 (64 nt), intron 25 (126 nt) and in intron 24 (91 nt), respectively.



A.	Δ13p(3)	Δ12p(3)	Δ10q(3309)	Δ(8-10)	Δ(8-9)	Δ8	Δ7p(3)	Δ4q(22)	Δ4	Δ1q(6)
Δ1q(6)										3667
Δ4									29757	194
Δ4q(22)								1063	0	278
Δ7p(3)							4607	335	485	1195
Δ8						669	152	49	264	124
Δ(8-9)					21480	0	1560	302	17218	1151
Δ(8-10)				11847	0	0	173	58	11095	209
Δ10q(3309)			4025	0	1924	138	1255	244	554	1022
Δ12p(3)		3	0	3	0	0	0	0	3	0
Δ13p(3)	8511	3	1375	730	5976	121	936	196	5976	757

B.	Δ(17-18)	Δ12	Δ(6-7)	Δ3
Δ3				44595
Δ(6-7)			45108	43453
Δ12		43623	35981	35549
Δ(17-18)	33771	32590	26194	25818

Figure 3: Investigation of alternative splicing events in *BRCA1* and *BRCA2* genes. A. Matrix of pairwise combinations of *BRCA1* splicing events within the long reads B. Matrix of pairwise combinations of *BRCA2* splicing events counts within the long reads C. Circos plot of common junctions between SR and LR junctions for *BRCA1* gene D. Circos plot of common junctions between SR and LR junctions for *BRCA2* gene. Values are plotted on a logarithmic scale. The * represents the splicing events previously described by (Colombo et al., 2014) for *BRCA1* gene and (Fackenthal et al., 2016) for *BRCA2* gene.

Comparison of splicing junctions

Splicing junctions from both SRS and LRS (IsoSeq and LoRID) were compared for the 28 genes of the panel. A total of 2,739 junctions were found from SRS, compared to 8,464 (PacBio) and 1,195 (LoRID) junctions from LRS (Figure 4A).

Among junctions observed by SRS, 987 junctions were not detected by LRS but were supported by a lower average read count per sample (15.65 [2.75; 738.25]) compared with the 1,753 junctions identified by at least one long-read pipeline (10,275.46 [2.75; 538,850.5]). Among these common SRS-LRS junctions, LoRID mostly detected junctions supported by a high average read count (Figure 4B). While PacBio mainly detected junctions with a low average read count.

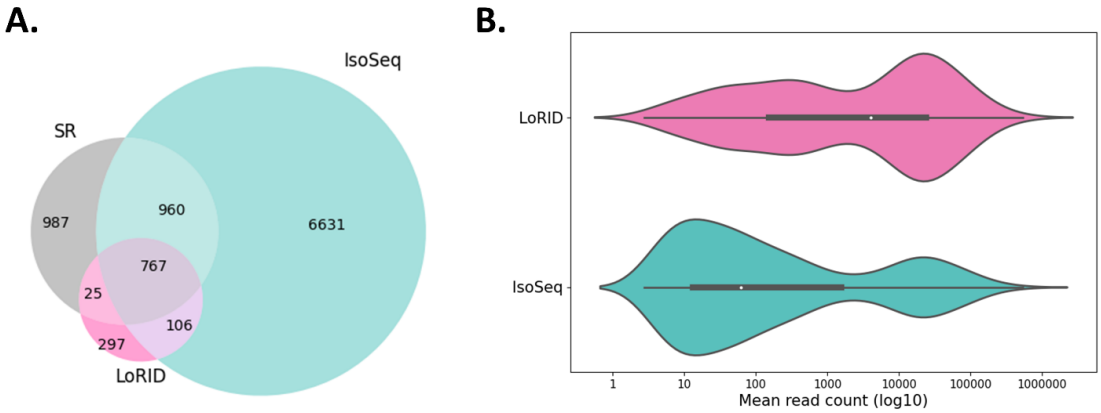


Figure 4: Comparison of splicing junctions between SRS and LRS for the whole gene panel. A. Venn diagram of splicing junctions detected by SRS and LRS B. Violin plot of mean read counts of common junctions between SRS and LRS junctions. Values are plotted on a logarithmic scale.

Common junctions between SRS and the LRS pipelines represented 767 splicing events. Among these common junctions, physiological junctions were the most represented.

Comparison of the expression values of common junctions between SRS and LRS

Correlation between expression values of the common splicing junctions from LRS (LoRID isoforms) and SRS was investigated. Expression values were correlated for the 28 genes panel (r = 0.724, p-value <0.001) between LRS and SRS (Figure 5).

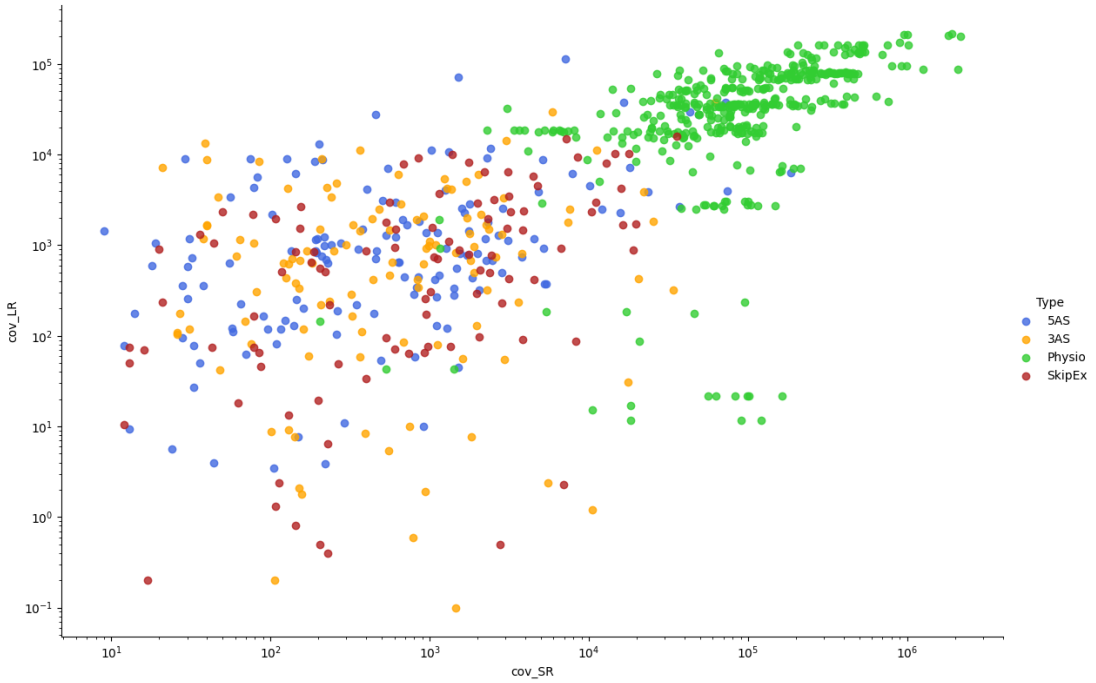


Figure 5: Expression values of common splicing junctions between short and long read sequencing. Values are plotted on a logarithmic scale. Junction are represented by types: 5AS=alternative 5' splice site / 3AS=alternative 3' splice site / Physio=physiological junction / SkipEx=exon skipping.

Detection of aberrant isoforms

Both technologies allowed the detection of the two positive controls first identified by DNAseq. They also allowed the characterization of the effects at the transcript level. The first control carried a 28 base genomic deletion of intron 14 of *BRCA1* (c.4676-31_4676-4del). Bioinformatics predictions of SPiP (33) were in favor of an abolition of the acceptor splice site. The second control was a duplication of exon 7 of *BRCA1* gene.

The BAM alignment files from LRS of these two positive controls were loaded into the IGV software to visualise the events. An intronic retention, due to the use of a new acceptor splice site 64 nt upstream the natural splice site in intron 14 was observed. The intronic retention embedded the genomic deletion (Figure 6A). This event was out-of-frame (p.Gly1560Tyrfs*5).

For the second patient, different insert sizes, corresponding to the exon 7 sequence, were observed at the end of exon 7 (Figure 6E). Several long reads were observed supporting both the exon 7 duplication and other proximal exons (exon 6 or exon 8). The duplicated exon was spliced normally. Then, we were able to assert that this duplication was a tandem duplication, leading to an out of frame event.

Sanger sequencing confirmed these aberrant isoforms on specific RT-PCR amplicons (Figure 6B, 6F) using primer pair 1 for the first patient and primer pair 2 for the second patient (Table S3).

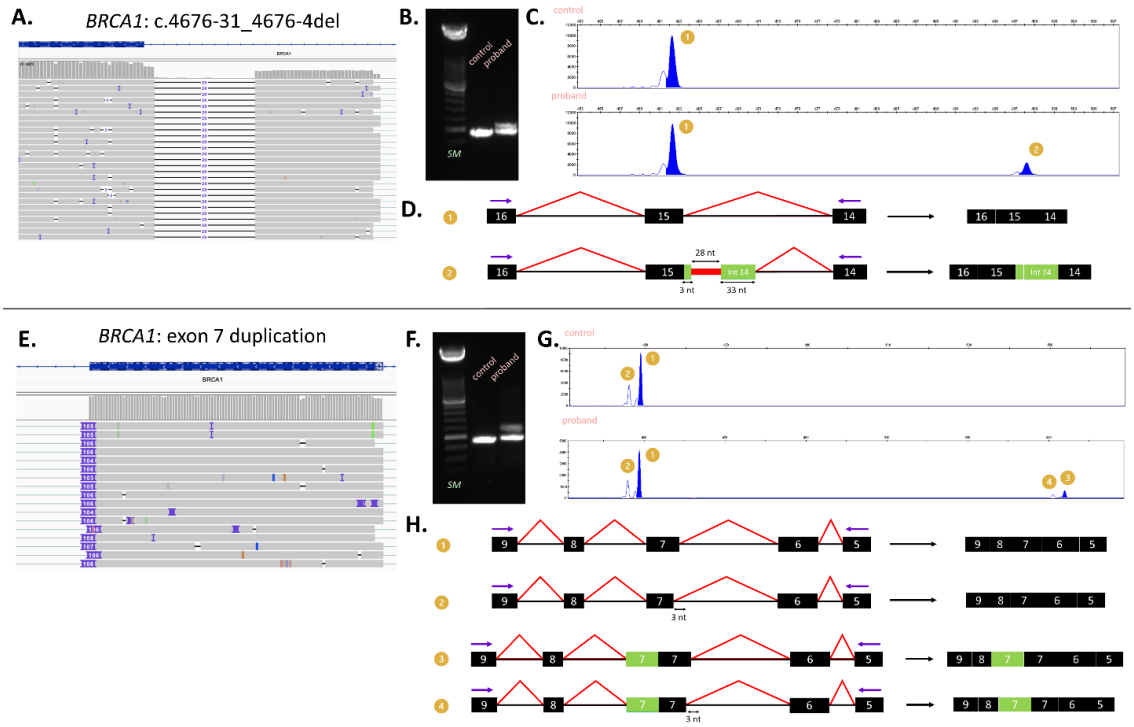


Figure 6: Validation on positive controls. A. Bam file from the LRS of the proband carrying the intronic retention in intron 14 of *BRCA1* B. RT-PCR, gel electrophoresis SM: molecular weight size marker C. Capillary electrophoresis of RT-PCR products D. Isoform structures of fragments obtained in C. black boxes: exons; green boxes: novel exons, black thin lines: introns; red thick line: deletion; red lines: splicing junctions; purple arrows: RT-PCR primers E. Bam file from LRS of the proband carrying the

exon 7 duplication of *BRCA1* F. RT-PCR, gel electrophoresis G. Capillary electrophoresis of the RT-PCR products H. Isoform structures of the fragments obtained in G.

Final table generated by LoRID pipeline allowed the complete identification and annotation of these two controls. In the first control, one isoform carrying the aberrant event without other abnormalities was found. According to our annotation algorithm, ▼14p(64) event corresponded to the 64 nt retention in intron 14. In the second patient, the aberrant event was annotated: ▼7q(95). Due to misalignment, the insert size did not exactly match the 106 nt of exon 7. Three different isoforms supporting this aberrant event were assembled. Of these three isoforms, one isoform only supported this aberrant event, while the other two also supported previously known alternative splicing events: Δ(8-9) and Δ10q(3309). These aberrant isoforms were also present in other patients, but with lower expression level. Indeed, the comparison of isoform expression metrics in the patient cohort allowed the identification of these aberrant isoforms for all these controls (Table S4).

Characterization of an unexplained hereditary

Two probands from a family with cancer history constituted the last two patients in this analysis. The first proband (III.1) developed breast cancer at the age of 39 and ovarian cancer at the age of 56. Her sister (III.2) developed ovarian cancer at 49. The mother (II.2) died of breast cancer at age 51. The grandmother (I) died of gynecological cancer (Figure 7A). Initial targeted DNA sequencing (34) screening of these two probands was inconclusive.

LoRID pipeline assembled two different isoforms carrying an aberrant event in the *BRCA1* gene. These two isoforms were present on all the samples but with a significantly higher expression value in the probands. The aberrant event was annotated as 12exo(104)[p416,q5269] by SOSTAR and corresponded to a pseudo-exon of 104 bp in intron 12 of *BRCA1*. Looking at the alignment files on IGV, we identified this pseudo-exon as well as an insertion of about 900 bp of a repeated sequence in intron 12 of *BRCA1* (Figure 7B). Asking the Dfam transposable element database

(https://www.dfam.org/home), we identified that this sequence corresponded to an SVA retrotransposon.

Long read sequencing was also performed on the high molecular weight DNA using the adaptive sampling method. Regions of interest, comprising 153 genes, were enriched with an average coverage of 50 reads per base, compared to 5 reads for the off-target region. The SVA retrotransposon was found in intron 12 of the *BRCA1* gene at a length of approximately 2,700 bp. Comparison between SVA sequence from RNA and SVA sequence from DNA was performed. The RNA sequence matched part of the complete DNA sequence.

Targeted RNA short read sequencing revealed two abnormal splicing junctions using SpliceLauncher tool. These two junctions were present in both probands and absent in the other samples in the run. They were both located in intron 12 of the *BRCA1* gene and were annotated as ▼12p(5373) starting at position c.4357+417 and ▼12q(5400) ending at position c.4358-390 of the *BRCA1* gene. These junctions were in favor of the presence of a cryptic event in this deep intronic region. A loss of expression of the *BRCA1* gene was computed by the DESeq2 tool (35) on these sequencing data.

RT-PCR with primer pair 3 located on *BRCA1* was performed (Table S3). Gel electrophoresis and Sanger sequencing of these amplicons confirmed the insertion of a sequence of approximately 1000 bp, present in the probands and absent in the controls (Figure 7C). This sequence included 102 bp of a cryptic exon, followed by the beginning of the SVA retrotransposon with the hexamer repeats. Another RT-PCR with primer pair 4 amplified only the full-length transcript with a polymorphism at position c.3548 T>C (Table S3). Sanger sequencing of this polymorphism revealed the absence of the mutated allele.

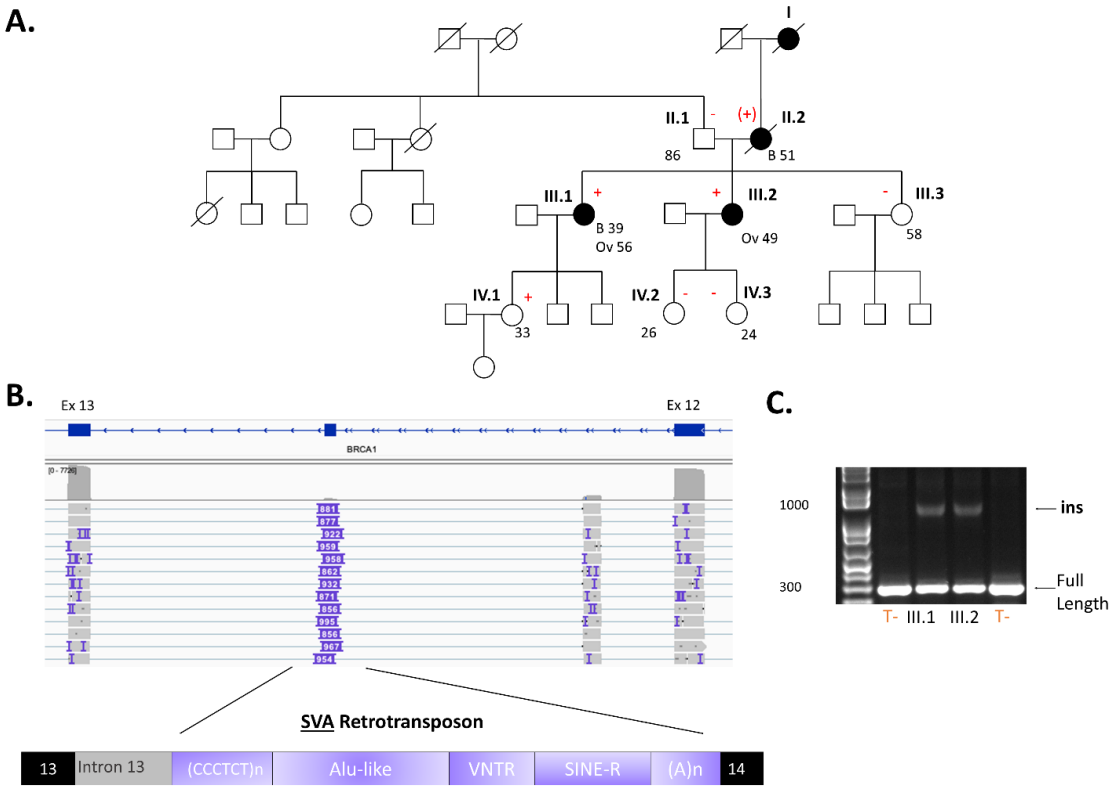


Figure 7: An unexplained hereditary case. A. Family pedigree of a family with breast and ovarian cancers. Black symbols indicate patients with breast (B) or ovarian (Ov) cancers. Ages are given as age at diagnosis for cancer patients and current age for living probands. The red '+' indicates the SVA retrotransposon insertion, the red '-' indicates the normal sequence at intron 13 B. Bam alignment file, displayed on IGV software, from ONT long read sequencing showing the pseudo exon and the SVA retrotransposon insertion in intron 13 of BRAC1 gene with the detailed schematic structure of the aberrant isoform C. RT-PCR gel electrophoresis showing an insertion of approximately 1000 bp in cancer probands (III.1; III.2) compared to two controls (T-).

Analysis of the ovarian tumor of the second proband (III.2) highlighted a genomic instability using the GIScar test (36). Specific RT-PCR on this ovarian tumor with primer located specifically on the SVA

retrotransposon and on exon 13 of *BRCA1* (primer pair 5, Table S3) identified the abnormal transcript. In addition, sequencing of this ovarian tumor showed a loss of heterozygosity on two polymorphisms.

A specific PCR test was designed with primer pair 6 (Table S3) for the identification of this SVA retrotransposon from DNA blood samples. In a first time, the unaffected father (II.1) and sister (III.3) were tested and found to be negative. The unaffected daughters (IV.1, IV.2, IV.3) were then tested. The test was positive for one of them. These results showed that this SVA insertion segregated with cancers in this family.

Discussion

We developed an innovative targeted capture enrichment for a panel of 28 genes involved in predisposition to breast and ovarian cancer suitable for long read RNAseq analysis. Looking at the high percentage of on target, our approach allowed an enrichment of these genes. Therefore, we achieved sufficient coverage rate to detect a collection of isoform structures. The study of *BRCA1* and *BRCA2* transcripts demonstrated the reliability of these data since all well-known alternative splicing events were found with an expression level coherent with the literature data (31,32). Common splicing junctions obtained by both LRS and SRS were the most expressed compared to junctions identified only by SRS. Additionally, we identified novel associations of these splicing events leading to novel isoforms. LRS simplified the detection of complex events such as pseudo-exon, while SRS detected the donor and acceptor splice site separately.

Regarding the positive controls, both PacBio and ONT platforms allowed the complete identification of aberrant transcripts in a single experiment. While previous experiments, based on RT-PCR or SRS, could only detect these events indirectly and required several steps of analysis. These results highlight the advantage in time and ease of LRS. Another advantage of the target enrichment was the possibility to sequence multiple patients at once, reducing sequencing cost. As a result, decreasing costs facilitate the implementation of LRS in both research and diagnostic laboratories. Multiplexing

allowed also the highlighting of aberrant isoforms by comparing the isoforms expression metrics between samples.

In this study, RNA from LCL samples were used. Indeed, a sufficient RNA quality (RIN > 9) is required for any RNA LRS. This constraint could be a limit for *ex vivo* samples such as PAXgene sample, where RNA is partially decayed. Also, the design of probes played a major role in the final isoform length. Previous designs with overlapping probes or a 1x tailing were tried but resulted in increased isoforms fragmentation. This phenomenon could be explained by mechanical constraints applied to the cDNA by two probes.

Our protocol was compatible with both long read technologies (PacBio or ONT). Nanopore sequencing allowed to sequence longer fragments than PacBio sequencing. PacBio provides an all-in-one pipeline (IsoSeq) to analyze sequencing data with read consensus generation and high-quality isoforms auto-assembly. While IsoSeq assembled more isoforms, isoforms assembled by LoRID pipeline were the most relevant. Indeed, isoforms identified by LoRID were the longest and the most expressed. This pipeline is consistent with a diagnostic approach to detect aberrant isoforms with good confidence and without background noise. The versatility of LoRID pipeline allows SOSTAR to be used separately with any GTF file comprising isoforms assembled by other tools from any LRS data. SOSTAR facilitates interpretation of LRS data by providing a descriptive and a human understandable annotation of isoform structures.

LoRID pipeline helped us to identify a mobile element in a family with an unexplained hereditary, where current techniques failed. This mobile element corresponded to an SVA retrotransposon. Characterization of this element allowed us to develop a specific PCR test suitable for genetic counselling of the family. This long read approach contributes to optimize preventive and therapeutic care. Combination of LoRID and other ONT pipelines could be an interesting strategy to detect aberrant isoforms with good confidence, and a comprehensive collection of isoforms. Such an

insertion of an SVA retrotransposon into intron 12 of *BRCA1* was previously described in the literature (37). This suggested a fragile region in the *BRCA1* gene into which mobile elements could be inserted.

In conclusion, we validated a new protocol of targeted enrichment by probes suitable for ONT and PacBio platforms on both negative and positive controls. We provide a pipeline, suitable for diagnostic purposes, to detect and annotate aberrant isoforms from LRS data. This pipeline elucidates an unexplained hereditary by characterizing a complex event. This proof-of-concept offers new opportunities in RNA structure exploration for research and molecular diagnostic purposes.

Declaration of interests

All authors except N.S. and D.B. declare that they have no competing interests. N.S. is employed by SeqOne Genomics for the time period October 2020 to present in the context of a public-private PhD project (CIFRE fellowship #2020/0103) partnership between INSERM and SeqOne Genomics. D.B. is employed by SeqOne Genomics as Head Bioinformatics.

Acknowledgement

We are grateful to the *Site de Recherche Intégrée sur le Cancer* (SiRIC, PhD Sylvain Baulande) of Curie Institut and the *Génotypage et Séquençage en Auvergne* (GENTYANE, PhD Charles Poncet) of *Centre de Clermont Auvergne Rhône Alpes* for performing the PacBio sequencing. We also thank Nolan Talhi and Peter Verhasselt, from Integrated Dna Technologies (IDT®), for their help during the design of capture probes.

Nanopore commercial

PacBio commercial

Data and code availability

Supplementary data are available online at: https://github.com/LBGC-CFB/LoRID

Funding

This work was supported by a grant from the French *Cancéropôle Nord-Ouest* (CNO) n° 2018/06.

Ethical declaration

All patients had signed an informed consent.

References

1. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet [Internet]. déc 2008 [cité 28 sept 2018];40(12):1413-5. Disponible sur: https://www.nature.com/articles/ng.259

2. Navaratnam DS, Bell TJ, Tu TD, Cohen EL, Oberholtzer JC. Differential Distribution of Ca2+-Activated K+ Channel Splice Variants among Hair Cells along the Tonotopic Axis of the Chick Cochlea. Neuron [Internet]. 1 nov 1997 [cité 15 juill 2019];19(5):1077-85. Disponible sur: http://www.sciencedirect.com/science/article/pii/S0896627300803980

3. Rosenblatt KP, Sun ZP, Heller S, Hudspeth AJ. Distribution of Ca2+-Activated K+ Channel Isoforms along the Tonotopic Gradient of the Chicken's Cochlea. Neuron [Internet]. 1 nov 1997 [cité 15 juill 2019];19(5):1061-75. Disponible sur: http://www.sciencedirect.com/science/article/pii/S0896627300803979

4. Bonnal SC, López-Oreja I, Valcárcel J. Roles and mechanisms of alternative splicing in cancer — implications for care. Nat Rev Clin Oncol [Internet]. août 2020 [cité 20 déc 2022];17(8):457-74. Disponible sur: https://www.nature.com/articles/s41571-020-0350-x

5. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am J Hum Genet [Internet]. 4 janv 2018 [cité 20 déc 2022];102(1):11-26. Disponible sur: https://www.sciencedirect.com/science/article/pii/S0002929717304548

6. Rogalska ME, Vivori C, Valcárcel J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. Nat Rev Genet [Internet]. 16 déc 2022 [cité 20 déc 2022];1-19. Disponible sur: https://www.nature.com/articles/s41576-022-00556-8

7. Cheung R, Insigne KD, Yao D, Burghard CP, Wang J, Hsiao YHE, et al. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. Mol Cell [Internet]. 3 janv 2019 [cité 16 mars 2020];73(1):183-194.e8. Disponible sur: http://www.sciencedirect.com/science/article/pii/S1097276518308979

8. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet Med [Internet]. juin 2020 [cité 2 sept 2020];22(6):1005-14. Disponible sur: https://www.nature.com/articles/s41436-020-0766-9

9. Truty R, Ouyang K, Rojahn S, Garcia S, Colavin A, Hamlington B, et al. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. Am J Hum Genet [Internet]. 1 avr 2021 [cité 12 mai 2022];108(4):696-708. Disponible sur: https://www.sciencedirect.com/science/article/pii/S0002929721000902

10. Bournazos AM, Riley LG, Bommireddipalli S, Ades L, Akesson LS, Al-Shinnag M, et al. Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. Genet Med [Internet]. 1 janv 2022 [cité 19 oct 2022];24(1):130-45. Disponible sur: https://www.sciencedirect.com/science/article/pii/S1098360021041289

11. Goyenvalle A, Vulin A, Fougerousse F, Leturcq F, Kaplan JC, Garcia L, et al. Rescue of Dystrophic Muscle Through U7 snRNA-Mediated Exon Skipping. Science [Internet]. 3 déc 2004 [cité 23 juill 2019];306(5702):1796-9. Disponible sur: https://science.sciencemag.org/content/306/5702/1796

12. Meulemans L, Mesman RLS, Caputo SM, Krieger S, Guillaud-Bataille M, Caux-Moncoutier V, et al. Skipping nonsense to maintain function: the paradigm of BRCA2 exon 12. Cancer Res [Internet]. 1 janv 2020 [cité 17 févr 2020]; Disponible sur: https://cancerres.aacrjournals.org/content/early/2020/02/11/0008-5472.CAN-19-2491

13. Lopez-Perolio I, Leman R, Behar R, Lattimore V, Pearson JF, Castéra L, et al. Alternative splicing and ACMG-AMP-2015-based classification of PALB2 genetic variants: an ENIGMA report. J Med Genet [Internet]. 19 mars 2019 [cité 5 avr 2019];jmedgenet-2018-105834. Disponible sur: https://jmg.bmj.com/content/early/2019/03/19/jmedgenet-2018-105834

14. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am J Hum Genet [Internet]. 7 mars 2019 [cité 4 janv 2022];104(3):466-83. Disponible sur: https://www.sciencedirect.com/science/article/pii/S0002929719300126

15. Davy G, Rousselin A, Goardon N, Castéra L, Harter V, Legros A, et al. Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. Eur J Hum Genet EJHG. oct 2017;25(10):1147-54.

16. Adamson SI, Zhan L, Graveley BR. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. Genome Biol [Internet]. 1 juin 2018 [cité 14 janv 2019];19(1):71. Disponible sur: https://doi.org/10.1186/s13059-018-1437-x

17. Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods [Internet]. déc 2013 [cité 20 déc 2022];10(12):1177-84. Disponible sur: https://www.nature.com/articles/nmeth.2714

18. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods [Internet]. déc 2019 [cité 20 déc 2022];16(12):1297-305. Disponible sur: https://www.nature.com/articles/s41592-019-0617-2

19. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature [Internet]. août 2022 [cité 21 déc 2022];608(7922):353-9. Disponible sur: http://www.nature.com/articles/s41586-022-05035-y

20. Treutlein B, Gokce O, Quake SR, Südhof TC. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proc Natl Acad Sci [Internet]. 1 avr 2014 [cité 25 juill 2019];111(13):E1291-9. Disponible sur: https://www.pnas.org/content/111/13/E1291

21. Jong L de, Cree S, Lattimore V, Wiggins G, Spurdle A, Investigators kConFab, et al. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. bioRxiv [Internet]. 30 juill 2017 [cité 23 nov 2017];169755. Disponible sur: https://www.biorxiv.org/content/early/2017/07/30/169755

22. Schwenk V, Leal Silva RM, Scharf F, Knaust K, Wendlandt M, Häusser T, et al. Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome. J Med Genet. 2 janv 2023;jmg-2022-108931.

23. Hardwick SA, Bassett SD, Kaczorowski D, Blackburn J, Barton K, Bartonicek N, et al. Targeted, High-Resolution RNA Sequencing of Non-coding Genomic Regions Associated With Neuropsychiatric Functions. Front Genet [Internet]. 2019 [cité 28 juin 2023];10. Disponible sur: https://www.frontiersin.org/articles/10.3389/fgene.2019.00309

24. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet. déc 2017;49(12):1731-40.

25. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res [Internet]. 4 janv 2016 [cité 25 janv 2019];44(D1):D733-45. Disponible sur: https://academic.oup.com/nar/article/44/D1/D733/2502674

26. Davy G, Rousselin A, Goardon N, Castéra L, Harter V, Legros A, et al. Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. Eur J Hum Genet [Internet]. oct 2017 [cité 23 août 2019];25(10):1147-54. Disponible sur: https://www.nature.com/articles/ejhg2017116

27. Leman R, Harter V, Atkinson A, Davy G, Rousselin A, Muller E, et al. SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data. Bioinformatics [Internet]. 1 mars 2020 [cité 10 mars 2020];36(5):1634-6. Disponible sur: https://academic.oup.com/bioinformatics/article/36/5/1634/5588411

28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma Oxf Engl. 15 sept 2018;34(18):3094-100.

29. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 16 déc 2019;20(1):278.

30. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. janv 2011;29(1):24-6.

31. Colombo M, Blok MJ, Whiley P, Santamariña M, Gutiérrez-Enríquez S, Romero A, et al. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. Hum Mol Genet. 15 juill 2014;23(14):3666-80.

32. Fackenthal JD, Yoshimatsu T, Zhang B, Garibay GR de, Colombo M, Vecchi GD, et al. Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples. J Med Genet. 1 août 2016;53(8):548-58.

33. Leman R, Parfait B, Vidaud D, Girodon E, Pacot L, Le Gac G, et al. SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing. Hum Mutat. déc 2022;43(12):2308-23.

34. Castéra L, Krieger S, Rousselin A, Legros A, Baumann JJ, Bruet O, et al. Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. Eur J Hum Genet EJHG. nov 2014;22(11):1305-13.

35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

36. Leman R, Muller E, Goardon N, Chentli I, Tranchant A, Legros A, et al. 2022-RA-935-ESGO Development of an academic genomic instability score for ovarian cancers. Int J Gynecol Cancer [Internet]. 1 oct 2022 [cité 11 août 2023];32(Suppl 2). Disponible sur: https://ijgc.bmj.com/content/32/Suppl_2/A280.1

37. Walsh T, Casadei S, Munson KM, Eng M, Mandell JB, Gulsuner S, et al. CRISPR-Cas9/long-read sequencing approach to identify cryptic mutations in BRCA1 and other tumour suppressor genes. J Med Genet. déc 2021;58(12):850-2.

Supplementary Figures

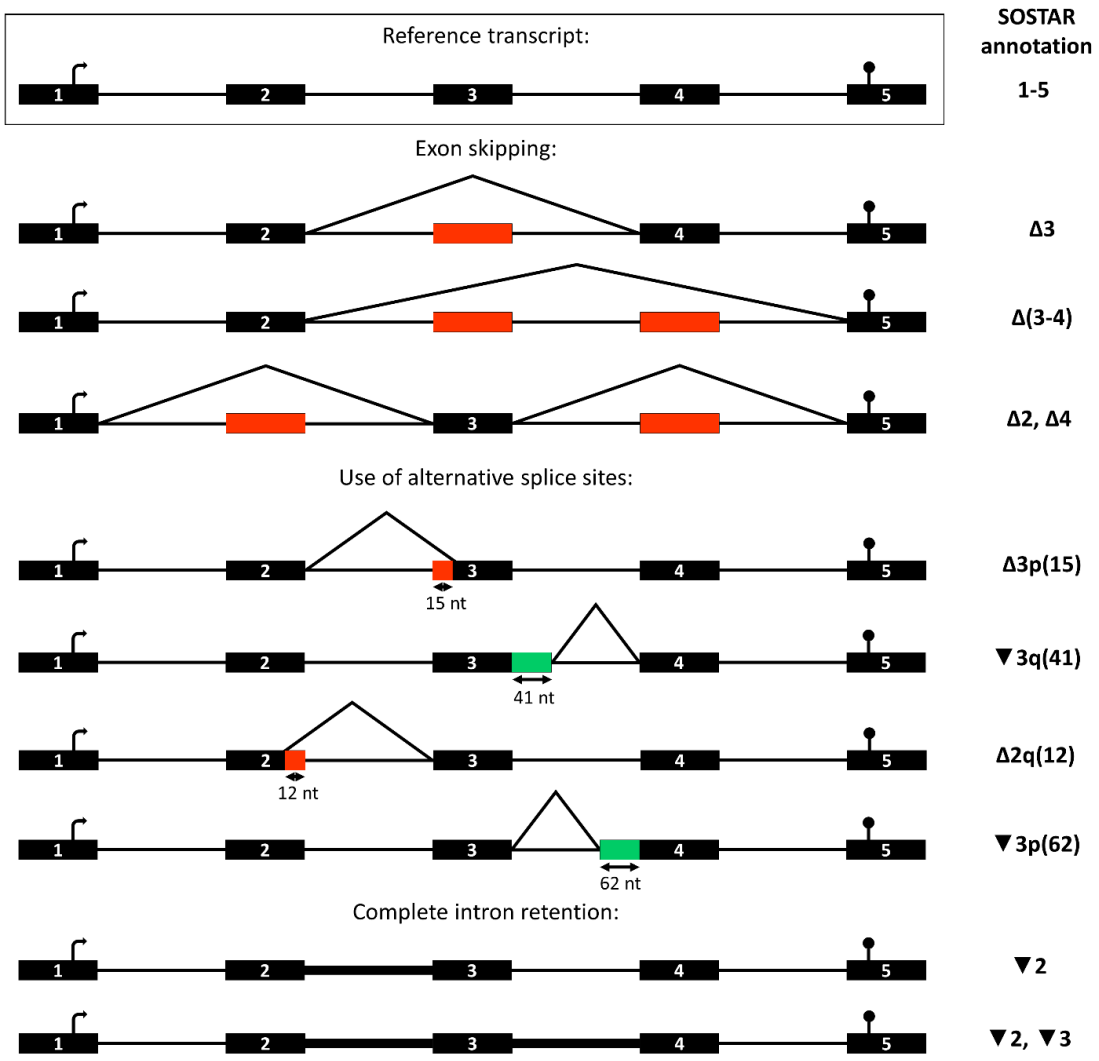


Figure S1: Mechanisms of the transcript diversity with the corresponding SOSTAR annotation. Black boxes: exons, black lines: intron, red boxes: exon (or part of exon) skipping, green boxes: novel exon (or part of exon).

Figure S2: Gene panel capture probe design. Reference transcripts are represented in blue, and the capture probes in red.

Supplementary Tables

Table S1: Highly predisposed families selection criteria. yo = years old.

	Number of generation	**Degrees**	**Number of cancers***	**Type and age of cancers** at least:
a	≥ 2 generations	First-degree or second-degree through a male relative	2 cases	- 1 triple negative breast cancer ≤ 40 yo
b			2 cases	- 2 ovarian cancers with at least one ovarian cancer ≤ 50 yo
c			3 cases	- 1 breast cancer ≤ 40 yo or - 1 ovarian cancer ≤ 50 yo
d			4 cases	- 1 breast cancer ≤ 50 yo or - 1 ovarian cancer ≤ 60 yo
e			5 cases	- 1 breast or ovarian cancer without age criteria

* Number of cancers in the restricted HBOC spectrum (breast, ovarian)

Table S2: Custom panel gene list with associated reference transcript NMs. bp = base pairs.

Genes	Number of probes	Reference transcript NMs	Transcript lengths
ATM	12	NM_000051.4	12,915 bp
BAP1	8	NM_004656.4	3,600 bp
BRCA1	10	NM_007294.4	7,088 bp
BRCA2	13	NM_000059.3	11,954 bp
BRIP1	14	NM_032043.3	8,182 bp
CDH1	16	NM_004360.5	4,811 bp
CHEK2	20	NM_007194.4	1,844 bp
MLH1	12	NM_000249.4	2,494 bp
MLH3	14	NM_001040108.2	7,820 bp
MRE11	17	NM_005591.4	6,841 bp
MSH2	20	NM_000251.3	3,115 bp
MSH6	12	NM_000179.3	4,265 bp
MUTYH	11	NM_001048171.2	1,858 bp
NBN	15	NM_002485.5	4,622 bp
PALB2	10	NM_024675.4	4,008 bp
PMS1	14	NM_000534.5	3,156 bp
PMS2	12	NM_000535.7	5,093 bp
PTEN	13	NM_000314.8	8,515 bp
RAD50	21	NM_005732.4	8,722 bp
RAD51	10	NM_001164269.2	2,335 bp
RAD51B	11	NM_133509.5	2,679 bp
RAD51C	16	NM_058216.3	2,562 bp
RAD51D	12	NM_002878.4	9,966 bp
RINT1	15	NM_021930.6	2,860 bp
STK11	14	NM_000455.5	3,293 bp
TP53	11	NM_000546.6	2,512 bp
XRCC2	5	NM_005431.2	4,771 bp
XRCC3	9	NM_001100119.2	2,604 bp

Table S3: Primer pair designations and sequences. ex = exon, F = forward, R = reverse, bp = base pairs.

Primer pair designations	Primer localizations	Sequences	Product lengths
1	BRCA1_ex14_F BRCA1_ex16_R	GTTGTTGATGTGGAGGAGCA TGATGTGGTGTTTTCTGGCA	433 bp
2	BRCA1_ex5_F BRCA1_ex9_R	TACGAGATTTAGTCAACTTG ATTTCATCCCTGGTTCCTTG	420 bp
3	BRCA1_ex12_F BRCA1_ex14_R	AGCCAGCCTTCTAACAGCTA GGGCATTTAGAAGGGGATGAC	239 bp
4	BRCA1_ex10_R BRCA1_ex13_R	GACATTAAGGAAAGTTCTGCT CCTTCTGGATTCTGGCTTATAGGG	908 bp
5	SVA_F BRCA1_ex13_R	CTGTCTGGGATGTGAGGA AAGGCCTTCTGGATTCTGG	119 bp
6	BRCA1_ex12_F SVA_R	CCTTGAGGACCTGCGAAA CTCCCCACATCTCAGACGAT	2,018 bp

5. Discussion générale

5.1 Conclusions sur le développement de DrugOrder

Le développement de DrugOrder a permis, au travers de l'étude d'un large panel de gènes, la réalisation d'un profil génétique de la tumeur mettant en avant les variants actionnables. L'actionnabilité d'un variant s'évalue après une étape d'annotation rendue possible par l'utilisation de nombreuses bases de données qui fournissent, entre autres, de l'information sur l'impact du variant dans l'oncogenèse ainsi que les thérapies disponibles associées à un variant, pour un phénotype tumoral donné. Au travers du *framework* inclus dans DrugOrder, la classification des associations thérapie-variant est basée sur 4 critères : l'impact du variant sur le gène, la similarité du variant du patient par rapport aux variants présents dans la base de données thérapeutique, la similarité du cancer du patient comparé aux cancers décrits pour le variant dans les bases thérapeutiques et le niveau d'évidence clinique de la thérapie associée au variant dans la base de données. Cette classification a montré, au travers de comparaisons avec des rapports réalisés par une méthode commerciale (FMI®) et par un biologiste, que DrugOrder est capable de fournir une liste bien ordonnée d'associations thérapie-variant. Au-delà de mettre en évidence les thérapies possédant des AMM, DrugOrder permet de mener à l'identification de thérapies d'évidence clinique plus faible, comme la recommandation de la thérapie AZD8186 pour les pertes de fonction du gène *PTEN* [198] ou du Lucitanib pour les amplifications des gènes de la famille FGF [199].

Bien qu'il existe un bon recouvrement pour les thérapies approuvées par la FDA entre les bases de données comme CIViC, OncoKB, CKB ou MM, il y a peu de concordances pour les autres thérapies ne faisant pas encore consensus dans les circuits de prescription ou en manque de preuves cliniques. L'hétérogénéité des bases de données nécessite alors d'utiliser plusieurs bases de données afin de couvrir l'ensemble des solutions thérapeutiques proposées pour un variant.

Le *framework* de classification inclus dans DrugOrder nécessite des informations présentes dans diverses bases de données mais ne requiert pas l'utilisation d'une base de données spécifique. Toutefois nous n'avons pas comparé les performances de ce dernier, en utilisant d'autres bases de données que MM (OncoKB, COSMIC, CIViC, CKB), pouvant fournir des annotations différentes. De plus, nous n'avons pas établi de logique pour l'intégration de plusieurs bases de données thérapeutiques pouvant amener de la (i) redondance (informations identiques entre les bases de données) ou (ii) des différences dans les informations fournies.

(i) La redondance se manifeste au travers de thérapies présentes dans plusieurs bases de données. DrugOrder pourrait intégrer des identifiants uniques pour décrire chaque association thérapie-variant et traiter la redondance. Ces identifiants pourraient être composés du nom du gène, du nom de la mutation du variant, du nom de la thérapie et du type de cancer. Ce système serait applicable sur l'ensemble des bases testées comme COSMIC, CIViC, CKB ou MM.

(ii) Des associations peuvent posséder des informations différentes en fonction des bases de données. Cela peut correspondre à des différences sur l'ontologie, le niveau d'évidence clinique ou le nom de la thérapie. Pour l'algorithme de DrugOrder, ces différences seraient considérées comme des associations différentes. Ainsi il est à prévoir des méthodes complémentaires capable de choisir l'association la plus pertinente. Il pourrait, par exemple, sembler pertinent de favoriser l'association disposant du niveau d'évidence clinique la plus forte ou la plus récente, selon un niveau de confiance *a priori* sur la qualité d'une base.

DrugOrder a montré sa capacité de prioriser les associations thérapie-variant pour les SNV, CNV et fusions de gènes. DrugOrder est un outil qui peut aider les biologistes à choisir les meilleures options thérapeutiques et pouvant identifier des variants dont l'actionnabilité peut être discutée durant des Réunion de Concertation Pluridisciplinaire (RCP), nécessaires pour la prise en charge d'un patient et de son suivi.

5.2 Conclusions sur le développement de LoRID

Avec le développement de LoRID nous avons montré que l'assemblage de transcrits de l'ARNm en *long-reads* à partir d'un panel de gènes, permet d'obtenir la structure complète d'une isoforme alternative de l'ARN. LoRID a permis d'identifier, après un protocole innovant d'enrichissement d'ARNm de gènes d'intérêt, des isoformes complètes en grande profondeur. Jusqu'alors seules les jonctions de ces isoformes étaient décrites par le séquençage en *short-reads*. L'assembleur utilisé dans

LoRID, StringTie, a identifié les isoformes non physiologiques attendues, présentes dans les échantillons contrôles positifs. LoRID a également permis de résoudre un cas d'hérédité manquante chez un patient présentant un syndrome HBOC, en mettant en évidence un transcrit anormal.

Ainsi l'utilisation de StringTie au travers de LoRID a démontré sa pertinence pour une utilisation en diagnostic, avec les procédures de séquençage employées. LoRID, et grâce à l'assemblage proposé par StringTie, les isoformes assemblées sont plus longues, moins diversifiées et supportées par plus de *reads* par rapport au pipeline Iso-Seq. Cette diminution du nombre d'isoformes peut s'expliquer par l'étape d'assemblage de StringTie, présente dans LoRID et absente d'Iso-Seq. L'assemblage peut en effet utiliser des *reads* homologues issus de molécules d'ARN différentes, pour la construction de l'isoforme [200, 176].

La présence, dans l'ADN d'un des échantillons, de l'insertion d'un retrotransposon SVA sur le gène *BRCA1* a montré certaines limites de LoRID. Les régions faiblement couvertes en *reads* ne sont plus présentes lors de l'assemblage d'isoformes, aboutissant à la disparition de certains épissages alternatifs présents à l'alignement (exemple de l'intronisation présent en amont du retrotransposon). Une amélioration des alignements ou un reparamétrage de LoRID pourrait être nécessaire pour mieux caractériser ces événements et gagner en sensibilité. En effet, l'utilisation de deux étapes d'alignement (la première est utilisée pour assembler les transcrits qui sont utilisés en guide pour la seconde) semble montrer plus de justesse sur la détection finale des isoformes (voir Figure 5.1). Les combinaisons d'évènements d'épissage sont moins nombreuses dans le cas d'une double étape d'alignement par rapport au nombre de combinaisons présentes avec une seule étape d'alignement. Deux étapes d'alignement semblent favoriser une meilleure caractérisation de certaines compositions d'isoformes (exemple de la $\Delta 4$ avec $\Delta(8-9)$ ou $\Delta(8-10)$).

A.

	Δ13p(3)	Δ12p(3)	Δ10q(3309)	Δ(8-10)	Δ(8-9)	Δ8	Δ7p(3)	Δ4q(22)	Δ4	Δ1q(6)
Δ1q(6)										3065
Δ4									1059	160
Δ4q(22)								614	0	175
Δ7p(3)							3247	177	145	941
Δ8						326	97	22	16	111
Δ(8-9)					4344	0	1304	199	499	1176
Δ(8-10)				644	0	0	113	31	302	98
Δ10q(3309)			5205	0	2520	129	1654	265	259	1534
Δ12p(3)		2007	1034	63	796	73	569	109	151	582
Δ13p(3)	4679	947	2377	152	1898	147	1466	282	358	1359

B.

	Δ13p(3)	Δ12p(3)	Δ10q(3309)	Δ(8-10)	Δ(8-9)	Δ8	Δ7p(3)	Δ4q(22)	Δ4	Δ1q(6)
Δ1q(6)										3667
Δ4									29757	194
Δ4q(22)								1063	0	278
Δ7p(3)							4607	335	485	1195
Δ8						669	152	49	264	124
Δ(8-9)					21480	0	1560	302	17218	1151
Δ(8-10)				11847	0	0	173	58	11095	209
Δ10q(3309)			4025	0	1924	138	1255	244	554	1022
Δ12p(3)		3	0	3	0	0	0	0	3	0
Δ13p(3)	8511	3	1375	730	5976	121	936	196	5976	757

Figure 5.1 – Combinaison des événements d'épissage du gène *BRCA1*.
A. Paires d'événements d'épissage combinées détectées en *long-reads* en utilisant une étape d'alignement. **B.** Paires d'événements d'épissage combinées détectées en *long-reads* en utilisant deux étapes d'alignement.

Des étapes de correction d'erreurs des *reads* [201, 202], induits par des INDEL aléatoires lors d'un séquençage ONT, pourraient améliorer la détection de transcrits alternatifs caractérisés par l'utilisation d'un site d'épissage éloigné de quelques bases du site utilisé par le transcrit de référence. La présence possible de ce type d'erreurs de séquence au niveau des sites d'épissage, peut induire à l'assemblage, des isoformes fausses positives ou inversement de fausses négatives, si, le bruit de fond d'erreurs, masque l'utilisation d'un site alternatif. De même, condenser en un seul transcrit ne différant que par des anomalies (réduction ou allongement) des *Untranslated Region* (UTR), pourrait améliorer les capacités de détection de LoRID, en limitant la diversité des isoformes décrites, probablement sans conséquence biologique.

StringTie peut être paramétré en diminuant le nombre de *reads* devant supporter un assemblage, ce qui pourrait augmenter la diversité des transcrits, augmentant probablement les chances d'observer une isoforme fausse positive. Une augmentation de l'abondance des isoformes détectées pourrait favoriser l'identification de nouveaux épissages alternatifs de transcrits. Néanmoins, cela nécessiterait de nombreuses vali-

dations en aval, par des RT-PCR *Long Range* spécifiques de chaque événement, afin d'identifier les transcrits vrais positifs des transcrits faux positifs.

Aussi, StringTie fournit des valeurs d'expression comme le *Transcripts Per kilobase Million* (TPM) une métrique de comptage de *reads* normalisée sur la taille du gène et sur la profondeur de séquençage. L'étude des valeurs d'expression fournies par LoRID a permis de corréler l'expression des *reads* au niveau des jonctions des sites d'épissage en *short-reads* avec l'expression des transcrits en *long-reads*, laissant penser qu'une analyse quantitative des transcrits pleine longueur peut être envisagée.

LoRID permettrait d'explorer de nouvelles interprétations du transcriptome de l'individu, au travers d'expression différentielle, puisque l'information du transcrit par million est donnée par le pipeline. Néanmoins ces données d'expression doivent encore faire l'objet d'un travail de normalisation et d'analyses biostatistiques qui entreront dans de prochaines versions du pipeline.

L'intégration de l'outil SOSTAR permet de faciliter l'interprétation en écrivant de manière lisible la variation vis-à-vis d'un transcrit de référence. La nomenclature proposée par SOSTAR décrit les épissages alternatifs du transcrit permettant d'identifier des sauts d'exon, des rétentions d'intron ou la présence de nouveaux sites d'épissage alternatif. SOSTAR est un module pouvant être utilisé de manière indépendante à LoRID. Toutefois, il est adapté à un format non standard de fichier de sortie, défini lors du développement de LoRID.

LoRID en combinaison du séquençage *long-read* ciblant les messagers d'un panel de gène peut être proposé pour des approches RNAseq dans le contexte des maladies génétiques et notamment des prédispositions aux cancers du sein et de l'ovaire. L'utilisation de *long-reads* ouvre de nouvelles opportunités dans le diagnostic moléculaire notamment dans l'exploration de la structure de l'ARN jusqu'alors limité par les techniques de séquençage *short-reads* disponibles. Au final, LoRID décrit lors d'un séquençage ciblé en *long-reads*, une grande diversité des transcrits assemblés. C'est une première étape pour aider à l'interprétation des isoformes afin notamment de comprendre leur structure et d'exprimer leur abondance, au sein d'un individu ou leur distribution dans des populations spécifiques (population de patients malades comparée à une population saine).

5.3 Perspectives

Des perspectives à court et moyen terme sont envisagées pour DrugOrder et LoRID.

DrugOrder est capable d'associer des thérapies en fonction de signatures mais ces résultats ne sont pas intégrés à la hiérarchisation pour le moment. DrugOrder permet l'identification de thérapies pour des signatures mutationnelles :

— La signature *Microsatellite Instability* (MSI), permet de détecter des déficiences *Missmatch Repair* (MMR) caractérisées par une accumulation d'erreurs particulièrement sur les séquences répétées de l'ADN et notamment dans les microsatellites. [203, 204, 205].

— La signature *Tumor Mutation Burden* (TMB), décrite comme le nombre de mutations par mégabase dans une tumeur. Le TMB identifie les cancers avec une forte charge mutationnelle (10 mut/Mb) [206, 207]. Les signatures TMB et MSI président une réponse aux immunothérapies [207].

— La signature *Homologous Recombination Deficiency* (HRD), est associée à une défaillance du système de recombinaison homologue. L'emploi d'inhibiteur de PARP tel que l'Olaparib est recommandé pour les patients, dont la tumeur présente un statut positif à la signature HRD, dit déficient HRD [208].

Aujoud'hui des outils bioinformatiques validés cliniquement, existent pour la détection des MSI [209, 210], la définition du TMB [211] et la signature HRD [212, 213]. Ces résultats pourraient être inclus dans le *framework* de DrugOrder afin d'améliorer la hiérarchisation.

Aussi, il pourrait être envisagé que le séquençage en longs fragments des transcrits de gènes (ciblés ou non), d'une tumeur congelée, avec l'utilisation de LoRID, permette sa caractérisation au travers de profils d'expression pouvant être typiques de cancers [214, 215] ou même aider à l'identification d'un statut HRD [216]. En génétique constitutionnelle, l'établissement de profils d'isoformes des messagers de gènes spécifiques, s'exprimant dans les lymphocytes, pourrait permettre l'établissement de signatures spécifiques d'une prédisposition aux cancers.

DrugOrder et LoRID vont faire l'objet de validations sur de plus larges cohortes. Pour DrugOrder, nous prévoyons l'analyse de plus de 300 patients séquencés à l'aide du panel *TruSight Oncology* (TSO) 500 d'Illumina. Les conclusions thérapeutiques données en clinique seront indiquées dans un rapport clinique, ce qui pourra aider à une validation complémentaire de DrugOrder, afin de comparer les conclusions d'un biologiste avec celles de l'outil, sur un jeu de données plus large. De nouveaux prélèvements de patients présentant une forte prédisposition sans anomalies génomiques caractérisées, seront analysés par RNAseq ciblés en *long-reads* au laboratoire. La validation de LoRID sur ce grand volume de données permettra de valider la robustesse de LoRID pour la description des isoformes alternatives. Il devrait également

permettre de résoudre de nouveaux cas d'hérédité manquante et de tenter d'établir des signatures moléculaires comme précédemment évoquées.

Enfin, l'utilisation de données plus volumineuses peut offrir de nouvelles opportunités en intelligence artificielle. Les données issues du TSO 500 ainsi que leurs conclusions cliniques peuvent permettre la construction d'un modèle afin d'améliorer les prédictions de DrugOrder, dans lequel les scores calculés seraient issus d'un modèle défini par des étapes de *machine learning*. L'intégration du *machine learning* dans DrugOrder permettrait une meilleure validation du DOscore et surtout une plus grande fiabilité de la classification des associations ayant une évidence clinique moindre. Pour l'heure, nous souhaitons proposer une version de DrugOrder capable de directement mettre en évidence les thérapies les plus pertinentes pour le patient, tout en offrant la possibilité d'explorer les associations moins décrites ou dans des phases cliniques moins avancées. L'emploi de l'intelligence artificielle avec LoRID pourrait ouvrir de nouvelles possibilités [217, 218, 219]. L'annotation effectuée dans LoRID, les métriques d'expression de gènes, les annotations de transcrits avec SOSTAR pourraient servir au développement d'un modèle, à condition de pouvoir obtenir un jeu de données de référence suffisamment volumineux.

6. Conclusion générale et personnelle

6.1 Conclusion générale

Le travail de cette thèse a permis le développement de deux outils, DrugOrder et LoRID résolvant des problématiques dans le cadre, respectivement, du diagnostic moléculaire en génétique tumorale et des syndromes de prédisposition aux cancers. Les deux outils ont pu faire l'objet de la rédaction de publications en cours de soumission. DrugOrder et LoRID peuvent s'intégrer dans la routine clinique des laboratoires de diagnostic moléculaire des cancers. Ces deux outils tendent à faciliter l'interprétation d'événements complexes, soit par la nature de l'événement soit par la diversité de l'information autour de l'événement. L'annotation fournie par les deux outils, devrait fournir un gain de temps lors de l'interprétation faite par les biologistes, devant faire face à une augmentation du nombre de tests moléculaires d'année en année. Ainsi DrugOrder a été mis en production sur la plateforme SeqOne et rendu disponible pour les utilisateurs, remplissant les objectifs d'industrialisation de la thèse CIFRE. LoRID s'intègre également dans une démarche de routine clinique et son développement a permis de respecter les conditions requises pour son utilisation médicale : tests, développement par versions fixes au travers de technologies comme Docker et Singularity et validation sur une cohorte du laboratoire. LoRID sera de plus disponible en libre accès dès la publication de l'article. S'inscrivant dans une démarche *open science*, il pourra faire l'objet de retours de la part de la communauté scientifique avec une totale transparence sur le code qui le compose.

Il reste toutefois de nombreux défis à relever sur ces deux outils, au centre de la discussion de cette thèse. Des approches itératives seront effectuées sur chacun de ces outils afin d'améliorer le processus de validation, leurs méthodes respectives pour la détection et l'annotation des événements actionnables ou des isoformes alterna-

tives de l'ARN ainsi que l'exploration de nouveaux événements pouvant améliorer le diagnostic moléculaire des patients. DrugOrder et LoRID ont été conçus de sorte à ordonner l'information plutôt que la filtrer, ainsi l'expert biologiste ou clinicien dispose d'une information peu transformée, pour mener à bien son diagnostic.

6.2 Conclusion personnelle

Au travers de cette thèse CIFRE j'ai pu effectuer un travail en étroite collaboration entre le milieu académique et industriel. Cela m'a conforté dans l'idée que ces deux milieux, de part leur complémentarité, ont beaucoup à apporter à la recherche moderne au travers de projets collaboratifs comme ce sujet de thèse. J'ai pu me sensibiliser à de nombreuses problématiques de cancers, apprendre de nouveaux domaines d'expertise (gestion de projet, programmation python, procédés d'industrialisation, développement continu, médecine personnalisée) et participer à des RCP afin de comprendre le besoin réel, de nos jours, en diagnostic des cancers. Cette thèse m'a permis de produire deux publications, de mettre en production industrielle DrugOrder et de valoriser LoRID au travers d'une démarche *open source*. J'ai pu participer aux congrès comme les Assises de Génétique Humaine, de l'*Europen Society of Human Genetics* (ESHG), des Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) ou de la Société Française de Médecine Prédictive et Personnalisée (SFMPP), afin d'enrichir ma culture scientifique sur la génétique et génomique humaine, la médecine personnalisée et la bioinformatique tout en présentant mes travaux à la communauté scientifique. Même si au cours de ces 3 dernières années, j'ai pu consolider mes connaissances dans ce vaste monde qu'est la génétique des cancers, la seule certitude que j'ai est illustrée par la maxime de Platon : "Je ne sais qu'une chose, c'est que je ne sais rien", ce qui me motive à poursuivre dans ce domaine. J'espère que ce travail contribuera à améliorer les connaissances dans le domaine de la recherche et du diagnostic des cancers. Je remercie encore tous les acteurs qui ont permis de mener ce projet à bien, à mes côtés.

Références

[1] J. D. Watson et F. H. C. Crick. « Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid ». In : *Nature* 171.4356 (avr. 1953). Number : 4356 Publisher : Nature Publishing Group, p. 737-738.

[2] Rosalind E. Franklin et R. G. Gosling. « Molecular Configuration in Sodium Thymonucleate ». In : *Nature* 171.4356 (avr. 1953). Number : 4356 Publisher : Nature Publishing Group, p. 740-741.

[3] Deanna M. Church et al. « Modernizing Reference Genome Assemblies ». In : *PLOS Biology* 9.7 (2011). Publisher : Public Library of Science, e1001091.

[4] Arshia Rehman, Saeeda Naz et Imran Razzak. « Leveraging big data analytics in healthcare enhancement : trends, challenges and opportunities ». In : *Multimedia Systems* 28.4 (1er août 2022), p. 1339-1371.

[5] Apostolia M. Tsimberidou et al. « Review of Precision Cancer Medicine : Evolution of the Treatment Paradigm ». In : *Cancer treatment reviews* 86 (juin 2020), p. 102019.

[6] Solip Park, Fran Supek et Ben Lehner. « Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits ». In : *Nature Communications* 9.1 (4 juill. 2018). Number : 1 Publisher : Nature Publishing Group, p. 2601.

[7] Stephen J. Chanock. « How the germline informs the somatic landscape ». In : *Nature Genetics* 53.11 (nov. 2021). Number : 11 Publisher : Nature Publishing Group, p. 1523-1525.

[8] Roger McLendon et al. « Comprehensive genomic characterization defines human glioblastoma genes and core pathways ». In : *Nature* 455.7216 (oct. 2008). Number : 7216 Publisher : Nature Publishing Group, p. 1061-1068.

[9] Matthew H. Bailey et al. « Comprehensive Characterization of Cancer Driver Genes and Mutations ». In : *Cell* 173.2 (avr. 2018). Publisher : Cell Press, 371-385.e18.

[10] P. C. Nowell. « The clonal evolution of tumor cell populations ». In : *Science (New York, N.Y.)* 194.4260 (1er oct. 1976), p. 23-28.

[11] Mel Greaves et Carlo C. Maley. « Clonal evolution in cancer ». In : *Nature* 481.7381 (jan. 2012). Number : 7381 Publisher : Nature Publishing Group, p. 306-313.

[12] Antonio Passaro et al. « Overcoming therapy resistance in EGFR-mutant lung cancer ». In : *Nature Cancer* 2.4 (avr. 2021). Number : 4 Publisher : Nature Publishing Group, p. 377-391.

[13] *Inherited genes and cancer types*. Cancer Research UK. 2 juin 2015. URL : https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/inherited-cancer-genes-and-increased-cancer-risk/inherited-genes-and-cancer-types (visité le 19/07/2023).

[14] *Quelques chiffres - Cancer du sein*. URL : https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Quelques-chiffres (visité le 19/07/2023).

[15] *Cancer de l'ovaire : les points clés - Cancer de l'ovaire*. URL : https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-de-l-ovaire/Les-points-cles (visité le 20/07/2023).

[16] Y. Miki et al. « A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1 ». In : *Science (New York, N.Y.)* 266.5182 (7 oct. 1994), p. 66-71.

[17] R. Wooster et al. « Identification of the breast cancer susceptibility gene BRCA2 ». In : *Nature* 378.6559 (21 déc. 1995), p. 789-792.

[18] Jessica Moretta et al. « [The French Genetic and Cancer Consortium guidelines for multigene panel analysis in hereditary breast and ovarian cancer predisposition] ». In : *Bulletin Du Cancer* 105.10 (oct. 2018), p. 907-917.

[19] Anna Öfverholm et al. « Extended genetic analysis and tumor characteristics in over 4600 women with suspected hereditary breast and ovarian cancer ». In : *BMC Cancer* 23.1 (10 août 2023), p. 738.

[20] Susan M. Domchek et Mark E. Robson. « Update on Genetic Testing in Gynecologic Cancer ». In : *Journal of Clinical Oncology* 37.27 (20 sept. 2019), p. 2501-2509.

[21] Reiko Yoshida. « Hereditary breast and ovarian cancer (HBOC) : review of its molecular characteristics, screening, treatment, and prognosis ». In : *Breast Cancer (Tokyo, Japan)* 28.6 (2021), p. 1167-1180.

[22] Sandrine M. Caputo et al. « Classification of 101 BRCA1 and BRCA2 variants of uncertain significance by cosegregation study : A powerful approach ». In : *The American Journal of Human Genetics* 108.10 (7 oct. 2021). Publisher : Elsevier, p. 1907-1923.

[23] Mattia Garutti et al. « Hereditary Cancer Syndromes : A Comprehensive Review with a Visual Tool ». In : *Genes* 14.5 (30 avr. 2023), p. 1025.

[24] *Le dispositif national d'oncogénétique - L'organisation de l'offre de soins*. URL : https://www.e-cancer.fr/Professionnels-de-sante/L-organisation-de-l-offre-de-soins/Oncogenetique-et-plateformes-de-genetique-moleculaire (visité le 03/09/2023).

[25] Karin Kast et al. « Prevalence of BRCA1/2 germline mutations in 21 401 families with breast and ovarian cancer ». In : *Journal of Medical Genetics* 53.7 (juill. 2016), p. 465-471.

[26] *Les prédispositions génétiques - Oncogénétique et plateformes de génétique moléculaire*. URL : https://www.e-cancer.fr/Professionnels-de-sante/L-organisation-de-l-offre-de-soins/Oncogenetique-et-plateformes-de-genetique-moleculaire/Les-predispositions-genetiques (visité le 03/09/2023).

[27] Naomi Wilcox et al. « Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk ». In : *Nature Genetics* 55.9 (sept. 2023). Number : 9 Publisher : Nature Publishing Group, p. 1435-1439.

[28] Yoshitaka Sakamoto, Sarun Sereewattanawoot et Ayako Suzuki. « A new era of long-read sequencing for cancer genomics ». In : *Journal of Human Genetics* 65.1 (2020), p. 3-10.

[29] *Le Plan cancer 2003-2007 - Les Plans cancer*. URL : https://www.e-cancer.fr/Institut-national-du-cancer/Strategie-de-lutte-contre-les-cancers-en-France/Les-Plans-cancer/Le-Plan-cancer-2003-2007 (visité le 03/09/2023).

[30] Debyani Chakravarty et al. « OncoKB : A Precision Oncology Knowledge Base ». In : *JCO Precision Oncology* 1 (nov. 2017). Publisher : Wolters Kluwer, p. 1-16.

[31] Malachi Griffith et al. « CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer ». In : *Nature Genetics* 49.2 (fév. 2017). Number : 2 Publisher : Nature Publishing Group, p. 170-174.

[32] Zbyslaw Sondka et al. « The COSMIC Cancer Gene Census : describing genetic dysfunction across all human cancers ». In : *Nature Reviews Cancer* 18.11 (nov. 2018). Number : 11 Publisher : Nature Publishing Group, p. 696-705.

[33] Sara E. Patterson et al. « The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies ». In : *Human Genomics* 10 (16 jan. 2016), p. 4.

[34] Sharon E. Plon et al. « Sequence variant classification and reporting : recommendations for improving the interpretation of cancer susceptibility genetic test results ». In : *Human mutation* 29.11 (nov. 2008), p. 1282-1291.

[35] Hongyan Li et al. « Classification of variants of uncertain significance in BRCA1 and BRCA2 using personal and family history of cancer from individuals in a large hereditary cancer multigene panel testing cohort ». In : *Genetics in Medicine* 22.4 (2020), p. 701-708.

[36] Sue Richards et al. « Standards and guidelines for the interpretation of sequence variants : a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. » In : *Genetics in medicine : official journal of the American College of Medical Genetics* 17.5 (mai 2015). Publisher : NIH Public Access, p. 405-24.

[37] Steven M. Harrison, Leslie G. Biesecker et Heidi L. Rehm. « Overview of specifications to the ACMG/AMP variant interpretation guidelines ». In : *Current protocols in human genetics* 103.1 (sept. 2019), e93.

[38] Emmanuelle Masson et al. « Expanding ACMG variant classification guidelines into a general framework ». In : *Human Genomics* 16 (16 août 2022), p. 31.

[39] Guy Froyen et al. « Standardization of somatic variant classifications in solid and haematological tumours by a two-level approach of biological and clinical classes : An initiative of the belgian compermed expert panel ». In : *Cancers* 11.12 (2019).

[40] Marilyn M. Li et al. « Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer ». In : *The Journal of Molecular Diagnostics : JMD* 19.1 (jan. 2017), p. 4-23.

[41] F. Roila et al. « Guideline update for MASCC and ESMO in the prevention of chemotherapy- and radiotherapy-induced nausea and vomiting : results of the Perugia consensus conference ». In : *Annals of Oncology* 21 (1er mai 2010). Publisher : Elsevier, p. v232-v243.

[42] P. Andrew Futreal et al. « A CENSUS OF HUMAN CANCER GENES ». In : *Nature reviews. Cancer* 4.3 (mars 2004), p. 177-183.

[43] Bert Vogelstein et al. « Cancer Genome Landscapes ». In : *Science (New York, N.Y.)* 339.6127 (29 mars 2013), p. 1546-1558.

[44] John N. Weinstein et al. « The Cancer Genome Atlas Pan-Cancer Analysis Project ». In : *Nature genetics* 45.10 (oct. 2013), p. 1113-1120.

[45] « International network of cancer genome projects ». In : *Nature* 464.7291 (15 avr. 2010), p. 993-998.

[46] Ege Ülgen et O. Uğur Sezerman. « driveR : a novel method for prioritizing cancer driver genes using somatic genomics data ». In : *BMC Bioinformatics* 22.1 (24 mai 2021), p. 263.

[47] Musalula Sinkala. « Mutational landscape of cancer-driver genes across human cancers ». In : *Scientific Reports* 13.1 (7 août 2023). Number : 1 Publisher : Nature Publishing Group, p. 12742.

[48] David Tamborero et al. « Comprehensive identification of mutational cancer driver genes across 12 tumor types ». In : *Scientific Reports* 3 (2 oct. 2013), p. 2650.

[49] Francisco Martínez-Jiménez et al. « A compendium of mutational cancer driver genes ». In : *Nature Reviews Cancer* 20.10 (oct. 2020). Number : 10 Publisher : Nature Publishing Group, p. 555-572.

[50] Alfred G. Knudson. « Mutation and Cancer : Statistical Study of Retinoblastoma ». In : *Proceedings of the National Academy of Sciences of the United States of America* 68.4 (avr. 1971), p. 820-823.

[51] David E. Comings. « A General Theory of Carcinogenesis ». In : *Proceedings of the National Academy of Sciences of the United States of America* 70.12 (déc. 1973), p. 3324-3328.

[52] Junya Toguchida et al. « Complete Genomic Sequence of the Human Retinoblastoma Susceptibility Gene ». In : *Genomics* 17.3 (1er sept. 1993), p. 535-543.

[53] D. Malkin et al. « Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms ». In : *Science (New York, N.Y.)* 250.4985 (30 nov. 1990), p. 1233-1238.

[54] Kun Tan, Dwayne G. Stupack et Miles F. Wilkinson. « Nonsense-mediated RNA decay : an emerging modulator of malignancy ». In : *Nature Reviews Cancer* 22.8 (août 2022). Number : 8 Publisher : Nature Publishing Group, p. 437-451.

[55] Li-Hui Wang et al. « Loss of Tumor Suppressor Gene Function in Human Cancer : An Overview ». In : *Cellular Physiology and Biochemistry : International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology* 51.6 (2018), p. 2647-2693.

[56] M. W. Anderson et al. « Role of proto-oncogene activation in carcinogenesis ». In : *Environmental Health Perspectives* 98 (nov. 1992), p. 13-24.

[57] Carlo M. Croce. « Oncogenes and Cancer ». In : *New England Journal of Medicine* 358.5 (31 jan. 2008). Publisher : Massachusetts Medical Society _eprint : https ://doi.org/10.1056/NEJMra072367, p. 502-511.

[58] P. Nowell, D. Hungerford et P. Nowell. « A minute chromosome in human chronic granulocytic leukemia ». In : *Science* (1960).

[59] J. D. Rowley. « Letter : A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining ». In : *Nature* 243.5405 (1er juin 1973), p. 290-293.

[60] Herbert T. Abelson et Louise S. Rabstein. « Lymphosarcoma : Virus-induced Thymic-independent Disease in Mice1 ». In : *Cancer Research* 30.8 (1er août 1970), p. 2213-2222.

[61] Ann Marie Pendergast. « The Abl family kinases : mechanisms of regulation and signaling ». In : *Advances in Cancer Research* 85 (2002), p. 51-100.

[62] J. Groffen et al. « Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22 ». In : *Cell* 36.1 (jan. 1984), p. 93-99.

[63] Sandra M. Swain, Mythili Shastry et Erika Hamilton. « Targeting HER2-positive breast cancer : advances and future directions ». In : *Nature Reviews Drug Discovery* 22.2 (fév. 2023). Number : 2 Publisher : Nature Publishing Group, p. 101-126.

[64] Olga Martínez-Sáez et Aleix Prat. « Current and Future Management of HER2-Positive Metastatic Breast Cancer ». In : *JCO Oncology Practice* 17.10 (oct. 2021). Publisher : Wolters Kluwer, p. 594-604.

[65] Mary Luz Uribe, Ilaria Marrocco et Yosef Yarden. « EGFR in Cancer : Signaling Mechanisms, Drugs, and Acquired Resistance ». In : *Cancers* 13.11 (1er juin 2021), p. 2748.

[66] Mohammed A. S. Abourehab et al. « Globally Approved EGFR Inhibitors : Insights into Their Syntheses, Target Kinases, Biological Activities, Receptor Interactions, and Metabolism ». In : *Molecules* 26.21 (4 nov. 2021), p. 6677.

[67] Domenico Salvatore, Massimo Santoro et Martin Schlumberger. « The importance of the RET gene in thyroid cancer and therapeutic implications ». In : *Nature Reviews Endocrinology* 17.5 (mai 2021). Number : 5 Publisher : Nature Publishing Group, p. 296-306.

[68] Ashleigh Porter et Deborah J. Wong. « Perspectives on the Treatment of Advanced Thyroid Cancer : Approved Therapies, Resistance Mechanisms, and Future Directions ». In : *Frontiers in Oncology* 10 (25 jan. 2021), p. 592202.

[69] John H. Strickler et al. « Diagnosis and Treatment of ERBB2-Positive Metastatic Colorectal Cancer : A Review ». In : *JAMA oncology* 8.5 (1er mai 2022), p. 760-769.

[70] Nannan Wang et al. « Emerging Role of ERBB2 in Targeted Therapy for Metastatic Colorectal Cancer : Signaling Pathways to Therapeutic Strategies ». In : *Cancers* 14.20 (jan. 2022). Number : 20 Publisher : Multidisciplinary Digital Publishing Institute, p. 5160.

[71] Valerie A. Schneider et al. *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.* Pages : 072116 Section : New Results. 30 août 2016.

[72] *Assembly Terminology - Genome Reference Consortium.* URL : `https://www.ncbi.nlm.nih.gov/grc/help/definitions/` (visité le 13/09/2023).

[73] Sergey Nurk et al. « The complete sequence of a human genome ». In : *Science* 376.6588 (avr. 2022). Publisher : American Association for the Advancement of Science, p. 44-53.

[74] Arang Rhie et al. « The complete sequence of a human Y chromosome ». In : *Nature* (23 août 2023). Publisher : Nature Publishing Group, p. 1-11.

[75] S Henikoff et J G Henikoff. « Amino acid substitution matrices from protein blocks. » In : *Proceedings of the National Academy of Sciences of the United States of America* 89.22 (15 nov. 1992), p. 10915-10919.

[76] Peter Ebert et al. « Haplotype-resolved diverse human genomes and integrated analysis of structural variation ». In : *Science (New York, N.Y.)* 372.6537 (2 avr. 2021), eabf7117.

[77] Peter D. Stenson et al. « The Human Gene Mutation Database : building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine ». In : *Human Genetics* 133.1 (2014), p. 1-9.

[78] Medhat Mahmoud et al. « Structural variant calling : the long and the short of it ». In : *Genome Biology* 20.1 (20 nov. 2019), p. 246.

[79] Alba Sanchis-Juan et al. « Complex structural variants in Mendelian disorders : identification and breakpoint resolution using short- and long-read genome sequencing ». In : *Genome Medicine* 10 (7 déc. 2018), p. 95.

[80] Michal Levy-Sakin et al. « Genome maps across 26 human populations reveal population-specific patterns of structural variation ». In : *Nature Communications* 10 (4 mars 2019), p. 1025.

[81] Rebecca I. Torene et al. « Mobile element insertion detection in 89,874 clinical exomes ». In : *Genetics in Medicine* 22.5 (mai 2020). Number : 5 Publisher : Nature Publishing Group, p. 974-978.

[82] Ryan E. Mills et al. « Which transposable elements are active in the human genome ? » In : *Trends in Genetics* 23.4 (1er avr. 2007). Publisher : Elsevier, p. 183-191.

[83] Geòrgia Escaramís, Elisa Docampo et Raquel Rabionet. « A decade of structural variants : description, history and methods to detect structural variation ». In : *Briefings in Functional Genomics* 14.5 (1er sept. 2015), p. 305-314.

[84] Chie Kikutake et Mikita Suyama. « Possible involvement of silent mutations in cancer pathogenesis and evolution ». In : *Scientific Reports* 13.1 (10 mai 2023). Number : 1 Publisher : Nature Publishing Group, p. 7593.

[85] D.Gareth R. Evans et al. « A Dominantly Inherited 5Ú Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer ». In : *American Journal of Human Genetics* 103.2 (2 août 2018), p. 213-220.

[86] Pascaline Gaildrat et al. « Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants ». In : *Methods in Molecular Biology (Clifton, N.J.)* 653 (2010), p. 249-257.

[87] Adam Shlien et David Malkin. « Copy number variations and cancer ». In : *Genome Medicine* 1.6 (16 juin 2009), p. 62.

[88] Beata Nowakowska. « Clinical interpretation of copy number variants in the human genome ». In : *Journal of Applied Genetics* 58.4 (nov. 2017). Publisher : Springer, p. 449-457.

[89] Kenzui Taniue et Nobuyoshi Akimitsu. « Fusion Genes and RNAs in Cancer Development ». In : *Non-Coding RNA* 7.1 (4 fév. 2021), p. 10.

[90] Ruth Nussinov, Chung-Jung Tsai et Hyunbum Jang. « A New View of Activating Mutations in Cancer ». In : *Cancer Research* 82.22 (15 nov. 2022), p. 4114-4123.

[91] Jen Jen Yeh et al. « KRAS/BRAF mutation status and ERK1/2 activation as biomarkers for MEK1/2 inhibitor therapy in colorectal cancer ». In : *Molecular cancer therapeutics* 8.4 (avr. 2009), p. 834-843.

[92] L. T. Chow et al. « An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA ». In : *Cell* 12.1 (sept. 1977), p. 1-8.

[93] Chun-Hao Su, Dhananjaya D et Woan-Yuh Tarn. « Alternative Splicing in Neurogenesis and Brain Development ». In : *Frontiers in Molecular Biosciences* 5 (12 fév. 2018), p. 12.

[94] Jiawei Ouyang et al. « The role of alternative splicing in human cancer progression ». In : *American Journal of Cancer Research* 11.10 (15 oct. 2021), p. 4642-4667.

[95] Luke Frankiw, David Baltimore et Guideng Li. « Alternative mRNA splicing in cancer immunotherapy ». In : *Nature Reviews Immunology* 19.11 (nov. 2019). Number : 11 Publisher : Nature Publishing Group, p. 675-687.

[96] Sophie C. Bonnal, Irene López-Oreja et Juan Valcárcel. « Roles and mechanisms of alternative splicing in cancer — implications for care ». In : *Nature Reviews Clinical Oncology* 17.8 (août 2020). Number : 8 Publisher : Nature Publishing Group, p. 457-474.

[97] Yuanjiao Zhang et al. « Alternative splicing and cancer : a systematic review ». In : *Signal Transduction and Targeted Therapy* 6.1 (24 fév. 2021). Number : 1 Publisher : Nature Publishing Group, p. 1-14.

[98] Robert F. Stanley et Omar Abdel-Wahab. « Dysregulation and therapeutic targeting of RNA splicing in cancer ». In : *Nature Cancer* 3.5 (mai 2022). Number : 5 Publisher : Nature Publishing Group, p. 536-546.

[99] Eric Wang et Iannis Aifantis. « RNA Splicing and Cancer ». In : *Trends in Cancer* 6.8 (août 2020), p. 631-644.

[100] Carolyn Horton et al. « Mutational and splicing landscape in a cohort of 43,000 patients tested for hereditary cancer ». In : *npj Genomic Medicine* 7.1 (25 août 2022). Number : 1 Publisher : Nature Publishing Group, p. 1-6.

[101] Dimitra Bouzarelou et al. « Reclassification of Splicing Gene Variants in Hereditary Cancer : Cases Report and Literature Review ». In : *In Vivo* 37.4 (1er juill. 2023). Publisher : International Institute of Anticancer Research Section : Review, p. 1432-1444.

[102] F. Sanger, S. Nicklen et A. R. Coulson. « DNA sequencing with chain-terminating inhibitors ». In : *Proceedings of the National Academy of Sciences* 74.12 (déc. 1977). Publisher : Proceedings of the National Academy of Sciences, p. 5463-5467.

[103] Ruqin Kou et al. « Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations ». In : *PLOS ONE* 11.1 (11 jan. 2016). Publisher : Public Library of Science, e0146638.

[104] Johnny A. Sena et al. « Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis ». In : *Scientific Reports* 8.1 (3 sept. 2018). Number : 1 Publisher : Nature Publishing Group, p. 13121.

[105] Ashley Byrne et al. « Realizing the potential of full-length transcriptome sequencing ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 374.1786 (25 nov. 2019), p. 20190097.

[106] Mark J. P. Chaisson et al. « Multi-platform discovery of haplotype-resolved structural variation in human genomes ». In : *Nature Communications* 10 (16 avr. 2019), p. 1784.

[107] Dafni A. Glinos et al. « Transcriptome variation in human tissues revealed by long-read sequencing ». In : *Nature* 608.7922 (août 2022). Number : 7922 Publisher : Nature Publishing Group, p. 353-359.

[108] Aaron M. Wenger et al. « Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome ». In : *Nature Biotechnology* 37.10 (oct. 2019). Number : 10 Publisher : Nature Publishing Group, p. 1155-1162.

[109] Glennis A. Logsdon, Mitchell R. Vollger et Evan E. Eichler. « Long-read human genome sequencing and its applications ». In : *Nature Reviews Genetics 2020 21 :10* 21.10 (juin 2020). Publisher : Nature Publishing Group, p. 597-614.

[110] Yunhao Wang et al. « Nanopore sequencing technology, bioinformatics and applications ». In : *Nature Biotechnology* 39.11 (nov. 2021). Number : 11 Publisher : Nature Publishing Group, p. 1348-1365.

[111] Huanle Liu et al. « Accurate detection of m6A RNA modifications in native RNA sequences ». In : *Nature Communications* 10.1 (9 sept. 2019). Number : 1 Publisher : Nature Publishing Group, p. 4079.

[112] Hengyun Lu, Francesca Giordano et Zemin Ning. « Oxford Nanopore MinION Sequencing and Genome Assembly ». In : *Genomics, Proteomics & Bioinformatics* 14.5 (oct. 2016), p. 265-279.

[113] Rory Bowden et al. « Sequencing of human genomes with nanopore technology ». In : *Nature Communications* 10.1 (23 avr. 2019). Number : 1 Publisher : Nature Publishing Group, p. 1869.

[114] Ying Chen et al. « Efficient assembly of nanopore reads via highly accurate and intact error correction ». In : *Nature Communications* 12.1 (4 jan. 2021). Number : 1 Publisher : Nature Publishing Group, p. 60.

[115] Isabel S. Naarman-de Vries et al. *Adaptive Sampling as tool for Nanopore direct RNA-sequencing*. Pages : 2022.10.14.512223 Section : New Results. 18 oct. 2022.

[116] Julien Lagarde et al. « High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing ». In : *Nature genetics* 49.12 (déc. 2017), p. 1731-1740.

[117] Simon A. Hardwick et al. « Targeted, High-Resolution RNA Sequencing of Non-coding Genomic Regions Associated With Neuropsychiatric Functions ». In : *Frontiers in Genetics* 10 (12 avr. 2019), p. 309.

[118] Vincent Schwenk et al. « Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome ». In : *Journal of Medical Genetics* 60.8 (août 2023), p. 747-759.

[119] Holly LaDuca et al. « A clinical guide to hereditary cancer panel testing : evaluation of gene-specific cancer associations and sensitivity of genetic testing criteria in a cohort of 165,000 high-risk patients ». In : *Genetics in Medicine* 22.2 (fév. 2020). Number : 2 Publisher : Nature Publishing Group, p. 407-415.

[120] A. Bayle et al. « ESMO Study on the Availability and Accessibility of Biomolecular Technologies in Oncology in Europe ». In : *Annals of Oncology* 0.0 (3 juill. 2023). Publisher : Elsevier.

[121] Hyunchul Jung, Kang Seon Lee et Jung Kyoon Choi. « Comprehensive characterisation of intronic mis-splicing mutations in human cancers ». In : *Oncogene* 40.7 (fév. 2021). Number : 7 Publisher : Nature Publishing Group, p. 1347-1361.

[122] Jin-Sun Ryu et al. « Exon splicing analysis of intronic variants in multigene cancer panel testing for hereditary breast/ovarian cancer ». In : *Cancer Science* 111.10 (oct. 2020), p. 3912-3925.

[123] Anna Stengel et al. « Whole transcriptome sequencing detects a large number of novel fusion transcripts in patients with AML and MDS ». In : *Blood Advances* 4.21 (10 nov. 2020), p. 5393-5401.

[124] Grégoire Davy et al. « Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer ». In : *European Journal of Human Genetics* 25.10 (oct. 2017). Number : 10 Publisher : Nature Publishing Group, p. 1147-1154.

[125] Erin E. Heyer et al. « Diagnosis of fusion genes using targeted RNA sequencing ». In : *Nature Communications* 10.1 (27 mars 2019). Number : 1 Publisher : Nature Publishing Group, p. 1388.

[126] Heng Li et Richard Durbin. « Fast and accurate short read alignment with Burrows–Wheeler transform ». In : *Bioinformatics* 25.14 (15 juill. 2009), p. 1754-1760.

[127] Alexander Dobin et al. « STAR : ultrafast universal RNA-seq aligner ». In : *Bioinformatics* 29.1 (jan. 2013), p. 15-21.

[128] Heng Li. « Minimap2 : pairwise alignment for nucleotide sequences ». In : *Bioinformatics* 34.18 (15 sept. 2018), p. 3094-3100.

[129] Heng Li et al. « The Sequence Alignment/Map format and SAMtools ». In : *Bioinformatics* 25.16 (août 2009), p. 2078-2079.

[130] Yury A. Barbitoff et al. « Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery ». In : *BMC Genomics* 23.1 (22 fév. 2022), p. 155.

[131] Chang Xu. « A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data ». In : *Computational and Structural Biotechnology Journal* 16 (6 fév. 2018), p. 15-24.

[132] Daniel C Koboldt et al. « VarScan 2 : somatic mutation and copy number alteration discovery in cancer by exome sequencing. » In : *Genome research* 22.3 (mars 2012). Publisher : Cold Spring Harbor Laboratory Press, p. 568-76.

[133] Zhongwu Lai et al. « VarDict : A novel and versatile variant caller for next-generation sequencing in cancer research ». In : *Nucleic Acids Research* 44.11 (2016). Publisher : Oxford University Press, e108-e108.

[134] Andy Rimmer et al. « Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications ». In : *Nature Genetics* 46.8 (août 2014). Number : 8 Publisher : Nature Publishing Group, p. 912-918.

[135] Scott L Carter et al. « Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples ». In : *Nature Biotechnology* 31.3 (mars 2013). Publisher : Nature Publishing Group, p. 213-219.

[136] Sangtae Kim et al. « Strelka2 : fast and accurate calling of germline and somatic variants ». In : *Nature Methods* 15.8 (août 2018). Number : 8 Publisher : Nature Publishing Group, p. 591-594.

[137] Daniel P. Cooke, David C. Wedge et Gerton Lunter. « A unified haplotype-based method for accurate and comprehensive variant calling ». In : *Nature Biotechnology* 39.7 (juill. 2021), p. 885-892.

[138] Aaron McKenna et al. « The genome analysis toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data ». In : *Genome Research* 20.9 (sept. 2010). Publisher : Cold Spring Harbor Laboratory Press, p. 1297-1303.

[139] Erik Garrison et Gabor Marth. *Haplotype-based variant detection from short-read sequencing*. 20 juill. 2012. arXiv : `1207.3907[q-bio]`.

[140] Ryan Poplin et al. *Scaling accurate genetic variant discovery to tens of thousands of samples*. Pages : 201178 Section : New Results. 24 juill. 2018.

[141] Petr Danecek et al. « The variant call format and VCFtools ». In : *Bioinformatics* 27.15 (août 2011). Publisher : Oxford University Press, p. 2156-2158.

[142] Jana Marie Schwarz et al. « MutationTaster2 : mutation prediction for the deep-sequencing age ». In : *Nature Methods* 11.4 (avr. 2014). Number : 4 Publisher : Nature Publishing Group, p. 361-362.

[143] Pauline C. Ng et Steven Henikoff. « Predicting Deleterious Amino Acid Substitutions ». In : *Genome Research* 11.5 (1er mai 2001). Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab, p. 863-874.

[144] Jonathan Frazer et al. « Disease variant prediction with deep generative models of evolutionary data ». In : *Nature* 599.7883 (nov. 2021). Number : 7883 Publisher : Nature Publishing Group, p. 91-95.

[145] Kishore Jaganathan et al. « Predicting Splicing from Primary Sequence with Deep Learning ». In : *Cell* 176.3 (24 jan. 2019). Publisher : Elsevier, 535-548.e24.

[146] Jennifer Harrow et al. « GENCODE : The reference human genome annotation for The ENCODE Project ». In : *Genome Research* 22.9 (sept. 2012), p. 1760-1774.

[147] Raphaël Leman et al. « SPiP : Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing ». In : *Human Mutation* 43.12 (déc. 2022), p. 2308-2323.

[148] William McLaren et al. « The Ensembl Variant Effect Predictor ». In : *Genome Biology* 17.1 (6 juin 2016), p. 122.

[149] Nilah M. Ioannidis et al. « REVEL : An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants ». In : *The American Journal of Human Genetics* 99.4 (6 oct. 2016). Publisher : Elsevier, p. 877-885.

[150] Philipp Rentzsch et al. « CADD : predicting the deleteriousness of variants throughout the human genome ». In : *Nucleic Acids Research* 47 (D1 8 jan. 2019), p. D886-D894.

[151] Martin Kircher et al. « A general framework for estimating the relative pathogenicity of human genetic variants ». In : *Nature Genetics* 46.3 (mars 2014). Number : 3 Publisher : Nature Publishing Group, p. 310-315.

[152] *Calculated consequences*. URL : `https://grch37.ensembl.org/info/genome/variation/prediction/predicted_data.html` (visité le 25/07/2023).

[153] Philipp Rentzsch et al. « CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores ». In : *Genome Medicine* 13.1 (22 fév. 2021), p. 31.

[154] Tatiana A. Gurbich et Valery Vladimirovich Ilinsky. « ClassifyCNV : a tool for clinical annotation of copy-number variants ». In : *Scientific Reports* 10.1 (23 nov. 2020). Number : 1 Publisher : Nature Publishing Group, p. 20375.

[155] Daniel R. Schrider et al. « Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans ». In : *PLOS Genetics* 9.1 (24 jan. 2013). Publisher : Public Library of Science, e1003242.

[156] Siwei Chen et al. *A genome-wide mutational constraint map quantified from variation in 76,156 human genomes*. Pages : 2022.03.20.485034 Section : New Results. 10 oct. 2022.

[157] *dbSNP Overview*. URL : `https://www.ncbi.nlm.nih.gov/projects/SNP/get_html.cgi?whichHtml=overview` (visité le 25/07/2023).

[158] Melissa J Landrum et al. « ClinVar : improving access to variant interpretations and supporting evidence ». In : *Nucleic Acids Research* 46 (D1 4 jan. 2018), p. D1062-D1067.

[159] Melissa S. Cline et al. « BRCA Challenge : BRCA Exchange as a global resource for variants in BRCA1 and BRCA2 ». In : *PLoS Genetics* 14.12 (26 déc. 2018), e1007752.

[160] *The TP53 Database | ISB-CGC*. URL : `https://tp53.isb-cgc.org/` (visité le 04/09/2023).

[161] Kelvin César de Andrade et al. « The TP53 Database : transition from the International Agency for Research on Cancer to the US National Cancer Institute ». In : *Cell Death & Differentiation* 29.5 (mai 2022). Number : 5 Publisher : Nature Publishing Group, p. 1071-1073.

[162] Véronique Geoffroy et al. « AnnotSV : An integrated tool for structural variations annotation ». In : *Bioinformatics* 34.20 (oct. 2018). Sous la dir. de Bonnie Berger. Publisher : Oxford University Press, p. 3572-3574.

[163] Nuala A. O'Leary et al. « Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation ». In : *Nucleic Acids Research* 44 (Database issue 4 jan. 2016), p. D733-D745.

[164] Joanna S Amberger et al. « OMIM.org : leveraging knowledge across phenotype–gene relationships ». In : *Nucleic Acids Research* 47 (D1 8 jan. 2019), p. D1038-D1043.

[165] Jeffrey R. MacDonald et al. « The Database of Genomic Variants : a curated collection of structural variation in the human genome ». In : *Nucleic Acids Research* 42 (Database issue 1er jan. 2014), p. D986-D992.

[166] Ilkka Lappalainen et al. « dbVar and DGVa : public archives for genomic structural variation ». In : *Nucleic Acids Research* 41 (Database issue jan. 2013), p. D936-D941.

[167] John G Tate et al. « COSMIC : the Catalogue Of Somatic Mutations In Cancer ». In : *Nucleic Acids Research* 47 (D1 8 jan. 2019), p. D941-D947.

[168] *MolecularMatch database*. URL : `https://www.molecularmatch.com/` (visité le 01/12/2022).

[169] Philip A. Ewels et al. « The nf-core framework for community-curated bioinformatics pipelines ». In : *Nature Biotechnology* 38.3 (mars 2020). Number : 3 Publisher : Nature Publishing Group, p. 276-278.

[170] Alex H. Wagner et al. « A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer ». In : *Nature Genetics* 52.4 (avr. 2020). Number : 4 Publisher : Nature Publishing Group, p. 448-457.

[171] Sebastian Uhrig et al. « Accurate and efficient detection of gene fusions from RNA sequencing data ». In : *Genome Research* 31.3 (1er mars 2021). Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab, p. 448-460.

[172] Brian J. Haas et al. « Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods ». In : *Genome Biology* 20.1 (21 oct. 2019), p. 213.

[173] Nadia M Davidson, Ian J Majewski et Alicia Oshlack. « JAFFA : High sensitivity transcriptome-focused fusion gene detection ». In : *Genome Medicine* 7.1 (11 mai 2015), p. 43.

[174] Richard I. Kuo et al. « Illuminating the dark side of the human transcriptome with long read transcript sequencing ». In : *BMC Genomics* 21.1 (30 oct. 2020), p. 751.

[175] Mihaela Pertea et al. « StringTie enables improved reconstruction of a transcriptome from RNA-seq reads ». In : *Nature Biotechnology* 33.3 (mars 2015). Number : 3 Publisher : Nature Publishing Group, p. 290-295.

[176] Sam Kovaka et al. « Transcriptome assembly from long-read RNA-seq alignments with StringTie2 ». In : *Genome Biology* 20.1 (16 déc. 2019), p. 278.

[177] Alison D. Tang et al. « Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns ». In : *Nature Communications* 11.1 (18 mars 2020). Number : 1 Publisher : Nature Publishing Group, p. 1438.

[178] PacBio. *Introduction of the Iso-Seq method : state of the art for full-length transcriptome sequencing*. PacBio. 3 mai 2018. URL : `https://www.pacb.com/blog/introduction-of-the-iso-seq-method-state-of-the-art-for-full-length-transcriptome-sequencing/` (visité le 04/09/2023).

[179] Jie Sun et al. « Germline Mutations in Cancer Susceptibility Genes in a Large Series of Unselected Breast Cancer Patients ». In : *Clinical Cancer Research* 23.20 (12 oct. 2017), p. 6113-6119.

[180] Preethi Srinivasan et al. « The context-specific role of germline pathogenicity in tumorigenesis ». In : *Nature Genetics* 53.11 (nov. 2021). Number : 11 Publisher : Nature Publishing Group, p. 1577-1585.

[181] Tao Qing et al. « Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden ». In : *Nature Communications* 11.1 (15 mai 2020). Number : 1 Publisher : Nature Publishing Group, p. 2438.

[182] Nadine Tung et al. « Potential pathogenic germline variant reporting from tumor comprehensive genomic profiling complements classic approaches to germline testing ». In : *npj Precision Oncology* 7.1 (11 août 2023). Number : 1 Publisher : Nature Publishing Group, p. 1-11.

[183] Franklin Gaylis et al. « Low Penetrance Germline Genetic Testing : Role for Risk Stratification in Prostate Cancer Screening and Examples From Clinical Practice ». In : *Reviews in Urology* 22.4 (2020), p. 152-158.

[184] Anusha Vaidyanathan et Virginia Kaklamani. « Understanding the Clinical Implications of Low Penetrant Genes and Breast Cancer Risk ». In : *Current Treatment Options in Oncology* 22.10 (23 août 2021), p. 85.

[185] Kadir C. Akdemir et al. « Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer ». In : *Nature Genetics* 52.3 (mars 2020). Number : 3 Publisher : Nature Publishing Group, p. 294-305.

[186] Andreas Hochhaus et al. « Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia ». In : *New England Journal of Medicine* 376.10 (9 mars 2017). Publisher : Massachusetts Medical Society, p. 917-927.

[187] David B. Fogel. « Factors associated with clinical trials that fail and opportunities for improving the likelihood of success : A review ». In : *Contemporary Clinical Trials Communications* 11 (7 août 2018), p. 156-164.

[188] Zikai Zhang et al. « Assessing clinical trial failure risk factors and reasons in gastric cancer ». In : *BMC Gastroenterology* 22 (30 nov. 2022), p. 496.

[189] Thomas Tursz et al. « Implications of personalized medicine—perspective from a cancer center ». In : *Nature Reviews Clinical Oncology* 8.3 (mars 2011). Number : 3 Publisher : Nature Publishing Group, p. 177-183.

[190] Htoo A. Wai et al. « Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance ». In : *Genetics in Medicine* 22.6 (juin 2020). Number : 6 Publisher : Nature Publishing Group, p. 1005-1014.

[191] Rebecca Truty et al. « Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation ». In : *The American Journal of Human Genetics* 108.4 (1er avr. 2021), p. 696-708.

[192] Adam M. Bournazos et al. « Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants ». In : *Genetics in Medicine* 24.1 (1er jan. 2022), p. 130-145.

[193] Aurélie Goyenvalle et al. « Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping ». In : *Science (New York, N.Y.)* 306.5702 (3 déc. 2004), p. 1796-1799.

[194] Laëtitia Meulemans et al. « Skipping Nonsense to Maintain Function : The Paradigm of BRCA2 Exon 12 ». In : *Cancer Research* 80.7 (2 avr. 2020), p. 1374-1386.

[195] Irene Lopez-Perolio et al. « Alternative splicing and ACMG-AMP-2015-based classification of PALB2 genetic variants : an ENIGMA report ». In : *Journal of Medical Genetics* 56.7 (1er juill. 2019). Publisher : BMJ Publishing Group Ltd Section : Cancer genetics, p. 453-460.

[196] Tamara Steijger et al. « Assessment of transcript reconstruction methods for RNA-seq ». In : *Nature Methods* 10.12 (déc. 2013). Number : 12 Publisher : Nature Publishing Group, p. 1177-1184.

[197] Fergal J Martin et al. « Ensembl 2023 ». In : *Nucleic Acids Research* 51 (D1 6 jan. 2023), p. D933-D941.

[198] Atish D. Choudhury et al. « A Phase I Study Investigating AZD8186, a Potent and Selective Inhibitor of PI3Kβ/δ, in Patients with Advanced Solid Tumors ». In : *Clinical Cancer Research* 28.11 (1er juin 2022), p. 2257-2269.

[199] Mingxiang Liao et al. « Population Pharmacokinetic Modeling of Lucitanib in Patients with Advanced Cancer ». In : *European Journal of Drug Metabolism and Pharmacokinetics* 47.5 (2022), p. 711-723.

[200] Mihaela Pertea et al. « Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown ». In : *Nature Protocols* 11.9 (sept. 2016). Number : 9 Publisher : Nature Publishing Group, p. 1650-1667.

[201] Leandro Lima et al. « Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data ». In : *Briefings in Bioinformatics* 21.4 (15 juill. 2020), p. 1164-1181.

[202] Haowen Zhang, Chirag Jain et Srinivas Aluru. « A comprehensive evaluation of long read error correction methods ». In : *BMC Genomics* 21 (Suppl 6 21 déc. 2020), p. 889.

[203] Liisa Chang et al. « Microsatellite Instability : A Predictive Biomarker for Cancer Immunotherapy ». In : *Applied Immunohistochemistry and Molecular Morphology* 26.2 (2018). Publisher : Lippincott Williams and Wilkins, e15-e21.

[204] Ronald J. Hause et al. « Classification and characterization of microsatellite instability across 18 cancer types ». In : *Nature Medicine* 22.11 (nov. 2016). Publisher : Nature Publishing Group, p. 1342-1350.

[205] Martina Amato et al. « Microsatellite Instability : From the Implementation of the Detection to a Prognostic and Predictive Role in Cancers ». In : *International Journal of Molecular Sciences* 23.15 (jan. 2022). Number : 15 Publisher : Multidisciplinary Digital Publishing Institute, p. 8726.

[206] Michael J. Fusco, Howard (Jack) West et Christine M. Walko. « Tumor Mutation Burden and Cancer Treatment ». In : *JAMA Oncology* 7.2 (1er fév. 2021), p. 316.

[207] Paolo Manca et al. « Tumor mutational burden as a biomarker in patients with dMMR/MSI-H metastatic colorectal cancer treated with immune checkpoint inhibitors ». In : *European Journal of Cancer* 0.0 (31 mars 2023). Publisher : Elsevier.

[208] Giorgio Valabrega et al. « Differences in PARP Inhibitors for the Treatment of Ovarian Cancer : Mechanisms of Action, Pharmacology, Safety, and Efficacy ». In : *International Journal of Molecular Sciences* 22.8 (19 avr. 2021), p. 4203.

[209] A. Echle et al. « Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application ». In : *ESMO Open* 7.2 (2 mars 2022), p. 100400.

[210] Beifang Niu et al. « MSIsensor : microsatellite instability detection using paired tumor-normal sequence data ». In : *Bioinformatics* 30.7 (1er avr. 2014), p. 1015-1016.

[211] Lijing Yao et al. « ecTMB : a robust method to estimate and classify tumor mutational burden ». In : *Scientific Reports* 10.1 (18 mars 2020). Number : 1 Publisher : Nature Publishing Group, p. 4983.

[212] Raphaël Leman et al. « Validation of the Clinical Use of GIScar, an Academic-developed Genomic Instability Score Predicting Sensitivity to Maintenance Olaparib for Ovarian Cancer ». In : *Clinical Cancer Research* (26 sept. 2023), OF1-OF11.

[213] Doga C. Gulhan et al. « Detecting the mutational signature of homologous recombination deficiency in clinical samples ». In : *Nature Genetics* 51.5 (mai 2019). Publisher : Nature Publishing Group, p. 912-919.

[214] Yumei Feng et al. « Evidence for a transcriptional signature of breast cancer ». In : *Breast Cancer Research and Treatment* 122.1 (juill. 2010), p. 65-75.

[215] Hao Cai et al. « A qualitative transcriptional signature to reclassify estrogen receptor status of breast cancer patients ». In : *Breast Cancer Research and Treatment* 170.2 (juill. 2018), p. 271-277.

[216] Guillaume Beinse et al. « Discovery and validation of a transcriptional signature identifying homologous recombination-deficient breast, endometrial and ovarian cancers ». In : *British Journal of Cancer* 127.6 (5 oct. 2022), p. 1123-1132.

[217] Mostafa Abbas et Yasser EL-Manzalawy. « Machine learning based refined differential gene expression analysis of pediatric sepsis ». In : *BMC Medical Genomics* 13.1 (28 août 2020), p. 122.

[218] Anupama Jha et al. « Identifying common transcriptome signatures of cancer by interpreting deep learning models ». In : *Genome Biology* 23.1 (17 mai 2022), p. 117.

[219] Tulika Kakati et al. « DEGnext : classification of differentially expressed genes from RNA-seq data using a convolutional neural network with transfer learning ». In : *BMC Bioinformatics* 23.1 (6 jan. 2022), p. 17.

Développement de méthodes bioinformatiques d'analyses combinatoires du patrimoine génétique de la tumeur et de son hôte : intérêt en diagnostic moléculaire et recherche transversale en cancérologie.

Résumé

Le séquençage à haut débit *short-read* d'ADN a favorisé une médecine personnalisée par la génomique, permettant d'identifier des anomalies moléculaires qui augmentent le risque de cancer ou guident la prise en charge thérapeutique. L'interprétation de ces variants est cruciale, notamment pour les cancers, où la diversité des indications thérapeutiques nécessite des panels de gènes larges et des résultats rapides pour répondre aux contraintes de prise en charge des patients.

Nous avons développé DrugOrder un outil qui permet d'associer l'ensemble des thérapies disponibles pour des variants (SNV, CNV et fusions de gènes), dits actionnables. DrugOrder propose une classification basée sur le niveau de preuve par rapport à la thérapie considérée et l'impact du variant dans l'oncogénèse. DrugOrder est également capable d'identifier une similarité entre un nouveau variant non décrit et un variant équivalent déjà présent dans une base de données. DrugOrder a été évalué sur deux jeux de données, simulés et réels pour lesquels un ensemble de variants actionnables était préalablement connus. Cet outil a montré sa capacité à prioriser ces variants actionnables avec une haute efficacité.

En complément du séquençage *short-read*, de nouvelles technologies innovantes de séquençage *long-read* de l'ADN ou de l'ARN offrent la possibilité d'explorer des évènements complexes, composés de variants structuraux ou d'isoformes alternatives d'ARNm. Par conséquent, il est nécessaire de constituer des méthodologies d'analyses intégratives permettant de détecter ces événements à partir des données issues de séquençage *long-read*.

Avec le développement de LoRID, nous avons résolu une partie de ce défi pour les patientes atteintes du syndrome *Hereditary Breast and Ovarian Cancer* (HBOC). LoRID est un pipeline complet qui permet l'assemblage d'isoformes alternatives de l'ARN ainsi que leur annotation via un module appelé SOSTAR. Ce dernier permet de décrire les différents événements d'épissage composant une isoforme donnée, tout en quantifiant l'abondance de chaque isoforme. Notre outil a pu identifier un rétrotransposon SVA dans une famille atteinte du syndrome HBOC sans anomalies moléculaires connues, expliquant ainsi le trait génétique de ce syndrome pour cette famille.

Mots clés :
Séquençage haut-débit, Bioinformatique, Diagnostic moléculaire, ADN, ARN, Prédisposition aux cancers, Médecine de précision.

Abstract

High-throughput sequencing has encouraged personalised medicine through genomics, making it possible to identify molecular anomalies that increase the risk of cancer or guide therapeutic management. The interpretation of these variants is crucial, particularly for cancers, where the diversity of therapeutic indications requires large gene panels and rapid results to meet patient management constraints.

We have developed DrugOrder, a tool that combines all available therapies for actionable variants (SNV, CNV and gene fusions). DrugOrder offers a classification based on the level of evidence for the therapy under consideration and the impact of the variant in oncogenesis. DrugOrder is also capable of identifying similarities between a new, undescribed variant and an equivalent variant already present in a database. DrugOrder was evaluated on two datasets, simulated and real, for which a set of actionable variants was previously known. The tool demonstrated its ability to prioritise these actionable variants effectively.

In addition to short-read sequencing, new innovative technologies for long-read sequencing of DNA or mRNA offer the possibility of exploring complex events composed of structural variants or alternative isoforms of mRNA. Consequently, there is a need to develop integrative analysis methodologies that can detect these events from long-read sequencing data.

With the development of LoRID, we have solved part of this challenge for patients with Hereditary Breast and Ovarian Cancer (HBOC) syndrome. LoRID is a complete pipeline that enables the assembly of alternative RNA isoforms and their annotation with a module called SOSTAR. The latter can be used to describe the different splicing events making up a given isoform, while quantifying the abundance of each isoform. Our tool was able to identify a SVA retrotransposon in a family suffering from HBOC syndrome with no known molecular anomalies, thus explaining the genetic trait of this syndrome in this family.

Keywords :
High-throughput sequencing, Bioinformatics, Molecular diagnostics, DNA, RNA, Cancer predisposition, Precision medicine

Sample	Location	Event	Gene	Chrom	Start	NM	HGVS dna	HGVS protein	VAF (%)	Drug	Level (ASCO)	Level (OncoKI	Level (NH)	Source
RCP01	Lung	Fusion	ALK			NM_004304				Alectininb	A	1	1	OncoKB
RCP01	Lung	Fusion	ALK			NM_004304				Brigatinib	A	1	1	OncoKB
RCP01	Lung	Fusion	ALK			NM_004304				Ceritinib	A	1	1	OncoKB
RCP01	Lung	Fusion	ALK			NM_004304				Crizotinib	A	1	1	OncoKB
RCP01	Lung	SNV	ATM	11	108098364	NM_000051	c.15dup	p.N6Ter	48,4	Olaparib	D	4	2	OncoKB
RCP01	Lung	SNV	RAD51D	17	33434136	NM_002878	c.350_351del	p.C117SfsTer36	37,7	Olaparib	D	4	3	OncoKB
RCP02	Lung	Fusion	ROS1			NM_002944				Crizotinib	A	1	1	OncoKB
RCP02	Lung	Fusion	ROS1			NM_002944				Entrectinib	A	1	1	OncoKB
RCP02	Lung	Fusion	ROS1			NM_002944				Ceritinib	A	2	2	OncoKB
RCP02	Lung	Fusion	ROS1			NM_002944				Lorlatinib	A	2	2	OncoKB
RCP02	Lung	Fusion	ROS1			NM_002944				Reprotectinib	B	3a	3	OncoKB
RCP03	Lung	SNV	KRAS	12	25380276	NM_004985	c.182A>T	p.Q61L	9,72	Tremetinib	D	4	1	OncoKB
RCP03	Lung	SNV	KRAS	12	25380276	NM_004985	c.182A>T	p.Q61L	9,72	Cobimetinib	D	4	1	OncoKB
RCP03	Lung	SNV	KRAS	12	25380276	NM_004985	c.182A>T	p.Q61L	9,72	Binimetinib	D	4	1	OncoKB
RCP03	Lung	SNV	STK11	19	1219336	NM_000455	c.388G>T	p.E130Ter	8,63	Bemcentinib + P	D	4	2	OncoKB
RCP04	Lung	SNV	CDKN2A	9	21971108	NM_000077	c.250G>A	p.D84N	5,79	Abemaciclib	D	4	1	OncoKB
RCP04	Lung	SNV	CDKN2A	9	21971108	NM_000077	c.250G>A	p.D84N	5,79	Palbociclib	D	4	1	OncoKB
RCP04	Lung	SNV	CDKN2A	9	21971108	NM_000077	c.250G>A	p.D84N	5,79	Ribociclib	D	4	1	OncoKB
RCP04	Lung	SNV	ATM	11	108151710	NM_000051	c.3403-12_3403-2delinsATTTCTTTTTAT	None	59,8	Olaparib	D	4	2	OncoKB
RCP05	Lung	Fusion	ROS1			NM_002944				Crizotinib	A	1	1	OncoKB
RCP05	Lung	Fusion	ROS1			NM_002944				Entrectinib	A	1	1	OncoKB
RCP05	Lung	Fusion	ROS1			NM_002944				Ceritinib	A	2	2	OncoKB
RCP05	Lung	Fusion	ROS1			NM_002944				Lorlatinib	A	2	2	OncoKB
RCP05	Lung	Fusion	ROS1			NM_002944				Reprotectinib	B	3a	3	OncoKB
RCP06	Lung	SNV	EGFR	7	55249002	NM_005828	c.2303_2311dup	p.S768_D770dup	8,62	Amivantanamb	A	1	1	OncoKB
RCP06	Lung	SNV	EGFR	7	55249002	NM_005828	c.2303_2311dup	p.S768_D770dup	8,62	Mobocertinib	A	1	1	OncoKB
RCP06	Lung	SNV	PTEN	10	89720649	NM_000314	c.802-2_804delinsTTGGAA	None	4,23	GSK2636771	D	4	2	OncoKB
RCP06	Lung	SNV	PTEN	10	89720649	NM_000314	c.802-2_804delinsTTGGAA	None	4,23	AZD8186	D	4	2	OncoKB
RCP07	Lung	Fusion	ALK			NM_004304				Alectininb	A	1	1	OncoKB
RCP07	Lung	Fusion	ALK			NM_004304				Brigatinib	A	1	1	OncoKB
RCP07	Lung	Fusion	ALK			NM_004304				Ceritinib	A	1	1	OncoKB
RCP07	Lung	Fusion	ALK			NM_004304				Ceritinib	A	1	1	OncoKB
RCP07	Lung	Fusion	ALK			NM_004304				Crizotinib	A	1	1	OncoKB
RCP07	Lung	SNV	FGFR1	8	38275855	NM_015850	c.1315G>T	p.V439F	44,7	AZD4547	D	4	2	OncoKB
RCP07	Lung	SNV	FGFR1	8	38275855	NM_015850	c.1315G>T	p.V439F	44,7	Erdafitinib	D	4	2	OncoKB
RCP07	Lung	SNV	FGFR1	8	38275855	NM_015850	c.1315G>T	p.V439F	44,7	Debio1347	D	4	2	OncoKB
RCP07	Lung	SNV	FGFR1	8	38275855	NM_015850	c.1315G>T	p.V439F	44,7	Infigratinib	D	4	2	OncoKB
RCP07	Lung	SNV	ARID1A	1	27023696	NM_006015	c.802C>T	p.Q268Ter	8,88	PLX2853	D	4	3	OncoKB
RCP07	Lung	SNV	ARID1A	1	27023696	NM_006015	c.802C>T	p.Q268Ter	8,88	Tazemetostat	D	4	3	OncoKB
RCP07	Lung	SNV	PTEN	10	89720648	NM_000314	c.802-3_816delinsAGGACAAAATGTTTCAA	None	5,8	GSK2636771	D	4	4	OncoKB
RCP07	Lung	SNV	PTEN	10	89720648	NM_000314	c.802-3_816delinsAGGACAAAATGTTTCAA	None	5,8	AZD8186	D	4	4	OncoKB
RCP09	Lung	SNV	CDKN2A	9	21971000	NM_000077	c.358G>T	p.E120Ter	30,9	Abemaciclib	D	4	1	OncoKB
RCP09	Lung	SNV	CDKN2A	9	21971000	NM_000077	c.358G>T	p.E120Ter	30,9	Palbociclib	D	4	1	OncoKB
RCP09	Lung	SNV	CDKN2A	9	21971000	NM_000077	c.358G>T	p.E120Ter	30,9	Ribociclib	D	4	1	OncoKB
RCP11	Bladder	SNV	FGFR3	4	1806119	NM_000142	c.1138G>A	p.G380R	6,45	Erdafinib	B	3a	1	OncoKB
RCP11	Bladder	SNV	FGFR3	4	1806119	NM_000142	c.1138G>A	p.G380R	6,45	Debio1347	D	4	2	OncoKB
RCP11	Bladder	SNV	FGFR3	4	1806119	NM_000142	c.1138G>A	p.G380R	6,45	Infigratinib	D	4	2	OncoKB
RCP11	Bladder	SNV	FGFR3	4	1806119	NM_000142	c.1138G>A	p.G380R	6,45	AZD4547	D	4	2	OncoKB
RCP11	Bladder	SNV	ERBB2	17	37868208	NM_004448	c.929C>T	p.S310F	8,71	Trastuzumab De	D	4	3	OncoKB
RCP11	Bladder	SNV	ERBB2	17	37868208	NM_004448	c.929C>T	p.S310F	8,71	Ado-Trastuzuma	D	4	3	OncoKB
RCP11	Bladder	SNV	ERBB2	17	37868208	NM_004448	c.929C>T	p.S310F	8,71	Neratinib	D	4	3	OncoKB
RCP11	Bladder	SNV	ERBB2	17	37868208	NM_004448	c.929C>T	p.S310F	8,71	Pertuzumab + D	D	4	3	OncoKB
RCP12	Bladder	SNV	FGFR2	10	123279677	NM_000141	c.755C>T	p.S252L	15,2	AZD4547	D	4	1	OncoKB
RCP12	Bladder	SNV	FGFR2	10	123279677	NM_000141	c.755C>T	p.S252L	15,2	Erdafinib	D	4	1	OncoKB
RCP12	Bladder	SNV	FGFR2	10	123279677	NM_000141	c.755C>T	p.S252L	15,2	Debio1347	D	4	1	OncoKB
RCP12	Bladder	SNV	FGFR2	10	123279677	NM_000141	c.755C>T	p.S252L	15,2	Infigratinib	D	4	1	OncoKB
RCP12	Bladder	SNV	ARID1A	1	27023805	NM_006015	c.911C>A	p.S304Ter	5,45	PLX2853	D	4	2	OncoKB
RCP12	Bladder	SNV	ARID1A	1	27023805	NM_006015	c.911C>A	p.S304Ter	5,45	Tazemetostat	D	4	2	OncoKB

RCP13	Bladder	SNV	FGFR3		4	1807889	NM_000142	c.1948A>G	p.K650E	47,3	Erdafinib	B	3a		1	OncoKB
RCP13	Bladder	SNV	FGFR3		4	1807889	NM_000142	c.1948A>G	p.K650E	47,3	Debio1347	D		4	2	OncoKB
RCP13	Bladder	SNV	FGFR3		4	1807889	NM_000142	c.1948A>G	p.K650E	47,3	Infigratinib	D		4	2	OncoKB
RCP13	Bladder	SNV	FGFR3		4	1807889	NM_000142	c.1948A>G	p.K650E	47,3	AZD4547	D		4	2	OncoKB
RCP13	Bladder	SNV	KDM6A	X		44918644	NM_021140	c.1129_1130dup	p.N377KfsTer63	47	Tazemetostat	D		4	3	OncoKB
RCP13	Bladder	SNV	PIK3CA		3	178952085	NM_006218	c.3140A>G	p.H1047R	23,6	RLY-2608	D		4	4	OncoKB
RCP13	Bladder	SNV	PIK3CA		3	178952085	NM_006218	c.3140A>G	p.H1047R	23,6	LOXO-783	D		4	4	OncoKB
RCP15	Bladder	SNV	KDM6A	X		44929164	NM_021140	c.2264del	p.T755RfsTer19	27,8	Tazemetostat	D		4	1	OncoKB
RCP17	Bladder	SNV	PTEN		10	89725053	NM_000314	c.1036T>A	p.Y346N	30,3	GSK2636771	D		4	1	OncoKB
RCP17	Bladder	SNV	PTEN		10	89725053	NM_000314	c.1036T>A	p.Y346N	30,3	AZD8186	D		4	1	OncoKB
RCP17	Bladder	SNV	BRCA1		17	41223175	NM_007294	c.4755dup	p.E1586RfsTer36	37,6	Niraparib	D		4	2	OncoKB
RCP17	Bladder	SNV	BRCA1		17	41223175	NM_007294	c.4755dup	p.E1586RfsTer36	37,6	Olaparib + Beva	D		4	2	OncoKB
RCP17	Bladder	SNV	BRCA1		17	41223175	NM_007294	c.4755dup	p.E1586RfsTer36	37,6	Olaparib	D		4	2	OncoKB
RCP17	Bladder	SNV	BRCA1		17	41223175	NM_007294	c.4755dup	p.E1586RfsTer36	37,6	Rucaparib	D		4	2	OncoKB
RCP17	Bladder	SNV	BRCA1		17	41223175	NM_007294	c.4755dup	p.E1586RfsTer36	37,6	Talazoparib	D		4	2	OncoKB
RCP19	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	21,7	Erdafitinib	A		1	1	OncoKB
RCP19	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	21,7	Debio1347	D		4	2	OncoKB
RCP19	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	21,7	Infigratinib	D		4	2	OncoKB
RCP19	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	21,7	AZD4547	D		4	2	OncoKB
RCP20	Bladder	SNV	ARID1A		1	27101373	NM_006015	c.4656_4657insGGGGGGGGGGGGG	p.P1553GfsTer17	7,97	PLX2853	D		4	1	OncoKB
RCP20	Bladder	SNV	ARID1A		1	27101373	NM_006015	c.4656_4657insGGGGGGGGGGGGG	p.P1553GfsTer17	7,97	Tazemetostat	D		4	1	OncoKB
RCP21	Bladder	Fusion	FGFR3				NM_000142				Erdafitinib	A		1	1	OncoKB
RCP21	Bladder	SNV	KDM6A	X		44937702	NM_021140	c.2891_2894del	p.N964TfsTer5	26,4	Tazemetostat	D		4	2	OncoKB
RCP21	Bladder	SNV	KDM6A	X		44937686	NM_021140	c.2874_2887del	p.Q958HfsTer16	26,7	Tazemetostat	D		4	2	OncoKB
RCP22	Bladder	SNV	KDM6A	X		44949124	NM_021140	c.3685C>T	p.Q1229Ter	40,2	Tazemetostat	D		4	1	OncoKB
RCP22	Bladder	SNV	MTOR		1	11204745	NM_004958	c.4832G>A	p.R1611Q	19,6	Everolimus	D		4	2	OncoKB
RCP22	Bladder	SNV	MTOR		1	11204745	NM_004958	c.4832G>A	p.R1611Q	19,6	Temsirolimus	D		4	2	OncoKB
RCP23	Bladder	SNV	TSC1		9	135786500	NM_000368	c.1030G>C	p.A344P	34,9	Everolimus	D		4	1	OncoKB
RCP23	Bladder	SNV	KDM6A	X		44949167	NM_021140	c.3728C>A	p.P1243Q	4,48	Tazemetostat	D		4	2	OncoKB
RCP24	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	30	Trastuzumab De	D		4	1	OncoKB
RCP24	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	30	Ado-Trastuzuma	D		4	1	OncoKB
RCP24	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	30	Neratinib	D		4	1	OncoKB
RCP24	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	30	Pertuzumab + D	D		4	1	OncoKB
RCP24	Bladder	SNV	ARID1A		1	27057961	NM_006015	c.1669C>T	p.Q557Ter	37,1	PLX2853	D		4	2	OncoKB
RCP24	Bladder	SNV	ARID1A		1	27057961	NM_006015	c.1669C>T	p.Q557Ter	37,1	Tazemetostat	D		4	2	OncoKB
RCP25	Bladder	SNV	KDM6A	X		44969323	NM_021140	c.4006-1G>A	None	51,8	Tazemetostat	D		4	1	OncoKB
RCP25	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	14	Trastuzumab De	D		4	2	OncoKB
RCP25	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	14	Ado-Trastuzuma	D		4	2	OncoKB
RCP25	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	14	Neratinib	D		4	2	OncoKB
RCP25	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	14	Pertuzumab + D	D		4	2	OncoKB
RCP25	Bladder	SNV	KRAS		12	25398284	NM_015083	c.35G>A	p.G12D	25,2	Trametinib	D		4	3	OncoKB
RCP25	Bladder	SNV	KRAS		12	25398284	NM_015083	c.35G>A	p.G12D	25,2	Cobimetinib	D		4	3	OncoKB
RCP25	Bladder	SNV	KRAS		12	25398284	NM_015083	c.35G>A	p.G12D	25,2	Binimetinib	D		4	3	OncoKB
RCP25	Bladder	SNV	KRAS		12	25398284	NM_015083	c.35G>A	p.G12D	25,2	RMC-6236	D		4	3	OncoKB
RCP25	Bladder	SNV	ARID1A		1	27100919	NM_006015	c.4201C>T	p.Q1401Ter	29,1	PLX2853	D		4	4	OncoKB
RCP25	Bladder	SNV	ARID1A		1	27100919	NM_006015	c.4201C>T	p.Q1401Ter	29,1	Tazemetostat	D		4	4	OncoKB
RCP25	Bladder	SNV	RAD54L		1	46739413	NM_003579	c.1607_1610+42del	None	39,6	Olaparib	D		4	5	OncoKB
RCP25	Bladder	SNV	PIK3CA		3	178936091	NM_006218	c.1633G>A	p.E545K	39,9	RLY-2608	D		4	6	OncoKB
RCP25	Bladder	SNV	PIK3CA		3	178936091	NM_006218	c.1633G>A	p.E545K	39,9	LOXO-783	D		4	6	OncoKB
RCP25	Bladder	SNV	PIK3CA		3	178938934	NM_006218	c.2176G>A	p.E726K	12	RLY-2608	D		4	6	OncoKB
RCP25	Bladder	SNV	PIK3CA		3	178938934	NM_006218	c.2176G>A	p.E726K	12	LOXO-783	D		4	6	OncoKB
RCP26	Endometrium	SNV	PTEN		10	89624245	NM_000314	c.19G>T	p.E7Ter	6,3	GSK2636771	D		4	1	OncoKB
RCP26	Endometrium	SNV	PTEN		10	89624245	NM_000314	c.19G>T	p.E7Ter	6,3	AZD8186	D		4	1	OncoKB
RCP26	Endometrium	SNV	PTEN		10	89711899	NM_000314	c.517C>T	p.R173C	9,4	GSK2636771	D		4	1	OncoKB
RCP26	Endometrium	SNV	PTEN		10	89711899	NM_000314	c.517C>T	p.R173C	9,4	AZD8186	D		4	1	OncoKB
RCP26	Endometrium	SNV	ARID1A		1	27106354	NM_006015	c.5965C>T	p.R1989Ter	9,05	PLX2853	D		4	2	OncoKB
RCP26	Endometrium	SNV	ARID1A		1	27106354	NM_006015	c.5965C>T	p.R1989Ter	9,05	Tazemetostat	D		4	2	OncoKB
RCP26	Endometrium	SNV	ARID1A		1	27105550	NM_006015	c.5161C>T	p.R1721Ter	6,22	PLX2853	D		4	2	OncoKB
RCP26	Endometrium	SNV	ARID1A		1	27105550	NM_006015	c.5161C>T	p.R1721Ter	6,22	Tazemetostat	D		4	2	OncoKB

RCP36	Lung	Fusion	ALK			NM_004304				Lorlatininb	A		1	1	OncoKB	
RCP37	Lung	Fusion	RET			NM_020975				Selpercatinib	A		1	1	OncoKB	
RCP37	Lung	Fusion	RET			NM_020975				Pralsetinib	A		1	1	OncoKB	
RCP38	Salivary gland	Fusion	NTRK3			NM_001012338				Entrectinib	A		1	1	OncoKB	
RCP38	Salivary gland	Fusion	NTRK3			NM_001012338				Larotrectinib	A		1	1	OncoKB	
RCP38	Salivary gland	SNV	ARID1A		1	27056207	NM_006015	c.1204_1205del	p.S402AfsTer220	10,5 PLX2853	D		4	2	OncoKB	
RCP38	Salivary gland	SNV	ARID1A		1	27056207	NM_006015	c.1204_1205del	p.S402AfsTer220	10,5 Tazemetostat	D		4	2	OncoKB	
RCP39	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	20,1 Erdafitinib	A		1	1	OncoKB	
RCP39	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	20,1 Debio1347	D		4	2	OncoKB	
RCP39	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	20,1 Infigratinib	D		4	2	OncoKB	
RCP39	Bladder	SNV	FGFR3		4	1806099	NM_000142	c.1118A>G	p.Y373C	20,1 AZD4547	D		4	2	OncoKB	
RCP39	Bladder	SNV	ARID1A		1	27097814	NM_006015	c.3404del	p.P1135LfsTer26	25,6 PLX2853	D		4	3	OncoKB	
RCP39	Bladder	SNV	ARID1A		1	27097814	NM_006015	c.3404del	p.P1135LfsTer26	25,6 Tazemetostat	D		4	3	OncoKB	
RCP41	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>A	p.S310Y	6,77 Trastuzumab De	D		4	1	OncoKB	
RCP41	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>A	p.S310Y	6,77 Ado-Trastuzuma	D		4	1	OncoKB	
RCP41	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>A	p.S310Y	6,77 Neratinib	D		4	1	OncoKB	
RCP41	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>A	p.S310Y	6,77 Pertuzumab + D	D		4	1	OncoKB	
RCP42	Bladder	SNV	KDM6A	X		44894201	NM_021140	c.593_619+17del	None	18,5 Tazemetostat	D		4	1	OncoKB	
RCP42	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	17,7 Trastuzumab De	D		4	2	OncoKB	
RCP42	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	17,7 Ado-Trastuzuma	D		4	2	OncoKB	
RCP42	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	17,7 Neratinib	D		4	2	OncoKB	
RCP42	Bladder	SNV	ERBB2		17	37868208	NM_004448	c.929C>T	p.S310F	17,7 Pertuzumab + D	D		4	2	OncoKB	
RCP42	Bladder	SNV	ARID1A		1	27057727	NM_006015	c.1435C>T	p.Q479Ter	24,6 PLX2853	D		4	3	OncoKB	
RCP42	Bladder	SNV	ARID1A		1	27057727	NM_006015	c.1435C>T	p.Q479Ter	24,6 Tazemetostat	D		4	3	OncoKB	
RCP43	Breast	SNV	PIK3CA		3	178922301	NM_006218	c.1070G>T	p.R357L	5,13 Alpelisib + Fulve	A		2	1	OncoKB	
RCP43	Breast	SNV	BRCA2		13	32907465	NM_000059	c.1850C>A	p.S617Ter	4,1 Olaparib	B	3a		2	OncoKB	
RCP43	Breast	SNV	BRCA2		13	32907465	NM_000059	c.1850C>A	p.S617Ter	4,1 Talazoparib	B	3a		2	OncoKB	
RCP43	Breast	SNV	BRCA2		13	32907465	NM_000059	c.1850C>A	p.S617Ter	4,1 Niraparib	D		4	3	OncoKB	
RCP43	Breast	SNV	BRCA2		13	32907465	NM_000059	c.1850C>A	p.S617Ter	4,1 Rucaparib	D		4	3	OncoKB	
RCP43	Breast	SNV	TSC2		16	2129381	NM_000548	c.3236C>A	p.S1079Ter	4,13 Everolimus	D		4	4	OncoKB	
RCP43	Breast	SNV	TSC2		16	2129381	NM_000548	c.3236C>A	p.S1079Ter	4,13 ABI-009	D		4	4	OncoKB	
RCP43	Breast	SNV	ARID1A		1	27023307	NM_006015	c.413C>A	p.S138Ter	4,26 PLX2853	D		4	5	OncoKB	
RCP43	Breast	SNV	ARID1A		1	27023307	NM_006015	c.413C>A	p.S138Ter	4,26 Tazemetostat	D		4	5	OncoKB	
RCP43	Breast	SNV	PIK3CA		3	178922301	NM_006218	c.1070G>T	p.R357L	5,13 RLY-2608	D		4	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Ado-Trastuzuma	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Neratinib	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Pertuzumab	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Trastuzumab	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Lapatinib	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Margetuximab	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Trastuzumab De	A		1	1	OncoKB	
RCP44	Breast	CNV	ERBB2			NM_004448				Tucatinib	A		1	1	OncoKB	
RCP45	Bladder	SNV	KDM6A	X		44879905	NM_021140	c.494G>A	p.R165Q	4,29 Tazemetostat	D		4	1	OncoKB	
RCP47	Bladder	SNV	TSC1		9	135786889	NM_000368	c.980del	p.P327LfsTer4	6,25 Everolimus	D		4	1	OncoKB	
RCP48	Thyroid	SNV	CDK12		17	37657693	NM_015083	c.2609+1G>A	None	6,45 Olaparib	D		4	1	OncoKB	
RCP50	Lung	SNV	MET		7	116412043	NM_000245	c.3028G>T	p.D1010Y	45 Capmatinib	A		1	1	OncoKB	
RCP50	Lung	SNV	MET		7	116412043	NM_000245	c.3028G>T	p.D1010Y	45 Tepotinib	A		1	1	OncoKB	
RCP50	Lung	SNV	MET		7	116412043	NM_000245	c.3028G>T	p.D1010Y	45 Crizotinib	A		2	2	OncoKB	
RCP50	Lung	SNV	ATM		11	108224555	NM_000051	c.8734A>G	p.R2912G	31 Olaparib	D		4	3	OncoKB	
RCP51	Lung	SNV	MET		7	116412043	NM_000245	c.3028G>C	p.D1010H	65,5 Capmatinib	A		1	1	OncoKB	
RCP51	Lung	SNV	MET		7	116412043	NM_000245	c.3028G>C	p.D1010H	65,5 Tepotinib	A		1	1	OncoKB	
RCP51	Lung	SNV	MET		7	116412043	NM_000245	c.3028G>C	p.D1010H	65,5 Crizotinib	A		2	2	OncoKB	
RCP54	Bladder	SNV	KDM6A	X		44942817	NM_021140	c.3397C>T	p.Q1133Ter	25,4 Tazemetostat	D		4	1	OncoKB	
RCP54	Bladder	SNV	ARID1A		1	27106925	NM_006015	c.6536_6552del	p.S2179NfsTer40	13,1 Tazemetostat	D		4	2	OncoKB	
RCP54	Bladder	SNV	ARID1A		1	27106925	NM_006015	c.6536_6552del	p.S2179NfsTer40	13,1 PLX2853	D		4	2	OncoKB	
RCP57	Bladder	SNV	KRAS		12	25398285	NM_004985	c.34G>A	p.G12S	32,6 Trametinib	D		4	1	OncoKB	
RCP57	Bladder	SNV	KRAS		12	25398285	NM_004985	c.34G>A	p.G12S	32,6 Cobimetinib	D		4	1	OncoKB	
RCP57	Bladder	SNV	KRAS		12	25398285	NM_004985	c.34G>A	p.G12S	32,6 Binimetinib	D		4	1	OncoKB	
RCP58	Bladder	SNV	FGFR3		4	1803568	NM_000142	c.746C>G	p.S249C	13,4 Erdatinib	A		1	1	OncoKB	

RCP58	Bladder	SNV	FGFR3	4	1803568	NM_000142	c.746C>G	p.S249C	13,4	Debio1347	D		4	2	OncoKB
RCP58	Bladder	SNV	FGFR3	4	1803568	NM_000142	c.746C>G	p.S249C	13,4	Infigratinib	D		4	2	OncoKB
RCP58	Bladder	SNV	FGFR3	4	1803568	NM_000142	c.746C>G	p.S249C	13,4	AZD4547	D		4	2	OncoKB
RCP58	Bladder	SNV	TSC1	9	135786489	NM_000368	c.1041G>A	p.W347Ter	62,6	Everolimus	D		4	3	OncoKB
RCP58	Bladder	SNV	PIK3CA	3	178936091	NM_006218	c.1633G>A	p.E545K	25,8	RLY-2608	D		4	4	OncoKB
RCP58	Bladder	SNV	PIK3CA	3	178936091	NM_006218	c.1633G>A	p.E545K	25,8	LOXO-783	D		4	4	OncoKB
RCP59	Duodenum	SNV	CDKN2A	9	21974721	NM_000077	c.106del	p.A36RfsTer17	23,3	Abemaciclib	D		4	1	OncoKB
RCP59	Duodenum	SNV	CDKN2A	9	21974721	NM_000077	c.106del	p.A36RfsTer17	23,3	Palbociclib	D		4	1	OncoKB
RCP59	Duodenum	SNV	CDKN2A	9	21974721	NM_000077	c.106del	p.A36RfsTer17	23,3	Abemaciclib, Pa	D		4	1	OncoKB
RCP60	Kidney	SNV	HRAS	11	533874	NM_005343	c.182A>T	p.Q61L	31,1	Tipifarnib	B	3a		1	OncoKB
RCP60	Kidney	SNV	ARID1A	1	27100383	NM_006015	c.4096dup	p.Q1366PfsTer79	9,52	Tazemetostat	D		4	2	OncoKB
RCP60	Kidney	SNV	ARID1A	1	27100383	NM_006015	c.4096dup	p.Q1366PfsTer79	9,52	PLX2853	D		4	2	OncoKB
RCP61	Skin	SNV	CHEK2	22	29121266	NM_007194	c.409C>T	p.R137Ter	5,68	Olaparib	D		4	1	OncoKB
RCP66	Ovary	SNV	BRCA1	17	41234451	NM_7294	c.4327C>T	p.R1443Ter	37,6	Niraparib	A		1	1	OncoKB
RCP66	Ovary	SNV	BRCA1	17	41234451	NM_7294	c.4327C>T	p.R1443Ter	37,6	Olaparib + Beva	A		1	1	OncoKB
RCP66	Ovary	SNV	BRCA1	17	41234451	NM_7294	c.4327C>T	p.R1443Ter	37,6	Olaparib	A		1	1	OncoKB
RCP66	Ovary	SNV	BRCA1	17	41234451	NM_7294	c.4327C>T	p.R1443Ter	37,6	Rucaparib	A		1	1	OncoKB
RCP66	Ovary	SNV	BRCA1	17	41234451	NM_7294	c.4327C>T	p.R1443Ter	37,6	Talazoparib	D		4	2	OncoKB
RCP71	Breast	SNV	ARID1A	1	27023290	NM_006015	c.400del	p.A134RfsTer98	35,1	PLX2853	D		4	1	OncoKB
RCP71	Breast	SNV	ARID1A	1	27023290	NM_006015	c.400del	p.A134RfsTer98	35,1	Tazemetostat	D		4	1	OncoKB
RCP77	Breast	SNV	PALB2	16	23634426	NM_024675	c,2680G>T	p.E954Ter	87,2	Olaparib	D		4	1	OncoKB
RCP77	Breast	SNV	PALB2	16	23634426	NM_024675	c,2680G>T	p.E954Ter	87,2	Rucaparib	D		4	1	OncoKB
tPD02	CUP	SNV	PTEN	10	89623708	NM_001304717	c.2T>C	p.L1?	4,34	GSK2636771, A2	D		4	1	OncoKB