Les défis du séquençage à haut débit dans l'exploration génétique des cancers du sein et de l'ovaire.

Etienne Muller

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité Aspects moléculaires et cellulaires de la biologie
- Génétique du Cancer -

Préparée au sein de l'Université de Rouen

Les défis du séquençage à haut-débit dans l'exploration des prédispositions génétiques aux cancers du sein et/ou de l'ovaire

Présentée et soutenue par
Etienne MULLER

Thèse soutenue publiquement le 12 décembre 2017 devant le jury composé de		
Mme Marie-Pierre BUISINE	PU-PH, CHRU de Lille	Rapporteur
M. Claude HOUDAYER	MCU-PH, Institut Curie, Paris	Rapporteur
M. Pascal GAUDUCHON	Professeur émérite, Université de Caen Normandie	Examinateur
M. Stéphane BEZIEAU	Professeur, CHU de Nantes	Examinateur
M. Laurent CASTERA	Pharmacien spécialiste des CLCC, Centre François Baclesse, Caen	Co-encadrant de Thèse
M. Thierry FREBOURG	Professeur, CHU de Rouen	Directeur de Thèse

Thèse dirigée par le Professeur Thierry FREBOURG, Unité INSERM 1245

Thèse co-encadrée par le Docteur Laurent CASTERA, Unité INSERM 1245

Remerciements

Je remercie tout d'abord l'ensemble du jury de me faire l'honneur de juger ce travail et de me permettre de le présenter.

Je remercie mon directeur de thèse, le Professeur Thierry Frebourg, pour m'avoir accueilli au sein de l'unité INSERM U1245, l'attention et les conseils qu'il a su apporter pendant les différentes phases de ce travail.

Cette thèse a été effectuée sous la co-direction de Laurent Castéra, à qui j'adresse mes remerciements les plus sincères pour m'avoir orienté vers le domaine de la bioinformatique, pour son attention, ses conseils et sa grande disponibilité aux cours de ses quatre dernières années.

Je remercie Mr Dominique Vaur, Directeur du Laboratoire de Biologie et de Génétique du Cancer du Centre François Baclesse, pour avoir eu la possibilité d'évoluer pendant ces quatre années au sein de son équipe, me permettant de découvrir le monde de la cancérologie et de la génétique médicale, ainsi que pour son soutien et ses conseils.

Je remercie Agathe Ricou et Sophie Krieger pour leurs remarques constructives et leur aide apportée tout au long de ces quatre années.

Je remercie l'équipe de bio-informatique du laboratoire, Antoine, Baptiste et Germain, pour les conseils et l'aide précieuse qu'ils ont pu prodiguer à leur « stagiaire » dans la réalisation de ce projet.

Merci à Valentin Harter, Jean-Christophe Thery, Pierre Fermey, Sophie Coutant et Camille Charbonnier pour leur collaboration.

Je remercie également tout le reste de l'équipe du laboratoire pour leur accueil, leur bonne humeur et leur soutien moral : Angélina, Manuella, Hafsa, Céline, Julien, Aurore, Caroline, Florian, Robin, Nicolas, Jean-Pierre, Nadine, Olivia, Chan, Arnaud, Gauthier, Raphael, Grégoire, Anne-Laurence, Sosthene, Laurence.

Enfin, merci à ma famille et mes amis, toujours là pour me soutenir, me changer les idées et pour avoir fait que ces dernières années d'études se soient passées aussi bien.

Table des Matières

Liste des Figures

Liste des Tables

Abréviations

ACP : Analyse en composante principale
ACR : American College of Radiology
ACMG : American College of Medical Genetics and Genomics
ADN : Acide Désoxyribonucléique
ADNc : ADN circulant
AF : Allele Frequency
ARN : Acide Ribonucléique
ARNm : ARN messager
ASCII : American Standard Code for Information Exchange
A-T : ataxie-télangiectasie
ATM : *ATM serine/threonine kinase*
ATR : *ATR serine/threonine kinase*
ATP : Adenosine Triphosphate
BAM : Binay Alignment/Map Format
BARD1 : *BRCA1 associated RING domain 1*
BCAC : Breast Cancer Association Consortium
BED : Browser Extensible Data
BI-RADS : Breast Imaging Reporting And Data System
BLAST : Basic Local Alignment Search Tool
BRIDGES : Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm
BQSR : Base quality Score Recalibration
BRIP1 : *BRCA1 interacting protein C-terminal helicase 1*
BRAF : *v-Raf murine sarcoma viral oncogene homolog B*
BRCA1 : *Breast Cancer 1*
BRCA2 : *Breast Cancer 2*
BRCT : BRCA1 C-Terminal
BRIDGES : Breast cancer RIsk after Diagnostic Gene Sequencing
BWA : Burrows-Wheeler Aligner
CCD : Charge Coupled Device
CDH1 : *Cadherin 1*
CDK : Cyclin-Dependant kinase
CIViC : Clinical Interpretation Of Variants in Cancer
ChAM : Chromatin Association Motif
CHEK2 : *Checkpoint kinase 2*
cnLOH : copy neutral Loss Of Heterozygosity
CNV : Copy Number Variation
DBD : DNA-binding Domain
Del : Deletion
DHPLC : Denaturing High Performance Liquid Chromatography

DoCM : Database of Curated Mutations
DP : Depth
DSB : Double Strand Break
Dup : Duplication
ENIGMA : Evidence-based Network for the Interpretation of Germline Mutants Alleles
ExAC : Exome Agregation Consortium
FANCA : *Fanconi Anemia Complementation Group A*
FANCL : *Fanconi Anemia Complementation Group L*
FANCM : *Fanconi Anemia Complementation Group M*
FREX : French Exome Project
GATK : Genome Analysis ToolKit
GGC : Groupe Génétique et Cancer
gVCF : genomic Variant Call Format
GWAS : Genome Wide Association Study
HBOC : Hereditary Breast and Ovarian Cancer
HC : High Confidence
HER2 : *ERBB2 Receptor Tyrosine Kinase 2*
HGMD : Human Gene Mutation Database
HGVS : Human Genome Variation Society
HMM : Hidden Markov Model
HNPCC : Hereditary Non-Polyposis Colorectal Cancer
HPO: Human Phenotype Ontology
HR : Hazard Ratio
HR : Homologous Recombination
HRM : High Resolution Melt
HSF : Human Splicing Finder
HSP : High-scoring Segment Pairs
IARC : International Agency For Research on Cancer
IC : Intervalle de Confiance
IGF2 : *Insulin Like Growth Factor 2*
Indel : Insertion / Deletion
InSiGHT : International Society for Gastrointestinal Hereditary Tumours
IRM : Imagerie par Résonance Magnétique
Kb : Kilobase
LFS : Li Fraumeni Syndrome
LOD : Log ODds Score
LoF : Loss of Function
LOVD : Leiden Open Variation Database
MAF : Minor Allele Frequency
MCA : Multiple Correspondance Analysis
MLH1 : *MutL homolog 1*
MMR : Mismatch Repair
MRE11A : MRE11 homolog, double strand break repair nuclease
MSH2 : *MutS homolog 2*

MSH6 : *MutS homolog 6*
mTOR : mechanistic target of Rapamycin kinase
Myc : Myc proto-oncogene
NBN : *Nibrin*
NES : Nuclear Export Sequence
NF1 : *Neurofibromin 1*
NGS : Next Generation Sequencing
NHEJ : Non-Homologous End Joining
nlcRNA : Non Long-Coding RNA
NLS : Nuclear Localization Sequence
NMD : Nonsense-Mediated mRNA Decay
OMIM : Online Mendelian Inheritance in Man
OR : Odd Ratio
PA : Personnes-années
PALB2 : *Partner and Localizer of BRCA2*
PARP : Poly(ADP-ribose) Polymerase
PCA: Principal Component Analysis
PCR : Polymerase Chain Reaction
PMS2 : *PMS1 Homolog 2*
PRS: Polygenic Risk Score
PTEN : *Phosphatase and Tensin Homolog*
PV : Pathogenic Variant
RAD51 : *RAD51 recombinase*
RAD51B : *Rad51 paralog B*
RAD51C : *Rad51 paralog C*
RAD51D : *Rad51 paralog D*
RB : Retinoblastoma Protein
RB1 : *Retinoblastoma 1*
RING : Really Interesting New Gene
SAM : Sequence Alignment/Map Format
SCD : Serine Containing Domain
SD : Standard Deviation
SMRT : single-molecule real-time sequencing
SNP : Single Nucleotide Polymorphism
SNV : Single Nucleotide Variation
SSF : Splicing Sequences Finder
STK11 : *Serine/threonine kinase 11*
TAD: Topological Association Domain
TCGA : The Cancer Genome Atlas
TD : Tower Domain
TNM : Taille de la tumeur / Nombre de ganglions lymphatiques atteints / Métastases
TP53 : *Tumor Protein p53*
VEP : Variant Effect Predictor
VCF : Variant Call Format

VQSLOD : Variant-Quality Score Log-Od
VSI : Variant de signification inconnue
VUS : Variant of Unknown Signifiance
WRN : Werner Syndrome RecQ like Helicase
WT : Wild Type
XRCC2 : X-Ray repair Cross Complementing 2
XRCC3 : X-Ray repair Cross Complementing 3
ZMW : zero-mode waveguide

Préambule

Les cancers du sein et de l'ovaire sont les cancers les plus fréquents chez la femme, et les deuxièmes plus fréquents tous cancers confondus. Parmi eux, 5 à 10% sont liés à une prédisposition génétique héréditaire, transmise sur un mode autosomique dominant. Chez les cas sélectionnés par les consultations d'oncogénétique pour présenter une forte probabilité de prédisposition génétique, des variants pathogènes des gènes de *BRCA1* et *BRCA2* ne sont retrouvés que dans 9 à 12% des cas. Les variants pathogènes d'autres gènes sont également responsables d'un sur-risque de cancer du sein ou de l'ovaire ou suspectés de l'être, tels que par exemple *TP53*, *PTEN*, *STK11* et *CDH1* (forte pénétrance), *ATM*, *BRIP1* et *CHEK2* (faible pénétrance), *BARD1* et les paralogues de *RAD51* (suspectés). Malgré cela, une grande partie des cas de prédisposition à ces cancers reste inexpliquée. Une première hypothèse pour tenter d'expliquer la part d'hérédité manquante, a proposé que la prédisposition au cancer soit la résultante d'une combinaison de variants fréquents en population. Cette hypothèse a principalement été testée par des études de GWAS (Genome-Wide Association Study), qui ne résolvent qu'une part faible de l'hérédité manquante reposant sur l'existence de variants associés à de faibles risques.

L'avènement du séquençage à haut-débit (NGS : Next generation Sequencing) dans les laboratoires permet aujourd'hui d'explorer de nouvelles pistes dans la résolution de cette hérédité manquante, notamment avec l'hypothèse du « Common Disease - Rare Variant », qui propose que la prédisposition soit provoquée par de multiples variants rares. Les débits d'analyse fournis par le NGS donnent accès à l'ensemble des variants du génome humain, permettant d'étudier de nouveaux variants jamais visualisés auparavant dans les études de GWAS effectuées en puce SNP (Single Nucleotide Polymorphism), même à haute densité. La rareté en population des variants étudiés (MAF : Minor Allele Frequency < 0.1 %) complique néanmoins la mise en évidence d'un effet d'association, nécessitant des études cas-témoins comprenant de grandes populations pour démontrer un effet significatif. Plusieurs solutions

permettent d'augmenter la puissance de l'analyse (burden test), ou encore la sélection de phénotypes extrêmes.

Le NGS va aussi permettre d'explorer une seconde hypothèse, qui est celle d'une prédisposition tissulaire locale, liées à la présence de néo-mutations en mosaïque difficilement détectables par les techniques conventionnelles. La sensibilité atteinte par les technologies de séquençage à haut-débit, associée à une analyse bio-informatique dédiée, pourrait ainsi permettre de rechercher des néo-mutations post-zygotiques en mosaïque faiblement représentées sur les gènes de prédisposition au syndrome, pouvant expliquer une part supplémentaire de l'hérédité manquante.

Avant de pouvoir répondre à ces hypothèses, cette technologie impose de surmonter des défis à la fois en biologie, en informatique et en statistiques, afin de pouvoir disposer de données pertinentes et interprétables.

L'obtention de variants vraisemblablement réels (vrais positifs) par séquençage à haut-débit dépend ainsi d'un traitement complexe et adapté, donc chaque étape (alignement, recalibration, variant-calling) nécessite d'être maîtrisée. Il s'agit d'un premier défi majeur du NGS. En effet, la détection des variants notamment par les logiciels de « variant-calling » est une étape critique, permettant de discriminer les vrais positifs des faux-positifs et d'éviter les biais d'interprétation. Cette étape permet aussi de mettre en évidence les variants en mosaïque, difficiles à détecter du fait de leur faible représentation dans le matériel génétique séquencé.

Le deuxième grand défi du séquençage à haut-débit réside dans l'interprétation des variants identifiés qui sont découverts toujours plus nombreux et souvent uniques. Des outils bio-informatiques d'annotation et de prédiction du caractère pathogène peuvent être utilisés dans l'aide à l'interprétation, mais leur maniement doit être réalisé avec précaution et de manière

éclairée, du fait de leur précision relative. Des bases de données locus-spécifiques sont disponible publiquement permettant un partage des connaissances entre les laboratoires et facilitant l'identification des variants pathogènes. Ces bases doivent néanmoins être elles-aussi soumises à un regard critique, la qualité de leur maintien, les critères de classification des variants pouvant différer d'une base à l'autre. La classification des variants est un travail critique pour réaliser des études cas-témoins testant la part des variants rares et non encore connus. En effet dans un objectif de tester l'association des variants d'un gène ou d'une région génomique définie afin d'identifier de nouveaux gènes ou de nouveaux locus impliqués par des tests statistiques en agrégation de variants, une hiérarchisation précise des variants permet d'assurer le maintien d'une puissance statistique suffisante pour conclure. Ces analyses nécessitent néanmoins un nombre important de cas et de témoins pour obtenir une analyse significative. L'agrégation des variants sur des ensembles de gènes impliqués dans un même processus biologique est un moyen efficace de gagner en puissance et le fait d'étudier des cas sélectionnés pour le caractère extrême de leur phénotype permet d'augmenter celle-ci et donc d'autoriser l'analyse sur des populations éventuellement plus restreintes. Aussi, l'accessibilité à une population témoin appariée est une difficulté importante dès lors que les cas ont été colligés dans le cadre de leur diagnostic moléculaire après requalification de leurs données. L'obtention d'un nombre de témoins suffisants devient alors problématique, mais peut être compensée par l'utilisation de bases de données de variants en accès publique libre, réunissant les résultats de plusieurs milliers d'exomes ou de génomes. Ces bases de données de variants vont ainsi permettre d'écarter les variants les plus fréquents (polymorphismes) considérés comme non-pathogènes dans un contexte de maladie mendélienne rare, afin de se focaliser sur les variants les plus rares. L'utilisation de ces bases pour reconstruire des populations de témoins nécessite l'utilisation de nouveaux tests statistiques, l'information donnée par les bases de données de séquençage ne concernant que la fréquence des variants

en population (sans possibilité de récupérer l'haplotype complet de l'individu). Cette approche nécessite néanmoins une grande prudence.

Les travaux présentés dans ce manuscrit vont ainsi explorer la part d'hérédité manquante caractérisant les cancers du sein et de l'ovaire, en utilisant les possibilités offertes par le séquençage à haut-débit.

Avec l'étude d'une population de 5 131 cas index atteints par le syndrome HBOC (*Hereditary Breast And Ovarian Cancer*), nous avons estimé le risque induit par les variants pathogènes rares de 34 gènes connus pour leur implication, ou suspectés de l'être, dans les formes héréditaires du cancer du sein et de l'ovaire. Une nouvelle approche statistique a été développée afin de pouvoir tester l'association des variants rares d'un gène avec le syndrome, en utilisant comme contrôle les données agrégées de la base ExAC, ainsi que les données individuelles de séquençage d'exome des individus d'une population de témoins d'origine française issue d'une collaboration avec le centre national de génotypage (Projet du consortium French Exome Project, FREX).

La participation des néo-mutations retrouvées en mosaïque dans le syndrome a été explorée grâce à un logiciel développé dans le cadre de cette thèse : outLyzer. Ce logiciel a été développé afin de mettre en évidence les variants faiblement représentés dans les milieux hétérogènes en se basant sur une analyse statistique et locale du bruit de fond de séquençage. Profitant de la collection d'ADN tumoraux pour lesquels des variants dont les fréquences d'allèle muté étaient connues, outLyzer a été validé après séquençage d'un panel de 22 gènes par NGS (enrichissement par capture). Son efficacité a été comparée à une sélection de logiciels existants. Après démonstration de sa bonne sensibilité et spécificité et le développement d'une amélioration prenant en compte le biais de l'équilibre des *reads* en *Forward* /*Reverse* généré par les méthodes de capture, outLyzer a permis la recherche de

variants en mosaïque dans 34 gènes dans une population de 1 750 patientes atteintes du syndrome des cancers du sein et de l'ovaire héréditaires.

Enfin, l'exploration d'un plus grand nombre de gènes impliqués dans le cancer pourrait permettre également d'explorer l'hérédité manquante des cancers du sein et de l'ovaire. Des publications récentes, basées sur des analyses pan-génomiques, ont pu montrer que les variants somatiques impliqués dans la tumorigénèse étaient finalement relativement limités et se regroupaient dans les mêmes voies de signalisation. De plus, quasiment la moitié des gènes de prédisposition décrits jusqu'à présent sont aussi affectés par des variants présents dans les tumeurs, confirmant que les voies de signalisation mises en jeux dans les cancers héréditaires et sporadiques sont au moins partiellement chevauchantes. Nous avons ainsi développé une analyse par séquençage à haut-débit de 201 gènes sélectionnés pour leur implication dans le cancer au niveau constitutionnel ou somatique. De plus nous avons favorisé l'étude de phénotypes extrêmes caractérisés par un âge précoce d'apparition d'un cancer, afin d'augmenter la probabilité de mettre en évidence des évènements fortement pénétrants et augmenter la puissance de notre analyse. 105 patientes diagnostiquées pour un cancer du sein avant 31 ans et 13 patientes diagnostiquées pour un cancer de l'ovaire avant 41 ans ont été incluses dans cette étude.

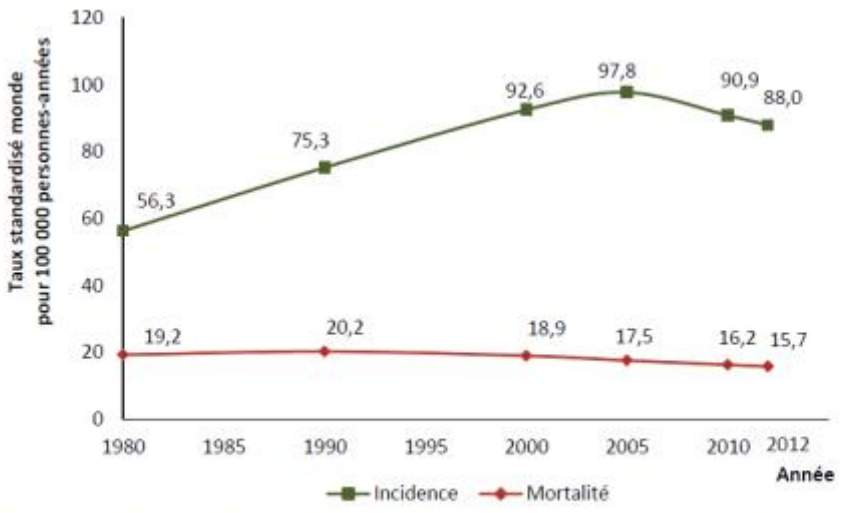
Introduction

I / Etiologie des prédispositions génétiques aux cancers du sein et de l'ovaire

A. Epidémiologie

1. Les cancers du sein

En France métropolitaine, le cancer du sein est une maladie quasiment exclusivement féminine présentant une incidence estimée à 54 062 cas en 2015, soit 94.7 pour 100 000 (InVS : Institut de Veille Sanitaire), provoquant 11 913 décès la même année (les cas de cancer chez l'homme ne représentent qu'1% de la pathologie). Cette pathologie demeure le cancer le plus fréquent chez la femme et le 2ème plus fréquent tous cancers confondus. Son incidence a fortement augmenté entre 1980 et 2012 (+1,4% par an en moyenne) avec néanmoins une baisse sur la période 2005-2012 (-1,5% par an) (Figure 1).



Source : Partenariat Francim/HCL/Santé publique France/INCa [Binder-Foucard F, 2013].
Traitement : INCa 2016

Figure 1: Evolution de l'incidence et de la mortalité par cancer du sein de 1980 à 2012 en France métropolitaine.

Les facteurs qui pourraient expliquer cette baisse comprennent notamment une saturation du dépistage et une diminution de la prescription de traitements hormonaux de la ménopause[1].

Malgré cela le groupe des cancers les plus jeunes (dans la tranche 30-39 ans) montre une augmentation constante de l'incidence, passant de 32.8 en 1980 à 54.09 nouveaux cas pour 100 000 personnes par an (PA) en 2012. La mortalité est restée relativement stable sur la même période, avec une baisse plus significative en regardant la période 2005-2012 (-1.5% par an). Ainsi en 2012, l'âge médian au diagnostic était de 63 ans, et l'âge médian de décès de 73 ans. Le taux de survie atteint aujourd'hui 87 % à 5 ans 76 % à 10 ans.

Les projections de l'Institut de Veille Sanitaire pour 2015 estiment le nombre de nouveaux cas à 54 062, correspondant à 31% des cancers diagnostiqués chez la femme[2]. Le nombre de décès est lui estimé à 11913 (soit 14,6 décès pour 100 000 PA) sur cette même année, correspondant à 18.2% des décès par cancer chez la femme.

Au niveau international, le nombre de nouveaux cas de cancers du sein a été estimé à 1,67 million en 2012 (43,1 nouveaux cas pour 100 000 PA), toujours en tant que 2ème cancer le plus fréquent et en tant que 1er chez la femme[3] (soit 25% des cancers féminins). La France montre une des plus fortes incidences en Europe (89,7 nouveaux cas pour 100 000 PA), supérieure à celle de l'Union Européenne (80,3 nouveaux cas pour 100 000 PA) [2]. Le taux de mortalité sur la même année au niveau mondial a été estimé à 521907 décès (12,9 décès pour 100 000 PA) soit 14,7% des décès par cancer chez la femme. En France, le taux de mortalité est proche de la moyenne européenne (15,5 décès pour 100 000 PA).

2. Les cancers de l'ovaire

En France, le cancer de l'ovaire est la 7ème cause de cancers et la 4ème cause de décès par cancer chez la femme. Son incidence a été évaluée à 4615 cas en 2012, soit 7.6 pour 100 000 personnes[4], provoquant par ailleurs 3140 décès la même année.

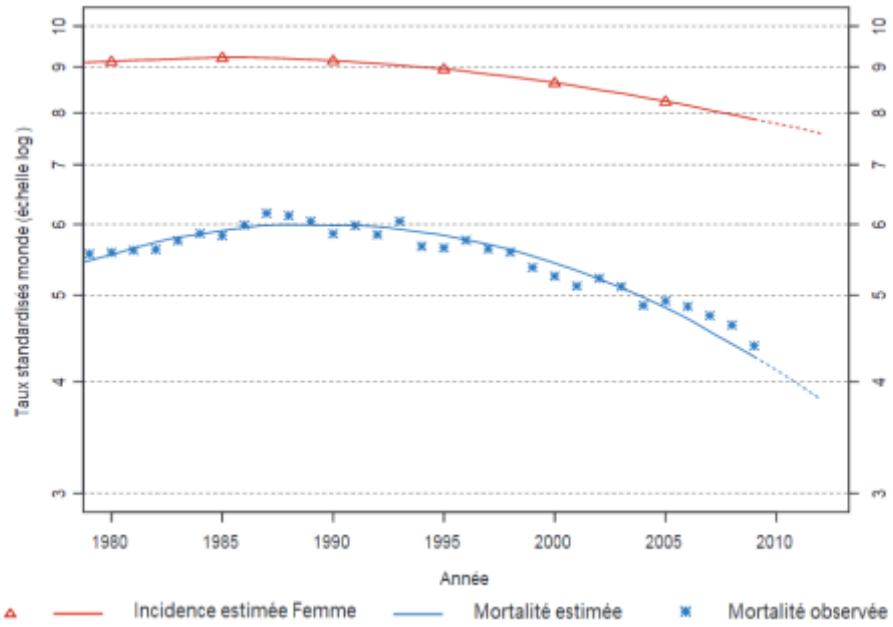


Figure 2: Estimation de l'incidence et de la mortalité des cancers de l'ovaire en France sur la période 1980-2012.

L'incidence de ce cancer diminue depuis 1990, avec une accélération de la diminution depuis 2005 (Figure 2). Le taux de mortalité suit une évolution assez semblable, avec une accélération de la diminution à partir de 2005. Cette diminution peut s'expliquer notamment par la prise de contraceptifs oraux, reconnus comme facteur protecteur dans le cancer de l'ovaire[5]. Ce cancer est de très mauvais pronostique, dû à un diagnostic souvent tardif et donc à une maladie régionalement étendue. En effet l'âge médian au diagnostic est de 66 ans et l'âge médian au décès est de 76 ans. Les taux de survie à 5 ans et à 10 ans sont respectivement de 40 % et 32 %, avec peu d'évolution dans le temps.

B. Contexte d'apparition

Le cancer est une maladie d'origine génétique liée à l'accumulation de variants génétiques dans les cellules, leur conférant des avantages de prolifération ou de survie. L'apparition de cette maladie est difficilement prévisible, et modulée par l'influence de facteurs environnementaux pouvant favoriser la transformation de cellules saines en cellules malignes. Néanmoins l'étude de l'anamnèse familiale rapporte parfois un nombre plus élevé de cancers que l'incidence nationale. Cette agrégation familiale suggère l'existence d'une composante héréditaire responsable d'une prédisposition à développer la pathologie. Le cancer du sein et de l'ovaire peut donc apparaitre soit dans des contextes sporadiques généralement chez la personne âgée probablement en dehors d'un contexte de prédisposition, soit chez des personnes plus jeunes possiblement avec une histoire familiale de cancers du même type signant la transmission d'un trait génétique.

1. Cancers sporadiques non liés à une prédisposition génétique

Un cancer sera qualifié de sporadique dès lors qu'aucun contexte héréditaire n'est suspecté. Généralement, la transformation du tissu normal en cancer fait suite dans ce cas à l'accumulation de modifications génétiques acquises par certaines cellules au cours de la vie d'un individu (on parle alors de variants somatiques et de variants tumoraux lorsque la transformation maligne est aboutie). Les cancers sporadiques sont caractérisés par un âge d'apparition plus tardif[6] au cours de la vie de l'individu et favorisée par deux types de facteurs dans le contexte du cancer du sein:

- Les facteurs intrinsèques :
    - une densité plus élevée du tissu mammaire[7] (proportion de tissu fibreux et glandulaire plus importante).

- L'existence de lésions bénignes[8] (hyperplasie lobulaire ou canalaire atypique, carcinome lobulaire ou canalaire *in situ*)
- Un âge précoce d'apparition des premières menstruations[9] (avant 12 ans)
- Un âge tardif de ménopause[10] (après 55 ans)

- Environnementaux :
  - L'obésité[11], particulièrement après la ménopause, liée à une production plus importante d'œstrogènes par les tissus graisseux
  - L'âge de la femme à la naissance du premier enfant[12] : un âge tardif d'obtention du premier enfant (> 30 ans) est associé à une augmentation du risque d'apparition d'un cancer du sein
  - Les hormonothérapies substitutives pour lutter contre les effets de la ménopause[13], particulièrement les associations œstrogène-progestérone
  - L'allaitement[14], en tant que facteur protecteur cette fois-ci, particulièrement lorsque celui-ci est prolongé

2. Cancers liés à une prédisposition génétique

Un cancer apparaissant dans un contexte de prédisposition génétique trouve son origine dans la préexistence dans l'ensemble (ou une grande majorité) des cellules de l'organisme de l'individu d'un variant génétique pathogène du trait transmis par ses parents (variant pré-zygotique), ou apparaissant précocement lors du processus embryologique (cf. Introduction, § V.D) On parle alors de variants constitutionnels. Certains variants génétiques augmentent donc le risque de développer un cancer chez les personnes porteuses par rapport à la population générale, conduisant à une agrégation des cas de cancers chez les familles où ceux-ci sont transmis entre les générations. Ces cas d'agrégations familiales ont été rapportés dès la

deuxième partie du XIXème siècle, avec la description d'une famille présentant 16 cas de cancers sur 4 générations, dont 10 cancers du sein[15]. A la même époque, Gregor Mendel définit les bases de la génétique moderne expliquant la transmission héréditaire par des expériences de croisements sur les pois[16]. Les découvertes successives dans le domaine de la génétique au cours du XXème siècle et l'observation des différences entre cancers d'apparition sporadique et ceux développés dans un contexte familial (notamment sur le rétinoblastome) conduiront ainsi Knudson à proposer un modèle de développement des cancers héréditaires en 2 étapes (« two hits »)[17]. Selon ce modèle, le développement d'un cancer, qu'il soit sporadique ou héréditaire, nécessite qu'une cellule acquiert au moins 2 évènements mutationnels sur le même locus génomique. La principale différence entre les 2 types de cancers proviendra du fait que dans le cas des cancers héréditaires, le premier variant est hérité, et présent dans toutes les cellules de l'individu (il y a transmission de la prédisposition). Le deuxième évènement est acquis, intervenant selon la théorie de Knudson par hasard, et déclenche le processus de tumorigénèse. Dans les cancers sporadiques, l'origine des 2 « hits » est uniquement somatique et liée au hasard. Ainsi, la probabilité qu'une cellule présente les deux événements déclencheurs du cancer est bien plus élevée en cas de prédisposition que ne le voudrait le simple hasard, et en conséquence, les cancers dans un contexte de prédisposition génétique sont d'apparition plus précoce et peuvent être multiples, alors qu'un cancer sporadique est d'apparition plus tardive et unique du fait de la faible probabilité d'apparition de 2 évènements somatiques sur le même locus génomique.

En accord avec cette le modèle de Knudson , l'étude des cas de cancers du sein et de l'ovaire en situation d'agrégation familiale ont pu mettre en évidence les gènes majeurs de prédisposition au cancer du sein et de l'ovaire, répondant à un modèle autosomique dominant à pénétrance élevée mais incomplète [18]. Les premiers gènes dont l'inactivation a été associée aux cancers du sein et de l'ovaire ont été *BRCA1*[19] et *BRCA2*[20], définissant le syndrome

HBOC, et suivis plus récemment par *PALB2*[21]. L'étude de ces gènes et de leurs variants contribue à la compréhension du syndrome, dans l'objectif d'améliorer la prise en charge médicale des patients atteints.

C. Diagnostic et traitements

1. Cancer du sein

Les éléments cliniques évocateurs qui orientent le clinicien vers un diagnostic de cancer du sein sont : rétractation ou inflammation de la peau, écoulement spontané au niveau du mamelon, maladie de Paget ou encore adénopathie axillaire. Néanmoins dans 90% des cas, la découverte d'un cancer du sein est fortuite, au détour d'un examen de dépistage. Cet examen est recommandé et organisé pour les femmes de 50 à 74 ans, et est renouvelé tous les 2 ans, ou en cas de suspicion suite à un élément évocateur. Il consiste dans un premier temps en une mammographie bilatérale à la recherche d'une lésion évocatrice d'un cancer. Les résultats de cet examen sont classés en 6 catégories, en se basant sur le système BI-RADS (Breast Imaging Reporting And Data System) de l'American College of Radiology[22] (ACR) :

- ACR 0 : classification d'attente, des investigations complémentaires sont nécessaires
- ACR 1 : mammographie normale
- ACR2 : anomalies bénignes ne nécessitant ni surveillance ni examen complémentaire
- ACR 3 : anomalie probablement bénigne pour laquelle une surveillance à court terme (3 à 6 mois) est conseillée
- ACR 4 : anomalie indéterminée ou suspecte
- ACR 5 : anomalie évocatrice d'un cancer

Cet examen pourra être complété par une échographie mammaire bilatérale en cas de contre-indication à la mammographie, d'image douteuse ou de seins denses. Dans le cas d'images ACR 3, la réalisation d'une biopsie percutanée mammaire pourra être discutée. Cette biopsie

sera en revanche systématique en cas d'images ACR 4 ou 5, notamment afin de confirmer le diagnostic, et effectuer une description précise de la tumeur. Cette description va notamment comprendre :

- Une classification histologique, s'attardant sur la morphologie et la vitesse de multiplication des cellules par rapport à des cellules saines. Cette classification donnera un grade (1, 2 ou 3) en fonction de la croissance plus ou moins forte des cellules.
- Le stade TNM (Taille de la tumeur / Nombre de ganglions lymphatiques atteints / présence potentielle de Métastases), évalué à partir des données d'imagerie et d'analyse en laboratoire.
- Le statut des récepteurs hormonaux : certaines cellules cancéreuses étant dépendantes des œstrogènes et de la progestérone pour leur propagation, la présence (tumeur dite hormono-dépendante) ou l'absence de ces récepteurs pourra conditionner l'accès aux traitements par hormono-thérapie.
- Le statut du récepteur HER2 (*Human epidermal growth factor receptor 2*): la surexpression de ce récepteur pourra aussi conditionner l'accès à une thérapie ciblée : le trastuzumab (Herceptin® : Anticorps monoclonal bloquant les récepteurs HER2).

Une tumeur négative pour les récepteurs hormonaux et les récepteurs HER2 sera ainsi qualifiée de triple négative. En fonction de l'agressivité et du stade du cancer diagnostiqué, une thérapeutique adaptée sera proposée au patient. L'arsenal thérapeutique pourra comprendre :

- Des chimiothérapies et hormono-thérapies néo-adjuvantes : chimiothérapie permettant de réduire la taille de la tumeur en vue d'un acte chirurgical ou d'une radiothérapie.
- La chirurgie mammaire : Mastectomie partielle (tumorectomie) ou totale

- La chirurgie axillaire : technique du ganglion sentinelle (ablation du premier ganglion lymphatique le plus proche de la tumeur pour vérification anatomopathologique de la présence de cellules cancéreuses. Permet de réserver le curage axillaire aux tumeurs qui le nécessitent) et curage axillaire.
- La radiothérapie : de la glande mammaire, du lit tumoral, de la paroi thoracique ou encore des ganglions de la chaîne mammaire interne et sus-claviaires.
- Les chimiothérapies conventionnelles : proposées en cas de risque de récidive important ou encore pour des cancers métastatiques
- Les thérapies ciblées : médicaments dirigés contre des cibles moléculaires particulièrement sensibles dans les cellules cancéreuses, tels que les anticorps anti-HER2 (Trastuzumab), les inhibiteurs de la dimérisation d'HER2 (Pertuzumab), les inhibiteurs de la tyrosine kinase d'HER2 (Lapatinib), les inhibiteurs de l'angiogénèse (Bevacizumab), ou les inhibiteurs de la voie de transduction mTOR (Everolimus).

Les patientes porteuses d'un facteur de prédisposition génétique connu se verront aussi proposer un parcours de soins et une surveillance mammaire adaptés (cf. Introduction, § I.D).

2. Cancer de l'ovaire

Il n'existe pas à l'heure actuelle, à l'image du cancer du sein, de programme généralisé de dépistage du cancer de l'ovaire. Ce cancer provoque peu de symptômes et les signes d'appel sont peu spécifiques : douleurs abdominales, augmentation du volume de l'abdomen, constipation, pollakiurie, métrorragies… Un dépistage est néanmoins recommandé par une échographie pelvienne annuelle chez les femmes porteuses d'une prédisposition génétique. Une annexectomie bilatérale préventive peut aussi être recommandée chez ces femmes à partir de 40 ans.

La détection de ce cancer est de ce fait souvent réalisée de manière trop tardive, avec des localisations métastatiques déjà présentes dans trois quarts des cas au moment du diagnostic. Celui-ci est établi par examen clinique complet incluant notamment examen abdominal, touchers pelviens, palpation des aires ganglionnaires, mesure du poids et interrogatoire qui doit préciser les antécédents personnels et familiaux de cancer et les comorbidités. Le diagnostic sera complété par une échographie abdomino-pelvienne à la recherche d'une anomalie évocatrice. Il sera secondé en cas de doute par une tomodensitométrie de la zone abdomino-pelvienne en cas de doute. La confirmation sera apportée par une analyse anatomopathologique de la masse pelvienne suspecte.

La prise en charge thérapeutique de ce cancer est orientée sur deux axes principaux : La chirurgie et la chimiothérapie. Concernant les stades les plus précoces, l'intervention standard consiste en une hystérectomie totale (chez la femme ménopausée ou ne voulant plus d'enfant). Concernant les stades les plus avancés, la résection complète est préconisée, potentiellement appuyée par une chimiothérapie néo-adjuvante. Les chimiothérapies adjuvantes peuvent aussi être proposées sur les tumeurs de stade intermédiaire, mais associées aussi à une chirurgie. De plus l'arsenal thérapeutique est aujourd'hui élargi par la mise sur le marché de l'olaparib, thérapie ciblée indiquée dans les cancers de l'ovaire en récidive chez les patientes porteuses d'un variant inactivateur sur les gènes de *BRCA1* ou *BRCA2* et sensibles aux chimiothérapies à base de Platine. L'olaparib, inhibiteur de PARP (Poly(ADP-ripbose) polymérase), fonctionne selon un mécanisme de létalité synthétique (Figure 3), nécessitant l'inactivation conjointe de deux cibles moléculaires pour provoquer la mort cellulaire.

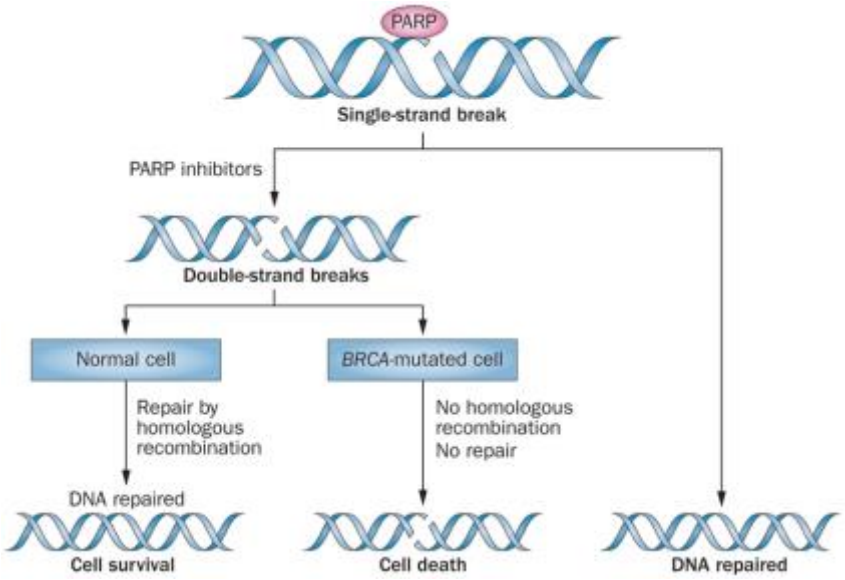


Figure 3: Le rôle des inhibiteurs de PARP dans la létalité synthétique[23]

Les PARP sont des enzymes jouant un rôle clé dans la réparation des cassures ADN simple-brin. Un échec dans la réparation de ces cassures simple-brin conduit à une cassure double-brin durant la phase de réplication de l'ADN. L'inhibition de ces enzymes peut donc conduire à d'autres dommages à l'ADN. Pourtant, les cassures double-brin peuvent être réparées par un autre mécanisme : le système de recombinaison homologue, dont BRCA1 et BRCA2 sont des éléments fondamentaux, (cf. Introduction, § II). Une cellule mutée sur les gènes *BRCA1* ou *BRCA2* possèdera donc un système de réparation par recombinaison homologue déficient. La non-réparation des cassures doubles brin entrainera une instabilité chromosomique et un arrêt du cycle cellulaire conduisant finalement à la mort cellulaire par apoptose.

D. Prise en charge des cas de cancer du sein et de l'ovaire dans un contexte de prédisposition : dispositif d'oncogénétique

L'origine héréditaire de 5 à 10 % des cancers du sein (Institut National du Cancer, 2016) demande une prise en charge à la fois familiale et personnelle adaptée réalisée grâce au dispositif national d'oncogénétique. En France, ce dispositif s'articule autour de 139 sites de consultation dans 94 villes réparties sur l'ensemble du territoire, et de 25 laboratoires en charge de la réalisation des tests diagnostiques. L'objectif de ce dispositif national va être d'identifier les personnes possédant une prédisposition génétique au cancer, qu'elles soient des personnes atteintes de cancer (cas index) ou des membres indemnes de leur famille (apparentés). En effet la présence d'un variant pathogène responsable d'une prédisposition augmente le risque de développer (les risques induit sont différents en fonction des gènes et sont précisés dans l'Introduction, § IV):

- Un cancer du sein à un âge jeune (avant la ménopause)
- Un cancer du sein bilatéral
- Un cancer de l'ovaire (essentiellement à partir de 40 ans)

L'orientation vers une consultation d'oncogénétique ainsi sera proposée pour une personne présentant un des signes évocateurs suivants :

- Critères individuels :
    - Un cancer du sein d'apparition précoce (avant l'âge de 36 ans), ou avant 50 ans si triple négatif
    - Un cancer du sein bilatéral avant 40 ans
    - Un cancer du sein de type médullaire
    - Un cancer du sein chez un homme
    - Un cancer de l'ovaire diagnostiqué avant l'âge de 60 ans

- Critères familiaux :
    - Au moins 3 cas de cancer du sein et / ou de l'ovaire dans la même branche parentale unis par des liens de premier ou second degré (quel que soit l'âge de diagnostic)
    - Deux cas de cancer du sein unis par des liens de premier ou second degré passant par un homme, avec dans au moins un cas un diagnostic avant 40 ans, ou une atteinte portée avant l'âge de 50 ans, l'autre avant 70 ans
    - Deux cas de cancer du sein chez des apparentés au premier degré dont au moins un cas est masculin
    - Un cas de cancer du sein et un cas de cancer de l'ovaire chez des apparentés au premier degré, ou chez la même personne

Ces indications orienteront le patient vers une consultation d'oncogénétique qui se chargera de recueillir ses informations médicales, de reconstituer son histoire personnelle et familiale, de construire l'arbre généalogique de la famille, et d'estimer la probabilité de prédisposition. Au regard de ces informations, le clinicien évaluera la nécessité de prescrire un test génétique à la recherche d'une prédisposition.

Aujourd'hui une extension des indications est réalisée pour tout cancer de l'ovaire, afin de pouvoir envisager un traitement par inhibiteurs de PARP[24].

En France, 45 430 consultations d'oncogénétiques ont été réalisées en 2015 dans le cadre du syndrome seins-ovaires, avec 14 700 cas-index ayant bénéficié d'un test génétique. Parmi ces derniers, 1 610 se sont révélés être porteurs d'un variant prédisposant au syndrome (soit 10,95 %).

La découverte d'un variant responsable d'une prédisposition déclenche pour le cas index un suivi médical adapté, comprenant un examen clinique semestriel, auquel s'ajoute un suivi annuel par imagerie associant IRM et mammographie. En effet, le risque de développer un

cancer du sein controlatéral est multiplié par 4 (soit 20 à 40 % de risque à 10 ans) en cas de variant de *BRCA1* ou *BRCA2*. Un dépistage précoce (tumeur de moins d'1 cm sans envahissement ganglionnaire) permet un taux de survie à 5 ans supérieur à 90 %. Le dépistage du cancer des annexes est effectué de manière annuelle par un examen clinique gynécologique et une échographie pelvienne, malgré une efficacité limitée, la mesure prophylactique la plus efficace restant l'annexectomie bi-latérale. Une mastectomie bi ou controlatérale prophylactique pourra aussi être envisagée en fonction du pronostic de cancer, ce geste chirurgical permettant de réduire le risque de développer un deuxième cancer de près de 95%[25].

Les apparentés d'un patient atteint peuvent alors bénéficier d'un test génétique ciblé si l'événement génétique responsable est identifié au cours du diagnostic moléculaire. Ainsi en 2015, 5 302 tests pré-symptomatiques ont été réalisés, identifiant un variant en lien avec le syndrome chez 2 284 personnes. Les apparentés porteurs d'un variant pathogène bénéficient d'un suivi médical plus régulier à partir de 20 ans, avec un examen clinique semestriel, auquel s'ajoute un suivi annuel par imagerie associant IRM et mammographie à partir de 30 ans, et d'un examen clinique pelvien annuel. En fonction du risque (évalué au cas par cas), une mastectomie bilatérale préventive et/ou une annexectomie bilatérale peut être discutée.

La prise en charge oncogénétique comprend aussi une évaluation et un soutien psychologique, suite à l'annonce du diagnostic.

Depuis 2016, le GGC (Groupe Génétique et Cancer) a validé *PALB2* comme gène de prédisposition aux cancers du sein et de l'ovaire, au même titre que *BRCA1* et *BRCA2*, suite aux conclusions d'une étude réalisée chez 154 familles[21]. Cette étude montre néanmoins une différence de risque importante entre les porteuses sans histoire familiale et celles possédant deux apparentés au premier degré ayant eu un cancer du sein diagnostiqué avant 50 ans (cf.

Introduction, § I.B.2), suggérant l'implication d'autres facteurs génétiques ségrégant dans les familles et/ou de facteurs environnementaux confondants.

Malgré tout, les variants inactivateurs retrouvés sur les gènes de *BRCA1*, *BRCA2* ou encore *PALB2* n'expliquent qu'une partie des histoires familiales de cancer du sein et de l'ovaire et d'autres facteurs génétiques pouvant être associés à la pathologie doivent être recherchés et évalués pour leur utilité clinique.

II / Base moléculaire du syndrome HBOC

A. Structure du gène *BRCA1*

Le gène *BRCA1* est situé sur le bras long du chromosome 17 (localisation chromosomique : 17q21.31). Il est constitué de 23 exons dont 22 sont codants, sur un locus de 81 kb (1 kb = 1 kilobase = 1000 bases). La transcription de ce gène fournit un transcrit principal de 7.2 kb, codant pour une protéine de 1 863 acides aminés.

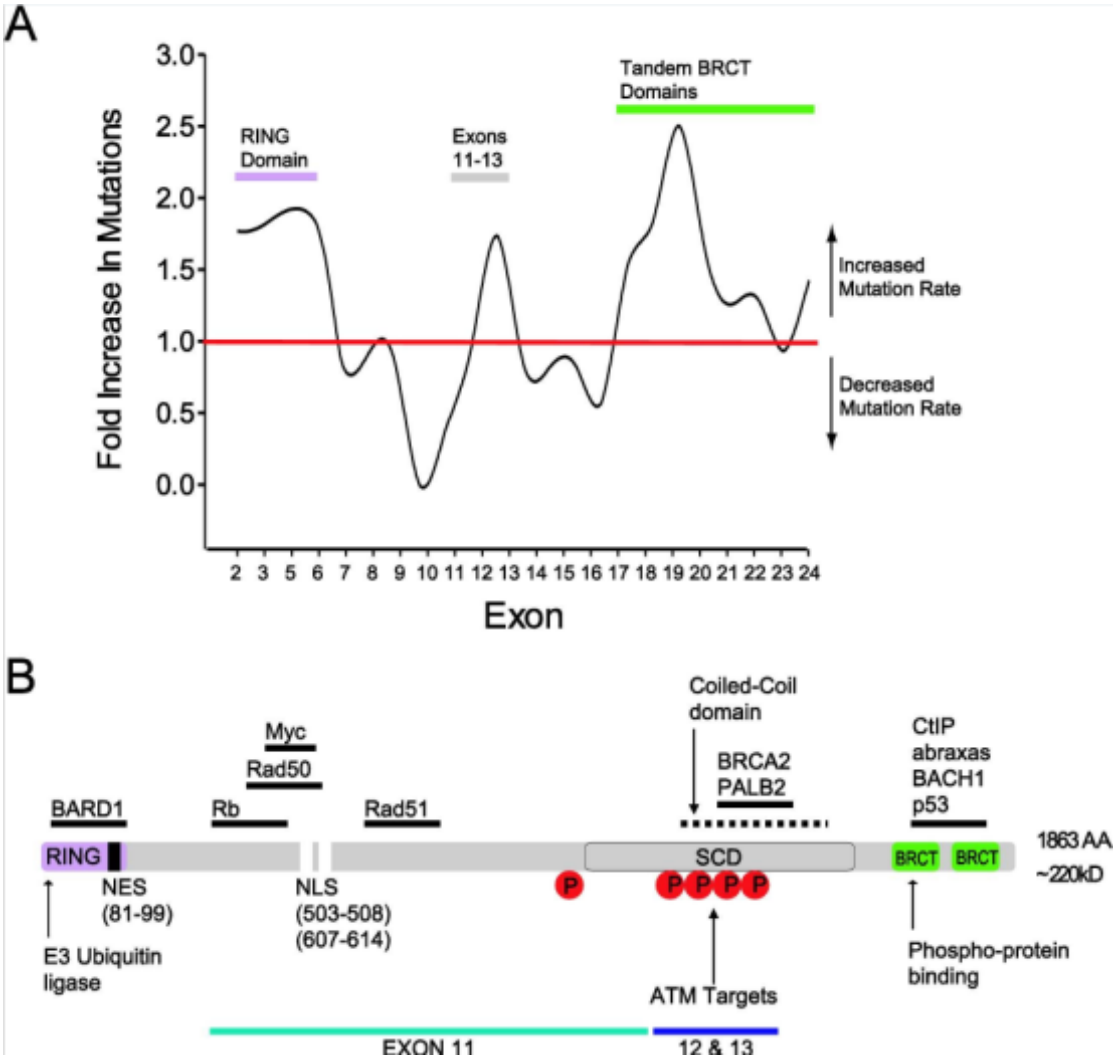


Figure 4: Domaines fonctionnels de la protéine BRCA1 et taux de mutation associés[26] **A**. taux de mutations cliniquement relevantes selon les régions de BRCA1 **B**. Représentation des différents domaines protéiques contenus dans BRCA1. NES : Nuclear Export Sequence. NLS : Nuclear Localization Sequence. SCD : Serine Containing Domain.

La protéine BRCA1 comporte plusieurs domaines « clé », essentiels à sa fonction. Elle possède notamment 2 domaines hautement conservés : RING et BRCT (Figure 4B). Le domaine RING (Really Interesting New Gene), situé en position N-terminale, constitue une structure tridimensionnelle de type « doigt de zinc » qui comprend sept cystéines et une histidine conservées, lui permettant de coordonner deux ions Zn^{2+} stabilisant la structure. Le domaine RING forme une structure globulaire responsable de l'activité E3-ubiquitine ligase de BRCA1[27]. Cette structure va aussi permettre l'interaction avec son partenaire BARD1 (BRCA1 Associated RING Domain 1) via son propre domaine RING. L'activité ubiquitine ligase de BRCA1 sera ainsi fortement augmentée par la formation de cet hétérodimère[28]. La formation de ce complexe va en même temps enfouir la séquence d'export nucléaire (NES : Nuclear Export Sequence), localisée sur l'hélice C-terminale du domaine RING de BRCA1 et BARD1[29], dans une région hydrophobe du complexe. Il en résultera une rétention nucléaire des 2 protéines. L'ubiquitine ligase va intervenir dans plusieurs processus cellulaires différents, via la dégradation des protéines-cibles ubiquitinylées : altération dans l'activation des gènes[30], réparation de l'ADN[31] ou encore remodelage de la chromatine[32]. Ce domaine est donc fondamental pour la fonction de la protéine. Plusieurs études ont ainsi montré que les variants touchant les points clé de cette structure, tels que les résidus cystéine de liaison au Zn^{2+}, pouvaient affecter le repliement du domaine et diminuer l'activité de la ligase[33], conduisant à une augmentation du risque de cancer. Les agents alkylants à base de sels de Platine peuvent aussi avoir les mêmes conséquences, à travers la formation d'adduits via l'atome de Pt sur l'Histidine 117 de BRCA1[34]. Ces différentes observations soulignent l'importance du domaine RING dans l'activité suppressive de tumeurs de BRCA1.

Le domaine BRCT (BRCA1 C-terminal) est un domaine hautement conservé retrouvé dans de nombreuses autres protéines, la plupart impliquées dans la réparation des dommages à l'ADN. Sur BRCA1, ce domaine est retrouvé en tandem et va moduler les interactions entre BRCA1

et les protéines phosphorylées par ATM et ATR (kinases activées par des dommages à l'ADN[35]). Ce domaine possède aussi la capacité à moduler la liaison à l'ADN ainsi que les interactions avec certaines protéines non phosphorylées[36]. Il a en effet été montré la capacité de ce domaine à se lier directement aux cassures d'ADN double-brin (DSB : Double Strand Break).

Une autre région clé de la protéine est contenue dans les exons 11 à 13 de *BRCA1*, qui couvrent environ 65 % de la séquence complète du gène. Même s'ils ne forment pas une structure protéique spécifique, ils contiennent de nombreux domaines d'interaction avec d'autres protéines impliquées dans un large éventail de voies de signalisation, telles que Myc (Facteur de transcription), Rb (Retinoblastoma Protein, contrôle de la progression dans le cycle cellulaire) ou PALB2 (réparation de l'ADN). Le nombre de variants pathogènes intervenant dans ces régions, souvent inactivateurs, souligne l'importance de ces régions pour l'activité suppressive de tumeurs de *BRCA1* (Figure 4**A**).

BRCA1 possède dans cette région deux signaux de localisation nucléaire (NLS : Nuclear Localization sequences) reconnus par la machinerie de l'importine-alpha, afin de réguler le transport de BRCA1 du cytosol vers le noyau. Les variants survenant sur ces régions vont aboutir à une localisation sub-cellulaire de BRCA1 altérée en faveur d'une concentration cytosolique. Les variants des NLS diminueraient donc l'activité suppressive de tumeurs de BRCA1 par une perte de l'activité de réparation de l'ADN.

Cette région contient aussi des zones d'interaction avec Rb[37], RAD50[31] et RAD51[38], c-Myc[39] ou encore PALB2[40], impliquant BRCA1 dans le contrôle du cycle cellulaire, la réparation de l'ADN par jonction d'extrémités non-homologues (NHEJ : Non-Homologous End Joining) ou par recombinaison homologue (HR : Homologous Recombination).

Cette région possède enfin un domaine dénommé SCD (Serine Containing Domain) qui contient une concentration de sites de phosphorylation potentiels (pouvant être phosphorylés par les kinases ATM / ATR). ATM et ATR étant activées par des dommages à l'ADN, celles-ci vont recruter BRCA1 sur les sites de cassures double-brin via sa phosphorylation[41]. La mutation de ces sites de phosphorylation pourrait ainsi perturber le recrutement de BRCA1 et sa fonction de réponse aux dommages à l'ADN.

BRCA1 est ainsi en interaction avec de nombreux partenaires impliqués dans les voies de réparation de l'ADN, différents oncogènes ou gènes suppresseurs de tumeurs, ou encore dans la régulation du cycle cellulaire[42].

B. Structure du gène *BRCA2*

Le gène *BRCA2* est localisé sur le bras long du chromosome 13 (Localisation chromosomique : 13q13.1). Ce gène comporte 27 exons dont 26 sont codants, sur un locus génomique de 84 kb. La transcription de ce gène fournit un transcrit de 10,98 kb, codant pour une protéine de 3418 acides aminés.

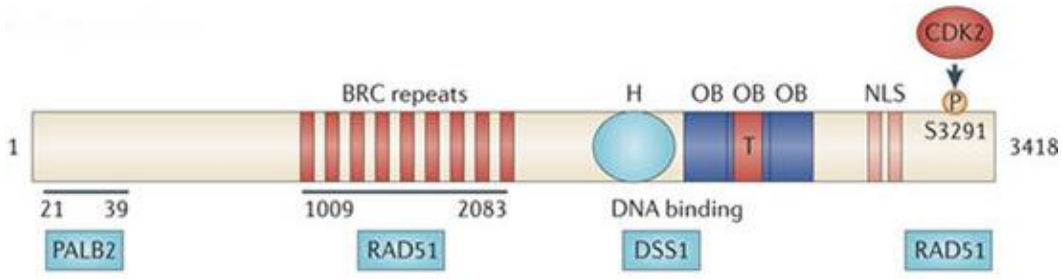


Figure 5: Domaines fonctionnels de la protéine BRCA2[43] H : Helical domain ; OB : Oligonucleotide Binding ; T : Tower Domain ; NLS : Nuclear Localization Domain

A l'inverse de BRCA1 qui possède plusieurs fonctions différentes, BRCA2 va principalement jouer un rôle dans la voie de réparation de l'ADN par recombinaison homologue. Il contient

un domaine de liaison à l'ADN (DBD : DNA-binding Domain) capable de se lier aux ADN simple-brin et aux ADN double-brin[44]. Le DBD contient 5 composants (Figure 5) : une hélice alpha (H), 3 domaines de liaison à des oligonucléotides (OB) qui sont les modules de liaison à l'ADN simple brin, et un domaine TD (Tower Domain) chargé de la liaison aux ADN double-brin.

La protéine contient aussi un motif répété 8 fois, les BRC repeats, chargés de l'interaction avec RAD51[45], et impliquant les 2 protéines dans la réparation des cassures double-brin par recombinaison homologue. Cette interaction est régulée par un site de phosphorylation activé par les CDK[46] (Cyclin-Dependant Kinase), situé dans la région C-terminale, région qui contient aussi deux NLS.

BRCA2 contient aussi dans sa région N-terminale un domaine de liaison à PALB2, dont l'association est requise pour que BRCA2 puisse se localiser au niveau nucléaire[47].

C. Structure du gène *PALB2*

Le gène de *PALB2* (Partner and Localizer of BRCA2) est localisé sur le bras court du chromosome 16 (localisation chromosomique : 16p12.2). Ce gène comporte 13 exons répartis sur un locus génomique de 38,2 kb, codant pour une protéine de 1186 acides aminés.

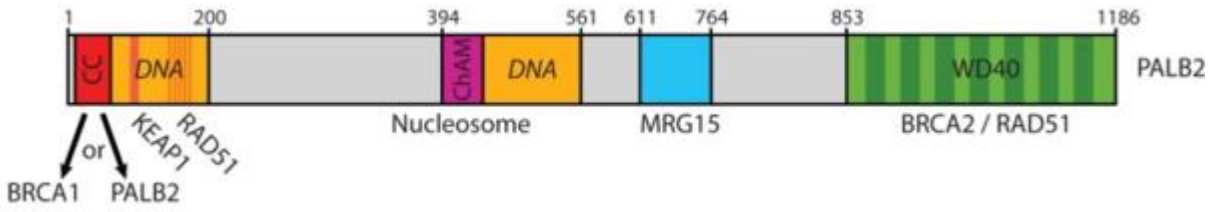


Figure 6: Domaines fonctionnels de la protéine PALB2[48]

PALB2 va pouvoir interagir avec BRCA1 par l'intermédiaire de sa super-hélice N-terminale (Figure 6, CC = Coil-coiled) et avec d'autres protéines PALB2 pour s'oligomériser, et avec BRCA2 par son domaine WD40 en position C-terminale[48], les domaines étant importants pour une recombinaison homologue optimale. Un troisième domaine, ChAM (Chromatin Association Motif) situé au centre de la protéine, est fondamental pour localiser PALB2 au niveau de la chromatine. PALB2 possède enfin un domaine d'interaction avec l'ADN et va promouvoir la formation du filament de RAD51 sur l'ADN simple-brin[49].

D. Implication de BRCA1, BRCA2 et PALB2 dans la réparation des cassures double-brin par recombinaison homologue.

Les cassures double-brin de l'ADN représentent une des formes les plus cytotoxiques de lésion à l'ADN, caractérisées par une cassure simultanée des 2 brins d'ADN complémentaires. Ces cassures peuvent être provoquées par des sources exogènes telles que les thérapies anti-cancéreuses ou encore les radiations ionisantes[50–52], mais aussi apparaître lors de la réplication.

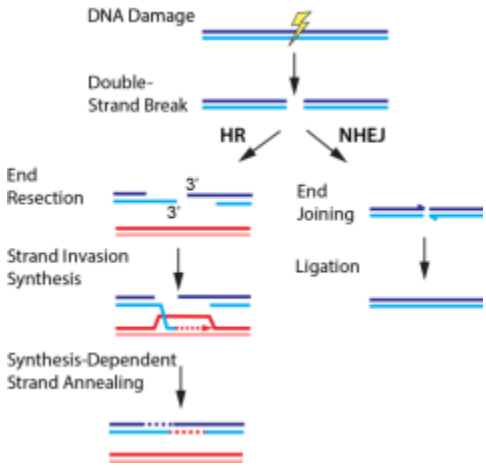


Figure 7: Mécanismes de réparation des cassures double-brin (Finkelsteinlab.org) HR (gauche) : Homologous Recombination ; NHEJ (droite) : Non Homologous End-Joining

Deux mécanismes cellulaires existent pour réparer ces cassures double-brin (Figure 7) : le NHEJ et le HR. Le NHEJ est un système non conservatif conduisant à une modification de l'information génétique[53], mais pouvant intervenir à n'importe quelle phase du cycle cellulaire. A l'inverse le système de réparation par recombinaison homologue va utiliser la chromatide sœur intacte comme modèle afin d'initier une réparation la plus fidèle possible. Cette réparation ne peut donc être réalisée qu'en phase S et G2 du cycle cellulaire.

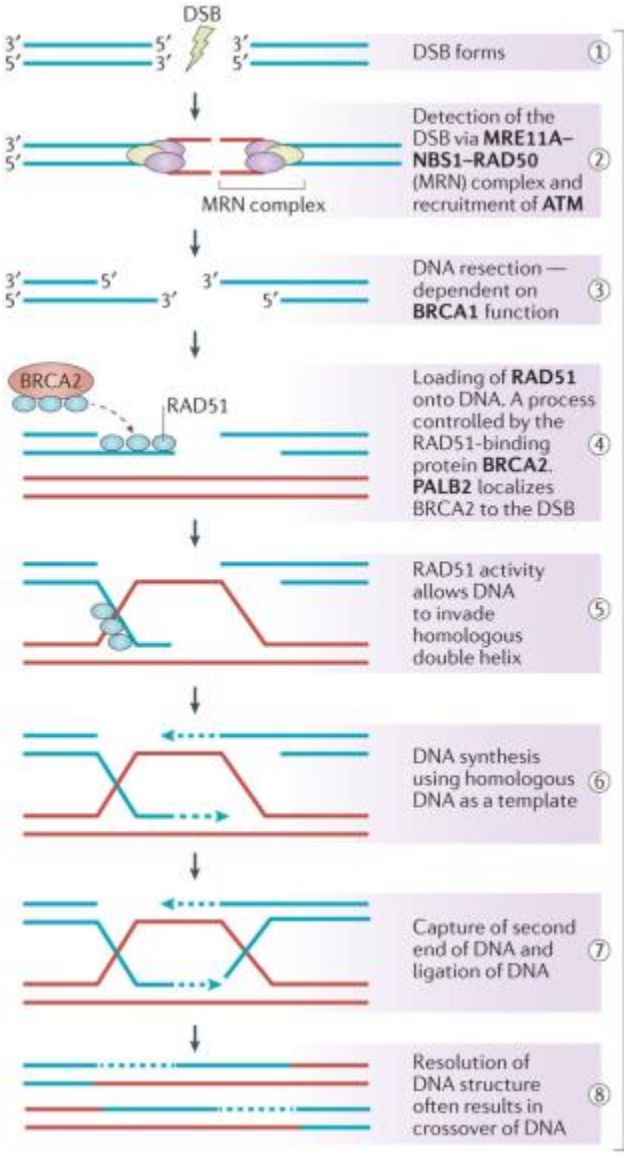


Figure 8: Etapes de la réparation des cassures double-brin par recombinaison homologue[54]

BRCA1 et BRCA2 sont aujourd'hui reconnues comme les protéines au centre de la voie de la recombinaison homologue, via un processus hautement régulé (Figure 8). Ainsi après l'apparition de la cassure double brin (Etape 1), le complexe MRN (MRE11A/RAD50/NBS1) va détecter la lésion et se lier à ses extrémités (Etape 2). Le complexe MRN pourra ainsi entamer la résection des extrémités 5' de la cassure sous le contrôle de BRCA1 et recruter ATM (Etape 3). Cette étape aboutit à l'exposition de 2 régions ADN simple brin de chaque côté de la cassure. BRCA2 va ensuite pouvoir diriger les recombinases RAD51 sur les régions ADN simple brin (étape 4). Ce processus sera aussi régulé par PALB2, qui va co-localiser et stabiliser BRCA2 sur le site de la lésion. Les différentes protéines de RAD51 fixées sur les ADN simple brin vont former un filament nucléoprotéique ayant la propriété d'envahir la double hélice d'ADN (étape 5) intacte et homologue. Les ADN polymérases vont utiliser la séquence ADN homologue comme modèle et l'ADN simple brin comme amorce afin de synthétiser le brin complémentaire (étape 6). Les ADN ligases et les endonucléases vont ensuite résoudre la structure ADN complexe (étapes 7 et 8) formée par les étapes précédentes.

III / Variants génétique à l'origine des prédispositions au cancer du sein et de l'ovaire

A. Types de variants

Les prédispositions décrites précédemment sont en partie provoquées par des variants de la séquence nucléotidique, pouvant toucher des éléments « clés » des processus cellulaires, perturbant ainsi leur fonctionnement.

Ces variants de séquence peuvent être classées en 2 grandes catégories :

1. Les variants de petite taille

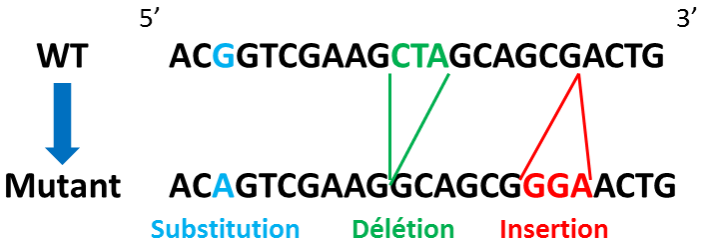


Figure 9: Réarrangements de petite taille. WT : Wild Type (Séquence de référence) / Mutant : Séquence mutée

Ces variants décrivent la modification d'un seul jusqu'à quelques dizaines de nucléotides (Figure 9), avec des conséquences variables en fonction du type et de la région touchée.

Les substitutions concernent la modification d'un seul nucléotide, avec 3 effets possibles si le variant intervient dans une séquence codante (Exon) :

- Faux-sens : La modification de la séquence nucléotidique induit une modification du codon dans lequel elle intervient, avec le recrutement d'un acide aminé différent au moment de la traduction.

- Non-sens : La modification fera apparaître un codon STOP de manière prématurée, pouvant aboutir à la synthèse d'une protéine tronquée. La grande majorité de ces transcrits anormaux sont pris en charge par un système spécifique, le *Nonsens-mediated mRNA Decay* (NMD), chargé de détruire les transcrits anormaux.
- Isosémantique (synonyme) : la modification du codon induite par le variant sera sans conséquence sur la traduction du fait de la redondance du code génétique. Il pourra par contre avoir un impact éventuel sur l'épissage des ARNm.

Les Indels sont des modifications impliquant l'insertion ou la délétion de plusieurs nucléotides successifs simultanément (jusqu'à quelques dizaines). Si le nombre de bases insérées ou délétées est différent d'un multiple de 3 (nombre de bases impliquées dans un codon), le variant provoquera un décalage du cadre de lecture avec l'apparition d'un codon STOP prématurément dans la séquence (variant dit « frameshift »). Autrement, la modification aura pour unique conséquence l'insertion ou la suppression d'un ou plusieurs acides aminés dans la séquence peptidique au moment de la traduction (variant dit « in-frame »).

L'effet des variants introniques est difficile à prévoir mais certains peuvent être responsables d'anomalies de l'épissage. De la même façon, un variant situé dans la région 5'UTR (*Untranslated Region*, en amont du codon d'initiation de la traduction) peut modifier la séquence consensus Kosak (reconnue comme site d'initiation de la traduction par le ribosome) ou encore perturber la formation de structures secondaires complexes régulant la traduction.

2. Les variants de grande taille

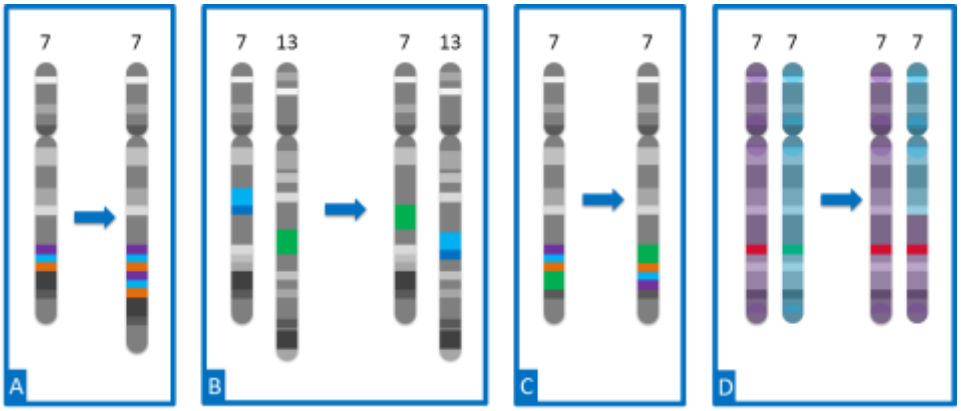


Figure 10: Réarrangements complexes. (A) Amplification ; (B) Translocation ; (C) Inversion ; (D) Perte d'hétérozygotie

Les réarrangements complexes vont concerner des séquences nucléotidiques beaucoup plus longues, de plusieurs centaines de bases jusqu'à des régions chromosomiques entières.

- Amplification (Figure 10A): Augmentation du nombre de copies d'un ou plusieurs gènes, voire de régions chromosomiques entières
- Délétions / gain : Perte ou ajout d'une région chromosomique allant de la taille d'un exon à un ou plusieurs gènes, voire de régions chromosomiques entières
- Translocations (Figure 10B): Déplacement de régions chromosomiques entières entre 2 chromosomes non homologues pouvant aboutir à la juxtaposition de parties de gènes habituellement séparées, conduisant à la création de gènes de fusion
- Inversions (Figure 10C) : Changement d'orientation d'une région chromosomique entière
- Perte d'hétérozygotie (Figure 10D): Perte d'un des 2 allèles d'un organisme diploïde pouvant conduire à une hémizygotie (1 seul allèle sur un locus donné) ou encore au remplacement de l'allèle manquant par une copie de l'allèle restant (cnLOH : copy neutral Loss Of Heterozygosity)

B. Interprétation des variants

La détection des variants au sein d'une séquence nucléotidique est aujourd'hui grandement facilitée, notamment grâce aux technologies de séquençage à haut-débit (cf. Introduction, § VI.B)[55]. Le défi auquel doivent faire face les médecins et généticiens aujourd'hui réside donc de plus en plus dans l'interprétation[56] des variants plutôt que dans leur détection.

Les variants dont l'impact fonctionnel est le plus aisé à anticiper sont les variants non-sens et *frameshift*, puisqu'ils vont aboutir la plupart du temps à la synthèse d'une protéine tronquée et donc non fonctionnelle, ainsi que ceux associés à une forte agrégation familiale. Ces variants sont d'ailleurs ceux retrouvés le plus fréquemment associés au syndrome HBOC[57], concernant jusqu'à 70 % des variants de BRCA1 et 90% des variants de BRCA2[58]. L'imputation de ces variants au syndrome doit néanmoins faire l'objet de précaution. En effet un variant *frameshift* peut pouvoir par exemple provoquer une anomalie de l'épissage qui si elle entraine l'épissage de l'exon porteur du *frameshift* peut représenter un phénomène de sauvegarde si l'exon épissé est en phase et en dehors de tout domaine fonctionnel utile[59]. Aussi l'impact fonctionnel d'une majorité des variants faux-sens, synonymes et introniques est beaucoup plus difficile à évaluer. Les variants faux-sens et isosémantiques peuvent provoquer des modifications de domaines protéiques clés ou encore impacter les sites d'épissage des ARNm. Les variants introniques peuvent également eux aussi perturber la machinerie d'épissage. En conséquence une partie des variants identifiés au cours du diagnostic moléculaire sont des variants de signification inconnue (VSI / VUS : « Variants of Unknown Signifiance »). Ceux-ci représentent 30 % des évènements mis en évidence sur *BRCA1* et *BRCA2*[56] et chacun d'eux sont un véritable défi pour le diagnostic .

La détermination de la pathogénicité d'un VSI nécessite idéalement des études fonctionnelles afin de prouver son caractère délétère et comprendre le mécanisme moléculaire à l'origine de

sa pathogénicité. Cependant, devant la multiplicité des fonctions de BRCA1 et BRCA2 et la complexité de mise en place de tels tests dans chaque laboratoire, il apparaît difficile d'effectuer une confirmation fonctionnelle pour chaque VSI dans un contexte clinique. Ainsi la création de consortiums internationaux permet en partie de répondre à ces difficultés d'interprétation, par le partage des connaissances de laboratoires experts et la mise en place d'un système de standardisation et la classification et des variants à interpréter.

Le consortium ENIGMA[60] (Evidence-based Network for the Interpretation of Germline Mutants Alleles) se focalise sur la signification clinique des variants découverts sur les gènes de *BRCA1* et *BRCA2* et les autres gènes susceptibles de conférer un risque de développer un cancer du sein. Ce consortium a mis en place un système standardisé de classification des variants, basé sur les systèmes suivants :

- Un système de 5 classes utilisé pour l'évaluation quantitative de la pathogénicité des variants utilisant un modèle de vraisemblance multifactoriel[61], se basant principalement sur des informations en provenance d'études de co-ségrégation, de co-occurrence, de cas-contrôles ou encore l'étude des histoires personnelles et familiales. Ces 5 classes comprennent : les variants non pathogènes (classe 1), probablement non pathogènes (classe 2), de signification inconnue (classe 3 = VSI), probablement pathogènes (classe 4), pathogènes (classe 5).
- Le système de 3 classes pour l'interprétation des possibles variants d'épissage développé par le consortium ENIGMA[62]
- Les recommandations de l'ACMG (American College of Medical Genetics and Genomics) pour l'interprétation des séquences variantes[63]

En France, le GGC se charge de l'élaboration, l'organisation et la mise en place des bonnes pratiques de prise en charge des patients et de leur famille. Parmi ses missions, ce groupe a

ainsi développé une base de données focalisée sur les gènes *BRCA1* et *BRCA2*, *BRCA Share*^{TM}, permettant à chaque laboratoire de partager ses connaissances sur les variants découverts dans le cadre d'une prise en charge des cancers héréditaires du sein et de l'ovaire[64]. Les informations contenues dans cette base sont rigoureusement contrôlées afin d'atteindre le même niveau de rigueur, et les VSI régulièrement interprétés par les généticiens des 16 laboratoires appartenant à l'initiative GGC. Le développement et le maintien de bases de données de variants de grade clinique est aujourd'hui un enjeu majeur en génétique.

IV / Bases génétiques des prédispositions

L'étude des familles présentant une forte agrégation de cancers du sein et de l'ovaire a historiquement permis l'identification de *BRCA1* et *BRCA2* comme gènes majeurs de susceptibilité dans la prédisposition aux cancers du sein et de l'ovaire. La présence de variants inactivateurs sur ces gènes augmente donc fortement la probabilité d'apparition de la maladie (mais pas systématiquement : on parle alors de pénétrance incomplète). Depuis, l'étude de l'héritabilité de ces prédispositions a permis d'identifier d'autres facteurs de risque dans l'apparition de ce syndrome, le risque induit étant relié à la prévalence de ces facteurs dans la population[65] (Figure 11). On estime aujourd'hui que 5 à 10 % des cancers du sein et de l'ovaire surviennent dans un contexte de prédisposition héréditaire[66].

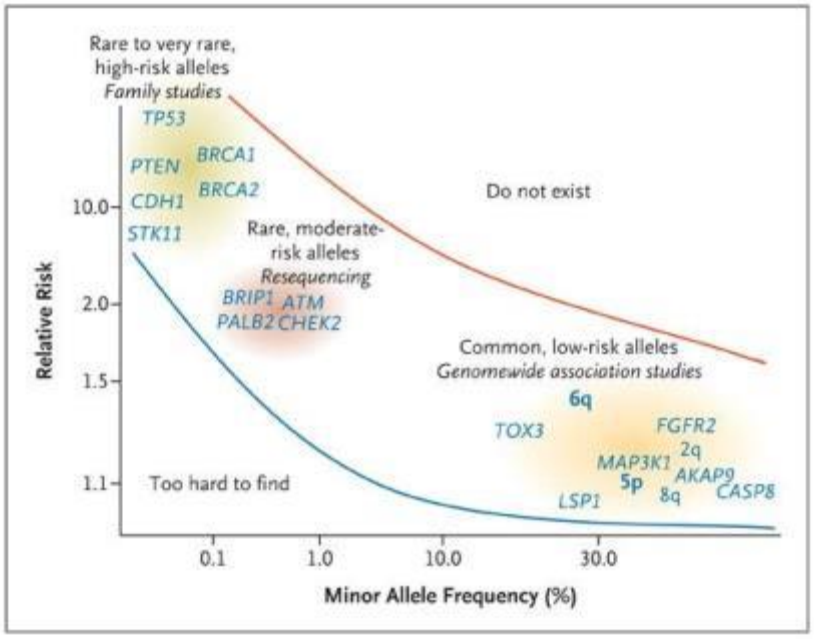


Figure 11: Présentation des facteurs de risque en fonction de leur prévalence dans la population générale[65]

Le lien entre la pathologie et les variants pathogènes d'un gène peut être décrit grâce à deux outils statistiques différents : (i) l'*Odd Ratio* (OR) qui représente la variation de la proportion de gens malades entre groupe « témoin » et groupe « cas », (et décrit une association de la

variable étudiée à la pathologie et par des études généralement rétrospectives) (ii) le risque relatif, qui représente le rapport des incidences de la maladie entre personnes exposées ou non exposées au facteur de risque (par des études de cohortes généralement prospectives). Grâce à ces outils, les allèles (version d'un gène touchée par un variant donné) de susceptibilité aux cancers du sein et/ou de l'ovaire ont pu être classés en 3 groupes majeurs, en fonction du risque associé et de leur prévalence dans la population générale (Figure 11).

Le groupe des allèles à haut risque est constitué généralement d'allèles rares en dehors d'éventuels isolats génétiques (Fréquence en population générale = Minor Allele Frequency = MAF < 0.5%) et de forte pénétrance. Ce groupe rassemble notamment, aux côtés de *BRCA1* et *BRCA2*, des gènes tels que *PTEN*, *TP53*, *CDH1*. Il faut également citer les gènes responsables du syndrome de Lynch étant donné le sur-risque de cancer de l'ovaire provoqué par leur inactivation. Le risque relatif de cancer associé à ces variants est au moins 4 fois plus important que dans la population générale. Les variants pathogènes de *PALB2* ont été montrés associés à une augmentation du risque de développer un cancer du sein, mais cette augmentation est variable en fonction des publications, le classant ainsi de haut risque à risque intermédiaire[21,67]. La plupart des allèles de ce groupe ont pu être mis en évidence par des études de liaison génétique (cf. Introduction, § V.A).

La deuxième catégorie, intermédiaire, comprend des allèles avec une MAF comprise entre 0.5 et 1%, les risques relatifs induits plus modérés, compris entre 2 et 4. Des gènes de cette catégorie tels que *BRIP1* dans les cancers de l'ovaire, *BARD1* et *FANCM* dans les cancers du sein paraissent intéressants à inclure dans un suivi médical guidé par la génétique, mais le risque induit est encore trop faible ou contradictoire entre les publications pour pouvoir être inclus dans le diagnostic moléculaire du syndrome HBOC. Les gènes de cette catégorie ont été mis en évidence par l'étude des partenaires interagissant avec les gènes de haut-risque cités précédemment dans des études cas-témoins[68,69].

La dernière catégorie regroupe des allèles de faible pénétrance et relativement communs dans la population, et induisant un risque relatif de 1,2 à 1,5 fois celui de la population générale de développer un cancer du sein et / ou de l'ovaire. L'association de ces évènements au syndrome est difficile à démontrer puisqu'elle nécessite des études incluant plusieurs milliers voire dizaines de milliers d'individus[70].

A. *BRCA1* et *BRCA2* : Syndrome HBOC

Les gènes *BRCA1* et *BRCA2* sont les 2 premiers à avoir été associés au syndrome de prédisposition au cancer du sein et de l'ovaire. Leur mise en évidence a fait suite à l'observation de familles présentant de nombreux cas de cancers du sein et de l'ovaire[18]. Ces deux gènes ont dès lors été intensément étudiés, démontrant que leur inactivation augmentait considérablement le risque de développer un cancer du sein ou de l'ovaire en comparaison avec la population générale.

Il est aujourd'hui établi que les porteuses de variants pathogènes inactivant la fonction *BRCA1* ont un risque cumulé de 72 % (IC 95%, 65-79%) de déclarer un cancer du sein à l'âge de 80 ans, et de 44 % (IC 95%, 36-53%) pour le cancer de l'ovaire[71]. Le risque de développer un cancer du sein contralatéral 20 ans après le premier diagnostic est estimé à 40 % (IC 95%, 35-45 %). Dans le cancer du sein, l'incidence moyenne (standardisée sur l'âge) observée chez les personnes porteuses d'un variant sur *BRCA1* est de 16.6 (IC 95%, 14.7-18.7) pour 1 000 PA. Cependant, les tranches d'âge les plus jeunes, 21-30 ans et 31-40 ans, possèdent l'incidence la plus élevée, avec des taux standardisés de 73.7 (IC 95%, 42.9-126.8) et 46.2 (IC 95%, 37.3-57.1) pour 1 000 PA respectivement. Concernant le cancer de l'ovaire, celle-ci est de 49.6 (IC 95%, 40.0-61.5) pour 1 000 PA, sans variation notable par tranche d'âge.

Concernant les porteuses de variants sur *BRCA2*, les risques cumulés à 80 ans sont de 69 % (IC 95%, 61-77%) pour le cancer du sein, 17 % (IC 95%, 11-25%) pour le cancer de l'ovaire,

le risque de cancer contralatéral 20 ans après le premier diagnostic étant estimé à 26% (IC95%, 20-33%)[71]. Dans le cancer du sein, l'incidence moyenne (standardisée sur l'âge) observée chez les personnes porteuses d'un variant sur *BRCA2* est de 12.9 (IC 95%, 11.1-15.1) pour 1 000 PA. Les mêmes observations que pour les porteuses d'un variant *BRCA1* sont faites par tranche d'âge, les tranches d'âge 21-30 ans et 31-40 ans possédant les taux standardisés les plus élevés (60.8 (IC95%, 25.5-144.9) et 20.3 (IC 95%,13.5-30.5) pour 1 000 PA respectivement). Concernant le cancer de l'ovaire, celle-ci est de 13.7 (IC 95%, 9.1-20.7) pour 1 000 PA.

De plus, l'histoire familiale (au moins 2 apparentés du premier ou du second degré diagnostiqués pour un cancer du sein) semble augmenter encore le risque de cancer du sein chez les personnes porteuses d'un variant *BRCA1*, avec un *Hazard Ratio* (HR) estimé à 1.99 (IC95%, 1.41-2.82). La même observation est réalisée pour les personnes porteuses d'un variant BRCA2, avec HR de 1.91 (95% IC, 1.08-3.37). Aucune différence n'est en revanche constatée pour les cancers de l'ovaire.

Même si les cas de cancer du sein chez l'homme ne représentent qu' 1 % des cas de cancers du sein, un variant inactivateur sur *BRCA1* ou *BRCA2* favorise aussi l'apparition de la pathologie. Ainsi les porteurs d'un variant sur *BRCA1* auront un risque cumulé à 70 ans de 1,2 % de développer un cancer du sein, et 6,8 % pour *BRCA2*[72]. La présence d'une prédisposition augmente aussi le risque de cancer de la prostate, mais surtout avec les variants de *BRCA2*: à 80 ans, un individu prédisposé aura un risque cumulé de 20 % de développer un cancer de la prostate[73]. L'association avec les variants de *BRCA1* est moins évidente, avec un risque relatif 1,8 fois plus important de développer un cancer de la prostate avant 65 ans.

B. *TP53*: Syndrome de Li-Fraumeni

Ce syndrome a été caractérisé pour la première fois en 1969 à travers 4 familles dans lesquelles un nombre anormalement élevé d'enfants ont développé un rhabdomyosarcome ainsi qu'un nombre élevé de tumeurs apparues chez les apparentés du premier ou du second degré[74]. Ce syndrome, transmis sur un mode autosomique dominant, se caractérise une agrégation familiale de cancers d'apparition précoce ou le développement de tumeurs multiples chez le même individu, comprenant principalement les sarcomes des tissus mous, les cancers du sein développés avant la ménopause, les tumeurs du cerveau, les carcinomes adrénocorticaux et les leucémies.

Un lien a été établi en 1990 entre la pathologie et le gène *TP53*[75], gène suppresseur de tumeurs majeur. Le produit de ce gène, le facteur de transcription p53 (aussi appelé la « gardienne du génome »[76] en raison de son rôle dans la conservation de l'intégrité de l'ADN) possède de multiples fonctions, avec des implications dans l'arrêt du cycle cellulaire, la réparation de l'ADN, la stabilité génomique, la senescence, la différenciation cellulaire, l'autophagie, l'angiogenèses ou encore le métabolisme[77]. Ainsi 49 % des femmes et 21% des hommes portant un variant de *TP53* développeront un cancer avant l'âge de 30 ans[78].

Concernant les cancers du sein, le risque relatif pour les porteurs d'un variant de TP53 est estimé à 6.4 (IC 95%, 4.3-9.3)[79].

C. *PTEN* : Syndrome de Cowden

Ce syndrome, de transmission autosomique dominante, porte le nom de la famille chez qui il a été décrit pour la première fois en 1963[80]. Il se caractérise par des hamartomes multiples se formant sur la peau, la poitrine, la thyroïde, le tractus gastro-intestinal, l'endomètre et le cerveau, ainsi que par un risque accru de développer des tumeurs malignes (principalement au

niveau du sein, de l'endomètre et de la thyroïde). Des variants constitutionnels touchant le gène *PTEN* (Phosphatase and Tensin Homolog) ont ensuite pu être associés à ce syndrome[81].

Le gène *PTEN* est un gène suppresseur de tumeur avec une fonction de PI (phosphoinositide) 3-phosphatase, pouvant inhiber la prolifération cellulaire, la survie et la croissance par inactivation de la voie de signalisation PI 3-kinase dépendante[82]. Des variants inactivateurs de ce gène sont retrouvés chez 80 % des patients diagnostiqués pour ce syndrome[83].

Le cancer du sein reste la tumeur la plus fréquemment associée à ce syndrome, avec un risque cumulé de 25 à 50 % de développer un cancer du sein au cours de la vie d'un individu porteur[84].

A. PALB2

PALB2 fut initialement découvert par l'étude des partenaires de *BRCA2*, par des expériences de co-immunoprécipitation[85]. Des variants bi-alléliques de ce gène ont par la suite été rapidement identifiés et associés à l'anémie de Fanconi[86]. Son implication dans la prédisposition aux cancers du sein a été investiguée en parallèle par des études de ségrégation familiale, évaluant en premier lieu le risque relatif induit par l'inactivation de ce gène à 2.3 (IC 95%, 1.4-3.9)[87].

L'association d'une prédisposition au cancer du sein avec des variants inactivant le gène *PALB2* a ensuite été réévaluée plus récemment, suggérant une augmentation du risque de développer un cancer du sein de 8 à 9 fois avant 40 ans, 6 à 8 fois entre 40 et 60 ans, et 5 fois après 60 ans. Le risque cumulé pour une femme porteuse d'un variant constitutionnel de *PALB2* de déclarer un cancer du sein est de 35 % (IC 95%, 26-46%) à 70 ans en l'absence de

tout contexte familial, et de 58 % (IC 95% , 50-66%) pour celles dont 2 apparentés au premier degré ou plus ont déclaré un cancer du sein avant 50 ans[21].

Cette augmentation du risque reste néanmoins difficile à estimer étant donné la rareté des variants étudiés, les divergences de risque pouvant être observées entre les publications[67].

B. *STK11* : Syndrome de Peutz-Jeghers

Le syndrome de Peutz-Jeghers est une maladie de transmission autosomique dominante caractérisée par le développement de multiples polypes hamartomateux dans le tractus intestinal, une pigmentation cutanée caractéristique (lentiginose) et une augmentation du risque de développer un cancer colorectal, pancréatique, mammaire et ovarien[88]. Cette pathologie a été associée à des variants inactivateurs constitutionnels sur le gène de *STK11* (Serine Threonine Kinase 11)[89].

STK11 est caractérisé comme gène suppresseur de tumeur, avec des implications dans l'arrêt du cycle cellulaire, l'apoptose, les voies de signalisation de Wnt et du TGF-β, ou encore la polarité cellulaire[90].

Ainsi le risque cumulé de cancer du sein chez les patientes diagnostiquées pour ce syndrome est de 31 % à 60 ans[91].Concernant le cancer de l'ovaire, le risque cumulé est estimé à 21% à 65 ans[92].

C. *MLH1* / *MSH2* / *MSH6* / *PMS2* : Syndrome de Lynch

Le syndrome de lynch, de transmission autosomique dominante, se caractérise par une prédisposition à développer principalement des cancers colorectaux sans polypose (HNPCC :

Hereditary Non-Polyposis Colorectal Cancer) et de l'endomètre, mais aussi de l'ovaire, de l'estomac, du tractus urinaire, du pancréas ou encore de l'intestin[93].

Ce syndrome de prédisposition est provoqué par l'altération de la voie de réparation des mésappariements de l'ADN (MMR : MisMatch Repair), touchant principalement les gènes *MLH1*, *MSH2*, *MSH6* et *PMS2*. L'association de ces gènes au syndrome a d'abord été faite par l'observation d'instabilités microsatellitaires dans les HNPCC[94], montrant l'implication du système MMR dans la pathologie (via des résultats initialement observés chez la levure[95]). Le clonage de ces quatre gènes a ensuite conduit à leur association à la pathologie.

Les défauts de cette voie de réparation sont ainsi estimés responsables de 10 à 15% des cas de cancers de l'ovaire héréditaires[96]. Le risque cumulé sur la vie de développer un cancer de l'ovaire a été estimé à 8% (IC 95%, 5.8-10.3)[97], ces risques nécessitant d'être affinés, notamment par gène.

D. *CDH1* : Cancers gastriques héréditaires diffus

Même si la majorité des cancers gastriques est d'origine sporadique, 1 à 3 % de ces cancers résultent d'un syndrome de prédisposition sur un mode autosomique dominant, augmentant aussi le risque de développer un cancer du sein lobulaire[98]. Des variants pathogènes portés par le gène suppresseur de tumeurs *CDH1* (E-Cadherin 1) ont pu être associés à la pathologie, montrant que le risque cumulé de développer un cancer gastrique chez l'homme était de 67% et de 83 % chez la femme[99]. Concernant le cancer du sein, le risque cumulé chez les porteuses d'un variant pathogène sur *CDH1* est de 39 %.

CDH1 code pour une glycoprotéine membranaire essentielle aux jonctions d'adhérence entre cellules épithéliales[100]. Elle joue un rôle clé dans la différenciation et la polarisation des

cellules épithéliales durant le développement embryonnaire et est impliquée dans la régulation de la migration, la prolifération, l'apoptose et la différenciation cellulaire. La perte d'adhésion médiée par l'E-cadhérine va ainsi caractériser la transition d'une lésion bénigne vers une tumeur invasive, métastatique[101].

E. ATM

Les variants inactivateurs du gène *ATM* (Ataxia Telangiectasia Mutated) ont d'abord été caractérisés comme étant à l'origine de l'ataxie-télangiectasie (A-T), maladie autosomique récessive multi systémique caractérisée par une ataxie cérébelleuse, une télangiectasie oculo-cutanée, une radiosensibilité, et une prédisposition aux tumeurs lymphoïdes chez l'enfant et aux immunodéficiences[102]. Néanmoins il a aussi été observé le développement de tumeurs épithéliales plus tardives chez les patients porteurs de cette pathologie[103]. La plupart des patients A-T étant hétérozygotes composites, il a donc été suggéré que les porteurs hétérozygotes, apparentés des patients A-T, pouvaient avoir un risque de cancer augmenté.

ATM est une protéine kinase impliquée dans la réparation des cassures double-brin, nécessaire à l'activation du checkpoint de dommages à l'ADN et à une décondensation locale de la chromatine[104], après recrutement par le complexe RMN. Il a ainsi été montré qu'un variant hétérozygote sur ATM pouvait augmenter le risque relatif de développer un cancer du sein par 2 en comparaison avec la population générale[105], et jusqu'à 5 avant 50 ans. Le risque cumulé de développer un cancer du sein a ainsi été estimé à 60 % à 80 ans[106], mais reste à affiner.

F. CHEK2

CHEK2 (Checkpoint Kinase 2) est une kinase du checkpoint G2 du cycle cellulaire. Activée par phosphorylation par ATM, CHEK2 pourra ensuite interagir avec p53[107] et BRCA1[108] dans la réponse aux cassures double-brin de l'ADN, et va aussi bloquer l'entrée en mitose de la cellule[109].

L'implication de ce gène a d'abord été suggérée par la découverte de variants inactivateurs hétérozygotes chez des familles diagnostiquées pour le syndrome de Li-Fraumeni sans variant identifié sur *TP53*[110]. Son implication dans la prédisposition a ensuite été montrée au cours de l'exploration des prédispositions aux cancers du sein et de l'ovaire non attribuables à *BRCA1* ou *BRCA2*, par une analyse de liaison génétique sur une famille prédisposée[111]. Le variant le plus régulièrement décrit et étudié dans ces familles prédisposées reste le c.1100delC, l'implication d'autres variants du gène dans la prédisposition restant incertaine[112].

Les personnes porteuses du variant c.1100delC ont ainsi à 80 ans un risque cumulé de développer une tumeur ER-positive ou négative de 20 % et 30 % respectivement, contre 9 % et 2 % dans la population générale[113].

G. BRIP1 / FANCJ / BACH1

BRIP1 (BRCA1-Interacting Protein 1) est une hélicase interagissant avec les domaines BRCT de BRCA1, se co-localisant avec celle-ci sur les sites de dommage à l'ADN pour contribuer au processus de réparation[114]. L'interaction entre BRCA1 et BRIP1 est en effet requise pour la réparation des cassures double-brin, notamment pour le passage du checkpoint de la phase G2/M du cycle cellulaire[115].

Les variants inactivateurs du gène *BRIP1* ont tout d'abord été associés au groupe de complémentation J de l'anémie de Fanconi[116] (chez des porteurs bi-alléliques), avant d'être décrits à l'état hétérozygote comme conférant un risque relatif de 2 de provoquer un cancer du sein[69]. Cette prédisposition a été remise en cause récemment par une étude Cas-Témoins portant sur plusieurs dizaines de milliers d'individus, ne montrant pas d'augmentation du risque de développer un cancer du sein significative[117]. Les variants inactivateurs de *BRIP1* semblent néanmoins conférer un risque accru de cancer de l'ovaire, avec un risque cumulé à 80 ans de 5,8 % (IC 95%, 3.6-9.1%) de développer un cancer de l'ovaire[68].

H. Les paralogues de RAD51

A la fin des années 1980, l'analyse de lignées cellulaires sensibles aux radiations ionisantes et aux UV a mis en évidence 2 gènes responsables d'un déficit dans le mécanisme de la recombinaison homologue[118,119] : *XRCC2*[120] et *XRCC3*[121]. L'analyse des séquences de *XRCC2* et *XRCC3* a montré une homologie significative avec la recombinase RAD51[122]. Trois autres paralogues de RAD51 ont été identifiés par des analyses de bases de données de séquences[123–125] : *RAD51B*, *RAD51C*, *RAD51D*. Ces 5 paralogues possèdent de 20 à 30 % d'homologie avec RAD51, principalement dans le domaine de liaison et d'hydrolyse de l'ATP[126]. Ils vont ainsi former deux complexes protéiques distincts : l'un contenant RAD51B, RAD51C, RAD51D et XRCC2 (complexe BCDX2), l'autre contenant RAD51C et XRCC3 (complexe CX3). Ces deux complexes vont agir en amont (BCDX2) et en aval (XC3) du recrutement de RAD51[127].

Parmi tous les paralogues de RAD51, seule une implication de *RAD51C*[128] et *RAD51D*[129] dans la prédisposition aux cancers du sein et de l'ovaire a pu être suggérée, par l'identification de variants inactivateurs dans des cas de cancers familiaux sans variant sur *BRCA1* ou

BRCA2. Les variants inactivateurs de RAD51C confèreraient un risque relatif de développer un cancer de l'ovaire de 5.2 (IC 95%, 1.1-24) et ceux de RAD51D de 12 (IC95%, 1.5-90)[130], ces risques nécessitant confirmation du fait de l'incertitude observée. En revanche aucun risque significatif n'a pu être mis en évidence pour ces deux gènes concernant le cancer du sein.

En revanche, le risque de développer un cancer en lien avec un variant de *RAD51B*[130], *XRCC2*[131] ou *XRCC3*[132] n'est pas encore clairement établi, malgré l'identification de variants constitutionnels potentiellement inactivateurs sur ces gènes.

I. BARD1

BARD1 (BRCA1-Associated RING Domain Protein) est une protéine qui partage des similarités à la fois structurales et fonctionnelles avec BRCA1, notamment un domaine RING en position N-terminale ainsi que deux domaines BRCT. L'hétérodimérisation de BARD1 et BRCA1 via leur domaine RING est essentielle pour l'activité ubiquitine-ligase du complexe formé et la stabilisation des 2 protéines[28].

Les variants faux-sens de *BRCA1* touchant le domaine d'interaction entre les 2 protéines sont connus pour être délétères, empêchant leur intération[133]. De plus *le knock-out* de *BARD1* chez la souris provoque une létalité embryonnaire précoce, avec un phénotype proche de celui provoqué par un knock-out de *BRCA1*, soulignant ainsi l'effet fonctionnel similaire provoqué par leur extinction[134]. De nouveaux variants probablement causaux sont régulièrement identifiés dans ce gène [135], mais aucune étude n'a pu encore valider le risque induit dans les cancers du sein et/ou de l'ovaire.

V / Les moyens de l'exploration de l'hérédité manquante

Tous les facteurs de prédisposition présentés précédemment n'expliquent qu'une partie des syndromes héréditaires du cancer du sein et de l'ovaire, interrogeant sur les causes d'apparition de la pathologie chez les patients sans facteur de prédisposition connu identifié.

Les premiers facteurs de risque aux cancers du sein et de l'ovaire ont pu être mis en évidence par des études de liaison génétique (cf. Introduction, § V.A), mettant en lumière les gènes de prédisposition *BRCA1* et *BRCA2* dont l'inactivation confère un haut risque de développer la pathologie. La pathologie ségrége sur un mode autosomique dominant à pénétrance élevée, nécessitant un nombre relativement faible de participants à une étude pour démontrer un effet significatif. Les premiers liens de causalité entre locus et pathologie ont ainsi pu être démontrés sur 214 familles pour *BRCA1*[136], et 15 familles (exemptes d'anomalies de *BRCA1*) pour *BRCA2*.

Les gènes associés à un risque modéré de développer un cancer du sein et / ou de l'ovaire (variants de pénétrance modérée) ont ensuite été majoritairement mis en évidence par des approches « gène-candidat » grâce à des études cas-témoins (focalisées sur un gène) ou des études familiales, comme *CHEK2*[111], *BRIP1*[69] ou encore *ATM*[137]. Du fait du risque induit modéré et de la faible fréquence en population de ces variants (0,5 - 1 %), les études démontrant leur implication dans le syndrome nécessitent la mise en place de populations d'étude de taille plus importantes (plusieurs centaines voire plusieurs milliers d'individus)[138].

A. Les études de liaison génétique

Cette méthodologie, étudiant des familles dans lesquelles ségrége la pathologie, a pour objectif de localiser la région génomique associée à une pathologie en démontrant une co-ségrégation de la maladie avec des marqueurs génétiques dont la localisation chromosomique est connue. La région génomique identifiée sera ainsi plus susceptible de contenir des variants génétiques causaux[139]. Cette région pourra ensuite explorée plus finement par des techniques de clonage positionnel[140] (Figure 12), aboutissant *in fine* à la détection du gène responsable et à l'identification de ses variants.

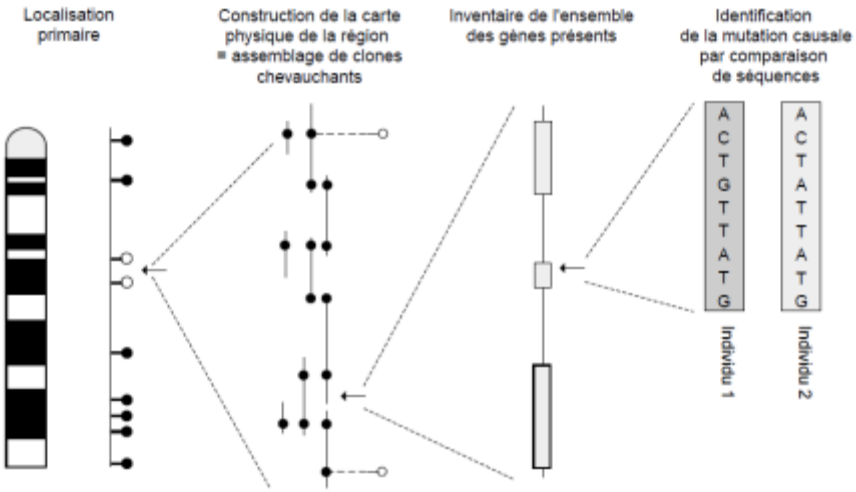


Figure 12: Etapes d'Identification d'un gène candidat par étude de liaison génétique[140]

B. Les études d'association pan-génomiques (GWAS)

L'ensemble des cas de prédisposition aux cancers du sein et de l'ovaire n'étant pas expliqués par la présence de variants à pénétrance forte ou modérée, le déterminisme génétique des cancers pourrait être expliqué par un modèle polygénique complexe impliquant une combinaison de plusieurs variants fréquents dans la population générale et à faible pénétrance, sous influence environnementale. Ce modèle supporte la théorie du « Common Disease /

Common Variant » (« maladie commune / variant commun »), proposant d'explorer la part génétique des maladies (maladies psychiatriques, métaboliques ou encore cancers).

Cette hypothèse a été testée par modélisation par Antoniou et al.[141], montrant que l'intégration d'une composante polygénique dans des familles à haut risque prédisait le mieux les risques de développer un cancer du sein chez les familles non porteuses de variants sur *BRCA1* ou *BRCA2*. Ces modélisations suggéraient par ailleurs que la composante polygénique pouvait aussi modifier le risque relatif chez les porteurs d'une prédisposition.

L'exploration de ce type de variants a pu être réalisée par des études d'association pangénomiques, ou GWAS, se focalisant sur les polymorphismes, ou SNPs, disséminés sur l'ensemble du génome et présents à plus de 5 % dans la population générale. Le nombre de SNPs dans un génome humain est estimé à environ 10 millions[142]. Malgré cela, les études de GWAS ne vont pas utiliser l'ensemble des SNPs dans leur analyse, mais plutôt tirer parti du fait que des régions génomiques vont avoir tendance à être héritées ensemble, des allèles adjacents étant transmis de manière non indépendante de génération en génération. Cette association non aléatoire d'allèles sur des loci proches, appelée déséquilibre de liaison, permet d'utiliser certains SNPs comme marqueurs pour les autres SNPs proches. Ainsi l'utilisation du déséquilibre de liaison permet de réduire le nombre de SNPs pour caractériser un génome à 500 000[143]. L'objectif de ce type d'études est ainsi de déterminer si certains SNPs sont retrouvés plus fréquemment dans une population de cas par rapport à une population contrôle.

Ces analyses ont ainsi pu mettre en évidence plus de 80 SNPs associés au cancer du sein, n'impliquant malgré tout que 16 % des cancers du sein héréditaires[70]. Le risque relatif induit par ces polymorphismes est de plus en moyenne relativement faible, évalué à environ 1,3 par rapport à la population générale[144].

Les études de GWAS n'ont donc réussi qu'à expliquer une part minoritaire de l'hérédité manquante des cancers du sein et de l'ovaire, et n'impliquant que des variants de faible pénétrance.

De plus, la mise en place de ces études a nécessité la création de consortiums internationaux, afin d'obtenir des preuves de significativité suffisantes. En effet, la puissance d'une étude d'association va dépendre de deux paramètres principaux : (i) la fréquence en population des allèles qui prédisposent à la maladie (ii) la pénétrance de ces allèles. Ainsi, comme le montre la Figure 13, associer des allèles à la maladie avec des risques relatifs de l'ordre de 1,2 ou 1,3 (tels que ceux associés aux SNPs dans les cancers du sein) nécessite des populations cas et contrôle supérieures à 10 000 individus.

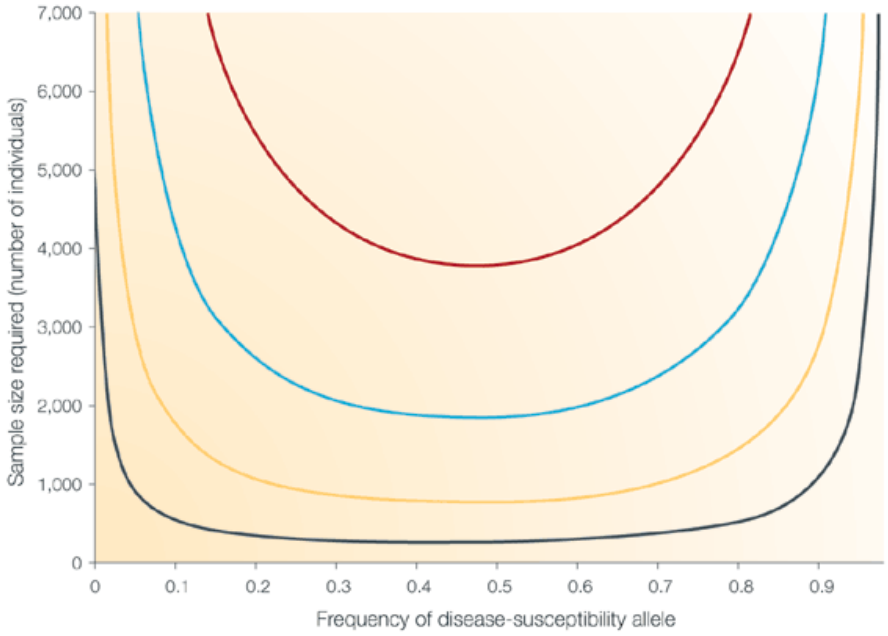


Figure 13: Effets de la fréquence allélique sur la taille requise pour une population en GWAS[145] Le nombre de cas et de controles requis pour une étude en GWAS est représenté en fonction de la fréquence allélique en population des variants associés, induisant un risque relatif de 1,2 (rouge), 1,3 (bleu), 1,5 (jaune) et 2 (noir). La puissance statistique est de 80 % avec un niveau de significativité $p < 10^{-6}$, assumant un modèle multiplicatif pour les effets des allèles et une corrélation parfaite du déséquilibre de liaison.

Le Breast Cancer Association Consortium (BCAC) a ainsi rassemblé 45 290 cas et 41 880 contrôles dans son étude[146] pour associer 41 loci à risque, tandis ce que l'équipe de

Michailidou et al. a effectué une méta analyse rassemblant 120 000 individus pour mettre en évidence 15 loci de susceptibilité au cancer du sein supplémentaires[70].

C. Approches « gènes candidats » et études cas témoins par séquençage extensif

Les études en GWAS n'ont eu qu'une efficacité limitée pour expliquer la part d'hérédité manquante, notamment du fait de la focalisation sur des polymorphismes fréquents (> 5 % en population générale). Ce défaut d'héritabilité pouvait être comblé par la recherche de variants constitutionnels rares voire privés (< 0.1% de fréquence allélique) mais avec une pénétrance plus importante, hypothèse déjà confortée par les travaux réalisés dans le cadre de la prédisposition aux cancers du sein et de l'ovaire par des approches « gène candidat ». Cette approche repose sur une bonne connaissance de la physiologie et de la fonction du gène étudié ainsi que de ses partenaires impliqués dans la même voie métabolique. L'étude des partenaires de *BRCA1* et *BRCA2* ou des gènes impliqués dans les voies de réparation de l'ADN avait déjà permis d'identifier de forts candidats au syndrome de prédisposition, dont *CHEK2*, *BRIP1*, l'inactivation de ces gènes étant caractérisées par de multiples et rares variants inactivants associés à une pénétrance modérée. Malgré cela, en prenant en compte les gènes de pénétrance forte et modérée découverts par les études familiales et les études gènes-candidats, ainsi que les variants de faible pénétrance découverts par les études de GWAS, l'ensemble des évènements n'expliquaient toujours qu'environ un tiers des prédispositions aux cancers du sein et de l'ovaire[147], suggérant que d'autres gènes pouvaient être impliqués. Ces candidats potentiels sont probablement de pénétrance modérée, le risque n'étant pas assez fort pour une identification par étude en agrégation familiale, et les variants impactants étant trop rares pour être mis en évidence par des études de GWAS.

L'arrivée dans les laboratoires des séquenceurs de nouvelle génération (NGS) à la fin des années 2000 a permis de s'affranchir de ces contraintes, notamment grâce à leur débit d'analyse. Ces appareils sont actuellement capables de séquencer l'équivalent de plusieurs génomes humains avec une résolution de l'ordre de la paire de bases en une seule expérimentation. Ainsi le séquençage d'exomes entiers ou de l'ensemble des membres d'une voie de signalisation permet d'envisager une approche naïve et non supervisée plutôt qu'une approche par gène-candidat dans la découverte de nouveaux éléments prédisposants.

Le gène *FANCM* a ainsi pu être associé à un sur-risque de développer un cancer du sein, d'abord par un séquençage de l'exome de 24 patients provenant de 11 familles à risque, identifiant 22 variants délétères sur 21 gènes des voies de réparation de l'ADN. Les variants candidats ont ensuite été recherchés par séquençage ciblé sur une population d'étude de réplication comprenant 3166 patientes atteintes d'un cancer du sein, 569 d'un cancer de l'ovaire, et 2090 contrôles. Le variant non-sens c.5101C>T (Q1701X) a notamment pu être mis en évidence, particulièrement enrichi chez les patientes avec un cancer du sein triple-négatif (OR = 3,56 ; 95% CI = 1,81-6,98 ; P=0,0002)[148].

A la différence des études de GWAS, les études Cas-témoins réalisées par séquençage à haut-débit permettent donc d'explorer la part des variants rares impliqués dans le syndrome de prédisposition. Néanmoins cette exploration requiert toujours une puissance statistique importante, en considérant des populations de plusieurs dizaines de milliers d'individus afin de démontrer une association significative (Figure 14) pour les variants les plus rares.

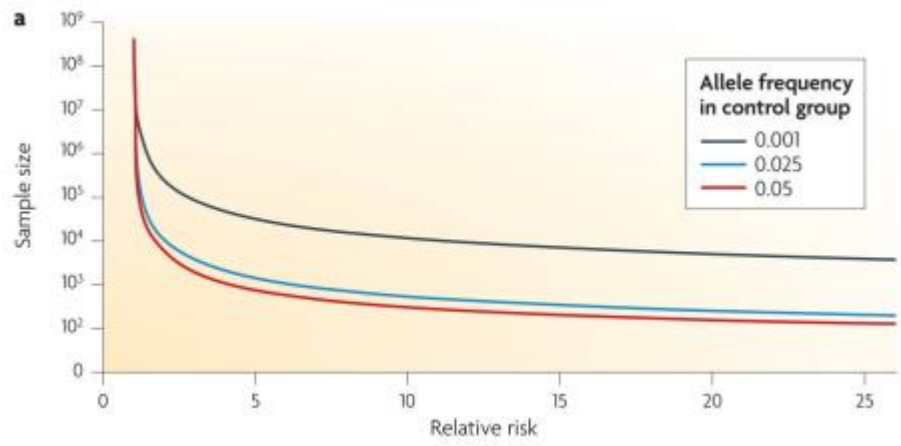


Figure 14: taille de la population nécessaire pour démontrer une association significative d'un allèle à une pathologie, en fonction de du risque relatif induit[149]

La puissance de ces analyses peut être augmentée en agrégeant les variants d'intérêt par gène[150]. Enrichir la population Cas en variants pathogéniques permettrait aussi de diminuer la puissance statistique nécessaire, par une sélection rigoureuse des patients analysés (ex : sélection de phénotypes extrêmes, augmentant la probabilité de détecter un évènement causal).

D. Impact potentiel des néo-mutations dans le syndrome

Le génome d'un individu n'est pas immuable et accumule des variants tout au long de la vie d'un individu, phénomène appelé « instabilité génétique ». Ce processus résulte d'erreurs produites au cours de la réplication cellulaire ou encore de dommages causés à l'ADN, malgré les systèmes de réparation du génome. La plupart des variants issus de ce phénomène apparaissent de manière aléatoire, produisant ainsi un effet généralement défavorable. Par sélection naturelle, les génomes comportant de tels évènements seront la plupart du temps non transmis à cause du décès prématuré de l'individu porteur du variant causal. Certains variants vont au contraire donner un avantage sélectif à l'individu porteur, permettant la propagation du variant au sein de la population. Cette sélection darwinienne est à l'origine de la grande

complexité et diversité génomique des populations et permet aussi l'adaptation des individus à leur environnement.

Dans le cas où le variant *de novo* apparait dans les gamètes d'un des 2 parents, la descendance portera le variant de manière constitutionnelle, sans que celui-ci soit retrouvé dans le patrimoine génétique des parents. On parle alors de variant pré-zygotique, estimés à 38 par individu[151]. Il est donc possible d'imaginer que certains de ces variants pourront se révéler pathogènes et induire le développement d'une pathologie, en l'absence de tout contexte familial. Il a ainsi été montré que ces variants *de novo* étaient retrouvés enrichis dans les altérations de type « perte de fonction » dans plusieurs pathologies telles que l'autisme ou encore les déficiences intellectuelles[152,153]. Concernant les patients atteints par la Neurofibromatose de type 1, 74 % des variants délétères identifiés chez les patients sur le gène *NF1* sont des néo-mutations[154]. Ce phénomène est aussi décrit dans les cancers du sein et de l'ovaire, cependant avec des taux beaucoup plus faibles (0,3 % en moyenne pour les gènes *BRCA1* et *BRCA2*)[155].

Mais cette instabilité génétique touche aussi les lignées somatiques, et pourra créer une diversité génétique au sein du même individu. Si le variant *de novo* apparait pendant les premiers stades du développement embryonnaire (variant post-zygotique), l'individu concerné portera le variant en mosaïque, et l'étendue de cette mosaïque sera fonction de la précocité d'apparition du variant au cours du développement embryonnaire. Ainsi un individu en mosaïque pourra potentiellement présenter un pourcentage variable de cellules mutées dans différents organes. Ce modèle a été particulièrement illustré dans le cadre du rétinoblastome, cancer déclenché par une inactivation bi-allélique du gène *RB1*. Les formes bilatérales (concernant environ 40% des proposants) sont en effet décrites comme issues majoritairement d'un variant constitutionnel hétérozygote de *RB1*, soit hérité (10 % des cas), soit *de novo*. Cependant, il a ainsi pu être montré que 30% des formes bilatérales et 6% des formes

unilatérales diagnostiquées initialement négatives étaient issues d'un variant de *RB1* en mosaïque[156].

Il est donc possible d'imaginer une prédisposition locale, pour un organe ou un tissu, qui puisse favoriser l'apparition d'autres cancers tels que les cancers du sein et de l'ovaire. La mosaïque peut aussi se retrouver au niveau du tissu hématopoïétique et ainsi se retrouver détectable dans le flux sanguin. De rares cas ont d'ailleurs été déjà documentés, impliquant des variants en mosaïque de *BRCA1*[157,158], et retrouvés jusqu'à 5% dans l'ADN leucocytaire. Des analyses plus sensibles et une investigation étendue à d'autres gènes de prédisposition, candidats également, pourraient ainsi permettre d'expliquer certains cas sporadiques sans variant prédisposant identifié par des techniques conventionnelles.

VI / Les méthodes de biologie moléculaire d'exploration du syndrome

A. Méthodes historiques

Le diagnostic du syndrome HBOC est réalisé par la mise en évidence de variants pathogènes sur les gènes de prédisposition associés, notamment *BRCA1*, *BRCA2* et plus récemment *PALB2*.

La technique de référence demeure le séquençage Sanger, méthode par synthèse enzymatique, qui permet une lecture précise des fragments d'ADN d'intérêt, base par base. Cette méthode est aujourd'hui automatisée, les machines les plus récentes (ex : Séquençeur ABI 3730XL) étant capables de séquencer jusqu'à 2100 kb (kilobases / 1kb = 1000 bases) par jour, par fragments de 400 à 900 paires de bases. La mise en place d'une telle méthode dans le cadre d'un laboratoire de diagnostic demande néanmoins des moyens importants, chaque fragment ADN d'intérêt nécessitant une analyse spécifique. A titre d'exemple, le séquençage de toutes les régions codantes de *BRCA1* nécessite environ 35 fragments différents, à réaliser pour chaque patiente analysée.

Des analyses de pré-criblage ont ainsi été mises en place, moins chères et plus rapides, permettant de sélectionner précisément les régions à séquencer : la HRM (High Resolution Melt) ou analyse des courbes de fusion à haute résolution, et la DHPLC (Denaturing High Performance Liquid Chromatography) ou chromatographie liquide à haute performance sur gel dénaturant. Ces deux méthodes, après une première étape d'amplification par PCR, vont se baser sur la formation d'hétéroduplexes (dénaturation de l'ADN double-brins puis hybridation entre brins mutés et brins sauvages) et d'homoduplexes (hybridation entre brins sauvages ou entre brin mutés). En HRM, les homoduplexes et hétéroduplexes sont marqués

par des colorants intercalants fluorescents qui ne se fixent qu'aux ADN double-brins. Une chauffe très progressive des amplicons va ensuite engendrer une dénaturation des brins d'ADN (à une température de fusion Tm), qui perdront simultanément en fluorescence. Il sera ainsi possible de suivre en temps réel la dénaturation des brins d'ADN via une chute de la fluorescence. Les hétéroduplexes, thermiquement moins stables que les homoduplexes du fait de leur mismatch, auront un Tm et une courbe de fusion légèrement différente de celle des homoduplexes, permettant de les distinguer et ainsi de conclure à la présence d'un variant sur l'amplicon analysé. En DHPLC, les ADN sont chauffés à une température de dénaturation déterminée informatiquement, entraînant un désappariement partiel des hétéroduplexes, au niveau du variant. Ce désappariement conduit à une délocalisation partielle des charges négatives des hétéroduplexes, qui seront élués plus rapidement que les homoduplexes dans la colonne, chargée positivement.

Ainsi les régions présentant des variants mis en évidence par les analyses de pré-criblage sont ensuite ré-analysés par séquençage Sanger afin de caractériser précisément le variant mis en cause.

B. La révolution technologique : Les séquenceurs de 2ème génération

Le premier projet de séquençage du génome humain s'est appuyé sur la méthode de séquençage Sanger. Cette entreprise, officiellement lancée en 1990, nécessita la mise en place d'un consortium international d'une vingtaine de centres de recherche à travers le monde. Après onze ans de travail collaboratif, et plus de 3 milliards de dollars, une première version du génome humain (3,2 milliards de paires de bases) fut publiée dans la revue Nature[159], et complétée deux ans plus tard[160].

1. Explosion du débit de séquençage

La première décennie du XXIème siècle a ensuite connu une véritable rupture technologique avec l'apparition des séquenceurs de deuxième génération (NGS) à haut-débit. Ces nouvelles technologies de séquençage massivement parallèle ont permis d'augmenter le débit d'analyse de façon très importante, passant de quelques milliers de paires de bases séquencées à plusieurs milliards (Figure 15).

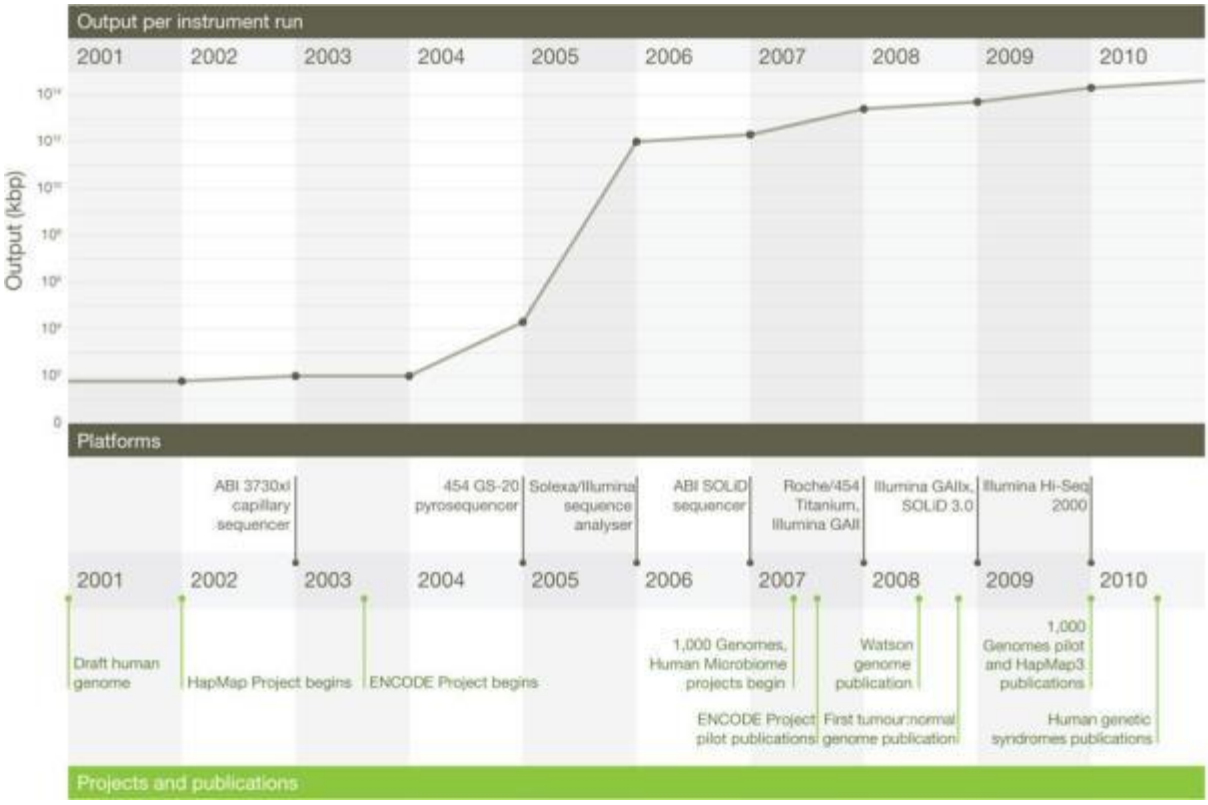


Figure 15: Evolution des débits de séquençage au cours des années 2000[161]

Cette génération de machines est aujourd'hui capable de générer un volume de données équivalent à plusieurs génomes humains complets ou à plus d'une centaine d'exomes en moins de 48 heures (ex : NovaSeq, Illumina).

Les séquenceurs de deuxième génération ont aussi permis dans un deuxième temps de baisser de manière drastique les coûts de séquençage (Figure 16), permettant leur démocratisation dans les laboratoires.

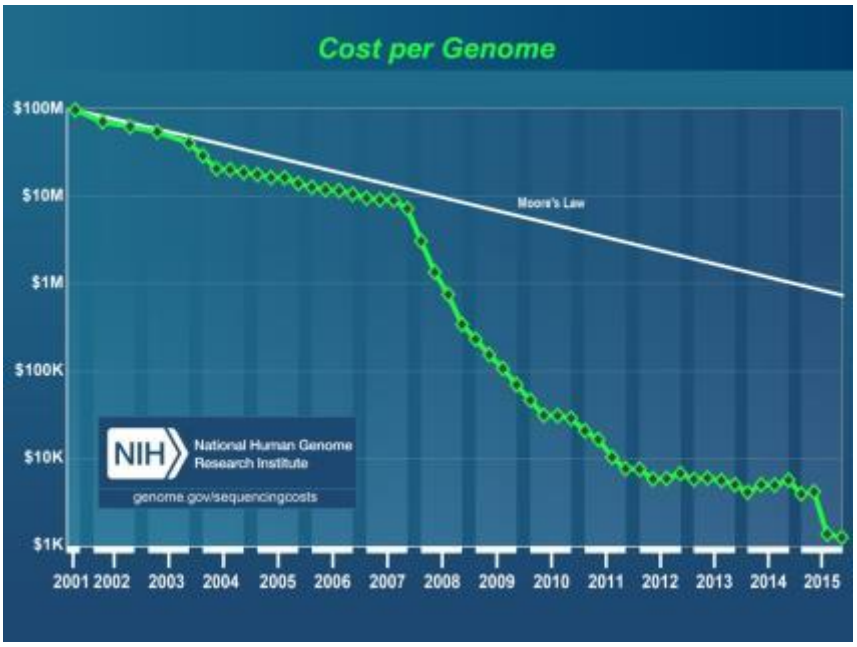


Figure 16: Evolution du coût de séquençage d'un génome complet (www.genome.gov/sequencingcostdata)

2. Les séquenceurs de fragments d'ADN courts

Plusieurs technologies de séquençage à haut-débit sont aujourd'hui accessibles aux laboratoires. Les approches de séquençage se focalisant sur la lecture de fragments ADN courts (de 50 à 300 pb : *short-read sequenci*ng) vont rassembler 2 grandes catégories : le séquençage par ligation (SOLiD) et le séquençage par synthèse (Illumina, Ion Torrent). Dans les approches par ligation, une séquence ADN sonde liée à un fluorophore va pouvoir s'hybrider aux fragments d'ADN à séquencer grâce à une ADN ligase. La couleur de l'émission lumineuse du fluorophore va indiquer l'identité de ou des bases liguées. Dans les approches par synthèse, une polymérase va former le brin complémentaire de la séquence ADN à lire, l'incorporation d'un nucléotide étant signalée par une émission lumineuse (fluorophore) ou par un changement dans la concentration ionique du milieu réactionnel (technique détaillée ci-après). Ces deux techniques nécessitent que l'ADN à étudier soit amplifié de manière clonale sur une surface solide : soit en émulsion sur billes (SOLiD, Ion

Torrent) soit sur phase solide (Illumina). L'obtention de milliers de copies identiques d'un fragment d'ADN permet une amplification du signal lors du séquençage, le distinguant du bruit de fond.

3. Les séquenceurs de fragments d'ADN longs

Le génome comprenant beaucoup d'éléments complexes trop longs pour être résolus par les technologies de *short-read sequencing* d'autres approches ont pu voir le jour avec pour objectif de séquencer des fragments de plusieurs milliers de paires de bases (jusqu'à 200 kb : *long-read sequencing*), permettant de mettre en évidence des variants structuraux de grande taille de l'ADN, ou encore des régions fortement répétées. Elles permettent également de reconstruire des haplotypes complets. Ici aussi, deux types de méthodes se confrontent : le séquençage en temps réel de molécules uniques (*single-molecule real-time sequencing* = SMRT) et l'approche synthétique basée sur les technologies de *short-read sequencing* pour construire des *reads* longs *in-silico*.

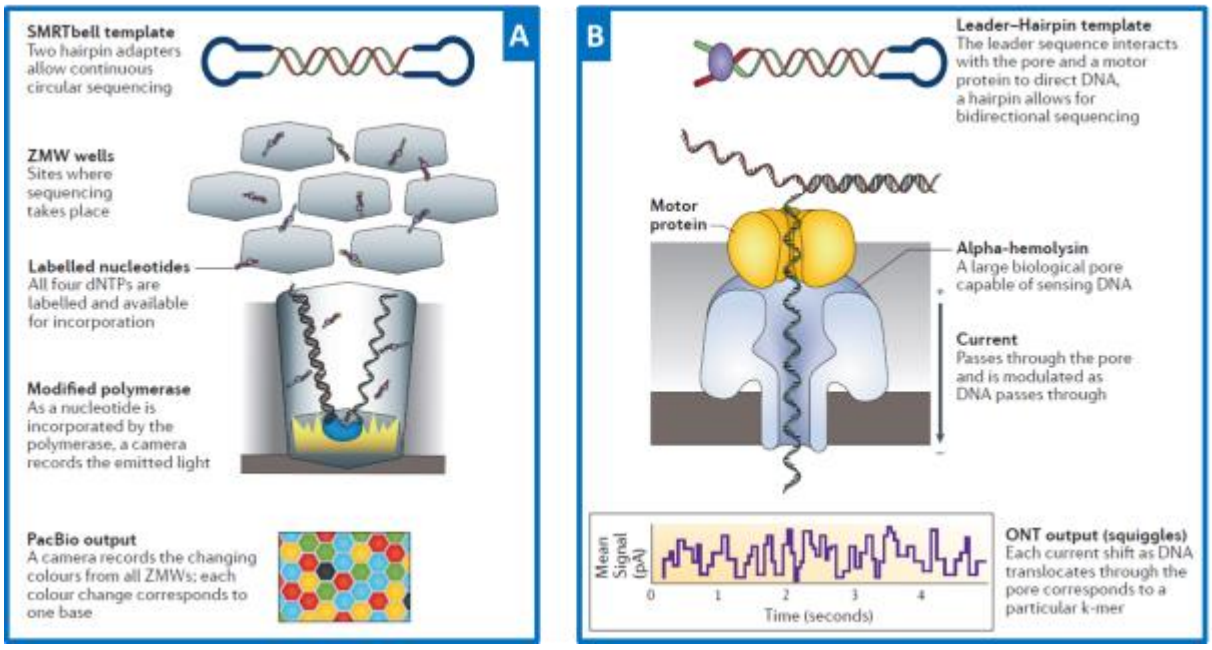


Figure 17: Séquençage en temps réel de molécules uniques[162] (A). Technologie utilisée par les séquenceurs PacBio (Pacific Biosciences) (B) Technologie utilisée par les séquenceurs Nanopore (Oxford Nanopore technologies).

La particularité du SMRT sequencing réside dans le fait que celui-ci n'a pas besoin d'une population clonale ou de fragments ADN amplifiés pour générer un signal détectable. Les séquenceurs PacBio (Pacific Biosciences) possèdent des supports de lecture de l'ADN particuliers, comprenant des milliers de puits de quelques zeptolitres à fond transparent dénommés *zero-mode waveguides*[163] (ZMW), au fond desquels sont fixés une polymérase (Figure 17A). Le brin d'ADN peut ainsi progresser dans le ZMW. L'incorporation des nucléotides se faisant toujours au même endroit du fait de la fixation de l'enzyme, le système peut se focaliser sur une seule molécule. L'incorporation des nucléotides est visualisée en direct par un laser et un système de caméras. De plus le système utilise des ADNs circularisés permettant à la molécule d'être lue de multiples fois par la polymérase. Les séquenceurs de type Nanopore (Oxford Nanopore Technologies) vont eux analyser directement la composition de la molécule d'ADN simple brin plutôt que de monitorer l'incorporation de nucléotides lors de la synthèse du brin complémentaire (Figure 17B). Les ADN simple brin à analyser vont ainsi à travers des protéines transmembranaires possédant un pore en position

centrale (les nanopores), et sur lesquelles est appliquée une tension électrique. Au fur et à mesure que l'ADN traverse le pore grâce à une protéine « moteur » secondaire, le courant électrique passant à travers le pore est modulé. Les changements de tension sont caractéristiques de la séquence ADN passant à travers le pore, pouvant être interprétés comme des k-mers. Plutôt que d'avoir 4 signaux possibles (fonction du nucléotide passant à travers le pore), les séquenceurs Nanopore possèdent plus de 1 000 types de signaux enregistrés, un pour chaque k-mer possible.

Les approches de *long-read* synthétiques vont utiliser des « codes-barres » (barcoding) afin d'associer les fragments séquencés sur des séquenceurs de reads courts. L'objectif ici va être de répartir de grands fragments d'ADN (jusqu'à 10 kb) dans différents puits, de manière à n'obtenir que très peu de molécules dans chaque puits (environ 3 000). Les molécules seront ensuite fragmentées et barcodées. Après séquençage, les données sont ensuite rassemblées par code-barres tout en sachant que les fragments présentant le même code-barres sont dérivés du même grand fragment original.

4. Comparaison des technologies

	Technologie De séquençage	Longueur Des Reads	Capacité par Run	Profil d'erreur	Durée d'un Run	Coût par Gb
ABI 3730xl	Séquençage par terminaison de chaîne	400 – 900 pb	2 100 kb	0,001 %	1 - 3h	2 400 000 $
Illumina NextSeq (High output)	Séquençage par synthèse	75 pb (PE) 150 pb (PE)	50-60 Gb 100-120 Gb	< 1 % Substitution	18 h 29 h	41 $ 33 $
SOLiD 5500xl	Séquençage par ligation	50 pb (SE)	320 Gb	≤ 0,1 % Biais AT	10 j	70 $
Ion PGM 318	Séquençage par synthèse	400 pb (SE)	1 – 2 Gb	1% Indels	7,3 h	450 – 800 $
Illumina Hiseq 2500	Séquençage par synthèse (**synthetic long-read**)	~ 100 kb Longueur synthétique	450 – 500 Gb	0,1 % Substitution	6 j	30 $
Oxford Nanopore MK1 MinION	SMRT	Jusqu'à 200 kb	Jusqu'à 1,5 Gb	~ 12 % indels	48 h	750 $
Pacific Biosciences RSII	SMRT	20 kb	500 Mb – 1 Gb	13 % lecture unique ≤ 1% lecture circulaire	4 h	1 000 $

Figure 18: Exemples de capacité de différentes plateformes de séquençage[162] Vert: Séquençage Sanger ; Bleu : Séquenceurs de fragments courts ; Violet : Séquenceurs de fragments longs ; SMRT : Single-Molecule Real-Time sequencing ; pb : paires de base ; PE : Pair-End ; SE: Single-End; kb = kilobase = 1 000 pb ; Gb = Gigabase = 1 000 000 000 pb. Le séquenceur Illumina Hiseq est ici présenté dans le contexte des long-reads synthétiques.

Comme illustré précédemment, les technologies de séquençage de nouvelle génération constituent un progrès considérable dans la lecture de l'ADN, notamment en termes de coût et de rapidité (Figure 18). Il conviendra de plus de choisir judicieusement la technologie employée en fonction de l'application. Les séquenceurs de fragments courts offrent une plus grande précision de lecture, permettant une détection plus sensible des variants de petite taille (de 1 à quelques dizaines de nucléotides), avantage précieux dans des domaines tels que la cancérologie pour la détection des variants somatiques. Les séquenceurs de fragments longs, malgré un taux d'erreurs plus important, mettront plus facilement en évidence des réarrangements structuraux de grande taille, ou pourront lire des transcrits ARNm entiers.

5. Les applications pour l'exploration des variants génomiques

Les échantillons d'ADN doivent subir plusieurs étapes préliminaires de préparation, permettant optionnellement d'enrichir les régions d'intérêt (si le séquençage ne concerne pas le génome entier) et de les rendre compatibles pour une prise en charge par l'instrument d'analyse.

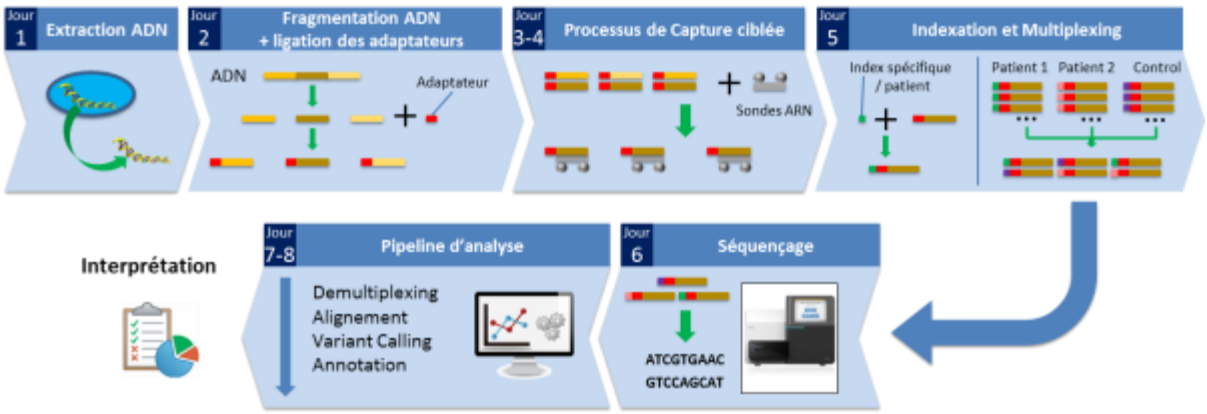


Figure 19: Etapes successives d'enrichissement des échantillons d'ADN par capture. Après extraction de l'ADN (Jour 1), celui-ci sera fragmenté puis lié à des adaptateurs de séquençage (Jour 2). Les fragments d'ADN d'intérêt (gènes cible) vont être capturés grâce à des sondes d'ARN biotinylées complémentaires et fixation sur billes magnétiques (Jour 3 et 4). Les fragments d'intérêt vont être liés à un index spécifique de chaque patient (indexation) afin de pouvoir mélanger l'ADN de tous les échantillons (multiplexing) au sein du même run de séquençage (Jour 5). Les fragments d'ADN d'intérêt sont lus pas le séquenceur (jour 6) et les données brutes de séquençage transformées par le pipeline d'analyse bioinformatique.

L'enrichissement des régions d'intérêt peut notamment se faire soit par une approche de type « amplicon » ou par capture des régions d'intérêt.

L'approche par amplicon est constituée d'une PCR initiale multiplex par échantillon, les amorces encadrant les régions d'intérêt à séquencer. Dans la technique par capture (Figure 19), des sondes nucléotidiques d'une centaine de paires de bases vont être utilisées, complémentaires des régions d'intérêt. Ces sondes sont généralement dessinées de manière chevauchante, afin que la même région d'intérêt puisse être capturée par plusieurs sondes différentes (tiling). De plus, la taille des sondes autorise de nombreux « mismatchs », permettant la capture des séquences variantes. Ces sondes sont généralement liées à un résidu

de biotine permettant leur fixation sur système magnétique, avec pour conséquence la capture sélective des régions ADN d'intérêt. Quelle que soit la technique, les fragments d'ADN sont ensuite indexés (fixation d'un code-barre oligonucléotidique différent par échantillon) avant séquençage.

Les approches par amplicon nécessitent une petite quantité d'ADN, mais leur principale limite va résider dans l'homogénéité des données produites. En effet le nombre de *reads* (séquences ADN lues par le séquenceur) couvrant les régions d'intérêts peut varier dans des proportions plus importantes que dans les approches par capture, ceci étant lié aux rendements différentiels des PCR[164]. L'information se retrouve dupliquée par l'amplification par PCR, et le risque qu'une erreur de polymérase soit représentée à des fréquences alléliques telles qu'elle induit l'identification un variant faux positif n'est pas négligeable. A l'inverse, les approches par capture nécessiteront une plus grande quantité d'ADN, mais procurent une couverture beaucoup plus homogène. Les techniques d'enrichissement par capture séquenceront néanmoins plus de régions non désirées dans l'analyse, du fait des rendements d'enrichissement inférieurs à 100%. De plus les régions de faible complexité (peu de diversité dans la composition en nucléotides) disséminées dans le génome sont difficilement capturable par ces techniques et les régions homologues (similarité de séquence entre plusieurs régions), similaires aux régions génomiques ciblées sont sources d'ambiguïtés et de perte de sensibilité et spécificité. Le rapport entre le nombre de *reads* recouvrant des régions ciblées dans l'analyse et le nombre de reads total est ainsi qualifié de « on-target ».

6. Focus sur les séquenceurs par termination réversible (séquençage par synthèse)

Dans le cadre de cette thèse nous ne détaillerons que la méthode de séquençage par synthèse utilisant des terminateurs réversibles (Illumina), méthode aujourd'hui la plus utilisée notamment en génétique des populations, celle-ci ayant été utilisée exclusivement pour tous les travaux présentés ci-après.

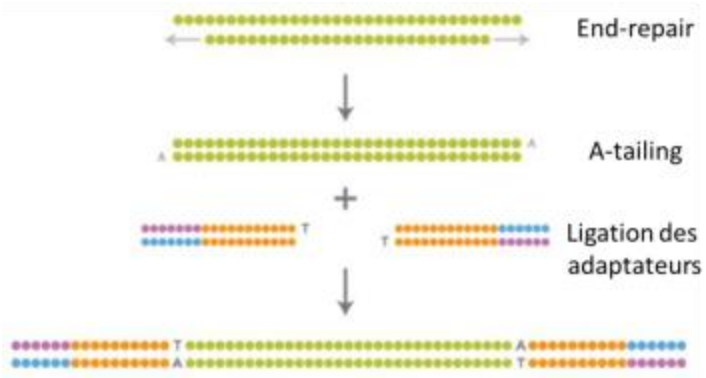


Figure 20: Préparation des ADNs avant séquençage.

L'ADN, après avoir été extrait, est tout d'abord fragmenté (fragmentation enzymatique ou par sonication) afin d'obtenir des séquences d'environ 200 pb. Ces fragments vont ensuite subir plusieurs étapes de préparation successives avant séquençage (Figure 20) :

- Réparation des extrémités (End-Repair) : les extrémités des fragments d'ADN vont être complétées afin que les fragments d'ADN soient totalement double-brin (à bouts francs), indispensable à la suite du processus.
- Adénylation (A-tailing) : Une adénine sera ajoutée à l'extrémité des fragments, afin de prévenir la ligation de fragments d'ADN entre eux.
- Ligation des adaptateurs : des adaptateurs spécifiques vont être ajoutés aux extrémités des fragments, notamment grâce à leur thymine supplémentaire en 3', complémentaire de l'adénine ajoutée à l'étape précédente. Les adaptateurs sont de courtes séquences ADN qui serviront à se lier de manière complémentaire à des amorces nécessaires au

processus de séquençage. Des index moléculaires pourront aussi être ajoutés (« code-barres » nucléotidique de 6 à 8 paires de bases) permettant de différencier le matériel génétique de chaque échantillon, et autorisant ainsi l'analyse de plusieurs échantillons différents dans la même manipulation de séquençage.

Ces fragments d'ADN modifiés (appelés librairies) pourront être sélectionnés en fonction des régions d'intérêt (cf. Introduction, § VI.B.5), puis transférés sur le séquenceur, dont le processus d'analyse va se dérouler en plusieurs étapes.

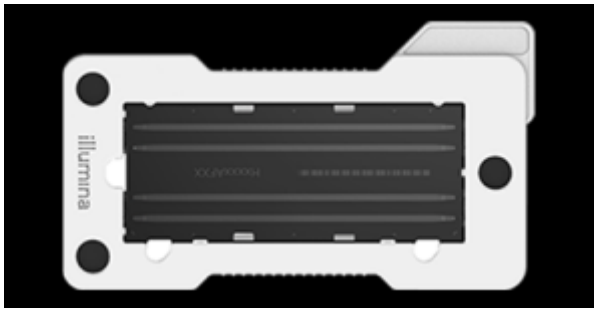


Figure 21: Flow-cell pour séquençeur NextSeq (Illumina)

Etape 1 : Les librairies vont être déposées sur une *flow cell* (Figure 21), une plaque de verre spécifique à la machine, à la surface de laquelle sont fixées de courtes séquences ADN simple-brin : les *primers*. Les fragments d'ADN préparés aux étapes précédentes et dénaturés vont pouvoir se répartir et se fixer aux *primers* de manière complémentaire grâce à leurs adaptateurs.

Etape 2 : Les fragments vont subir une amplification clonale par un procédé dit de « bridge PCR » (PCR en pont, Figure 22). L'extrémité encore libre du fragment d'ADN fixé sur la *Flow Cell* va se lier à une amorce de PCR fixée elle-aussi sur la surface de la *Flow Cell* et ainsi former un pont. Une fois ce pont formé, le brin complémentaire va pourvoir être synthétisé. A l'image d'une PCR classique, plusieurs cycles d'amplification sont réalisés, permettant d'obtenir à l'emplacement de chaque molécule initiale un « massif » de molécules

identiques fixées sur la plaque (un *cluster*). Cette méthode permet d'amplifier en parallèle un très grand nombre de molécules d'ADN différentes en produisant des *clusters* monoclonaux, permettant une lecture du signal par des caméras haute-résolution.

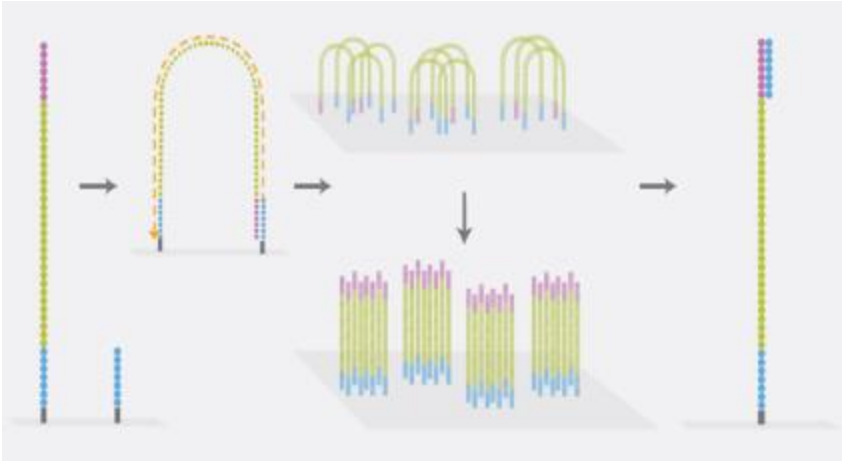


Figure 22: Amplification clonale en "Bridge PCR" (Illumina)

Etape 3 : Les fragments vont être « lus » par le séquenceur en exploitant une technique de séquençage par synthèse basée sur l'utilisation de terminateurs réversibles (Figure 23). Les nucléotides nécessaires au séquençage sont modifiés par la liaison d'un marqueur fluorescent (différent pour chacun des 4 nucléotides) et d'un terminateur réversible (Remarque : Le Séquenceur NextSeq possède une chimie fonctionnant avec seulement 2 fluorochromes différents : rouge et vert. Une adénine émettra ainsi un signal lumineux composé de 50% de rouge et 50% de vert, une cytosine de 100% de rouge, une thymine de 100% de vert, et une guanine n'aura aucune intensité). A chaque cycle de séquençage sont ajoutés les 4 nucléotides modifiés et une ADN polymérase. Ainsi, à chaque cycle de polymérisation, un seul nucléotide complémentaire sera ajouté à la séquence en cours de lecture. L'excitation par un laser permettra dans un second temps l'émission fluorescente de la couleur correspondante au nucléotide ajouté, permettant ainsi sa lecture. Le regroupement par *cluster* des fragments permet d'augmenter l'intensité du signal lumineux émis pour qu'il soit détectable par une caméra CCD qui enregistrera l'ensemble des signaux.

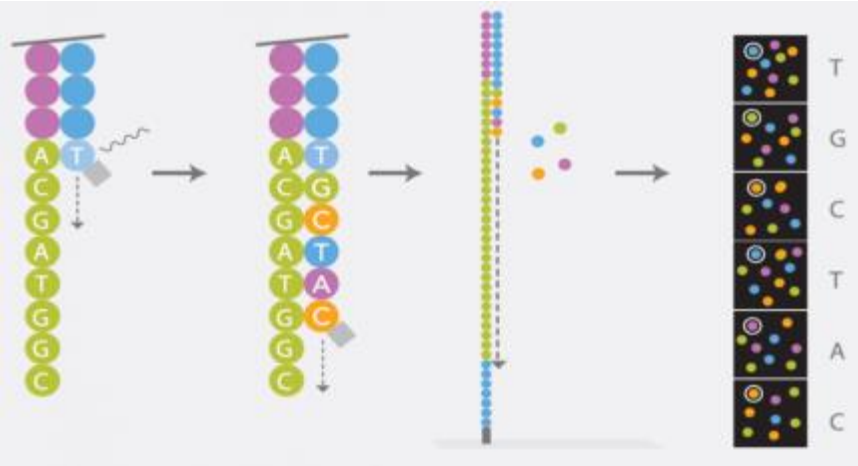


Figure 23:Séquençage par synthèse basé sur l'utilisation de terminateurs réversibles (Illumina)

L'interprétation du signal lumineux par la machine nécessite les étapes suivantes afin d'obtenir une lecture des nucléotides la plus efficiente possible :

- La normalisation de l'intensité du signal lumineux (Figure 24) : les fluorochromes de chaque nucléotide vont avoir des intensités lumineuses différentes. Une normalisation est donc effectuée sur les 12 premières paires de bases séquencées afin de ramener l'intensité des nucléotides à une valeur moyenne.

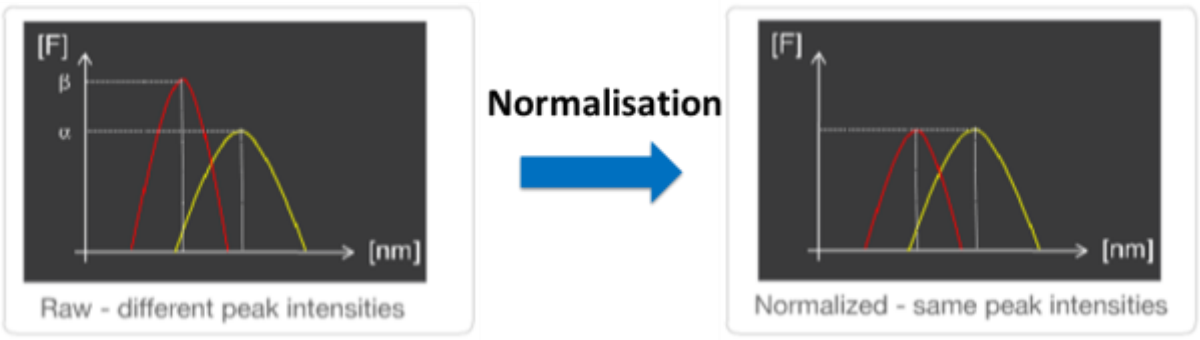


Figure 24: Normalisation des intensités lumineuses pendant le séquençage (https://support.illumina.com).

- La correction du bruit de fond lumineux (Figure 25) : L'échec d'incorporation d'une base pendant un cycle (*Phasing*) ou à l'inverse l'incorporation de plusieurs bases au

cours d'un cycle (*Prephasing*) va créer un signal lumineux impur au sein d'un *cluster*. Ce bruit de fond sera aussi pris en compte dans la lecture des nucléotides.

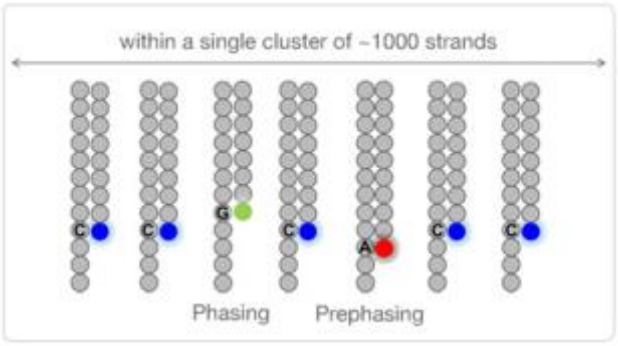


Figure 25: Bruit de fond généré dans un cluster (https://support.illumina.com).

- Identification des bases (*Base calling*) : La détection des bases séquencées est effectuée après normalisation de l'intensité du signal lumineux. La base qui possèdera la plus forte intensité sera appelée (Figure 26).

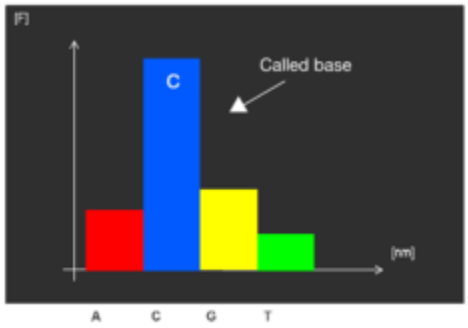
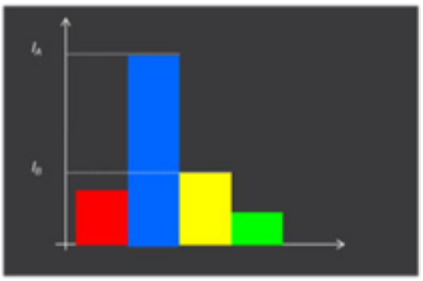


Figure 26: Base Calling (https://support.illumina.com).

- Filtration des *clusters* de mauvaise qualité (Figure 27): Cette évaluation de la qualité des *clusters* est effectuée sur la pureté du signal sur les 25 premières bases séquencées. Cette pureté est déterminée par le ratio de la plus grande intensité lumineuse sur la somme de la plus forte et de la seconde plus forte intensité lumineuse (CHASTITY). Pour passer ce critère qualité, un *cluster* doit avoir un score CHASTITY supérieur à 0,6 sur 24 des 25 premières bases séquencées.



$$C = \frac{I_A}{I_A + I_B}$$

Figure 27: Filtration des clusters de mauvaise qualité (https://support.illumina.com)

- Attribution d'un score qualité pour chaque base séquencée : à partir des profils d'intensité, des ratios signal / bruit, la comparaison des critères de fiabilité des bases séquencées à une table pré-calculée (Q-table) va permettre d'attribuer un score qualité à chaque base séquencée : le score PHRED. Ce score reflète la probabilité d'erreur d'identification d'une base sur le séquenceur.

$$\boldsymbol{score\ PHRED = Q = -10 \times \log(P)}$$

Ex : Score PHRED = 10 ➔ probabilité d'erreur de lecture de la base de 10%

Score PHRED = 20 ➔ probabilité d'erreur de lecture de la base de 1%

Score PHRED = 30 ➔ probabilité d'erreur de lecture de la base de 0.1%

Après acquisition de l'image, le marqueur fluorescent et le terminateur réversible sont excisés, terminant le cycle d'incorporation et préparant les fragments pour le prochain cycle. Plusieurs dizaines de cycles sont ainsi réalisés, produisant des séquences de 75 à 300 paires de bases successives (les *reads*).

Ces séquenceurs peuvent aussi réaliser du « paired-end sequencing». Une fois la première lecture du fragment d'ADN terminée (incorporation d'un nombre de bases déterminé correspondant au nombre de cycles sur la machine), le brin complémentaire va être synthétisé par une nouvelle étape de *bridge-amplification*. Une nouvelle lecture sera ensuite lancée sur

ce brin complémentaire. Ainsi les 2 extrémités du fragment d'ADN d'intérêt vont être séquencées (Figure 28).

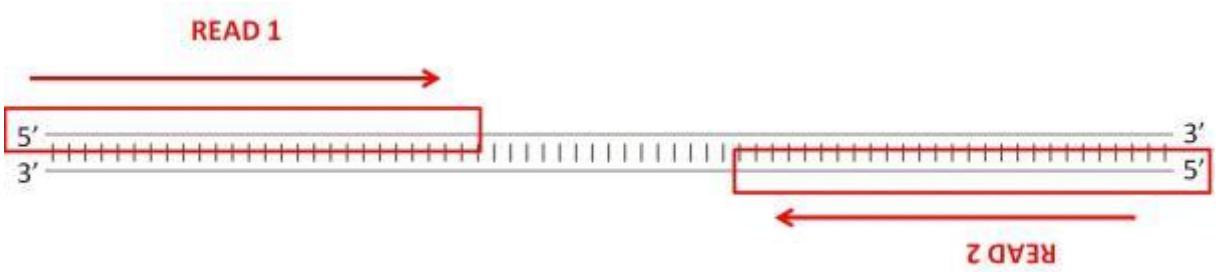


Figure 28: Illustration du Paired-End Sequencing. Read 1 = read Forward / Read 2 = Read Reverse.

Les séquenceurs dits « de paillasse » utilisant cette technologie peuvent générer plusieurs dizaines de millions de *reads* en 24h, correspondant à autant de fragments d'ADN. Les machines les plus performantes utilisant cette méthode génèrent aujourd'hui jusqu'à 20 milliards de *reads* sur 2 jours, permettant de séquencer le génome complet de plusieurs dizaines d'individus (NovaSeq 6000, Illumina). L'exploitation des séquenceurs de 2ème génération en médecine permet ainsi une exploration génétique des individus à grande échelle, en relation avec leur pathologie, et dans des délais compatibles avec leur prise en charge thérapeutique.

VII / Les défis du séquençage à haut-débit

Les séquenceurs de 2ème génération ont permis une augmentation considérable des débits d'analyse, ouvrant la voie à une exploration du génome jusqu'ici inaccessible, mais s'accompagnent de plusieurs défis à la fois technologiques, méthodologiques et scientifiques.

A. Le traitement des données

Dans divers domaines scientifiques (Astronomie, Biologie, Physique…), l'évolution des technologies et de la quantité de données qu'elles produisent a contraint les scientifiques à avoir recours à une assistance informatique afin de rendre interprétables les informations générées. Divers domaines biologiques ont nécessité un soutien informatique afin de résoudre des problèmes complexes. De cette nécessité est né un nouveau domaine scientifique, la bioinformatique, à la frontière de la biologie, de la physique et des mathématiques, avec des domaines d'application variés, tels que la biologie structurale (étude de la conformation tridimensionnelle des macromolécules biologiques), la biologie des systèmes (modélisation des réseaux et interactions constituant un système biologique), ou encore la génomique (étude de la structure et des fonctions du génome). Appliquée à l'analyse des génomes, la bioinformatique apporte un soutien précieux dans l'ordonnancement, la modélisation et l'interprétation de l'information génétique brute produite par les séquenceurs de 2ème génération. Aujourd'hui, la démocratisation du NGS dans les laboratoires de génétique moléculaire nécessite donc parallèlement un investissement important dans cette discipline.

La chaine de traitement (« pipeline ») bioinformatique des données de séquençage à haut-débit comporte généralement quatre étapes principales :

- Le *base-calling* : lecture des bases constituant les fragments d'ADN séquencés (étape réalisée directement sur le séquenceur, voir Introduction, § VI.B.6)
- Alignement : Détermination de la position des *reads* sur un génome de référence
- Re-calibration : étape optionnelle modifiant les données de qualité du *base-calling* visant à supprimer les biais, pouvant améliorer la sensibilité et la spécificité
- *Variant-calling* : détection des variants parmi les fragments séquencés

1. Les informations utilisées en NGS

Les séquenceurs de 2ème génération lisent l'ADN par fragments, et donnent cette information sous forme de séquences de quelques dizaines de paires de bases : les *reads*. Ces *reads* vont être regroupés dans des fichiers informatiques au formatage spécifique et normalisé : les fichiers FASTQ, qui contiendront aussi pour chaque base séquencée le score PHRED associé (Figure 29).

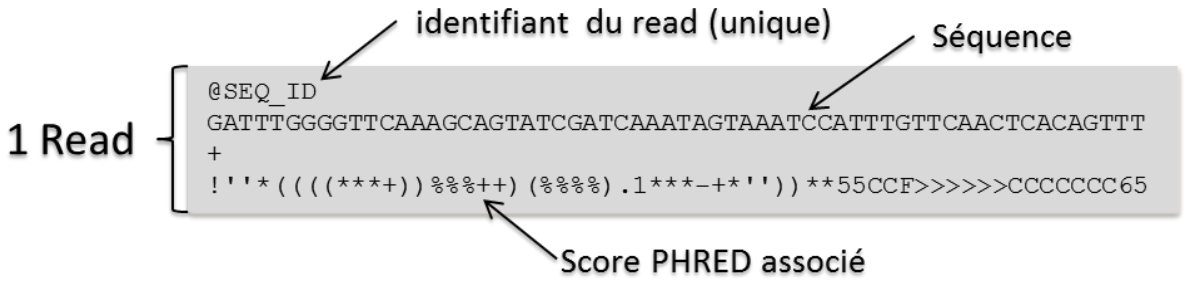


Figure 29: Contenu d'un fichier FASTQ

Ces valeurs de score sont codées dans le fichier FASTQ en format ASCII (American Standard Code For Information Interchange), chaque caractère correspondant à une valeur numérique (Ex : « ! » = 33).

2. Pré-traitement des données

Plusieurs modifications peuvent être apportées aux données brutes de séquençage avant traitement, en fonction de l'application.

Plusieurs échantillons différents peuvent être inclus dans un même *run* de séquençage, en étant préalablement indexés pendant la préparation des échantillons (cf. Introduction, § VI.B.6). Cet index va ainsi servir à réattribuer les *reads* à chaque échantillon, au cours d'une étape appelée le démultiplexage (Figure 30).

Figure 30: Démultiplexage

Il sera aussi possible de supprimer les données de mauvaise qualité en se basant sur les scores PHRED (*Sequence Quality Trimming*) ou encore de supprimer les séquences des adaptateurs pour ne conserver que les données des séquences d'ADN (*Adaptator Clipping*). On distingue généralement le *Soft-Clipping* et le *Hard-Clipping*. Dans le *Soft-Clipping*, les séquences des adaptateurs sont « masquées », c'est-à-dire conservées dans la séquence du *read*, mais non utilisées pour la suite du traitement bio-informatique. Le *Hard-Clipping* va supprimer définitivement les adaptateurs de la séquence du *read*. Le *soft-clipping* pourra ainsi avoir un intérêt lorsque les bases « clippées » sont en réalité sur un autre endroit du génome, signant un *read* chimérique (*read* s'alignant sur deux portions différentes du génome, indiquant un variant structural).

3. L’alignement des séquences : *mapping*

Une fois les données brutes de séquençage pré-traitées, la première étape va consister à assembler les *reads* les uns par rapport aux autres afin de reconstruire la séquence ADN originale. L’ADN original ayant été séquencé aléatoirement, les *reads* contiendront des séquences chevauchantes. Deux méthodes différentes peuvent être utilisées pour y parvenir :

- Assemblage *de novo* : les *reads* sont assemblés uniquement alignés les uns par rapport aux autres, sans apport de données externes. Cette méthode est généralement utilisée pour séquencer des génomes nouveaux, dont la séquence consensus est encore inconnue.
- Assemblage par alignement sur génome de référence : chaque *read* est aligné sur un génome de référence. Cette méthode est très utilisée en génétique humaine, dont la séquence consensus est établie. Cette séquence est régulièrement mise à jour, le génome humain de référence le plus utilisé étant actuellement la version hg19.

Le détail des algorithmes d’alignement sera traité ici en prenant comme exemple l’alignement sur un génome de référence, les activités de génétique humaine et les travaux présentés ici utilisant cette méthode d’assemblage.

Ainsi, l’alignement le long d’un génome de référence va consister à comparer la similarité entre chaque *read* et une référence, afin de déterminer à quelle partie du génome le *read* ressemble le plus (Figure 31, haut), pour attribuer à chaque *read* des coordonnées génomiques.

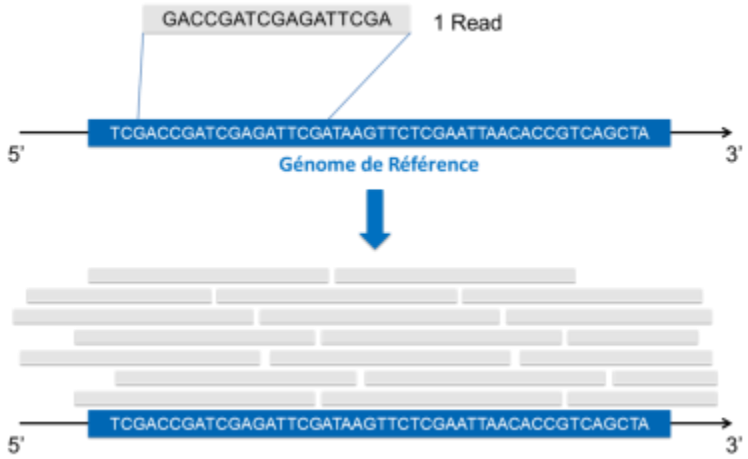


Figure 31: alignement des reads le long d'un génome de référence

Une coordonnée génomique sera définie par un numéro de chromosome, suivi de la position de la base dans ce chromosome (la première base d'un chromosome étant celle de l'extrémité 5' du bras court, la dernière celle de l'extrémité 3' du bras long). Cela donne par exemple pour le gène *BRCA1* chr17:41.196.312-41.277.500.

Chaque position génomique d'intérêt est lue plusieurs fois, puisque différents fragments d'ADN contenant cette position auront été séquencés. En conséquence, plusieurs *reads* vont se superposer à la même position génomique (Figure 31, bas). Les principales difficultés de cette étape vont provenir :

- Des variants présents sur les *reads*, modifiant la similarité avec la séquence de référence (en augmentant la différence entre les deux)
- Des régions de faible complexité : ces régions vont pouvoir avoir une composition biaisée en nucléotides (ex : prédominance de G et de C : région GC riche), des motifs répétés, des séquences palindromiques, ou encore une combinaison de ces différents facteurs. Ces régions seront de fait difficiles à aligner correctement car les *reads* seront positionnables à plusieurs endroits (régions dites de faible mappabilité).

Deux types d'approche sont principalement utilisés en génétique pour aligner les séquences de manière non supervisée, chacun avec ses utilisations préférentielles en fonction du problème posé : les méthodes basées sur la programmation dynamique, et celles basées sur les algorithmes heuristiques. Les méthodes de programmation dynamique vont garantir de trouver l'alignement optimal mais seront coûteuses en temps de calcul, alors que les algorithmes heuristiques seront plus efficients mais sans garantir de trouver le meilleur alignement. Ces méthodes restent néanmoins toutes basées sur un score d'alignement, l'alignement ayant le score le plus haut déterminant la meilleure position d'une séquence par rapport à l'autre.

a. Méthodes de programmation dynamique

Alignement global : Méthode de Needleman-Wunsch

Cet algorithme a été publié en 1970 par Saul Needleman et Christian Wunsch[165] et permet d'aligner n'importe quelle séquence, qu'elle soit nucléotidique ou protéique, par rapport à une autre. Le terme global signifie ici que l'algorithme va impliquer toute la longueur de la séquence à aligner. Cet outil est plus adapté pour aligner des séquences de longueur relativement similaires avec un grand degré de similarité au départ.

D'un point de vue mathématique et algorithmique, l'alignement va résider dans l'établissement d'une correspondance nucléotide-nucléotide entre deux séquences, de sorte que l'ordre des nucléotides dans chaque séquence soit préservé. De plus, un « gap », c'est-à-dire une correspondance entre un nucléotide et rien sur l'autre séquence, est autorisé sur n'importe quelle séquence. A chaque alignement sera associé un score établi sur un ensemble de règles, notamment des scores de substitution de nucléotides et des pénalités pour les *gaps*. Le score d'alignement va correspondre à la somme des scores de substitution et des pénalités

de *gap*. L'illustration de la démarche de l'algorithme est inspirée et adaptée à partir de la ressource Wikipedia (https://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm).

Si l'on prend comme exemple deux séquences :

1 : GCATGCT

2 : GATTACA

La première étape consiste à construire une matrice, dite « matrice de substitution », à partir des deux séquences (Figure 32).

		G	C	A	T	G	C	T
G								
A								
T								
T								
A								
C								
A								

Figure 32: Matrice de substitution

Cette matrice de substitution va ensuite devoir être remplie avec un système de scores, choisis au préalable. Si l'on prend un système de scores basique :

- *Match* : nucléotide identique entre les 2 séquences ➔ +1
- *Mismatch* : nucléotides différents entre les 2 séquences ➔ -1
- Indel : Insertion ou délétion. Un nucléotide s'aligne avec un gap sur l'autre séquence ➔ -1

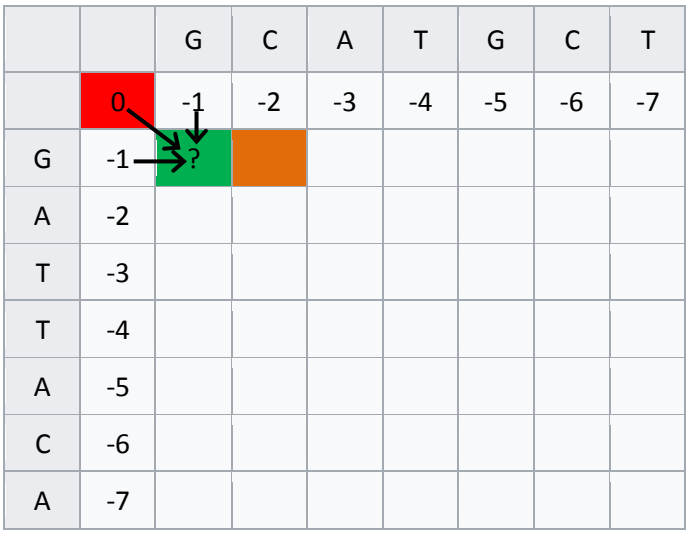


Figure 33: remplissage d'une matrice de substitution (Algorithme Needleman-Wunsch)

La position de départ pour remplir cette matrice est indiquée par la case rouge (Figure 33), qui possède un score à 0. Les scores pour les autres cases sont ensuite calculés par ligne, de gauche à droite pour chaque case (excepté pour la première ligne et la première colonne, remplies au préalable). Le remplissage de la matrice se fait de manière systématique en fonction des règles décrites ci-dessous.

Pour calculer le score d'une case, il faut se baser sur les scores des cases déjà remplies à gauche, au-dessus, ou à gauche au-dessus de la case d'intérêt. Si un score est calculé à partir d'une case de gauche (1ère ligne) ou de la case au-dessus cela représente un indel dans l'alignement (avec un score de -1 dans le système de scores décrit précédemment). C'est pour cette raison que la première ligne et la première colonne sont remplies de -1 à -7, puisque les scores de case ne peuvent être remplis qu'à partir de la case de gauche (1ère ligne) ou de la case au-dessus (1ère colonne). Le score à inscrire dans la case sera le meilleur score (le plus haut) que l'on puisse obtenir à partir des cases de gauche, du haut et à gauche au-dessus de la case d'intérêt.

Si l'on prend l'exemple de la case verte (Figure 33), 3 options sont possibles :

- A partir de la case du dessus : ce qui représente un indel = **-1** (indel) + **-1** (score de la case du dessus) = **-2**
- A partir de la case de gauche : ce qui représente un indel : **-1** (indel) + **-1** (score de la case du dessus) = **-2**
- A partir de la case en haut à gauche : même nucléotide G-G = +**1** (*match*) + **0** (score de la case en haut à gauche) = **+1**

Le meilleur score à choisir pour la case verte est donc +1.

La case suivante à remplir sera donc la case orange, pour laquelle on applique le même principe. Cette case aura donc un score de 0 (*mismatch* G-C + score de la case de gauche = **-1** + **+1**).

Ces règles vont ainsi permettre de remplir l'ensemble de la matrice de substitution (Figure 34).

		G	C	A	T	G	C	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	+1	0	-1	-2	-3	-4	-5
A	-2	0	0	+1	0	-1	-2	-3
T	-3	-1	-1	0	+2	+1	0	0
T	-4	-2	-2	-1	+1	+1	0	+1
A	-5	-3	-3	-1	0	0	0	0
C	-6	-4	-2	-2	-1	-1	+1	0
A	-7	-5	-3	-1	-2	-1	0	0

Figure 34: Matrice de substitution complète (Algorithme Needleman-Wunsch)

Le score obtenu en bas à droite (Figure 34, case verte) représente le score du meilleur alignement possible. Pour déterminer l'alignement il faut ensuite repartir de la dernière case (verte) et déterminer grâce à quelle case a été obtenu ce score. Ici, le score de la case verte a

été obtenu grâce à la case au-dessus à gauche (Figure 34, orange). Il faudra donc ensuite repartir de la case orange et déterminer comment ce score a été obtenu, et remonter jusqu'à la case rouge.

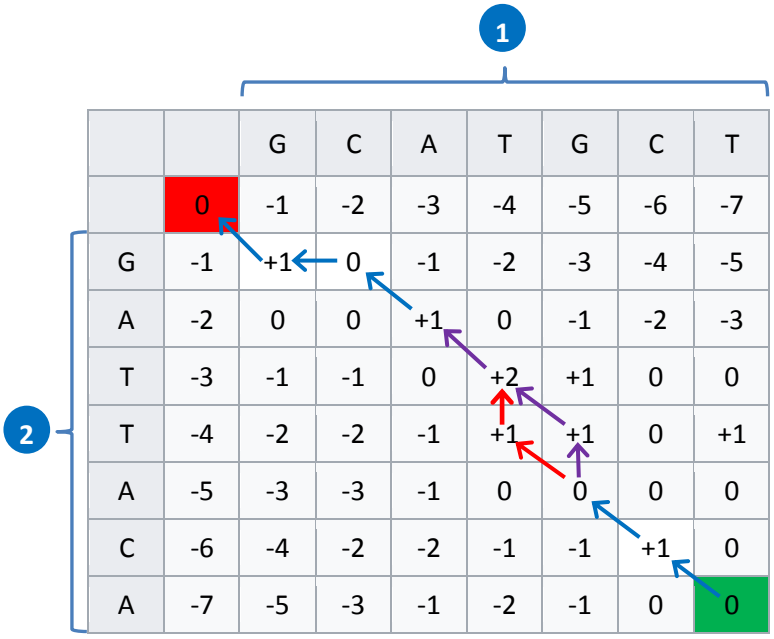


Figure 35: Alignement optimal à partir d'une matrice de substitution (Algorithme Needleman-Wunsch)

Comme le montre la Figure 35, plusieurs « chemins » sont possibles, et donc plusieurs alignements possibles. Une fois les meilleurs chemins possibles déterminés, la lecture se fait à partir de la case rouge. Un déplacement en diagonale représente un *match* ou un *mismatch*, un déplacement vers la droite l'ouverture d'un *gap* dans la séquence 2, un déplacement vers le bas l'ouverture d'un *gap* dans la séquence 1. Les 2 alignements possibles sont visibles sur la Figure 36.

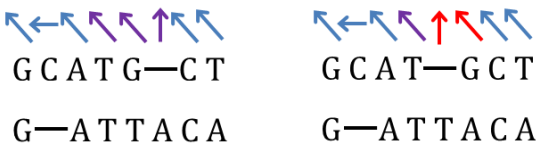


Figure 36: Alignements possibles obtenus par l'algorithme Needleman-Wunsch

Cet algorithme est généralement affiné par 2 paramètres différents, afin de se rapprocher d'une réalité biologique.

- La dissociation entre ouverture de *gap* et prolongation de *gap* : attribuer un score élevé à une ouverture de *gap* et un score plus faible à une prolongation de *gap* permettra de fusionner plusieurs indels en une seule :

- L'utilisation de matrices de similarité pour affiner l'attribution de scores pour les *mismatchs* : toutes les substitutions ne sont pas équivalentes, certaines pouvant par exemple être plus fréquentes que d'autres (ratio transition/transversion). On pourra donc attribuer un score en fonction de la substitution observée :

	A	**G**	**C**	**T**
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

Les matrices de similarité les plus élaborées sont aujourd'hui construites pour des alignements de séquences protéiques, telles que les matrices BLOSUM (BLOcks Substitution Matrix).

Même si cet algorithme garantit un alignement optimal, cette recherche de la meilleure solution possible se fait au détriment du temps de calcul. De plus, les séquences à comparer doivent être relativement apparentées et de taille similaire. Ces caractéristiques rendent l'algorithme inadapté à l'alignement de nombreuses et courtes séquences le long d'un génome de référence, tel qu'en séquençage à haut-débit.

Cet algorithme d'alignement est aujourd'hui utilisé principalement pour effectuer des alignements de bout en bout dans le cadre de la comparaison de séquences protéiques ou nucléotidiques pour mettre en évidence des relations structurales ou évolutionnaires. Un exemple d'outil utilisant cet algorithme est disponible sur le site de l'European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) : http://www.ebi.ac.uk/Tools/psa/.

Alignement local : Algorithme de Smith-Waterman

Cet algorithme a été publié par Temple F. Smith et Michael S. Waterman 1981[166]. Il est apparenté à l'algorithme de Needleman-Wunsch, utilisant aussi des matrices de similarité afin de déterminer le meilleur alignement. Mais à la différence du précédent, cet algorithme va aussi rechercher des alignements locaux, autorisant un alignement partiel entre deux séquences. Il permettra ainsi d'aligner des séquences de tailles très différentes.

Dans cette méthode, les scores négatifs sont ramenés à 0, mais les matrices de substitution et les pénalités de *gap* sont toujours utilisées. Ainsi, si l'on prend ces deux séquences à aligner :

- TGTTACGG
- GGTTGACTA

Dans cet exemple, les *matchs* ont un score de +3, les *mismatchs* un score de -3, et les *gaps* un score de -2.

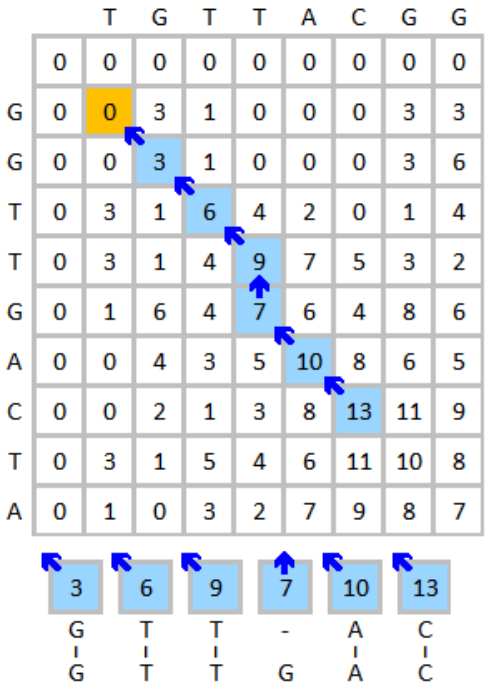


Figure 37: Illustration de l'algorithme de Smith-Waterman[167]

La matrice de similarité (Figure 37) se remplit de la même manière qu'avec l'algorithme précédent. Une fois complétée, il faut repartir du score le plus élevé dans la matrice, puis remonter le « chemin » par lequel le score a été obtenu. Le meilleur alignement obtenu ici n'a donc été fait que sur une partie des séquences à comparer, mais cette méthode garantit l'alignement optimal. Si l'on repart du coin inférieur droit de la matrice, l'algorithme permettra de définir l'alignement optimal complet entre les deux séquences, alignement qui peut être différent du meilleur alignement local.

Avec cette méthode, des séquences de taille différentes peuvent être alignées (toujours de manière optimale), avec en plus la possibilité de trouver le meilleur alignement localement. Néanmoins, l'utilisation de cet algorithme nécessite toujours de compléter l'ensemble de la matrice de similarité, avec des temps de calcul toujours trop importants pour la problématique du séquençage à haut-débit (alignement de multiples séquences sur un génome de référence).

Cet algorithme est principalement implémenté dans la suite logicielle FASTA[168], conçue pour l'alignement de séquences ADN et protéiques, présentée pour la première fois en 1985 et dont la dernière version (v36) est sortie en 2010.

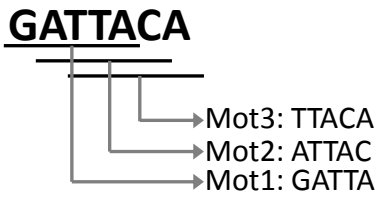
b. Méthodes heuristiques

Ces méthodes sont particulièrement utilisées pour traiter les données de séquençage à haut-débit du fait de leur rapidité d'exécution et de la quantité de séquences à aligner sur une référence.

BLAST : Basic Local Alignment Search Tool

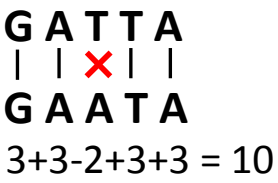
BLAST est un algorithme heuristique d'alignement dont la première version a été publiée en 1990[169]. Il se base sur l'utilisation d'une table de hashage, qui est une structuration spécifique des données permettant d'associer rapidement une clé (requête) à une valeur (réponse) via un index. Plusieurs étapes successives sont ainsi réalisées par BLAST pour réaliser un alignement relativement précis et beaucoup plus rapide comparé aux méthodes de programmation dynamique, approximant un alignement de Smith-Waterman.

Etape 1 : La séquence à aligner est découpée en mots d'une longueur « k » définie (généralement k = 11 pour les séquences nucléotidiques, dans l'exemple k=5).



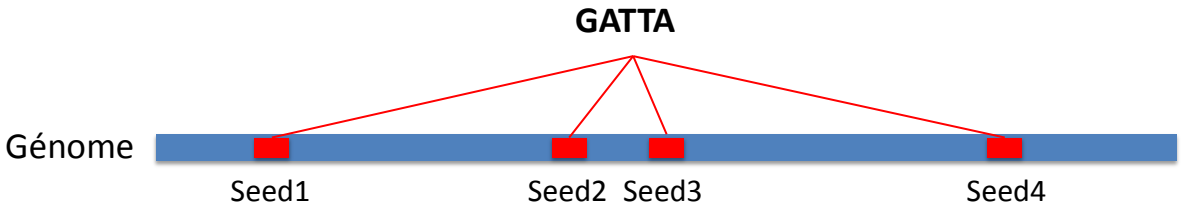
Etape 2 : Chaque mot généré à l'étape précédente sera comparé à tous les mots possibles de la longueur correspondante. Dans cet exemple, chaque mot sera ainsi comparé à 4^5 = 1024 mots

possibles de 5 lettres formés à partir des lettres A, C, G, T. Une matrice de substitution va être utilisée pour donner un score (s) à chaque comparaison (ex : *match* = +3 / *mismatch* = -2).

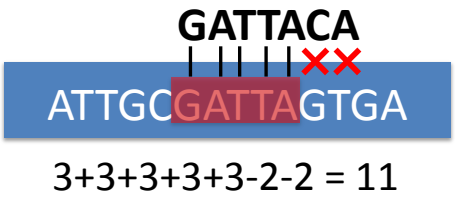


Ne seront gardées que les comparaisons avec les scores les plus hauts, supérieurs à un seuil défini au préalable.

Etape 3 : Les mots restants vont ensuite être comparés à une table de hashage (ex : le génome humain) afin de trouver les endroits où ces mots vont correspondre de manière exacte (sans aucun *mismatch*). Ces endroits vont constituer des points de départ (appelés « seeds ») pour tenter d'aligner le reste de la séquence d'intérêt.



Etape 4 : Les *matchs* exacts vont être étendus en alignant le reste de la séquence d'intérêt à partir du *seed*, pour former des *high-scoring Segment Pairs* (HSP). Un nouveau score (S) sera calculé sur chacun de ces HSP, à partir de matrices de substitution comme précédemment.



Ne seront gardés que les HSP dont le score est supérieur à un second seuil lui aussi fixé au préalable.

Etape 5 : BLAST va évaluer la vraisemblance de l'alignement obtenu en s'appuyant sur la loi de distribution des valeurs extrêmes de Gumbel[170]. Cette loi va permettre de calculer la probabilité qu'une séquence aléatoire ait un score S plus élevé que la séquence observée. Ainsi plus cette probabilité sera faible, meilleur sera l'alignement.

Cet algorithme est aujourd'hui l'un des programmes bio-informatiques les plus utilisés avec de nombreuses versions dérivées sorties depuis la première publication, toutes adaptées à un problème spécifique (alignement de séquences protéiques, nucléotidiques, ARN, alignement de domaines structuraux, alignements multiples, comparaisons inter-espèces…). Sa facilité d'accès (avec une utilisation possible en ligne : https://blast.ncbi.nlm.nih.gov/Blast.cgi) et son efficacité en font aujourd'hui un outil incontournable en génétique.

L'utilisation de tables de hashage pour résoudre le problème de l'alignement de millions de reads est aussi reprise par de nombreux autres logiciels dont ELAND[171], SOAP[172], MAQ[173] ou encore Novoalign (http://www.novocraft.com/products/novoalign/), chacun présentant des optimisations sur la création ou l'optimisation des *seeds*, notamment par l'autorisation de *mismatchs* lors de la création des *seeds*.

BWA : Burrows-Wheeler Aligner

Cet aligneur[174] beaucoup plus récent (2009) est basé en réalité sur la transformée de Burrows-Wheeler, algorithme mathématique publié en 1994[175], destiné au départ à la compression des données. Cette méthode avait au départ comme objectif de réorganiser les données afin que les caractères identiques initialement éloignés les uns des autres aient une plus grande probabilité de se retrouver adjacents. Son nouveau domaine d'application fut trouvé face à la problématique du séquençage à haut-débit et à l'alignement de millions de *reads* sur un génome de référence, avec une efficacité et une rapidité d'analyse considérablement

améliorée. L'application de cet algorithme à la problématique de l'alignement servira ici à créer un arbre des suffixes, sorte d'arbre décisionnel permettant d'orienter l'alignement au fur et à mesure de la lecture de la séquence à aligner.

Si l'on considère une chaine de caractères, ou un génome de référence dans le cadre du séquençage à haut-débit[176] :

X = GOOGOL$

Le « $ » sert de terminateur, signifiant qu'il n'apparaitra qu'en fin de séquence. De plus ce signe sera considéré comme précédant toutes les autres lettres dans l'ordre alphabétique ($ < G < L < O). Cette séquence va être placée en première ligne d'un tableau carré (nombre de colonnes = nombre de lignes = nombre de lettres constituant la séquence) (Figure 38, gauche). Chaque ligne suivante sera ensuite constituée de la même chaine de caractères pour laquelle sera effectué un décalage d'une lettre vers la gauche. La lettre qui sortira du tableau sera replacée en dernière position (Figure 38, centre).

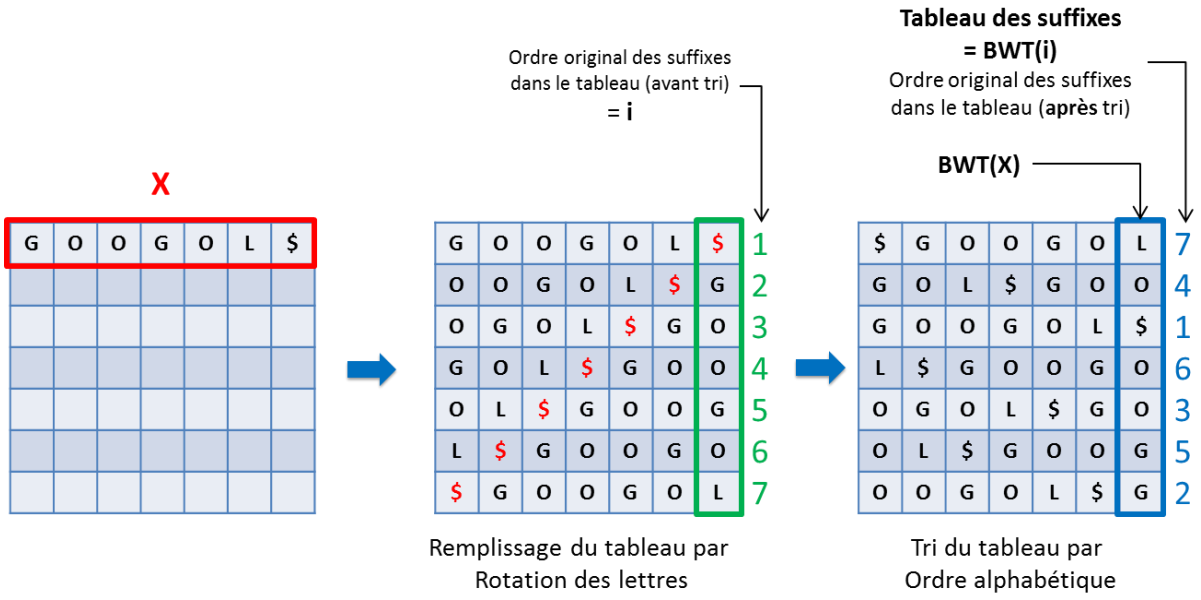


Figure 38: Illustration de la transformée de Burrows-Wheeler

Une fois le tableau entièrement rempli, les lignes seront triées par ordre alphabétique (Figure 38, droite). Ainsi, après ces manipulations de données, tous les éléments avec le même suffixe auront tendance à être proches les uns des autres (Figure 38, cadre bleu). La transformation du mot GOOGOL$ par cet algorithme donnera ainsi :

BWT(X) = LO$OOGG

A cette transformée pourra être associé le tableau des suffixes (ici BWT(i) = [7 ; 4 ; 1 ; 6 ; 3 ; 5 ; 2]) qui correspond à l'ordre original de chaque suffixe avant le tri par ordre alphabétique.

On utilisera ensuite une propriété de la transformée de Burrows-Wheeler appelée le *LF Mapping*. En reprenant la séquence X = GOOGOL$, nous allons pouvoir attribuer un indice à chaque lettre, correspondant au nombre de fois où la lettre a été vue précédemment dans la séquence (avec une numérotation commençant à 0). Pour X, cela donnerait :

$G_0O_0O_1G_1O_2L_0$

La première occurrence de G dans la séquence devient G_0 (1ère lettre dans la séquence donc pas de lettres précédentes), la troisième occurrence de O devient O_2 (2 « O » retrouvés dans le début de la séquence)… En revanche le signe « $ »n'est pas annoté. Cet exemple appliqué à l'ensemble du tableau décrit précédemment donnerait la Figure 39:

$	G_0	O_0	O_1	G_1	O_2	L_0
G_1	O_2	L_0	$	G_0	O_0	O_1
G_0	O_0	O_1	G_1	O_2	L_0	$
L_0	$	G_0	O_0	O_1	G_1	O_2
O_1	G_1	O_2	L_0	$	G_0	O_0
O_2	L_0	$	G_0	O_0	O_1	G_1
O_0	O_1	G_1	O_2	L_0	$	G_0

Figure 39: Illustration du LF Mapping

La propriété du *LF Mappi*ng énonce que la N^{ème} occurrence d'un caractère « x » dans la dernière colonne (Last column = L) se retrouve au même rang dans la première colonne (First Column = F). Ainsi, dans cet exemple, les caractères « O » (cadres rouges) dans la dernière colonne sont rangés dans cet ordre : O_1 ➔ O_2 ➔ O_0 ou encore 1➔2➔0. Le même classement sera retrouvé dans la première colonne. Il en est de même pour les caractères « G » (cadres verts).

Si l'on applique le LF Mapping à la séquence X modifiée par la transformée de Burrows-Wheeler :

$$\mathbf{BWT(X) = L_0O_0\$O_1O_2G_0G_1}$$

Les rangs qui lui correspondent sont [0 ; 0 ;1 ; 2 ; 0 ; 1]. On lui associera aussi la première et la dernière colonne de la Figure 38 (droite) (LF = *LF mapping* appliqué à la dernière colonne) :

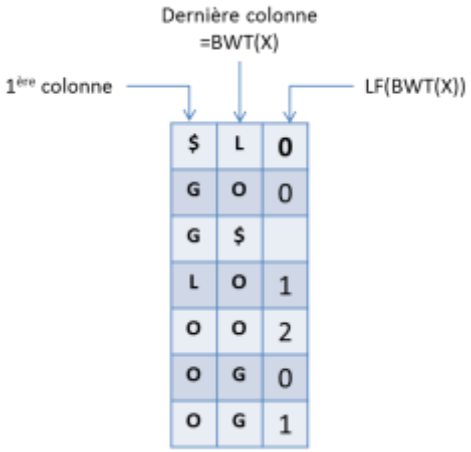


Figure 40: Table d'index construite à partir de la transformée de Burrows-Wheeler

La table d'index (Figure 40) étant construite, elle pourra servir à la recherche de séquences. C'est cette table qui sert en réalité de table des suffixes ici. Si l'on prend comme exemple la séquence W = « GO » (Figure 41).

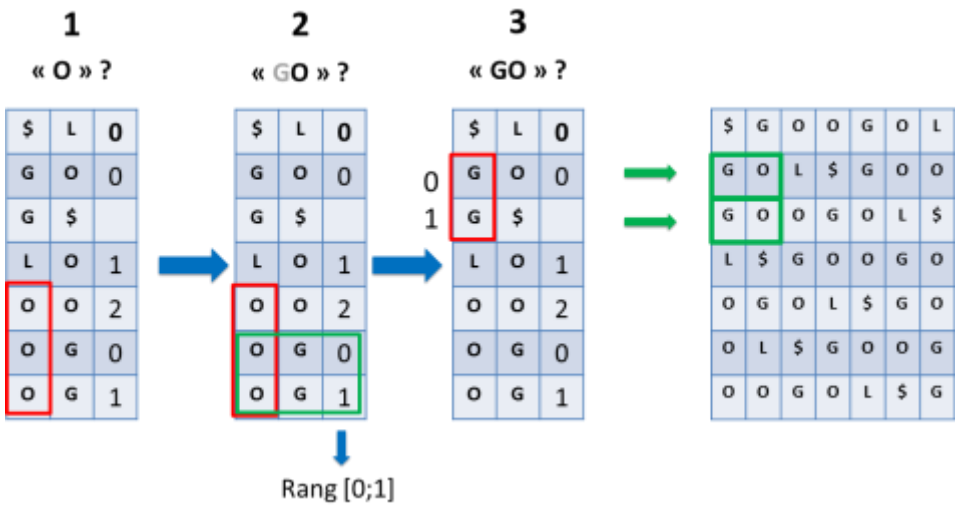


Figure 41: Exemple d'alignement en utilisant la méthode de Burrows-Wheeler

Etape 1 : Rechercher les lignes de la première colonne qui commencent par la dernière lettre de W, ici « O » (Cadre rouge, premier tableau)

Etape 2 : Dans W, la lettre qui précède « O » est « G ». Ainsi dans la deuxième colonne on recherchera les lignes qui correspondent aussi à « G » (Cadre vert, 2ème tableau). La troisième colonne donnera un intervalle [0 ;1] qui servira à l'étape suivante.

Etape 3 : On recommence l'étape 1, mais avec l'avant-dernière lettre de W : « G » (toute la séquence à aligner est ainsi explorée lettre par lettre en commençant par la dernière lettre de la séquence). On recherche ainsi les lignes de la première colonne qui correspondent à « G » (Cadre rouge, 3ème tableau). On utilisera aussi l'information extraite à l'étape précédente, à savoir l'intervalle. Il faudra donc se restreindre à n'utiliser que les lignes de 0 à 1 dans toutes celles contenant un « G ».

Dans cet exemple la recherche s'arrête là puisque la séquence à rechercher a été explorée en entier. Avec une séquence plus longue les étapes continueraient à s'enchainer. Si l'on reprend le tableau trié (Figure 41, 4ème tableau), l'algorithme a correctement identifié les 2 lignes commençant par la séquence recherchée. La phrase « GO » est donc présente deux fois dans

la séquence. Afin d'identifier la position de ces occurrences dans la phrase, il faudra se référer au tableau des suffixes ajoutés à l'index (Figure 42).

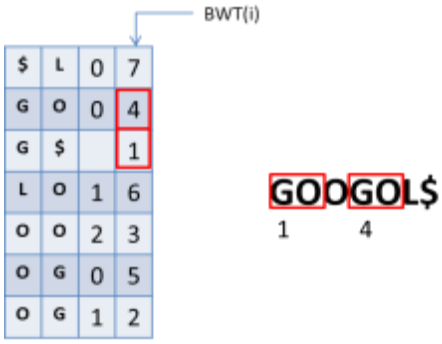


Figure 42: Identification de la position dans la séquence de référence

La table d'index identifie que la séquence « GO » est située en première et en quatrième position du mot GOOGOL. Ce principe et cette table d'index peuvent être reproduits à grande échelle sur un génome complet (fragmenté en blocs afin de rester efficace), répondant ainsi efficacement à la problématique de millions de *reads* produits par des séquenceurs à haut-débit. Ainsi, une fois la table d'index construite, l'alignement de séquences sur le génome de référence sera beaucoup plus efficace puisqu'au fur et à mesure de la lecture de la séquence à aligner, l'algorithme déterminera directement les régions correspondantes sur le génome de référence, sans avoir à fouiller sur l'ensemble de la séquence.

L'efficacité de la méthode implémentée BWA, reprise par d'autres outils tels que Bowtie[177], en fait aujourd'hui l'une des plus utilisée pour traiter les données de séquençage à haut-débit.

Ainsi, en fonction de la problématique posée, le choix de l'outil d'alignement va s'avérer crucial en terme d'efficacité (temps d'analyse) mais surtout de précision, afin d'éviter la génération d'erreurs d'alignements provoquées par des erreurs de séquençage ou des régions de faible complexité[178]. Les différents algorithmes utilisés pour aligner des données de

séquençage à haut-débit montrent d'ailleurs une précision différente, variant en fonction de la taille des reads alignés (Figure 43).

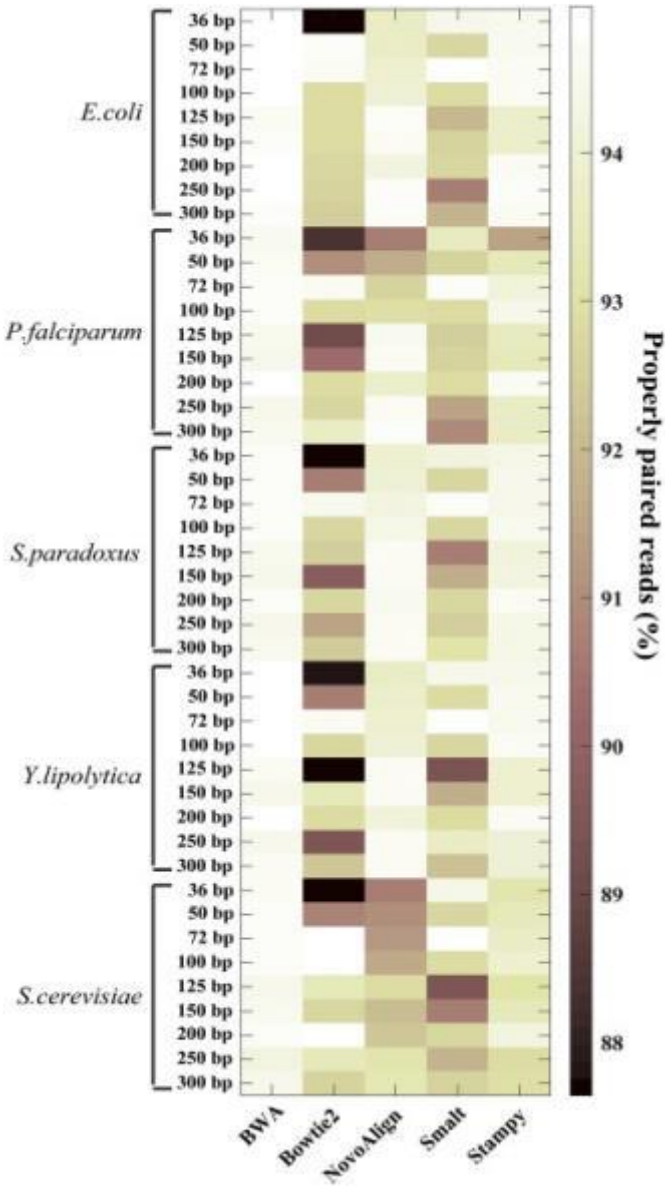


Figure 43: Comparaison de la précision de 5 logiciels d'alignement[179] La proportion de reads correctement alignés (représentée par des graduations de couleur) est visualisée pour les logiciels BWA, Bowtie2, Novoalign, Smalt et Stampy (en abscisse), testée sur les données provenant de 5 génomes différents en fonction de la longueur des reads alignés.

4. Données produites après alignement

L'alignement est une étape cruciale dans l'interprétation des données brutes de séquençage à haut-débit, une erreur dans le *mapping* d'un *read* pouvant générer des artéfacts considérés à tort comme des variants et conduisant à des erreurs d'interprétation.

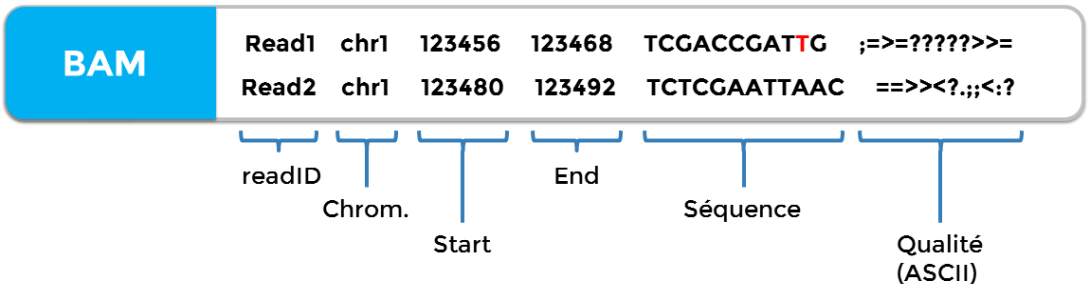


Figure 44: Constitution d'un fichier BAM

Ces différents outils d'alignement vont produire un autre type de fichiers : généralement un fichier texte SAM (*Sequence Alignment/Map Format*), contenant pour chaque read les coordonnées génomiques sur lesquelles il a été aligné. Ce fichier pourra ensuite être encodé au format BAM (*Binary Alignment/Map Format*) qui consiste en une représentation binaire compressée et sans perte du fichier SAM. Chaque ligne du fichier représentera ainsi un *read* ainsi que les informations qui lui sont associées (Figure 44). Les données stockées dans ce fichier ont ainsi une orientation que l'on pourrait qualifier « d'horizontale » (Figure 45, cadre bleu).

A chaque ligne du fichier sont retrouvées les informations suivantes (de gauche à droite) :

- readID : identifiant unique pour chacun des *reads* de l'échantillon
- le chromosome sur lequel le *read* a été aligné
- Start : la position génomique sur laquelle la dernière base du *read* a été alignée
- End : la position génomique sur laquelle la dernière base du *read* a été alignée
- Séquence : la séquence en nucléotides contenue dans le *read*

- Qualité : Score PHRED encodé au format ASCII

Figure 45: Orientation des données de séquençage en fonction du format utilisé. Cadre bleu : Fichier BAM ; Cadre vert : fichier pileup

Un deuxième format existe pour stocker les données de séquençage alignées : le format pileup. Ce format est créé à partir d'un fichier BAM, la conversion se faisant grâce à la suite logicielle SAMTOOLS par exemple[180]. Dans ce format, chaque ligne du fichier représente toutes les bases des reads alignées sur une position génomique (Figure 46), d'où une orientation que l'on pourrait qualifier de « verticale » (Figure 45, cadre vert).

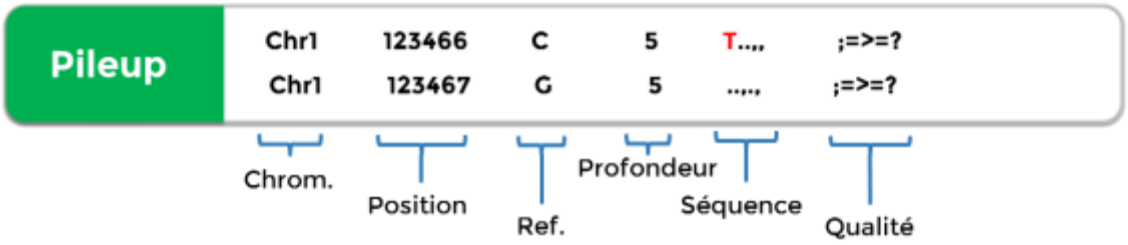


Figure 46: Constitution d'un fichier Pileup

A chaque ligne du fichier sont retrouvées les informations suivantes (de gauche à droite) :

- Le numéro de chromosome sur lequel le *read* a été aligné
- La position génomique analysée
- L'allèle de référence
- Le nombre de *reads* empilés sur cette position génomique (= profondeur de séquençage)
- Le contenu des *reads* empilés sur cette position génomique (« . » = base identique à l'allèle de référence, sur un *read* en orientation *Forward* ; « , » = base identique à

l'allèle de référence en position *Reverse* ; « A, C, G, T » = base différente de l'allèle de référence en orientation *Forward* ; « a, c, g, t » = base différente de l'allèle de référence en position *Reverse*)

5. Recalibration des données

Une fois les *reads* alignés le long d'un génome de référence, ceux-ci vont pouvoir être retravaillés, triés, marqués ou éliminés afin de rendre la détection des variants plus efficace et de supprimer un maximum de faux-positifs.

Plusieurs suites logicielles existent, chacune contenant une multitude d'outils dédiés à l'analyse de données de séquençage à haut-débit, parmi lesquelles Samtools, Picard (https://broadinstitute.github.io/picard/) ou encore GATK[181,182]. Ici sera pris comme exemple GATK, suite logicielle consensus utilisée en génétique des populations (excepté le marquage des *reads* dupliqués, réalisé par Picard), même si chaque suite logicielle présente des outils similaires. Les étapes présentées ici sont ainsi issues des « Best Practices » éditées par le Broad Institute (Créateur de la suite GATK), qui est un guide d'utilisation de référence en génétique.

a. Mark Duplicates : marquage des reads dupliqués

Durant le processus de séquençage, notamment pendant la préparation des échantillons avant passage sur le séquenceur, des étapes de PCR sont nécessaires. Ces étapes vont intentionnellement créer des copies de chaque fragment d'ADN à séquencer. Le séquenceur va donc lire des fragments d'ADN identiques, informations redondantes pour l'analyse : les duplicats de PCR.

Un autre type de duplicat peut être créé, cette fois-ci pendant le séquençage lui-même : le duplicat optique. Le séquenceur identifiera mal un seul cluster et l'interprétera comme plusieurs, dupliquant une information unique.

L'étape de marquage des *reads* dupliqués va ainsi identifier les *reads* qui apparaissent comme dupliqués et soit les supprimer, soit leur appliquer une marque spécifique, signifiant pour certains algorithmes qu'ils ne doivent pas être pris en compte pour le reste de l'analyse. Pour cela il comparera les séquences des positions en 5' des reads par échantillon. Tous les *reads* avec des séquences identiques en 5' seront identifiés comme un groupe de duplicats. Pour différencier le read primaire et les reads dupliqués, l'outil choisira ensuite celui qui cumulera le meilleur score-qualité par base (Score PHRED).

b. Réalignement local

Les logiciels d'alignement vus précédemment travaillent sur un génome complet pour effectuer leurs alignements. Ces alignements peuvent ensuite être affinés dans certaines zones comportant un nombre important de mismatchs sur les *reads*, généralement provoqués par la présence d'une insertion ou d'une délétion dans le génome de l'échantillon. Ce travail local va permettre de supprimer un certain nombre de faux-positifs dus à des erreurs d'alignement, mais surtout d'améliorer la détection des insertions et délétions (Figure 47).

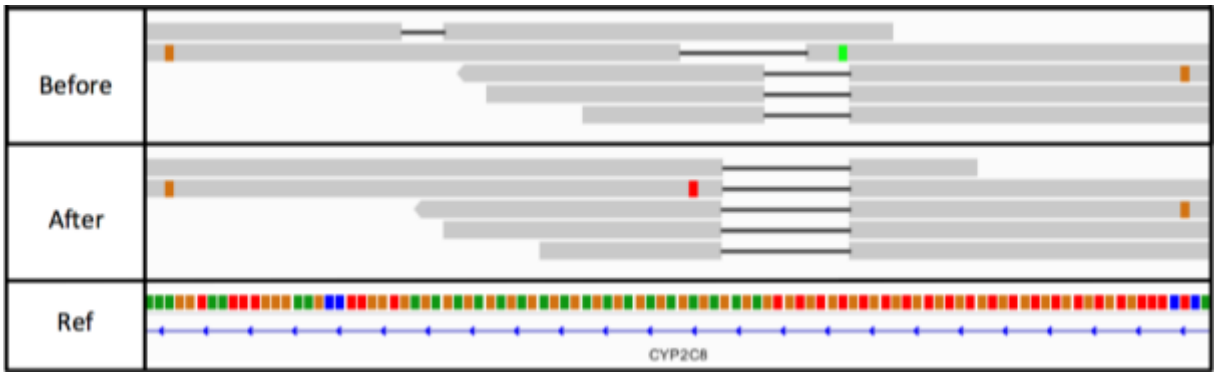


Figure 47: Réalignement local autour des Indels (http://gatkforums.broadinstitute.org)

c. Recalibration des scores-qualité des bases

Le score-qualité (PHRED) associé à chaque base lors de la lecture par le séquenceur peut être influencé par différentes sources d'erreur systématiques (biais provoqués par des machines, la chimie utilisée dans le séquençage ou la préparation des échantillons…), conduisant à sous-estimer ou surestimer ce score. Le BQSR (Base Quality Score Recalibration) est un processus basé sur l'apprentissage automatique (*machine learning*) qui va compenser ces erreurs systématiques en tentant de les modéliser.

Le BQSR est composé de deux étapes principales : Une première étape d'analyse et de construction d'un modèle de co-variation basé sur les données de séquençage à modifier, et un jeu de variants connus (généralement des polymorphismes). Les variants connus sont utilisés pour masquer les bases au niveau des sites attendus de variation. Dans une deuxième étape, ce modèle de co-variation va ainsi permettre de recalibrer les données de séquençage afin de tenter d'éliminer le bruit de fond et mettre en évidence les variants réels (Figure 48).

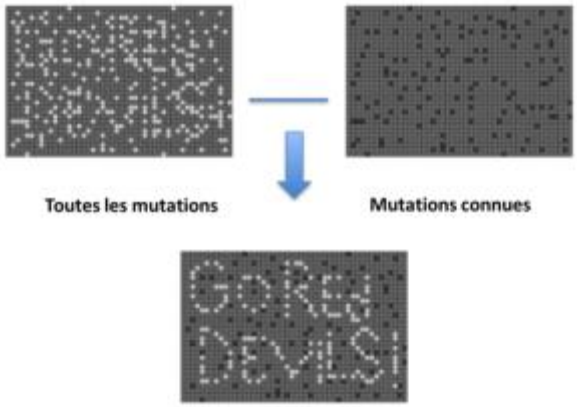


Figure 48: Illustration du principe du BSQR (GATK Talk, Pretoria, 2015)

6. Détection des variants : Variant-Callers

Une fois les données de séquençage (c'est-à-dire les *reads*) alignées, marquées et recalibrées, elles pourront être utilisées afin de mettre en évidence les variants de la séquence nucléotidique par rapport à une séquence de référence. De manière conceptuelle, les variants peuvent être représentés à la manière de la Figure 49 dans le cadre du séquençage à haut-débit. Un variant homozygote, présent sur les deux allèles de toutes les cellules de l'individu séquencé, sera présent sur tous les *reads* de la région génomique où est présent le variant. La fréquence allélique attendue du variant sera alors proche de 100 %. Dans le cas d'un variant hétérozygote, le variant sera présent sur environ 50 % des *reads* (dans toutes les cellules d'un individu, mais sur un allèle sur deux). Concernant les variants somatiques, ceux-ci ne seront présents que dans une fraction des cellules de l'individu et donc du matériel génétique séquencé. La fréquence allélique sera alors variable.

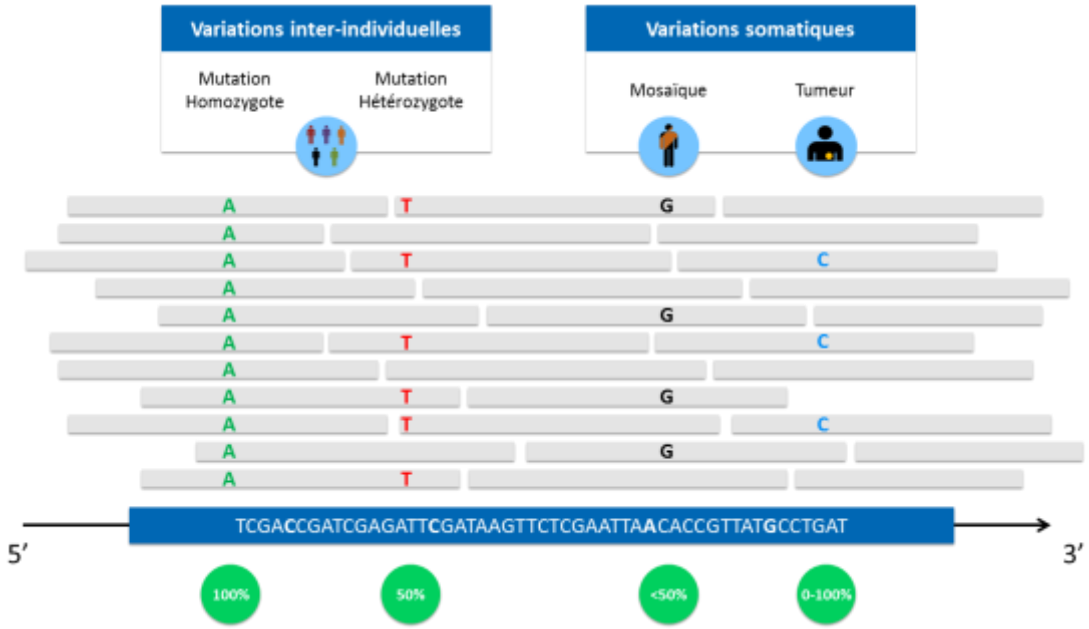


Figure 49: représentation des différents types de variants sur les données de séquençage à haut-débit

Chaque logiciel de détection des variants utilise le format BAM ou le format Pileup. Des dizaines de logiciels sont disponibles, avec des algorithmes associés à des règles de décision propres, en fonction du type d'évènement à rechercher et des utilisations. Le site Omictools (https://omictools.com, plateforme classant et référençant les outils bioinformatiquess disponibles afin de les rendre facilement accessibles à la communauté scientifique) a ainsi répertoriés 150 outils de *variant-calling* différents, classés par catégories (détection des variants constitutionnels, somatiques, indels, CNV…). Seront présentés ici quatre logiciels majeurs utilisés en génétique constitutionnelle (HaplotypeCaller) et en génétique somatique (Varscan, MuTect, Lofreq), chacun avec une méthode de détection différente.

Les variants détectés par ces différents outils seront toutes retranscrites sous un format standardisé : le fichier VCF[183] (Variant Call Format). Ce fichier est composé de deux parties principales :

- l'en-tête ou « header » : contient la description des informations fournies avec chaque variant détecté par le logiciel

- le corps du VCF : chaque ligne décrira un variant détecté par le logiciel dans l'échantillon. Seront retrouvées principalement le chromosome, la position génomique, l'allèle de référence, l'allèle alternatif, le score qualité associé au variant, et des informations supplémentaires associées (profondeur de séquençage à la position identifiée, fréquence allélique…)

Le gVCF (pour genomic VCF, produit par HaplotypeCaller) constitue une alternative à ce format, et contient les informations pour toutes les positions génomiques de l'échantillon étudié, que celles-ci soient mutées ou non. De plus la confiance avec laquelle chaque position génomique a été étudiée est estimée. Ce format sera ainsi privilégié pour les analyses en génétique des populations puisqu'il facilite une meilleure prise en compte des données manquantes ou issues d'un appel de variant de faible qualité et permet la recalibration des variants entre les patients et les différentes populations d'étude.

a. HaplotypeCaller : Haplotypage

HaplotypeCaller est un variant-caller intégré dans la suite logicielle GATK, dédié à la détection des variants constitutionnels et utilisé comme référence pour les études de population. Il prend en charge les fichiers de format BAM. Sa méthode de détection est constituée de quatre opérations principales, illustrées dans la Figure 50.

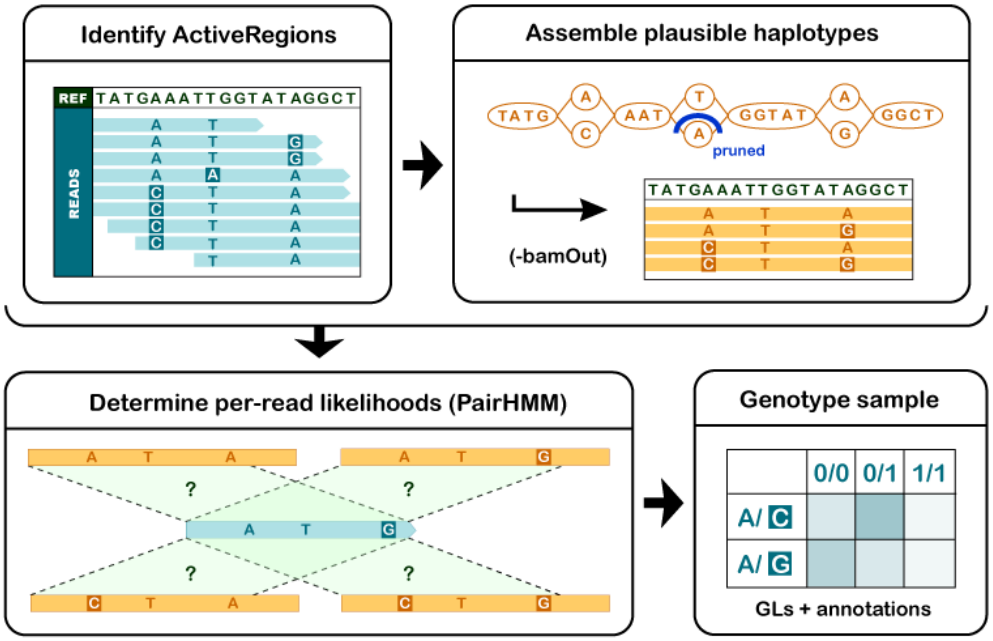


Figure 50: Etapes de détection des variants dans HaplotypeCaller (https://software.broadinstitute.org/gatk/)

Etape 1 : détection des régions actives (Figure 50, « Identify activeRegions »)

HaplotypeCaller va commencer par identifier des régions dites « actives », pour lesquelles une différence significative avec la référence est mise en évidence, par une mesure de l'entropie (degré d'incertitude) des données.

La première étape va consister à calculer un profil d'activité brute, par position génomique. Ce profil reflète la probabilité qu'une position génomique soit mutée, en fonction des reads alignés sur cette position.

Les profils d'activité vont ensuite être lissés grâce à un algorithme qui va effectuer une moyenne de ces profils.

Ces profils moyennés vont servir à définir les régions actives, à l'intérieur desquelles seront effectuées les recherches de variants. Toutes les régions pour lesquelles le profil dépassera un seuil prédéfini seront identifiées comme actives.

Etape 2 : Détermination des haplotypes possibles (Figure 50 : « Assemble plausible Haplotypes »)

Cette étape va permettre de reconstruire les séquences possibles des différents segments réels d'ADN présents dans l'échantillon original. Pour cela le programme va construire un graphe de Bruijn (graphique orienté qui va permettre de représenter toutes les séquences possibles, chaque possibilité étant pondérée par le nombre de reads qui la porte), en prenant la séquence de référence comme modèle. Chaque variation par rapport à la séquence de référence sera représentée par un nœud, une possibilité différente. Le programme pourra ensuite estimer les haplotypes les plus vraisemblables.

Etape 3 : évaluation de la pertinence de chaque haplotype et de chaque variant (Figure 50 : « Determine per-read likelihood (PairHMM) »)

Après que le programme ait déterminé tous les haplotypes possibles, il déterminera quels sont les plus pertinents en alignant chaque *read* sur les différents haplotypes déterminés (dont le génome de référence) en utilisant l'algorithme PairHMM[184]. Cet algorithme utilise le modèle des chaines cachées de Markov (HMM : Hidden Markov Model). En génétique moléculaire, les modèles de Markov permettent d'estimer quelle est la probabilité qu'une séquence d'intérêt puisse descendre d'une séquence ancestrale particulière. L'application de ce modèle dans ce contexte va permettre d'estimer la vraisemblance du *read* testé en regard d'un haplotype donné, tout en prenant en compte les informations de qualité de séquençage (Scores PHRED). Cela permettra de construire une matrice de vraisemblance des haplotypes en fonction des *reads*.

Cette matrice va ensuite être utilisée pour estimer la vraisemblance d'allèles potentiels sur un site candidat. Cette étape est appelée la marginalisation des allèles, et produira une nouvelle

matrice regroupant la vraisemblance de chaque allèle par read, au niveau de chaque site candidat.

Etape 4 : Assignation des génotypes par échantillon (Figure 50 : « Genotype Sample »)

La matrice générée à l'étape précédente sera utilisée ici pour estimer le génotype le plus probable en fonction des haplotypes, en utilisant le théorème de Bayes pour calculer la vraisemblance de chaque génotype possible. Ce théorème permet de reconstituer une probabilité totale (ici la probabilité que le génotype évalué existe) en fonction de ses probabilités conditionnelles (les variations possibles dans ce génotype), sachant quels variants existent réellement dans ce génotype. C'est ce génotype le plus probable qui sera retourné comme résultat pour chaque position génomique variant du génome de référence.

Analyse de populations :

L'avantage majeur d'HaplotypeCaller va résider dans sa capacité à prendre en charge une population entière d'individus, plutôt que de les traiter de manière unitaire (Figure 51). Il pourra ainsi exploiter l'information des populations d'étude entières et de manière comparative, par une analyse détaillée de tous les sites variants ou non-variants sur l'ensemble du jeu de données. L'objectif de cette analyse est de supprimer des faux-positifs causés par des erreurs d'alignement ou des artefacts de séquençage, mais aussi de récupérer certains variants ayant pu dans un premier temps être rejetés à cause d'un défaut de couverture ou de qualité de séquençage dans certains individus (notion de « missingness » ou information manquante liée aux régions génomiques non ou mal séquencées pour certains individus de la population étudiée). De cette manière les analyses en cas témoins évitent de conclure à tort sur l'éventuel enrichissement d'un variant dans une population qui serait non exploré dans l'autre population.

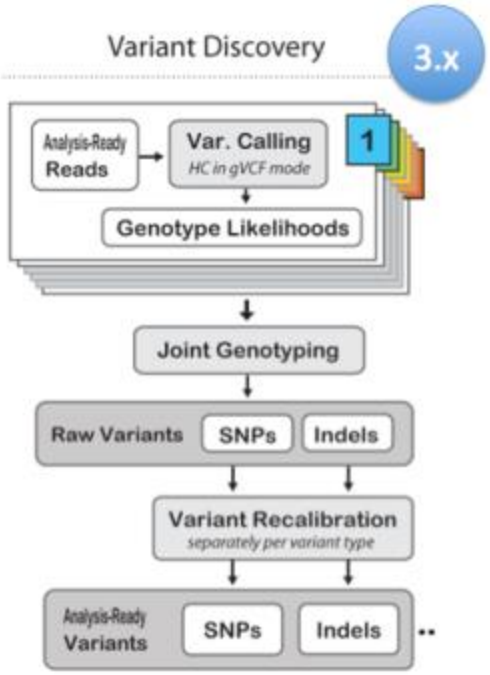


Figure 51: analyse de population par HaplotypeCaller

La première étape de cette analyse particulière va consister, pour chaque échantillon de la population d'étude, à générer un fichier gVCF.

Dans un second temps, toutes les informations de tous les gVCFs de la population seront fusionnées, par une analyse appelée « Joint Genotyping ». Cet outil va ainsi recalculer la vraisemblance des génotypes dans la population, « re-génotyper » les nouvelles données fusionnées et les ré-annoter, les informations de toutes les populations d'étude étant rassemblées dans un seul fichier.

Enfin, une étape de recalibration des variants va être effectuée : le VQSR (Variant Quality Score recalibration). Un nouveau score va être calculé pour chaque variant du jeu de données, appelé le VQSLOD (Variant-Quality Score Log-Odds). L'objectif de ce score va être de pouvoir filtrer simplement les variants en définissant un seuil de décision qui résumera le meilleur compromis sensibilité / spécificité. Ce score prendra en compte, entre autres informations, le contexte nucléotidique (la séquence) autour du variant, la profondeur de séquençage, le nombre de *reads* recouvrant chaque allèle, le nombre de *reads* en position

Forward ou en position *Reverse*. La méthode BQSR utilise des algorithmes d'apprentissage automatique (« machine learning ») afin de reconnaître les variants vrais des faux positifs, et ainsi intégrer dans un seul score toutes les informations citées ci-dessus. L'apprentissage est réalisé en s'appuyant sur des ressources de variants déjà connus (1000Genomes, HapMap) et hautement qualifiés afin de sélectionner un sous-ensemble de variants présents dans le jeu de données à analyser (et ayant une haute probabilité d'être des vrais positifs), sous-ensemble qui servira de jeu d'apprentissage.

Sa mise à disposition gratuite dans un cadre académique en fait aujourd'hui un logiciel incontournable dans le domaine de la génomique.

b. MuTect

MuTect[185] est un autre logiciel développé par le Broad Institute dédié à l'analyse des variants somatiques, plus particulièrement dans le domaine de la cancérologie. En effet HaplotypeCaller montre de très bonnes performances analytiques sur les variants constitutionnels, mais l'hétérogénéité du tissu tumoral ainsi que la qualité des échantillons à analyser (tissus inclus en paraffine) font baisser considérablement sa sensibilité.

MuTect a donc été développé dans ce contexte, afin de pouvoir détecter des variants faiblement représentés (sous-populations clonales diluées parmi les cellules saines). Ce logiciel va aussi pouvoir distinguer les variants strictement somatiques (issus de la tumeur) des variants constitutionnels de l'individu en analysant en regard des données issues d'ADN tumoral et d'ADN constitutionnel provenant du même individu.

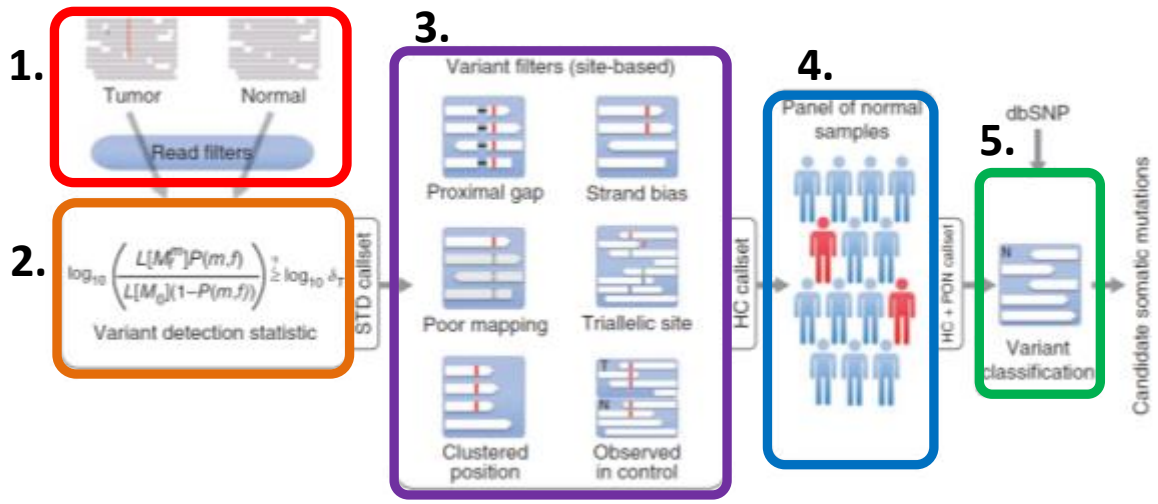


Figure 52: Principe de fonctionnement de MuTect

Pour que ce logiciel fonctionne, il faut nécessairement lui fournir les données de séquençage issues d'ADN tumoral et de tissu sain provenant du même individu, qu'il traitera en parallèle (à partir de fichiers BAM).

La première étape (Figure 52-1) consiste à filtrer les données de séquençage en fonction de critères-qualité (notamment le score PHRED et le score d'alignement).

Les variants candidats vont ensuite être mis en évidence en utilisant une classification bayésienne (basée sur le théorème de Bayes) (Figure 52-2). Le théorème de Bayes permet, en plus de définir une probabilité totale à partir de probabilités conditionnelles, de trouver les paramètres de la loi ayant généré cette probabilité. Un classificateur bayésien permettra ainsi de reconnaitre un objet (ici un variant) en fonction de ses paramètres.

Sur chaque site potentiellement muté, le logiciel va créer deux modèles :

$$M_0 = Pas\ de\ variant\ au\ site\ considéré\ (Base\ non\ référencée = bruit)$$

$$M_f^m = variant\ m\ existant\ réellement\ au\ site\ considéré, avec\ une\ fraction\ allélique\ f$$

Le modèle M_0 est ainsi équivalent au modèle M_f avec f = 0.

La vraisemblance de ces modèles va tout d'abord être évaluée :

$$L\left(M_f^m\right) = P(\{b_i\} \mid \{e_i\}, r, m, f)$$

$$= \prod_{i=1}^{d} P(b_i \mid e_i, r, m, f)$$

$d = Nombre\ de\ reads\ alignés\ sur\ le\ site\ considéré$

$b_i = Base\ du\ read\ qui\ couvre\ le\ site\ considéré$

$e_i = Probabilité\ d'erreur\ de\ lecture\ de\ cette\ base\ (associée\ au\ score\ PHRED)$

$r = allèle\ de\ référence\ au\ site\ considéré$

$m = variant\ au\ site\ considéré$

$f = fraction\ allélique\ du\ variant\ observée\ dans\ l'échantillon\ tumoral$

La vraisemblance des modèles peut donc être traduite comme le produit des probabilités, pour chaque base du site considéré, d'obtenir la base b_i en sachant e_i, r, m et f.

Ces probabilités sont calculées de la manière suivante :

$$P(b_i \mid e_i, r, m, f) = \begin{cases} f^{e_i}/3 + (1-f)(1-e_i) & \text{if } b_i = r \\ f(1-e_i) + (1-f)^{e_i}/3 & \text{if } b_i = m \\ {}^{e_i}/3 & \text{otherwise} \end{cases}$$

Une fois la vraisemblance des deux modèles établie pour le site considéré, elles seront utilisées dans un classificateur bayésien, ici nommé le LOD Score (pour Log ODds Score) (Figure 52-2). Ce LOD Score est basé sur un ratio entre les deux modèles :

$$LOD_T(m,f) = \log_{10}\left(\frac{L(M_f^m)P(m,f)}{L(M_0)(1-P(m,f))}\right) \geq \log_{10}\delta_T$$

Ainsi si le LOD Score dépasse un seuil de décision fixe ($\log_{10}\delta_T$) pré-établi et déterminé dans la publication initiale de MuTect (compromis entre sensibilité et spécificité), le variant m au site considéré est déclaré comme vraisemblable, et pourra passer aux étapes suivantes.

L'étape suivante est facultative (Figure 52-3) et permettra de sélectionner les variants dits de grande confiance (HC = High Confidence). Elle consiste en un ensemble de six filtres :

- **Proximal Gap** : détection d'une insertion ou d'une délétion à proximité du variant considéré, ce gap pouvant être à l'origine d'un variant artéfactuel.
- **Strand Bias** : Dans le cas d'un séquençage Paired-end, la variant doit théoriquement être distribué de manière équilibrée entre les *reads Forward* et les *reads Reverse*. Un déséquilibre pourrait signer un faux-positif.
- **Triallelic site** : quand le logiciel évalue un site, il considère tous les sites alternatifs possibles comme des variants candidats. Si plusieurs variants candidats passent les filtres, le site sera rejeté puisque le site considèrera qu'il est très peu probable qu'un site tri-allélique apparaisse dans un échantillon tumoral.
- **Clustered Position** : Certaines erreurs d'alignement vont provoquer l'apparition d'allèles alternatifs à une distance constante du début ou de la fin d'un *read*. Ces erreurs seront détectées par la mesure de la distance du variant par rapport aux extrémités du *read*. Les faux-positifs générés par ces erreurs d'alignement seront ainsi supprimées.
- **Observed in Control** : Les variants aussi retrouvés dans le tissu sain seront supprimées.

L'étape 4 (Figure 52-4) est elle aussi facultative. C'est aussi une étape de filtration des variants afin de réduire le nombre de faux-positifs et d'identifier d'éventuels polymorphismes. Pour cela MuTect va dans un premier temps être lancé sur les données de tout un groupe de tissus sains, afin de ne retenir que les variants vus dans plus d'un tissu sain. Ces sites seront ainsi identifiés comme à rejeter.

L'étape 5 (Figure 52-5) va dans un premier temps classer les évènements en fonction de leur origine : constitutionnel (présent dans le tissu sain associé), somatique (non présent dans le tissu sain), variant (présent dans la tumeur mais statut indéterminé dans le tissu sain). Le classificateur bayésien (LOD Score) présenté à l'étape 2 sera réutilisé, mais en comparant la vraisemblance des données observées à un modèle ou le variant est présent à l'état hétérozygote ou absent dans le tissu sain. A cette étape sera aussi utilisée une base de données de variants constitutionnels connus comme prior de probabilité pour l'évaluation du caractère constitutionnel d'un variant.

Ce logiciel fait partie d'une génération de logiciels récents, comme SomaticSniper[186] ou Strelka[187], basés sur des statistiques bayésiennes et destinés à l'identification de variants somatiques dans les tumeurs. Ils fonctionnent pour leur majorité en mode pairé, c'est-à-dire avec la nécessité d'avoir de manière concomitante des données de séquençage issues d'un tissu sain apparié. Cela leur permet d'obtenir de bonnes performances analytiques avec un bon compromis sensibilité / spécificité, mais avec la contrainte de séquencer en plus un tissu sain apparié.

c. Varscan : Méthode Heuristique

Varscan est une méthode publiée pour la première fois en 2009[188] et destinée à la détection des variants constitutionnels. Une deuxième version a été publiée en 2012[189], élargissant sa

détection aux variants somatiques avec la possibilité de prendre en charge à la fois un échantillon tumoral ainsi que le tissu sain apparié. Le principe de cette méthode est relativement simple, basé sur une série de règles élémentaires et empiriques, et permettant une détection rapide et relativement efficace des variants (méthode heuristique), à partir d'un fichier pileup.

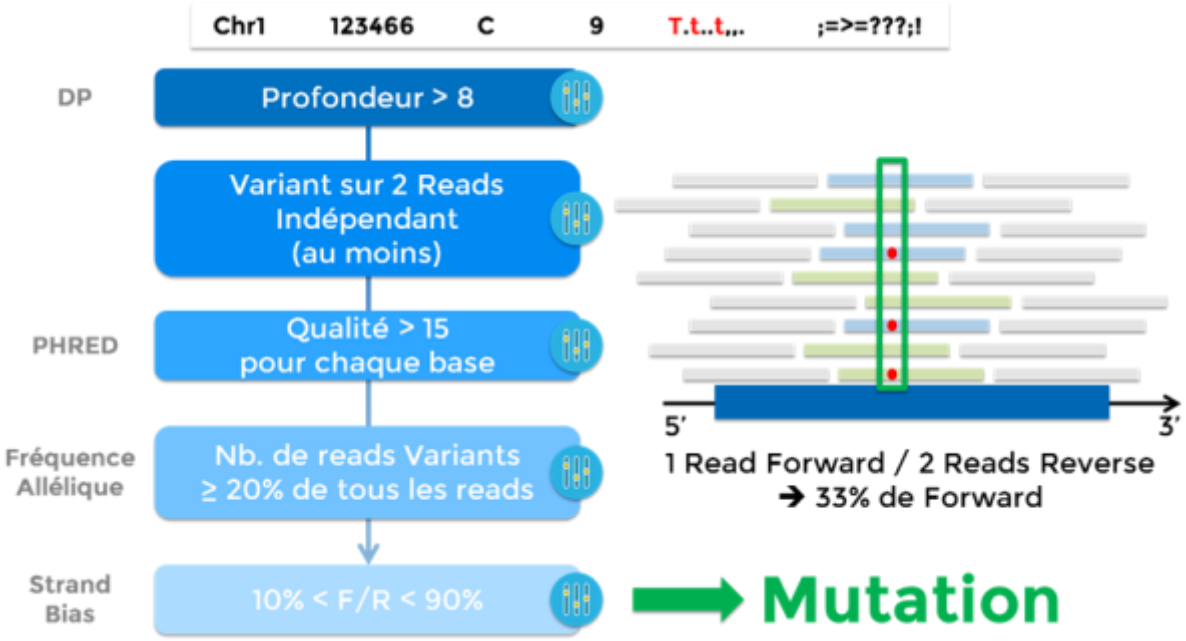


Figure 53: Principe de détection des variants par Varscan

Le logiciel analysera pour chaque ligne d'un fichier pileup les *reads* superposés sur la position génomique considérée (Figure 53, en haut), un variant devant remplir toutes les critères répertoriés dans la Figure 53 afin d'être retenue. Chaque critère peut être modifié dans les paramètres du logiciel.

Dans le mode pairé, Varscan utilisera les données de tumeur et de tissu sain apparié afin de déterminer la nature du variant. En utilisant un test exact de Fischer, il compare le nombre de *reads* portant l'allèle sauvage et le nombre de *reads* portant le variant dans la tumeur et dans le tissu sain. Si la p-value du test de Fischer dépasse un seuil préétabli (par défaut 0,10), le

variant est alors qualifié de somatique (si le tissu sain n'est pas porteur de variant) ou en perte d'hétérozygotie (si le tissu sain est hétérozygote).

Les principaux avantages de ce logiciel résident dans sa rapidité d'exécution, une compréhension facile de son processus analytique et dans la possibilité d'effectuer des recherches de variants somatiques même en l'absence de tissu sain apparié. Malgré tout, ce logiciel nécessite que l'utilisateur effectue au préalable la conversion des données de séquençage au format pileup. De plus, la spécificité du logiciel a tendance à chuter pour la détection d'évènements très faiblement représentés (fréquence allélique < 5 %).

d. Lofreq : Loi binomiale

Lofreq est un logiciel publié en 2012[190], directement conçu pour détecter les variants nucléotidiques dans une population cellulaire hétérogène, et est donc adapté à l'analyse tumorale. Il utilise pour ses calculs les données de séquençage au format pileup, mais effectue la conversion lui-même à partir d'un fichier BAM.

Ainsi, pour chaque position génomique à séquencer, Lofreq prend en compte toutes les bases des *reads* alignés sur cette position, et les considère comme résultant d'un test de Bernouilli (succès = base de référence / échec = base variante). Chaque base est considérée de manière indépendante et associée à une probabilité d'erreur dérivée du PHRED Score. Le nombre de variants (K) sur le nombre total (N) de bases alignées sur la position génomique considérée est donné par une distribution binomiale, dans laquelle chaque test de Bernouilli a une probabilité de succès différente. Ainsi, pour chaque base empilée sur la position génomique considérée, la formule suivante est employée de manière itérative :

$$Pr_n(X = k) = Pr_{n-1}(X = k)(1 - P_n) + Pr_{n-1}(X = k - 1)P_n$$

$\boldsymbol{Pr_n(X = k)}$ = Probabilité d'observer *k* variants dans les *n* premières bases

$\boldsymbol{P_n}$= probabilité d'erreur pour la $n^{ième}$ base

La *p-value* associée au variant est alors calculée comme la somme de toutes les probabilités précédentes, telle que montrée sur la Figure 54.

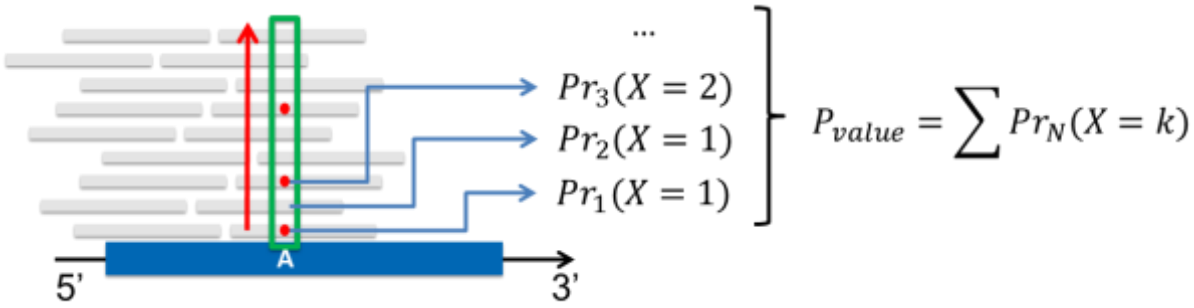


Figure 54: Evaluation de la vraisemblance d'un variant par Lofreq

Par défaut, la *p-value* doit être inférieure à 1 % pour qu'un variant soit reconnu comme vraisemblable. Les résultats peuvent aussi être filtrés en précisant une profondeur minimum de séquençage pour rechercher un variant, ou encore le score PHRED minimum (par défaut : 20) pour prendre un variant en compte dans l'analyse. Le déséquilibre en *reads Forward* / *Reverse* peut aussi être évalué.

Lofreq a aussi la possibilité réaliser une analyse pairée (échantillon tumoral / échantillon sain). Les variants non vus dans le tissu sain sont testés à nouveau (par un test binomial modifié) afin de déterminer si l'absence de détection est due à une profondeur de séquençage insuffisante. Ces tests serviront à déterminer la nature somatique ou constitutionnelle d'un variant.

Ce logiciel présente de bonnes performances en termes de sensibilité et de spécificité mais nécessite une profondeur de couverture importante ainsi que des données de séquençage de bonne qualité.

B. Interprétation des données

Malgré la complexité des processus nécessaires au traitement des données de séquençage à haut-débit, la détection des variants génomiques est aujourd'hui de mieux en mieux maîtrisée, particulièrement en ce qui concerne les évènements de petite taille. Le plus grand défi associé au séquençage à haut-débit et à l'étude du génome à grande échelle réside maintenant dans l'interprétation des variants identifiés aux étapes précédentes, afin de déterminer l'impact fonctionnel de ces évènements et leur implication dans la pathologie.

1. Prédiction de l'effet fonctionnel des variants

Hormis les variants induisant une perte de fonction, la conséquence fonctionnelle de la grande majorité des variants identifiés par séquençage à haut-débit demeure difficilement prédictible, notamment pour les variants faux-sens et les variants impactant l'épissage des ARN pré-messagers. La réalisation de tests fonctionnels pour tous les variants identifiés apparaît à l'heure actuelle impossible à la vue de la quantité d'informations produites, notamment dans un cadre diagnostique. Différents outils ont ainsi été créés afin d'apporter une aide à l'interprétation, par une évaluation *in silico* de l'effet fonctionnel que pourraient avoir les variants identifiés en séquençage à haut-débit.

a. Align GVGD

Align GVGD[191] est un outil de prédiction développé initialement pour estimer le caractère délétère des variants de signification inconnue identifiés sur les gènes de *BRCA1* et *BRCA2*. Align GVGD utilise 2 paramètres principaux afin de déterminer le grade de pathogénicité :

- Le score de Grantham : Ce score va refléter la distance évolutionnaire entre deux acides aminés, un score élevé reflétant une substitution délétère pour la fonction protéique. Ce score est calculé en fonction de la composition atomique, de la polarité et du volume moléculaire des acides aminés.
- Les alignements multiples : Align AGVGD va réaliser des alignements des séquences protéiques de 9 à 15 espèces différentes afin d'évaluer la conservation des acides aminés à travers les espèces, un variant survenant sur une position très conservée ayant une plus grande probabilité d'être délétère.

Les scores AGVGD permettent de définir des probabilités premières (« prior » de probabilités) de causalité (Figure 55), définis à partir de variants de *BRCA1* et *BRCA2* dont l'effet fonctionnel était connu au préalable. Ces *prior* de probabilités sont la base du modèle multifactoriel utilisé aujourd'hui pour la classification des variants et qui complète cette estimation de la causalité par des données de coségrégation, de structure familiale (similitude d'une famille à une famille « type » dans laquelle ségrége un variant pathogène de *BRCA1* ou *BRCA2*. Cet algorithme a été récemment retravaillé pour inclure les variants d'épissage dans la prédiction de pathogénicité[192].

Align-GVGD grade	Prior Probability	95% CI
C65	0.81	(0.61-0.95)
C35-C55	0.66	(0.34-0.93)
C15-C25	0.29	(0.09-0.56)
C0	0.03	(0.00-0.06)
Outside functional domains	0.02	(0.00-0.04)
Splicing consensus site alteration	0.96	(0.91-1.00)
Intronic variants outside the consensus dinucleotides	0.26	(0.15-0.39)

Figure 55: Priors de probabilité utilisés par AGVGD[193]

Aujourd'hui des alignements calibrés pour le logiciel existent pour 15 protéines différentes, dont BRCA1, BRCA2 et PALB2.

b. SIFT

Le programme SIFT[194] (« Sorting Intolerant From Tolerant ») se base sur les homologies de séquence au sein d'une famille de protéines afin de déterminer le caractère pathogène d'un variant. Ainsi, si dans l'alignement d'une famille protéique une position ne contient que l'acide aminé Isoleucine, le logiciel émet l'hypothèse que cet acide aminé demeure essentiel pour la fonction protéique. La substitution par n'importe quel autre acide aminé serait ainsi prédite comme pathogène. Si une position ne contient que des acides aminés hydrophobes (isoleucine, leucine et valine), SIFT présume que cette position ne supporte que des acides aminés hydrophobes. Un changement par un acide aminé avec des propriétés physico-chimiques différentes (chargé, polaire…) sera ainsi considéré comme pathogène.

SIFT, pour chaque substitution d'acide aminé, va : (i) considérer les séquences de toutes les protéines de la même famille (ii) sélectionner les séquences avec la plus grande ressemblance et partageant une fonction similaire (par l'intermédiaire de PSI-BLAST) (iii) effectuer un alignement multiple de ces différentes séquences (iv) calculer la probabilité que le variant observé apparaisse à cette position[195]. Selon SIFT, une probabilité inférieure à 5 % signera le caractère délétère du variant (« Deleterious »). Autrement, le variant sera qualifié de bénin (« Benign »).

c. PolyPhen-2

PolyPhen-2[196] va lui aussi se baser sur les homologies de séquences pour effectuer des prédictions, mais aussi sur des critères structuraux au niveau protéique. Il réalisera une recherche de séquences homologues puis un alignement multiple de ces séquences, à partir desquels il extraira 8 paramètres : le PSIC[197] (Position Specific Independant Count) Score de l'allèle sauvage, la différence entre le PSIC score de l'allèle sauvage et de l'allèle mutant, le changement de volume entre l'acide aminé sauvage et variant, la position du variant dans un domaine protéique, le nombre de résidus différents observés à la position étudiée, la congruence de l'allèle mutant dans l'alignement multiple, l'identité de séquence du plus proche homologue déviant de l'allèle sauvage, et si le variant est survenu dans un contexte CpG. Trois paramètres structuraux seront aussi extraits : la surface accessible de l'acide aminé, le facteur beta cristallographique, et les modifications de la surface accessible pour les résidus enfouis. L'ensemble de ces paramètres sont ensuite pris en compte par un classificateur bayésien afin de déterminer la probabilité que le variant soit délétère pour la fonction protéique.

Le classificateur nécessite un jeu de données « d'entrainement » afin de déterminer les critères de prédiction d'un variant délétère. Deux jeux de données différents ont été utilisés afin d'entrainer PolyPhen :

- HumDiv : comprend tous les allèles pathogéniques avec un effet fonctionnel connu, à l'origine d'une maladie mendellienne, présents dans la base UniPortKB, ainsi que toutes les différences observées entre les séquences protéiques humaines et leurs plus proches homologues, supposées non-délétères.

- HumVar : comprend tous les variants identifiés comme pathogènes dans UniProtKB, ainsi que tous les SNPs non synonymes (MAF > 1%) sans implication dans une pathologie.

PolyPhen-HumDiv est ainsi destiné à l'identification d'allèles rares sur des loci potentiellement impliqués dans des maladies complexes, là où même les allèles avec un effet faiblement délétère doivent être traités comme pathogènes.

PolyPhen-HumVar est lui destiné à l'étude de maladies mendéliennes, requérant de distinguer les variants avec un effet délétère fort de tous les autres variants, y compris ceux possédant un effet moyennement délétère.

d. MutationTaster

MutationTaster[198] va effectuer plusieurs analyses en parallèle afin d'évaluer les variants qui lui sont soumis:

- SNPs : les variants identifiés dans la base de données dbSNP (base de données référençant des polymorphismes) sont recherchés dans le projet HapMap (projet dont l'objectif est d'établir une cartographie des haplotypes du génome humain). Si un des 3 génotypes possibles est observé dans au moins une population HapMap (sans notion de fréquence), le variant est automatiquement classé en tant que polymorphisme.
- Analyse de conservation évolutionnaire: la séquence protéique et nucléotidique du gène impacté par le variant est alignée avec les séquences de dix autres espèces afin de déterminer le statut de conservation de l'acide aminé ou du nucléotide, en partant du principe que plus un élément est conservé, plus celui-ci doit avoir une fonction

importante dans la protéine (un variant étant dans ce cas-là considérée comme plus délétère).

- Analyse d'épissage : MutationTaster utilise un logiciel de prédiction de l'effet des variants sur l'épissage, NNSplice, qui évalue l'impact du variant sur les sites d'épissage.
- Analyse du signal de polyadénylation : MutationTaster analyse les variants résidant dans la zone des signaux de polyadénylation, en utilisant l'outil *polyadq* pour les repérer.
- Analyse des séquences consensus Kozak : MutationTaster évalue l'impact des variants touchant la séquence consensus Kozak, motif essentiel à l'initiation de la traduction.
- Caractéristiques protéiques : à l'aide de la base de données SwissProt (qui contient des informations et annotations fonctionnelles des protéines triées par des évaluateurs), MutationTaster évalue si le changement d'acide aminé provoqué par le variant affecte la fonction protéique (en fonction du domaine impacté).
- La longueur de la protéine : MutationTaster évalue si la protéine produite est étendue, tronquée ou si le système NMD est susceptible de se déclencher (en déterminant si un codon STOP prématuré apparaît en 5'de la dernière limite intron-exon ± 50pb). Si le logiciel estime que le système NMD va se déclencher, il classe automatiquement le variant comme pathogène.

Tous ces paramètres sont ensuite pris en compte par un classificateur bayésien (entraîné au préalable sur un jeu de données dont le caractère pathogène des variants est connu), afin de déterminer l'impact fonctionnel des variants évalués.

e. MaxEntScan

MaxEntScan[199] est un logiciel de prédiction entièrement dédié à l'analyse de l'impact des variants sur l'épissage. Il se base sur la notion d'entropie maximale, qui correspondra ici à la probabilité qu'une séquence consensus soit utilisée (reconnue par la machinerie d'épissage ou spliceosome), en fonction de contraintes définies au préalable. Les contraintes utilisées par MaxEntScan ont été définies suite à l'analyse d'un jeu de données d'apprentissage, constitué de l'ensemble des sites consensus d'épissage du génome humain. Deux contraintes ont ainsi pu être définies suite à l'apprentissage : l'existence d'un motif GT en position +1 ➔ +2 d'un site donneur (défini ici de -3 à +6), et l'existence d'un motif AG en position -1 ➔ -2 d'un site accepteur (défini ici de -20 à +3). MaxEntScan attribue ainsi un score aux sites consensus à partir de cette entropie maximale.

f. SSF : Splice SiteFinder

SSF[200] est un autre algorithme de prédiction de l'impact des variants sur les sites d'épissage, utilisant des matrices de positions pondérées. Ces matrices correspondent à la répartition de chaque nucléotide (fréquence à laquelle ils sont retrouvés) pour chaque position d'un site consensus d'épissage. Deux matrices ont été déterminées, une matrice pour les sites donneurs (définis ici de -3 à +6) et les sites accepteurs (définis ici de -14 à +1). SSF calcule ensuite un score utilisant la différence entre la somme des fréquences de la séquence testée et la somme des fréquences minimales, pondérée par la différence entre la somme des fréquences minimales et la somme des fréquences maximales. HSF[201] (Human Splicing Finder) fonctionne sur le même principe, les matrices de position pondérées ayant été calculées sur l'ensemble des sites consensus du génome humain.

g. Comparaison des performances des différents outils

Les différents outils décrits ci-dessus présentent tous des approches différentes, ayant de plus été « entrainés » sur des jeux de données différents. Plusieurs études se sont ainsi attachées à évaluer si ces outils pouvaient fournir des différences au niveau de l'interprétation de la pathogénicité d'un variant [202–204].

Peng-Wei et al.[203] ont ainsi pu observer des problèmes d'évaluation avec SIFT et Polyphen-2. Sur un jeu de 13 572 SNPs non-sysnonymes, 7 % de ces variants n'ont pas été évalués par SIFT, et 9% n'ont pas été évalués par PolyPhen (Prédiction manquantes). Ainsi, même sur des variants intervenant sur les régions codantes du génome, les outils n'évalueront pas les mêmes évènements. Dong et al.[202] ont observé le même phénomène en évaluant cette fois 9 outils différents.

La comparaison des valeurs prédictives laisse aussi apparaître des discordances entre les différents outils. En testant 141 variants dont la pathogénicité était connue au préalable, Flanagan et al.[205] ont pu ainsi calculer une sensibilité de 69 % et 68 % respectivement pour SIFT et Polyphen dans la détection de variants pathogènes. Ces outils sont en revanche apparus comme peu spécifiques (13 et 16%). La détection des variants « perte-de-fonction » a à chaque fois été plus efficace que la détection des variants « gain-de-fonction ». Des résultats similaires ont été retrouvés par Dong et al., avec des taux de détection de vrais positifs (variants réellement délétères) de 93 %, 90 % et 92 % pour SIFT, PolyPhen et MutationTaster, mais des taux de vrais négatifs de 64 %, 39 % et 46 %. Cette équipe a néanmoins pu démontrer que les prédictions pouvaient se révéler plus précises en combinant les scores de plusieurs outils. La probabilité qu'un vrai positif ait un score plus élevé que pour un vrai négatif est ainsi de 92 % pour le score combiné, SIFT, PolyPhen et MutationTaster atteignant 78 %, 76 % et 71 % respectivement. Concernant les algorithmes de prédiction

d'épissage, la sensibilité et la spécificité restent aussi variables en fonction des outils, les meilleurs résultats ayant aussi été obtenus par une association de MaxEntScan et SSF[206].

Les outils de prédiction de pathogénicité peuvent ainsi être utiles dans l'aide à l'interprétation des variants et peuvent aider à hiérarchiser les évènements mis en évidence par séquençage à haut-débit, mais doivent être utilisés avec précaution du fait de leur faible spécificité, leur utilisation par une approche combinatoire semblant néanmoins plus efficace. De plus, la classification des variants en fonction de leur pathogénicité est une étape-clé dans les études cas-témoins. En effet des erreurs de prédiction auront tendance à diminuer un effet d'association par l'inclusion de variants non réellement pathogènes, ou à l'inverse par l'exclusion de variants prédits comme neutres malgré leur caractère délétère.

2. Bases de données

Le NGS a déclenché une intensification de l'exploration génomique à grande échelle. Les informations produites par ces technologies ont elles-aussi augmenté exponentiellement, nécessitant d'être classées, hiérarchisées et facilement accessibles afin d'être compréhensibles et utilisables aussi bien dans un cadre de recherche que médical. L'usage de bases de données, généralement accessibles en ligne, permet un accès et un partage rapide des connaissances disponibles pour l'interprétation des données de séquençage.

a. Les bases de données descriptives

Les bases de données descriptives sont des outils compilant des informations d'interprétation sur des gènes ou des variants de séquences dans des gènes connus pour être associés à une pathologie (reliant génotype et phénotype). Les informations qu'elles contiennent peuvent

être fournies par la littérature scientifique ou des laboratoires et sont généralement évaluées par des comités d'experts, constituant ainsi une aide précieuse dans une utilisation médicale.

Une des bases les plus anciennes et aussi une des plus importantes aujourd'hui est la base de données OMIM[207] (Online Mendelian Inheritance in Man). Cette base fait le lien entre les maladies mendéliennes et les gènes qui en sont à l'origine. La première version a été publiée en 1968 (à l'époque dénommée MIM), et est disponible en ligne depuis 1987. Cette base contient aujourd'hui des informations sur toutes les maladies mendéliennes connues et sur plus de 15 000 gènes.

Aujourd'hui des dizaines de bases existent. Certaine peuvent être gène-centrées, telles que BRCAshare[208], BIC[209] (focalisées sur les gènes de *BRCA1* et *BRCA2*), ou proposée par le IARC[210] (International Agency For Research on Cancer) base centrée sur *TP53*, et fournissent des informations de pathogénicité pour chaque variant décrit dans la base.

InSiGHT[211] (International Society for Gastrointestinal Hereditary Tumours) se focalise sur les tumeurs gastro-intestinales héréditaires, en référençant les variants sur tous les gènes associés à la pathologie, ainsi que l'évaluation de leur pathogénicité.

Les bases HGMD[212] (Human Gene Mutation Database), ClinVar[213] et LOVD[214] (Leiden Open Variation Database) sont plus généralistes, proposant de regrouper tous les variants génomiques identifiés en relation avec les pathologies humaines. HGMD se focalise sur les évènements publiés responsables de pathologies héréditaires et vérifiés manuellement par des curateurs. ClinVar comprend à la fois des variants somatiques et constitutionnels, la soumission des variants dans la base étant libre, mais avec 5 niveaux de confiance dans l'interprétation donnée (allant d' « aucune interprétation donnée » à « utilisation en pratique diagnostique »). LOVD enfin est un système de bases de données gène-centrées, agrégeant à la fois des bases externes et proposant un système dans lequel l'utilisateur pourra entreposer

les variants qu'il a identifié en y ajoutant des interprétations ou encore des données clinico-biologiques.

Parmi ces bases de données, certaines telles que DoCM[215] (Database of Curated Mutations) ou CIViC[216] (Clinical Interpretation Of Variants in Cancer), s'attachent à associer les variants aux réponses prédictives à un traitement ou à des thérapies ciblées disponibles ou en essai clinique.

Ces bases constituent une ressource précieuse dans l'interprétation des variants géniques dans un cadre médical, mais doivent néanmoins être utilisées avec précaution. Les critères définissant un variant pathogène vont pouvoir varier d'une base à l'autre. La base de données ClinVar fournit par exemple différents niveaux de certitude dans l'interprétation des variants, allant d'un variant soumis dans la base sans interprétation associée jusqu'aux variants évalués par un comité d'expert ou encore validés en pratique clinique.

D'autre part il a été montré que la classification d'un variant pouvait diverger entre différentes bases, rendant l'interprétation conflictuelle. Vail et al.[217] ont ainsi comparé l'interprétation de 2017 variants différents de *BRCA1* et *BRCA2* identifiés chez 24 650 patients dans leur laboratoire, à travers 5 bases de données locus-spécifique différentes (BIC, ClinVar, HGMD, LOVD et UMD). Seuls 66 % de leurs variants étaient présents dans au moins 1 base, et 6 % identifiés dans les 5.

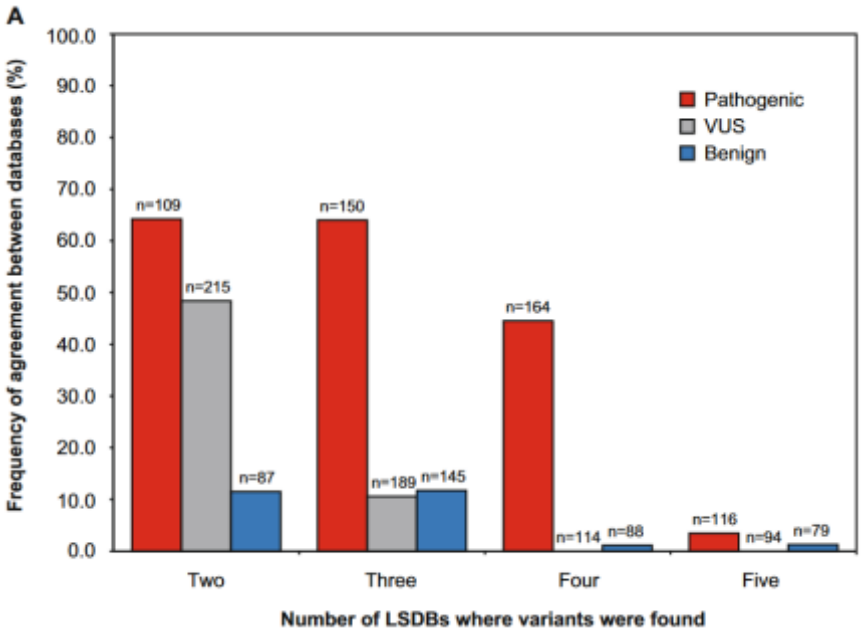


Figure 56: Concordance entre les classifications de pathogénicité de variants de BRCA1 et BRCA2 à travers 5 bases de données locus-spécifiques[217]. Le taux de concordance est représenté en fonction du nombre de bases de données en accord (BIC, ClinVar, HGMD, LOVD, UMD). LSDB : Locus Specific DataBase, VUS : Variants of Unknown Signification.

La Figure 56 décrit dans la même étude le taux de concordance entre les bases de données en fonction du type de variant caractérisé, montrant moins de 5% de concordance entre toutes les bases, quel que soit le type de variant étudié. L'utilisation de ce type de bases doit donc être faite avec précaution afin de ne pas commettre d'erreur d'interprétation.

b. Les bases de données exoniques et génomiques

Le développement du séquençage à haut-débit a aussi permis l'exploration du génome entier, les appareils les plus récents ayant une capacité de séquencer le génome complet de 50 individus par jour (Hiseq X Ten, Illumina). La description du génome ou de l'ensemble des séquences codantes d'un individu et notamment de ses variants (environ 3,3 millions de SNVs par génome[218]) établit un lien avec la biologie de l'individu et son histoire au sein d'une population. L'analyse des variants constitutionnels et de leur fréquence au sein d'une population est aussi précieuse dans le domaine médical, permettant de filtrer les variants les

plus fréquents en population (polymorphismes), pour se focaliser uniquement sur les variants rares.

La première grande base de données recensant les variants génomiques provenant de diverses espèces fut dbSNP[219] (Single Nucleotide Polymorphisme Database), créée en 1998. Cette base a pour vocation d'être un entrepôt de variants génétiques, dans lequel chaque laboratoire peut déposer de nouvelles informations, qui peuvent ensuite être complétées par des laboratoires additionnels. dbSNP a aussi rassemblé les premières informations de fréquence en population des variants identifiés, facilitant le travail d'interprétation des laboratoires. Néanmoins la procédure de soumission trop permissive a conduit à des informations incomplètes ou inexactes. Musumeci et al.[220] ont ainsi évalué en 2010 que la moitié des variants reportés dans dbSNP n'étaient en réalité que des candidats, non validés dans une population, en grande partie dus à la présence d'une séquence paralogue. Cette équipe a aussi démontré que plus de 8 % des SNVs bi-alléliques codants étaient en réalité artéfactuels.

La première étude de grande ampleur, antérieure à l'apparition du NGS, fut le projet HapMap[221], un consortium international dont l'objectif a été de déterminer les haplotypes (groupe d'allèles transmis de manière concomitante) les plus fréquents dans le génome humain (SNVs avec une fréquence > 1 % en population). Cette étude, lancée en 2002, se fonde sur l'hypothèse que, à l'inverse des maladies mendéliennes rares (pour lesquelles un variant nucléotidique sera à l'origine de la pathologie), les maladies « communes » telles que le diabète, les maladies cardiaques, la dépression ou l'asthme proviennent d'une combinaison de différentes variants génomiques associés à des facteurs environnementaux. La mise en évidence des facteurs génétiques impliqués dans ces maladies nécessite théoriquement d'étudier la séquence génomique complète d'une population d'individus malades, en regard d'individus sains. L'accès à l'ensemble du génome des individus étant au lancement de

l'étude irréalisable, le consortium a pris le parti de n'étudier que les SNPs avec une fréquence allélique en population supérieure à 1 %, en se basant sur le fait que deux allèles situés sur les loci différents (mais relativement proches) pourront être associés de manière préférentielle (la fréquence d'association des allèles considérés est différente de la fréquence attendue pour des allèles indépendants et associés aléatoirement : déséquilibre de liaison). La description d'une série de SNPs permet ainsi d'estimer l'haplotype de l'individu pour une région considérée. Le génotype des individus malades et sains peut ensuite être comparé afin de déterminer les régions génomiques impliquées dans la pathologie. Le consortium a ainsi génotypé 269 individus d'origines géographiques diverses, permettant de décrire les fréquences alléliques des SNPs évalués par population. Les haplotypes décrits par HapMap peuvent ainsi servir de population saine dans des études cas-témoins. Néanmoins la focalisation sur les SNPs connus au moment de l'étude sans exploration complète et naïve de la séquence génomique laisse dans l'ombre une grande partie des variants présents sur le génome.

Le projet 1000 genomes[222], débuté en 2008, est le premier projet de séquençage du génome complet à grande échelle, permettant une exploration des variants génomiques à l'échelle d'une population, grâce aux technologies de séquençage à haut-débit. 3 900 individus en provenance de 30 populations différentes ont finalement été séquencés (génome complet à faible profondeur, exome complet à forte profondeur et puces à ADN denses), mettant en évidence 84,7 millions de SNPs, 3,6 millions d'indels et 60 000 variants structuraux. Les fréquences alléliques des variants en population ont pu être évaluées en population mais aussi par sous-population (jusqu'à 1 % à travers le génome et 0,1 - 0,5 % sur les régions exoniques), ainsi que les haplotypes et les déséquilibres de liaison. Ce projet a ainsi constitué un outil de choix dans les études de GWAS pour les imputations de variants, et a permis d'envisager des études cas-témoins par séquençage à haut débit géographiquement appariées.

L'étude ESP[223] (Exome Sequencing Project) s'est ensuite focalisée sur les maladies cardiaques, pulmonaires et sanguines, et a ainsi séquencé l'exome de 6 515 individus américains (classés en populations) afin de caractériser les variants génétiques associés à ces pathologies. Elle propose également une mise à disposition de ses données, avec la fréquence en population (et en sous-population d'origine africaine ou européenne) des variants identifiés.

Dernièrement, le consortium ExAC[224] (Exome Aggregation Consortium) a agrégé et normalisé les données de séquençage d'exome provenant de 14 consortia différents (populations malades et saines), dont 1000 genomes et le TCGA (The cancer Genome Atlas), totalisant ainsi les données exomiques de 60 706 individus. L'ampleur de la population étudiée a permis une augmentation substantielle de la résolution dans l'analyse des variants avec une très faible fréquence en population. Sur les 7 404 909 exoniques identifiés dans la population, 99 % ont une fréquence en population inférieure à 1 %, la moitié sont des variants identifiés une seule fois dans la population, et 72 % sont absents des populations 1000 Genomes et ESP. L'ensemble des résultats sont disponibles à destination de la communauté scientifique. Ce projet s'est ensuite développé pour former le *Genome Aggregation Database* (gnomAD), une coalition d'investigateurs dont l'objectif est de rassembler les données d'exome mais aussi de génome d'un maximum de projets de séquençage de grande ampleur, et qui regroupe aujourd'hui les résultats de 123 136 exomes et 15 496 génomes. Ce projet ainsi constitue une référence précieuse et puissante pour l'étude des populations.

3. Annotation des variants

Une première étape dans l'interprétation des variants génomiques consiste à traduire l'impact fonctionnel que peuvent avoir ces variants. Cette tâche en bioinformatique est réalisée par un

annotateur, dont le rôle est d'apporter des informations aux variants afin de la contextualiser dans un génome et sur un transcrit en général. Ils associent au variant génomique le type de région impactée par le variant (intronique, exonique, intergénique, site d'épissage…) et la conséquence au niveau protéique (insertion d'un codon stop prématuré, synonyme, non-synonyme…). Ils peuvent aussi agréger les informations des algorithmes de prédiction *in silico*, des bases de données locus-spécifiques ou encore les fréquences alléliques observées dans la population générale, tirées des bases de données exomiques et génomiques.

De nombreux outils existent, dévolus à cette tâche, parmi lesquels SNPeff[225], Variant Effect Predictor[226] (VEP), Oncotator[227], Annovar[228] ou encore Alamut-Batch[229]. Ils fournissent tous une conversion du variant selon la nomenclature HGVS (Human Genome Variation Society), convention internationale pour la description des variants génomiques. Néanmoins malgré l'utilisation de cette nomenclature, des discordances peuvent apparaître. En effet, la conséquence au niveau protéique pourra diverger en fonction du transcrit étudié, les logiciels ne possédant pas forcément la même base de transcrits pour leur annotation. Il a ainsi été montré qu'en utilisant deux jeux de transcrits différents (REFSEQ et ENSEMBL), Annovar ne montrait que 85% d'annotations concordantes, ce taux de convergence tombant à 44% en se focalisant sur les variants « perte de fonction » (variants introduisant un codon STOP prématuré dans la séquence et variants touchant les sites consensus d'épissage)[230]. En comparant l'annotation d'Annovar et VEP utilisant le jeu de transcrits ENSEMBL, la convergence des deux outils est de 86% concernant les variants exoniques.

Ces informations peuvent se révéler très utiles pour la détermination de la pathogénicité d'un variant, mais sont à utiliser avec précaution et en traçant la version de bases utilisées. En effet les divers annotateurs peuvent utiliser des versions différentes des logiciels de prédiction ou des bases de données, laissant parfois apparaître des discordances dans l'évaluation d'un variant avec les mêmes outils.

4. Etude Cas/Témoins adaptées aux données de séquençage à haut débit

L'association d'un variant avec une pathologie peut être prouvée par la démonstration de son enrichissement dans une population de malades, en regard d'une population saine (étude cas-témoins).

L'analyse d'effet d'association avec la pathologie variant par variant est une approche standard dans les études de GWAS (Test de variant uniques : Figure 57A). Celle-ci est adaptée dès lors que les variants restent suffisamment fréquents pour conserver une puissance statistique. Lorsque les variants sont rares de grands effectifs de test sont indispensables. Il est admis que les variants comme prédisposant à l'apparition d'un cancer sont des variants retrouvés rarement dans la population générale en dehors de tout isolat génétique. L'analyse de variants de faible fréquence est donc rendue difficile du fait de la taille des populations d'étude requise. Ainsi pour démontrer une association avec un odds ratio de 1,4 et une puissance de 80 % (probabilité de détecter une différence lorsque celle-ci existe réellement), la taille de la population nécessaire est de 6 400, 54 000 et 540 000 pour un variant avec une fréquence allélique de 0,1 / 0,01 / 0,001 %[231]. Des études d'association des variants de *PALB2*, *ATM* et *CHEK2* avec le syndrome de prédisposition au cancer du sein et de l'ovaire ont été réalisées récemment[67] grâce à des puces de génotypage, elles ont nécessité la mise en place de populations de cas et de témoins comprenant plus de 40 000 individus et n'explorent pas la part des nouveaux variants. Ainsi, à titre d'exemple le variant *PALB2* c.1592delT a été décrit avec un OR 3.44 mais avec un intervalle de confiance large (IC 95% 1.39-8.52), ce qui montre les limites de ces approches.

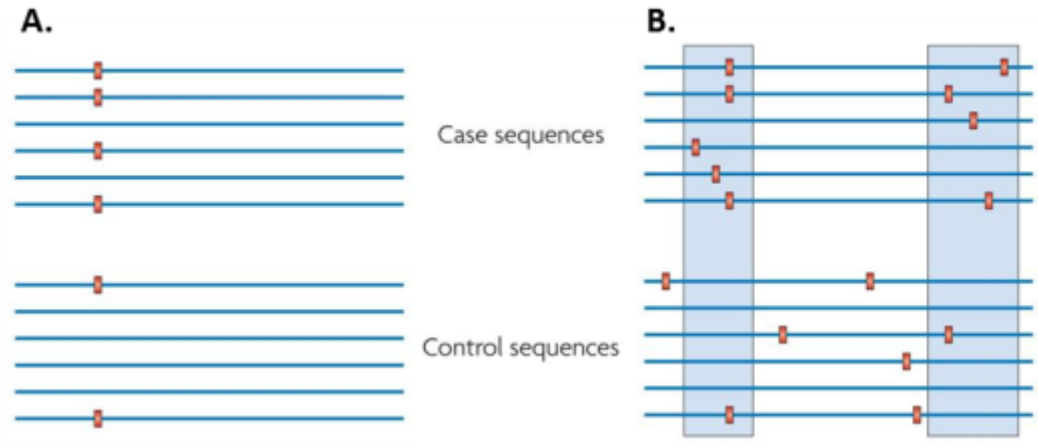


Figure 57: Visualisation des variants dans les études cas-témoins. **A.** Variant unique plus fréquents chez les cas que chez les contrôles : tests de variants uniques **B.** variants multiples situés sur le même locus ou le même gène, plus fréquents chez les cas que chez les contrôles : Tests en agrégation.

Ainsi, plutôt que de tester un seul variant individuellement, il est possible de tester l'association cumulative de plusieurs variants dans un gène ou un locus donné, permettant d'associer la région génomique à la pathologie (Tests en agrégation : Figure 57B). L'avantage de cette méthode est d'augmenter la puissance statistique quand plusieurs variants différents peuvent être associés à la pathologie (Figure 58). Egalement une autre manière d'augmenter la puissance statistique est de théoriquement enrichir la population des cas en variants provoquant des effets forts (en augmentant les OR attendus, le nombre de cas et de témoins nécessaires diminue). Cette hypothèse d'étude peut être employée en sélectionnant la population des cas sur un critère extrême de leur phénotype (forte pénétrance du trait génétique).

	Tests de variants uniques	Tests en agrégation
10 variants / RR = 2 pour tous / AF = 0,005	0,05	0,86
10 variants / RR = 2 pour tous / AF variables	0,20	0,85
10 variants / RR = 2 en moyenne/ AF = 0,005	0,11	0,97

Figure 58: puissance des tests statistiques des variants uniques comparés aux tests en agrégation[232]. Les estimations sont réalisées sur une des données simulées comportant 250 cas et 250 témoins. RR = Risque Relatif ; AF = Allele Frequency.

Plusieurs types de tests en agrégation existent : les Burden Tests, les tests des composantes de la variance, ou encore les tests combinés (Figure 59).

Type de méthode	Méthodes disponibles	Caractéristiques principales	Discrimination des facteurs de risques / protecteurs	Type de variants étudiés
Burden test	ARIEL test, RWAS, CAST, CMC method, MZ Test, WSS, aSum, Step-up, EREC test, VT, KBAC method, RBT	rassemble tous les variants dans un seul score, en partant de l'hypothèse que tous les variants testés sont causaux et associés avec le trait étudié	Non	Variants causaux (ex: variants "Perte de Fonction" LoF)
Test des composantes de la variance	C-Alpha test, SKAT, SSU test, KBAT	Autorise l'étude simultanée de facteurs de risques et protecteurs	Oui	Tous, avec une stratégie de pondération possible
Test combiné	SKAT-O, EMMPAT, Fisher method, MiST	Combine les résultats de deux tests complémentaires ou plus	Oui	Tous, avec une stratégie de pondération possible
Autres tests	LASSO, EC	Prend en compte la rareté du signal	Non	Tous, avec une stratégie de pondération possible

Figure 59: Descriptions des méthodes utilisées pour les études d'association de variants rares[233]. ARIEL accumulation of rare variants integrated and extended locus-specific, aSum data-adaptive sum test, CAST cohort allelic sums test, CMC combined multivariate and collapsing, EC exponential combination, EPACTS efficient and parallelisable association container toolbox, EREC estimated regression coefficient, GRANVIL gene- or region-based analysis of variants of intermediate and low-frequency, KBAC kernel-based adaptive cluster, MiST mixed-effects score test for continuous outcomes, MZ Morris and Zeggini, RBT replication-based test, Rvtests rare-variant tests, SKAT sequence kernel association test, SSU sum of squared score, VAT variant association tools, VT variable threshold, WSS weighted-sum statistic

La réalisation de ces tests doit être rigoureuse sur tous les aspects de l'analyse afin d'éviter les biais. L'ensemble des analyses bio-informatiques doivent être contrôlées sur plusieurs critères-qualité « clé », tels que la profondeur de séquençage, les ratios transitions / transversions ou encore le ratio de variants hétérozygotes. Des scores-qualité complémentaires pourront être évalués en fonction des outils utilisés (ex : VQSLOD pour HaplotypeCaller). L'objectif est de sélectionner les données de meilleure qualité, tout en permettant de supprimer un maximum de variants faux-positifs. Le taux de « missingness » (cf. Introduction, § VII.A.6.a, analyse de populations) est aussi évalué, afin de constituer une analyse homogène et équilibrée des régions génomiques entre population cas et population témoin.

Enfin les variants testés pour l'association pourront être catégorisées, sur la base de l'impact au niveau protéique (variant non-sens, *frameshift*, faux-sens…), de la fréquence dans la population générale, ou encore en fonction des prédictions bio-informatiques de pathogénicité réalisées par des outils tels que SIFT, PolyPhen ou MutationTaster (Notion de pondération). Cette catégorisation permet dans certains cas de pondérer les effets.

Aussi la fréquence des variants rares peut varier en fonction du groupe ethnique étudié, nécessitant un appariement rigoureux entre population cas et population témoin afin d'éviter l'introduction de biais de stratification, pouvant fausser les effets d'association. Un biais de stratification est issu de la constitution d'une population d'étude non-homogène, comprenant plusieurs sous-populations aux caractéristiques génétiques différentes, avec des risques différents. Le résultat d'un test d'association pourrait ainsi conclure à tort à une association. La stratification de la population étudiée en fonction du groupe ethnique est généralement réalisée par des analyses en composante principale (PCA : Principal Component Analysis), méthode statistique qui, appliquée ici, permet de discriminer les populations en fonction du génotype de variants fréquents.

Ces analyses sont d'un intérêt majeur pour découvrir de nouveaux facteurs de prédisposition génétique au cancer du sein et de l'ovaire et peuvent être adaptées à l'analyse de panels de gènes complets par séquençage à haut-débit (exome ou panel de gènes candidat) afin d'explorer la part des variants rares dans le syndrome de prédisposition au cancer du sein et de l'ovaire [234,235]. Mais la principale problématique de ce type d'analyse demeure encore la taille des populations étudiées pour démontrer un effet d'association significatif, malgré l'utilisation de tests en agrégation. En effet les variants identifiés peuvent avoir une fréquence allélique très inférieure à 0,1 % en population générale, voire être des singletons. Aussi la conduite de ces tests nécessite d'obtenir une population de témoins de taille relativement équivalente. Pour rassembler de telles populations d'étude il semble judicieux de mettre à

profit l'extraordinaire accumulation des données de génomiques produites lors des diagnostics moléculaires initiaux[236] pour constituer une population de cas. Egalement, l'utilisation de bases de données telles qu'ExAC ou gnomAD peut être envisagée avec une extrême prudence pour reconstituer et simuler des populations de témoins. Il s'agit donc d'un véritable défi qui demande de re-calibrer entre elles des données d'origine hétérogènes afin de pouvoir les comparer sans introduire des biais qui seraient responsables d'interprétations erronées.

Résultats

I / Estimation des risques de cancers induits par les variants pathogènes de 34 gènes identifiés par NGS chez 5131 familles présentant une prédisposition au cancer du sein et de l'ovaire

A. Présentation de l'étude

Depuis ces deux dernières décennies, le diagnostic moléculaire des prédispositions au cancer du sein et de l'ovaire est basé principalement sur l'identification des variants constitutionnels inactivateurs des gènes de *BRCA1*[237] et *BRCA2*[20]. Ces variants sont responsables d'environ 10 % des présentations personnelles ou familiales évocatrices d'une prédisposition au cancer du sein et de l'ovaire. Si les études de liaison ont permis l'identification de *BRCA1* et *BRCA2*, elles n'ont pu mettre en évidence l'ensemble des facteurs génétiques liés à ce syndrome. De nombreuses études d'association pangénomiques basées sur la mise en œuvre de « SNP array », permettant la comparaison de la fréquence de polymorphismes communs (fréquence allélique > 5%) entre une population cas et une population témoin appariée, ont cherché d'autres facteurs de prédisposition. Ces études de GWAS supportent la théorie du « common disease / common variant », proposant d'expliquer la composante génétique des maladies fréquentes (schizophrénie, troubles bi-polaires, diabète, hypertension artérielle, cancers…) par une combinaison d'allèles fréquents mais de risque faible. Malgré tout, ces études n'ont pu mettre en évidence que des SNPs associés à des odds-ratios dépassant rarement 1,5[70], et la combinaison de ces SNPs à risque ne peut expliquer l'ensemble des situations cliniques évocatrices d'un risque génétique élevé. Cette part d'héritabilité manquante est probablement

liée à l'existence de variants de fréquence allélique faible (< 0,1 %, dits variants rares) que les techniques de *SNP array* n'étaient pas en mesure de détecter, ce qui a été confirmé par certaines études employant l'approche gènes-candidats dans des familles très sélectionnées mais sans pour autant déterminer finement les niveaux de risques associés aux variants pathogènes identifiés[148].

Aujourd'hui, grâce aux technologies innovantes du NGS, il est possible de caractériser les variants rares sur de grandes populations en séquençant soit un nombre de gènes restreints (panel) soit un exome voire le génome complet. Le NGS permet donc d'envisager des études d'association en exploitant la part des variants rares grâce à des méthodes statistiques adaptées, par des tests en agrégation de variants multiples. Nous avons donc étudié les données rétrospectives issues du génotypage de la grande majorité des gènes candidats connus dans les prédispositions génétiques au cancer du sein et de l'ovaire chez 5131 patientes ayant bénéficié d'une analyse moléculaire (réalisée au Laboratoire de Biologie et de Génétique du Cancer (LBGC), Centre François Baclesse) dans le cadre d'une consultation d'oncogénétique.

Ce travail a ainsi eu pour objectif de réaliser des études d'association basées sur l'agrégation et la caractérisation de variants rares afin d'estimer les niveaux de risque de cancer induits par la présence de ces évènements dans tous les gènes candidats du panel de gènes utilisé dans le cadre du diagnostic moléculaire de prédispositions au cancer du sein et de l'ovaire.

Cette étude a tiré profit de différentes sources de données induisant la comparaison de données hétérogènes. Les données des populations témoin utilisées proviennent en effet de l'exploitation des données mise à disposition par le consortium ExAC, de la réanalyse des données des individus du projet FREX (French Exome Project) et de la réanalyse des données de la population des cas issues du diagnostic. Le défi majeur de ce travail a ainsi consisté à rendre ces données comparables de manière à contraindre les biais en employant une

méthodologie permettant une maitrise des données manquantes de part et d'autres des jeux de données (Notion de « missingness »), une filtration rigoureuse de la qualité des données, et la prise en compte de la précision de l'incidence dans chaque population.

Afin de prendre en compte la nature hétérogène des données, une nouvelle méthodologie statistique adaptée aux données de séquençage haut-débit et à la problématique des variants rares a été mise en place.

La contribution à ce travail dans le cadre de cette thèse a consisté à participer (i) à la mise en place du pipeline bioinformatique dédié à l'analyse des données de séquençage fournies par le consortium FREX et à l'entrainement de celui-ci, (ii) à la définition de l'algorithme automatisé de hiérarchisation des variants et à l'entrainement de celui-ci, (iii) à la mise en place et adaptation (non présentée dans le cadre de cette thèse) du modèle statistique de l'étude.

B. Descriptif de la méthodologie originale employée pour l'étude : Estimation par simulation en population générale de la probabilité de porter un variant pathogène et calcul d'Odd Ratio

Les données du consortium ExAC (release0.3.1 nonTCGA subset) fournissent la fréquence observée d'un variant dans des populations européennes non finlandaises. Cependant les génotypes complets ne sont pas disponibles, ne permettant pas d'évaluer le nombre précis de variants portés par un individu donné. Cette base de données ne peut donc pas être utilisée en tant que population témoin pour une étude cas-témoins classique (calcul d'*odd-ratios*, tests de comparaisons, tests d'agrégation (ex : Burden Tests)). De plus, les données de séquençage des

individus inclus dans cette base sont elles-mêmes hétérogènes, une partie des régions génomiques pouvant être mieux séquencées chez certains individus que d'autres. Les fréquences alléliques présentées sont ainsi calculées à partir d'un nombre d'individus différents, et donc avec une précision différente.

Par un calcul de probabilité, en émettant l'hypothèse que les variants rares sont indépendants les uns des autres (sans déséquilibre de liaison), il est cependant possible de comparer les probabilités de porter un variant rare pathogène entre la population Cas (P_{cas}) et cette population de référence ($P_{témoin}$) sur une région génomique donnée (ex : par gène ou par voie de signalisation).

Pour chaque variant rare sélectionné, la fréquence allélique est considérée comme une variable aléatoire distribuée selon une loi Beta dont les paramètres sont déterminés à partir des observations de la base ExAC[238].

Ces modélisations des variants permettent de prendre en compte ces niveaux de précision différents. Les distributions Beta associées à chaque variant illustrent bien que plus le nombre d'individus sur lequel a été calculé la MAF est faible, plus la distribution est large, et donc la précision faible (Figure 60A).

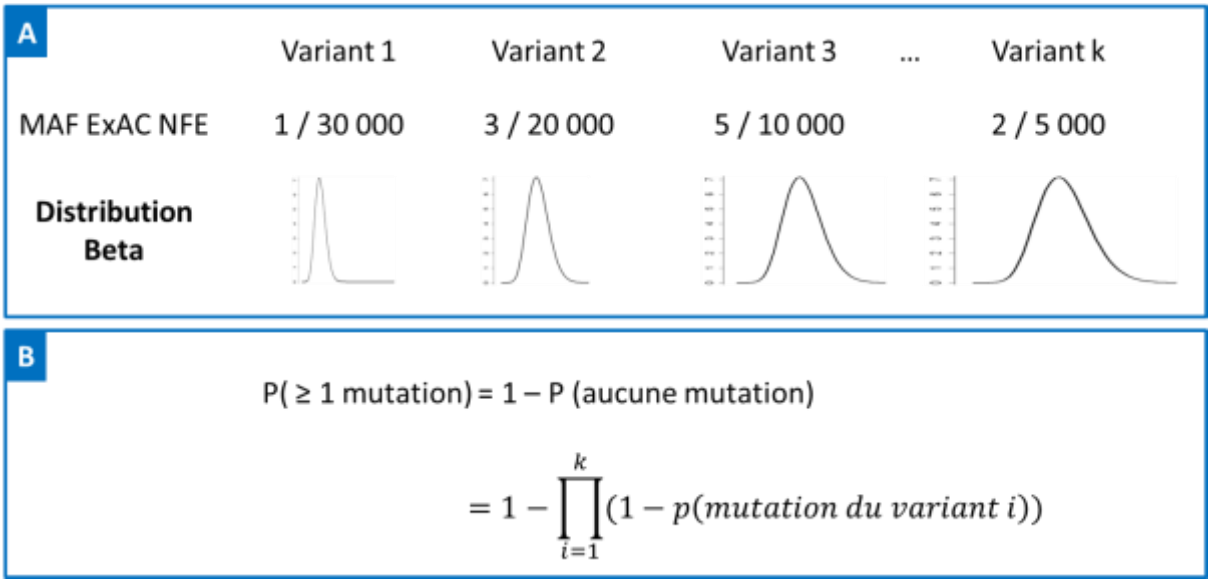


Figure 60: Modélisation des MAF par des lois Beta. (A) Représentation de la précision de l'estimation par la distribution Beta. (B) Calcul de la probabilité de porter au moins un variant rare délétère pour un individu.

L'hypothèse d'indépendance des variants rares délétères permet ainsi d'estimer la probabilité de porter au moins un variant rare délétère pour un individu issu de cette population de référence selon la formule de la Figure 60B.

$P_{témoin}$ et son intervalle de confiance sont obtenus par des simulations de toutes ces variables aléatoires (i.e. distribution beta). A chaque simulation, P(porter le variant i) est simulée selon la loi Beta déterminée pour le variant i et un calcul de la probabilité de porter un variant rare pathogène chez les témoins, $P_{témoin}$ = P(≥ 1 variant), est ainsi réalisé. Sur toutes les simulations réalisées, la valeur moyenne de $P_{témoin}$ donne l'estimation de la probabilité de porter au moins un variant rare pathogène pour un individu issu de la population ExAC. Les quantiles d'ordre 0.025 et 0.975 donnent les bornes de son intervalle de confiance.

En appliquant parallèlement cette méthode de calcul à la population des cas, chaque simulation permet de disposer d'une observation de $P_{témoin}$ et P_{cas}. Pour chaque simulation, il

est donc possible de calculer un odd-ratio de l'enrichissement en variants rares pathogènes de la population cas par rapport à la population témoin :

$$OR = \frac{p_{Cas}/(1 - p_{Cas})}{p_{Temoins}/(1 - p_{Temoins})}$$

On obtient donc autant d'observations d'OR que de simulations. Et, de la même façon que précédemment pour P_{cas} et P_{temoin}, la moyenne des OR obtenus par simulation donnent une estimation de celui-ci et les quantiles d'ordre 0.025 et 0.975 son intervalle de confiance.

Dans cette première étude, ces estimations ont été réalisées en utilisant la population ExAC non-TCGA NFE (NFE : Non-Finnish European) comme population Témoin, 100 000 simulations étant réalisées pour chaque estimation.

C. Publication

Landscape of pathogenic variations in a panel of 34 genes and cancer risk estimation from 5131 HBOC families

Running title: Estimation of risk cancer with HBOC genes panel test

Laurent Castéra[1,2,*], Valentin Harter[1,3], Etienne Muller[1,2], Sophie Krieger[1,2,4], Nicolas Goardon[1], Agathe Ricou[1], Antoine Rousselin[1], Germain Paimparay[1] , Angelina Legros[1], Olivia Bruet[1], Céline Quesnelle[1], Florian Domin[1], Chankannira San[1], Baptiste Brault[1], Robin Fouillet[1], Caroline Abadie[5], Odile Béra[6], Pascaline Berthet[7], French Exome Project consortium[#], Thierry Frébourg [2,13], Dominique Vaur[1,2]

[1] Laboratory of Cancer Biology and Genetics, Comprehensive Cancer Center François Baclesse, Caen, France;

[2] Inserm U1245, Rouen University, Normandy Centre for Genomic and Personalized Medicine, France;

[3] Northwest Data Center (CTD-CNO), Comprehensive Cancer Center François Baclesse, Caen, France

[4] Caen University, France;

[5] Department of Genetics, Comprehensive Cancer Center Eugène Marquis, Rennes, France ;

[6] Department of Genetics, CHU, Fort de France, France

[7] Department of Genetics, Comprehensive Cancer Center François Baclesse, Caen, France ;

[8] UMR1078, Brest, France

[9] Institut du Thorax, UMR 1087 Nantes, France

[10] Centre National de Génotypage, Evry, France

[11] UMR 744, Lille, France

[12] UMR 1219, Bordeaux, France ;

[13] Department of Genetics, University Hospital, Rouen, France

[#] Principal co investigators of French Exome Project consortium: Emmanuelle Génin, UMR1078, Brest; Richard Redon, Institut du Thorax, UMR 1087 Nantes, Jean-François Deleuze, Centre National de Génotypage, Evry, Dominique Campion, Inserm U1245, Rouen, Jean-Charles Lambert, UMR 744, Lille ; Jean-François Dartiques UMR 1219, Bordeaux

ABSTRACT

Integration of gene panels in the diagnostic of Hereditary Breast and Ovarian Cancer (HBOC) requires a careful evaluation of the risk associated with variants detected in each gene. We analyzed by NGS 34 genes in 5131 index cases suspected to present HBOC. Using ExAC data sets and 571 individuals from the French Exome Project (FREX), the probability than an individual from ExAC carries a pathogenic variation was simulated and compared to the estimated frequency within the HBOC population. Odds-ratio conferred by pathogenic variants within *BRCA1*, *BRCA2*, *PALB2*, *RAD51C*, *RAD51D*, *ATM*, *BRIP1*, *CHEK2* and MSH6 were estimated to 13.22[10.01-17.22], 8.61[6.78-10.82], 8.22[4.91-13.05], 4.54[2.55-7.48], 5.23[1.46-13.17], 3.20[2.14-4.53], 2.49[1.42-3.97], 1.67[1.18-2.27] and 2.50[1.12-4.67], respectively. Variations within *RAD51C*, *RAD51D*, and *BRIP1* were associated with ovarian cancer family history (OR = 11.36[5.78-19.59], 12.44[2.94-33.30] and 3.82[1.66-7.11]). *PALB2* variants were associated with bilateral breast cancer (OR=16.17[5.48-34.10]) and *BARD1* variants with triple negative breast cancer (OR=11.27[3.37-25.01]). Burden tests performed in patients and FREX population confirmed the association of pathogenic variants of *BRCA1*, *BRCA2*, *PALB2* and *RAD51C* with HBOC. Our results validate the integration of *PALB2*, *RAD51C*, *RAD51D* in the molecular diagnostic of HBOC and support that the other genes are involved in an oligogenic determinism.

Key words: HBOC, genetic risk estimation, panel gene sequencing

INTRODUCTION

The development of high throughput sequencing has created an unprecedented revolution of the molecular diagnosis of inherited diseases. In this context, Hereditary Breast and Ovarian Cancer syndrome (HBOC) represents an interesting paradigm. Indeed, molecular diagnosis of

HBOC was initially restricted to the identification of germline pathogenic variants (PV) within the 2 main genes BRCA1 and BRCA2[20,237]. Today, medical laboratories, that have integrated in their practice bioinformatics expertise quality and insurance (i.e. ISO 15189) exigencies for next generation sequencing (NGS), are performing molecular diagnosis of HBOC by NGS analysis gene panels[57,239–241]. As other teams, we previously reported that in families suspected to present HBOC, almost 50% of PVs were identified within other genes than *BRCA1* and *BRCA2*[57]. The involvement in breast cancers of other genes such as *TP53*, whose germline mutations cause the Li-Fraumeni Syndrome (LFS) and represent the main cause of very-early breast cancers occurring before 31 years of age, has been clearly established[242]. In contrast, the imputation of germline mutations within other genes and the level of risk cancer induced by these PVs are a matter of debate and the use of gene panels in HBOC diagnosis a subject of controversy in medical laboratories[243,244]. Indeed, integration of gene panels in the diagnostic of HBOC requires a careful evaluation of the risk associated with the genetic variants detected in each gene, considering the potential impacts on the mutation carrier clinical management. For some of the non BRCA genes, progressive accumulation of segregation, statistical and functional data knowledge has refined the risk evaluation[236]. For instance, PVs in *PALB2* have clearly been shown to be associated with an increased risk of breast cancer, but according to the publications, the level of risk varies from moderated to high-risk in the same magnitude than that conferred by *BRCA2* PVs[67,245,246]. Consequently, the appropriate clinical management of *PALB2* PV mutation carriers is still debated. In the same manner, the levels of ovarian cancer risk associated to *RAD51C* and *RAD51D* PVs appear to be sufficient to propose prophylactic salpingo-oophorectomy after 50 years of age, although the estimated risk confident intervals are large and the medical community hesitates to systematically recommend risk-reducing surgery[130,247]. In contrast, the estimated risks associated to germline PVs detected in other genes commonly integrated in

gene panel diagnostic for HBOC, such as *BRIP1* for ovarian risk[68], or *BARD1*[248] and *FANCM*[249] for breast cancer risk, are either to low or contradictory to use them in order to guide clinical management and genetic counseling within the families.

This imprecision of the estimated levels of risks is, in part, explained by the low incidence of PVs detected within these genes. Consequently, case/control studies requiring thousands of cases and appropriate controls are required to document significant enrichment of PVs in cases versus controls. A very large association study, performed on 41 611 breast cancer white patients and Exome Aggregation Consortium reference appropriate controls, recently reported that beside the *BRCA*, *CDH1*, *PTEN*, and *TP53* genes only pathogenic variants in *PALB2* were associated to a high breast cancer risk (OR, 7.46). Variants within *ATM*, *BARD1*, *CHEK2* and *RAD51D* were found to be associated to moderately increased risks with OR estimated to 2.78; 2.16; 1.48; and 3.07 respectively[246]. Although this study provides very useful information for diagnostic laboratories, it might expose to the risk of the genetic heterogeneity of the analyzed samples. If case-control studies must take into account the stratification bias induced by the geographical ancestral origin of the controls[224,246,250] , they also need to take into account the potential bias resulting from the allelic frequency provided by ExAC dataset. Indeed, allelic frequencies in the ExAC database are deduced from the adjusted counts obtained after quality filtration of the sequenced alleles. Comparison of these allelic frequencies in controls versus that observed in patients, without taking into account the heterogeneity of this measurement, may impact the risk estimation.

To analyze the association with HBOC of variants detected within 34 genes among 5131 French consecutive patients, we developed a model which allows calculating the probability that a case or a control is carrier of a variant. To limit the bias and provide accurate estimates of the risks associated with the detected variants, we restricted the analysis to patients from

French origin and we performed burden tests, using genomic individual data from 571 French control genotypes sequenced by the French Exome (FREX) Consortium.

PATIENTS AND METHODS

Patients and controls

Personal criteria	
	Number of patients*
Breast carcinoma < 36 years	541
Triple negative breast carcinoma (TNBC) < 51 years	394
Medullary breast carcinoma < 61 years	19
Male breast carcinoma before < 71 years	63
Ovary adenocarcinoma < 71 years	656
Or Familial criteria	
Two breast carcinomas in first or second-degree relatives (with a transmitting male), with at least one cancer < 51 years and the other before 71 years	
Three breast carcinomas in first or second-degree relatives, with at least one < 61 years	
Breast carcinoma < 51 with first-degree relatives with either prostate cancer < 61 years or pancreatic cancer < 61 years or ovary adenocarcinoma < 71 years	
Other familial presentations suggestive of HBOC validated by a multidisciplinary team	

Table S1. Clinical presentation of the patients analyzed in this study

All patients analyzed in this study have been seen in the context of genetic counseling and fulfilled at least one of the criteria presented in the supplementary Table S1. A consecutive series of 5131 French patients was studied. For each patient, informed consent for genetic analysis was obtained. We collected genomic data from 574 control individuals originated from 6 French regions and sequenced in the framework of the FREX project (www.france-genomique.org/spip/spip.php?article158). All individuals analyzed in this study were tested for their geographic origin, using multiple correspondence analysis (MCA) of common SNPs

within 34 HBOC genes (Supplementary Table S2), with a MAF>5%, as determined on the 1000 Genome phase 3 data. After filtration, 281 common variants were thus included. In order to compute posterior probabilities of the individual's geographical origin, a linear discriminant model was adjusted on the first relevant axis from the MCA. Individuals with a posterior probability of non-European origin > 95% were excluded from analyzes (arbitrary threshold).

Table S2: Distribution of variants. C5: pathogenic variants; C4: probably pathogenic variants; SD: Strictly damaging - variants predicted as pathogenic by three bioinformatics prediction algorithms.

Gene	Number of patients included	Number of C5 and C4 variants	Number of SD variants
BRCA2	4409	174	36
BRCA1	4406	166	26
MUTYH	4147	98	22
CHEK2	4374	51	65
ATM	4408	46	137
PALB2	4173	37	31
RAD51C	4309	23	11
TP53	4180	21	5
BRIP1	4408	21	81
MSH2	3800	11	27
MSH6	4408	11	23
MLH1	4340	10	99
RAD50	4399	10	20
RAD51D	4011	9	14
NBN	3617	8	62
PMS2	4130	8	29
PMS1	4408	8	40
BARD1	3667	7	23
MRE11A	4408	7	30
MLH3	4395	6	42
RINT1	2373	3	8
XRCC3	2567	2	10
XRCC2	3729	2	6
CDH1	3988	2	12
RAD51B	4387	2	9
FAM175A	2413	1	28
BAP1	3774	1	5
HOXB13	2389	0	12
INHA	2367	0	5
CDKN2A	1010	0	2
INHBA	2354	0	1
STK11	2386	0	1
RAD51	4406	0	1
PTEN	4408	0	0

Sample preparation, enrichment, Next Generation Sequencing and bioinformatic analyses

DNA from patients was extracted from peripheral blood, using the Agencourt® Genfind™ V2 on Biomek FX workstations (Beckman, Villepinte, France). DNA was sonicated using a Covaris S2 (Covaris, Inc. MS, USA). The sample preparation was performed with SPRIworks HT-High Throughput (Beckman, Villepinte, France). Illumina adapters were replaced by indexed adapters (NEB, Milano, Italy), previously published by Huentelman's team[251]. We used Agilent SureSelect to design two different SureSelect solution library baits (Agilent, Santa Clara, California), covering a variable number of HBOC genes (Supplementary Table 2). For each gene, exons and part of the introns were covered by the capture. The SureSelect enrichment process was performed after combining equimolarly indexed samples according to Kenny et al. and published before[57]. The current protocol was robotized on two Biomek FX workstations dedicated to the pre- and post-PCR zone. Libraries were then sequenced on Miseq or NextSeq500 (Illumina, San Diego, USA) using the paired-end program. Data from patients was generated with CASAVA v.1.8 suite from Illumina. Variants from the 5131 VCFs generated by CASAVA were excluded if DP <20 or QUAL < 100 for SNV or QUAL < 500 from Indels, ensuring sensitivity and specificity, as previously described before[57]. For one gene, individual data from the 5131 patients were excluded if at least one exonic position did not reach 20x of coverage. Also if more than 20% of the patients were not covered with at least 20x on the exons, these specific regions were excluded from the final analysis corresponding to <1% of the targeted region (i.e. 5 exons). FREX genomic data were generated using Agilent SureSelect Human All Exon kits (Agilent Technologies, Santa Clara, CA, USA). Final libraries were sequenced on a HiSeq2000 or 2500 (Illumina, San Diego, CA, USA) with paired ends[250]. Individual gVCF were generated by the FREX consortium. Best practices edited by the Broad Institute were used to determine the genotypes. Variant from

FREX with <5% of missing data were kept when VQSR values reached 99.5 for SNV and 99% for Indels.

Probability calculation and statistical analyses

Genomic regions of interest (ROI) were determined as the interceptions of the covered regions by exome and HBOC gene panels (Supplementary Table 2) and variants detected within these ROI were included in the study. The adjusted allele count provided by ExAC for the NFE population in the ExAC non-TCGA subset data (ExAC.r0.3.nonTCGA.sites.vep.vcf.gz) and corresponding to individuals with genotype quality (GQ) >= 20 and depth (DP) >= 10 was used. Pathogenic ExAC variants classified in the non-PASS category, as defined below, in the VCF, were checked manually.

All the variants detected in patients, and ExAC and FREX databases were annotated using ANNOVAR (v2015-12-14 database 2014-10) and Alamut BATCH version (v1.4.2 database 1.4.-2015.11.02) GRCH37 (Interactive BioSoftware, Rouen, France). Analyses were restricted to variants with a MAF <1% in at least one population (patients, ExAC NFE, FREX). Annotated variants were categorized using the ACMG criteria[252]. For genes with alternative transcripts, the most pathogenic prediction was considered. Variants were classified as definitively pathogenic and were called hereafter "class 5" (C5) if (i) the corresponding clinical significance of variants detected was based on consensus data integrated in the French UMD-*BRCA1*/*BRCA2* databases, the IARC *TP53* Database or the Insight database for MMR genes, or (ii) they introduced a premature termination codon (PTC) or (iii) they affect the canonical AG/GT splice sites. Variants were considered as probably pathogenic and were called hereafter "class 4" (C4) if (i) they were defined as probably pathogenic in previous database or (ii) they abolished the translation initiation codon or (iii) If

MaxEntScan and SpliceSiteFinder scores predicted a complete abolition of the donor or the acceptor splice site even if a value was missing (e.i "GC site" not predicted by MaxEntScan")[206]. For the other variants, we defined another class designed strictly damaging (SD) variants, corresponding either to (i) missense variants predicted as pathogenic according to MutationTaster, SIFT and Polyphen-2 (Hum Div) algorithms (ii) or variants outside the canonical AG/GT splice sites but predicted to alter splicing (50 % decrease according to MaxEntScan and SpliceSiteFinder scores, or a complete abolition with one of these algorithms and a decrease with the other).

We then evaluate the enrichment in PVs within 34 HBOC genes. Let $P_{patient}$, P_{FREX} and P_{ExAC} the probabilities that an individual, respectively from the 5131 patients, FREX and ExAC NFE population samples, carries a pathogenic variation (C5 or C4; C5, C4 or SD; or SD only). $P_{patient}$ and P_{FREX} were estimated as proportions with a usual Casagrande-Pike confidence interval. P_{ExAC} is equal to {1 – the probability of an ExAC individual to not carry a pathogenic variation}. Assuming the independence of these rare pathogenic variations, i.e. no linkage disequilibrium, P_{ExAC} can be formulated as:

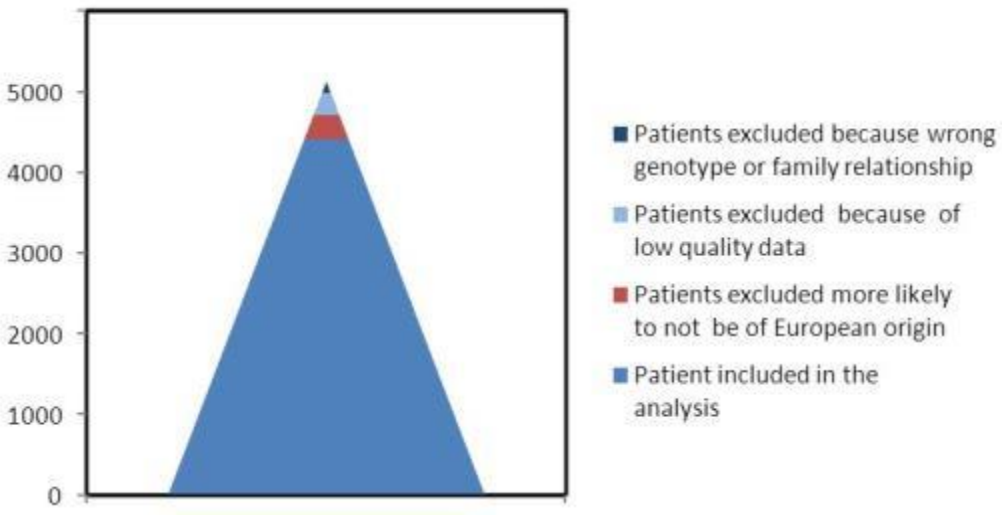
$$\left\{1 - \prod_{i=1}^{k} = 1 - \frac{AC\_NFE_i - Hom\_NFE_i}{AN\_NFE_i/2}\right\}$$

Where $\frac{AC\_NFE_i - Hom\_NFE_i}{AN\_NFE_i/2}$ is the estimated probability that an individual from the ExAC NFE population carries the i^{th} variation between the k pathogenic variants identified in each analyzed gene. With ExAC notation, AN_NFE_i , AC_NFE_i et Hom_NFE_i are respectively the number of alleles considered for the adjusted count, the number of alternative alleles and the number of homozygous individuals counted in the ExAC NFE population for this i^{th} variant. The interval of confidence of P_{ExAC} was empirically determined with modeling each probability $\frac{AC\_NFE_i - Hom\_NFE_i}{AN\_NFE_i/2}$ using Beta distribution with (AC_NFE_i - Hom_NFE_i ;

AN_NFEi/2-(AC_NFE_i and Hom_NFE_i)) as parameters[238]. In the same manner, the Patient/ExAC OR was empirically determined by modeling jointly P_{ExAC} and $P_{patient}$ and calculating $\frac{p_{patient}\cdot(1-p_{ExAC})}{(1-p_{patient})\cdot p_{ExAC}}$ for each simulation. 100,000 simulations were performed for each estimate. The Patient / FREX OR and its confidence interval were estimated in a conventional manner. Two complementary burden tests, corresponding to the cohort allelic sum test (CAST) and weighted-sum test (WST), were used to test the association between HBOC and pathogenic variants, considering the FREX population as the reference.

RESULTS

From 5131 patients suspected to present HBOC, 270 were excluded because the quality of sequencing did not fulfill quality criteria and 160 others were excluded because they were either related to another index case or because clinical criteria did not meet prerequisite for the statistical study (Supplementary Figure S1).



Supplementary Figure S1: Representation of the proportion of selected patients included in analysis

Multiple correspondence analysis (MCA) showed a superposition of the projections of the genotypes from analyzed patients and FREX controls (Figure S1 and Figure 1).

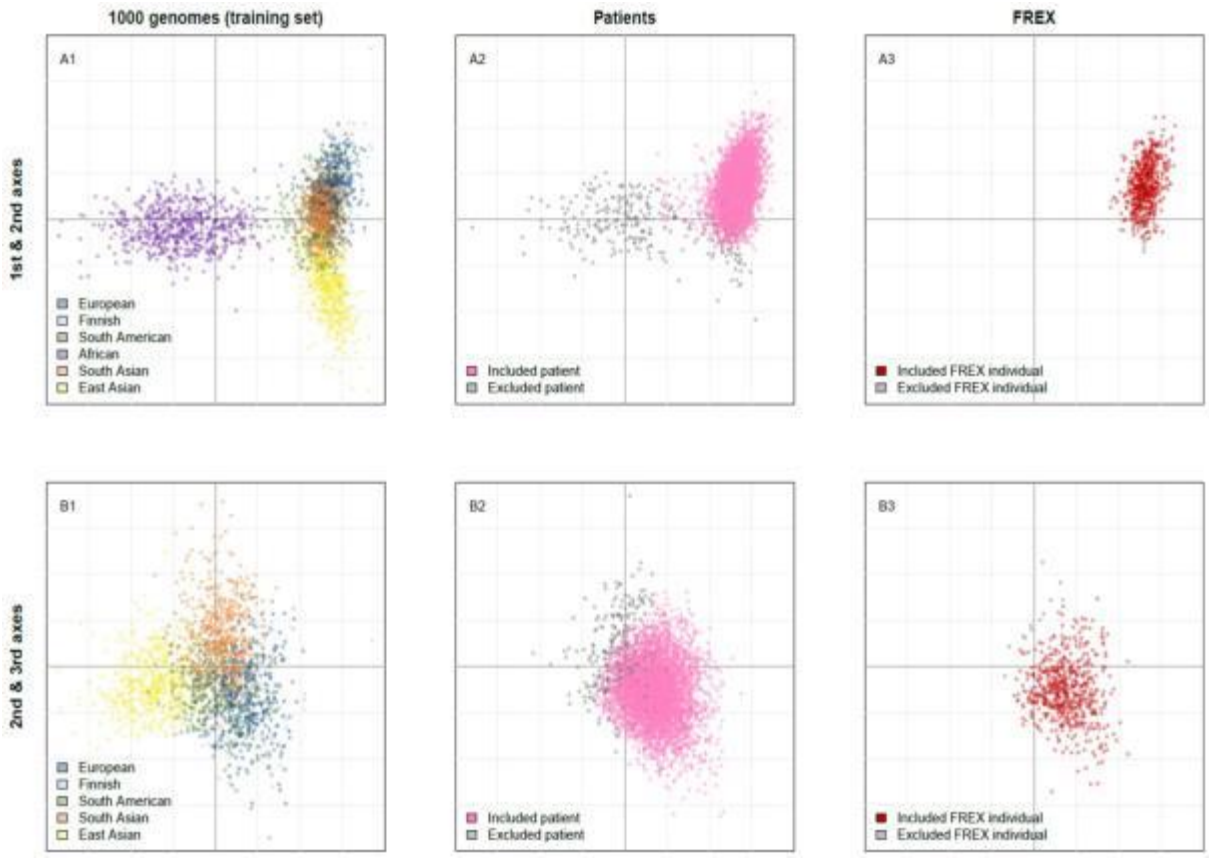


Figure 1: Multiple correspondence analysis and selection of patients aand FREX individuals based on ancestral geographic origins. Projections of 1000 genomes, patients and FREX individual data, respectively on the first discriminant plane for scatter plot A1, A2, A3, on the 2nd and 3rd axes plane for scatter plots B1, B2, B3. Discriminant axes were fitted on the MCA of 1000 genomes training set. Patients and FREX individuals grey dots were predicted non-European origin according to this model (p > 95%) and excluded from the analysis.

MCA detected and excluded from the following analysis 292 additional patients and 12 FREX individuals with a high probability not to be from European origin, according 1000 genome training dataset. In the 4409 remaining patients, the frequency of C5 and C4 variants within *BRCA2*, *BRCA1*, *CHEK2*, *ATM*, *PALB2*, *RAD51C*, *TP53*, and *BRIP1* was 3.9%, 3.7%, 1.1%, 1.0%, 0.9%, 0.5%, 0.5%, 0.5%, respectively (Supplementary Table S2 and Figure 2A).

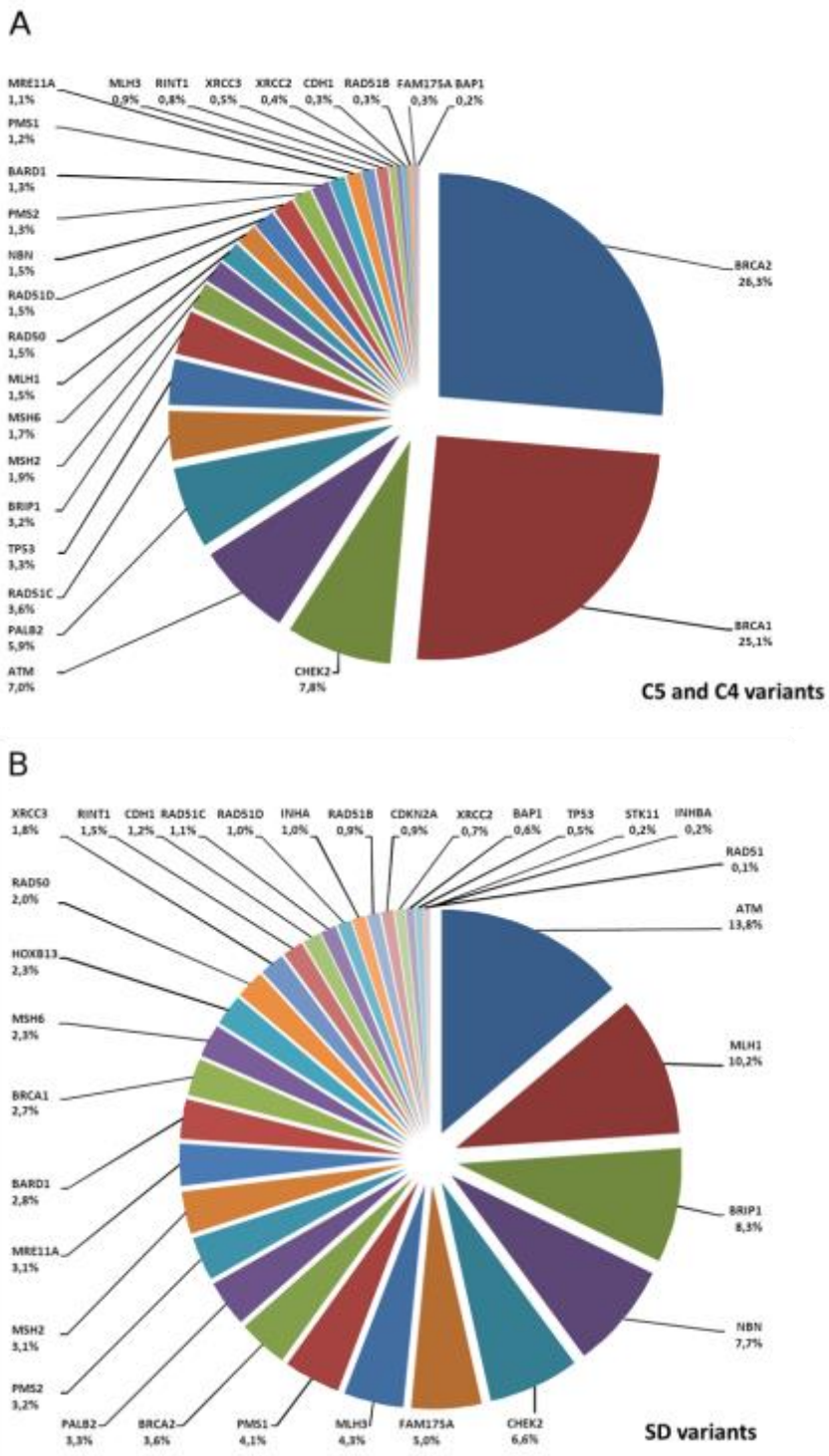


Figure 2: Relative distribution of variants detected with NGS in patients. Percentages were based on the number of times the genes was sequenced, depending on the version of the capture design. (A) Distribution of C5 and C4 variants (B) Distribution of SD variants.

C5 and C4 variants in *MSH2*, *MSH6*, *MLH1* and *PMS2* were found in less than 0.3% of the patients. Incidence of PVs in *MUTYH* was 2.4%, a value in agreement with the frequency of recurrent PVs in the French population[25]. To avoid confusing results due to this high

incidence, the *MUTYH* gene was not considered in the further analyses. Almost half (48.6%) of C5 and C4 variants were detected within other genes than *BRCA1* or *BRCA2*, confirming the genetic heterogeneity in HBOC. Half of the strictly damaging (SD) variants were detected within 6 genes: *ATM*, *MLH1*, *BRIP1*, *NBN CHEK2* and *FAM175A* (Figure 2B). An earlier age at diagnostic of breast cancer was observed in patients who carried a C5 or C4 variants, comparing with non-carrier patients within the same gene, not only for *BRCA1* (mean age = 41.8 years, p<0.001), or for *BRCA2* (mean age = 44.6 years, p<0.05) but also for *CHEK2* (mean age = 42.2 years, p<0.01) and *ATM* (mean age = 42.5 years, p<0.05) (Supplementary Table S3).

Table S3: Distribution of C5 and C4 variations in a subset of genes according to the clinical presentation of the families. C5: pathogenic variants; C4: probably pathogenic variants; TNBC: Triple-Negative Breast Cancer

		C5 and C4 variants	*BRCA1*	*BRCA2*	*PALB2*	*BRIP1*	*CHEK2*	*ATM*	*RAD51B*	*RAD51C*	*RAD51D*	*TP53*
All families	Number of patients	+	166	174	37	21	50	46	2	23	9	21
		-	4240	4235	4136	4387	4324	4362	4385	4286	4002	4159
	Index case with breast cancer (mean age [number of patient])	+	41.8 [n=113]	44.6 [n=136]	45.1 [n=33]	43.5 [n=13]	42.2 [n=43]	42.5 [n=39]	67 [n=1]	39.1 [n=7]	45.2 [n=5]	47.5 [n=15]
		-	46.6 [n=3227]	46.5 [n=3206]	46.5 [n=3107]	46.5 [n=3328]	46.5 [n=3274]	46.5 [n=3302]	46.4 [n=3323]	46.5 [n=3252]	46.6 [n=3011]	46.5 [n=3136]
		p-value	<1e-3	0.05	NS	NS	0.01	0.03	NC	NS	NS	NS
	Index case with ovarian cancer (mean age [number of patient])	+	55.4 [n=46]	59.5 [n=39]	58 [n=2]	63 [n=7]	60.8 [n=4]	57 [n=4]	42 [n=1]	60.1 [n=14]	66.2 [n=4]	59 [n=5]
		-	58.7 [n=702]	58.5 [n=709]	58.6 [n=723]	58.5 [n=741]	58.5 [n=738]	58.5 [n=744]	58.5 [n=744]	58.5 [n=723]	58.6 [n=689]	58.6 [n=715]
		p-value	0.05	NS	NC	NS	NC	NC	NC	NS	NC	NS
Families with at least one ovarian cancer cases	Number of patients	+	78	53	4	9	12	11	1	16	6	6
		-	1158	1184	1175	1227	1211	1225	1230	1194	1123	1178
Families with only ovarian cancers	Number of patients	+	12	14	1	3	3	0	1	7	2	1
		-	407	405	405	416	412	419	415	407	394	407
TNBC index cases	Number of patients	+	54	26	5	1	3	4	1	2	0	3
		-	499	528	525	553	550	550	550	540	510	526
Index cases with bilateral breast cancer	Number of patients	+	20	17	6	3	5	4	0	2	0	2
		-	345	348	341	362	359	361	364	358	334	344
Families with at least one male breast cancer	Number of patients	+	3	16	0	0	0	1	0	0	0	0
		-	133	120	129	136	136	135	136	130	123	130
Male breast cancer index cases	Number of patients	+	0	8	0	0	0	2	0	0	0	0

We then analyzed the association of PVs with HBOC, by determining the OR from the calculated probabilities to carrier a PV in patients and controls. According to our model, the simulated confidence interval derived from OR provides an estimation of a potential association of PVs with HBOC and indirectly with a risk of cancer. First, to quantify a possible bias related to the simulations, we evaluated the enrichment in rare synonymous (MAF<1%) variants, not classified as C5 C4 or SD, as a validation step. Assuming that these synonymous variants should not increase cancer risk, an association of HBOC and rare synonymous variants for a gene would indicate a putative bias. We found indeed an enrichment for *ATM* (OR 1.40[1.17-1.64]), *MLH1* (OR 1.66[1.08-2.38]), *RAD50* (OR 1.66[1.25-2.15]), and *RAD51* (OR 2.33[1.42-3.55]) (Supplementary Table S4).

Table S4 : Evaluation of the association between HBOC and rare synonymous variants comparing the patient population and Exac population

	Patients		**Exac**	**Odd Ratio [IC95%]**
Genes	Number of cases	$P_{patient}$ = probability to be carrier*	$P_{patient}$ = Simulated probability to be carrier*	All Families
ATM	4408	3.95%	2.86%	**1.40[1.17-1.64]**
BAP1	3774	2.01%	1.95%	1.04[0.80-1.31]
BARD1	3667	1.15%	0.91%	1.27[0.89-1.72]
BRCA1	4406	1.32%	1.45%	0.91[0.68-1.18]
BRCA2	4409	3.54%	4.28%	0.82[0.69-0.97]
BRIP1	4408	0.61%	0.60%	1.02[0.65-1.49]
CDH1	3988	3.74%	3.17%	1.19[0.99-1.41]
CDKN2A	1010	0.30%	0.20%	1.51[0.30-3.76]
CHEK2	4374	0.16%	0.15%	1.10[0.41-2.24]
FAM175A	2413	0.21%	0.14%	1.62[0.46-3.69]
HOXB13	2389	0.13%	0.15%	0.84[0.17-2.10]
INHA	2367	0.30%	0.75%	0.40[0.15-0.78]
INHBA	2354	1.06%	1.10%	0.97[0.61-1.41]
MLH1	4340	0.74%	0.45%	**1.66[1.08-2.38]**
MLH3	4395	0.43%	0.68%	0.64[0.38-0.98]
MRE11A	4408	0.98%	1.18%	0.83[0.59-1.12]
MSH2	3800	4.16%	3.33%	**1.26[1.06-1.49]**
MSH6	4408	1.81%	1.96%	0.93[0.72-1.16]
MUTYH	4147	0.65%	0.58%	1.14[0.73-1.66]
NBN	3617	1.33%	1.01%	1.33[0.95-1.77]
PALB2	4173	1.89%	1.54%	1.24[0.96-1.56]
PMS1	4408	0.57%	0.47%	1.22[0.76-1.81]
PMS2	4130	0.87%	1.16%	0.75[0.51-1.05]
PTEN	4408	0.20%	0.18%	1.19[0.51-2.21]
RAD50	4399	1.57%	0.95%	**1.66[1.25-2.15]**
RAD51	4406	0.59%	0.26%	**2.33[1.42-3.55]**
RAD51B	4387	0.41%	0.35%	1.20[0.63-2.00]
RAD51C	4309	0.21%	0.10%	2.10[0.87-4.07]
RAD51D	4011	0.10%	0.14%	0.78[0.18-1.92]
RINT1	2373	0.46%	0.47%	1.00[0.49-1.72]
STK11	2386	1.22%	1.72%	0.71[0.45-1.04]
TP53	4180	0.22%	0.36%	0.61[0.26-1.15]
XRCC2	3729	0.19%	0.096%	2.03[0.74-4.17]
XRCC3	2567	0.39%	1.19%	0.33[0.15-0.60]

Consequently, OR derived from subsequent analysis for these 4 genes have to be interpreted carefully. OR for C5 and C4 variants within *BRCA1* and *BRCA2* were respectively estimated on the entire HBOC series to 13.22[10.01-17.22] and 8.61[6.78-10.82] (Table 1, Figure 3), compatible with knowledge of this two genes[253].

Table 1: Evaluation of the association between HBOC and C5 or C4 variants based on the comparison of the patient population and Exac population

	Patients		Exac controls	Odd Ratio [IC95%]						
Genes	Number of cases	$P_{patient}$ = probability to be carrier	P_{ExAC}= Simulated probability to be carrier	All Families	Families with only ovarian cancers	Families with at least one ovarian cancer case	Families with only breast cancer	Index cases with TNBC	Index cases with bilateral breast cancer	Age of diagnostic <45 years
ATM	4408	1.04%	0.33%	**3.20 [2.14-4.53]**	*NC*	**2.73 [1.29-4.77]**	**3.38 [2.18-4.94]**	2.21 [0.58-4.98]	3.36 [0.88-7.58]	**4.75 [2.88-7.21]**
BAP1	3774	0.026%	0.037%	0.80 [0.017-3.22]	*NC*	2.78 [0.062-11.28]	*NC*	*NC*	9.61 [0.21-38.78]	2.11 [0.05-8.44]
BARD1	3667	0.19%	0.10%	2.00 [0.74-4.10]	*NC*	2.00 [0.23-5.86]	1.99 [0.60-4.40]	**11.27 [3.37-25.01]**	6.92 [0.79-20.32]	2.28 [0.44-5.83]
BRCA1	4406	3.70%	0.29%	**13.22 [10.01-17.22]**	**10.16 [4.96-17.51]**	**22.56 [16.08-30.66]**	**9.72 [7.07-13.07]**	**36.59 [25.05-51.56]**	**18.94 [10.66-30.19]**	**16.05 [11.49-21.87]**
BRCA2	4409	3.95%	0.48%	**8.61 [6.78-10.82]**	**7.26 [3.83-11.99]**	**9.40 [6.65-12.80]**	**8.31 [6.40-0.63]**	**10.34 [6.43-15.41]**	**10.26 [5.74-16.33]**	**10.72 [7.98-14.08]**
BRIP1	4408	0.48%	0.20%	**2.49 [1.42-3.97]**	3.77 [0.74-9.40]	**3.82 [1.66-7.11]**	1.98 [0.96-3.46]	0.94 [0.02-3.56]	4.33 [0.87-10.80]	**2.49 [1.02-4.73]**
CDH1	3988	0.05%	0.011%	6.82 [0.49-28.32]	*NC*	*NC*	9.50 [0.69-38.96]	*NC*	*NC*	9.01 [0.15-43.87]
CHEK2	4374	1.14%	0.69%	**1.67 [1.18-2.27]**	1.06 [0.21-2.58]	1.44 [0.72-2.42]	**1.77 [1.20-2.48]**	0.79 [0.16-1.92]	2.03 [0.64-4.24]	**2.46 [1.57-3.58]**
FAM175A	2413	0.041%	0.12%	0.36 [0.01-1.37]	*NC*	*NC*	0.52 [0.01-1.96]	*NC*	*NC*	*NC*
MLH1	4340	0.23%	0.15%	1.59 [0.71-2.91]	1.66 [0.041-6.23]	1.68 [0.33-4.23]	1.54 [0.58-3.06]	2.51 [0.30-7.23]	3.82 [0.46-10.98]	1.24 [0.25-3.13]
MLH3	4395	0.14%	0.15%	0.93 [0.32-1.91]	*NC*	*NC*	1.29 [0.45-2.65]	1.23 [0.03-4.62]	1.87 [0.046-7.07]	0.81 [0.09-2.32]
MRE11A	4408	0.16%	0.095%	1.74 [0.63-3.57]	**10.57 [2.66-25.15]**	**4.44 [1.33-9.91]**	0.69 [0.078-2.03]	2.00 [0.05-7.60]	*NC*	0.65 [0.02-2.49]
MSH2	3800	0.29%	0.23%	1.34 [0.56-2.62]	2.44 [0.27-7.30]	1.71 [0.42-4.15]	1.19 [0.42-2.53]	0.95 [0.02-3.62]	*NC*	0.97 [0.18-2.54]
MSH6	4408	0.25%	0.10%	**2.50 [1.12-4.67]**	2.39 [0.061-9.15]	3.24 [0.82-7.60]	2.21 [0.81-4.49]	1.80 [0.044-6.79]	2.75 [0.069-10.38]	2.37 [0.60-5.55]
NBN	3617	0.22%	0.17%	1.35 [0.54-2.64]	*NC*	0.59 [0.014-2.24]	1.66 [0.62-3.31]	1.29 [0.032-4.92]	*NC*	1.81 [0.47-4.19]
PALB2	4173	0.89%	0.11%	**8.22 [4.91-13.05]**	2.27 [0.056-8.67]	3.12 [0.80-7.28]	**10.25 [6.03-16.40]**	**8.75 [2.63-19.24]**	**16.19 [5.48-34.22]**	**8.18 [4.00-14.47]**
PMS1	4408	0.16%	0.12%	1.38 [0.51-2.78]	2.08 [0.05-7.88]	0.71 [0.02-2.68]	1.64 [0.56-3.45]	1.58 [0.04-6.03]	2.40 [0.06-9.09]	2.07 [0.53-4.79]
PMS2	4130	0.19%	0.17%	1.16 [0.47-2.24]	1.49 [0.04-5.64]	1.04 [0.12-2.99]	1.21 [0.42-2.50]	1.16 [0.029-4.39]	1.77 [0.04-6.66]	1.90 [0.58-4.12]
RAD50	4399	0.18%	0.34%	0.54 [0.23-1.01]	*NC*	0.24 [0.006-0.90]	0.66 [0.26-1.26]	*NC*	*NC*	1.06 [0.38-2.12]
RAD51B	4387	0.046%	0.044%	1.12 [0.12-3.46]	5.94 [0.14-23.58]	2.00 [0.05-7.95]	0.78 [0.018-3.09]	4.47 [0.10-17.90]	*NC*	*NC*
RAD51C	4309	0.53%	0.12%	**4.54 [2.55-7.48]**	**14.62 [5.39-29.52]**	**11.36 [5.78-19.59]**	1.92 [0.71-3.85]	3.15 [0.37-9.15]	4.75 [0.54-13.89]	**3.10 [1.07-6.46]**
RAD51D	4011	0.22%	0.052%	**5.23 [1.46-13.17]**	**11.84 [1.09-40.00]**	**12.44 [2.94-33.30]**	2.42 [0.36-7.39]	*NC*	*NC*	3.08 [0.28-10.51]
RINT1	2373	0.13%	0.14%	0.90 [0.18-2.26]	*NC*	1.00 [0.025-3.76]	0.86 [0.10-2.49]	1.94 [0.048-7.34]	4.08 [0.09-15.46]	*NC*
TP53	4180	0.43%	0.28%	1.56 [0.85-2.52]	0.88 [0.02-3.32]	1.53 [0.47-3.28]	1.57 [0.79-2.71]	1.38 [0.16-3.96]	2.10 [0.24-6.01]	1.37 [0.48-2.80]
XRCC2	3729	0.054%	0.052%	1.12 [0.12-3.40]	*NC*	*NC*	1.56 [0.17-4.80]	*NC*	*NC*	1.48 [0.03-5.82]
XRCC3	2567	0.039%	0.13%	0.34 [0.01-1.41]	*NC*	*NC*	0.48 [0.01-1.99]	*NC*	*NC*	*NC*

NC : Not Computable (absence of pathogenic variants in a population or a sub-population)

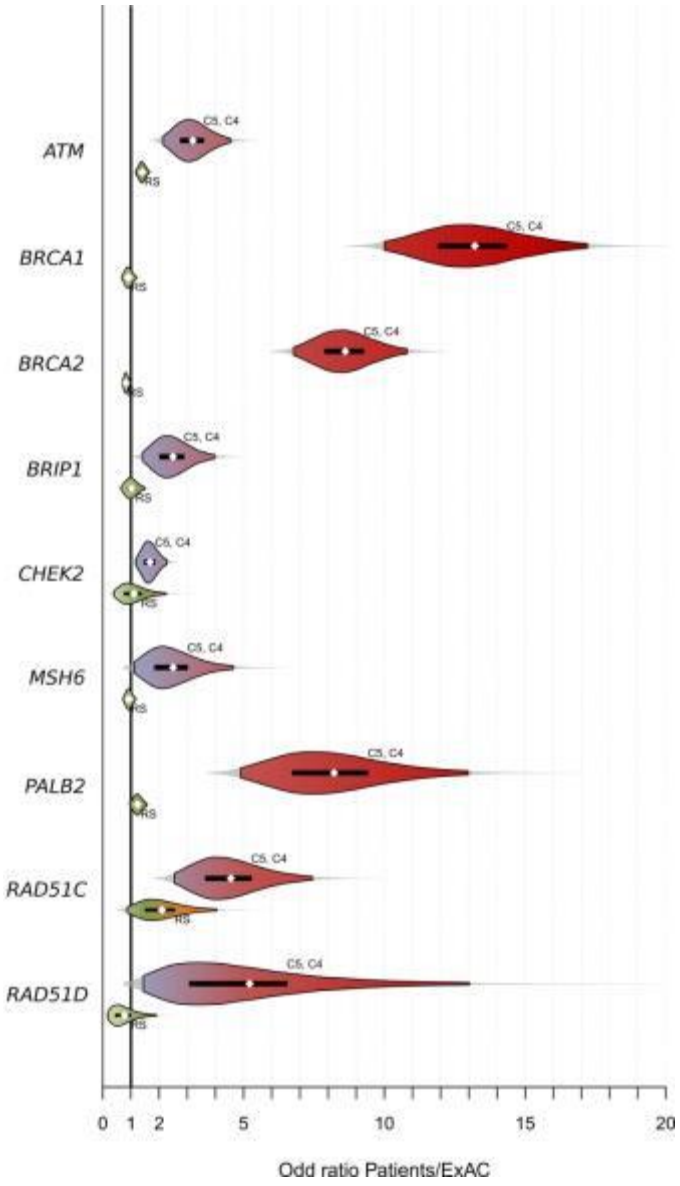


Figure 3: Distribution of OR evaluating the association between HBOC and C5 or C4 variants in the patient population and ExAC population. Violin plots with Gaussian kernel density estimation representing Patient / ExAC odd ratios distribution among 100,000 simulations. Only results comparing C5 and C4 variants with a significant odd ratio greater than 1 are represented. The violin labeled C5, C4 pathogenic variations; the ones labeled RS stand for the Patient / ExAC odd ratios for rare synonymous variations as a base enrichment value (validation step). The violin colored area stand for the 95% confidence interval of the odd ratio. The black box represents the 1st and 3rd quartiles interval. The white point gives an average of the simulation.

In families with only breast cancer, OR conferred by these variants were 9.72[7.07-13.07] and 8.31[6.40-0.63]. When a case of ovarian cancer in HBOC families was reported, *BRCA1* PVs showed a higher level of association than *BRCA2* PVs (OR 22.56[16.08-30.66] vs OR 9.40[6.65-12.80]). If a TNBC index case was reported, PVs within *BRCA1* were associated

with high OR (OR = 36.59[25.05-51.56]) such as PVs within *BRCA2*, but at a lower level (OR=10.34[6.43-15.41]). PVs within *BRCA1* and *BRCA2* were also found associated in families with bilateral breast cancer index case (respectively OR 18.94[10.66-30.19] and 10.26[5.74-16.33]) and with and index case with an early age of diagnosis (respectively OR 16.05[11.49-21.87] and 10.72[7.98-14.08]). OR conferred by *PALB2* PVs was estimated to 8.22[4.91-13.05], a value similar to that found with *BRCA2* PVs. In HBOC families presenting only breast cancers, this OR was estimated to 10.25[6.03-16.40]), but no difference in the probability to carry a *PALB2* PV was detected between HBOC families with at least one ovarian case and controls. *PALB2* PVs were associated with families with the index case present a TNBC, bilateral breast cancer or an early age at diagnosis (OR = 8.75[2.63-19.24]; 16.19[5.48-34.22]; 8.18[4.00-14.47] respectively). We detected a significant association of *RAD51C* and *RAD51D* PVs with HBOC patients (respectively, 4.54[2.55-7.48] and 5.23[1.46-13.17]), but this effect was only maintained when testing families with ovarian cancer (11.36[5.78-19.59] and 12.44[2.94-33.30]). In HBOC families presenting at least one ovarian cancer case, PVs within *BRIP1* were in favor of a mild association (OR 3.82[1.66-7.11]) and surprisingly PVs within *MRE11A* are associated with OR equal to 10.57[2.66-25.15]). PVs in *CHEK2* and *BRIP1* were associated with HBOC with low OR (respectively 1.67[1.18-2.27] and 2.49[1.42-3.97]). We found also a slight association (OR 2.50[1.12-4.67]) of PV within *MSH6* and HBOC as suggested previously[254]. More unexpected, PV in *BARD1* were associated only with HBOC family where the index case presented a TNBC (OR = 11.27[3.37-25.01]).

We then performed the same analysis after including the SD variants. This extension dramatically reduced the association or abolished it (supplementary Table S5). When the analysis was restricted to SD variants (Table S6), we detected a weak association of SD variants of *BRCA1*, *BRCA2*, *BRIP1*, *MRE11A* (significant OR from 1.58 to 2.27). Even more

unexpected, was the moderated association found between SD variant in *BARD1* and HBOC (OR = 3.31[1.93-5.22]).

Table S5: Evaluation of the association between HBOC and C5, C4 or SD variants comparing the patient population and Exac population

	Patients		**Exac**	**Odd Ratio [IC95%]**
Genes	Number of cases	$P_{patient}$ = probability to be carrier *	$P_{patient}$ = Simulated probability to be carrier *	All Families
ATM	4408	4.11%	2.77%	**1.51[1.27-1.77]**
BAP1	3774	0.16%	0.18%	0.92[0.32-1.90]
BARD1	3667	0.82%	0.29%	**2.84[1.80-4.20]**
BRCA1	4406	4.29%	0.56%	**8.06[6.45-9.96]**
BRCA2	4409	4.74%	0.93%	**5.35[4.41-6.43]**
BRIP1	4408	2.31%	1.41%	**1.66[1.32-2.06]**
CDH1	3988	0.33%	0.39%	0.85[0.44-1.42]
CDKN2A	1010	0.20%	0.12%	1.78[0.21-5.16]
CHEK2	4374	2.61%	2.17%	1.21[0.97-1.48]
FAM175A	2413	1.16%	1.28%	0.91[0.59-1.30]
HOXB13	2389	0.50%	1.24%	0.40[0.21-0.67]
INHA	2367	0.21%	0.28%	0.78[0.23-1.72]
INHBA	2354	0.042%	0.29%	0.15[0.0039-0.56]
MLH1	4340	2.49%	1.89%	**1.32[1.06-1.62]**
MLH3	4395	1.09%	1.41%	0.78[0.56-1.03]
MRE11A	4408	0.84%	0.53%	**1.60[1.08-2.25]**
MSH2	3800	0.97%	0.91%	1.08[0.73-1.52]
MSH6	4408	0.77%	0.92%	0.84[0.57-1.17]
MUTYH	4147	2.89%	2.01%	**1.45[1.18-1.76]**
NBN	3617	1.94%	2.27%	0.85[0.65-1.08]
PALB2	4173	1.63%	0.73%	**2.25[1.68-2.93]**
PMS1	4408	1.07%	0.89%	1.20[0.86-1.61]
PMS2	4130	0.90%	1.32%	0.68[0.47-0.94]
RAD50	4399	0.64%	1.02%	0.62[0.41-0.89]
RAD51	4406	0.023%	0.092%	0.26[0.0063-0.98]
RAD51B	4387	0.25%	0.19%	1.36[0.64-2.41]
RAD51C	4309	0.79%	0.37%	**2.16[1.42-3.11]**
RAD51D	4011	0.45%	0.36%	1.27[0.71-2.02]
RINT1	2373	0.46%	0.38%	1.23[0.60-2.13]
STK11	2386	0.042%	0.16%	0.30[0.0067-1.23]
TP53	4180	0.55%	0.39%	1.42[0.85-2.17]
XRCC2	3729	0.21%	0.15%	1.49[0.60-2.87]
XRCC3	2567	0.43%	0.57%	0.78[0.35-1.45]

Table S6: Evaluation of the association between HBOC and SD variants comparing the patient population and Exac population

	Patients		Exac controls	Odd Ratio [IC95%]						
Genes	Number of cases	$P_{patient}$ = proba-bility to be carrier	P_{ExAC}= Simulated probability to be carrier	All Families	Families with only ovarian cancers	Families with at least one ovarian cancer case	Families with only breast cancer	Index cases with TNBC	Index cases with bilateral breast cancer	Age of diagnostic <45 years
ATM	4408	3.06%	2.44%	**1.26 [1.04-1.52]**	0.98 [0.47-1.70]	1.03 [0.69-1.44]	**1.36 [1.09-1.65]**	0.66 [0.30-1.17]	0.78 [0.31-1.48]	**1.45 [1.09-1.87]**
BAP1	3774	0.13%	0.14%	0.98 [0.30-2.14]	3.96 [0.46-11.47]	2.06 [0.40-5.20]	0.55 [0.064-1.59]	NC	NC	0.52 [0.01-1.97]
BARD1	3667	0.63%	0.19%	**3.31 [1.93-5.22]**	4.25 [0.84-10.58]	**4.58 [1.97-8.49]**	**2.82 [1.43-4.79]**	3.39 [0.68-8.49]	**5.26 [1.05-13.2]**	1.92 [0.60-4.11]
BRCA1	4406	0.59%	0.26%	**2.27 [1.40-3.46]**	1.83 [0.22-5.21]	1.87 [0.67-3.76]	**2.43 [1.40-3.84]**	**3.50 [1.10-7.40]**	3.18 [0.64-7.88]	**2.52 [1.20-4.41]**
BRCA2	4409	0.82%	0.45%	**1.84 [1.23-2.61]**	2.16 [0.57-4.82]	**3.12 [1.75-4.94]**	1.35 [0.78-2.09]	1.63 [0.44-3.64]	3.11 [0.99-6.53]	1.88 [0.99-3.07]
BRIP1	4408	1.84%	1.21%	**1.53 [1.18-1.93]**	1.19 [0.43-2.34]	**2.03 [1.34-2.88]**	1.33 [0.97-1.76]	**2.12 [1.14-3.43]**	1.60 [0.63-3.03]	**1.58 [1.06-2.22]**
CDH1	3988	0.28%	0.38%	0.74 [0.36-1.28]	2.07 [0.42-5.10]	1.19 [0.37-2.48]	0.56 [0.20-1.12]	0.53 [0.013-1.97]	NC	0.54 [0.11-1.31]
CDKN2A	1010	0.20%	0.098%	2.12 [0.25-6.23]	NC	NC	3.06 [0.35-9.00]	NC	NC	2.95 [0.075-11.31]
CHEK2	4374	1.49%	1.49%	1.00 [0.75-1.29]	0.65 [0.17-1.44]	0.77 [0.41-1.23]	1.09 [0.80-1.45]	0.12 [0.003-0.45]	1.31 [0.52-2.46]	1.25 [0.83-1.77]
FAM175A	2413	1.16%	1.16%	1.00 [0.65-1.43]	1.27 [0.34-2.82]	1.32 [0.65-2.23]	0.87 [0.50-1.34]	0.45 [0.055-1.28]	0.48 [0.012-1.77]	0.71 [0.28-1.34]
HOXB13	2389	0.50%	1.21%	0.41 [0.21-0.69]	0.30 [0.008-1.13]	0.57 [0.18-1.18]	0.35 [0.14-0.65]	0.22 [0.006-0.82]	NC	0.098 [0.0024-0.36]
INHA	2367	0.21%	0.28%	0.80 [0.23-1.77]	1.42 [0.034-5.43]	0.54 [0.013-2.03]	0.91 [0.23-2.15]	NC	NC	0.46 [0.011-1.76]
INHBA	2354	0.042%	0.29%	0.15 [0.01-0.56]	NC	NC	0.21 [0.01-0.79]	NC	2.03 [0.05-7.62]	NC
MLH1	4340	2.26%	1.75%	**1.30 [1.03-1.61]**	0.69 [0.22-1.43]	1.22 [0.79-1.76]	**1.33 [1.02-1.69]**	1.04 [0.49-1.80]	1.44 [0.65-2.56]	1.32 [0.92-1.81]
MLH3	4395	0.96%	1.26%	0.76 [0.54-1.03]	0.38 [0.045-1.06]	0.51 [0.22-0.93]	0.86 [0.58-1.19]	0.72 [0.23-1.50]	1.32 [0.48-2.61]	0.57 [0.29-0.95]
MRE11A	4408	0.68%	0.44%	**1.58 [1.02-2.30]**	1.67 [0.34-4.10]	1.89 [0.88-3.30]	1.47 [0.87-2.25]	0.84 [0.10-2.36]	1.27 [0.15-3.60]	1.79 [0.93-2.99]
MSH2	3800	0.68%	0.68%	1.01 [0.64-1.48]	1.56 [0.42-3.49]	1.23 [0.55-2.20]	0.92 [0.52-1.45]	0.91 [0.19-2.23]	1.90 [0.50-4.27]	1.04 [0.49-1.81]
MSH6	4408	0.52%	0.82%	0.64 [0.40-0.95]	0.88 [0.18-2.15]	0.59 [0.22-1.17]	0.66 [0.38-1.02]	0.44 [0.05-1.24]	NC	0.58 [0.25-1.06]
NBN	3617	1.71%	2.11%	0.81 [0.61-1.04]	0.52 [0.14-1.14]	0.73 [0.41-1.13]	0.85 [0.61-1.12]	0.90 [0.41-1.58]	0.81 [0.26-1.67]	0.88 [0.56-1.27]
PALB2	4173	0.74%	0.62%	1.20 [0.79-1.72]	2.42 [0.87-4.78]	1.52 [0.74-2.60]	1.08 [0.64-1.65]	1.53 [0.49-3.19]	0.94 [0.11-2.61]	1.02 [0.48-1.78]
PMS1	4408	0.91%	0.78%	1.18 [0.82-1.62]	0.93 [0.19-2.25]	1.05 [0.49-1.82]	1.23 [0.81-1.75]	1.41 [0.51-2.79]	1.43 [0.38-3.18]	1.00 [0.52-1.64]
PMS2	4130	0.70%	1.15%	0.61 [0.40-0.87]	0.65 [0.13-1.59]	0.38 [0.12-0.78]	0.70 [0.44-1.03]	0.33 [0.04-0.94]	1.03 [0.28-2.29]	0.77 [0.42-1.25]
RAD50	4399	0.45%	0.69%	0.67 [0.40-1.01]	0.70 [0.083-1.96]	0.83 [0.33-1.58]	0.60 [0.31-0.99]	0.53 [0.06-1.48]	0.80 [0.10-2.26]	0.61 [0.24-1.15]
RAD51	4406	0.023%	0.092%	0.26 [0.006-0.98]	2.72 [0.07-10.48]	0.91 [0.02-3.45]	NC	NC	NC	NC
RAD51B	4387	0.21%	0.14%	1.47 [0.62-2.75]	NC	0.58 [0.014-2.19]	1.81 [0.73-3.51]	2.60 [0.30-7.47]	NC	1.27 [0.25-3.18]
RAD51C	4309	0.26%	0.25%	1.04 [0.50-1.83]	**3.98 [1.04-9.00]**	2.03 [0.71-4.09]	0.66 [0.21-1.39]	2.27 [0.46-5.60]	NC	0.99 [0.26-2.24]
RAD51D	4011	0.35%	0.31%	1.15 [0.60-1.92]	1.68 [0.20-4.77]	1.76 [0.62-3.54]	0.92 [0.38-1.71]	1.30 [0.15-3.67]	0.99 [0.03-3.74]	1.09 [0.35-2.30]
RINT1	2373	0.34%	0.24%	1.45 [0.60-2.72]	NC	1.81 [0.36-4.47]	1.30 [0.41-2.74]	1.17 [0.03-4.40]	2.45 [0.06-9.22]	1.54 [0.31-3.82]
STK11	2386	0.042%	0.16%	0.31 [0.007-1.23]	NC	NC	0.43 [0.01-1.74]	NC	NC	NC
TP53	4180	0.12%	0.11%	1.11 [0.33-2.44]	NC	NC	1.55 [0.47-3.41]	NC	NC	1.75 [0.34-4.48]
XRCC2	3729	0.16%	0.096%	1.75 [0.58-3.71]	2.93 [0.07-11.26]	2.05 [0.23-5.99]	1.63 [0.42-3.83]	2.31 [0.06-8.79]	3.61 [0.09-13.9]	0.78 [0.019-2.96]
XRCC3	2567	0.39%	0.43%	0.94 [0.40-1.80]	0.86 [0.021-3.30]	0.96 [0.18-2.48]	0.93 [0.33-1.92]	2.08 [0.39-5.35]	3.60 [0.67-9.29]	0.78 [0.15-2.00]

NC : Not Computable (Patients and Methods)

In order to consolidate simulations performed with the ExAC controls, data from FREX controls were used to perform CAST and WST burden tests (Table 2). Tests showed sufficient power to confirm association between C5 or C4 *BRCA1*, *BRCA2*, *PALB2* and *RAD51C* PVs with HBOC. Absence of PV in the FREX dataset due to the limited size of our sample hampered to calculate an OR in *PALB2*, *RAD51C* and *RAD51D*. Indeed no were

detected. No other significant association was detected using the FREX data testing C5 and C4 variants.

Table 2: Concordant significant association detected by CAST test and WST test between patients population and FREX population

	C5 and C4 variants			C5, C4 or SD variants			SD variants		
Genes	OddRatio [IC95%]	CAST pvalue	WSS pvalue	OddRatio [IC95%]	CAST pvalue	WST pvalue	OddRatio [IC95%]	CAST pvalue	WST pvalue
BRCA1	21.44 [3.00 - 153.40]	**<1e-6**	**<1e-6**	25.01 [3.50 - 178.81]	**<1e-6**	**<1e-6**			
BRCA2	7.61 [2.42 - 23.92]	**<1e-3**	**<1e-6**	4.59 [2.03 - 10.38]	**<1e-3**	**<1e-6**			
FAM175A				*NC*	**<0.01**	**<1e-6**	*NC*	**<0.01**	**<1e-6**
MLH3				*NC*	**<0.01**	**<1e-6**	*NC*	**0.01**	**<0.01**
PALB2	*NC*	**0.02**	**<1e-6**						
PMS1				*NC*	**<0.01**	***<1e-6***	*NC*	**0.02**	**0.02**
RAD51C	*NC*	**0.1**	**0.05**						

NC : Not Computable (absence of pathogenic variants in a population or a sub-population)

DISCUSSION

We showed in this study that it is feasible to re-evaluate carefully the risks induced by pathogenic variants within genes sequenced by diagnostic laboratories, using heterogeneous control datasets, including recalibrated data from the ExAC consortium, 1000 genome project and appropriate controls such as the FREX dataset. Using a precautious workflow of analysis based on the control of stratification, on the accuracy of the clinical data and on the quality of NGS data, we demonstrated an enrichment of PVs in patients and we could replicate the data by performing burden tests using the appropriate FREX control dataset (Table 2). This confirms that it is possible, with these cautions, to evaluate disease risk by comparing allelic frequencies observed in large series of patients to that deduced from massive reference datasets, such as ExAC and now GNOMAD[246]. Due to our recruitment of selected families with high genetic risks of cancer, the evaluation of breast cancer or ovarian cancer risk for a carrier of PV, without considering a familial model might have been overestimated[236]. Odd Ratio defined here described the level of association of PVs with HBOC and the comparison of significant OR may estimate the level of risk of cancer induced by PVs of each gene, because some of them are today perfectly known (i.e. *BRCA1* and *BRCA2*).

The OR that we calculated from this large series of patients validates the integration, besides *BRCA1* and *BRCA2*, of *PALB2*, *RAD51C*, *RAD51D* in the molecular diagnostic of HBOC. In contrast, the lower OR values observed for the other genes suggest that they could be involved in an oligogenic determinism of breast and ovarian cancers (Table 1). We confirmed that *PALB2* PVs are associated to a high risk of breast cancer in the same magnitude than *BRCA2* PVs[245]. This level of breast cancer risk justifies the integration of *PALB2* in genetic testing of patients referred for HBOC and the use of *PALB2* PVs in genetic counseling. This confirms that the clinical management of *PALB2* mutation carriers should be the same than that recommended for *BRCA2* mutation carriers, restricted to breast cancer risk according

NCCN guidelines. Indeed in this series, OR found in families with only breast cancer cases, conferred by PVs detected within these 2 genes, were equivalent, but no significant association of *PALB2* PVs was found with families with at least one ovarian cancer case. Interestingly, PVs within *PALB2* were found in this study associated with HBOC families with index case with TNBC, bilateral breast cancer or a younger age of diagnosis. This observation prompts us, in order to prioritize re-analyzes of index cases without detectable *BRCA1* or *BRCA2* mutations, to screen for *PABL2* alterations, first, index cases with at least TNBC, bilateral breast cancer or with age at diagnosis below 45 and to index cases with breast cancer with first-degree relatives presenting the same criteria. OR conferred in families with ovarian cancer cases by PVs detected within the *RAD51C* and *RAD51D* were found comparable to those conferred by *BRCA2* PVs. This confirm that it is justified to propose prophylactic salpingo-oophorectomy to menopausal carrier women[128,130,255,256].

Our study also confirmed the moderated cancer risks induced by *CHEK2* and *BRIP1* PVs. Considering the OR for breast cancer conferred by *CHEK2* PVs (1.77[1.20-2.48]), and for ovarian cancer induced by *BRIP1* PVs 3.82[1.66-7.11], it appears indeed inappropriate to base clinical management on these variants[67,257,258]. Our results are consistent with previous publication describing the absence of significant risk of breast cancer and moderate risk of ovarian cancer with PV of *BRIP1*[68,259,260]. Interestingly, PV in *BARD1* was found associated with ovarian cancer and bilateral breast cancer (Table 1). Such effect was already suspected and combined with our data, this could suggest that PV of *BARD1* could be responsible for risk of cancer, maybe high risk, but restricted to a particular phenotype[68]. Generally, refinement of level of risk should take into account the precise tumor phenotype. More effort and larger population studies are needed to collect greater data set to precise risk in subgroup of tumor. Also, our results suggesting a high ovarian cancer risk in *MRE11A* PV mutation carriers need to be confirmed by other studies. Interestingly, a previous study had suggested

modification of the ovarian risk cancer with rare haplotypes of *MRE11A*[261]. Lastly, we did not find any association of breast cancer with *TP53* variants whereas PVs within in this gene clearly confer high risk of early-onset of breast cancer. This apparently unexpected result is coherent with previous studies which did not detect significant association or report association with large OR confident interval, from about 2 to 100[244,246]. The low frequency of *TP53* PVs could explain this result, but we hypothesized that a bias of our recruitment might explain this unexpected result. Indeed early-onset breast cancer cases linked due to *TP53* PVs have a high probability to present Li Fraumeni syndrome (LFS) and these patients were not referred to our laboratory for HBOC, since the LFS diagnosis was strongly suspected or has already been established.

We also confirmed the moderate risk conferred by PV detected within *ATM*. Nevertheless, the fact that we detected an association with rare synonymous variants (Table S5)[67,137] limits the validity of this observation. Similarly, we found an effect with synonymous variants of *MSH2*, *RAD50*, and *RAD51*. These effects could be induced by a stratification bias or could be due to a true signal of association underlining low level of risk. According to odd-ratio found, PV within *CHEK2*, *ATM* and *BRIP1*, but perhaps also variants within *FAM175A*, *MLH3*, *PMS1* (Table 3), are in favor of moderate risks. These results, and more generally moderate or low risk factors, seem individually not useful for care management but they would have more clinical utility if they will be integrated in an oligogenic risk evaluation model which would improve accuracy of polygenic risk score[262].

The estimation of risk in this study confirms the relevance for HBOC diagnosis of the panel gene sequencing strategy including not only *BRCA1* and *BRCA2* but also *PALB2*, *RAD51C* and *RAD51D*. Based on this study, we consider that the refinement of risk in the "other genes" suspected to be involved in HBOC does not allow their routine analysis in the context

of medical laboratories but prefigures the development of oligogenic risk scores, not only based on frequent variants but also including rare PV variants of these genes.

ACKNOWLEDGEMENTS

The authors would like to thank the high-throughput sequencing platform of Basse-Normandie SéSAME (Sequencing for Health, Agronomy, the Sea and the Environment) for technological support. The authors are grateful to Camille Charbonnier- Le Clezio for critical review of the manuscript and the Cancéropole Nord-Ouest and the Institut National du Cancer (INCa) for funding

NOTES

No conflicts of interest are declared.

D. Discussion des résultats

Les résultats de cette étude démontrent que l'étude de variants pathogènes rares, et l'utilisation de grandes bases de données de séquençage telles qu'ExAC en tant que témoins, semblent être pertinente pour estimer les risques associés au syndrome HBOC. Les risques estimés ici pour les variants pathogènes des gènes de *BRCA1* et *BRCA2*, mais aussi pour *PALB2*, *ATM*, ou encore *CHEK2* sont similaire à ceux décrits dans la littérature[246,263], ce qui suggère que la méthode statistique utilisée est cohérente.

Les tests d'agrégation réalisés par les méthodes statistiques classiques (Burden tests) mettent en évidence la problématique de la disponibilité d'une population de témoins appariés suffisamment importante dans cette étude (574 témoins de la population FREX). Les effets d'association ne sont significatifs (malgré des intervalles de confiance très importants liés au manque de puissance de l'analyse) que sur *BRCA1*, *BRCA2*, *PALB2* et *RAD51C*, ce qui confirme néanmoins l'intérêt dans le diagnostic de *PALB2* et *RAD51C*. L'utilisation de bases de données compilant les résultats du séquençage de plusieurs milliers d'individus (telles qu'ExAC) en tant que population de témoins permet d'estomper cette difficulté, dans le cadre d'une utilisation prudente. Mais l'utilisation des données d'ExAC expose à des biais de stratification évidents. Ici, nous avons sélectionné les informations de séquençage en provenance de la population européenne non-finlandaise (NFE) afin de se rapprocher un maximum de notre population d'étude, et en utilisant le jeu de données excluant les informations de séquençage issues du TCGA, afin d'éviter tout biais de recrutement dans la cohorte témoin. Dans la volonté de contraindre au maximum le biais de stratification, l'origine des individus a été déterminée par une analyse en composante principale (ACP) comparant la population de cas, la population FREX et 1000 genomes Project.

Aussi, la sélection et la classification des variants étudiés apparaissent cruciales pour ne pas introduire de biais dans l'analyse. En effet dans cette étude, les variants étudiées pour tester un effet d'association ont été sélectionnés par 2 analyses bio-informatiques différentes. Les variants des cas ont été identifiées par la suite logicielle CASAVA, et par le pipeline d'analyse BWA-GATK pour les données FREX et ExAC. La sélection des variations sur la base de leur qualité de séquençage a été faite de la manière la plus rigoureuse possible, afin de supprimer un maximum de faux-positifs. GATK utilise le VQSLOD (cf. Introduction, § VII.A.6.a), qui estime la qualité du variant détecté en prenant en considération l'ensemble des populations étudiées. CASAVA à l'inverse réalise l'appel de variants patient par patient et le calcul de ses scores qualité est indépendant des autres individus étudiés. Notre étude s'expose donc à des divergences dans la sélection des variants et à d'éventuels biais, néanmoins nous avons pris le parti de favoriser la sensibilité de sélection des variants chez les témoins et la spécificité de la sélection chez les cas afin que les biais éventuels soient tournés vers les témoins évitant ainsi la mise en évidence d'effets confondants. Egalement le test élaboré ici a une tendance à surestimer la probabilité qu'un individu de la cohorte ExAC porte un variant rare pathogène par rapport à un patient de la cohorte du LBGC, minimisant ainsi les risques de mettre en évidence un faux effet d'association. Les données individuelles ne sont pas mises à disposition dans ExAC, de sorte que nous avons émis l'hypothèse d'indépendance des variants rares pathogènes (i.e. sans déséquilibre de liaison). De plus la fréquence en population des variants considérées est tellement faible (singletons pour une grande majorité d'entre eux) que le déséquilibre serait impossible à estimer. Ainsi, si certains variants pathogènes rares étaient en déséquilibre de liaison, la probabilité qu'un individu soit porteur d'un variant pathogène serait inférieure à notre estimation puisque l'ensemble des variants seraient portés par moins d'individus.

Avec ces précautions, les risques estimés pour les gènes rendus au titre du diagnostic, notamment *BRCA1* et *BRCA2*, sont compatibles avec ceux de la littérature existante validant indirectement notre démarche[236].

Les résultats présentés dans cette étude démontrent ainsi qu'il est possible d'estimer les risques associés à la présence de variants pathogènes dans les gènes impliqués dans le syndrome de prédisposition au cancer du sein et de l'ovaire grâce à une méthodologie rigoureuse et l'utilisation de nouveaux outils statistiques et l'exploitation de bases de données mises en place par les consortiums internationaux. Des approches émettant les mêmes hypothèses de travail que les nôtres, pouvant employer une méthodologie autre (Test exact de Fisher), ont également été utilisées par d'autres équipes, estimant des risques similaires[246]. Notre laboratoire avait précédemment démontré qu'une extension de l'exploration génétique du syndrome grâce au panel de gènes utilisé permettait de retrouver des variations délétères chez plus de 15 % des patients suspectés d'avoir le syndrome[57]. Aujourd'hui nous nous confirmons l'intérêt pour le diagnostic de rechercher les variant de *PALB2*, *RAD51C* et *RAD51D*. L'élargissement du spectre d'analyse contribue donc à expliquer l'hérédité manquante dans le syndrome HBOC. Les risques faibles ou modérés retrouvés pour les gènes du panel autres que *BRCA1*, *BRCA2, PALB2, RAD51C* ou *RAD51D*, peuvent en revanche suggérer un mécanisme oligogénique, comprenant des gènes et des évènements non examinés lors de l'étude réalisée.

II / Exploration de l'implication des néo-mutations en mosaïque dans le syndrome de prédisposition au cancer du sein et de l'ovaire

Les analyses présentées lors de l'étude précédente ont donc permis d'estimer les risques associés à la présence de variants pathogènes de 34 gènes impliqués dans le syndrome de prédisposition au cancer du sein et de l'ovaire (HBOC), ou suspectés de l'être. Il parait assez clair que l'extension du diagnostic moléculaire à d'autres gènes permet d'expliquer une partie de l'hérédité manquante mais de manière faiblement contributive (le pourcentage de détection de variants participant à la définition du risque est doublé), de sorte que d'autres mécanismes génétiques induisant une prédisposition au cancer doivent être explorés. Comme piste possible nous avons choisi d'évaluer la proportion de néo-mutations en mosaïque possiblement responsables d'une augmentation du risque de cancer du sein et de l'ovaire. Les variants en mosaïque sont possiblement détectables par séquençage à haut-débit mais impliquent l'emploi d'une méthodologie bioinformatique dédiée sensible mais surtout très spécifique si cette technique doit pouvoir détecter des variants très faiblement représentés (ratio d'allèle variant inférieur à 5%). En effet des variants très faiblement représentés seront très difficiles à confirmer par une autre méthode. Dans l'étude précédente, l'implication des néo-mutations en mosaïque n'a pu été évaluée, la méthodologie utilisée n'étant pas suffisamment sensible pour détecter ce type d'évènements. Cependant, de rares cas de variants en mosaïque ont déjà été décrits dans la littérature, nous laissant supposer que la recherche systématique des variants en mosaïque pourrait expliquer certaines histoires de cancer évocatrices d'une prédisposition[157,158]. Nous proposons ici une étude systématique d'une population importante de patientes afin de déterminer l'incidence des variants en mosaïques dans les gènes de l'étude précédente. Une sous-population de patients issus de

l'étude précédente a été analysée de nouveau avec un pipeline bioinformatique dédié à cette étude, et, puisque la mise en évidence de variants en mosaïque nécessite l'utilisation d'outils de détection particulièrement sensibles et spécifiques, il a été développé dans le cadre de cette thèse un logiciel nouveau : OutLyzer. OutLyzer[264] est un *variant-caller* qui implémente une estimation statistique du bruit de fond de séquençage pour mettre en évidence les variants faiblement représentés. Il faut noter que la détection des variants en mosaïque est la même problématique que celle rencontrée lors du séquençage de tumeurs, qui nécessite de détecter des variants parmi plusieurs populations clonales différentes ou « diluées » par de l'ADN de tissu sain contaminant. Dans un objectif d'application clinique immédiate, OutLyzer a été validé après séquençage de 130 tumeurs parfaitement qualifiées et pouvant servir de témoins portant des variants faiblement à très faiblement représentés (confirmés par Sanger ou Pyroséquençage). Un panel de 22 gènes étudiés en tant que biomarqueurs utiles à la prise en charge thérapeutique des patients atteint de cancer ont été séquencés. La sensibilité et la spécificité d'OutLyzer a été comparée à une sélection de logiciels existants. Nous avons choisi pour la comparaison HaplotypeCaller, référence pour la recherche de variants constitutionnels. Egalement, Varscan2 et Lofreq2 ont été choisis pour leur capacité à détecter les variants faiblement représentés.

Après validation du logiciel (cf. Résultats, § II.A correspondant à l'article publié), une analyse été réalisée en utilisant ce logiciel qui a été mis à jour pour améliorer ses performances, afin d'évaluer l'incidence des variants en mosaïque à partir d'une population de 1750 patientes atteintes d'un cancer du sein ou de l'ovaire (cf. Résultats, § II.B).

A. outLyzer : Logiciel de détection des variants génétiques dans les milieux hétérogènes à partir des données de séquençage à haut-débit

Research Paper

OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice

Etienne Muller[1,2], Nicolas Goardon[1], Baptiste Brault[1], Antoine Rousselin[1], Germain Paimparay[1], Angelina Legros[1], Robin Fouillet[1], Olivia Bruet[1], Aurore Tranchant[1], Florian Domin[1], Chankannira San[1], Céline Quesnelle[1], Thierry Frebourg[2,3,4], Agathe Ricou[1], Sophie Krieger[1,2,5], Dominique Vaur[1,2], Laurent Castera[1,2]

[1]Department of Cancer Biology and Genetics, CCC François Baclesse, Genomic and Personalized Medicine in Cancer and Neurological Disorders Unit, Caen, France

[2]Inserm U1079, Genomic and Personalized Medicine in Cancer and Neurological Disorders Unit, Rouen, France

[3]Genetic Department, Rouen University Hospital, Genomic and Personalized Medicine in Cancer and Neurological Disorders Unit, Rouen, France

[4]Rouen University, France

[5]Caen University, France

Correspondence to: Laurent Castéra, **email:** l.castera@baclesse.unicancer.fr

Keywords: *variant-caller, somatic mutation, bioinformatics, oncology, precision medicine*

Received: *May 26, 2016* ***Accepted:*** *October 11, 2016* ***Published:*** *November 04, 2016*

ABSTRACT

Highlighting tumoral mutations is a key step in oncology for personalizing care. Considering the genetic heterogeneity in a tumor, software used for detecting mutations should clearly distinguish real tumor events of interest that could be predictive markers for personalized medicine from false positives. OutLyzer is a new variant-caller designed for the specific and sensitive detection of mutations for research and diagnostic purposes. It is based on statistic and local evaluation of sequencing background noise to highlight potential true positive variants. 130 previously genotyped patients were sequenced after enrichment by capturing the exons of 22 genes. Sequencing data were analyzed by HaplotypeCaller, LofreqStar, Varscan2 and OutLyzer. OutLyzer had the best sensitivity and specificity with a fixed limit of detection for all tools of 1% for SNVs and 2% for Indels. OutLyzer is a useful tool for detecting mutations of interest in tumors including low allele-frequency mutations, and could be adopted in standard practice for delivering targeted therapies in cancer treatment.

INTRODUCTION

The advent of Next Generation Sequencing (NGS) during the last decade has been a true revolution both in research and diagnostic laboratories. NGS allows DNA to be read with greater speed and convenience than ever before. The most recent sequencing devices produce the equivalent of several entire genomes in a few hours, while the sequencing of the first human genome took about ten years [1]. NGS is now used daily in clinical laboratories, in many fields of medicine, and particularly in oncology and the diagnosis of cancer predisposition [2]. The molecular typing of tumors in the context of precision medicine could benefit from this technology. For example, somatic mutations in the *EGFR* gene are now commonly sought in lung cancer [3], *KRAS* mutations in colorectal cancer [4] and *BRCA1* and *BRCA2* in ovarian cancer [5]. Thanks to the panel gene sequencing approach, NGS technologies optimize and simplify laboratory processes to the extent that it is today possible to sequence the majority of medical targets of interest in one experiment, regardless of tumor type. When associated with dedicated bioinformatics tools, NGS can explore tumoral heterogeneity and characterize intra-tumoral clonal subpopulations [6]. The identification of sub-clones possibly carrying sensitive or resistance mutations to targeted therapies appears to be a key challenge for patient support in the context of personalized medicine.

The fine characterization of the mutation profile of a tumor with NGS for clinical purposes is a challenge. Diagnostic laboratories therefore have to meet a number of constraints to satisfy the high level of sensitivity and specificity needed for diagnostic tests. Tumoral tissue may include many cell subpopulations, so cells carrying a mutation of interest may be poorly represented in a tumor sample (i.e. low allele-frequency tumor mutations). Moreover, tumor cells can be harvested together with

healthy tissue, thereby reducing the number of mutated alleles by dilution. In view of these constraints, a highly sensitive process is required to avoid false negative results. The analysis of sequencing can itself be misleading owing to a PCR reaction bias during sample preparation [7] or to sequencer reading errors [8]. Low level mutations may also be difficult to distinguish from a noise background generated by such technical limitations. Consequently, ensuring a high specificity is critical in diagnostic testing to avoid false positives.

Dedicated bioinformatics tools can help to ensure good sensitivity and specificity. Detection of mutations is a key step in bioinformatics analysis and is performed by variant-calling software. An example of the numerous variant-callers currently available is HaplotypeCaller in the GATK suite [9]. It is a reference in genotyping germline genomes but its sensitivity can dramatically decrease when faced with low level mutations. Others like Varscan2 [10] and LofreqStar [11] have been designed especially for tumor sample analysis and the detection of low level mutations but are efficient mainly for comparing matched healthy and tumor samples. In many diagnostic laboratories a matched healthy sample is not available for analysis owing to ethical considerations, organizational difficulties or legal constraints. Furthermore, even if it were to be available, sequencing would be twice as expansive owing to the need to sequence two different samples for the same patient.

Here we present OutLyzer, a new variant-caller which was validated in a local diagnostic setting to fit ISO15189 quality requirements. It has been designed for non-matched tumoral sample analysis and it is based on statistic and local evaluation of sequencing background noise. It was validated by analyzing paired-ends Illumina data from the targeted resequencing of a gene panel enriched by capture from colorectal, lung, ovarian and breast cancer paraffin-embedded tumors already genotyped during initial diagnostic of cancer. Its analytic performances were compared to those of Varscan2, LofreqStar and also to the well-known HaplotypeCaller (Pubmed: 2222 citations). It produces a powerful, simple and comprehensive analysis with an assessment of sensitivity limits for use in routine practice.

RESULTS

After sequencing, targeted regions were covered with an average depth of 2111× and 99.46% of nucleotides were covered with a depth > 200×. The 130 samples were analyzed by four different variant-callers, including OutLyzer, to highlight both Single Nucleotide Variations (SNVs) and Insertion-Deletion (Indels) events.

A total of 12747 SNVs with an allele ratio higher than 1% was identified on coding regions (Figure 1A) and 53 indels with an allele ratio higher than 2% (Figure 1B). SNVs and Indels were processed in two separate benchmark analyses. Regarding SNVs, most mutations detected by all variant-callers were from a probable germline origin with an allele ratio around 50 (heterozygous) or 100 % (homozygous). Among the 30 SNVs detected by both HaplotypeCaller and Varscan, 28 represented one same recurrent event located in an area with mapping issues associated with poor quality metrics. The 16 SNVs detected only by HaplotypeCaller also had a low Phred Score with mapping issues, just like the 60 SNVs, corresponding to 10 unique variants detected by Lofreq alone. Other SNVs found by OutLyzer only, Lofreq and Varscan, OutLyzer and Varscan, or Lofreq and Varscan and OutLyzer together were low allele-ratio events with good quality metrics. SNV found by Varscan only showed lower quality metrics and some were highly recurrent events between samples (190 variants for 5 unique variants).

To enhance the clarity of indel analysis, comparative data were firstly cleaned manually to remove two recurrent false-positives with a low allele ratio (detected by all variant-callers) that were induced by mapping issues and were present in most patients. All the expected events in this dataset were detected by all variant-callers, so their levels of performance were similar (Figure 1B). The 5 indels detected by all of them except HaplotypeCaller were mutations recovered with a low allele ratio and probably of somatic origin. Among the 2 indels identified only by Varscan, one was an artefact caused by a homozygous deletion on the adjacent nucleotide, and the other was a low allele-ratio deletion also identified by OutLyzer but below the limit fixed at 2%. The last indel identified only by Lofreq was also a low allele-ratio deletion supported by only 4 reads. Sensitivity and specificity were then calculated to evaluate and compare performances of each variant-caller.

Sensitivity was evaluated on all events previously genotyped, including 51 SNV and 27 indels from 1.3% to 93% of allele-ratio (Figure 2). OutLyzer had the best sensitivity with Varscan in identifying 100 % of the tested mutations, while, as awaited, HaplotypeCaller performed least well owing to a loss of sensitivity in the detection of low allele-frequency variants (Figure 2 and Figure 3). To evaluate the impact of coverage on outLyzer sensitivity, we used a sample built from the DNA of 11 tumors harbouring already known mutations. The 11 DNA were mixed in order to obtain a unique sample of DNA with 11 low allele-ratio mutations (from 1 to 10% of allele-ratio). This mixed sample was sequenced 10 times in independent experiments (reproducibility tests) and BAM files obtained were used to simulate different coverage conditions. For each BAM some "reads" were randomly selected *in silico* to divide by 2, 4 and 10 the initial depth of coverage, in order to obtain five ranges of depth of coverage on the genomic loci of the mutations (< 150×, 150–300×, 300–600×, 600–1000×, > 1000×). The sensitivity was calculated for each range of depth of coverage (Figure 4). As expected, a low coverage condition (< 150×) demonstrated a loss of sensitivity and is harmful for the

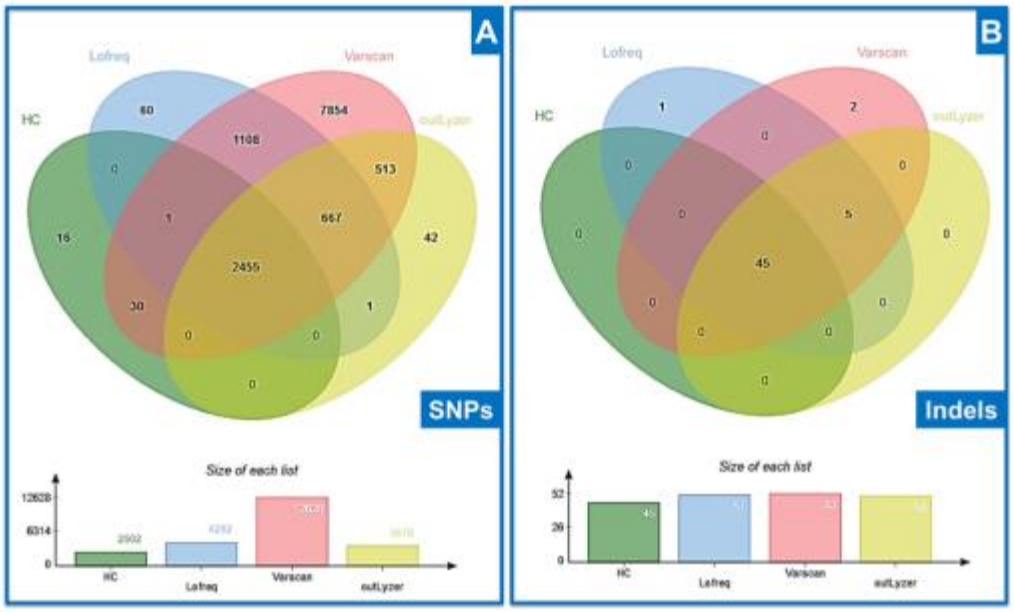


Figure 1: Venn diagram of mutations found by each variant-caller tested. In both figures, HC = HaplotypeCaller (**A**) Comparison of SNVs identified (**B**) Comparison of Indels identified.

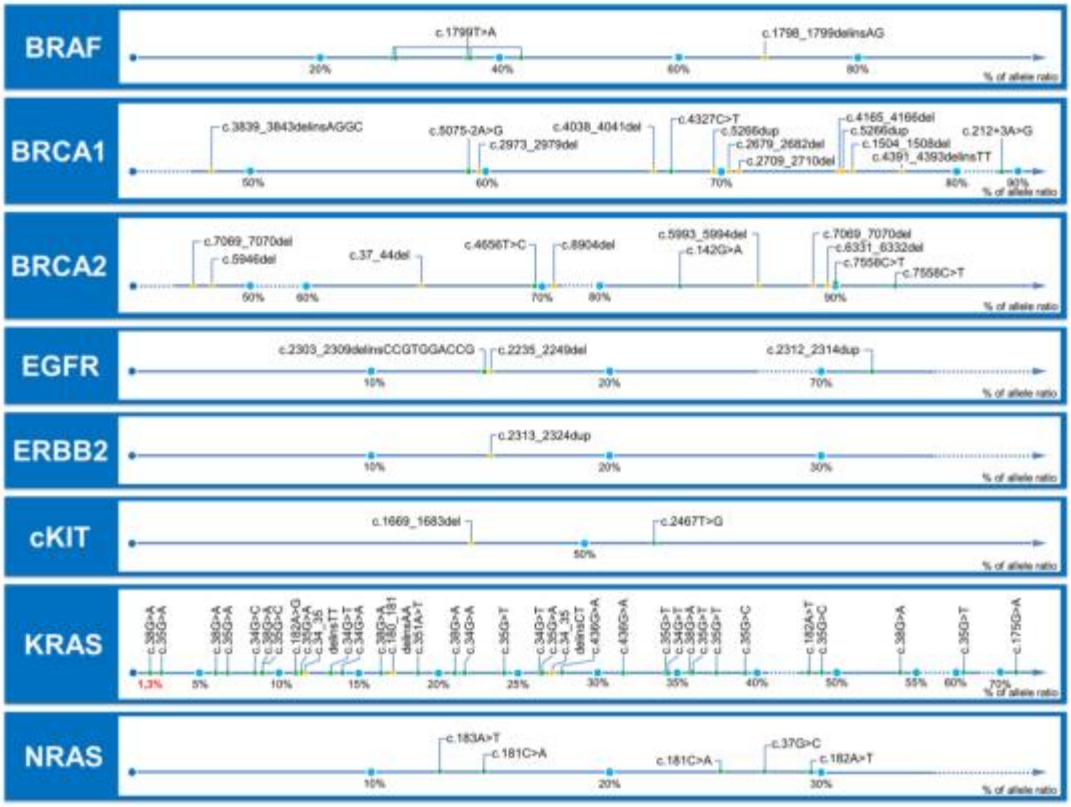


Figure 2: Allele-ratio of mutations tested for sensitivity evaluation. Mutations are gathered by gene, and positioned along an axis which represents the observed allele ratio. SNV are represented with a green dot and indels with a yellow dot. All mutations are annotated with HGVS nomenclature. These mutations were detected previously in the time of the diagnosis by specific methods (see Materials and Methods).

detection of low allele-ratio mutations. A coverage in the range of 150–300× is sufficient to detect mutations with 5% of allele-ratio with a good sensitivity. Increasing the coverage enhance the sensitivity up to detect mutations about 1 or 2% of allele-ratio with a coverage about 1000×.

It is difficult to test with a specific method all mutations detected by the variant-callers in order to determine their true or false positive nature, specifically regarding low allele-ratio variants. Consequently, the evaluation of the specificity was therefore restricted to the analysis of *KRAS* mutational hotspots (codon 12 and 13) for which a target specific method was already available for routine diagnostic purposes. The variant-callers detected 35 mutations in *KRAS* mutational hotspots. The corresponding samples were analyzed by Cold-PCR followed by pyrosequencing to evaluate the false or true positive nature of the mutations and then the specificity was calculated (Figure 5). OutLyzer also showed the best specificity with HaplotypeCaller by making no mistakes on the identified mutations, thereby establishing a specificity of 100% on these mutational hotspots (Figure 3). To illustrate performances of each variant-caller, a commercial sample (HorizonDX) in which a number of mutations with various allele ratios

	HaplotypeCaller	Lofreq	Varscan	outLyzer
Sensitivity (%)	87.2	98.7	100	100
Specificity (%)	100	99	94	100

Figure 3: Sensitivity and specificity evaluation. Sensitivity was calculated by testing previously genotyped samples harboring known mutations discovered in the time of diagnosis with contemporary validated methods which were mutation-specific (see Materials and Methods and Figure 2 for description of mutations). Specificity is calculated in KRAS codons 12 and 13 for which a sufficiently sensitive method was available (Cold-PCR followed by Pyro-sequencing). All variants detected in these specific regions by the variant-callers tested were checked for false or true positive nature.

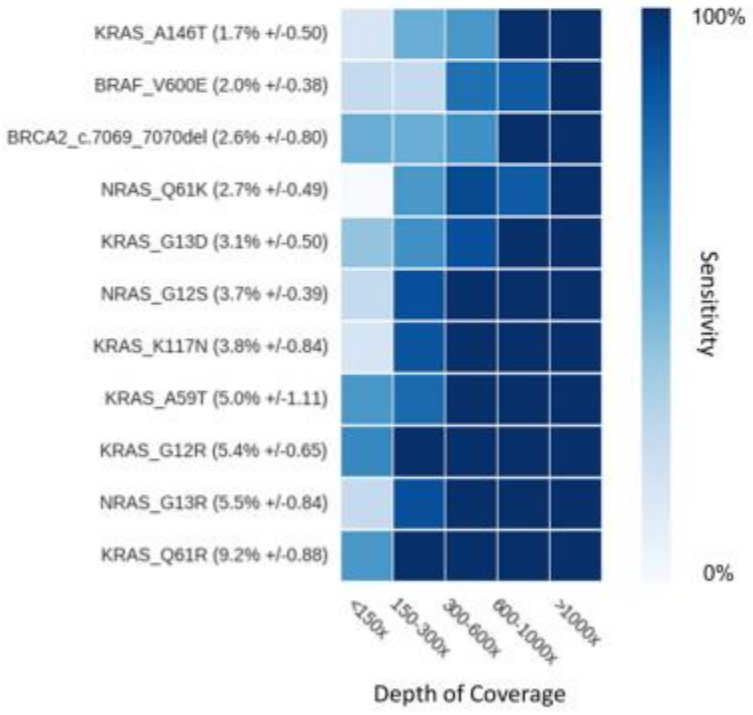


Figure 4: Impact of coverage on sensitivity. Each mutation of interest (y-axis) is ranked according to its average allele-ratio in ascending order from top to bottom. Sensitivity is calculated for each mutation at each coverage category (x-axis), and represented by color variations from 0 to 100% as shown on color bar on the right.

were known was sequenced and analyzed independently from other samples (Figure 6). Only OutLyzer and Varscan were able to detect all expected mutations. As awaited, HaplotypeCaller did not identify the lowest allele frequency mutations (Supplementary Table S1).

DISCUSSION

Highlighting mutational events with a very low allele frequency is challenging but essential in oncology in the search for somatic mutations characterizing tumor

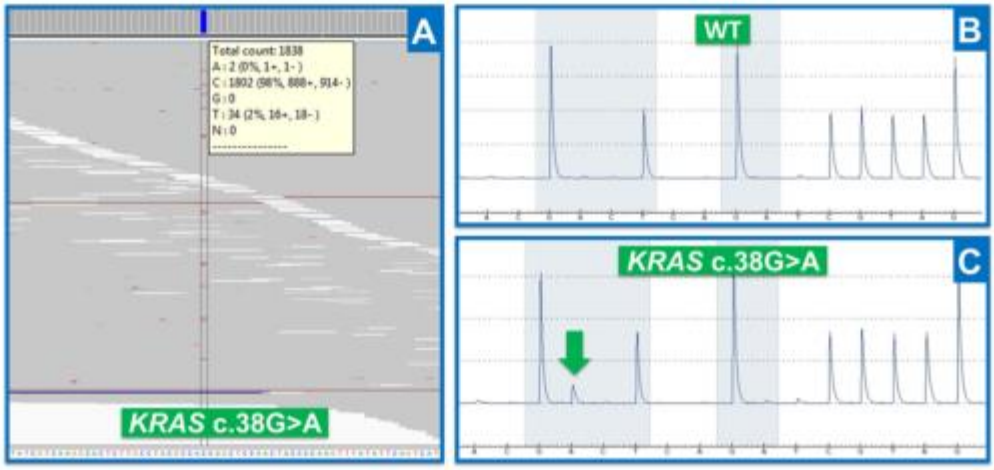


Figure 5: Results obtained by NGS and target-specific method. (**A**) NGS results (IGV visualization): reads aligned along a reference genome, illustrating *KRAS* c.38G > A mutation for Thera41 patient (codon 13). Data are represented in genomic orientation. (**B**) Pyrosequencing results obtained for a healthy patient on *KRAS* codons 12 and 13 (Wild Type). Data are represented in transcript orientation. (**C**) Pyrosequencing results obtained for Thera41 patient (Supplementary Table S1) with *KRAS* c.38G > A mutation (green arrow). Data are represented in transcript orientation.

	Allele ratio	Haplotype Caller	Lofreq	Varscan	outLyzer
BRAF c.1799T>A	10.2%	✗	✓	✓	✓
cKIT c.2447A>T	10.3%	✗	✓	✓	✓
EGFR c.2235_2249del	1.2%	✗	✓	✓	✓
EGFR c.2573T>G	3.8%	✗	✓	✓	✓
EGFR c.2369C>T	1.1%	✗	✗	✓	✓
EGFR c.2155G>A	26.7%	✓	✓	✓	✓
KRAS c.38G>A	14.5%	✓	✓	✓	✓
KRAS c.35G>A	5.4%	✗	✓	✓	✓
NRAS c.181C>A	11.5%	✓	✓	✓	✓
PIK3CA c.3140A>G	17.5%	✓	✓	✓	✓
PIK3CA c.1633G>A	10.5%	✗	✓	✓	✓
BRCA2 c.5073del	33.8%	✓	✓	✓	✓
MET c.713del	7.4%	✗	✓	✓	✓
BRCA2 c.5351del	39.3%	✓	✓	✓	✓
BRCA1 c.4327C>T	27.4%	✓	✓	✓	✓

Figure 6: Performance of all variant-callers on HorizonDX sample. Red cross means that mutation was not detected by the corresponding software.

heterogeneity [6]. Next Generation Sequencing (NGS) technologies have become powerful tools for helping to diagnose pathologies and establishing therapeutic strategies. However, it may be difficult to interpret their results and to differentiate false and true positives. Many tools are available to detect somatic mutations, each with its advantages and drawbacks. Tools such as HaplotypeCaller [9] were initially designed for discovering germline mutations so they are very specific but lack sensitivity, especially for detecting low-allele-frequency mutations. Other variant-callers have emerged to address this issue. Varscan [10] and Lofreq [11] had better performance here with almost all the tested mutations. However, in our hands they either generated false positive or they lose some sensitivity in case of low quality DNA samples. Others such as MuTect [12] or JointSNVMix [13], which were not tested here but have been reported to have both good sensitivity and specificity, need a paired healthy sample for mutation detection analysis. This may be problematic for the abovementioned reasons.

OutLyzer largely offsets these defects without the need for a matched healthy tissue sample by adjusting locally its sensitivity threshold depending on the sample sequencing quality, in the same way as a biologist inspects aligned sequencing data and assesses quality and background noise. This leads to better performance in terms of sensitivity and specificity and is critical in a clinical context, with all the quality constraints imposed on diagnostic laboratories nowadays. Evaluating the performance of variant-callers remains challenging, largely because of the amount of data produced by NGS technology. This revolution allows us to explore larger genomic regions with hopefully greater sensitivity. But all events discovered with this technology cannot be checked and compared systematically with conventional methods, such as Sanger sequencing, pyrosequencing or digital PCR. Therefore it appears critical to exactly know the theorical limit of each technology used. In order to understand outLyzer's limitations in its usage, in addition to variant-calling analysis, outLyzer is able to provide the analysis detection limits in the form of a minimum detectable allele ratio. This detection threshold is produced per patient, either for each region specified in an associated BED (Browser Extensible Data) File or for a specific genomic position, such as *KRAS* mutational hotspots. In a second operating mode, it is also able to provide immediately the major sequencing features for a specified genomic position, including sequencing depth, reference and alternative allele (if present), the number of forward and reverse reads that carry this alternative allele, the average sequencing quality, and an estimation of local sequencing background noise.

OutLyzer estimates the sum of several background noise sources, including sequencing mistakes, errors generated by sample preparation and bioinformatics analysis, based on an outlier detection algorithm. Here, detection of outliers is used to highlight true mutations but outlying data have to be used carefully depending on the experimentation type so as to fit the analysis as well as possible [14]. With a high depth of coverage on the diagnostic genomic regions, dedicated bioinformatics tools can help us to ensure a good sensitivity and specificity. Depending of the sensitivity required, the depth of coverage can be adapted (Figure 4), but a high sensitivity for detection of mutations with an allele-ratio of 1% requires at least 1000× of coverage for a robust analysis. Such depth of coverage can be helpful to explore tumoral heterogeneity, particularly with a low tumor cellularity, or to detect mosaic mutations or circulating tumor DNA mutations.

Despite the fact that bioinformatic, statistical and computational methods are constantly evolving, a hurdle they face is the nature of processed data, which contains errors non-distinguishable from real biological events. An important source of false positives lies in the sample preservation method, the FFPE (Formalin-Fixed, Paraffin-Embedded), which is responsible for DNA modifications considered as artifacts unrated to pathology [15]. This limitation may be not overpassed by bioinformatics improvements. Therefore, alternative sample preparation before sequencing could limit bias to enhance the detection of low allele frequency mutations. For example, sample preparation protocols based on the use of a random index can simulate the double-stranded sequencing of a unique DNA fragment with a suitable bioinformatics analysis. Such protocols could eliminate PCR and sequencing errors and distinguish true somatic mutations occurring on both strands from errors generated during the analytical process [16].

Today sequencing technologies make it possible to explore numerous diseases and characterize many genetic abnormalities. Software such as OutLyzer could prove useful by highlighting mainly true positive mutations in the sequencing background and focusing only on the most relevant information. Outlyzer sources are available on https://github.com/EtieM/outLyzer.

MATERIALS AND METHODS

OutLyzer implementation

OutLyzer is a tool written in Python [17] programming language that runs on Linux. It requires the SAMTOOLS [18] suite and additional python libraries: *scpiy*, *numpy*, *subprocess*, and *multiprocessing*. It was tested on Fedora 21 and CentOS 6.7 Linux distribution with SAMTOOLS 1.2 and Python 2.7.8 versions.

OutLyzer has two main operating modes: (i) as a classical variant-caller; (ii) as a tool to evaluate local quality metrics. By giving it a chromosomal for a given sample, it quickly evaluates whether a mutation is present at this position, specifies all raw sequencing information and evaluates local background noise.

OutLyzer uses BAM (Binary Alignment Map) files as input for analysis. It is preferable to use the BWA [19] / GATK [9] bioinformatic pipeline to produce BAM files, according to the Broad Institute recommendations. BAM files are converted into the pileup data format using SAMTOOLS [18]. For each genomic position of targeted regions, the number and type of alternative events on forward or reverse reads and the associated PHRED score are stored in memory for a defined genomic region to be analyzed together (ex: one exon) in the subsequent statistical steps.

For each region stored in memory, OutLyzer evaluates background noise locally by using Thompson's Tau Test [20]. This test is a statistical method for deciding whether to keep or discard a suspected outlier in a sample of a single variable. At each iteration, the sample mean $\bar{x}$ and standard deviation S are calculated. Then for each point of the sample, the absolute value of the deviation is calculated:

$$\delta_i = |d_i| = |x_i - \bar{x}|$$

The data point most suspected as a possible outlier is the data point with the maximum value of δ_i. The value of the modified Thompson τ is calculated from the critical value of Student's t PDF (Probability Density Function):

$$\tau = \frac{t_{\alpha/2} * (n-1)}{\sqrt{n} * \sqrt{(n-2) + t_{\alpha/2}^2}}$$

Where n is the number of data points in the sample. Student's t value is based on an α risk set by default at 0.001 (adjustable in settings in OutLyzer) and df (degree of freedom) = n–2. The removal or retention of a potential outlier is evaluated by the decision rule: if $\delta_i > \tau * S$, data point is rejected from sample because considered as an outlier. Otherwise, the data point is kept. The test is performed until no outlier is found in the sample. Considering a window of 200 bp centered on potential mutations (adjustable according to user preferences), for each genomic position, the number of reads containing alternative bases is added to a list which forms a local sample (Figure 7A and 7B). First, all values equal to zero (genomic position that does not contain a mutation) are removed from the sample. Then Thompson's Tau test is performed to remove the largest outlier in the sample and the test is performed again until no outlier is found. The largest data point in the remaining list is then used to define the background noise locally (Figure 7B). Back to the aligned data, if the number of reads supporting a potential variant is higher than the local background noise previously defined, the potential variant moves to the filtration step. Otherwise it is considered as part of the background noise (Figure 7C). For each candidate variation kept in the statistical step, several filtration steps are performed, all of which are configurable: (i) variation rate must be greater than twice the background noise, (ii) variation must have an average PHRED score greater than 20, (iii) the average PHRED score must have a standard deviation below 7, (iv) the forward/reverse balance of variants should be between 30 and 70%. Variations that meet all these criteria are written in a vcf (Variant Call Format) file.

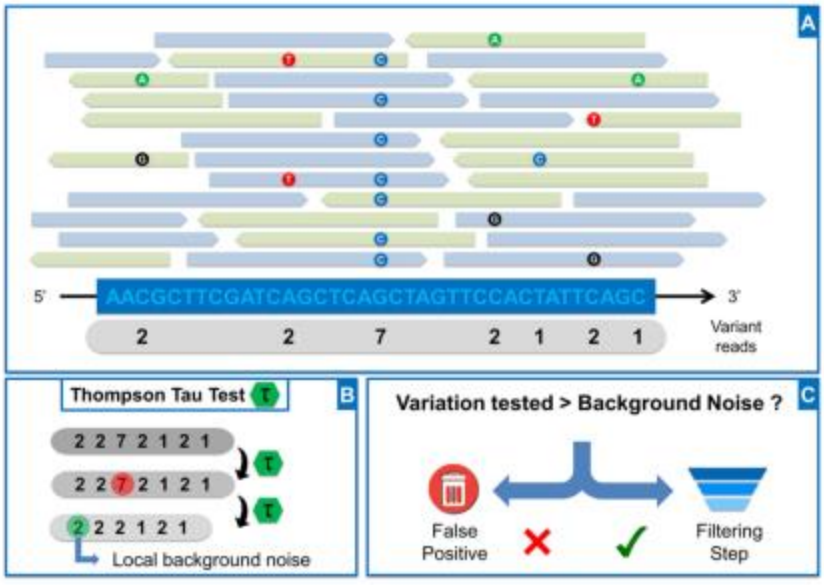


Figure 7: OutLyzer analysis. (**A**) Representation of reads aligned along a reference genome. For each genomic position, the number of variant reads is counted and stored in a list (grey banner) (**B**) Application of Thompson Tau test on list obtained in A (**C**) The number of reads carrying the potential variant is compared to local background noise to evaluate whether the event is a false positive. If the variant is above background noise, it will pass through a filtration step based on sequencing quality, and including the reads forward-reverse balance, the average PHRED score of mutated bases, and the standard deviation of average PHRED score.

Tumor samples

Paraffin-embedded tumor samples from 130 patients with colorectal (70 samples), ovarian (50 samples), lung (4 samples), breast (4 samples), skin (1 sample) and stomach (1 sample) cancer, were selected for high-throughput sequencing analysis (Supplementary Table S1). All patients had been previously sequenced by the target-specific techniques in the time of the initial diagnosis (Sanger sequencing, Cold-PCR followed by Pyrosequencing, SNaPSHOT).

Additionally, the Horizon DX Quantitative Multiplex Reference Standard (Horizon Discovery Group, Cambridge Research Park, Waterbeach, Cambridge, UK) was added to the dataset containing miscellaneous SNVs (Single Nucleotide Variations), insertion and deletion events at various allele frequency.

Sequencing analysis

DNAs were sequenced for a panel comprised of 22 genes (Figure 8). Agilent SureDesign (Agilent, Santa Clara, CA, USA) was used to create library baits covering the exonic regions of these genes. Regions of interest were captured with the SureSelect XT Protocol (Agilent, Santa Clara, CA, USA) and sequenced on Illumina Miseq (Illumina, San Diego, CA, USA) using the paired-end 2 × 150 bp program. Bio-informatic analysis was performed with the CASAVA Suite v1.8 for demultiplexing, followed by BWA 0.7.12 for alignment and GATK v3.3 pipeline to produce BAM files, according to the Broad Institute recommendations. The variant-calling step was carried out by HaplotypeCaller, Lofreq v2.1.1 and Varscan v2.3.7 for comparison with OutLyzer (settings are described in the Supplementary information S1). Only SNVs and Indels with an allele ratio respectively greater than 1% and 2% were compared. Venn Diagram representations were designed with jVenn [21].

ACKNOWLEDGEMENTS

We thank the SéSAME (Séquençage pour la santé, l'Agronomie, la Mer et l'Environnement) sequencing platform, Aude Lamy (CHRU Rouen, France) , Christian

Gene	Exons sequenced
AKT1	4
ALK	20, 22 - 25
BRAF	11, 15
BRCA1	All
BRCA2	All
EGFR	18 - 21
ERBB2	All
ERBB4	10, 12
FGFR2	7, 12, 14
FGFR3	7, 10, 15
HRAS	4
KIT	2a - 9, 11, 12
KRAS	2 - 4
MAP2K1	2
MET	2, 14 - 20
NRAS	2 - 4
PDGFRA	12, 14, 18
PIK3CA	10, 21
PTEN	5, 7, 8
SMAD4	9, 12
TGFBR2	4
TP53	All

Figure 8: Genes included in sequencing analysis. Genes have been selected to establish a NGS panel gene strategy in order to characterize solid tumors in clinical practice.

Bastard (Centre Henri Becquerel, Rouen) and Alexandra Lespagnol (CHRU Rennes, France) for providing some tumor samples.

CONFLICTS OF INTEREST

None declared.

FUNDING

The authors would like to thank the Institut National du Cancer (INCa) for funding.

REFERENCES

1. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–45.
2. Castéra L, Krieger S, Rousselin A, Legros A, Baumann JJ, Bruet O, Brault B, Fouillet R, Goardon N, Letac O. Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. Eur J Hum Genet. 2014; 22:1305–13.
3. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye F, Lindeman N, Boggon T. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004; 304:1497–500.
4. Kothari N, Schell MJ, Teer JK, Yeatman T, Shibata D, Kim R. Comparison of KRAS mutation analysis of colorectal cancer samples by standard testing and next-generation sequencing. J Clin Pathol. 2014; 67:764–7.
5. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, Mortimer P, Swaisland H, Lau A, O'Connor M, Ashworth A, Carmichael J, Kaye S. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med. 2009; 361:123–34. doi:10.1056/NEJMoa0900212.
6. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013; 501:355–64.
7. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing datasets. bioRxiv. 2014; 8375.
8. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen M. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012; 30:434–9.
9. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303.
10. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller C, Mardis E, Ding L, Wilson R.. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22:568–76.
11. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd M, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012; gks918.
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander E, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013; 31:213–9. doi:10.1038/nbt.2514.
13. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, Marra M, Aparicio S, Shah S. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinforma Oxf Engl. 2012; 28:907–13. doi:10.1093/bioinformatics/bts053.
14. Altman N, Krzywinski M. Points of Significance: Analyzing outliers: influential or nuisance? Nat Meth. 2016; 13:281–2.
15. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem. 2015; 61:64–71.
16. Marx V. Cancer: hunting rare somatic mutations. Nat Meth. 2016; 13:295–9.
17. Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.
18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25:2078–9.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25:1754–60.
20. Cimbala JM. Outliers. Penn State Univ 2011. http://www.mne.psu.edu/cimbala/me345/Lectures/Outliers.pdf.
21. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014; 15:293. doi:10.1186/1471-2105-15-293.

B. Evaluation de l'incidence des variants en mosaïque à partir d'une population de 1750 patientes prédisposées pour le cancer du sein et/ou de l'ovaire

1. Matériels et Méthodes

a. Patientes

Les 1750 patientes analysées dans cette étude constituent un sous-échantillon de la population utilisée pour la première étude de cette thèse.

b. Séquençage et analyse bio-informatique

Les FASTQ de séquençage de 34 gènes sont les mêmes que pour l'étude précédente. Les données issues du séquençage ont été analysées en utilisant l'algorithme d'alignement BWA[174] (algorithme mem, version 0.7.12), les fichiers BAM ont ensuite été recalibrés avec les suites logicielles Picard (version 1.119) GATK[181] (version 3.3) selon les recommandations fournies par le Broad Institute.

Une version modifiée d'outLyzer a ensuite été utilisée, afin de mieux prendre en compte le déséquilibre entre reads Forward et Reads Reverse pouvant être rencontré dans les méthodes d'enrichissement par capture notamment aux extrémités des régions cibles (régions non explorées dans l'étude des tumeurs). En effet, des pertes de sensibilités sur ces régions motivant la mise à jour du logiciel ont été notées lors des premiers tests sur du matériel leucocytaire issus de jeux de données d'entrainement et d'échanges inter-laboratoires.

OutLyzer implémente dans sa première version, parmi ses critères de filtration, un critère de qualité évaluant le déséquilibre observé entre *reads Forward* et *Reverse* (« FR balancing »)

identifié comme à l'origine de cette perte de sensibilité. Théoriquement, la distribution entre *reads* alignés en *Forward* et *Reverse* doit être équilibrée (50% de chaque), les deux brins de l'ADN étant séquencés sans sélection préalable. Toute variation observée doit donc aussi en théorie être représentée de manière équilibrée sur les deux sens. La version originale d'outLyzer était paramétrée pour ne sélectionner que les variants montrant un ratio *Forward*/*Reverse* compris entre 30 et 70 %. Ce critère est finalement trop stringent et pas assez adaptable quand il s'agit d'appeler les variants observés aux extrémités des régions génomiques capturées[265] là où des biais d'amplification liés à la fixation des amorces de PCR peuvent perturber cet équilibre *Forward*/*Reverse*. Une modification du logiciel a donc été apportée afin de prendre en compte ce déséquilibre et adapter le FR balancing en fonction du ratio observé sur les *reads* non mutés (Facteur de Correction D). Dans la nouvelle version du logiciel, le FR balancing des variants est donc re-calculé et corrigé en pondérant la valeur observée sur les alignements par le ratio de *reads Forward*/*Reverse* d'allèles non mutés selon la formule suivante, sur la position génomique étudiée :

$$FR_{balancing} = \frac{Alt_F - (Alt_F \times D)}{Alt_F + Alt_R} \qquad avec\ D = \frac{Ref_F}{Ref_F + Ref_R} - 0.5$$

$Ref_F = Nombre\ de\ reads\ Forward\ portant\ l'allèle\ sauvage$

$Ref_R = Nombre\ de\ reads\ Reverse\ portant\ l'allèle\ sauvage$

$Alt_F = Nombre\ de\ reads\ Forward\ portant\ l'allèle\ muté$

$Alt_R = Nombre\ de\ reads\ Reverse\ portant\ l'allèle\ muté$

$D = Facteur\ de\ correction$

Les seuils de filtre sont conservés (30-70%) et sont appliqués sur la valeur du FR balancing corrigée. Cette version a été développée après publication de la première version, et a été validée sur le jeu de données utilisé dans la publication initiale. Aucune perte de spécificité n'a été mise en évidence et l'algorithmique filtrant les déséquilibres de sens de séquençage retrouve les variants non détectés lors des tests préliminaire avec la première version du logiciel (données non montrées).

Les variants identifiés par outLyzer v.2 ont ensuite été annotées avec Alamut-Batch (Interactive Biosoftware) et sélectionnées selon les critères suivants : (i) ratio d'allèles variants compris entre 5 et 30 % (ii) variant identifié présent de manière unique dans l'ensemble de la cohorte (singleton) puisque nous émettons l'hypothèse que l'apparition d'un événement post-zygotique est un événement aléatoire ayant une probabilité infiniment faible d'être retrouvé aux mêmes positions génomiques dans notre série de patients. Seuls les variants les plus probablement pathogéniques ont été étudiés, en ne sélectionnant que les variants (i) responsables de l'apparition d'un codon Stop prématuré (variation non-sens / *frameshift*) (ii) référencés en tant que variants pathogènes dans les bases de données UMD BRCA1 / BRCA2, IARC pour TP53 et la base Insight pour les gènes MMR (iii) provoquant une abolition complète des sites physiologiques de l'épissage prédite par les algorithmes MaxEntScan ou SplicingSequencesFinder.

Afin d'assurer une spécificité très élevée de notre analyse et comme nous avons démontré une spécificité estimée à 100% sur le jeu de validation avec des variants pouvant montrer des ratio d'allèle variants de l'ordre de 1,5%, les variants utilisés pour cette étude ont été filtrés pour ne sélectionner que ceux ayant un ratio d'allèle variant supérieur à 5%.

2. Résultats

Parmi les 1750 patients inclus dans la cohorte explorée, 7 variants délétères en mosaïque ont été mis en évidence (Table 1), à une fréquence allélique moyenne de 6.7% (± 1.22). Les deux variants identifiés sur *ATM* sont portés par des patientes prises en charge dans le cadre d'un cancer de l'ovaire, et chez qui avait été mis en évidence un variant pathogène à l'état hétérozygote de *BRCA1* lors de la prise en charge oncogénétique. Un variant d'épissage de *BAP1* et un variant frameshift de *CHEK2* ont été découverts chez des patientes dans un contexte d'apparition sporadique. Enfin, un variant de *MLH1* et 2 variants de *TP53* ont été identifiés, sans variant de prédisposition constitutionnel connu chez les individus concernés.

Table 1: Variants en mosaïque identifiés parmi 1750 cas index pris en charge dans le cadre d'un cancer du sein et de l'ovaire. [1] BRCA1 c.3228_3229del (p.Glu1077Alafs*8). [2] BRCA1 c.5311_5332+1del (p.Pro1771Ilefs*15)

Gène	cNomen	pNomen	Catégorie	Allèle Ratio (%)	Cancer	Age de diagnostic	Présence d'une mutation constitutionnelle	Contexte familial
ATM	c.5180del	p.Val1727Alafs*11	Frameshift	6.5	Ovaire	58	BRCA1[1]	Cancer du sein
ATM	c.451del	p.Ser151Leufs*2	Frameshift	8.2	Sein /Ovaire	73 / 63	BRCA1[2]	Cancer du sein et de l'ovaire
BAP1	c.67+2T>A	p.?	Epissage	6.0	Ovaire	64	Ø	Sporadique
CHEK2	c.879del	p.Ile294Serfs*24	Frameshift	8.1	Sein /Ovaire	52 / 66	Ø	Sporadique
MLH1	c.2084C>A	p.Ser695*	Non-sens	5.5	Sein BL TN	47 /54	Ø	Cancer du sein
TP53	c.728dup	p.Met243Ilefs*21	Frameshift	7.2	Sein BL	67 / 79	Ø	Cancer du sein
TP53	c.483del	p.Ile152Serfs*8	Frameshift	5.1	Ovaire	70	Ø	Cancer du sein et de l'ovaire

3. Discussion

Les critères de sélection utilisés ici pour la mise en évidence de néo-mutations en mosaïque avaient pour objectif de ne conserver que les variants les plus vraisemblables (vrai positifs) permettant ainsi d'éviter la mise en œuvre d'une confirmation par une technique de référence afin d'estimer l'incidence de ces événements. Ainsi, sur le panel séquencé, l'incidence des mosaïques détectées atteint 0.4% concernant les variants « perte de fonction ». Ce taux de néo-mutations en mosaïque reste néanmoins difficile à interpréter, étant donné les rares cas décrits dans les cancers du sein et de l'ovaire. Démontrer un enrichissement de ce type de variants dans notre population de patientes prédisposées pour les cancers du sein et de l'ovaire nécessiterait des études cas-témoins de plusieurs milliers d'individus correctement appariés, étant donné la rareté des évènements[266] et l'influence de l'âge sur l'apparition des variants en mosaïque, une augmentation de l'incidence des variants en mosaïque avec l'âge pouvant biaiser l'analyse[267]. Cet enrichissement pourrait néanmoins être simulé en utilisant cette fois encore les informations mises à disposition par le consortium ExAC[268]. En effet le consortium ExAC a aussi évalué l'intolérance des gènes à l'apparition de variants de type perte de fonction, en comparant par gène le nombre de variants observés et le nombre de variants attendus (à partir d'une probabilité théorique de mutation). Les probabilités théoriques de néo-mutation par gène et par classe de variant (synonyme, faux-sens, non-sens et épissage) ont été calculées par une adaptation de la méthode de Samocha et al., basée sur l'observation des divergences entre génome humain et macaque[152]. L'utilisation de ces informations leur a permis d'estimer, par gène, une probabilité d'intolérance aux pertes de fonction (pLI) causées par les variants non-sens, frameshift et les variants d'épissage sur les sites canoniques. Les probabilités de néo-mutation, pondérées par la pLI, pourraient permettre de créer un modèle de probabilité, à l'image de celui développé dans la première étude de cette thèse, simulant

l'apparition d'une néo-mutation chez un individu considéré comme non-atteint de cancer afin de la comparer à celle observée dans notre population de cas.

L'impact des variants en mosaïque dans le syndrome de prédisposition pourrait cependant être sous-estimé, du fait de la difficulté de leur détection. La limite de détection de 5% utilisée ici constitue déjà une sensibilité rarement atteinte dans les études de génétique constitutionnelle mais il est probable que d'autres variants soient présents à de très faibles taux sur l'ADN de leucocytes circulants, encore inférieurs aux limites de détection fixées. Dans une deuxième analyse que nous avons réalisée, le seuil de détection a été abaissé à un ratio d'allèle variant à 2.5% tout en augmentant la stringence en fixant cette fois un nombre minimal de reads portants l'allèle variant à 8. Les résultats (données non montrés) sont plus incertains et nécessiteraient une confirmation des variants identifiés par une approche ciblée sensible (ex : PCR digitale). Cette étude met en évidence 27 variants en mosaïque sur la série des 1750 patientes précédemment étudiées. Ainsi, pour les patientes chez qui ont été identifiés de nouveaux variants en mosaïque (entre 2.5 et 5%), aucun variant clairement responsable d'une prédisposition au cancer à l'état hétérozygote n'a été identifié. De plus, de manière surprenante, 50 % de ces patientes identifiées présentaient une forme sporadique de cancer. Ces résultats demandent donc une investigation approfondie, par des techniques très sensibles, mais aussi par l'analyse de prélèvements tissulaires d'origines différentes. Ces résultats laissent penser que la participation des variants en mosaïque responsables de l'apparition de cancer reste très probablement encore sous-estimée.

III / Evaluation du risque de cancer du sein et de l'ovaire précoce induit par les variants constitutionnels rares des gènes impliqués dans le développement du cancer

A. Présentation de l'étude

Une grande partie des situations évocatrices d'une prédisposition au cancer du sein et de l'ovaire reste aujourd'hui inexpliquée. Jusqu'à présent, si l'exploration du génome a pu mettre en évidence de nombreux facteurs de risque de forte, moyenne ou faible pénétrance (cf. Introduction, § IV), certaines situations extrêmes de par leur précocité au diagnostic, la multiplicité des cancers individuels ou dans une famille sans facteurs prédisposants évidents laissent penser qu'il existe encore à découvrir des facteurs génétiques très fortement pénétrants rares voire complètement privés.

La révolution du séquençage à haut-débit a depuis permis de mettre en évidence la grande variabilité du génome humain, illustrée par le fait que 99% des variants retrouvés dans la base de données ExAC montrent une fréquence allélique inférieure à 1% (cf. Introduction, § VII.B.2.b). Cette technologie permet donc maintenant d'explorer la part des variants rares voire privés (MAF < 0.1%). Certains de ces variants sont présents de manière unique dans les bases de données (singletons) et dans le cas d'un variant prédit comme pathogène (cf. Introduction, § VII.B.1), l'absence de récurrence de ces variants avec le trait pathologique pose le défi de leur interprétation puisque la récurrence d'un variant rare avec la pathologie est un bon élément de preuve de causalité. La démonstration de la pathogénicité de tels variants nécessite la mise en place d'études cas-témoins adaptées (cf. Introduction, § VII.B.4)

pouvant requérir de grandes populations, afin de démontrer un effet significatif si celles-ci sont réalisées sur des données d'exomes et encore plus de génomes.

Alternativement, l'identification de facteurs de prédisposition peut aussi être réalisée par des études d'exomes comparatifs « en trio », entre un cas index et ses parents non-affectés, dans le cas où l'hypothèse est celle de la présence d'une néo-mutation pathogène[269]. Néanmoins le défi majeur de cette approche réside dans l'interprétation biologique des variants identifiés (pouvant toucher des gènes dont la fonction biologique est inconnue), la démonstration de leur pathogénicité étant particulièrement laborieuse.

Entre ces deux alternatives d'autres moyens d'étude peuvent être envisagés. De récentes publications, basées sur des analyses pangénomiques, ont montré que le nombre de gènes porteurs de variants acquis responsables du développement tumoral était relativement limité, ces gènes appartenant de plus aux mêmes voies de signalisation biologiques[270–272]. De plus, près de la moitié des gènes de prédisposition au cancer identifiés jusqu'ici sont aussi affectés par des variants acquis dans les tumeurs, confirmant ainsi que les voies de signalisation affectées par les cancers héréditaires et sporadiques sont au moins partiellement chevauchantes[270–272]. Ainsi, afin d'explorer la part d'hérédité manquante chez des individus fortement suspectés de présenter un risque augmenté de cancer du sein ou de l'ovaire, sans variant génétique de prédisposition connu, nous avons réalisé le séquençage de 201 gènes sélectionnés pour leur implication prouvée ou supposée dans le développement du cancer tant au niveau constitutionnel que tumoral. Cette analyse de séquençage a été conçue pour obtenir une forte profondeur de séquençage afin d'également explorer l'hypothèse de la présence de variants pathogènes en mosaïque tel que cela a été réalisé dans le deuxième travail de cette thèse (étude en cours, non montrée ici). Les individus ont été sélectionnés pour la précocité d'apparition de la pathologie, la précocité d'un cancer étant un fort élément de suspicion de prédisposition. L'enrichissement de ce phénotype doit donc théoriquement favoriser la mise

en évidence d'évènements causaux induisant des risques élevés et donc augmenter la puissance de l'étude (cf. Introduction, § VII.B.4). 105 patientes diagnostiquées pour un cancer du sein avant 31 ans et 13 patientes ayant été diagnostiquées pour un cancer de l'ovaire avant 41 ans et sans variant pathogène identifié sur *BRCA1*, *BRCA2* ou *TP53* ont été incluses dans cette étude. L'enrichissement en variants pathogènes dans cette population de cas a été évalué par deux méthodologies statistiques différentes : le calcul de la probabilité d'apparition d'un variant pathogène chez les cas et les témoins avec comme référence les données ExAC (comme cela a été défini dans la première étude de cette thèse) et une étude cas témoin en agrégation avec comme référence les données individuelles d'exomes de donneurs non atteints de cancer fournis par le consortium France Exome Project (FREX).

B. Matériels et Méthodes

1. Patientes et contrôles

Le recrutement des patientes a été effectué dans le cadre d'une prise en charge oncogénétique, et pour lesquelles un consentement éclairé a été obtenu. Les critères de sélection ont été les suivants : (i) un diagnostic de cancer du sein avant 31ans (105 patientes) ou (ii) un diagnostic de cancer de l'ovaire avant 41 ans (13 patientes) et (iii) aucun variant pathogène identifié sur les gènes de *BRCA1*, *BRCA2* et sont exclues les cancers du sein précoces dans un contexte de syndrome de Li Fraumeni avec un variant pathogène de TP53 identifié le cas échéant. Parmi celles-ci, une collection d'échantillons issue d'une étude précédente (EXAL) a été requalifiée afin d'être utilisée dans l'étude (22 patientes). Les données individuelles d'exomes de 522 individus non atteints de cancer ont été obtenues dans le cadre d'une collaboration avec le consortium France Exome Project (FREX). Les individus de cette étude ont été testés pour leur origine géographique par une double ACP utilisant les variants fréquents (>5% de

fréquence sur l'ensemble du jeu de données et pourcentage de valeurs manquantes <5%) et les variants peu fréquents (<5% de fréquence sur l'ensemble du jeu de données et pourcentage de valeurs manquantes <5%) ainsi qu'une analyse du nombre de singletons présents pour chaque patient. Sur ces analyses, un patient est considéré comme mal apparié en cas de valeurs d'ACP au-delà de la moyenne ± 3 SD (Standard Deviation = Ecart-type) (Variants fréquents) et ± 2 SD (Variants peu fréquents).

2. Séquençage à haut-débit
    a. Sélection des gènes analysés

Une première liste de 717 gènes suspectés pour leur implication dans le cancer au niveau constitutionnel ou tumoral a été élaborée afin de les hiérarchiser avant sélection des candidats à partir de :

- Panels de gènes publiés dans la littérature :
    - Cancer5000 (219 gènes) issu d'une analyse de saturation sur 21 types de tumeurs[271]
    - Panel HPV (226 gènes) issu d'une cartographie tumorale portant sur les cancers du col de l'utérus[273]
    - Panel IMPACT (279 gènes) portant sur l'identification de facteurs génétiques dans le cancer de la prostate (http://www.impact-study.co.uk/)
- Liste de gènes publiés par l'INCa dans le cadre du diagnostic des formes héréditaires de cancer (86 gènes)
- Panel utilisé dans le diagnostic des prédispositions aux cancers du sein et de l'ovaire au Laboratoire de Biologie et de Génétique du Cancer du Centre François Baclesse (34 gènes)

- Panels de gènes commercialisés
    - ThermoFisher : Comprehensive Cancer Panel (407 gènes)
    - Illumina: TruSeq Amplicon Cancer Panel (48 gènes)
    - Qiagen (24 gènes)

Chacun des gènes de la liste ainsi constituée a ensuite été évalué :

- pour son implication dans le cancer au niveau somatique ou constitutionnel par une revue de la littérature (0 : aucune implication ; 1 : discutable ; 2 indéniable)
- sa présence dans les 100 premiers gènes identifiés pour leur implication dans le cancer, selon la base de données COSMIC (http://cancer.sanger.ac.uk/cosmic/classic)
- un lien avec le cancer dans la base de données OMIM

Une hiérarchisation de ces gènes a ensuite été établie, en fonction d'un score calculé selon la Figure 61.

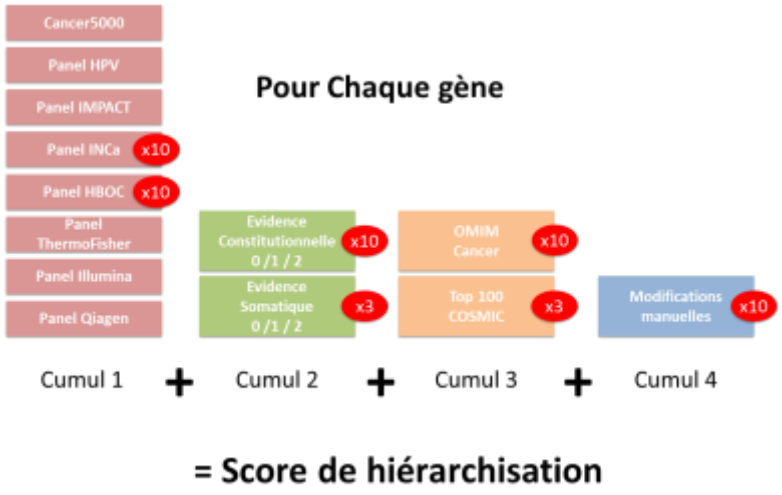


Figure 61: Calcul du score de hiérarchisation servant à la sélection des gènes à inclure dans l'étude. Pour chaque gène, un score est calculé en fonction de sa présence dans chaque liste ou catégorie. Les différents items sont pondérés en fonction de leur pertinence dans l'exploration des syndromes de prédisposition aux cancers du sein et de l'ovaire.

Après pondération de la profondeur de séquençage souhaitée avec la taille des régions exoniques de chaque gène, les 201 gènes montrant les scores les plus élevés ont été inclus dans l'analyse (Table 2) pour pouvoir séquencer 20 patients par *run* de nextSeq avec une cartouche *High Output* (2x150pb), ce format permettant d'obtenir une profondeur de séquençage suffisante pour l'exploration des mosaïques (ancillaire, analyse en cours).

Table 2: liste des gènes inclus dans l'étude.

AIP	*CCND1*	*FANCA**	*IGF2*	*MYCN*	*RAD50**	*SPRED1*
AKT1	*CDH1*	*FANCB**	*IKZF1*	*NBN*	*RAD51**	*STAG2*
AKT2	*CDK4*	*FANCC**	*INHA*	*NF1*	*RAD51B**	*STK11*
AKT3	*CDKN1A*	*FANCD2**	*INHBA*	*NF2*	*RAD51C**	*SUFU*
ALK	*CDKN1B*	*FANCE**	*INSR*	*NKX2-1*	*RAD51D**	*TCF7L2*
AMER1	*CDKN2A*	*FANCF**	*IRF4*	*NRAS*	*RAF1*	*TERT**
APC	*CDKN2B*	*FANCG**	*IRS1*	*NSD1*	*RB1*	*TGFBR2*
ARAF	*CDKN2C*	*FANCI**	*JAK1*	*NTRK1*	*RECQL4**	*TMEM127*
ARID1A	*CHEK2**	*FANCL**	*JAK2*	*PALB2**	*RET*	*TNFAIP3*
ARID5B	*CREBBP*	*FANCM**	*JAK3*	*PAX7*	*RINT1*	*TP53**
*ATM**	*CTNNA1*	*FAS*	*KIT*	*PDGFB*	*RNASEL*	*TP63*
*ATR**	*CTNNB1*	*FAT1*	*KMT2B*	*PDGFRA*	*RUNX1*	*TRIM33*
*ATRX**	*CYLD*	*FBXW7*	*KMT2C*	*PDGFRB*	*SBDS*	*TSC1*
AURKA	*DDB1**	*FGFR1*	*KRAS*	*PHOX2B*	*SDHA*	*TSC2*
AXIN2	*DDB2**	*FGFR2*	*MALT1*	*PIK3CA*	*SDHAF2*	*VHL*
BAP1	*DICER1*	*FGFR3*	*MAP3K1*	*PIK3CD*	*SDHB*	*WRN**
BARD1	*DNMT3A*	*FGFR4*	*MAX*	*PIK3R1*	*SDHC*	*WT1*
BCL2	*EGFR*	*FH*	*MDM2*	*PIK3R2*	*SDHD*	*XPA**
BCL2L11	*EGLN1*	*FHIT*	*MDM4*	*PML*	*SH2D1A*	*XPC**
BCL9	*EPCAM*	*FLCN*	*MEN1*	*PMS2**	*SLX4**	*XRCC2**
BIRC5	*EPHB2*	*GATA2*	*MET*	*POLE**	*SMAD3*	*XRCC3**
*BLM**	*ERBB2*	*GNA11*	*MGMT**	*POLH**	*SMAD4*	
BMPR1A	*ERCC1**	*GNAQ*	*MITF*	*POT1*	*SMARCA4*	
BRAF	*ERCC2**	*GNAS*	*MLH1**	*PPM1D*	*SMARCB1*	
*BRCA1**	*ERCC3**	*H3F3A*	*MPL*	*PRKAR1A*	*SMARCE1*	
*BRCA2**	*ERCC4**	*HNF1A*	*MRE11A**	*PTCH1*	*SMC1A*	
BRIP1	*ERCC5**	*HOXB13*	*MSH2**	*PTEN*	*SMO*	
BUB1B	*EXT1*	*HRAS*	*MSH6**	*PTPN11*	*SOS1*	
CARD11	*EXT2*	*IDH1*	*MUTYH**	*PTPRD*	*SPEN*	
CASP8	*FAM157A*	*IDH2*	*MYC*	*RAD21**	*SPOP*	

* Gènes appartenant aux voies de réparation de l'ADN.

b. Préparation des échantillons, enrichissement et séquençage à haut-débit

L'ADN des patients est extrait des leucocytes circulants sanguins, en utilisant le kit d'extraction Agencourt® Genfind V2 sur l'automate de préparation Biomek FX (Beckman, Villepinte, France). L'ADN a été fragmenté par sonication sur Covaris S2 (Covaris, Inc MS, USA). Le logiciel SureDesign (Agilent, Santa Clara, CA, USA) a été utilisé afin de créer les librairies de sondes de capture couvrant les régions exoniques (+ 25 pb dans les introns) de tous les transcrits RefSeq (NCBI) connus des 201 gènes d'intérêt, et modifiées afin de renforcer le nombre de sondes couvrant les zones GC riches. Les régions d'intérêt ont été capturées avec le protocole SureSelect XT (Agilent, Santa Clara, CA, USA). Les librairies ont ensuite été séquencées sur NextSeq 500 (Illumina, San Diego, USA).

c. Analyse bio-informatique

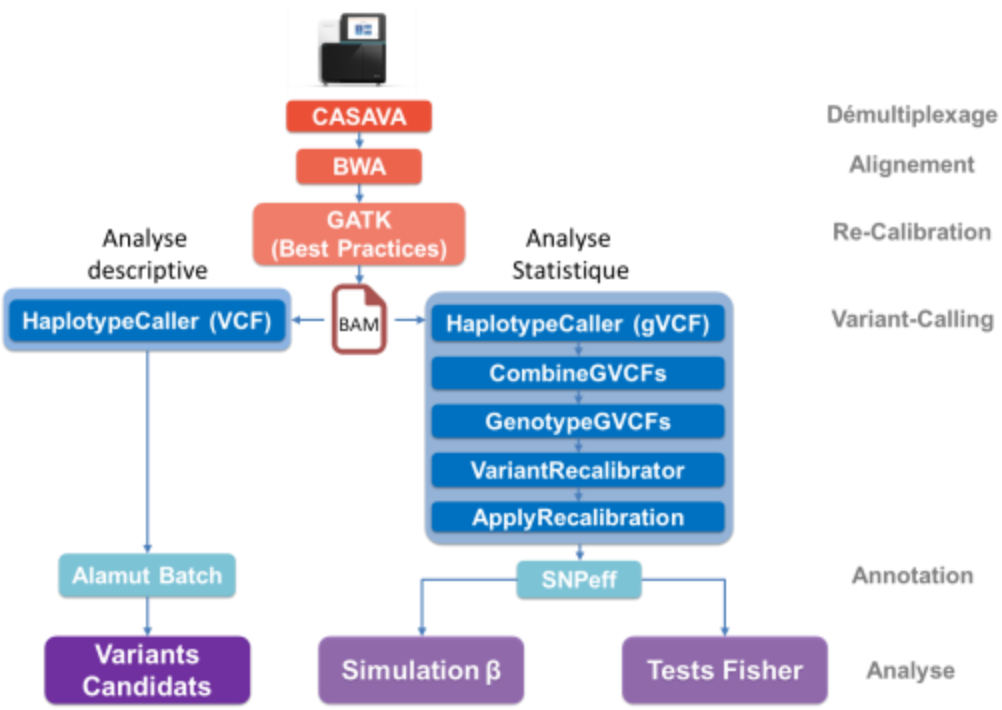


Figure 62: description du pipeline bio-informatique conduisant à une analyse descriptive et une analyse statistique.

Le démultiplexage des échantillons a été réalisé par la suite logicielle CASAVA v1.8 (Illumina). L'étape d'alignement a été réalisée par BWA 0.7.12, suivie d'une recalibration par la suite logicielle GATK (selon les recommandations du Broad Institute) afin de produire les fichiers BAM.

La suite de l'analyse se divise en 2 parties principales : pour une analyse descriptive et une analyse statistique (Figure 62).

Dans l'analyse descriptive, le variant-calling est réalisé par HaplotypeCaller (GATK) afin de produire un fichier VCF (ne contenant que les positions variantes) par patient, annoté ensuite par Alamut-Batch. Les variants reportés sont dits « Perte de Fonction » (LoF : Loss of Function), et sélectionnés pour être (i) des variants Frameshift (ii) des variants nonsense (iii) des variants affectant un site canonique d'épissage et avec (iv) une MAF < 0.1% (ExAC Non-TCGA, population Européenne non-Finlandaise = NFE).

Pour l'analyse statistique, le variant-calling est réalisé par HaplotypeCaller (GATK) afin de produire un gVCF (cf. Introduction §VII.6.a), par patient et par témoin de la population FREX. Les gVCFs sont ensuite combinés dans un seul fichier (CombineGVCF) et l'ensemble des données re-génotypées (GenotypeGVCFs). La recalibration des variants est enfin calculée à la fois pour les substitutions et les Indels. Les variants sont ici annotés par SNPeff.

d. Filtration des données pour l'analyse statistique

<u>Evaluation de la qualité de séquençage par individu:</u>

Chaque individu à prendre en compte dans l'analyse a été évalué selon la qualité de séquençage, en étudiant les SNVs passant le filtre VQSLOD (PASS) et aux génotypes GQ > 20 (données fournies par HaplotypeCaller):

- Pourcentage de génotypes manquants par individu (« missingness »)
- Le ratio de transitions / transversions (ti/tv)
- Le ratio de variants homozygotes / hétérozygotes
- Le nombre de singletons

Sélection des variants en fonction de la qualité de séquençage :

5 critères de qualité ont été fixés pour inclure les variants dans les analyses d'association :

- VQSLOD > 99.5% pour les SNVs
- VQSLOD > 99% pour les Indels
- Test d'équilibre d'Hardy-Weinberg (appliqué aux contrôles uniquement) : exclusion du variant en cas de p-valeur associée inférieure à 10^{-7}
- Balance allélique (calculée par GATK, sur SNV uniquement) : exclusion si ABHom < 90% / ABHet < 25% ou > 75% permettant l'exclusion de variants en mosaïque qui ne peuvent pas être mis en évidence avec une sensibilité équivalente entre les données de panel en forte profondeur (cas) et les données d'exomes (témoins)
- Pourcentage de génotype manquant sur l'ensemble des données : exclusion des variants pour lesquels plus de 5% des individus n'ont pas d'information.
- P-valeur du test de Fisher comparant le nombre de génotypes manquants chez les cas et chez les contrôles (sont exclus les variants dont la p-valeur est inférieure à 10^{-7})

Classification des variants pour les tests d'association :

3 Classes de variants ont été considérées pour les analyses statistiques décrites comme suit :

- LOF : Nonsense, frameshift et sites canoniques d'épissage

- LOF +SD : LOF ou les variants faux-sens (« Strictly Damaging ») prédits délétères par 3 algorithmes sur 3 (SIFT, PolyPhen2 et MutationTaster)
- MS : variants faux-sens prédits délétères par au moins 2 logiciels sur 3 algorithmes (« Missense »)

e. Analyses statistiques

Toutes les analyses ont été effectuées sur les variants recalibrés (cf. ci-avant).

Il a été réalisé une estimation par simulation en population générale de la probabilité de porter un variant pathogène et un calcul d'*odd ratio*. Ces estimations ont été réalisées tel que décrit dans la première publication présentée dans cette thèse (cf. Résultats, § I), en utilisant la population ExAC non-TCGA NFE comme population Témoin, 100 000 simulations étant réalisées pour chaque estimation. De plus une comparaison entre la population de cas de cancer précoces et la population FREX a été réalisée par un test exact de Fisher permettant de comparer les proportions de sujets portant au moins un variant rare pathogène.

C. Résultats

Les analyses de stratification de la population réalisées par PCA ont permis d'exclure 8 patientes sur la base de leur origine ethnique (n'étant pas d'origine européenne). L'analyse avec HaplotypeCaller de chacun des 110 cas index restants a permis de mettre en évidence 21 variants rares pathogènes représentés dans la Table 3.

Table 3: Description des variants pathogènes identifiés dans l'analyse descriptive. BL : bilatéral, TN : Triple négatif

Gène	cNomen	pNomen	Type de variation	Type de cancer	Présentation familiale
TP53	c.782+1G>A	p.?	Epissage	Sein BL	Sporadique
TP53	c.626_627del	p.Arg209Lysfs*6	Frameshift	Sein	Sporadique
PALB2	c.1972G>T	p.Glu658*	Nonsense	Sein	Cancer du sein
PALB2	c.2325dup	p.Phe776Ilefs*26	Frameshift	Sein TN BL	Cancer du sein
ATM	c.497-1G>A	p.?	Epissage	Sein	Cancer Prostate
ATM	c.8011-2A>C	p.?	Epissage	Sein	Cancer du sein
ATR	c.6417_6427del	p.Gln801Hisfs*13	Frameshift	Ovaire	Sporadique
BARD1	c.1741C>T	p.Gln581*	Nonsense	Sein	Cancer du sein
CHEK2	c.1229del	p.Thr410Metfs*15	Frameshift	Sein	Sporadique
CHEK2	c.720del	p.Val241Phefs*7	Frameshift	Sein	Cancer du sein
NBN	c.657_661del	p.Lys219Asnfs*15	Frameshift	Sein	Cancer Prostate
XRCC3	c.782_783del	p.Glu261Glyfs*29	Frameshift	Sein	Cancer du sein
PMS2	c.2007-2A>G	p.?	Epissage	Sein	Cancer du sein
FANCM	c.3975T>G	p.Tyr1325*	Nonsense	Sein	Sporadique
FANCL	c.1111-1114dup	p.Thr372Asnfs*13	Frameshift	Ovaire	Cancer du sein
FANCA	c.3558dup	p.aArg1187Glufs*28	Frameshift	Sein TN	Sporadique
WRN	c.3309+1G>A	p.?	Epissage	Sein + Ostéosarcome	Sporadique
IGF2	c.97C>T	p.Gln33*	Nonsense	Sein	Cancer du sein
INSR	c.2479C>T	p.Gln827*	Nonsense	Sein	Cancer du sein
PIK3R2	c.1978_1979del	p.Val660Serfs*150	Frameshift	Sein	Sporadique
RNASEL	c.793G>T	p.Glu265*	Nonsense	Sein	Sporadique

Parmi les patientes analysées présentant un variant pathogène décrit dans la Table 3, 12 présentaient un cancer précoce dans un contexte familial, et 9 patientes dans un contexte sporadique. Parmi ces cancers sporadiques, 2 variants pathogènes de *TP53* décalant le cadre de lecture ont été rapportés Ces patientes n'ayant pas été analysées pour ce gène dans le cadre de la prise en charge oncogénétique, le caractère en néo-mutation de ces variants n'a pas été évalué.

Afin de réaliser des études d'association entre variants rares pathogènes identifiés dans le panel de gènes et l'apparition précoce d'un cancer du sein ou de l'ovaire, nous avons employé deux méthodologies distinctes : (i) après estimation par simulation en population générale (ExAC) de la probabilité de porter un variant pathogène, un *odd ratio* a été calculé en regard de la probabilité dérivée des cas de cancer précoce (ii) la réalisation de Tests Exacts de Fisher comparant la population des cas à la population du french Exome Project (Table 4).

		Calcul d'Odd ratio à partir du modèle en simulation		**Test Exact de Fisher**		
		OR	IC 95%	OR	IC 95%	p-value
Voies de réparation de l'ADN	LOF	2.02	1.11-3.25	2.96	1.18-7.00	0.015
	LOF+SD	1.12	0.79-1.55	1.12	0.72-1.74	0.59
	MS	0.86	0.62-1.17	1.08	0.69-1.73	0.74
Autres voies signalisation	LOF	1.91	0.96-3.43	1.73	0.69-3.93	0.168
	LOF+SD	0.79	0.54-1.12	1.1	0.71-1.71	0.67
	MS	0.58	0.42-0.79	1.09	0.58-2.18	0.87

Table 4: Résultats des tests d'association. OR : Odd Ratio, IC : Intervalle de confiance, LOF : Loss Of Function, SD : Strictly Damaging, MS : Missense

L'analyse gène par gène ne montre aucun résultat significatif, quelle que soit la méthodologie envisagée (données non montrées). Ces tests ont été réalisés par la suite en agrégeant les variants en deux groupes distincts, séparant les gènes des voies de réparation à l'ADN et les gènes d'autres voies de signalisation, selon la liste de la Table 2.

Une association significative et concordante entre les deux méthodologies a été observée entre les variants rares de type LoF des gènes des voies de réparation de l'ADN et les cancers

précoces, avec des OR de 2.02 [1.11-3.25] pour la méthode en simulation et de 2.96 [1.18-7.00] pour le test exact de Fisher. Cet effet d'association est aboli dès que les faux-sens prédits comme pathogènes sont pris en compte avec les variants LoF (LOF+SD) ou analysés seuls (MS), pouvant signer une faible spécificité des algorithmes prédictifs utilisés. Une tendance à l'association des variants LOF présents dans les gènes non liés aux voies de réparation à l'ADN avec les phénotypes de cancer précoce est mise en évidence uniquement avec la méthodologie en simulation (OR = 1.91 [0.96-3.43]). Aucun autre effet significatif n'a été observé.

D. Discussion

Dans cette étude préliminaire, nous avons mis en évidence que 20% des patientes atteintes de cancers précoces du sein ou de l'ovaire étaient porteuses d'un variant pathogène dans un des 201 gènes séquencés. Plusieurs de ces variants ont été mis en évidence dans des gènes déjà connus pour leur implication dans le syndrome, certains présentant une pénétrance élevée établie, d'autres, à l'initiation de cette étude, induisant un niveau de risque de cancer mal connu. Ainsi, deux patientes qui avaient présenté un cancer du sein précoce sont porteuses d'un variant provoquant l'apparition d'un codon stop prématuré sur *TP53*, mais n'avaient pas bénéficié de l'analyse de ce gène au moment de leur prise en charge oncogénétique. En effet la recherche des variants de *TP53* n'était pas une indication formelle dans les cancers du sein jeunes au moment du dessin de l'étude comme elle peut l'être aujourd'hui chez les femmes atteintes de cancer du sein avant 31 ans même en l'absence de cancer pédiatrique. Il est intéressant de noter que les variants de *TP53* identifiés ici sont responsables d'une perte de fonction, ceci étant cohérent avec les observations précédemment publiées montrant que les variants de type perte de fonction de *TP53* semblent constituer une catégorie d'altérations

plus faiblement pénétrantes que les variants de type faux-sens pouvant être associés à un effet dominant négatif. Les variants perte de fonction de *TP53* sembleraient donc induire préférentiellement des histoires de cancer de l'adulte jeune éventuellement avec un tropisme de cancer sénologique sans que des cas de cancers typiques du syndrome de Li Fraumeni soient présents[242]. Dans le cas des deux patientes de notre étude, la question du retour des données de recherche cliniquement utiles à la prise en charge personnelle et familiale des individus inclus se pose, et est prévue par le protocole de recherche clinique soutenant cette étude (RCB : 2015-A01703-46). Aussi chez ces deux patientes, le caractère sporadique de l'apparition du cancer conforte le fait que certaines histoires de cancers sont provoquées par des néo-mutations pré ou post-zygotiques. Cependant la recherche de ces variants chez leurs parents n'a pas été réalisée dans cette étude et le caractère en néo-mutation de ces deux variants pathogènes de *TP53* n'a pas été démontré. De la même façon, la présence d'un variant pathogène de *PALB2* n'était pas un critère d'exclusion au moment de la conception de l'étude. L'implication des variants pathogènes de *PALB2* dans la prédisposition au cancer du sein en tant que facteur de risque élevé n'était pas encore établie et était fortement débattue[21,274]. La recherche des variants de *PALB2* est aujourd'hui inclue dans le diagnostic des prédispositions aux cancers du sein, l'association des variants pathogènes de ce gène avec les histoires familiales de cancer du sein étant confirmée, notamment dans la première publication présentée dans cette thèse avec un OR de 8.22 [4.91-13.05]. Les deux variants inactivateurs retrouvés chez des patientes dans un contexte familial de cancer du sein laissent ainsi supposer le caractère causal de ces évènements. Néanmoins la présentation très précoce de ces cancers suggère que d'autres facteurs ségrégant dans la famille du côté paternel ou maternel pourraient modifier le niveau de risque chez nos deux cas index et expliquer la précocité du trait. Ainsi un modèle oligogénique pourrait être envisagé.

L'identification de variants pathogènes sur *ATM*, *BARD1*, *CHEK2* et de manière moins établie de *PMS2* confirme leur implication dans les syndromes de prédisposition au cancer du sein ou de l'ovaire (cf. Introduction, § IV). Les niveaux de risque modérés auxquels ils sont habituellement associés[246,263] et tels que nous les avons estimés dans le premier article de cette thèse (3.20 [2.14-4.53] ; 2.00 [0.74-4.10] ; 1.67 [1.18-2.27] ; 1.16[0.47-2.24] respectivement) ne semblent pas expliquer à eux seuls la précocité d'apparition du cancer chez les patientes étudiées. Egalement la causalité du variant inactivateur identifié sur *NBN* (Nibrin) malgré son rôle dans le complexe MRN (avec *MRE11A* et *RAD50*) et dans la réparation des cassures double-brin ne semble pas pouvoir à lui seul expliquer la précocité d'apparition du cancer. Habituellement décrits dans le rare syndrome de Nijmegen, caractérisé par une ataxie-télangiectasie associée à l'inactivation bi-allélique du gène, celui-ci a déjà été évoqué dans les sur-risques de cancer et notamment du sein[275]. En effet, l'inactivation des éléments de ce complexe a pu être associée à un risque intermédiaire de développer un cancer du sein (OR=2.88, p=0.009)[276], la part de chacun des éléments dans ce sur-risque restant à préciser. Aussi, trois patientes sont porteuses d'un variant pathogènes sur *FANCA*, *FANCL* et *FANCM*, gènes responsables dans un contexte d'inactivation biallélique de l'anémie de Fanconi chez l'enfant. Cette maladie autosomique récessive se caractérise par de multiples anomalies congénitales, une insuffisance médullaire et une prédisposition au cancer. Les variants monoalléliques de *FANCM* ont déjà pu être associés à une susceptibilité au cancer du sein[249]. Des études récentes montrent que les variants pathogènes de *FANCM* sont associés à un OR de 3.75 [1.00-12.85] dans les tumeurs triple-négatives, pouvant laisser discuter de l'intégration de ce gène dans la prise en charge génétique des cancers du sein[249]. Ces observations renforcent donc le fait que la présence d'un variant inactivateur à l'état hétérozygote sur un des gènes responsables du syndrome de Fanconi est à l'origine d'un sur-risque de développer un cancer du sein, mais les niveaux de risques induits doivent être

précisés et probablement intégrés dans la définition d'un risque génétique global probablement multigénique. Enfin, le paralogue de *RAD51*, *XRCC3*, n'a pas été montré comme pouvant induire à lui seul un sur-risque de cancer, malgré son implication dans les mécanismes de recombinaison homologue (cf. Introduction, § IV.L). Nous émettons l'hypothèse que les variants de ces gènes participent à la modulation du risque et que notre étude n'a pas trouvé l'élément causal du phénotype extrême chez les porteurs de ces variants ou l'ensemble des éléments causaux responsables entrant en jeux dans un modèle oligogénique. Ce modèle oligogénique pourrait intégrer les facteurs génétiques responsables de risques faibles, modérés et forts[262], dont certains encore probablement inconnus, et mieux rendre compte des cas présentant une très forte probabilité de prédisposition au cancer du sein et de l'ovaire.

Notre étude a également mis en évidence la présence de variants pathogènes sur des gènes non répertoriés dans la pathologie mais responsables de maladies génétiques rares à transmission autosomique récessive :

WRN (*Werner Syndrome RecQ like helicase*) est un gène codant pour une hélicase participant à la réparation des cassures double-brin par recombinaison homologue et par jonction des extrémités non-homologues. L'inactivation homozygote ou hétérozygote composite de ce gène provoque le syndrome de Werner (avec une incidence comprise entre 1 / 1 000 000 et 1 / 10 000 000 naissances), caractérisé un vieillissement prématuré, ainsi que le développement de cancers multiples à un âge précoce[277] dont le cancer du sein.

IGF2 (*Insulin Like Growth Factor 2*) est un gène codant pour une hormone impliquée dans la prolifération, la croissance, la différenciation et la survie cellulaire. Le gène IGF2 est soumis à empreinte et exclusivement exprimé à partir de l'allèle paternel. La perte de l'empreinte maternelle de ce gène, conduisant à une activation bi-allélique, est une anomalie génétique à

l'origine des tumeurs de Wilms[278] (nephroblastome), cancer du rein d'apparition extrêmement précoce chez l'enfant.

La mise ne évidence de ces variants est peut-être de l'ordre de la découverte fortuite, « incidentalome » détectant par hasard les conductrices saines de ces pathologies rares. Il faut néanmoins s'interroger sur la pertinence du dépistage du conjoint en cas de projet parental et ceci doit être discuté par les investigateurs de l'étude. A l'image des variants pathogènes des gènes du Fanconi, il est aussi possible que les conductrices puissent posséder un réel risque accru de développer un cancer du sein, ce risque n'ayant cependant jamais été évalué dans ces situations extrêmes et rares. En effet la sélection de phénotypes extrêmes dans notre étude a pu conduire à la mise en évidence de facteurs génétiques de prédisposition ultra-rares et marginaux, dont la causalité sera alors difficile à démontrer par les approches statistiques.

Les études d'association ont été réalisées en agrégeant les variants pathogènes dans des voies de signalisation afin de gagner en puissance. Ces tests confirment les observations précédentes en démontrent un enrichissement en variants LOF dans les gènes impliqués dans les voies de réparation de l'ADN. Les effets d'associations ont été confirmés par les deux approches statistiques employées, confortant ainsi que le choix des gènes candidats séquencés dans cette étude est pertinent pour mettre en évidence des variants causaux dans notre population d'étude. Il existe une tendance à l'association dans les gènes non liés aux voies de réparation de l'ADN détectée par la méthode en simulation qui pourrait suggérer l'implication des gènes d'autres voies de signalisation, mais cette hypothèse doit être émise avec prudence étant donné que cette tendance n'est pas retrouvée par les tests de Fisher. L'annulation de l'effet observé lors de la prise en compte des variants faux-sens prédits comme délétères avec les variants LoF (catégorie LOF+SD) tend à faire douter de la spécificité des algorithmes de prédiction de pathogénicité utilisés, appelant à la prudence avec laquelle leurs résultats doivent être interprétés. Ces résultats interdisent toute conclusion sur le caractère pathogène

des variants faux-sens trouvés dans un contexte diagnostique et incitent à ne contraindre l'utilisation des algorithmes de prédiction qu'à un rôle de hiérarchisation simple. Globalement, les effets d'association, même lorsqu'ils sont significatifs, sont faibles, malgré l'agrégation par voie de signalisation et la puissance de notre étude est insuffisante pour réaliser les agrégations de variants au niveau du gène. Pour gagner en puissance, l'augmentation du nombre de cas étudiés est une solution mais qui montrera ses limites dès lors que nous souhaitons limiter l'étude aux phénotypes extrêmes (par définition plus rares). La mise en évidence de nouveaux facteurs de risque pourrait pourtant profiter d'une sélection encore plus rigoureuse des patientes incluses dans l'étude, excluant par exemple les patientes porteuses de facteurs de prédisposition les plus à risque déjà décrits dans le syndrome en réalisant un génotypage ciblé dédié à la sélection. L'équation est donc complexe dès lors qu'il s'agit d'allier la volonté de découvrir de nouveaux facteurs de risques fortement pénétrants, de préciser les risques induits et la puissance nécessaire à la résolution de ces objectifs. Une possibilité est le séquençage de l'ensemble des régions codantes du génome (Exome) sur notre population d'étude en agrégeant les variants dans des voies de signalisation plus précises et complètes. Eventuellement les interactions entre gènes et entre voies de signalisation pourraient être réalisées de manière non supervisées dans le but de maximiser les effets mis en évidence en y associant la mise en œuvre de réseaux dynamiques et des procédures d'apprentissage. La prise en compte des interactions gène-gène par des approches non-supervisées pourrait améliorer la compréhension de la pathologie[279]. Ces méthodes pourraient ainsi être utilisées dans l'évaluation du risque oligogénique.

Discussion générale

L'arrivée du séquençage à haut-débit a constitué une véritable rupture technologique et une révolution dans la manière d'explorer les maladies génétiques. Appliqué à l'étude des prédispositions au cancer du sein et de l'ovaire (HBOC), les débits massifs d'analyse produits par cette technologie autorisent aujourd'hui une exploration simultanée de l'ensemble des facteurs de risque génétiques connus voire encore inconnus au moment du diagnostic moléculaire lorsque l'on atteint l'échelle de l'exploration du génome entier ou de l'ensemble de ses régions codantes (Exome). Il faut néanmoins évaluer leur impact sur les niveaux de risque de plus en plus précisément et surtout l'association des risques entre eux à partir de l'étude de dizaines de milliers d'individus si le modèle oligogénique est exploré. L'interprétation de la quantité fantastique d'informations produites par ces automates constitue un ensemble de défis encore grandissants à la fois informatiques, statistiques et biologiques, et repose sur la capacité de constituer et maintenir des cohortes parfaitement phénotypées. Ces études reposeront sur des consortiums internationaux à l'image du Breast Cancer Association Consortium (BCAC) et son projet « Breast cancer RIsk after Diagnostic Gene Sequencing » (BRIDGES).

Les études réalisées sur des populations de plus en plus importantes, en panels de gènes connus ou suspectés d'être impliqués dans le syndrome de prédisposition au cancer du sein et de l'ovaire, pourraient permettre des calculs de risque individuels de plus en plus précis. La connaissance fine de ces risques est d'autant plus importante qu'il existe une explosion des tests génétiques proposés par de nombreuses grandes sociétés de biotechnologie qui se basent sur des approches de séquençage de panels de gènes larges sans que l'intérêt clinique de certains de ces gènes soit démontré, et sans réelle explication de la limite de ces tests au patient parfois assimilé à un simple consommateur. L'inclusion de dizaines de milliers d'individus dans des populations de cas comparées à des populations de témoins permet néanmoins d'affiner les niveaux de risque induits par les variants pathogènes dans les gènes

de ces panels. Ainsi, les études d'association récentes de Slavin et al. et de Couch et al. retrouvent des risques de cancer similaires à ceux que nous avons estimé dans le premier travail de cette thèse pour des gènes tels que *PALB2* (OR=6.95[3.71-12.7] vs OR=7.46[5.12-11.19]), *ATM* (OR=3.28[2.06-5.21] vs OR=2.78[2.22-3.62]) ou encore CHEK2 (OR=1.6[1.03-2.51] vs OR=1.48[1.31-1.67]) respectivement[246,263]. A court terme, il est donc probable que grâce à la formidable accumulation des données de ces niveaux de risque induits par les variants de ces gènes soient résolus. Néanmoins, il est possible que ces niveaux de risque doivent encore être affinés au regard d'informations clinico-biologiques beaucoup plus précises, sous réserve que la puissance statistique nécessaire aux études pour pouvoir différencier les risques en fonction de sous-groupes de patients soit atteignable. A titre d'exemple, nous avons ainsi pu observer une forte association des variants pathogènes de *BARD1* exclusivement dans les familles où le cas index présentait un cancer du sein triple-négatif (OR = 11.27 [3.27-25.01] IC 95% (cf. Résultats, § I.C), et il est probable que l'affinement des phénotypes puisse faire ressortir des effets d'association dans des sous-populations homogènes sélectionnées pour un phénotype précis. Il existe donc probablement des facteurs génétiques qui modulent les risques mais également l'expressivité de la maladie, comme le démontrent Kuchenbaeker et al. en évaluant les risques cumulatifs de développer un cancer du sein contralatéral 20 ans après le premier diagnostic chez les porteuses de variants sur *BRCA1* et *BRCA2*, ainsi que les différences de risque chez les patientes avec ou sans contexte familial[71].

Malgré ces avancées, une part importante des facteurs de risque génétiques à l'origine du syndrome de prédisposition au cancer du sein et de l'ovaire reste non-identifiée. La recherche de nouveaux facteurs de prédisposition génétique a donc été étendue à des panels de gènes plus larges, dans l'objectif d'identifier de nouveaux gènes portant des variants rares et induisant un risque fort à modéré. Comme dans notre étude, une série de 255 patientes avec

une histoire familiale de cancer du sein et de l'ovaire a été séquencée sur un panel de 94 gènes, dans laquelle les auteurs décrivent la mise en évidence de variants inactivateurs sur des gènes en grande majorité déjà connus pour leur implication dans le syndrome[280]. Aussi, une autre analyse d'un panel de 264 gènes impliqués dans les modifications épigénétiques sur 656 cas index provenant de familles non mutées sur les gènes de *BRCA1* et *BRCA2* a ainsi mis en évidence le gène *CHD8* (*Chromatin Helicase DNA Binding protein 8*), codant pour une protéine jouant un rôle dans la régulation transcriptionnelle, le remodelage épigénétique et la régulation de la synthèse de l'ARN. L'association à un sur-risque de développer un cancer du sein, évaluée dans un second temps par une analyse de ségrégation dans les familles porteuses, est cependant relativement faible (OR=2.48 [1.11-6.67]) et peu significative (p=0.05)[281]. Notre propre étude, proposant le séquençage d'un panel de 201 gènes chez des individus présentant les phénotypes extrêmes, identifie une part importante de gènes déjà connus pour leur implication dans le syndrome, avec des risques associés généralement décrits comme faibles à modérés, n'expliquant probablement pas à eux seuls la précocité d'apparition de la maladie. Notre étude sur 201 gènes a suggéré un enrichissement en variants délétères sur les voies de réparation à l'ADN, mais la prise en compte de l'ensemble des régions codantes pourrait permettre d'établir de meilleures définitions des voies de signalisation et augmenter la puissance de nos études en agrégation. Au final, ces études laissent penser, sans exclure qu'un facteur génétique fortement pénétrant reste à découvrir dans chacun des individus testés, que c'est la cumulation des risques identifiés (ou à identifier) qui pourrait expliquer les phénotypes extrêmes. Dès lors la mise en place de modèles pouvant intégrer l'ensemble de ces facteurs de risque, mais aussi protecteurs, semble pouvoir devenir une approche performante afin d'évaluer les niveaux de risque individuels.

Différents modèles de prédiction des risques de développer un cancer du sein ont déjà été créés, tels que le Breast Cancer Risk Assessement Tool[282] , le modèle du Breast Cancer

Surveillance Consortium[283,284], et le Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA)[285]. Ces modèles sont basés sur la démographie (âge, ethnie), l'histoire reproductive, le statut ménopausique, l'histoire familiale, le type tumoral ou encore la densité mammographique. Des scores de risque polygéniques (PRS : Polygenic Risk Score) ont aussi été créés, basés sur la combinaison de variants fréquents (SNPs) [286,287] identifiés lors d'études en GWAS, et qui pourraient permettre d'affiner et rendre compte d'une modulation du risque pour adapter les stratégies de prévention et de dépistage. La prise en compte dans ces modèles des variants pathogéniques rares, identifiés sur des gènes induisant un risque faible à modéré de développer un cancer du sein ou de l'ovaire, pourrait permettre d'évoluer vers un modèle multifactoriel aboutissant à une évaluation personnalisée des risques encore plus précise en prenant en compte l'ensemble des modulateurs du risque[262]. La qualité de ces modèles sera ainsi dépendante des informations qu'ils peuvent intégrer. Plus le nombre de facteurs de risques sera exhaustif, plus le modèle sera précis.

L'extension de la recherche des variants rares mais aussi de l'ensemble des facteurs de prédisposition à risque faible ou modéré après séquençage d'un exome voire d'un génome complet pourrait sembler être le meilleur moyen d'accéder à l'ensemble des facteurs génétiques utiles à ces calculs de risque mais encore avec la réserve que ces facteurs soient validés par l'étude de grandes populations. L'accessibilité au génome ouvre des perspectives nouvelles. En effet il est fortement probable que les variants à inclure dans ces modèles de risque ne soient pas inclus dans des régions géniques codantes. Les études d'association pangénomiques ont pu montrer qu'une large majorité des variants génétiques associés aux pathologies étaient situés dans des régions introniques ou intergéniques, représentant 88% des SNPs étudiés[288]. Ces régions contiennent des éléments fonctionnels non codants tels que des insulateurs, des *enhancers*, des opérons ou encore des *silencers*, jouant un rôle actif dans la

régulation transcriptionnelle. La perturbation de ces éléments régulateurs a déjà été décrite comme pouvant induire une prédisposition à des cancers cérébraux, en modifiant l'affinité de la liaison aux facteurs de transcription[289]. L'interaction des éléments cis-régulateurs et de leurs gènes cibles, parfois à distance, est favorisée par leur rassemblement dans une même région chromosomique : la TAD (Topological Association Domain), au sein de laquelle les interactions entre séquences par la formation de boucles de la chromatine (interactions intra-domaines) sont plus fréquentes qu'avec des séquences à l'extérieur de la TAD. La perturbation de ces régions, soit par la mutation de ses extrémités, soit par un réarrangement structural créant un point de cassure à l'intérieur d'une TAD, pourrait aussi être à l'origine d'une activation oncogénique[290]. Il existe donc de grandes régions génomiques sur lesquelles peuvent apparaître des variants rares responsables de prédispositions au cancer. Aussi, le séquençage de l'ensemble du génome donne accès à l'ensemble des évènements mutationnels et notamment un accès aux réarrangements structuraux et aux variations du nombre de copies (CNV : Copy Number Variation) non biaisées par des méthodes de capture de régions ciblées[291] (du fait d'une couverture homogène le long du génome), ce type d'évènements pouvant aussi être associé à des risques de cancer élevés ou à des modulations de risque modérées[292,293]. La prise en compte du niveau d'expression des gènes codants ou non codants dans l'étude des prédispositions au cancer doit aussi être envisagée. En effet l'analyse de certains polymorphismes identifiés par GWAS comme associés à une susceptibilité au cancer ont révélé que ces SNPs affectaient des sites régulateurs de la transcription (enhancer) de long ARN non codants (lncRNA : Long Non-Coding RNA)[294,295]. Plus récemment, Betts et al. ont pu démontrer le rôle de 2 lncRNA dans la prédisposition au cancer du sein, via leur modulation de la réponse aux dommages à l'ADN[296]. La découverte de ces facteurs potentiels de prédisposition fait cependant suite à l'analyse de SNPs fréquents en population, suggérant que d'autres éléments cis-régulateurs pourraient être impactés par la présence de variants plus

rares. L'étude du génome n'est peut-être pas suffisante pour rendre compte de l'ensemble des évènements signant un sur-risque de cancer. Ces travaux laissent suggérer également que les signatures d'expression génique pourraient être des marqueurs d'une prédisposition au cancer.

L'interprétation des informations produites par de telles analyses demeure un véritable défi et la complexité des analyses à mettre en œuvre devront aussi rendre compte des interactions gène-gène et gène-environnement. Le calcul d'un risque oligogénique (interactions gène-gène) pourra être affiné par la prise en compte des voies de signalisation impactées par les évènements identifiés par séquençage. L'intégration de ces informations permettra d'aborder plusieurs modèles d'interaction entre voies ou au sein d'une même voie de signalisation[297], tel qu'envisagé aujourd'hui dans les études de GWAS (Figure 63).

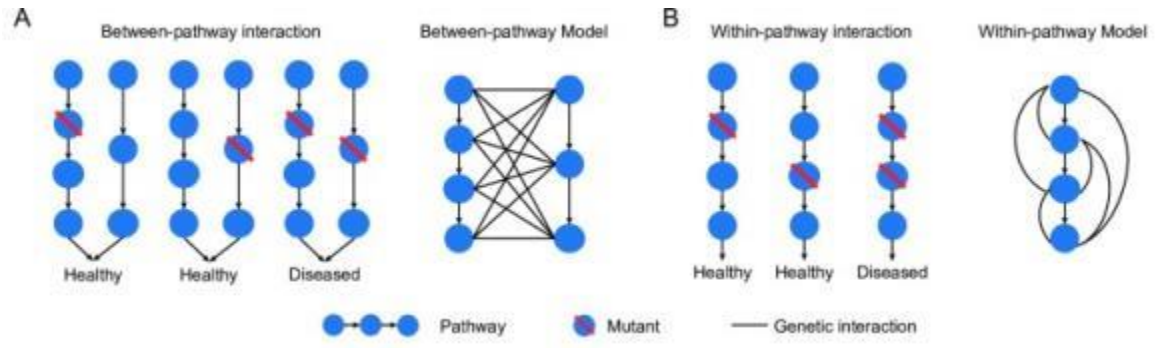


Figure 63: Modèles d'interaction des voies de signalisation[297] (A) Interactions et modèle entre voies de signalisation. Deux voies de signalisation partagent une fonction nécessaire pour le maintien d'un état de bonne santé. La présence d'un variant génétique dans une des deux voies de signalisation ne provoque pas l'apparition de la maladie, celle-ci résultant de l'inactivation conjointe des deux voies. (B) Interactions et modèle au sein d'une même voie de signalisation. L'inactivation de la voie et l'apparition de la pathologie nécessitera la perte de fonction conjointe de membres de la même voie.

Dans ces études, la prise en compte du phénotype ou de la certitude du génotype avec les données de génomique pourrait jouer un rôle déterminant dans la définition précise des risques de prédisposition. Une description précise et standardisée des composantes phénotypiques de la pathologie associée à des niveaux de certitude, en utilisant par exemple le système de classification du Human Phenotype Ontology (HPO) Project[298], pourra ensuite servir à modéliser des associations phénotype-génotype de manière non-supervisée[299]. Ces

approches trouvent leur sens dans l'exploration des maladies complexes telles que la prédisposition au cancer du sein. Ces méthodes analytiques pourraient aussi intégrer les facteurs environnementaux, estimés être à l'origine de l'acquisition de 15 à 20% des variants drivers dans les cancers du sein et de l'ovaire[300]. Si les interactions gène-environnement ont été démontrées avec des effets significatifs avec des variants fréquents, il est probable qu'elles existent en association à de forts risques de prédisposition au cancer du sein et de l'ovaire avec des variants rares[301].

Ces approches intégratives sont d'un enjeu majeur en médecine, à tel point que L'état français a lancé en 2016 un plan national dénommé « France Médecine Génomique 2025 ». Ce plan vise à intégrer la médecine génomique dans le parcours de soins du système de santé, à travers l'établissement de douze plateformes de génomique sur le territoire français. L'objectif à terme est de mettre en œuvre un parcours de soins générique avec un accès privilégié et commun à tous les patients atteints par un cancer, une maladie rare ou commune. En 2020, 235 000 génomes par an seront traités, focalisés sur les cancers et maladies rares. La montée en puissance du dispositif prendra en considération les maladies communes par la suite, ainsi que les analyses d'exomes et de transcriptomes en fonction des besoins. Ce plan devra relever des défis technologiques (traitement des données massives = « Big Data »), médicaux (interprétation, prise en charge personnalisée), organisationnels (dossier électronique médical du patient), éthiques et règlementaires (données incidentales), mais démontre bien l'importance que prend la génomique dans la prise en charge médicale.

A l'avenir, nous émettons l'hypothèse que la caractérisation du génome sur ADN circulant (ADNc) pourrait apporter des éléments pour guider la prise en charge prophylactique des patients en détectant les individus à haut-risque de cancer. Une grande profondeur de séquençage du génome entier sur ADNc (non tumoral à ce stade de prise en charge) donnerait accès de manière beaucoup plus précise aux néo-mutations en mosaïque pouvant s'accumuler

dans les différents tissus de l'organisme au cours de la vie d'un individu. Il a déjà été montré que le nombre de mosaïques augmentait avec l'âge[302,303]. Il est ainsi possible d'imaginer calculer une charge mutationnelle à partir des variants présents en mosaïque dans l'organisme, celle-ci évoluant avec l'âge de l'individu. L'observation d'une charge mutationnelle élevée à un âge précoce pourrait être associée à un sur-risque de développer un cancer, déclenchant un suivi médical adapté. L'ensemble des variants en mosaïque décrits chez l'individu pourraient aussi servir à l'établissement de signatures mutationnelles, à l'instar de celles développées par Alexandrov et al.[304] dans les tumeurs, et pourraient permettre d'identifier le mécanisme responsable de l'accumulation accélérée des néo-mutations tissulaires. La mise en œuvre d'un suivi multipliant les séquençages pourrait également être envisagée au cours de la vie de l'individu et un seuil pourrait déclencher des circuits de surveillance adaptés. Ces circuits basés sur du séquençage de génome à très forte profondeur voire utilisant du séquençage de molécule unique sont aujourd'hui hors de portée. En effet le coût des analyses et le volume des données ne permettent pas aujourd'hui la mise en œuvre d'une telle prise en charge mais l'évolution des techniques, de l'informatique et de la percée des intelligences artificielles laisse penser que ces approches pourraient être testées à moyen terme.

Les technologies de séquençage à haut-débit ont révolutionné le monde de la médecine moderne. Appliquées au décryptage des prédispositions dans les cancers du sein et de l'ovaire, elles constituent un outil d'analyse d'une puissance jamais atteinte, ayant déjà démontré leur intérêt dans l'exploration de l'hérédité manquante. La quantité et la complexité des informations générées nécessitent constamment d'imaginer de nouvelles approches et de développer des méthodes innovantes, permettant l'interprétation des mécanismes biologiques régissant des pathologies complexes. L'intégration de ces outils dans une prise en charge médicale intégrée constituera ainsi un pilier d'une médecine préventive et personnalisée.

Bibliographie

1. Molinié, F. *et al.* Trends in breast cancer incidence and mortality in France 1990–2008. *Breast Cancer Res. Treat.* **147,** 167–175 (2014).
2. *Les Cancers en France, édition 2016*. (Institut national du Cancer, 2017).
3. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136,** (2015).
4. *Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012*. (2013).
5. Beral, V., Doll, R., Hermon, C., Peto, R. & Reeves, G. Collaborative Group on Epidemiological Studies of Ovarian C. Ovarian cancer and oral contraceptives: collaborative reanalysis of data from 45 epidemiological studies including 23,257 women with ovarian cancer and 87,303 controls. *Lancet* **371,** 303–14 (2008).
6. Stratton, M. R. Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. *The Lancet* **349,** 1505–1510 (1997).
7. McCormack, V. A. & dos Santos Silva, I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol. Prev. Biomark.* **15,** 1159–1169 (2006).
8. Morrow, M., Schnitt, S. J. & Norton, L. Current management of lesions associated with an increased risk of breast cancer. *Nat. Rev. Clin. Oncol.* **12,** 227–238 (2015).
9. Brinton, L. A., Schairer, C., Hoover, R. N. & Fraumeni, J. F. Menstrual factors and risk of breast cancer. *Cancer Invest.* **6,** 245–254 (1988).
10. Kelsey, J. L., Gammon, M. D. & John, E. M. Reproductive factors and breast cancer. *Epidemiol. Rev.* **15,** 36–47 (1993).
11. Calle, E. E., Rodriguez, C., Walker-Thurmond, K. & Thun, M. J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults. *N Engl J Med* **2003,** 1625–1638 (2003).
12. Ewertz, M. *et al.* Age at first birth, parity and risk of breast cancer: A meta-analysis of 8 studies from the nordic countries. *Int. J. Cancer* **46,** 597–603 (1990).

13. Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. *The Lancet* **362,** 419–427 (2003).

14. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *The Lancet* **360,** 187–195 (2002).

15. Broca, P. *Traité des tumeurs*. **2,** (P. Asselin, 1869).

16. Bateson, W. & Mendel, G. *Mendel's principles of heredity*. (University press, 1913).

17. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci.* **68,** 820–823 (1971).

18. Claus, E. B., Risch, N. & Thompson, W. D. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am. J. Hum. Genet.* **48,** 232 (1991).

19. BRCA, S. G. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 7 (1994).

20. Wooster, R., Bignell, G., Lancaster, J. & Swift, S. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789 (1995).

21. Antoniou, A. C., Foulkes, W. D. & Tischkowitz, M. Breast-cancer risk in families with mutations in PALB2. *N. Engl. J. Med.* **371,** 1651–1652 (2014).

22. American College of Radiology. BI-RADS Committee. *Breast imaging reporting and data system*. (American College of Radiology, 1998).

23. Sonnenblick, A., De Azambuja, E., Azim Jr, H. A. & Piccart, M. An update on PARP inhibitors [mdash] moving to the adjuvant setting. *Nat. Rev. Clin. Oncol.* **12,** 27–41 (2015).

24. Blin, J. & Nowak, F. Cancer de l'ovaire et inhibiteur de PARP: Parcours des patientes en génétique oncologique. *Oncologie* 1–8 (2017).

25. Rebbeck, T. R. *et al.* Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. *J. Clin. Oncol.* **22,** 1055–1062 (2004).

26. Clark, S. L., Rodriguez, A. M., Snyder, R. R., Hankins, G. D. & Boehning, D. Structure-function of the tumor suppressor BRCA1. *Comput. Struct. Biotechnol. J.* **1,** 1–8 (2012).

27. Lorick, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl. Acad. Sci.* **96,** 11364–11369 (1999).

28. Hashizume, R. *et al.* The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. *J. Biol. Chem.* **276,** 14537–14540 (2001).

29. Henderson, B. R. Regulation of BRCA1, BRCA2 and BARD1 intracellular trafficking. *Bioessays* **27,** 884–893 (2005).

30. Scully, R. *et al.* BRCA1 is a component of the RNA polymerase II holoenzyme. *Proc. Natl. Acad. Sci.* **94,** 5605–5610 (1997).

31. Zhong, Q. *et al.* Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. *science* **285,** 747–750 (1999).

32. Bochar, D. A. *et al.* BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell* **102,** 257–265 (2000).

33. Brzovic, P. S. *et al.* Binding and recognition in the assembly of an active BRCA1/BARD1 ubiquitin-ligase complex. *Proc. Natl. Acad. Sci.* **100,** 5646–5651 (2003).

34. Atipairin, A., Canyuk, B. & Ratanaphan, A. The RING heterodimer BRCA1–BARD1 is a ubiquitin ligase inactivated by the platinum-based anticancer drugs. *Breast Cancer Res. Treat.* **126,** 203–209 (2011).

35. Mohammad, D. H. & Yaffe, M. B. 14-3-3 proteins, FHA domains and BRCT domains in the DNA damage response. *DNA Repair* **8,** 1009–1017 (2009).

36. Yamane, K., Katayama, E. & Tsuruo, T. The BRCT regions of tumor suppressor BRCA1 and of XRCC1 show DNA end binding activity with a multimerizing feature. *Biochem. Biophys. Res. Commun.* **279,** 678–684 (2000).

37. Aprelikova, O. N. *et al.* BRCA1-associated growth arrest is RB-dependent. *Proc. Natl. Acad. Sci.* **96,** 11866–11871 (1999).

38. Scully, R. *et al.* Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* **88,** 265–275 (1997).

39. Wang, Q., Zhang, H., Kajino, K. & Greene, M. I. BRCA1 binds c-Myc and inhibits its transcriptional and transforming activity in cells. *Oncogene* **17,** (1998).

40. Sy, S. M., Huen, M. S. & Chen, J. PALB2 is an integral component of the BRCA complex required for homologous recombination repair. *Proc. Natl. Acad. Sci.* **106,** 7155–7160 (2009).

41. Traven, A. & Heierhorst, J. SQ/TQ cluster domains: concentrated ATM/ATR kinase phosphorylation site regions in DNA-damage-response proteins. *Bioessays* **27,** 397–407 (2005).

42. Deng, C.-X. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Res.* **34,** 1416–1426 (2006).

43. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12,** 68–78 (2012).

44. Yang, H., Li, Q., Fan, J., Holloman, W. K. & Pavletich, N. P. The BRCA2 homologue Brh2 nucleates RAD51 filament formation at a dsDNA-ssDNA junction. *Nature* **433,** 653 (2005).

45. Carreira, A. *et al.* The BRC repeats of BRCA2 modulate the DNA-binding selectivity of RAD51. *Cell* **136,** 1032–1043 (2009).

46. Esashi, F., Christ, N., Gannon, J. & Liu, Y. CDK-dependent phosphorylation of BRCA2 as a regulatory mechanism for recombinational repair. *Nature* **434,** 598 (2005).

47. Oliver, A. W., Swift, S., Lord, C. J., Ashworth, A. & Pearl, L. H. Structural basis for recruitment of BRCA2 by PALB2. *EMBO Rep.* **10,** 990–996 (2009).

48. Buisson, R. & Masson, J.-Y. PALB2 self-interaction controls homologous recombination. *Nucleic Acids Res.* **40,** 10312–10323 (2012).

49. Buisson, R. *et al.* Cooperation of breast cancer proteins PALB2 and piccolo BRCA2 in stimulating homologous recombination. *Nat. Struct. Mol. Biol.* **17,** 1247–1254 (2010).

50. Chen, J. & Stubbe, J. Bleomycins: towards better therapeutics. *Nat. Rev. Cancer* **5,** 102–112 (2005).

51. Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol. Cell* **40,** 179–204 (2010).

52. Thompson, L. H. Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: the molecular choreography. *Mutat. Res. Mutat. Res.* **751,** 158–246 (2012).

53. Rodgers, K. & McVey, M. Error-prone repair of DNA double-strand breaks. *J. Cell. Physiol.* **231,** 15–24 (2016).

54. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16,** 110–120 (2016).

55. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155,** 27–38 (2013).

56. Frebourg, T. The challenge for the next generation of medical geneticists. *Hum. Mutat.* **35,** 909–911 (2014).

57. Castéra, L. *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.* **22,** 1305–1313 (2014).

58. Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *J. Mammary Gland Biol. Neoplasia* **9,** 221–236 (2004).

59. de la Hoya, M. *et al.* Combined genetic and splicing analysis of BRCA1 c.[594-2A> C; 641A> G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum. Mol. Genet.* **25,** 2256–2268 (2016).

60. Spurdle, A. B. *et al.* ENIGMA—Evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum. Mutat.* **33,** 2–7 (2012).

61. Goldgar, D. E. *et al.* Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Hum. Mutat.* **29,** 1265–1272 (2008).

62. Walker, L. C. *et al.* Evaluation of a 5-Tier Scheme Proposed for Classification of Sequence Variants Using Bioinformatic and Splicing Assay Data: Inter-Reviewer Variability and Promotion of Minimum Reporting Guidelines. *Hum. Mutat.* **34,** 1424–1431 (2013).

63. Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10,** 294–300 (2008).

64. Caputo, S. *et al.* Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. *Nucleic Acids Res* **40,** D992-1002 (2012).

65. Foulkes, W. D. Inherited susceptibility to common cancers. *N. Engl. J. Med.* **359,** 2143–2153 (2008).

66. Kobayashi, H., Ohno, S., Sasaki, Y. & Matsuura, M. Hereditary breast and ovarian cancer susceptibility genes. *Oncol. Rep.* **30,** 1019–1029 (2013).

67. Southey, M. C. *et al.* PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. *J. Med. Genet.* **53,** 800–811 (2016).

68. Ramus, S. J. *et al.* Germline Mutations in the BRIP1, BARD1, PALB2, and NBN Genes in Women With Ovarian Cancer. *J. Natl. Cancer Inst.* **107,** (2015).

69. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38,** 1239 (2006).

70. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47,** 373–380 (2015).

71. Kuchenbaecker, K. B. *et al.* Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *Jama* **317,** 2402–2416 (2017).

72. Tai, Y. C., Domchek, S., Parmigiani, G. & Chen, S. Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.* **99,** 1811–1814 (2007).

73. Liede, A., Karlan, B. Y. & Narod, S. A. Cancer risks for male carriers of germline mutations in BRCA1 or BRCA2: a review of the literature. *J. Clin. Oncol.* **22,** 735–742 (2004).

74. Li, F. P. & Fraumeni, J. F. Soft-tissue sarcomas, breast cancer, and other NeoplasmsA familial syndrome? *Ann. Intern. Med.* **71,** 747–752 (1969).

75. Malkin, D. *et al.* Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *science* 1233–1238 (1990).

76. Lane, D. P. p53, guardian of the genome. *Nature* **358,** 15–16 (1992).

77. Pflaum, J., Schlosser, S. & Müller, M. p53 family and cellular stress responses in cancer. *Front. Oncol.* **4,** (2014).

78. Hwang, S.-J., Lozano, G., Amos, C. I. & Strong, L. C. Germline p53 mutations in a cohort with childhood sarcoma: sex differences in cancer risk. *Am. J. Hum. Genet.* **72,** 975–983 (2003).

79. Ruijs, M. W. *et al.* TP53 germline mutation testing in 180 families suspected of Li–Fraumeni syndrome: mutation detection rate and relative frequency of cancers in different familial phenotypes. *J. Med. Genet.* **47,** 421–428 (2010).

80. Lloyd, K. M. & Dennis, M. Cowden's DiseaseA Possible New Symptom Complex with Multiple System Involvement. *Ann. Intern. Med.* **58,** 136–142 (1963).

81. Nelen, M. R. *et al.* Germline mutations in the PTEN/MMAC1 gene in patients with Cowden disease. *Hum. Mol. Genet.* **6,** 1383–1387 (1997).

82. Leslie, N. R. & Downes, C. P. PTEN function: how normal cells control it and tumour cells lose it. *Biochem. J.* **382,** 1–11 (2004).

83. Uppal, S., Mistry, D. & Coatesworth, A. Cowden disease: a review. *Int. J. Clin. Pract.* **61,** 645–652 (2007).

84. Pilarski, R. *et al.* Cowden syndrome and the PTEN hamartoma tumor syndrome: systematic review and revised diagnostic criteria. *J. Natl. Cancer Inst.* **105,** 1607–1616 (2013).

85. Xia, B. *et al.* Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol. Cell* **22,** 719–729 (2006).

86. Xia, B. *et al.* Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. *Nat. Genet.* **39,** 159 (2007).

87. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39,** 165 (2007).

88. Beggs, A. *et al.* Peutz–Jeghers syndrome: a systematic review and recommendations for management. *Gut* **59,** 975–986 (2010).

89. Hemminki, A., Markie, D., Tomlinson, I. & Avizienyte, E. A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature* **391,** 184 (1998).

90. Schumacher, V. *et al.* STK11 genotyping and cancer risk in Peutz-Jeghers syndrome. *J. Med. Genet.* **42,** 428–435 (2005).

91. Hearle, N. *et al.* Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. *Clin. Cancer Res.* **12,** 3209–3215 (2006).

92. Giardiello, F. M. *et al.* Very high risk of cancer in familial Peutz–Jeghers syndrome. *Gastroenterology* **119,** 1447–1453 (2000).

93. Kastrinos, F. & Stoffel, E. M. History, genetics, and strategies for cancer prevention in Lynch syndrome. *Clin. Gastroenterol. Hepatol.* **12,** 715–727 (2014).

94. Thibodeau, S. N., Bren, G. & Schaid, D. Microsatellite instability in cancer of the proximal colon. *Sci.-N. Y. THEN Wash.-* **260,** 816–816 (1993).

95. Boland, C. R. Evolution of the nomenclature for the hereditary colorectal cancer syndromes. *Fam. Cancer* **4,** 211–218 (2005).

96. Xiao, X., Melton, D. W. & Gourley, C. Mismatch repair deficiency in ovarian cancer—molecular characteristics and clinical implications. *Gynecol. Oncol.* **132,** 506–512 (2014).

97. Nakamura, K. *et al.* Features of ovarian cancer in Lynch syndrome. *Mol. Clin. Oncol.* **2,** 909–916 (2014).

98. Fitzgerald, R. C. *et al.* Hereditary diffuse gastric cancer: updated consensus guidelines for clinical management and directions for future research. *J. Med. Genet.* **47,** 436–444 (2010).

99. Pharoah, P. D., Guilford, P. & Caldas, C. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* **121,** 1348–1353 (2001).

100. Wijnhoven, B., Dinjens, W. & Pignatelli, M. E-cadherin—catenin cell—cell adhesion complex and human cancer. *Br. J. Surg.* **87,** 992–1005 (2000).

101. Pećina-Šlaus, N. Tumor suppressor gene E-cadherin and its role in normal and malignant cells. *Cancer Cell Int.* **3,** 17 (2003).

102. Lavin, M. F. & Shiloh, Y. The genetic defect in ataxia-telangiectasia. *Annu. Rev. Immunol.* **15,** 177–202 (1997).

103. Morrell, D., Cromartie, E. & Swift, M. Mortality and Cancer Incidence in 263 Patients With Ataxia-Telangiectasia 2. *J. Natl. Cancer Inst.* **77,** 89–92 (1986).

104. Kruhlak, M. J. *et al.* Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *J Cell Biol* **172,** 823–834 (2006).

105. Thompson, D. *et al.* Cancer risks and mortality in heterozygous ATM mutation carriers. *J. Natl. Cancer Inst.* **97,** 813–822 (2005).

106. Goldgar, D. E. *et al.* Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res.* **13,** R73 (2011).

107. Shieh, S.-Y., Ahn, J., Tamai, K., Taya, Y. & Prives, C. The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites. *Genes Dev.* **14,** 289–300 (2000).

108. Jong-Soo, L., Collins, K. M., Brown, A. L., Chang-Hun, L. & Chung, J. H. hCds-1 mediated phosphorylation of BRCA1 regulates the DNA damage response. *Nature* **404,** 201 (2000).

109. Matsuoka, S., Huang, M. & Elledge, S. J. Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. *Science* **282,** 1893–1897 (1998).

110. Bell, D. W. *et al.* Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science* **286,** 2528–2531 (1999).

111. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2* 1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* **31,** 55 (2002).

112. Schutte, M. *et al.* Variants in CHEK2 other than 1100delC do not make a major contribution to breast cancer susceptibility. *Am. J. Hum. Genet.* **72,** 1023–1028 (2003).

113. Schmidt, M. K. *et al.* Age-and Tumor Subtype–Specific Breast Cancer Risk Estimates for CH EK 2* 1100delC Carriers. *J. Clin. Oncol.* **34,** 2750–2760 (2016).

114. Cantor, S. B. *et al.* BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. *Cell* **105,** 149–160 (2001).

115. Yu, X., Chini, C. C. S., He, M., Mer, G. & Chen, J. The BRCT domain is a phospho-protein binding domain. *Science* **302,** 639–642 (2003).

116. Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. *Nat. Genet.* **37,** 934 (2005).

117. Easton, D. F. *et al.* No evidence that protein truncating variants in BRIP1 are associated with breast cancer risk: implications for gene panel testing. *J. Med. Genet.* **53,** 298–309 (2016).

118. Johnson, R. D., Liu, N. & Jasin, M. Mammalian XRCC2 promotes the repair of DNA double-strand breaks by homologous recombination. *Nature* **401,** 397 (1999).

119. Pierce, A. J., Johnson, R. D., Thompson, L. H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13,** 2633–2638 (1999).

120. Tambini, C. E. *et al.* TheXRCC2DNA Repair Gene: Identification of a Positional Candidate. *Genomics* **41,** 84–92 (1997).

121. Tebbs, R. S. *et al.* Correction of chromosomal instability and sensitivity to diverse mutagens by a cloned cDNA of the XRCC3 DNA repair gene. *Proc. Natl. Acad. Sci.* **92,** 6354–6358 (1995).

122. Liu, N. *et al.* XRCC2 and XRCC3, new human Rad51-family members, promote chromosome stability and protect against DNA cross-links and other damages. *Mol. Cell* **1,** 783–793 (1998).

123. Albala, J. S. *et al.* Identification of a novel humanRAD51Homolog, RAD51B. *Genomics* **46,** 476–479 (1997).

124. Dosanjh, M. K. *et al.* Isolation and characterization of RAD51C, a new human member of the RAD51 family of related genes. *Nucleic Acids Res.* **26,** 1179–1184 (1998).

125. Pittman, D. L., Weinberg, L. R. & Schimenti, J. C. Identification, Characterization, and Genetic Mapping ofRad51d, a New Mouse and HumanRAD51/RecA-Related Gene. *Genomics* **49,** 103–111 (1998).

126. Genois, M.-M. *et al.* Roles of Rad51 paralogs for promoting homologous recombination in Leishmania infantum. *Nucleic Acids Res.* **43,** 2701–2715 (2015).

127. Chun, J., Buechelmaier, E. S. & Powell, S. N. Rad51 paralog complexes BCDX2 and CX3 act at different stages in the BRCA1-BRCA2-dependent homologous recombination pathway. *Mol. Cell. Biol.* **33,** 387–395 (2013).

128. Meindl, A. *et al.* Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat Genet* **42,** 410–414 (2010).

129. Loveday, C. *et al.* Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat. Genet.* **43,** 879–882 (2011).

130. Song, H. *et al.* Contribution of Germline Mutations in the RAD51B, RAD51C, and RAD51D Genes to Ovarian Cancer in the Population. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **33,** 2901–2907 (2015).

131. Park, D. *et al.* Rare mutations in XRCC2 increase the risk of breast cancer. *Am. J. Hum. Genet.* **90,** 734–739 (2012).

132. He, X.-F. *et al.* Association between the XRCC3 polymorphisms and breast cancer risk: meta-analysis based on case–control studies. *Mol. Biol. Rep.* **39,** 5125–5134 (2012).

133. Wu, L. C. *et al.* Identification of a RING protein that can interact in vivo with the BRCA1 gene product. *Nat. Genet.* **14,** 430–440 (1996).

134. McCarthy, E. E., Celebi, J. T., Baer, R. & Ludwig, T. Loss of Bard1, the heterodimeric partner of the Brca1 tumor suppressor, results in early embryonic lethality and chromosomal instability. *Mol. Cell. Biol.* **23,** 5056–5063 (2003).

135. Irminger-Finger, I., Ratajska, M. & Pilyugin, M. New concepts on BARD1: Regulator of BRCA pathways and beyond. *Int. J. Biochem. Cell Biol.* **72,** 1–17 (2016).

136. Easton, D., Bishop, D., Ford, D. & Crockford, G. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **52,** 678 (1993).

137. Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* **38,** 873–875 (2006).

138. Mavaddat, N., Antoniou, A. C., Easton, D. F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. *Mol. Oncol.* **4,** 174–191 (2010).

139. Labbe, A. De la statistique à la génétique: identifier les gènes responsables de maladies complexes. *Bull. AMQ* **53,** (2013).

140. Eggen, A. Cartographie fine d'un gène et clonage positionnel. *Prod. Anim. HS 2000 133-1362000* (2000).

141. Antoniou, A. *et al.* A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* **86,** 76 (2002).

142. What are single nucleotide polymorphisms (SNPs)? - Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/primer/genomicresearch/snp. (Accessed: 17th August 2017)

143. Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. Mapping complex disease loci in whole-genome association studies. *Nature* **429,** 446 (2004).

144. Easton, D. F. & Eeles, R. A. Genome-wide association studies in cancer. *Hum. Mol. Genet.* **17,** R109–R115 (2008).

145. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6,** 109–118 (2005).

146. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45,** 353–361 (2013).

147. Snape, K. *et al.* Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res. Treat.* **134,** 429–433 (2012).

148. Kiiski, J. I. *et al.* Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proc. Natl. Acad. Sci.* **111,** 15172–15177 (2014).

149. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11,** 773–785 (2010).

150. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44,** 623–630 (2012).

151. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).

152. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).

153. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344 (2014).

154. Van Minkelen, R. *et al.* A clinical and genetic overview of 18 years neurofibromatosis type 1 molecular diagnostics in the Netherlands. *Clin. Genet.* **85,** 318–327 (2014).

155. Golmard, L. *et al.* Breast and ovarian cancer predisposition due to de novo BRCA1 and BRCA2 mutations. *Oncogene* **35,** 1324–1327 (2016).

156. Chen, Z. *et al.* Enhanced sensitivity for detection of low-level germline mosaic RB1 mutations in sporadic retinoblastoma cases using deep semiconductor sequencing. *Hum. Mutat.* **35,** 384–391 (2014).

157. Delon, I. *et al.* A germline mosaic BRCA1 exon deletion in a woman with bilateral basal-like breast cancer. *Clin. Genet.* **84,** 297–299 (2013).

158. Friedman, E. *et al.* Low-level constitutional mosaicism of a de novoBRCA1 gene mutation. *Br. J. Cancer* **112,** 765 (2015).

159. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

160. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

161. Mardis, E. R. A decade/'s perspective on DNA sequencing technology. *Nature* **470,** 198–203 (2011).

162. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

163. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).

164. Samorodnitsky, E. *et al.* Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum. Mutat.* **36,** 903–914 (2015).

165. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

166. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).

167. Smith Waterman algorithm.

168. Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227,** 1435–1441 (1985).

169. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

170. Ortet, P. & Bastien, O. Where Does the Alignment score Distribution shape come from? *Evol. Bioinforma.* **6,** 159 (2010).

171. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *nature* **456,** 53–59 (2008).

172. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714 (2008).

173. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).

174. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

175. Burrows, M. & Wheeler, D. J. A block-sorting lossless data compression algorithm. (1994).

176. Patwardhan, R. burrows-wheeler alignment algorithm. Available at: https://www.google.fr/search?q=burrows-wheeler+alignment+algorithm+googol&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&dcr=0&ei=b4LjWd2LCOj-8AeQ65ugBA. (Accessed: 15th October 2017)

177. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

178. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12,** 443 (2011).

179. Thankaswamy-Kosalai, S., Sen, P. & Nookaew, I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* (2017).

180. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

181. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

182. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

183. *hts-specs: Specifications of SAM/BAM and related high-throughput sequencing file formats*. (samtools, 2017).

184. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. (Cambridge university press, 1998).

185. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

186. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28,** 311–317 (2012).

187. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28,** 1811–1817 (2012).

188. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285 (2009).

189. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

190. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* gks918 (2012).

191. Tavtigian, S. V. *et al.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43,** 295–305 (2006).

192. Vallée, M. P. *et al.* Adding in silico assessment of potential splice aberration to the integrated evaluation of brca gene unclassified variants. *Hum. Mutat.* **37,** 627–639 (2016).

193. Lindor, N. M. *et al.* A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Hum. Mutat.* **33,** 8–21 (2012).

194. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

195. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11,** 863–874 (2001).

196. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

197. Sunyaev, S. R. *et al.* PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* **12,** 387–394 (1999).

198. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7,** 575–576 (2010).

199. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11,** 377–394 (2004).

200. Shapiro, M. B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15,** 7155–7174 (1987).

201. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37,** e67–e67 (2009).

202. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24,** 2125–2137 (2014).

203. Wei, P., Liu, X. & Fu, Y.-X. Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study. in **5,** S20 (BioMed Central, 2011).

204. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32,** 358–368 (2011).

205. Flanagan, S. E., Patch, A.-M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomark.* **14,** 533–537 (2010).

206. Houdayer, C. *et al.* Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum Mutat* **33,** 1228–1238 (2012).

207. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).

208. Béroud, C. *et al.* BRCA Share: a collection of clinical BRCA gene variants. *Hum. Mutat.* **37,** 1318–1328 (2016).

209. Online Research Resources Developed at NHGRI. *Online Research Resources Developed at NHGRI* Available at: https://research.nhgri.nih.gov/. (Accessed: 20th September 2017)

210. Bouaoun, L. *et al.* TP53 variations in human cancers: new lessons from the IARC TP53 database and genomics data. *Hum. Mutat.* **37,** 865–876 (2016).

211. Plazzer, J.-P. *et al.* The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam. Cancer* **12,** 175–180 (2013).

212. Stenson, P. D. *et al.* Human gene mutation database (HGMD®): 2003 update. *Hum. Mutat.* **21,** 577–581 (2003).

213. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2013).

214. Fokkema, I. F. *et al.* LOVD v. 2.0: the next generation in gene variant databases. *Hum. Mutat.* **32,** 557–563 (2011).

215. Ainscough, B. J. *et al.* DoCM: a database of curated mutations in cancer. *Nat. Methods* **13,** 806 (2016).

216. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49,** 170–174 (2017).

217. Vail, P. J. *et al.* Comparison of locus-specific databases for BRCA1 and BRCA2 variants reveals disparity in variant classification within and among databases. *J. Community Genet.* **6,** 351–359 (2015).

218. Shen, H. *et al.* Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. *PLoS One* **8,** e59494 (2013).

219. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

220. Musumeci, L. *et al.* Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat.* **31,** 67–73 (2010).

221. Gibbs, R. A. *et al.* The international HapMap project. (2003).

222. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68 (2015).

223. Exome Variant Server. Available at: http://evs.gs.washington.edu/EVS/. (Accessed: 22nd September 2017)

224. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* (2015).

225. De Baets, G. *et al.* SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* **40,** D935–D939 (2012).

226. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17,** 122 (2016).

227. Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36,** (2015).

228. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).

229. Alamut Batch: logiciel d'annotation haut-débit de variants. *Interactive Biosoftware* Available at: http://www.interactive-biosoftware.com/fr/alamut-batch/. (Accessed: 7th September 2017)

230. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6,** 26 (2014).

231. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

232. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).

233. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18,** 77 (2017).

234. Buys, S. S. *et al.* A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes. *Cancer* **123,** 1721–1730 (2017).

235. Norquist, B. M. *et al.* Inherited mutations in women with ovarian carcinoma. *JAMA Oncol.* **2,** 482–490 (2016).

236. Easton, D. F. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372,** 2243–2257 (2015).

237. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994).

238. Lee, J. C. & Sabavala, D. J. Bayesian Estimation and Prediction for the Beta-Binomial Model. *J Bus Econ Stat* **5,** 357–367 (1987).

239. Buys, S. S. *et al.* A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes. *Cancer* (2017). doi:10.1002/cncr.30498

240. Shirts, B. H. *et al.* Improving performance of multigene panels for genomic analysis of cancer predisposition. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18,** 974–981 (2016).

241. Susswein, L. R. *et al.* Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18,** 823–832 (2016).

242. Bougeard, G. *et al.* Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **33,** 2345–2352 (2015).

243. Li, J. *et al.* Targeted massively parallel sequencing of a panel of putative breast cancer susceptibility genes in a large cohort of multiple-case breast and ovarian cancer families. *J. Med. Genet.* **53,** 34–42 (2016).

244. Thompson, E. R. *et al.* Panel Testing for Familial Breast Cancer: Calibrating the Tension Between Research and Clinical Care. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **34,** 1455–1459 (2016).

245. Antoniou, A. C. *et al.* Breast-cancer risk in families with mutations in PALB2. *N. Engl. J. Med.* **371,** 497–506 (2014).

246. Couch, F. J. *et al.* Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer. *JAMA Oncol.* (2017). doi:10.1001/jamaoncol.2017.0424

247. Norquist, B. M. *et al.* Inherited Mutations in Women With Ovarian Carcinoma. *JAMA Oncol.* **2,** 482–490 (2016).

248. Young, E. L. *et al.* Multigene testing of moderate-risk genes: be mindful of the missense. *J. Med. Genet.* **53,** 366–376 (2016).

249. Neidhardt, G. *et al.* Association Between Loss-of-Function Mutations Within the FANCM Gene and Early-Onset Familial Breast Cancer. *JAMA Oncol.* (2016). doi:10.1001/jamaoncol.2016.5592

250. Nicolas, G. *et al.* SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. *Mol. Psychiatry* **21,** 831–836 (2016).

251. Craig, D. W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5,** 887–893 (2008).

252. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17,** 405–424 (2015).

253. Mavaddat, N. *et al.* Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. *J. Natl. Cancer Inst.* **105,** 812–822 (2013).

254. Win, A. K. *et al.* Risks of colorectal and other cancers after endometrial cancer for women with Lynch syndrome. *J. Natl. Cancer Inst.* **105,** 274–279 (2013).

255. Loveday, C. *et al.* Germline RAD51C mutations confer susceptibility to ovarian cancer. *Nat Genet* **44,** 475–476 (2012).

256. Sopik, V., Akbari, M. R. & Narod, S. A. Genetic testing for RAD51C mutations: in the clinic and community. *Clin. Genet.* **88,** 303–312 (2015).

257. Leedom, T. P. *et al.* Breast cancer risk is similar for CHEK2 founder and non-founder mutation carriers. *Cancer Genet.* **209,** 403–407 (2016).

258. Muranen, T. A. *et al.* Genetic modifiers of CHEK2*1100delC-associated breast cancer risk. *Genet. Med. Off. J. Am. Coll. Med. Genet.* (2016). doi:10.1038/gim.2016.147

259. Easton, D. F. *et al.* No evidence that protein truncating variants in BRIP1 are associated with breast cancer risk: implications for gene panel testing. *J. Med. Genet.* **53,** 298–309 (2016).

260. Rafnar, T. *et al.* Mutations in BRIP1 confer high risk of ovarian cancer. *Nat. Genet.* **43,** 1104–1107 (2011).

261. Rebbeck, T. R. *et al.* Modification of ovarian cancer risk by BRCA1/2-interacting genes in a multicenter cohort of BRCA1/2 mutation carriers. *Cancer Res.* **69,** 5801–5810 (2009).

262. Shieh, Y. *et al.* Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *J. Natl. Cancer Inst.* **109,** (2017).

263. Slavin, T. P. *et al.* The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. *NPJ Breast Cancer* **3,** (2017).

264. Muller, E. *et al.* OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. *Oncotarget* **7,** 79485–79493 (2016).

265. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13,** 666 (2012).

266. Ruark, E. *et al.* Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* **493,** 406 (2013).

267. Fernández, L. C., Torres, M. & Real, F. X. Somatic mosaicism: on the road to cancer. *Nat. Rev. Cancer* **16,** 43–55 (2016).

268. Song, W. *et al.* Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genet. Med.* **18,** 850 (2016).

269. Tournier, I. *et al.* Germline Mutations of Inhibins in Early-Onset Ovarian Epithelial Tumors. *Hum. Mutat.* **35,** 294–297 (2014).

270. Vogelstein, B. *et al.* Cancer genome landscapes. *science* **339,** 1546–1558 (2013).

271. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

272. Rahman, N. Realizing the promise of cancer predisposition genes (vol 505, pg 302, 2014). *Nature* **510,** 176–176 (2014).

273. Muller, E. *et al.* Genetic profiles of cervical tumors by high-throughput sequencing for personalized medical care. *Cancer Med.* **4,** 1484–1493 (2015).

274. Decker, B. *et al.* Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks. *J. Med. Genet.* jmedgenet-2017 (2017).

275. Zhang, G., Zeng, Y., Liu, Z. & Wei, W. Significant association between Nijmegen breakage syndrome 1 657del5 polymorphism and breast cancer risk. *Tumor Biol.* **34,** 2753–2757 (2013).

276. Damiola, F. *et al.* Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study. *Breast Cancer Res.* **16,** R58 (2014).

277. Shamanna, R. A., Croteau, D. L., Lee, J.-H. & Bohr, V. A. Recent Advances in Understanding Werner Syndrome. *F1000Research* **6,** (2017).

278. Bjornsson, H. T. *et al.* Epigenetic specificity of loss of imprinting of the IGF2 gene in Wilms tumors. *J. Natl. Cancer Inst.* **99,** 1270–1273 (2007).

279. Sokolov, A., Carlin, D. E., Paull, E. O., Baertsch, R. & Stuart, J. M. Pathway-based genomics prediction using generalized elastic net. *PLoS Comput. Biol.* **12,** e1004790 (2016).

280. Tedaldi, G. *et al.* Multiple-gene panel analysis in a case series of 255 women with hereditary breast and ovarian cancer. *Oncotarget* **8,** 47064 (2017).

281. Li, J. *et al.* Panel sequencing of 264 candidate susceptibility genes and segregation analysis in a cohort of non-BRCA1, non-BRCA2 breast cancer families. *Breast Cancer Res. Treat.* 1–13 (2017).

282. Rockhill, B., Spiegelman, D., Byrne, C., Hunter, D. J. & Colditz, G. A. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J. Natl. Cancer Inst.* **93,** 358–366 (2001).

283. Tice, J. A. *et al.* Using Clinical Factors and Mammographic Breast Density to Estimate Breast Cancer Risk: Development and Validation of a New Predictive ModelUsing Clinical Factors and Mammographic Breast Density to Estimate Breast Cancer Risk. *Ann. Intern. Med.* **148,** 337–347 (2008).

284. Tice, J. A. *et al.* Breast density and benign breast disease: risk assessment to identify women at high risk of breast cancer. *J. Clin. Oncol.* **33,** 3137–3143 (2015).

285. Lee, A. *et al.* BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *Br. J. Cancer* **110,** 535–545 (2014).

286. Dite, G. S. *et al.* Breast cancer risk prediction using clinical models and 77 independent risk-associated SNPs for women aged under 50 years: Australian Breast Cancer Family Registry. *Cancer Epidemiol. Prev. Biomark.* **25,** 359–365 (2016).

287. Mavaddat, N. *et al.* Prediction of breast cancer risk based on profiling with common genetic variants. *JNCI J. Natl. Cancer Inst.* **107,** (2015).

288. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* **106,** 9362–9367 (2009).

289. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528,** 418 (2015).

290. Valton, A.-L. & Dekker, J. TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36,** 34–40 (2016).

291. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Hum. Genet.* **135,** 359–362 (2016).

292. Park, R. W., Kim, T.-M., Kasif, S. & Park, P. J. Identification of rare germline copy number variations over-represented in five human cancer types. *Mol. Cancer* **14,** 25 (2015).

293. Sapkota, Y., Narasimhan, A., Kumaran, M., Sehrawat, B. S. & Damaraju, S. A Genome-Wide Association Study to Identify Potential Germline Copy Number Variants for Sporadic Breast Cancer Susceptibility. *Cytogenet. Genome Res.* **149,** 156–164 (2016).

294. Jendrzejewski, J. *et al.* The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci.* **109,** 8646–8651 (2012).

295. Chung, S. *et al.* Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.* **102,** 245–252 (2011).

296. Betts, J. A. *et al.* Long Noncoding RNAs CUPID1 and CUPID2 Mediate Breast Cancer Risk at 11q13 by Modulating the Response to DNA Damage. *Am. J. Hum. Genet.* **101,** 255–266 (2017).

297. Wang, W. *et al.* Pathway-based discovery of genetic interactions in breast cancer. *PLoS Genet.* **13,** e1006973 (2017).

298. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42,** D966–D974 (2013).

299. Greene, D., Richardson, S., Turro, E. & BioResource, N. Phenotype similarity regression for identifying the genetic determinants of rare diseases. *Am. J. Hum. Genet.* **98,** 490–499 (2016).

300. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355,** 1330–1334 (2017).

301. Rudolph, A., Chang-Claude, J. & Schmidt, M. K. Gene–environment interaction and risk of breast cancer. *Br. J. Cancer* **114,** 125–133 (2016).

302. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44,** 651–658 (2012).

303. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44,** 642–650 (2012).

304. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

Titre de la Thèse :

Les défis du séquençage à haut-débit dans l'exploration des prédispositions génétiques aux cancers du sein et/ou de l'ovaire

Résumé :

Les cancers du sein et de l'ovaire apparaissent dans 5 à 10% dans un contexte de prédisposition génétique, dont seule une faible part est expliquée par la présence d'un variant pathogène sur les gènes *BRCA1*, *BRCA2* et *PALB2*. Le séquençage à haut-débit permet d'explorer cette hérédité manquante, mais représente un nouveau défi à la fois informatique, statistique et biologique. Trois approches utilisant cette nouvelle technologie ont été employées pour rechercher de nouveaux facteurs de prédisposition.

En premier lieu, les risques associés à 34 gènes connus ou suspectés d'être impliqués dans les prédispositions ont été estimés à partir de l'analyse de 5 131 cas index et le développement d'une nouvelle approche statistique. Aussi la participation des néo-mutations en mosaïque dans le syndrome a été explorée à partir de 1 750 cas index issus de l'étude précédente, avec un logiciel de détection des variants faiblement représentés développé spécifiquement: outLyzer. Enfin, l'exploration par séquençage de l'hérédité manquante a été étendue à un panel de 201 gènes impliqués dans le cancer, à partir de 118 patientes sélectionnées pour la précocité d'apparition de leur maladie, élément fortement évocateur d'un facteur de prédisposition.

Les résultats de ces travaux ont permis de valider la pertinence de l'étude de *PALB2*, *RAD51C* et *RAD51D* pour la prise en charge des patients, et suggèrent aussi une implication sous-estimée des variants en mosaïque. Cependant il reste encore très probablement d'autres facteurs génétiques fortement pénétrants à découvrir mais dont la modulation du risque répond à un modèle oligogénique.

Mots clés : prédisposition héréditaires au cancer du sein et/ou de l'ovaire, séquençage de nouvelle génération, séquençage à haut-débit, bio-informatique, mosaïque, variant-calling

Title of the Thesis:

Challenges of Next Generation Sequencing in the exploration of genetic predispositions to breast and/or ovarian cancers

Abstract :

Breast and ovarian cancers appear in 5 to 10% of cases in a context of genetic predisposition, of which only a small proportion is explained by the presence of a pathogenic variant on the BRCA1, BRCA2 and PALB2 genes. High throughput sequencing can explore this missing heredity, but represents a new challenge both in computing, statistics and biology. Three approaches using this new technology have been used to investigate new predisposition factors.

First, the risks associated with 34 known or suspected genes involved in predispositions were estimated from the analysis of 5,131 index cases and the development of a new statistical approach. Also, the participation of mosaic neo-mutations in the syndrome was explored from 1,750 index cases from the previous study, with a software developed specifically for detecting poorly represented variants: outLyzer. Finally, the exploration by sequencing of the missing heredity was extended to a panel of 201 genes involved in cancer, from 118 patients selected for the early onset of their disease, a highly suggestive element of a predisposition factor.

The results of this work validated the relevance of the PALB2, RAD51C and RAD51D study for patient management, and also suggested an underestimated involvement of mosaic variants. However, there are still very likely other highly penetrating genetic factors to be discovered, but whose risk modulation is based on an oligogenic model.

Keywords: hereditary predisposition to breast and/or ovarian cancer, Next Generation Sequencing, High Throughput Sequencing, bioinformatics, mosaic, variant-calling

Laboratoire: Laboratoire de Biologie et de Génétique du Cancer, Unité INSERM U1245, Centre François Baclesse, 3 Av. du Général Harris, 14076 Caen