الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الإخوة منتوري قسنطينة ١
Frères Mentouri Constantine I University
Université Frères Mentouri Constantine I

جامعة الاخوة منتوري قسنطينة ١
كلية علوم الطبيعة والحياة
قسم البيولوجيا التطبيقية

Université Frères Mentouri Constantine1
Faculté des Sciences de la Nature et de la Vie
Département de Biologie Appliquée

Mémoire présenté en vue de l'obtention du diplôme de Master

Domaine : Sciences de la Nature et de la Vie
Filière : biotechnologie
Spécialité : Bioinformatique

N° d'ordre :
N° de série :

Intitulé :

Une approche *de novo* pour l'assemblage des génomes

Présenté par : AMOUCHE Selma
BENLAMRI Aya Sara

Le 18/06/2023

Jury d'évaluation :

Président : TAMAGOULT. M — Université Frères Mentouri, Constantine 1
Examinateur : CHEHILI. H — Université Frères Mentouri, Constantine 1
Encadreur : HAMIDECHI. M. A — Université Frères Mentouri, Constantine 1

Année universitaire
2022 - 2023

Remerciements

Nous remercions **DIEU** tout puissant de nous avoir accordé la santé et la volonté d'entamer et de terminer ce modeste travail.

Nous tenons à exprimer notre profonde gratitude envers notre encadreur de mémoire, le Professeur Mohamed Abdelhafid HAMIDECHI, pour son expertise, ses conseils précieux, ses avis éclairés et son accompagnement durant notre préparation de ce mémoire. Il est difficile de trouver les mots pour exprimer pleinement notre reconnaissance et notre profonde gratitude envers le professeur Mohamed Abdelhafid HAMIDECHI.

Nos remerciements s'adressent également à notre Président de jury et nos enseignants Mr **TEMAGOULT Mahmoud**, et Mr **CHEHILI Hamza**, Président et Examinateur de ce travail, qui nous font l'honneur d'évaluer nos efforts et nos contributions. Nous sommes pleinement conscientes de la lourde tâche qu'ils ont acceptée avec une extrême bienveillance, malgré leurs occupations respectives, notamment en cette fin d'année.

Nous exprimons notre reconnaissance envers tous nos enseignants de BioInformatique de notre département, depuis le niveau de Licence jusqu'à ce stade d'enseignement Master. Nous espérons qu'ils trouvent dans nos efforts un fruit tant attendu durant toutes ces années de formation. Merci à vous tous nos chers enseignants !

Dédicace

Avec tout mon respect et ma profonde gratitude, je dédie ma remise de diplôme et ma joie à mon paradis, ma lune, et au fil d'espoir qui illumine mon chemin jusqu'à la source de mon bonheur, à mon soutien qui a toujours été à mes côtés pour me soutenir et m'encourager, vos sacrifices, vos conseils avisés et votre amour indéfectible m'ont donné la force de poursuivre mes rêves et d'atteindre de nouveaux sommets : mon papa.

À celle qui m'a fait une femme, à la prunelle de mes yeux, ma source de vie, d'amour et d'affection, ta patience infinie et ta capacité à me soutenir dans mes choix m'ont donné la confiance nécessaire pour aller de l'avant, la représentation de la tendresse : maman.

À mes frères et sœurs, je les remercie du fond du cœur d'être à mes côtés et de me soutenir. Leur présence bienveillante et leurs conseils éclairés me donnent la force et le courage de poursuivre.

À mon binôme, Tu es mon ami avant d'être mon binôme, nous avons formé une équipe dynamique et solidaire. Notre travail commun a été marqué par une synergie unique et une volonté commune de donner le meilleur de nous-mêmes, Je suis profondément reconnaissant pour notre amitié et notre travail fructueux.

BENLAMRI AYA SARA

Dédicace

Aujourd'hui est un jour spécial, c'est la fin de mes études et le début d'une nouvelle étape dans ma vie. C'est avec une immense joie et une pointe de nostalgie que je prends la plume pour vous écrire cette dédicace de mémoire.

Maman et papa, vous êtes mes piliers, mes guides et mes plus grands supporters. Votre amour inconditionnel, votre soutien constant et vos encouragements ont été les moteurs qui m'ont permis de persévérer dans les moments difficiles. Votre confiance en moi m'a donné la force de poursuivre mes rêves et d'atteindre ce jour tant attendu. Ce mémoire est dédié à vous, en reconnaissance de tout ce que vous avez fait pour moi. Je vous suis infiniment reconnaissante et je ne pourrai jamais assez-vous remercier.

Je voudrais exprimer toute ma gratitude à mes frères et sœurs bien-aimés. Votre présence dans ma vie est un cadeau précieux qui apporte joie, soutien et amour inconditionnel.

Mon binôme, nous avons parcouru un long chemin ensemble. Les heures passées à travailler côte à côte, à partager nos idées, nos doutes et nos réussites, resteront gravées dans ma mémoire. Ta persévérance, ton dévouement et ton esprit collaboratif ont été une source d'inspiration constante. Je te remercie du fond du cœur pour notre amitié et notre partenariat fructueux.

Mes amis, Votre présence joyeuse, vos encouragements sincères et votre soutien inconditionnel ont illuminé mes journées les plus sombres. Les moments que nous avons partagés resteront à jamais gravés dans ma mémoire.

AMOUCHE SELMA

RÉSUMÉ

Le but de ce travail est de proposer une approche *ab initio* pour la reconstruction d'un génome en utilisant des algorithmes appropriés. Notre approche vise à améliorer la précision et la rapidité de la reconstruction de génome, l'analyse comparative, et l'exploitation des informations génomiques complémentaires. Notre application, basée sur la programmation en langage python a révélé que la théorie d'assemblage est tout à fait fiable et donne des résultats satisfaisants pour ces premiers essais.

Mots clés : Séquençage NGS ; Chevauchement ; Assemblage de génome ; Graph de De Bruijn.

ABSTRACT

The aim of this work is to propose an ab initio approach for genome reconstruction using appropriate algorithms. The main objective is to overcome current challenges related to genome reconstruction, such as the presence of repeating regions, sequencing errors and limitations of existing technologies. Our approach aims to improve the accuracy and speed of genome reconstruction, comparative analysis, and exploitation of complementary genomic information. The idea is to combine these different approaches to obtain more reliable and complete results, thus allowing genomes to be reconstructed in a more accurate way.

Keywords: NGS sequencing; Overlap; Genome assembly; De Bruijn graph.

ملخص

الهدف من هذا العمل هو اقتراح نهج مبدئي لإعادة بناء الجينوم باستخدام الخوارزميات المناسبة. الهدف الرئيسي هو التغلب على التحديات الحالية المتعلقة بإعادة بناء الجينوم، مثل وجود مناطق متكررة، وأخطاء التسلسل وقيود التقنيات الحالية. يهدف نهجنا إلى تحسين دقة وسرعة إعادة بناء الجينوم والتحليل المقارن واستغلال المعلومات الجينومية التكميلية. الفكرة هي الجمع بين هذه الأساليب المختلفة للحصول على نتائج أكثر موثوقية وكاملة، وبالتالي السماح بإعادة بناء الجينوم بطريقة أكثر دقة.

. الكلمات المفتاحية: تسلسل الجيل التالي; التداخل; تجميع الجينوم ; الرسم البياني دي بروين

TABLE DES MATIERES

PARTIE 2 : PARTIE EXPERIMENTALE

LISTES DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

ASCII: **American Standard Code for Information Interchange**

BLAST: Basic Local Alignment Search Tool

CCD : Charge-Coupled Device (dispositif à transfert de charge)

dNTPs : Désoxyribonucléotides triphosphates

NGS : Séquençage nouvelle génération (Next Generation Sequencing)

OLC : Overlap-Layout-Consensus

Introduction

INTRODUCTION

L'algorithme de reconstruction de génome est essentiel en bio-informatique car vise à reconstituer la séquence complète d'un génome à partir de lectures de séquençage partielles ou chevauchantes. Avec l'avènement des technologies de NGS, la quantité de données génomiques générées a explosé, nécessitant des méthodes sophistiquées pour les analyser et en extraire des informations significatives. Nous examinerons en théorie, deux principales stratégies : l'assemblage assisté par référence et l'assemblage *de novo*. La reconstruction de génome est une étape clé de cette analyse, car elle permet de comprendre la structure, la fonction et l'évolution des génomes.

En développant de nouvelles méthodes et en améliorant les approches existantes, les chercheurs contribuent à l'avancement des connaissances en génomique et ouvrent de nouvelles perspectives pour la recherche médicale, agricole et environnementale. Comprendre les principes et les limites de l'algorithme de reconstruction de génome est crucial pour garantir des analyses précises et fiables, et pour exploiter pleinement le potentiel des données génomiques massives générées par les technologies de séquençage de nouvelle génération.

Cependant, il est à noter que les méthodes d'assemblage constituent un challenge réel et avéré face aux bioinformaticiens en tenant compte de la taille des données génomiques extériorisées à partir d'un séquençage NGS classique. En effet toutes les étapes d'assemblage génomique sont étroitement corrélée aux algorithmes bioinformatiques tels que celui de Needleman & Wunsch ou celui de Smith & Waterman pour ne citer que ces deux-là !

Dans notre tentative, nous avons testé, d'une part, l'approche des alignements globaux pour détecter les meilleurs chevauchements (overlapping) et d'autre part, l'algorithme d'alignement local pour estimer, en fin d'assemblage, la fiabilité du génome assemblé.

Partie 1

Synthèse bibliographique

CHAPITRE 1 : STRUCTURE DES GENOMES

Le génome est le patrimoine génétique stocké dans le noyau d'un organisme. Il est contenu dans chacune de ses cellules sous forme de chromosomes. Le support matériel du génome est l'ADN, sauf chez certains virus à ARN (Centre national de la recherche scientifique).

La structure du génome est organisée en plusieurs régions. Les régions codantes contiennent les gènes qui sont responsables de la synthèse des protéines. Les régions non-codantes contiennent des séquences qui sont responsables de la régulation de l'expression des gènes.

Un génome contient toutes les informations nécessaires au développement et au fonctionnement d'un individu (Eric Green, 2023).

1. FONDEMENT DE LA GENOMIQUE

1.1.La taille du génome est corrélée à la complexité de l'organisme

La taille du génome varie significativement entre les différents organismes. Comme un plus grand nombre de gènes apparaît nécessaire à la formation d'un organisme plus complexe, il n'est pas surprenant que la taille du génome soit à peu près reliée à l'apparente complexité d'un organisme. Bien qu'il existe une corrélation entre la taille du génome et la complexité de l'organisme, elle est loin d'être parfaite. De nombreux organismes présentant une complexité similaire ont des génomes de tailles bien différentes : la drosophile possède un génome 25 fois plus petit que la sauterelle, et le génome du riz est 40 fois plus petite que celui du blé. Ces exemples mettent en avant que le nombre de gènes, plutôt que la taille du génome, sera relié à la complexité d'un organisme, et Cela est encore plus évident lorsque l'on observe la densité génique de ces génomes. Le tableau suivant donne la taille totale du génome et le nombre de gènes présents chez un certain nombre d'organismes dont le génome a été entièrement séquencé (Watson,J, 2009).

Tableau1 : variations du nombre de gènes et taille de génome chez différents organismes (wikipédia, génome).

Organismes	Taille de génome (Mpb)	Nombre de gènes
Haemophilus influenzae	1,8	1800
Escherichia coli	4,6	4300
Levure	12,1	6000
Drosophile	150,0	14500
Nématode	110,0	21000
Arabette	110,0	25500
Souris	2700,0	22000
Homme	22000	22000

1.2 . Structure et fonction des chromosomes

L'association d'ADN protéines qui constitue les chromosomes est appelée chromatine. Elle contient des protéines basiques et des protéines acides histones et non histones. Ces dernières forment des nucléosomes autour desquels s'enroule l'ADN (Pierce, M., 2012).

Les cellules sont engendrées par des cellules, et la seule manière de créer de nouvelles cellules est la division de cellules préexistantes (Bry, D., 2012).

En métaphase, lors de la division cellulaire, chaque chromosome est constitué de deux chromatides sœurs qui sont jointes au niveau de la région centromérique. Chaque bras de chromatide est constitué d'une seule fibre enroulée. La fibre est constituée d'ADN double-brin étroitement enroulé et de protéines. Un processus ordonné de condensation dépendant d'enroulements et de torsions intervient dans la transition de l'état de chromatine interphasique vers l'état de chromosome mitotique plus condensé. Il est estimé que, pendant la transition de l'interphase à la prophase, la longueur de l'ADN qui constitue la fibre chromatinienne est contractée. Sur les microphotographies électroniques, on note la séparation bien nette entre les chromatides sœurs constituant chaque chromosome. Elles sont jointes uniquement par le centromère qu'elles ont en commun avant l'anaphase (figure 1) (William, S., 2006).

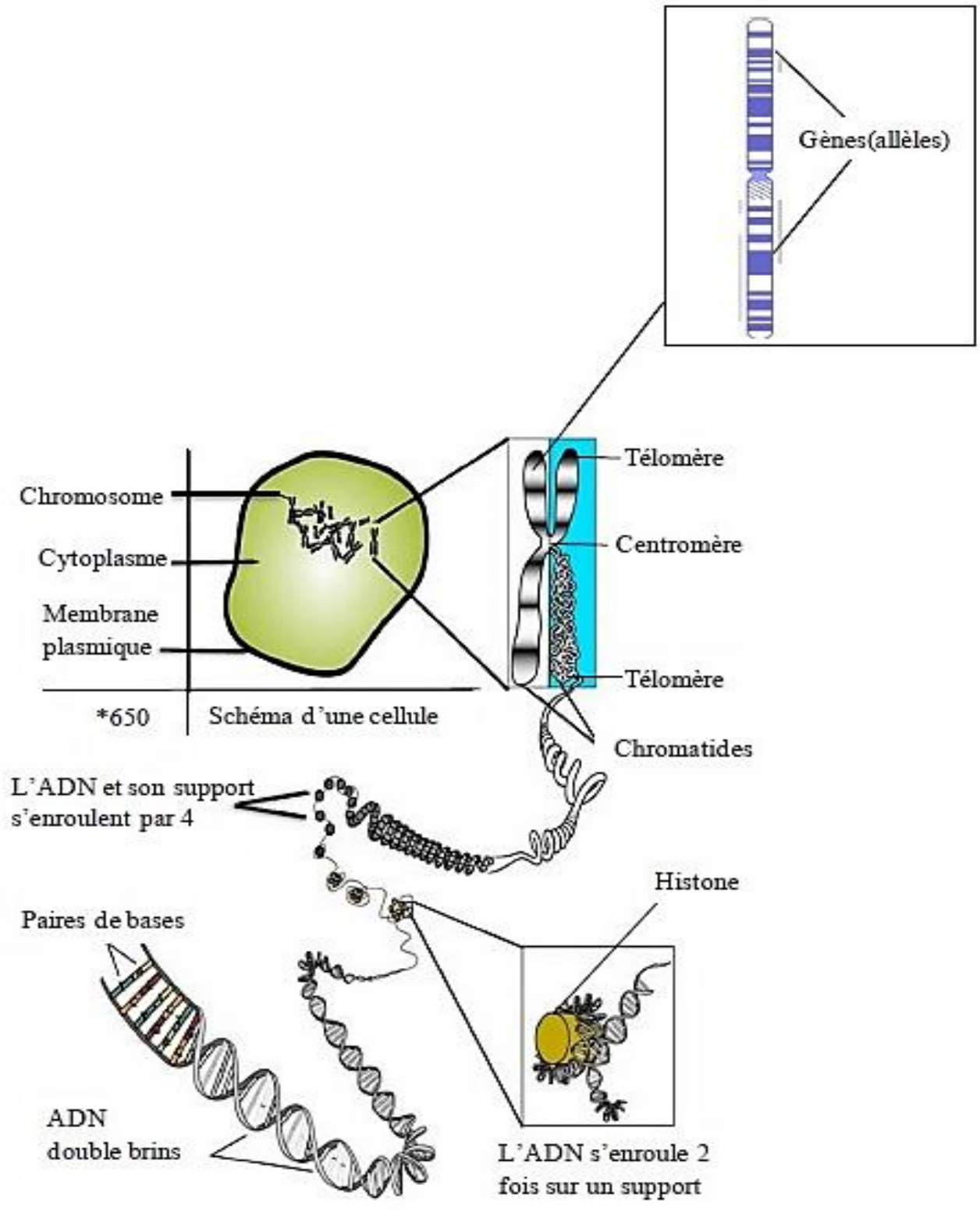


Figure 1 : Description de la structure d'un chromosome (Wikipédia)

1.3 Le cycle de division cellulaire

Le cycle de division cellulaire continue se compose de deux grandes étapes (figure 2) : la phase de croissance (interphase) et la phase de division (mitose ou méiose). Pendant l'interphase, la cellule se prépare pour la division en se développe et en répliquant son ADN. Pendant la mitose ou la méiose, la cellule se divise en deux cellules filles identiques (Ronald Dery, 2021).

L'interphase se décompose en trois phases différentes :

- La phase G1 (gap, environ 12 heures) : correspond à la phase de croissance optimale.la cellule augmente eu taille et crée les organites qui lui manquent.
- La phase S (environ 8 heures) : la duplication de l'ADN.
- La phase G2 (environ 3 heures) : synthèse des protéines et correspond à la période qui sépare la fin de la duplication de l'ADN et la division cellulaire.

La mitose (ou phase M) est la phase de division cellulaire. Elle dure une à deux heures.

Les variations de durée du cycle cellulaire se font essentiellement sur la phase G1, la durée des autres phases étant relativement constante (Thamad, D., 2021).

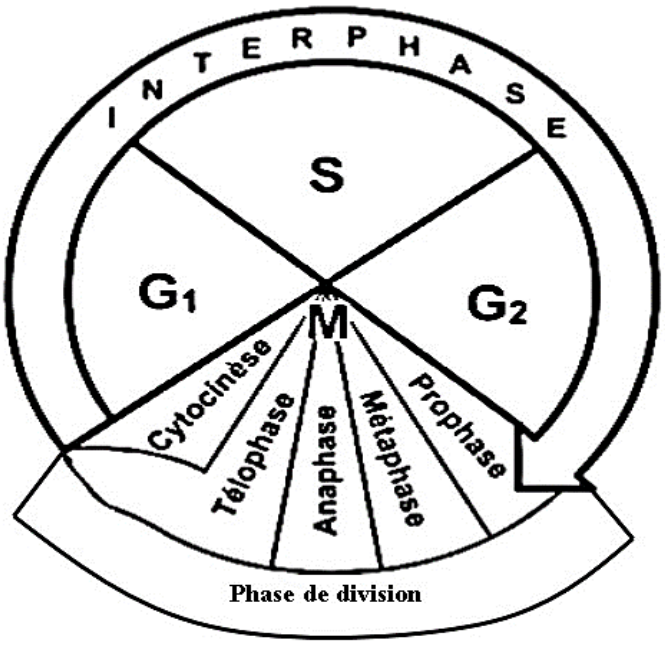


Figure 2 : Cycle cellulaire (Ronald Dery, 2021).

2. TECHNIQUES DE SEQUENCAGE DU GENOME

Le séquençage est la détermination de l'ordre exact des paires de bases dans un génome, et permet de décrypter le code génétique qui se cache dans chaque cellule.

2.1. Principe général du séquençage génomique

Le principe général est basé sur la fragmentation aléatoire du génome en fragments d'ADN de quelques centaines de paires de bases, car la plupart des techniques de séquençage ne permettent de lire qu'une centaine (300-900) de paires de bases. Cependant, le génome humain contient 2,9 milliards de paires de bases, ce qui rend impossible la lecture de l'ensemble du génome en une seule fois. Les fragments d'ADN obtenus sont appelés lectures ou séquences unitaires ou reads, qui sont ensuite alignées pour reconstruire la séquence complète du génome en utilisant les chevauchements entre les fragments (Gilles Furelaud, 2004).

2.2. Le séquençage des génomes : Il existe plusieurs techniques de séquençage

1- La méthode enzymatique de Sanger : utilise des amorces spécifiques et des nucléotides modifiés pour étendre des fragments d'ADN en chaînes complémentaires. Cette technique a permis la détermination de la séquence de l'ADN de nombreux organismes, y compris le génome humain (Boeck, S., 2021).

Le fonctionnement de technique de sanger se fait par extraction d'un fragment d'ADN de l'échantillon. Ensuite, ce fragment est chauffé pour forcer l'ADN à se dérouler. Les deux brins de la double hélice se séparent alors en brins individuels.

La prochaine étape consiste à baisser la température et à ajouter une amorce d'ADN. Celle-ci est une courte séquence monocaténaire. Elle s'attache au brin d'ADN à séquencer.

La température est augmentée, puis, ajout des nucléotides libres et l'ADN polymérase. Les nucléotides libres contiennent l'une des quatre dNTPs. Commence par la séquence d'amorce, l'ADN polymérase construit un brin d'ADN complémentaire (ou inverse). Elle le fait en ajoutant un nucléotide à la fois.

Quatre réactions de séquençage différentes doivent se produire, une pour chacun des quatre types de nucléotides. Pour obtenir ces réactions, il faut ajouter au mélange les quatre didésoxyribonucléotides triphosphates (ddNTPs). Ces derniers sont des nucléotides de synthèse, ils se comportent comme des inhibiteurs de la réplication. Ces versions indiquent la terminaison d'une chaîne. Chacun de ces nucléotides spéciaux est étiqueté avec un colorant différent, ainsi seront visible lorsqu'ils sont exposés aux rayons UV.

Lorsque l'ADN polymérase atteint un nucléotide de terminaison de chaîne, elle arrête la séquence d'ADN. L'ADN polymérase ajoute les nucléotides modifiés de façon aléatoire. De nombreuses séquences d'ADN de différentes longueurs sont donc produites.

Les segments d'ADN subissent une électrophorèse sur gel de polyacrylamide-urée dénaturant. Celle-ci permet de séparer les fragments d'ADN de différentes longueurs. Pour ce faire, les fragments d'ADN doivent être ajoutés à un excipient de gel de polyacrylamide, puis y faire passer un courant électrique. Cela fait en sorte que les segments s'alignent dans le gel en fonction de leur taille. Les petits fragments se déplacent davantage que les gros. Lorsque les fragments ont fini de se déplacer, le gel est examiné à l'aide d'un appareil de radiographie ou d'une lampe UV.

Le gel peut être lu en regardant les bandes foncées dans chaque colonne. Il y a une colonne pour chaque type de nucléotide (G, C, A, T). La séquence des nucléotides peut être déterminée en examinant la séquence des bandes, La figure ci-dessous présente un aperçu résumé. (Parlons sciences, 2020).

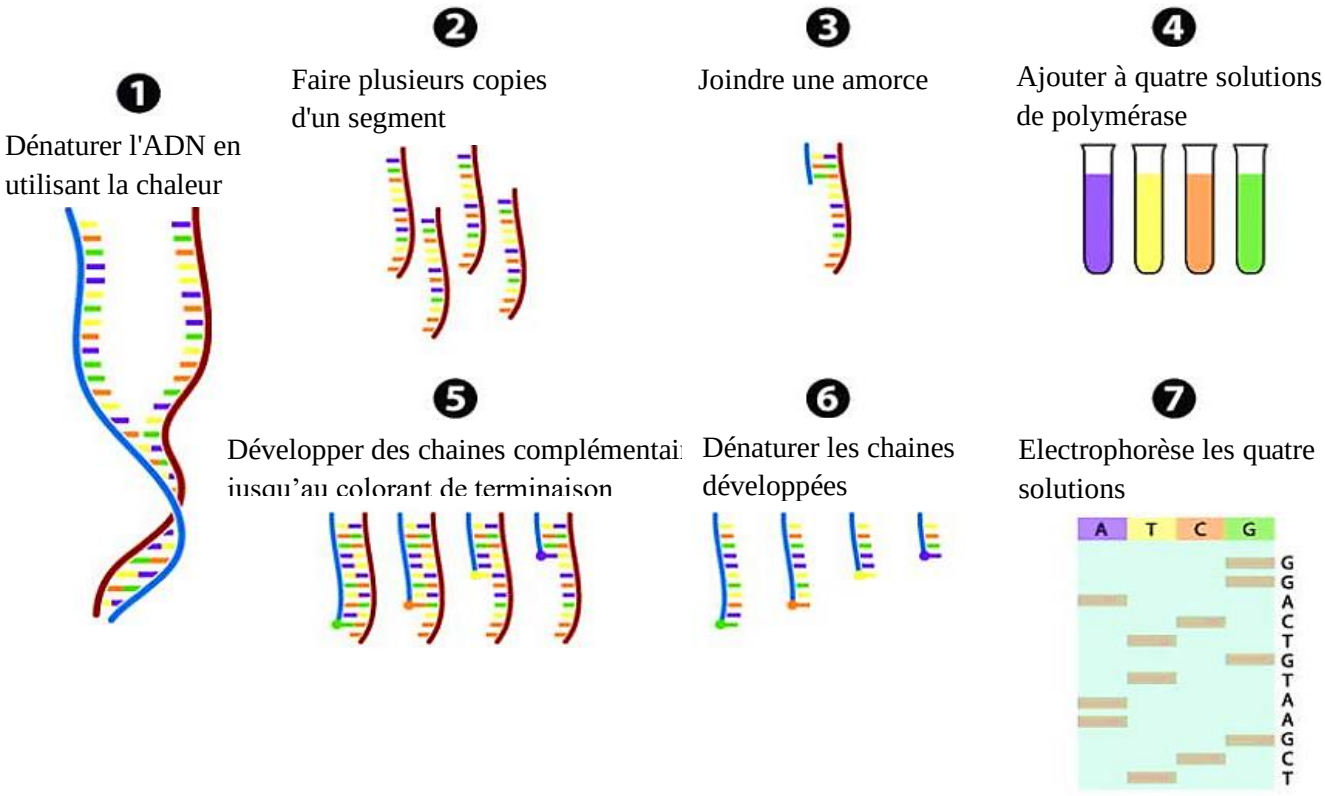


Figure 3 : Méthode de séquençage de Sanger (researchGate, 2007)

3- Technologie NGS : Également appelée séquençage à haut débit, c'est la révolution biotechnologique actuelle, et permet de séquencer de grandes quantités d'ADN en des temps records.

Les technologies NGS présentent trois étapes communes :

- La préparation de banques (librairies) : sont créées en utilisant une fragmentation aléatoire de l'ADN en petits morceaux de l'ordre de centaines (150-600) de paires de bases. Cette fragmentation est réalisée en utilisant des endonucléases. Les fragments d'ADN sont adaptés avec des séquences d'amorces et des marqueurs par des oligonucléotides qui liés aux ses extrémités.
- L'amplification : la banque est amplifiée par PCR ou des méthodes d'amplification clonale comme amplification en pont qui utilisée par la plateforme Illumina et Amplification par PCR en microplaques utilisée par la plateforme Ion Torrent.
- Le séquençage : l'ADN est séquencé en utilisant différentes approches en fonction de la technologie (plateformes) utilisée.

3-1. Les plateformes NGS

Il existe un certain nombre de plateformes NGS différentes utilisant différentes technologies de séquençage comme : Illumina (Solexa), Ion torrent (Proton / PGM), Roche 454 et SOLiD. Chaque plateforme NGS utilise une technologie de séquençage différente. La qualité des données, du débit et de la longueur des 'reads' sont indiqués dans la figure suivante (Renaud blervaque, 2013) :

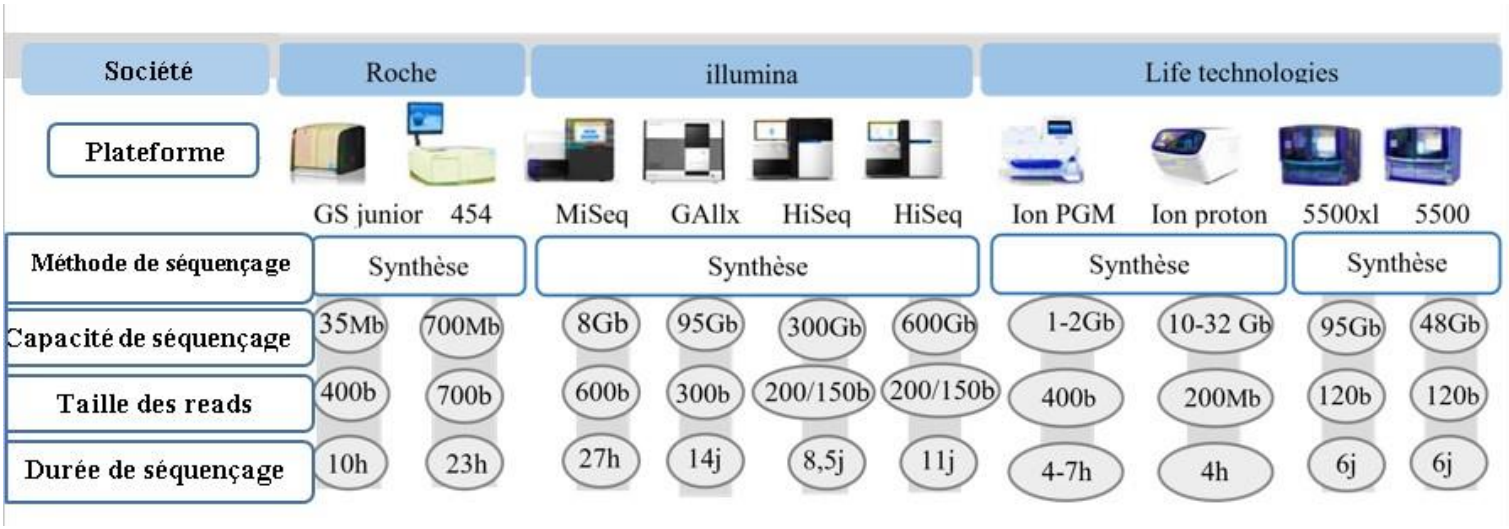


Figure 4 : Comparaison globale de plateformes de séquençage à haut débit (Renaud blervaque, 2013).

3-2. Séquençage Illumina /Solexa

La technologie la plus utilisée sur le marché du NGS est le séquenceur Ilumina/Solexa Génome Analyzer (GA). Le séquenceur adopte la technologie du séquençage par synthèse (SBS).

1. Préparation de la librairie : les échantillons d'ADN sont fragmentés de manière aléatoire, Cette fragmentation est généralement réalisée à l'aide d'endonucléases, qui coupent l'ADN au niveau de séquences spécifiques appelées sites de restriction. Cela produit des fragments d'ADN de différentes tailles, qui sont ensuite ligaturés avec des adaptateurs d'index spécifiques à chaque échantillon, ils sont des séquences courtes d'oligonucléotides synthétiques qui contiennent des séquences d'ancrage permettant de fixer les fragments d'ADN à une surface solide.
2. Amplification (PCR) : 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.
3. Séquençage : cette étape consiste à déterminer chaque nucléotide des séquences.

Illumina utilise l'approche de séquençage par synthèse basé sur l'incorporation par une polymérase modifiée de désoxyribonucléotides portant des terminateurs réversibles couplés à des fluorophores. Ici, les quatre nucléotides modifiés, les amorces de séquençage et les ADN polymérases sont ajoutés sous forme de mélange (mix) et les amorces sont hybridées aux séquences. Ensuite, Chaque nucléotide est marqué avec un fluorescent spécifique et incorporé un par une grâce à un groupe 3'-hydroxyle inactif, évitant ainsi les répétitions et garantissant la précision du séquençage. 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 résumé visuel peut être observée dans la figure 2.

4. Analyse des données : enfin, les données de séquençage sont analysées pour identifier les séquences d'ADN ou d'ARN et leurs modifications éventuelles. Cela peut inclure l'assemblage de séquences, l'alignement à un génome de référence et L'analyse des variantes génétiques qui consiste à rechercher des différences entre le génome de référence et le génome de l'organisme étudié.

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 (Martin Krahn , 2016).

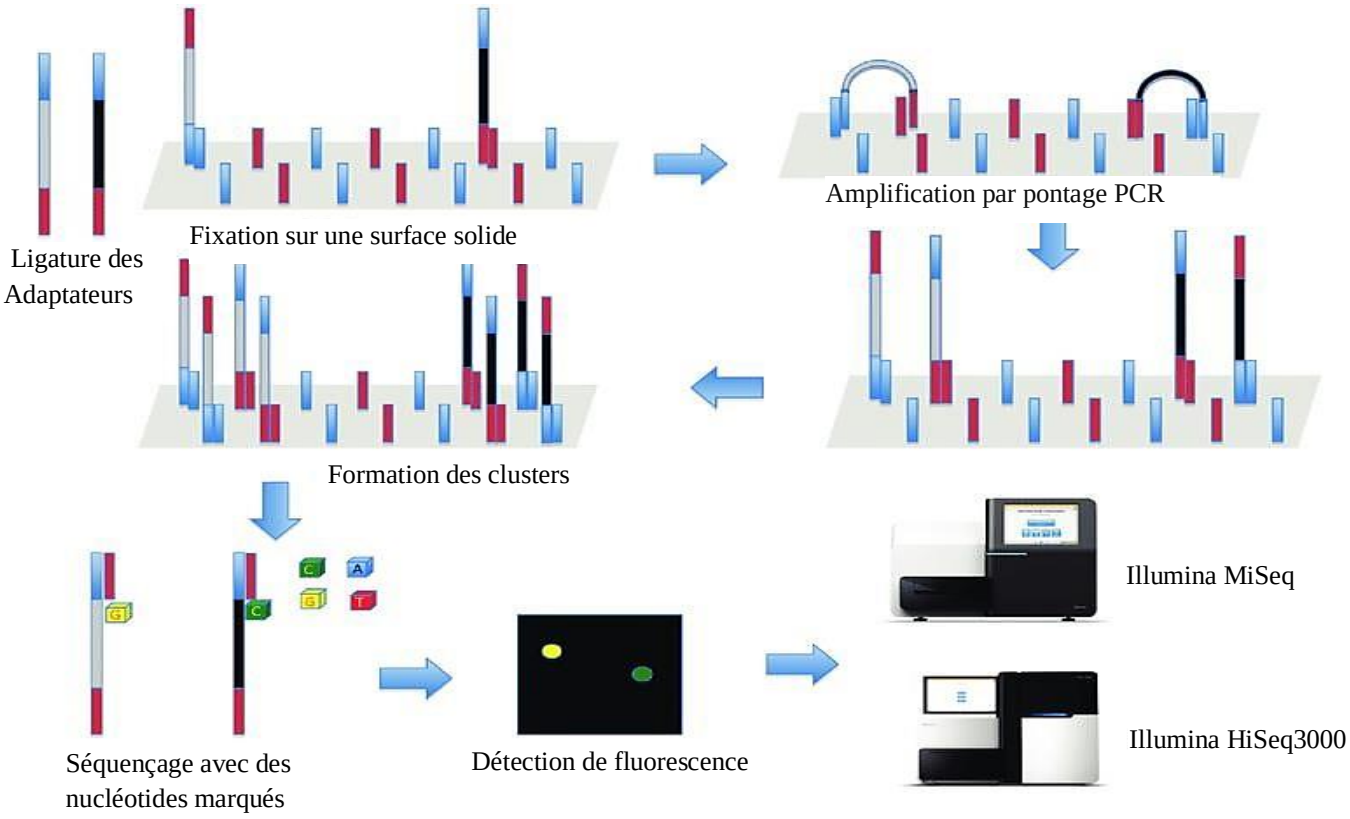


Figure 5 : Principe général de la technologie de Illumina (researchGate, 2017).

La figure suivante montre la plaque de verre rectangulaire utilisé dans le séquençage Illumina :

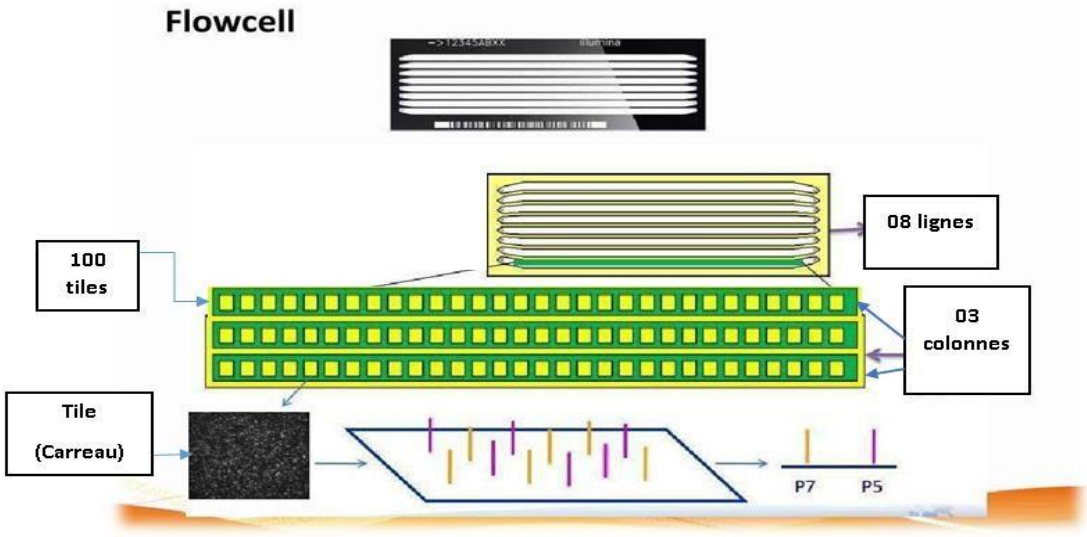


Figure 6 : Cellule de flux Illumina (Séquençage d'extrémité appariée Illumina, 2018)

3-3. Le séquençage Ion Torrent

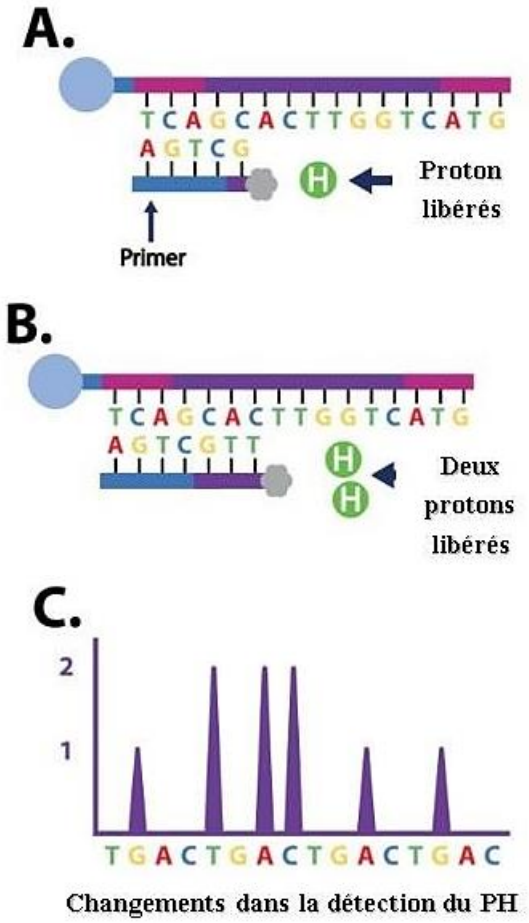
Ion Torrent basé sur un circuit intégré capable de détecter les variations de pH dans des cellules contenant des billes sur lequel est fixé l'ADN à séquencer.

Le processus de séquençage Ion Torrent commence par la préparation de la librairie d'ADN, qui consiste à fragmenter par l'utilisation d'enzymes de restriction l'ADN cible en fragments de taille réduite (100-400), à ajouter des adaptateurs pour identifier les échantillons et à amplifier les fragments pour obtenir suffisamment de matériau pour le séquençage.

Ensuite, les fragments d'ADN sont fixés sur une surface solide (généralement une micropuce) et une solution contenant des nucléotides est ajoutée chacun seul, Chaque nucléotide a été synthétisé avec des modifications chimiques spécifiques pour permettre leur incorporation dans la chaîne d'ADN en fonction des règles de complémentarité de Watson Crick.

Lorsque la polymérase ajoute un nucléotide à la chaîne nouvellement synthétisée, un ion H^+ est libéré en tant que sous-produit de la réaction chimique. Cette libération d'ions H^+ est détectée par des capteurs de pH intégrés à la surface de la micropuce, ce qui permet de déterminer la base ajoutée à la chaîne d'ADN.

Le processus de synthèse et de détection de l'ion H^+ est répété pour chaque position dans la chaîne d'ADN, ce qui permet de déterminer la séquence complète de l'échantillon d'ADN cible. Ce processus de séquençage Ion Torrent permet de produire des millions des reads (environ 200 -400 pb) en parallèle, ce qui permet d'obtenir des résultats de séquençage à haut débit et à coût relativement faible. La figure suivante le montre.



A. L'ajout d'un des quatre dNTP à l'échantillon de réaction entraîne l'incorporation d'une guanine dans le brin d'allongement et la libération d'un proton.

B. L'ajout d'un des quatre dNTP entraîne l'incorporation de deux thymines dans le brin d'allongement et la libération de deux protons.

C. Détection des changements de pH qui fournit des informations sur la séquence d'ADN cible.

Figure 7 : Séquençage Ion Torrent. (ION TORRENT SEQUENCING, 2021).

CHAPITRE 2 : ANALYSE DES DONNÉES NGS

C’est une méthode avancée de détection des variations génétiques à haut débit. Elle permet de générer les données de séquençage de millions de fragments d'ADN et comprend trois étapes :

- L'analyse primaire : appel de bases
- L'analyse secondaire : contrôle qualité des données
- L'analyse tertiaire : assemblage des lectures (reads) par rapport à un génome de référence (mapping) ou non (*de novo*).

1. Appel de bases (base calling) : processus par lequel un ordre de nucléotides dans une matrice est déduit lors d'une réaction de séquençage. Les platesformes de séquençage de nouvelle génération qui utilisent des terminateurs réversibles marqués par fluorescence ont une couleur unique pour chaque base. Ceux-ci sont incorporés dans le brin complémentaire de la matrice d'ADN et capturés avec une caméra CCD (charge-coupled device) sensible. Ces images sont transformées en signaux qui sont utilisés pour déduire l'ordre des nucléotides, également connu sous le nom d'appel de base. (Industrialisation des procédures d’analyses de données de séquençage pan-génomiques constitutionnelles, 2022). Un programme informatique pour accomplir ce travail est Phred qui permet l'identification d'une séquence de nucléobases à partir de données de "traces" de fluorescence générées par un séquenceur d'ADN automatisé qui utilise l'électrophorèse et la méthode du colorant 4-fluorescent.

Le résultat est stocké sous la forme d'un fichier FASTQ. La qualité de chaque base est stockée sous forme de caractères ASCII représentant sa qualité. Les scores de qualité sont calculés en fonction de la probabilité qu'une base soit incorrecte.

Le score de qualité Phred est une mesure logarithmique de cette probabilité, exprimée en échelle de Phred. La formule pour calculer le score de qualité Phred est :

$$Q = \text{Phred} = -10 * \log_{10}(P)$$

Si on veut estimer la probabilité d'erreur (p) on prend :

$$P = 10^{-(Q//10)}$$

Où P est la probabilité qu'une base soit incorrecte. Ainsi, plus la valeur Phred est élevée, plus la qualité de la base est élevée et moins la probabilité qu'elle soit incorrecte est élevée.

Le tableau suivant indique le score de qualité Phred et l'estimation de leur précision :

Tableau 2 : score de qualité Phred (contrôle qualité des données de séquençage illumina 2022)

Score de qualité phred	Probabilité qu'une base soit mal identifiée (Error)	Précision de l'identification de la base (1-Error)
10	1/10	90%
20	1/100	99%
30	1/1000	99,9%
40	1/10000	99,99%
50	1/100000	99,999%

Chaque read est représentée sous format FASTQ par quatre lignes de texte :

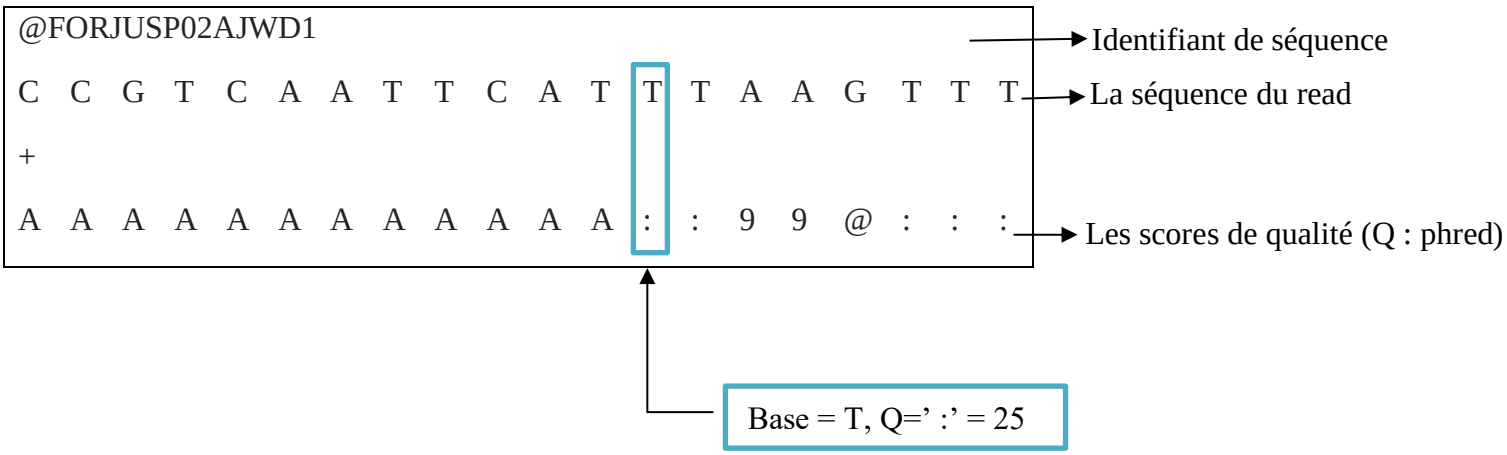
1. La première ligne commence par un symbole "@". C'est une ligne d'information qui identifie la séquence
2. La deuxième ligne contient la séquence d'ADN du read, avec chaque base nucléotidique représentée par une lettre (A, C, G ou T).
3. La troisième ligne commence par un symbole "+" suivi de l'identifiant de séquence correspondant (qui peut être identique à celui de la première ligne).
4. La quatrième ligne contient les scores de qualité associés à chaque base nucléotidique du read, représentés par des caractères ASCII qui correspondent à une valeur numérique.

ASCII (American Standard Code for Information Interchange) représente des caractères en texte brut, tels que des fichiers de texte :

Tableau 3 : Scores de qualité – Analyse NGS (learn.gencore.bio.nyu.edu)

Q	ASCII	Q	ASCII	Q	ASCII	Q	ASCII
0	!	12	-	23	8	34	C
1	"	13	.	24	9	35	D
2	#	14	/	25	:	36	E
3	$	15	0	26	;	37	F
4	%	16	1	27	<	38	G
5	&	17	2	28	=	39	H
6	‘	18	3	29	>	40	I
7	(	19	4	30	?	41	J
8	)	20	5	31	@		
9	*	21	6	32	A		
10	+	22	7	33	B		
11	,						

Exemple d'un fichier FASTQ résultant de la NGS :



2. Contrôle qualité des données et pré-process

Évaluer la qualité des données issues d'un séquençage haut-débit est la première étape à effectuer avant de se lancer dans des analyses bio-informatiques. On appelle cette étape « pre-processing ». Elle est utilisée pour exclure les lectures de mauvaise qualité qui auraient pu apparaitre lors du séquençage. La sortie des séquenceurs se compose de lectures brutes organisées en fichiers texte au format FASTQ, où chaque lecture est annotée avec sa qualité score. Il est toujours conseillé d'effectuer la qualité contrôle (CQ) sur ces fichiers. Pour s'assurer que l'analyse en aval produit des données fiables et appels à haute confiance (Andrews, S. 2010)

Processus de Traitement des données brutes de séquençage :

Etape 1 : Evaluation de la qualité des données brutes /enlèvements des bases

Plusieurs outils de contrôle de la qualité ont été développés, visant à fournir une qualité complète, profils comprenant des statistiques de base telles que le total, nombre de lectures et leur longueur, contenu GC, scores de qualité par base et par séquence (Van Der Auwera, 2013).

FastQC est un outil qui donne la qualité sous forme d'un graphique représente la qualité (score Phred) de chaque base pour tous les reads. À chaque position du read, la qualité de tous les reads est représentée sous la forme d'un boxplot. La médiane est en rouge, la moyenne en bleu. Le code couleur indique les scores de très bonne qualité (en vert), bonne qualité (en orange) et mauvaise (en rouge) (Olivier Rué.2022).

Etape 2 : Nettoyage des données brutes

Éliminations des séquences contaminantes : incluent les adaptateurs et les primer de séquençage qui sont des séquences non génomiques peuvent poser un problème lors de l'alignement. Les outils couramment utilisés sont : Trimmomatic, Cutadapt, etc.

Etape 3 : Elimination des parties de lectures de mauvaise qualité

Éliminer la fin des lectures qui compose des bases ayant une qualité inférieure à un seuil donné, souvent déterminé à l'aide de logiciels (Dijon, 2017).

La figure suivante résume les étapes du Processus de Traitement des données brutes de séquençage :

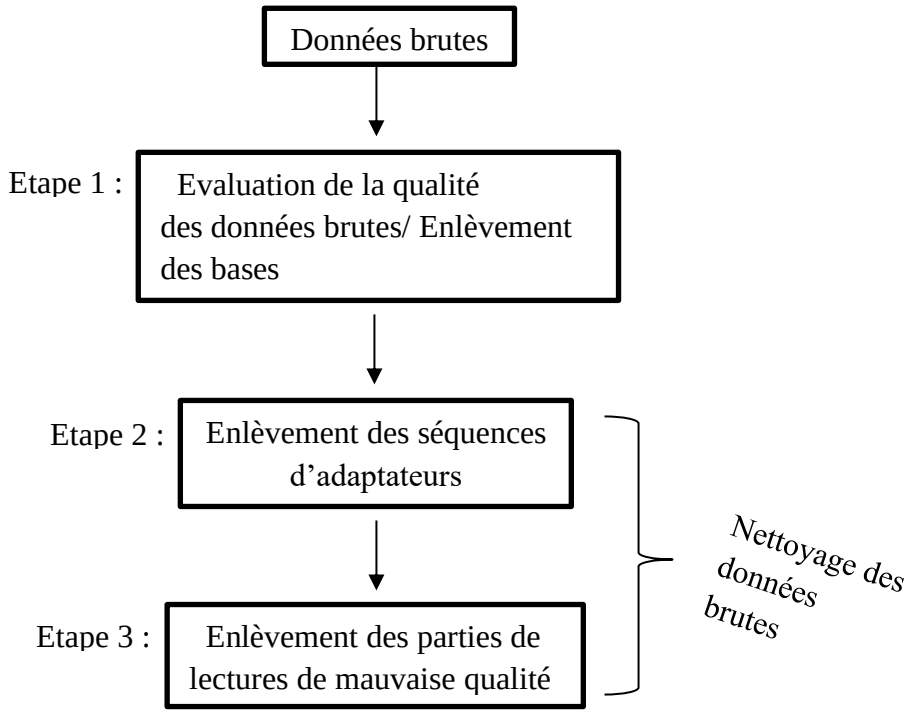
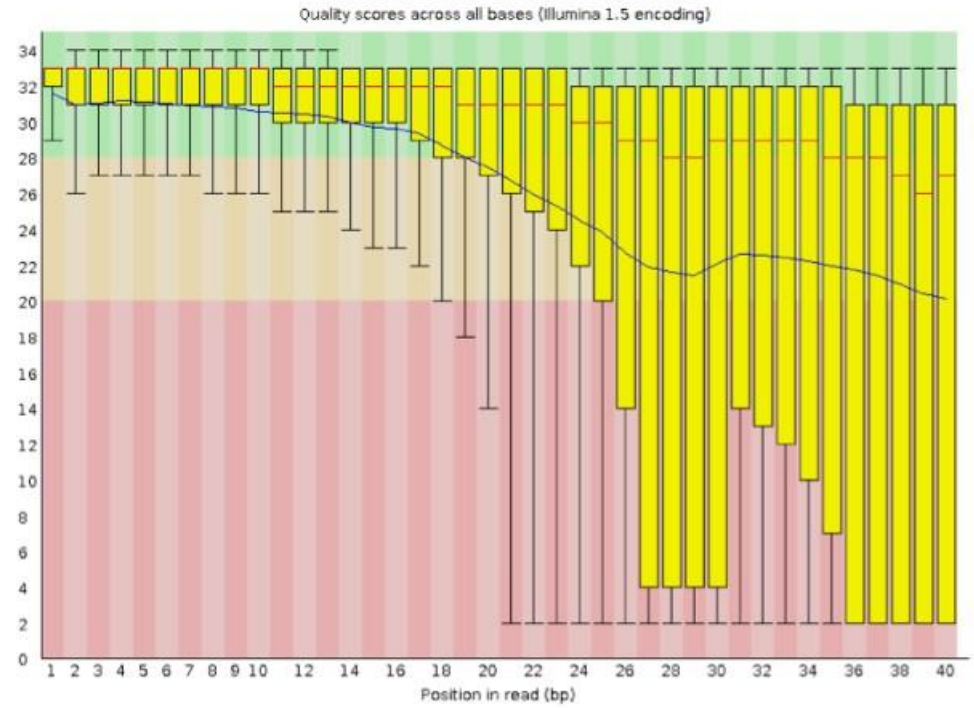


Figure 8 : Les types de processus de contrôle qualité

Le graphique suivant indique s'il faut trimmer les reads, et à partir de quelle position le faire : les séquences ayant des scores élevés sont conservées, tandis que celles ayant des scores faibles sont supprimées (figure 9).

A :



B :

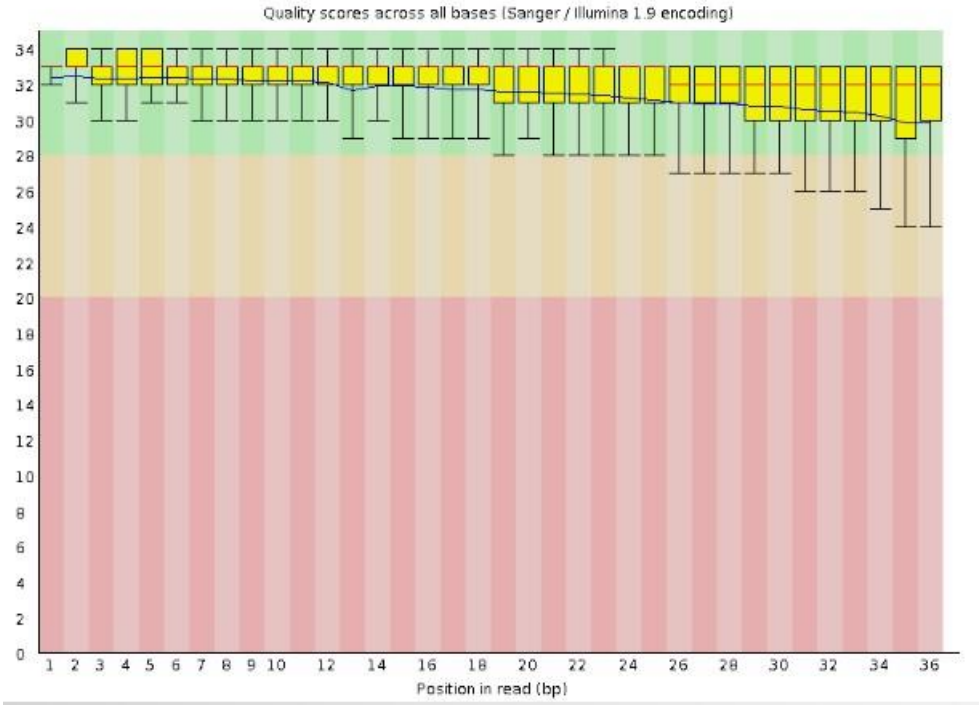


Figure 9 : A : Données de mauvaise qualité / B : Données de bonne qualité

3. METHODES D'ASSEMBLAGE DES GENOMES

Le processus de création d'un génome complet commence une fois que les lectures brutes de qualité suffisante ont été sélectionnées. Cette étape d'assemblage consiste à aligner et à fusionner les reads qui chevauchent partiellement ou totalement afin de former des contigs (séquences partielles) qui sont ensuite reliés entre eux pour former des scaffolds (séquences complètes) (figure 10).

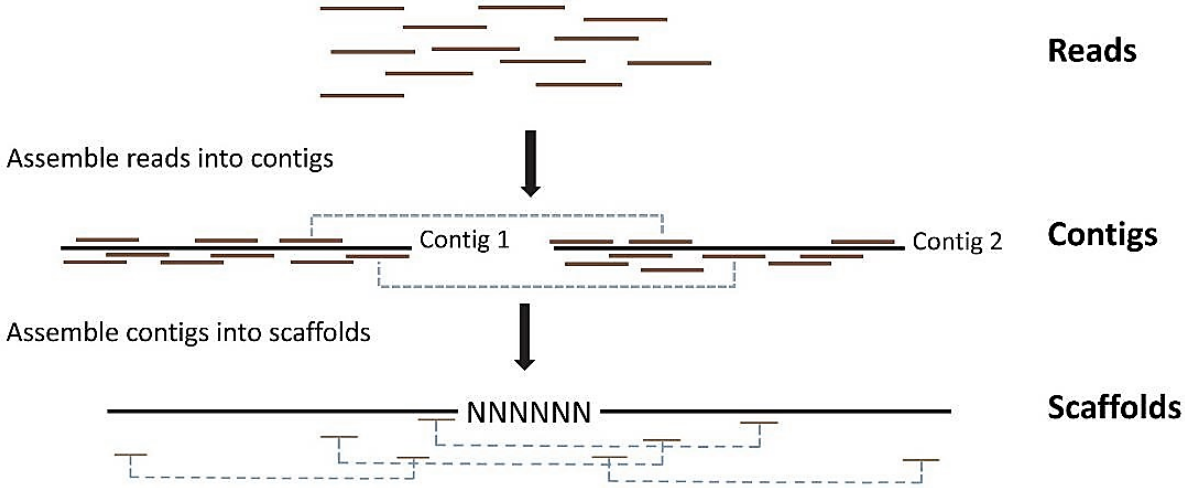


Figure 10 : Représentation schématique pour assembler les lectures en contigs et les contigs en scaffolds (Johnson MT., 2012).

Le tableau suivant indique l'assemblage du génome reposant sur la théorie des graphes.

Tableau 4 : Assemblage du génome (théorie des graphes).

Génome cible	ATTTGCGCAGAGAGCATTAGCTTGGCCCTAAAG
Reards	ATTTGC AGAGACCTAAG TTAGCTTGGC AAAG TGCGC AGA TGGCCCTAA
Over-lapping	ATTTGC AGAGACCTAAG TTAGCTTGGC AAAG TGCGC AGA TGGCCCT AA
Contigs	ATTTGCGCAGAGACCTAAG GAP 3 pd TTAGCTTGGCCCTAAAG

On distingue deux types d'assemblage des génomes :

- L'assemblage assisté par référence (mapping)
- L'assemblage *de novo* (*ab initio*) : sans génome de référence

a- Assemblage par rapport à un génome de référence (mapping)

Une approche qui consiste à aligner les lectures de séquençage d'un organisme contre un génome de référence préexistant. Cette méthode nécessite évidemment une séquence de référence très proche du génome que nous avons séquencé, (dans les cas où aucune référence n'est disponible, il faut utiliser les assemblages *de novo)*.

L'alignement des lectures brutes contre le génome de référence à l'aide d'un algorithme d'alignement, tel que Bowtie2, BLAST. Les lectures qui s'alignent sur le génome de référence sont appelées "lectures mappées", qui sont ensuite utilisées pour assembler des contigs, qui chevauchent les uns les autres. Les lectures qui ne s'alignent pas sur le génome de référence sont appelées "reads non mappés". Les modèles Markov peuvent être utilisés pour aider à aligner ces lectures non mappées en utilisant une approche d'assemblage *de novo.* Dans ce contexte, un modèle Markov est utilisé pour prédire la séquence de nucléotides la plus probable pour une région donnée de la lecture non mappée, en fonction des séquences précédentes dans la lecture et des probabilités de transition entre les différents nucléotides. Ces prédictions peuvent ensuite être utilisées pour assembler les lectures non mappées avec d'autres lectures qui chevauchent la même région du génome (Théroux,J,, 2015).

b- L'assemblage *de novo* (*ab initio*)

Il s'agit de résoudre un puzzle sans son modèle. Les fragments d'ADN qui sont chevauchants permettent petit à petit de reconstruire un contig. L'assemblage des contigs entre eux permet d'obtenir un scaffold. Cette technique est très couteuse en termes de calcul. Des algorithmes bioinformatiques comme le graphe de Bruijn, permettent de résoudre ce problème. Cette méthode est principalement employée pour reconstruire des génomes non connus (Dortet, L., 2017).

Le nombre de fois où une base d'ADN est couverte par une lecture de séquençage dans un assemblage de novo est nommé la couverture (coverage), en donne un exemple de couverture dans la figure suivante (figure11) (David G. ,2014) :

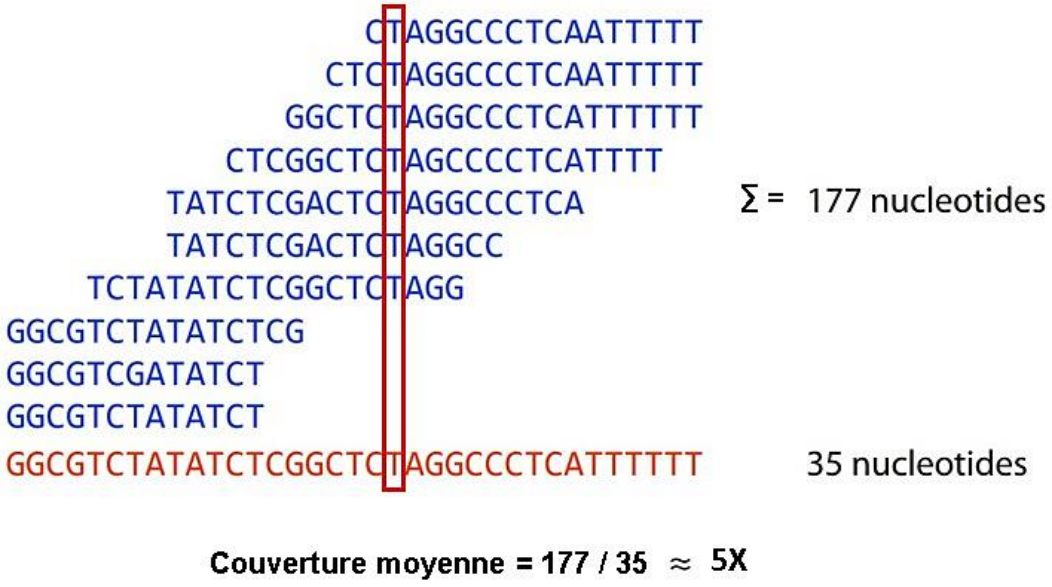


Figure 11 : exemplaire d'une couverture moyenne

- Les algorithmes d'assemblage des génomes :

1. Méthode de chemin Eulérien : est un chemin qui traverse chaque arête d'un graphe exactement une fois. Un chemin eulérien peut exister ou non dans un graphe donné, selon le degré (nombre d'arêtes incidentes à un sommet) de chaque sommet. Si tous les sommets d'un graphe ont un degré pair, un chemin eulérien existe. Si exactement deux sommets ont un degré impair, il existe un chemin eulérien qui commence à l'un des sommets de degré impair et se termine à l'autre. Les approches graphiques de De Bruijn peuvent être résolues même avec de grands ensembles de données complexes.

2. Méthode de chemin Hamiltonien : est un chemin dans un graphe qui passe par chaque nœud exactement une fois. Dans le contexte de l'assemblage de génome, les nœuds représentent les fragments de séquence d'ADN et les arêtes représentent les chevauchements entre ces fragments. Le but est de trouver un chemin Hamiltonien dans le graphe d'overlap qui passe par tous les fragments de séquence, ce qui permet de reconstituer le génome.

3. Graphe de De Bruijn (k-mers)

Le graphe de De Bruijn est un type de graphe utilisé en bio-informatique pour assembler les séquences d'ADN à partir de données de séquençage de nouvelle génération. Dans ce graphe, chaque séquence est divisée en k-mers, qui sont représentés par des nœuds dans le graphe. Les k-mers se chevauchent d'une base, ce qui signifie qu'il y a une arête reliant deux nœuds s'ils partagent une k-1 mer commune. Le graphe de de Bruijn permet de représenter les données de séquençage de manière compacte et est utile pour détecter les recouvrements et assembler les séquences en des séquences plus longues et continues (contigs).

4. Overlap-Layout-Consonsus (OLC)

Cette méthode se base sur la construction d'un graphe de chevauchement. Elle se divise en trois étapes : Overlap, Layout et Consensus (figure 2).

- La première étape consiste à comparer les reads entre eux afin de trouver des paires de reads qui se chevauchent. Le graphe de chevauchement est construit au fur et à mesure : les noeuds représentent les reads et deux noeuds seront connectés ensembles par une arrête s'il y a chevauchement entre ces reads.
- La deuxième étape consiste à rechercher dans le graphe le chemin passant par tous les nœuds permettant de reconstruire une séquence contiguë (chemin Hamiltonien). Cette étape permet de déterminer l'ordre dans lequel les reads seront assemblés en contigs.
- La troisième étape consiste, après correction des erreurs, à trouver une séquence consensus à partir des contigs obtenus à l'étape précédent

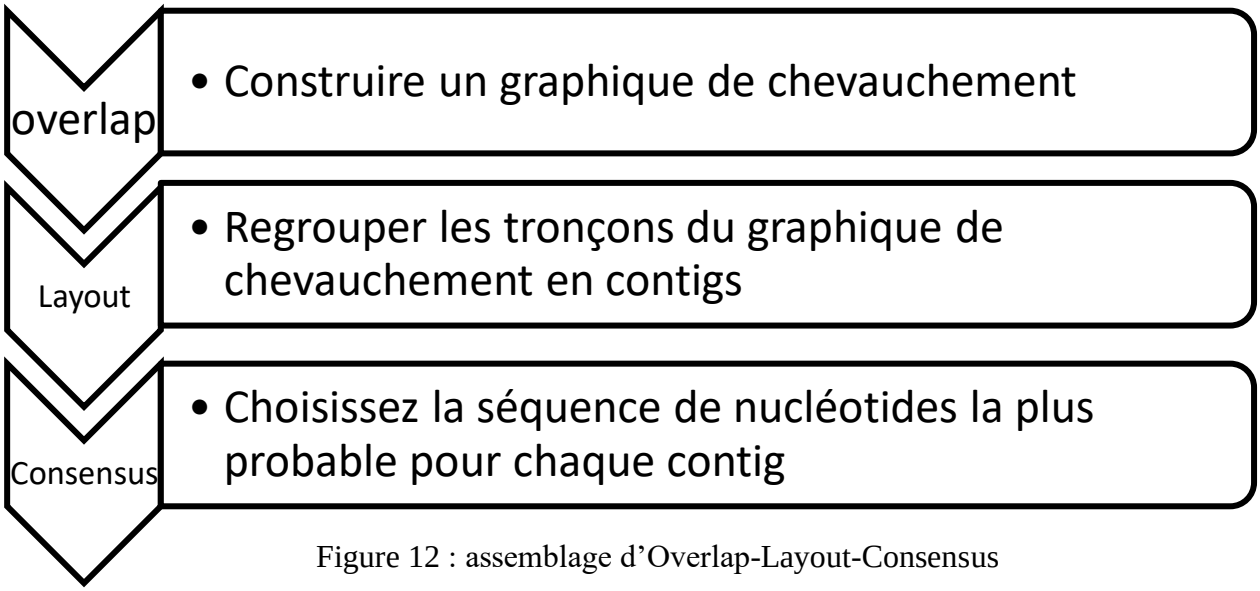


Figure 12 : assemblage d'Overlap-Layout-Consensus

Dans un exemple suivant, on suppose la séquence suivante : CGATTCTAAGT

Lors de son séquençage, les reads : CGATTCTA, TTCTAAGT, GATTCTAA sont obtenus. La figure suivante montre le principe de fonctionnement de la méthode OLC (Martin Ayling ,2020)(Figure 13).

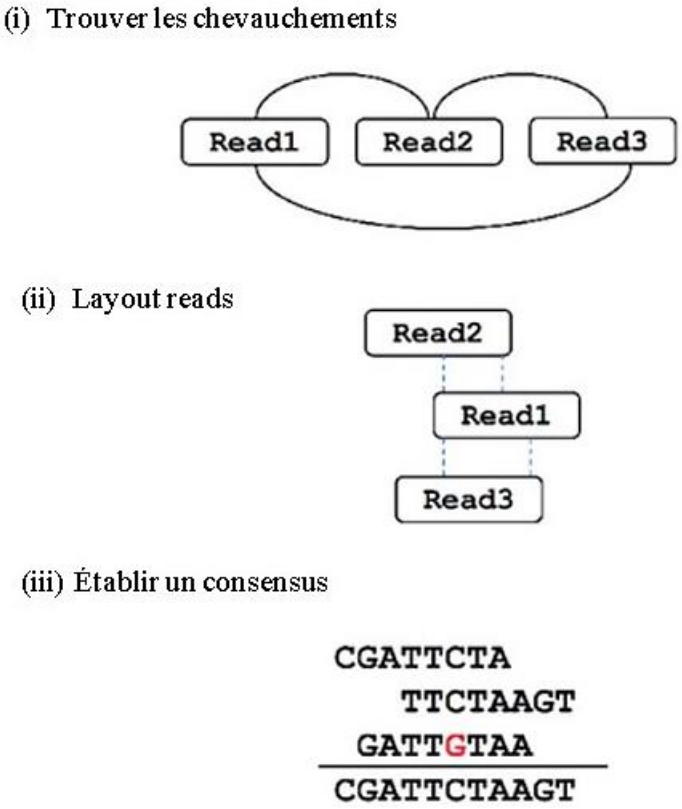


Figure 13 : Exemplaire de l'assemblage Overlap, Layout, Consensus

PARTIE 2
Partie expérimentale

1. MATÉRIEL

1.1. Données biologiques

Le type de données utilisé sont des séquences nucléiques (ADN)

1.2. Configuration de la machine : Les détails des caractéristiques de l'ordinateur utilisé sont répertoriés dans le tableau ci-dessous :

Tableau 5 : Description de la configuration utilisée pour l'application informatique.

Ordinateur	Caractéristiques
Processeur	Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.19 GHz
Mémoire RAM	8,00 Go
Système d'exploitation	Windows 10 Professionnel
Type de système	Système d'exploitation 64 bits, processeur x64
Version du système	22H2

1.3. Software

- Python : Le langage de programmation Python a été créé par Guido van Rossum à la fin des années 1980 et sa première version publique, Python 0.9.0, est sortie en 1991. Depuis lors, Python a connu une adoption croissante et est devenu l'un des langages de programmation les plus populaires dans de nombreux domaines tels que le développement web, l'analyse de données, l'intelligence artificielle, la bioinformatique, etc.

En étant un langage interprété et polyvalent, ce qui signifie que le code source est exécuté ligne par ligne par un interpréteur Python, cela permet un développement plus rapide et une itération plus facile lors de la création de programmes (python.org ,2020).

Bibliothèques Python utilisées :

- random : intégrée qui fournit des fonctionnalités pour générer des nombres aléatoires dans un contexte de programmation. Si vous avez besoin de vrais nombres aléatoires (random.org, 2023).
- Biopython : une bibliothèque de programmation en Python spécialisée dans la bioinformatique. Elle fournit des outils et des modules pour l'analyse de séquences biologiques, la manipulation de structures moléculaires, l'alignement de séquences,

l'interaction avec des bases de données biomoléculaires, et bien plus encore (biopython.org).

Le tableau suivant montre les outils et bibliothèques utilisés pour notre travail, avec les versions correspondantes :

Tableau 6 : Outils et bibliothèque utilisé

Outils / bibliothèques	Version
Python	Python 3.11.0
biopython	1.81

2. MÉTHODES

L'assemblage du génome est une étape clé dans le séquençage de l'ADN, où des petits fragments sont combinés pour reconstruire la séquence complète initiale. Les étapes du processus d'assemblage proposées dans ce travail sont :

1. La séquence d'origine (génome) est définie et répliquée (par exemple cinq fois) dans une liste. La longueur de la séquence est également imprimée :

   ['GTGTCACTTTCGCTGCGTGTCTTGCCCGAT', 'GTGTCACTTTCGCTGCGTGTCTTGCCCGAT', 'GTGTCACTTTCGCTGCGTGTCTTGCCCGAT', 'GTGTCACTTTCGCTGCGTGTCTTGCCCGAT', 'GTGTCACTTTCGCTGCGTGTCTTGCCCGAT']

   Longueur de séquence : 30 pb.

   Cette polymérisation *in silico* est une étape qui permet d'avoir, virtuellement, plusieurs copies du génome initial, car, *in vitro*, le génome est purifié dans le tube à essai en des milliers d'exemplaires.

2. Importer le module **random** pour générer des nombres aléatoires. Cette étape va permettre de couper aléatoirement les différentes copies de notre génome initial. Cette étape est justifiée par le fait que, *in vitro*, l'hydrolyse des copies du génome a lieu de façon aléatoire soit avec des ultrasons soit avec des jeux d'enzymes de restriction (nucléases).

   'cut_sequence' : est une fonction qui lit le génome et tient compte des longueurs de lecture minimale et maximale en entrée. Elle coupe aléatoirement le génome en sous-séquences (lectures ou reads) de longueurs variables (min_read_length et max_read_length) et renvoie une liste des lectures générées.

3. **'filter_reads'** : est une fonction qui prend une liste de lectures pour les filtrer si elles sont en duplicate, en triplicate etc., et si elles ont une taille en nombre de nucléotides très petite (par exemple cinq nucléotides). Elle renvoie une nouvelle liste contenant les lectures filtrées avec des longueurs ≥ 5 pb. Cette longueur limite de 5pb est choisie par nos soins pour avoir des reads qui feront au moins 15% de la taille initiale ; ce qui permettra d'éviter au maximum le problème de gaps interne et terminaux à la fin de l'assemblage.

4. **'find_overlap'** : elle est définie pour trouver le chevauchement (appariement, ou identité) entre deux lectures. Elle prend une paire de reads en entrée et renvoie la longueur du chevauchement.
   - Elle parcourt toutes les lectures et détermine le meilleur (optimal) chevauchement entre elles. Si un chevauchement valide est trouvé, le contig est alors créé en combinant les deux lectures dans le sens 5' vers 3'.
   - Elle continue de former des scaffolds en déterminant les meilleurs chevauchements entre les contigs. Le processus est similaire à l'étape d'assemblage des contigs, à la différence que des scaffolds sont désormais créés à la place des contigs.

5. **'assemble_scaffolds'** : permet de fusionner deux scaffolds en fonction du meilleur chevauchement et de les fusionner jusqu'à ce qu'il ne reste qu'une seule séquence, appelée « séquence assemblée » laquelle sera affichée avec sa longueur en pb.

6. '**coverage_percentage**' : Calcule le pourcentage de couverture totale en divisant la longueur du génome final assemblé par la longueur du génome initial et en multipliant par 100 (`% coverage = (l assembly / L init) * 100`). Le pourcentage de couverture est alors imprimé. Cette mesure de pourcentage de couverture permet d'évaluer à quel point la séquence assemblée est couverte par la séquence d'origine et donne alors une première estimation sur la qualité de cet assemblage. En effet si la couverture est faible, cela signifie que le génome assemblé est nettement inférieur au génome initial ; ce qui se traduit par de graves pertes d'informations cliniques et une perte considérable de la cridibilité de l'application et même du diagnostic médical.

7. Provenant de la bibliothèque 'biopython', le module '**Align**' effectue un alignement de séquence local par paires à l'aide de Align.PairwiseAligner. L'alignement local est une approche plus précise et plus spécifique permettant de trouver des régions identiques (ou du moins similaires) du génome assemblé comparativement au génome initial.

La figure suivante représente le principe global du fonctionnement du code (figure 14).

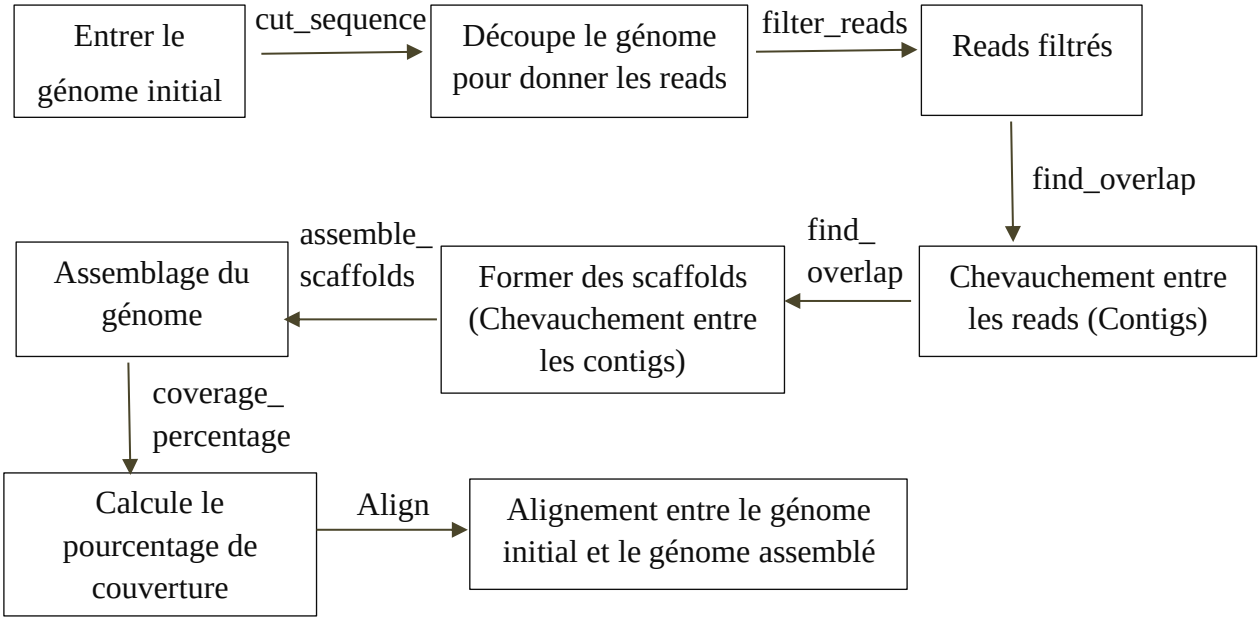


Figure 14 : processus d'assemblage du génome et de détection de chevauchements

Les résultats d'assemblage sont basés sur l'approche d'assemblage *de novo*, dans laquelle les lectures sont d'abord alignées pour former des contigs. Les contigs sont ensuite fusionnés pour construire des scaffolds, et ces derniers sont à leur tour assemblés pour générer la séquence finale, représentant le génome assemblé. Cette méthode permet la reconstruction du génome sans dépendre d'une référence génomique préexistante.

3. RÉSULTATS

Dans cette partie, nous examinerons les lectures découpées (reads) à partir des génomes, en mettant l'accent sur les lectures filtrées et les chevauchements détectés entre elles. Nous présenterons également les contigs assemblés et les scaffolds obtenus à partir des chevauchements inter-contigs. Ces résultats nous permettront de mieux comprendre la structure des génomes d'ADN assemblés et d'évaluer l'efficacité de l'application d'assemblage employée à cet effet.

Le génome initial est polymérisé (dans cet exemple : cinq fois ; voir partie méthode : 'GTGTCACTTTCGCTGCGTGTCTTGCCCGAT'), puis coupé aléatoirement ; les fragments obtenus sont stockés dans une liste. La longueur de la séquence génomique prise comme exemple est calculée et affichée.

La liste suivante montre les reads (n = 22) obtenues de longueurs variables après coupure aléatoire :

```
reads:
read 1: GTGTCACT
read 2: TTCGCTG
read 3: CGTG
read 4: TCTTGCCCG
read 5: AT
read 6: GTGTCAC
read 7: TTTCGCTGCG
read 8: TGTCTT
read 9: GCCCGAT
read 10: GTGTCA
read 11: CTTTC
read 12: GCTGC
read 13: GTGTCTTGCC
read 14: CGAT
read 15: GTGTCA
read 16: CTTTCGCTG
read 17: CGTGTCT
read 18: TGCCCGAT
read 19: GTGTCACTT
read 20: TCGCTGCGTG
read 21: TCTTGCCCGA
Read 22 : T
```

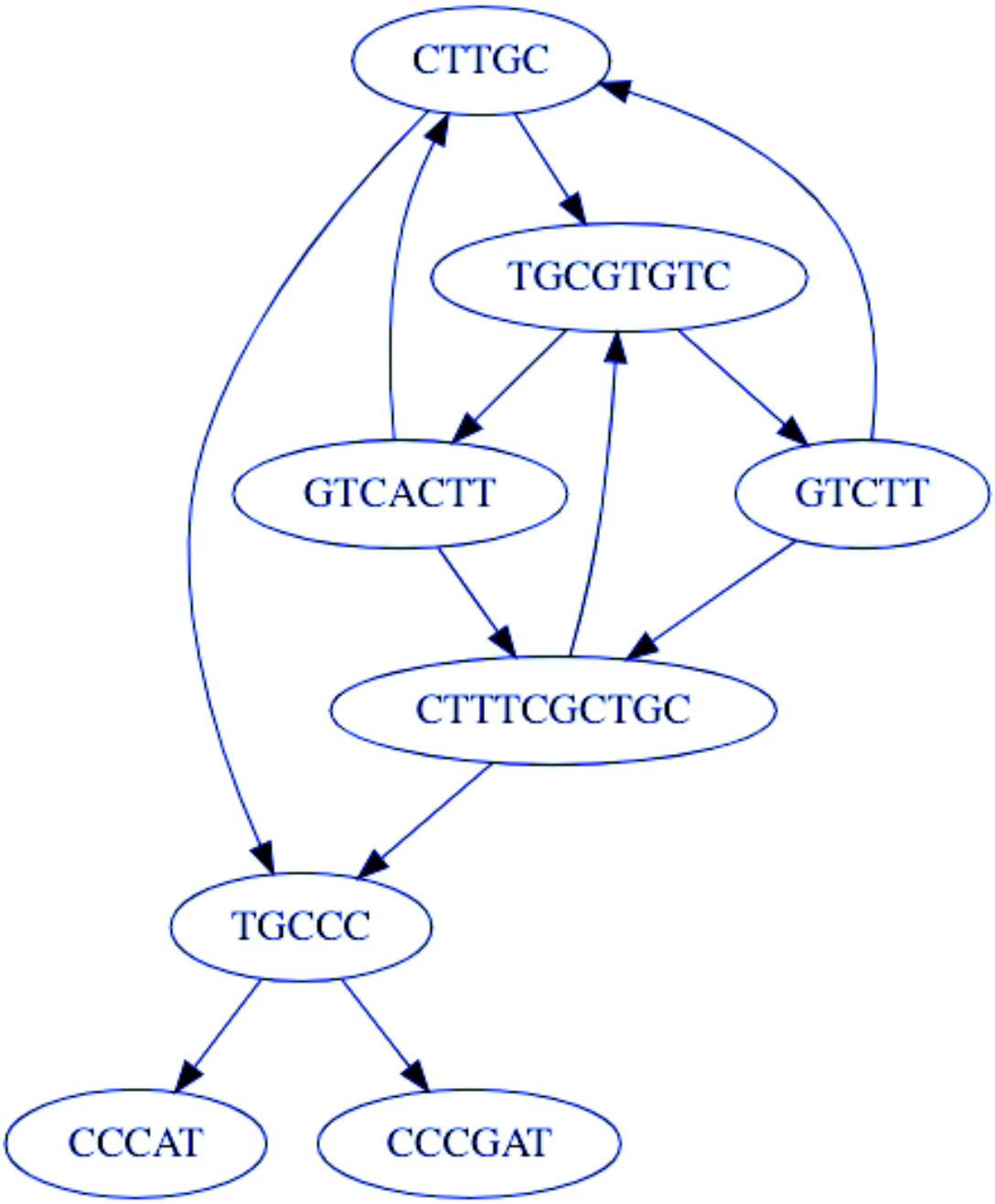
Nous constatons la présence de courtes séquences mononucléotidique comme le read n°22 représenté par le nucléotide Thymine ou encore des reads de longueur courte de quatre nucléotides tels que le read n°14 avec la succession tétranucléotidique égale à CGAT.

Le résultat suivant affiche les reads filtrées après avoir supprimé les séquences répétées et les lectures de moins de cinq nucléotides.

```
reads après filtrage:
read_filtrées 1: GTGTCACT
read_filtrées 2: TTCGCTG
read_filtrées 3: TCTTGCCCG
read_filtrées 4: GTGTCAC
read_filtrées 5: TTTCGCTGCG
read_filtrées 6: TGTCTT
read_filtrées 7: GCCCGAT
read_filtrées 8: GTGTCA
read_filtrées 9: CTTTC
read_filtrées 10: GCTGC
read_filtrées 11: GTGTCTTGCC
read_filtrées 12: CTTTCGCTG
read_filtrées 13: CGTGTCT
read_filtrées 14: TGCCCGAT
read_filtrées 15: GTGTCACTT
read_filtrées 16: TCGCTGCGTG
read_filtrées 17: TCTTGCCCGA
```

Les reads résultant de cette étape de filtration révèlent une longueur moyenne de 7,82 pb et une variance de 3,03 donc un écart-type de 1,74. Cette longueur moyenne des reads filtrées représente 26% de la taille initiale de notre génome test. C'est valeur très appréciable compte tenu des longueurs de reads effectivement utilisées sur les plateformes NGS qui correspondraient à des valeurs avoisinant les 900pb/34.491pb dans le cas du chromosome 22 humain ; soit 2,61%.

Le graphe de DeBruijn (https://0petya.github.io/debruijn-assembler/) résultant de ces reads avec une valeur de k = 4 est le suivant :



Il est à constater que sur ce graphe nous ne pouvons avoir de chemin Eulerien car certains nœuds ne sont pas équilibrés et montrent un nombre d'entrées différents du nombre de sorties.

Après avoir parcouru tous les reads, le code affiche les informations sur les contigs assemblés. Il itère sur la liste des contigs et récupère les informations pour chaque contig.

Contig 1: TTCGCTGCGTG (read 2: TTCGCTG - read 16: TCGCTGCGTG),
Chevauchement de 6 nucléotides : TCGCTG
Contig 2 : TCTTGCCCGAT (read 17 : TCTTGCCCGA - read 14 : TGCCCGAT),
Chevauchement de 7 nucléotides : TGCCCGA
Contig 3: TTTCGCTGCGTG (read 5: TTTCGCTGCG - read 16: TCGCTGCGTG),
Chevauchement de 8 nucléotides: TCGCTGCG
Contig 4: TGTCTTGCCCG (read 6: TGTCTT - read 3: TCTTGCCCG),
Chevauchement de 4 nucleotides: TCTT
Contig 5: CTTTCGCTGCG (read 12: CTTTCGCTG - read 5: TTTCGCTGCG),
Chevauchement de 8 nucléotides : TTTCGCTG
Contig 6 : GCTGCCCGAT (read 10 : GCTGC - read 14 : TGCCCGAT),
Chevauchement de 3 nucléotides : TGC
Contig 7: GTGTCTTGCCCG (read 11: GTGTCTTGCC - read 3: TCTTGCCCG),
Chevauchement de 7 nucleotides: TCTTGCC
Contig 8: CGTGTCTTGCC (read 13: CGTGTCT - read 11: GTGTCTTGCC),
Chevauchement de 6 nucléotides : GTGTCT
Contig 9 : GTGTCACTTTC (read 15 : GTGTCACTT - read 9 : CTTTC),
Chevauchement de 3 nucléotides : CTT
Contig 10: TCGCTGCGTGTCT (read 16: TCGCTGCGTG - read 13: CGTGTCT),
Chevauchement de 4 nucléotides : CGTG

L’assemblage des contigs pour visualiser et d'analyser les scaffolds obtenus. Les informations affichées fournissent des détails sur la structure et les connexions entre les contigs.

Scaffold 1 : TTCGCTGCGTGTCT (contig 1 : TTCGCTGCGTG - contig 11 : TCGCTGCGTGTCT), Chevauchement de 10 nucléotides : TCGCTGCGTG
Scaffold 2 : TTTCGCTGCGTGTCT (contig 3 : TTTCGCTGCGTG - contig 11 : TCGCTGCGTGTCT), Chevauchement de 10 nucléotides : TCGCTGCGTG
Scaffold 3 : TGTCTTGCCCGAT (contig 4 : TGTCTTGCCCG - contig 2 : TCTTGCCCGAT), Chevauchement de 9 nucléotides : TCTTGCCCG
Scaffold 4 : CTTTCGCTGCGTG (contig 8 : CTTTCGCTGCG - contig 3 : TTTCGCTGCGTG), Chevauchement de 10 nucléotides : TTTCGCTGCG
Scaffold 5 : GTGTCTTGCCCGAT (contig 7 : GTGTCTTGCCCG - contig 2 : TCTTGCCCGAT), Chevauchement de 9 nucléotides : TCTTGCCCG
Scaffold 6 : CGTGTCTTGCCCG (contig 9 : CGTGTCTTGCC - contig 7 : GTGTCTTGCCCG), Chevauchement de 10 nucléotides : GTGTCTTGCC
Scaffold 7 : GTGTCACTTTCGCTGCG (contig 10 : GTGTCACTTTC - contig 5 : CTTTCGCTGCG), Chevauchement de 5 nucléotides : CTTTC
Scaffold 8 : TCGCTGCGTGTCTTGCC (contig 11 : TCGCTGCGTGTCT - contig 9 : CGTGTCTTGCC), Chevauchement de 7 nucléotides : CGTGTCT

Conséquemment, la détection des chevauchements entre les scaffolds va permettre d’assembler le génome.

Séquence assemblée : GTGTCACTTTCGCTGCGTGTCTTGCCCGAT
Longueur de la séquence assemblée : 30

Le résultat comparé avec la séquence d’origine et le pourcentage élevé indique une bonne couverture, ce qui est préférable lors de l'assemblage de séquences génomiques.

Pourcentage de couverture : 100.0

L’alignement suggère que les deux séquences sont hautement similaires, La valeur du score 30.0 indique la similarité totale entre les deux séquences alignées. En alignement local, le score est généralement calculé en attribuant des points pour les correspondances et des pénalités pour les écarts ou les mutations.

Score 30.0

```
G T G T C A C T T T C G C T G C G T G T C T T G C C C G A T
| | | | | | | | | | | | | | | | | | | | | | | | | | | |
G T G T C A C T T T C G C T G C G T G T C T T G C C C G A T
```

Conclusion

CONCLUSION

L'étude des génomes dans les applications cliniques et autres agonomiques et/ou environnementales est essentielle pour comprendre le fonctionnement et les caractéristiques d'un organisme et de leurs écosystèmes, grâce à des techniques de NGS, y compris l'appel des bases, le contrôle de qualité et le prétraitement des données, ainsi que les méthodes d'assemblage du génome. Ces connaissances permettent de comprendre et d'interpréter les résultats des analyses de séquençage NGS afin de mener à bien les analyses génomiques approfondies.

L'approche *de novo* d'assemblage des génomes constitue une percée majeure dans le domaine de la génomique. C'est une méthode puissante pour reconstruire des séquences génétiques complètes sans dépendre de références préexistantes surtout dans le où génome est rare voire absent des bases de données. Cette approche permet de relever des défis complexes tels que l'identification de nouvelles variants génétiques (mutants) et la découverte de séquences non référencées. Grâce aux avancées technologiques et bioinformatiques, l'assemblage *de novo* du génome ouvre de nouvelles perspectives pour l'étude de la diversité génétique, la compréhension des maladies complexes et l'évolution des espèces. Cependant, des défis demeurent, notamment en termes de coût, de complexité et de gestion des données massives. Néanmoins, cette approche continue de stimuler la recherche et de contribuer à notre compréhension globale de la complexité et de la richesse du génome.

Quant à notre tentative pratique nous estimons que les résultats peuvent être nettement améliorés voire corrigés en associant les compétences et partageant les tâches entre les différentes disciplines concernées par les défis de la bioinformatique appliquée.

Indubitablement, nous prévoyant l'amélioration de cet essai par l'ajout de sous programmes de lectures et d'exploitations de fichiers initiaux de type fastq pour tester et valoriser l'application sur de reelles données massives, d'éliminer les contigs et les scaffolds en répétition sans pour autant perdre de l'information génétique et surtout présenter un sou programme capable de générer des graphes d'assemblage basés sur la théorie des graphes.

Références Bibliographique

RÉFÉRENCES

Andrews, S. Bittencourt, Bittencourt A (2010). FastQC: a quality control tool for high throughput sequence data

apollo-institute. ion torrent sequencing. (2021). URL:https://apollo-institute.org. Consulté le : 10.04.2023, à 18h00min.

Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. Brief Bioinform. 2020 Mar 23;21(2):584-594. doi: 10.1093/bib/bbz020. PMID: 30815668; PMCID: PMC7299287.

Biopython, 2023. URL : https://biopython.org/. Consulté le: 14.06.2023.

Biorigami. Renaud blervaque. (2013). Séquençage haut-débit de deuxième génération : Principes et caractéristiques.

Bray, D., Alberts, B., Hopkin, K., Johnson, A. (2012). L'essentiel de la biologie cellulaire. éd. *Médecine-sciences Lavoisier*, deuxième édition. pp : 510-12.

Centre national de la recherche scientifique. Aucune. URL : https://www.cnrs.fr. Consulté le : 08.03.2023, à 10h30min.

Chadi Saad. Caractérisation des erreurs de séquençage non aléatoires : application aux mosaïques et tumeurs hétérogènes. Médecine humaine et pathologie. Université de Lille, 2018. Français.

Dijon, Stéphanie Le Gras. (2017). Control qualité des données brutes, nettoyage des données. Manipulation des fichiers. FASTQ.

Dortet, Laurent & Bonnin, Rémy & Naas, Thierry. (2017). Impact du séquençage d'ADN à haut débit sur la surveillance des épidémies de bactéries multi-résistantes aux antibiotiques. Feuillets de biologie. P:354.

Gauthier, Michel. (2007). Simulation of polymer translocation through small channels: A molecular dynamics study and a new Monte Carlo approach.

Génome. Wikipédia.URL: http://fr.wikipedia.org. Consulté le: 08.02.2023, à 15h30min.

Johnson MT, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, et *all.* Evaluating methods for isolating total RNA and predicting the success of sequencing

phylogenetically diverse plant transcriptomes. PLoS One. 2012 ;7(**11**) : p :50226. doi: 10.1371 pub 2012 Nov 21.

Le cycle cellulaire. Ronald Dery. URL: https://youtu.be/fQ1CSETTOqo. Consulté le : 08.03.2023, à 22h55min.

Martin Krahn, Nicolas Lévy, et Marc Bartoli. Le séquençage de nouvelle génération (*Next-Generation Sequencing*, ou NGS) appliqué au diagnostic de maladies monogéniques hétérogènes. 2016 :(**13**): pp :31-33.doi : https://doi.org/10.1051/myolog/201613008.

Michel A Quail ,Iwanka Kozarewa ,Frances Smith ,Aylwyn Scally ,Philippe J, Stephens et *al*. (2008). Améliorations apportées par un grand centre de génomique au système de séquençage Illumina. Méthodes naturelles, 12. pp :1005-10.

National human genom research institute. Eric green. URL: https://www.genome.gov. Consulté le : 09.03.2023, à 21h30min.

Parlons sciences. (2020). Séquençage de sanger. URL : https://parlonssciences.ca. Consulté le : 03.03.2023, à 20h20min.

Pierce, M. B. A. (2012). L'essentiel de la génétique : Concepts and connections. éd. *Pierce*, première édition. Pp :260-267.

Planet vie. Gilles Furelaud, Yann Esnault. (2004). Le séquençage des génomes.URL : https://planet-vie.ens.fr. Consulté le: 03.03.2023, à 10h00min.

Pr. Christopher Burge, Pr. David Gifford et Pr. Ernest Fraenkel, 2014. Fondements De La Biologie Computationnelle Et Des Systèmes

Quentin Testard. Industrialisation des procédures d'analyses de données de séquençage pangénomiques constitutionnelles. Biologie du développement. Université Grenoble Alpes.(2021)

Random Navigator — Anaconda documentation, 2020. URL : https:// random.org,. Consulté le : 06.06.2023.

Sherwood, RI., Hashimoto, T., O'donnell, CW.,Lewis, S., et *all*. (2014). Découverte de facteurs de transcription pionniers directionnels et non directionnels en modélisant l'amplitude et la forme du profil de la DNase. *Biotechnologie de la nature, 32* (2),pp : 171-8.

Théroux, Jean-François. (2015). Développement de méthodes d'assemblage de génomes de novo adaptées aux bactéries endosymbiotes. : doi : https://doi.org/1866/13126

Thomas, D., William, C. (2021). Biologie cellulaire. éd. *Campus masson*. pp :20-6.

Tutorials migale inrae. Olivier Rué. (2022) Contrôle qualité des données de séquençage Illumina.https://tutorials.migale.inrae.fr. Consulté le : 05.04.2023.

Van der Auwera GA, et al. (2013). From FastQ data to high confdence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics.

Watson, J., Levine, M., Bell, S., Losick, R., Gann, A., Baker, T. (2012). Biologie moléculaire du gène. éd. *Pearson*, sixième édition. p .105.

Welcome to Python.org, 2020. Python.org. URL: https://www.python.org. Consulté le : 29.05.2023.

William, S., Michael, R. Charlotte, A. (2006). Génétique.éd. *Education France/Pearson*, Huitième édition. P :272.

Année universitaire : 2022-2023	**Présenté par :** AMOUCHE Selma BENLAMRI Aya Sara

Thème :
Une approche *de novo* pour l'assemblage des génomes

Mémoire pour l'obtention du diplôme de Master en bioinformatique

Le but de ce travail est de proposer une approche *ab initio* pour la reconstruction de génome en utilisant des algorithmes appropriés. L'objectif principal est de surmonter les défis actuels liés à la reconstruction de génome, tels que la présence de régions répétées, les erreurs de séquençage et les limitations des technologies existantes. Notre approche vise à améliorer la précision et la rapidité de la reconstruction de génome, l'analyse comparative, et l'exploitation des informations génomiques complémentaires. L'idée est de combiner ces différentes approches pour obtenir des résultats plus fiables et plus complets, permettant ainsi de reconstruire des génomes de manière plus.

Mots-clefs : Séquençage NGS ; Chevauchement ; Assemblage de génome ; Graph de De Bruijn.

Encadreur : HAMIDECHI. M. A — Université Frères Mentouri, Constantine 1
Présidant : TAMAGOULT.M — Université Frères Mentouri, Constantine 1
Examinateur : CHEHILI.H — Université Frères Mentouri, Constantine 1