ACADEMIE DE MONTPELLIER

UNIVERSITÉ DE MONTPELLIER

MASTER « Statistiques pour les Sciences de la Vie. »

Master 1 SSV – Année 2020-2021

MEMOIRE sur le stage

Comparaison des outils bio-informatiques pour la découverte de variants génétiques

Effectué du 01/04/2021 au 31/08/2021

• à : UMR DGIMI-INRAE-UNIV DE MONTPELLIER

• sous la direction de Kiwoong NAM

• par Messaoud LEFOUILI

• soutenu le 15/06/2021

• devant la commission d'examen

J.N. BACRO, C. REYNES, R. SABATIER

INRAE

DGIMI
Diversité, Génomes & Interactions
Microorganismes Insectes

Remerciements

Mes remerciements vont à mon maître de stage, Kiwoong Nam. Durant ces 2 mois passés dans l'unité, son soutien et son suivi m'ont permis d'acquérir de nombreuses connaissances. Je suis évidemment reconnaissant pour sa disponibilité et son implication tout au long de mon stage, ainsi que pour sa relecture de ce document.

Beaucoup de remerciements à d'ALENCON Emmanuelle, la responsable de l'équipe EHA, pour m'avoir bien accueillie dans son équipe et offert l'opportunité de découvrir le monde de la recherche. Pour son aide et son soutien constant, pour sa lecture et pour toute correction apportée à ce rapport.

Je remercie toute l'équipe EHA pour l'agréable accueil au sein de cette unité.

Mes remerciements vont également aux membres du commission d'examen J.N. BACRO, C. REYNES, et R. SABATIER pour l'évaluation du travail.

SOMMAIRE

Liste des tables

Liste des figures

Introduction

En raison de la baisse des coûts des technologies de séquençage de nouvelle génération et de l'augmentation des capacités de production de données, les analyses des données de polymorphisme de la séquence du génome en entier deviennent courantes dans l'étude de nombreux organismes. Cette approche a fait de la génétique des populations une discipline guidée par les données, elle lui permet de jouer un rôle essentiel dans les analyses de l'écologie moléculaire et de la biologie de la conservation, où elle fournit un cadre pour comprendre la distribution de la variabilité génétique entre les populations et pour déduire l'histoire démographique des populations naturelles à partir de données moléculaires. Elle joue également un rôle central dans les études de l'évolution moléculaire, en fournissant une base pour comprendre les contributions de la mutation, de la dérive génétique et de la sélection naturelle dans l'évolution des gènes, des génomes et des espèces ; et dans les changements démographiques (Pool et al., 2010).

L'équipe de recherche "Epigénétique, holocentrisme et adaptation de l'insecte" qui fait partie de l'UMR DGIMI appartenant à l'INRAE et en collaboration avec l'Agence Nationale pour la recherche Allemande, mène un projet qui vise à comprendre les origines de la divergence entre deux souches de *Spodoptera frugiperda*. Kiwoong Nam et al veulent utiliser l'approche du polymorphisme de séquence du génome entier pour identifier les locus divergents entre les souches. Dans le cadre de cette étape du projet, une sous-étude est actuellement mise en œuvre pour essayer de déduire la meilleure stratégie/pipeline pour identifier des SNV (pour Single Nucleotide Variant, c'est à dire les variants à un seul nucléotide) concrets dans le but d'une caractérisation précise et non biaisée de cette divergence génétique entre les souches.

La technologie de séquençage Illumina est actuellement la plus utilisée pour les études de séquençage du génome entier visant à identifier les SNV. Elle produit des lectures relativement précises, mais leur longueur est limitée, généralement à moins de 300 paires de bases (pb). Ces lectures courtes et précises sont bien adaptées à l'identification des SNV et des petits indels (pour insertion délétion), mais sont moins rentables pour l'assemblage de novo, le phasage des haplotypes et la détection des variants structurels, qui nécessitent tous des informations sur des séquences plus longues. Le séquençage en temps réel d'une seule molécule (SMRT), également connu sous le nom de séquençage PacBio (Pacific Biosciences, nom des producteurs de la technologie), ainsi que le séquençage Oxford Nanopore, constituent deux alternatives prometteuses qui permettent aux scientifiques de surmonter les limites de lecture d'Illumina, mais ils ont souffert de problèmes de

précision. PacBio a récemment introduit une technologie de séquençage appelée séquençage par consensus circulaire (CCS). Cette technologie permet d'obtenir une séquence consensuelle à partir de plusieurs passages d'une seule molécule modèle, produisant ainsi des lectures précises à partir de sous-lectures individuelles susceptibles d'erreurs. Ces lectures précises sont appelées lectures longues à haute fidélité (HiFi) et ont une longueur moyenne de 13,5 kilobases (kb). Wenger et al., ont utilisé le séquençage par consensus circulaire (CCS) pour produire des lectures longues du génome humain avec une précision moyenne de 99,8 %. Ils ont affirmé que la méthode CCS permet d'égaler ou de dépasser la capacité du séquençage à lecture courte pour détecter les petits variants et les variants structurels.

Bien que les technologies de séquençage de nouvelle génération (NGS) telles qu'Illumina et PacBio s'améliorent constamment, transformer les lectures de séquences brutes en informations biologiquement significatives reste une étape difficile. Les technologies NGS souffrent encore de nombreux défis techniques qui rendent difficile l'obtention d'un enregistrement complet et précis de la variation des séquences. Compte tenu de ces défis, plusieurs nouveaux outils bio-informatiques ont été développés pour identifier les variants de séquence à partir des données NGS. Ces outils d'analyse jouent un rôle dans l'étude des variations et permettent d'inférer des informations aussi importantes que le type et la qualité de la lecture (Pirooznia et al., 2014).

L'objectif du projet sur lequel je travaille dans le cadre de mon stage est de comparer différentes stratégies basées sur la technologie de séquençage et les pipelines d'analyse pour identifier des variants de séquence (Variant Calling) à partir de données de séquences du génome entier. L'objectif est de déterminer quelle stratégie est optimale pour le projet d'étude en laboratoire des insectes afin d'identifier les locus divergents entre différentes souches ou des locus ciblés par sélection naturelle.

Présentation de l'organisme d'accueil :

I. Présentation de INRAE

INRAE, l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement est un organisme de recherche français résultant de la fusion en Janvier 2020 entre l'Inra, l'Institut national de la recherche agronomique et l'Irstea, l'Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture. Grâce à cette fusion, il représente une masse importante d'infrastructures de recherche (www.inrae.fr/nous-connaitre).

L'objectif de l'INRAE s'articule autour de six thèmes principaux : Changement climatique et risques, Agroécologie, Biodiversité, Alimentation, Santé globale, Bioéconomie, Société et territoires. Il s'appuie sur des approches systémiques et des recherches et innovations interdisciplinaires pour produire et diffuser des connaissances permettant de répondre aux enjeux de l'agriculture, de l'alimentation et de l'environnement (www.inrae.fr/nous-connaitre).

II. Présentation de l'UMR DGIMI

L'UMR Diversité, Génomes, Interactions Microorganismes-Insectes (DGIMI), est une unité mixte de recherche créée en 2011. Elle est située sur le campus du Triolet de l'Université de Montpellier (UM) et a pour tutelle l'Université de Montpellier (UM) et l'INRAE. Cette UMR comporte 4 équipes : Biologie intégrative des interactions hôte-parasitoïde (B2iHP), Dynamique des interactions densovirus-insectes (DIDI), Epigénétique, holocentrisme et adaptation de l'insecte (EHA), et Biologie intégrative des interactions bactéries-insectes-nématodes entomopathogènes (BIBINE) (www6.montpellier.inrae.fr/dgimi/).

Les recherches de cette unité visent principalement à comprendre les mécanismes d'interaction entre les ravageurs des cultures et leur environnement biotique. Elles sont axées sur la mise en évidence des stratégies adaptatives des agents pathogènes et des parasites par rapport à l'insecte hôte et des stratégies adaptatives de l'insecte vis-à-vis de son hôte (plante ou insecte) ou de ses ennemis naturels. Les approches utilisées dans ces études comprennent aussi bien l'étude des mécanismes moléculaires que des études plus systémiques au niveau de la population.

III. Présentation de l'équipe de recherche Epigénétique, holocentrisme et adaptation de l'insecte

L'équipe Epigénétique, Holocentrisme et Adaptation de l'Insecte (EHA) a été formée en 2006. Cette équipe de recherche se concentre sur les études de la structure et de la fonctionnalité des génomes holocentriques appliquées aux insectes nuisibles pour les plantes. L'un des projets principaux de l'équipe est axé sur la variation génétique et la régulation épigénétique dans le génome holocentrique du lépidoptère ravageur *Spodoptera frugiperda*, principalement en relation avec les

stades de développement et l'adaptation à la plante hôte. Des études de génomique des populations sont également réalisées pour l'identification de l'évolution adaptative chez cette espèce.

Cadre de l'étude

Ce projet de stage fait partie d'une étude plus vaste qui vise à identifier les régions génomiques sous sélection positive chez l'un et l'autre des souches de l'insecte d'intérêt en utilisant une analyse de séquençage du génome entier pour déterminer les variants génétiques qui ont permis leur divergence. Ceci devrait permettre à terme d'inférer leur histoire évolutive. Actuellement, l'approche la plus commune utilise des lectures de séquençage Illumina des individus étudiés qui sont ensuite alignées à un génome de référence et utilisées pour identifier les variants génétiques entre individus. Le problème est que les lectures courtes Illumina ont de nombreuses limitations qui peuvent affecter l'analyse finale.

I. Qu'est-ce que la découverte des variants génétiques, « variant calling » en anglais ?

La découverte de variants est le processus qui consiste à identifier les différences ou les variations génétiques entre l'échantillon et la séquence du génome de référence. L'entrée typique est un ensemble de lectures alignées sous le format BAM ou équivalent , qui est parcouru par le logiciel de découverte des variants pour identifier les variants de séquence.(Roy et al., 2018). La Figure01 illustre de façon simplifiée les différentes étapes de la découverte de variants. Chaque étape du pipeline peut affecter la performance globale de l'identification des variants génétiques notamment la technologie de séquençage, l'outil d'alignement, l'outil de la découverte de variant et la méthode de phasage de l'haplotype. Ceci va avoir un impact sur l'interprétation qui en découle.

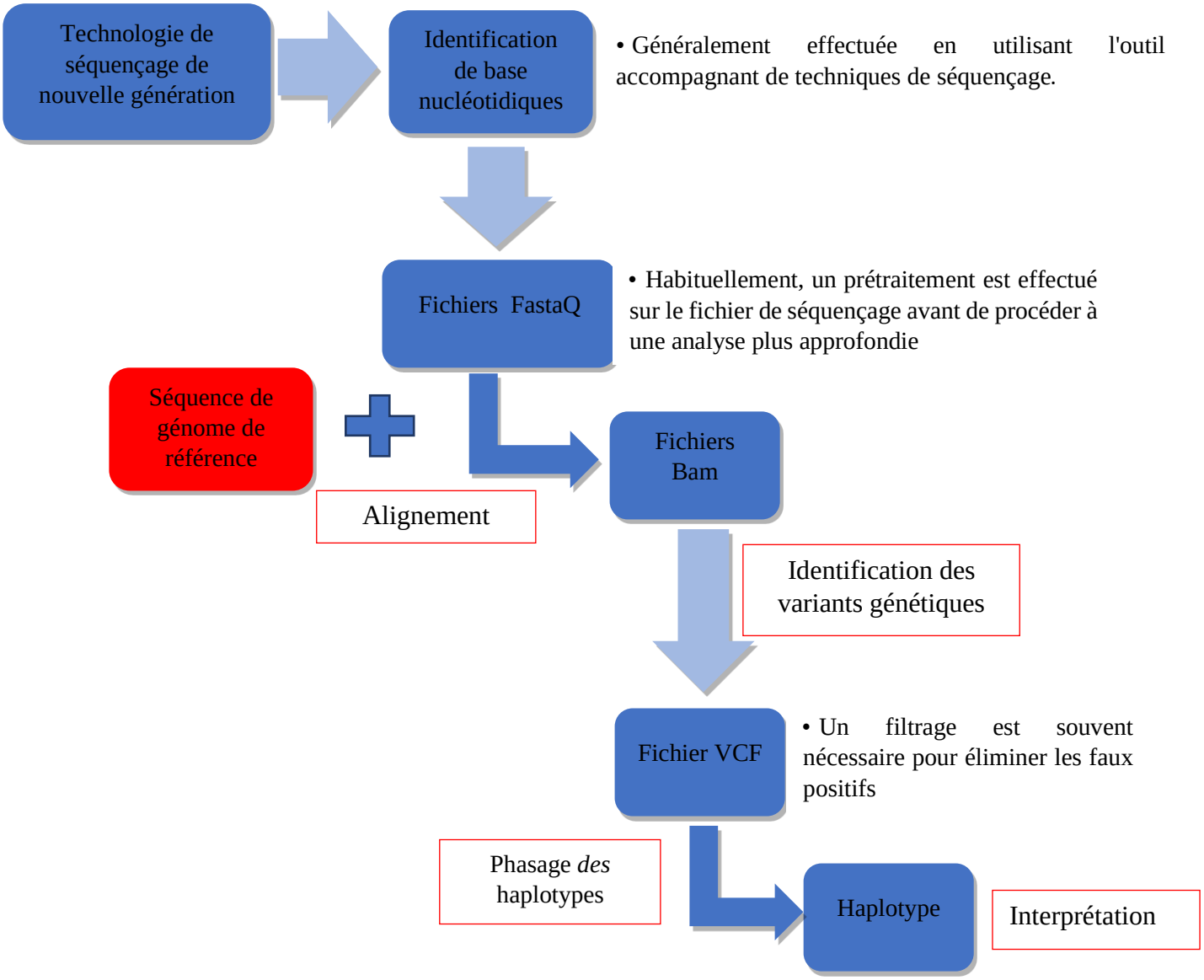


Figure 1. Pipeline bio-informatique pour l'identification de variants génétiques après le séquençage de nouvelle génération.

IV. Avantages et inconvénients de l'utilisation de séquences courtes

Bien que le séquençage Illumina permette de réaliser des analyses rapides, peu coûteuses et précises du génome entier avec un débit important, la faible longueur des lectures (<300 pb) pose généralement de nombreux problèmes qui peuvent affecter les analyses bio-informatiques ultérieures. Les lectures relativement courtes ne peuvent pas reconstruire entièrement les régions répétitives, ce qui conduit à des assemblages fragmentés incomplets et parfois erronés (Yuan et al., 2017) (Amarasinghe et al., 2020). En outre, cette longueur limitée ne permet pas aux algorithmes d'alignement de séquence de traiter efficacement la polyploïdie (Yuan et al., 2017). Le séquençage à l'aide de la technologie Illumina est aussi parfois basé sur l'amplification multiple des séquences

d'ADN segmentées, entraînant des quantités biaisées de différentes régions en raison d'une couverture inégale selon que le contenu en GC est élevé ou faible, ce qui peut conduire à une interprétation erronée (Roberts et al., 2021).

En pratique également, il existe de nombreux scénarios importants en génétique où une lecture courte de l'ADN est inadéquate. L'alignement de lectures courtes est généralement confronté à de nombreux défis autour des variations de structures génomiques communes (tableau 01).

Tableau 1. Les conséquences des variations de structure génomique sur l'alignement des courtes et longues lectures de séquençage (Pollard et al., 2018)

Variation de structure génomique	Lectures courtes	Lectures longues
Grande insertion	Les lectures au bord de l'insertion seront identifiées comme chimèriques Les lectures à l'intérieur de l'insertion seront soit non alignées, soit alignées de manière incorrecte.	Les lectures couvriront l'insertion ou auront suffisamment de contexte pour être identifiées comme séquence insérée.
Grande délétion	Les lectures couvrant la délétion peuvent être mal alignées ou seule une des lectures considérées s'aligne parce que la longueur mesurée de référence indique que la taille de l'insert dévie de la distribution prévue	Les lectures couvriront l'écart et la plupart auront suffisamment de contexte pour identifier la délétion.
Copy number variation (CNV)	Dans le cas où les CNV ont une longueur inférieure à celle de la lecture, ils seront correctement identifiés. Les lectures plus courtes peuvent être regroupées et sembler avoir une profondeur de séquençage accrue ou être identifiées comme étant mal alignées.	Les CNV ont une longueur toujours inférieure à celle de la lecture, ils seront donc correctement identifiés
Inversion	Les lectures seront soit représentées comme un alignement primaire sur le segment inversé, soit identifiées comme chimériques autour du bord de l'inversion avec une réduction de la profondeur.	Les lecteurs couvriront l'inversion

V. Les avantages et inconvénients des séquences longues

Les lectures longues couvrent et permettent d'identifier des variations auxquelles les lectures courtes ne permettent pas d'avoir accès. Les lectures de multi-kilobases peuvent permettre de couvrir divers segments répétitifs, ce qui minimise les ruptures d'assemblage et augmente la complétude de l'assemblage par rapport aux séquences à lecture courte. (Yuan et al., 2017). Les lectures longues peuvent ainsi améliorer la certitude d'alignement, l'identification des isoformes de transcription et la détection des variants structurels. En outre, le séquençage de molécules natives, qu'il s'agisse d'ADN ou d'ARN, élimine le biais d'amplification tout en préservant les modifications des bases. Ces capacités, ainsi que les progrès continus en matière de précision, de débit et de réduction des coûts,

ont commencé à faire du séquençage à long terme une option pour un large éventail d'applications en génomique pour les organismes modèles et non modèles.(Amarasinghe et al., 2020).

En revanche, les séquençages à lecture longue ont leurs propres inconvénients. Par exemple, ils nécessitent un ADN de haute qualité et de poids moléculaire élevé, ce qui est encore difficile dans de nombreux cas (Schwessinger and Rathjen, 2017). Ils coûtent également plus cher que le séquençage Illumina. Et il n'y a pas beaucoup de travaux qui abordent leur utilisation pour les études de population, contrairement à Illumina.

Méthodes :

I. Génération de données simulées :

Pour comparer les différentes stratégies, nous avons utilisé des données simulées. Les données simulées nous aident à mieux comprendre les ensembles de données spécifiques. Leur avantage est qu'elles nous permettent de générer autant de données que souhaité dans des scénarios contrôlés avec des paramètres prédéfinis dont les vraies valeurs sont connues. L'utilisation de données simulées est également utile pour la conception d'expériences telles que l'estimation de la profondeur de couverture requise pour l'assemblage du génome et la détection de variants génétiques(Escalona et al., 2016).

Afin d'évaluer dans quelle mesure les différents pipelines peuvent affecter la découverte de variants génétiques, nous avons simulé plusieurs autosomes de l'insecte modèle *Drosophila melanogaster*. Un fichier FASTA de la séquence de référence (Release 6) du génome de *Drosophila melanogaster* a été téléchargé de NCBI (Les numéros d'accès sont NT_033778.4, NT_037436.4, NC_004353.4, NT_033777.3, NT_033779.5 pour chaque bras chromosomique). Nous avons utilisé le programme Slim2 pour simuler différentes mutations (variants) dans les génomes d'une population. Slim2 a la capacité de simuler une grande variété de scénarios évolutifs complexes (Haller and Messer, 2017). Nous avons utilisé les paramètres suivants : Taux de mutation =2,8 × 10^{-9}, taux de recombinaison= 2.0× 10^{-8}, Type : mutation neutre, Taille de la population :10^{6}. Pour des raisons de capacité computationnelle, ces paramètres ont été réduits à 2,8×10^{-6}, 2,0×10^{-5} et 10^{3} pour le taux de mutation, le taux de recombinaison et la taille de la population, respectivement. Cette remise à l'échelle génère les mêmes données simulées.

Slim2 a simulé des mutations pendant 3000 générations. Nous avons pris au hasard 40 individus. Afin de générer des génomes plus réalistes, c'est-à-dire de transformer la longue séquence unique du génome en génomes diploïdes, nous avons joint le premier chromosome de chacun des deux individus. A la fin, nous avions 20 génomes diploïdes. Les génomes diploïdes sont ensuite passés par le simulateur Neat-genreads (“Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data | BMC Bioinformatics | Full Text,” n.d.) afin de

générer des lectures courtes Illumina avec une couverture de 40X et par le simulateur Sim-it pour générer de lectures longues HiFi avec une couverture de 5X.

D'autres scénarios de simulation sont prévus pour les prochains mois du stage. Nous utiliserons le même processus pour la simulation mais avec des paramètres différents : taux de mutation, couverture (profondeur) de séquençage et taux d'erreurs de séquençage. Le but est d'essayer de comprendre l'effet de ceux-ci sur la performance et la précision des différents pipelines pour la découverte de variants génétiques et d'identifier la stratégie la plus adaptée pour tolérer la divergence génétique de la population et les erreurs de séquençage.

VI. Pipelines étudiés :

Dans la présente étude, nous visons à évaluer la meilleure combinaison de technologies de séquençage, d'aligneur de séquences et d'outils de découverte de variants pour détecter les variations de nucléotides uniques (SNV). Pour atteindre cet objectif, nous avons effectué une analyse comparative des données simulées de lectures Illumina et de lectures HiFi. Nous avons utilisé plusieurs pipelines pour traiter ces deux types de données (Figure 02). Nous avons choisi GATK (McKenna et al., 2010) et BCFtools (Danecek et al., 2021) les deux programmes les plus utilisés pour le traitement et l'analyse des données de séquençage pour la découverte de variants (Danecek et al., 2021 ; Robinson et al., 2017). Nous allons également utiliser une nouvelle approche développée et suggérée par mon encadrant Kiwong Nam : HPVC (HiFi Phased Variant Caller).

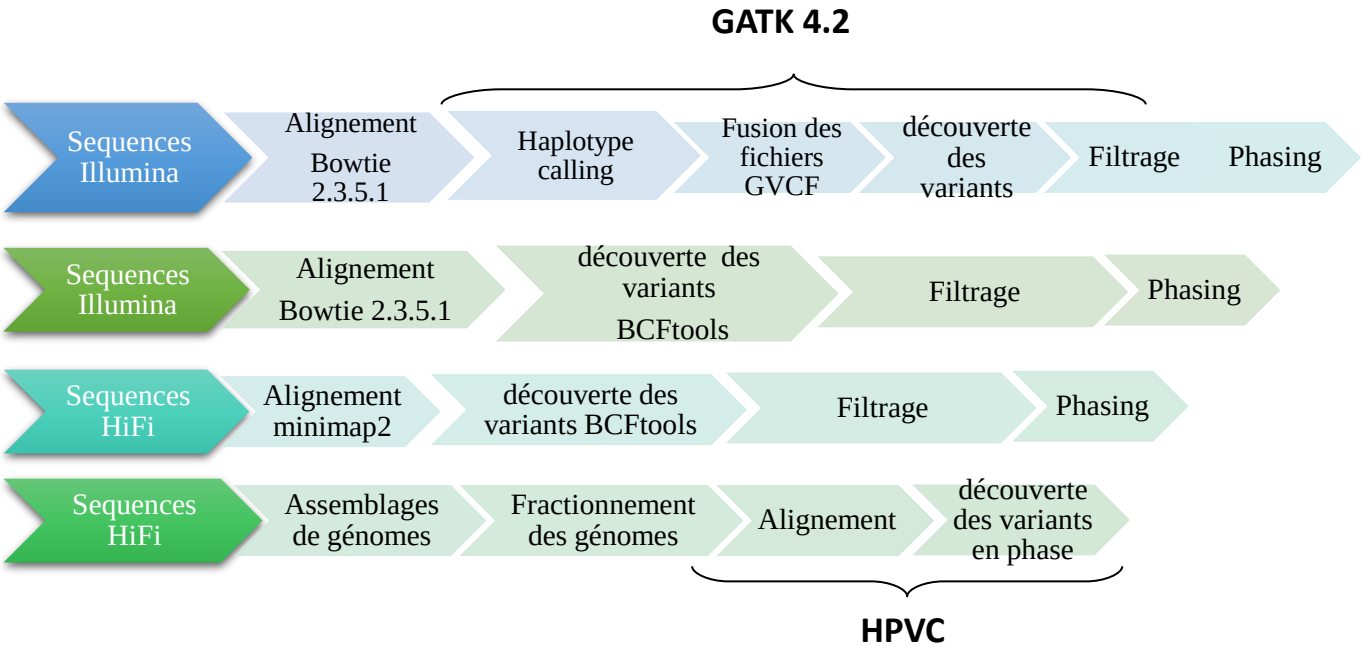


Figure 2.les pipelines bio-informatiques étudiés et évalués pour la découverte de variants dans notre étude.

VII. Evaluation de la performance des pipelines de découverte de variants.

Les variants déterminés par les pipelines ont été comparés sur la base de leur position avec l'emplacement des variants déjà connus qui ont été générés par Slim2.

Dans ce but, j'ai écrit un programme java pour analyser le fichier VCF et capturer l'emplacement des variants. Ces emplacements ont été comparés au bon emplacement (position réelle connue) qui ont été générés par Slim2, afin de déterminer si chaque variant représente un vrai positif ou un faux positif. Pour évaluer la performance des différents pipelines, nous avons calculé le nombre de vrais positifs (TP), de faux positifs (FP) et de faux négatifs (FN) ; ceux-ci ont été définis sur la base des positions et non du type de mutation de la base; TP est un vrai variant positif qui existe dans l'ensemble de données du fichier de position correcte et qui est également détectée par le pipeline; FP est un faux variant positif qui n'existe pas dans le fichier de position correcte et qui est détecté par le pipeline; FN est un faux variant négatif qui existe dans le fichier de position correcte et qui n'est pas détectée par le pipeline.

La performance du pipeline de découverte de variants a été estimé statistiquement comme suit :

$$sensibilité(Se) = \frac{TP}{(TP+FN)}$$

$$False\ Discovery\ Ratio = taux\ de\ fausse\ découverte = \frac{FP}{(TP+FP)}$$

$$F_{score} = \frac{2TP}{(2TP+FP+FN)}$$

De nombreuses études ont utilisé ces paramètres pour comparer les performances des pipelines bio-informatiques (Andreu-Sánchez et al., 2021; Chen et al., 2019; Kumaran et al., 2019; Zhao et al., 2020).

Résultats et discussion :

I. Taux d'alignement:

Le taux d'alignement moyen sur le génome de référence en utilisant bowtie2 pour les lectures Illumina des 20 individus simulés était de 77,05 %. Alors qu'il a atteint 99,95 % pour l'alignement des lectures HiFi en utilisant minimap2.

II. Découverte des variants :

1 Les résultats avant filtrage :

Le tableau 02 représente les résultats des différents pipelines. Les trois pipelines ont généré une proportion élevée de faux positifs qui représentent plus de 40% des variants identifiés. Ces chiffres nous montrent qu'un filtrage pourrait être nécessaire pour tous les pipelines. En regardant le nombre de vrais positifs et la sensibilité (taux de vrais positifs), nous pouvons voir que PacBio CCS a identifié le plus grand nombre total de variants avec également le plus grand taux de vrais positifs de 99,73%. En comparant les résultats des lectures Illumina, nous pouvons observer que le pipeline BCFtools est légèrement plus performant que le pipeline GATK avec plus de vrais positifs et moins de faux positifs.

Pour comparer la performance globale des trois pipelines, nous avons utilisé le score F qui permet de capturer à la fois la sensibilité et la précision du pipeline dans la détection des variants génétiques. Cela nous a montré que le pipeline basé sur les lectures longues a mieux fonctionné que les autres pipelines, car il a montré le score F le plus élevé. Par rapport aux autres pipelines, ce pipeline a permis d'identifier le plus grand nombre de vrais variants 4.339.012 qui représentent 99,73% des variants existants.

Tableau 2. Evaluation des pipelines avant filtrage

Technologie	**Aligneur**	**Coverage**	**Genotyper**	**TP**	**FP**	**FN**	**Se**	**TFD**	**F_{score}**
Illumina	Bowtie2	40 X	GATK	3.908.611	3.546.406	442.165	89,84%	47,57%	66,22%
Illumina	Bowtie2	40 X	BCFtools	3.935.828	2.761.101	414.948	90,46%	41,23%	71,25%
PacBio CCS	Minimap2	5 X	BCFtools	4.339.012	3.380.650	11.764	99,73%	43,79%	71,89%

TFD : $taux\ de\ fausse\ découverte$, Se : Sensibilité

2 Filtrage :

Comme nous l'avons vu sur les résultats, la découverte de variants génétiques a typiquement généré beaucoup d'erreurs (False positive). Une quantité aussi importante de faux positifs peut affecter l'analyse en aval. C'est pourquoi les chercheurs effectuent souvent un filtrage pour réduire les faux positifs. Idéalement, on aimerait éliminer tous les faux positifs tout en maintenant un niveau de sensibilité aussi élevé que possible afin que les mutations d'intérêt, en particulier les mutations rares, ne soient pas perdues dans le processus de filtrage. Cet aspect nous incite à vouloir déterminer l'impact du filtrage sur les performances des différents pipelines.

Pipelines avec GATK

Les développeurs de GATK suggèrent deux approches de filtrage, le Hard Filtering et le VQSR (Variant Quality Score Recalibration) : le Hard Filtering est basé sur le choix manuel de seuils spécifiques pour un ou plusieurs paramètres qui décrivent chaque variant, à partir desquels nous éliminons tout variants dont la valeur est supérieure ou inférieure aux seuils fixés. VQSR utilise des algorithmes d'apprentissage automatisés pour apprendre à partir des données quels sont les profils d'annotation des variants connus (vrais positifs) et des mauvais variants (faux positifs) dans un ensemble de données particulier. Cela permet au filtrage d'extraire les variants en fonction de la façon dont ils se regroupent selon plusieurs dimensions. Bien que les meilleures pratiques de GATK recommandent l'utilisation de VQSR, c'est rarement possible, surtout pour les organismes non

modèles, car cette approche nécessite un grand nombre de variants et des ressources de variants connus bien cataloguées.

Nous avons testé les deux approches. Pour le Hard Filtering, nous avons utilisé les paramètres recommandés avec les seuils recommandés (QD<2, FS>60, MQ<40, MQRankSum < -12.5, ReadPosRankSum< - 8), (voir en annexe 01 , pour plus de détails sur ces paramètres). Pour le filtrage VQSR, nous avons simulé un autre ensemble de données, effectué la découverte de variants et sélectionné uniquement les vrais positifs pour ce jeu de données. Nous avons fait la même chose pour les données d'entraînement mais avec la différence d'utiliser l'ensemble des vrais positifs et des faux positifs. Ces deux ensembles nous ont permis de recalibrer et de filtrer les variants identifiés dans le jeu de données simulées initial. Le tableau 02 représente le résultat du filtrage GATK.

Tableau 3. Evaluation de pipelines GATK pour les lecteurs Illumina après le Filtrage

Technologie	Genotyper	Filtrage	TP	FP	FN	Se	TFD	F_{score}
Illumina	GATK	Aucune	3.908.611	3.546.406	442.165	89,84%	47,57%	66,22%
Illumina	GATK	Hard Filtering	3.532.106	2.761.101	818.670	81,18%	44,80%	65,71%
Illumina	GATK	VQSR	3.901.076	2.882.233	449.700	89,66%	42,49%	70,07%

TFD : $taux\ de\ fausse\ découverte$, Se : Sensibilité

Le résultat global révèle que les deux approches de filtrage ne sont pas suffisantes ; le taux de fausses découvertes est toujours supérieur à 40%. Autrement dit, plus de 40% des variants détectés sont des faux positifs. VQSR a permis d'améliorer légèrement les performances, dans l'ensemble, le filtre VQSR a éliminé 671 708 variants, dont 98,87 % (664 173 variantes) sont des faux positifs, ce qui a entraîné la perte de seulement 7535 vrais positifs. Cela a entraîné une diminution du taux de fausses découvertes de 5% tout en maintenant presque le même niveau de sensibilité de 89%. En ce qui concerne le Hard Filtering, la réduction du nombre de FP s'est faite au prix de la perte de nombreux vrais positifs. Le Hard filtering a éliminé 1 161 810 dont un tiers d'entre eux sont de vrais positifs (32,40% des variants éliminer). Cela a conduit à une diminution à la fois de la sensibilité et du taux de fausses découvertes. Ceci avec presque une même performance globale entre sensibilité et taux de fausses découvertes (presque le même F_{score}). Ce résultat était attendu, car le Hard Filtering repose sur une dimension linéaire pour le filtrage. Les mauvaises performances de VQSR peuvent également s'expliquer par la quantité de données disponibles. Selon l'équipe de GATK, VQSR fonctionne mieux pour le séquençage d'exomes entiers avec un minimum de 30 échantillons. Cette quantité de données nécessaires est très peu pratique, seuls quelques laboratoires disposent de ce type de ressources. Sur cette base, nous avons décidé d'étudier plus en profondeur les paramètres utilisés dans le Hard

Filtering de GATK pour essayer d'établir de meilleurs critères de filtrage. Nous avons construit une distribution de densité des valeurs de ces paramètres (Figure 3).

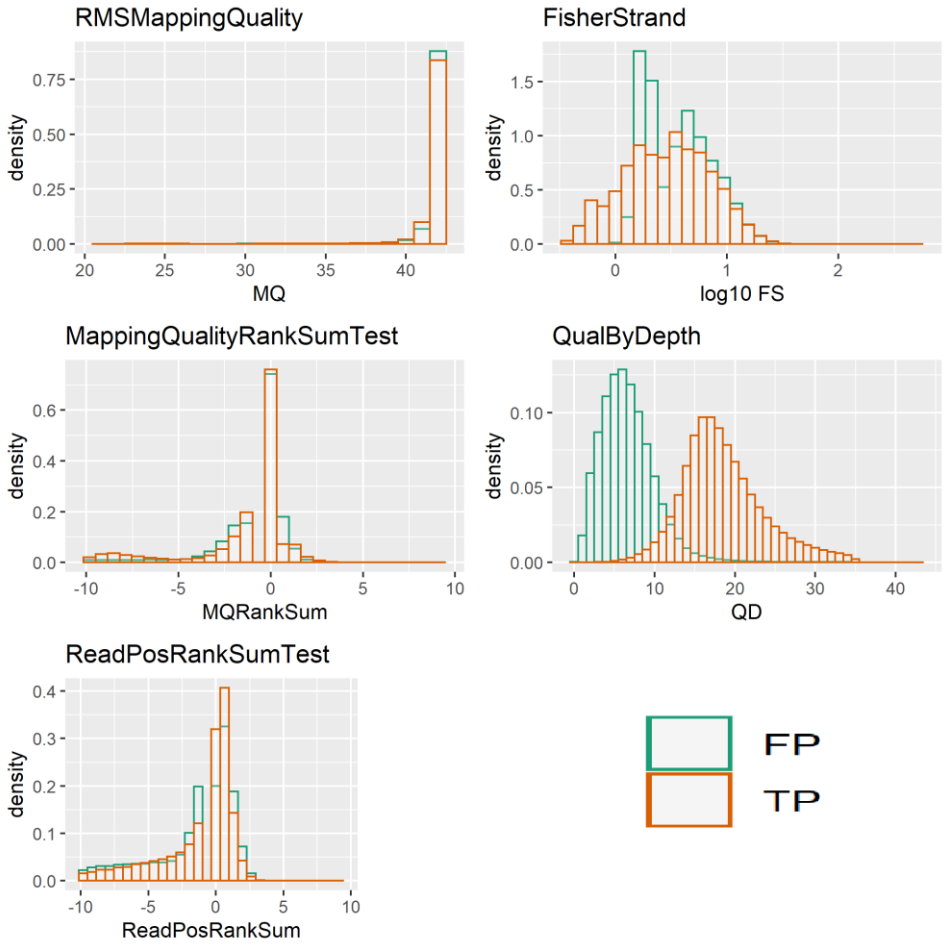


Figure 3. Distribution des différents paramètres qui caractérisent les variants générés par le pipeline GATK des lectures Illumina

Sur les cinq paramètres, seul le score QD (**Quality By Depth)** a permis de distinguer les vrais positifs des faux positifs, ce qui indique le rôle important que le QD peut jouer dans le filtrage. Le QD est le score QUAL -la probabilité qu'un enregistrement de variants soit vrai- normalisé par la profondeur d'allèle pour une variation génétique dans chaque échantillon étudié. Il augmente lorsqu'il y a plus de preuves de l'existence d'un variant à cette position. Si la qualité par profondeur est faible, on en déduit que les preuves de l'existence d'un variant sont faibles en proportion du nombre de lectures disponibles.

Sur la base de cette observation, nous avons appliqué différents filtrages en utilisant différentes valeurs du score QD (tableau annexe 02). En utilisant ces résultats, nous avons pu construire une courbe ROC afin d'identifier le seuil qui nous donne le meilleur équilibre entre sensibilité et spécificité pour une meilleure performance d'un test de classification binaire (voir annexe 01 pour

plus d'information sur le courbe ROC). Dans une courbe ROC le point le plus proche du coin supérieur gauche assure le meilleur équilibre entre sensibilité et spécificité, donc la meilleure performance pour le test. Ce point correspond à QD> 12. Ces résultats pourraient être adaptés uniquement à l'ensemble de données que nous avons utilisé. Nous pourrions simplement suggérer l'hypothèse que le score QD est crucial pour le filtrage, et que la distribution bimodale de QD peut être due aux vrais positifs et aux faux positifs. Plus d'études avec des données simulées et empiriques sont nécessaires pour aider à choisir des critères réalistes concernant le filtrage avec QD SCORE.

Pipelines BCFtools

Pour les pipelines BCFtools, il est généralement recommandé d'effectuer un filtrage rigoureux basé sur le score QUAL de chaque variant. Ce score représentant la précision de chaque découverte de variant. Il s'agit en fait d'un score à l'échelle de Phred, soit la probabilité qu'un variant soit vrai. Bien que le fichier VCF généré par GATK contienne aussi ce score, il a été publié qu'il n'est pas informatif dans le cas de la découverte de variant par GATK, et il n'est pas utilisé pour GATK contrairement à BCFtools. Nous avons confirmé ce fait en regardant la distribution du score QUAL pour les trois fichiers. Dans BCFtools, nous pouvons observer comment les vrais positifs sont plus concentrés sur les valeurs élevées de QUAL. Ce n'est pas le cas pour GATK.

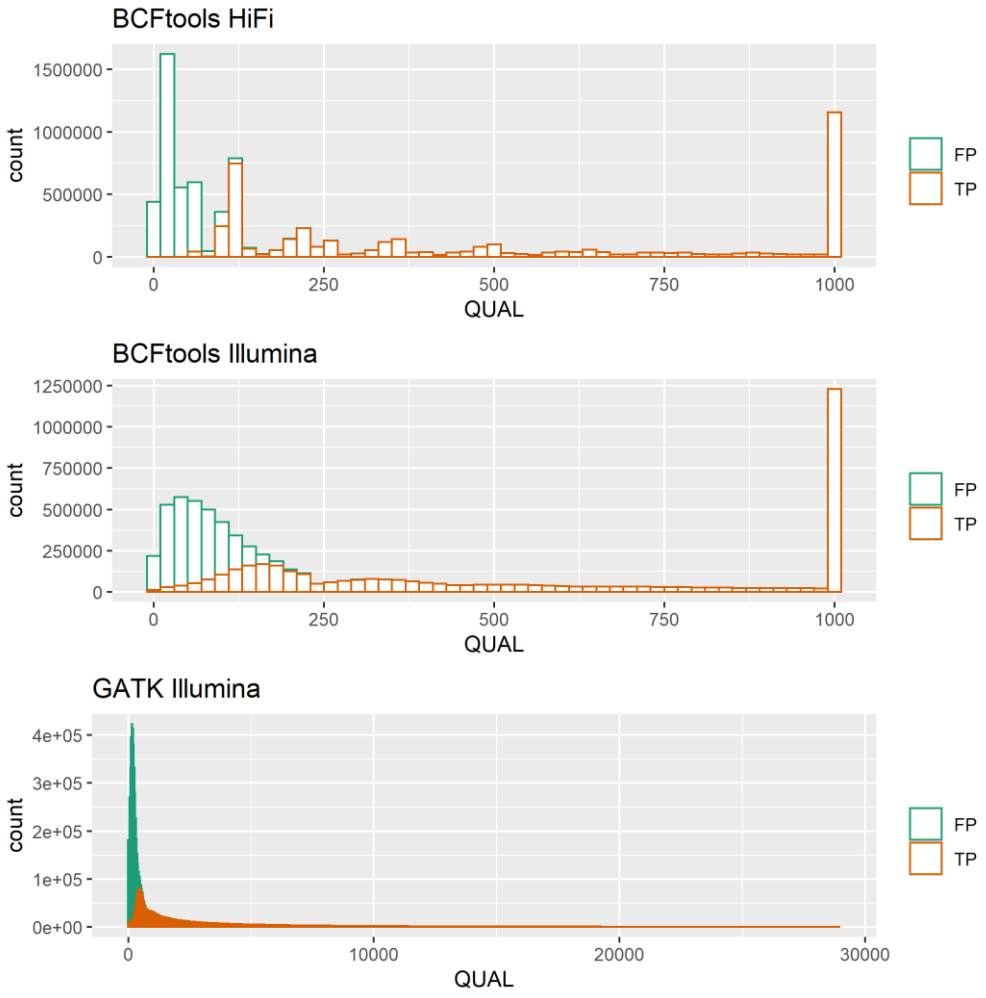


Figure 4. Distribution de QUAL score des variants identifiésr dans les fichiers VCF génersé par les différents pipelines étudiés

En utilisant différents seuils du score QUAL, nous avons effectué un filtrage multiple (tableau Annexe 03 et Annexe 04). En utilisant ces données, nous avons construit la courbe ROC pour Illumina+BCFtools, Illumina+GATK et HiFi+BCFtools (Figure. 05)

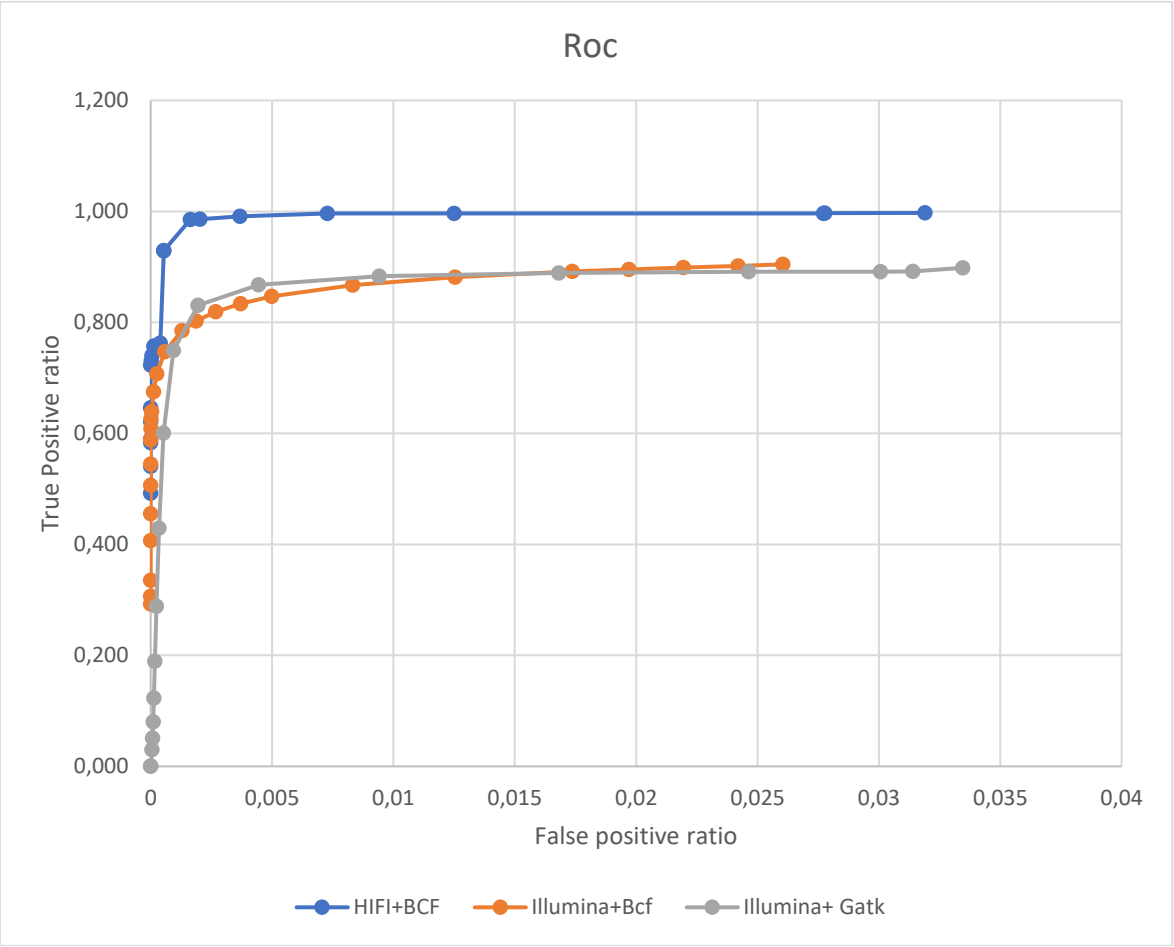


Figure 5. Courbe roc des différents pipelines étudiés

Dans une courbe ROC, l'aire sous la courbe peut être interprétée comme la probabilité que, parmi deux sujets choisis au hasard, un positif et un négatif, la valeur du paramètre discriminant soit plus élevée pour le positif que pour le négatif. Par conséquent, une AUC de 0,5 (y=x) indique que le test est non-informatif. Une augmentation de l'AUC indique une amélioration des capacités discriminatoires, avec un maximum de 1,0. c'est-à-dire plus la courbe s'éloigne de la diagonale (y=x) et s'approche de point (0,1) plus le test est performant en terme de discrimination. Sur la base dess courbes ROC dans la figure 05 Illumina+BCFtools et Illumina+GATK sont équivalents en termes de performance. HiFi+BCFtools est le meilleur pipeline.

La courbe ROC nous permet également de déterminer le meilleur seuil de filtrage. Nous avons appliqué ces filtrages optimaux (table04).

Tableau 4. Résultats des découvertes des variants après le filtrage avec les seuils optimaux déterminés par les courbes ROC.

Technologie	Genotyper	Paramètre de filtrage	TP	FP	FN	Sensibilité	TFD	F_{score}
Illumina	GATK	QD>12	3.546.406	207.418	735.304	83,10%	5,43%	88,47%
Illumina	BCFtools	QUAL >120	3.564.185	283.564	786.591	81,92%	7,37%	86,95%
PacBio CCS	BCFtools	QUAL >90	4.287.705	173.494	63.071	98,55%	3,89%	97,32%

Les résultats du tableau 04 nous montrent que le filtrage des variants nous permet d'améliorer radicalement les résultats de la découverte des variants. Le taux de fausses découvertes a diminué de 47,57%, 41.23%, et 43,79 à 5.43%, 7.37% et 3.89% pour les pipelines Illumina+GATK ,Illumina+BCFtools, HiFi+BCFtools, respectivement . Cette diminution importante de TFD n'a causé qu'une faible diminution de Sensibilité de 6.74 %, 8.54% et 1,18% pour Illumina+GATK ,Illumina+BCFtools,et HiFi+BCFtools, respectivement. Les performances des pipelines basés sur Illumina sont équivalentes avec une différence de Fscore de seulement 1,79%. Les lectures longues ont montré une meilleure performance sur la base d'un plus grand nombre de vrais positifs détectés (98,55%) et seulement 3,89% de taux de fausse découverte.

Compétences acquises pendant le stage

D'un point de vue technique, ce stage m'a permis d'apprendre à utiliser différents outils d'analyse bio-informatique sous l'environnement shell Bash, ainsi qu'à utiliser un cluster de calcul haute performance (Genotoul), ce qui est une première pour moi. J'ai appris les différentes étapes qui nous permettent de passer de données brutes de séquençage à des informations interprétables dans le domaine de la biologie. J'ai appris à utiliser les outils les plus courants dans les analyses génétiques : Bowtie, Minimap Samtools, BCFtools et GATK. J'ai également appris les différents formats de fichiers standardisés en bio-informatique : Fasta, FastaQ, Sam, et Vcf.

Ce stage m'a également donné l'occasion d'apprendre les commandes essentielles de Bash et de mettre en pratique ce que j'ai appris en programmation java en écrivant des programmes pour un objectif spécifique ; je me suis essentiellement familiarisé avec l'écriture de syntaxes de programmation simples qui m'ont permis d'extraire et de comparer des données à partir de fichiers texte. Il m'a permis également d'apprendre à utiliser les outils de visualisation des données,

notamment la bibliothèque "ggplot2" en R, afin de mieux comprendre une grande quantité de données, et à utiliser R pour analyser les données.

Ce stage m'a également permis d'acquérir une meilleure connaissance de la recherche en général. En assistant aux différentes présentations du laboratoire, à la réunion mensuelle de l'équipe, et en interagissant avec différents membres du laboratoire qui m'ont accordé une partie de leur temps pour m'expliquer le travail général de l'équipe et les projets spécifiques sur lesquels ils travaillent. La participation à la réunion de l'équipe a été une expérience particulièrement enrichissante, car elle m'a permis d'observer les interactions entre scientifiques, non seulement sur des sujets scientifiques, mais aussi sur l'organisation et la planification d'un laboratoire de recherche. Cela a été pour moi l'occasion de faire une présentation de mon travail de stage aux membres du groupe.

Par ailleurs ce stage m'a permis d'entrer en contact avec différents domaines de recherche en biologie, allant de la génétique des populations, de la génomique et de l'évolution jusqu'à l'étude des insectes.

Mes perspectives pour la suite du stage

Je dois poursuivre le travail avec les lectures HiFi, en effet le processus d'analyse pour la découverte de variants de longues lectures prend beaucoup de temps. Cette étape peut donc prendre un certain temps. Entre-temps, nous pourrions générer d'autres données de simulation pour nous aider à faire une meilleure inférence sur différentes conditions, telles que des tailles de population, des taux de mutation et des taux d'erreur de séquençage différents, afin de mesurer leur effet sur les performances des différents pipelines.

Je dois procéder à la découverte de variants à partir de longues lectures en utilisant l'approche directe, ainsi que tester l'approche suggérée par mon encadrant Kiwong Nam qui consiste à identifier des variants à partir d'assemblages de séquences construits à partir de lectures HiFi en utilisant un programme qu'il a écrit.

Je vais suivre les mêmes étapes et utiliser les mêmes métriques que celles utilisées pour les pipelines d'Illumina pour évaluer les performances de ces deux approches, pour finalement comparer les résultats entre les différentes stratégies.

Conclusion

L'objectif de ce stage est d'essayer d'identifier les meilleures stratégies pour la découverte de variants, nous comparons et évaluons plusieurs stratégies basées sur différentes technologies de séquençage et outils bio-informatiques. L'objectif est de tester si les stratégies qui s'appuient sur des lectures longues nous aideraient à surmonter les difficultés liées aux lectures courtes. Idéalement, nous voulons un pipeline qui nous permette d'identifier les variants rares dans une population sans les confondre avec des faux positifs. Cela représente un grand défi car ces variants se manifestent dans une population avec une très faible fréquence et ils sont habituellement confondus avec des erreurs d'analyse dans le séquençage, dans l'identification de base, dans l'alignement et dans la découverte des variants.

Détecter toutes les variations génétiques, même les plus rares, et éviter les faux positifs est important car cela nous permettrait d'améliorer l'identification de la base génétique de problèmes complexes en biologie, tels que l'inférence des changements démographiques des populations naturelles et la compréhension des contributions des mutations, de la dérive génétique et de la sélection naturelle dans l'évolution des gènes, des génomes et des espèces.

Nous utilisons le comptage des vrais positifs (TP) et des faux positifs (FP), la sensibilité, le taux de fausses découvertes et le score F pour évaluer les performances des différents pipelines. La mesure de ces paramètres est possible car nous utilisons des données simulées avec des variants connus pour effectuer notre analyse. Nos premiers résultats montrent une amélioration prometteuse de la performance de la découverte de variants génétique en utilisant les lectures longues de séquençage. Nous avons montré que le pipeline BCFtools qui utilise les lectures HiFi avec une couverture du séquençage de 5X est plus performant que les autres pipelines des lectures Illumina. Il a nous permis de détecter la plupart des vrais variants tout en ayant une quantité très faible de faux positifs, Mais cela n'est possible que grâce à un bon filtrage qui est considéré comme aussi nécessaire que les autres étapes car il a radicalement amélioré les résultats. Nous avons également constaté que le filtrage basé sur QUAL fonctionne très bien sur les variants générés par BCFtools mais trouver le bon seuil pour le filtrage reste un problème difficile. Nous avons également constaté que QD est un paramètre critique dans le filtrage des variants générés par le pipeline GATK.

Bien que ces comparaisons aient été effectuées sur des données simulées, cette première analyse nous a permis de nous faire une idée des différences entre les méthodes et de leurs performances. Nous devons encore tester l'approche suggérée par mon superviseur, Kiwoong Nam. Nous pourrions également générer d'autres données simulées avec différentes conditions, telles que des tailles de population, des taux d'erreur de séquençage et des taux de mutation différents, afin d'évaluer les capacités des différents pipelines à les gérer.

Bibliographie

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biology 21, 30. https://doi.org/10.1186/s13059-020-1935-5

Andreu-Sánchez, S., Chen, L., Wang, D., Augustijn, H.E., Zhernakova, A., Fu, J., 2021. A Benchmark of Genetic Variant Calling Pipelines Using Metagenomic Short-Read Sequencing. Front. Genet. 12. https://doi.org/10.3389/fgene.2021.648229

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., Song, Y.-Q., 2011. Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56, 406–414. https://doi.org/10.1038/jhg.2011.43

Chen, J., Li, X., Zhong, H., Meng, Y., Du, H., 2019. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep 9, 9345. https://doi.org/10.1038/s41598-019-45835-3

Escalona, M., Rocha, S., Posada, D., 2016. A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet 17, 459–469. https://doi.org/10.1038/nrg.2016.57

Haller, B.C., Messer, P.W., 2017. SLiM 2: Flexible, Interactive Forward Genetic Simulations. Molecular Biology and Evolution 34, 230–240. https://doi.org/10.1093/molbev/msw211

Hsieh, F., Turnbull, B.W., 1996. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. The Annals of Statistics 24, 25–40.

Kumaran, M., Subramanian, U., Devarajan, B., 2019. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics 20, 342. https://doi.org/10.1186/s12859-019-2928-9

Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R., Zandi, P.P., 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics 8, 14. https://doi.org/10.1186/1479-7364-8-14

Pool, J.E., Hellmann, I., Jensen, J.D., Nielsen, R., 2010. Population genetic inference from genomic sequence variation. Genome Res 20, 291–300. https://doi.org/10.1101/gr.079509.108

Roberts, H.E., Lopopolo, M., Pagnamenta, A.T., Sharma, E., Parkes, D., Lonie, L., Freeman, C., Knight, S.J.L., Lunter, G., Dreau, H., Lockstone, H., Taylor, J.C., Schuh, A., Bowden, R., Buck, D., 2021. Short and long-read genome sequencing methodologies for somatic variant detection; genomic analysis of a patient with diffuse large B-cell lymphoma. Scientific Reports 11, 6408. https://doi.org/10.1038/s41598-021-85354-8

Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K.V., Wang, C., Carter, A.B., 2018. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics 20, 4–27. https://doi.org/10.1016/j.jmoldx.2017.11.003

Schwessinger, B., Rathjen, J.P., 2017. Extraction of High Molecular Weight DNA from Fungal Rust Spores for Long Read Sequencing, in: Periyannan, S. (Ed.), Wheat Rust Diseases: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 49–57. https://doi.org/10.1007/978-1-4939-7249-4_5

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Yuan, Y., Bayer, P.E., Batley, J., Edwards, D., 2017. Improvements in Genomic Technologies: Application to Crop Genomics. Trends in Biotechnology 35, 547–558. https://doi.org/10.1016/j.tibtech.2017.02.009

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., Hovig, E., 2020. Accuracy and efficiency of germline variant calling pipelines for human genome data. Sci Rep 10, 20222. https://doi.org/10.1038/s41598-020-77218-4

Hsieh, F., Turnbull, B.W., 1996. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. The Annals of Statistics 24, 25–40.

Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R., Zandi, P.P., 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics 8, 14. https://doi.org/10.1186/1479-7364-8-14

Pool, J.E., Hellmann, I., Jensen, J.D., Nielsen, R., 2010. Population genetic inference from genomic sequence variation. Genome Res 20, 291–300. https://doi.org/10.1101/gr.079509.108

Roberts, H.E., Lopopolo, M., Pagnamenta, A.T., Sharma, E., Parkes, D., Lonie, L., Freeman, C., Knight, S.J.L., Lunter, G., Dreau, H., Lockstone, H., Taylor, J.C., Schuh, A., Bowden, R., Buck, D., 2021. Short and long-read genome sequencing methodologies for somatic variant detection; genomic analysis of a patient with diffuse large B-cell lymphoma. Scientific Reports 11, 6408. https://doi.org/10.1038/s41598-021-85354-8

Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K.V., Wang, C., Carter, A.B., 2018. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics 20, 4–27. https://doi.org/10.1016/j.jmoldx.2017.11.003

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Yuan, Y., Bayer, P.E., Batley, J., Edwards, D., 2017. Improvements in Genomic Technologies: Application to Crop Genomics. Trends in Biotechnology 35, 547–558. https://doi.org/10.1016/j.tibtech.2017.02.009

Annexe

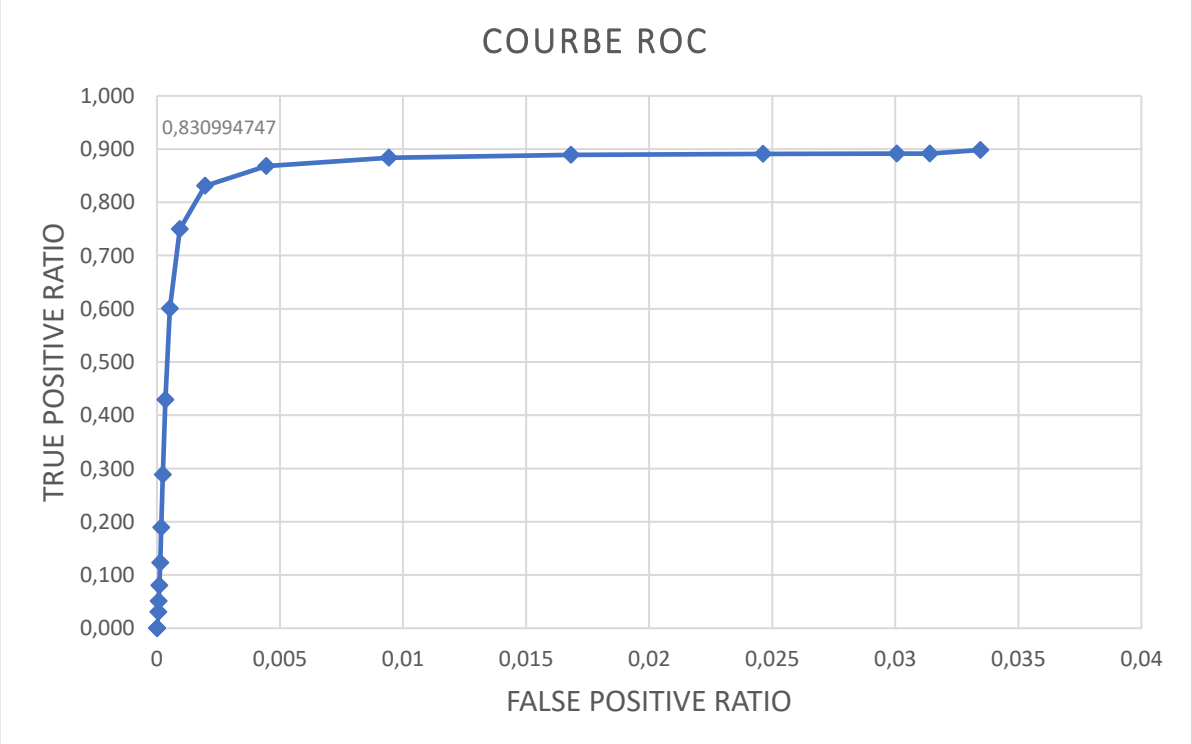
Annexe 01 : Glossaire des termes :

Identification des bases nucleotides **Base calling**	L'identification de base est le processus par lequel l'ordre des nucléotides dans une matrice est déduit lors d'une réaction de séquençage.
Alignment	Le processus de rangement des séquences d'ADN, ARN à une séquence de référence. Le but est d'aligner les lectures ou leurs parties à leur "véritable origine" dans le génome de référence, tout en tolérant un certain nombre de désaccords pour permettre la détection des variations de sous-séquence, en déduisant à la fin leurs positions et les variants survenus. Pour atteindre cet objectif, chaque lecture est alignée sur la ou les régions les plus similaires dans la référence, déterminées par les scores d'alignement (Bao et al., 2011).
Phasing	Le phasage est l'une des étapes d'analyse possibles qui suit la découverte de variants. Ce processus applique des méthodes statistiques sur les données de génotypage (résultats de la découverte des variants) pour inférer l'haplotype de chaque copie homologue d'un chromosome. Cette information sur l'haplotype nous permet d'étudier l'ascendance génétique ou l'histoire démographique, d'imputer les génotypes non observés, de détecter la sélection ou les variants causaux (Delaneau et al., 2013).
QD score **(QualByDepth)**	Le score de qualité relatif à un variant divisé par la profondeur de lecture non filtrée à cette position (AD : allele depth). Pour un seul échantillon, l'HaplotypeCaller calcule le QD en prenant QUAL/AD. Pour des échantillons multiples, HaplotypeCaller et GenotypeGVCFs calculent le QD en prenant QUAL/AD des échantillons avec une découverte de génotype non homozygote aux références (GATK Website).
FisherStrand (FS)	Il s'agit de la probabilité à l'échelle de Phred qu'il y ait un biais de brin sur le site. Le biais de brin nous indique si l'allèle alternatif a été vu plus ou moins souvent sur le brin direct ou inverse que l'allèle de référence. Lorsqu'il n'y a pas ou peu de biais de brin sur le site, la valeur FS sera proche de 0 (GATK Website).
RMSMappingQuality (MQ)	Estimation de la qualité globale d'alignement des lectures soutenant un découverte de variant, en moyenne sur tous les échantillons d'une cohorte (GATK Website).

MappingQualityRankSumTest (MQRankSum)	Compare les qualités de cartographie des lectures soutenant l'allèle de référence avec celles soutenant l'allèle alternatif. Le résultat idéal est une valeur proche de zéro, ce qui indique qu'il y a peu ou pas de différence. Une valeur négative indique que les lectures soutenant l'allèle alternatif ont des scores de qualité de cartographie plus faibles que celles soutenant l'allèle de référence. Inversement, une valeur positive indique que les lectures soutenant l'allèle alternatif ont des scores de qualité de cartographie plus élevés que celles soutenant l'allèle de référence (GATK Website).
ReadPosRankSumTest (ReadPosRankSum)	Teste s'il existe des preuves de biais dans la position des allèles au sein des lectures qui les supportent, entre l'allèle de référence et l'allèle alternatif. Le fait de voir un allèle uniquement près des extrémités des lectures indique une erreur, car c'est là que les séquenceurs ont tendance à faire le plus d'erreurs. Le résultat idéal est une valeur proche de zéro, ce qui indique qu'il y a peu ou pas de différence dans l'emplacement des allèles par rapport aux extrémités des lectures. Une valeur négative indique que l'allèle alternatif est trouvé aux extrémités des lectures plus souvent que l'allèle de référence. Inversement, une valeur positive indique que l'allèle de référence se trouve plus souvent aux extrémités des lectures que l'allèle alternatif (GATK Website).
Roc	RECEIVER OPERATING CHARACTERISTIC: The RECEIVER OPERATING CHARACTERISTIC (ROC) describe the performance of binary classification test used to discriminate between two stats based on a variable measured on a continuous scale. This curve allows summarizing performance over a range of trade-offs between true positive (TP) and false positive (FP) error ratios. It is a plot of sensitivity (the ability of the model to predict an event correctly) versus 1-specificity at various threshold of a continuous variable that allows the classification.(Hsieh and Turnbull, 1996).

Annexe 02 : Résultat de filtrage dans le pipeline GATK Illumina :

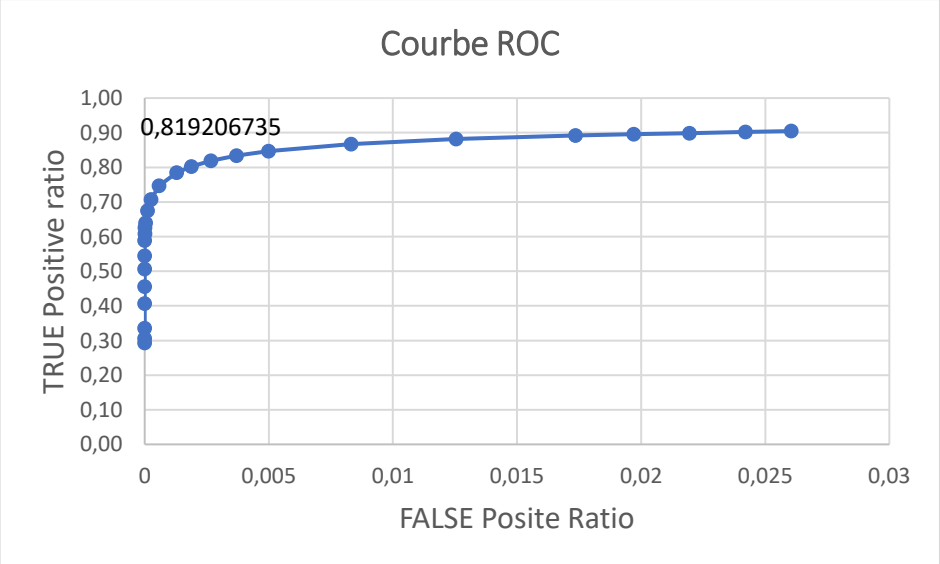
	False positive ratio	True positive ratio	F_{score}
Sans filtrage	0,033460587	0,898	66,215%
QD>1	0,031402251	0,892	67,126%
QD>2	0,030061179	0,892	67,960%
QD>4	0,024627852	0,891	71,547%
QD>6	0,016819313	0,889	77,364%
QD>8	0,009420747	0,883	83,620%
QD>10	0,004446361	0,868	87,828%
QD>12	0,001957003	0,831	88,47%
QD>14	0,000932392	0,749	84,582%
QD>16	0,000530458	0,600	74,425%
QD>18	0,00033903	0,429	59,698%
QD>20	0,000240443	0,289	44,596%
QD>22	0,00017853	0,189	31,705%
QD>24	0,000133318	0,123	21,874%
QD>26	0,00010254	0,080	14,788%
QD>28	7,69241E-05	0,051	9,665%
QD>30	5,13834E-05	0,030	5,844%
QD>35	3,00979E-06	0,000	0,091%
QD>40	1,88701E-08	0,000	0,001%



Courbe Roc selon QD score de pipelines GATK Illumina

Annexe 03 : Résultat de filtrage dans le pipeline BCFtools Illumina :

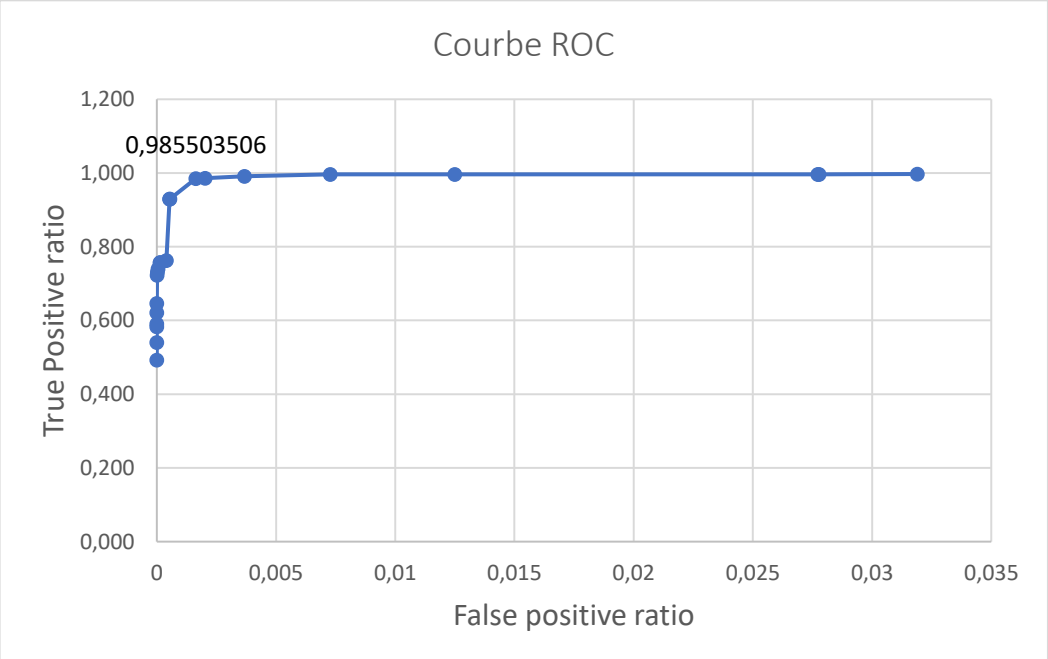
Filtrage	false positive ratio	True positive ratio	F score
aucune	0,02605118	0,90	0,71251504
QUAL>10	0,024204774	0,90	0,72394219
QUAL>20	0,021950387	0,90	0,73864687
QUAL>30	0,019704039	0,90	0,75391214
QUAL>40	0,017362311	0,89	0,77048016
QUAL>60	0,012546859	0,88	0,80615255
QUAL>80	0,008324024	0,87	0,83795015
QUAL>100	0,004986349	0,85	0,86048847
QUAL>110	0,003701123	0,83	0,8668273
QUAL>120	0,002675446	0,82	0,86946981
QUAL>130	0,001880145	0,80	0,86853258
QUAL>140	0,001291067	0,78	0,86431823
QUAL>160	0,000579483	0,75	0,8482792
QUAL>180	0,000257294	0,71	0,82564593
QUAL>200	0,000117938	0,67	0,8044088
QUAL>225	3,51834E-05	0,64	0,77971285
QUAL>250	1,82757E-05	0,63	0,76962553
QUAL>275	8,73687E-06	0,61	0,75661856
QUAL>300	3,94386E-06	0,59	0,74084535
QUAL>350	8,77461E-07	0,54	0,70497055
QUAL>400	3,30227E-07	0,51	0,6726482
QUAL>500	2,83052E-07	0,46	0,62580537
QUAL>600	2,35877E-07	0,41	0,57856352
QUAL>800	1,98136E-07	0,34	0,50228987
QUAL>900	1,50961E-07	0,31	0,46901157
QUAL>950	1,13221E-07	0,29	0,45298165



Courbe Roc selon le score QUAL de pipelines BCFtools Illumina

Annexe 04 : Résultats de filtrage dans le pipeline BCFtools HiFi

	false positive ratio	True positive ratio	F score
sans filtrage	0,031896668	0,997	0,71894856
QUAL>10	0,027772155	0,997	0,74571916
QUAL>20	0,027717743	0,997	0,74596968
QUAL>30	0,012498042	0,996	0,86612011
QUAL>40	0,007274665	0,996	0,91676499
QUAL>60	0,003678734	0,991	0,95273557
QUAL>80	0,002027851	0,986	0,96888954
QUAL>90	0,001636928	0,986	0,97315415
QUAL>100	0,000548121	0,929	0,95677702
QUAL>110	0,00053828	0,929	0,95675376
QUAL>120	0,000401443	0,763	0,86058113
QUAL>130	0,000144272	0,758	0,86037111
QUAL>140	0,000136507	0,757	0,85974766
QUAL>160	5,08645E-05	0,740	0,85024607
QUAL>180	2,15308E-05	0,731	0,84441916
QUAL>200	7,44427E-06	0,723	0,83919651
QUAL>225	2,32103E-06	0,646	0,78517519
QUAL>250	9,34072E-07	0,621	0,76638832
QUAL>275	2,07572E-07	0,590	0,74247724
QUAL>300	9,43507E-08	0,583	0,73692837
QUAL>350	9,43507E-09	0,541	0,70184632
QUAL>400	0	0,492	0,65989939



Courbe Roc selon le score QUAL de pipelines BCFtools HiFi

Résumé

L'objectif de ce stage est de comparer différents pipelines pour l'identification de variants génétiques à partir de séquences des génomes entiers. Cette identification de variants génétiques est importante en biologie, notamment en génétique des populations, car elle nous permet de comprendre les contributions des mutations, de la dérive génétique et de la sélection naturelle à l'évolution des gènes, des génomes et des espèces, ainsi qu'aux changements démographiques. Dans notre étude, nous utilisons des données simulées pour comparer quatre pipelines de la découverte de variants, deux basés sur les lectures Illumina et deux basés sur les longues lectures HiFi, y compris un nouveau pipeline développé par mon superviseur Kiwong Nam. Nous voulons vérifier si les stratégies qui utilisent les longues lectures sont plus performantes.

Jusqu'à présent, nos résultats ont montré que parmi les trois pipelines déjà étudiés, celui utilisant les longues lectures Hifi et BCFtools pour la découverte de variants avait la sensibilité la plus élevée tout en maintenant un faible nombre de FP pour les SNVs. Les résultats ont également montré l'importance de l'étape de filtrage dans les pipelines, notamment sur la base du score Quality By Depth (QD) pour le pipeline Gatk et du score QUAL pour les pipelines BCFtools.

Mots clés : Pipelines de la découverte de variants, données simulées, séquençage du génome entier, variants d'un seul nucléotide, filtrage des variants, lectures Hifi, lectures Illumina.

The goal of this internship is to compare different pipelines for the identification of genetic variants from whole genome sequences. The identification of genetic variants is important in biology, especially in population genetics, because it allows us to understand the contributions of mutations, genetic drift, and natural selection to the evolution of genes, genomes, and species, as well as to demographic changes. In our study, we use simulated data to compare 4 variant calling pipelines, two based on Illumina reads and two based on long HiFi reads, including a new pipeline developed by my supervisor Kiwong Nam. We want to highlight whether the strategies that use the long reads perform better.

So far, our results have shown that among the three pipelines already studied, the one using long Hifi and BCFtools reads for variant calling had the highest sensitivity while maintaining a low number of FPs for SNVs. The results also showed us the importance of the filtering step in the pipelines, especially based on the Quality by Depth (QD) score for the Gatk pipeline and the Qual score for the BCFtools pipelines.

Key words: Variant Calling pipelines, Simulated data, Whole Genome sequencing, SNVs, Variants Filtering, Hifi reads, Illumina reads.