République Algérienne Démocratique et Populaire Ministère

De l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الإخوة منتوري قسنطينة 1
Frères Mentouri Constantine I University
Université Frères Mentouri Constantine 1

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

Un Workflow pour l'implémentation d'un pipeline de traitement des données générées par le séquençage NGS

Présenté par :

HAMOUDA Rania **HAZOURLI Lina**

Devant le jury :

Président du jury : DAAS.M.S	Université Frères Mentouri- Constantine 1
Encadrant : CHEHILI.H	Université Frères Mentouri- Constantine 1
Examinateur : KELLOU.K	Université Frères Mentouri- Constantine 1

Année universitaire 2021-2022

Remerciement

الحمد لله الذي بنعمته تتم الصالحات، وبفضله تتنزل الخيرات والبركات وبتوفيقه تتحقق المقاصد والغايات.

Un grand merci à nos professeurs, **HAMIDECHI MOHAMED ABDELHAFID,** Monsieur **KELLOU KAMEL** et Monsieur **TEMAGOULT MAHMOUD**, et Monsieur **DAAS**. Merci à vous pour votre qualité d'enseignement vos conseils et votre disponibilité.

A notre encadrant monsieur **CHEHILI HAMZA** notre professeur mais pas que… merci pour vos efforts, vos conseils, votre temps et votre encouragement merci de nous pousser toujours à faire mieux et à dépasser nos limites MERCI.

On présente notre respect et remerciement pour Monsieur **BENSAADA MOUSTAFA** et le staff **du CHU Saadna Abdenour de Sétif** d'avoir accepté notre visite pour un petit stage et de nous donner la chance d'assisté au premier séquençage en Algérie.

On présente notre remerciement à **Monsieur SAYAD MOHAMED EL AMINE** et Monsieur **SERRAR RAID** pour leur accompagnement leur paissance et leur partage.

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

MERCI à nos professeurs de la BIOINFORMATIQUE pour leur paissance leur investissement et encouragement.

Dédicaces

En tout premier lieu, je remercie le bon Dieu, tout puissant, de m'avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés.

À

Mes meilleures personnes dans ma vie chère mère et père

À

Ma grand-mère

À

Mon gentil frère

À

Ma sœur douce et adorable

À

Tous les amis qui m'ont aidé Narimane et Abderahmane

À

Lina, chère amie avant d'être binôme

HAMOUDA Rania

Dédicaces

بسم الله الرحمان الرحيم

A mes chers parents, pour leur sacrifice, leur amour, leur tendresse, leur soutien et leur prière tout au long de mon parcours.

A mes amis, ma deuxième famille, merci pour votre fidélité, votre soutien vos encouragements : (HANINE, ASSALA, HADIL, HAZAR).

A la mémoire de ma chère Hanane, que ce travail restera une Charité continue pour ton âme.

A ma grande famille (HAZOURLI, NAHOUI), merci pour votre présence et soutien merci pour votre amour.

A mes professeurs, merci pour vos encouragements, votre disponibilité merci pour LE SAVOIR que vous m'avez donné.

Je remercie mon binôme HAMOUDA Rania pour la patience, l'effort et la compréhension merci à toi Rania. Merci à nous.

A tous ceux qui, par un mot, m'ont donné la force de continuer…

HAZOURLI Lina

Résumé

Le mot « cancer » désigne des maladies très différentes les unes des autres. Le cancer se traduit par une multiplication anarchique de cellules qui forment une tumeur maligne. Avec le temps, la tumeur cancéreuse peut se propager dans l'organisme et former des métastases.

La détection des mutations somatiques présente un grand potentiel dans le traitement du cancer et a été un domaine de recherche très actif au cours des dernières années, en particulier depuis la percée de la technologie de séquençage de nouvelle génération (NGS). Notre étude permet de faire une liaison entre le domaine de l'oncologie et le séquençage, dans l'espoir de fournir un guide pratique pour sélectionner le pipeline approprié pour des applications spécifiques. Nous avons concentré notre étude sur la détection des variants somatiques.

À la fin, 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 le séquenceur Ion Torrent, on a aussi pu avoir accès à des données réelles ici en Algérie dans le laboratoire de diagnostic du CHU Sétif.

Mots clés : Tumeur, NGS, oncologie, séquençage, pipeline, variants somatiques.

Abstract

The word "cancer" refers to diseases that are very different from each other. Cancer results in an anarchic multiplication of cells which form a malignant tumor. Over time, the cancerous tumor can spread throughout the body and form metastases.

Somatic mutation detection has great potential in cancer therapy and has been a very active area of research in recent years, especially since the breakthrough of next-generation sequencing (NGS) technology. Our study bridges the field of oncology and sequencing, hoping to provide a practical guide to selecting the appropriate pipeline for specific applications. We focused our study on the detection of somatic variants.

In the end, we created our own pipeline using free access software and tools to test sequences that are already sequenced by Ion Torrent, we were also able to access real data here in Algeria in the diagnostic laboratory of the CHU Sétif.

Keywords: Tumor, NGS, oncology, sequencing, pipeline, somatic variants.

ملخص

تشير كلمة "سرطان" إلى أمراض مختلفة تمامًا عن بعضها البعض. يؤدي السرطان إلى تكاثر فوضوي للخلايا التي تشكل ورمًا خبيثًا. بمرور الوقت ، يمكن أن ينتشر الورم السرطاني في جميع أنحاء الجسم ويشكل النقائل.

يتمتع اكتشاف الطفرات الجسدية بإمكانيات كبيرة في علاج السرطان وقد كان مجالًا نشطًا جدًا للبحث في السنوات الأخيرة ، خاصةً منذ طفرة تقنية الجيل التالي من التسلسل (**NGS**). تربط دراستنا مجال علم الأورام والتسلسل ، على أمل تقديم دليل عملي لاختيار خط الأنابيب المناسب لتطبيقات محددة. ركزنا دراستنا على اكتشاف المتغيرات الجسدية.

في النهاية ، أنشأنا خط الأنابيب الخاص بنا باستخدام برامج وأدوات الوصول المجاني لاختبار التسلسلات التي تم تسلسلها بالفعل بواسطة **Ion Torrent** ، وتمكنا أيضًا من الوصول إلى بيانات حقيقية هنا في الجزائر في مختبر التشخيص في **CHUSétif**.

الكلمات المفتاحية: الورم ، **NGS** ، علم الأورام ، التسلسل ، خط الأنابيب ، المتغيرات الجسدية.

Sommaire :

Sommaire

Liste des abréviations :

ADN : Acide désoxyribonucléique.

ARN : Acide ribonucléique.

BAM : Binaire-SAM.

Bp : Paire de bases.

BWA : Burrows-Wheeler-Aligner.

CNV : Variation du nombre de copies.

CWL : Langage de flux de travail commun.

Fast QC : Contrôle qualité rapide.

Indels : Insertions/suppressions.

GATK : Boîte à outils d'analyse du génome.

GUI : Interface utilisateur graphique.

NCBI : Centre national d'information sur la biotechnologie.

NGS : Séquençage de nouvelle génération.

SAM : Carte d'alignement de séquence.

SNV : Variante mononucléotidique.

SV : Variation structurelle.

SRA : Archive de lecture de séquence.

VCF : Variante de format d'appel.

Liste des figures :

Liste des tableaux :

Introduction Générale :

Le profilage moléculaire des biopsies tumorales joue un rôle de plus en plus important non seulement dans la recherche sur le cancer, mais aussi dans la prise en charge clinique des patients atteints de cancer. Les approches multi-omiques promettent d'améliorer les diagnostics, les pronostics et les traitements personnalisés. Pour tenir cette promesse d'oncologie de précision, des méthodes bio-informatiques appropriées pour gérer, intégrer et analyser des données volumineuses et complexes sont nécessaires [1].

L'amélioration continue, la plus grande disponibilité et la baisse des coûts du séquençage de nouvelle génération (NGS) ont permis aux principaux centres anticancéreux du monde entier d'offrir une oncologie personnalisée basée sur le NGS. L'objectif est de dresser le profil des aberrations génétiques des tumeurs telles que les variants mononucléotidiques (SNV), les variants du nombre de copies (CNV), les insertions et les délétions (indels), les variants structurels (SV). Ces approches peuvent être organisées soit sous la forme d'un seul comité institutionnel des tumeurs moléculaires (MTB), où les aberrations génétiques détectées seront évaluées pour tout traitement potentiel correspondant, soit sous la forme d'un essai collectif, dans lequel des altérations génétiques prédéfinies sont attribuées à des bras de traitement correspondants [2].

Les technologies de séquençage de nouvelle génération sont basées sur le déchiquetage de l'ADN en petits fragments et la détermination de la séquence nucléotidique dans ces fragments. Ces courtes lectures sont cartographiées sur le génome de référence pour identifier la séquence génomique de l'échantillon. Dans des échantillons de cancer, des séquences génomiques de tissus tumoraux et non tumoraux appartenant au même individu sont comparées pour trouver des mutations somatiques spécifiques au cancer.

Dans cette étude, nous discutons des exigences spécifiques des méthodes et logiciels bio-informatiques qui surviennent dans le cadre de l'oncologie, en raison d'un environnement réglementaire plus strict et du besoin de procédures rapides. Nous décrivons le flux de travail (workflow) d'un comité des tumeurs moléculaires et le support bio-informatique spécifique dont il a besoin (de l'analyse primaire des données brutes de profilage moléculaire à la génération automatique d'un rapport clinique et sa livraison aux oncologues décisionnaires).

De tels flux de travail ont été mis en œuvre à de divers degrés dans de nombreux essais cliniques, ainsi que dans des comités de tumeurs moléculaires dans des centres anticancéreux spécialisés et des hôpitaux universitaires du monde entier.

Ces deux étapes (alignement et appel de variantes) constituent les deux étapes importantes des analyses de séquençage du cancer, et de nombreux algorithmes ont été développés pour ces tâches. Ces algorithmes sont combinés dans des pipelines logiciels, qui prennent des données de séquençage brutes en entrée et produisent des changements spécifiques au cancer en sortie. Différents algorithmes de mappage et de découverte de variantes ont des hypothèses et des priorités différentes. Par conséquent, le nombre et le type de variantes identifiées par ces algorithmes peuvent varier considérablement. Cela fait des tests détaillés le workflow construits pour une utilisation efficace des technologies de séquençage. Et aussi, le plus important est de faciliter toutes les analyses en fabriquant un workflow qui comprend toutes les étapes du pipeline. L'objectif de ce travail est le développement d'un workflow constituée d'un ensemble de pipeline pour automatiser l'appel des variants.

Ce manuscrit est organisé en quatre chapitres. Le premier donne des informations sur le séquençage de l'ADN et son utilité dans l'oncologie de précision. Le deuxième chapitre aborde les méthodes computationnelles pour l'oncologie de précision. Quant au troisième chapitre, il présente notre contribution en décrivant le workflow et les données utilisées. Les résultats et la discussion sont détaillés dans le quatrième chapitre.

Chapitre 1 :

Séquençage d'ADN et l'oncologie

Introduction

Au cours des six dernières années, nous avons assisté à une révolution dans les technologies de séquençage qui a déjà eu un impact profond sur notre compréhension de la génétique et de la biologie du génome. Dans un contexte de recherche, le NGS a été largement mis en œuvre pour le séquençage du génome. Ces efforts de recherche ont ouvert la voie au développement de nouveaux protocoles (contextes moléculaires et bio-informatiques) et ont joué un rôle déterminant dans la compréhension des principales forces et faiblesses de cette technologie [3, 4].

Comme le cancer est une maladie génétique causée par des mutations héréditaires ou somatiques, les nouvelles technologies de séquençage de l'ADN auront un impact important sur la détection, la gestion et le traitement de la maladie.

Dans cette partie de notre mémoire, on va discuter des différents types et moyens de séquençage avec les différentes générations et aussi avoir une idée générale sur le cancer.

1- Technologies de séquençage de l'ADN :

La séquence de l'ADN contient l'information nécessaire aux êtres vivants, déterminer cette séquence est donc utile aussi bien pour les recherches visant à savoir comment vivent les organismes que pour des sujets appliqués, en médecine elle peut être utilisée pour identifier, diagnostiquer et potentiellement trouver des traitements à des maladies génétiques [5].

La séquence de l'ADN constitue en quelque sorte l'anatomie d'un génome, elle indique les formules des protéines, sa connaissance est cruciale pour la biologie. Le séquençage de l'ADN est une technique qui a révolutionné la biologie moléculaire [5].

1-1- Séquençage de l'ADN première génération :

Le séquençage des acides nucléiques est une méthode permettant de déterminer l'ordre exacte des nucléotides présents dans une molécule d'ADN ou d'ARN donnée. Au cours de la dernière décennie, l'utilisation du séquençage des acides nucléiques a augmenté, car la capacité de séquençage est devenue accessible aux laboratoires de recherches et aux laboratoires cliniques. La première grande incursion dans le séquençage de l'ADN a été le Projet du génome humain, un projet de 3 milliards de dollars sur 13 ans, achevé en 2003. Le projet du génome humain a été réalisé avec le séquençage de première génération, connu sous le nom de séquençage Sanger. Sanger sequencing (méthode de terminaison de chaîne), a été considéré comme l'étalon-or pour le séquençage des acides nucléiques pour les deux décennies et demie suivantes (Sanger et coll., 1977) [6]. Le tableau suivant présente la différence entre les deux méthodes de la première génération :

Sanger Séquençage	**Maxam Gilbert**
La méthode de séquençage Sanger a été introduite après la méthode de séquençage Maxam Gilbert.	Le séquençage Maxam Gilbert est la première technique développée pour le séquençage de l'ADN.
Usage	
Le séquençage Sanger est couramment utilisé pour le séquençage.	Cette méthode est rarement utilisée.
Utilisation de produits chimiques dangereux	
L'utilisation de produits chimiques dangereux est limitée par rapport à la méthode de Maxam Gilbert.	Il utilise des produits chimiques dangereux.
Étiquetage pour la détection	
Le séquençage de Sanger utilise des ddNTP marqués de manière radioactive ou fluorescente.	Cette méthode utilise du P radioactif 32 pour marquer les extrémités des fragments d'ADN.

Tableaux 1 : La différence entre la méthode Sanger et Maxam Gilbert [7].

1-2- Séquençage de nouvelle génération NGS :

Depuis l'achèvement de la première séquence du génome humain, la demande de méthodes de séquençage moins coûteuses et plus rapides a considérablement augmenté. Cette demande a mené au développement de méthodes de séquençage de deuxième génération ou séquençage de nouvelle génération (NGS). La technologie de séquençage massivement parallèle facilite le séquençage à haut débit, qui permet de séquencer un génome entier en moins d'une journée. Au cours de la dernière décennie, plusieurs plateformes ont fourni un séquençage à faible coût et à haut débit. Nous soulignons ici deux des plates-formes les plus couramment utilisées dans la recherche et les laboratoires cliniques aujourd'hui, la Technologie Ion Torrent Personal Genome Machine (PGM) et l'Illumina MiSeq. La création de ces plateformes et d'autres plateformes de NGS a rendu le séquençage accessible à un plus grand nombre de laboratoires, augmentant rapidement la quantité de recherche et de diagnostics cliniques effectués avec le séquençage des acides nucléiques [8].

Le séquençage est réalisé au sein d'automate dont le fonctionnement repose sur différentes technologies brevetées que l'on peut séparer en deux groupes (Figure 1) :

- Les technologies à lectures courtes (short reads)
- Les technologies à lecture longue (long reads)

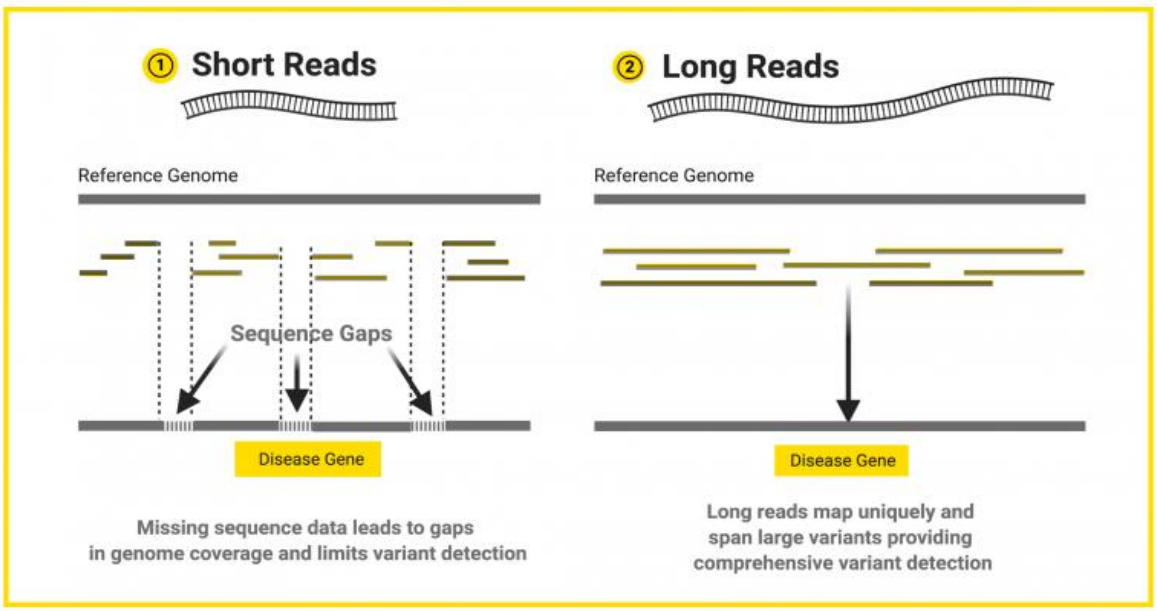

Figure 1 : La différence entre short et long reads [8].

Les applications de NGS semblent presque sans fin, permettant des avancées rapides dans de nombreux domaines liés aux sciences biologiques. Le séquençage du génome humain est en cours pour identifier les gènes et les éléments régulateurs impliqués dans les processus pathologiques.

Ce ne sont que quelques-unes des applications générales qui commencent à effleurer la surface de ce que le NGS peut offrir au chercheur et au clinicien. Comme le NGS continue de gagner en popularité, il est inévitable qu'il y ait d'autres applications novatrices [9].

2- Le cancer :

Le corps humain est complexe et présente, globalement, 3 niveaux d'organisation qui sont :

- **Les organes** (par exemple, le cœur, le cerveau, les poumons, etc.) qui ont tous un rôle différent et souvent fondamental dans le fonctionnement du corps.
- **Les tissus** qui composent les organes (par exemple, les muscles, les glandes, etc.) et structurent le corps (le squelette, la peau).
- **Les cellules,** enfin, qui sont l'unité de base des tissus, et qui présentent elles même une organisation microscopique et complexe.

Véritables petites usines autonomes, les cellules cohabitent et communiquent de façon harmonieuse afin de préserver l'architecture et les fonctions propres à chaque organe ou système. Elles sont constituées :

- D'une enveloppe membrane équipée de petites structures (récepteurs) permettant la communication avec l'environnement.
- De différents équipements internes qui sont essentiels au maintien en vie de la cellule et lui permettent de jouer son rôle dans le corps.
- D'un noyau situé également à l'intérieur de la cellule et qui contient toute l'information génétique que la cellule utilise pour savoir comment fonctionner, 46 chromosomes organisés en 23 paires et constitués de « gènes ».

2-1-La cellule cancéreuse :

C'est une cellule qui devient totalement indisciplinée, suite à une agression ou un dommage et liée à une modification de la structure d'un gène ; c'est ce qu'on appelle une «

mutation ». Parfois, l'agression est violente et courte. Le plus souvent, elle est de faible intensité, mais s'étend sur une longue période.

Cette altération intime de la cellule constitue la base même de tous les cancers. La cellule n'arrête plus de se multiplier, et reste en vie dans un organe où habituellement les cellules meurent et se renouvellent rapidement. Cette prolifération va aboutir à la formation de la tumeur, qui, en se développant, arrive à détruire les cellules normales avoisinantes.

Une cellule cancéreuse se multiplie beaucoup, elle commence à former un regroupement de cellules. Une tumeur devient dangereuse (maligne) lorsqu'elle commence à s'infiltrer, c'est-à-dire que les cellules cancéreuses, au lieu de rester groupées les unes aux autres, commencent à former des extensions vers des zones voisines [10].

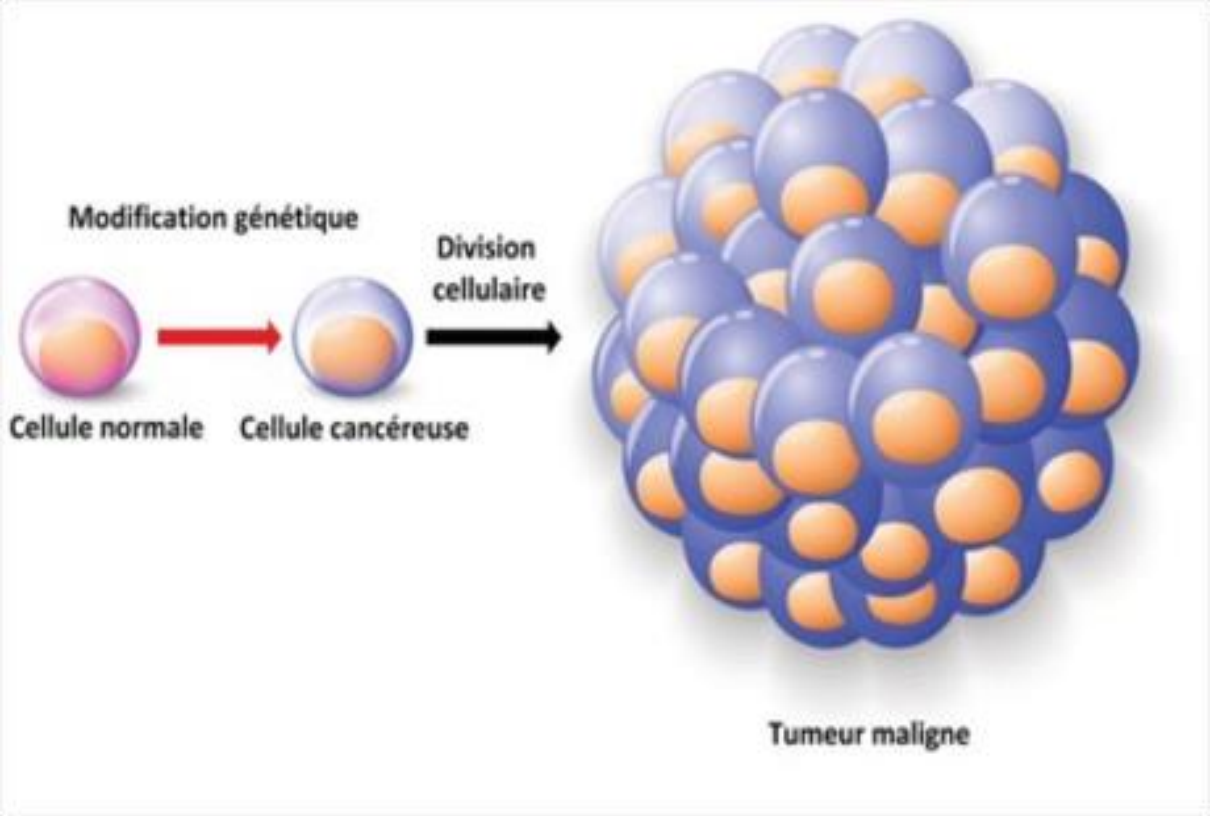

Figure 2 : Transformation les cellules cancéreuses.

2-2- L'oncologie et le séquençage à haut débit :

Avec le développement et l'amélioration de nouvelles technologies de séquençage, le séquençage de nouvelle génération (NGS) a été appliqué de plus en plus dans la recherche en génomique du cancer au cours de la dernière décennie. Plus récemment, le NGS a été adoptée

en oncologie clinique pour faire progresser le traitement personnalisé du cancer. Le NGS est utilisé pour identifier les mutations cancéreuses nouvelles et rares, détecter les porteurs de mutations cancéreuses familiales et fournir une justification moléculaire pour une thérapie ciblée appropriée. Par rapport au séquençage traditionnel, le NGS présente de nombreux avantages, comme la capacité de séquencer complètement tous les types de mutations pour un grand nombre de gènes (des centaines à des milliers) en un seul test à un coût relativement faible. Toutefois, des défis importants, en particulier en ce qui concerne l'exigence de tests plus simples, d'un débit plus souple, d'un délai d'exécution plus court et, surtout, d'une analyse et d'une interprétation plus faciles des données, devra être surmonté pour traduire le NGS au chevet des patients atteints de cancer. Dans l'ensemble, le dévouement continu à l'application du NGS dans la pratique de l'oncologie clinique nous permettra de faire un pas de plus vers la médecine personnalisée. Les technologies NGS ont permis une détection efficace et précise de mutations somatiques nouvelles et rares [4, 5].

Conclusion :

Les techniques de séquençage de nouvelle génération se sont à présent bien démocratisées. Le catalogue de séquenceurs s'est étoffé et offre à ce jour un large panel de machines adaptées aux différentes structures (laboratoire de recherche, milieu hospitalier...) et aux applications choisies. Grâce aux séquenceurs de deuxième génération, il est aujourd'hui possible de produire un très grand nombre de séquences, en un temps record et à un coût très faible. En moins de vingt ans, le coût du séquençage d'un génome humain a pu être divisé. Toutes ces avancées technologiques répondent aux nouvelles exigences d'une médecine plus personnalisée. En effet, l'utilisation des données de séquençage du génome du patient permet, dans un nombre croissant de cas, d'affiner le diagnostic médical et le pronostic, afin de choisir le traitement le plus adapté à la pathologie décelée.

Aujourd'hui, la prise en charge oncologique des tumeurs avancées ou métastatiques requiert en plus du diagnostic histologique, le statut mutations, reconnu comme facteur prédictif des nouvelles thérapies ciblées. Le NGS est devenu un outil incontournable dans cette approche thérapeutique. Actuellement limité aux cancers avancés, il est certain qu'il sera appelé à élargir son application en oncologie. Dans le chapitre suivant nous allons développer la relation entre l'oncologie et méthodes computationnelles.

Chapitre 2 :

Méthodes computationnelles pour l'oncologie de précision

Introduction :

L'oncologie de précision est un nouveau domaine de recherche et une approche des soins contre le cancer qui tire le parti technologie de séquençage à haut débit et pipelines bio-informatiques pour déterminer le diagnostic, le pronostic et le traitement des patients de manière personnalisée. Basée sur les caractéristiques spécifiques de la tumeur individuelle plutôt que sur le type de cancer [10].

Une caractéristique clé de l'oncologie de précision est l'altération actionnable, c'est-à-dire une altération génétique ou moléculaire. Altération majeure qui peut être directement ciblée par un médicament ou un biomarqueur indiquant une sensibilité à un médicament spécifique.

Le développement des technologies de séquençage à haut débit abordables au cours de la dernière décennie a alimenté les progrès de l'oncologie de précision, aujourd'hui prend de l'ampleur de plus en plus, intégrés dans les pratiques cliniques courantes.

Les deux principaux éléments permettant la précision de l'oncologie est la technologie qui produit les données et les pipelines de calcul pour analyser les données. Le séquençage de l'ADN permet de détecter les variations pathologiques du génome et des transcriptases des cellules cancéreuses [11].

1-Pipeline bio-informatique à partir de données NGS :

Le but de la bio-informatique est de donner un sens à des données générées par le séquençage de l'ADN. Les bio-informaticiens travaillent sur les données génomiques pour identifier les gènes, générer des hypothèses sur leurs fonctions et les appliquer en médecine et

dans d'autres disciplines. Dans cette partie, nous allons parler du fonctionnement des pipelines bio-informatiques et comment ils sont utilisés.

Un pipeline bio-informatique correspond à une chaîne de traitement informatisée de données biologiques (séquences, variants). Dans cette application, le pipeline bio-informatique doit permettre l'identification de variation de séquence d'un échantillon par rapport à une séquence de référence.

Le pipeline bio-informatique doit permettre, à partir des données primaires, la détection de la présence d'une variation (ou « variant ») qualitative de séquence nucléotidique par rapport à une séquence de référence, puis l'identification (position par rapport à la séquence de référence) et la qualification de cette variation. Ces données sont considérées comme les données secondaires (données II, analyses additionnelles obtenues à partir des séquences).

Les différentes étapes de cette analyse sont conduites dans des pipelines complexes qui diffèrent selon la méthode de séquençage utilisée. Même pour le même type de méthode de séquençage, de nombreux pipelines sont disponibles et il a été observé à plusieurs reprises que les résultats peuvent être différents [10]. L'analyse primaire peut être subdivisée en

- Traitement de fichier de séquençage brut.
- Mappage de lecture.
- Post-traitement d'alignement
- Appel de variantes.

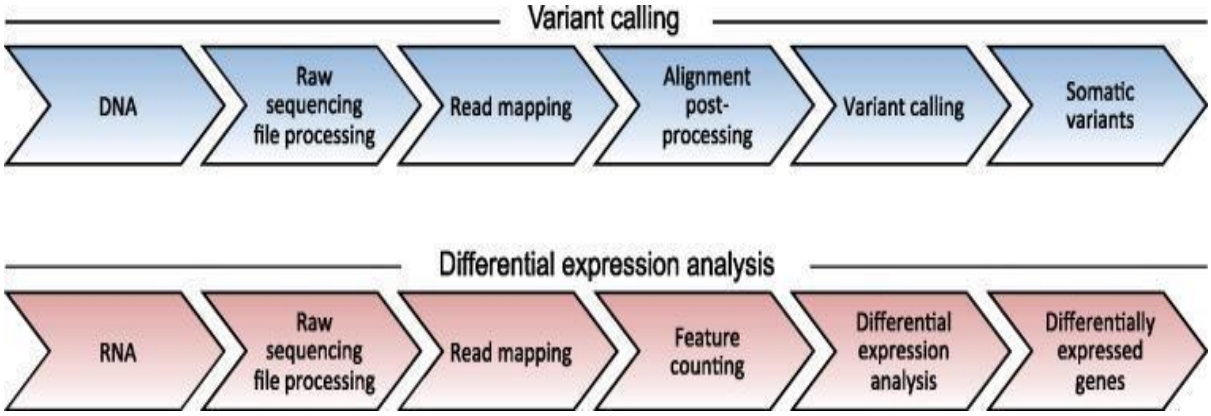

Figure 3 : Aperçu schématique des étapes d'analyse pour l'appel de variantes d'ADN et l'analyse de l'expression d'ARN.

Les variantes germinales sont des changements de nucléotides dans les germes ou les ovules et peuvent être transmises à un enfant des parents lors de la conception. Comme les variants se trouvent dans les cellules reproductrices, ils sont héréditaires mutations et peuvent être transmises aux générations futures. Les mutations germinales représentent environ 5 à 10 % des cancers.

Les variantes somatiques présentent un intérêt particulier parce qu'elles sont associées à diverses maladies humaines, y compris les cancers (Figure 4).

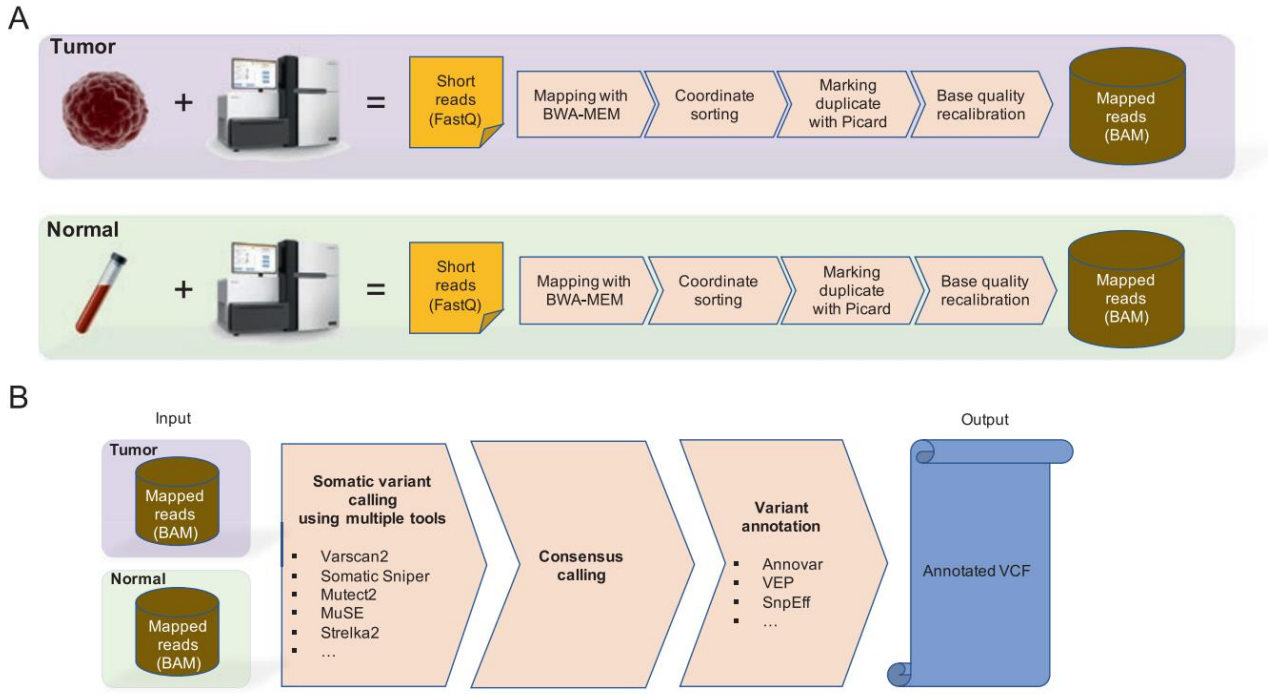

Figure 4 : Le flux de travail de l'appel de variante somatique d'échantillons appariés tumeur-normale. (a) Prétraitement des données, étapes de la préparation des échantillons à la cartographie des lectures courtes et étalonnage en version binaire de Sequence Alignment/ Fichiers Map (BAM) pour les échantillons tumoraux et normaux appariés. (b) Étapes d'appel et d'annotation de variantes à partir de paires fichiers BAM tumoraux normaux aux variantes somatiques annotées dans Format VCF [12].

2- Les logiciels pour l'oncologie de précision :

Un Workflow est un outil permettant d'exécuter un ensemble de processus de façon automatique. Ces « pipelines » sont très présents en bio-informatique (à défaut d'être très utilisés) car ils permettent aux chercheurs en biologie d'analyser leurs données (issues de séquences, génotypages) de façon relativement transparente et (quasiment) sans l'aide d'informaticiens (denrées rares dans la recherche).

L'oncologie de précision repose principalement sur la génétique et le profilage moléculaire des patients à partir de données de séquençage à haut débit. La nécessité de traiter et analyser de grands volumes de données a conduit au développement d'outils et de méthodes informatiques robustes. L'aspect le plus difficile dans la mise en œuvre d'une précision le flux de travail en oncologie implique une manipulation appropriée d'un grand volume de données, tout en assurant les résultats sont reproductibles et reproductibles [12]. À partir de là, nous abordons ce flux de travail pour les tumeurs (Tableau 2). L'oncologie de précision est un domaine de recherche innovante qui a introduit une nouvelle approche des traitements contre le cancer, où le diagnostic, le pronostic et la thérapie sont informés par des facteurs génétiques et moléculaires du profilage du patient individuel, plutôt qu'étant basé sur une approche unique [13].

Nom	**Description**	**Site web**
Nextflow	Langage spécifique au domaine	http://nextfow.io
Toil	Système de gestion de pipeline	https://toil.ucsc-cgl.org
Snakemake	Langage spécifique au domaine	https://snakemake.github.io
Bpipe	Langage spécifique au domaine	http://docs.bpipe.org
WDL	Langage de spécification de workflow	https://openwdl.org/
CWL	Langage de spécification de workflow	https://www.commonwl.org/

Tableau 2 : systèmes de gestion de flux de travail.

NextFlow[1] est un système de flux de travail populaire développé par Seqera Labs à Barcelone, en Espagne, conçu pour traiter l'instabilité numérique, l'exécution parallèle efficace, la tolérance aux erreurs, la provenance de l'exécution et la traçabilité. Semblable à CWL, ce langage spécifique au domaine (DSL) utilise des conteneurs logiciels pour créer des workflows reproductibles, permettant un pipeline rapide, développement par l'adaptation de l'existant pipeline écrit dans n'importe quel langage de script [7].

3-Identification des variants :

La variation du nombre de copies (CNV), qui est la suppression et la multiplication de segments d'un génome, est une altération génomique importante qui a été associée à de nombreuses maladies, y compris le cancer. Dans le cancer, les NVC sont majoritairement des aberrations somatiques qui surviennent au cours l'évolution du cancer. Les avancées des technologies de séquençage et l'arrivée du séquençage de nouvelle génération données (séquençage du génome entier et séquençage de l'exome entier ou séquençage ciblé) ont ouvert une opportunité de détecter les CNV avec une précision et une résolution supérieures. De divers des méthodes de calcul ont été développées pour la détection somatique des CNV, qui est une tâche difficile en raison de la complexité du cancer donnée de séquençage, niveau élevé de bruit et de biais dans le processus de séquençage, et la nature du big data de données de séquençage [14]. Néanmoins, la détection informatique de CNV dans les données de séquençage a abouti à la découverte d'action CNV spécifiques au cancer à utiliser pour guider les thérapies anticancéreuses, contribuant à progrès en oncologie de précision (Figure 5).

Variation du nombre de copies est une forme de variation structurelle d'une séquence d'ADN qui comprend l'amplification et délétion d'un segment particulier d'ADN. Il présente une mutation supérieure tôt que les polymorphismes mononucléotidiques (SNP) et affecte un plus grand fragment de génomes [15].

[1] https://nf-co.re/usage/installation

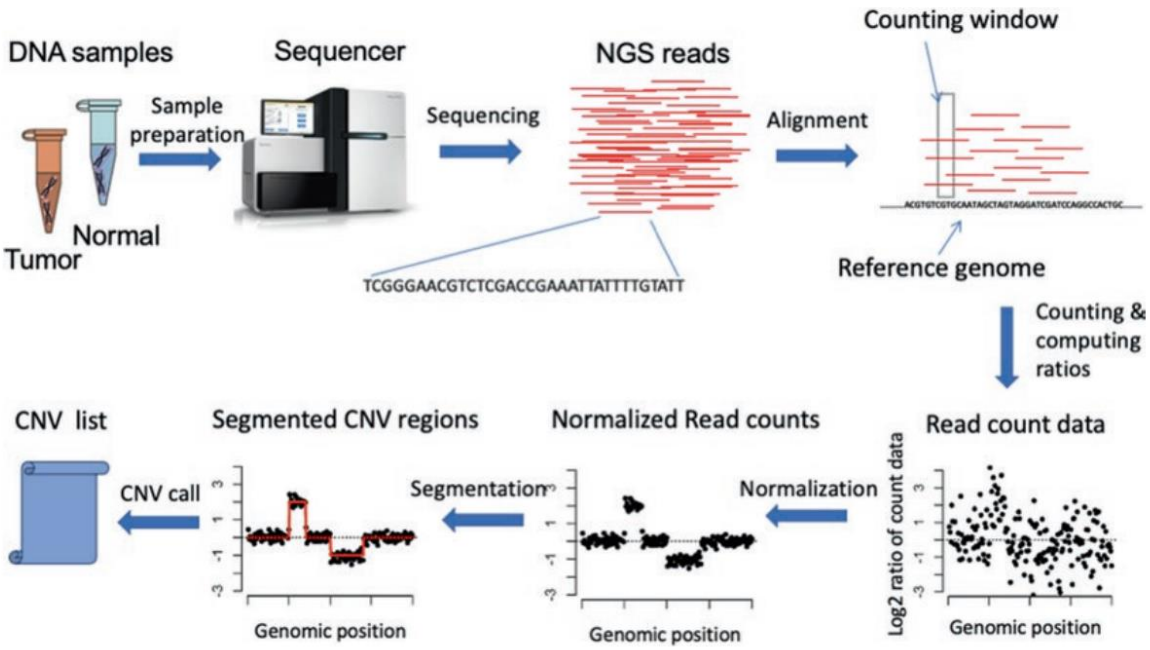

Figure 5 : Pipeline global pour détecter les CNV à l'aide des données de comptage de lecture

Les CNV comme variation génomique majeure et les biomarqueurs exploitables éprouvés jouent un rôle important dans l'oncologie de précision. Avec les technologies NGS émergentes et la diminution du coût de génération des données NGS, des analyses génomiques complètes sont devenues de plus en plus disponibles en raison des progrès du développement d'outils bio-informatiques accessibles et applicables pour la détection de variants génomiques et moléculaires à partir des données NGS [14].

Des études montrent que de nombreuses tumeurs spécifiques, variations moléculaires dans les gènes responsables du cancer (y compris les SNP, les CNV, les translocations et les gènes fusions) sont des biomarqueurs prédictifs éprouvés de réponse aux thérapies ciblées sélectives [10].

Conclusion :

L'oncologie de précision est un domaine en évolution rapide qui permet une traduction rapide de la recherche biomédicale découverte dans les soins cliniques contre le cancer. L'utilisation des mutations d'ADN est capable d'agir pour déterminer le pronostic et la thérapie elle est devenue une composante de divers essais cliniques et de la prise de décision clinique de routine. Il est également raisonnable de s'attendre à ce que les extensions futures des systèmes d'oncologie de précision incluent d'autres technologies de séquençage, telles que le

séquençage d'ADN, qui pourraient aider à mieux disséquer la complexité de l'hétérogénéité antitumorale et du microenvironnement tumoral, et d'autres types d'omiques [16].

Utilisation d'un langage approprié et spécifique au domaine pour le développement et l'exécution du flux de travail est une nécessité. L'ingénieur expert en bio-informatique nécessitera la combinaison de plusieurs outils et cette combinaison devra être sans couture. CWL, WDL, Snakemake et NextFlow offrent tous la portabilité et la flexibilité nécessaires pour les workflows d'oncologie de précision.

Chapitre 3 :

Matériels et méthodes

1- Matériels :

1-1- Les données :

Souvent, la première étape d'un flux de travail bio-informatique consiste à transférer les données avec lesquelles on souhaite travailler vers un ordinateur où on peut les utiliser. Nous utiliserons NCBI pour télécharger les données en tant que SRA. Aujourd'hui, nous allons travailler avec des données de séquences accessibles au public.

Nous étudions le cancer du sein qui commence dans les cellules du sein. Une tumeur cancéreuse (maligne) est un groupe de cellules cancéreuses qui peuvent envahir et détruire les tissus voisins. Il peut également se propager à d'autres parties du corps. Nous allons travailler avec deux échantillons dans cette expérience.

Séquence	SRR11805419	SRR11805381
Nombre de spots	964.6 K	9.5 K
Nombre de bases brutes	139.6 Mbp	1.46 Gbp
Taille	92 M	923.4 M
Contenu GC	42.3%	41.9%
Publié	2021-05-01	2021-05-01
Plateforme	Ion Torrent	Ion Torrent
ID	10869755	10869793
Nom	Normal	Tumeur

Tableau 3 : Données utilisées pour l'analyse workflow.

La deuxième donnée est GRCh38/hg38 (Genome Reference Consortium Human Build 38) fournit séquences alternées (« alt_sequences ») pour certaines régions génomiques pour lesquelles leur variabilité empêche représentation adéquate par une seule référence.

Ids	88331 [UID] 883148 [GenBank] 884148 [RefSeq]
Longueur totale de la séquence	3,099,734,149
Nombre total de chromosomes et de plasmides	24
Nombre de séquences de composants (WGS ou clone)	35,614

Tableau 4 : Données utilisées pour l'alignement et l'annotation des variantes

1-2-Le séquenceur :

Au cours de la préparation de notre mémoire, on a eu la chance d'être présent durant le séquençage pour la première fois en Algérie, exactement au CHU de Sétif (CHU Saadna Abdenour). Le séquençage de l'hôpital de Sétif est en relation avec notre thème de mémoire, il se passe précisément au centre de diagnostic avec l'aide des biologistes de laboratoire, des médecins de spécialité et aussi une doctorante en bio-informatique.

La visite du chu Sétif était pour voir le séquençage pour la première fois, mais plus précisément pour récupérer des données réelles du séquençage et avoir l'opportunité de les utiliser comme données pour notre pipeline. Mais avant de voir les résultats, on va parler un peu de la machine ou le séquenceur ION TORENT qui est utilisée au niveau de l'hôpital de Sétif.

Alimentée par des puces à semi-conducteurs, la technologie de séquençage de nouvelle génération d'Ion Torrent nous aide à mettre en œuvre un flux de travail rapide et simple qui répond à nos besoins de recherche dans de multiples applications, y compris les maladies héréditaires, l'oncologie, les maladies infectieuses, la génomique reproductive, l'identification humaine, l'agrogénomique et plus encore.

Le séquenceur utilisait par le chu de Sétif, c'est l'ion s5 sequencing. Les biologistes du laboratoire sont encore en formation continue sur les étapes du NGS et ils sont toujours dans la découverte de la machine. La plus importante étape c'est la préparation de l'ADN et aussi la préparation de la puce qui est le plus important gadget dans cette machine de l'ion s5

sequencing. Puce avec code à barres pour le suivi et le séquençage des échantillons avec les systèmes de séquençage Ion S5 et Ion S5 XL. La puce Ion 510 détecte électroniquement l'incorporation de base entraînée par la polymérase sans utilisation de fluorescence (Figure 6). En éliminant l'utilisation d'un système de détection optique, cette avancée dans la technologie de séquençage de nouvelle génération permet des temps de séquençage rapides en aussi peu que 2,5 heures pour le séquençage de 200 bp et 4 heures pour le séquençage de 400 bp.

Figure 6 : Puce utilisé dans le séquenceur ion s5 torrent.

On a remarqué que le NGS ion s5 sequencing contient plusieurs appareils, chaque appareil est une étape importante pour avoir un résultat sans faute. **Les appareils NGS ion s5 sequencing :**

Figure 7 : L'instrument Ion Torrent.

Figure 8 : ion s5 sequencing 1- Écran tactile, 2- Bouton d'alimentation, 3- cartouche de kit de réactifs de séquençage Ion S5™, 4- colliers de serrage Bouteille de solution de lavage ,5- Ion S5™. Réservoir de déchets situé derrière la bouteille de solution de lavage (à droite), 6- bouteilles de solution de nettoyage S5™ 7- Réservoir de déchets.

1-2-Les outils :

Notre flux de travail offre une approche complète pour le contrôle qualité des expériences NGS tumorales-normales et se compose de quatre outils distincts qui peuvent être utilisés indépendamment : ReadQC, MappingQC, VariantQC et SomaticQC.

Pour en savoir plus sur les outils que nous utilisons, nous devons connaître tous leurs avantages et inconvénients (Tableau 5).

- **SRA Tools :**

SRA (Sequence Read Archive) est un format défini par NCBI pour les données NGS. Chaque donnée soumise à NCBI doit être en format SRA. SRA Toolkit fournit des outils pour télécharger des données, convertir différents formats de données en format SRA, extraire des données SRA dans d'autres formats différents.

- **BWA :**

BWA est un progiciel permettant de cartographier des séquences à faible divergence par rapport à un grand génome de référence, tel que le génome humain. Il se compose de trois algorithmes : BWA-backtrack, BWA-SW et BWA-MEM.

BWA-MEM et BWA-SW partagent des fonctionnalités similaires telles que la prise en charge de la lecture longue et l'alignement fractionné, mais BWA-MEM, qui est le plus récent, est généralement recommandé pour les requêtes de haute qualité, car il est plus rapide et plus précis.

- **SAM Tools :**

La carte d'alignement des séquences (SAM) est un format texte à l'origine pour le stockage de séquences biologiques alignées sur une séquence de référence développée par Heng Li et Bob Handsaker et al. Il a été élaboré lorsque le Projet 1000 génomes a voulu s'éloigner du format du mappeur MAQ et a décidé de concevoir un nouveau format [3].

- **BAM :**

BAM est la représentation binaire compressée de SAM (Sequence Alignment Map), une représentation compacte et indexable des alignements de séquence nucléotidique. Le but de l'indexation est de récupérer rapidement les alignements qui chevauchent un emplacement précis sans avoir à les passer tous en revue [5].

- **GATK :**

La Genome Analysis Toolkit (GATK) est un ensemble d'outils bio-informatiques pour l'analyse des données de séquençage à haut débit (HTS) et de format d'appel variant (VCF). La boîte à outils est bien établie pour la découverte de la variante courte germinale à partir de données de séquençage du génome entier et de l'exome [7].

- **IGV :**

L'Intégrative Genomics Viewer (IGV) est un outil interactif à haute performance et facile à utiliser pour l'exploration visuelle des données génomiques. Il favorise l'intégration souple de tous les types communs de données et de métadonnées génomiques, générées par les chercheurs ou accessibles au public, chargées à partir de sources locales ou infonuagiques.

Algorithmes	les usages	Avantages	Inconvénients	lien hypertexte
BWA	Alignement de l'ADN	Alt-conscient et plus précis pour séquences avec variation.	/	http://bio-bwa.sourceforge.net/
samtools	Variante de la lignée germinale Appel	Sensibilité la plus élevée pour les SNV, largement utilisé, exécution lente, plus grande précision pour les indels.	Faible sensibilité pour les indels.	http://samtools.sourceforge.net/
picard	Marquage ou suppression doublons	Largement utilisés, les doublons peuvent être marqués ou supprimés.	/	https://broadinstitute.github.io/picard/
GATK	Variante de la lignée germinale appel	Sensibilité la plus élevée pour les indels, largement utilisé pour la lignée germinale variante d'appel.	Nécessite de nombreuses étapes pour une précision appelant, avec de nombreux compagnons programmes de filtrage amélioration de la précision, lenteur Durée.	https://software.broadinstitute.org/gatk/
Mutect2	Variante somatique appel	Sensibilité la plus élevée pour les SNV, largement utilisé, exécution lente.	Nécessite de nombreuses étapes pour une précision appelant avec de nombreux compagnons programmes de filtrage amélioration de la précision, lenteur Durée.	/

Tableau 5 : Les avantages et les inconvénients des algorithmes bio-informatiques [7].

2- Méthodes :

2-1-Description du pipeline:

Un pipeline typique de séquençage du cancer comprend les étapes suivantes ; prétraitement, contrôle de la qualité et découpage, mappage et appel de variantes [10]. Dans cette étude, nous avons créé un ensemble de pipelines pour lesquels des données séquencées brutes au format FASTQ ont été utilisées comme entrée. La sortie de chaque pipeline était une variante contenant le fichier VCF. Chaque workflow utilise différents algorithmes pour les étapes de mappage et d'appel de variantes. Toutes les autres étapes et outils dans les pipelines étaient identiques, à savoir le contrôle de la qualité des lectures via FastQC.

Nous avons développé un flux de travail modulaire pour le traitement NGS. Tout d'abord, les données de lecture brutes avec des estimations d'erreur de base bien calibrées au format fastq sont mappées au génome de référence. L'application de cartographie BWA est utilisée pour mapper les lectures à la référence du génome humain et générer un format de fichier de référence SAM/BAM indépendant de la technologie. Ensuite, les fragments en double sont marqués et éliminés avec Picard, la qualité du mappage est évaluée et les lectures mappées de faible qualité sont filtrées, et les informations de lecture appariées sont évaluées pour s'assurer que tous les partenaires -les informations de pair sont synchronisées entre chaque lecture [16]. Nous affinons ensuite les alignements initiaux par réalignement local et identifions les régions suspectes. Le recalibrage du score de qualité de base GATK est effectué. Enfin, l'appel de variants est effectué à l'aide des fichiers BAM recalibrés et réalignés [12].

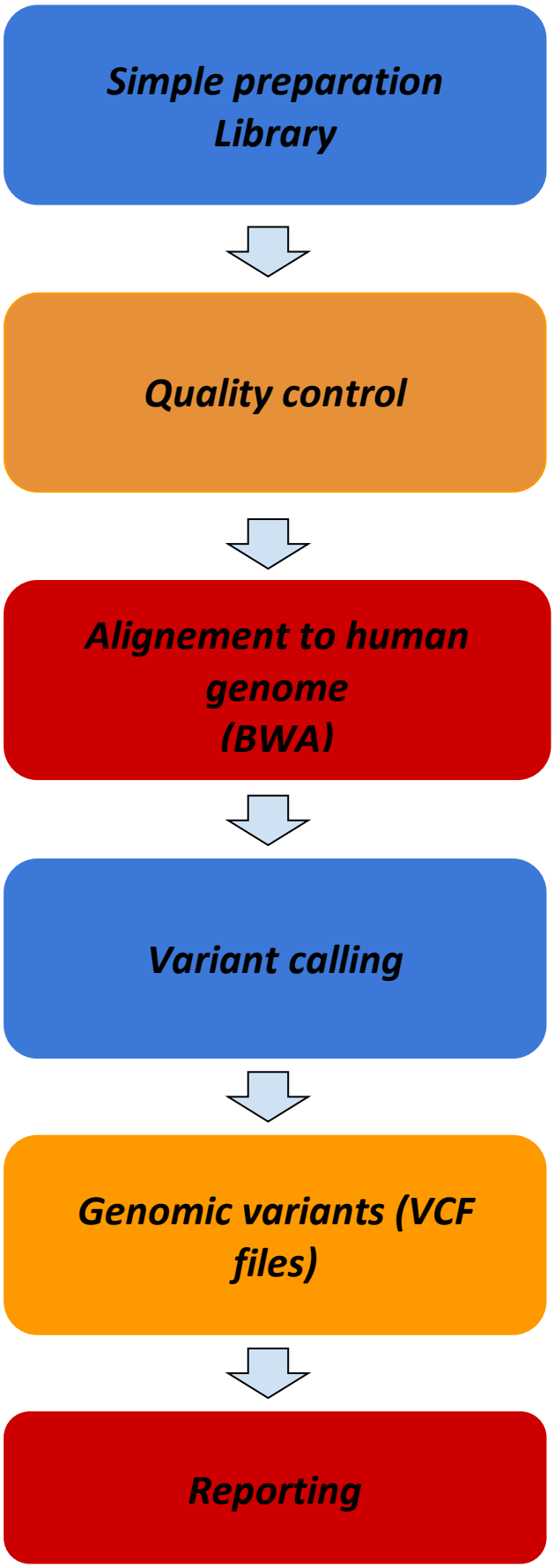

Figure 9 : un schéma typique de l'oncologie de précision.

2-1-1- Préparation (prétraitement) :

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 [17].

Préparation de la bibliothèque et le séquençage peut, en effet, introduire des biais, les erreurs et la contamination, qui, à leur tour, peuvent affecter l'identification des variantes et conduire à des résultats.

2-1-2- Contrôle de la qualité :

Le contrôle qualité et le prétraitement des données sont des étapes cruciales et peuvent aider à identifier et à atténuer les problèmes dans les analyses en aval. Plusieurs outils de contrôle de la qualité ont été développés dans le passé ans, 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 [18].

FastQC est un outil populaire et option CQ largement utilisée. Dans certains cas, lit doit être coupé pour supprimer les séquences d'adaptateur et des bases de mauvaise qualité.

D'autres outils peuvent, en outre, être utilisé pour estimer la pureté de la tumeur, qui est la proportion de cellules cancéreuses dans l'échantillon et, par conséquent, produit une estimation de contamination possible avec des cellules normales. En fait, peuvent se produire à différentes étapes du pipeline expérimental et d'analyse de données [17].

2-1-3- Alignement de séquence :

Une fois le CQ effectué et les lectures retraités et filtrés, l'étape suivante consiste à cartographier chaque paire de lecture ou de lecture individuelle au génome de référence (par exemple, humain GRCh38/hg38), qui est une séquence représentative du génome commun de l'espèce analysée sous-forme de chaîne, pour correctement identifier leurs origines.et les étapes de cette section et des sections suivantes peuvent être appliquées à tout type de données de séquençage d'ADN.

Nous effectuons un alignement ou une cartographie des lectures pour déterminer d'où proviennent nos lectures dans le génome. Certains outils sont mieux adaptés à des analyses NGS particulières. Nous utiliserons le Burrows Wheeler Aligner (BWA), qui est un progiciel permettant de cartographier des séquences à faible divergence par rapport à un grand génome de référence [18].

Le processus d'alignement comprend deux étapes :

- Indexation du génome de référence
- Alignement des lectures sur le génome de référence

Indexer le génome de référence

Notre première étape consiste à indexer le génome de référence à utiliser par BWA. L'indexation permet à l'aligneur de trouver rapidement des sites d'alignement potentiels pour les séquences de requête dans un génome, ce qui permet de gagner du temps lors de l'alignement. L'indexation de la référence ne doit être exécutée qu'une seule fois. La seule raison pour laquelle vous voudriez créer un nouvel index est si vous travaillez avec un génome de référence différent ou si vous utilisez un outil différent pour l'alignement [19].

Aligner les lectures sur le génome de référence

Le processus d'alignement consiste à choisir un génome de référence approprié pour cartographier nos lectures, puis à choisir un aligneur. Nous utiliserons l'algorithme BWA-MEM, qui est le plus récent et qui est généralement recommandé pour les requêtes de haute qualité, car il est plus rapide et plus précis [19].

E/S	Entité	Format
Entrer	Lectures non alignées soumises ou lectures alignées soumises.	FastQ ou BAM
Sortie	Lectures alignées.	SAM ou BAM

Tableau 6 : Les entrées et les sorties de l'alignement.

Les alignements sont habituellement sortis en le format textuel SAM (Sequence Alignment Map) ou sa version compressée appelée BAM (Carte d'alignement binaire) et les alignements de lecture sont ensuite traités pour atténuer les biais introduits par la préparation de la bibliothèque étapes, telles que l'amplification par PCR, et recalibrer le score de qualité de

base. La version binaire compressée de SAM est appelée un fichier BAM. Nous utilisons cette version pour réduire la taille et permettre l'indexation, ce qui permet un accès aléatoire efficace aux données contenues dans le fichier [19].

2-1-4- Appel de variants :

Une fois les fichiers BAM filtrés et nettoyés, ils sont prêts à être traités pour la découverte de variantes. C'est l'étape fondamentale où les lectures d'une tumeur et d'un échantillon normal correspondant du même individu sont comparées au génome de référence pour identifier les variations somatiques d'un seul nucléotide (SNV) et indels courts, c'est-à-dire une insertion ou une courte suppression de bases, qui sont ensuite généralement sorties dans VCF (Variant Call Format) fichiers texte. Le GATK Mutect2 est un flux de travail d'appelant de variante somatique, il peut être utilisé pour détecter les SNV et les indels dans un ou plusieurs échantillons de tumeur d'un même individu, avec ou sans échantillon normal correspondant [7].

Et les variantes somatiques se trouvent dans la tumeur, mais ni dans la normale contrôle, ni dans les génomes de référence. Ainsi, ils sont plus susceptibles d'avoir un impact sur le processus oncogénique.

Ensuite, GATK HaplotypeCaller est utilisé pour identifier les variants germinales dans l'échantillon, qui est sortis dans un fichier GVCF (Genomic VCF). Plusieurs Les GVCF de différents patients peuvent ensuite être consolidés et traités par GenotypeGVCF, qui effectue l'appel de variants conjoints de cohorte.

E/S	Entité	Format
Entrer	Lectures alignées	BAM
Sortie	Mutation somatique simple brute	VCF

Tableau 7 : Les entrées et les sorties de l’appel des variants.

2-1-5-Le filtrage des variants VCF:

Cette étape est fortement recommandée, car elle améliore la précision et la sensibilité de la détection des variantes, en particulier sur les sites à faible couverture ou de faible qualité, Le fichier VCF brut résultant est ensuite filtré à l'aide de différents outils dans GATK, tels que

VariantRecalibrator, pour supprimer les variants susceptibles d'être des faux positifs. Ces outils utilisent l'apprentissage automatique algorithmes qui sont entraînés sur de grands ensembles de données de variantes connues, puis appliquées à identifier les variantes susceptibles d'être réelles dans l'échantillon cible. Le VCF filtré est alors prêt pour annotation, évaluation et analyses en aval. Freebayes[2], VarScan[3] et DeepVariant[4] sont d'autres outils populaires pour la découverte de variantes germinales [3].

2-1-6- Visualisation et tabulation des données de la séquence de nouvelle génération IGV :

IGV est un navigateur autonome, qui a l'avantage d'être installé localement et d'offrir un accès rapide. Les navigateurs de génome, basés sur le Web, sont plus lents, mais offrent plus de fonctionnalités. Ils permettent non seulement une visualisation plus soignée et flexible, mais offrent également un accès facile à une multitude d'annotations et de sources de données externes. Cela facilite la mise en relation de nos données avec des informations sur les régions répétées, les gènes connus, les caractéristiques épigénétiques ou les zones de conservation inter-espèces, pour n'en nommer que quelques-uns [19].

Dans la piste VCF, chaque barre en haut du tracé indique la fraction d'allèle pour un seul locus. La deuxième barre montre les génotypes pour chaque locus dans chaque échantillon.

2-2- Automatisation d'un workflow d'appel de variante :

Le flux de travail d'appel de variants que nous venons d'effectuer comporte environ dix étapes où nous devions taper une commande dans notre terminal. La plupart de ces commandes sont assez longues. Si nous voulions faire cela pour tous nos fichiers de données. Et si nous avions 20 échantillons, ce serait 200 étapes !

Dans cette partie, nous aborderons l'automatisation. Nous préparons un script shell qui est un programme informatique qui présente une interface de ligne de commande qui nous permet de contrôler notre ordinateur à l'aide de commandes saisies avec un clavier au lieu de contrôler des interfaces utilisateur graphiques (GUI) avec une combinaison souris/clavier.

[2] https://github.com/freebayes/freebayes
[3] http://varscan.sourceforge.net/
[4] https://github.com/google/deepvariant

Il existe de nombreuses raisons de choisir le shell[5] et d'en savoir plus à son sujet :

- De nombreux outils bio-informatiques ne peuvent être utilisés que via une interface de ligne de commande ou disposent de fonctionnalités supplémentaires dans la version de ligne de commande qui ne sont pas disponibles dans l'interface graphique. C'est le cas par exemple de BLAST qui propose de multiples fonctions avancées accessibles uniquement aux utilisateurs sachant utiliser un shell.
- La coque rend votre travail moins ennuyeux. En bio-informatique, on doit souvent effectuer le même ensemble de tâches avec un grand nombre de fichiers. Apprendre le shell nous permettra d'automatiser ces tâches répétitives et nous laissera libre de faire des choses plus excitantes.
- De diverses tâches bio-informatiques nécessitent de grandes quantités de puissance de calcul et ne peuvent pas être exécutées de manière réaliste sur votre propre machine.

Notre workflow d'appel de variantes comporte les étapes suivantes :

- Indexer le génome de référence à utiliser par bwa et samtools.
- Aligner les lectures sur le génome de référence.
- Convertissez le format de l'alignement en BAM trié, avec quelques étapes intermédiaires.
- Détecter les variantes (SNPs , SVs , CNVs …).
- Filtrez et signalez les variantes dans VCF (format d'appel de variante).

Nous prenons toutes les commandes individuelles que nous avons écrites auparavant, les mettre dans un seul fichier, ajouter des variables afin que le script sache parcourir nos fichiers d'entrée et écrire dans les fichiers de sortie appropriés.

[5] https://github.com/cfarkas/Genotype-variants

Chapitre 4 :

Résultats et discussion

1- Résultats :

Dans cette partie, nous présentons les résultats que nous avons obtenus à partir des données récupérées du centre de Sétif (Tableau 8) qui représente le nombre de chromosomes que les bases mappées contiennent. Et aussi parler des données de résultats sur lesquelles nous avons travaillé (Tableau 3).

Name	Length	Mapped bases	Mean coverage	Standard deviation
chr1	249250621	0	0	0
chr2	243199373	53	0	0.0005
chr3	198022430	0	0	0
chr4	191154276	117	0	0.0008
chr5	180915260	0	0	0
chr6	171115067	0	0	0
chr7	159138663	142	0	0.0009
chr8	146364022	0	0	0
chr9	141213431	0	0	0
chr10	135534747	0	0	0
chr11	135006516	0	0	0
chr12	133851895	38	0	0.0005
chr13	115169878	0	0	0
chr14	107349540	0	0	0
chr15	102531392	0	0	0
chr16	90354753	0	0	0

Tableau 8 : Les statistiques sur les chromosomes qui récupérées du centre de Sétif.

FastQC possède un certain nombre de fonctionnalités qui peuvent nous donner une idée rapide des problèmes que nos données peuvent rencontrer, afin que nous puissions prendre ces problèmes en considération avant de poursuivre nos analyses. Plutôt que d'examiner les scores de qualité pour chaque lecture individuelle, FastQC examine la qualité collectivement pour toutes les lectures d'un échantillon. Les graphiques suivants représenteront tous les résultats du workflow :

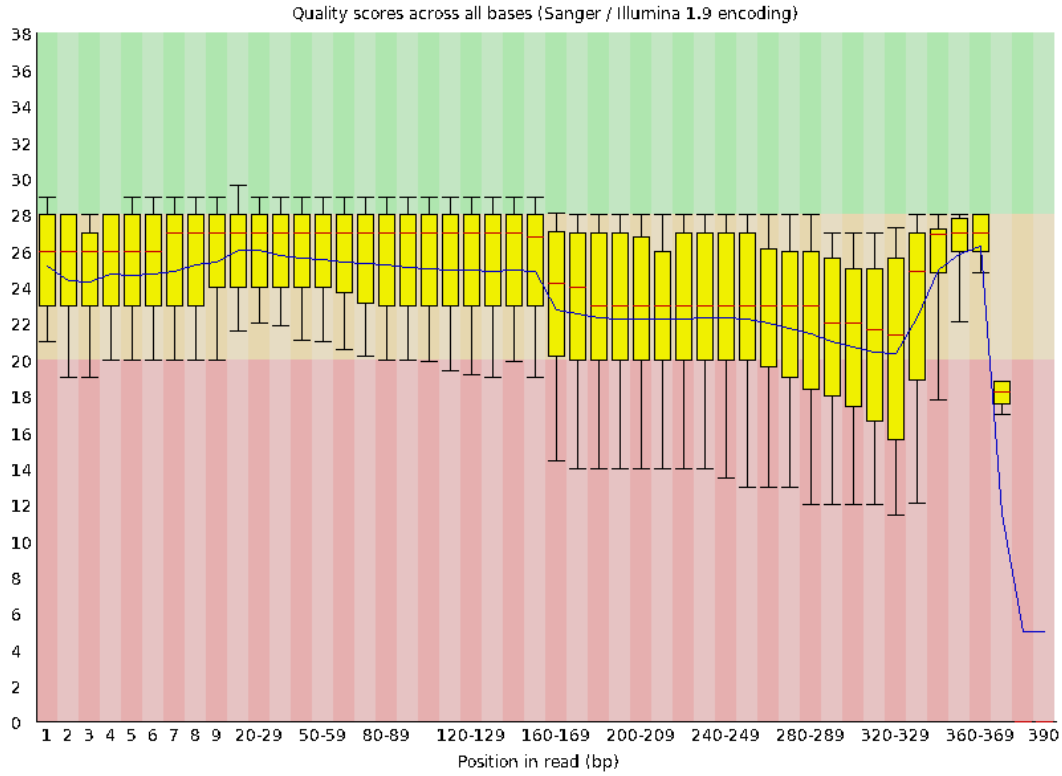

Figure 10 : Un graphe généré par FastQC qui indique un échantillon tumeur.

L'axe des x affiche la position de base dans la lecture et l'axe des y montre les scores de qualité. Pour chaque poste de cet échantillon, les valeurs de qualité ne descendent pas beaucoup plus bas que 32. Il s'agit d'un score de qualité élevé. L'arrière-plan du tracé est également codé par couleur pour identifier les scores de qualité bons (vert), acceptables (jaune) et mauvais (rouge).

Ici, nous voyons des positions au sein de la lecture dans lesquelles les cases couvrent une gamme beaucoup plus large. De plus, les scores de qualité chutent assez bas dans la plage « mauvais », en particulier à la fin des lectures. L'outil FastQC produit plusieurs autres tracés de diagnostic pour évaluer la qualité de l'échantillon [12].

Reference size	3,209,286,105
Number of reads	9,986,732
Mapped reads	9,971,175 / 99.84%
Unmapped reads	15,557 / 0.16%
Mapped paired reads	0 / 0%
Secondary alignments	0
Supplementary alignments	496,656 / 4.97%
Read min/max/mean length	45 / 390 / 155.63
Duplicated reads (flagged)	9,315,785 / 93.28%
Clipped reads	1,810,028 / 18.12%

Tableau 9 : Informations globales sur l'échantillon.

Mean	0.4374
Standard Deviation	71.1699

Tableau 10 : Couverture du génome.

Mean Mapping Quality	14.71

Tableau 11 : Qualité de la cartographie.

General error rate	0.72%
Mismatches	5,205,447
Insertions	4,102,941
Mapped reads with at least one insertion	23.66%
Deletions	2,663,252
Mapped reads with at least one deletion	19.99%
Homopolymer indels	63.84%

Tableau 12 : Décalages et indels.

Name	Length	Mapped bases	Mean coverage	Standard deviation
chr1	248956422	9730136	0.0391	26.0307
chr10	133797422	32714484	0.2445	55.5519
chr11	135086622	142364158	1.0539	114.0008
chr11_KI270721v1_random	100316	0	0	0
chr12	133275309	40688070	0.3053	67.7471
chr13	114364328	18756129	0.164	39.9984
chr14	107043718	10083512	0.0942	24.4962
chr14_GL000009v2_random	201709	0	0	0
chr14_GL000225v1_random	211173	405	0.0019	0.0505
chr14_KI270722v1_random	194050	140	0.0007	0.0269
chr14_GL000194v1_random	191469	32	0.0002	0.0129
chr14_KI270723v1_random	38115	0	0	0
chr14_KI270724v1_random	39555	0	0	0
chr14_KI270725v1_random	172810	0	0	0
chr14_KI270726v1_random	43739	0	0	0
chr15	101991189	592700	0.0058	4.6538

Tableau 13 : Statistiques sur les chromosomes.

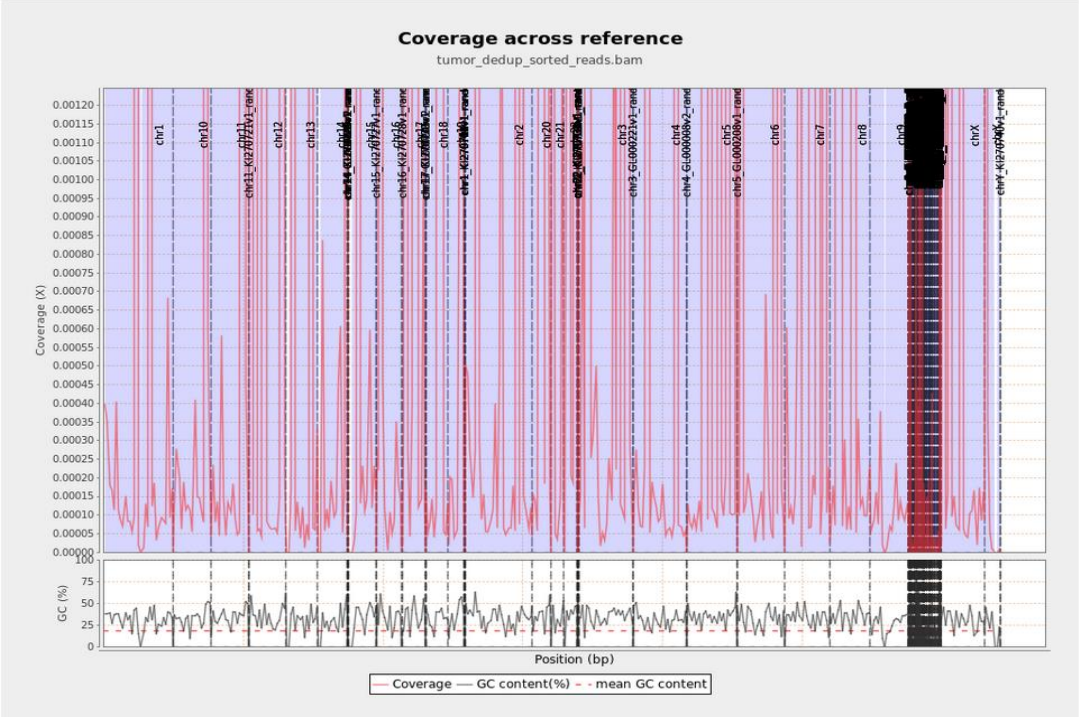

Figure 11 : Graphique représentant la couverture.

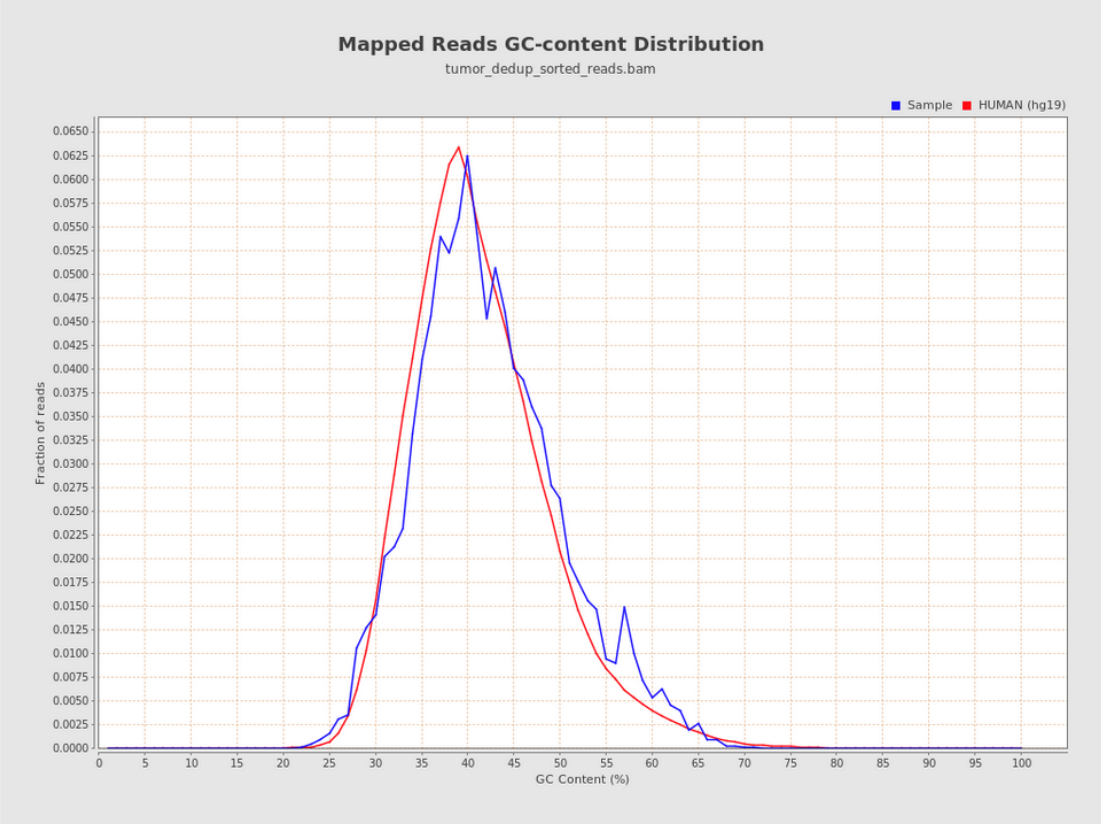

Figure 12 : Lectures mappées Distribution de contenu GC.

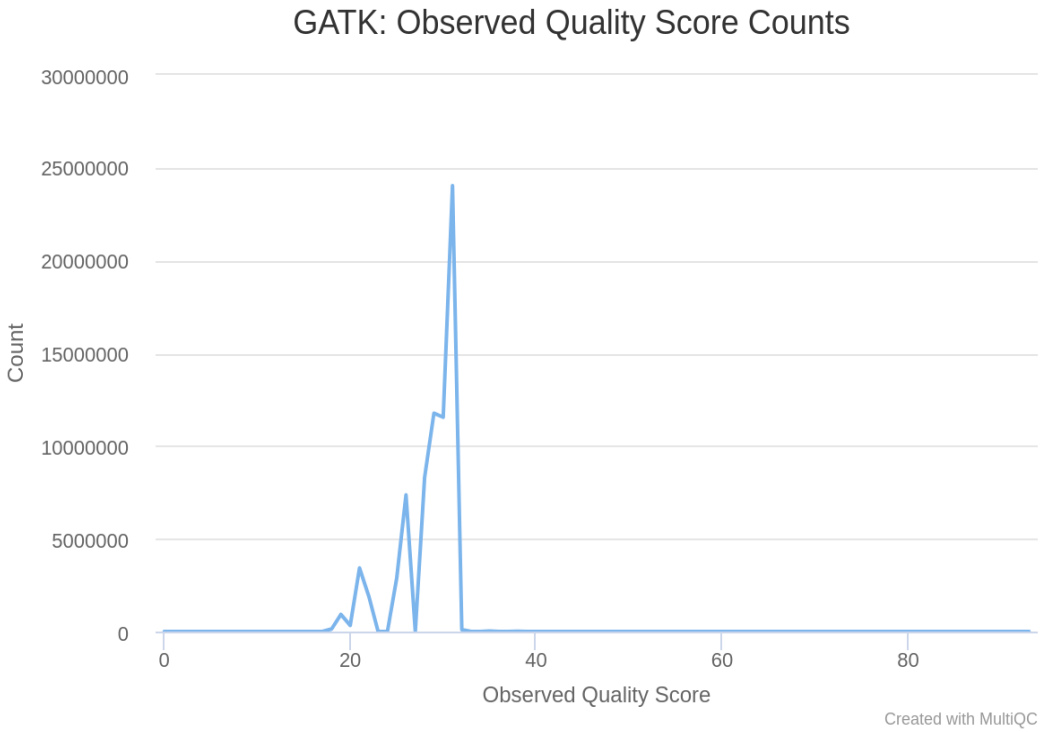

Figure 13 : Graphique représentant les scores de qualité observés.

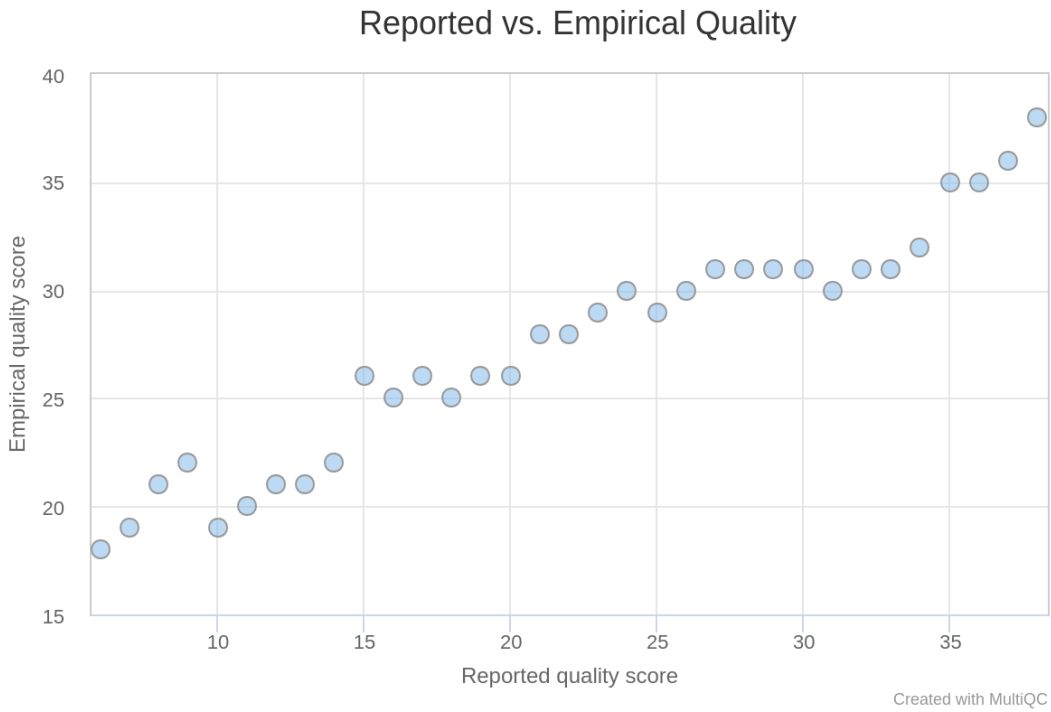

Figure 14 : Graphique représentant la qualité entre rapportée et empirique.

```
499 #CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_1
500 chr1    4127288         .       C       A       .       .       AS_SB_TABLE=0,0|9,1;DP=10;ECNT=1;MBQ=0,27;MFRL=0,0;MMQ=60,60;MPOS=136;POPAF=7.30;TLOD=30.12 GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,10:0.917:10:0,1:0,9:0,10:0,0,9,1
501 chr1    4127417         .       G       T       .       .       AS_SB_TABLE=0,0|7,4;DP=12;ECNT=1;MBQ=0,27;MFRL=0,0;MMQ=60,60;MPOS=10;POPAF=7.30;TLOD=30.32  GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,11:0.925:11:0,3:0,7:0,11:0,0,7,4
502 chr1    4677394         .       C       A       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=3;MBQ=0,28;MFRL=0,0;MMQ=60,60;MPOS=91;POPAF=7.30;TLOD=4.20
    GT:AD:AF:DP:F1R2:F2R1:FAD:PGT:PID:PS:SB       0|1:0,1:0.667:1:0,0:0,1:0,1:0|1:4677394_C_A:4677394:0,0,1,0
503 chr1    4677397         .       T       A       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=3;MBQ=0,28;MFRL=0,0;MMQ=60,60;MPOS=88;POPAF=7.30;TLOD=4.20
    GT:AD:AF:DP:F1R2:F2R1:FAD:PGT:PID:PS:SB       0|1:0,1:0.667:1:0,0:0,1:0,1:0|1:4677394_C_A:4677394:0,0,1,0
504 chr1    4677471         .       C       A       .       .       AS_SB_TABLE=0,0|0,0;DP=2;ECNT=3;MBQ=0,24;MFRL=0,0;MMQ=60,60;MPOS=39;POPAF=7.30;TLOD=5.17    GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,2:0.750:2:0,0:0,1:0,2:0,0,1,1
505 chr1    7054368         .       A       C       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,29;MFRL=0,0;MMQ=60,60;MPOS=50;POPAF=7.30;TLOD=3.08    GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,1:0.667:1:0,1:0,0:0,1:0,0,0,1
506 chr1    7576934         .       T       TA      .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,18;MFRL=0,0;MMQ=60,60;MPOS=27;POPAF=7.30;TLOD=3.73    GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,1:0.667:1:0,0:0,0:0,1:0,0,1,0
507 chr1    10622736        .       GA      G       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,21;MFRL=0,0;MMQ=60,60;MPOS=67;POPAF=7.30;RPA=4,3;RU=A;STR;TLOD=3.14
    GT:AD:AF:DP:F1R2:F2R1:FAD:SB  0/1:0,1:0.667:1:0,1:0,0:0,1:0,0,0,1
508 chr1    12384338        .       AG      A       .       .       AS_SB_TABLE=0,0|0,0;DP=2;ECNT=1;MBQ=0,24;MFRL=0,0;MMQ=60,60;MPOS=76;POPAF=7.30;RPA=2,1;RU=G;STR;TLOD=7.02
    GT:AD:AF:DP:F1R2:F2R1:FAD:SB  0/1:0,2:0.750:2:0,2:0,0:0,2:0,0,0,2
509 chr1    13548172        .       GC      G       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,22;MFRL=0,0;MMQ=60,60;MPOS=62;POPAF=7.30;RPA=2,1;RU=C;STR;TLOD=3.44
    GT:AD:AF:DP:F1R2:F2R1:FAD:SB  0/1:0,1:0.667:1:0,0:0,1:0,1:0,0,1,0
510 chr1    14941314        .       CT      C       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,23;MFRL=0,0;MMQ=60,60;MPOS=27;POPAF=7.30;RPA=4,3;RU=T;STR;TLOD=3.07
    GT:AD:AF:DP:F1R2:F2R1:FAD:SB  0/1:0,1:0.667:1:0,1:0,0:0,1:0,0,0,1
511 chr1    17002810        .       C       G       .       .       AS_SB_TABLE=0,0|0,0;DP=2;ECNT=1;MBQ=0,28;MFRL=0,0;MMQ=60,60;MPOS=43;POPAF=7.30;TLOD=6.08    GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,2:0.750:2:0,1:0,1:0,2:0,0,1,1
512 chr1    20303897        .       C       T       .       .       AS_SB_TABLE=0,0|2,2;DP=4;ECNT=1;MBQ=0,25;MFRL=0,0;MMQ=60,60;MPOS=39;POPAF=7.30;TLOD=10.80   GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,4:0.833:4:0,2:0,2:0,4:0,0,2,2
513 chr1    31477394        .       AG      A       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,13;MFRL=0,0;MMQ=60,60;MPOS=51;POPAF=7.30;RPA=3,2;RU=G;STR;TLOD=4.20
    GT:AD:AF:DP:F1R2:F2R1:FAD:SB  0/1:0,1:0.667:1:0,0:0,0:0,1:0,0,0,1
514 chr1    43966375        .       T       C       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=1;MBQ=0,29;MFRL=0,0;MMQ=60,60;MPOS=25;POPAF=7.30;TLOD=3.08    GT:AD:AF:DP:F1R2:F2R1:FAD:SB
    0/1:0,1:0.667:1:0,1:0,0:0,1:0,0,0,1
515 chr1    58048505        .       T       TC      .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=2;MBQ=0,24;MFRL=0,0;MMQ=60,60;MPOS=41;POPAF=7.30;RPA=2,3;RU=C;STR;TLOD=4.20
    GT:AD:AF:DP:F1R2:F2R1:FAD:PGT:PID:PS:SB       0|1:0,1:0.667:1:0,0:0,1:0,1:0|1:58048505_T_TC:58048505:0,0,1,0
516 chr1    58048508        .       T       C       .       .       AS_SB_TABLE=0,0|0,0;DP=1;ECNT=2;MBQ=0,28;MFRL=0,0;MMQ=60,60;MPOS=37;POPAF=7.30;TLOD=4.20
    GT:AD:AF:DP:F1R2:F2R1:FAD:PGT:PID:PS:SB       0|1:0,1:0.667:1:0,0:0,1:0,1:0|1:58048505_T_TC:58048505:0,0,1,0
```

Figure 15 : Des informations sur chacune des variations observées fichier VCF.

Le VCF est conçu pour être évolutif afin d'inclure des millions de locus avec des données de génotype et des annotations provenant de milliers d'échantillons. Il a adopté un balisage textuel, avec une indexation complémentaire, pour permettre une création facile de fichiers tout en maintenant un accès rapide aux données [3].

Les premières colonnes représentent les informations dont nous disposons sur une variation prévue :

Colonne	Information
CHROM	Contig emplacement où la variation se produit.
POS	Position dans le contig où la variation se produit.
ID	Un . jusqu'à ce que nous ajoutions des informations d'annotation.
REF	Génotype de référence.
ALT	Génotype de l'échantillon.
QUAL	Probabilité à l'échelle Phred que la variante observée existe sur ce site.
FILTER	Un . si aucun filtre de qualité n'a été appliqué, PASS si un filtre est réussi ou le nom des filtres auxquels cette variante a échoué.

Tableau 14 : Informations sur une variation prévue.

2-Discussion :

La médecine de précision repose sur un profilage détaillé des échantillons de patients. Les développements récents des technologies de séquençage ont permis de mesurer l'ADN et l'ARN avec une résolution sans précédent à des prix en baisse. Cela fait du séquençage un outil idéal pour étudier les maladies génétiques comme le cancer. Les données brutes obtenues par des dispositifs de séquençage massivement parallèles sont volumineuses, sujettes aux erreurs et hautement redondantes. Il ne peut être converti en informations utiles qu'avec une analyse bio-informatique appropriée, ce qui rend les meilleures pratiques d'analyse des données de séquençage cruciales pour l'utilisation efficace des technologies de séquençage.

Dans cette étude, nous avons étudié des pipelines de séquençage du cancer les plus populaires pour proposer un workflow qui automatise un pipeline qui comprend les outils qui donne les meilleurs résultats .par exemple : nous avons utilisé plusieurs logiciels d'appel des varaiants pour détecter un nombre maximum des variatns possible.

Concernant les données, nous avons utilisé des échantillons hétérogènes et homogènes à haute définition appartenant à une seule tumeur pour mesurer les performances des algorithmes de séquençage pour différents niveaux d'hétérogénéité dans un scénario réaliste. Pour les données récupérées du séquenceur du centre de Sétif, nous avons constaté qu'elles ne sont pas exploitables Ceci est dû à la partie avant séquençage qui demande plus de maitrise.

L'analyse de séquençage du cancer comprend deux étapes principales, à savoir l'alignement et la découverte de variantes. Dans notre travail, nous avons utilisé les algorithmes les plus populaires que nous avons pu obtenir pour les deux étapes.

Conclusion :

Les progrès rapides réalisés dans les technologies de séquençage de nouvelle génération (NGS), conjugués à la baisse spectaculaire des coûts, ont fait du NGS l'une des principales approches appliquées dans la recherche sur le cancer. En outre, il est de plus en plus utilisé dans la pratique clinique pour le diagnostic et le traitement du cancer, grâce à la bio-informatique, il est possible de faire une étude sur le gène et avoir exactement les réponses aux questions posées.

Notre étude sur le domaine de l'oncologie et nos recherches sur les logiciels et les différents outils qui nous donnent la main de créer un pipeline et notre réalisation de l'importance de la bio-informatique et la biologie pour trouver des solutions, nous on permet de réaliser notre objectif d'avoir un workflow qui nous donne les variants trouver dans les données séquencés avec les techniques de séquençage de la nouvelle génération.

On a eu plusieurs difficultés durant notre étude, mais nous avons eu le reflex et le bon encadrement pour les surmonter. On a eu des difficultés dans les mises à jour des logiciels et outils, le téléchargement et la récupération des données, et aussi sur l'environnement l'Unix mais notre objectif était toujours de réaliser ce projet pour aider à la progression du domaine de la bio-informatique et l'oncologie en Algérie, grâce aussi à notre visite au CHUdu Sétif.

Enfin, nous avons la volonté et l'esprit de développer notre travail encore plus :

- Développer nous-même des outils
- Travailler avec des données réelles grâce au séquençage effectuer en Algérie
- Réaliser un workflow pour d'autres domaines, non seulement l'oncologie
- Tester plusieurs d'autres outils.

Bibliographie

[1]. Laganà A, et al. (2018). Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. JCO Precis Oncol.

[2]. Welch JS, Link DC. (2011). Genomics of AML: clinical applications of next-generation sequencing. Hematology Am Soc Hematol Educ Program.

[3]. Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics, Volume 27, Issue 15, 1 August 2011, Pages 2156–2158, https://doi.org/10.1093/bioinformatics/btr330.

[4]. Di Tommaso P, et al. (2017). Nextfow enables reproducible computational workfows. Nat Biotechnol.

[5]. Pirooznia, M., Kramer, M., Parla, J. et al. (2014).Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics 8. https://doi.org/10.1186/1479-7364-8-14.

[6]. Piecing together the genome: the long and short of it all By Sarah Sharman, PhD, Science Writer.

[7]. do Valle, Í.F., Giampieri, E., Simonetti, G. et al. (2016). Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinformatics 17. https://doi.org/10.1186/s12859-016-1190-7.

[8]. Maxam, A. M et Walter Gilbert. (1973).“Une nouvelle méthode de séquençage de l'ADN.” Actes de la National Academy of Sciences. National Acad Sciences.

[9]. Erwin van Dijk, Claude. (2021). La révolution de la génomique : les nouvelles méthodes de séquençage et leurs applications, /www.clinisciences.com/achat/cat-sequencage-nouvelle-generation3452.htmlThermes.

[10]. Odle TG. (2017). Precision medicine in breast cancer. Radiol Technol.

[11]. Bødker JS, et al. (2013). Development of a precision medicine workfow in hematological cancers, Aalborg University Hospital, Denmark. Cancers (Basel).

[12]. Chen S, Zhou Y, Chen Y, Gu J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics.

[13]. Jäger N.(2020). Bioinformatics workffows for clinical applications in precision oncology. In: Seminars in cancer biology. Academic Press; 2021. https://doi. org/10.1016/j.semcancer.

[14]. Alkodsi, A., Louhimo, R., & Hautaniemi, S. (2015). Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Briefings in bioinformatics, 16(2), 242–254. https://doi.org/10.1093/bib/bbu004.

[15]. Garraway LA,(2013). Lander ES. Lessons from the cancer genome. Cell.

[16]. Jeffrey A. SoRelle, MD; Megan Wachsmann, MD, MSc; Brandi L. Cantarel, PhD (2020). Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays

[17]. Andrews, S. et al. (2010). FastQC: a quality control tool for high throughput sequence data.

[18]. 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.

[19]. Koboldt DC. (2020). Best practices for variant calling in clinical sequencing. Genome Med.

[20]. Xu C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and structural biotechnology journal, 16, 15–24. https://doi.org/10.1016/j.csbj.2018.01.003.

[21]. Kısakol, B., Sarıhan, Ş., Ergün, M. A., & Baysan, M. (2021). Detailed evaluation of cancer sequencing pipelines in different microenvironments and heterogeneity levels. Turkish journal of biology = Turk biyoloji dergisi, 45(2), 114–126. https://doi.org/10.3906/biy-2008-8.

[22]. Singer, J., Irmisch, A., Ruscheweyh, H. J., Singer, F., Toussaint, N. C., Levesque, M. P., Stekhoven, D. J., & Beerenwinkel, N. (2019). Bioinformatics for precision oncology. Briefings in bioinformatics, 20(3), 778–788. https://doi.org/10.1093/bib/bbx143.

[23].Yan-Fang Guan, Gai-Rui Li, Rong-Jiao Wang, Yu-Ting Yi, Ling Yang, Dan Jiang, Xiao-Ping Zhang, and Yin Peng..Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer.

[24].DePristo MA, et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

[25].Martin M. (2017). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J.

[26]. Osman, W., Laganà, A. (2022). Software Workflows and Infrastructures for Precision Oncology. In: Laganà, A. (eds) Computational Methods for Precision Oncology. Advances in Experimental Medicine and Biology, vol 1361. Springer, Cham. https://doi.org/10.1007/978-3-030-91836-1_2.

[27]. Aran D, Sirota M, Butte AJ. (2017). Systematic pancancer analysis of tumour purity. Nat Commun.

[28]. Lee S, et al. (2017). NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res.

[29]. Sun JX, et al. (2018). A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol.

[30]. Liu, X., Bienkowska, J. R., & Zhong, W. (2020). GeneTEFlow: A Nextflow-based pipeline for analysing gene and transposable elements expression from RNA-Seq data. PloS one, 15(8), e0232994. https://doi.org/10.1371/journal.pone.0232994.