الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

كلية علوم الطبيعة والحياة

Faculté des Sciences de la Nature et de la Vie

جامعة الإخوة منتوري قسنطينة1

Université Frères Mentouri Constantine 1

**Département de Biologie Appliquée**

**قسم البيولوجيا التطبيقية**

Mémoire en vue de l'obtention du Diplôme de Master en :
**Bioinformatique**

**THÈME**

# Pipeline basé sur des logiciels libres pour l'analyse des variants issus du séquençage NGS

Présenté par :

**BENFIFI Sara**

**ZELMAT Djamila**

Soutenu le : 04 - 10 - 2020

Devant le jury :

Président du jury : Dr. GHERBOUDJ Amira

Encadreur : Dr. CHEHILI Hamza

Examinateur : Pr. HAMIDECHI M. Abdelhafid

Année universitaire 2019-2020

## *Dédicace*

*A mes parents avant tout et pour tout ;*

*A dieu de maman prolongent vie ;*

*A mes frères et mes sœurs ;*

*A tous mes amis sans exception ;*

*A tous ceux que j'aime ;*

*A tous ceux qui sont proches et chères ;*

*Djamila.*

# *Dédicace*

*C'est avec un immense honneur c'est et une grande modestie que je dédie ce modeste travail à celui et celle qui m'ont donné la vie*

*A Ma mère.*

*Pour toute sa tendresse amour et affection qui ont été pour moi une lumière et un appui d'une valeur inestimable, je te prie mère de trouver ici le témoignage de mes sentiments les plus distingués et s'il y a quelqu'un monde envers qui je dois beaucoup, ça serait toi ma mère et quoique je fasse jamais je ne pourrai te revaloir ce que tu m'as donné avec cœur et âme.*

*A Mon Frère et Mes sœurs.*

*A mon époux.*

*Je te demande de bien vouloir trouver ici ma reconnaissance pour ta patience et ton encouragement, et ton grand soutien pour réaliser ce travail.*

*A tous mes amis et à tous ceux qui sont proche et chères.*

*Sara.*

# REMERCIEMENTS

Au terme de ce travail, qu'ils nous soient permis d'exprimer mes plus vifs remerciements à Docteur CHEHILI Hamza pour avoir accepté de diriger ce travail et avoir bien voulu y consacrer son temps, ses conseils et sa précieuse aide tout au long de la conduite de cette étude.

Nous remercions vivement Docteur GHERBOUDJ Amira d'avoir acceptée de présider le jury.

Nous sommes reconnaissons envers Professeur HAMIDECHI Mohamed Abdelhafid d'avoir honoré le jury en qualité d'examinateur.

Nous remercions également le Docteur BENSAADA Moustafa, pour ses réflexions pertinentes, ses orientations et ses précieux conseils.

Nous tenons à remercier le Docteur BOUKALIA Abdelbasset pour le temps précieux qu'il nous a consacré et pour son aide dans ce travail.

Enfin, notre reconnaissance s'adresse aussi à tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.

# RÉSUMÉ

Les techniques de séquençage de nouvelle génération (NGS) permettent une détection à haut débit d'une grande quantité de variations de manière rentable. Cependant, il existe encore des incohérences et des débats sur la manière de traiter et d'analyser ces « mégadonnées ». Pour extraire avec précision des informations cliniquement pertinentes des données génomiques, il est essentiel de choisir des outils d'appel de variants précis. Les outils actuellement disponibles ont une précision variable dans la prédiction des variants cliniques. Donc, il est important de savoir comment les sélectionner, les utiliser au mieux et surtout interpréter correctement leurs résultats. Dans ce travail, on propose un pipeline qui donne une meilleure combinaison d'outils appelants de variants pour détecter séparément les variants précises d'un seul nucléotide (SNV) et les CNV et les petites insertions et délétions (InDels).

Mots clés : NGS, Mégadonnées, Pipeline, Analyse, Détection des variants.

# ABSTRACT

Next Generation Sequencing (NGS) techniques allow high throughput detection of a large amount of variation in a cost effective manner. However, there are still inconsistencies and debates on how to handle and analyze this "big data". To accurately extract clinically relevant information from genomic data, it is essential to choose accurate variant calling tools. Currently available tools have varying accuracy in predicting clinical variants. So, it is important to know how to select them, to best use and to correctly interpret theirs results. In this work, a pipeline is proposed which provides a better combination of variant calling tools to separately detect precise single nucleotide variants (SNVs), CNVs, small insertions, and deletions (InDels).



Key words: NGS, big data, Pipeline, Analyze, variant discovery.

## ملخص

تسمح تقنيات تسلسل الجيل القادم (NGS) بالكشف عن الإنتاجية العالية لكمية كبيرة من التباين بطريقة فعالة من حيث التكلفة. ومع ذلك، لا تزال هناك تناقضات ومناقشات حول كيفية التعامل مع هذه "البيانات الضخمة" وتحليلها. لاستخراج المعلومات ذات الصلة سريريًا بدقة من البيانات الجينية، من الضروري اختيار الادوات التي تسمح بالكشف عن المتغيرات بدقة. الأدوات المتاحة حاليًا لها دقة متفاوتة في التنبؤ بالمتغيرات السريرية. لذلك، من المهم معرفة كيفية اختيارها، وأفضل استخدام لتفسير النتائج بشكل صحيح. في هذا العمل، تم اقتراح مخطط تحليلي يوفر مزيجًا أفضل من أدوات الاتصال المتغيرة للكشف بشكل منفصل عن المتغيرات الدقيقة للنيكليوتيدات المفردة (SNVs) و التباين في عدد نسخ الجين ( CNVs ) وعمليات الإدراج والحذف الصغيرة (InDels).


كلمات مفتاحية: تسلسل الجيل القادم، البيانات الضخمة، مخطط تحليلي، التحليل، اكتشاف المتغيرات.

# LISTES DES FIGURES

Figure 1: Principe de séquençage Sanger .....4
Figure 2: Séquençage à haut débit : Principe et caractéristique .....5
Figure 3: Profondeur et couverture de séquençage .....8
Figure 4 : Comparaison de quelques séquenceurs de différentes générations .....9
Figure 5 : les applications du NGS .....10
Figure 6 :Différents types de variants structuraux .....12
Figure 7: Illustration schématique du processus d'appel des variants. .....16
Figure 8: Un exemple de workflow de découverte de variantes à échantillon unique.....39
Figure 9 : Schéma du pipeline d'analyse des données NGS .....41
Figure 10 : bonnes pratiques pour les SNP germinales et inde dans le séquençage du génome entier .....42
Figure 11 :Flux de travail bioinformatique de séquençage de nouvelle génération.....43
Figure 12 :Schéma général du flux de travail bioinformatique pour les tests de séquençage de nouvelle génération(NGS).....44
Figure 13: Vue d'ensemble du flux de travail bioinformatique de séquençage de nouvelle génération (NGS) .....45
Figure 14 : Qualité par base .....59
Figure 15:Distribution de la qualité moyenne des lectures. .....60
Figure 16: Distribution moyenne des bases à travers les lectures.....60
Figure 17:Distribution taux de GC par lecture. .....61
Figure 18 :Distribution taux de GC par base.....61
Figure 19:Distribution des longueurs des lectures. .....62
Figure 20:Niveau de duplication de séquence.....63
Figure 21: Contenu de l'adaptateur.....63
Figure 22 : Schéma d'alignement sur un génome de référence. .....64
Figure 23: Affichage des variants avec Vastsifter. .....65

# LISTE DES TABLEAUX

**Tableau 1 :** Contrôle qualité et nettoyage des données de séquençage....18
**Tableau 2** : Alignement de séquence ....19
**Tableau 3 :** Alignement et analyse de données de Rna-seq. ....21
**Tableau 4** : Plateforme d'analyse intégrée. ....25
**Tableau 5 :** Autre logiciel . ....26
**Tableau 6**: Détection de variants....27
**Tableau 7** : Analyse de données de chip-seq ....28
**Tableau 8**: Analyse de données RRbs (Reduced Representation Bisulfite Sequencing). ....30
**Tableau 9**: Analyse de données de rad-seq ....31
**Tableau 10** : Manipulation et visualisation de fichiers ....32
**Tableau 11** : Clustering et visualisation....34
**Tableau 12**:Puces à ADN....35
**Tableau 13** : Plateforme d'analyse intégrée....36
**Tableau 14:** Base de données des variants....37
**Tableau 15**: Outils utilisé pour effectuer un filtre fonctionnel NGS....40
**Tableau 16** : Outils utilisé pour effectuer un filtre fonctionnel NGS....41
**Tableau 17** : Outils utilisé pour effectuer un filtre fonctionnel NGS....42
**Tableau 18** : Outils utilisé pour effectuer un filtre fonctionnel NGS....43
**Tableau 19**: Outils utilisé pour effectuer un filtre fonctionnel NGS....44
**Tableau 20** : les outils utilisé pour effectuer un filtre fonctionnel NGS ....45
**Tableau 21**: Données utilisées pour toutes les étapes d'analyse ....47
**Tableau 22** : Données utilisées pour l'étape d'alignement et l'étape d'analyse des variant. ....47
**Tableau 23**: Caractéristiques des différents outils informatiques utilisés....52
**Tableau 24:** logiciels utilisés selon les étapes du processus. ....55

# GLOSSAIRE

**Bioinformatique**

La bioinformatique est l'ensemble des méthodes qui convertissent les données biologiques en informations. [11]

**Couverture de séquençage/*Breadth of coverage***

Correspond au pourcentage du génome (ou de la séquence ciblée) couverte par les fragments séquencés. [5]

**Exome**

Partie du génome constitué par les exons, c'est-à-dire les parties des gènes qui sont exprimées pour synthétiser les produits fonctionnels sous forme de protéines. [5]

**Génome**

Totalité du matériel génétique porté par l'ensemble des chromosomes d'un organisme. [5]

**Profondeur de séquençage/*Depth of coverage***

Lecture d'une même base à partir de différents fragments; une profondeur de lecture de 30x Signifie que chaque base a été séquencée en moyenne 30 fois. [5]

**Séquençage ciblé**

Séquençage de régions codantes ou sélectionnées dans un sous-groupe de gènes relativement petit. [5]

**Variant**

Changement dans la séquence d'ADN par rapport à un génome de référence qui peut ou non avoir des conséquences fonctionnelles [5].

**Zone ciblée de gènes**

Site d'un gène au niveau duquel des mutations se produisent avec une fréquence anormalement élevée (chimères, maladies résiduelles, polymorphismes) [5].

**Variation du nombre de copies**

Variations structurelles génomiques qui font augmenter (amplification) ou diminuer (délétion) le nombre de copies d'un gène ou d'une région donnée.

**Librairies:** Production de molécules d'acides nucléiques à séquencer.

# ACRONYMES

- AND: Acide désoxyribonucléotide
- APR: *Area under Precision/Recall curve*
- ARN: Acide Ribonucléique
- ASCII:*American Standard Code for Information Interchange*
- BCL: binary base call
- BED format: *Browser Extensible Data*
- BA: Benign stand Alone
- BS: Benign Strong
- BP: Benign Poor
- CCD: *Charge-Coupled Device*
- CGH: *Comparative Genomic HybridizatioN*
- CNVs: *Copy Number Variation*
- CRT: *Cyclic Reversible Termination*
- Gb: Giga-Base
- GFF format: *General Feature Format*
- HTS: *High-Throughput Sequencing*
- ISP: *Ion Sphere Particles*
- IR: *ignore the reference allele*
- SNV: *Single nucleotide variant* mutation d'une base
- SNP: *single nucleotide polymorphism*
- Kb: Kilo-Base
- MAP: *Minor Allele Frequency* la fréquence d'Allele mineur
- Mb: *Méga-Base*
- MIP*: Molecular Inversion Probe*
- NGS: *Next-Generation Sequencing*
- 32P : Traceur Radioactif
- Pb : Paire De Base
- PCR : *Polymerase Chain Reaction*
- PCRem : PCR En Emulsion
- PGM: *Personal Genome Machine*
- PMT: Photomultiplicateur
- PVS: Pathogenic Very Strong

- PS: Pathogenic Strong
- PM: Pathogenic Moderate
- PP: Pathogenic Poor
- PTP: *Pico Titer Plate*
- RRBS: *Reduced Representation Bisulphite Sequencing*
- SMRT: single molecule Real time
- SNPs: *Single Nucleotide Polymorphisms*
- SNV: *Single Nucléotide variation*
- TVC : *Torrent Variant Caller*
- VAF : variant à faible ratio allélique (*variant allele fraction*)
- VCF: *variant calling fraction*
- WGS: *Whole-Genome Sequencing*
- WES: *Whole-Exome Sequencing*

# TABLE DES MATIERES

# Table des matières

REMERCIEMENTS ..... iii
RÉSUMÉ ..... iv
LISTES DES FIGURES ..... xi
LISTE DES TABLEAUX ..... xii
GLOSSAIRE ..... xiii
ACRONYMES ..... xiv
INTRODUCTION ..... xii
CHAPITRE 1 : Concepts de base sur le séquençage ..... xii
INTRODUCTION ..... 3
1. LES TECHNIQUES D'ANALYSE GENETIQUE ..... 3
1.1 Approche gène candidats ..... 3
1.2 Puce à ADN ..... 3
1.3 Séquençage ..... 3
2. VARIANTS GENETIQUES ..... 10
2.1 Types de variants génétique ..... 11
2.2 Classification et interprétation des variants génétiques ..... 13
3. Conclusion ..... 13
CHAPITRE 2 : Analyse bioinformatique des données issues du NGS ..... 14
INTRODUCTION ..... 14
.1 PROCESSUS GENERAL D'ANALYSE BIO-INFORMATIQUE ..... 14
1.1 Control qualité ..... 15
1.2 Pré traitement ..... 15
1.3 Alignement ..... 15
1.4 Appel des variants ..... 15
1.5 Annotation des variants ..... 16
1.6 Filtrage des variants ..... 17
2. PANORAMA DES LOGICIELS EXISTANTS ..... 17
3. ETAT DE L'ART DES PIPELINES EXISTANTS ..... 39
4. CONCLUSION ..... 46
CHAPITRE 3 : ..... 47
Matériels et méthodes ..... 47
1. MATÉRIEL ..... 47
1.1Données biologiques ..... 47

1.2. Software ..... **Erreur ! Signet non défini.**
2. MÉTHODES ..... 52
2.1 Aperçu global du pipeline développé ..... 52
2.2 Description détaillée du pipeline d'analyse bioinformatique développé ..... 56
CHAPITRE 4 ..... 60
Résultats et discussions ..... 60
1. RÉSULTATS ..... 59
2. DISCUSSION ..... 65
Conclusion ..... 67
Références ..... 68

# INTRODUCTION

# INTRODUCTION

La Bioinformatique est née d'une prise de conscience par les biologistes de ce que pouvait apporter l'Informatique à la Biologie pour le traitement répétitif de processus ou pour l'analyse de grandes quantités de données. Cet apport prend toute sa mesure lors du traitement des génomes ou des grandes banques de données mondiales.

L'analyse bioinformatique nécessite principalement une forte puissance de calcul avec des algorithmes et des outils disponibles publiquement ou dans le commerce nécessite une infrastructure de calcul appropriée en plus d'une compréhension au moins de base des technologies de séquençage.

Le séquençage du génome entier (WGS) et le séquençage de l'exome entier(WES) sont des technologies de séquençage de nouvelle génération (NGS) qui déterminent la séquence génomique complète et codante pour les protéines d'un organisme, respectivement. Le séquençage en profondeur des génomes améliore la compréhension de l'interprétation clinique des variations génomiques. L'analyse des données NGS permet de comprendre l'impact et l'importance des variations génomiques. Dans la littérature, il existe plusieurs approches bioinformatiques de pointe dans la détection de variants génomiques à base d'un pipeline précis.

Cependant, face aux nombreux types des pipelines existants, et possédant chacun ses propres caractéristiques avec des licences et des coûts différents, il est difficile de développer un pipeline d'analyse des données de séquençage issues des futures plateformes de séquençage installées récemment en Algérie (Ion torrent) ou les futures stations (illumina).

Dans le cadre de ce travail, nous proposons un pipeline d'analyse des données inspirés de plusieurs pipeline existants. Pour éviter d'être liés aux propriétaires des solutions, il est basé sur des logiciels libres de droit.

Le manuscrit est organisé en quatre chapitres. Le premier présente les concepts de base de séquençage, ceux-ci concernent en premier lieu les techniques d'analyse génétique, et les déférentes techniques de séquençage. Ce chapitre se termine par les déférentes variations génétiques qui existent et leur classification et interprétations.

Le deuxième chapitre traite l'analyse bioinformatique des données issues du NGS. Ensuite, il évoque le processus général d'analyse bio-informatique en citant les déférentes étapes qui existe. En plus, il dresse un état de l'art des pipelines existants. Ce deuxième chapitre se termine par une analyse des outils bioinformatiques existants.

Le troisième chapitre décrit un nouveau pipeline proposé pour le traitement des données NGS. Il contient les étapes nécessaires pour identifier les variants de manière fiable. Ce chapitre précise les logiciels à utiliser dans chaque étape du pipeline.

Le quatrième et dernier chapitre dresse une comparaison du pipeline developpé avec les pipelines existants en discutantles points communs et les différences selon des critères bien définis.

# CHAPITRE 1 :

# Concepts de base sur le séquençage

# INTRODUCTION

Depuis la description du structure d'ADN en 1955 jusqu'à nos jours, la biologie a connu plusieurs techniques d'analyse génétique. Dans ce premier chapitre, on cite l'approche gène candidats, puce à ADN et le séquençage. La première approche est bien souvent infructueuse, et a quasiment disparu au profit des deux dernières. Les variations génétiques prennent leur place dans ce chapitre pour présenter essentiellement leurs types, classification et interprétation.

# 1. LES TECHNIQUES D'ANALYSE GENETIQUE

Les techniques d'analyse génétique permettent d'analyser le matériel génétique d'un individu, ces techniques sont décrites ci-dessous

## 1.1 Approche gène candidats

L'approche "gènes candidats" consiste à rechercher des mutations chez un patient dans un ou plusieurs gènes cibles. Il existe principalement trois cas pour le choix des gènes cibles, la première méthode consiste à rechercher des mutations sur le gène hortologue humain par l'étude de gènes reliés à des phénotypes proches du phénotype étudiés dans différents modèles animaux et notamment murins. Un autre cas où les variants seront recherchés sur les gènes paralogues à un gène précédemment identifié avec l'idée important que leur structure proche implique une fonction similaire. Le dernier cas c'est l'étude de gènes connus comme étant des partenaires de gènes déjà identifiés dans cette pathologie [2].

## 1.2 Puce à ADN

Le concept de bio puce (= puce à ADN) est né dans les années 1990. Il repose sur une technologie pluridisciplinaire intégrant la micro-électronique, la chimie des acides nucléiques, l'analyse d'image set la bio-informatique [3].Les puces à ADN sont des lames de verre activé sur lesquelles sont déposés de nombreuses copies d'une séquence d'ADN spécifique d'un gène donné [4],Permettent des tests plus rapides, plus sensibles et plus spécifiques dans le but de mesurer le taux d'expression des transcrits provenant de plusieurs milliers de gènes lors d'une seule et unique expérience. Comme elle permit de déterminer des patterns d'expression de gènes à un statu physiologique donné. L'analyse des "signatures" d'expression a ainsi permis de caractériser plusieurs cancers. Ainsi, cette technologie a été utilisée afin de détecter des *single nucleotide polymorphisms (SNPs)*au sein de notre génome De même, l'utilisation des puces à ADN a permis la détection de *copy number variation (CNVs)* [2].

## 1.3 Séquençage

Le séquençage d'ADN consiste à déterminé l'enchaînement des nucléotides qui composent la séquence d'ADN [1]. On distingue deux types de séquençage, le séquençage intégral du génome (*whole- genome sequencing ou WGS*) qui détermine la composition en nucléotides du génome relativement à tout le matériel génétique d'une cellule et le séquençage intégral d'un exome(*whole-exome sequencing ou WES*) qui détermine la succession de nucléotide d' un fragment d'ADN donnée dans un exome représentant la totalité des régions de codage de protéines du génome en augmentant la couverture des régions d'intérêt par des technologies à haut débit[4].

### 1.3.1 Séquençage de première génération

Deux méthodes ont été développées dans les années 70, l'une par l'équipe de Walter Gilbert, aux États-Unis, et l'autre par celle de Frederick Sanger, au Royaume-Uni, L'approche de Sanger est une méthode par synthèse enzymatique sélective, tandis que celle de Maxam et Gilbert est une méthode par dégradation chimique sélective [12].

- Méthode de Maxam et Gilbert

Cette méthode est basée sur une dégradation chimique de l'ADN et utilise les réactivités différentes des quatre bases A, T, G et C, pour réaliser des coupures sélectives.

On peut décomposer ce séquençage chimique en six étapes successives :

- Marquage des extrémités des deux brins d'ADN à séquencer par un traceur radioactif ($^{32}P$).
- Isolement du fragment d'ADN à séquencer.
- Séparation et purification de brins.
- Modifications chimiques spécifiques.
- Clivage d'ADN au niveau de la modification par réaction avec une base la pipéridine.
- Analyse des toutes les produits des différentes réactions qui sont séparés par électrophorèse [12].
- Méthode Sanger

L'une des premières méthodes utilisées pour séquencer l'ADN et la méthode séquençage d'ADN rapide [1], a été développée en 1997 par le chercheur britannique Frederick Sanger. Cette technique basée sur une méthode *chain-termination* [2]. Cette technologie améliorée avec le

temps et largement utilisée depuis son invention, elle est considérée aujourd'hui comme le gold standard de la génétique médicale [5].

Cette technique est basée sur la synthèse du brin d'ADN complémentaire à partir d'un brin d'ADN matrice par réaction en cycle (dénaturation d'ADN, hybridation d'amorce et l'élongation) (figure 1.1).

La technique de Sanger permet de séquencer de façon fiable avec une précision élevée des petits fragments d'ADN (jusqu'à 400 à 900 pb), avec un faible débit. Le débit de cette technologie de séquençage et le cout par base est élevé, limite son utilisation à la détection de variants dans des régions de petite taille (1 kb) ou à la validation de variants détectés par une autre technique. De plus, il est impossible de détecter des variants minoritaires qui ont un faible ratio allélique, comme cela peut être le cas pour les variants tumoraux, car le seuil de détection d'un variant par séquençage Sanger est d'environ 15-20% d'allèle muté [1].

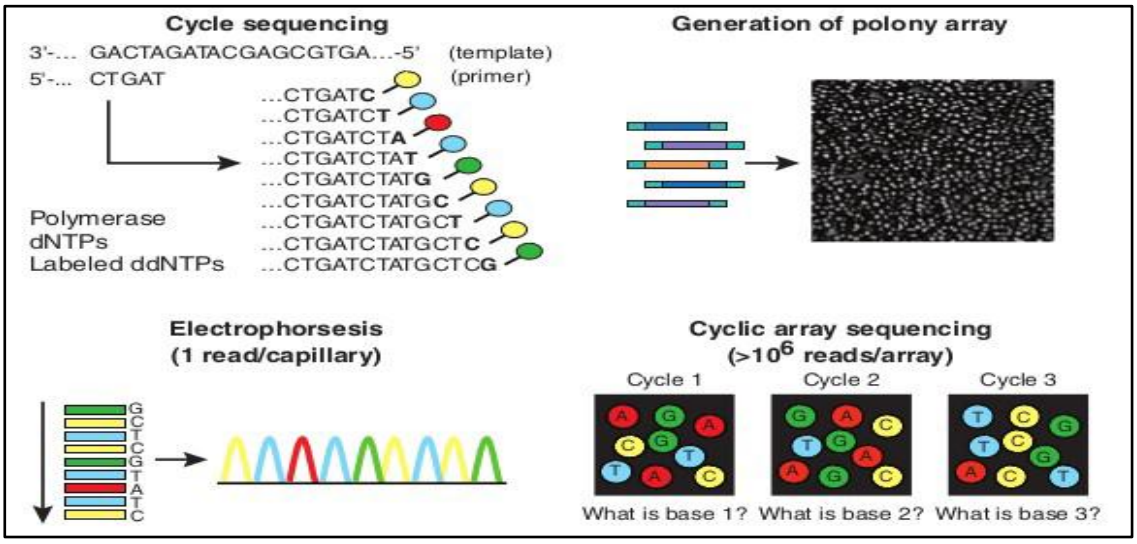



Figure 1: Principe de séquençage Sanger [5].

### 1.3.2 Séquençage de nouvelle génération

On désigne par séquençage haut débit *(HTS pour high-throughput sequencing)* aussi appelé NGS pour *next-generation sequencing* regroupe l'ensemble des technologies ou plateformes de séquençage développées depuis2005 produisant des millions de séquences en un run et à faibles coût [14]. Ces nouvelles technologies de séquençage massivement parallèle ont permis d'augmenter le débit d'analyse de façon très importante passant de quelques milliers de paires de bases séquencées à plusieurs milliards [15].

Les principales plateformes de séquençage à haut débit de 2eme génération actuellement utilisées sont présentées par 3 sociétés, Roche, Illumina et Life technologies. Ces séquenceurs vont travailler sur support solide, Illumina présente une capacité de séquençage largement supérieure aux deux autres, son HiSeqX remporte la palme, alors que Roche et Ion PGM de Life technologies permettent des lectures de plus grandes tailles (2x100pb à 2x300pb vs 400 à 700pb) Les séquenceurs SOLiD de Life technologies offrent une meilleure exactitude de séquençage et des délais plus courts, mais cela pour le même coût onéreux d'équipement que les machines Illumina qui surpassent de six fois la capacité de séquençage de son concurrent (Figure 2) [7].

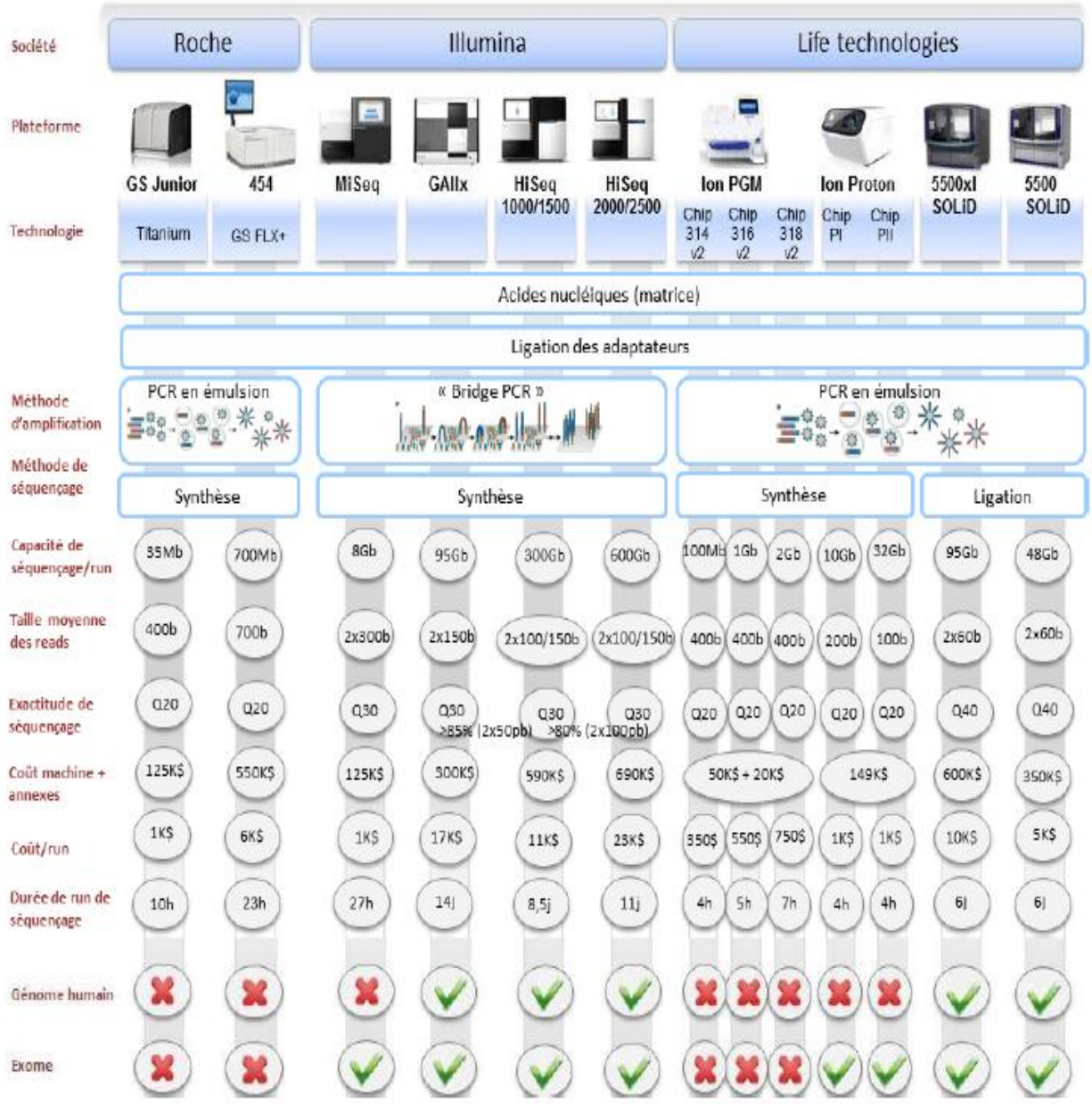



Figure 2: Séquençage à haut débit : Principe et caractéristique [7].

Dans le cadre de ce mémoire, on intéresse au séquençage NGS de 2eme génération, les différentes étapes de cette technique de la préparation de l'échantillon au séquençage sont décrites ci-dessous :

– La préparation de la librairie

La première étape consiste à préparer les fragments d'ADN à séquencer. Il existe deux grandes stratégies pour l'obtention de l'ADN à séquencer, selon la région d'intérêt à analyser.

1) Séquençage du génome entier : consiste à fragmenter et séquencer l'intégralité du génome. Il existe deux méthodes pour fragmenter l'ADN

– Fragmentation mécanique, par sonication (en utilisant des ultra-sons qui cassent l'ADN) ou par nébulisation.
– Fragmentation enzymatique, à l'aide des enzymes de restriction qui coupent l'ADN au niveau de sites de restriction.

La fragmentation permet d'obtenir des fragments d'ADN de taille compatible avec la technologie de séquençage. L'étape suivante de préparation de la librairie consiste à ajouter aux extrémités de ces fragments des adaptateurs permettant leur fixation sur le support de séquençage pour l'amplification et le séquençage [1].

2) Séquençage ciblé : Pour de nombreuses applications, il peut être intéressant de ne séquencer qu'une partie du génome et non pas son intégralité. Dans cette sous partie de génome ciblé on peut trouver par exemple : une région génomique spécifique à laquelle une pathologie a déjà été associée, l'ensemble des exons de certains gènes candidats, ou encore l'intégralité des exons de l'ensemble des gènes codant pour une protéine. Dans ce cas on parle alors de *whole exome sequencing* [2]

– L'amplification clonale

Dans la plupart des technologies, la phase de séquençage est précédée par une étape d'amplification de l'ADN. Les fragments attachés aux adaptateurs sont déposés et distribués aléatoirement sur le support de séquençage (*Flow Cells* dans le cas de la technologie illumina, *Ion Sphere Particles* (ISP) dans le cas de la technologie Ion-Torrent par exemple) qui contient des spots d'amplification ou centre de réaction, donc chaque spot être le représentant d'un unique fragment d'ADN. Chaque fragment d'ADN subit ensuite une amplification clonale, elle permet d'obtenir dans une région définie plusieurs milliers de copies du même fragment d'ADN, Appelés des clones. Cette étape assure que le signal émis lors du séquençage pourra être distingué du bruit. Ceux-ci seront ensuite séquencés parallèlement aux autres spots.

Une plateforme de séquençage peut gérer plusieurs millions de ces centres de réactions simultanément, séquençant ainsi plusieurs millions de molécules d'ADN en parallèle, donnant ainsi le nom de séquençage massif en parallèle à ces techniques [1] [2].

– Réaction de séquence

La réaction de séquence est l'étape suivant l'amplification. Elle consiste à déterminer l'ordre dans lequel se succèdent les nucléotides de l'ensemble des clones générés dans la phase d'amplification. Il existe plusieurs technologies de séquençage en fonction du séquenceur utilisé. On peut citer le séquençage par synthèse, par ligation ou bien le séquençage en temps réel (SMRT). Les séquenceurs de 2eme génération utilisent la technologie de séquençage par synthèse et par ligation.

– Séquençage et génération de données

La librairie préparée précédemment est séquencés par le séquenceur. Chaque base séquencée génère un signal qui lui est spécifique. Ce signal brut, qui peut être un signal lumineux dans le cas d'Illumina ou une différence de pH dans le cas d'Ion Torrent, est converti par le séquenceur en une séquence nucléotidique. Chaque séquence nucléotidique obtenue par séquençage d'un fragment d'ADN est appelée un read (une lecture). Ces reads sont enregistrés dans un fichier au format Fastq, contenant les séquences nucléotidiques et leurs scores de qualité [1].

Chaque position nucléotidique est séquencée plusieurs fois, et le nombre de fois où une base est séquencée correspond à la profondeur de séquençage (figure 3)

1) La profondeur de séquençage moyenne (P) : est le nombre moyen de reads qui couvrent une base. Par exemple, 30X veulent dire qu'en moyenne, une base est couverte par 30 reads. P = L.N/G, où L est la longueur moyenne des reads, N le nombre de reads et G la longueur du génome haploïde [1].
2) La couverture de séquençage (C) : correspond au pourcentage de la région d'intérêt bien couverte par des reads. Cx = Bx/G, avec Bx le nombre de bases couvertes par au moins x reads et G la longueur du génome haploïde [1].

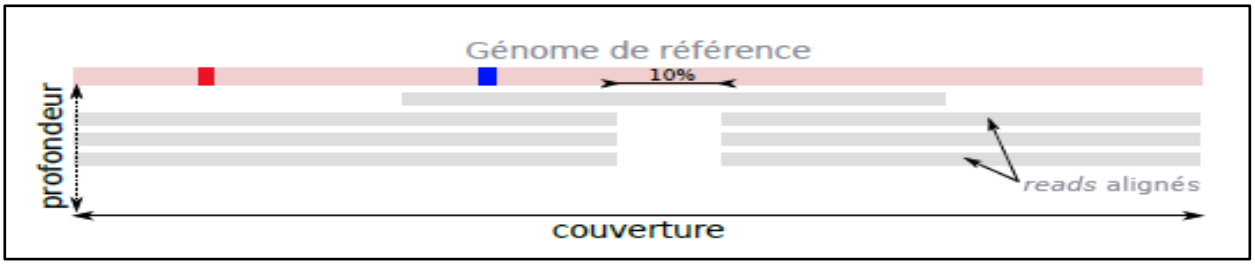



Figure 3: Profondeur et couverture de séquençage [1].

La profondeur de séquençage au niveau de la base rouge est égale à 3X, alors qu'au niveau de la base bleue elle est égale à 4X. La profondeur de séquençage moyenne est environ égale à 3X. La couverture de séquençage est de 90% (10% de bases sont mal couvertes), si on considère qu'une profondeur de 3X est suffisante [1].

– Le séquençage de 3eme génération

Une nouvelle génération de séquenceurs se profile et fait appel, entre autres, à une technologie de séquençage par nanopore, qui permet le séquençage direct de molécules uniques et rend ainsi superflu le marquage fluorescent, étant donné que chaque nucléotide émet un signal électrique spécifique [4].

Le génome comprenant beaucoup d'éléments complexes trop longs pour être résolus par les technologies de *short-read sequencing* d'autres approches ont pu voir le jour avec pour objectif de séquencer des fragments de plusieurs milliers de paires de bases (jusqu'à 200 kb : *long-read sequencing*), permettant de mettre en évidence des variants structuraux de grande taille de l'ADN, ou encore des régions fortement répétées. Elles permettent également de reconstruire des haplotypes complets. Ici aussi, deux types de méthodes se confrontent : le séquençage en temps réel de molécules uniques (*single-molecule real-time sequencing*= SMRT) et l'approche synthétique basée sur les technologies de *short-read sequencing* pour construire des *reads* longs *in-silico* [10].

### 1.3.3 Comparaison des technologies

Comme illustré précédemment, les technologies de séquençage de nouvelle génération constituent un progrès considérable dans la lecture de l'ADN, notamment en termes de coût et de rapidité (Figure 1.6). Il conviendra de plus de choisir judicieusement la technologie employée en fonction de l'application. Les séquenceurs de fragments courts offrent une plus grande précision de lecture, permettant une détection plus sensible des variants de petite taille (de 1 à quelques dizaines de nucléotides), avantage précieux dans des domaines tels que la cancérologie pour la détection des variants somatiques. Les séquenceurs de fragments longs, malgré un taux d'erreurs

plus important, mettront plus facilement en évidence des réarrangements structuraux de grande taille, ou pourront lire des transcrits ARNm entiers [10].

| | Technologie de séquençage | Longueur des *reads* | Débit d'un *run* | Taux d'erreurs | Durée d'un *run* | Coût par Gb |
|---|---|---|---|---|---|---|
| **ABI 3730xl** | Séquençage par terminaison de chaîne | 400 à 900 pb | 2100 Kb | 0.001% | 1-3 hr | 2 400 000 $ |
| **Illumina MiSeq V3** | Séquençage par synthèse | 300 pb (PE) | 15 Gb | 0.1%, substitution | 21-56 hr | 100 $ |
| **Ion Proton** | Séquençage par synthèse | 300 pb (SE) | 10 Gb | 1%, InDel | 2-4 hr | 80 $ |
| **Oxford Nanopore MinION** | ONT | 200 Kb | 1.5 Gb | ~12%, InDel | 48 hr | 750 $ |
| **Pacific Biosciences RSII** | SMRT | 20 Kb | 1 Gb | 13%, en lecture unique | 4 hr | 1 000 $ |

Figure 4 : Comparaison de quelques séquenceurs de différentes générations [1].

En rose, un séquenceur Sanger. En orange, deux séquenceurs NGS de seconde génération. En bleu, deux séquenceurs de troisième génération. SE : Single-End. PE : Paired-End. SMRT : Single-Molecule Real-Time [1].

### 1.3.4 Les application du NGS

Généralement, les applications de séquençage NGS se regroupe en 4 catégories : le séquençage (ou assemblage) de novo, le reséquençage, l'analyse du transcriptome (RNA-seq) et les analyses fonctionnelles (ChIP-Seq, MeDIP-Seq) [4]

– Le séquençage (ou assemblage) de novo

Permet de trouver la séquence génomique inconnu. La combinaison de plusieurs techniques d'analyse permet d'obtenir ainsi du matériel génomique de bonne qualité [4].

– Le reséquençage

Consiste a fournir des information pour connaitre les variations génomiques, le reséquençage indiqué lorsque la séquence du génome de référece et déjà connue. La séquence à l'étude est comparée à celui de référence par un séquençage à haut débit. « Cette approche peut remplacer les méthodes traditionnelles d'hybridation génomique comparative (CGH) et permettre de préciser les diagnostics soit de façon préventive, soit de caractériser une pathologie déjà déclarée. Il est ainsi possible, par exemple, de typer ou de suivre l'évolution des tumeurs cancéreuses chez les patients » [4].

– L'analyse du transcriptome (RNA-seq)

Cette approche qui est de plus en plus utilisé, consiste à séquencer tout l'ARN messager des cellules cancéreuses. L'analyse du transcriptome permet une appréciation des mutations, étant donné que les ARN messagers sont dérivés des exons et donne une idée des profils d'expression des gènes, autant sur l'abondance que sur la composition des gènes transcrits. Il est possible de mettre en évidence des variations structurelles, notamment les produits des fusions exprimés [4].

– Les analyses fonctionnelles (ChIP-Seq, MeDIP-Seq)

Dans le domaine de la génomique fonctionnelle, cherche à quantifier le nombre et le type d'éléments biologiques présents au lieu de chercher à connaître la séquence d'ADN des échantillons. « Il est, par exemple, possible de connaître les régions de l'ADN où se fixent les facteurs de transcription (ChIP-Seq) et de déterminer les modifications épigénétiques d'un génome en cartographiant ses sites de méthylation (MeDIP-Seq) » [4].

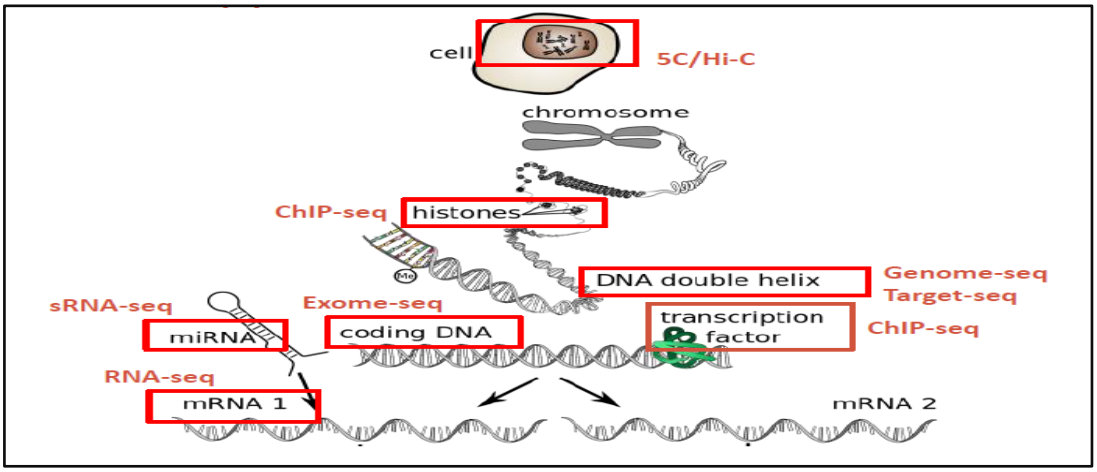



Figure 5 : Les applications du NGS [9].

## 2. VARIANTS GENETIQUES

Les variants génétiques sont des modifications de la séquence d'ADN et donc de l'information contenue dans cette séquence. La modification de la base est appelée <mutation>. Les conséquences de ces variations dépendent de la partie du génome touchée.
L'apparition des variants peut être spontanée ou bien induite. Les mutations spontanées peuvent survenir lors du processus de réplication d'ADN [1].

## 2.1 Types de variants génétique

Il existe plusieurs types des variants, distingués notamment selon la taille de la région affectée.

### 2.1.1 Variants ponctuels

Les variants ponctuels correspondent à des substitutions, délétions ou insertions ne touchant qu'un seul nucléotide ou un petit nombre de nucléotides.
Substitutions (*SNV, Single Nucleotide Variation*) : Ces variants correspondent au remplacement d'une base par une autre. Elles sont divisées en deux classes,

1) Les transitions et les transversions :

Une transition est un remplacement d'une base par une autre de la même catégorie chimique purine par une autre purine, par exemple A/G, ou bien pyrimidine par une autre pyrimidine, par exemple T/C. Par contre, la transversion est un remplacement d'une base par une autre d'une catégorie chimique différente purine par pyrimidine.

Les conséquences de ce type de variants dépendent de deux facteurs principaux, la position dans le gène (région codante ou non codante), et le type de la base remplaçante [1].
- Substitution silencieuse ou synonyme ou même-sens : la variation modifie la séquence d'un codon sans en modifier la signification, ce qui est possible grâce à la dégénérescence du code génétique.
- Substitution faux-sens : la variation modifie la séquence d'un codon qui code un acide aminé différent.
- Substitution non-sens : le codon est remplacé par un codon stop. Ces variants conduisent à la terminaison prématurée de la traduction, et à la production d'une protéine tronquée.

2) Insertions/Délétions (InDel) **:**

Ce sont des délétions ou des insertions d'une ou de plusieurs bases. Ce type de variants induit un décalage du cadre de lecture à partir de la position du variant si la taille du fragment inséré/délété n'est pas un multiple de trois, entraînant souvent l'apparition d'un codon stop prématuré.

### 2.1.2 Variants structuraux

Les variants structuraux (SV, Structural Variations) sont des variations qui affectent une grande région génique ou chromosomique. Ils sont définis comme étant des altérations génomiques impliquant des segments d'ADN ayant généralement une taille plus grande que 1 kb. Il en existe plusieurs types :

Les variations du nombre de copies (*CNV, Copy Number Variation*) : elles sont définies comme étant des segments d'ADN qui sont présents en un nombre de copies différent de celui du génome de référence. Il peut s'agir de duplication ou de délétion.

- Les inversions : l'orientation d'un segment d'ADN est inversée par rapport au reste du chromosome.

- Les translocations : ce sont des réarrangements touchant un ou plusieurs chromosomes et correspondant au déplacement d'un fragment de chromosome plus ou moins long. Il peut s'agir d'échanges réciproques de matériel chromosomique entre des chromosomes différents, qui peuvent être équilibrés. Les translocations peuvent être sans conséquence notamment si elles n'entraînent pas d'interruption de gènes. Si, par contre, le point de cassure d'une translocation se situe à l'intérieur d'un gène, la translocation provoque une interruption de ce gène et aura un effet délétère.

-Les insertions : Les insertions correspondent à l'introduction d'une séquence qui peut être une séquence endogène mobile, ou une séquence exogène virale.

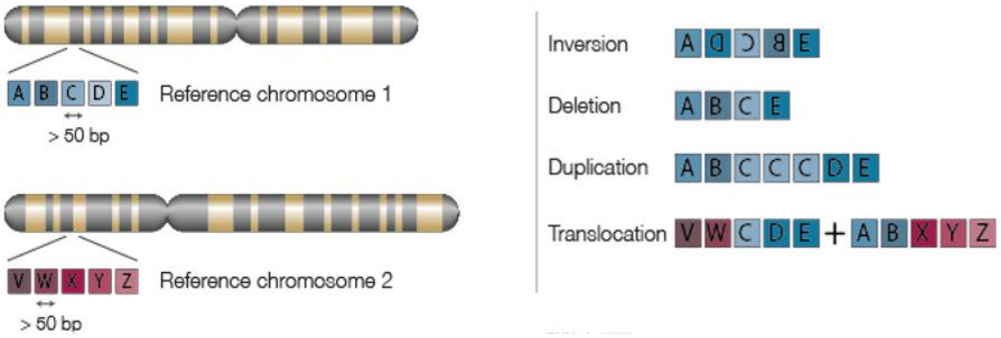



Figure 6 : Différents types de variants structuraux [81].

### 2.1.3 Variants germinaux et somatiques

Les variants peuvent survenir dans les cellules germinales ou dans les cellules somatiques.

-Variants germinaux: Un variant germinal est un variant présent dans les cellules germinales (ovocyte ou spermatozoïde). De ce fait, il sera transmis à la descendance et l'embryon sera porteur du variant dans toutes ses cellules.

-Variants somatique : Un variant somatique est un variant qui survient dans une cellule somatique (non germinale) et par définition n'est pas transmis à la descendance.

2.2 Classification et interprétation des variants génétiques

Dans le diagnostic génétique, la classification des variants constitue la base du jugement clinique, avant de classé ces variants il faut les interprétés en basant sur un faisceau d'arguments. Ces arguments ont un poids plus ou moins important dans l'interprétation du variant mis en évidence.

– PVS/PS/PM/PP : Argument très fort (*Pathogenic Very Strong*) / fort *(Pathogenic Strong*) / moyen (*Pathogenic Moderate*) / faible (*Pathogenic Poor*) en faveur de la pathogénicité du variant.
– BA/BS/BP : Argument suffisant (*Benign stand Alone*)/fort (*Benign Strong)* /faible (*Benign Poor*) en faveur du caractère bénin du variant.

L'interprétation des résultats est sous la responsabilité exclusive du biologiste et consiste à combiner ces arguments pondérés afin d'assigner une des 5 classes suivantes au variant étudié:

- Classe 1 : Variant bénin
- Classe 2 : Variant probablement bénin
- Classe 3 : Variant de signification inconnue
- Classe 4 : Variant probablement pathogène
- Classe 5 : Variant pathogène

## 3. Conclusion

Dans ce chapitre nous avons vu que le l'avènement des technologies de séquençage de nouvelle génération a grandement favorisé les progrès dans l'étude des maladies humaines aux niveaux génomique, transcriptomique et épigénétique, et que la région codante du génome est capturée et séquencée à un niveau profond, s'est avéré être une méthode rentable pour détecter des variantes pathogènes et découvrir des gènes cibles.

# CHAPITRE 2 : Analyse bioinformatique des données issues du NGS

## INTRODUCTION

La bioinformatique joue un rôle central dans l'analyse des données générées par le Séquençage haut débit. Le NGS a fait entrer la bioinformatique comme nouvelle compétence indispensable au sein des laboratoires de génétique moléculaire.

La capacité à détecter des variations génétiques de différents types est possible le plus souvent en combinant plusieurs programmes, libres ou commerciaux et dont chacun est capable de détecter un type d'événement mutationnel spécifique [7].

Alors, dans ce chapitre on décrit le cadre général de l'analyse des données de séquences. Ensuite, nous détaillons les étapes d'analyse des données. Ceci permettra de comprendre comment à partir des données issues de la machine de séquençage arriver à une liste de variantes ayant un sens pour le généticien. Puis, adresse un panorama des logiciels qu'on peut les utilisés dans le traitement des données du séquençage. On termine le chapitre par un état de l'art sur les pipelines existant.

## 1. PROCESSUS GENERAL D'ANALYSE BIO-INFORMATIQUE

L'analyse bioinformatique par technique NGS peut être découpée en déférentes étapes qui sont représentées dans le schéma suivant :

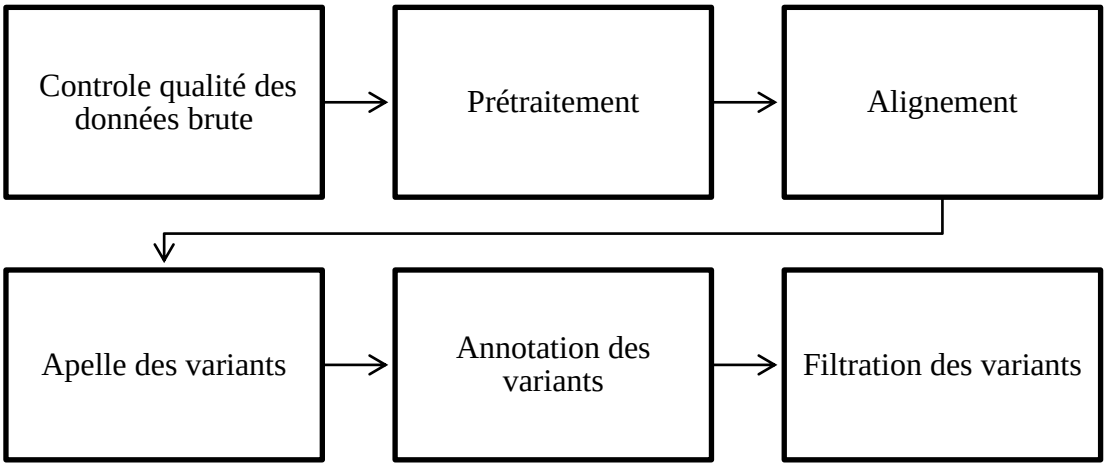



Figure 7 Flux de travail de base pour l'analyse des données issu de séquençage.

### 1.1 Control qualité

FASTQ et FASTA sont des formats standards pour représenter les données de séquence biologique. Le format FASTA est une représentation textuelle des séquences, qui commence par le nom de la séquence suivi de lignes de nucléotides ou d'acides aminés codés à une seule lettre. Le format FASTQ a été développé pour incorporer les scores de qualité de base à l'échelle Phred afin de faciliter l'évaluation de la qualité de la séquence. Il est largement accepté comme format de fichier standard pour les données brutes NGS. Plusieurs outils ont été développés pour évaluer la qualité des données brutes NGS [20].

Certains des outils couramment utilisés incluent FastQC [21], FastQScreen [22], FASTX-Toolkit [23], NGS QC Toolkit [24], PRINSEQ [25], QC-Chain [26] et récemment publié QC3 [27].

### 1.2 Pré traitement

La procédure de prétraitement standard comprend le retrait de l'adaptateur d'extrémité 3 'et le rognage des bases de faible qualité aux extrémités des lectures. Selon la conception de l'étude et l'utilisation des données, les lectures redondantes et les séquences indésirables telles que la contamination par les amorces, les adaptateurs ou d'autres espèces peuvent être supprimées à ce stade [20].

### 1.3 Alignement

-Il s'agit de l'étape d'alignement des *reads* sur une séquence de référence. Cet alignement est réalisé grâce à des algorithmes d'alignement qui positionnent chaque « *read*» sur une position génomique selon des analyses probabilistes. Il existe plusieurs aligneurs et certains sont plus adaptés aux SNV, d'autres aux CNV et d'autres aux Indels. Parmi ces outils, on peut citer : *BWA-MEM, Novoalign,* Bowtie et *MOSAIK*. Ces outils sont basés sur des algorithmes comme Burrows Wheeler. Comme sortie de ces outils, on obtient alors un fichier BAM (« *Binary Alignment Map* ») associant à chaque read ses coordonnées génomiques [2].

### 1.4 Appel des variants

L'appel des variants, ou *variant calling*, fait référence à l'ensemble des méthodes permettant d'identifier des SNVs ou des indels ou des CNV à partir des résultats de l'alignement. Cette étape est souvent différenciée de l'alignement. Cependant, les résultats de l'appel étant extrêmement dépendants de l'alignement, il est conseillé d'effectuer son appel, en tenant compte de l'aligneur choisi. Le variant est toute différence de séquence observée entre un individu et la séquence de référence utilisée (figure 8). Pour reprendre la comparaison avec la construction

d'un puzzle, cette étape consiste à détecter quelles sont les pièces qui présentent des différences avec le modèle [15] [2].


Figure 8: Illustration schématique du processus d'appel des variants [2].

Pour chaque position couverte, le pourcentage de read portant un allèle variant est analysé. Lorsqu'il est proche des 100%, l'appel est homozygote pour le variant, lorsqu'il est proche des 50% l'appel est hétérozygote.

Lorsqu'à une position donnée, peu de *reads* portent un variant, la cause est souvent une erreur de séquençage.

Il existe plusieurs logiciels d'appel des variants, Parmi les plus connus sont SAMtools, Genome Analysis ToolKit – Haplotyp eCaller (GATK-HC), Freebayes, SOAPsnp et Torrent Variant Caller (TVC).

Les quatre premiers peuvent être utilisés pour analyser des données provenant de tout type de plateforme de séquençage tandis que TVC a été développé spécifiquement pour les données provenant d'Ion Proton. La plupart de ces callers se basent soit sur des méthodes heuristiques, soit sur des méthodes probabilistes [2].

### 1.5 Annotation des variants

L'annotation des variantes est une autre étape critique de l'analyse WES / WGS Workflow. Le but de tous les outils d'annotation fonctionnels est d'annoter les informations de les effets / conséquences de la variante, y compris mais sans s'y limiter, énumérer quel gène(s) / transcript(s) sont affectés, détermination de la conséquence sur les protéines séquence, corrélation du variant avec des annotations génomiques connues (par exemple, séquence codante, séquence intronique, ARN non codant, régions régulatrices, etc.), et correspondant à des

variantes connues trouvées dans des bases de données de variantes (par exemple, dbSNP [46], 1000 Projet Génomes [47]ExAc [48], gnomAD [49], COSMIC [50], ClinVar [51], etc.).

La conséquence de chaque variante est exprimée par l'ontologie de séquence (SO) termes. La gravité et l'impact de ces conséquences sont souvent indiqués en utilisant qualificatifs (par exemple, faible, modéré, élevé) [18].

### 1.6 Filtrage des variants

Le nombre de variantes candidats est réduit à l'aide d'une stratégie de filtrage et de hiérarchisation en trois étapes pour générer une courte liste de mutations candidates pour la validation expérimentale. Le filtrage passe par les étapes suivantes :

-La première étape consiste à supprimer les appels de variantes moins fiables. Cela inclut les variantes à faible couverture, de faible qualité, biaisées par brin, situés dans des clusters SNV et / ou pris en charge par un alignement de lecture à faible confiance [79].

-La deuxième étape consiste à limiter les variantes à celles dont la fréquence de population est relativement faible, en supposant que les variantes communes sont moins susceptibles de provoquer la maladie que les rares.

-La troisième étape consiste à hiérarchiser les variantes par rapport à la maladie. En général, les SNV peuvent être ordonnés par leur effet codant, auquel cas les mutations d'épissage (SNV survenues au niveau des sites donneurs ou récepteurs d'épissage) et les mutations non-sens sont en général plus dommageables que les mutations faux-sens. Les indels, d'autre part, peuvent être ordonnés en fonction du fait qu'ils provoquent une interruption d'épissage ou un décalage de cadre de la séquence codante.

## 2. PANORAMA DES LOGICIELS EXISTANTS

L’étude des logiciels utilisés dans les étapes décrites ci-dessus, montre que chacun prend en charge une partie ou plusieurs des tâches nécessaires pour accomplir l’analyse des données. En effet, une étude approfondie est indispensable pour déterminer les caractéristiques et les fonctionnalités offertes par ces logiciels par rapport aux besoins.

Le tableau 1 récapitule les logiciels les plus importants dans le contrôle qualité et nettoyage des données de séquençage.

**Tableau 1 :** Contrôle qualité et nettoyage des données de séquençage

| logiciel | Fonctionnalités général | Format des fichiers entrés | Format des fichiers sortis | Site web | système | Licence |
|---|---|---|---|---|---|---|
| **FASTQC** | Logiciel permettant de faire un contrôle qualité du séquençage. | sam, bam et fastq. | / | http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ | Windows / linux / Mac | Libre de droit |
| **FASTQ SCREEN** | Permet d'aligner des séquences sur un ensemble de génomes/sequences.<br>FastQscreen Utilise l'aligneur Bowtie ou Bowtie 2. | fastq. | / | http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/ | Linux | Libre de droit |
| **CUTADAPT** | Logiciel permettant de supprimer les séquences des adaptateurs dans des données de séquençage | Fastq,<br>Fasta. | Fastq,<br>Fasta. | https://cutadapt.readthedocs.io/en/stable/ | Linux, MacOS | Libre de droit |
| **TRIMMOMATIC** | -Élimination des séquences d'adaptateurs Illumina.<br>-Retrait de la première base de chaque séquence et des Séquences de petite taille. | Fastq | Fastq | http://www.usadellab.org/cms/?page=trimmomatic | Linux, MacOS, Windows | Libre de droit |

| | | | | | | |
|---|---|---|---|---|---|---|
| | -Retrait des bases de mauvaise qualité. | | | | | |
| **Scythe** | -Eliminer les adaptateurs | Fastq | Fastq | http://github.com/vsbuffalo/scythe/issues | Linux | Libre de droit |
| **Sickle** | -Il supprimera également les lectures en fonction du seuil de longueur.<br>- Il prend les valeurs de qualité et fait glisser une fenêtre à travers elles dont la longueur est 0,1 fois la longueur de la lecture | Fastq | Fastq | | | |

Le tableau 2 récapitule les logiciels les plus importants pour alignement de séquence.

**Tableau 2** : Alignement de séquence

| logiciel | Fonctionnalités général | Format des fichiers entrés | Format des fichiers sortie | Site web | système | Licence |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| **BWA** | Logiciel d'alignement de « short-reads ».<br>Réaliser des Alignements sur des génomes de référence.<br>Réaliser trois types d'alignement :<br>1. BWA-backtrack.<br>2. BWA-SW.<br>3. BWA-MEM. | fastq. | sam | http://bio-bwa.sourceforge.net | Linux, MacOS, Windows | Libre de droit |
| **BOWTIE** | Logiciel d'alignement de « short reads ».<br>Le programme utilise une double-indexation du génome de référence. | fastq. | sam | http://bowtie-bio.sourceforge.net/index.shtml | Linux, MacOS, Windows | Libre de droit |
| **BOWTIE2** | Logiciel d'alignement de reads courts.<br>Capable de réaliser deux types d'alignement :<br>-End-to-end alignment<br>-Local alignment | fastq. | sam | http://bowtie-bio.sourceforge.net/Bowtie 2/index.shtml | Linux, MacOS, Windows | Libre de droit |

| | | | | | | |
|---|---|---|---|---|---|---|
| **NOVOALIGN** | Logiciel d'alignement de reads courts.<br>réaliser des alignements sur des génomes ambigus. | fastq. | / | http://www.novocraft.com/products/novoalign/ | Linux, MacOS, Windows | / |
| **SHRiMP2** | SHRiMP est un progiciel permettant d'aligner les lectures génomiques sur un génome cible. | **fasta** | **sam** | http://compbio.cs.toronto.edu/shrimp | Linux, MacOS. | Libre de droit |

Le tableau 3 récapitule les logiciels les plus importants pour Alignement et analyse de données de Rna-seq.

**Tableau 3 :** Alignement et analyse de données de Rna-seq.

| logiciel | Fonctionnalités général | Format des fichiers entrés | Site web | système | Licence |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **TOPHAT** | Est un logiciel d'alignement de séquence conçu spécialement pour le RNA-Seq.<br>Permet de réaliser des alignements sur les jonctions d'épissage et sur le génome.<br>TopHat utilise Bowtie pour faire l'alignement. | (gff ou gtf) | http://tophat.cbcb.umd.edu/ | Linux et MacOS | Libre de droit |
| **STAR** | STAR est un logiciel d'alignement de reads RNA-seq sur le génome de référence. (très rapidement) | (gff/gtf) | https://github.com/alexdobin/STAR/releases | Linux, MacOS | Libre de droit |
| **CUFFLINKS** | Permet de faire de l'assemblage de transcrits, de<br>mesurer leur abondance (FPKM) et tester si leur expression est<br>différentielle (Cuffdiff). | / | http://cufflinks.cbcb.umd.edu/ | Linux et Mac | Libre de droit |
| **HTSEQ-COUNT**<br>**(HTSEQ)** | -est un script appartenant au logiciel HTSeq.<br>- Il permet, de compter le nombre de reads s'alignant sur chaque élément (gènes, exons, …), à partir d'un fichier de reads alignés (format sam/bam) et d'un fichier<br>d'annotations (format gff/gtf), | / | http://www-huber.embl.de/users/anders/HTSeq/doc/count.html | Windows / Linux / Mac | Libre de droit |

| | | | | | |
|---|---|---|---|---|---|
| **FEATURECOUNTS** | est un programme faisant partie de la suite Subreads.<br>Il permet de compter les reads à partir d'un fichier de reads alignés (format sam/bam) et d'un fichier d'annotations (format gtf). | / | http://subread.sourceforge.net/ | Linux, MacOS | Libre de droit |
| **SAMTOOLS** | Samtools est un utilitaire permettant<br>de manipuler des fichiers au format sam (conversion au format<br>bam, tri, création d'index, statistiques sur l'alignement,<br>nettoyage de potentiels biais de PCR,...). | SAM | http://samtools.sourceforge.net/ | Linux et Mac | Libre de droit |
| **IGV** | Est un outil de visualisation pour l'exploration interactive de grands jeux de données génomiques. | / | http://www.broadinstitute.org/software/igv/ | Windows / Linux / Mac | Libre de droit |
| **EDGER** | Est un package Bioconductor pour des analyses<br>d'expression différentielle à partir de données de RNA-seq ou de DGE (Digital Gene Expression) avec réplicats biologiques. | fichiers contenant des comptages entiers non<br>normalisés. | http://www.bioconductor.org/packages/release/bioc/html/edgeR.html | Windows / Linux / Mac. Requiert l'installation préalable du<br>logiciel R (http://www.r-project.org/ ). | Libre de droit |

| | | | | | |
|---|---|---|---|---|---|
| **DESEQ ET DESEQ2** | Est un package Bioconductor permettant d'estimer la dépendance variance-moyenne dans des données de comptage<br>issues d'expériences de séquençage à haut débit comme le RNASeq.<br>DESeq2 est une évolution de DESeq dans laquelle le test exact est remplacé par un test utilisant le modèle linéaire généralisé. | fichiers contenant des comptages entiers<br>non normalisés | http://www.bioconductor.org/packages/release/bioc/html/DESeq.html<br>http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html | Windows / Linux / Mac. Requiert l'installation préalable du<br>logiciel R (http://www.r-project.org/ ). | Libre de droit |

Le tableau 4 récapitule les Plateforme d'analyse intégrée les plus importants

.

**Tableau 4** : Plateforme d'analyse intégrée.

| Logiciel | Fonctionnalités général | Site web | système | Licence |
|---|---|---|---|---|
| Galaxy | est une plateforme web qui offre un accès gratuit à de nombreux logiciels d'analyse NGS (manipulation de fichiers, alignement de séquence, analyse ChIP-Seq, analyse SNP, analyse RNA-seq...). | https://main.g2.bx.psu.edu/ | Interface web | Libre de droit |
| Mev | est une application qui permet de normaliser, d'analyser, et visualiser des données de puces à ADN. | http://www.tm4.org/mev/ | Windows / Linux / Mac 7 | Libre de droit |

**Tableau 5 :** Autre logiciel.

| logiciel | Fonctionnalités général | Site web | système | Licence |
|---|---|---|---|---|
| DIAGRAMME DE VENN | permet de comparer facilement des listes de gènes et d'obtenir les gènes des intersections. | Http://bioinfogp.cnb.csic.es/tools/venny/index.html<br>Http://bioinfo.genotoul.fr/jvenn/example.html | interface web | / |
| FORMATS DE FICHIERS | Une documentation sur les différents formats de fichier est<br>disponible sur UCSC | http://genome.ucsc.edu/FAQ/FAQformat.html | / | **Libre de droit** |
| OMICS | Site web proposant une classification des outils utilisés dans l'analyse de données « omics », et en particulier les données de séquençage hautdébit | Https://omictools.com/ | / | **Libre de droit** |

Le tableau 6 récapitule les logiciels les plus importants pour Détection de variants

**Tableau 6**: Détection de variants

| logiciel | Fonctionnalités général | Format des fichiers entrés | Site web | système | Licence |
|---|---|---|---|---|---|
| **SAMTOOLS MPILEUP /** | L'outil samtools mpilepup permet de convertir les reads alignés (fichiers BAM) en comptages par position génomique. | | http://www.htslib.org | Linux, MacOS | Libre de droit |
| **BCFTOOLS CALL** | L'outil BCFTools call met ensuite en œuvre une méthode statistique basée sur un modèle bayésien, afin d'identifier des sites variants par rapport à la référence (SNP et indels). | BAM | http://www.htslib.org | Linux, MacOS | Libre de droit |
| **GATK** | outils dont l'objectif premier est la<br>détection de variants Haplotype Caller et le génotype. | SAM , BAM, VCF | https://software.broadinstitute.org/gatk/ | Linux, MacOS | Libre de droit |
| **ANNOVAR** | ANNOVAR permet d'annoter fonctionnellement des variants en fonction de<br>différentes données disponibles dans les bases de données publiques. | BAM | http://annovar.openbioinformatics.org/en/latest/ | Linux, MacOS, Windows. Requiert l'installation de Perl. | Libre de droit |
| **TransVar** | TransVar est un annotateur à plusieurs voies pour les éléments génétiques et les variations | VCF | https://bioinformatics.mdanderson.org/transvar/ | Linux | Libre de droit |

| | | | | | |
|---|---|---|---|---|---|
| | génétiques | | | | |
| **VarSifter** | VarSifter est un programme conçu pour afficher la sortie de variation de séquençage massivement parallèle. Il permet un tri sur n'importe quel champ (ainsi que des combinaisons de champs), et un filtrage sur différents types d'informations (type de variante, héritage, etc.). De plus, il permet un filtrage personnalisé. | VCF | https://github.com/teerjk/VarSifter | Windows XP et XP 64 bits, Mac OS X et les distributions CentOS et Gentoo de GNU / Linux. . | Libre de droit |
| **FreeBayes** | FreeBayes est un détecteur de variantes génétiques conçu pour trouver de petits polymorphismes (SNP, indels, MNP et événements complexes). | BAM , VCF | http://github.com/ekg/freebayes/blob/master/src/bamfilte | Linux. | Libre de droit |
| snpEff | snpEff est un outil d'annotation et de prédiction d'effets. Il annote et prédit les effets des variantes génétiques (telles que les changements d'acides aminés). | VCF | https://github.com/pcingola/SnpEff/issues | Linux | Libre de droit |

Le tableau 7 récapitule les logiciels les plus importants pour Analyse de données de chip-seq.

**Tableau 7** : Analyse de données de chip-seq

| **Logiciel** | **Fonctionnalités général** | | **Site web** | **système** | **Licence** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **Macs** | Est un logiciel pour analyser des données issues d'expériences de ChIP-Seq chez les eukaryotes.<br>Utilisé pour identifier les sites de liaisons de facteurs de Transcription<br>Identifier des régions enrichies en modifications<br>d'histones. | bed, sam<br>ou bam). | MACS (version 1) : http://liulab.dfci.harvard.edu/MACS/<br>MACS2 : https://github.com/taoliu/MACS | Windows / Linux / Mac | Libre de droit |
| **Cisgenome v1.2** | Est un logiciel conçu pour l'analyse de données de ChIP (ChIP-chip, ChIP-Seq).<br>Il permet de visualiser et normaliser<br>les données.<br>D'identifier les régions enrichies (pics).<br>de calculer<br>le false discovery rate (FDR, taux de faux positifs), et de<br>contextualiser les résultats. | aln. | http://www.biostat.jhsph.edu/~hji/cisgenome/ | Windows / Linux / Mac | Libre de droit |
| **IGB** | Integrated Genome Browser: est un outil permettant de visualiser et d'explorer des données génomiques et d'annotation.<br>Il est utile notamment pour la visualisation de données de ChIP-Seq.<br>Il peut ouvrir des fichiers au format BAM et bar, wig… | BAM | http://bioviz.org/igb/ | Windows / Linux / Mac<br>9 | Libre de droit |

Le tableau 8 récapitule les logiciels les plus importants pour analyse des données RRBS

**Tableau 8**: Analyse de données RRbs (Reduced Representation Bisulfite Sequencing).

| Logiciel | Fonctionnalités général | Format des fichiers entrés | Site web | Système | Licence |
|---|---|---|---|---|---|
| **TRIM GALORE** | L'outil TrimGalore! utilise Cutadapt et FastQC afin de couper correctement les reads de type MspI. | | http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ | Linux | / |
| **BISMARK** | permet d'aligner des reads traités au bisulfite sur un génome de référence et de déterminer leur état de méthylation en une seule étape | fastq | http://www.bioinformatics.babraham.ac.uk/projects/bismark/ | Linux | / |
| **SEQMON** | permet de visualiser (et éventuellement analyser) des alignements. | | http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/ | Linux, MacOS, Windows | / |
| **METHYLSIG** | MethylSig est un package R permettant d'analyser des données de bis-seq (*whole-genome bisulfite sequencing*) ou de RRBS. | | http://sartorlab.ccmb.med.umich.edu/node/17 | Linux, MacOS, Windows. Requiert l'installation préalable du logiciel R (http://www.r-project.org/ ). | / |

Le tableau 9 récapitule les logiciels les plus importants Analyse de données de rad-seq

**Tableau 9**: Analyse de données de rad-seq

| logiciel | Fonctionnalités général | Site web | système | Licence |
|---|---|---|---|---|
| **STACKS** | Stacks est un logiciel conçu pour construire des *loci*à partir de reads courts. | http://catchenlab.life.illinois.edu/stacks/ | Linux, MacOS | / |

Le tableau 10 récapitule les logiciels les plus importants pour Manipulation et visualisation de fichiers

**Tableau 10** : Manipulation et visualisation de fichiers

| logiciel | Fonctionnalités général | Format des fichiers entrés | Site web | système | Licence |
|---|---|---|---|---|---|
| **FASTX-TOOLKIT**<br><br>**Fastxbarcode-splitter :**<br><br>**Fastx-trimmer:** | FASTX-Toolkit est un ensemble d'outils en ligne de commande pour manipuler des fichiers au format FastQ ou Fasta.<br><br>Permet de séparer des séquences issues de différents échantillons identifiables grâce à un barcode.<br><br>permet de raccourcir des reads, etc. | FastQ ou<br><br>Fasta | http://hannonlab.cshl.edu/fastx_toolkit/ | Linux, MacOS | Libre de droit |
| **SAMTOOLS** | permettant de manipuler des fichiers au format sam<br><br>(conversion au format bam (binaire correspondant), tri, création d'index,<br><br>statistiques sur l'alignement, nettoyage de potentiels biais de PCR, ...). | SAM | http://www.htslib.org/ | Linux, MacOS | Libre de droit |
| **SAMBAMBA** | permet de paralléliser les tâches, | | http://lomereiter.github.io/sambamba/ | Linux, MacOS | / |

| | | | | | |
|---|---|---|---|---|---|
| **BCFTOOLS** | Bcftools est un ensemble de programmes pour manipuler des fichiers de variants au format vcf ou bcf. | | http://www.htslib.org/ | Linux, MacOS | / |
| **PICARD TOOLS** | fournit un grand nombre de programmes (Java) pour manipuler des fichiers aux formats sam/bam/cram ou vcf permettent d'obtenir des statistiques sur les alignements | sam/bam/cram ou vcf | https://broadinstitute.github.io/picad/ | Linux, MacOS. Requiert l'installation de Python (2.7 pour la version 1, 2.8 pour la version 2) | Libre de droit |
| **IGV** | Integrative Genomics Viewer (IGV) est un outil de visualisation pour l'exploration interactive de grands jeux de données génomiques. permet de visualiser un grand nombre de formats de fichiers : fichiers bam (triés par position et indexés), bed, gff, vcf, P | Bam, bed, gff, vcf, | http://www.broadinstitute.org/software/igv/ | Linux, MacOS, Windows. | / |
| **BEDTOOLS** | Est un ensemble d'outils permettant de travailler sur des intervalles génomiques (intersections, fusion, comptage, …). | bam, bed, gff/gtf, vcf | http://bedtools.readthedocs.io/en/latest/ | Linux, MacOS | / |
| **SRA TOOLKIT** | Sequence Read Archive (SRA) est un service du NCBI permettant de stocker et de mettre à disposition de la communauté des chercheurs les séquences issues de séquençage haut-débit. est un ensemble d'outils permettant le téléchargement, la lecture ou l'écriture de fichiers depuis ou vers le format sra. | sra. | http://www.ncbi.nlm.nih.gov/sra | Linux, MacOS, Windows | / |

Le tableau 11récapitule les logiciels les plus importants pour Clustering et visualisation

**Tableau 11** : Clustering et visualisation

| logiciel | Fonctionnalités général | Site web | système | Licence |
|---|---|---|---|---|
| CLUSTER | Permet de mettre en oeuvre différentes méthodes d'analyses non supervisées.<br>Le logiciel permet d'appliquer divers traitements (centrage médian des gènes, transformation logarithmique..) avant classification.<br>Permet de traiter les données de puces à ADN. | http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm | Linux, MacOS, Windows. | / |
| JAVA TREEVIEW | permettant la visualisation sous forme de carte thermique (« heat map ») des données préalablement classées par le logiciel Cluster. | http://jtreeview.sourceforge.net/ | Linux, MacOS, Windows. | / |

**Tableau 12**:Puces à ADN

| logiciel | Fonctionnalités général | Site web | système | Licence |
|---|---|---|---|---|
| LIMMA | Limma est un package Bioconductor permettant de traiter et d'analyser des données de puces à ADN. | Site web :<br>http://www.bioconductor.org/packages/release/bioc/html/limma.html | Système :<br>Windows / Linux / Mac. Requiert l'installation préalable du<br>logiciel R (http://www.r-project.org/ ). | Libre de droit |
| SAM | Significance Analysis of Microarrays (SAM) est une méthode permettant d'identifier les gènes différentiellement exprimés lors d'une expérience de puce à ADN.<br>Permet d'estimer et de contrôler le taux de faux positifs (FDR). | http://www.bioconductor.org/packages/release/bioc/html/siggenes.html | Windows / Linux / Mac. Requiert l'installation préalable du<br>logiciel R (http://www.r-project.org/ ). | Libre de droit |

| | | | | |
|---|---|---|---|---|
| Cluster | Permet de mettre en oeuvre différentes méthodes d'analyses non supervisées.<br>Le logiciel permet d'appliquer divers traitements (centrage médian des gènes, transformation logarithmique..) avant classification.<br>Permet de traiter les données de puces à ADN. | http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm | Linux, MacOS, Windows. | Libre de droit |
| Java Treeview | permettant la visualisation sous forme de carte thermique (« heatmap ») des données préalablement classées par le logiciel Cluster. | http://jtreeview.sourceforge.net/ | Linux, MacOS, Windows. | Libre de droit |

**Tableau 13** : Plateforme d'analyse intégrée

| logiciel | Fonctionnalités général | Site web | système | Licence |
|---|---|---|---|---|
| **Galaxy** | est une plateforme web qui offre un accès gratuit à de nombreux logiciels d'analyse NGS(manipulation de fichiers,<br>alignement de séquence, analyse ChIP-Seq, analyse SNP, analyse RNA-seq...). | https://main.g2.bx.psu.edu/ | Interface web | Libre de droit |
| **Mev** | est une application qui permet de normaliser, d'analyser, et visualiser des données de puces à ADN. | http://www.tm4.org/mev/ | Windows / Linux / Mac<br>7 | Libre de droit |

Le tableau 14 récapitule les bases de données les plus importants

**Tableau 14:** Base de données des variants

| Base de données | Fonctionnalités | Site web | Lisence |
|---|---|---|---|
| **Db SNP** | Le dbSNP contient des variations de nucléotides uniques humains, des microsatellites et des insertions et suppressions à petite échelle ainsi que la publication, la fréquence de la population, les conséquences moléculaires et des informations de cartographie génomique et RefSeq pour les variations courantes et les mutations cliniques. | https://www.ncbi.nlm.nih.gov/snp/ | Libre de droit |
| **100genomes project** | L'International Genome Sample Resource (IGSR) a été créé pour garantir l'utilisation continue des données générées par le projet 1000 Genomes et pour étendre l'ensemble de données. Plus d'informations sont disponibles sur l'IGSR. | https://www.internationalgenome.org/ | Libre de droit |
| **gnomAD** | La base de données d'agrégation du génome (gnomAD) est une ressource développée par une coalition internationale de chercheurs, dans le but d'agréger et | https://gnomad.broadinstitute.org/ | Libre de droit |

| | | | |
|---|---|---|---|
| | d'harmoniser les données de séquençage des exomes et du génome à partir d'une grande variété de projets de séquençage à grande échelle, et de rendre les données de synthèse disponibles pour le communauté scientifique. | | |
| **ClinVar** | ClinVar est une archive publique et librement accessible de rapports sur les relations entre les variations humaines et les phénotypes hébergée par le National Center for Biotechnology Information (NCBI) et financée par un financement intra-muros des National Institutes of Health (NIH). Les chercheurs de ClinGen travaillent en étroite collaboration avec le NCBI concernant le développement et la fonctionnalité de ClinVar et pour soutenir le dépôt de données provenant de nombreuses sources; Les efforts de conservation de ClinGen amélioreront constamment les données au sein de ClinVar. | https://www.clinicalgenome.org/data-sharing/clinvar/ | Libre de droit |

## 3. ETAT DE L'ART DES PIPELINES EXISTANTS

La littérature contient plusieurs pipelines d'analyse des données issues des séquenceurs NGS. Chacun implémente les étapes du processus global différemment en exploitant des logiciels adéquats. Dans la suite de cette section, on présente les plus importants.

Pipeline1 [18] :

La figure suivante récapitule les étapes du pipeline, alors que, le tableau 14 donne les logiciels utilisés dans chaque étape.

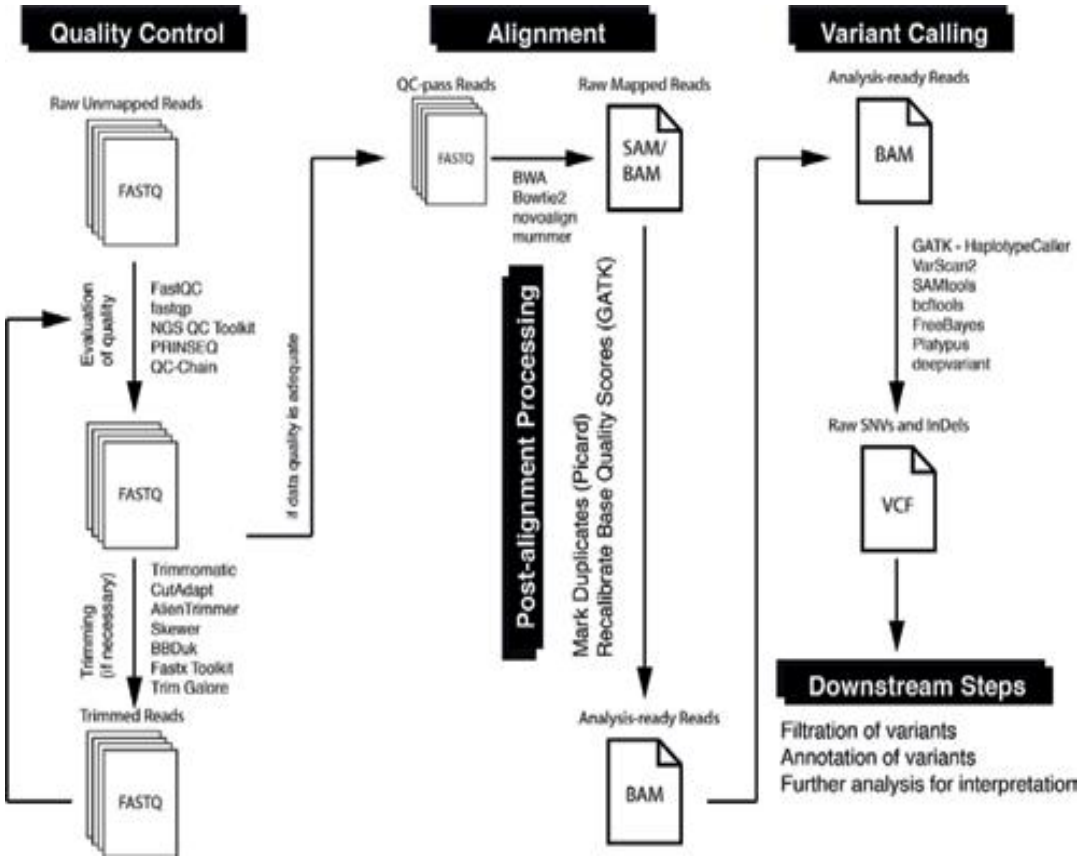



Figure 9: Un exemple de workflow de découverte de variantes à échantillon unique.

**Tableau 15**: Outils utilisé pour effectuer un filtre fonctionnel NGS

| Les étapes | | Outils utilisés |
|---|---|---|
| Contrôle de qualité | Évaluation de la qualité des données FASTQ | FastQC, NGSQC Toolkit, PRINSEQ, QC-chain, Fastqp |
| | Découpage des lectures de mauvaise qualité et retrait des adaptateurs (si nécessaire) | Trimmomatic, cutadapt, fastxToolkit |
| Alignement de séquence | | BWA, Bowtie2, Novoalign, mummer |
| Traitement poste- alignement | Marquage des doublons PCR | Picard, GATK |
| | Réétalonnage du score de qualité de base (BQSR) | |
| Découverte de variantes | | GATK , Varscan2,Samtools, Freebayes |
| Analyses en aval | Filtration des variations génomiques | |
| | Annotation des variations | SIFT, PolyPhen, CADD, VEST |
| | Interprétation / hiérarchisation des variations génomiques | |

Pipeline 2 [19] :

La figure suivante récapitule les étapes du pipeline, alors que, le tableau 15 donne les logiciels utilisés dans chaque étape.

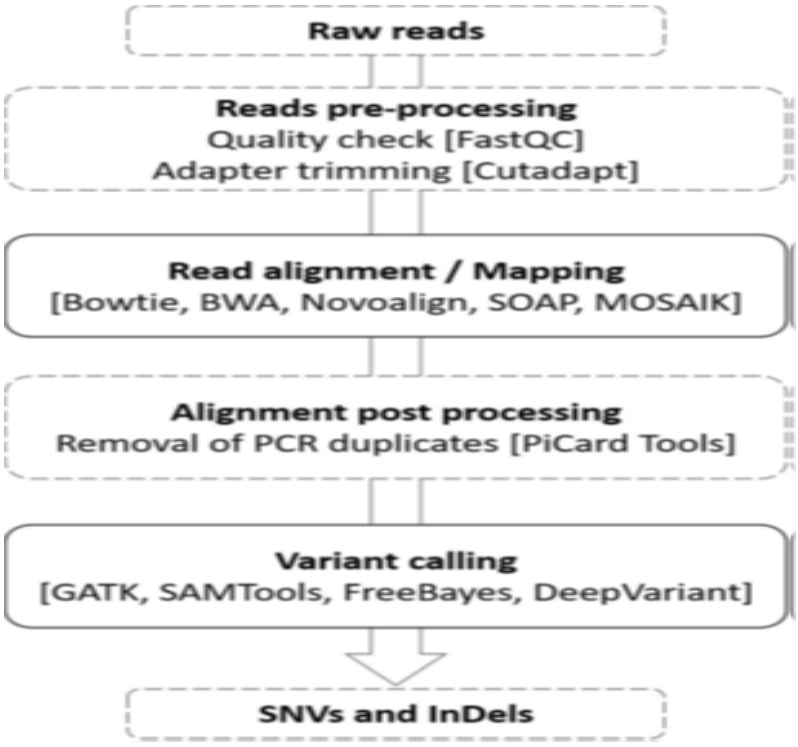



Figure 10 : Schéma du pipeline d'analyse des données NGS [19].

**Tableau 16** : Outils utilisé pour effectuer un filtre fonctionnel NGS

| Les étapes | | Outils utilisé |
|---|---|---|
| Prétraitement | Contrôle qualité | FastQC, |
| | Coup de l'adaptateur | Cutadapt |
| Alignement/mappage | | Bowtie, BWA, NovoAlign, SOAP,MOSAIK |
| Post-traitement d'alignement | | PiCard Tools. |
| Appel des variants | | GATK, SAMTools, FreeBayes, DeepVariant |

Pipeline 3 [86]:

La figure suivante récapitule les étapes du pipeline, alors que, le tableau 16 donne les logiciels utilisés dans chaque étape.

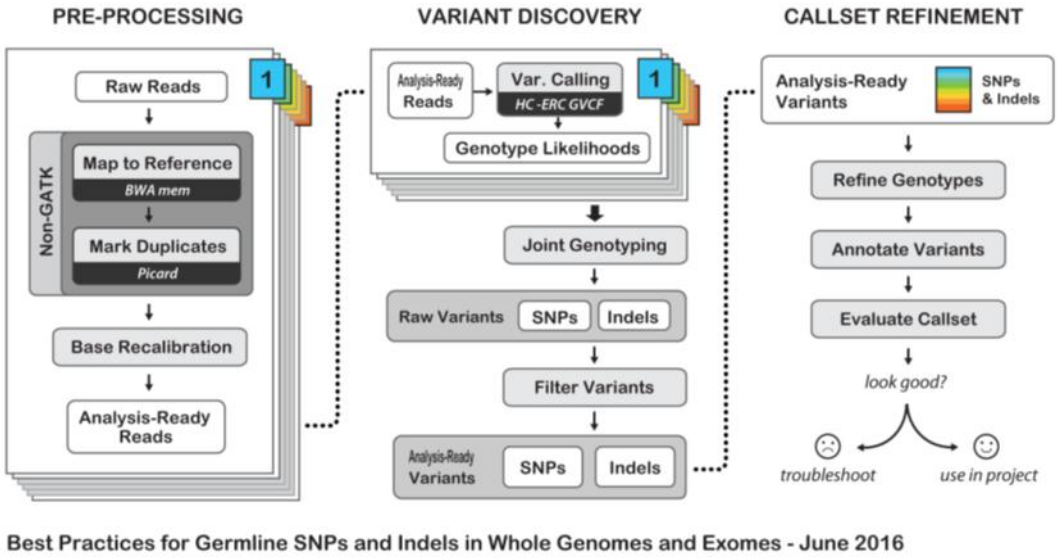



Figure 11 : bonnes pratiques pour les SNP germinales et inde dans le séquençage du génome entier [86].

**Tableau 17** : Outils utilisé pour effectuer un filtre fonctionnel NGS

| Les étapes | | Outils utilisé |
|---|---|---|
| Prétraitement | Alignement | BWA MEM |
| | Marquer les duplicat PCR | Picard |
| | Recalibration des bases | GATK |
| Ddécouverte de variantes | Appel des variants | GATK |
| | Génotypage conjoint | |
| | Filtration des variants | |
| Raffinement de l'appel | Affiner le génotype | SNPEFF |
| | Annotation des variants | |
| | Evaluer callset | |

Pipeline 4[84] :

La figure suivante récapitule les étapes du pipeline, alors que, le tableau 17 donne les logiciels utilisés dans chaque étape.

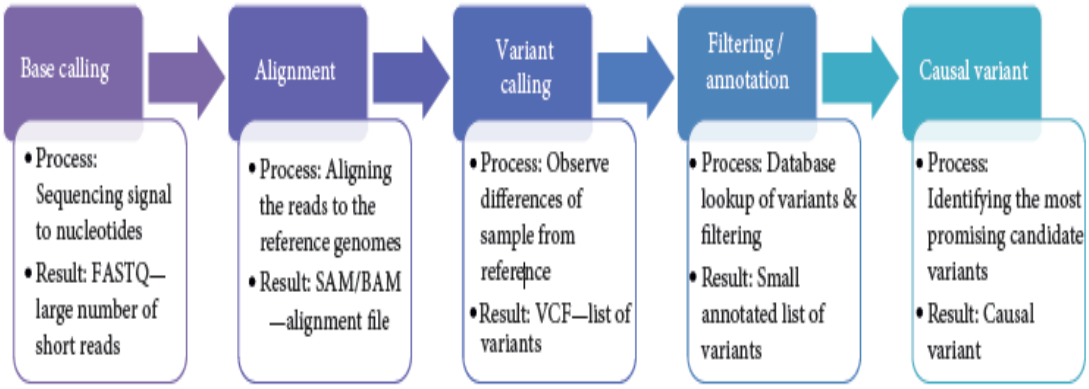



Figure 12 :Flux de travail bioinformatique de séquençage de nouvelle génération[84]

.

**Tableau 18** : Outils utilisé pour effectuer un filtre fonctionnel NGS

| Les étapes | Outils utilisé |
|---|---|
| alignement | BWA, Bowtie, Bwotie2 |
| Appel de varaints | GATK, VarScan |
| Filtration/annotation | SIFT, PolyPhen |

Pipeline 5 [86] :

La figure suivante récapitule les étapes du pipeline, alors que, le tableau 18 donne les logiciels utilisés dans chaque étape.

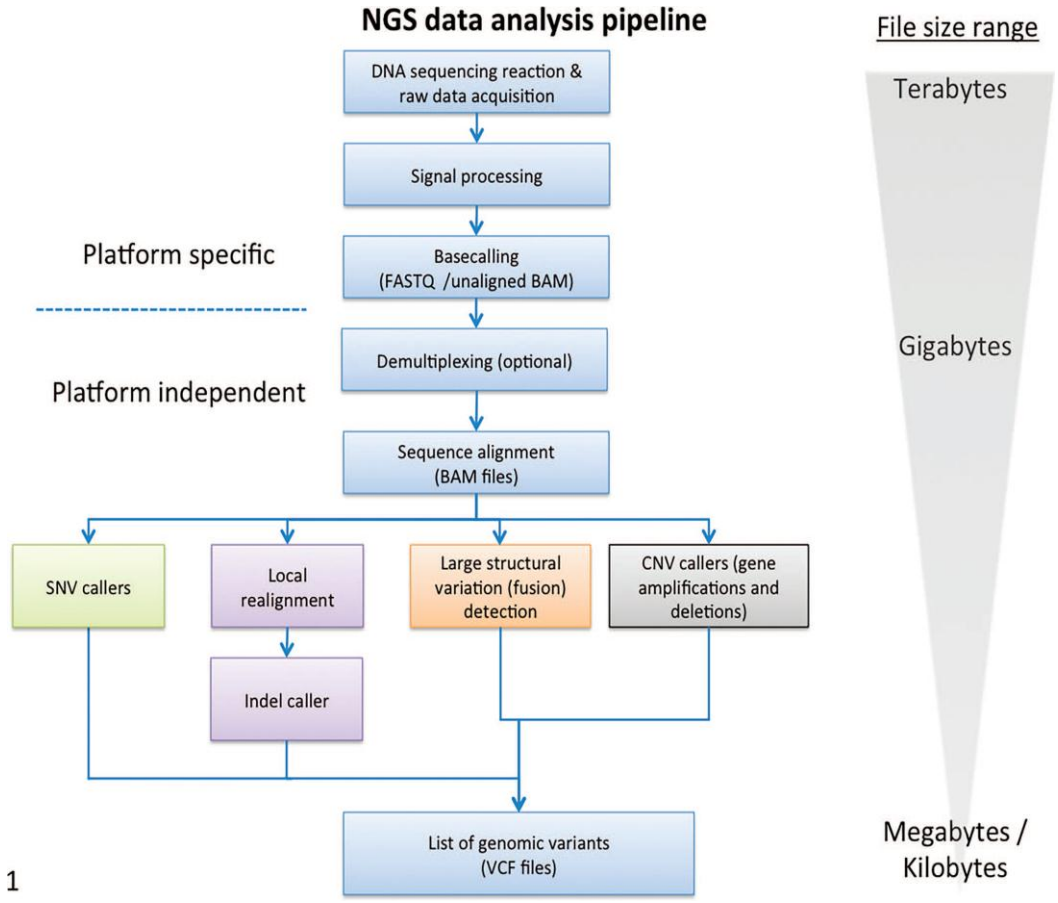



Figure 13 : Schéma général du flux de travail bioinformatique pour les tests de séquençage de nouvelle génération(NGS)[86] ;

**Tableau 19**: Outils utilisé pour effectuer un filtre fonctionnel NGS

| Les étapes | Outils utilisé |
|---|---|
| L'alignement | Bowtie 2 |
| Appel des SNV | Samtools |
| Réalignement local | GATK |
| Détection de grandes variations structurelle | GATK |
| Appel des CNV | GATK |

Pipeline 6 [85]:

La figure suivante récapitule les étapes du pipeline, alors que, le tableau 19 donne les logiciels utilisés dans chaque étape.

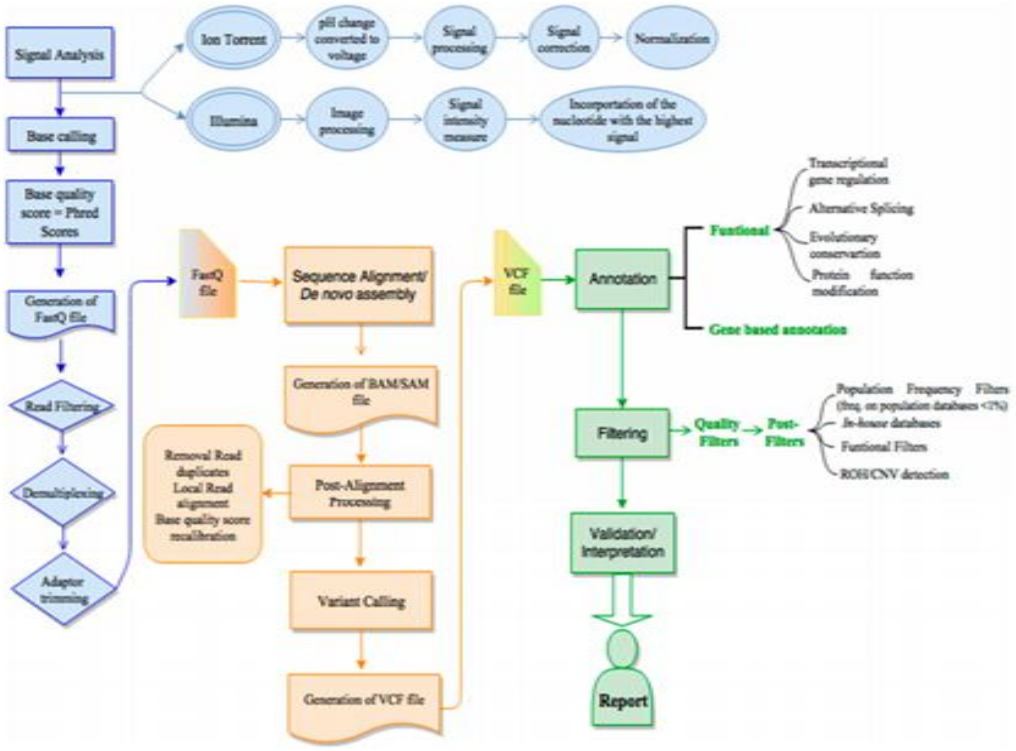



Figure 14: Vue d'ensemble du flux de travail bioinformatique de séquençage de nouvelle génération (NGS) [85].

**Tableau 20** : les outils utilisé pour effectuer un filtre fonctionnel NGS

| Les étapes | | Outils utilisé |
|---|---|---|
| Control qualité | Qualité des scores des bases | FastQC, NGS QC Toolkit |
| | Coupe des adaptateures | Trimmomatic, AdapterRemoval |
| Alignement | | BWA, Bowtie, Novoalign |
| Traitement post alignement | | SAMTools, GATK, Picard |
| Appel des variants | | SAMTools, GATK, Freebayes |
| Annotation des variants | | SIFT, CADD |
| Filtration des variants | | Mutation Taster, PolyPhen |

## 4. CONCLUSION

La vitesse de traitement est très importante, non seulement pour les diagnostics, mais aussi pour l'analyse et le partage des ressources de calcul. Le séquençage de nouvelle génération (NGS) est de plus en plus utilisé comme outil de dépistage génétique dans le diagnostic et réduire le temps écoulé entre la prise d'échantillons et le résultat, l'analyse des résultats des NGS ou de puces à ADN nécessitent le développement d'outils bioinformatiques, Avant toute utilisation de l'un des outils, il est important de se référer à sa documentation et de vérifier qu'il soit adapté au type de données à analyser.

En effet, après cette étude, nous présentons dans le chapitre suivant, un pipeline basé sur des logiciels libres tout en couvant les étapes du pipeline général.

# CHAPITRE 3 :

# Matériels et méthodes

# 1. MATÉRIEL

## 1.1Données biologiques

Deux ensembles de données nucléiques ont été utilisés :

- Le premier ensemble de données avec lequel nous allons travailler provient d'un ensemble de données illumina (Miseq) d'un organisme homosapiens. Comme les données issues de séquençage sont assez volumineuses (en raison des millions de lectures), nous n'utiliserons qu'un sous-ensemble des données d'origines pour les analyser. Cet échantillon a été téléchargé sur la banque de données NCBI sous format fichier fastq.

**Tableau 21**: Données utilisées pour toutes les étapes d'analyse

| Lecture | Spots | Bases | Taille | ID | Publié |
|---|---|---|---|---|---|
| SRR11212874 | 765,119 | 221.2M | 90.9Mb | 10239840 | 2020-03-25 |

- La deuxième donnée c'est le génome de référence qui a été téléchargée sur la banque de données NCBI sous format de fichier fasta.

**Tableau 22** : Données utilisées pour l'étape d'alignement et l'étape d'analyse des variant.

| IDs | description | Taille |
|---|---|---|
| [RefSeq] 2334371<br>[UID] 8687898<br>[GenBank] 8765528 | Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13) | 3.099.706 .404 bp |

## 1.2. Moyens informatique

### 1.2.1 Environnement de travail

- Google Colab

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud sans

donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur. [97].

Pour l'exécution de notre pipeline, Colab nous a réservé un espace de RAM de 12 Gb

- Linux Debian

Debian est une organisation composée uniquement de bénévoles, dont le but est de développer le logiciel libre et de promouvoir les idéaux de la communauté du logiciel libre. Le projet Debian a démarré en 1993, quand Ian Murdock a invité tous les développeurs de logiciels à participer à la création d'une distribution logicielle, complète et cohérente, basée sur le nouveau noyau Linux. Ce petit groupe d'enthousiastes, d'abord subventionné par la Free Software Foundation, et influencé par la philosophie GNU, a grandi pour devenir une organisation composée par environ 1010 développeurs Debian [98].

- Notebook Jupyter

Les notebooks Jupyter sont des cahiers électroniques qui, dans le même document, peuvent rassembler du texte, des images, des formules mathématiques et du code informatique exécutable. Ils sont manipulables interactivement dans un navigateur web.

Initialement développés pour les langages de programmation Julia, Python et R (d'où le nom Jupyter), les notebooks Jupyter supportent près de 40 langages différents.

cellule est l'élément de base d'un notebook Jupyter. Elle peut contenir du texte formaté au format Markdown ou du code informatique qui pourra être exécuté.

### 1.2.2 Software

- SRA Toolkit

La boîte à outils SRA et le kit de développement de système (SDK) SRA code source permette d'accéder par programme aux données hébergées dans SRA et de les convertir du format SRA aux formats suivants:

- ABI SOLiD natif (espace colorimétrique fasta / qual)
- fasta
- fastq
- sff
- sam (bam lisible par l'homme, aligné ou non aligné)
- Originaire d'Illumina

On peut également utiliser la boîte à outils pour convertir les formats listés ci-dessous au format SRA (non requis pour la soumission, mais on permettra d'utiliser la boîte à outils SRA pour archiver ou analyser nos données):

- paires fastq ou fasta / qual
- AB SOLiD-SRF
- AB SOLiD natif
- Illumina SRF
- Originaire d'Illumina
- sff
- Bam aligné

La boîte à outils SRA est disponible dans des versions compatibles avec les systèmes d'exploitation Linux, Windows et Mac [99].

- FastQC

FastQC est un programme conçu pour détecter les problèmes potentiels dans les ensembles de données de séquençage à haut débit. Il exécute un ensemble d'analyses sur un ou plusieurs fichiers de séquences brutes au format fastq ou bam et produit un rapport qui résume les résultats [100].

- Scythe

Scythe utilise une approche bayésienne naïve pour classer les sous-chaînes de contaminants en séquences de lecture. Il prend en compte des informations de qualité, ce qui peut le rendre robuste dans la sélection des adaptateurs 3', qui incluent souvent des bases de mauvaise qualité [102].

- Sckile

Sckile est un outil qui utilise des fenêtres coulissantes avec des seuils de qualité et de longueur pour déterminer quand la qualité est suffisamment basse pour couper l'extrémité 3' des lectures et détermine également quand la qualité est suffisamment élevée pour couper l'extrémité 5' des lectures. Il supprimera également les lectures en fonction du seuil de longueur. Il prend les valeurs de qualité et fait glisser une fenêtre à travers elles dont la longueur est 0,1 fois la longueur de la lecture. Si cette longueur est inférieure à 1, la fenêtre est définie pour être égale à la longueur de la lecture [103].

- Trimmomatic

Trimmomatic est une application java qui fournit des fonctions utiles pour gérer les lectures par paires [29].

- BWA

BWA est un progiciel permettant de cartographier des séquences faiblement divergentes contre un grand génome de référence, tel que le génome humain. Il se compose de trois algorithmes: BWA-backtrack, BWA-SW et BWA-MEM. Le premier algorithme est conçu pour la séquence Illumina lit jusqu'à 100 pb, tandis que les deux autres pour les séquences plus longues variaient de 70 pb à 1 Mbp. BWA-MEM et BWA-SW partagent des fonctionnalités similaires telles que la prise en charge de la lecture longue et l'alignement fractionné, mais BWA-MEM, qui est le dernier, est généralement recommandé pour les requêtes de haute qualité car il est plus rapide et plus précis. BWA-MEM a également de meilleures performances que BWA-backtrack pour les lectures Illumina 70-100bp [104].

- Samtools

Samtools est un ensemble d'utilitaires qui manipulent les alignements au format BAM. Il importe et exporte au format SAM (Sequence Alignment / Map), effectue le tri, la fusion et l'indexation, et permet de récupérer rapidement des lectures dans toutes les régions.

Samtools est conçu pour fonctionner sur un flux. Il considère un fichier d'entrée `- 'comme l'entrée standard (stdin) et un fichier de sortie` -' comme la sortie standard (stdout). Plusieurs commandes peuvent ainsi être combinées avec des tubes Unix. Samtools génère toujours des messages d'avertissement et d'erreur sur la sortie d'erreur standard (stderr).

Samtools est également capable d'ouvrir un fichier BAM (et non SAM) sur un serveur FTP ou HTTP distant si le nom du fichier BAM commence par `ftp: // 'ou` http: //'. Samtools vérifie le répertoire de travail actuel pour le fichier d'index et télécharge l'index en cas d'absence. Samtools ne récupère pas l'intégralité du fichier d'alignement sauf si on lui demande de le faire [105].

- Bcftools

BCFtools est un ensemble d'utilitaires qui manipulent les appels de variantes dans le Variant Call Format (VCF) et son équivalent binaire BCF. Toutes les commandes fonctionnent de manière transparente avec les VCF et les BCF, non compressés et BGZF.

La plupart des commandes acceptent VCF, bgzipped VCF et BCF avec le type de fichier détecté automatiquement même lors du streaming à partir d'un tube. Les VCF et BCF indexés

fonctionneront dans toutes les situations. Les flux VCF et BCF non indexés fonctionneront dans la plupart des situations, mais pas dans toutes. En général, chaque fois que plusieurs VCF sont lus simultanément, ils doivent être indexés et donc également compressés. (Notez que les fichiers avec des noms d'index non standard sont accessibles comme par exemple " bcftoolsview -r X:2928329 file.vcf.gz##idx##non-standard-index-name".)

BCFtools est conçu pour fonctionner sur un flux. Il considère un fichier d'entrée "-" comme l'entrée standard (stdin) et les sorties vers la sortie standard (stdout). Plusieurs commandes peuvent ainsi être combinées avec des tubes Unix [103]

- FreeBayes

FreeBayes est largement utilisé pour appeler des variantes dans les systèmes diploïdes. Cependant, il peut également être utilisé pour appeler des variantes dans des échantillons groupés où le nombre d'échantillons n'est pas connu. C'est le scénario exact que nous avons ici: dans notre échantillon, nous avons plusieurs génomes mitochondriaux (ou bactériens ou viraux), mais nous ne savons pas exactement combien. Ainsi, nous utiliserons l' --pooled-continuous option de FreeBayes pour générer des appels de variantes basés sur la fréquence ainsi que certaines autres options mises en évidence ci-dessous (l'outil est dans NGS: Variant Analysis → FreeBayes) [107].

- VCFTools

VCFtools est un package de programme conçu pour travailler avec des fichiers VCF, tels que ceux générés par le projet 1000 Genomes. Le but de VCFtools est de fournir des méthodes facilement accessibles pour travailler avec des données de variation génétique complexes sous la forme de fichiers VCF.

Cet ensemble d'outils peut être utilisé pour effectuer les opérations suivantes sur les fichiers VCF:

- Filtrer les variantes spécifiques
- Comparer des fichiers
- Résumer les variantes
- Convertir en différents types de fichiers
- Valider et fusionner les fichiers
- Créer des intersections et des sous-ensembles de variantes

VCFtools se compose de deux parties, un module perl et un exécutable binaire. Le module perl est une API Perl générale pour manipuler les fichiers VCF, tandis que l'exécutable binaire fournit des routines d'analyse générales [108].

**Tableau 23**: Caractéristiques des différents outils informatiques utilisés

| Outils / bibliothèques | Versions | Date de publication |
|---|---|---|
| FastQC | FastQC 0.11.9 | 08-01-19: |
| Scythe | Scythe 0.9 | |
| Sickle | Sickle 0.7.0 | 17/5/2020 |
| Trimmomatic | Trimmomatic 0.39 | |
| BWA | BWA 0.7.17 | 23-10- 2017 |
| Samtools | Samtools 1.11 | / |
| Bcf Tools | Bcf Tools 1.10.2 | 2020-05-28 |
| FreeBayes | FreeBayes 1.3.2 | 02-09-2020 |
| SnpEff | SnpEff 5.0 | 09- 08- 2020 |
| VCFTools | VCFTools 0.1.13 | |
| Varsifter | | / |

## 2. MÉTHODES

Cette partie décrit le pipeline développé pour l'analyse de données issues du NGS.

### 2.1 Aperçu global du pipeline développé

La quantité de donnés générées par le séquençage haut-débit a rendu l'analyse des résultats manuelle impossible. Il est nécessaire de développer des outils bioinformatiques dédiés et adaptés.

Le traitement des données se fait en une succession d'étapes, du traitement du signal de base lors du séquençage, à l'annotation finale des variants retrouvés. Certaines étapes sont réalisées directement par l'automate de séquençage, quand d'autres peuvent être réalisées de manière manuel

ou semi automatisée grâce à l'utilisation de scripts informatiques ou des logiciels libres de droit ou commerciaux.

Les principales étapes de l'analyse bioinformatique du pipeline proposé se résume dans la figure suivante :

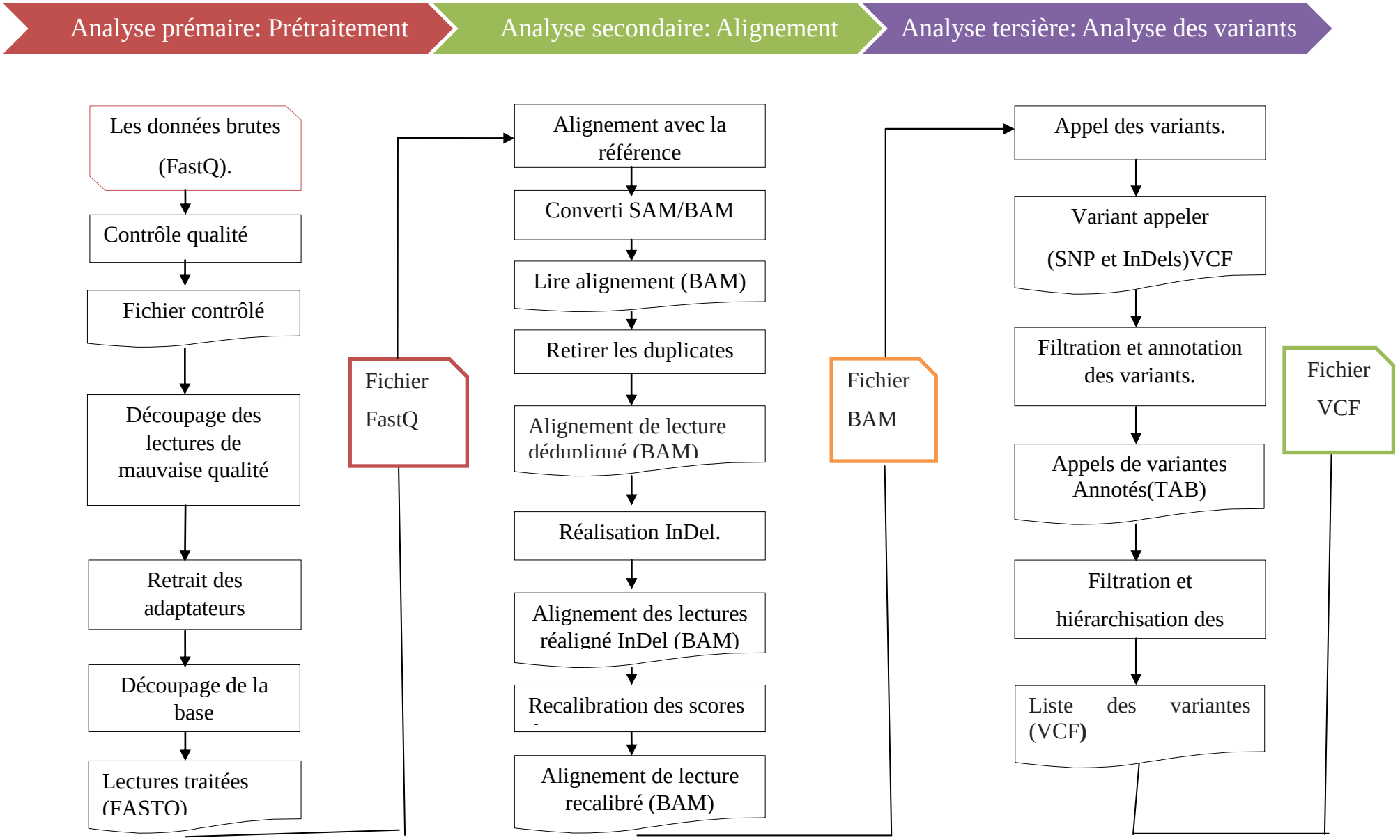



Figure 15 : pipeline bioinformatique d'analyse des données NGS

**Tableau 24:** logiciels utilisés selon les étapes du processus.

<table>
<tr><th colspan="3">Etapes de processus</th><th>outils</th><th>Format de fichier produit</th></tr>
<tr><td>Prétraitement</td><td colspan="2">Control qualité</td><td>FastQC</td><td rowspan="2">Format FastQ</td></tr>
<tr><td></td><td colspan="2">Découpage des bases de mauvaise qualité<br>Elimination des adaptateurs</td><td>Scythe, sickle<br>Trimmomatic</td></tr>
<tr><td rowspan="4">Alignement</td><td colspan="2">Alignement avec la référence</td><td>BWA</td><td rowspan="4">Format BAM</td></tr>
<tr><td colspan="2">Transformation SAM en BAM</td><td rowspan="3">Samtools<br>Samtools</td></tr>
<tr><td rowspan="2">Traitement<br>poste<br>alignement</td><td>Retirer les duplicat PCR</td></tr>
<tr><td>Réalisation Indel</td></tr>
<tr><td></td><td colspan="2">Recalibration des scores de base</td></tr>
<tr><td>Analyse des variants</td><td colspan="2">Appel des variants</td><td>BcfTools</td><td rowspan="3">Format VCF</td></tr>
<tr><td></td><td colspan="2">Filtrage et annotation des variants</td><td>SnpEff</td></tr>
<tr><td></td><td colspan="2">Filtration et héirachisation</td><td>varstifer</td></tr>
</table>

2.2 Description détaillée du pipeline d'analyse bioinformatique développé

Par rapport au pipeline général, celui proposé intègre le contrôle de qualité dans l'étape de prétraitement. Dans la suite, nous détaillons chaque étape :

1) Prétraitement

Cette étape inclut deux sous étapes qui sont :

-L'évaluation de la qualité des données FASTQ.

Pour réaliser l'évaluation de la qualité des données, nous avons exploité FastQC. Ce dernier est une application Java qui lit un ensemble de fichiers de séquence et produit à partir de chacun d'eux un rapport de contrôle de la qualité composé d'un certain nombre de modules différents [87], et génère de nombreux diagnostics de données utiles et des graphiques tels que la distribution du contenu GC, la distribution du score Phred le long des lectures, la distribution de la longueur de lecture et le niveau de duplication de séquence. Il détecte également les séquences sur représentées *qui peuvent être une indication de contamination de l'amorce ou de l'adaptateur. Grâce à un rapport de contrôle qualité des lectures brutes complet généré par FastQC [20].

Grâce à ces avantages on a choisir l'outil fastqc pour cette étape.

Concernant les tâches de découpage des lectures de mauvaise qualité et retrait des adaptateurs et de découpage de la base, le filtrage de lecture ou le découpage de l'adaptateur, nous exploité trois outils qui sont : Scythe, Sickle et Trimmomatic [29] [80]

Ces outils permettent le prétraitement générique énuméré ci-dessus En plus, chaque outil est équipé de ses propres fonctionnalités personnalisées.

2) Alignement

Après le prétraitement des données brutes, l'étape suivante d'alignement est composée de deux parties :

i. Alignement avec le génome de référence : Cette partie consiste à mapper les lectures sur le génome de référence et avec une efficacité et une précision élevées.

De nombreux outils différents ont été développés pour la cartographie des lectures courtes. Bowtie2 [30] et BWA [31] sont deux outils d'alignement de courtes lectures bien connus qui implémentent l'algorithme BWT. MOSAIK [32] SHRiMP2 [33] et Novoalign [87] sont des implémentations d'algorithmes SW avec une précision d'alignement accrue.

Afin de trouver la correspondance d'alignement optimale dans un temps de calcul acceptable nous avons utilisé l'outil BWA.

BWA est capable de réaliser trois types d'alignement :

BWA-backtrack : pour reads illumina < 100pb (plutôt conseillé pour les reads<70pb car BWA-MEM est plus performant pour les reads de 70 à 100bp).

BWA-SW : pour reads de 70bp à 1Mbp (non conseillé).

BWA-MEM : pour reads de 70bp à 1Mbp ; plus rapide et plus précis que BWA-SW car plus récent; meilleures performances que BWA-backtrack pour les reads Illumina de 70 à 100pb [87].

ii. Traitement Poste alignement

Une procédure de traitement post-alignement est divisée en trois étapes :

Elle comprend la suppression des doublons en lecture, le réalignement des Indel et le recalibrage du score de qualité de base (BQSR).

- Suppression des doublons en lecture

Lors de la préparation des échantillons pour le séquençage, des doublons de PCR apparaissent à l'étape d'amplification par PCR des fragments. Puisqu'ils partagent la même séquence et la même position d'alignement, ils peuvent conduire à des problèmes de détection de variantes. Par exemple, pendant l'appel SNV, des variantes faussement positives peuvent survenir car certains allèles peuvent être surreprésentés en raison des biais d'amplification. Pour surmonter ce problème, PCR duplique sont marqués d'une certaine balise à l'aide d'un algorithme (Mark Duplicates) disponible dans l'outil Picard [88]. SAMtools [31].

- Le réalignement indel

Après l'élimination des doublons, la deuxième étape consiste à identifier les régions génomiques contenant des indels et à améliorer la qualité de l'alignement dans la région cible. Deux algorithmes ont été développés pour accomplir cette tâche:

Le réalignement local des lectures espacées sur le génome de référence ou d'autres haplotypes candidats; L'assemblage local de novo des lectures alignées autour de la région cible suivie de la construction d'une séquence consensus pour la découverte indel.

Les programmes qui implémentent un algorithme ou un mélange des deux incluent SRMA [33] et IndelRealigner du GenomeAnalysisToolkit (GATK) [35].

- Le recalibrage du score de qualité de base (BQSR).

Le recalibrage du score de qualité de base (BQSR) est un apprentissage automatique. Modélise ces erreurs de manière empirique et réajuste les scores de qualité de base en conséquence. Grâce à ce recalibrage, une qualité de base plus précise et plus fiable les scores sont obtenus, ce qui améliore la fiabilité des étapes en aval dans d'autres analyses [18].

Après ces opérations de traitement des données post-alignement, un BAM prêt pour l'analyse fichier est obtenu.

L'outil le plus largement utilisé pour BQSR est fourni par le Génome Boîte à outils d'analyse (GATK) [13], NGSUtils, [40] qui fournit des fonctions similaires à celles de GATK, et BaseRecalibrator [35], et le package BioconductorReQON [41] qui utilise la régression logistique pour le recalibrage des scores de qualité de base, ReQON produit un ensemble de données de diagnostic et de tracés avant et après le recalibrage pour illustrer l'amélioration de la précision [20].

Comme BWA-MEM permet de faire l'alignement avec le génome de référence et le post alignement, nous avons opté pour l'utiliser dans les deux traitements. Pour l'indexation des données elle est réalisée par BWA.

3) Analyse des variants

Cette étape est subdivisée aux trois sous étapes suivantes :

i. Appel des variants

Dans cette sous étape nous utilisons l'outil bcftools pour identifier les variants.

ii. Filtration et annotation des variants

Après l'identification des variants, l'étape suivante consiste à déterminer lesquels de ces variants sont susceptibles de contribuer au processus pathologique d'étude. Cette sous étape combine deux processus, à savoir le filtrage et l'annotation [89].

- Le processus de filtrage supprime les variantes qui conviennent à des modèles génétiques spécifiques ou qui ne sont pas présents dans les tissus normaux.

Les outils utilisé pour cette étape sont SIFT [52], PolyPhen-2 [53], LRT [54], MutationTaster [55], MutationAssessor [56], FATHMM [57], GERP ++ [58], PhyloP [59], SiPhy [60], PANTHER-PSEP [61], CONDEL [62], CADD [63], CHASM [64],CanDrA [65] et VEST [66] et FreeBayes )[107] qui est utilisé dans notre pipeline.

-Le processus d'annotation est utilisé pour rechercher les informations sur les variants et identifier les variants adaptés au processus biologique.

Les variants génomiques peuvent alors être annotés à l'aide de divers outils. Les outils d'annotation fonctionnels les plus utilisés incluent, mais sans s'y limiter AnnoVar [70], SnpEff [71], Variant Effect Predictor (VEP) [72], GEMINI [73],VarAFT [74], VAAST [75], TransVar [76], MAGI [77], SNPnexus [78] et VarMatch [79].

Notre pipeline utilise SnpEff.

iii. Filtration et hiérarchisation des variants

L'étape suivante consiste à filtrer et à hiérarchiser les variants. Cette étape est utilisée pour hiérarchiser les variants par rapport à la maladie [89].

Différents types de variants génomiques, y compris les SNV, les indels, les CNV et les grands SV peuvent être détectsé à partir de l'échantillon en comparant les lectures alignées au génome de référence.

Les outils couramment utilisés pour la filtration des variants sont VAAST2, CADD, VarSifter, KGGseq, PLINK/SEQ, SPRING, Gnome [20].

Notre pipeline utilise VarSifter.

# CHAPITRE 4

# Résultats et discussions

# 1. RÉSULTATS

Dans ce chapitre, nous présentons les résultats d'implémentation du pipeline proposé. Cette présentation suit les étapes du processus. Elle commence par le traitement des données brutes à l'aide des outils FastQC, Scythe, Sickle, Trimmomatic, Ensuite, on passe à l'alignement des reads avec le génome de référence via l'outil BWA. La dernière partie consiste a détecter et analysé les variants.

1) Résultats du prétraitement

Après l'exécution de chaque outil, on lance FastQC pour récupérer un rapport détaillé sur nos données. La suite de cette section présente quelques éléments du rapport final de l'étape de prétraitement.

- Qualité par base

Ce graphique représente les scores qualités des bases en fonction de leur position dans les lectures.

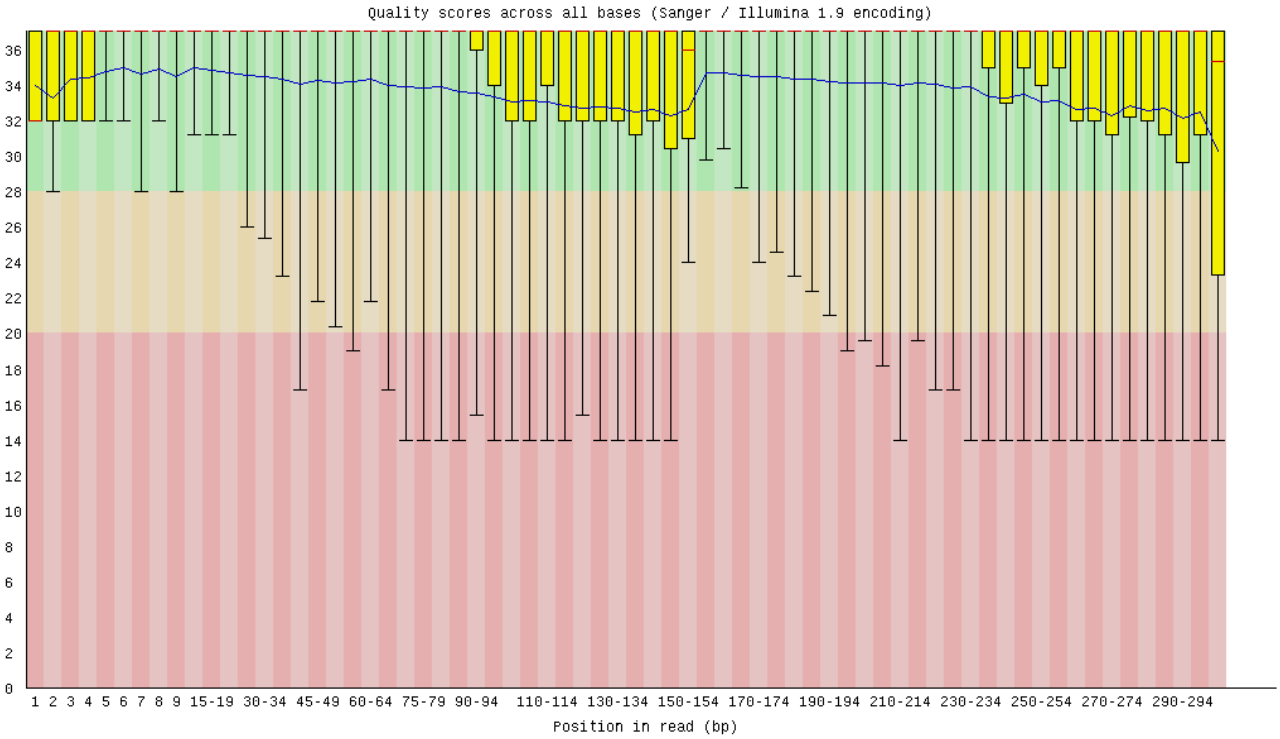



Figure 16 : Qualité par base

- Qualité par séquence :

Le graphique suivant correspond à la distribution de la qualité moyenne des lectures. Il est préférable que cette courbe soit centrée sur un score Phred le plus grand possible.

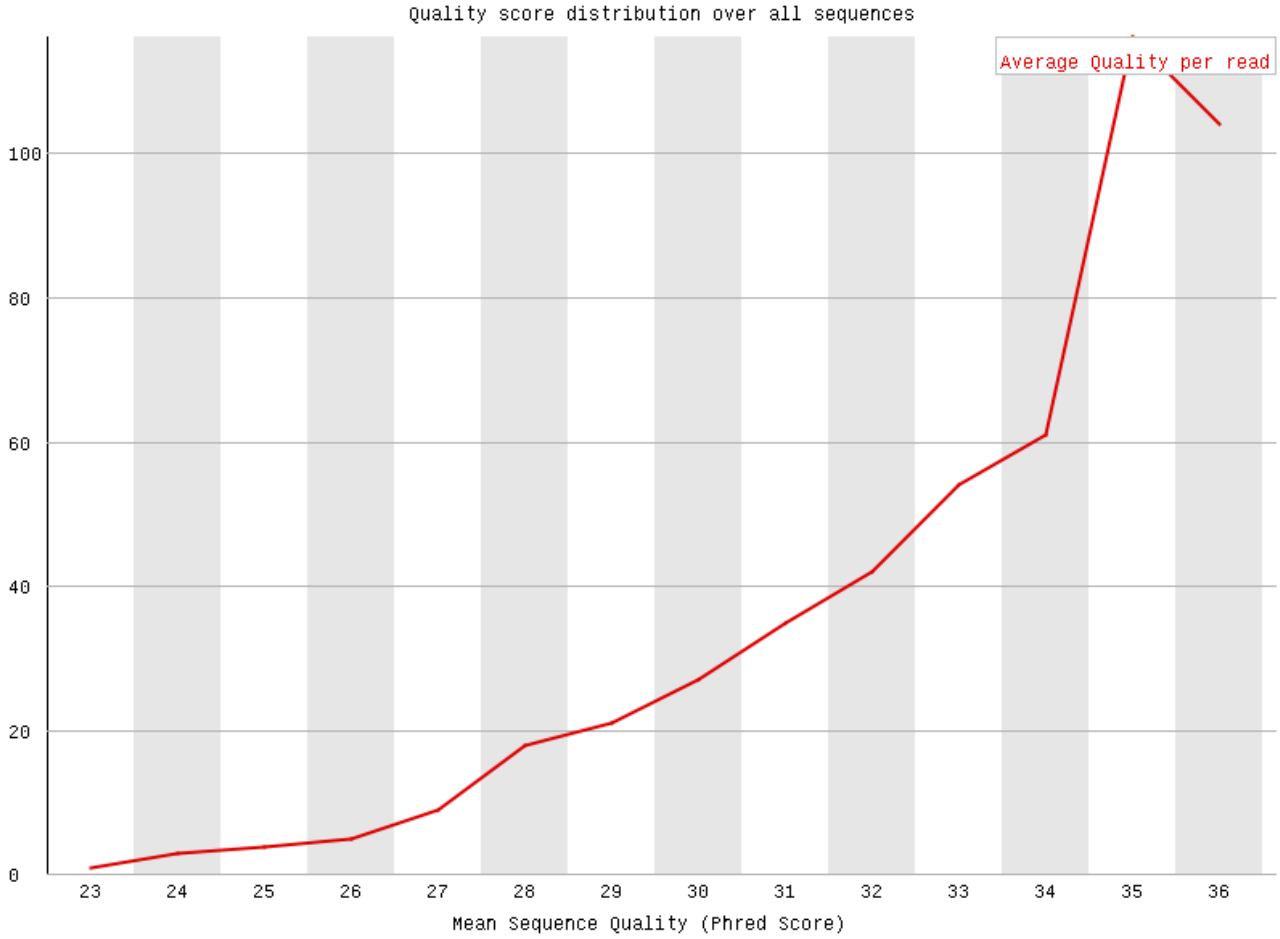



Figure 17:Distribution de la qualité moyenne des lectures.

- Distribution des bases:

Le graphique suiavnt représente la distribution moyenne des nucléotides selon leur position dans la lecture. Dépendante des séquences analysées, cette distribution doit être fixe entre les différentes séries d'un même panel

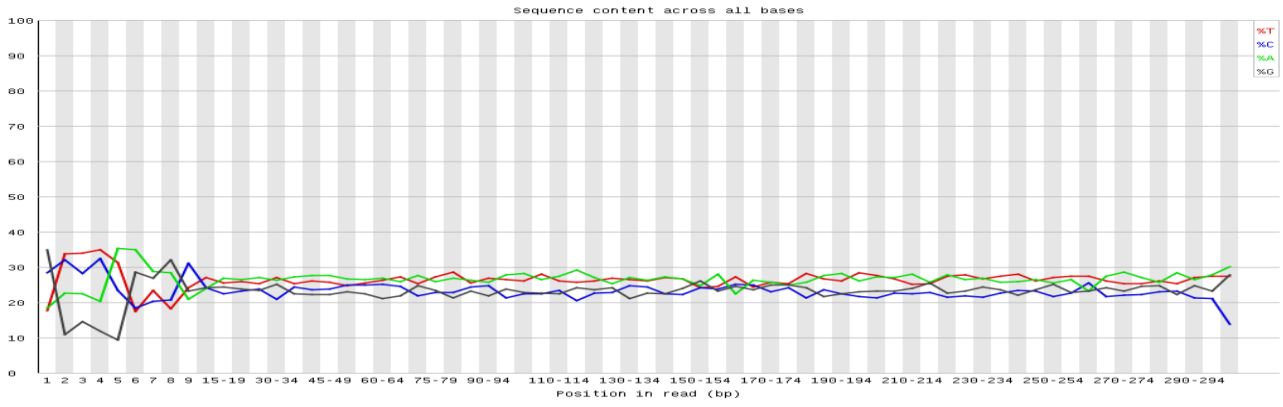



Figure 18: Distribution moyenne des bases à travers les lectures

- Contenu GC par séquence :

Le pourcentage de GC dans un amplicon influence la qualité et la quantité des lectures produites lors de l'étape de séquençage. Le graphique ci-dessous représente la distribution moyenne en GC par séquence et GC par base. Les taux doivent être constants.

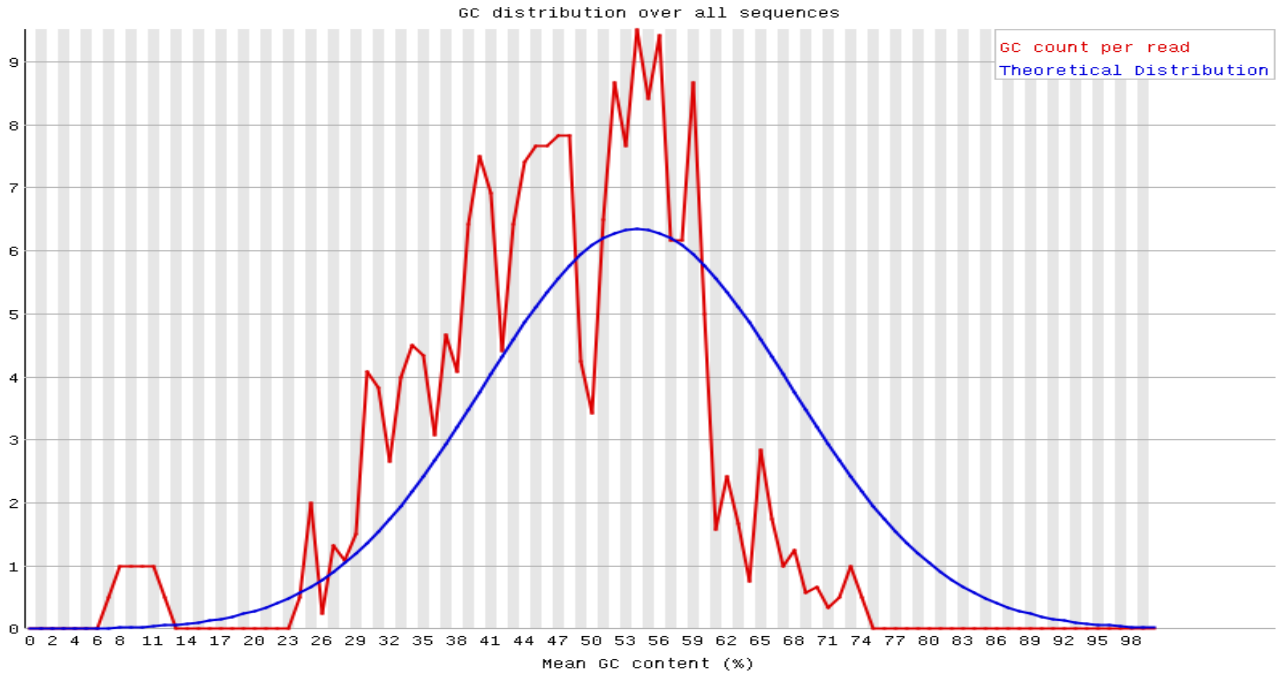



Figure 19:Distribution taux de GC par lecture.

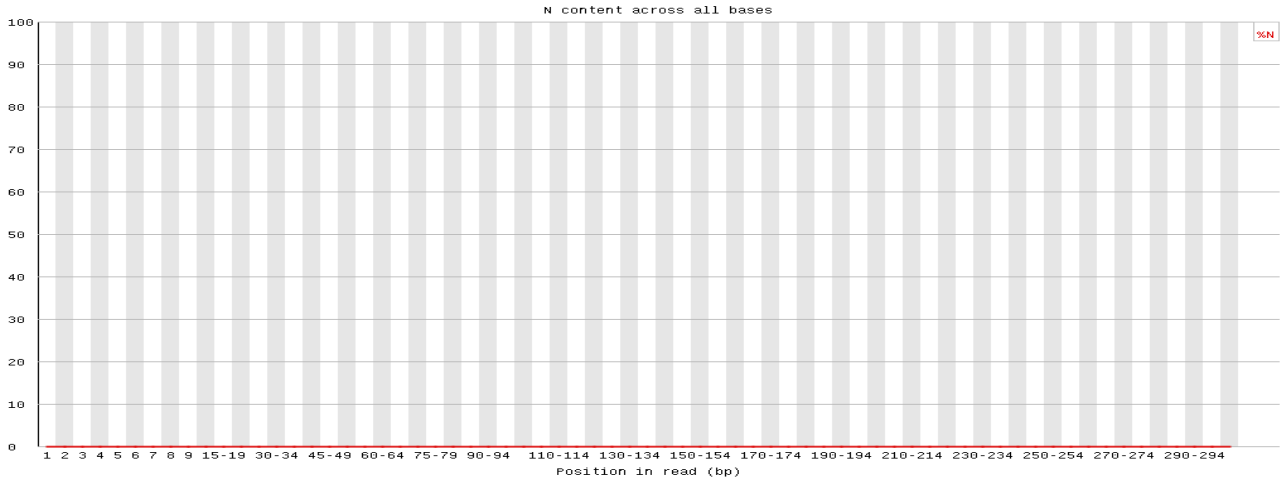



Figure 20 :Distribution taux de GC par base

- Distribution de longueur de séquence:

Ce graphique représente la distribution des longueurs de lecture. Si la présence de dimère d'amorce important on apercevrait un pic de faible taille ( signification mauvaise purification).

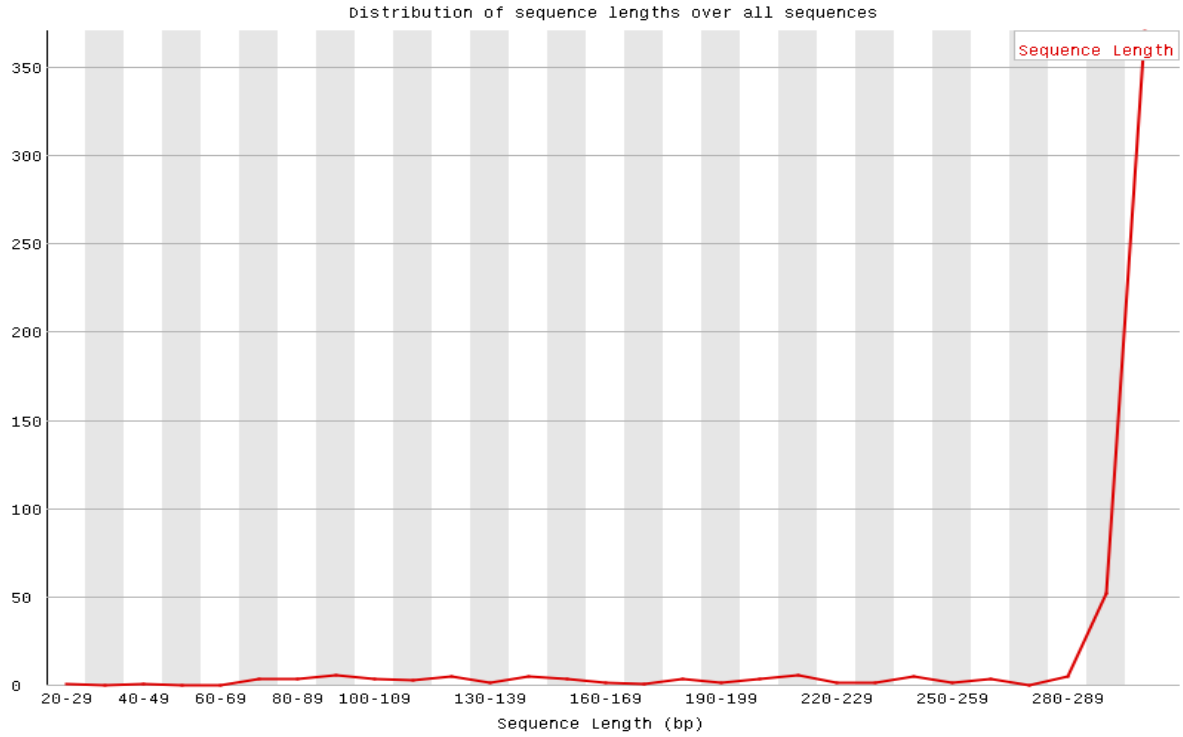



Figure 21:Distribution des longueurs des lectures.

- Niveau de duplication de séquence :

Le graphique suivant montre la proportion de la bebiliothèque qui est composée de séquences dans chacun des deffirents niveaux de duplication. Il ya deux ligne. La ligne bleu prend le jeu de séquences complet et montre comment ses niveaux de duplications sont repartis. Dans le graphique rouge les séquence sont dédupliquées et les proportions indiquées sont les proportions de l'ensemble dédupliqué qui provient des défférents niveaux de duplication dans les données d'origine.

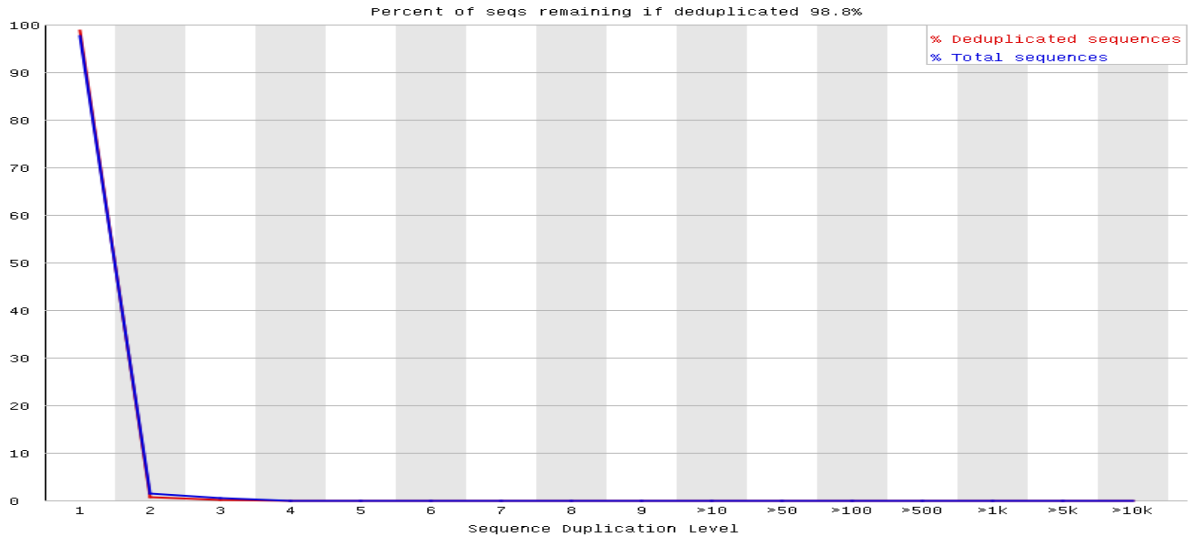
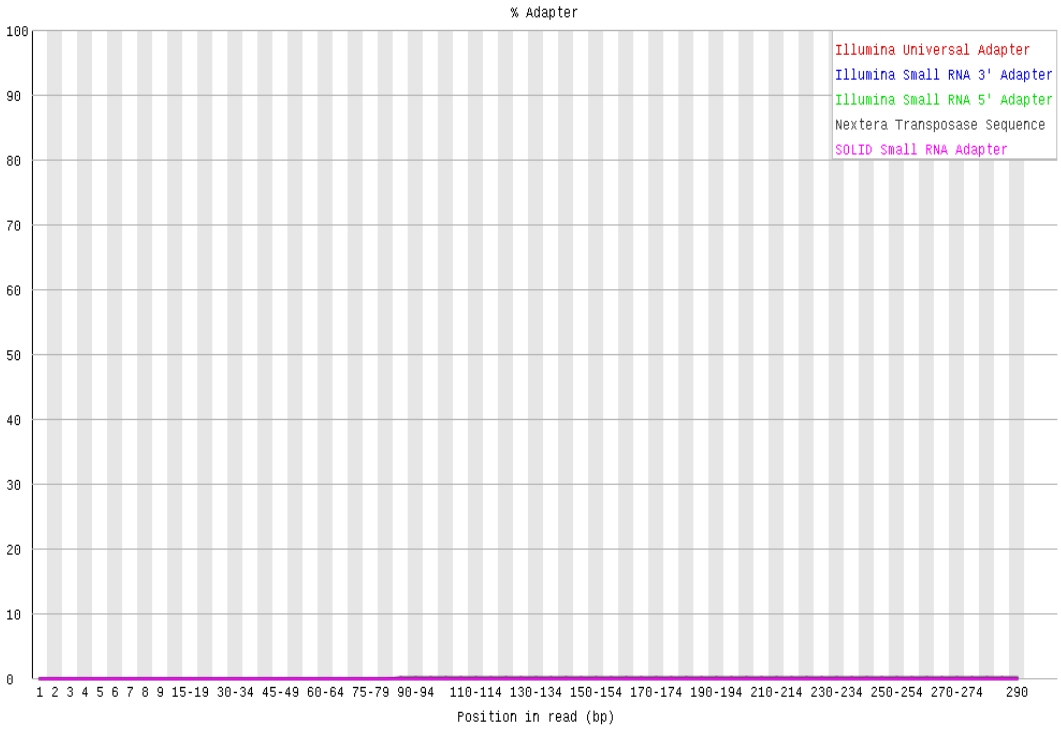



Figure 22:Niveau de duplication de séquence

- Contenu de l'adaptateur :

  La figure suivante montre qu'il n'y a plus d'adapeteurs.




-

Figure 23: Contenu de l'adaptateur

- Contenu Kmer :

Le module Kmer content effectuera une analyse générique de tous les Kmer de notre bébliothèque pour trouver ceux qui n'ont pas une couverture uniforme sur la longueur de notre lecture. Cela peut trouver un certain nombre de source de biais différentes dans la bébliothèque

qui peuvent inclure la présence de séquence d'addaptateur de lecture se développant a la fin de nos séquences.

## 2) Résultats de l'alignement

L'alignement est fait directement à la suite du prétraitement. Il réalisé sur le génome homosapiens (GRCh38) via l'outil BWA-Mem qui génère comme résultat un fichier BAM. La figure 22 donne un aperçu simplifié de ce dernier.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SRR11212874.835 | 16 | CM000663.2 | 171165450 | 17 | 44M | * | 0 | 0 | GGTGCCCGCCCCAGGGTAATGAGTGAGTTCTCACTCTATTAGTG F |
| SRR11212874.105 | 0 | CM000663.2 | 171249699 | 0 | 33M6S | * | 0 | 0 | GTGTATGTGTGTGTGTGAGAGAGAGAGACAGAGAGGAAG 6AAFAFFFF |
| SRR11212874.601 | 0 | CM000668.2 | 29889605 | 0 | 38M | * | 0 | 0 | GTAAGGAGGGAGATGGGGGTGTCATGTCTCTTAGGGTA =AAA//FFF |
| SRR11212874.437 | 0 | CM000668.2 | 29928686 | 0 | 92S86M5S | * | 0 | 0 | GCCATAATCAAGGTGATAAATCTGTCCTTCATTGTCACAGG |
| SRR11212874.665 | 272 | CM000668.2 | 29943814 | 0 | 151H150M | * | 0 | 0 | CAGCAGCCTTGGGACCGTGACTTTTCCTCTCAGGCCTTGTT |
| SRR11212874.234 | 0 | CM000668.2 | 29943882 | 5 | 151M149S | * | 0 | 0 | GGTCTGAGTCCAGCACTTCTGAGTCCCTCAGCCTCCACTCA |
| SRR11212874.643 | 272 | CM000668.2 | 29943929 | 0 | 148M151H | * | 0 | 0 | GACCAGAAGTCGCTGTTCCCTCCTCAGGGAATAGAAGATTA |
| SRR11212874.110 | 0 | CM000668.2 | 29943957 | 0 | 151S148M | * | 0 | 0 | GAGACAGCGTGGTGGGTCATATGTGTCTTGGGGGGGTCTGA |
| SRR11212874.638 | 272 | CM000668.2 | 29943971 | 0 | 147H151M | * | 0 | 0 | CCCAGGTGCCTGTGTCCAGGCTGGTGTCTGGGTTCTGTGCT |
| SRR11212874.668 | 0 | CM000668.2 | 29944012 | 0 | 71S71M | * | 0 | 0 | GTCATGGGACACTCCACCAGCATGCATGTGGCCATCTTGAGAATGGACA |

Figure 24 Schéma d'alignement sur un génome de référence.

## 3) Résultats de l'analyse des variants

A la fin des trois sous étapes d'analyse des vatriants, l'outil varSifter génère le fichier VCF et le représente sous forme de tableau qui permet de consulter les variants avec leurs caractéristiques (figure25).

À partir des deux figures, nous remarquons qu'il y a que deux types de variants dans les données utilisées, à savoir : InDel et SNP.

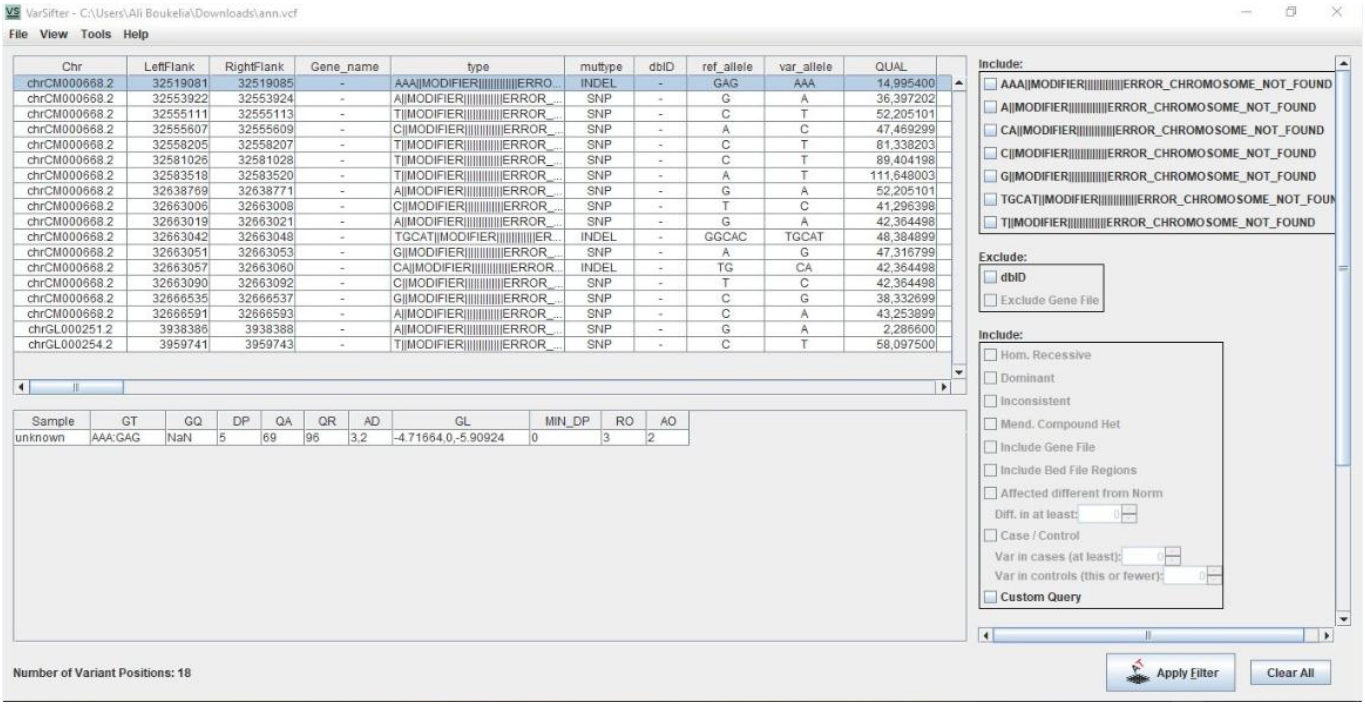



Figure 25: Affichage des variants avec Vastsifter.

## 2. DISCUSSION

L'analyse des données de séquençage de nouvelle génération est difficile car les ensembles de données sont volumineux. Les plates-formes de séquençage de deuxième génération ont des taux d'erreur élevés parce que chaque position dans le génome cible (exome, transcriptome, etc.) est séquencée plusieurs fois. Face à ces défis, Un pipeline bioinformatique correspond à une chaine de traitement informatisée de données biologiques (séquences, variants) est nécessaire.

Après les études qu'on a fait sur les pipelines qui existent on déduit que les étapes d'analyse bioinformatique sont sensiblement les mêmes, alors que le choix des outils à chacune de ces étapes diffère selon les besoins. En effet, notre pipeline intègre de nombreux logiciels d'analyses performants et libres de droit tout en ayant la capacité à détecter avec précision les SNV et InDels et CNV et SV.

Le travail accompli et les résultats obtenus dans l'étape de prétraitement prouvent que la combinaison de trois logiciels qui sont scythe, trimmomatic et sickle joue un rôle efficace dans l'amélioration de la qualité des données issue du NGS. Ainsi, les logiciels utilisés dans les deux dernières étapes permettent aussi de faire une annotation précise des variants pour donner à la fin un fichier VCF clair et compréhensible.

Donc, ce pipeline permet de dresser une synthèse des variants qui se trouvent dans les données séquencées en permettant ainsi une meilleur lecture et interprétation pour les professionnels de santé.

# Conclusion

# CONCLUSION

La réduction sans précédent du coût du séquençage à haut débit a permis de mener des études plus avancées sur les maladies humaines. Les défis du NGS sont passés de la production de données de séquençage à la gestion, l'analyse et la synthèse de ces données. Dans le cadre de ce travail, nous sommes intéressés à l'identification des variants qui se trouvent dans les données séquencées.

Une étude approfondie des outils existants nous a permis de dresser un panorama regroupant ceux les plus utilisés avec leurs caractéristiques. Ainsi, l'état de l'art sur les pipelines de détection des variants à partir des données séquencées, nous a permis de couvrir les tâches nécessaires pour arriver à notre objectif.

En effet, le pipeline développé dans le cadre de ce travail consiste à analyser les données NGS à partir d'un fichier fastq généré par un séquenceur jusqu'à à l'analyse des variants. Nous avons appliqué les outils de bioinformatique selectionnés pour l'analyse complète des données, y compris le prétraitement des données, l'alignement, le traitement post-alignement, l'appel de variantes, l'annotation et filtrations.

Comme perspective à ce travail, nous envisageons **:**

- Comparaison des résultats sortants de notre pipeline avec ceux des d'autres pipelines.
-Utilisé des données réelles issus du NGS en Algérie dans le processus.
-Etudier le temps de réponse des outils utilisées.
-Développer un Workflow basé sur des Apis pour automatiser le passage des données d'une étape à une autre.

# Références

1. Chadi S. Caractérisation des erreurs de séquençage non aléatoires, application aux mosaïques et tumeurs hétérogènes. Bio-informatique [q-bio.QM]. Université de Lille Nord de France, 2018 ,193p.

2. Thomas K. Bioinformatique et infertilité : analyse des données de séquençage haut-débit et caractérisation moléculaire du gène DPY19L2. Génétique humaine. Université Grenoble Français. NNT : 2017,323p.

3. Equipe Alain N UMR144 CNRS .Les Puces à ADN sur lames de verre : principes et méthodes de confection, d'application expérimentale et d'analyse des données**.** Extraits du mémoire de la thèse de doctorat : « Applications de la technologie des Puces à ADN à l'étude de la différenciation méiotique et des mécanismes de recombinaison chez la levure *Saccharomyces cerevisiae* » 2004.

4. Guylaine R., Gino B., Séquençage génétique des cancers. Validité et utilité cliniques des profils moléculaires obtenus à l'aide des technologies de séquençage de nouvelle génération (NGS),2015.66 p,1915-3104 INESSS (PDF),

5. Jean B. Introduction aux technologies de. Séquençage nouvelle génération. Séquençage de 2nde génération. & aperçu des technologies.de 3ème génération.2013.

6. Laurent C. Signification des variants génétiques à faible ratio allélique détectés par séquençage à haut débit dans le cadre du diagnostic moléculaire des prédispositions aux cancers du sein et de l'ovaire : mosaïque, hématopoïèse clonale ou ADN tumoral circulant.2018 /2019.

7. Buot C. Département de biochimie, Application clinique du séquençage de l'exome pour le diagnostic moléculaire des syndromes polymalformatifs. Sherbrooke, Québec, Canada .2015.

8. Amélie P. DU de séquençage haut-débit Module 1 : Séquençage nouvelle génération - Technologies & applications,principale méthodes d'enrichissement,2013.

9. Nicolas S. DU Séquençage Haut Débit et Maladies Génétiques Dijon, 2013.

10. Etienne M. Les défis du séquençage à haut débit dans l'exploration génétique des cancers du sein et de l'ovaire. Génétique humaine. Normandie Université, 2017.

11. NANCIE R. Bioinformatique des puces à ADN et application à L'analyse du transcriptome de Buchnera aphidicola.Sciences du Vivant, INSA de Lyon,2004,326p.
12. Séquençage de l'ADN. https://fr.wikipedia.org/wiki/S%C3%A9quen%C3%A7age_de_l%27ADN#
13. LE CROM S. Concept paper : le sequençage a haut débit méthodes et enjeux en médecine, pharmacologie et toxicologie, Version 1, 2011.
14. BLERVAQUE R. Séquençage haut-débit nouvelle génération: Principes et caractéristiques.2011, 759-769.
15. BOULOUARD Flavie. Signification des variants génétiques à faible ratio allélique détectés par séquençage à haut débit dans le cadre du diagnostic moléculaire des prédispositions aux cancers du sein et de l'ovaire : mosaïque, hématopoïèse clonale ou ADN tumoral circulant. médecine, université de caen.normandie, 2019, 107p
16. VIKTOR L. Basic bioinformatics - from fastq to variants Viktor Ljungström Department of Immunology, Genetics and Pathology Uppsala University pdf(28/08/2020)
17. Variant calling Guillaume Robert-Siegwald – Inovarion École de bioinformatique AVIESAN-IFB 2018..
18. Sezerman, Ulgen E., Seymen N. , Durasi. Bioinformatics Workflows for Genomic Variant Discovery, Interpretation and Prioritization.chapiter 2.In: Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variation,[en ligne]. Iran University of Medical Sciences: Ali Samadikuchaksaraei.2019. Disponible sur : (consulté le 28.02.2019). (29/08/2020).
19. Manojkumar K. Umadevi S. Bharanidharan D. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data.BMC Bioinformatics, 2019, Vol.20, n° 342, 1471-2105 (29/08/2020).
20. Bao R, Huang L, Andrade J, et al.Review of current methods,applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Informatics.2014;13(Suppl 2):67-82 (01/09/2020).
21. Andrews S. *FastQC*. Babraham Bioinformatics; 2012. Cambridge, UK.
22. Babraham Bioinformatics. *FastQ Screen*. Babraham Bioinformatics; 2013. Cambridge, UK.
23. Hannon Lab. *FASTX-Toolkit*. Hannon Lab; 2010. Cold Spring Harbor, NY.
24. Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*. 2012;7:e30619.

25. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27:863–4.
26. Zhou Q, Su X, Wang A, Xu J, Ning K. QC-chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One*. 2013;8:e60234.
27. Guo Y, Zhao S, Sheng Q , et al. Multi-perspective quality control of Illumina exome sequencing data using QC3. Genomics. 2014;103(5–6):323–8.
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2.
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
30. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81.
31. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
32. Zhou W, Chen T, Zhao H, et al. Bias from removing read duplication in ultra-deep sequencing experiments. Bioinformatics. 2014;30(8):1073–80.

33. Homer N, Nelson SF. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol. 2010;11:R99.

34. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.

35. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

36. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. Curr. Protoc. *Bioinform*. 2013;43:11.10.1–33 .

37. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. *Genome Res*. 2011;21:961–73.

38. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol*. 2011;12:R112.

39. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

40. Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics. 2013;29:494–6.

41. Cabanski CR, Cavin K, Bizon C, et al. ReQON: a bioconductor package for recalibrating quality scores from next-generation sequencing data. BMC Bioinformatics. 2012;13:22.

42. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z. Variant callers for next-generation sequencing data: a comparison study. PLoS One. 2013;8:e75619.

43. Garrison E, Marth G: Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio 2012.

44. Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*. 2012/13/8.
45. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
46. Sherry ST, Ward MH, Kholodov M,et al. dbSNP: The NCBI database of genetic variation. Nucleic Acids Research. 2001;29(1):308-311.
47. Auton A, Brooks LD, Durbin RM,et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
48. Lek M, Karczewski KJ, Minikel EV,et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature.2016;536(7616):285-291.
49. Available from: https://gnomad.b roadinstitute.org/.
50. Tate JG, Bamford S, Jubb HC, et al.COSMIC: The catalogue of somatic mutations in cancer. Nucleic Acids Research. 2019;47(D1):D941-D947.
51. Landrum MJ, Lee JM, Riley GR,et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2014;42(Database issue):D980-D985.
52. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003;31(13):3812-3814.
53. Adzhubei IA, Schmidt S, Peshkin L,et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010;7(4):248-249.
54. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Research.2009;19(9):1553-1561.
55. Schwarz JM, Rödelsperger C,Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods.2010;7:575-576.
56. Reva B, Antipin Y, Sander C.Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research. 2011;39(17):e118.
57. Shihab HA, Gough J, Cooper DN,et al. Predicting the functional,molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation. 2013;34(1):57-65.
58. Davydov EV, Goode DL, Sirota M,Cooper GM, Sidow A, Batzoglou S.Identifying a high fraction of the human genome to be under selective

constraint using GERP++. PLoS Computational Biology. 2010;6(12):e1001025.

59. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral 28 Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations substitution rates on mammalian phylogenies. Genome Research. 2010;20(1):110-121.
60. Garber M, Guttman M, Clamp M,Zody MC, Friedman N, Xie X.Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics.2009;25(12):i54-i62.
61. Tang H, Thomas PD. PANTHERPSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics. 2016;32(14):2230-2232.
62. González-pérez A, López-bigas N.Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score,Condel. American Journal of Human Genetics. 2011;88(4):440-449.
63. Rentzsch P, Witten D, Cooper GM,Shendure J, Kircher M. CADD:Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019;47(D1):D886-D894.
64. Wong WC, Kim D, Carter H,Diekhans M, Ryan MC, Karchin R.CHASM and SNVBox: Toolkit for detecting biologically important single nucleotide mutations in cancer.Bioinformatics. 2011;27(15):2147-2148.
65. Mao Y, Chen H, Liang H, Mericbernstam F, Mills GB, Chen K. CanDrA:Cancer-specific driver missense mutation annotation with optimized features. PLoS One. 2013;8(10):e77945.
66. Carter H, Douville C, Yeo G, Stenson PD, Cooper DN, Karchin R.Identifying Mendelian disease genes with the variant effect scoring tool.BMC Genomics. 2013;14(3):1-16.
67. O'leary NA, Wright MW, Brister JR,et al. Reference sequence (RefSeq)database at NCBI: Current status,taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016;44(D1):D733-D745.
68. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. Nucleic Acids Research. 2018;46(D1):D754-D761 [51] Mccarthy DJ, Humburg P, Kanapin

A, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Medicine. 2014;6(3):26.

69. Wang K, Li M, Hakonarson H.ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Research. 2010;38:e164.
70. Cingolani P, Platts A, Wang le L,et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.
71. Mclaren W, Gil L, Hunt SE, et al.The Ensembl variant effect predictor.Genome Biology. 2016;17(1):122.
72. Paila U, Chapman BA, Kirchner R,Quinlan AR. GEMINI: Integrative exploration of genetic variation and genome annotations. PLoS Computational Biology. 2013;9(7): e1003153.
73. Desvignes JP, Bartoli M, Delague V,et al. VarAFT: A variant annotation and filtration system for human next generation sequencing data. Nucleic Acids Research. 2018;46(W1):W545-W553.
74. Hu H, Huff CD, Moore B, Flygare S,Reese MG, Yandell M. VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genetic Epidemiology. 2013;37(6):622-634.
75. Zhou W, Chen T, Chong Z, et al.TransVar: A multilevel variant annotator for precision genomics.Nature Methods. 2015;12(11):1002-1003.
76. Leiserson MD, Gramazio CC, Hu J,Wu HT, Laidlaw DH, Raphael BJ.MAGI: Visualization and collaborative annotation of genomic aberrations.Nature Methods. 2015;12(6):483-484.
77. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C.SNPnexus: Assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic Acids Research. 2018;46(W1):W109-W113.
78. Sun C, Medvedev P. VarMatch: Robust matching of small variant datasets using flexible scoring schemes.Bioinformatics. 2017;33(9):1301-1308.

79. Li H: Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 2014:btu356.
80. Anthony M. Bolger1,2, Marc Lohse1 and Bjoern Usadel2,3,*. Trimmomatic: a flexible trimmer for Illumina sequence data. April 1, 2014. Genome analysis, Vol. 30 no. 15 2014, 2114–2120 P.
81. ALKAN, C., COE, B. P. & EICHLER, E. E. Genome structural variation discovery and genotyping. Nature reviews genetics 12, 363 (2011).
82. Rute pereira 1,2,* , jorge oliveira 2,3,y and mário sousa 1,2,y.3 january 2020. Bioinformatics and computational tools for Next-generation sequencing analysis in Clinical genetics.clinical medicine.30p.
83. Bian X, Zhu B, Wang M, et al.Comparing the performance of selected variant callers using synthetic data and genome segmentation. BMC.Bioinformatics. 2018;19(1):429.
84. Marisa P. Dolled-Filhart, Michael Lee Jr., Chih-wen Ou-yang,Rajini Rani Haraksingh, and Jimmy Cheng-Ho Lin. 22 November 2012. Computational and Bioinformatics Frameworks for. the Scientific World Journal. Volume 2013Article ID 730210.10 pages.
85. Somak Roy, MD; William A. LaFramboise, PhD; Yuri E. Nikiforov, MD, PhD; Marina N. Nikiforova, MD; Mark J. Routbort, MD, PhD;John Pfeifer, MD, PhD; Rakesh Nagarajan, MD, PhD; Alexis B. Carter, MD; Liron Pantanowitz, MD. Challenges and Strategies for Implementation in a Clinical Environment. Next-Generation Sequencing Informatics.2016.20P.
86. Yannis D. Alignement de séquences Manipulation de fichiers SAM/BAM. Bioinformatique.centre hospitaliare universitaire Dijon. *2019.31*
87. Logiciels conseillés par laplateforme.12/03/18, Version 3.
88. Available from: http://broadinstitute.github.io/picard/.
89. Tebani, A.; Afonso, C.; Marret, S.; Bekri, S. Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations. Int. J. Mol. Sci. **2016**, 17, 1555. [CrossRef] [PubMed].
90. Ohashi, H.; Hasegawa, M.; Wakimoto, K.; Miyamoto-Sato, E. Next-Generation Technologies for Multiomics Approaches Including Interactome Sequencing. BioMed Res. Int. **2015**, 2015, 1–9. [CrossRef] [PubMed].
91. Logiciels conseillés par laplateforme. 19/09/12. Version 1.

92. M.A drien BUISSON. mise au point d'une technique de séquençage haut-débit par technologie ion torrent™ pour l'étude des gènes nf1 et spred1. biologie medicale. l'hôpital edouard herriot (lyon), universite claude bernard - lyon i,2016,245p.
93. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioi nformatics. 2011. 27(15):2156-2158.
94. source :http://vcftools.sourceforge.net/VCF-poster.pdf.
95. Henri Michel , [en ligne] (28/9/2020).
96. Qu'est-ce que Debian ?.[enligne] https://www debian.org/releases/etch/arm/ch01s01.html.fr.(28/09/2020).
97. Bastien L. Python : tout savoir sur le principal langage Big Data et Machine Learning.en ligne. (28/09/2020).
98. National Center for Biotechnology Information, U.S. National Library of Medicine.Using the SRA Toolkit to convert .sra files into other formats.https://www.ncbi.nlm.nih.gov/books/NBK158900/.(28.092020).
99. https://github.com/s-andrews/FastQC.
100. sickle - A windowed adaptive trimming tool for FASTQ files using qualityhttps://github.com/najoshi/sickle.(30.09.2020).
101. https://sourceforge.net/projects/bio-bwa/.
102. Petr D. samtools – Utilities for the Sequence Alignment/Map (SAM) format.http://www.htslib.org/doc/samtools.html.(01.10.2020).
103. Petr D. Shane M. John.M.This documentation refers to the latest development version of BCFtools which can be downloaded from github.http://github.com/samtools/bcftools (01.10.2020).
104. Anton Nekrutenko , Alex Ostrovsky , Appel de variantes dans les systèmes non diploïdes.
105. http://vcftools.sourceforge.net/. (30.09.202)