République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Faculté des Sciences de la Nature et de la Vie

Université Frères Mentouri Constantine 1

Département de Biologie Appliquée

Mémoire en vue de l'obtention du Diplôme de Master en : **Bio-informatique**

THÈME

Traitement des données générées par le séquençage NGS du génome du SARS-CoV-2

Réaliser par :

MELIANI Imène

BOUDEMAGH Imène

Soutenu le : 23 - 09- 2021

Devant le jury :

Président : Pr. HAMIDECHI.M. A	Université Frères Mentouri- Constantine 1
Encadreur : Dr. CHEHILI.H	Université Frères Mentouri- Constantine 1
Co-Encadreur : Dr. BENSAADA.M	Université Abbas Leghrour- Kenchela
Examinateur : Dr. TEMAGOULT.M	Université Frères Mentouri- Constantine 1

Année universitaire 2020-2021

بسم الله الرحمن الرحيم :﴿وَمَا أُوتِيتُمْ مِنَ الْعِلْمِ إِلَّا قَلِيلًا﴾ {الإسراء: 85}.

MELIANI IMENE

Dedication

I'm thankful to ALLAH Almighty my creator, for all his blessings.

My great teacher and messenger, Mohammed (May Allah bless and grant him), who taught us the purpose of life

I am dedicating this work to the beloved people who have meant and continue to mean so much to me. Although they are no longer in this world, their memories continue to live inside me, my second mother LEILA TAIR who raised me and loved me, my grandfather's CHAABAN BOUDEMAGH and ABD EL RAHMAN TAIR. I love you all and miss you all beyond words. May Allah (SWT) grant you Jannah Firdaws.

My great parents BOUDEMAGH CHERIF and TAIR SOUAAD, who never stop giving of themselves in countless ways.who have never failed to give me financial and moral support, for giving all my needs during all this time.

To my grandmother HEDA and to LAMRI and all the TAIR family, my strong pillar, my source of inspiration, wisdom, knowledge and understanding, my symbol of love and giving. They have been the source of my strength throughout all this time.

To my little brother and sister Med AMINE and SALSABIL.

to my dearest friend NIHEL BENMANSOUR who has been a constant source of support and encouragement during the challenges of graduate school and life.

Especially dedicated to the teachers who helped and guided me to successfully complete the journey.

IMENE BOUDEMAGH

Remerciement

Un travail scientifique, étant le résultat de plusieurs efforts conjugués, tout apport pour l'améliorer est le bienvenu.

Nous remercions Dieu l'Éternel pour le souffle de vie et toutes les capacités physiques et intellectuelles utilisées pour réaliser ce travail.

Nos sincères gratitudes à Monsieur le Professeur **HAMIDECHI MOHAMED ABDELHAFID et** Monsieur **TEMAGOULT MAHMOUD** pour la qualité de leurs enseignements, conseils et intérêt incontestable qu'ils portent à tous les étudiants. Messieurs les jurys, vous nous faites un grand honneur en acceptant de juger notre travail.

A notre Docteur **CHEHILI HAMZA** votre compétence et votre encadrement ont toujours suscité votre profond respect. Nous vous remercions pour votre accueil et vos conseils. Veuillez trouver ici, l'expression de notre gratitude et de notre grande estime.

Nous présentons notre gratitude à monsieur le Docteur **BENSAADA MOUSTAFA**, pour avoir accepté la direction de ce travail de fin de cycle malgré ses diverses occupations.

Nous remercions le corps du département de BIOLOGIE APPLIQUÉE pour le support et la formation reçue tout au long de notre parcours de Master.

Nous remercions tous les enseignants de notre spécialité BIO-INFORMATIQUE pour leurs qualités scientifiques et pédagogiques.

Enfin, nous remercions tous ceux qui, de loin ou de près, ont procuré leurs apports pour la construction de cet édifice. Nous leur exprimons ici grandement notre reconnaissance.

Résumé

Depuis décembre 2019, une maladie à coronavirus 2019 (COVID-19) causée par le coronavirus 2 du syndrome respiratoire aigu sévère (SARS-CoV-2) est apparue à Wuhan et devient par la suite une épidémie mondiale. Cette dernière a stressé le monde entier, elle a obligé les chercheurs à faire des efforts pour étudier ce virus qui s'est propagé rapidement, provoquant une morbidité et mortalité importantes, ce qui a incité beaucoup de laboratoires de recherches et des universités à séquencer le génome de ce virus et disposer des pipelines spécifiques pour faire des analyses et des interprétations qui nous aides à lutter contre cette urgence sanitaire la plus difficiles.

Cette étude a comme objectif être capable au futur de traiter des données séquencées du SARS-CoV-2 en Algérie. Nous avons créé notre propre pipeline en utilisant des logiciels et outils libres d'accès pour faire des essais sur des séquences qui sont déjà séquencées par Illumina et Ion Torrent et traiter par d'autres pipelines et comparer les résultats après avoir détecté les variants.

Mots clés : COVID-19, SARS-CoV-2, pipelines, outils, variants.

Abstract

In December 2019, the coronavirus disease (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) appeared in Wuhan and subsequently became a global epidemic. It forced researchers to make efforts to study this virus that extended rapidly, causing significant morbidity and mortality. One of the solutions was to sequence the genome of this virus and dispose specific pipelines to make analyzes and interpretations which help to fight against this most difficult health emergency.

The objective of this study is to be able in the future to process sequenced data from SARS-CoV-2 in Algeria. We created our own pipeline using open-source software and tools to test sequences that are already sequenced by Illumina and Ion Torrent and process through other pipelines and compare the results after detecting the variants.

Keywords: COVIDE-19, SARS-CoV-2, pipelines, tools, variants.

ملخص

في ديسمبر 2019، ظهر فيروس كورونا (Covid-19) الناجم عن متلازمة الالتهاب التنفسية الحاد، (Sars-cov-2)في ووهان، وأصبح فيما بعد وباء عالميا، مما وضع العالم تحت ضغط كبير. أجبر الباحثين على بذل جهود كثيرة لدراسة هذا الفيروس الذي انتشر بسرعة وتسبب في العديد من الوفيات. مما جعلهم يقومون بسلسلة جينوم هذا الفيروس واجراء التحليلات التي تساعد على مكافحة هذه الحالة الصحية الطارئة.

الهدف من هذه الدراسة هو ان نكون قادرين مستقبلا على معالجة البيانات من Sars-cov-2 في الجزائر. فقمنا بإنشاء مخطط تحليلات خاص بنا باستخدام برامج يمكن تحميلها مجانا.

البيانات التي تم تسلسلها بالفعل بواسطةIllumina و، Ion torrent ومعالجتها من خلال مخططات تحليلية أخرى ومقارنة النتائج بعد اكتشاف سلسلة المتغيرات.

الكلمات المفتاحية : SARS COV-2 , COVID-19 ,مخططات تحليلية ، برامج ، المتغيرات.

Sommaire :

Liste des abréviations :

DNA: Deoxyribonucleic Acid

RNA: Ribonucleic Acid

GA: Genome Analyzer

SBS: Sequencing by Synthesis

PCR: Polymerase Chain Reaction

CCD: Coupled-Charge Device

PE: Paired-End

Bp: Base Pair

SMRT: Single-Molecule Real-Time

ONT: Oxford Nanopore Technology

CoV: Coronavirus

COVID-19: Coronavirus Disease appeared in 2019

SARS-CoV-2: the Severe Acute Respiratory Syndrome Coronavirus 2

CNRS: National Center for Scientific Research

NGS: Next Generation Sequencing

Fast QC: Fast Quality Control

GATK: Genome Analysis Toolkit

SRA: Sequence Read Archive

BWA: Burrows-Wheeler-Aligner

NCBI: National Center for Biotechnology Information

SBS: Sequencing By Synthesis

SVs: Visualizing Structural Variations

Liste des figures :

Liste des tableaux :

Glossaire :

ADN, nucléotide, gène, exome, séquençage, variantes... De quoi parlons-nous ?

Génome : Le génome humain est l'ensemble du matériel génétique d'un individu transporté par l'ADN.

ADN : L'ADN est le vecteur moléculaire de l'information génétique. Il est formé par la séquence de 4 nucléotides : ces 4 molécules sont constituées chacune avec une base différente, A, T, C, G. La séquence de nucléotides d'ADN constitue sa séquence. Séquencer l'ADN signifie lire la longueur linéaire de la séquence des nucléotides d'un fragment d'ADN donné

Gènes : Notre génome contient environ 23 000 gènes répartis sur les chromosomes. Les gènes sont présents en double, un sur chaque homologue chromosomique. Ces versions peuvent être identiques ou différentes : ce sont les allèles (ou des versions différentes d'un gène). Les gènes sont des morceaux d'ADN qui fabriquent des protéines (la dystrophine est une protéine). Dans chaque gène, les nucléotides sont liés entre eux d'une manière particulière, ce qui donne des informations spécifiques.

Exons : Les régions codantes des gènes sont appelées exons. C'est à partir des informations contenues dans les exons que la cellule fabrique des protéines : Nous disons qu'un gène « code » pour une protéine.

Exome : Il s'agit de toutes les régions codantes de notre génome (sur les 23 000 gènes). Ils constituent 34 millions de paires de bases d'ADN.

Introns : Les régions non codantes des gènes sont appelées introns. Ils sont situés entre les régions de codage. Ils ne permettent pas de produire des protéines mais d'avoir d'autres fonctions.

Variations de l'ADN : Environ 0,5 % du génome diffère d'un individu à l'autre, ce qui rend possibles des millions de variations. Ils peuvent apparaître dans les séquences codantes de gènes, d'exons ou d'autres régions. Certaines de ces variations sont fréquentes dans la population (fréquence supérieure à 1%) et ne sont pas pathogènes : on parle de polymorphisme. Ils rendent chaque individu unique. D'autres variantes appelées « anomalies ou mutations génétiques » sont beaucoup plus rares et causent des maladies. Il existe également de nombreuses variantes de l'ADN dont nous ne savons pas si elles sont responsables ou non d'une maladie : nous parlons de variantes génétiques.

Bibliothèque de fragments : Collection du brin entier des fragments d'ADN (à séquencer)

Gaps : La région non séquencée de l'ADN est appelée gap.

Canting: Une séquence continue de l'ADN assemblé.

Lecture : Les données de sortie provenaient de la machine séquenceur vers l'ordinateur pour une séquence particulière.

Couverture : Le nombre de fois que la machine de séquençage a couvert la séquence d'AND.

Introduction Générale

Introduction générale :

Le séquençage de nouvelle génération (NGS) est le mécanisme permettant de déterminer les séquences d'ARN et d'ADN à l'aide des machines et des techniques modernes. Le séquençage de l'ADN fournit des informations génétiques qui sont transportées dans un segment d'ADN particulier, ce qui aide à mieux comprendre l'industrie agricole, les relations évolutives entre les espèces, le diagnostic des maladies causées par des facteurs génétiques.... Avec le développement rapide de l'informatique et des sciences biologiques, le volume des données génomiques produites à partir des plateformes NGS est devenu massif.

Les données collectées concernent non seulement les génomes complets, mais aussi les CDs, d'exon, d'intron…etc provoquent des problèmes de gestion et de stockage des données. En outre, ces données doivent être structurées et analysées. Ainsi, des logiciels spéciaux et des systèmes informatiques rapides sont nécessaires pour traiter les immenses données.

En effet des bio-informaticiens spécialisés et formés sont devenus indispensables à l'analyse des données générées par le NGS.

Au cours des deux dernières années, la plupart des scientifiques n'avaient qu'un seul objectif, en savoir plus sur le SARS-CoV-2. Ce virus est devenu une pandémie en peu de temps, il a rendu chaque biologiste intéressé à résoudre son mystère. Par conséquent, les données et les méthodes de séquençage de cette pandémie sont si volumineuses et si variantes.

En raison de la diversité des variantes du SARS-CoV-2, un pipeline de bio-informatique est devenu nécessaire pour comprendre cette maladie. Après avoir étudié les outils de séquençage qui ont été utilisés, nous avons proposé un pipeline pour traiter les données générées par NGS.

Ce travail est divisé en quatre chapitres. Le premier parle de l'histoire du séquençage et de ses générations.

Dans le deuxième chapitre, nous présentons certains des outils qui ont des commandes spécifiques pour le virus, d'autres sont spécifiques à un séquenceur.

Dans le troisième chapitre, après avoir essayé et comparé différents outils et séquences, nous nous faisons une idée du chemin spécifique. En utilisant les derniers outils et versions pour obtenir de meilleurs résultats clairs, nous avons créé un pipeline général NGS SARS-CoV-2 qui est facile à installer, à utiliser, à comprendre, en donnant des résultats rapides et moins d'erreurs.

Dans le chapitre quatre, nous avons implémenté notre pipeline général NGS avec les deux séquenceurs, nous obtenons nos résultats et les discutons.

Chapitre 1 : Le séquençage et les nouvelles techniques

Introduction :

Le développement des technologies de séquençage de nouvelle génération ou NGS (Next Generation Sequencing) au cours des dix dernières années constitue une révolution technologique sans précédent. Alors que le séquençage d'un seul génome humain qui a nécessité dix ans de collaboration internationales de 1993 à 2003 et 2,7 milliards de dollars, il est désormais possible, dans quelques semaines ou quelques jours, pour séquencer l'ensemble de la région codante de 23000 gènes humains, l'exome, représentant 34 millions de paires de bases d'ADN a des coûts inférieurs à 1000 € (Kern, 2021).

Le séquençage de génomes humains entiers dans un but médical (3 milliards de paires de bases) a déjà été réalisé par quelques équipes aux États-Unis, Royaume-Uni, Pays-Bas et France (Verma, 2019).

Cette technologie est à l'origine d'une révolution médicale : la médecine de précision ou génomique dont l'objectif est d'optimiser le diagnostic, la prévention et le traitement des maladies humaines en fonction des variations génétiques individuelles… (Weinbrecht2, 2013).

1-Séquençage de l'ADN :

Le séquençage de l'ADN est une technique biochimique utilisée pour déterminer l'ordre correct des quatre blocs de construction chimiques appelés « bases » (adénine, thymine, cytosine et guanine) de la séquence d'ADN (acide désoxyribonucléique) (DNA Sequencing Fact Sheet, 2020).

L'ADN est le modèle d'un organisme, aucune compréhension de la fonction génétique ou de l'évolution ne pourrait être complète sans lui, il contient toute l'information génétique. Alors que l'ordre des acides nucléiques dans les chaînes de polynucléotides contient l'information sur les propriétés héréditaires et biochimiques de la vie terrestre. C'est pourquoi le séquençage de l'ADN indique le type d'information génétique contenues dans le segment de l'ADN. Il s'agit du niveau le plus fondamental de connaissance d'un gène ou d'un génome. (Griffiths, 2016) (James M Heather 1, 2016).

2-Différentes méthodes de séquençage de l'ADN :

L'histoire de l'ADN commence Watson et Crick ont résolu de façon célèbre la structure tridimensionnelle de l'ADN en 1953, à partir de données cristallographiques produites par Rosalind Franklin et Maurice Wilkins, ce qui a contribué à établir un cadre conceptuel pour la réplication de l'ADN et le codage des protéines dans l'acide nucléique. Cependant, la capacité de le lire n'a pas été possible pendant un certain temps. Les stratégies développées pour

déduire la séquence des chaînes de protéines ne semblaient pas s'appliquer facilement à l'acide nucléique, les molécules d'ADN étant beaucoup plus longues et constituées de moins d'unités plus semblables les unes aux autres. De nouvelles techniques devaient être développées (Barton E. Slatko, 2018) (PhD, 2021) (Jason R. Miller, 2010).

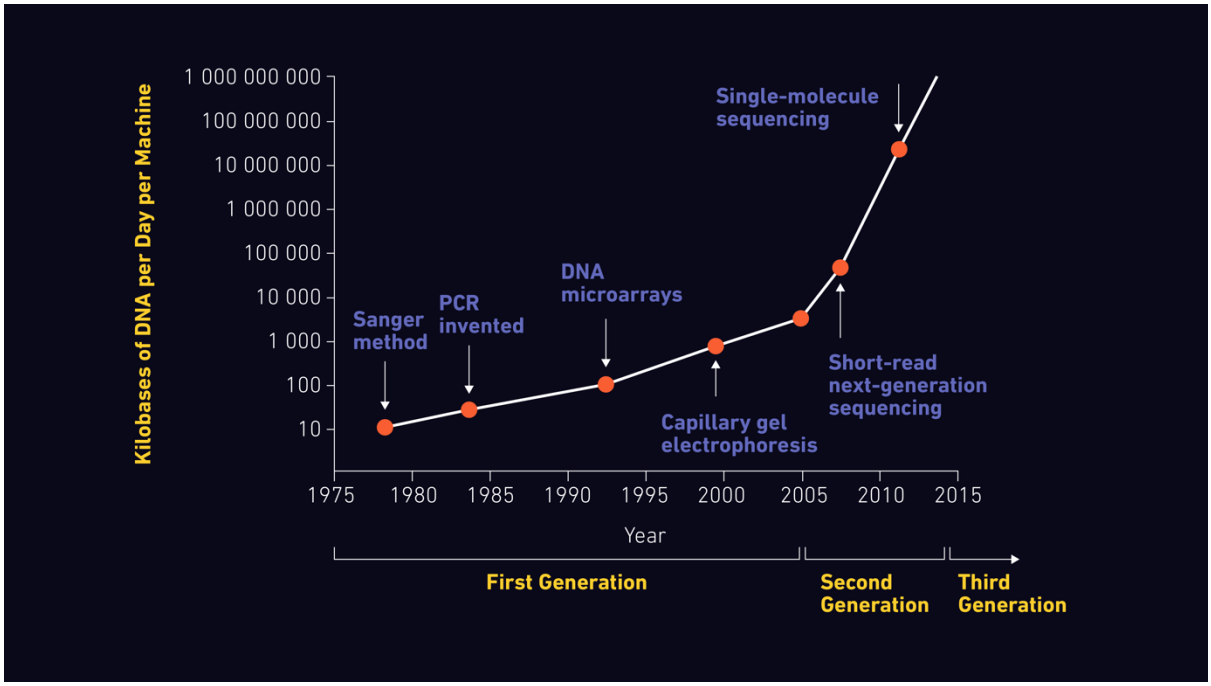


Figure 1 : L'évolution des méthodologies de séquençage (PhD, 2021).

2-1-Séquençage de Sanger :

Développée par Fred Sanger et ses collègues au milieu des années 1970, elle utilise des didésoxynucléotides pour la "terminaison de chaîne". Le séquençage de Sanger est connu sous le nom de méthode de terminaison de chaîne ou méthode des didésoxynucléotides ou encore méthode de séquençage par synthèse. Il consiste à utiliser un brin de l'ADN double brin comme matrice à séquencer. Ce séquençage utilise des nucléotides chimiquement modifiés appelés dideoxy-nucléotides (dNTPs). Ils sont marqués pour chaque base de l'ADN par ddG, ddA, ddT, et ddC. Les dideoxy-nucléotides utilisent les dNTP pour freiner l'élongation, une fois incorporés dans le brin d'ADN ils empêchent la poursuite de l'élongation et l'élongation est terminée. On obtient alors des fragments d'ADN terminés par un dNTP. Les fragments sont séparés selon leur taille à l'aide d'une plaque de gel où les bandes résultantes correspondant aux fragments d'ADN peuvent être visualisées par un système d'imagerie (rayons X ou lumière UV). La réalisation la plus emblématique de cette technologie de séquençage est le décodage du premier génome humain (M.HeatherBenjaminChain, 2016) (Mehdi Kchouk1, 2017).

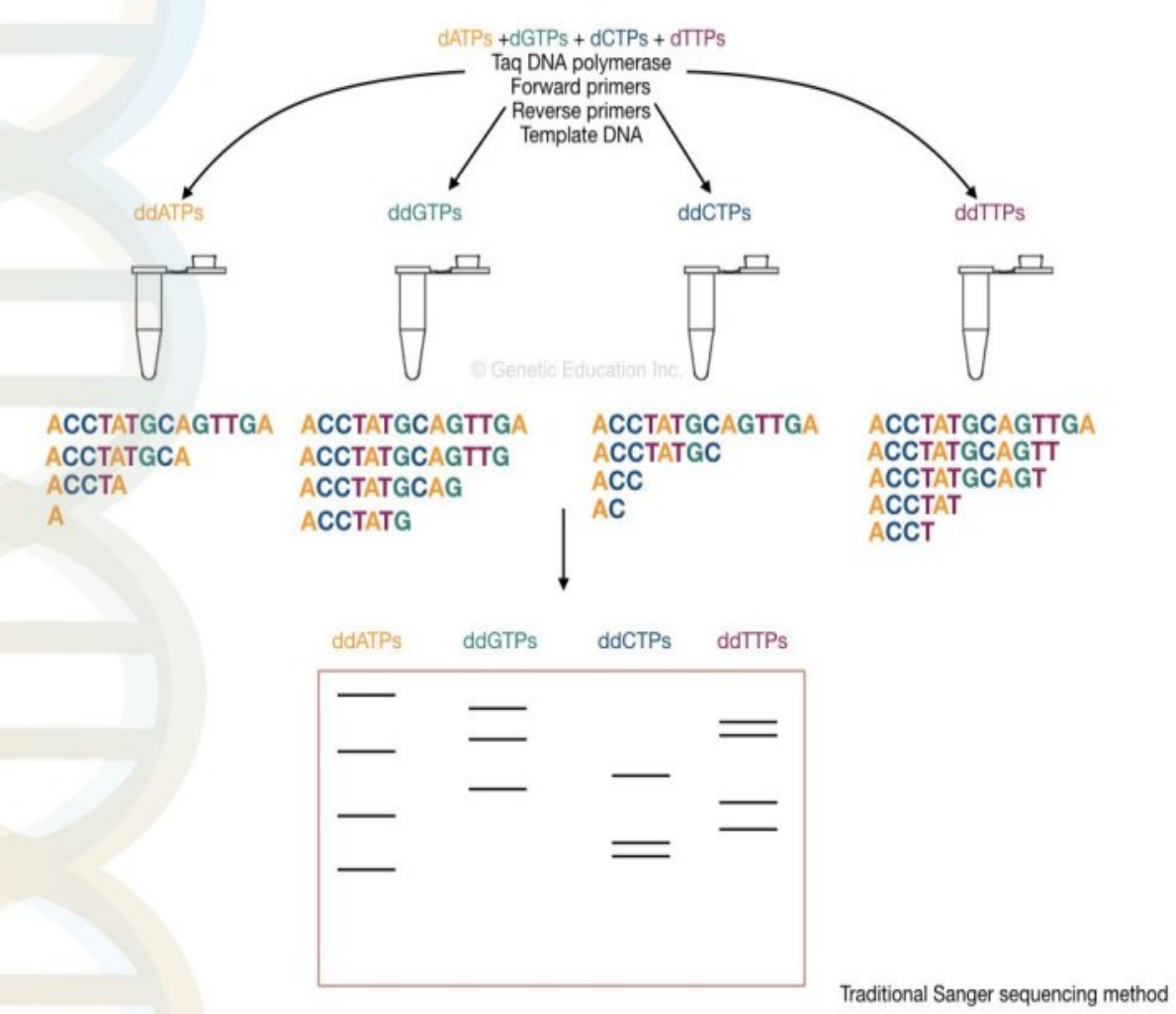


Figure 2 : Méthode de terminaison de chaîne de Sanger pour le séquençage de l'ADN (DNA Sequencing: History, Steps, Methods, Applications and Limitations, 2019).

2-2-Méthode MAXAM -GILBERT :

La méthode de Maxam et Gilbert a été développée en 1977. Elle est également appelée méthode de clivage chimique ; elle est connue sous le nom de méthode de dégradation chimique. Elle repose sur le clivage des nucléotides par des produits chimiques et est plus efficace avec les petits polymères de nucléotides. Le traitement chimique génère des cassures au niveau d'une petite proportion d'une ou deux des quatre bases nucléotidiques dans chacune des quatre réactions (C, T+C, G, A+G). Cette réaction donne des séries de fragments marqués qui vont se séparer selon leur taille par électrophorèse. Cette méthode est réalisée sans clonage d'ADN. Elle est également considérée comme dangereuse car elle utilise des produits chimiques toxiques et radioactifs. L'autobiographie est utilisée pour visualiser la séparation des fragments d'ADN. Grâce à l'extrémité 32P radio-marquée de l'ADN, les bandes d'ADN sont

visualisées par autoradiographie, comme le montre la figure ci-dessous (Mehdi Kchouk1, 2017) (G. Dorado, 2019).

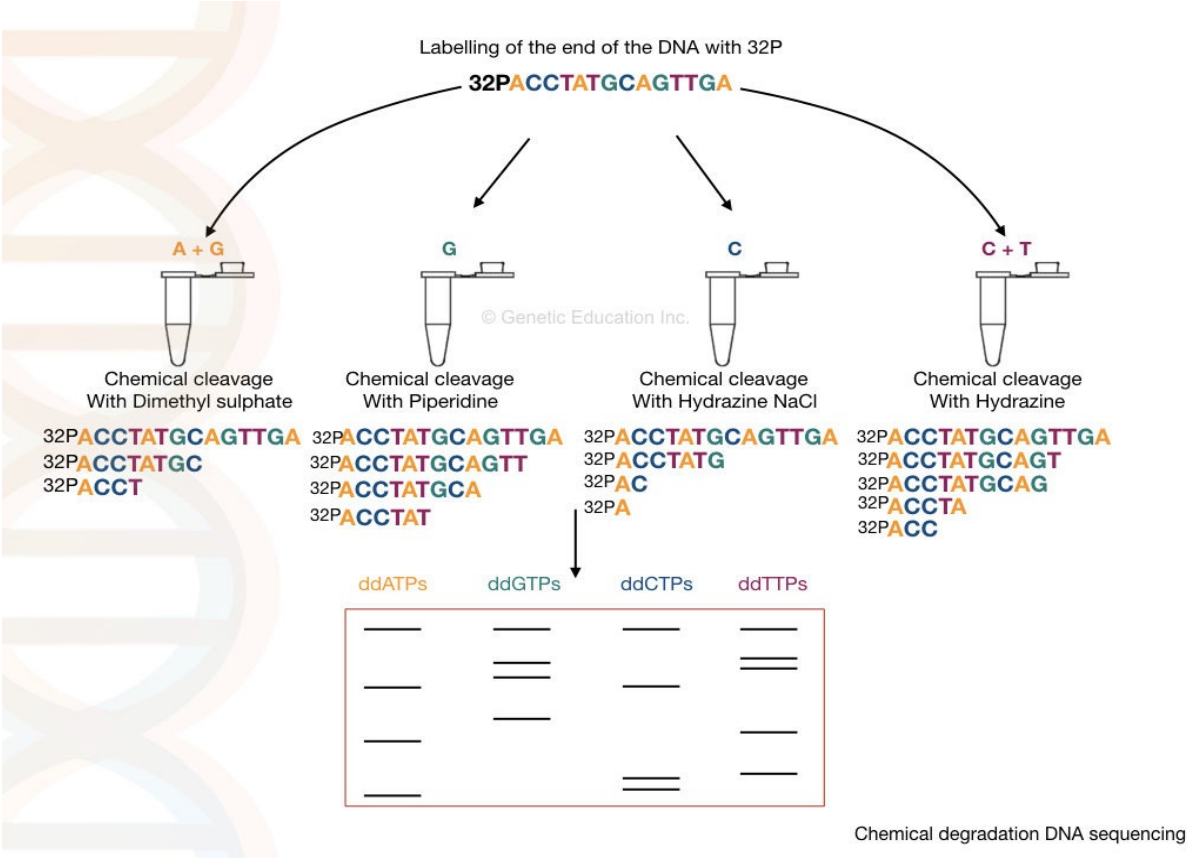


Figure 3 : Méthode de clivage chimique Maxam- Gilbert pour le séquençage de l'ADN (DNA Sequencing: History, Steps, Methods, Applications and Limitations, 2019).

2-3-La deuxième génération de séquençage :

Séquençage Roche/454 :

Le séquençage Roche/454 est une technique qui utilise le pyroséquençage, c'est-à-dire la détection du pyrophosphate libéré après chaque incorporation de nucléotide dans le nouveau brin d'ADN synthétique. La technique de pyroséquençage est une approche de séquençage par synthèse. Les échantillons d'ADN sont fragmentés de façon aléatoire et chacun d'entre eux est attaché à une bille qui porte des amorces ayant des oligonucléotides complémentaires aux fragments d'ADN, de sorte que chaque bille est associée à un seul fragment. Ensuite, chaque bille est isolée et amplifiée à l'aide d'une émulsion PCR qui produit environ un million de copies de chaque fragment d'ADN à la surface de la bille.

Les billes sont ensuite transférées sur une plaque contenant de nombreux puits appelée plaque picotitre (PTP) et on applique la technique de pyroséquençage qui consiste à activer une série de réactions en aval produisant de la lumière à chaque incorporation de nucléotide. En détectant l'émission de lumière aier chaque incorporation de nucléotide, la séquence du fragment d'ADN est déduite. L'utilisation de la plaque picotitre permet à des centaines de milliers de réactions de se produire en parallèle, augmentant considérablement le débit de séquençage. Le Roche/454 est capable de générer des lectures relativement longues qui sont plus faciles à mapper à un génome de référence. Les principales erreurs détectées lors du séquençage sont des insertions et des délétions dues à la présence de régions homopolymères (Lin Liu, 2012) (Erwin van Dijk, 2021).

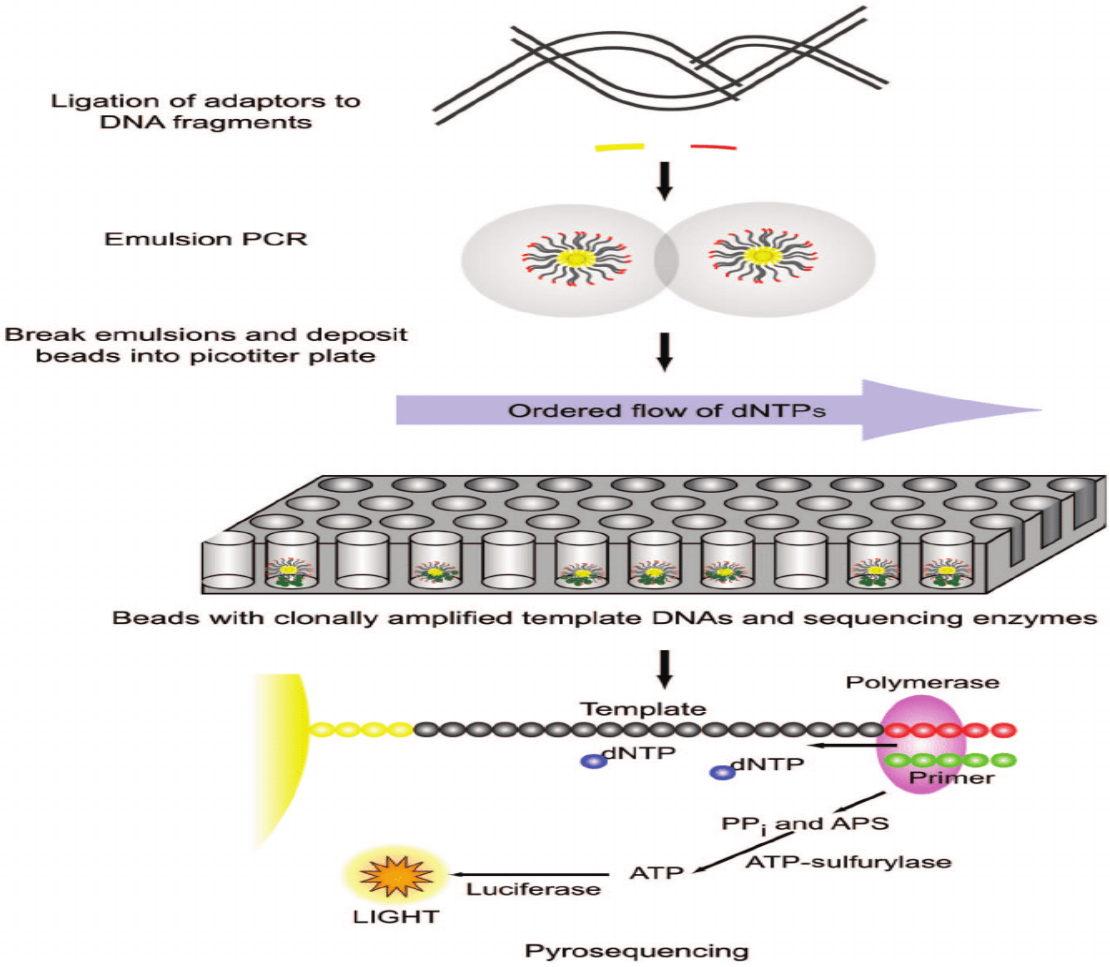


Figure 4 : Séquençage Roche 454 GS FLX (Karl V Voelkerding, 2009)

Séquençage Ion Torrent :

En 2010, Life Technologies a dévoilé la technologie de séquençage à semi-conducteurs Ion Torrent. Contrairement aux autres technologies de deuxième génération, elle n'utilise pas

de nucléotides marqués par fluorescence. Elle est basée sur la détection de l'ion hydrogène libérer lors du processus de séquençage. Concrètement, Ion Torrent utilise une puce qui contient un ensemble de micro-puits et dont chacun comporte une bille contenant plusieurs fragments identiques. L'incorporation de chaque nucléotide avec un fragment dans la perle, un ion hydrogène est libéré qui change le pH de la solution. Ce changement est détecté par un capteur fixé au fond du micro-puits et converti en un signal de tension qui est proportionnel au nombre de nucléotides incorporés.

Les séquenceurs Ion Torrent produisent des longueurs de lecture de 200 bp, 400 bp et 600 bp avec un débit qui peut atteindre 10 Gb pour le séquenceur ion proton. Les principaux avantages de cette technologie de séquençage sont axés sur les longueurs de lecture, qui sont plus longues que celles des autres séquenceurs SGS, et sur la rapidité du séquençage (entre 2 et 8 heures). L'inconvénient majeur est la difficulté d'interpréter les séquences homopolymères (plus de 6 pb) qui provoquent des erreurs d'insertion et de délétion (indel) avec un taux d'environ ~1% (Barton E. Slatko, 2018).

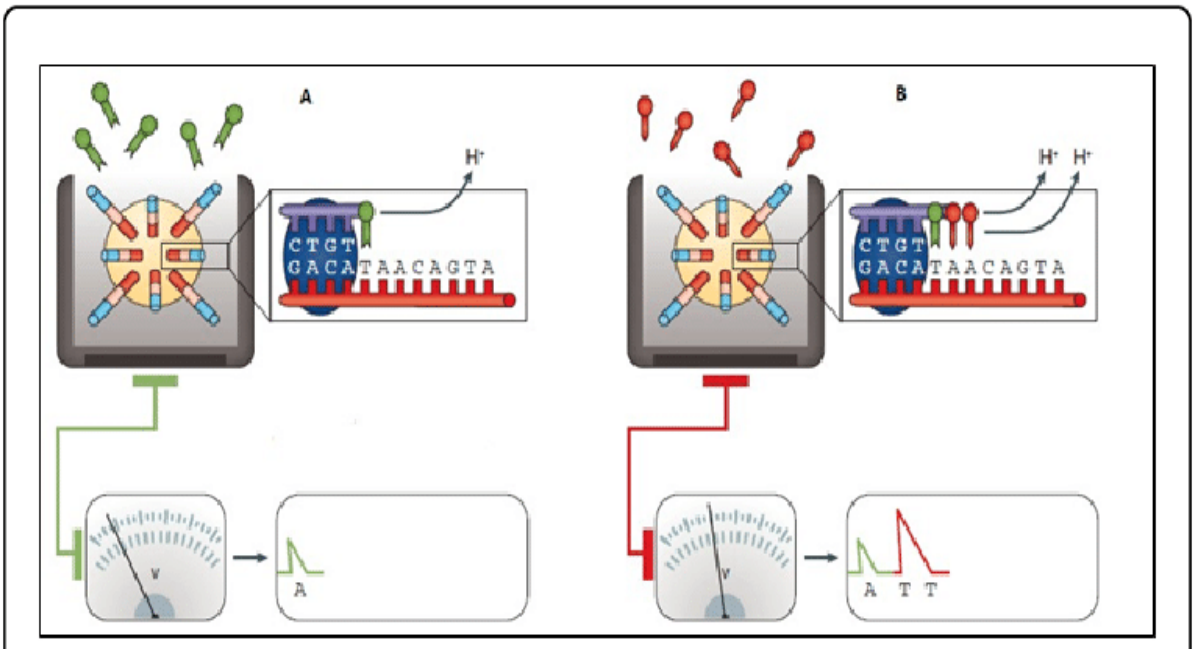


Figure 5 : Technologie de séquençage Ion torrent (Mehdi Kchouk1, 2017).

Séquençage Illumina/Solexa :

Actuellement, la technologie la plus utilisée sur le marché du NGS est le séquenceur Ilumina/Solexa Genome Analyzer (GA). Le séquenceur adopte la technologie du séquençage par synthèse (SBS). Au cours de la première étape, les échantillons d'ADN sont fragmentés de manière aléatoire, puis les molécules d'ADN avec des adaptateurs appropriés ligaturés à chaque extrémité sont utilisées comme substrats pour des réactions répétées de synthèse par

amplification sur un support solide (lame de verre) qui contient des séquences d'oligonucléotides complémentaires à un adaptateur ligaturé. Au cours de la deuxième étape, chaque séquence fixée sur le support solide est amplifiée par "amplification par pontage PCR" qui crée plusieurs copies identiques de chaque séquence ; un ensemble de séquences fabriquées à partir de la même séquence d'origine est appelé un cluster. Chaque cluster contient environ un million de copies de la même séquence originale. La dernière étape consiste à déterminer chaque nucléotide des séquences, Illumina utilise l'approche de séquençage par synthèse qui emploie des terminateurs réversibles dans lesquels les quatre nucléotides modifiés, les amorces de séquençage et les ADN polymérases sont ajoutés sous forme de mélange, et les amorces sont hybridées aux séquences.

Ensuite, des polymérases sont utilisées pour étendre les amorces en utilisant les nucléotides modifiés. Chaque type de nucléotide est marqué par un spécifique fluorescent afin que chaque type soit unique. Les nucléotides ont un groupe 3'-hydroxyle inactif qui garantit qu'un seul nucléotide est incorporé. Les clusters sont excités par un laser pour émettre un signal lumineux spécifique à chaque nucléotide, qui sera détecté par une caméra à dispositif à charge couplée (CCD) et des programmes informatiques traduiront ces signaux en une séquence de nucléotides.

Le processus se poursuit par l'élimination du terminateur avec le marqueur fluorescent et le démarrage d'un nouveau cycle avec une nouvelle incorporation. Les premiers séquenceurs Illumina/Solexa GA ont été capables de produire des lectures très courtes et ils avaient l'avantage de pouvoir produire des lectures courtes en paires (PE), dans lesquelles la séquence aux deux extrémités de chaque cluster d'ADN est enregistrée. Les données de sortie des derniers séquenceurs Illumina sont actuellement supérieures à 600 Gpb et les longueurs des lectures courtes sont d'environ 125 bp.

L'un des principaux inconvénients de la plateforme Illumina/Solexa est le besoin élevé de contrôle de la charge de l'échantillon, car une surcharge peut entraîner un chevauchement des clusters et une mauvaise qualité de séquençage. Le taux d'erreur global de cette technologie de séquençage est d'environ 1 %. Les substitutions de nucléotides sont le type d'erreurs le plus courant dans cette technologie, la principale source d'erreur étant due à une mauvaise identification du nucléotide incorporé (Christophe Audebert1, 2014) (Weinbrecht2, Next-Generation Sequencing: Methodology and Application, 2013) (Lin Liu, 2012).

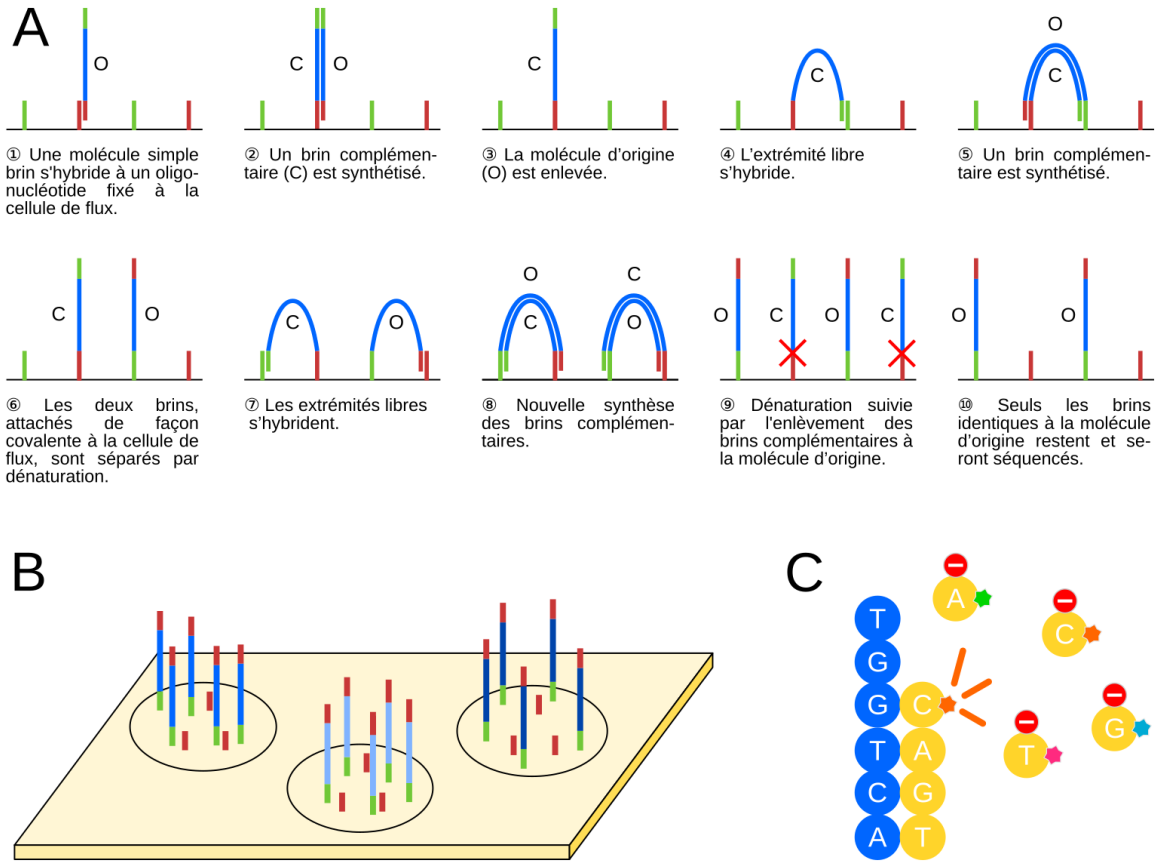


Figure 6 : Le principe de la technologie Illumina (Erwin van Dijk, 2021).

2-4-La troisième génération de séquençage :

Alors que le séquençage de nouvelle génération est de plus en plus utilisé et modifié, le séquençage de troisième génération apporte un nouvel éclairage sur le séquençage. Le séquençage de troisième génération présente deux caractéristiques principales. Premièrement, la PCR n'est pas nécessaire avant le séquençage, ce qui raccourcit le temps de préparation de l'ADN pour le séquençage. Deuxièmement, le signal est capturé en temps réel, ce qui signifie que le signal, qu'il soit fluorescent (Pacbio) ou électrique (Nanopore), est contrôlé pendant la réaction enzymatique d'ajout de nucléotides dans le brin complémentaire (James M Heather 1, 2016) (Mehdi Kchouk1, 2017).

Molécule unique en temps réel (SMRT) :

Il s'agit d'une méthode développée par Pacific Bioscience, qui utilise une enzyme modifiée et l'observation directe de la réaction enzymatique en temps réel. La cellule SMRT est constituée de millions de guides d'ondes en mode zéro (ZMW), dans lesquels sont intégrés un seul ensemble d'enzymes et une matrice d'ADN qui peut être détectée tout au long du processus. Au cours de la réaction, l'enzyme va incorporer le nucléotide dans le brin

complémentaire et cliver le colorant fluorescent précédemment lié au nucléotide. Ensuite, la caméra à l'intérieur de l'appareil va capturer le signal sous forme de film pour une observation en temps réel. Cela donnera non seulement le signal fluorescent mais aussi la différence de signal à long terme, ce qui peut être utile pour la prédiction de la variance structurelle dans la séquence, particulièrement utile dans les études épigénétiques telles que la méthylation de l'ADN (Barton E. Slatko, 2018) (Weinbrecht2, Next-Generation Sequencing: Methodology and Application, 2013).

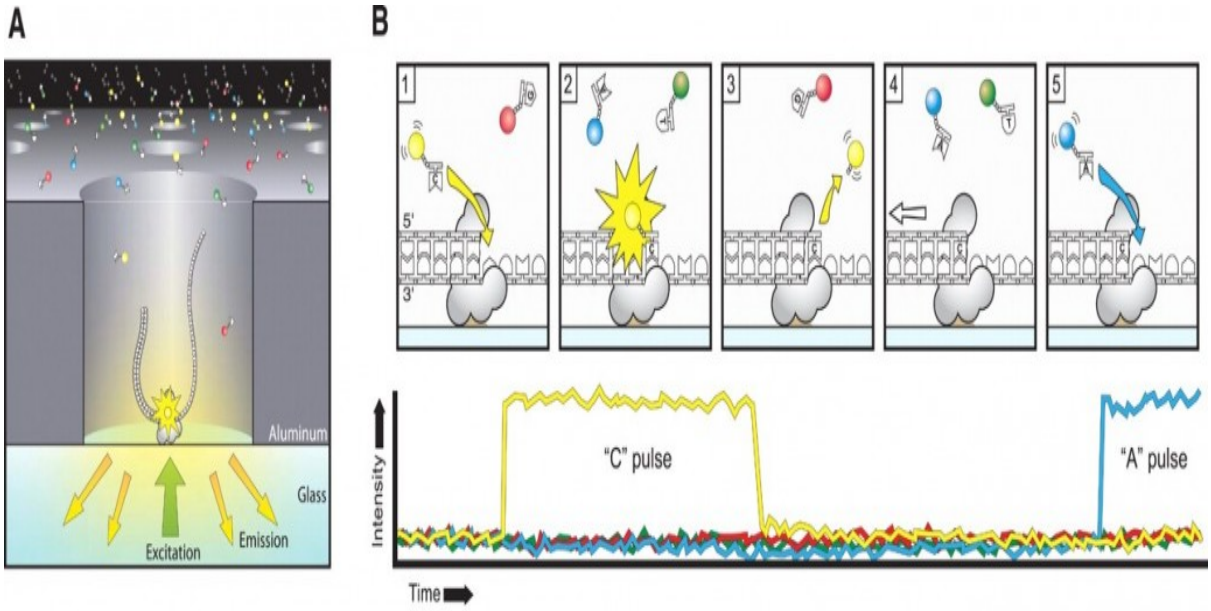


Figure 7 : Principe du séquençage "SMRT" (John Eid*, 2009).

Séquençage Oxford Nanopore :

Le séquençage Oxford Nanopore (ONT) a été développé comme une technique permettant de déterminer l'ordre des nucléotides dans une séquence d'ADN. En 2014, Oxford Nanopore Technologies a mis sur le marché le dispositif MinION qui promet de générer des lectures plus longues qui assureront une meilleure résolution structurelle du contenu des variantes et répétitions génomiques. Ces séquenceurs utilisent des nanopores protéiques dans une membrane polymère résistante à l'électricité, à travers laquelle des changements de courant caractéristiques se produisent lorsque chaque nucléotide passe dans le détecteur.

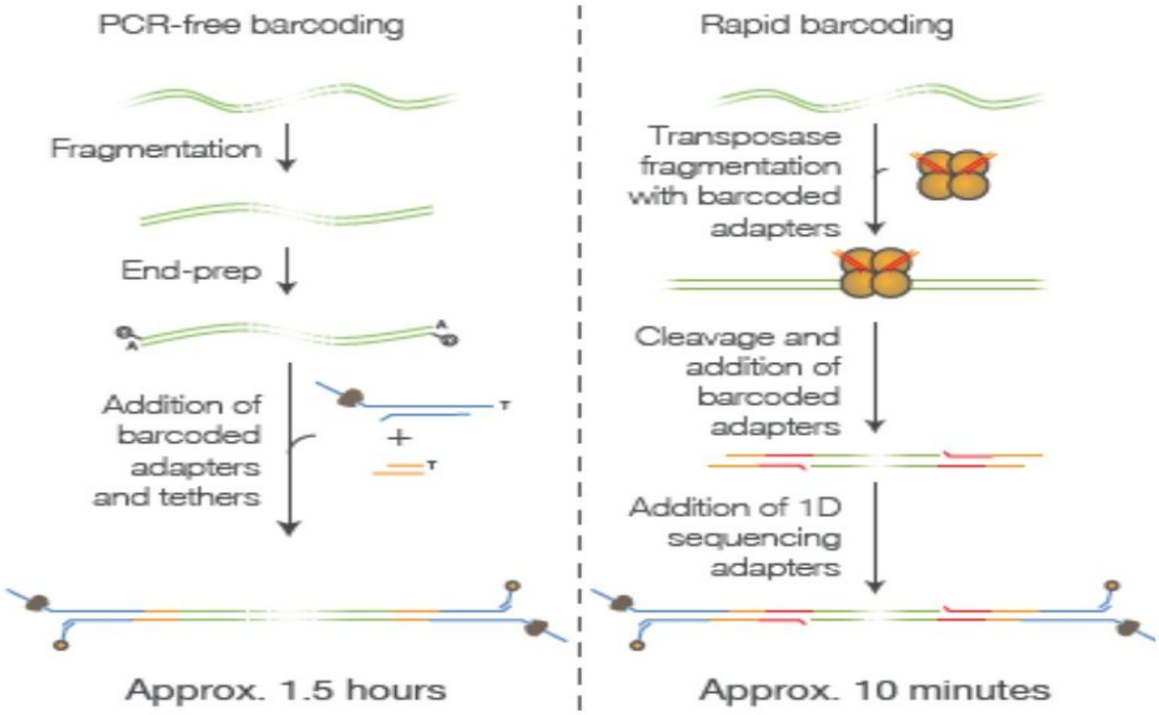


Figure 8 : Préparation des banques (ONT) (Séquençage à « longues lectures ONT », 2021).

3-Séquençage des coronavirus :

SARS-CoV-2 (coronavirus 2 du syndrome respiratoire aigu sévère) est un nouveau virus de la famille des Coronaviridae. Ce virus est à l'origine de la maladie infectieuse COVID-19. Le génome du SARS-CoV-2, c'est-à-dire l'ensemble de son information génétique, est composé de près de 30 000 nucléotides. Séquencer le virus revient à lire ces nucléotides, représentés par des lettres pour plus de clarté. Pour effectuer le séquençage, la réalisation d'un test PCR est nécessaire. En laboratoire, le matériel génétique est ensuite récupéré pour observer les séquences de nucléotides. Celles-ci peuvent ensuite être comparées aux autres séquences déjà répertoriées.

Cette technique permet de noter les différences éventuelles, d'identifier les variants et de les analyser. Cette discipline, relativement récente, est appelée phylodynamique. Elle avait déjà été utilisée par des chercheurs pour travailler sur Zika, Ebola ou le VIH. L'épidémie de Covid-19 prouve une fois de plus l'importance de la discipline. "L'idée de la phylodynamique, c'est que la façon dont les virus se propagent laisse des traces dans leur génome", explique au "Monde" Samuel Alizon, chercheur au CNRS (Larousserie, 2020).

Conclusion :

Grâce aux progrès rapides de la technologie de séquençage de l'ADN et à la réduction substantielle des coûts, ainsi qu'à l'augmentation substantielle du débit et de la précision. De

plus en plus d'organismes étant séquencés, un flot de données génétiques inonde le monde chaque jour. Mais une fois le séquençage terminé, on obtient des données de séquence brutes, ce qui les rend impossibles à lire. Elles doivent donc subir plusieurs étapes d'analyse : prétraitement des données pour éliminer les séquences adaptatrices et les lectures de mauvaise qualité, mise en correspondance des données avec un génome de référence ou alignement de novo des lectures de la séquence, et analyse de la séquence compilée.

L'analyse de la séquence peut inclure une grande variété d'évaluations bio-informatiques, y compris l'appel de variants génétiques pour la détection de SNP ou d'indels, la détection de nouveaux gènes ou d'éléments régulateurs, et l'évaluation des niveaux d'expression des transcriptions. Grâce à la diversité de ces informations utilisables, les scientifiques ont commencé à remarquer et à extraire les différences et les variantes dans ce qui était auparavant similaire.

Chapitre 2 : Les différentes variantes de Covid-19 et le processus de leur détermination

Introduction :

La plupart des personnes infectées par le virus COVID-19 souffrent d'une maladie respiratoire légère à modérée et se rétabliront sans nécessiter de traitement particulier. Les personnes âgées et celles qui présentent des pathologies sous-jacentes comme les maladies cardiovasculaires, le diabète, les maladies respiratoires chroniques et le cancer sont plus susceptibles de développer une maladie grave SARS-CoV-2, le virus à l'origine du COVID-19, a été identifié pour la première fois chez l'homme en décembre 2019. Au 6 novembre, il a touché plus de 48 millions de personnes, causant plus de 1, 2 millions de décès dans le monde. Bien que le virus soit considéré comme ancestralement lié aux chauves-souris, l'origine du virus et les hôtes intermédiaires du SARS-CoV-2 n'ont pas encore été identifiés. Malheureusement, plus le temps passe, plus ce virus se développe, en moins de 2 ans et il y a déjà plus d'une douzaine de ses variantes, même s'il n'y a pas de médicament permanent, mais nous avons réussi à en savoir beaucoup, chaque jour il y a une nouvelle découverte. L'analyse bioinformatique est la solution que les scientifiques ont utilisée pour étudier cette maladie (Clinkemaillié, 2021) (Mario Cannataro, 2021).

1-Variants du SARS-CoV-2 :

1-1-Les variantes :

Une variante est une nouvelle version du virus, légèrement différente, mais pas suffisamment pour que l'on estime qu'il s'agisse d'un nouveau virus (Vergnaud, 2021).

Un mutant (un variant) est un organisme ou une cellule qui présente une nouvelle caractéristique, acquise suite à l'apparition d'une modification dans son génome – on parle de « mutation ». Un tel événement peut être provoqué par certains agents chimiques ou physiques (les « mutagènes »), mais des mutations peuvent aussi survenir de manière spontanée, lors du processus qui permet à une cellule de se multiplier (Divers et variants – C'est quoi un mutant ?, 2021).

1-2-Les types des variantes :

Dans le séquençage de l'ADN, nous constatons qu'il existe cinq types de variants :

- Les variations nucléotidiques simples (SNV) (Justin M Zook, 2014).
- Les polymorphismes de nucléotides simples (SNP) (Michel, 2021) (Justin M Zook, 2014).
- Petites insertions et délétions (INDEL) (Nadine Hanna, 2005).

- Grands réarrangements chromosomiques - Variations structurelles (SV) (Gillet-Markowska, 2020).
- Variations du nombre de copies (CNV) (J.Tremblay1J.Raelson1F.Harvey1M.Ivanga1J.Chalmers2M.Woodward2S.Harrap3M.Marre4P.Hamet1, 2014) (Différents types de variantes : qu'est-ce que la variation génomique ?, 2016).

1-3-Les différentes variantes du SARS-CoV-2 :

Le coronavirus responsable du Covid-19 est un virus simple : c'est une coquille de protéines qui renferme du matériel génétique. De l'ARN dans ce cas, et non de l'ADN comme pour les humains. Le virus utilise des cellules humaines pour faire des copies de lui-même à partir de son ARN, une sorte de code qui recèle les instructions pour cette réplication (Vergnaud, 2021) (Robin, 2021).

Quatre variantes sont particulièrement étudiées :

- Le variant anglais, ou VOC2020, ou B.1.1.7, ou 20I/501Y.V1 :

Détecté pour la première fois au Royaume-Uni en octobre 2020, ce variant contient 17 mutations, dont 8 affectent la protéine spike. C'est cette protéine spike qui est utile au coronavirus pour se greffer aux récepteurs ACE2, qui permettent au virus d'infecter l'homme. C'est notamment sur la reconnaissance de cette protéine par notre organisme que sont basés les vaccins. La variante anglaise est connue pour provoquer une contagion accrue. Elle est également 64% plus mortelle que le virus classique, selon une étude publiée dans la revue BMJ (Robert Challen, 2021).

- La variante sud-africaine, ou B.1.351, ou 20H/501Y.V2 :

A la fin du mois d'avril, la part de la variante sud-africaine dans les contaminations en France était de 4,2%. Ce variant a été découvert dès le mois d'août 2020 en Afrique du Sud. Huit de ses mutations affectent également la fameuse protéine spike, dont la mutation N501Y, mais aussi les mutations E484K et K417N. Le variant sud-africain se distingue par une contagion plus élevée, sans être particulièrement plus mortel que le coronavirus classique. En revanche, les vaccins sont moins efficaces pour les protéger (Robert Challen, 2021) (Priam, 2021).

- La variante brésilienne ou 20J/501Y.V3 :

A la fin du mois d'avril, la part de contamination de la variante brésilienne en France était de 4,2%. Il existe deux variants brésiliens, nommés P1 et P2. Ils présentent les mutations

E484K et N501Y sur la protéine spike. Ce variant est inquiétant car il est non seulement très contagieux mais il semble pouvoir échapper à l'immunité conférée par une infection ou une vaccination antérieure. Il touche également une population plus jeune que d'habitude (Roberts, 2021).

- La variante indienne ou B.1.617 :

Le variant indien n'a pas été détecté en France métropolitaine à ce jour. Découvert en octobre 2020 dans le centre de l'Inde, le variant B.A.617 est responsable d'un foyer épidémique en Inde, où il est responsable de près de 10% des contaminations. Il a été classé dans la catégorie VOC (variant of concern). Ce variant résulte de quinze mutations. Ce sont deux d'entre elles qui inquiètent les chercheurs : les mutations L452R et E484Q, jamais vues ensemble auparavant, et qui concernent toutes deux la protéine spike. Le variant semble donc particulièrement contagieux et potentiellement résistant aux anticorps. Il existe de nombreux autres variants. En France, on a récemment parlé du variant dit " breton " : ce variant présente 9 mutations concernant la protéine spike. Si on parle de lui, c'est parce qu'il semble être beaucoup plus difficile à détecter à l'aide de tests PCR : un patient non détecté peut être plus à risque, car traité trop tard (Comité d'experts en vigie génomique du SRAS-CoV-2, 2021).

Figure 9 : le code génétique de chacune des variantes (Roberts, 2021)

2-Les pipelines d'analyse Des données NGS :

Un pipeline est un ensemble d'éléments de traitement de données connectés en série, où la sortie d'un élément est l'entrée du suivant.

Les pipelines de bio-informatique font partie intégrante du séquençage de prochaine génération (NGS). Le traitement des données de séquence brutes pour détecter les altérations

génomiques à un impact significatif sur la gestion de la maladie (authors, 2018) (Somak Roy, 2020).

2-1-Cadres des pipelines :

Afin d'automatiser le processus de création et de disposition des scripts pour former des flux de travail, des cadres de pipelines ont été créés. Cela a permis de réduire la charge des scientifiques qui devaient configurer manuellement des pipelines et les exécuter individuellement pour différents projets comportant des téraoctets de données. Les cadres de pipelines ont introduit de nouvelles fonctionnalités telles que les scripts reproductibles, le contrôle des versions et les fonctions de rapport (Leipzig, 2017).

Un cadre de pipeline (workflow) est construit pour les données, ce qui élimine de nombreuses étapes manuelles du processus de transition des données et permet un flux de données fluide, évolutif et automatisé d'une station à l'autre (Somak Roy 1, 2018).

Depuis l'apparition du nouveau COVID-19, tous les efforts sont consacrés à la recherche rapide de solutions à l'évolution rapide de la situation à l'aide de pipelines viraux.

La figure ci-dessous décrit le déroulement général d'un pipeline :

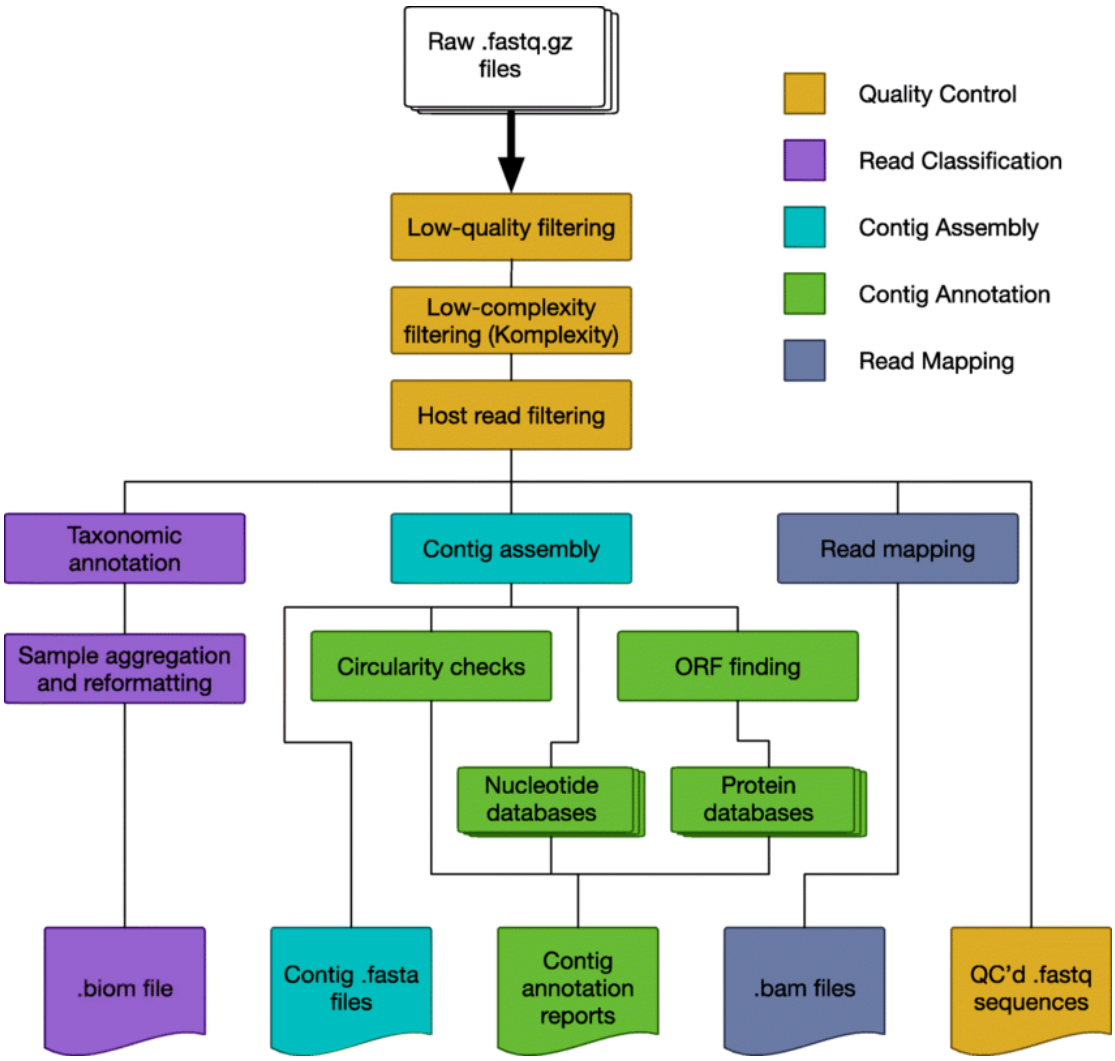


Figure 10 : entrées, sorties du processus analytique d'un pipeline (Erik L. Clarke, 2019)

3-Les différents flux de travail pour le COVID-19 :

Les outils bio-informatiques conçus explicitement pour le SARS-CoV-2 n'ont été développés que récemment, en réaction rapide à la nécessité de détecter, comprendre et traiter rapidement le COVID-19. Pour contrôler la pandémie actuelle de COVID-19, il est de la plus haute importance de comprendre l'évolution et la pathogenèse du virus (Somak Roy 1, 2018) (S.Masters, 2006).

3-1-V-PIPE :

V-pipe est un pipeline bio-informatique pour les données de séquençage viral. Il intègre divers outils informatiques pour l'analyse des données de séquençage viral à haut débit. Il permet l'analyse reproductible de la diversité génomique dans les populations de virus intra-hôte. V-pipe prend en entrée des données de lecture obtenues à partir d'une expérience de séquençage viral et produit, en une seule exécution du pipeline, divers fichiers de sortie

couvrant le contrôle de la qualité, l'alignement des lectures et l'inférence de la diversité génomique virale au niveau des variants nucléotidiques uniques et des haplotypes viraux. V-pipe comprend également une fonctionnalité d'analyse comparative fournissant un environnement normalisé pour les évaluations comparatives de différentes configurations de pipelines[1].

L'association V-pipe a développé une nouvelle méthode, appelée ngshmmalign, basée sur des modèles de Markov cachés de profil et adaptée aux petits génomes viraux très diversifiés pour l'alignement des lectures (Susana Posada-Céspedes, 2021).

3-2-Pipeline de détection des pathogènes ARN DRAGEN pour la détection et la surveillance des pathogènes viraux :

L'application Illumina® DRAGEN RNA Pathogen Detection App utilise une référence combinée humain plus le virus pour analyser les données sur les agents pathogènes et créer des FASTA de consensus. L'analyse est effectuée à l'aide d'une version personnalisée du pipeline DRAGEN RNA, qui est également disponible sur le matériel du serveur DRAGEN local[2].

4-Logiciels et outils de pipeline bio-informatique :

Nous présentons ici certains des logiciels que les bio-informaticiens utilisent manuellement pour créer leur propre pipeline :

Tableau 1 : la description des logiciels

[1] V-pipe. Sur : https://cbg-ethz.github.io/V-pipe/. Consulter : Mars 2021.
[2] DRAGEN RNA Pathogen Detection. Sur : https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-rna-pathogen-detection.html. Consulter : Mars 2021.

Logiciels	Description
Fast QC	FastQC vise à fournir un moyen simple d'effectuer des contrôles de qualité sur les données de séquence brutes provenant de pipelines de séquençage à haut débit. Il fournit un ensemble modulaire d'analyses que vous pouvez utiliser pour donner une idée rapide de vos données qui ont des problèmes dont vous devriez être conscient avant de faire toute analyse plus approfondie[3].
iSeqQC	Il s'agit d'un outil de contrôle de la qualité basé sur l'expression pour détecter les valeurs aberrantes produites par des effets de lot ou simplement en raison de la dissimilitude au sein d'un groupe phénotypique[4].
Trimmomatic	Est l'outil de prétraitement le plus flexible et le plus efficace, capable de traiter correctement les données paires. La valeur du prétraitement des lectures NGS est démontrée pour les tâches basées ou non sur des références. Il est démontré que Trimmomatic[5] produit des résultats qui sont au moins compétitifs, et dans de nombreux cas supérieurs, à ceux produits par d'autres outils, dans tous les scénarios testés (Anthony M. Bolger, Trimmomatic: a flexible trimmer for Illumina sequence data, 2014).
Bowtie	Bowtie est un outil d'alignement de courtes séquences d'ADN sur un génome de référence. L'algorithme Bowtie utilise la technique de la transformée de Burrows-Wheeler (BWT) et permet l'utilisation de plusieurs processeurs[6].
GATK	Le GATK est la norme de l'industrie pour l'identification des SNP et des indels dans les données d'ADN et d'ARNseq de la lignée germinale. Son champ d'application s'étend maintenant pour inclure l'appel de variantes courtes somatiques et pour s'attaquer au numéro de copie (CNV) et à la variation structurelle (SV). En plus des appelants de variantes eux-mêmes, le GATK comprend également de nombreux utilitaires pour effectuer des

[3] Fast QC. Sur : https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Consulter : Avril 2021.
[4] iSeqQC. Sur : https://github.com/gkumar09/iSeqQC et http://cancerwebpa.jefferson.edu/iSeqQC/ . Consulter : May 2021.
[5] Trimomatic. Sur : https://github.com/usadellab/Trimmomatic. Consulter : Mars 2021
[6] Bowtie. Sur : https://bioinformaticshome.com/tools/rna-seq/descriptions/Bowtie.html#:~:text=Bowtie%20is%20a%20tool%20for,the%20use%20of%20multiple%20CPUs. Consulter: May 2021.

	tâches connexes telles que le traitement et le contrôle de la qualité des données de séquençage à haut débit et regroupe la populaire boîte à outils Picard[7].
ANNOVAR	ANNOVAR est un outil logiciel efficace pour utiliser des informations de mise à jour pour annoter fonctionnellement les variantes génétiques détectées à partir de divers génomes (y compris le génome humain hg18, hg19, hg38, ainsi que souris, ver, mouche, levure et bien d'autres). Compte tenu d'une liste de variantes avec chromosome, position de départ, position de fin, nucléotide de référence et nucléotides observés[8].
Picard	Picard[9] est un ensemble d'outils en ligne de commande permettant de manipuler des données de séquençage à haut débit (HTS) et des formats tels que SAM/BAM/CRAM et VCF. La boîte à outils Picard est open-source sous la licence MIT et gratuite pour toutes les utilisations.
SeattlSeq	Le serveur d'annotation SeattleSeq fournit une annotation des SNV (variations mononucléotidiques) et des petits indels, connus et nouveaux. Cette annotation comprend les ID rs dbSNP, les noms de gènes et les numéros d'accession, les fonctions de variation (par exemple le malsens), la position des protéines et les changements d'acides aminés, les scores de conservation, les fréquences HapMap, les prédictions de PolyPhen et l'association clinique[10].
Freebayes	freebayes[11] est un détecteur bayésien de variants génétiques conçu pour trouver de petits polymorphismes, en particulier des SNP (polymorphismes mononucléotidiques), des indels (insertions et délétions), des MNP (polymorphismes multinucléotidiques) et des événements complexes (événements composites d'insertion et de substitution) plus petits que la longueur d'un alignement de séquençage à lecture courte. freebayes est basé sur les haplotypes, en ce sens qu'il appelle les variants sur la base des

[7] GATK. Sur: https://gatk.broadinstitute.org/hc/en-us. Consulter: May 2012.
[8] ANNOVAR. Sur: https://annovar.openbioinformatics.org/en/latest/. Consulter: Juin 2012.
[9] Picard. Sur: https://broadinstitute.github.io/picard/. Consulter: May 2021.
[10] SeattleSeq. Sur: https://snp.gs.washington.edu/SeattleSeqAnnotation154/HelpAbout.jsp. Consulter : Avril 2021.
[11] Freebayes. Sur : https://github.com/freebayes/freebayes. Consulter : Juin 2021.

	séquences littérales de lectures alignées sur une cible particulière, et non sur leur alignement précis.
TMAP	Le Torrent Mapping Alignment Program (TMAP)[12] est un logiciel d'alignement rapide et précis, optimisé spécifiquement pour les données Ion Torrent™. TMAP peut exécuter des cycles de cartographie/alignement de manière itérative, en appliquant différents algorithmes et paramètres aux lectures qui n'ont pas été alignées lors des itérations précédentes. Dans un flux de travail typique, une seule itération de cartographie/alignement est utilisée.
SnpEff	SnpEff[13] est un outil d'annotation de variants et de prédiction d'effets. Il annote et prédit les effets des variants génétiques (tels que les changements d'acides aminés). SnpEff a la capacité de fonctionner sur des systèmes Windows, Unix ou Mac. Outil rapide, il peut être intégré dans d'autres outils tels que Galaxy, GATK et GKNO, il peut être combiné avec d'autres boîtes à outils pour réduire les paramètres de prédiction des variantes, il liste également la façon dont la variante a été classée.
SnpSift	Il s'agit d'une boîte à outils qui complète l'outil d'annotation SnpEff et qui permet de filtrer et de manipuler les fichiers annotés. Une fois que la variante génomique est annotée, il est nécessaire de la filtrer afin de trouver les "variantes intéressantes / pertinentes". Étant donné la taille des fichiers de données, ce n'est pas une tâche triviale. SnpSift aide à effectuer cette manipulation et ce filtrage des fichiers VCF nécessaires à ce stade des pipelines de traitement des données[14].
VarSifter	Il s'agit d'un programme graphique Java[15] pour ordinateurs de bureau qui permet aux chercheurs ayant des compétences informatiques variées de trier, filtrer et passer en revue facilement et rapidement les données de variation de séquence. Une variété de filtres et un cadre d'interrogation personnalisé permettent un filtrage basé sur n'importe quelle combinaison d'informations

[12] TMAP. Sur : http://129.130.90.13/ion-docs/GUID-8059A3FE-EEBB-4416-B629-EAD74FA5FADE.html. Consulter : Juin 2021.
[13] SnpEff. Sur : https://pcingola.github.io/SnpEff/se_introduction/. Consulter: May 2021.
[14] SnpSift. Sur: https://pcingola.github.io/SnpEff/ss_introduction/. Consulter Avril 2021.
[15] VarSifter. Sur : https://research.nhgri.nih.gov/software/VarSifter/ / https://github.com/teerjk/VarSifter. Consulter : Juin 2021.

	sur les échantillons et les annotations. Il simplifie la lecture des données de variation à l'échelle de l'exome, soit dans un fichier texte délimité par des tabulations avec en-tête, soit dans un fichier VCF non compressé.

5-Quelques exemples d'utilisation des pipelines pour l'analyse des données du SARS-CoV-2 :

Exemple 1 :

Les données NGSm ont été analysées avec un pipeline interne. En bref, les lectures humaines de faible qualité (<Q30) et les lectures d'une longueur inférieure à cinquante nucléotides ont été éliminées par filtrage. Les lectures restantes ont été alignées sur le génome de référence du SARS-CoV-2 à l'aide de l'algorithme BWA-MEM (v0.7.15-r1140). Les séquences consensuelles ont été générées par une règle de majorité simple à l'aide d'un script PERL personnalisé. Ces séquences ont été utilisées comme référence de cartographie propre aux patients pour le réalignement ultérieur des lectures. La séquence consensuelle finale a été appelée à 10× en utilisant l'alignement sans clip.

Une inspection visuelle des alignements de lecture a été effectuée à l'aide d'Integrative Genomics Viewer (IGV) autour des régions présentant une baisse importante de la couverture afin de détecter et de définir les délétions potentielles. Pour l'approche par amplicon CleanPlex SARS-CoV-2, l'analyse des données a été réalisée conformément aux recommandations du fournisseur avec un pipeline développé par Sophia Genetics, comprenant une étape d'élagage des amorces Les données de séquençage de l'approche par capture RVOP ont été analysées à l'aide de la plateforme Bio-IT Dynamic Read Analysis for GENomics (DRAGEN ; v3.5.13 ; Illumina) d'Illumina. Enfin, les données de séquençage ONT ont été analysées en mettant en œuvre la bioinformatique recommandée développée par l'ARTIC. En bref, après l'exécution, l'appel de base a été effectué en utilisant les modèles de haute précision de Guppy (v3.5.2).

Comme les lectures chimériques sont la source prédominante d'erreurs d'assignation de codes-barres croisés ; ils ont suivi les recommandations d'ARTIC et démultiplexé en utilisant des paramètres Guppy_barcoder stricts pour s'assurer que les codes-barres sont présents à chaque extrémité du fragment. Le pipeline ARTIC a été exécuté en utilisant Minimap2 (v 2.17) pour l'alignement (Li 2018) et nanopolish (v0.13.2) pour l'appel de variants. Pour surmonter la limite de profondeur de 400x du pipeline bioinformatique ARTIC, ils ont généré indépendamment un graphique de couverture à partir des données BEDtools (v2.29.2), après

l'alignement Minimap2 des fichiers FASTQ filtrés générés par la commande guppyplex du pipeline bioinformatique ARTIC (Caroline Charre C. G., 2020).

Exemple 2 :

Ils ont utilisé MEGA 737 pour les alignements de séquences multiples afin de différencier les génomes du SARS-CoV-2 en fonction de leurs cadres de lecture ouverts (ORF). Un nettoyeur de séquences a été utilisé pour éliminer toutes les séquences ambiguës et de mauvaise qualité. La boîte à outils SeqKit38 a été utilisée pour intercepter les souches contenant des lacunes pour l'analyse des délétions. Les séquences contenant des codons d'arrêt internes ont été supprimées à l'aide de SEquence DAtaset builder. L'analyse de l'hétérogénéité des acides aminés a été réalisée avec Fingerprint, un outil d'analyse du profil des protéines basé sur le Web39. L'analyse des mutations d'acides aminés a été effectuée par un simple programme bio-python avec alignement par paires. Ils ont utilisé des diagrammes de Venn personnalisés pour créer les diagrammes de Venn. Enfin, les séquences alignées ont été visualisées à l'aide de Unipro-UGENE 1.26.1 pour visualiser les délétions par rapport au génome de référence40 (M. Rafiul Islam, 2019).

Exemple 3 :

Les 46 723 assemblages ont été alignés sur le génome de référence Wuhan-Hu-1 à l'aide de MAFFT v7.47152 mis en œuvre dans le pipeline d'alignement phylodynamique rapide fourni par Augur 6.3.0. Comme certains sites dans l'alignement ont été signalés comme des erreurs de séquençage putatives, ils ont suivi deux stratégies de masquage distinctes. La première stratégie de masquage est conçue pour tester l'impact de l'inclusion d'erreurs de séquençage putatives dans l'inférence phylogénétique, en masquant plusieurs sites du génome ainsi que les 55 premiers et les 100 derniers sites de l'alignement. Ils ont également employé une approche moins stricte, en suivant la stratégie de masquage employée par NextStrain, qui masque uniquement les positions 18 529, 29 849, 29 851 et 29 853 ainsi que les 130 premiers et les 50 derniers sites de l'alignement. Une liste complète des positions masquées est fournie dans les données supplémentaires. Cela a donné lieu à deux alignements masqués de 46 723 et 46 745 assemblages avec 12 706 et 12 807 SNP, respectivement.

Par la suite, pour les deux alignements, un arbre phylogénétique à maximum de vraisemblance a été construit en utilisant IQ-TREE 2.1.0 COVID release comme méthode de construction d'arbre53. Les phylogénies résultantes ont été visualisées et annotées à l'aide de ggtree v1.16.654. La numérotation des sites et la structure du génome sont fournies pour les

annotations disponibles (cadres de lecture ouverts (ORF) non chevauchants) en utilisant Wuhan-Hu-1 (NC_045512.2) comme référence (Lucy van Dorp, 2019).

Exemple 4 :

Les lectures Fastq démultiplexées, obtenues par séquençage shotgun ou par enrichissement de cible, ont été générées à partir de fichiers d'appel de base de séquençage brut à l'aide de BCL2Fastq v2.20.0, puis mappées au génome de référence Wuhan par BWA v0.7.17. Les statistiques d'alignement, telles que la couverture et les lectures mappées, ont été générées à l'aide de Picard 2.18.17. L'appel de variants a été effectué par GATK v3.8-1-0 et a été suivi par l'assemblage du génome du SARS-CoV-2 à l'aide de BCFtools v.1.3.1 (Divinlal Harilal, 2020).

Exemple 5 :

Les fichiers fastq ont été nettoyés avec Trimmomatic 0.36 puis soumis à une cartographie avec le génome de référence du SARS-CoV-2 en utilisant GS Reference Mapper 2.9, et une séquence consensus nommée hCoV-19_Morocco_OUA677_19_2020 a été obtenue. Les cadres de lecture ouverts (ORF) ont été prédits à l'aide de Generous Prime 2020.1 et annotés à l'aide de l'outil "CD-Search" de la base de données Conserved Domain. Les alignements de séquences avec le génome du SARS-CoV-2 ont été effectués avec CLC Genomics Workbench 20, en utilisant les outils "create Alignment 1.02" (Sanaâ Lemriss 1, 2020).

Conclusion :

La quantité de données générées exige des compétences en informatique et en bio-informatique pour gérer, analyser et interpréter l'énorme quantité de données NGS. Par conséquent, l'informatique et la bio-informatique NGS connaissent un développement considérable, qui ne peut avoir lieu qu'en raison de l'augmentation des capacités de calcul (matériel), ainsi que des algorithmes et des applications (logiciels) pour faciliter toutes les étapes nécessaires : du traitement des données brutes à l'analyse plus détaillée des données et à l'interprétation des variantes dans un contexte clinique. L'objectif global de chaque analyse est fondamentalement le même quelle que soit la plateforme NGS, cependant, chaque plateforme a ses propres particularités et spécificités, ses limites, ses caractéristiques et ses avantages. C'est donc en étudiant et en s'inspirant des expériences passées et en notant leurs points faibles que nous avons créé notre pipeline.

Chapitre 3 : matériels et méthodes

1-Matériels :

1-1-Les données :

Le premier fichier avec lequel nous allons travailler est celui de COVID du SARS qui ont été séquencés à partir de l'organisme coronavirus 2 du syndrome respiratoire aigu sévère, avec plus de 25 000 données de rangées avec différentes méthodes de séquençage. En raison de la variabilité du virus, nous avons choisi 2 types différents de variants de SARS-CoV-2 comme échantillons. Ces échantillons ont été téléchargés de la base de données NCBI au format SRA.

Tableau 2 : Données utilisées pour l'analyse

Séquence	SRR13704280	ERR6019891
Nombre de spots	1,950,373	3,237,847
Taille	585.1M	643.6M
Nombre de bases brute	189.5Mb	316.9Mb
Publié	2021-02-12	2021-06-09
ID	13220840	14770904
Instrument	Illimuna NovaSeq 6000	Ion Torrent S5
Contenu GC	40,5 %	38.8%
Disposition	Paire	Unique

La deuxième donnée est l'isolat de référence Wuhan-Hu-1 du coronavirus 2 du syndrome respiratoire aigu sévère, dont le génome complet a été téléchargé dans la base de données NCBI au format de fichier FASTA.

Tableau 3 : Données utilisées pour l'alignement et l'annotation des variantes

IDs	Description	Taille
[UID] 15851418 [GenBank] 15851418 [RefSeq] 15851438	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.	29,903 pb

1-2-Les environnements:

Unix/Linux: Ubuntu:

Linux est le précurseur des technologies à code source ouvert. Le système d'exploitation de bureau Ubuntu est développé par la communauté Ubuntu et l'écosystème Linux. Il s'agit du système d'exploitation à code source ouvert le plus populaire au monde, tant pour le développement que pour le déploiement. Il s'agit de la meilleure distribution Linux[16] de bureau, mais assurez-vous que Ubuntu[17] prend en charge un grand nombre de matériels. Les utilisateurs d'Ubuntu ont une liberté totale lorsqu'il s'agit de mettre à niveau leur matériel.

Galaxy :

Galaxy[18] est une plateforme ouverte, basée sur le web, pour la recherche biomédicale à forte intensité de données. Aucune expérience de programmation n'est requise pour télécharger facilement des données, exécuter des outils et des flux de travail complexes et visualiser les résultats. Galaxy capture les informations afin que tout utilisateur puisse répéter et comprendre une analyse computationnelle complète, des paramètres de l'outil à l'arbre de dépendance. Les utilisateurs partagent et publient leurs historiques, leurs flux de travail et leurs visualisations via le web.

2-Les outils :

2-1-SRA-Tools :

Sequence Read Archive (SRA)[19] stocke des données de séquence brutes provenant des technologies de séquençage de "nouvelle génération", notamment Illumina, 454, Ion Torrent, Complete Genomics, PacBio et Oxford Nanopores. En plus des données de séquence brutes.

[16] LINUX. Sur : https://www.linux.org. Consulter : Mars 2021.
[17] UBUNTU. Sur : https://ubuntu.com/desktop/developers. Consulter : Mars 2021.
[18] GALAXY. Sur : https://galaxyproject.org. / / http://www.bioinformatics.nl/galaxy. Consulter : Mars 2021.
[19] SRA-Tools. Sur : https://ncbi.github.io/sra-tools/ / https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement. Consulter : Mars 2021.

Le SRA Toolkit et SDK du NCBI est une collection d'outils et de bibliothèques permettant d'utiliser les données des INSDC Sequence Read Archives. Il peut être utilisé sur des systèmes Mac, Windows et Linux.

Tableau 4 : Entrées et sorties de SRA

Convertir le format SRA en	Convertir le format SRA à partir de
- ABI SOLiD natif - fasta - fastq - sff - sam -bam - Originaly from Illumina	- pairs fastq or fasta / qual - AB SOLiD -SRF - AB SOLiD natif - Illumina SRF - Originaly from Illumina - sff - Bam aligned

2-2-Fastp :

Un outil conçu pour fournir un prétraitement rapide tout-en-un pour les fichiers Fastq. Cet outil est développé en C++ avec support du multithreading pour offrir de hautes performances. Il peut être téléchargé gratuitement pour Windows (Bioconda), Mac ou Linux. Il peut lire les séquences en paire ou en simple extrémité[20].

Caractéristiques :

- Profilage de qualité complet avant et après le filtrage des données (courbes de qualité, contenu de base, KMER, Q20/Q30, rapport GC, duplication, contenu de l'adaptateur...).
- Filtrer les mauvaises lectures (trop mauvaise qualité, trop courte ou trop de N.…).
- Couper des bases de faible qualité par lecture dans ses 5' et 3' en évaluant la qualité moyenne à partir d'une fenêtre coulissante (comme Trimmomatic mais plus rapide).
- Couper toutes les lectures à l'avant et à la queue.

[20] Fastp. Sur : https://github.com/OpenGene/fastp. Consulter : Avril 2021.

- Adaptateurs de coupe. Les séquences d'adaptateur peuvent être détectées automatiquement, ce qui signifie que vous n'avez pas à entrer les séquences d'adaptateur pour les couper.
- Corriger les paires de bases non appariées dans les régions superposées des lectures d'extrémité appariées, si une base est de haute qualité tandis que l'autre est de très faible qualité.
- Trim polyg aux extrémités de 3', ce qui est couramment observé dans les données novaseq/nextseq. Coupez polyX aux extrémités de 3' pour éliminer les résidus polyX indésirables (c.-à-d. Résidus polyA pour les données d'ARNm-Seq).
- Prétraiter les données activées par l'identificateur moléculaire unique (UMI), déplacer l'UMI vers le nom de la séquence.
- Rapporter le résultat du format JSON pour une interprétation plus approfondie.
- Visualiser les résultats du contrôle de la qualité et du filtrage sur une seule page HTML (comme FASTQC, mais plus rapide et plus informatif).
- Diviser la sortie en plusieurs fichiers (0001. R1.gz, 0002. R1.gz...) Pour prendre en charge le traitement parallèle. Deux modes peuvent être utilisés, limitant le nombre total de fichiers fractionnés ou limitant les lignes de chaque fichier fractionné.
- Prennent en charge les lectures longues (données des appareils Pacbio / Nanopore).
- Soutenir la lecture de STDIN et l'écriture sur STDOUT.
- Soutenir l'entrée entrelacée.

Nous avons évalué la performance de la vitesse et de la qualité du fastp par rapport à fastQC. Les résultats indiquent que fastp est beaucoup plus rapide que ses homologues et fournit le filtrage des données de la plus haute qualité. En raison de sa vitesse élevée et de ses fonctions riches en contrôle et filtrage de la qualité des fichiers fastq, nous l'avons choisi.

2-3-BWA :

BWA[21] utilise l'algorithme de transformation de Burrows-Wheeler (un algorithme de transformation de données qui restructure les données pour les rendre plus compressibles), principalement utilisé pour Illumina, initialement développé pour préparer les données à des techniques de compression telles que bzip2, c'est un aligneur rapide et efficient performant pour les lectures courtes et longues. Il se compose de trois algorithmes : BWA-backtrack, BWA-SW et BWA-MEM. BWA fonctionne avec divers types de données de séquences d'ADN, bien que l'algorithme et le paramètre optimaux puissent varier. Il produit des

[21] BWA. Sur: https://github.com/lh3/bwa. Consulter: May 2021.

alignements au format SAM qui est pris en charge par plusieurs appelants SNP génériques tels que Samtools et GATK. Tous les algorithmes BWA fonctionnent avec un génome d'une longueur totale supérieure à 4 Go.

Bwa-mem Il produit un alignement identique au BWA et est ~1.3-3.1x plus rapide selon le cas d'utilisation, le jeu de données et la machine en fonctionnement. BWA-MEM est plus tolérant aux erreurs avec des séquences de requête plus longues, car la probabilité de manquer toutes les graines est faible. Comme il est montré ci-dessus, avec des paramètres autres que ceux par défaut.

2-4-Samtools :

Samtools[22] est un ensemble d'utilitaires qui manipulent les alignements dans les formats SAM (Sequence Alignment/Map), BAM et CRAM. Il convertit entre les formats, fait du tri, de la fusion et de l'indexation, peut récupérer rapidement les lectures dans n'importe quelle région et générer des alignements dans un format par position.

2-5-BCFtools :

BCFtools[23] est un ensemble d'utilitaires qui manipulent les appels de variante dans le format d'appel de variante (VCF) et son homologue 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, VCF bgzipped et BCF avec un type de fichier détecté automatiquement, même lors de la diffusion en continu à partir d'un tuyau. Le FVC et le FBC indexés fonctionneront dans toutes les situations. Les VCF et BCF non indexés et les flux 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.

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

2-6-IGV :

La visionneuse génomique intégrative (IGV)[24] est un outil interactif haute performance, facile à utiliser pour l'exploration visuelle des données génomiques. Il prend en charge l'intégration flexible de tous les types courants de données et de métadonnées génomiques,

[22] SAMtools. Sur: http://samtools.sourceforge.net / http://www.htslib.org/doc/samtools.html. Consulter: May 2021.
[23] BCFtools. Sur: https://samtools.github.io/bcftools/bcftools.html. Consulter : Juin 2021.
[24] IGV. Sur : http://software.broadinstitute.org/software/igv/home. Consulter : septembre 2021.

générées par les chercheurs ou accessibles au public, chargées à partir de sources locales ou cloud.

Nous l'avons choisi car IGV prend en charge le chargement flexible d'ensembles de données locaux et distants, et est optimisé pour fournir une visualisation et une exploration de données hautes performances sur les systèmes de bureau standard, il fournit aussi un support étendu pour la visualisation des variantes stockées au format VCF. Ce format permet l'encodage des appels de variante (SNP, indels et réarrangements génomiques) ainsi que les informations de génotype à l'appui pour des échantillons individuels. Les échantillons peuvent également être annotés avec des informations sur les attributs, y compris des informations généalogiques et familiales. L'IGV utilise ces annotations pour regrouper, trier et filtrer les échantillons.

3-Méthodes :

L'objectif global de chaque analyse est fondamentalement le même quelle que soit la plateforme NGS. Cependant, chaque plateforme a ses propres particularités et spécificités. Pour des raisons de simplicité, nous nous sommes concentrés sur deux des principales plateformes commerciales de 2e génération : Illumina, Ion torrent.

3-1-Illumina :

Pour la plateforme Illumina[25], le principe de détection du signal repose sur la fluorescence. Par conséquent, l'appel des bases est apparemment beaucoup plus simple, se fait directement à partir des mesures d'intensité du signal fluorescent résultant des nucléotides incorporés au cours de chaque cycle. Illumina affirme que sa technologie SBS fournit le plus haut pourcentage de lectures sans erreur.

Alta-Cyclic et Bustard comptent parmi les premiers appelants de bases pour la plateforme d'Illumina. À l'heure actuelle, il existe de nombreux autres appelants de bases qui se différencient par les méthodes statistiques et informatiques utilisées pour déduire la base correcte. Malgré cette variabilité, l'algorithme d'appel de base le plus largement utilisé est le Bustard et plusieurs algorithmes d'appel de base ont été construits en utilisant le Bustard comme point de départ. Globalement, l'algorithme de Bustard est basé sur la conversion des signaux de fluorescence en données de séquence réelles.

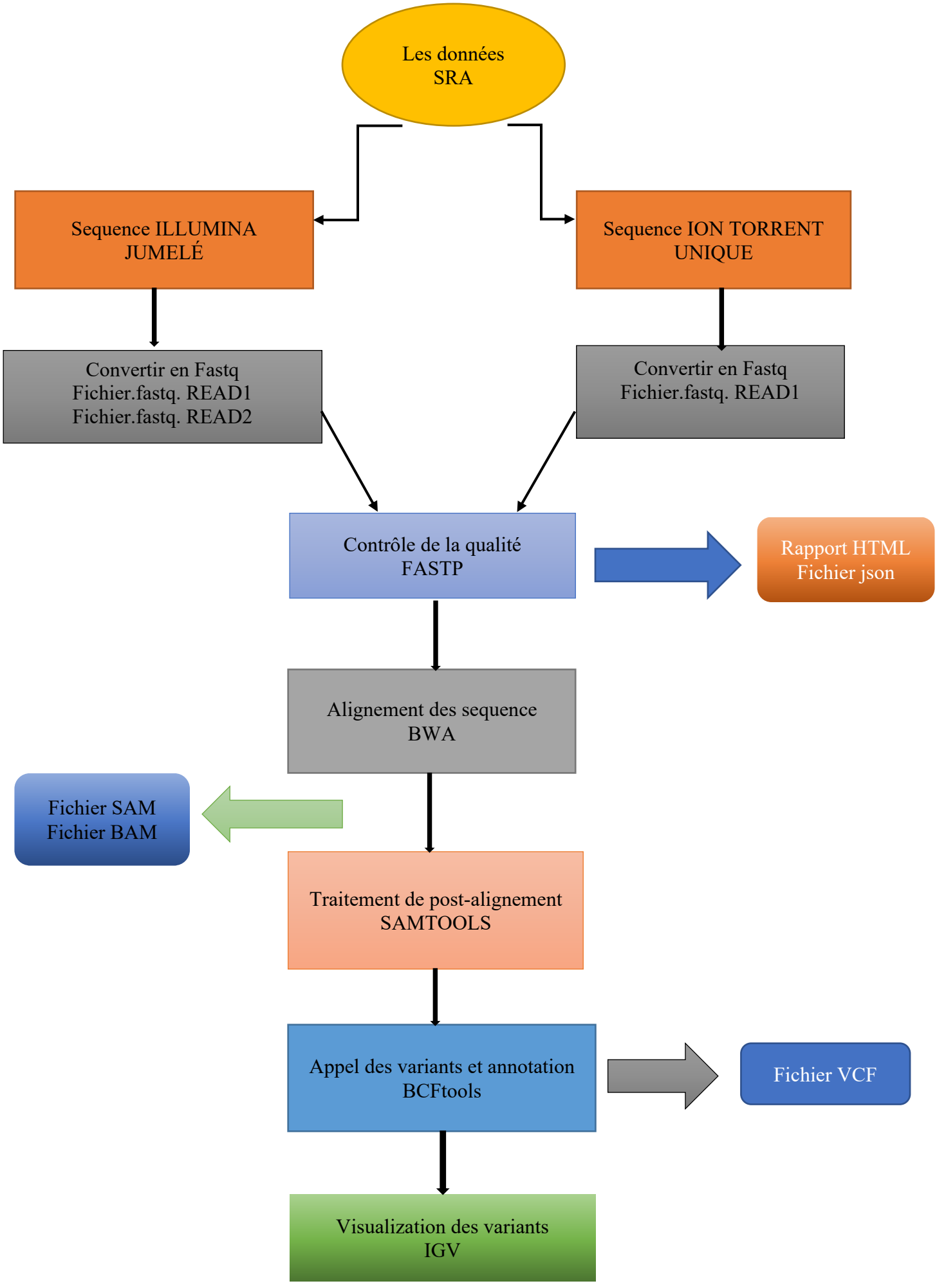
3-2-Ion Torrent :

[25] Illumina. Sur : https://www.ncbi.nlm.nih.gov/probe/docs/distrillumina/. Consulter : Avril 2021.

Ion Torrent est l'une des plateformes les plus fréquemment utilisées dans la recherche et l'industrie des soins de santé. Malgré ses nombreux avantages, les artefacts spécifiques à la plateforme compliquent la séparation efficace des vrais variants et des erreurs, en particulier pour les variants à faible fréquence allélique. L'utilisation du séquençage massivement parallèle permet de générer des milliers de méga-bases d'informations de séquence par jour, ouvrant ainsi la voie à de nouvelles études de recherche qui étaient auparavant difficiles à réaliser de manière pratique. Alimentée par des puces à semi-conducteurs, la technologie de séquençage de nouvelle génération d'Ion Torrent permet de mettre en œuvre un flux de travail rapide et simple qui répond aux besoins de la recherche dans de nombreuses applications, notamment les maladies héréditaires, l'oncologie, les maladies infectieuses et bien d'autres encore (Sunguk Shin 1, 2018).

4-Le processus du pipeline :

Les processus de pipeline sont les étapes nécessaires à l'analyse des données que nous avons, après avoir obtenu les données, nous devons faire le prétraitement (QC, Trimming, Filetring...), l'alignement, l'appel et l'annotation des variants, et enfin nous les visualisons.
Le schéma suivant résume le processus du pipeline proposé :


Les données
SRA
Sequence ILLUMINA
JUMELÉ
Sequence ION TORRENT
UNIQUE
Convertir en Fastq
Fichier.fastq. READ1
Fichier.fastq. READ2
Convertir en Fastq
Fichier.fastq. READ1
Contrôle de la qualité
FASTP
Rapport HTML
Fichier json
Alignement des sequence
BWA
Fichier SAM
Fichier BAM
Traitement de post-alignement
SAMTOOLS
Appel des variants et annotation
BCFtools
Fichier VCF
Visualization des variants
IGV

4-1-Analyse primaire :

L'analyse primaire des données consiste à détecter et à analyser les données brutes (analyse du signal), à cibler la génération de lectures de séquençage lisibles (appel de base) et à noter la qualité des bases. Les sorties typiques de cette analyse primaire sont des fichiers FASTQ pour illumina et ion torrent après avoir télécharger les données brutes avec SRAtoolkit.

4-2-Contrôle de la qualité : Filtrage et découpage des lectures :

Le contrôle de la qualité[26] dans toute technologie de séquençage à haut débit est une étape critique qui, si elle est négligée, peut compromettre une expérience et les conclusions qui en découlent. Un certain nombre de méthodes existent pour identifier les biais pendant le séquençage ou l'alignement. Pour cette étape, nous avons utilisé Fastp vs fastQC et nous avons comparé les résultats et on a décidé de choisir Fastp (Gaurav Kumar, 2020) (Anthony M. Bolger, Trimmomatic: a flexible trimmer for Illumina sequence data, 2014).

4-3-Alignement de séquences :

La méthode d'assemblage préférentielle lorsque le génome de référence est connu est l'alignement par rapport au génome de référence. Un algorithme de cartographie essaiera de localiser un emplacement dans la séquence de référence qui correspond à la lecture, en tolérant un certain nombre de mésappariements pour permettre la détection des variations de sous-séquences. Pour cette étape, nous avons utilisé BWA-aligner. (Nuno A Fonseca 1, 2012) (Knut Reinert 1, 2015) (Assembly algorithms for next-generation sequencing data, 2010).

4-4-Traitement de post-alignement :

Le traitement post-alignement est recommandé avant d'effectuer l'appel de variante. Son objectif est d'augmenter la précision de l'appel de variante et la qualité du processus en aval, en réduisant les artefacts d'appel de base et d'alignement. Des outils sont nécessaires pour le filtering (suppression) des lectures dupliquées, le réalignement local intensif (principalement près des INDEL) et le recalibrage du score de qualité des bases, et nous avons donc utilisé Samtools pour cela (Shulan Tian, 2016).

4-5-Appel des variants :

L'étape d'appel des variants a pour objectif principal d'identifier les variants en utilisant le fichier BAM post-traité. Plusieurs outils sont disponibles pour l'appel des variants, certains identifient les variants en se basant sur le nombre d'appels de base de haute confiance qui ne

[26] Le contrôle de la qualité. Sur : https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html. Consulter : Avril 2021.

correspondent pas à la position du génome de référence d'intérêt. D'autres utilisent des méthodes statistiques bayésiennes, de vraisemblance ou d'apprentissage automatique qui utilisent des paramètres de facteurs, tels que les scores de qualité des bases et de la cartographie, pour identifier les différences entre les variantes. Ici, nous avons utilisé le logiciel BCFtools (Kuhn, Stange, Herold, Thiede, & Roeder, 2018).

4-6-Annotation des variants :

L'annotation des variants est le processus d'attribution d'informations fonctionnelles aux variants d'ADN. De nombreux types d'informations peuvent être associés aux variants, depuis les mesures de conservation de la séquence jusqu'aux prédictions sur l'effet d'un variant sur la structure et la fonction des protéines. L'un des niveaux les plus fondamentaux de l'annotation des variants, consiste à catégoriser chaque variant en fonction de sa relation avec les séquences codantes du génome et de la façon dont il peut modifier la séquence codante et affecter le produit du gène. Pour cela nous avons utilisé des commandes spécifiques dans BCFtools (Gregory M Cooper 1 & Eric D Green, 2005) (Prateek Kumar, 2009) (Jana Marie Schwarz, 2010).

4-7-Filtrage, hiérarchisation et visualisation des variantes :

La visualisation des variations est une étape critique pour trouver des associations entre les SVs et les traits ou les maladies. Nous avons utilisé l'outil IGV (Kasahara, 2020).

Chapitre 4 : résultats et discussion

1-Résultats :

Après avoir choisi les outils et tracer le chemin de travail, nous avons mis en œuvre notre pipeline. Nous présenterons dans ce chapitre nos résultats pour les séquences Illumina et Ion Torrent. Nous suivrons le pipeline depuis le début du processus, avec les séquences brutes, le prétraitement, les alignements, jusqu'à l'annotation et la visualisation des données. Nous expliquerons également nos choix pour chaque outil/logiciel.

1-1-Résultats du prétraitement :

Nous avons utilisé fastp pour les deux séquences Illumina et Ion Torrent pour fournir un prétraitement rapide pour les fichiers fastq.

La sortie est un fastq /fastq.gz (par choix), un rapport Json et un rapport Html qui nous donne l'avant et l'après utilisation de fastp. Nous avons capturé certains des résultats du rapport Html et du fichier Json.

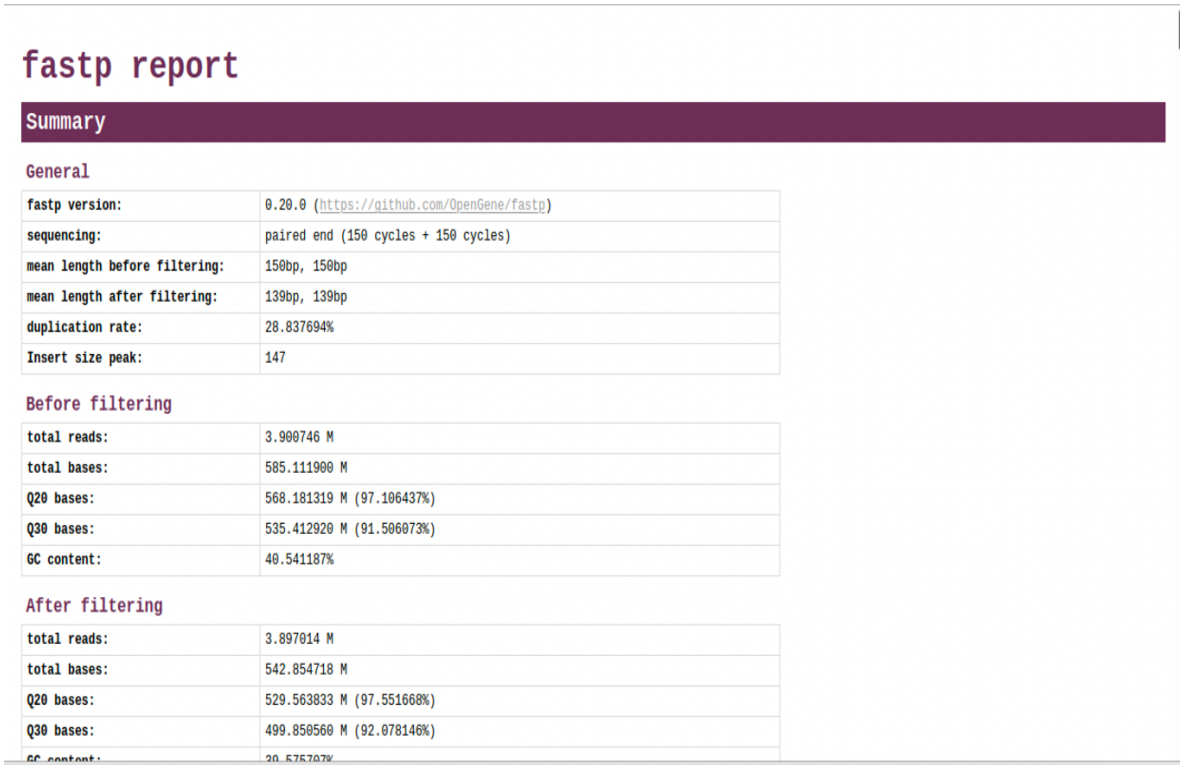
Fichier Json :

Figure 11 : fichier Json pour Illumina.


```
{
	"summary": {
		"before_filtering": {
			"total_reads":3237847,
			"total_bases":643584031,
			"q20_bases":580971556,
			"q30_bases":33764610,
			"q20_rate":0.902713,
			"q30_rate":0.0524634,
			"read1_mean_length":198,
			"gc_content":0.387743
		},
		"after_filtering": {
			"total_reads":3230059,
			"total_bases":642981575,
			"q20_bases":580694805,
			"q30_bases":33761984,
			"q20_rate":0.903128,
			"q30_rate":0.0525085,
			"read1_mean_length":199,
			"gc_content":0.387752
		}
	},
	"filtering_result": {
		"passed_filter_reads": 3230059,
		"low_quality_reads": 7788,
		"too_many_N_reads": 0,
		"too_short_reads": 0,
		"too_long_reads": 0
	},
	"duplication": {
		"rate": 0.288653,
		"histogram": [26767,2466,881,432,231,164,105,81,55,48,31,17,12,14,6,7,11,7,9,9,3,5,1,1,1,2,1,5,3,2,19],
		"mean_gc":
```


Figure 12 : fichier Json pour Ion Torrent.

Rapport HTML :



fastp report

Summary

General

fastp version:	0.20.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (150 cycles + 150 cycles)
mean length before filtering:	150bp, 150bp
mean length after filtering:	139bp, 139bp
duplication rate:	28.837694%
Insert size peak:	147

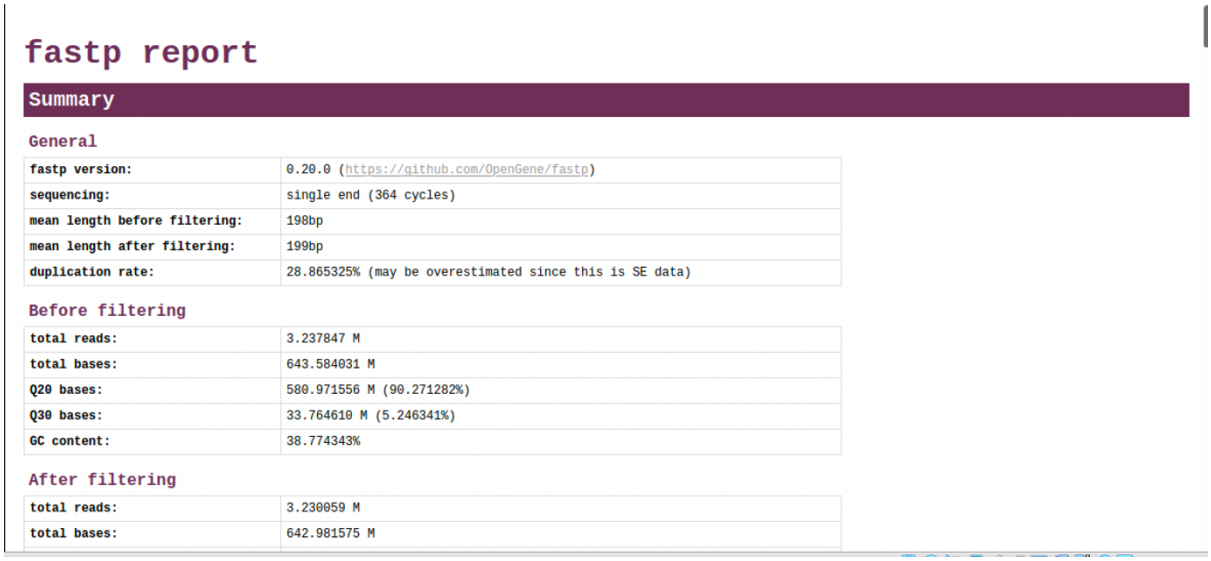
Before filtering

total reads:	3.900746 M
total bases:	585.111900 M
Q20 bases:	568.181319 M (97.106437%)
Q30 bases:	535.412920 M (91.506073%)
GC content:	40.541187%

After filtering

total reads:	3.897014 M
total bases:	542.854718 M
Q20 bases:	529.563833 M (97.551668%)
Q30 bases:	499.850560 M (92.078146%)
GC content:	39.575707%

Figuer 13 : rapport HTML pour Illumina.



fastp report

Summary

General

fastp version:	0.20.0 (https://github.com/OpenGene/fastp)
sequencing:	single end (364 cycles)
mean length before filtering:	198bp
mean length after filtering:	199bp
duplication rate:	28.865325% (may be overestimated since this is SE data)

Before filtering

total reads:	3.237847 M
total bases:	643.584031 M
Q20 bases:	580.971556 M (90.271282%)
Q30 bases:	33.764610 M (5.246341%)
GC content:	38.774343%

After filtering

total reads:	3.230059 M
total bases:	642.981575 M

Figure 14 : rapport HTML pour Ion Torrent.

La qualité des lectures que le Fastp détecte avant et après filtrage est présentée ci-dessous pour les données Illumina et pour Ion Torrent.

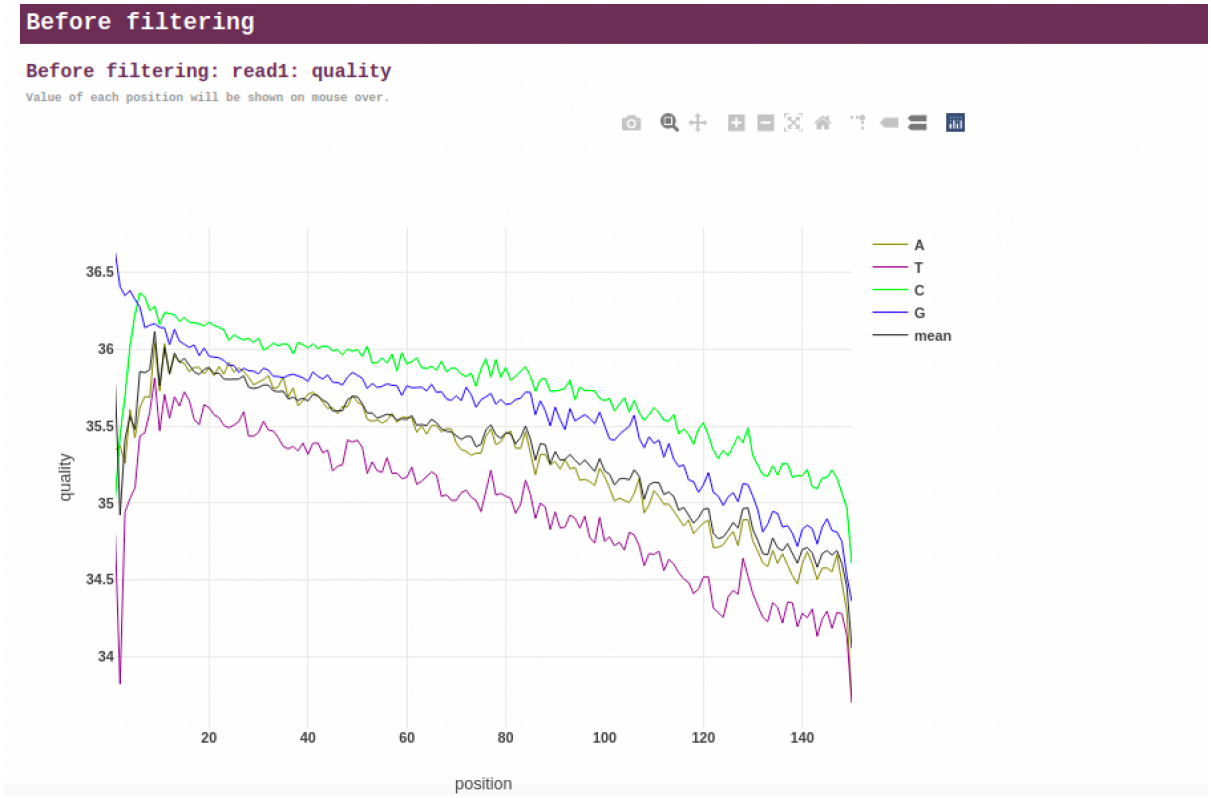


Figure 15 : qualité de lecture Illumina avant filtrage.

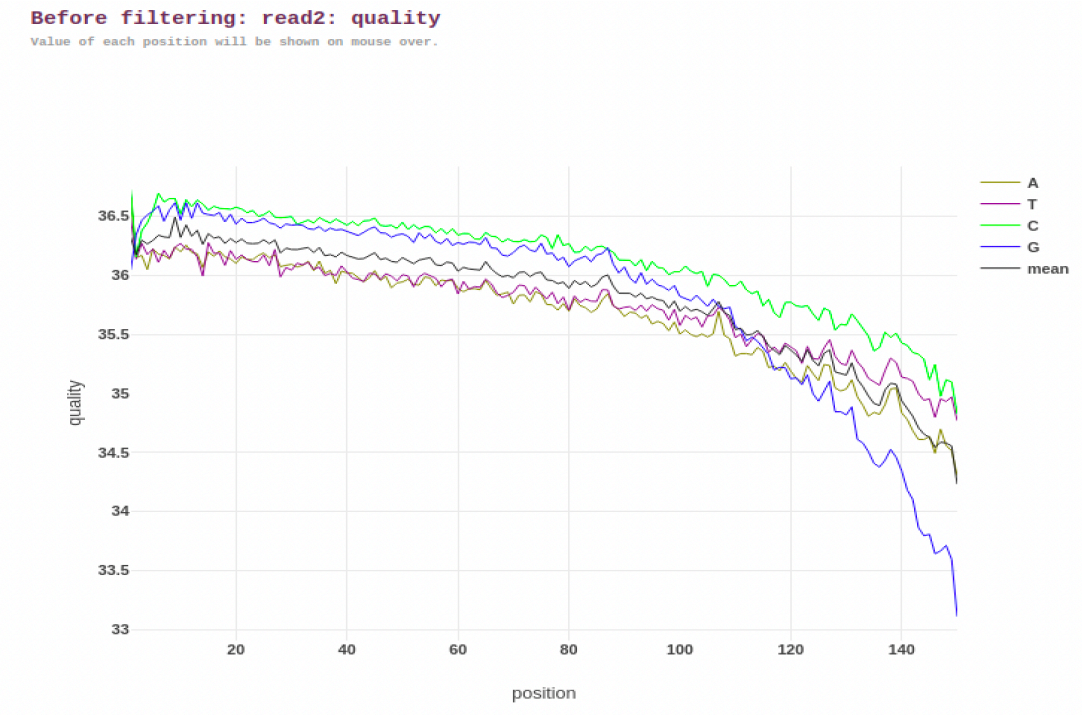


Figure 16 : qualité de lecture Illumina avant filtrage.

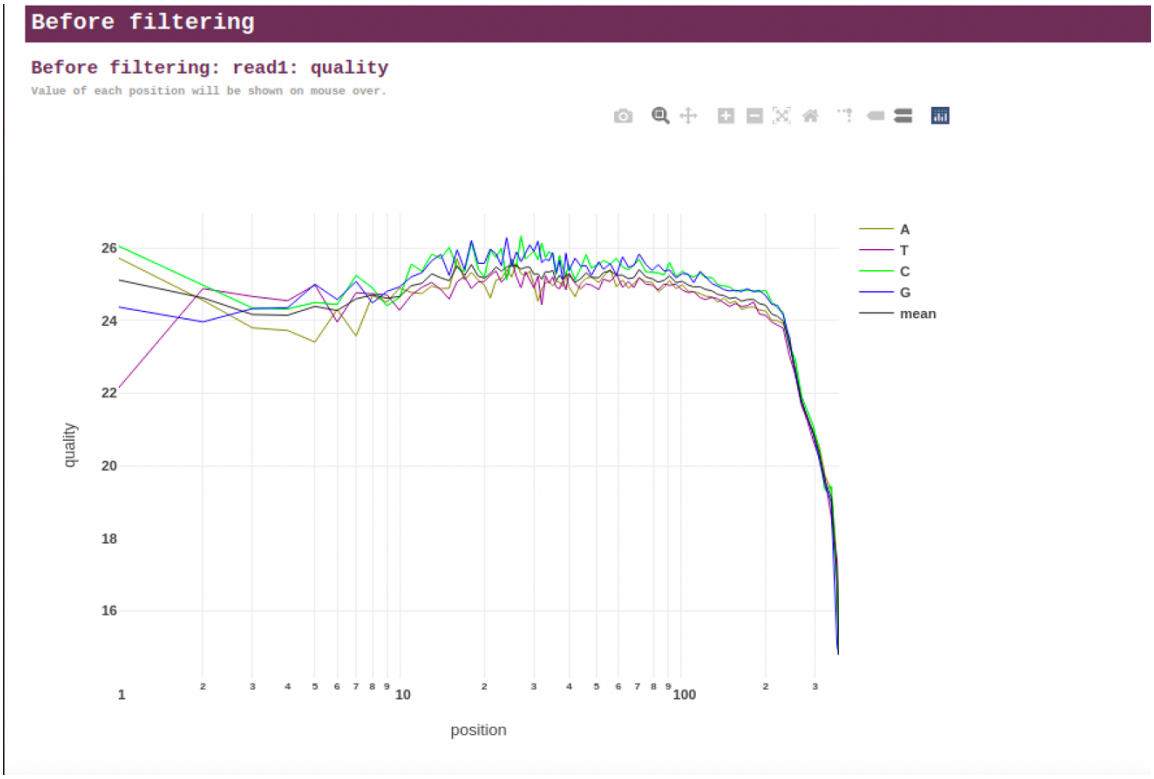


Figure 17 : qualité de lecture Ion Torrent avant filtrage.

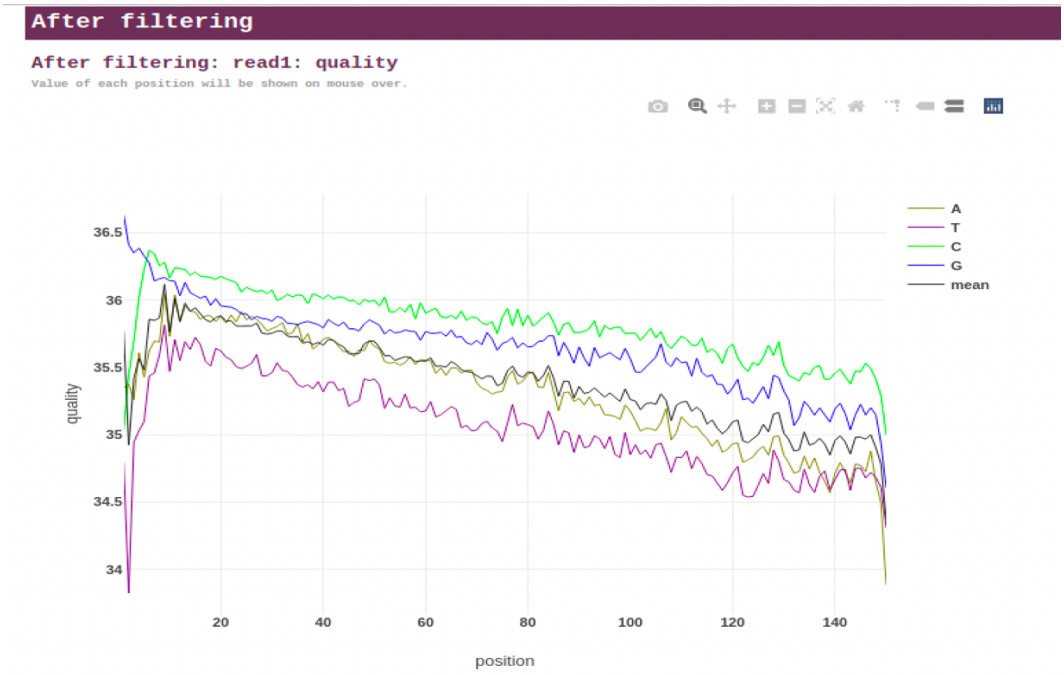


Figure 18 : qualité de lecture Illumina après filtrage.

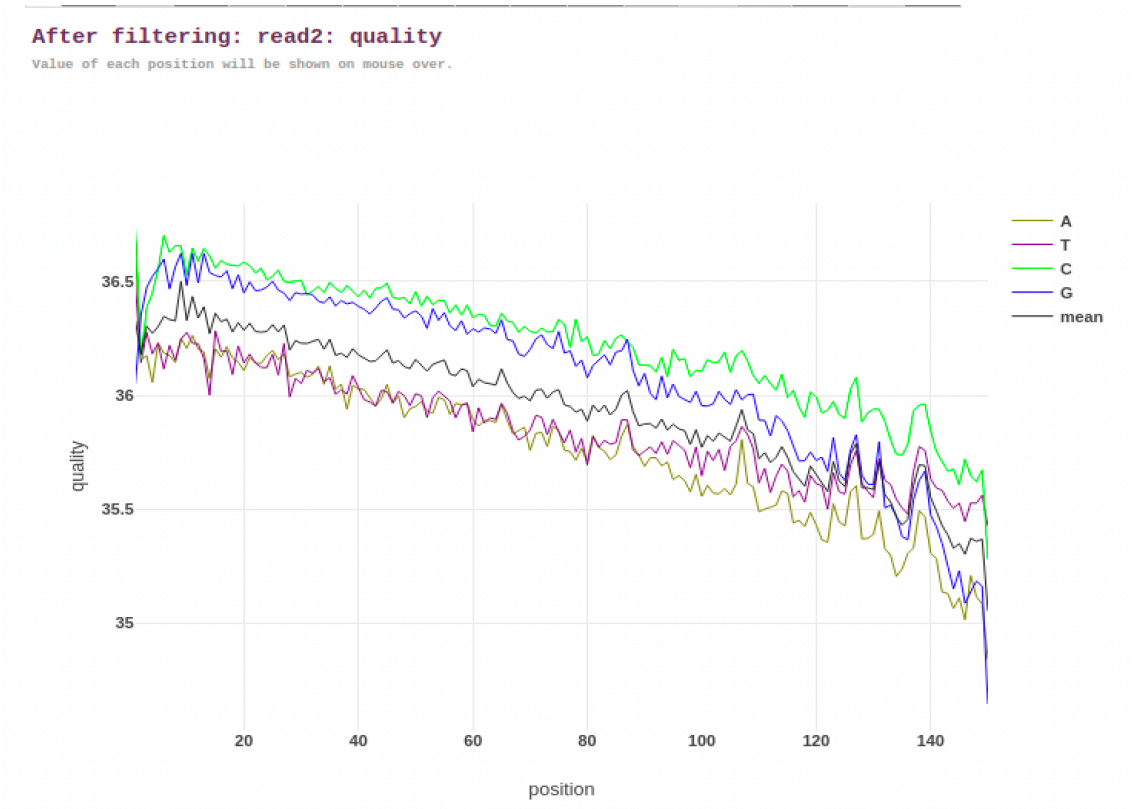


Figure 19 : qualité de lecture Illumina après filtrage.

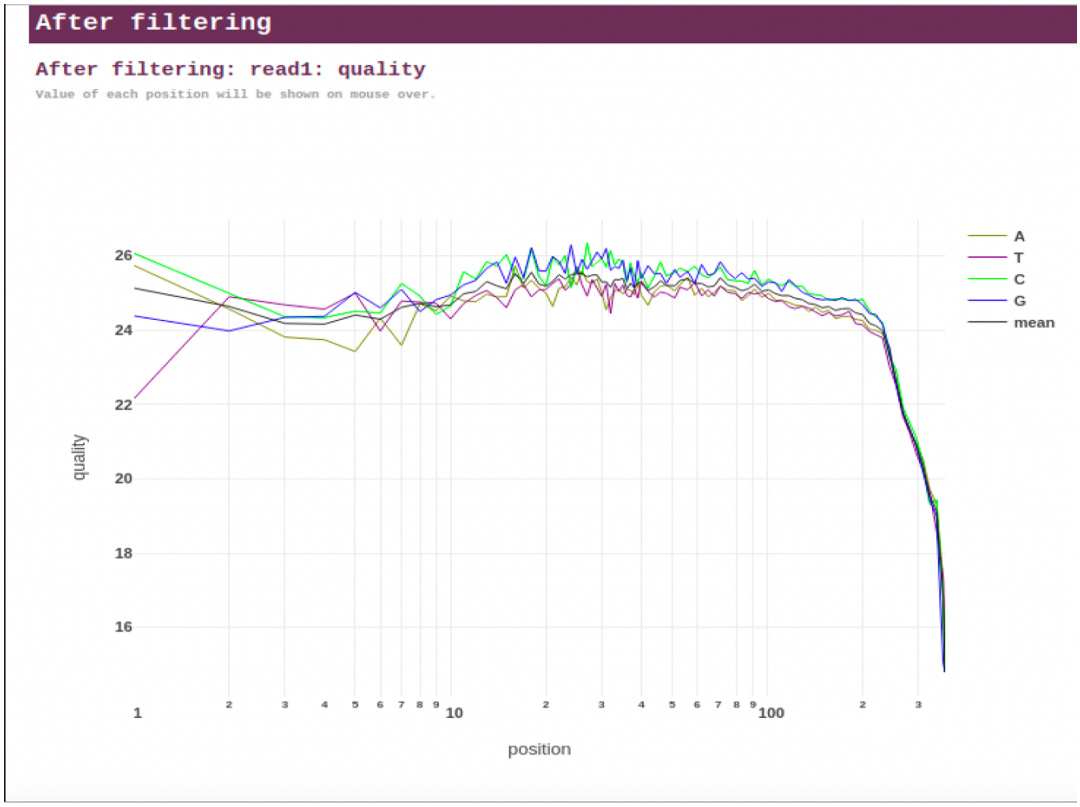


Figure 20 : qualité de lecture Ion Torrent après filtrage.

Le fastp peut détecter et couper le plolyG dans les queues de lecture, il active automatiquement l'élagage de la queue polyG avant et après filtrage. Le problème de la queue polyG peut entraîner un grave problème de séparation du contenu en bases, ce qui signifie que A et T ou C et G ont des rapports de contenu en bases sensiblement différents.

Remarque : la queue polyG est un problème courant observé dans les séries des séquenceurs, qui sont basées sur la chimie bicolore.

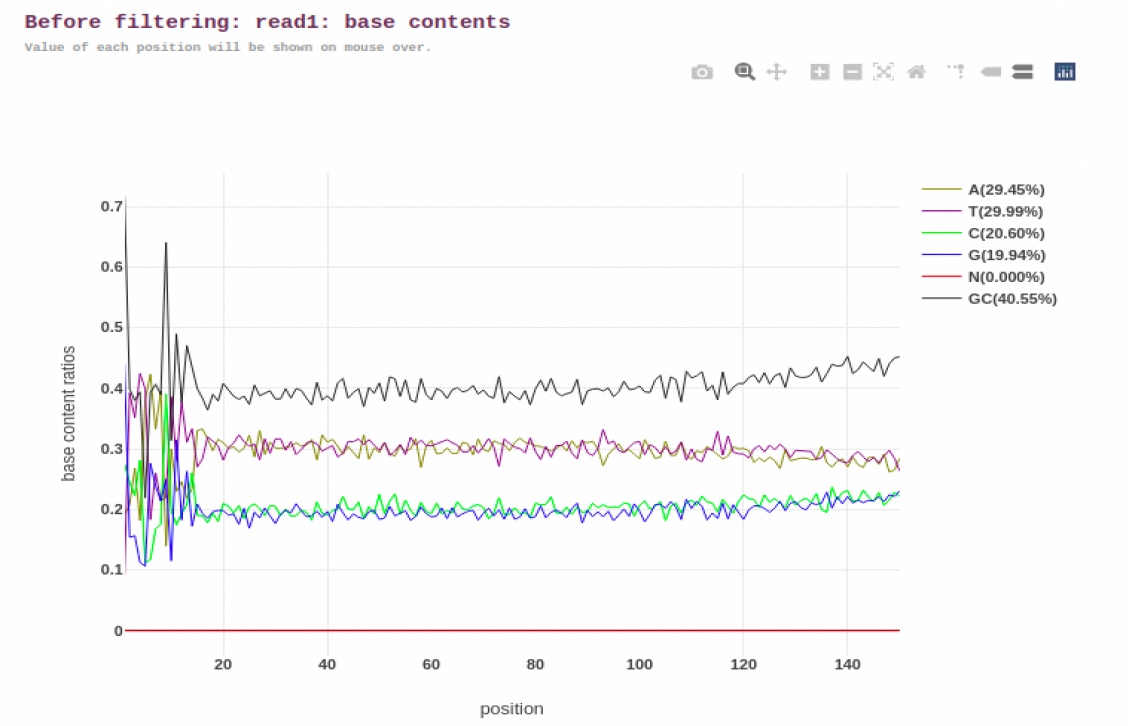


Figure 21 : contenu de base avant filtrage Illumina

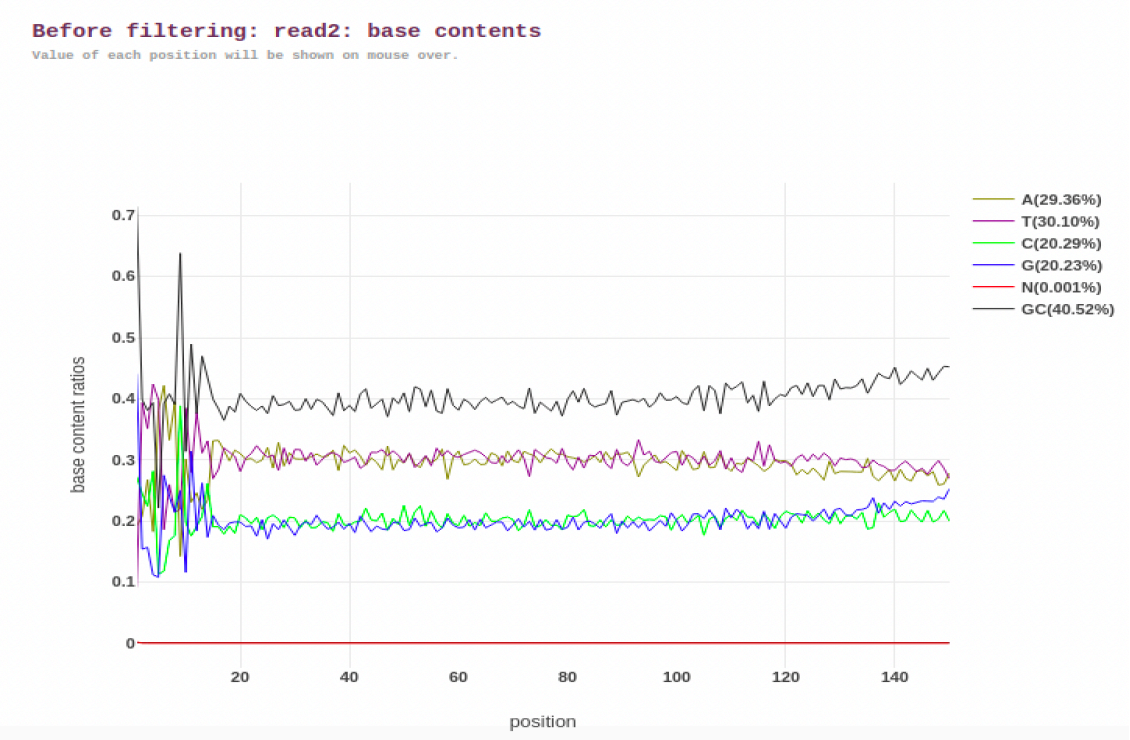


Figure 22 : contenu de base avant filtrage Illumina

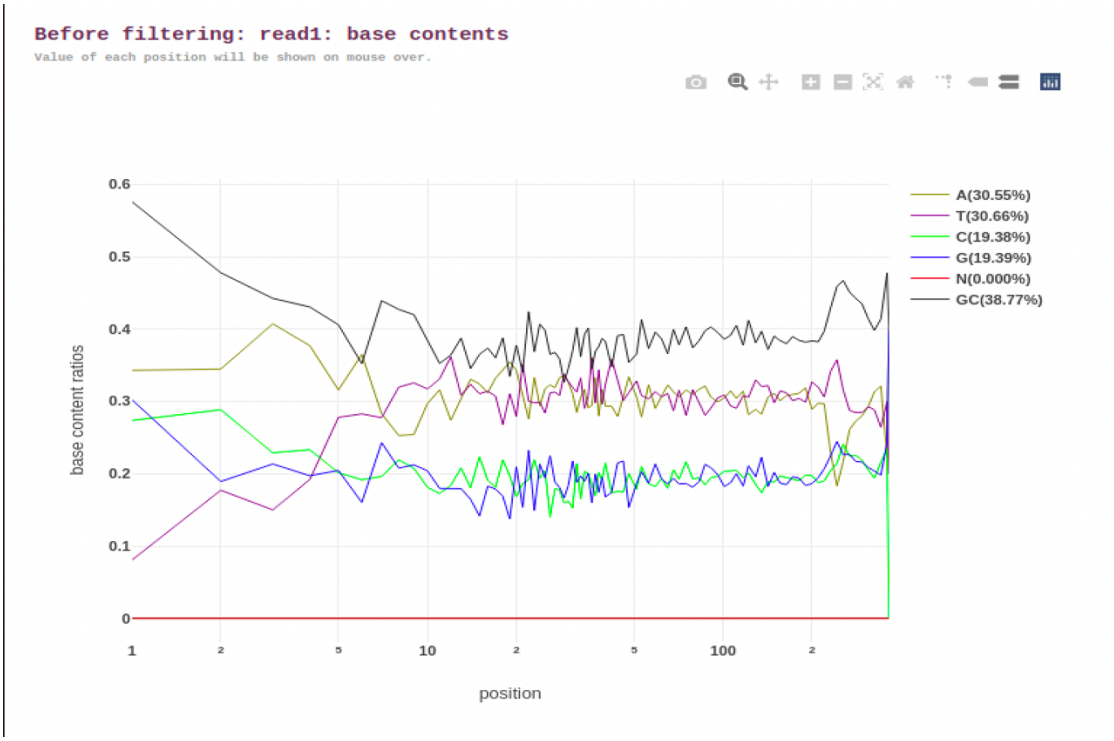


Figure 23 : contenu de base avant filtrage Ion Torrent

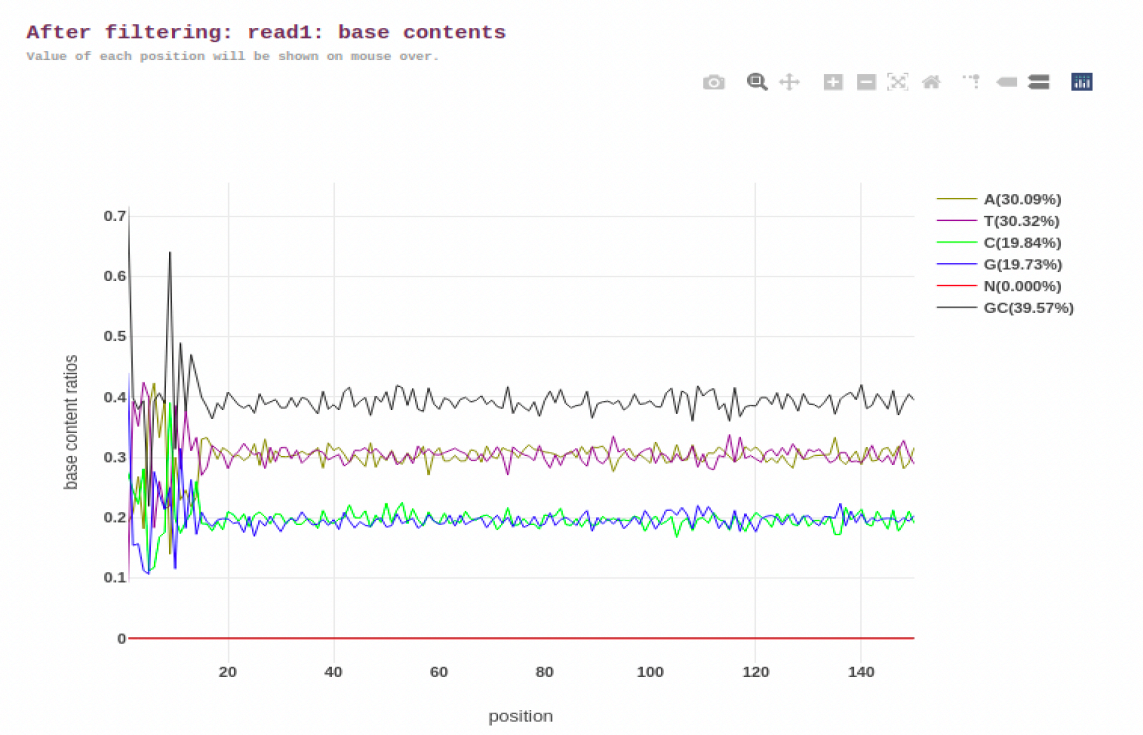


Figure 24 : contenu de base après filtrage Illumina

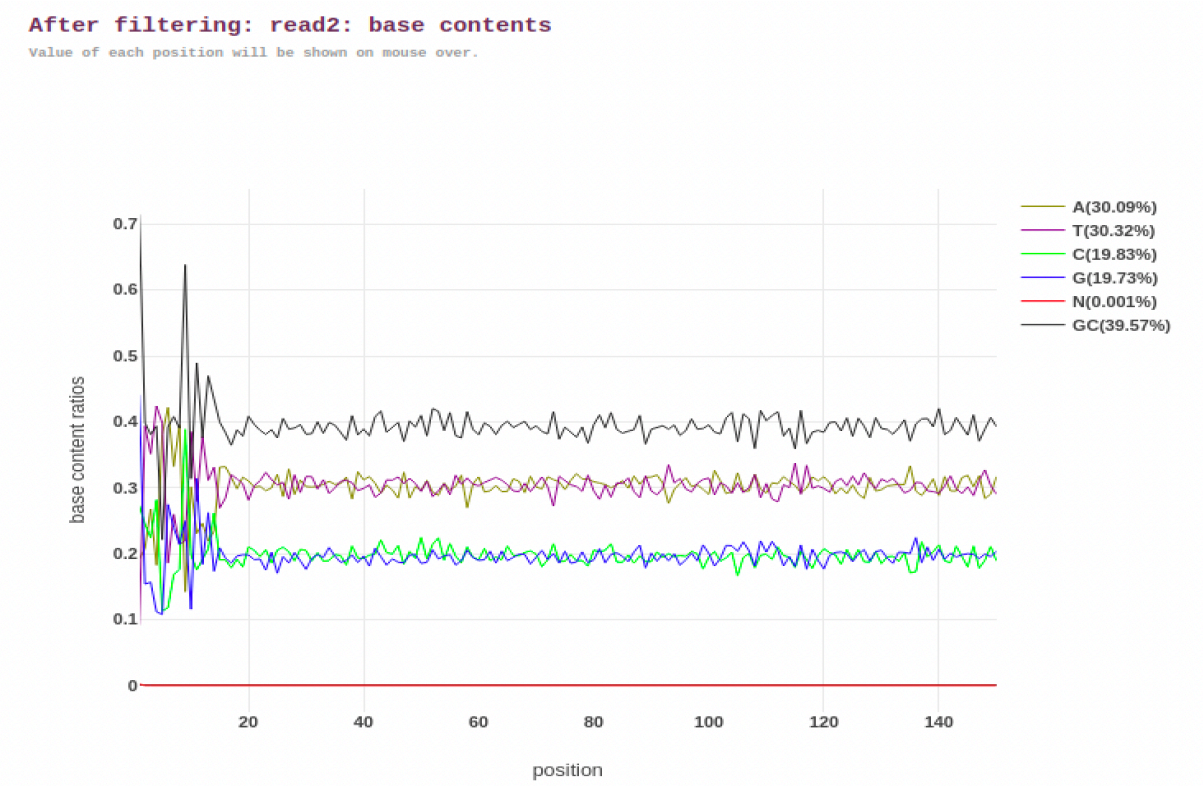


Figure 25 : contenu de base après filtrage Illumina

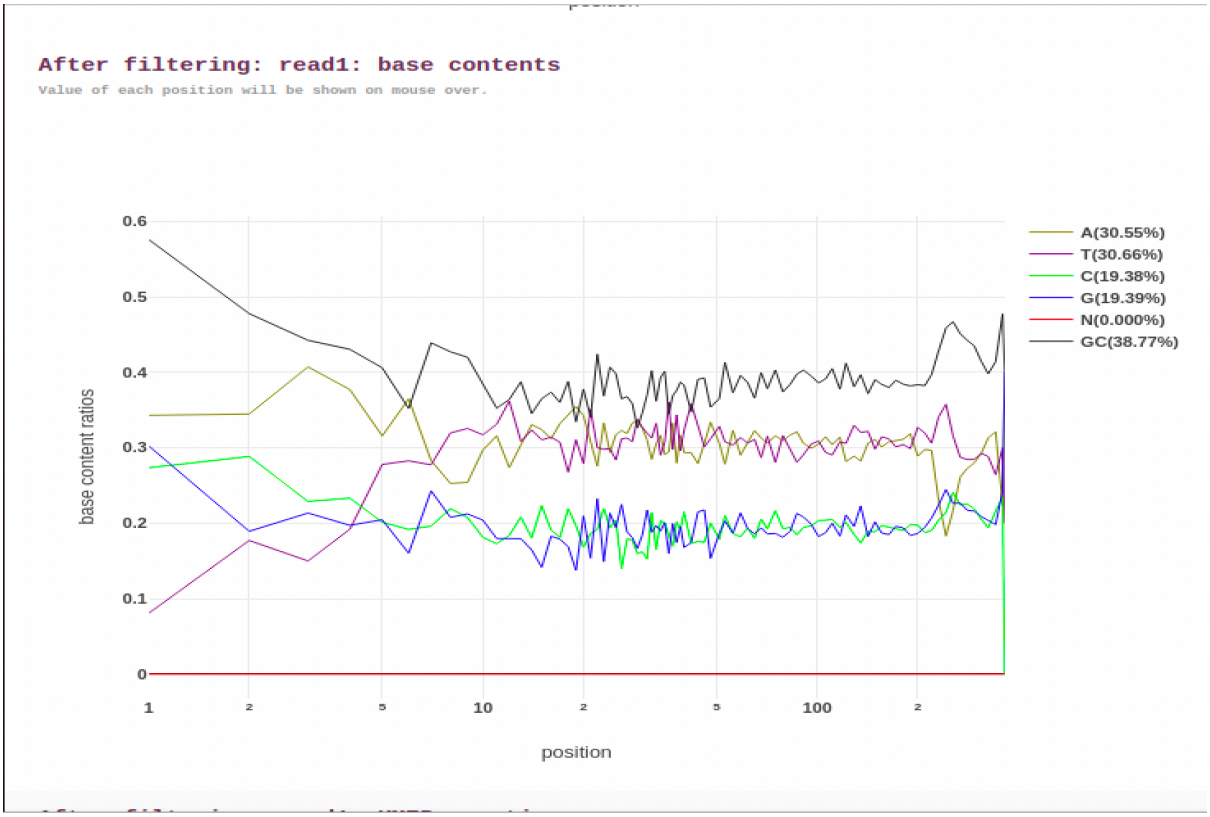


Figure 26 : contenu de base après filtrage Ion Torrent

Les tableaux de contenu K-mer ici sont les résultats de balayage de toutes les possibilités Kmer pour trouver celles qui n'ont pas une couverture uniforme sur la longueur de notre lecture. Un fond plus sombre signifie un nombre plus important. Le compte sera affiché au passage de la souris (dans le rapport html).

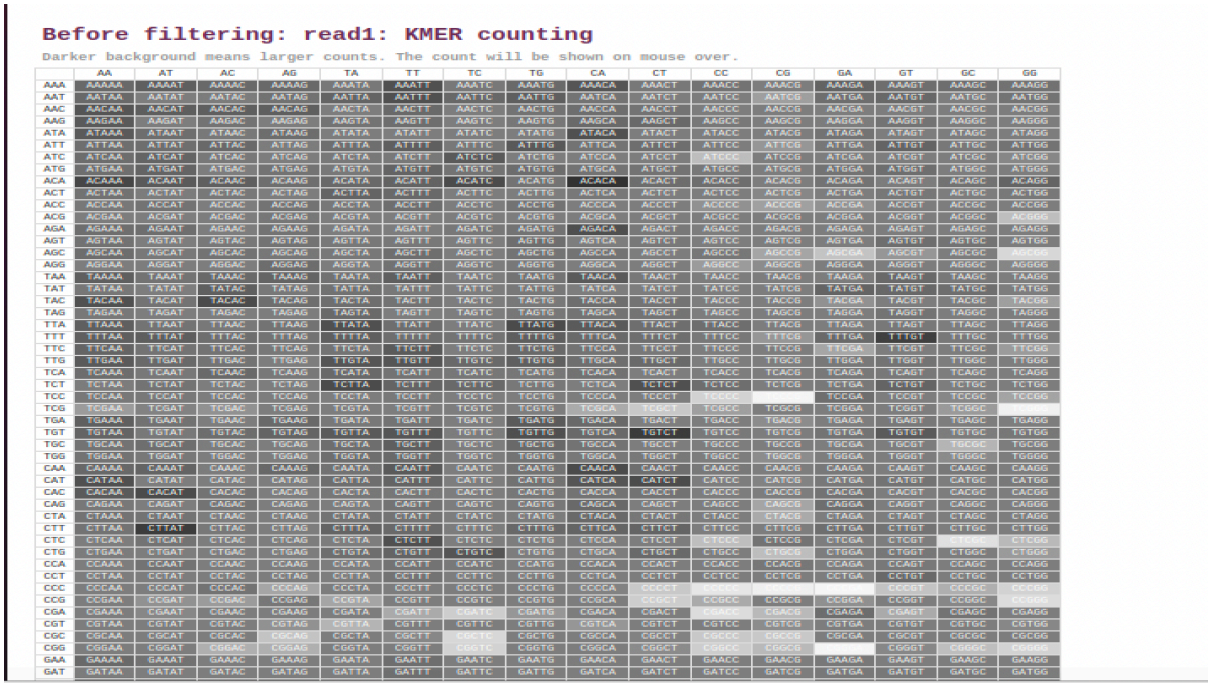


Figure 27 : contenu KMER avant filtrage Illumina

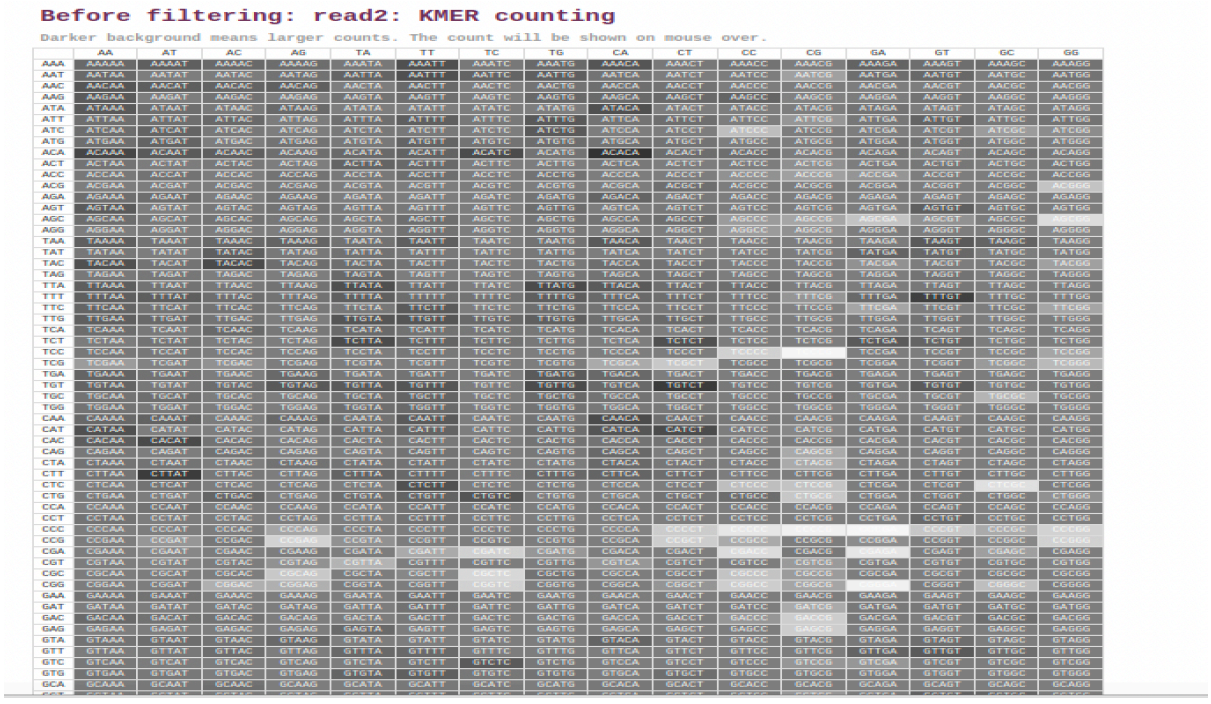


Figure 28 : contenu KMER avant filtrage Illumina

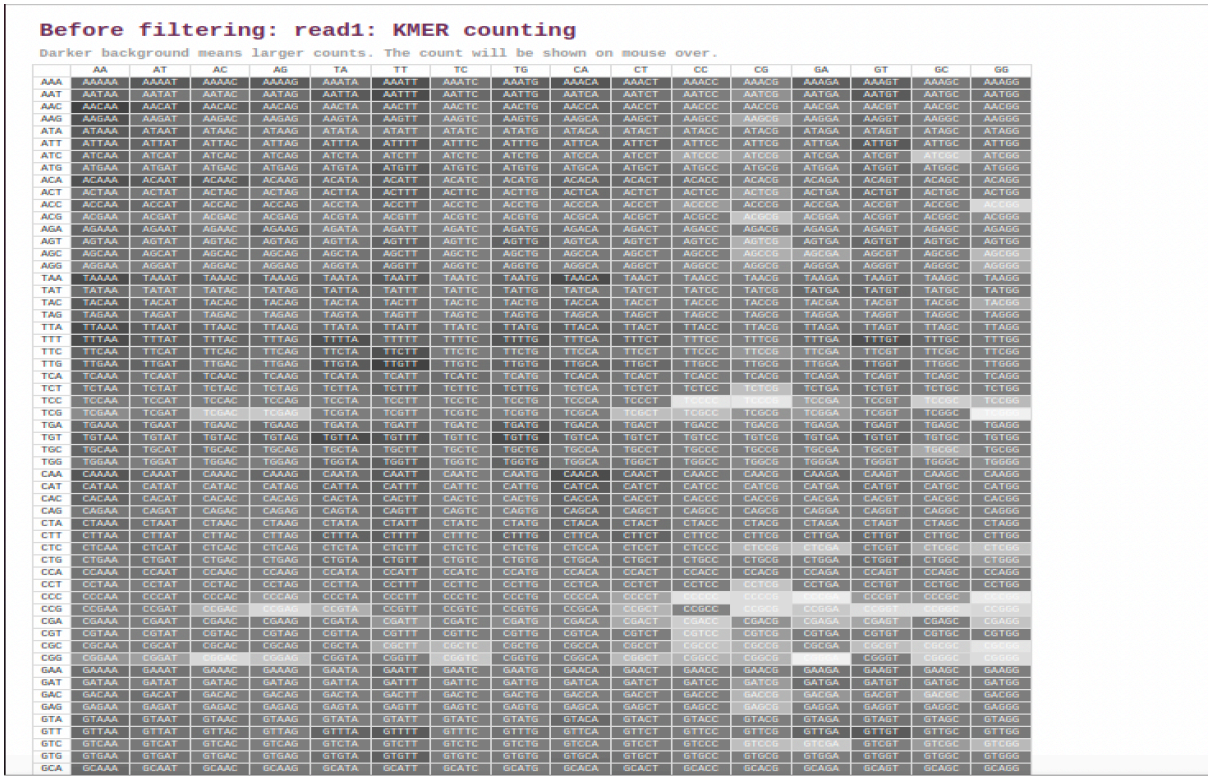


Figure 29 : contenu KMER avant filtrage Ion Torrent

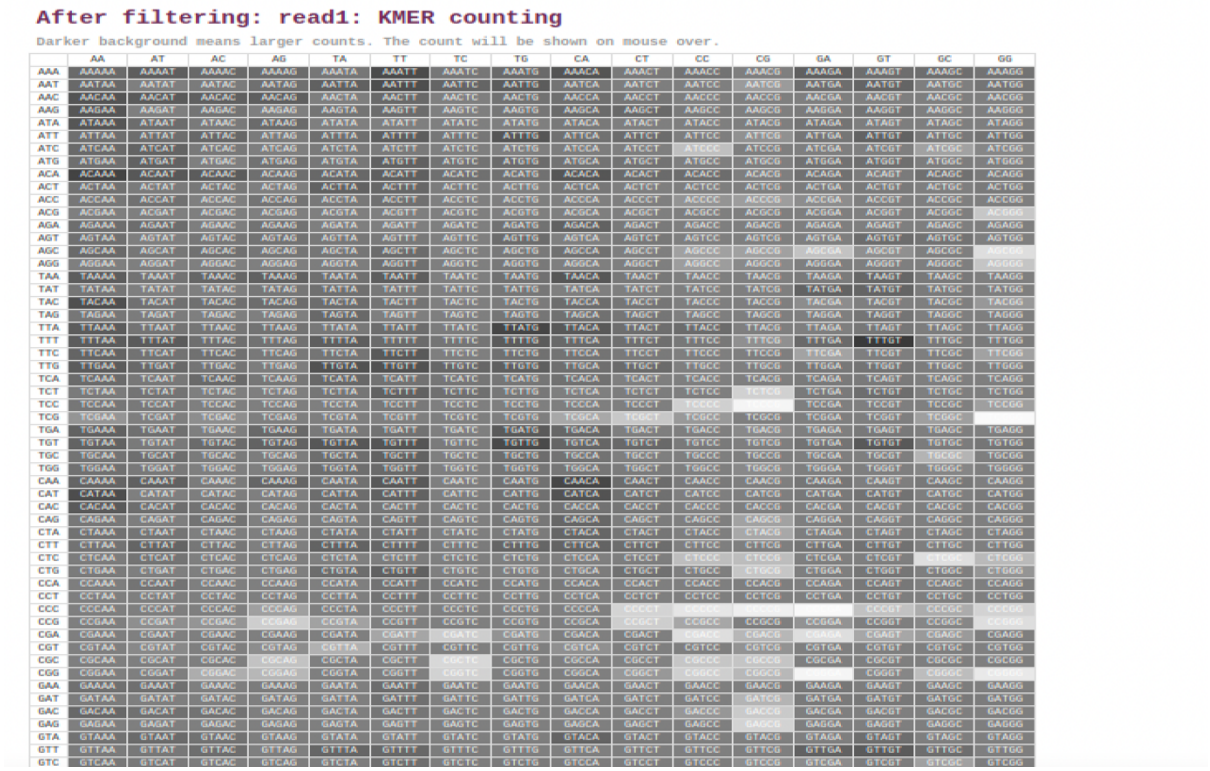


Figure 30 : contenu KMER après filtrage Illumina

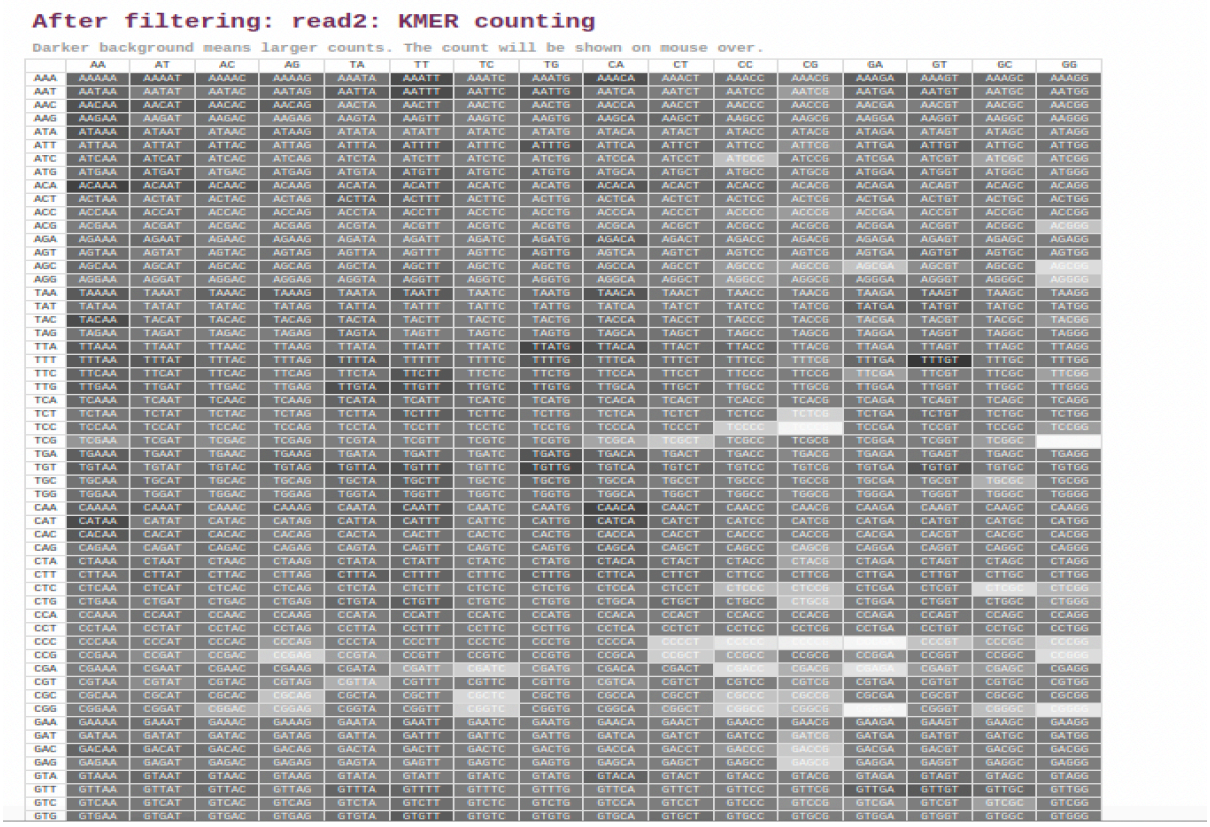


Figure 31 : contenu KMER après filtrage Illumina

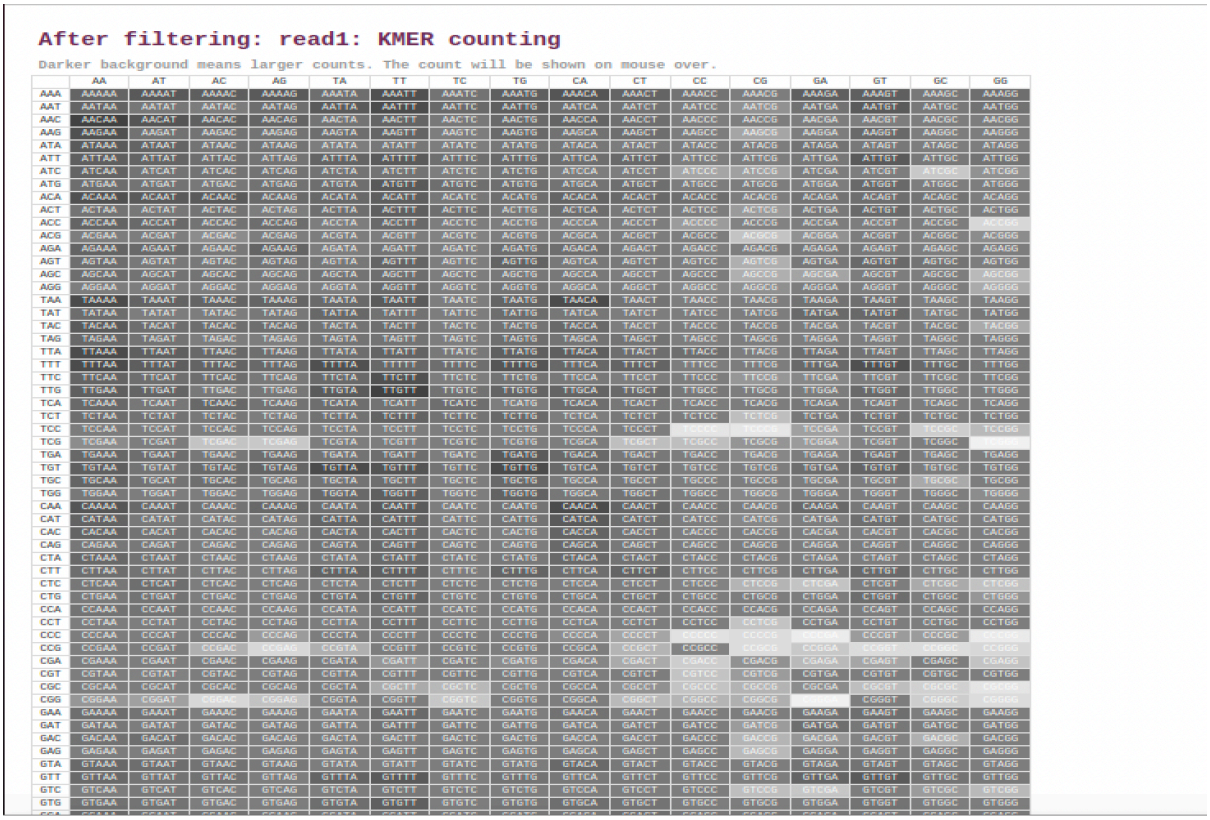


Figure 32 : contenu KMER après filtrage Ion Torrent

1-2-Résultats de l'alignement :

Pour faire l'alignement on a lancer BWA et Samtools pour des résultats plus rapides et meilleurs. Nous avons utilisé le génome complet de l'isolat Wuhan-Hu-1 du coronavirus 2 du syndrome respiratoire aigu sévère comme génome de référence. En sortie, nous avons obtenu un fichier BAM.

Figure 33 : fichier BAM Illumina

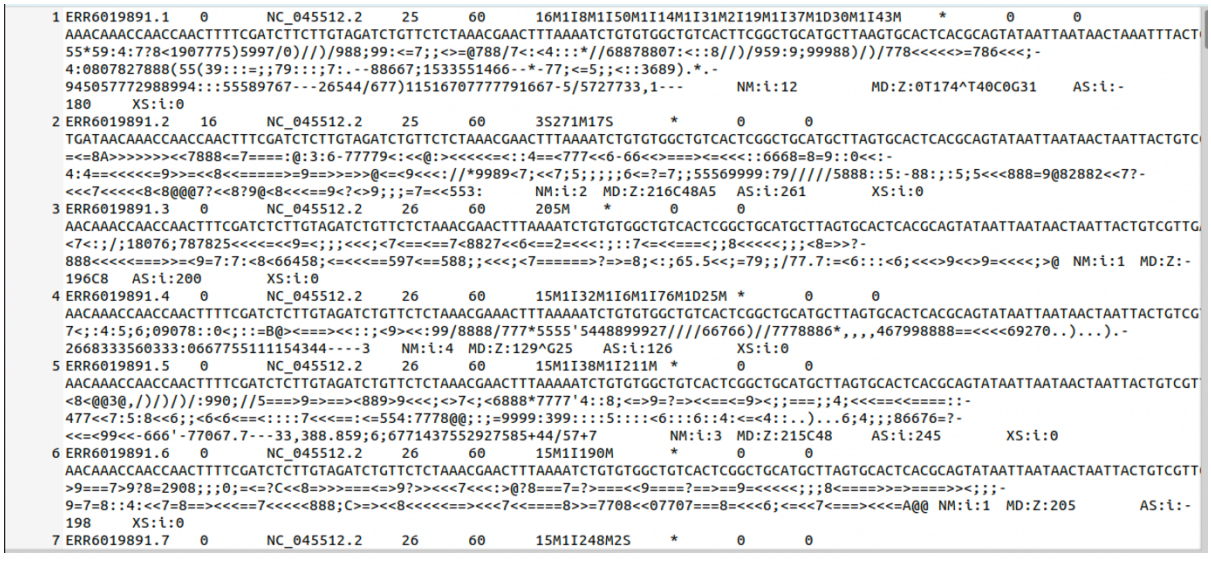


Figure 34 : fichier BAM Ion Torrent

1-3-Résultats de l'analyse des variants :

Après avoir utilisé les outils Samtools et BCFtools et obtenu notre fichier au format VCF (la commande *vcfutils.pl varFilter* est spécifique pour obtenir le fichier VCF), nous avons utilisé l'outil de visualisation IGV pour générer les données et représenter les variants de notre fichier sous la forme suivante qui le rend plus informatif et plus lisible.

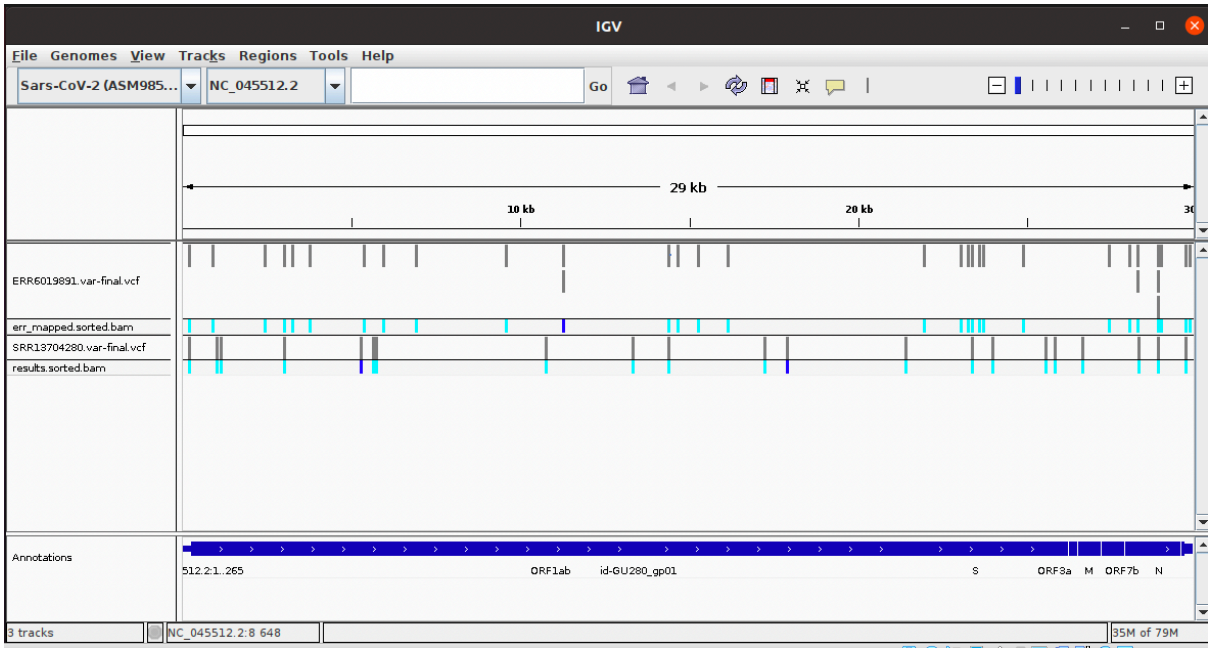


Figure 35 : la fenêtre de visualisation IGV

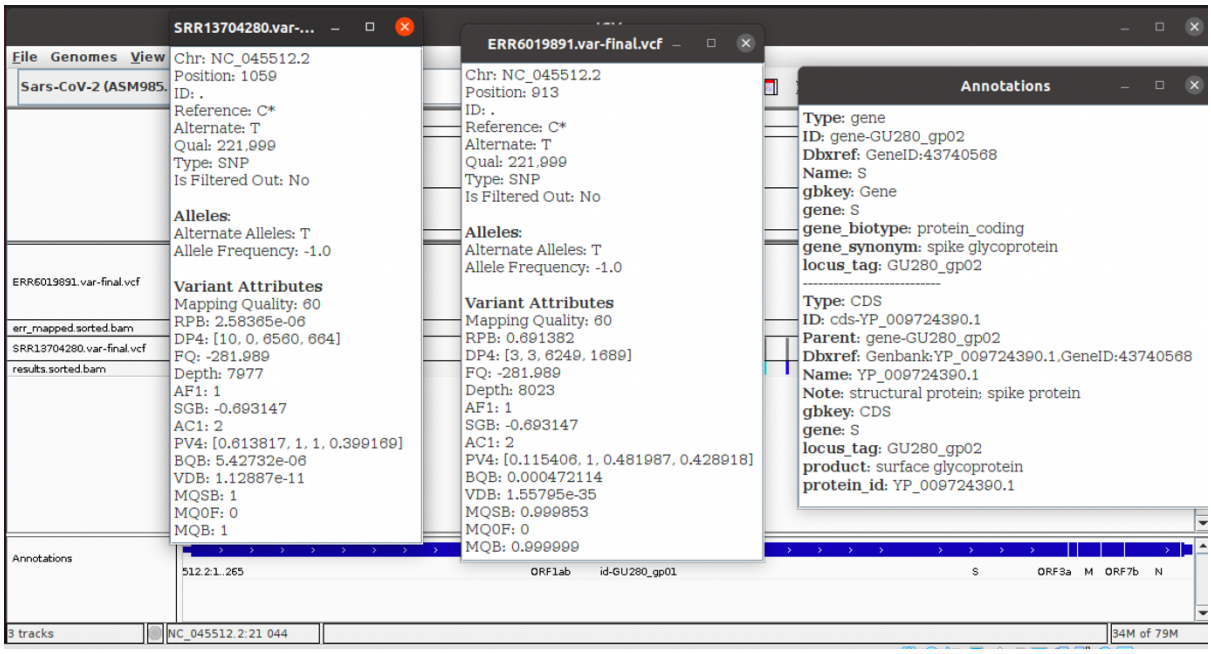


Figure 36 : visualisation d'annotation et des variants des séquences

2-Discusion :

Grâce au NGS, les chercheurs ont identifié la séquence complète du génome de SARS-CoV-2. Il s'agit du plus grand génome de tous les virus à ARN, composé d'environ 29 926 nucléotides. Après de nombreuses recherches, nous avons fait nos choix en fonction de nos besoins et de nos équipements. Nous avons choisi d'avoir nos échantillons à partir de

séquenceurs Illumina et Ion Torrent pour de multiples raisons, Illumina est le séquenceur le plus utilisé pour le séquençage du COVID19 donc nous avions beaucoup de choix de séquences, en plus la plupart des logiciels Illumina ont le droit libre et accessible. Pour Ion Torrent c'était un défi, il n'y avait pas beaucoup de données sur NCBI, la plupart de ses logiciels nécessitaient une méthode de travail différente de celle d'Illumina, mais nous avons réussi à les obtenir. A la fin, nous avons pu faire fonctionner le pipeline des deux séquenceurs. Nous avons d'abord créé un pipeline pour chaque séquenceur puis nous avons remarqué que la plupart des outils et des logiciels étaient en fait similaires dans leurs bases et non seulement pouvaient fonctionner pour leurs séquenceurs mais donnaient aussi des résultats proches. Pour le prétraitement, nous avons utilisé Fastp au lieu de trois outils différents et nous avons réussi à obtenir de meilleurs résultats. Ensuite nous avons utilisé BWA et Samtools comme meilleurs outils d'alignement, l'utilisation de BCFtools et IGV nous a fourni une annotation précise des variantes pour donner à la fin un fichier VCF clair et compréhensible.

Nous avons donc créé une chaîne de production générale qui contient des outils généraux pour Ion Torrent et Illumina, qui peut fonctionner avec les deux. Nous avons également utilisé un nombre minimum d'outils et de logiciels de droit libre pour faciliter l'utilisation, réduire les erreurs, montrer la performance de la vitesse des outils, réduire les fichiers de transition, garder plus d'espace pour le stockage de la machine et finalement avoir les résultats dans un fichier lisible.

Conclusion

Conclusion :

La pandémie actuelle a poussé les chercheurs à s'intéresser davantage à l'étude du SARS-CoV-2. Des universités et des laboratoires de recherche ont séquencé le génome du COVID19 pour l'analyser et répondre à plusieurs questions. Les chercheurs ont pu percer en si peu de temps de nombreux secrets sur le SARS-CoV-2. Grâce à la bio-informatique, ils ont pu disposer d'outils et de méthodes pour analyser les données et accélérer la recherche.

Après une étude approfondie des outils et des logiciels à utiliser pour créer un pipeline capable de traiter les données de séquençage et de détecter les variants, nous sommes arrivés à notre objectif de proposer un pipeline basé sur des outils gratuits.

Les difficultés que nous avons rencontrées ne nous ont pas empêchés d'avancer et de réaliser notre travail. Nous avons essayé plusieurs outils et avons surmonté de nombreux problèmes, notamment les changements de versions et les mises à jour des logiciels.

En effet, notre pipeline a été conçu pour analyser les données NGS "Illumina" et "Ion Torrent" de SARS-CoV-2 à partir d'un fichier SRA jusqu'à l'analyse des variantes, en utilisant des outils bio-informatiques bien choisis.

Comme perspective, nous visons les objectifs suivants :

- Tester plusieurs autres outils.
- Tester d'autres données.
- Développer une interface graphique web et mobile.
- Développer nous-mêmes des outils.
-Interpréter les données NGS.
- Créer un pipeline SARS-CoV-2 pour les données de séquençage en Algérie.

Bibliographie

Bibliographie

Kern, J. (2021, février 21). Il y a 20 ans, l'Homme séquençait son génome pour la première fois. *FUTURA SANTÉ*.

Weinbrecht2, A. G. (2013). Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology, 133*.

Verma, N. G. (2019). Le séquençage de nouvelle génération et son application : Autonomiser la santé publique au-delà de la réalité. *Technologie microbienne pour le bien-être de la société*(13 septembre 2019), 313-341.

DNA Sequencing Fact Sheet. (2020, august 16). *NATIONAL HUMAN GENOME RESEARCH INSTITUTE*.

Griffiths, A. J. (2016, aout 16). DNA sequencing. *Britannica*.

James M Heather 1, B. C. (2016, january). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*, 1-8.

Barton E. Slatko, A. F. (2018, apr). Overview of Next Generation Sequencing Technologies. *Curr DProtoc Mol Biol, 122*(2019 apr 1).

PhD, A. G. (2021, march 17). An Overview of Next-Generation Sequencing. *Technology Networks Genomics Research*.

M.HeatherBenjaminChain, J. (2016, january). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*, 1-8.

Mehdi Kchouk1, 3. J.-F. (2017). Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine, 9*, 1-8.

DNA Sequencing: History, Steps, Methods, Applications and Limitations. (2019, december 01). *Genetic Education*.

G. Dorado, .. P. (2019). Maxam-Gilbert Chemical-Degradation Sequencing. *Encyclopedia of Biomedical Engineering, 3*.

Lin Liu, 1. Y. (2012). Comparison of Next-Generation Sequencing Systems. *BioMed Research International, 2012*.

Erwin van Dijk, C. T. (2021, 01 08). La révolution de la génomique : les nouvelles méthodes de séquençage et leurs applications. *PLANET VIE*.

Karl V Voelkerding, S. D. (2009, march). Next-Generation Sequencing: From Basic Research to Diagnostics. *PubMed, 55*, 641-58.

Christophe Audebert1, 4. D. (2014, decembre). Le séquençage haut-débit. *Med Sci, 40*(24 decembre 2014), 1144-1151.

Weinbrecht2, A. G. (2013). Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology, 133*, 1-4.

John Eid*, A. F. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science, 323*(2 jan 2009), 133-138.

Séquençage à « longues lectures ONT ». (2021). *France Génomique*.

Larousserie, D. (2020). La phylodynamique, l'autre traque du coronavirus. *L'OBS HORS-SÉRIE*(20 avril 2020).

Clinkemaillié, T. (2021). Covid : pourquoi le séquençage du virus est primordial dans la lutte contre l'épidémie. *Les Echos*(23 févr 2021).

Mario Cannataro, A. H. (2021, march). Bioinformatics helping to mitigate the impact of COVID-19 – Editorial. *Briefings in Bioinformatics, 22*(08 march 2021), 613-615.

Vergnaud, V. (2021). Qu'est-ce qu'un variant? *Le Journal du Dimanche*(10 juillet 2021).

Robin, C. (2021). Variants, mutations… quelles différences et conséquences sur l'épidémie ? *Capital*(15 janvier 2021).

Divers et variants – C'est quoi un mutant ? (2021). *Inserm*(14 janviers 2021).

Michel, V. (2021). Polymorphisme génétique et variation. *Encyclopédie de l'Environnement*.

Justin M Zook, B. C. (2014). L'intégration d'ensembles de données de séquence humaine fournit une ressource d'appels de base SNP et de génotypes indel. *Nature Biotechnologie*(16 février 2014), 246-251.

Nadine Hanna, B. P. (2005). Mécanismes et conséquences des mutations. *Med Sci*(2005 november).

Gillet-Markowska, A. (2020). Etude quantitative des variations structurelles des chromosomes chez Saccharomyces cerevisiae. (10 decembre 2020).

J.Tremblay1J.Raelson1F.Harvey1M.Ivanga1J.Chalmers2M.Woodward2S.Harrap3M.Marre4 P.Hamet1. (2014, march). O32 La variabilité du nombre de copies de gènes est associée à l'hypertension dans le diabète de type 2. *ScienceDirect, 40*.

Différents types de variantes : qu'est-ce que la variation génomique ? (2016, september 16). *Genomics Education Programme*.

Robert Challen, E. B.-P.-A. (2021). Risque de mortalité chez les patients infectés par le CoV-2 du SRAS variante préoccupante 202012/1 : étude de cohorte appariée. *The BMJ*(25 février 2021).

Priam, E. (2021, avril 28). Variant sud-africain du Covid-19 : ce qu'il faut savoir sur le variant d'Afrique du Sud. *Dotissimo*.

Roberts, M. (2021). What are the Delta, Gamma, Beta and Alpha Covid variants? *BBC News online*.

Comité d'experts en vigie génomique du SRAS-CoV-2. (2021). État de situation sur le variant B.1.617 du SRAS-CoV-2 (émergent d'Inde) et recommandations pour en rehausser la surveillance au Québec. *Institut National de Santé Publique Québec*, 1-8.

authors, S. R. (2018, january 01). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines. *THE JOURNAL OF MOLECULAR DIAGNOSTICS, 20*(november 14,2017).

Leipzig, J. (2017, may). A review of bioinformatic pipeline frameworks. *Brief Bioinform, 18* (2016 mar 24), 530–536.

Somak Roy 1, C. C.-S. (2018, jan). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn, 20*, 4-27.

Somak Roy, M. (2020, mar 01). AACC.org // Clinical Laboratory News // All Articles // Next-Generation Sequencing Bioinformatics Pipelines Next-Generation Sequencing Bioinformatics Pipelines. *Clinical Laboratory News*.

Erik L. Clarke, L. J.-J. (2019). Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome, 7*(22 march 2019).

S.Masters, P. (2006). The Molecular Biology of Coronaviruses. *Advences in virus research, 66*, 193-292.

Susana Posada-Céspedes, D. S. (2021, june 15). V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics , 37*(20 january 2021), 1673-1680.

Anthony M. Bolger, M. L. (2014, August 1). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(01 April 2014), 2114–2120.

Caroline Charre, C. G. (2020, July). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evolution, 6*(05 October 2020).

Anthony M. Bolger, M. L. (2014, august 1). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics , 30*(01 april 2014), 2114-2120.

Caroline Charre, C. G. (2020, july). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evolution, 6*(05 october 2020).

M. Rafiul Islam, M. N. (2019). Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Scientific Reports, 10*(19 august 2019).

Lucy van Dorp, D. R. (2019). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nature Communications, 11*(25 november 2019).

Divinlal Harilal, S. R.-A. (2020, november). SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission. *Clinical Chemistry, 66*(28 october 2020), 1450-1458.

Sanaâ Lemriss 1, A. S. (2020, jul). Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Causing a COVID-19 Case in Morocco. *Microbiol Resour Announc, 9*.

Sunguk Shin 1, H. L. (2018, avril 1). AIRVF: a filtering toolbox for precise variant calling in Ion Torrent sequencing. *Bioinformatics, 34*, 1232-1234.

Jason R. Miller, S. K. (2010, juin). Assembly Algorithms for Next-Generation Sequencing Data. *Genomics, 6*(6 mars 2010), 315-327.

Gaurav Kumar, A. E. (2020). iSeqQC: a tool for expression-based quality control in RNA sequencing. *BMC Bioinformatics, 21*(13 fevrier 2020).

Anthony M. Bolger, M. L. (2014, aout 1). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(1 avril 2014), 2114-2120.

Nuno A Fonseca 1, J. R. (2012, dec 15). Tools for mapping high-throughput sequencing data. *Bioinformatics , 24*(11 octobre 2012).

Knut Reinert 1, B. L. (2015). Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet, 16*(4 may 2015).

Assembly algorithms for next-generation sequencing data. (2010, jun). *Genomics, 6*(6 mar 2010).

Shulan Tian, H. Y. (2016). Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics, 17*(3 octobre 2016).

Kuhn, M., Stange, T., Herold, S., Thiede, C., & Roeder, I. (2018). Finding small somatic structural variants in exome sequencing data: a machine learning approach. *Computational Statistics, 33*(sep 2018), 1145-1158.

Gregory M Cooper 1, E. A., & Eric D Green, S. B. (2005, jul). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res, 7*(17 juin 2005).

Prateek Kumar, S. H. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols, 4*(25 june 2009), 1073-1081.

Jana Marie Schwarz, C. R. (2010, aug). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods, 8*.

Kasahara, T. T. (2020). Visualization tools for human structural variations identified by whole-genome sequencing. *Journal of Human Genetics volume, 65*(30 octobre 2019), 49-60.