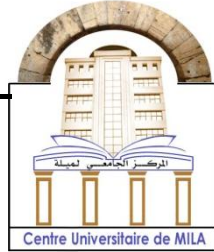


الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique



N°Ref :.....

Centre Universitaire Abdelhafid BOUSSOUF- Mila

Institut des Sciences de la Nature et de la Vie

Département des Sciences Biologiques et Agricoles

**Mémoire préparé en vue de l'obtention du diplôme de
Master**

Domaine : Sciences de la Nature et de la Vie

Filière : Sciences Biologiques

Spécialité : Biochimie Appliquée

Thème :

Séquençage de l'exome humain par NGS

Présenté par :

➤ **KOKO Souha**

Devant le jury :

Nom et Prénom : BOUCHERIT Hanane MCA

Présidente

Nom et Prénom : MERZOUG Amina MCA

Examinatrice

Nom et Prénom : CHEKROUD Karim MCA

Promoteur

Année Universitaire : 2024/2025

Remerciements

Nous remercions tout d'abord Dieu Tout-Puissant de nous avoir accordé santé, bien-être, courage et volonté pour mener à bien ce modeste projet.

*Je remercie sincèrement le professeur **CHEKROUD.K** pour ses conseils et son aide précieuse dans la réalisation de ce projet.*

Je remercie également sincèrement tous les membres du comité d'arbitrage : le rapporteur de cette thèse, qui nous a fait l'honneur de siéger au comité de discussion.

Et tous nos professeurs qui nous ont soutenus tout au long de la formation.

Je remercie tous les membres du laboratoire pour leurs contributions directes et indirectes.

Nous exprimons également notre gratitude à tous mes camarades de classe.

Je remercie également toutes les personnes qui m'ont aidé, directement ou indirectement, tout au long de la réalisation de ce travail.

Dédicace

*Je dédie ce mémoire à mes chers parents, **Koko Elhade** et **Kerkeb Saïda**, pour leur amour, leurs sacrifices et leur soutien indéfectible tout au long de mon parcours.*

*À mon encadrant, **Monsieur Chekroud Karim**, pour ses précieux conseils, sa disponibilité et son accompagnement tout au long de ce travail.*

*Aux familles **Koko**, pour leur appui, leur affection et leur présence constante.*

À mes amis de cette année, pour les moments partagés, leur soutien et leur bonne humeur qui m'ont portée jusqu'au bout.

À vous tous, ma profonde reconnaissance et mes sincères remerciements.

Séquençage de l'exome humain par NGS

Résumé

Le séquençage de l'exome humain (Whole Exome Sequencing, WES) est une approche ciblée du séquençage génomique, qui consiste à analyser spécifiquement les régions codantes du génome, appelées exons. Bien que ces régions ne représentent qu'environ 1 à 2 % du génome total, elles concentrent près de 85 % des mutations responsables de maladies génétiques connues. Grâce à l'émergence des technologies de séquençage haut débit (NGS – Next-Generation Sequencing), le séquençage de l'exome est devenu un outil clé dans les domaines de la recherche biomédicale et du diagnostic génétique.

Ce mémoire présente les différentes étapes techniques et bioinformatiques du séquençage de l'exome humain, de l'extraction de l'ADN à l'interprétation des variants. Il décrit la préparation des bibliothèques d'ADN, l'enrichissement des régions exoniques, le séquençage à haut débit (notamment via la plateforme MGI), ainsi que le traitement des données (fichiers FASTQ, alignement, appel de variants, annotation et interprétation des fichiers VCF).

Les applications médicales du WES sont nombreuses, notamment dans le diagnostic des maladies rares, les troubles neurologiques ou encore certaines pathologies oncologiques. Toutefois, cette approche présente aussi des limites, comme l'incapacité à détecter certaines mutations structurelles ou régulatrices non couvertes par l'exome.

Ainsi, le séquençage de l'exome humain par NGS constitue une méthode efficace, rapide et relativement économique, offrant de grandes perspectives pour la médecine personnalisée et la compréhension des mécanismes génétiques des maladies humaines.

Mots-clés : Séquençage de l'exome, NGS, ADN, diagnostic génétique, analyse bioinformatique, MGI, variants, VCF.

Human exome sequencing by NGS

Abstract

Whole exome sequencing (WES) is a targeted approach to genomic sequencing that involves specifically analyzing the coding regions of the genome, called exons. Although these regions represent only approximately 1 to 2% of the total genome, they account for nearly 85% of the mutations responsible for known genetic diseases. Thanks to the emergence of high-throughput sequencing (NGS – Next-Generation Sequencing) technologies, exome sequencing has become a key tool in biomedical research and genetic diagnosis.

This thesis presents the various technical and bioinformatics steps of human exome sequencing, from DNA extraction to variant interpretation. It describes the preparation of DNA libraries, the enrichment of exonic regions, high-throughput sequencing (particularly via the MGI platform), as well as data processing (FASTQ files, alignment, variant calling, annotation, and interpretation of VCF files).

The medical applications of WES are numerous, particularly in the diagnosis of rare diseases, neurological disorders, and certain oncological pathologies. However, this approach also has limitations, such as the inability to detect certain structural or regulatory mutations not covered by the exome.

Thus, human exome sequencing using NGS is an efficient, rapid, and relatively economical method, offering great potential for personalized medicine and understanding the genetic mechanisms of human diseases.

Keywords: Exome sequencing, NGS, DNA, genetic diagnosis, bioinformatics analysis, MGI, variants, VCF.

تسلسل الإكسوم البشري بواسطة NGS

ملخص

تسلسل الإكسوم الكامل (WES) هو نهج مستهدف للتسلسل الجينومي الذي يحلل بشكل خاص مناطق الترميز في الجينوم، والتي تسمى الإكسونات. وعلى الرغم من أن هذه المناطق تمثل حوالي 1 إلى 2% فقط من إجمالي الجينوم، فإنها تتركز فيها ما يقرب من 85% من الطفرات المسؤولة عن الأمراض الوراثية المعروفة. بفضل ظهور تقنيات التسلسل عالية الإنتاجية (NGS – Next-Generation Sequencing)، أصبح تسلسل الإكسوم أداة رئيسية في مجالات البحث الطبي الحيوي والتشخيص الجيني.

تقدم هذه الأطروحة المراحل التقنية والمعلوماتية الحيوية المختلفة لتسلسل الإكسوم البشري، من استخراج الحمض النووي إلى تفسير المتغيرات. يصف هذا الكتاب إعداد مكتبات الحمض النووي، وإثراء المناطق الخارجية، والتسلسل عالي الإنتاجية (خاصة عبر منصة MGI)، بالإضافة إلى معالجة البيانات (ملفات FASTQ)، والمحاذاة، واستدعاء المتغيرات، والشرح والتفسير لملفات VCF.

إن التطبيقات الطبية لـ WES عديدة، وخاصة في تشخيص الأمراض النادرة والاضطرابات العصبية وبعض الأمراض الأورامية. ومع ذلك، فإن هذا النهج له أيضًا حدود، مثل عدم القدرة على اكتشاف بعض الطفرات الهيكلية أو التنظيمية التي لا يغطيها الإكسوم.

وبالتالي، فإن تسلسل الإكسوم البشري بواسطة NGS يشكل طريقة فعالة وسريعة واقتصادية نسبيًا، مما يوفر أفقًا عظيمة للطب الشخصي وفهم الآليات الجينية للأمراض البشرية.

كلمات المفتاحية: تسلسل الإكسوم، NGS، الحمض النووي، التشخيص الوراثي، تحليل المعلومات الحيوية، MGI، المتغيرات، VCF

Abréviations

- **ADN** Acide Désoxyribonucléique
- **BRCA1** Breast cancer
- **CCNF** La cycline-F spécifique de la mitose G2
- **CFTR** *pour l'anglais cystic fibrosis trans membrane conductance regulator*
- **Le NGS** *Next-Generation Sequencing*
- **PCR** Polymerase Chain Reaction
- **SLC12A3** *Solute Carrier Family 12 Member 3*
- **SNP** Single Nucleotide Polymorphisms
- **TPH1** La tryptophane hydroxylase 1
- **TYRO3** Le récepteur de la tyrosine-protéine kinase
- **VCF** *Variant Call Format*
- **WES** *Whole exome sequencing*
- **WGS** *Whole genome sequencing*

Liste des tableaux

Tableau N°	Titre	Page N°
1	la taille des fragments isolés à partir de 100µL d'ADN fragmenté en fonction de la quantité de billes magnétiques	30
2	mélange pour la réparation d'extrémité et de queue A	31
3	Le mélange de ligature de l'adaptateur	32
4	le mélange Amplification par PCR	34
5	Préparation du mélange d'hybridation	36
6	Préparation du mélange de capture	36
7	mélanger et lancer la PCR	38
8	le mélange de réaction de circularisation	39

Liste des figures

Figure N°	Titre	Page N°
1	Représentation simplifiée d'un gène d'eucaryote (les exons sont les séquences présentes dans le transcrit mature, alors que les introns sont retirés pendant l'épissage). Ce gène peut aussi bien être codant que non codant.	8
2	Épissage d'un pré-ARN messenger : excision des introns et suture des exons	8
3	Souris avec différentes couleurs de pelage	8
4	Il repose sur trois étapes clés Le séquençage Sanger	15
5	Séquence a déterminer marquée	17
6	Tailles des séquences lues (Reads) après « Trimming »	45
7	Nombre moyen de lectures (reads) par base.	46
8	%GC des lectures (reads)	46

Table des matières

Remerciements	
Dédicace	
Résumé	
Abstract	
ملخص	
Abréviations	
Liste des tableaux	
Liste des figures	
Table des matières	
Introduction	1

Partie I

Chapitre I. Information générale les Génomes

I. Le génome	6
I.1. Définition du génome	6
I.1.1. Le Génome Humain	6
I.2. Comparaison des Génomes chez les Organismes Vivants	6
I.2.1. Chez les virus	6
I.2.2. Chez les procaryotes (bactéries et archées)	6
I.2.3. Chez les eucaryotes	7
I.2.4. Chez l'être humain (organisme eucaryote)	7
II. Le Gène	7
II.1. Gènes et évolution	8
III. Les Mutations Génétiques	10
III.1. Définition de mutation	10
III.2. Types de mutations génétiques	10

III.2.1. Mutations ponctuelles	10
III.2.2. Délétions et duplications	11
III.2.3. Conversion génique	11
III.2.4. Fusion de gènes	11
III.2.5. Insertions	12
III.2.6. Inversions	12
III.2.7. Mutations perturbant l'épissage	12
III.2.8. Délétions et insertions de petite taille	12
III.2.9. Mutations dynamiques ou instables	12
III.3. Impact des mutations sur les caractéristiques héréditaires	13

Chapitre II. Le Séquençage de l'ADN

Le Séquençage de l'ADN	15
I. Les méthodes de Séquençage ADN	15
I.1. Le Séquençage de Sanger : Une Méthode Fondamentale	15
I.1.1. Principes du Séquençage	15
I.2. Le séquençage Maxam-Gilbert	16
I.2.1. Principe de la méthode	16
II. Les séquenceurs à haut débit	17
II.1. Séquençage par hybridation	17
II.1.1. Principe de l'hybridation de l'ADN	17
II.2. Le séquençage Illumina	18
II.2.1. Principe du séquençage Illumina	18
II.3. Le séquençage Nanopore	18
II.3.1. Principe	19
II.4. Séquençage à courtes lectures par DNB	19
II.4.1. Principe	19

III. Analyse des Exons et Séquençage de l'Exome	19
III.1. Importance et limites du séquençage de l'exome	20
III.2. Séquençage de l'exome vs. Séquençage du génome entier	20
III.3. Analyse des Données de Séquençage	20
IV. Méthodes pour améliorer la qualité des données	21
IV.1. Identification et stockage des variants génétiques	21

Chapitre III. Maladie héréditaire

I. Maladie héréditaire	23
II. Types de maladies héréditaires	24
II.1. Maladies monogéniques (mutation d'un seul gène)	24
II.2. Maladies chromosomiques (anomalies dans les chromosomes)	24
II.3. Maladies multifactorielles (combinaison de gènes et de facteurs environnementaux)	24
III. Transmission des maladies héréditaires	25
IV. Prévalence	25
V. Dépistage et diagnostic	25
VI. Qui sont les personnes à risque ?	25
VII. Quels traitements ?	26
VIII. Qu'est-ce que l'analyse de l'Exome?	26
VIII.1. Types d'analyses génétiques	26
VIII.2. Types d'Exomes	26

Partie II : Matériel et méthodes

Matériel et Méthodes	29
I. Méthodes	29
I.1. Echantillon biologique	29
I.2. Extraction et purification de l'ADN	29
I.3. Fragmentation de l'ADN	29

I.3.1. Sélection des Fragments d'ADN en fonction de leurs tailles	30
I.3.2. Etapes de la sélection des fragments	30
I.4. Réparation de l'extrémité des fragments et ajout de la queue « A »	31
I.5. Ligature de l'adaptateur	32
I.5.1. Mélange de ligature d'adaptateur	32
I.5.2. Programme de ligature d'adaptateur	33
I.5.3. Nettoyage de l'ADN ligaturé par un adaptateur	33
I.6. Amplification par PCR	34
I.6.1. Nettoyage du produit PCR	35
I.7. Hybridation et capture	35
I.8. Elution	37
I.9. PCR Post-Capture	38
I.10. Dénaturation et circularisation simple brin	39
I.11. Circularisation monocaténaire	39
I.12. Digestion	40
I.12.1. Nettoyage du produit de digestion	41
II. Séquençage sur la plateforme MGI	41
III. Données brutes : fichiers FASTQ	42
IV. Analyse bioinformatique	42
IV.1. Alignement des lectures	42
IV.2. Appel des variants (Variant Calling)	42
V. Fichiers VCF	42
VI. Analyse diagnostique	42
VI.1. Annotation des variants	42
VI.2. Interprétation clinique	42

Partie III : Résultats et discussion

I. Résultats	45
I.1. Matériel biologique de départ	45
I.2. Préparation de la librairie	45
I.3. Paramètre de séquençage	45
I.4. Analyse bioinformatique	46
I.5. Données brutes : fichiers FASTQ	46
I.6. Alignement des lectures	47
I.7. Appel des variants (Variant Calling)	47
I.8. Annotation des variants	47
I.9. Fichiers VCF	48
I.10. Interprétation clinique	48
II. Discussion	48
Conclusion Générale	51
Références	53
Annexes	

Introduction

Introduction

La génétique est la science qui étudie les gènes afin de comprendre leur structure, leur fonction et leur transmission. Les gènes sont les unités fondamentales de l'hérédité : ils permettent aux organismes vivants de transmettre leurs caractéristiques d'une génération à l'autre. Ainsi, un enfant ressemble souvent à ses parents car il hérite de leurs gènes. La génétique vise à identifier ces traits héréditaires et à expliquer les mécanismes de leur transmission **(1)**.

Certains caractères héréditaires sont visibles, tels que la couleur des yeux ou la taille, tandis que d'autres, comme le groupe sanguin ou la prédisposition à certaines maladies, sont moins apparents. Certains traits sont strictement génétiques, alors que d'autres résultent de l'interaction entre les gènes et l'environnement. Par exemple, un enfant génétiquement prédisposé à être grand peut rester de petite taille s'il souffre de malnutrition. Ainsi, les caractéristiques phénotypiques d'un individu résultent souvent d'une interaction complexe entre facteurs génétiques et environnementaux. Le développement de certaines pathologies telles que les cancers ou les maladies cardiovasculaires illustre cette interaction : leur survenue dépend à la fois de mutations héréditaires et de facteurs environnementaux, tels que l'alimentation, le mode de vie ou l'exposition à des agents toxique **(1)**.

Les gènes sont constitués d'ADN (acide désoxyribonucléique), une molécule formée d'une séquence de nucléotides

— adénine (A), thymine (T), guanine (G) et cytosine (C)

organisés selon un ordre précis. L'ADN porte l'information génétique nécessaire à la synthèse des protéines, molécules essentielles au fonctionnement cellulaire. Le code génétique est le système de correspondance entre la séquence de nucléotides et la séquence des acides aminés constituant les protéines. Il permet la traduction de l'information génétique en structures fonctionnelles **(2)**

Les gènes présentent des variations interindividuelles appelées allèles. Ces variations peuvent entraîner des différences phénotypiques telles que la couleur des cheveux ou la susceptibilité à certaines maladies. Les mutations génétiques, qui correspondent à des modifications aléatoires de la séquence d'un gène, peuvent engendrer de nouveaux allèles. Ces mutations constituent l'un des mécanismes clés de la variation génétique sur laquelle agit la sélection naturelle, jouant un rôle fondamental dans l'évolution des espèces **(2)**

1. Gènes et hérédité

Les gènes sont des segments d'ADN codant pour des acides ribonucléiques (ARN) ou des protéines. Lors de la reproduction sexuée, chaque parent transmet une copie de ses gènes à sa descendance. Les cellules humaines possèdent deux copies de chaque gène (allèles), une d'origine maternelle et l'autre d'origine paternelle. La fusion d'un ovule et d'un spermatozoïde lors de la fécondation forme un zygote, qui possède ainsi un patrimoine génétique complet, mélange unique des gènes parentaux **(2)**.

2. Réplication de l'ADN et mutations

Avant chaque division cellulaire, l'ADN est dupliqué par un processus appelé réplication, assurant la transmission fidèle du matériel génétique aux cellules filles. La molécule d'ADN est constituée de deux brins enroulés en hélice. Lors de la réplication, ces brins sont séparés, et chaque brin sert de matrice à la synthèse d'un nouveau brin complémentaire **(3)**.

Bien que la réplication soit un processus très précis, des erreurs peuvent survenir. Ces erreurs peuvent provoquer des mutations, c'est-à-dire des changements dans la séquence de nucléotides. Certaines mutations sont neutres, d'autres peuvent être bénéfiques ou, au contraire, délétères, en altérant la fonction des protéines codées **(3)**.

Si ces mutations surviennent dans les cellules germinales, elles peuvent être transmises à la descendance et contribuer à la diversité génétique de la population. Cette diversité explique les différences entre individus, même au sein d'une même famille **(3)**.

3. Les gènes et la synthèse des protéines

Les cellules, unités de base de tout être vivant, utilisent les gènes comme instructions pour fabriquer les protéines. Ces dernières jouent des rôles variés : catalyse des réactions chimiques (enzymes), signalisation cellulaire, structure, transport, réparation de l'ADN, etc.

Une protéine est formée d'une chaîne d'acides aminés dont l'ordre est déterminé par la séquence du gène correspondant. Une fois la chaîne synthétisée, elle se replie dans une conformation tridimensionnelle spécifique, essentielle à sa fonction. Certaines protéines possèdent des sites actifs spécifiques leur permettant d'interagir avec d'autres molécules de manière sélective **(3)**.

Le processus de synthèse protéique commence par la transcription : la séquence d'un gène est copiée en ARN messager (ARNm). Cet ARNm est ensuite traduit en chaîne d'acides aminés selon le code génétique **(3)**.

Les allèles, qui diffèrent souvent par une ou plusieurs bases nucléotidiques, peuvent engendrer des modifications de la protéine codée, affectant ainsi ses propriétés ou sa fonction. Ces variations expliquent une grande partie de la diversité biologique entre individus **(3)**.

4. Réplication de l'ADN et mutations

Avant chaque division cellulaire, l'ADN est dupliqué par un processus appelé réplication, assurant la transmission fidèle du matériel génétique aux cellules filles. La molécule d'ADN est constituée de deux brins enroulés en hélice. Lors de la réplication, ces brins sont séparés, et chaque brin sert de matrice à la synthèse d'un nouveau brin complémentaire **(2)**.

Bien que la réplication soit un processus très précis, des erreurs peuvent survenir. Ces erreurs peuvent provoquer des mutations, c'est-à-dire des changements dans la séquence de nucléotides. Certaines mutations sont neutres, d'autres peuvent être bénéfiques ou, au contraire, délétères, en altérant la fonction des protéines codées **(2)**.

Si ces mutations surviennent dans les cellules germinales, elles peuvent être transmises à la descendance et contribuer à la diversité génétique de la population. Cette diversité explique les différences entre individus, même au sein d'une même famille **(2)**.

Partie I

Chapitre I. Information générale les Génomes

I. Le génome

I.1. Définition du génome

Le Génome : L'Ensemble du Matériel Génétique

Le génome désigne l'intégralité du matériel génétique d'un organisme. Il comprend l'ensemble de l'ADN (acide désoxyribonucléique) qui contient les instructions nécessaires à la croissance, au développement et au fonctionnement de l'organisme. Il regroupe non seulement tous les gènes, mais aussi des séquences non codantes, essentielles à la régulation de l'expression des gènes.

La taille et la structure du génome varient selon les espèces. Par exemple, le génome humain comprend environ 3 milliards de paires de bases, réparties sur 23 paires de chromosomes, tandis que celui d'une bactérie peut ne contenir que quelques millions de paires de bases, organisées sur un chromosome circulaire unique.

En plus de l'ADN nucléaire, certains organismes possèdent de l'ADN mitochondrial, comme chez l'être humain, où il joue un rôle crucial dans la production d'énergie cellulaire **(5)**.

I.1.1. Le Génome Humain

Le génome humain regroupe l'ensemble des séquences d'acides nucléiques constituant l'ADN. Il est réparti entre 23 paires de chromosomes situées dans le noyau des cellules (génome nucléaire) et une petite molécule d'ADN présente dans chaque mitochondrie (génome mitochondrial) **(5)**.

I.2. Comparaison des Génomes chez les Organismes Vivants

I.2.1. Chez les virus

Le génome viral peut être constitué d'une ou plusieurs molécules d'ADN (virus à ADN) ou d'ARN (virus à ARN ou ribovirus). Il peut être simple ou double brin et est protégé par une capside protéique. Certains scientifiques considèrent les virus non pas comme de véritables organismes, mais comme des parasites moléculaires **(6)**.

I.2.2. Chez les procaryotes (bactéries et archées)

Le génome est généralement formé d'une molécule d'ADN circulaire. Certaines bactéries possèdent également un génome extrachromosomique contenu dans des plasmides ou des épisomes. Toutefois, certaines bactéries, comme les actinomycètes, ont des génomes linéaires **(6)**.

I.2.3. Chez les eucaryotes

On distingue plusieurs types de génomes :

- Le génome nucléaire : Situé dans le noyau, il constitue l'essentiel du génome des eucaryotes (animaux, plantes, champignons, etc.).
- Les génomes non nucléaires, présents dans certains organites cellulaires :
 - Le génome mitochondrial : Situé dans les mitochondries, il est composé de plusieurs molécules circulaires d'ADN mitochondrial (ADNmt), semblables à l'ADN bactérien.
 - Le génome chloroplastique : Présent dans les chloroplastes des eucaryotes photosynthétiques (plantes et algues).
 - Les génomes cytoplasmiques : Chez certains eucaryotes (comme la levure), on trouve également de petits plasmides et des particules virales (7).

I.2.4. Chez l'être humain (organisme eucaryote)

Le génome nucléaire est réparti sur 46 chromosomes, comprenant 22 paires d'autosomes et une paire de chromosomes sexuels : XX chez la femme et XY chez l'homme (6).

II. Le Gène

Un gène est une séquence d'ADN contenant l'information nécessaire à la production d'une protéine spécifique, qui influencera ensuite un caractère particulier ou l'une de ses variantes. Ces gènes déterminent de nombreux traits, tels que la couleur des yeux et des cheveux, le groupe sanguin, ou encore la taille d'un individu. En d'autres termes, ils sont les instructions génétiques responsables du développement et du fonctionnement de notre organisme. (8).

Tout l'ADN d'un organisme ne contient pas uniquement des gènes. Une partie, souvent qualifiée d'"ADN non codant" ou parfois d'"ADN poubelle", ne joue pas directement un rôle dans la production de protéines. Cependant, certaines de ces séquences sont essentielles, car elles servent de régulateurs en permettant la fixation de protéines spécifiques ou en protégeant l'ADN contre la dégradation (8).

Les gènes eux-mêmes comportent des parties non codantes. Lors de la transcription de l'ADN en ARN, certaines sections appelées introns sont éliminées avant que la séquence ne soit traduite en protéine. En revanche, les exons, qui contiennent l'information codante, sont conservés et assemblés pour produire la protéine finale (9).

Dans le génome humain, un gène contient en moyenne 9 exons et 8 introns, bien que cette répartition varie considérablement. Certains gènes, comme ceux codant pour les histones, ne possèdent qu'un seul exon et aucun intron (9).

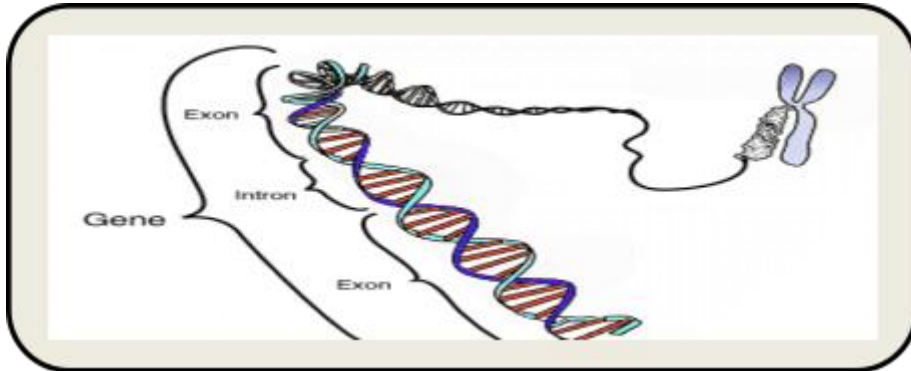


Figure N° 1 : Représentation simplifiée d'un gène d'eucaryote (les exons sont les séquences présentes dans le transcrit mature, alors que les introns sont retirés pendant l'épissage). Ce gène peut aussi bien être codant que non codant.(41).

Chez l'humain, seuls environ 3 % des gènes sont dépourvus d'introns et ne comportent qu'un seul exon.

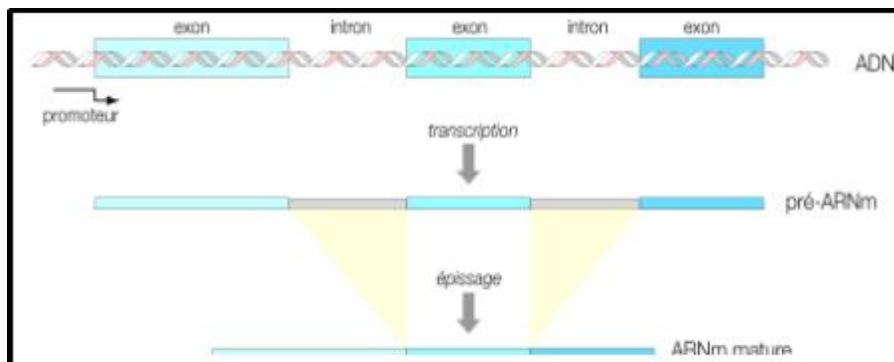


Figure N° 2 : Épissage d'un pré-ARN messenger : excision des introns et suture des exons (42).

II.1. Gènes et évolution



Figure N° 3 : Souris avec différentes couleurs de pelage (40).

L'évolution d'une population d'organismes se produit lorsqu'un trait devient plus ou moins fréquent au fil du temps. Par exemple, toutes les souris vivantes sur une île forment une seule population, avec certaines ayant une fourrure blanche et d'autres une fourrure grise. Si, au fil des générations, le nombre de souris blanches augmente tandis que celui des souris grises diminue, on observe alors une évolution de la couleur de la fourrure au sein de cette population. En génétique, ce phénomène correspond à une modification de la fréquence allélique **(10)**.

La fréquence des allèles dans une population peut changer soit de manière aléatoire, par un phénomène appelé dérive génétique, sélection naturelle.

Dans l'exemple précédent, si l'île devenait progressivement plus froide et superficielle de neige, les souris à fourrure blanche auraient un avantage, car elles seraient moins visibles pour les prédateurs, contrairement aux souris grises qui seraient plus exposées. Par conséquent, la proportion de souris blanches augmenterait au fil des générations, tandis que celle **(10)**.

Les mutations génétiques génèrent de nouveaux allèles, qui codent pour des protéines aux propriétés inédites. Par exemple, si une île est peuplée uniquement de souris grises, des mutations peuvent survivre et donner naissance à des allèles responsables

L'évolution résultant de la combinaison entre ces mutations aléatoires et la sélection naturelle, qui fait adaptation, biologie évolutive du développement. **(10)**.

Certaines maladies sont **héréditaires** et se transmettent de génération en génération, tandis que d'autres, comme les **maladies infectieuses**, résultent de l'interaction avec l'environnement. Certaines affections peuvent également être causées par une combinaison de ces deux facteurs **(10)**.

Les **troubles génétiques** sont des maladies provoquées par une mutation dans un unique gène et transmises au sein des familles. Parmi elles, on retrouve la **maladie de Huntington**, la **mucoviscidose** (ou **fibrose kystique**) et la **myopathie de Duchenne**. La mucoviscidose, par exemple, est due à des mutations du gène **CFTR** et se transmet selon un mode **récessif** **(11)**.

Certaines maladies sont également influencées par la génétique, mais les gènes hérités des parents ne déterminent pas directement leur apparition ; ils modifient simplement le risque de les développer. La transmission de ces maladies est souvent **complexe**, impliquant plusieurs gènes et des facteurs environnementaux **(11)**.

Par exemple, le **cancer du sein** présente un risque jusqu'à **50 fois plus élevé** dans certaines familles que dans d'autres. Cette différence s'explique probablement par la présence de nombreux allèles, chacun ayant un impact minime sur le risque global. Des gènes comme **BRCA1** et **BRCA2** ont été identifiés comme jouant un rôle clé, mais d'autres restent encore inconnus **(11)**.

Toutefois, bien que la prédisposition génétique soit un facteur important, des éléments **environnementaux** comme le **surpoids**, la **consommation excessive d'alcool** et le **manque d'exercice** peuvent également accroître le risque. Ainsi, le développement d'un cancer du sein résulte d'une interaction complexe entre la **génétique** et **l'environnement**, ce qui rend sa prévision particulièrement difficile. **(11)**.

III. Les Mutations Génétiques

III.1. Définition de mutation

Les mutations sont des modifications de la séquence d'ADN qui peuvent survenir n'importe où dans le génome. Si une mutation a lieu en dehors d'une séquence codante, dans l'ADN non codant, elle a généralement peu ou pas d'effet sur le phénotype, car elle ne modifie pas directement un gène codant une protéine. On parle alors de mutation neutre. En revanche, lorsqu'une mutation affecte un gène, elle peut impacter le phénotype de l'organisme **(14)**.

Les mutations peuvent se produire au niveau des exons, qui sont les régions codantes de l'ADN responsables de la production des protéines. Les protéines, essentielles au fonctionnement de l'organisme, ne représentent qu'environ 1,5 % du génome humain **(14)**.

Il existe 3 grandes classes de mutations :

-Génomiques : portent sur une altération du nombre total des chromosomes. :

Chromosomiques : portent sur une altération de la structure d'un ou de plusieurs chromosomes.

-Géniques : altération d'une séquence d'ADN transmissible ou de gène **(14)**.

III.2. Types de mutations génétiques

Il existe plusieurs types de mutations affectant l'ADN :

III.2.1. Mutations ponctuelles

Les représentent la principale cause des maladies génétiques.

Ces mutations se manifestent sous forme d'additions, de substitutions ou de suppressions de bases. Elles peuvent avoir une origine exogène, qu'elle soit physique, comme les rayons X et les rayons UV, ou chimique, via l'exposition à des agents mutagènes. Elles peuvent également être d'origine endogène, notamment à la suite d'erreurs survenant lors de la réplication ou de la réparation de l'ADN **(14)**.

Parmi les substitutions de bases d'origine endogène, on distingue deux types :

Les transitions : remplacement d'une base purique par une autre base purique, ou d'une base pyrimidique par une autre base pyrimidique **(14)**.

Les transversions : remplacement d'une base purique par une base pyrimidique, ou inversement **(14)**.

III.2.2. Délétions et duplications

Les délétions résultent de l'excision d'un segment d'ADN, suivie du rétablissement de la continuité de la double hélice. La taille du fragment supprimé peut varier d'une seule base à plusieurs millions de bases, pouvant même concerner un chromosome entier.

Les duplications, quant à elles, correspondent à la répétition d'un fragment d'ADN de taille variable **(15)**.

Ces deux types d'altérations génétiques — délétions et duplications — sont généralement induites par des recombinaisons dites illégitimes, causées par un mauvais alignement des séquences au cours de la méiose, conduisant à un crossing-over inégal. Ce phénomène se produit entre des séquences similaires mais non identiques **(15)**.

III.2.3. Conversion génique

C'est un transfert non réciproque d'une information de séquence. -L'échange peut se faire entre une paire de séquence non allélique, dans ce cas on parle de conversion génique interlocus c'est à dire entre deux gènes différents occupant des loci différents. -Ou entre une paire de séquence allélique, dans ce cas on parle de conversion génique interallélique entre les deux allèles du même gène. Dans cette situation, la séquence donneuse n'est pas altérée, elle reste inchangée, mais la séquence acceptrice recevant une partie copiée de la séquence donneuse sera modifiée **(15)**.

III.2.4. Fusion de gènes

C'est un réarrangement qui se produit lorsqu'une double cassure se produit entre deux gènes, il s'en suit une transposition d'origine dans l'autre gène, soit dans le même

chromosome ou dans deux chromosomes différents. La transposition intéressants deux chromosomes différents est appelée translocation **(15)**.

III.2.5. Insertions

Sont la conséquence de l'introduction d'une séquence qui peut être un transposon ou une séquence virale dans un gène. Transposon : séquence d'ADN capable de changer de localisation dans le génome exemple : les séquences ALU, séquences LINE **(15)**.

III.2.6. Inversions

Ce sont des changements tête bêche de segments plus ou moins long d'ADN exemple : L'inversion du gène du facteur IIX est à l'origine de 50 % des cas d'hémophile **(15)**.

III.2.7. Mutations perturbant l'épissage

Il existe une zone de consensus a niveau des jonctions exons introns qui peut être sujette à des mutations. Les mutations de cette zone donnent un saut d'exon ou une absence d'épissage si la mutation siège au niveau du site A de branchement **(15)**.

III.2.8. Délétions et insertions de petite taille

La délétion ou l'insertion d'une base ou d'un nombre qui n'est pas multiple de trois bases entraîne un décalage du cadre de lecture de la protéine (Frame Shift) ce qui provoque l'apparition d'un codon STOP prématurément en aval. La délétion ou l'insertion d'un nombre multiple de trois bases n'entraîne pas un décalage du cadre de lecture de la protéine. **(15)**.

III.2.9. Mutations dynamiques ou instables

A l'état normale, notre génome est riche en séquences répétées de type microsatellites. Elles sont stables lorsqu'elles sont transmises de génération en génération. **(15)**.

Les microsatellites, également appelés Short Tandem Repeats (STRs), sont de courtes séquences d'ADN composées de 1 à 6 nucléotides, répétées en tandem un certain nombre de fois (ex. : (CA)_n, (GATA)_n).

Ces séquences sont dispersées dans tout le génome et représentent une source importante de variabilité génétique entre les individus.

Chez un individu sain, elles sont généralement stables au cours des divisions cellulaires et lors de la transmission héréditaire. Cette stabilité est assurée par les mécanismes de réparation de l'ADN, qui corrigent les erreurs de réplication **(15)**.

Exemple : Maladie de Huntington. Une expansion du triplet CAG à l'origine d'une expansion d'un acide aminé qui est répété de façon anormale.

III.3. Impact des mutations sur les caractéristiques héréditaires

Dans certains cas, une seule mutation dans un gène suffit à déterminer une caractéristique héréditaire. Dans d'autres cas, plusieurs gènes sont impliqués dans une même caractéristique.

Les gènes ne déterminent pas uniquement l'apparence physique, mais jouent aussi un rôle fondamental dans le fonctionnement des cellules. En effet, ils contiennent les instructions nécessaires à la production des protéines, qui sont essentielles à la structure et aux fonctions biologiques de l'organisme (17).

Chapitre II. Le Séquençage de l'ADN

Le Séquençage de l'ADN

Le séquençage de l'ADN représente une avancée révolutionnaire en génomique, permettant d'analyser avec précision la succession des nucléotides dans une séquence génétique. Cette technologie a considérablement transformé la recherche scientifique et médicale, en facilitant la compréhension des génomes et l'identification de mutations associées à diverses maladies.

Différentes méthodes de séquençage ont été développées au fil du temps, chacune avec ses spécificités et niveaux de précision. On distingue notamment les techniques classiques, qui ont posé les bases du séquençage génétique, et les technologies modernes à haut débit, qui offrent une analyse plus rapide et à grande échelle (18).

I. Les méthodes de Séquençage ADN

I.1. Le Séquençage de Sanger : Une Méthode Fondamentale

Le séquençage de Sanger, également appelé "méthode de terminaison de chaîne", a été mis au point en 1977 par Frederick Sanger, scientifique distingué par deux prix Nobel. Cette approche a marqué une avancée majeure dans la cartographie des génomes, posant les fondements des techniques de séquençage actuelles (19).

Le séquençage Sanger peut être réalisé manuellement ou, plus couramment, de façon automatisée à l'aide d'un séquenceur (Figure N°05). Il repose sur trois étapes clés, qui seront détaillées dans les sections suivantes. (18).

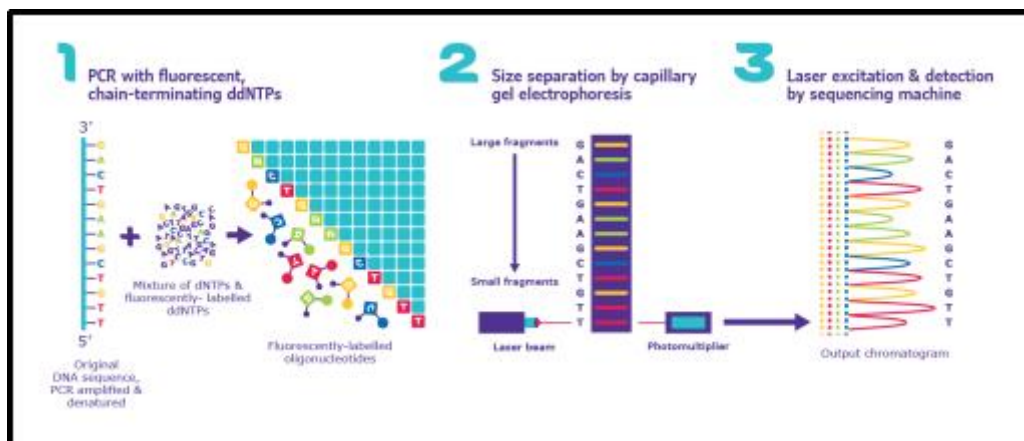


Figure N° 4 : Il repose sur trois étapes clés Le séquençage Sanger (43).

I.1.1. Principes du Séquençage

Le séquençage de Sanger, ou méthode de terminaison de chaîne, est une technique permettant de déterminer l'ordre des nucléotides dans une molécule d'ADN. Il repose sur

l'incorporation de didésoxynucléotides (ddNTP) marqués, qui stoppent la synthèse de l'ADN à des positions spécifiques.

Les fragments ainsi générés sont ensuite séparés par électrophorèse en gel ou par capillaire, permettant de reconstituer la séquence des bases grâce à l'analyse des signaux fluorescents ou radioactifs. Cette méthode, bien que plus lente que les technologies modernes, reste une référence pour le séquençage de courtes séquences avec une grande précision. LODE, Laurence. Etude moléculaire de l'évolution clonale de TP53 des Syndromes Myélodysplasiques avec del (5q): conséquences sur la résistance au traitement et la progression du cancer (19).

I.2. Le séquençage Maxam-Gilbert

Le séquençage Maxam-Gilbert, aussi appelé méthode de séquençage chimique, est une technique développée par Walter Gilbert et Alan Maxam en 1976. Faisant partie de la première génération de méthodes de séquençage de l'ADN, elle fut l'une des premières techniques largement utilisées par les scientifiques, notamment en raison de sa capacité à être appliquée directement sur de l'ADN purifié (20).

I.2.1. Principe de la méthode

Cette technique repose sur la clivage chimique spécifique des bases de l'ADN. Les fragments d'ADN sont marqués à une extrémité, puis exposés à des traitements chimiques qui induisent des coupures à des bases spécifiques (adénine, guanine, cytosine ou thymine).

Les fragments obtenus sont ensuite séparés par électrophorèse en fonction de leur taille. L'analyse du schéma de migration permet alors de reconstituer la séquence de l'ADN. Bien que cette méthode ait été largement remplacée par le séquençage de Sanger et les technologies de nouvelle génération (NGS), elle reste un jalon historique dans le développement du séquençage de l'ADN (20).

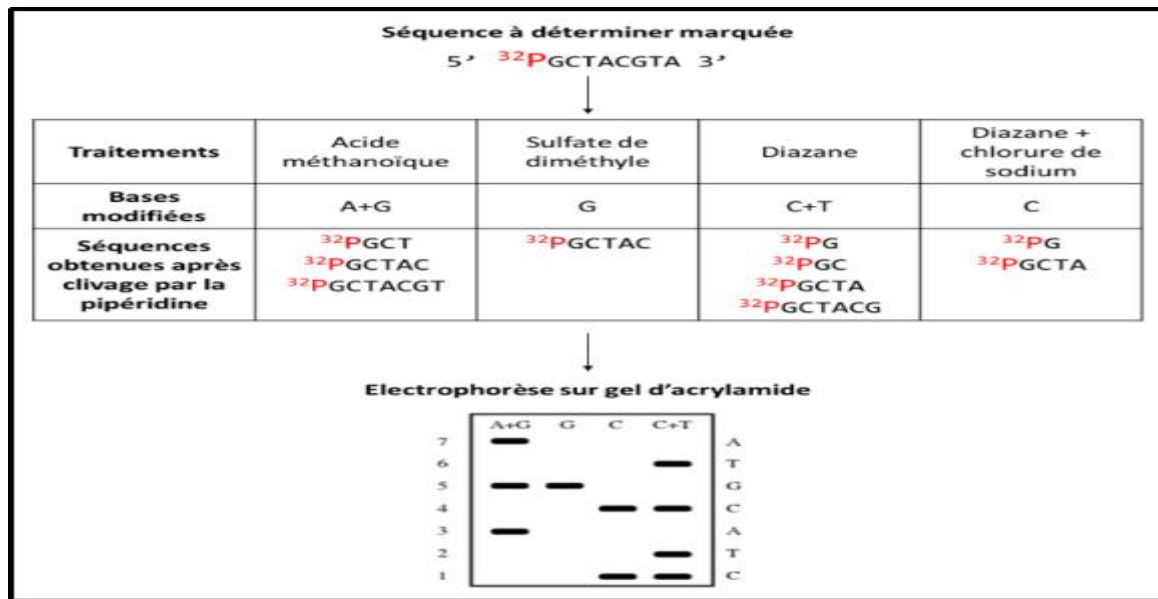


Figure N° 5 : Séquence a déterminer marquée (44).

II. Les séquenceurs à haut débit

II.1. Séquençage par hybridation

Le séquençage par hybridation repose sur l'utilisation de puces à ADN, contenant des centaines à plusieurs milliers d'oligonucléotides, selon la génération de la puce. L'ADN à analyser est d'abord fragmenté, puis incubé sur la puce où il s'hybride aux oligonucléotides complémentaires.

La lecture de la puce, qui consiste à détecter les oligonucléotides hybridés, permet d'obtenir un spectre de séquence, représentant la composition de l'ADN en sous-séquences de taille définie. Un traitement informatique de ces données permet ensuite de reconstituer la séquence complète (21).

II.1.1. Principe de l'hybridation de l'ADN

L'hybridation repose sur le principe d'appariement des bases complémentaires des acides nucléiques. Cette propriété est exploitée dans diverses techniques de biologie moléculaire, notamment pour détecter la présence de séquences spécifiques d'ADN ou d'ARN. C'est également le principe fondamental utilisé dans des méthodes comme le Northern blot ou le Southern blot pour la visualisation des acides nucléiques (21).

II.2. Le séquençage Illumina

Le séquençage Illumina, ou séquençage par synthèse (SBS), est une technologie de séquençage de nouvelle génération (NGS) qui permet l'analyse parallèle de millions de fragments d'ADN. Ce procédé repose sur quatre étapes principales :

1. Préparation de la bibliothèque : fragmentation de l'ADN et ligation d'adaptateurs spécifiques aux extrémités.
2. Génération de grappes (cluster generation) : amplification isotherme des fragments sur une surface solide (flow cell) par amplification en pont (bridge amplification).
3. Séquençage : incorporation de nucléotides fluorescents réversibles permettant la lecture base par base.
4. Analyse bioinformatique : traitement des données brutes pour reconstruire la séquence complète **(23)**.

Cette technologie est largement utilisée en génomique pour sa précision, sa rapidité et sa capacité à produire un volume important de données à faible coût par base **(22)**.

II.2.1. Principe du séquençage Illumina

Le séquençage Illumina repose sur le même principe fondamental que le séquençage Sanger, à savoir l'incorporation de nucléotides marqués par fluorescence. Toutefois, à la différence de la méthode Sanger, qui ne peut traiter qu'un fragment à la fois, la technologie Illumina est massivement parallèle, permettant le séquençage simultané de millions de fragments en une seule analyse **(23)**.

Illumina propose une gamme de plateformes de séquençage adaptées aux besoins spécifiques:

- MiSeq, MiniSeq : compacts, adaptés au séquençage ciblé ou aux petits projets.
- NextSeq, NovaSeq : conçus pour des volumes de données importants, idéals pour les études de grande envergure et les projets cliniques **(24)**.

II.3. Le séquençage Nanopore

Le séquençage Nanopore est une technologie innovante et évolutive permettant l'analyse directe et en temps réel de longs fragments d'ADN ou d'ARN. Son principe repose sur la détection des variations de courant électrique lorsqu'un acide nucléique traverse un nanopore protéique. Le signal généré est ensuite interprété pour reconstituer la séquence spécifique de l'ADN ou de l'ARN **(25)**.

II.3.1. Principe

Le séquençage par nanopores repose sur la détection des variations de courant électrique lorsqu'un brin d'ADN ou d'ARN traverse un nanopore immergé dans une solution conductrice sous l'effet d'une tension appliquée. Ce processus permet le passage des nucléotides un par un à travers le nanopore, à l'image d'un fil passant dans le chas d'une aiguille, facilitant ainsi la lecture de la séquence en temps réel (25).

II.4. Séquençage à courtes lectures par DNB

La technologie développée par MGI repose sur le séquençage par synthèse (SBS) de pelotes d'ADN simple brin (DNB - DNA NanoBalls). La préparation des bibliothèques implique une circularisation suivie d'une réplication en cercle roulant (rolling circle replication) de fragments d'ADN ou d'ADNc préalablement indexés. Ce processus génère des DNBs contenant 300 à 500 copies de chaque fragment (Figure A). Cette approche réduit les biais d'amplification et permet d'obtenir des taux de duplication et d'erreur de séquençage très faibles (24).

II.4.1. Principe

Dans cette méthode, un dNTP modifié portant un groupement bloquant en 3' est intégré à chaque cycle. L'identification des nucléotides repose sur l'utilisation d'anticorps fluorescents, spécifiquement dirigés contre ces groupements bloquants. Une fois la reconnaissance effectuée, le complexe groupement bloquant-anticorps est clivé, permettant ainsi d'initier un nouveau cycle de séquençage (23).

Le procédé HotMPS suit le protocole standard, mais utilise des dNTP-H, qui portent des chromophores bloquants en 3'. Après incorporation de la base, une enzyme-H clive le chromophore, libérant l'extrémité 3' pour la prochaine étape d'incorporation.

La lecture du brin complémentaire (paired-end) s'effectue via le système MDA (Multiple Displacement Amplification - Dean FB, 2002), garantissant une amplification efficace et une meilleure précision du séquençage (24).

III. Analyse des Exons et Séquençage de l'Exome

Le séquençage de l'exome entier (Whole Exome Sequencing - WES) est une technique qui consiste à analyser uniquement les régions codantes du génome, c'est-à-dire les gènes qui produisent des protéines. Ce processus se déroule en deux étapes principales :

Ciblage des séquences exoniques : Cette phase vise à isoler uniquement l'exome humain. En génétique humaine, ces régions ciblées représentent environ 60 millions de paires de bases, soit environ 1 % du génome humain.

Séquençage et analyse : L'ADN est ensuite séquençé à l'aide de plateformes de séquençage à haut débit, permettant d'identifier des mutations ou variations génétiques. (26).

III.1. Importance et limites du séquençage de l'exome

Le séquençage de l'exome permet de détecter des variations génétiques affectant la séquence des protéines, pouvant entraîner des maladies telles que l'athérosclérose, la maladie d'Alzheimer et bien d'autres pathologies. Puisque de nombreuses maladies résultent de mutations dans les régions codantes du génome, l'étude de l'exome offre un aperçu approfondi des prédispositions génétiques à certaines maladies.

Cependant, il est maintenant reconnu que des variations situées en dehors des exons peuvent également influencer l'expression des gènes et la production de protéines, contribuant ainsi à des troubles génétiques. Le séquençage de l'exome ne permet pas de détecter ces variations, ce qui constitue sa principale limitation (27).

III.2. Séquençage de l'exome vs. Séquençage du génome entier

Historiquement, l'avantage du séquençage de l'exome était son coût plus faible et son efficacité temporelle par rapport au séquençage du génome entier. Il permettait d'analyser un grand nombre de gènes en un temps réduit et à moindre coût (28).

Cependant, les progrès technologiques ont rendu le séquençage du génome entier (Whole Genome Sequencing - WGS) plus accessible et abordable, offrant une analyse plus complète de l'ADN. Contrairement au séquençage de l'exome qui ne couvre que 1 % du génome, le séquençage du génome entier permet d'identifier 100 % des variations génétiques, y compris celles situées dans les régions non codantes, qui peuvent également influencer l'apparition de maladies (29).

III.3. Analyse des Données de Séquençage

Même lorsqu'il ne concerne que 1 % du génome humain, le séquençage de l'exome génère un volume important de données nécessitant des analyses complexes. Pour traiter ces données, plusieurs logiciels spécialisés permettent d'aligner et d'assembler les séquences obtenues (28).

IV. Méthodes pour améliorer la qualité des données

Afin d'optimiser l'interprétation des résultats et garantir leur fiabilité, différentes stratégies sont utilisées :

- Comparaison des variantes génétiques identifiées par séquençage avec celles obtenues via génotypage par matrice, lorsque cela est possible.
- Analyse des SNP (Single Nucleotide Polymorphisms) codants, en les confrontant aux données de personnes atteintes de la maladie étudiée et dont le génome entier a été séquencé.
- Validation des SNP codants en les comparant avec les résultats du séquençage Sanger des individus référencés dans la base HapMap **(29)**.

IV.1. Identification et stockage des variants génétiques

Les outils d'analyse génomique permettent d'identifier les mutations présentes dans les données de séquençage et de générer différents types de fichiers essentiels pour l'interprétation des résultats :

- Fichiers FASTQ : contiennent les séquences brutes issues du séquençage.
- Fichiers BAM/CRAM : formats compressés pour l'alignement des séquences sur un génome de référence.
- Fichiers VCF (Variant Call Format) : listent les variantes génétiques détectées.

Grâce à ces analyses et au traitement approfondi des données, le séquençage de l'exome permet d'identifier des mutations pouvant être liées à diverses maladies génétique **(29)**.

***Chapitre III. Maladie
héréditaire***

I. Maladie héréditaire

Il existe près de 6000 maladies génétiques différentes dans le monde. Elles se déclenchent par la présence d'un ou plusieurs gènes ou chromosomes défectueux.

Une maladie génétique résultante génétique résultant d'une ou plusieurs anomalies affectant un ou plusieurs chromosomes, entraînant un dysfonctionnement de certaines cellules de l'organisme. Ces cellules produisent des protéines dont la structure et l'activité sont déterminées par l'information génétique contenue dans les gènes. Lorsqu'un gène est altéré, il peut perturber le fonctionnement cellulaire et provoquer l'apparition d'une maladie à tout âge de la vie **(30)**.

Lorsqu'une anomalie concerne un chromosome, on parle d'anomalie chromosomique. Cela peut se traduire par la présence d'un chromosome ou d'un fragment supplémentaire sur une paire, comme dans le cas de la trisomie, ou par l'absence d'un chromosome, appelée monosomie **(31)**.

Lorsqu'une anomalie affecte un gène de l'ADN, on parle de mutation. Ces mutations perturbent le fonctionnement de l'organisme et peuvent être monogéniques (impliquant un seul gène) ou polygéniques (concernant plusieurs gènes). Si le gène muté est dominant, il suffit d'une seule copie altérée pour déclencher la maladie, dite dominante. En revanche, si le gène est récessif, la maladie ne se manifeste que si l'enfant hérite de la mutation des deux parents, on parle alors de maladie récessive.

Les maladies génétiques ne sont pas contagieuses : elles ne se transmettent ni par l'air ni par contact **(31)**.

Les maladies génétiques peuvent se manifester à tout âge. Comme le souligne l'Organisation mondiale de la santé, leur gravité varie : certaines peuvent être mortelles avant la naissance, tandis que d'autres nécessitent une prise en charge permanente. Elles peuvent apparaître dès la naissance ou durant la petite enfance, comme la mucoviscidose ou la myopathie de Duchenne, mais aussi se déclarer à l'âge adulte, comme la chorée de Huntington **(30)**.

Les maladies génétiques sont qualifiées de dominantes ou récessives, selon que l'allèle impliquée est dominant ou non. Chaque individu possède deux allèles par gène, et la transmission de la maladie dépend de cette caractéristique **(30)**.

Les maladies génétiques peuvent également être classées selon la localisation du gène impliqué. Si l'anomalie se trouve sur une paire de chromosomes sexuels, on parle de maladie gonosome. En revanche, si elle affecte une paire de chromosomes non sexuels (chromosomes homologues), elle est qualifiée d'autosomale. Ainsi, on distingue les maladies autosomales récessives, comme la phénylcétonurie, et les maladies gonosomales récessives, telles que l'hémophilie.

Les maladies génétiques peuvent être aussi bien bénignes ou peu invalidantes, comme le daltonisme, que particulièrement sévères. Leur point commun est qu'elles sont généralement présentes tout au long de la vie et, dans certains cas, transmissibles à la descendance, car inscrites dans le patrimoine génétique de l'individu.

Les maladies rares, définies par une prévalence inférieure à un cas pour 2 000 personnes, représentent environ 80 % des maladies génétiques, soit près de 6 000 pathologies.

Toutes les maladies génétiques ne sont pas nécessairement héréditaires, car une mutation peut survivre spontanément dans un gamète sans être présente dans le génome du parent. En revanche, toute maladie héréditaire repose essentiellement sur une altération génétique (30).

II. Types de maladies héréditaires

II.1. Maladies monogéniques (mutation d'un seul gène)

- Mucoviscidose
- Drépanocytose
- Myopathie de Duchenne
- Hémophilie

II.2. Maladies chromosomiques (anomalies dans les chromosomes)

- Trisomie 21 (syndrome de Down)
- Syndrome de Turner
- Syndrome de Klinefelter

II.3. Maladies multifactorielles (combinaison de gènes et de facteurs environnementaux)

- Diabète de type 2
- Hypertension
- Certains cancers

III. Transmission des maladies héréditaires

- Autosomique dominante : Un seul parent porteur suffit pour transmettre la maladie (ex. Maladie de Huntington).
- Autosomique récessive : Les deux parents doivent être porteurs du gène muté (ex. Mucoviscidose).
- Liée au chromosome X : Transmise via le chromosome sexuel X, souvent plus fréquente chez les hommes (ex. Hémophilie) (34).

IV. Prévalence

Chaque année, au moins 7,6 millions d'enfants naissent dans le monde avec une malformation génétique ou congénitale sévère. Dans les pays développés, ces maladies génétiques et anomalies congénitales représentent la deuxième cause de mortalité chez les nourrissons et les jeunes enfants, avec une prévalence à la naissance pouvant atteindre 25 à 60 pour 1 000 naissances (35).

V. Dépistage et diagnostic

Face à certains symptômes, un médecin peut prescrire une analyse de l'ADN et des chromosomes afin de détecter une anomalie génétique. Ce test, réalisé à partir d'une prise de sang dans un laboratoire spécialisé, s'inscrit dans un réseau dédié au dépistage des maladies génétiques. Selon l'Agence de la Biomédecine, plus de 1 000 maladies génétiques peuvent être diagnostiquées en France (35).

En cas de diagnostic un accompagnement médical et psychologique est essentiel. Le patient bénéficie d'un suivi par divers professionnels de santé, tels que des kinésithérapeutes, psychologues, neurologues et généticiens cliniciens (35).

VI. Qui sont les personnes à risque ?

Selon l'Organisation mondiale de la santé (OMS), toute personne peut être concernée par une maladie liée à une mutation génétique. Cependant, certains facteurs sociaux et culturels, comme les mariages consanguins, augmentent le risque de transmission. Par ailleurs, le risque d'anomalies chromosomiques chez l'enfant est plus élevé lorsque la mère a plus de 35 ans (35).

VII. Quels traitements ?

Certaines maladies génétiques fréquentes, comme la thalassémie, la mucoviscidose, l'hémophilie ou la phénylcétonurie, peuvent être prises en charge de manière efficace. Un traitement adapté permet non seulement d'améliorer la qualité de vie des patients, mais aussi d'augmenter leur espérance de vie (35).

VIII. Qu'est-ce que l'analyse de l'Exome?

VIII.1. Types d'analyses génétiques

Le génome humain est composé d'environ 20 000 gènes. Ces gènes contiennent l'information nécessaire à la synthèse des protéines, essentielles au bon fonctionnement de l'organisme. Cette information est portée par des segments appelés exons, séparés par des régions non codantes appelées introns. L'ensemble des exons du génome constitue ce que l'on nomme l'exome (30).

Lorsqu'une modification de l'ADN survient dans les exons, elle peut entraîner une maladie. Ce changement est appelé mutation. Pour établir un diagnostic précis d'une maladie d'origine génétique, il est essentiel d'identifier la mutation en cause.

Jusqu'à récemment, cette identification était un processus complexe, long et coûteux. La technologie disponible ne permettait d'analyser que de petits fragments d'ADN un par un, ce qui pouvait prendre plusieurs mois et représenter un coût élevé. La difficulté était encore plus grande pour les maladies impliquant des mutations sur plusieurs gènes ou dont l'origine génétique était inconnue (31).

Aujourd'hui, grâce à la technologie de séquençage massif (NGS), il est possible d'analyser simultanément un grand nombre de gènes ou de régions spécifiques, comme les exons. Cette avancée permet d'obtenir des résultats en un temps réduit et à un coût bien plus abordable, facilitant ainsi l'identification des mutations responsables de nombreuses maladies (31).

VIII.2. Types d'Exomes

L'analyse de l'exome permet d'identifier une mutation génétique selon différentes approches:

- **Exome clinique ciblé** : cette analyse se concentre sur les exons d'un ensemble de gènes connus, susceptibles d'expliquer les symptômes du patient. Elle est recommandée lorsqu'une maladie peut être causée par plusieurs gènes différents, comme

l'ostéogénèse imparfaite, la maladie de Charcot-Marie-Tooth ou certaines formes de surdité (36).

- **Exome clinique complet** : En général, un WES classique complet couvre environ **19 000 à 22 000 gènes** codants. Cela correspond à environ **180 000 à 220 000 exons**, selon la plateforme et le kit de capture utilisés. Elle est indiquée lorsque les symptômes sont non spécifiques ou lorsqu'il est difficile de sélectionner un groupe précis de gènes, comme dans le cas de l'autisme, de l'épilepsie, du handicap intellectuel ou des malformations congénitales multiples.

Le séquençage de l'exome est particulièrement utile dans les cas suivants :

- L'analyse de gènes de grande taille, difficiles à étudier avec les méthodes classiques.
- L'identification des causes génétiques de maladies complexes impliquant plusieurs gènes.
- La recherche de mutations dans des maladies peu étudiées, dont l'origine génétique est inconnue, ou présentant des symptômes non spécifiques (39).

En plus des maladies, certaines conditions pathologiques peuvent également bénéficier de ces analyses, notamment celles liées à la fertilité, aussi bien chez la femme que chez l'homme. C'est le cas de l'échec ovarien précoce (FOP), qui se caractérise par l'arrêt prématuré des menstruations avant l'âge de 40 ans. Dans la majorité des cas, sa cause reste inconnue. Toutefois, grâce aux avancées du séquençage génétique, il est désormais possible d'identifier de nombreux gènes potentiellement impliqués dans cette condition.

Ces analyses génèrent une quantité importante de données qui doivent être comparées à de nombreuses bases de données et publications scientifiques afin d'obtenir un résultat fiable. Dans de nombreux cas, l'identification de la mutation responsable nécessite l'étude de plusieurs membres d'une même famille affectée (par exemple, les deux parents) ou de plusieurs personnes atteintes de la même maladie. Ainsi, l'analyse de l'exome est une démarche complexe qui exige l'expertise d'un laboratoire spécialisé et d'une équipe qualifiée (33).

Partie II : Matériel et méthodes

Matériel et Méthodes

Instruments :

- Séquenceur d'ADN NGS : DNB seq G99, MGI
- Spectrophotomètre à fluorescence : QUBIT4, Thermofisher scientific
- Thermocycleur : NEOS 48 q PCR, LINEAR
- Sonicateur : M220 Focused-ultrasonicator, COVARIS
- PSM 080 : TMLab
- Rack magnétique : MAgSorb 16, BIOSAN

I. Méthodes

I.1. Echantillon biologique

3-5 mL de sang total de patient sont prélevés dans un tube hépariné. Le sang est centrifugé pendant 10 min à 1000g. 200µL de la couche correspondant aux cellules lymphocytaires est récupérée dans un tube eppendorf.

I.2. Extraction et purification de l'ADN

L'extraction est réalisée à partir de 200µL d'échantillon biologique humain.

Lyse cellulaire : Les cellules sont lysées à l'aide de tampons contenant des agents chaotropiques et des détergents (ex. : guanidinium thiocyanate + Triton X-100).

Ajout des billes magnétiques : Les billes sont ajoutées au lysat. L'ADN s'y fixe en présence du sel chaotrope.

Séparation magnétique : L'échantillon est placé à proximité d'un aimant. Les billes (avec l'ADN lié) sont attirées contre la paroi du tube. Le surnageant est éliminé.

Lavages : Les billes sont lavées plusieurs fois avec des tampons d'éthanol pour éliminer les impuretés (protéines, sels, détergent).

Élution de l'ADN : L'ADN est ensuite libéré des billes en utilisant un tampon faiblement salin ou de l'eau ultra-pure, les billes sont retirées à l'aide de l'aimant.

I.3. Fragmentation de l'ADN

Le sonicateur Covaris M 220 est utilisé pour fragmenter un volume de 55 µL d'échantillon d'ADN génomique. Les fragments produits ont une taille comprise entre 300 et 500 pb.

I.3.1. Sélection des Fragments d'ADN en fonction de leurs tailles

La sélection de la taille des fragments d'ADN est réalisée par billes magnétiques et ce en jouant sur la proportion des billes par rapport au volume de l'échantillon d'ADN. Deux sélections successives permettent d'isoler les fragments à taille souhaitées. **Le tableau N°01** montre la taille des fragments isolés à partir de 100µL d'ADN fragmenté en fonction de la quantité de billes magnétiques :

Tableau N° 1 : la taille des fragments isolés à partir de 100µL d'ADN fragmenté en fonction de la quantité de billes magnétiques

Taille cible (ADN pb)	180	230	280	335	420	550
1 ^{ère} selections par billes mag	100	90	80	70	60	50
2 ^{ème} selections par billes mag	50	20	20	20	20	20

Dans notre cas, nous avons ciblé des fragments de 420 pb ce qui correspond à un mix de 100µL d'échantillons + 60 µL de billes pour la première sélection, et 100µL d'échantillon + 20 µL de billes pour la seconde sélection.

I.3.2. Etapes de la sélection des fragments

Transférer l'ADN fragmenté dans un nouveau tube 1.5mL puis centrifuger brièvement. Compléter jusqu'à 100µL avec du tampon TE.

Ajouter 60 µL de billes magnétiques dans le tube à échantillon. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber à pendant 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique jusqu'à ce que le liquide soit clair. Transférer ensuite délicatement le surnageant dans un nouveau tube à centrifuger de 1,5 ml. Conserver le surnageant et jeter les billes.

Ajouter des billes d'ADN propres au tube à centrifuger contenant le surnageant. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension. Incuber à température ambiante pour 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique jusqu'à ce que le liquide soit clair. Retirer et éliminer soigneusement le surnageant.

Tout en maintenant le tube à centrifuger sur le support magnétique, ajoutez de l'éthanol à 80% dans chaque tube pour laver les billes et la paroi du tube. Attendez. Retirez et jetez soigneusement le surnageant.

Répétez l'étape et essayez de retirer tout le liquide du tube.

Conservez le tube sur le support magnétique. Ouvrez le bouchon du tube et laissez sécher les billes à l'air libre jusqu'à ce qu'elles ne soient plus humides ni brillantes. Aucune fissure ne doit apparaître à la surface des billes.

Retirez le tube du support magnétique et ajoutez 32 μL de TE pour éluer l'ADN. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber le tube température ambiante pour 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique pendant 2 à 5 minutes jusqu'à ce que le liquide soit clair. Transférer délicatement. 30 μL du surnageant dans un nouveau tube à centrifuger de 1,5 ml.

Quantification de l'ADN

La Quantification est réalisée avec le spectro-fluoromètre Qubit.

I.4. Réparation de l'extrémité des fragments et ajout de la queue « A »

Transférez l'échantillon d'ADN dans un nouveau tube PCR de 0,2 ml et ajoutez du tampon TE pour un volume total de 39,5 μL .

Préparez le mélange de réparation d'extrémité et de queue A sur la glace :

Tableau N° 2 : mélange pour la réparation d'extrémité et de queue A

	Composants	volume
1	ERAT Buffer	6 μL
2	Mélange d'enzyme ERAT	4,5 μL
	Total	10,5 μL

Transfert 10,5 μ L du mélange dans le tube PCR de 0,2 ml de l'étape précédente. Vortexer 3 fois et centrifuger brièvement.

Placez le tube PCR de 0,2 ml dans le thermocycleur et exécutez le programme.

	Température	Temps
1	Couvercle chauffant (105°C)	Oui
2	4°C	30 minutes
3	65°C	15minutes
4	20 °C	Pause

Centrifuger brièvement pour recueillir la solution au fond du tube.

I.5. Ligature de l'adaptateur

Ajouter 5 μ L de l'adaptateur UDB au tube d'échantillon. Mélangez 3 fois au vortex, centrifuger brièvement et placer sur la glace.

Préparez le mélange de ligature de l'adaptateur. Agiter 6 fois au vortex puis centrifuger brièvement et placer sur la glace.

Tableau N° 3 : Le mélange de ligature de l'adaptateur

	Composant	Volume
1	ADN ligase	4U1
2	Tampon de ligature	21uL
	Total	25uL

I.5.1. Mélange de ligature d'adaptateur

Pipeter lentement 25 μ L du mélange de ligature de l'adaptateur dans le tube d'échantillon. Vortexer 6 fois, centrifuger brièvement et placer sur la glace.

Le mélange de ligature de l'adaptateur est très visqueux. Pipeter lentement et avec précaution.

Placez le tube PCR dans le thermocycleur. Exécutez le programme dans les conditions suivantes.

	Température	Temps
1	Couvercle chauffant (50 C°)	Oui
2	23 C°	30 minutes
3	4 C°	Pause

I.5.2. Programme de ligature d'adaptateur

Une fois le programme terminé, centrifugez brièvement le tube PCR et placez-les sur la glace.

Ajouter 20 µL Tampon TE, pour un volume total de 100 µL et transférer dans un nouveau tube à centrifuger de 1,5 ml.

I.5.3. Nettoyage de l'ADN ligaturé par un adaptateur

Ajouter 50 µL de billes magnétiques propres dans chaque tube à échantillon. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber le tube à échantillon à température ambiante pendant 5 minutes

Centrifuger brièvement le tube à échantillon et le placer sur le support magnétique pendant 2 à 5 minutes jusqu'à ce que le liquide soit clair. Retirer et éliminer soigneusement le surnageant.

Gardez le tube sur le support magnétique, ajoutez 200 µL d'éthanol à 80 % dans le tube pour laver les billes et la paroi du tube. Attendre 30 secondes puis retirez et jetez soigneusement le surnageant.

Répétez l'étape de lavage une fois. Essayez d'éliminer tout le liquide du tube. S'il reste du liquide sur la paroi du tube, centrifugez-le brièvement et placez-le sur le support magnétique pour séparation. Éliminez tout le liquide à l'aide d'une pipette à faible volume.

Laissez le tube sur le support magnétique. Ouvrez le bouchon du tube et laissez sécher les billes à l'air libre. Température ambiante jusqu'à ce qu'il n'y ait plus d'humidité ni de brillance à la surface des perles. Aucune fissure ne doit apparaître à la surface des perles.

Retirez le tube du support magnétique et ajoutez 40 μL de tampon TE pour éluer l'ADN. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber le tube à température ambiante pour 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique pendant 2 à 5 minutes jusqu'à ce que le liquide soit clair. Transférer délicatement 38 μL du surnageant dans un nouveau tube PCR de 0,2 mL.

I.6. Amplification par PCR

Réaliser le mélange selon le tableau suivant :

Tableau N° 4 : le mélange Amplification par PCR

	Composants	Volume
1	produit de ligation	38 μl
2	Mélange d'enzymes PCR	50 μl
3	Mélange d'amorces PCR UDB	12 μL
	Total	100 μL

Placez le tube PCR dans le thermocycleur. Exécutez le programme dans les conditions suivantes :

	Température	Temps	Cycles
1	Couvercle chauffant 105 c°	Oui	
2	95 c°	3 minutes	1
3	98 c°	20 sec	x20
4	60 c°	15 sec	
5	72 c°	30 sec	
6	72 c°	10 minutes	1
7	4 c°	Pause	

Une fois le programme terminé, centrifugez brièvement le tube et transférez-le dans un nouveau tube à centrifuger de 1,5 ml.

I.6.1. Nettoyage du produit PCR

Ajouter 100 μ L de billes magnétiques dans chaque tube d'échantillon. Mélanger avec un vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber le tube à échantillon à Température ambiante pour 5 minutes.

Centrifuger brièvement le tube à échantillon et le placer sur le support magnétique pendant 2 à 5 minutes jusqu'à ce que le liquide soit clair. Retirer et éliminer soigneusement le surnageant.

Gardez le tube sur le support magnétique, ajoutez 200 μ L d'éthanol à 80 % dans chaque tube pour laver les billes et la paroi du tube. Attendre 30 secondes, puis retirez et jetez soigneusement le surnageant.

Répétez l'étape de lavage une fois. Essayez d'éliminer tout le liquide du tube. S'il reste du liquide sur la paroi du tube, centrifugez-le brièvement et placez-le sur le support magnétique pour séparation.

Éliminez tout le liquide à l'aide d'une pipette à faible volume.

Laissez le tube sur le support magnétique. Ouvrez le bouchon du tube et laissez sécher les billes à l'air libre à température ambiante jusqu'à ce qu'il n'y ait plus d'humidité ni de brillance à la surface des billes. Aucune fissure ne doit apparaître à la surface des billes.

Retirez le tube du support magnétique et ajoutez 32 μ L de tampon TE pour éluer l'ADN. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber le tube à température ambiante pour 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique pendant 2 à 5 minutes jusqu'à ce que le liquide soit clair. Transférer délicatement. 30 μ L du surnageant dans un nouveau tube à centrifuger de 1,5 ml.

Quantifier l'ADN double brin avec le kit d'analyse Qubit dsDNA-HS. Le rendement des produits PCR doit être supérieur à 1 pmol.

I.7. Hybridation et capture

Cette étape permet de sélectionner (hybridation) et d'isoler (capture) les fragments d'ADN correspondant aux séquences exomiques.

Préparation du mélange d'hybridation :

Préparer le mélange suivant dans un tube à PCR de 0,2 mL :

Tableau N° 5 : Préparation du mélange d'hybridation

Components	Volume
Hyb Buffer 1	10 uL
Hyb Buffer	0,4 uL
Hyb Buffer	4 uL
Hyb Buffer	5,6 uL
Total	20 uL

Incuber le mélange à 65°C pendant 5 minutes.

Préparation du mélange de capture :

Préparer le mélange de capture suivant dans un nouveau tube 0.2 mL.

Tableau N° 6 : Préparation du mélange de capture

Components	Volume
NF water	1,5 uL
Blok 5	0,5 uL
MGI Exome VS Probe	5 uL
Total	7 uL

Placer le mélange dans un thermocycleur avec le programme suivant :

Temperature	Time
105 °C Heated lid	on
65 °C	2 min
65 °C	Hold

Les deux mélanges (Hybridation et capture) doivent être gardés à une température de 65°C

Transférer 20µL du mélange d'hybridation aux 7 µL du mélange de capture. Mélanger à l'aide d'une pipette.

Placer le tube dans le thermocycleur et lancer le programme suivant pendant 24 heures :

Tempurature	Time
105 °C Heated lid	On
65 °C	2 min
65 °C	Hold

I.8. Elution

Pour la réaction d'hybridation, ajouter 1,8mL de Wash Buffer II dans un nouveau tube de 2mL. Placer le mélange dans le thermo-mixeur réglé à 65°C. Puis ajouter 50µL de Billes magnétiques et mélanger l'ensemble.

Ajouter 200µL de Binding buffer au tube contenant le mélange. Placer le tube contre le rack magnétique pour 2-5 minutes et jeter le liquide transparent.

Ajouter 200µL de Binding buffer et resuspendre les billes.

Placer le mélange dans un agitateur rotatif et faire pivoter pendant 30 minutes à température ambiante.

Placer à nouveau les tubes contre un rack magnétique, laisser agir pendant 2-5 minutes et éliminer le surnageant.

Réaliser un lavage avec 500 µL de Wash Buffer.

Réaliser un second lavage avec le Wash Buffer II préchauffé à 65°C. Mélanger à l'aide de l'agitateur magnétique (1000rpm) pendant 10 secondes. Incuber à 65°C pendant 7 minutes.

Placer contre le rack magnétique et éliminer le surnageant.

Répéter cette étape une seconde fois.

Ajouter 44µL de NF water au tube et resuspendre le mélange et transférer le mélange (avec les billes magnétiques) dans un nouveau tube 0,2 mL.

I.9. PCR Post-Capture

Ajouter 65µL de mixture PCR au tube de l'étape précédente (44µL avec billes magnétiques), bien mélanger et lancer la PCR avec le programme suivant :

Tableau N° 7 : mélanger et lancer la PCR

Tempurature	Time	Cycles
105 °C Heated lid	On	—
95 °C	3 min	1
98 °C	20 sec	
60 °C	15 sec	x
72 °C	30 sec	
72 °C	10 min	1
4 °C	Hold	—

A l'issu de la PCR, placer le tube contre le rack magnétique et récupéré le surnageant dans une nouveau tube 1,5mL.

I.10. Dénaturation et circularisation simple brin

Le kit de circularisation MGIEasy (réf. : 1000020570) est utilisé pour la préparation des DNB.

Dénaturation :

En fonction de la concentration des produits PCR, ajoutez 300 ng des produits PCR dans un nouveau tube PCR de 0,2 ml. Si le volume est inférieur à 48 µL, ajoutez du tampon TE pour obtenir un volume total de 48 µL.

Placez le tube PCR dans le thermocycleur. Exécutez le programme dans les conditions suivantes :

	Température	Temps
1	Couvercle chauffant a 100 °C	Oui
2	95 °C	3 minutes
3	4 °C	10 minutes

Après la réaction, centrifuger brièvement le tube et placer sur la glace .

I.11. Circularisation monocaténaire

Préparez le mélange de réaction de circularisation dans un nouveau tube PCR de 0,2 mL sur la glace. Vortexez 3 fois puis centrifuger brièvement et placer sur la glace :

Tableau N° 8 : le mélange de réaction de circularisation

	Réactif	Volume
1	Tampon d'attelle a doubles code-barres	11,6 uL
2	Ligase rapide d'ADN	0,5 uL
	Total	12,1 µL

Ajouter 12,1 µL du mélange réactionnel de circularisation dans le tube à échantillon. Vortexer 3 fois puis centrifuger brièvement et placer sur la glace.

Placez le tube PCR dans le thermocycleur et exécutez le programme dans les conditions suivantes :

	Températures	Temps
1	Couvercle chauffant 42c°	Oui
2	37C°	10 Minutes
3	4C°	Pause

Une fois le programme terminé, placez le tube PCR sur la glace puis centrifuger brièvement et passer immédiatement à l'étape suivante.

I.12. Digestion

Préparer le mélange de digestion dans un tube PCR de 0,2 mL sur la glace . Vortexez 3 fois, centrifuger brièvement et placer sur la glace.

Ajouter 4 µL du mélange de digestion dans le tube à échantillon. Vortexer 3 fois, centrifuger brièvement, puis placer sur la glace.

Placez le tube PCR dans le thermocycleur. Exécutez le programme dans les conditions suivantes :

	Température	Temps
1	Couvercle chauffant 42c	Oui
2	37C°	10 minutes
3	4C°	Pause

Une fois le programme terminé, centrifugez brièvement le tube et ajoutez immédiatement 7,5µL de tampon d'arrêt de digestion dans le tube à échantillon. Mélanger au vortex 3 fois puis centrifuger brièvement et placer sur la glace.

I.12.1. Nettoyage du produit de digestion

Mélangez soigneusement les billes magnétiques de nettoyage d'ADN. Ajoutez 130 μ L de billes d'ADN propres dans le tube à échantillon. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber à température ambiante pour 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique pendant 5 minutes jusqu'à ce que le liquide soit clair. Retirer et jeter soigneusement le surnageant.

Réaliser un nettoyage avec 160 μ L d'éthanol à 80 %.

Placez-le sur le support magnétique pour séparation. Éliminez tout le liquide à l'aide d'une pipette à faible volume.

Conservez le tube sur le support magnétique. Ouvrez le bouchon du tube et laissez sécher les billes à l'air libre à température ambiante jusqu'à ce qu'elles ne présentent plus d'humidité ni de brillance. Aucune fissure ne doit apparaître à la surface des billes.

Retirez le tube du support magnétique et ajoutez 25 μ L de tampon TE pour éluer l'ADN. Mélanger au vortex jusqu'à ce que toutes les billes soient en suspension.

Incuber à température ambiante pour 5 minutes.

Centrifuger brièvement le tube et le placer sur le support magnétique pendant 5 minutes jusqu'à ce que le liquide soit clair. Transférer délicatement 24 μ L du surnageant dans un nouveau tube à centrifuger de 1,5 mL.

Contrôle qualité du produit de digestion :

Quantifier le ssCir avec le kit d'analyse Qubit ssDNA. Les produits finaux de la digestion enzymatique (ssDNA,ng) / les produits d'entrée de la PCR (dsADN, 300 ng) doivent être $\geq 7\%$.

II. Séquençage sur la plateforme MGI

Les DNBs sont chargés sur des flow cells spécifiques au séquenceur MGI G-99.

Le séquençage est réalisé en utilisant la technologie DNBSEQ, qui repose sur la synthèse d'ADN par ancrage de sondes combinatoires. Cette méthode permet une lecture efficace et précise des séquences d'ADN.

III. Données brutes : fichiers FASTQ

Les données de séquençage sont initialement générées sous forme de fichiers FASTQ, qui contiennent les séquences d'ADN lues ainsi que les scores de qualité associés à chaque base. Ces fichiers sont essentiels pour les analyses bioinformatiques ultérieures.

IV. Analyse bioinformatique

IV.1. Alignement des lectures

Les lectures des fichiers FASTQ sont alignées sur un génome de référence (GRCh38) en utilisant un outils interne au serveur du séquenceur équivalent à BWA ou Bowtie2. Cette étape génère des fichiers BAM, qui contiennent les informations d'alignement.en.mgi-tech.com

IV.2. Appel des variants (Variant Calling)

Les fichiers BAM sont ensuite utilisés pour identifier les variations génétiques (SNPs, indels, etc.) par rapport au génome de référence. Cette étape est réalisée avec un outils interne au serveur du séquenceur équivalent à GATK, Sentieon ou DeepVariant. Les résultats sont enregistrés dans des fichiers VCF (Variant Call Format).

V. Fichiers VCF

Les fichiers VCF contiennent des informations détaillées sur les variants identifiés, y compris leur position dans le génome, le type de variant, la qualité de l'appel et d'autres annotations pertinentes. Ces fichiers sont essentiels pour les analyses ultérieures et la recherche clinique.

VI. Analyse diagnostique

VI.1. Annotation des variants

Les variants identifiés sont annotés pour déterminer leur impact potentiel sur les gènes et les protéines. Des outils comme SnpEff ou VEP (Variant Effect Predictor) sont utilisés pour cette annotation.

VI.2. Interprétation clinique

Les variants annotés sont ensuite interprétés en fonction de leur pertinence clinique, en se référant à des bases de données telles que ClinVar ou gnomAD. Cette étape est cruciale pour identifier les variants pathogènes potentiels et établir des liens avec des phénotypes ou des

maladies spécifiques. L'annotation des variants et l'interprétation clinique sont pris en charge par l'équipe médicale.

Partie III : Résultats et discussion

I. Résultats

I.1. Matériel biologique de départ

Extrait d'ADN, sang humain hépariné :

Concentration : 25 ng/ μ L,

Quantité totale : 0,7 μ g

I.2. Préparation de la librairie

Quantité d'ADN utilisée : 50ng

Kit de préparation : MGIEasy Fast FS Exome Library Prep

I.3. Paramètre de séquençage

2x100 pb

Contrôle qualité du séquençage : Q30 = 92,03 %

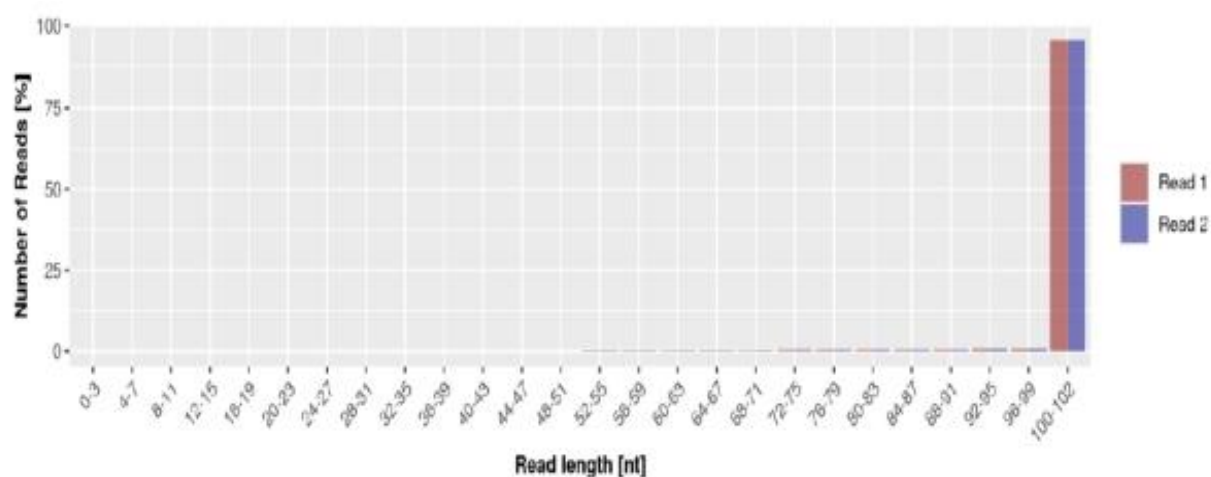


Figure N° 6 : Tailles des séquences lues (Reads) après « Trimming »

La figure n°7 indique que plus de 98% des séquences lues ont une taille comprise entre 98 et 102nt.

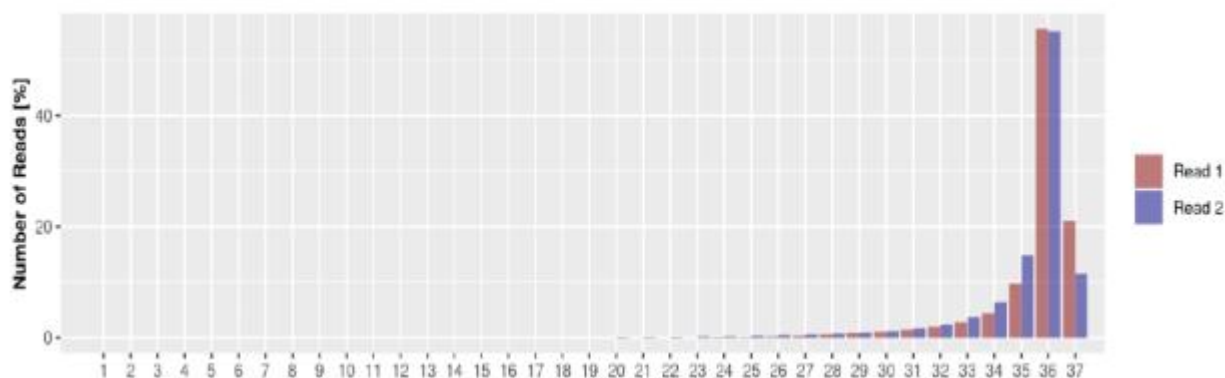


Figure N° 7 : Nombre moyen de lectures (reads) par base.

La figure n°8 montre que plus de 80% des bases ont été lues 34-37 fois en moyenne.

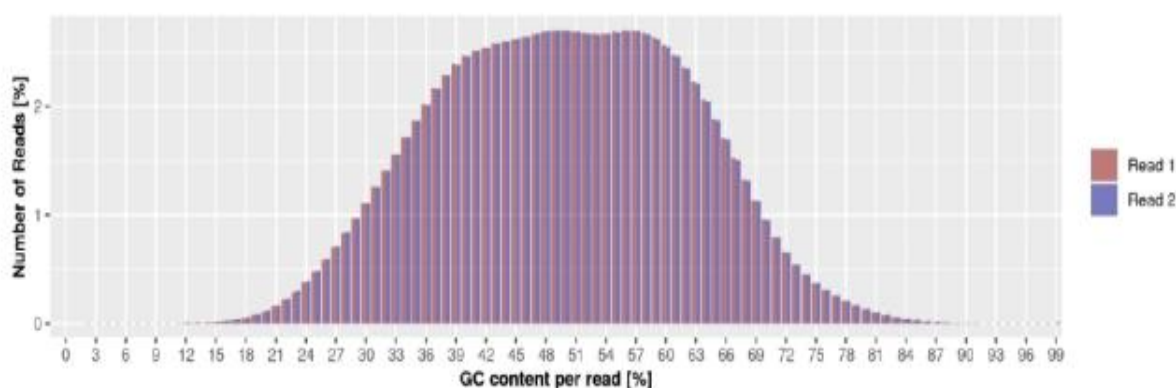


Figure N° 8 : %GC des lectures (reads)

La figure n°9 montre une distribution gaussienne du pourcentage GC des séquences lues centrée à 50%.

I.4. Analyse bioinformatique

Le démultiplexage des séquences et l'ajustage des adaptateurs (trimming) ont été réalisés avec un module associé au serveur interne du séquenceur G-99.

I.5. Données brutes : fichiers FASTQ

Les données de séquençage sont initialement générées sous forme de fichiers FASTQ, qui contiennent les séquences ADN lues ainsi que les scores de qualité associés à chaque base.

Ces fichiers sont essentiels pour les analyses bioinformatiques ultérieures.

Taille de fragments lus (reads) : **43,05 Millions**

Nombre de bases lues : **16,278 Gb**

I.6. Alignement des lectures

Les lectures des fichiers FASTQ sont alignées sur un génome de référence (GRCh38) via un module interne au serveur du séquenceur G-99. Cette étape génère un fichier BAM (Binary Alignment Map) contenant les informations d'alignement.

I.7. Appel des variants (Variant Calling)

Les fichiers BAM sont ensuite utilisés pour identifier les variations génétiques (SNPs, indels et CNVs) par rapport au génome de référence. Des outils tels que GATK, Sentieon ou Deep Variant sont couramment utilisés pour cette étape. Les résultats sont enregistrés dans des fichiers VCF (Variant Call Format).

Un logiciel interne au serveur du séquenceur G-99 a été utilisé pour la détection des variants.

Appel des variants a été effectué dans les régions cibles avec +/- 30 paires de bases.

Les variants ont été annotés à partir de diverses bases de données publiques. Les variations du nombre de copies (CNV ou copy number variations) ont été détectées en comparant le nombre de lectures chevauchant les régions génomiques cibles « Coverage » avec le nombre attendu dans une cohorte échantillons de référence. Pour analyser les CNVs sur les chromosomes X et Y, le sexe du patient (si non spécifié par le client) a été estimé à partir de la couverture afin de déterminer le nombre attendu de copies.

La qualité des fichiers FASTQ a été analysée avec FastQC (version 0.11.5-cegat) (Andrews 2010). Les tracés ont été créés avec ggplot2 (Wickham 2009) dans R (version 3.6.1) (R Core Team 2015).

Les résultats de cette étape d'analyse sont les suivants :

Nombre SNVs : **48604**

Nombre d'Indels : **9528**

Nombre de CNVs : **491**

I.8. Annotation des variants

Les variants identifiés sont annotés pour déterminer leur impact potentiel sur les gènes et les protéines. Les outils SnpEff et VEP (Variant Effect Predictor) sont utilisés pour cette annotation.

I.9. Fichiers VCF

A l'issue de l'analyse bioinformatique, des fichiers VCF (Variant call format) sont générés contenant des informations détaillées sur les variants identifiés, y compris leur position dans le génome, le type de variant, la qualité de appel et autres annotations pertinentes. Les fichiers VCF peuvent être consultés avec un outil d'édition de texte tels que bloc note ou Word et peuvent contenir un nombre de variants allant de 50000 jusqu'à 10000.

Ces fichiers sont essentiels pour les analyses ultérieures et la recherche clinique. Un exemple d'un fichier VCF est annexé à ce manuscrit.

I.10. Interprétation clinique

Cette partie de l'analyse est réalisée par le médecin généticien. Les variants annotés sont interprétés en fonction de leur pertinence clinique, en se référant à des bases de données telles que ClinVar ou gnomAD. Cette étape est cruciale pour identifier les variants pathogènes potentiels et établir des liens avec des phénotypes ou des maladies spécifiques.

II. Discussion

L'analyse de l'exome est un test important pour le diagnostic de nombreuses maladies rares et héréditaires. Dans de nombreux cas, les informations fournies par ce test suffisent pour poser le diagnostic et permettent l'établissement d'un conseil génétique adéquat.

Dans ce travail, nous avons réalisé le séquençage d'un exome dans le but de fournir au médecin généticien une liste de variants génétiques susceptibles d'être à l'origine de la pathologie du malade.

L'ADN extrait à partir d'un prélèvement sanguin du malade a subi une série de modifications afin de le préparer pour le séquençage dans la plateforme MGI G-99. Les différents points de contrôle qualité réalisés autorisent le passage d'une étape à la suivante.

La quantité d'ADN de départ (0,7 µg) et sa qualité évaluée par le Qubit sont satisfaisantes pour la suite des expérimentations.

Le score final du séquençage (Q30=93,03%) est supérieur au seuil de validation des résultats fixé par le fabricant MGI qui est de l'ordre de 90%.

Le nombre de « reads » (43,05M) et de bases lues (16,278Gb) correspond à la moyenne observée habituellement avec la technologie utilisée. Plus de 80% des bases ont été lues à plus de 34x ce qui est considéré comme acceptable.

La courbe de distribution du contenu en %GC est également dans les normes avec une forme gaussienne centrée à 50%.

Le nombre de variantes totaux (60k) est comparable au nombre obtenu habituellement par des analyses WES similaires.

L'ensemble des analyses bioinformatiques ont été réalisés au sein du serveur intégré au séquenceur G-99.

Le fichier VCF final contenant une liste de variants de plus de 55000 variants a été livré au médecin généticien qui prendra en charge la dernière étape de L'analyse et établir un compte rendu final. Tout le travail technique a été effectué sous la supervisons du médecin généticien compte tenu de l'aspect médical de l'analyse.

Conclusion

Conclusion Générale

Le séquençage de l'exome par les technologies de séquençage de nouvelle génération (NGS) constitue une révolution technologique et médicale dans l'étude du génome humain. En ciblant les régions codantes — soit environ 1 à 2 % du génome — cette approche permet d'identifier efficacement les mutations génétiques responsables de nombreuses maladies, tout en réduisant les coûts et les volumes de données à analyser par rapport au séquençage du génome entier.

La préparation des bibliothèques d'ADN, incluant l'extraction, la fragmentation, l'enrichissement exomique et la création de DNB (DNA Nanoballs), est une étape essentielle qui conditionne la qualité des résultats obtenus. Grâce à des plateformes performantes comme celles développées par MGI, le processus est devenu plus automatisé, plus rapide et plus fiable.

L'analyse bioinformatique, qui comprend la génération de fichiers FASTQ, l'alignement sur un génome de référence, le variant calling et l'interprétation des variants via les fichiers VCF, est tout aussi déterminante. Elle nécessite des outils robustes, des bases de données actualisées et une expertise multidisciplinaire pour interpréter correctement les résultats, notamment dans un contexte clinique.

Sur le plan médical, le séquençage de l'exome a transformé le diagnostic des maladies génétiques rares, permis des avancées en oncogénétique et ouvert la voie à une médecine personnalisée. Toutefois, certaines limites subsistent, telles que la couverture incomplète de certains exons, la détection difficile de certains types de variants (structuraux, répétitifs) et la gestion des variants de signification incertaine (VUS).

En somme, le séquençage de l'exome par NGS s'impose comme un outil central en génomique moderne, à la croisée de la biologie moléculaire, de la bioinformatique et de la médecine de précision. Son développement continu, associé à une meilleure standardisation et à une interprétation clinique rigoureuse, en fera un levier essentiel pour répondre aux défis diagnostics et thérapeutiques du futur.

Références

Références

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2014). *Biologie moléculaire de la cellule* (6e éd.). De Boeck Supérieur.
2. Aron, S., & Passera, L. (2009). *Les sociétés animales : évolution de la coopération et organisation sociale*. De Boeck Supérieur.
3. Ayad, N. (2024). *Étude de l'effet préventif et cardio-protecteur de Cinnamomum zeylanicum* [Thèse de doctorat, Université Ibn Khaldoun].
4. Bechkri, S. *Cours de Génétique*. [Support pédagogique].
5. Boushaba, N. (2024). *Génétique Humaine*. [Ouvrage universitaire].
6. Debré, P., & Gonzalez, J.-P. (2013). *Vie et mort des épidémies*. Odile Jacob.
7. Dessalles, J.-L., Gaucherel, C., & Gouyon, P.-H. (2016). *Le fil de la vie : la face immatérielle du vivant*. Odile Jacob.
8. Edelman, G. M. (2008). *Biologie de la conscience*. Odile Jacob.
9. Flint, S. J., Enquist, L. W., Racaniello, V. R., & Skalka, A. M. (2015). *Principles of virology* (4th ed.). ASM Press.
10. Gobat, J.-M., Aragno, M., & Matthey, W. (2010). *Le sol vivant : bases de pédologie, biologie des sols*. EPFL Press.
11. Grzych, G. (2017). *Génétique*. De Boeck Supérieur.
12. Jacquard, A. (1991). *Inventer l'homme*. Éditions Complexe.
13. Lewin, B. (2013). *Genes XI*. Jones & Bartlett Learning.
14. Madigan, M. T., Bender, K. S., Buckley, D. H., Sattley, W. M., & Stahl, D. A. (2018). *Brock biology of microorganisms* (15th ed.). Pearson.
15. Petsko, G. A., Ringe, D., & Sanlaville, C. (2008). *Structure et fonction des protéines*. De Boeck Supérieur.
16. Strachan, T., & Read, A. P. (2018). *Human molecular genetics* (5th ed.). Garland Science.
17. Voet, D., & Voet, J. G. (2016). *Biochimie*. De Boeck Supérieur.

18. Cassinari, K. (2023). *Étude de la régulation de l'expression des gènes pour l'interprétation de variations génomiques* [Thèse de doctorat, Normandie Université].

Articles scientifiques

19. Adams, M. D., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*.
20. Adzhubei, I. A., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*.
21. Bamshad, M. J., et al. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*.
22. Bayley, H. (2015). Nanopore sequencing: from imagination to reality. *Clinical Chemistry*, 61(1), 25–31.
23. Chen, X., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8), 1220–1222.
24. Chiang, C., et al. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, 49(5), 692–699.
25. Choi, Y., et al. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, 7(10), e46688.
26. Cock, P. J. A., et al. (2010). The Sanger FASTQ file format for sequences with quality scores. *Nucleic Acids Research*, 38(6), 1767–1771.
27. Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617), 286–290.
28. Drmanac, R., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961), 78–81.
29. Fonseca, N. A., et al. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24), 3169–3177.
30. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.
31. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359.

32. Lek, M., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291.
33. Li, H., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
34. Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), 589–595.
35. Margulies, M., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–380.
36. Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature*, 470(7333), 198–203.
37. Rausch, T., et al. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), i333–i339.
38. Stark, Z., et al. (2019). Integrating genomics into healthcare: a global responsibility. *The American Journal of Human Genetics*, 104(1), 13–20.
39. WHO. (2023). *Genomic Resource Centre: Genetic Diseases*. World Health Organization.
40. Adaptée de Alberts, B. et al. (2014). .
41. Alberts, B., Johnson, A., Lewis, J., et al. (2014). *Biologie moléculaire de la cellule* (6^e éd.). De Boeck..
42. Adaptée de Alberts, B., Johnson, A., Lewis, J., et al. (2014). *Biologie moléculaire de la cellule* (6^e éd.). De Boeck.
43. Adaptée de Bechkri, S. (Cours de Génétique, Université d’Alger, 2022).
44. Adaptée de Smith et al., 2020.

Annexes

Annexe – Pipeline bioinformatique pour l’analyse du séquençage de l’exome (WES)

🔗 Objectif :

Analyser les données brutes issues du séquençage (fichiers FASTQ) pour obtenir des fichiers de variants annotés (VCF), interprétables pour le diagnostic.

Étapes du pipeline :

Étape	Outil	Commande / Description
1. Contrôle qualité des lectures	FastQC	fastqc sample_R1.fastq sample_R2.fastq
2. Nettoyage des lectures (trimming)	Trimmomatic	trimmomatic PE sample_R1.fastq sample_R2.fastq [...]
3. Alignement au génome de référence (hg38)	BWA-MEM	bwa mem -t 8 hg38.fa sample_R1_trimmed.fq sample_R2_trimmed.fq > sample.sam
4. Conversion SAM → BAM	Samtools	samtools view -Sb sample.sam > sample.bam
5. Tri et indexation des BAM	Samtools	samtools sort sample.bam -o sample_sorted.bamsamtools index sample_sorted.bam
6. Marquage des duplicats	Picard	picard MarkDuplicates I=sample_sorted.bam O=sample_dedup.bam M=metrics.txt
7. Appel de variants	GATK HaplotypeCaller	gatk HaplotypeCaller -R hg38.fa -I sample_dedup.bam -O sample_raw.vcf
8. Filtrage des variants	GATK VariantFiltration	gatk VariantFiltration -R hg38.fa -V sample_raw.vcf -O sample_filtered.vcf [...]
9. Annotation des variants	ANNOVAR / VEP	table_annoar.pl sample_filtered.vcf humandb/ - buildver hg38 -out sample_annotated

Structure des fichiers en sortie :

- sample_fastqc.html : rapport qualité
- sample_sorted.bam / .bai : fichier aligné
- sample_filtered.vcf : fichier de variants filtrés
- sample_annotated.txt : variants annotés cliniquement

Remarque :

- Le pipeline est généralement automatisé via un script shell ou un workflow (Snakemake, Nextflow).
- Il est essentiel d'utiliser une base de données de référence à jour (ClinVar, gnomAD) pour l'interprétation clinique.

Exemple fichier VCF :

##fileformat=VCFv4.2
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##FILTER=<ID=HARDTOVALIDATE,Description="QD < 2.0 FS > 60.0 MQ < 40.0 HaplotypeScore > 13.0 MappingQualityRankSum < -12.5 ReadPosRankSum < -8.0">
##FILTER=<ID=LOWQUAL,Description="Set if not true: %QUAL>10">
*
*
*
*
##samtoolsVersion=1.12+htslib-1.12
##source=HaplotypeCaller

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	6332746 5_S33.F REEBAY ES	633274 65_S33 .GATK	63327465 _S33.SAM TOOLS
chrM	73	.	A	G	15853.10	PASS	AB=0;ABP=0;AC=6;AC1=2;AF=1.00;AF1=1;AN=6;AO=683;CIGAR=1X;DP=1785;DP4=0,0,319,1;DPB=683;DPRA=0;EPP=35.4425;EPPR=0;ExcessHet=3.0103;FQ=-281.989;FS=0;GQ_MEAN=1481;GTI=0;LEN=1;MEANALT=1;MLEAC=2;MLEAF=1;MQ0F=0.0106762;MQM=54.757;MQMR=0;MQSB=1;NCC=0;NS=1;NUMALT=1;ODDS=951.444;PAIRED=0.814056;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=24174;QD=32.09;QR=0;REGION=.;RO=0;RPL=217;RPP=200.131;RPPR=0;RPR=466;RUN=1;SAF=379;SAP=20.8939;SAR=304;SGB=-0.693147;SOR=1.089;SRF=0;SRP=0;SRR=0;TYPE=snp;VDB=7.48586e-13;set=Intersection;technology.illumina=1	GT:A D:AO :DP: GQ:P L:QA :QR: RO	1/1:0,68 3:683:68 3::2062 9,2056,0 :24174:0 :0	1/1:0,4 94::49 4:99:15 867,14 81,0	1/1:::320 :99:255,2 55,0
chrM	195	.	T	C	36371.10	PASS	AB=0;ABP=0;AC=6;AC1=2;AF=1.00;AF1=1;AN=6;AO=1475;BaseQRankSum=2.214;CIGAR=1X;DP=4200;DP4=0,0,689,154;DPB=1479;DPRA=0;EPP=4.81373;EPPR=7.35324;ExcessHet=3.0103;FQ=-281.989;FS=0;GQ_MEAN=3777;GTI=0;LEN=1;MEANALT=2;MLEAC=2;MLEAF=1;MQ=60;MQ0F=0.000722022;MQM=59.9675;MQMR=60;MQRankSum=0;MQSB=0.99208;NCC=0;NS=1;NUMALT=1;ODDS=2036.78;PAIRED=0.935593;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=49984;QD=27.74;QR=21;REGION=.;RO=2;RPL=820;RPP=43.0905;RPPR=3.0103;RPR=655;RUN=1;ReadPosRa	GT:A D:AO :DP: GQ:P L:QA :QR: RO	1/1:2,14 75:1475: 1479::4 4951,44 28,0:499 84:21:2	1/1:1,1 310::1 311:99: 36385, 3777,0	1/1:::843 :99:255,2 55,0

							nkSum=- 0.807;SAF=756;SAP=5.02572;SAR=719;SGB=- 0.693147;SOR=0.257;SRF=1;SRP=3.0103;SRR=1; TYPE=snp;VDB=0.0101606;set=Intersection;tec hnology.illumina=1				
chr1	145 167 4	.	T	C	21.04	PASS	AC=1;AC1=1;AF=0.500;AF1=0.499998;AN=2;BQ B=0.450401;DP=12;DP4=3,2,0,3;FQ=24.0321;M Q=60;MQOF=0;MQB=0.900802;MQSB=0.90080 2;PV4=0.196429,0.104859,1,1;REGION=.;RPB=0 .150134;SGB=- 0.511536;VDB=0.0295886;set=SAMTOOLS	GT:D P:GQ :PL	./.	./.	0/1:8:54:5 1,0,167
chr1	145 170 3	.	G	T	39.64	PASS	AB=0.555556;ABP=3.25157;AC=2;AC1=1;AF=0.5 00;AF1=0.5;AN=4;AO=5;BQB=0.372138;CIGAR= 1X;DP=18;DP4=1,3,0,4;DPB=9;DPRA=0;EPP=13. 8677;EPPR=5.18177;FQ=44.981;GTI=0;LEN=1;M EANALT=1;MQ=60;MQOF=0;MQB=0.992367;M QM=60;MQMR=60;MQSB=1;NS=1;NUMALT=1; ODDS=9.12696;PAIRED=1;PAIREDR=1;PAO=0;P QA=0;PQR=0;PRO=0;PV4=1,0.157798,1,1;QA=1 50;QR=139;REGION=.;RO=4;RPB=0.124046;RPL =0;RPP=13.8677;RPPR=11.6962;RPR=5;RUN=1;S AF=0;SAP=13.8677;SAR=5;SGB=- 0.556411;SRF=1;SRP=5.18177;SRR=3;TYPE=snp; VDB=0.0320192;set=FREEBAYES- SAMTOOLS;technology.illumina=1	GT:A D:AO :DP: GQ:P L:QA :QR: RO	0/1:4,5: 5:9:..111 ,0,101:1 50:139:4	./.	0/1:..:8:7 5:72,0,93
chr1	145 170 3	.	G	GAG GGA GGG CCG GTG TTG GTG AGG GCG TCT GGT CGT CCG T	171.60	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=- 0.253;DP=9;ExcessHet=3.0103;FS=0;GQ_MEAN =29;MLEAC=1;MLEAF=0.5;MQ=60;MQRankSum =0;NCC=0;QD=28.6;REGION=.;ReadPosRankSu m=-0.21;SOR=2.303;set=GATK	GT:A D:DP :GQ: PL	./.	0/1:1,5: 6:29:17 9,0,29	./.
chr1	145 319 9	.	G	A	570.50	PASS	AB=0.731707;ABP=22.1298;AC=2;AC1=1;AF=0.5 00;AF1=0.5;AN=4;AO=30;BQB=0.952821;CIGAR =1X;DP=125;DP4=14,11,19,7;DPB=41;DPRA=0;E PP=26.4622;EPPR=7.94546;FQ=82.0157;GTI=0;L EN=1;MEANALT=1;MQ=25;MQOF=0.511905;M QB=0.0199485;MQM=35.2667;MQMR=34.6364 ;MQSB=0.849282;NS=1;NUMALT=1;ODDS=5.13 41;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0; PRO=0;PV4=0.249161,1,1,1;QA=1006;QR=285;R EGION=.;RO=11;RPB=0.368424;RPL=14;RPP=3.2 9983;RPPR=3.20771;RPR=16;RUN=1;SAF=14;SA P=3.29983;SAR=16;SGB=- 0.692976;SRF=2;SRP=12.6832;SRR=9;TYPE=snp; VDB=0.0101692;set=FREEBAYES- SAMTOOLS;technology.illumina=1	GT:A D:AO :DP: GQ:P L:QA :QR: RO	0/1:11,3 0:30:41.. :669,0,1 04:1006: 285:11	./.	0/1:..:51: 99:242,0,1 09
chr4	322 547	.	T	C	136.64	PASS	AB=0.454545;ABP=3.20771;AC=3;AC1=1;AF=0.5 00;AF1=0.5;AN=6;AO=5;BQB=0.406082;BaseQR ankSum=-	GT:A D:AO :DP:	0/1:6,5: 5:11:..:10 9,0,149:	0/1:6,5: :11:99: 144,0,1	0/1:..:9:8 6:83,0,146

	8						0.131;CIGAR=1X;DP=33;DP4=2,3,1,3;DPB=11;DPRA=0;EPP=3.44459;EPPR=8.80089;ExcessHet=3.0103;FQ=56.0154;FS=0;GQ_MEAN=144;GTI=0;LEN=1;MEANALT=1;MLEAC=1;MLEAF=0.5;MQ=60;MQOF=0;MQB=0.974597;MQM=60;MQMR=60;MQRankSum=0;MQSB=0.924584;NCC=0;NS=1;;RPB=0.730948;RPL=4;RPP=6.91895;RPPR=8.80089;RPR=1;RUN=1;ReadPosRankSum=0.175;SAP=1;SAP=6.91895;SAR=4;SGB=-0.556411;SOR=1.179;SRF=2;SRP=4.45795;SRR=4;TYPE=snp;VDB=0.286646;set=Intersection;technology.illumina=1	GQ:PL:QA:QR:RO	154:199:6	68	
--	---	--	--	--	--	--	---	----------------	-----------	----	--