Popular Democratic Republic of Algeria
Ministry of High Education and Scientific Research
Abbes LAGHROUR University- Khenchela
Natural and Life Sciences Faculty
Molecular and Cellular BiologyDepartment

N° de série : .....

MÉMOIRE DE FIN D'ÉTUDESDE MASTER ACADÉMIQUE

Domaine : Sciences de la Nature et de la Vie

Filière :Sciences Biologiques

Spécialité : Génétique

Présenté par :

DJEMEL Raouia KASSEM Rayen

Thème

ANALYSE BIOINFORMATIQUE DES VARIANTS GÉNÉTIQUES ASSOCIÉS AUX MALADIES HÉRÉDITAIRES : ÉTUDE DES MUTATIONS DES GÈNES *BRCA1* ET *BRCA2*

Mémoire soutenu publiquement le : …… juin 2025

Devant le jury composé de :

Dr. HAMADA Youcef	Président	Université Abbes LAGHROUR-Khenchela
Dr. BOUTARFI Zakaria	Examinateur	Université Abbes LAGHROUR-Khenchela
Pr. HAMIDECHI M.Abdelhafid	Encadrant	Université Abbes LAGHROUR-Khenchela

Année Universitaire 2024/2025

Remerciements

Au terme de la rédaction de ce mémoire, nous remercions :
Dieu qui nous a toujours données la force de passer à travers les épreuves et les découragements et qui nous ont aidées à mener à terme cette recherche.
Toute personne ayant participé de près ou de loin à la réalisation de ce travail.
Nous tenons à remercier sincèrement notre directeur de mémoire Monsieur HAMIDECHI Mohamed Abdelhafid, que tous les mots ne suffisent pas à exprimer nos profondes gratitudes pour la confiance que vous nous avez accordée en acceptant de superviser ce travail, s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi pour l'inspiration, la gentillesse, l'aide et le temps qu'il a bien voulu nous consacrer
et sans qui ce mémoire n'aurait jamais vu le jour.
Nos remerciements s'adressent également à notre chef d'option Monsieur
HAMADA Youcef, pour sa générosité, ses efforts à bien nous diriger dans notre formation, pour son aide, son encouragement, ces conseils et la grande patience dont il a su faire preuve malgré ses charges professionnelles, vraiment un chapeau pour vous.
Nous remercions très sincèrement Monsieur BOUTARFI Zakaria pour avoir accepté de faire partie de ce jury. Son regard attentif et ses remarques pertinentes contribueront certainement à enrichir la qualité de ce mémoire.
Nous tenons très chaleureusement à remercier notre honorable Directeur de l'université Abbes LAGHEROUR KHENCHELA et tous les staffs administratifs de notre faculté.
Un grand merci également à tous les enseignants et les enseignantes de UNIVERSITE de
KHENCHELA, pour leur amabilité et leur patience qui ont permis d'affiner notre problématique.
Enfin nous réitérons nos vifs remerciements à tous ceux ou celles qui nous ont apporté tout
Leur soutien de près ou de loin et à nos ami(e)s de la promotion2023/2025 pour leursencouragements ;à toutes ces personnes nous dirons :

Sincèrement MERCI

Dédicace

Avec un énorme plaisir, un cœur ouvert et une immense joie que je dédie ce modeste travail à :

Mes chers parents ;

Mon père «MOHAMMED» L'homme de ma vie et mon exemple éternel, ma mère «ADJROUD DALILA» La lumière de mes jours, la source de mes efforts et la flamme de mon cœur.Aucun dédicace, aucun mot ne pourrait exprimer à leur juste valeur la gratitude et l'amour.que je vous porte. Vous représentez pour moi le symbole de la bonté par excellence, la source de tendresse, le secret de ma force et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi.Votre prière et votre bénédiction m'ont été d'un grand secours pour mener à bien mes études, rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien-être.Chaque ligne de ce mémoire, chaque mot et chaque lettre vous exprime la reconnaissance, le respect, l'estime et le merci d'être mes parents. Que DIEU vous garde pour moi «INCHAE ALLAH».

Je dédie ce travail : Avec tous mes vœux de bonheur, santé et réussite spécialement à ma perle sœur «DJIHANE HANINE» ;

À mes chers frères :«YACINE» et «TAHA» en témoignant de l'attachement, de l'amour et de l'affection que jeporte pour eux ;À tous ceux qui m'aiment et que j'aime ;Je dédie ce travail à ceux qui ont toujours cru en moi et m'ont soutenue corps et âme pour venir au bout de mes ambitions.

DJEMEL Raouia

Dédicace

Celui qui a dit "je suis à elle ", l'a obtenue.
Le voyage n'était pas court et il ne devrait pas l'être, le rêve n'était pas proche et le chemin n'était pas pavé de facilités. Mais je l'ai fait et je l'ai obtenu.

Je dédie ce modeste travail à mes très chers parents qui m'ont guidé durant les moments les plus pénibles de ce long chemin, ma mère « **Houria** » qui a été à mes côtés et m'a soutenu durant toute ma vie, et mon père « **Ali** » qui a sacrifié toute sa vie afin de me voir devenir ce que je suis, merci mes parents.

J'espère avoir répondu aux espoirs que vous avez fondés en moi, je vous rends hommage par ce travail en guise de ma reconnaissance éternelle et de mon infini amour.

Mon cher frère « **Zaki** » et ma sœur « **Amina** » qui n'ont jamais cessé de me conseiller, encourager et soutenir tout au long de mes études. Que Dieu vous protège et vous offre la chance, la réussite et le bonheur qui puissent exister.

Et, finalement, c'est un moment de plaisir de dédier ce projet de fin d'étude à ma famille, mes cousin (es) et mes amis (es) merci pour leur amour et encouragement.

Sans oublier ma chère camarade « SARA » merci pour ton soutien ,Beaucoup de bonheur et de réussite dans ta vie.

Résumé

Ce travail porte sur l'analyse différentielle de l'expression des deux allèles (nommés A et B dans ce mémoire) des gènes *BRCA1* et *BRCA2* à partir de données issues de puces à ADN Affymetrix. L'objectif est d'identifier d'éventuels déséquilibres d'expression allélique pouvant refléter des altérations génomiques, comme des délétions ou des amplifications. Après normalisation des intensités et sélection des sondes spécifiques, un test de Mann-Whitney a permis de comparer les niveaux d'expression entre les deux allèles. Les résultats révèlent, chez plusieurs échantillons, une expression déséquilibrée, suggérant des événements génomiques cliniquement pertinents. Cette approche peut contribuer à la médecine personnalisée en précisant les profils moléculaires des tumeurs.

Mots clés : *Expression allélique – BRCA1/2 – Analyse différentielle – Puces Affymetrix.*

الملخص

، وذلك باستخدام بيانات *BRCA2*و *BRCA1* من الجينين Bو A يتناول هذا البحث تحليلًا تفاضليًا للتعبير الجيني للأليلين يهدف هذا العمل إلى تحديد اختلالات محتملة في التعبير الأليلي قد تعكس .(Affymetrix) مستخرجة من شرائح الحمض النووي تغيّرات جينومية مثل الحذوفات أو التكبيرات. بعد القيام بعملية تطبيع شدة الإشارات واختيار المجسات المناسبة، تم استخدام اختبار لمقارنة مستويات التعبير بين الأليلين. أظهرت النتائج، في عدد من العينات، وجود تعبير غير (Mann-Whitney) مان-ويتني متوازن، مما يشير إلى أحداث جينومية ذات أهمية سريرية. يمكن أن تسهم هذه المقاربة في تطوير الطب الشخصي من خلال توضيح الخصائص الجزيئية للأورام.

الكلمات المفتاحية:.Affymetrix التحليل التفاضلي – شرائح –BRCA1/2 – التعبير الأليلي

Abstract

This thesis focuses on the differential analysis of the expression of alleles A and B of the *BRCA1* and *BRCA2* genes, using data from Affymetrix DNA microarrays. The objective is to identify potential allelic expression imbalances that may reflect genomic alterations, such as deletions or amplifications. After normalization of signal intensities and selection of specific probes, a Mann-Whitney test was used to compare expression levels between the two alleles. The results reveal imbalanced expression in several samples, suggesting clinically relevant genomic events. This approach may contribute to personalized medicine by refining the molecular profiling of tumors.

Keywords: Allelic expression – *BRCA1/2* – Differential analysis – Affymetrix microarrays

INTRODUCTION

Les maladies génétiques représentent un ensemble de pathologies causées par des altérations du matériel génétique. Ces affections, parfois rares mais souvent graves, affectent des millions d'individus dans le monde et englobent un large spectre de troubles allant des anomalies métaboliques aux cancers génétiques. L'avancée des technologies de séquençage à haut débit et l'émergence de la bioinformatique ont permis une meilleure compréhension de l'origine génétique de ces maladies, facilitant l'identification des variants responsables. Comprendre les bases moléculaires des maladies génétiques s'avère donc essentiel pour anticiper leur évolution et améliorer la prise en charge des patients comme cela a été proposé dans la conclusion de Abdelhamid et al. (2021).

L'analyse des variants génétiques est essentielle pour comprendre la susceptibilité aux maladies génétiques, prédire les réponses aux traitements et orienter les approches de médecine personnalisée. Elle permet de détecter des mutations pathogènes ou protectrices influençant directement l'expression ou la fonction des gènes impliqués dans diverses pathologies (Zhou et al., 2021).

Les gènes BRCA1 et BRCA2 jouent un rôle crucial dans la réparation de l'ADN par recombinaison homologue, un mécanisme essentiel au maintien de la stabilité génomique. Leur analyse permet d'identifier les individus à haut risque, facilitant ainsi le dépistage précoce, les mesures préventives et les décisions thérapeutiques ciblées, notamment l'utilisation des inhibiteurs de PARP (PolyADP-RibosePolymerase), une famille d'enzymes impliquées dans la réparation de l'ADN, notamment des cassures simple brin). L'étude approfondie de ces gènes est donc essentielle pour les stratégies de médecine personnalisée (Tung et al., 2022 ; Wendt et al., 2023).

Ces contextes scientifique et médical justifient pleinement notre intérêt à étudier, en profondeur, les mutations génétiques impliquées dans ces pathologies, notamment celles affectant les gènes BRCA1 et BRCA2 :

- Objectif 1 : Identifier les allèles présentant des mutations connues des gènes *BRCA1* et *BRCA2* associées aux maladies génétiques à partir de bases de données publiques telles que dbSNP et *Ensembl*.

- Objectif 2 : Analyser l'effet de quelques mutations à l'aide d'outils bioinformatiques afin d'évaluer leur impact potentiel sur la fonction des protéines BRCA.

1. LES MALADIES D'ORIGINE GÉNÉTIQUE

Les maladies monogéniques sont des maladies causées par la mutation d'un seul gène. Elles suivent généralement un mode de transmission mendélienne (autosomique dominant, autosomique récessif ou lié au chromosome X). Exemples : mucoviscidose, drépanocytose ; alors que les maladies polygéniques résultent de l'interaction de plusieurs gènes associés à des facteurs environnementaux. Ces maladies ont souvent une base génétique complexe, rendant leur étude et leur prédiction plus difficiles. Exemples : Cancer du sein,diabète de type 2, asthme, maladies cardiovasculaires (Moutschen, 2022 ; Yaakoubi, 2024).

La recombinaison homologue, prépondérante durant la phase S, est une voie de réparation très conservée dans le règne vivant qui utilise des séquences du génome semblable à celle affectée, en particulier provenant du chromosome homologue de la paire, en principe identique. Si cette séquence est indisponible, les deux extrémités de la cassure partent à la recherche de régions du génome qui leur ressemblent, afin que des protéines reconstruisent ensuite la partie manquante à partir du chromosome intact et reconnectent les brins (Salaun, 2021).

1.1. Mécanismes moléculaires des maladies génétiques : Les maladies génétiques résultent d'altérations du matériel génétique affectant l'expression ou la fonction normale des gènes. Ces altérations peuvent être des mutations ponctuelles, des délétions, des duplications ou des réarrangements chromosomiques plus complexes (Cooper, 2021). Au niveau moléculaire, une mutation peut entraîner la production d'une protéine défectueuse ou l'absence totale de cette protéine, perturbant ainsi les voies métaboliques, de signalisation cellulaire ou de régulation du cycle cellulaire (Nussbaum et al., 2021).

D'autres mécanismes incluent les mutations affectant les sites de régulation génétique, tels que les promoteurs ou enhancers, modifiant ainsi la transcription des gènes sans altérer directement la séquence codante (MacArthur et al., 2014). Certaines maladies, comme la dystrophie musculaire de Duchenne, sont dues à de grandes délétions entraînant la perte de fonction d'un gène entier (Darras et al., 2022). De plus, des anomalies dans les mécanismes de réparation de l'ADN, comme dans le syndrome de Lynch, favorisent l'accumulation de mutations somatiques conduisant à des maladies génétiques complexes telles que le cancer (Chen et al., 2022).

Avec l'avènement du séquençage haut débit, l'identification précise des variants pathogènes a permis de mieux comprendre les liens entre mutations génétiques et manifestations cliniques. Cette

compréhension ouvre la voie à des approches thérapeutiques innovantes telles que les thérapies géniques ou l'édition génomique par CRISPR-Cas9 (Doudna & Charpentier, 2020).

1.2. Exemples de maladies associées à des variants pathogènes: De nombreuses maladies génétiques sont directement liées à des variants pathogènes affectant des gènes spécifiques. Par exemple, la fibrose kystique est causée principalement par des mutations du gène *CFTR*, la plus fréquente étant la délétion ΔF508. De même, les mutations du gène *BRCA1* ou *BRCA2* augmentent considérablement le risque de développer des cancers du sein et de l'ovaire. Dans les maladies neurologiques, la mutation du gène *HTT* est responsable de la maladie de Huntington, caractérisée par une expansion anormale de triplets CAG. En cardiologie, les mutations dans le gène *MYH7* sont associées à la cardiomyopathie hypertrophique. D'autres maladies comme la drépanocytose, causée par une mutation du gène de la bêta-globine (*HBB*), illustrent l'impact majeur qu'un simple variant peut avoir sur la physiologie humaine. L'identification précise de ces variants pathogènes permet aujourd'hui d'améliorer le diagnostic précoce et de développer des thérapies ciblées (Richards et al., 2022).

Des mutations pathogènes dans l'un ou l'autre de gènes BRCA1/2 entraînent une altération de la réparation de l'ADN ; ce qui augmente considérablement le risque de cancers, notamment du sein et de l'ovaire. Les porteurs d'une mutation *BRCA1* ont un risque de développer un cancer du sein pouvant atteindre 65 % avant l'âge de 70 ans, tandis que pour *BRCA2*, ce risque est estimé à 45 %. Ces mutations peuvent être des délétions, des insertions ou des substitutions ponctuelles, chacune pouvant altérer la fonction de la protéine codée. L'étude bioinformatique de ces mutations est aujourd'hui essentielle pour affiner la prédiction de leur impact clinique (Tung et al., 2022).

2. LES GÈNES BRCA1 ET BRCA2

2.1. Localisation chromosomique et structure des gènes：Les gènes *BRCA1* et *BRCA2*(Figure 1) sont deux acteurs majeurs dans la préservation de la stabilité génétique. *BRCA1* est localisé sur le chromosome 17, précisément en 17q21, tandis que *BRCA2* se trouve sur le chromosome 13, en 13q12.3. *BRCA1* couvre environ 80 kb et contient 24 exons, alors que *BRCA2* est encore plus vaste, s'étendant sur environ 84 kb et composé de 27 exons. Les deux gènes codent pour de grandes protéines nucléaires impliquées dans la réparation des cassures double-brin de l'ADN par recombinaison homologue. La complexité de leur structure exon-intron contribue à la diversité des variants observés, dont certains sont pathogènes et associés à un risque élevé de cancers héréditaires du sein, de l'ovaire, mais aussi d'autres types de tumeurs (Roy et al., 2012).

Les protéines codées par ces deux gènes de prédisposition au cancer du sein, interviennent dans une voie commune de protection du génome. Cependant, ces deux protéines interviennent à des stades différents de la réponse aux dommages de l'ADN et de la réparation de l'ADN. BRCA1 est une protéine qui répond aux dommages du DNAet intervient à la fois dans l'activation des points de contrôle et dans la réparation de l'ADN, tandis que BRCA2 est un médiateur du mécanisme central de la recombinaison homologue. Les liens entre les deux protéines sont mal compris, mais ils doivent exister pour expliquer la forte similitude de la prédisposition au cancer chez l'humain, liée aux mutations de ces gènes.

BRCA1 est une protéine polyvalente qui interagit avec les suppresseurs de tumeurs, les protéines de réparation de l'ADN et les régulateurs du cycle cellulaire par l'intermédiaire de ses divers domaines fonctionnels et joue ainsi divers rôles dans de multiples voies de réparation de l'ADN (notamment la recombinaison homologue, jonction d'extrémité non homologue et réparation du simple brin) et dans la régulation des points de contrôle.

BRCA1 contient un domaine RING amino-terminal doté d'une activité ubiquitine ligase E3 (qui catalyse l'ubiquitylation de la protéine) et un domaine BRCT qui facilite la liaison aux phosphoprotéines. De nombreuses mutations héréditaires du gène BRCA1 associées au cancer ont été découvertes dans les domaines RING et BRCT, indiquant que les deux domaines sont impliqués dans la suppression du cancer du sein et de l'ovaire. L'activité de l'ubiquitine ligase E3 de BRCA1 est renforcée lorsqu'elle est associée au domaine RING de sa protéine partenaire, la protéine du domaine RING 1 associée à BRCA1 (BARD1). L'hétérodimère BRCA1–BARD1 génère des chaînes de polyubiquitine au niveau des liaisons K6 non conventionnelles qui ne semblent pas signaler la dégradation de la protéine, mais pourraient plutôt médier des événements de signalisation en aval par des mécanismes encore mal connus10–13. La fonction suppressive de tumeur de l'activité de l'ubiquitine ligase E3 a été récemment remise en question par l'observation que, dans un modèle murin knock-in exprimant un mutant de BRCA1 déficient en E3-ligase, le développement de tumeurs était inhibé dans la même mesure que lorsque BRCA1 de type sauvage était exprimé.

Contrairement aux activités multifonctionnelles de BRCA1, la fonction principale de BRCA2 se situe dans la recombinaison homolgue. BRCA2 assure le recrutement de la recombinase RAD51 dans les cassures du DNA double brin ; le recrutement de RAD51 est non seulement essentiel à la recombinaison homolgue, mais est également responsable de la fonction antitumorale de ce processus de réparation. BRCA2 contient un domaine de liaison à l'ADN qui se lie à l'ADN simple brin et à l'ADN double brin et huit répétitions BRC qui se lient à RAD51. Le DBD contient cinq composants :

un domaine hélicoïdal α de 190 acides aminés, trois replis de liaison aux oligonucléotides qui sont des modules de liaison à l'ADN, et un domaine tour qui se lie à l'ADN.

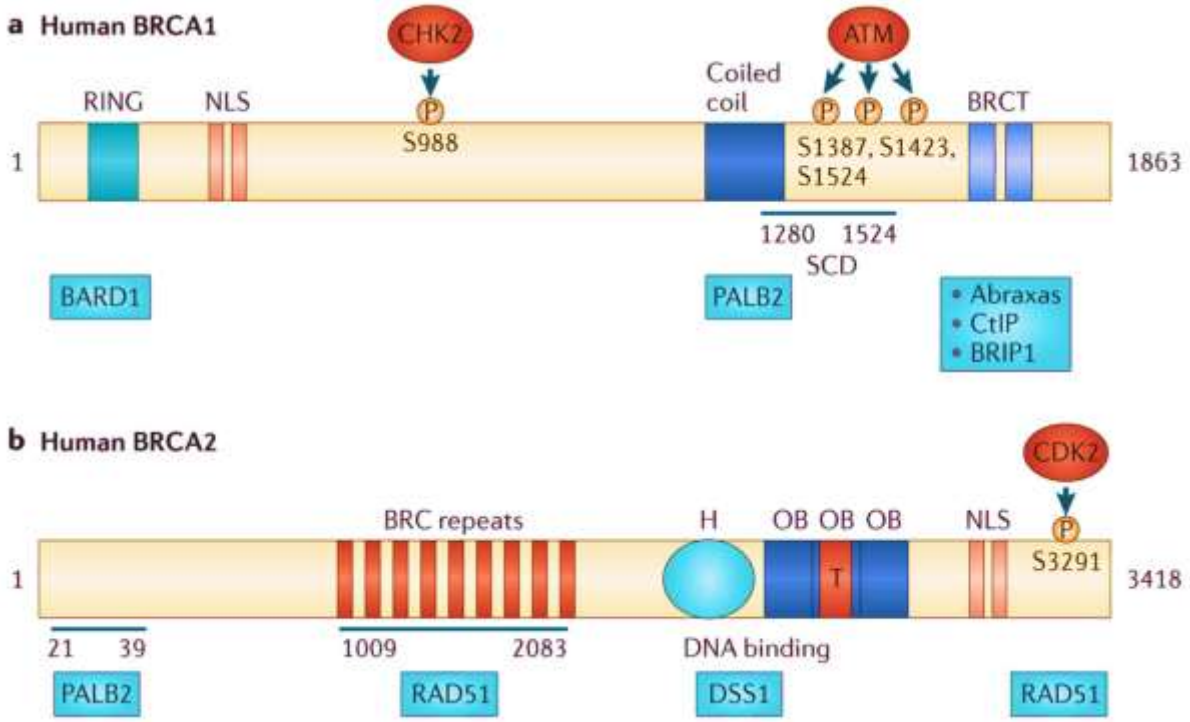

Figure 1 : a) L'extrémité amino-terminale de BRCA1 contient un domaine RING associé à la protéine BARD1 (BRCA1-associated RING domain protein 1) et à une séquence de localisation nucléaire (NLS). L'extrémité C-terminale de BRCA1 contient : un domaine spiralé qui s'associe au partenaire et localisateur de BRCA2 (PALB2) ; un domaine de cluster SQ/TQ (SCD) qui contient environ dix sites potentiels de phosphorylation et s'étend sur les résidus d'acides aminés 1280 à 1524 ; et un domaine BRCT qui lie l'abraxas phosphorylé par ATM, la protéine interagissant avec CtBP (CtIP) et l'hélicase C-terminale 1 interagissant avec BRCA1 (BRIP1). b) : L'extrémité N-terminale de BRCA2 se lie à PALB2 au niveau des acides aminés 21 à 39. BRCA2 contient huit répétitions BRC entre les acides aminés 1009 et 2083 qui se lient à RAD51. Le domaine de liaison à l'ADN de BRCA2 contient un domaine hélicoïdal (H), trois replis de liaison aux oligonucléotides (OB) et un domaine tour (T), qui pourraient faciliter la liaison de BRCA2 à l'ADN simple brin et double brin. L'extrémité C-terminale de BRCA2 contient un site de phosphorylation sur la séquence NLS et une kinase dépendante des cyclines (CDK) en S3291 qui se lie également à RAD51. La région centrale de BRCA1 contient un site de phosphorylation CHK2 sur S988 (Roy et al., 2012).

2.2. Types de mutations fréquentes : Les mutations les plus fréquentes observées dans ces gènes sont des mutations ponctuelles, des insertions ou délétions de quelques nucléotides, qui provoquent souvent un décalage du cadre de lecture (frameshift). Ce décalage conduit généralement à l'apparition d'un codon stop prématuré, produisant une protéine tronquée et non fonctionnelle. Les mutations non-sens (création directe d'un codon stop) sont également fréquentes et responsables d'une perte totale d'activité de la protéine. En parallèle, des réarrangements génomiques de grande ampleur affectant plusieurs exons (grandes délétions ou duplications) sont surtout rapportés pour BRCA1, et dans une moindre mesure pour BRCA2. Ce type d'altération est particulièrement délétère car il peut

abolir l'expression du gène ou perturber gravement la structure protéique. Des analyses génétiques indiquent que la majorité des mutations pathogènes de BRCA1 et BRCA2 détectées dans les cohortes de patients atteints de cancers héréditaires appartiennent aux catégories frameshift et non-sens (Boussios et al., 2022). L'identification précise de ces mutations est aujourd'hui au cœur du diagnostic génétique et des stratégies de prévention ciblée.

3. APPROCHES BIOINFORMATIQUES DANS L'ANALYSE DES VARIANTS

3.1. Bases de données : Les bases de données bioinformatiques jouent un rôle essentiel dans l'identification, l'annotation et l'interprétation des variants génétiques. dbSNP (Database of Single NucleotidePolymorphisms) est une ressource majeure développée par le NCBI, regroupant des millions de variants nucléotidiques simples (SNPs), ainsi que d'autres types de variations génétiques, tels que les insertions/délétions (indels). Elle fournit des informations sur la fréquence des allèles, leur position chromosomique, ainsi que leur éventuelle implication dans certaines pathologies. ClinVar, également hébergée par le NCBI, est une base de données complémentaire qui associe des variants génétiques à des phénotypes cliniques. Elle intègre des annotations fournies par des laboratoires de diagnostic, des chercheurs et d'autres institutions, et attribue un niveau de confiance à l'interprétation clinique des variants (bénin, probablement pathogène, incertain, etc.). Enfin, Ensembl constitue une plateforme intégrée qui offre des annotations génomiques riches, en reliant les variants aux gènes, transcrits, effets prédits (via VEP – Variant EffectPredictor), et en fournissant des données comparatives entre espèces. Sa version de 2024 permet l'annotation des variants dans les régions non codantes, telles que les ARN longs non codants (IncRNA) et les régions UTR, facilitant ainsi la priorisation des variants dans des contextes cliniques complexes. L'exploitation conjointe de ces bases permet une meilleure compréhension du potentiel fonctionnel et clinique des variants identifiés, en s'appuyant sur des données standardisées, validées et continuellement mises à jour (Phan, 2025).

3.1.1. Cas de la dbSNP *:* Parmi les bases de données bioinformatiques utilisées pour l'analyse des variants génétiques, dbSNP constitue une ressource incontournable (Figure 2). Développée par le NCBI, cette base regroupe un vaste ensemble de polymorphismes nucléotidiques simples (SNPs) ainsi que d'autres types de variations, tels que les insertions et délétions courtes. La dernière version de dbSNP, publiée en mars 2025, contient plus de 1,17 milliard de variants référencés (rsID) et intègre des données issues de grands projets tels que 1000 Genomes, TOPMed, gnomAD et ALFA. Cette mise à jour renforce l'utilité de dbSNP comme ressource centrale pour l'analyse des fréquences alléliques et la cartographie des variants humains.

dbSNP est utilisée comme source principale pour l'identification et la collecte des variants présents dans les régions génomiques d'intérêt, en lien avec la pathologie étudiée. Chaque variant est associé à un identifiant unique (rsID), à sa localisation précise sur le génome de référence (GRCh38), à la nature de la mutation (ex. : transition ou transversion), ainsi qu'à des données sur la fréquence allélique dans différentes populations. Cette base fournit également des informations sur la validation expérimentale des variants, leur conservation évolutive et leurs éventuelles implications fonctionnelles. Grâce à son accès facile via des interfaces web ou par téléchargement de jeux de données complets, dbSNPpermet une première étape de criblage et de sélection des variants candidats, avant leur annotation fonctionnelle plus approfondie à l'aide d'autres outils comme Ensembl ou ClinVar. Son utilisation systématique garantit une standardisation des analyses et une reproductibilité des résultats dans les études de génétique humaine (Phan et al., 2024)..

Par ailleurs, dbSNP fournit des indicateurs de qualité comme le statut de validation (par séquençage Sanger, soumission multiple, etc.) et des liens fonctionnels vers d'autres bases (ClinVar, Gene, OMIM). L'interrogation de la base est réalisée via le portail NCBI Variation Viewer ainsi que par extraction directe des fichiers VCF disponibles en téléchargement. Cette étape a permis de générer un catalogue préliminaire de variants bruts, servant de base à l'annotation fonctionnelle ultérieure avec des outils tels que Ensembl VEP et à la priorisation des variants potentiellement pathogènes.

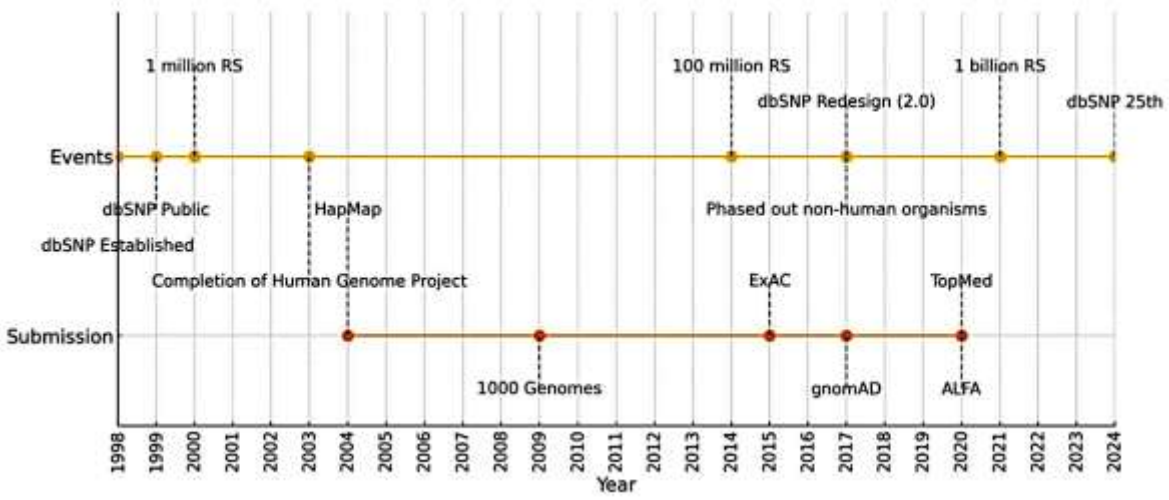

Figure 2 : Chronologie des étapes clés de la dbSNP et des soumissions génomiques. Cette chronologie classe les événements clés de l'histoire du dbSNP et des soumissions génomiques associées. Notamment, l'achèvement du Projet Génome Humain en 2003 a entraîné une augmentation significative des soumissions de dbSNP. Les « événements » incluent la création dedbSNP en 1998, l'atteinte du million de rs en 2000, la refonte de 2017 et le 25ème anniversaire en 2024. Les « soumissions » présentent des contributions majeures telles que HapMap (2004), 1000 Genomes (2009), ExAC (2015), GnomAD (2017) et TopMed (2020) (13-17). La figure illustre l'ordre chronologique et l'importance de ces étapes (Phan et al., 2024).

3.2. Outils de prédiction des effets des mutations *:* Ils jouent un rôle crucial dans l'interprétation des variants génétiques, notamment pour évaluer leur potentiel pathogène. Parmi ces outils, CADD (Combined Annotation DependentDepletion) et PolyPhen-2 (PolymorphismPhenotyping v2) sont largement utilisés.

CADD[1] est un cadre d'annotation intégratif qui combine diverses informations pour évaluer la délétion potentielle des variants génétiques. La version la plus récente, CADD v1.7, intègre des modèles de langage protéique, des réseaux de neurones convolutifs pour les régions régulatrices, ainsi que d'autres scores au niveau des nucléotides pour améliorer les prédictions à l'échelle du génome. Cette version a démontré une amélioration significative des performances par rapport à la version précédente, notamment en augmentant l'AUROC de 0,981 à 0,982 pour les variants cliniquement annotés dans ClinVar (Schubach, 2024).

PolyPhen-2 est un outil qui prédit l'impact des mutations non synonymes sur la structure et la fonction des protéines. Il utilise des informations sur la séquence, la structure et l'évolution pour estimer la probabilité qu'une mutation soit délétère. Des études récentes ont montré que PolyPhen-2,

[1]https://cadd.gs.washington.edu/home?utm_source=chatgpt.com

bien qu'efficace, peut être amélioré par des approches d'apprentissage automatique plus avancées. Par exemple, l'intégration de modèles de langage protéique spécifiques aux maladies a montré une amélioration de plus de 5 % de l'AUC dans la prédiction des variants pathogènes (Zhan et Zhang, 2023).

3.3. Notion de variant bénins, pathogènes, ou incertains : Dans le cadre de l'interprétation bioinformatiquedes données génomiques, les variants identifiés par séquençage sont classés en différentes catégories selon leur potentiel impact clinique : bénin (benign), probablement bénin (likelybenign), de signification incertaine (variant of uncertainsignificance, VUS), probablement pathogène (likelypathogenic) et pathogène (pathogenic). Cette classification repose sur les recommandations établies par l'American College of MedicalGenetics and Genomics (ACMG) et l'Association for MolecularPathology (AMP) (Richards et al., 2015).

Les variants bénins n'ont pas d'impact significatif sur la fonction des gènes et sont fréquents dans la population générale. À l'inverse, les variants pathogènes sont associés à une perturbation fonctionnelle démontrée d'un gène et à une corrélation claire avec un phénotype pathologique. Les VUS, quant à eux, représentent une catégorie intermédiaire pour laquelle les données disponibles sont insuffisantes ou contradictoires, rendant impossible une interprétation fiable à des fins cliniques. Cette incertitude peut résulter d'un manque d'informations sur la fréquence allélique dans les bases de données populationnelles (comme gnomAD), d'une absence de données fonctionnelles, ou de résultats discordants dans les prédictions bioinformatiques (ex. : SIFT, PolyPhen-2, CADD). Plusieurs études récentes ont souligné l'importance de la reclassification dynamique des VUS au fil du temps, grâce à l'accumulation de nouvelles données cliniques, fonctionnelles ou populationnelles (Walsh et al., 2024 ; Pleasant et al., 2025). Ainsi, la catégorisation des variants est un processus évolutif qui nécessite une intégration rigoureuse de multiples lignes de preuves pour garantir une interprétation fiable dans le contexte de la médecine génomique.

4. LES PRINCIPAUX TYPES DE PUCES ADN

L'étude du profil d'expression génique des cellules et des tissus est devenue un outil majeur de découverte en médecine. Les expériences sur puces à ADN permettent de décrire les variations d'expression à l'échelle du génome, tant en santé qu'en maladie. De plus, une étude impartiale et systématique du profil d'expression génique devrait permettre d'établir une nouvelle taxonomie des syndromes obstétricaux et gynécologiques. Ainsi, une nouvelle ère s'ouvre où les processus et troubles de la reproduction pourraient être caractérisés à l'aide d'outils moléculaires et d'empreintes génétiques.

La conception, l'analyse et l'interprétation des expériences sur puces à ADN requièrent des connaissances spécialisées qui ne font pas partie du cursus standard de notre discipline. Les auteursont décrit les types d'études réalisables avec des expériences sur puces à ADN (comparaison de classes, prédiction de classes, découverte de classes) et ont abordé les questions clés relatives à la conception expérimentale, au prétraitement des données et aux méthodes de sélection des gènes. Les types courants de représentation des données sont illustrés. Les pièges potentiels dans l'interprétation des expériences de microarray, ainsi que les atouts et les limites de cette technologie, sont mis en évidence (Tung, 2006).

Les puces à ADN peuvent mesurer simultanément le niveau d'expression de milliers de gènes dans un échantillon d'ARNm particulier. Un tel profilage d'expression à haut débit peut être utilisé pour comparer le niveau de transcription des gènes dans des conditions cliniques afin de : 1) identifier des biomarqueurs diagnostiques ou pronostiques ; 2) classer les maladies (par exemple, les tumeurs avec un pronostic différent qui sont indiscernables par examen microscopique) ; 3) surveiller la réponse au traitement ; et 4) comprendre les mécanismes impliqués dans la genèse des processus pathologiques. Pour ces raisons, les puces à ADN sont considérées comme des outils importants pour la découverte en médecine clinique (Schena, 2000 ; Zhang, 2005).

Une puce à ADN peut être considérée comme une analyse de Southern ou de Northern blot parallèle à grande échelle (au lieu d'un gel, les sondes sont fixées sur une surface inerte, qui deviendra la puce à ADN). L'ARNm est extrait des tissus ou des cellules, rétrotranscrit et marqué avec un colorant (généralement fluorescent), puis hybridé sur la puce, comme illustré à la figure 3. L'hybridation et les lavages sont réalisés dans des conditions de stringence élevée afin de minimiser le risque d'hybridation croisée entre gènes similaires. L'étape suivante consiste à générer une image par imagerie par fluorescence induite par laser. Le principe de la quantification des niveaux

d'expression est que la quantité de fluorescence mesurée à chaque emplacement spécifique de la séquence est directement proportionnelle à la quantité d'ARNm de séquence complémentaire présente dans l'échantillon analysé. Ces expériences ne fournissent pas de données sur le niveau d'expression absolu d'un gène particulier (concentrations réelles d'ARNm), mais sont utiles pour comparer le niveau d'expression entre différentes conditions (figure 3) et gènes sain ou malade (Knudsen, 2004).

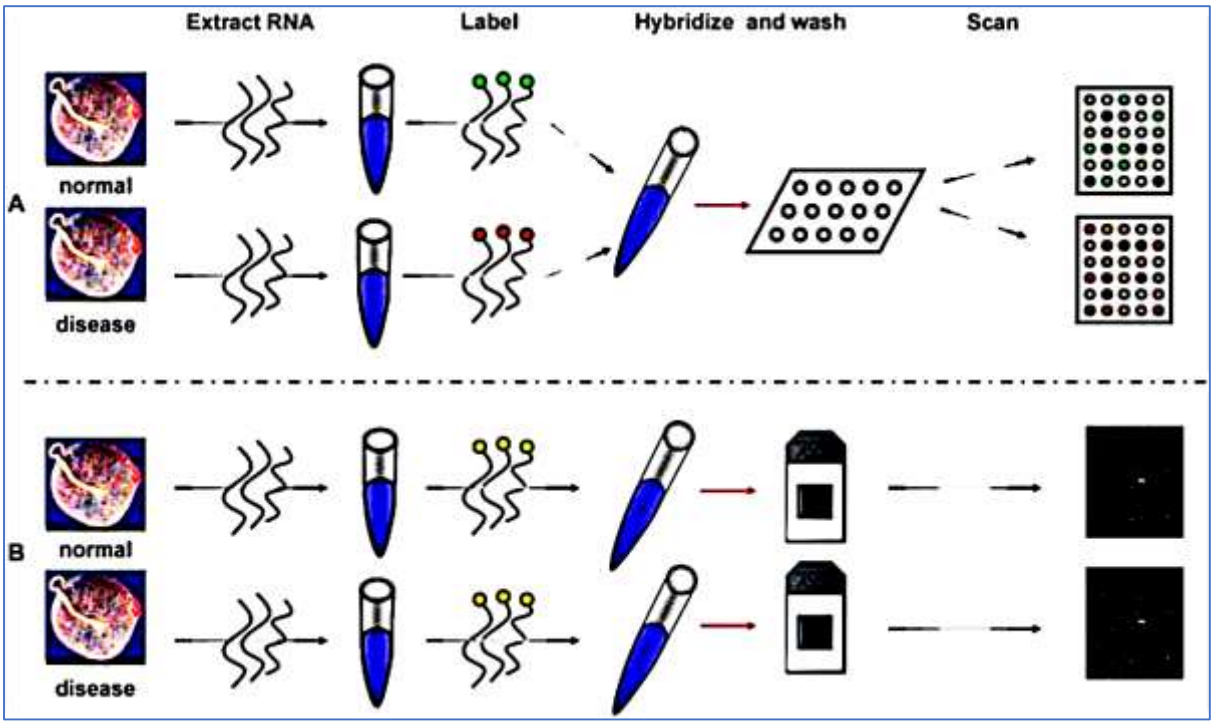

Figure 3 : Représentation schématique des étapes de la réalisation de microarrays. A (le panneau supérieur) illustre la technologie à deux canaux, tandis que B (le panneau inférieur) illustre la technologie à canal unique. Dans le panneau A, les ARNm normaux et pathologiques sont marqués avec deux colorants différents, mélangés puis hybridés sur le même microarray. Après lavage, le microarray est balayé à deux longueurs d'onde différentes pour produire deux images : une pour l'échantillon normal et une pour celui malade. Dans le panneau B (monocanal), chaque échantillon est marqué avec le même colorant fluorescent, mais hybridé indépendamment sur des microarrays différents (Tarca et al., 2006).

4.1. Les types de microarrays *:* les micropuces à ADN (ou puces à ADN) peuvent être classées selon trois critères :

i) La longueur des sondes : en fonction de leur longueur, les puces à ADN peuvent être classées en "puces à ADNc", qui utilisent de longues sondes de plusieurs centaines ou milliers de paires de bases, et en "puces à oligonucléotides", qui utilisent des sondes courtes (généralement de 50 pb ou moins).

ii) Les méthodes de fabrication comprennent : le « dépôt ouspotting» de séquences précédemment synthétisées et la « synthèse *in situ* ». Généralement, les puces à ADNc sont fabriquées par dépôt, tandis que les puces à oligonucléotides sont fabriquées par des technologies *in situ*. Ces technologies comprennent : la « photolithographie » (par exemple, Affymetrix, Santa Clara, Californie), l'« impression par jet d'encre » (par exemple, Agilent, Palo Alto, Californie) et la « synthèse électrochimique » (par exemple, Combimatrix, Mukilteo, Washington).

iii) *Les puces monocanal* analysent un seul échantillon à la fois, tandis que les « puces multicanaux » peuvent analyser deux échantillons ou plus simultanément. La puce AffymetrixGeneChip est un exemple de puce oligonucléotidique monocanal.

En pratique, le terme « sonde » désigne la séquence nucléotidique fixée à la surface du microarray. Dans les expériences sur microarray, le terme « cible » désigne ce qui est hybridé aux sondes.

4.2. Les types d'études conduites avec une puce à ADN *:* il existe trois principaux types d'applications des puces à ADN en médecine (Khatri, 2005 ; Tarca et al., 2006):

1- La première consiste à identifier les différences d'expression entre des groupes prédéfinis d'échantillons. Il s'agit d'une expérience de « comparaison de classes », par exemple, l'identification de gènes différentiellement exprimés dans les placentas de femmes enceintes normales et de femmes atteintes de prééclampsie.

2- Une deuxième application, la « prédiction de classe », consiste à identifier l'appartenance à une classe d'un échantillon en fonction de son profil d'expression génétique (Intelligence Artificielle). Par exemple, il serait possible de prédire si une patiente présente (ou développera) une prééclampsie en fonction de son profil d'expression sanguine. Cela nécessite la construction d'un classificateur (un modèle mathématique) capable d'analyser le profil d'expression génétique d'un échantillon et de prédire son appartenance à une classe. Ce classificateur est construit à partir d'un ensemble représentatif d'échantillons dont l'appartenance à une classe est connue (par exemple, des femmes ayant une grossesse normale et celles développant ultérieurement une prééclampsie). Ce classificateur sera ensuite utilisé pour évaluer la probabilité de développer une prééclampsie chez les patientes non incluses dans sa construction.

3-Le troisième type d'application consiste à analyser un ensemble donné de profils d'expression génétique afin de découvrir des sous-groupes partageant des caractéristiques communes. Cette

application est appelée « découverte de classes ». Par exemple, les profils d'expression d'un grand nombre de femmes atteintes de prééclampsie seront mesurés afin d'identifier des sous-groupes de patientes présentant un profil d'expression génétique similaire. Cet effort vise à générer une taxonomie moléculaire de la maladie. Autrement dit, combien de types moléculaires de prééclampsie (sous-groupes) existent dans un échantillon de femmes atteintes de la maladie ?

4.3. Le pré-traitement des données : Une fois les puces à ADN hybridées, les images obtenues servent à générer un ensemble de données. Cet ensemble de données doit être prétraité avant l'analyse et l'interprétation des résultats.

Le prétraitement est une étape qui extrait ou améliore les caractéristiques significatives des données et les prépare à l'application des méthodes d'analyse. Un exemple typique de prétraitement consiste à calculer le logarithme des valeurs d'intensité brutes. La « normalisation » est un type particulier de prétraitement effectué pour tenir compte des différences systématiques entre les ensembles de données. Un exemple de normalisation consiste à modifier les valeurs d'intensité brutes afin de compenser les différences d'efficacité des colorants lors d'expériences sur puces à ADN à deux canaux utilisant Cy3 (vert) et Cy5 (rouge).

La correction de bruit de fond (Backgroud) est conçue pour ajuster l'hybridation non spécifique, c'est-à-dire l'hybridation de transcrits d'échantillons (cibles) dont les séquences ne correspondent pas parfaitement à celles des sondes sur la puce. Sur les puces à points, l'hybridation non spécifique incluse dans les valeurs d'intensité brutes peut être estimée à partir du niveau de fluorescence à proximité immédiate de la sonde (Yang, 2002).

PARTIE 2 : MATÉRIEL ET MÉTHODES

1. SOURCE DES DONNÉES ET TRAITEMENT

Dans le cadre de cette étude, nous avons exploité des données d'expression génique issues de puces à ADN de type AffymetrixCytoScan HD, ciblant spécifiquement les gènes *BRCA1* et *BRCA2*.

Les types de fichiers liés aux données d'expression génique Affymetrix sont généralement classés en deux catégories principales : les fichiers bruts et les fichiers normalisés. Les fichiers bruts contiennent les données directes provenant des scanners de puces à ADN, tandis que les fichiers normalisés sont le résultat du traitement des données brutes pour éliminer les biais et les erreurs systématiques (bruits de fond). Les étapes simplifiées du pipeline d'obtention des valeurs d'expression génique sont :

. CEL files (bruts) : fichier binaire ou texte

↓

Contrôle qualité

↓

Correction du bruit de fond

↓

Normalisation intra- et inter-puces

↓

Résumé des probes

↓

Matrice d'intensités normalisées

Un fichier de type '.CEL' présente le contenu simplifié suivant :

```
[CEL]
Version=3
Cols=2560
Rows=2560
TotalProbes=6553600
DatHeader= AFFYMETRIX_CYTOSCAN_HD
Algorithm=Percentile

[INTENSITY]
Index              X    Y    Intensity    StdDev    Pixels
1007-s-at.         12   52   9500          150       16
1053_at.           18   71   8200          110       16
117_at             22   98   700            98       16
121_at              8    6   21000         125       16
```

Chaque ligne correspond à une sonde sur la puce (coordonnées X, Y), avec l'intensité mesurée(figure 4). La colonne 'Pixels' indique le nombre de pixels utilisés pour calculer l'intensité moyenne de fluorescence d'une sonde (ou 'probe') sur 16 pixels (souvent disposés en carré 4x4) sur la puce à cet emplacement. Plus il y a de pixels, plus la mesure d'intensité est robuste (moins de bruit).

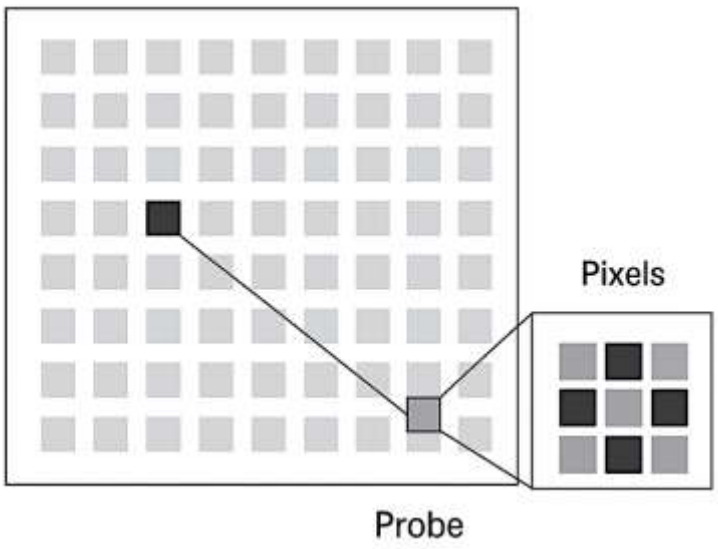

Figure 4 : Structure d'un spot de sonde (probe) sur une puce Affymetrix : organisation des pixelx de détection.

Ainsi, comme le montre la figure 4, chaque sonde correspond à une courte séquence d'ADN fixée à la surface de la puce et est constituée de plusieurs pixels (souvent 16) qui capturent le signal de fluorescence émis par les molécules hybridées. L'intensité de la sonde est obtenue en agrégeant les signaux de ses pixels. Chaque sonde cible une région spécifique du génome (gène, SNP, ou segment CNV:CopyNumber Variation) et plusieurs sondes peuvent couvrir un même gène pour améliorer la fiabilité de la détection.

Les fichiers '.CSV' sont ensuite générés par l'utilisateur bioinformaticien via un logiciel tel que 'Expression Consol' ou 'Transcriptome Analysis Console'. Le contenu final, après normalisation du fichier '.CEL' donne :

ProbSet_ID,Gene_Symbol,Chromosome,Normalized_Intensity, Log_2_Expression
1007_s_at GAPDH, 12, 30200, 14.55
1053_at, ACTB, 7, 18850, 13.2
117_at, BRCA1, 17, 9100, 12.2
121_at, BRCA2, 13, 21500, 13.72

Nos données d'expression géniquesont issues à partir des plate-formes'CancerGenome Atlas' (TCGA) et de'International Cancer Genome Consortium (ICGC)/Pan-Cancer Analysis of WholeGenomes' (PCAWG) car ces données ont été validées par l'étudede Hoadley (2018). La source de traitement de nos données est principalement la base de données génomique EMBL (EuropeanMolecularBiologyLaboratory)[2].

Ces données, au format CSV, ont, ensuite, été transforméesau format Excel pour traitements et analyses sur les logiciels tels que GraphPadPrism v.10.0 ou SPPS v.23. Elles concernent deux types de tumeurs : les cancers épithéliaux et les cancers stromaux.

Le tableau suivant (Tableau 1)regroupeles données liées à l'expression des deux gènes : *BRCA1*et BRCA2, avec leurs deux allèles notés A et B dans cette étude.

[2] ebi.ac.uk/

Tableau 1 : Échantillon des données de l'expression génique des deux allèles (A et B) de BARC1 et BRCA2 (Exemple pour N= 10 individus).

Individu	*BRCA1*_A	*BRCA1*_B	BRCA2_A	BRCA2_B
1	2,616	2,197	2,488	2,632
2	2,932	2,444	3,528	4,631
3	2,285	2,031	2,277	2,296
4	2,503	2,529	2,086	2,646
5	2,244	2,708	2,367	2,455
6	1,935	2,333	2,465	2,171
7	2,231	2,817	2,591	3,329
8	2,209	2,367	2,570	2,671
9	1,916	2,459	3,253	5,295
10	2,386	2,123	2,619	2,494

Les traitements statistiques ont été effectués par le logiciel GraphPadPrism v 10.0 et MS-Excel alors que les annotations bioinformatiquesont été réalisées sur les plateformes. Le test de Mann-Whitney (test U) a été utilisé dans la comparaison des moyennes d'expression génique (au lieu du test de Student) car nos données d'expression génique, pour les quatre allèles ne suivent pas une loi Normale selon le test effectué de Kolmogorow-Smirnov.

2. ANALYSE DES DONNÉES D'EXPRESSION DES GÈNES *BRCA1* ET *BRCA2*

2.1. Description statistique des données d'expression géniques : Nous avons retenu, pour la partie descriptive de nos données, les paramètres suivants : Taille de l'échantillon, minimum, 25e percentile, médiane, 75e percentile, maximum, moyenne et l'étendue. À partir de ces paramètres descriptifs, les seuils retenus dans cette analyse sont déterminés empiriquement. Ainsi, le 90èmepercentile a été retenu pour l'analyse de la différence d'expression entre deux allèles du même gène. Ce seuil est maintenu pour les cas suivants :

Allèle BRCA1_A : 1,571 ; Allèle BRCA1_B :2,027 ; Allèle BRCA2_A : 1,852 ; Allèle BRCA2_B : 1,648.

Notons que pour les cas de BRCA1, la valeur minimale de 1,571 a été retenue au lieu de celle de l'allèle B de 2,027 afin de limiter les faux positifs et de ne pas surestimer les pertes, car si nous choisissons la valeur la plus élevée (exemple 2,07 pour BRCA1), nous risquerons de conclure à tort que BRCA1_A (1,571) est sous-exprimé. Ainsi, nous considérons une perte allélique que si l'expression génique descend en dessous de ce niveau faible. La même approche est maintenue pour BRCA2, où la valeur de 1,648 a été maintenue.

2.2. Analyse du déséquilibre d'expression allélique : Afin de caractériser le statut des allèles des deux gènes *BRCA1* et BRCA2, nous avons interprété les valeurs d'intensité du signal obtenues à partir des données d'expression génique issues des puces AffymetrixCytoScan HD. Ces intensités d'expression génique ont été exploitées pour proposer un statut qualitatif (figure 5)basé sur des seuils empiriques, définis comme suit :

- Délétion : intensité < 1,7
- Normal : intensité ≥ 1,7 et <2,3
- Gain : intensité ≥ 2,3 et <4
- Amplification : intensité ≥ 4

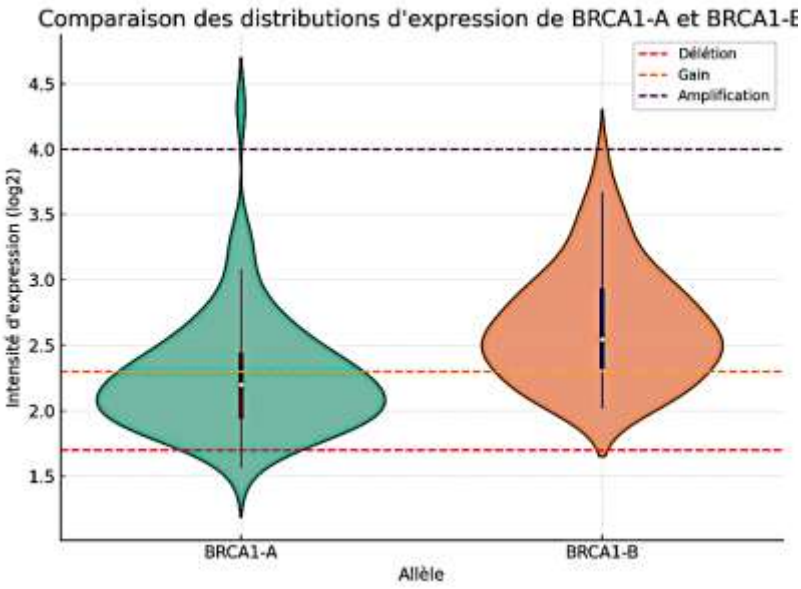

Figure 5 : Diagramme Violin Plot des deux allèles BRCA1_A et BRCA1_B.

Ces seuils d'expression génique (délétion, normal, gain, amplification) ont été définis de manière empirique, en s'appuyant sur la distribution des intensités $\log_2$ observées dans l'ensemble des données. Ils permettent une classification qualitative utile pour la détection d'anomalies d'expression. Ces bornes sont cohérentes avec celles rencontrées dans la littérature sur les profils transcriptomiques tumoraux.

Par exemple, une valeur d'intensité de 2,201 est interprétée comme un gain, celle de 2,685 correspond à un profil normal, tandis que 4,316 indique une amplification. Ce classement permet de détecter les altérations potentielles du nombre de copies géniques, souvent associées à des processus tumoraux.

Les statuts délétion, gain, normal et amplification, obtenus à partir de nos données, ont une signification biologique et pathologique importante, surtout dans le contexte du cancer(Tableau 2):

1. Délétion (<1,7) : Une diminution significative du signal d'expression ou de la copie du gène.

- Biologiquement : perte d'un allèle (hémizygotie) ou des deux (homozygotie).
- Pathologiquement :
    - Peut entraîner une perte de fonction de gènes suppresseurs de tumeurs (ex. : BRCA1, TP53).
    - Favorise la mutagenèse ou l'instabilité génomique. Ainsi, une délétion de BRCA1 peut conduire à une incapacité à réparer l'ADN, favorisant le développement de cancers du sein.

2. Normal (1,7 – 2,3) : Nombre de copies conforme à la normale (diploïdie attendue).

- Biologiquement : le gène est présent en deux copies (1 par chromosome homologue).
- Pathologiquement :
    - Aucune implication directe.
    - Sert de témoin pour détecter les déséquilibres ailleurs.

3. Gain (2,3 - 4) : une légère augmentation du nombre de copies du gène.

- Biologiquement : Duplication partielle ou légère amplification.
- Pathologiquement :
    - Peut entraîner une surexpression modérée du gène.
    - Souvent précoce ou bénigne, mais peut préparer le terrain à une amplification ultérieure.

4. Amplification (>4) : Multiplication de segments d'ADN (ex. : amplicons).

- Biologiquement : le gène est présent en deux copies (1 par chromosome homologue).
- Pathologiquement :
    - Très souvent oncogénique.
    - Gènes amplifiés deviennent hyperactifs. Par exemple, une amplification de BRCA2 peut parfois refléter une compensation ou une dérégulation dans certaines tumeurs.

Seuils sont indicatifs variant légèrement selon les plateformes, la normalisation et la distribution des données.

Tableau 2 : Seuils d'expression allélique de classification des différents profils

Statut	Log_2 ratio	Copie estimée	Implication possible
Délétion	<-0,3	<1,7	Perte partielle ou totale de fonction, gène suppresseur
Normal	-0,3 - +0,3	~1,7 – 2,3	Aucune altération apparente. Pas de variation significative.
Gain	+0,3 - +0,7	2,3 - 4	Surexpression légère, état instable.
Amplification	> +0,7	>4	Activation oncogène, mauvais pronostic.

Pour justifier les seuils empiriques que nous avons utilisés pour interpréter les intensités d'expression des gènes *BRCA1* et *BRCA2* à partir des données issues des puces AffymetrixCytoScan HD, il est pertinent de nous référer à des études récentes qui ont employé des approches similaires.

Une étude notable est celle de Zhu et al. (2024), publiée dans *GenomeMedicine*, qui a analysé les altérations du nombre de copies (CNAs) dans des tumeurs associées à des mutations bi-alléliques de *BRCA1* et *BRCA2*. Les auteurs ont utilisé des données de puces à ADN haute densité pour identifier des régions génomiques présentant des délétions ou des amplifications récurrentes, en se basant sur des variations significatives des intensités de signal. Bien que l'étude ne spécifie pas de seuils

numériques précis, elle illustre l'utilisation de variations d'intensité pour détecter des altérations du nombre de copies dans ces gènes.

De plus, l'outil Rawcopy, un package R pour l'analyse des données issues des puces Affymetrix, propose une approche pour estimer les variations du nombre de copies en calculant des ratios $\log_2$ des intensités de signal. Bien que Rawcopy ne fournisse pas de seuils universels, il permet aux utilisateurs de définir des seuils adaptés à leurs données spécifiques pour catégoriser les pertes, gains ou amplifications.

En l'absence de seuils standardisés dans la littérature, il est courant que les chercheurs définissent des seuils empiriques basés sur la distribution des intensités dans leurs propres ensembles de données. Ces seuils peuvent être ajustés en fonction des caractéristiques spécifiques des échantillons analysés et des objectifs de l'étude.

De plus, il est à noter que les seuils que ses seuils proposés dans notre étude sont des valeurs empiriques approximatives couramment utilisées dans certaines analyses de données de nombre de copies (CNVs), issues de puces de type Affymetrix. Toutefois, ces seuils ne sont pas universels ni tirés d'une publication unique, mais plutôt basés sur des pratiques générales dans le domaine de la cytogénétique moléculaire et l'analyse de données d'expression/copie avec des outils comme GISTIC (Genomic Identification of SignificantTargets In Cancer), Rawcopy (analyse de variations du nombre de copies : CNVs), ou CNVkit (Détection et visualisation des variations du nombre de copies : CNVs à partir de données NGS.

2.3. Analyse comparative des distributions d'expression génique de BRCA1 et BRCA2 : Afin de comparer les niveaux d'expression des deux gènes BRCA1 et BRCA2, nous avons appliqué le test non paramétrique de Mann-Whitney (également appelé Wilcoxon rank-sum test). Ce test a été choisi en raison de la nature non-gaussienne des données d'intensité issues des fichiers .CEL (échelle $\log_2$), et de la présence de valeurs extrêmes. Le test évalue la différence de distribution entre les deux groupes d'allèles pour chaque gène indépendamment.

Pour chaque gène, les intensités d'expression de l'allèle A ont été comparées à celles de l'allèle B sur l'ensemble des échantillons. L'hypothèse nulle stipule qu'il n'existe pas de différence significative entre les distributions des deux allèles. Une valeur de $p<0,05$ a été considérée comme statistiquement significative. L'analyse a été réalisée à l'aide du logiciel GraphPadPrism 10.0.

2.4. Détection de la LOH par expression allélique : Afin d'affiner l'interprétation des variations du nombre de copies, synonyme d'une perte d'hétérozygotie (LOH : Loss Of Heterozogosity), au niveau des deux gènes*BRCA1/2*, nous avons calculé la différence absolue entre les intensités mesurées des deux allèles pour chaque échantillon :

LOH = (Intensité_Expression_Allèle_A) – (Intensité_Expression_Allèle_B).

Cette démarche permet de détecter un déséquilibre d'expression entre les deux copies du gène, pouvant révéler une perte mono-allélique (perte d'hétérozygotie), même lorsque la valeur moyenne globale reste dans la plage de normalité. En effet, dans certains cas, une intensité globale "normale" peut masquer la perte d'un allèle compensée par une surexpression ou une duplication de l'autre. Le calcul de cette différence entre allèles permet donc une meilleure détection des altérations asymétriques, particulièrement utiles dans les contextes tumoraux où les profils d'expression sont souvent hétérogènes. Cette approche complémentaire améliore ainsi la sensibilité de détection des altérations structurelles impliquant *BRCA1/2*. Les seuils retenus dans cette analyse sont :

- Si la différence absolue entre les deux intensités dépasse un certain seuil empirique (>1,8, selon le niveau de bruit attendu), nous considérons qu'il existe une asymétrie significative entre les allèles d'où la conclusion : Perte d'hétérozygotie.
- Si au contraire, les deux intensités sont proches l'une de l'autre (faible différence), cela ne suggère qu'aucun des deux allèles n'est perdu ou altéré de façon asymétrique⇒conclusion : Pas de perte.

2.5. Quantification globale et relative de l'expression génique : L'estimation de l'expression génique totale a été réalisée en additionnant les valeurs d'expression génique des deux allèles pour les gènes BRCA1/2.

Afin d'estimer le niveau d'expression global des gènes BRCA1 et BRCA2, nous avons calculé la somme des intensités de signal des deux allèles (Allèle A + Allèle B) pour chaque échantillon. Cette mesure permet d'évaluer l'expression totale du gène, indépendamment de la contribution relative de chaque allèle. Un seuil empirique fixé à 5 en échelle $\log_2$ a été utilisé pour distinguer les cas de sous-expression (≤5) des cas de surexpression ou gain (>5).

Parallèlement, le ratio d'expression allélique (Allèle A / (Allèle A + Allèle B)) a été déterminé pour détecter d'éventuels déséquilibres d'expression entre les deux allèles. Cette double approche

(somme et ratio) permet d'identifier des événements tels que la perte d'hétérozygotie fonctionnelle ou des altérations transcriptionnelles, même en l'absence de variations génomiques détectables.

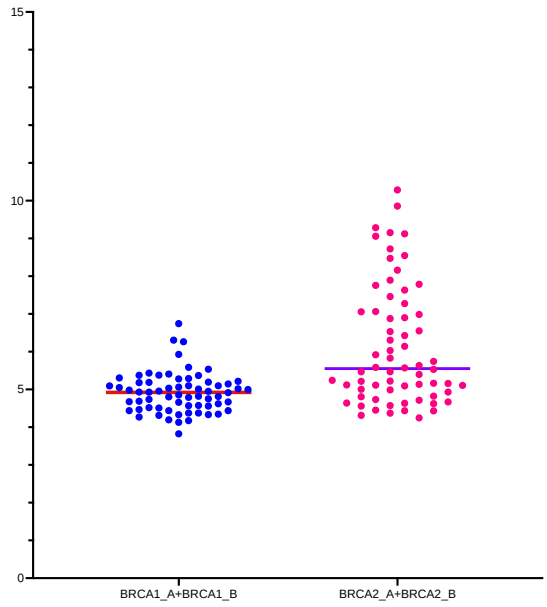

Figure 6 : Diagramme de dispersion des intensités individuelles des deux gènes BRCA1 et BRCA2

L'expression mathématique retenue pour mesurer le ratio est :

$$Ratio\ d'équilibre\ allélique = \frac{Allèle\ A}{Allèle\ A + Allèle\ B}$$

D'après l'équation proposée pour le calcul du ratio d'équilibre allélique, nous jugerons l'expression et la prédominance d'un allèle par rapport à l'autre selon les conditions :

- **Si Ratio < 0,5 → Allèle B prédomine** (L'allèle B est plus exprimé).
- **Si Ratio > 0,5 → Allèle A prédomine** (L'allèle B est moins exprimé).

RÉSULTATS ET DISCUSSION

1. STATISTIQUES DESCRIPTIVES

L'analyse biostatistique descriptive (tableau 3) révèle que les données d'expression génique des quatre allèles ne suivent pas une loi Normale (figure ppp) absolue. Le test de normalité de Kolmogorov-Smirnov le confirme avec KS-distance de l'allèle A = 0,155 et p=0,0004 et pour l'allèle B, la KS-distance = 0,1182 et p=0,0228 (plus proche de la normale que l'allèle A).Concernant le gène BRCA2, l'allèle BRCA2_A exprime une KS-distance de 0,2005 (p<0,0001) et l'allèle BRCA2_B montre une KS-distance de 0,1883 (p<0,0001).

Tableau 3 : Statistiques descriptives des expressions géniques pour les quatre allèles.

	BRCA1_A	BRCA1_B	BRCA2_A	BRCA2_B
Minimum	1,57	2,03	1,85	1,65
25% Percentile	1,96	2,32	2,31	2,38
Médiane	2,20	2,55	2,57	2,90
75% Percentile	2,44	2,92	2,92	4,09
Maximum	4,32	3,93	6,17	7,19
Étendue (max-min)	2,75	1,90	4,32	5,54
10% Percentile	1,84	2,12	2,19	2,05
90% Percentile	2,78	3,28	3,59	5,46
Moyenne	2,26	2,65	2,79	3,35
Écart-type	0,442	0,434	0,766	1,34

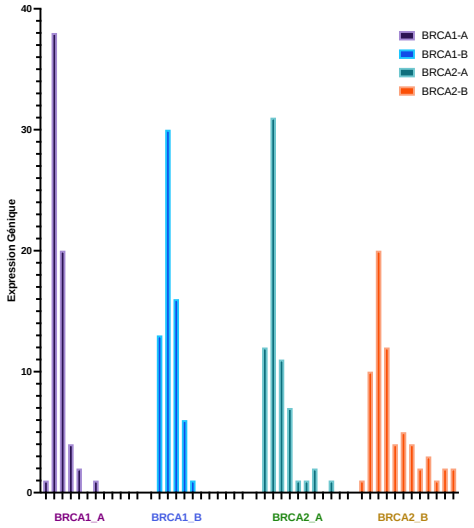

Figure 7 : Histogramme de distributions des valeurs d'expression génique pour les quatre allèles.

Les résultats du test Mann-Whitney révèlent une différence d'expression génique entre les deux allèles BRCA1_A et BRCA1_B (figure 8) et a donné une valeur U = 996 ($p<0,0001$).

Ce résultat signifie qu'il y a une différence très hautement significative entre les expressions des deux allèles (A et B) de BRCA1 ; ce qui suggère que l'allèle B de BRCA1 est considérablement plus exprimé que l'autre allèle et que le gène BRCA1 reflète une régulation différentielle avec un déséquilibre d'expression allélique.

Concernant le gène BRCA2, la comparaison d'expression allélique a révélé un test U = 1665 et p = 0,0193. Ce résultat indique une différence statistiquement significative entre les niveaux d'expression des allèles BRCA2_A et BRCA2_B, suggérant une préférence d'expression allélique ou une régulation différentielle de l'allèle B.

Globalement, en considérant les quatre allèles des deux gènes, la différence d'expression est plus marquée pour BRCA1 ($p<0,0001$) que pour BRCA2 ($p=0,0139$). Ce résultat indique que BRCA1 présente un déséquilibre d'expression allélique plus important dans les conditions étudiées.

Dans leur étude, Hou et al., (2023) sur le déséquilibre allélique dans le cancer du poumon, les auteurs confirment que l'allèle muté (ou altéré) peut prédominer en expression. Ils rapportent également que les gènes porteurs de mutation tumorale montrent une expression allélique préférentielle, plutôt qu'équilibrée comme dans le cas de notre étude.

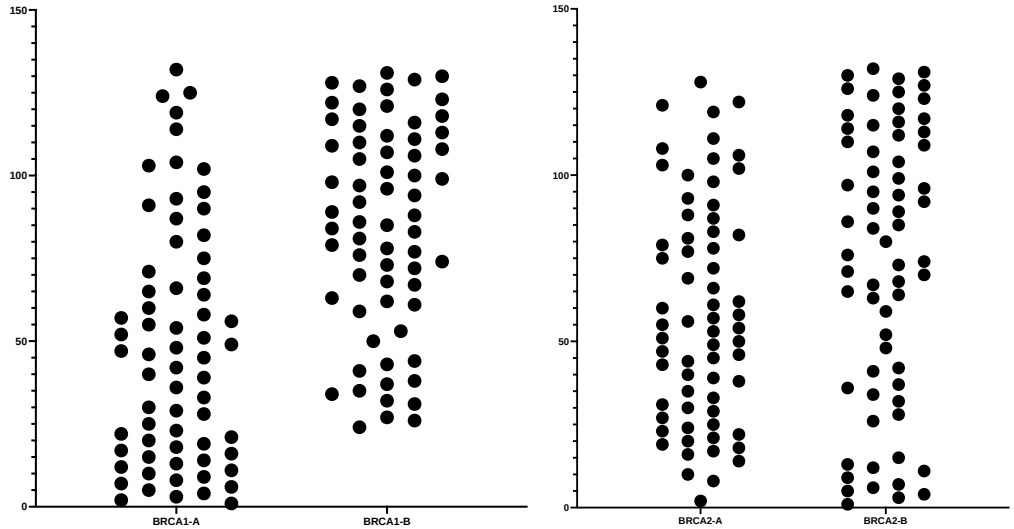

Figure 8 : Comparaison des niveaux d'expression génique chez BRCA1 et BRCA2

L’expression génique de BRCA1 et BRCA2 montre différents comportements entre leurs allèlesrespectifs(figure 9), avec un gradient d’expression croissant (BRCA1_A < BRCA1_B< BRCA2_A < BRCA2_B). Cela indique que BRCA2 est globalement plus exprimé que BRCA1 dans les deux allèles. L’allèle BRCA2_B semble avoir la plus grande variabilité, suggérant des sous-groupes ou une distribution possiblement bimodale ; cependant, BRCA1_1 montre une expression plus homogène.

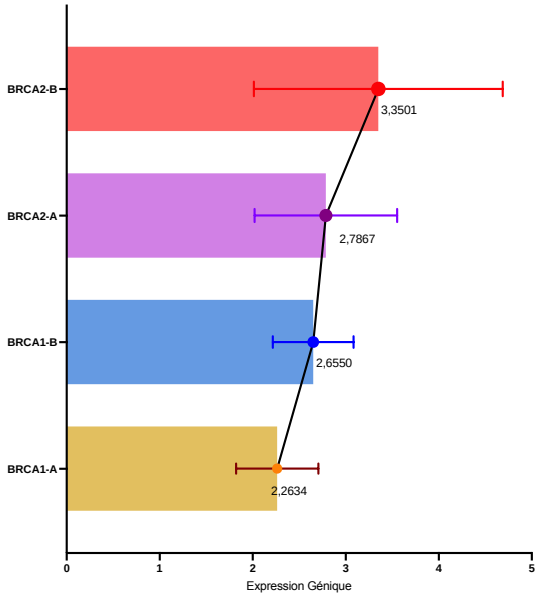

Figure 9 : Valeurs moyennes et comparaison entre les niveaux d’expression croissants.

La surexpression observée chez BRCA2_B pourrait être liée à un déséquilibre allélique ou un polymorphisme.

Une étude par Irani et Rafidazeh (2020) a exploré plus en détail les profils d'expression de BRCA1 et BRCA2 dans le carcinome épidermoïde œsophagien (CEO) par une analyse rétrospective. Leurs résultats ont indiqué que 63,3 % des tissus de CEO de grade intermédiaire et élevé présentaient une immunoréactivité cytoplasmique de BRCA1 modérée à forte, tandis que seulement 28,3 % présentaient une expression nucléaire de BRCA1. De même, l'expression de BRCA2 était entièrement cytoplasmique, 55,01 % des tissus de grade intermédiaire et élevé présentant une immunoréactivité cytoplasmique modérée à forte.

Gedeonova et al. (2025) affirment que des études récentes suggèrent que les profils d'expression des gènes BRCA1 et BRCA2 varient selon le grade tumoral, reflétant potentiellement leur rôle dans la progression tumorale.

2. ANALYSE DU DÉSÉQUILIBRE D'EXPRESSION GÉNIQUE

Analyse du déséquilibre d'expression génique : L'analyse du profil de chacun des quatre allèles (délétion, normal, gain et amplification) révèle un état de délétion chez chacun des deux allèles BRCA1_A et BRCA2_B. Les cas de délétion sont donc rares, avec seulement un cas pour chaque gène, affectant alternativement un seul allèle. Cela suggère une perte partielle et isolée d'expression (monoallélique), mais pas une inactivation complète.

L'état 'Normal' est plus observé chez BRCA1. En effet BRCA1_A est majoritairement exprimé (44 cas) dans un état 'Normal', bien plus que l'allèle BRCA1_B (16 cas). Pour le gène BRCA2, les deux allèles sont proches (15 et 12 cas respectivement), mais l'expression globale est plus faible que BRCA1 (44+16 > 15+12). Ce résultat pourrait refléter une expression de régulation différentielle des allèles, ou une désactivation partielle de BRCA2 dans les tissus étudiés.

L'état 'Gain' révèle un équilibre global entre les deux gènes (17+38 ≈ 35+22). L'allèle BRCA1_B montre un gain d'expression plus marqué (38 cas). Pour BRCA2, les deux allèles gagnent en expression, surtout BRCA2_A (35 cas).

L'état d'amplification est plus prononcé chez BRCA2 (16+31=47 cas) comparativement au gène BRCA1 (4+12=16 cas). Les deux gènes montrent une forte surexpression, surtout BRCA2_B (pic à 31). Ce résultat pourrait être associé à des mécanismes de compensation ou à des évènements oncogéniques ciblant cet allèle.

En fin, ces profils suggèrent un déséquilibre allélique fonctionnel, possiblement lié à la progression tumorale.

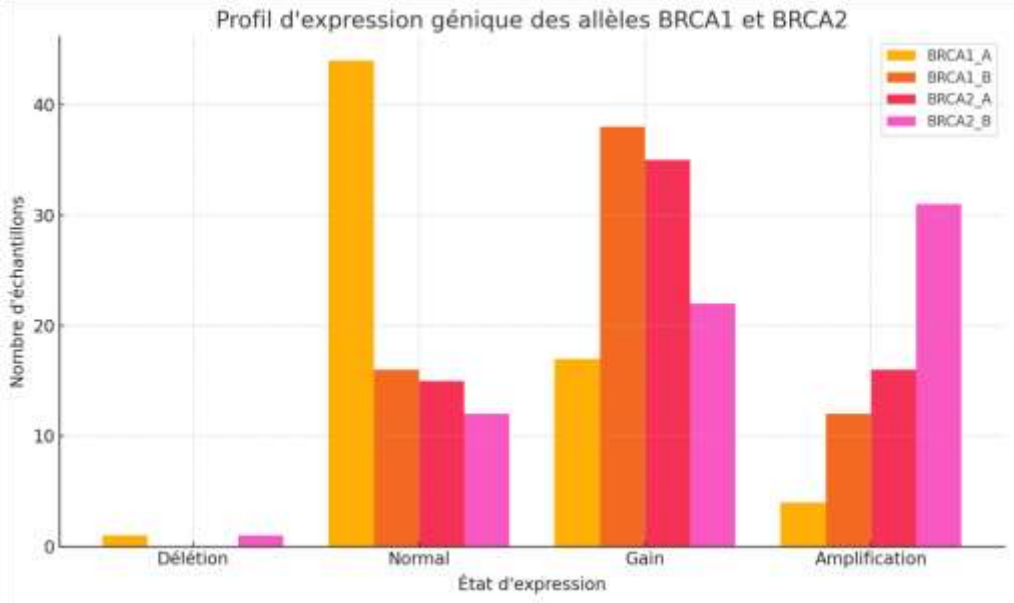

Figure 10 : Statut d'expression génique des allèles A et B des gènes BRCA1 et BRCA2.

La figure 11 illustre une répartition bimodale avec un regroupement central et deux groupes déviants, compatibles avec une altération de l'équilibre allélique chez certains échantillons.La zone entre les seuils 0,4 et 0,6 est synonyme d'une expression génique équilibrée entre les deux allèles A et B du gène BRCA1. Ces seuils empiriques de 0,4 et 0,6 sont indiqués pour suggérer des limites possibles de classification :

- Ratios < 0,4 : déséquilibre fort en faveur de l'autre allèle (possibilité de perte d'expression ou altération).
- Ratios > 0,6 : situation inverse, potentielle surexpression ou déséquilibre significatif.

De plus, la figure montre clairement :

- Un groupe "normal" avec une expression équilibrée des deux allèles,
- et deux groupes "anormaux" où un allèle est exprimé beaucoup plus ou beaucoup moins que l'autre, ce qui est cohérent avec une altération génétique chez certains patients.

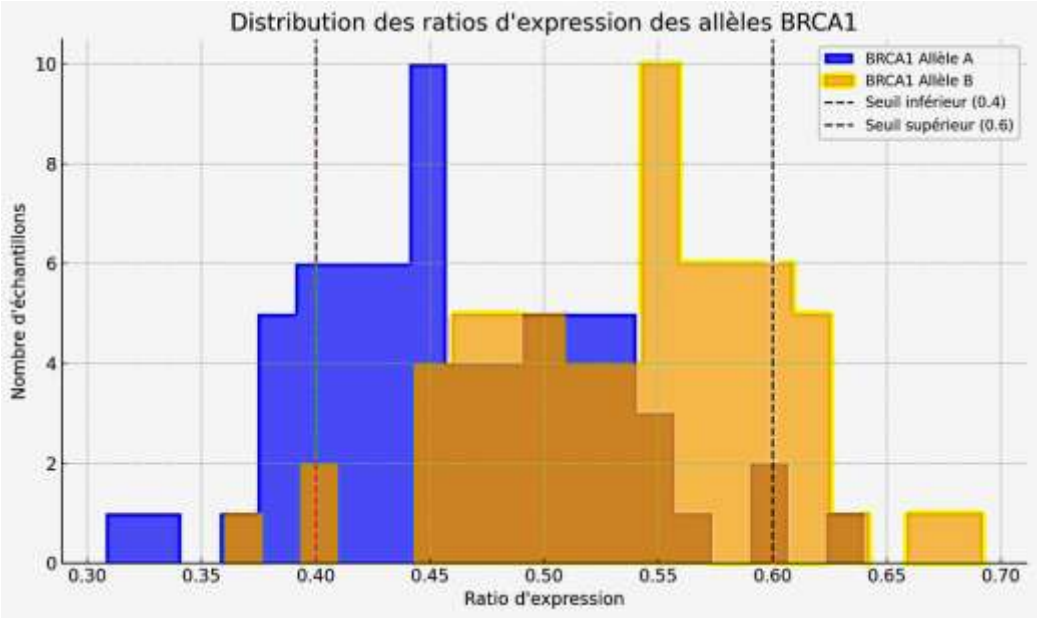

Figure 11 : Variabilité de l'équilibre allélique de BRCA1 dans les échantillons tumoraux analysés.

L'interprétation biologique de la bimodalité repose sur les points suivants :

1. Présence de deux populations d'échantillons : sur les 66 cas analysés, il y a 37 avec un ratio <0,5(donc allèle B prédomine) et 29 ont un ratio > 0,5, donc l'allèle A prédomine.

2. Déséquilibre allélique : Certains échantillons expriment préférentiellement un seul allèle, ce qui pourrait être dû à des mutations ou délétions affectant un allèle.

3. Zone d'équilibre (entre 0,4 et 0,6)

- La majorité des échantillons se situent dans cette zone, ce qui suggère une expression équilibrée des deux allèles dans la plupart des cas.
- Cependant, le fait que les deux pics soient proches des seuils 0,4 et 0,6 indique que plusieurs échantillons montrent une tendance au déséquilibre, sans franchir les seuils extrêmes.

Le gène BRCA2, révèle une répartition unimodale (figure 12), avec une légère tendance vers les valeurs inférieures à 0,5 ; ce qui indique une expression globalement plus élevée de l'allèle B. Toutefois, aucune bimodalité n'est visible.

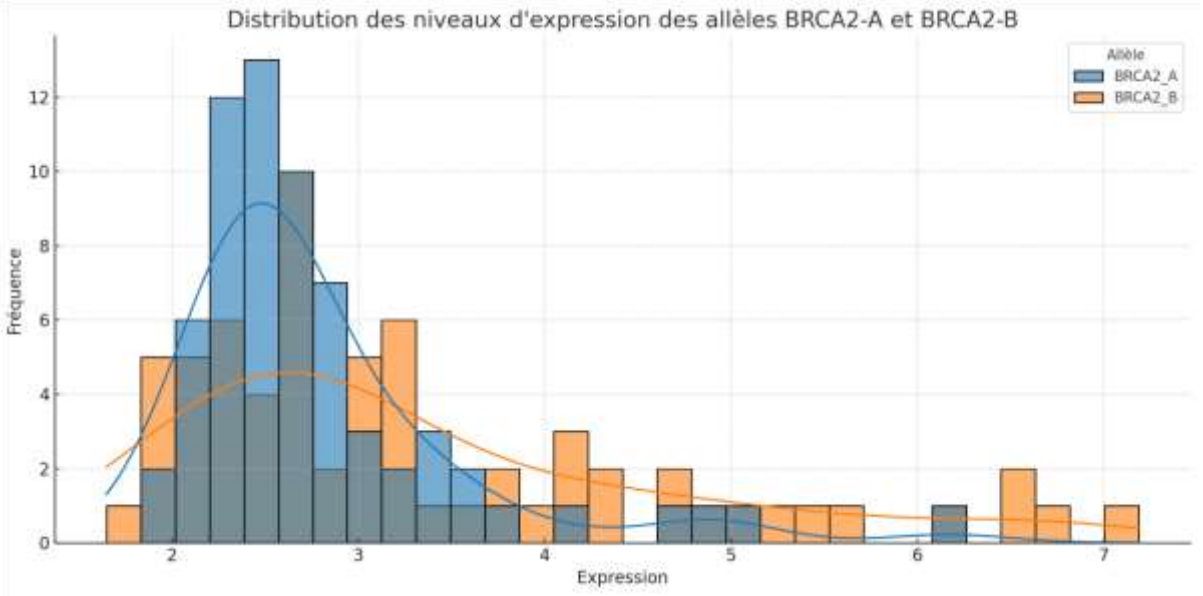

Figure 12 : Distribution des niveaux d'expression des allèles BRCA2_A et BRCA2_B.

Nous y observons une certaine bimodalité chez l'allèle BRCA2-B, suggérant qu'il pourrait exister deux sous-groupes d'échantillons présentant des profils d'expression différents.

L'analyse comparative des distributions d'expression génique de BRCA1/2 révèle une différence entre les deux allèles BRCA1_A et BRCA1_B selon le test non paramétrique de Mann-Whitney. La p-value extrêmement faible ($p<0,0001$) indique une différence très significative entre les distributions des intensités d'expression de l'allèle A et de l'allèle B du gène BRCA1. Cela pourrait signifier que l'allèle A et l'allèle B de BRCA1 sont exprimés de manière significativement différente dans les échantillons analysés ; d'où un déséquilibre allélique qui serait dû à une perte d'allèle (LOH). La p-value des deux allèles de BRCA2 est également significative ($p < 0,05$), bien que moins extrême que pour BRCA1. Les deux allèles de BRCA2 présentent eux aussi une différence significative, mais cette différence est moins marquée que pour BRCA1. Ces résultats indiquent une expression différentielle des allèles A et B pour les deux gènes BRCA, ce qui serait biologiquement pertinent, surtout dans un contexte de cancer.

La perte d'hétérozygotie (LOH) a été analysée par la mesure de la différence absolue de l'expression génique des deux allèles pour un même gène. Pour le gène BRCA1, Trois (03) pertes d'hétérozygotie ont été observées sur un total de 66 patients (4,5%). Cependant, le gène BRCA2 a montré 15 pertes (22,7%) au sein de la même population.

La perte d'hétérozygotie (LOH) désigne la disparition d'un des deux allèles d'un gène, souvent en lien avec un mécanisme tumoral. Chez un individu hétérozygote sain, les deux allèles sont présents et normalement exprimés. La perte d'un allèle (généralement le non muté) dans les cellules tumorales peut entraîner une perte de fonction si le second est déjà muté, ce qui est typique des gènes suppresseurs de tumeur, comme BRCA1 et BRCA2. Le taux de LOH pour BRCA2 est nettement plus élevé (22,7 >>4,5). Ce résultat pourrait suggérer que BRCA2 est plus fréquemment impliqué dans des altérations de type LOH dans notre cohorte de tumeurs pulmonaires, résultant d'une plus grande instabilité génomique touchant BRCA2, qui serait alors plus impliqué dans le développement tumoral.

La quantification globale d'expression allélique a révélé un taux de surexpression génique dee deux allèles A et B de BRCA1 égal à 40,91% (27/66) alors que chez BRCA2, le taux a été calculé à 72,73% (48/66). Cette surexpression est un signal d'activation génétique, c'est-à-dire une activation transcriptionnelle accrue du gène BRCA2 comparativement à BRCA1. Le gène BRCA2 semble alors plus dérégulé que BRCA1 dans notre cohorte. Ce résultat rejoint l'observation précédente où nous avons noté plus de pertes d'hétérozygotie dans BRCA2 suggérant qu'il était un acteur génétique central dans cette tumeur pulmonaire.

CONCLUSION

L'objectif principal de ce travail était d'analyser les déséquilibres d'expression allélique des gènes *BRCA1* et *BRCA2* à partir de données d'intensité issues de puces à ADN, dans le contexte de tumeurs humaines. Nos analyses statistiques et exploratoires ont permis de mettre en évidence des différences significatives entre les niveaux d'expression des allèles A et B de ces deux gènes, traduisant une régulation différentielle potentiellement liée à des mécanismes biologiques altérés en situation tumorale.

L'étude descriptive montre que les niveaux d'expression ne suivent pas une loi normale, ce qui a justifié l'utilisation de tests non paramétriques pour comparer les allèles. Ainsi, le test de Mann-Whitney a mis en évidence une différence très hautement significative entre les allèles *BRCA1_A* et *BRCA1_B* ($p < 0,0001$), et une différence significative entre *BRCA2_A* et *BRCA2_B* ($p = 0,0193$). Ces résultats suggèrent un déséquilibre d'expression allélique plus marqué pour *BRCA1* que pour *BRCA2* dans les échantillons analysés.

Par ailleurs, l'analyse du statut d'expression (normal, gain, amplification, délétion) révèle que l'allèle *BRCA2_B* est particulièrement sujet à une forte variabilité d'expression et à une fréquence élevée d'amplification (31 cas sur 66). De même, 72,7 % des cas montrent un gain ou une amplification pour *BRCA2,* contre 42,4 % pour *BRCA1*. Ces observations sont cohérentes avec une activité transcriptionnelle accrue et une possible implication oncogénique, notamment via la recombinaison homologue dans la réparation de l'ADN.

La mise en évidence d'une bimodalité de l'expression allélique de *BRCA1* et d'un déséquilibre significatif des ratios entre allèles souligne l'existence de deux sous-groupes de tumeurs : l'un exprimant majoritairement l'allèle A, l'autre l'allèle B. Ce phénomène est moins marqué pour *BRCA2,* qui présente une distribution globalement unimodale mais avec une variabilité importante pour *BRCA2_B.*

Ces résultats appuient l'hypothèse d'une régulation différentielle des allèles dans certains cancers, possiblement en lien avec des mutations, des polymorphismes affectant la stabilité de l'ARNm, ou encore des mécanismes épigénétiques ciblant sélectivement un allèle.

Perspectives

Bien que cette étude ait permis d'identifier des déséquilibres allélique d'expression dans les gènes *BRCA1* et *BRCA2*, elle constitue une étape exploratoire, dont plusieurs prolongements peuvent être envisagés :

1. **Validation par RT-qPCR ou RNA-seq** : Il serait pertinent de compléter cette analyse par une validation indépendante des résultats sur un sous-échantillon de tumeurs, en utilisant des techniques de quantification ciblée ou globale de l'expression génique.
2. **Corrélation avec les données cliniques** : Intégrer des paramètres cliniques (grade tumoral, stade, réponse au traitement) permettrait de mieux comprendre l'impact fonctionnel du déséquilibre allélique sur la progression tumorale.
3. **Recherche de variants cis-régulateurs (cis-eQTL)** : L'analyse des séquences génomiques proches des allèles surexprimés pourrait révéler des éléments de régulation (promoteurs, enhancers) responsables du déséquilibre observé.
4. **Extension à d'autres gènes de réparation de l'ADN** : Il serait envisageable d'étendre l'analyse à d'autres gènes impliqués dans la voie BRCA (PALB2, RAD51, etc.) pour cartographier les déséquilibres alléliques à l'échelle d'un réseau fonctionnel.

Finalement, cette étude met en lumière un déséquilibre d'expression allélique significatif des gènes *BRCA1* et *BRCA2* dans un contexte tumoral, en particulier une surexpression marquée de l'allèle *BRCA2_B* et un profil bimodal pour *BRCA1*. Ces observations soulignent l'importance de considérer l'expression différentielle des allèles comme un marqueur potentiel dans la compréhension des mécanismes tumoraux et dans la médecine personnalisée. Ce travail constitue un socle de réflexion utile pour des recherches futures alliant génomique, transcriptomique et bioinformatique intégrée.

RÉFÉRENCES BIBLIOGRAPHIQUES

Abdelhamid S., El-Mesallamy H., Abdelazizi H., Zekri A. 2021. Prognostic impact of BRCA1 and BRCA2 mutations on long-term survival outcomes in Egyptian female breast cancer patients. *Biology*. **10**:566-580. DOI:10.3390/biology10070566.

Chen L., Ye L., Hu B. 2022. Hereditary colorectal cancer syndromes: molecular genetics and precision medicine. *Biomedicines*. **10** (12) : 3207. DOI:10.3390/biomedicines10123207.

Cooper G. M. 2021. *The Cell: A Molecular Approach*. Sinauer Associates. 8th Edition. ISBN:1605357073.

Darras B. T., Urion D. K., Ghosh P. S. 2022. Dystrophinopathies. *GeneReviews*. University of Washington, Seattle. ISSN:2372-0697.

Dorschner, M.O., Amendola, L.M., et al. (2023). *Variant reclassification in clinical exome sequencing: longitudinal trends and implications for patient care*. **Genetics in Medicine**, 25(6), 100654. https://doi.org/10.1016/j.gim.2023.100654

Doudna J. A., Charpentier E. 2020. The new frontier of genome engineering with CRISPR-Cas9. *Science*. **367** (6481) : 1260-1262. https://doi.org/10.1126/science.aba5511.

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. 2018. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*.**173**(2):291-304 e6.

Hou Q., Shang L., Chen X., Luo Q., Wei L., Zhang C. 2023. Convergent evolution of allele-specific gene expression that leads to non-small cell lung cancer in different human populations. *J. Appl. Genetics*. **65** : 493-504. https://doi.org/10.1007/s13353-023-00813-4.

Irani S., Rafidazeh M. 2020. BRCA1/2 expression patterns in different grades of oral squamous cell carcinoma. *Middle East J. Cancer*. **11**:390-8. DOI:10.30476/mejc.2020.81282.0

Khatri P., Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. 2005. *Bioinformatics*.**21**:3587–95. DOI: 10.1093/bioinformatics/bti565.

Knudsen S. 2004.Guide to analysis of DNA microarray data. John Wiley & Sons, Inc.; Hoboken, NJ.

MacArthur D. G., Manolio T. A., Dimmock D. P., Rehm H. L., Shendure J., Abecasis G. R. et al. 2014. Guidelines for investigating causality of sequence variants in human disease. Nature. 508 : 469-476. DOI:10.1038/nature13127.

Moutschen M. 2022. Maladies auto-inflammatoires monogéniques : une introduction sur un mode translationnel aux inflammasomopathies. *Rev. Méd. Liège*. **77** (5-6) : 392-398. PMID:35657199.

Nussbaum R. L., McInnes R. R., Huntington F. W. 2022. *Thomson & Thomson's genetics in medicine*. Elsevier. 8th Edition. 512 p.

Phan L., Zhang H., Wang Q., Villamarin R., Hefferon T. Ramanathan A., Kattman B. 2025. The evolution of dbSNP : 25 years of impact in genomic research. *Nucleic Acids Research*. **53** (D1) : D925-D931. DOI: 10.1093/nar/gkae977.

Pleasant V., Boggan J., Richards B., Milliron K. J., Purrington K. S., Simon M. et al. 2025. Reclassification of variants of uncertain significance by race, ethnicity, and ancestry for patients at risk for breast cancer. *Frontiers in Oncology*. **15** : 1455509. DOI : 10.3389/fonc.2025.1455509.

Richards S., Aziz N., Bale S., Bick D., Das S., Gastier-Foster J. et al. 2015. Standards and guidelines for the interpretation of sequence variants : a joint consensus recommendation of the American College of Medical Genetics and Genomics ans the Association for Molecular Pathology. *Genetics in Medicine*. **17**:405-423. DOI:10.1038/gim.2015.30.

Roy R., Chun J., Powell S. N. 2012. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer*. **12** (1) : 68-78. DOI : 10.1038/nrc3181.

Salaun H., Saint-Ghislain M., Bellesoeur A., Beuzeboc P., Neuzillet C., Diéras V. et al. 2021. Homologous recombination deficiency and PARP inhibitors in therapeutics.*Bulletin du Cancer*. **109** (1) : 76-82.

Schena M. 2000. Microarray biochip technology. Eaton Publishing; Sunnyvale, CA.

Schubach M., Maass T., Nazaretyan L., Röner S., Kircher M. 2024. CADD v1.7:using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wine variant predictions. *Nucleic Acids Research*. **52** (D1):D1143-D1154. DOI: 10.1093/nar/gkad989.

Tarca A. L., Romero B., Draghici S. 2006. Analysis of microarray experiments of gene expression profiling. Am. J. Obstet. Gynecol. **195** (2) : 373-388. DOI:10.1016/j.ajog2006.07.001.

Tung N., Lin N. U., Kidd, J., et al. 2022. Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing. *JAMA Oncology*, **8** (1) : 1–8. DOI:10.1001/jamaoncol.2021.4530

Walsh N., Cooper A., Dockery A., O'Byrne J. J. 2024. Variant reclassification and clinical implications. Journal Medicine Genetics. 61 (3) : 207-211. DOI:10.1136/jmg-2023-109488.

Wendt C., Eickhoff J., Hu C. 2023. The clinical significance of BRCA1 and BRCA2 mutations beyond breast and ovarian cancer. *Cancer Research*. **83** (5) : 901–909. DOI : 10.1158/0008-5472.CAN-22-3569

Yaakoubi M. C. *HASH (0x2ad856121090)*. 2024. *Thèse de doctorat*. institution/ensta.

Yang Y. H., Buckley M. J., Dudoit S., Speed T. P. Comparison of methods for image analysis on cDNA microarray data. J. Comput Graph Stat. **11**:108-36.

Zhan H., Zhang Z. 2023. ProPath:Disease-Specific Protein Language Model for Variant Pathogenicity. arXiv. 8:1-13. DOI: 10.48550/arXiv.2311.03429.

Zhang X, Jafari N, Barnes RB, Confino E, Milad M, Kazer RR. 2005. Studies of gene expression in human cumulus cells indicate pentraxin 3 as a possible marker for oocyte quality. Fertil Steril. **83**:1169–79. DOI: 10.1016/j.fertnstert.11.030.

Zhou W., Zhao Y., Palmer C. 2021. Genetic variant analysis in the era of precision medicine. *Nat. Rev. Gen*. **22** (11) : 669-687. DOI:1038/s41576-021-00374-2.

Zhu Y., Pei X., Novaj A., Setton J., Bronder D., Darakhshan F. et al. 2024. Large-scale copy number alterations are enriched for synthetic viability in BRCA1/BRCA2tumors.*Genome Medicine*.**16** : Article. 108. DOI: https://doi.org/10.1186/s13073-024-01371-y