Précision accrue des appels des petits variants germinaux avec la plateforme DRAGEN[MC]

Plusieurs algorithmes d'amélioration de la précision permettent de détecter les petits variants avec une sensibilité et une précision élevées, tout en maintenant les standards de la plateforme DRAGEN en matière de vitesse de traitement

Introduction

Les progrès réalisés en matière de technologie de séquençage nouvelle génération (SNG) font en sorte que le volume de données de séquençage croît de manière exponentielle. Cette croissance a entraîné une demande pour des méthodes analytiques rapides et efficaces, qui maintiennent les standards élevés de précision en matière d'appel des variants. La plateforme Bio-IT DRAGEN (Dynamic Read Analysis for Genomics [Lecture d'analyse dynamique pour la génomique]) d'Illumina procure une analyse secondaire hautement précise et ultrarapide des données de séquençage nouvelle génération. Elle utilise la technologie hautement reconfigurable du réseau prédiffusé programmable par l'utilisateur (FPGA) pour accroître considérablement la rapidité de l'analyse secondaire des données de SNG, y compris le mappage, l'alignement et l'appel des variants.

Les caractéristiques fondamentales de la plateforme DRAGEN permettent de relever les défis courants liés à l'analyse génomique, comme les longs délais de traitement et le grand volume de données. La plateforme DRAGEN allie la rapidité, la souplesse, la précision et la rentabilité. La possibilité de reprogrammer la plateforme permet à Illumina d'améliorer les algorithmes en vue de nouvelles applications de SNG. La rapidité de la plateforme permet aux concepteurs de reproduire rapidement la conception d'un algorithme à l'aide de méthodes de calcul intensif impossibles à réaliser avec des modèles provenant uniquement de logiciels traditionnels. Ainsi, la précision de la plateforme DRAGEN a grandement évolué au fil des nouvelles versions, et DRAGEN offre maintenant une excellente solution pour l'appel de petits variants germinaux dans le séquençage du génome entier.

La présente note d'application décrit les améliorations récentes apportées à la plateforme Bio-IT DRAGEN d'Illumina en matière d'analyse secondaire, et démontre la rapidité et la précision de trois ensembles de données de séquençage du génome entier accessibles au public. Une analyse comparative de la plateforme DRAGEN v3.2.8 et d'autres pipelines, y compris BWA-MEM + GATK4 et DRAGEN v2, sont présentées (figure 1). Les résultats de l'appel des variants de chaque pipeline sont comparés à un ensemble de variants de référence représentatifs de la réalité afin d'identifier les faux positifs (FP) et les faux négatifs (FN). Les indicateurs utilisés pour comparer les pipelines sont la durée complète des analyses, ainsi que des indicateurs de précision tels que le rappel, la précision, les faux positifs et les faux négatifs. La combinaison de la rapidité, de la précision et d'un large éventail d'applications accessibles permet à la plateforme DRAGEN de révolutionner le domaine de l'analyse génomique.

Précision accrue avec les algorithmes de la plateforme DRAGEN v3

La plateforme DRAGEN v3 a recours aux algorithmes les plus récents pour détecter les polymorphismes de nucléotide simple, ainsi que les insertions et les délétions (indels), ce qui procure une meilleure précision et une plus grande sensibilité analytique. Les améliorations ont été apportées à quatre aspects de l'appel des variants : le modèle d'erreur des indels propres à l'échantillon, les modèles mathématiques rigoureux des erreurs d'empilement corrélées, une approche optimisée afin de représenter de manière exhaustive un nombre exponentiel de candidats haplotypes dans les régions riches en variants ou comportant du bruit, une augmentation colonne par colonne de la liste des événements générés par l'assemblage du graphe de De Bruijn. Ces mises à jour ont permis d'obtenir de modestes gains sur le plan de la rapidité, tout en augmentant les standards de précisions, comparativement aux pipelines évalués dans le présent document. L'annexe présente davantage de renseignements sur chacune des améliorations apportées aux algorithmes.

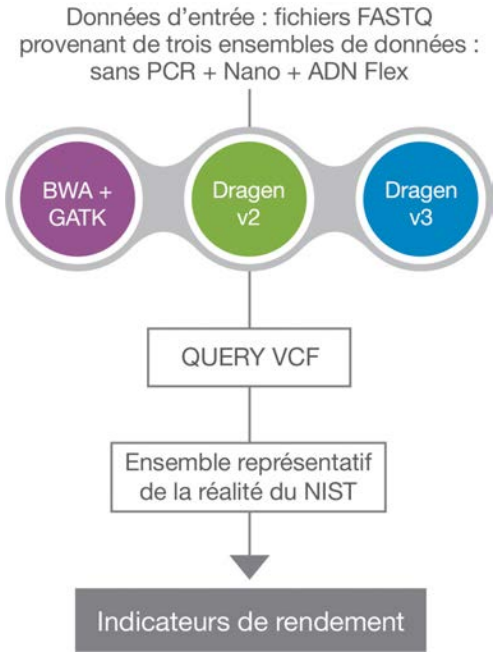

Figure 1 : Élaboration d'une analyse comparative : Des fichiers FASTQ provenant de trois ensembles de données ont été analysés à l'aide de trois pipelines d'analyse pour générer des fichiers de requête VCF. L'outil d'évaluation de l'appel des variants (VCAT) a ensuite été utilisé pour identifier les vrais positifs, les faux positifs et les faux négatifs en fonction d'une comparaison entre les appels des variants et l'ensemble des variants de référence du NIST, représentatifs de la réalité.

Méthodes

Les pratiques exemplaires recommandées en matière d'analyse comparative ont été suivies rigoureusement[1]. Pour démontrer la rapidité et la précision de la plateforme DRAGEN v3, une analyse comparative a été effectuée à l'aide de trois ensembles de données provenant de différentes préparations de librairie, générées à partir de l'échantillon NA12878 (figure 1). En résumé, le fichier FASTQ de chaque ensemble de données a été utilisé à titre de données d'entrée pour l'analyse secondaire des pipelines indépendants (DRAGEN v3.2.8, DRAGEN v2 et BWA + GATK[2]). Les fichiers VCF résultants de chaque pipeline (QUERY VCF) ont été téléversés dans un projet de BaseSpace^{MC} Sequence Hub. L'outil d'évaluation de l'appel des variants (VCAT v3.1.1 avec la version Hap.py 0.3.10) a été utilisé pour comparer chaque fichier QUERY VCF à un variant de référence représentatif de la réalité, dans le but d'identifier les vrais et les faux appels de variants. Les résultats ont été recueillis et tabulés à des fins de comparaison entre les pipelines. Les données d'entrée, les résultats d'analyse et les outils d'évaluation sont tous accessibles gratuitement dans le projet BaseSpace[3]. Davantage de renseignements sur les méthodes sont présentés dans l'annexe.

Résultats des analyses comparatives

Les résultats de la comparaison de la durée et de la précision des analyses démontrent que la plateforme DRAGEN offre une solution puissante pour l'analyse secondaire des données de séquençage nouvelle génération.

Précision de la plateforme DRAGEN : FP, FN, rappel et précision

Bien que la plateforme DRAGEN v2 se classait parmi les chefs de file de l'industrie en matière de solution informatique, plusieurs modifications (voir la section sur les méthodes algorithmiques) permettant d'améliorer significativement la précision de la plateforme DRAGEN v3 ont été apportées. Les résultats de l'analyse comparative ont aussi démontré que les améliorations apportées à la plateforme DRAGEN v3 la rendent supérieure aux autres pipelines, y compris la version antérieure de la plateforme, pour tous les indicateurs de précision analysés dans l'étude.

Lorsque l'indicateur des FP et des FN a été évalué pour la détection des SNV, la plateforme DRAGEN v3 a obtenu un rendement significativement plus élevé en matière de précision que les pipelines BWA + GATK4 et DRAGEN v2, et ce, pour les trois ensembles de données (figure 2). Lorsque l'indicateur des FP et des FN a été évalué pour la détection des indels, la plateforme DRAGEN v3 a obtenu un meilleur rendement que le pipeline BWA + GATK4 et a démontré une amélioration par rapport à la plateforme DRAGEN v2 pour les trois ensembles de données (figure 3).

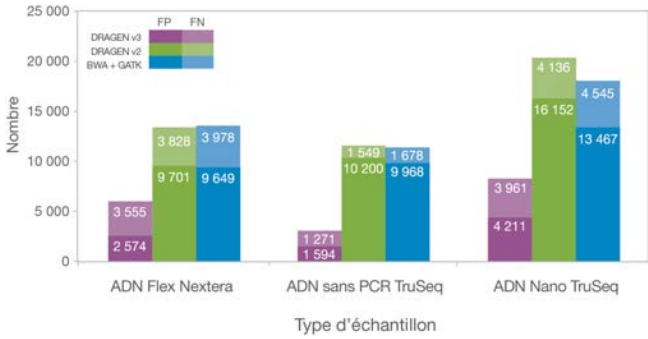

Figure 2 : Faux positifs et faux négatifs avec détection des SNV : Des fichiers de données brutes (FASTQ) provenant de trois ensembles de données ont été analysés à l'aide de trois pipelines indépendants. Chaque ensemble de données (ADN sans PCR TruSeq, ADN Flex Nextera et ADN Nano TruSeq) a été généré à partir de l'échantillon d'ADN NA12878, et l'appel des variants (VCF) de chaque pipeline d'analyse a été comparé à l'ensemble du NIST, représentatif de la réalité (provenant aussi de l'échantillon NA12878), pour identifier les faux positifs et les faux négatifs.

Lors de l'évaluation des indicateurs de précision et de rappel, l'amélioration des algorithmes de la plateforme DRAGEN v3 a été clairement démontrée pour la détection des SNP et des indels. Les valeurs de la précision et du rappel sont constamment au-dessus de 99 % pour chaque pipeline et pour chaque ensemble de données de détection des SNV (Tableau 1). Pour la détection des SNP, la plateforme DRAGEN v2 a offert un rendement comparable à l'application BWA + GATK4. La plateforme DRAGEN v3 a quant à elle démontré une amélioration significative en matière de rappel et de précision pour les autres pipelines. Pour la détection des indels, la plateforme DRAGEN v2 a démontré une plus grande précision que l'application BWA + GATK4, alors que la plateforme DRAGEN v3 a obtenu un rendement accru comparativement à la plateforme DRAGEN v2 pour le rappel et la précision (tableau 2).

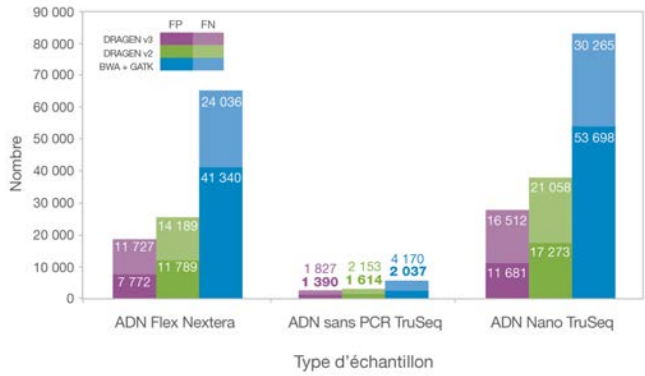

Figure 3 : Faux positifs et faux négatifs avec détection des indels : Des fichiers de données brutes (FASTQ) provenant de trois ensembles de données ont été analysés à l'aide de trois pipelines indépendants. Chaque ensemble de données (ADN sans PCR TruSeq, ADN Flex Nextera et ADN Nano TruSeq) a été généré à partir de l'échantillon d'ADN NA12878, et l'appel des variants (VCF) de chaque pipeline d'analyse a été comparé à l'ensemble du NIST, représentatif de la réalité (provenant aussi de l'échantillon d'ADN NA12878), pour identifier les faux positifs et les faux négatifs.

Tableau 1 : Sensibilité et précision de la détection des SNV

	Précision			Rappel		
Ensembles de données	DRAGEN v3	DRAGEN v2	BWA + GATK	DRAGEN v3	DRAGEN v2	BWA + GATK
ADN sans PCR TruSeq	99,95 %	99,68 %	99,69 %	99,96 %	99,95 %	99,95 %
ADN Flex Nextera	99,92 %	99,70 %	99,70 %	99,89 %	99,88 %	99,88 %
ADN Nano TruSeq	99,87 %	99,50 %	99,58 %	99,88 %	99,87 %	99,86 %

Tableau 2 : Sensibilité et précision de la détection des indels

	Précision			Rappel		
Ensembles de données	DRAGEN v3	DRAGEN v2	BWA + GATK	DRAGEN v3	DRAGEN v2	BWA + GATK
ADN sans PCR TruSeq	99,71 %	99,66 %	99,58 %	99,62 %	99,55 %	99,13 %
ADN Flex Nextera	98,37 %	97,54 %	91,53 %	97,56 %	97,05 %	95,01 %
ADN Nano TruSeq	97,56 %	96,39 %	89,37 %	96,57 %	95,63 %	93,71 %

Rapidité de la plateforme DRAGEN

Les données de comparaison de la durée des analyses de la plateforme DRAGEN ont été recueillies pour les solutions sur site et en nuage. Pour les solutions sur site, la plateforme DRAGEN v3 a été comparée à l'application BWA + GATK avec l'analyse des deux pipelines effectuée sur le même serveur. Pour la solution en nuage, la plateforme DRAGEN v3, utilisée dans BaseSpace Sequence Hub, a été comparée à l'application BWA+ GATK, utilisée dans Terra[4].

La plateforme DRAGEN a permis d'accélérer le processus de mappage et l'appel des variants, lesquels peuvent être analysés de façon indépendante. Bien que la capture n'en ait pas été effectuée, il est intéressant de noter qu'en amont de l'analyse secondaire, la plateforme DRAGEN prend également en charge la conversion des fichiers BCL2FASTQ, ce qui permet d'améliorer considérablement la rapidité et l'efficacité tout en produisant des fichiers FASTQ identiques. Il est aussi intéressant de noter que la plateforme DRAGEN génère automatiquement une liste exhaustive des indicateurs de CQ en matière de mappage et d'appel de variants, en n'augmentant pas, ou très peu, la durée d'analyse. Sur ce point, DRAGEN se démarque des autres pipelines qui s'appuient sur les outils d'analyse lents de fournisseurs tiers (p. ex., Samtools, Picard) pour obtenir des indicateurs de CQ à la durée d'analyse significativement plus élevée.

Lorsque les vitesses d'exécution des pipelines d'analyse exécutés sur le même serveur sur site ont été mesurées, la plateforme DRAGEN v3 s'est montrée significativement plus rapide que l'application BWA + GTAK, avec des gains en rapidité se situant entre 16 et 18× (figure 4).

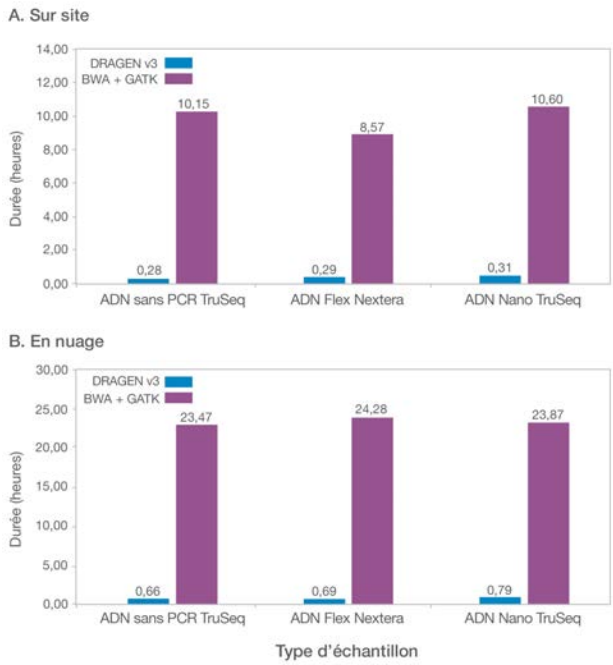

Figure 4 : Durée de l'analyse sur site comparée à la durée de l'analyse en nuage : (A) La plateforme DRAGEN v3 ainsi que l'application BWA + GATK s'exécutent sur le même serveur sur site. (B) La plateforme DRAGEN v3 est utilisée sur BaseSpace Sequence Hub, et l'application BWA + GATK est utilisée sur Terra.

Lorsque les vitesses d'exécution des pipelines d'analyse exécutés en nuage ont été mesurées, la plateforme DRAGEN v3 sur BaseSpace Sequence Hub s'est montrée significativement plus rapide que l'application BWA + GTAK sur Terra, avec des gains en rapidité se situant entre 13 et 16×.

Résumé

Alors que les applications de génomiques s'intéressent à la caractérisation précise des régions difficiles du génomes et à la mesure des appels à basse fréquence allélique à partir d'échantillons à niveau de bruit élevé, la plateforme DRAGEN se démarque, preuve à l'appui, comme étant la plateforme convenant le mieux au traitement des données du séquençage nouvelle génération du futur, tant en matière d'efficacité et que de précision.

La rapidité de la plateforme DRAGEN permet aux chercheurs de gérer l'augmentation de débit des instruments de séquençage nouvelle génération, et, de manière tout aussi importante, permet des itérations rapides offrant une amélioration en continu des algorithmes afin de procurer un niveau élevé de précision.

Annexe

Description détaillée des nouveaux algorithmes

Modèle d'erreur PCR propre à un échantillon

L'un des défis de l'appel des variants est de différencier les erreurs d'indels des vrais variants. Pour cela, les outils d'appel de variants utilisent souvent un modèle de Markov caché (MMC), qui modélise le comportement statistique des erreurs d'indels, dans le cadre de calculs de probabilités. Le MMC comporte habituellement des paramètres de saisie, une pénalité pour la création d'un écart (GOP) et une pénalité pour le maintien d'un écart (GCP), qui sont directement reliés au taux d'erreur des indels (taux d'erreur des indels = f(GOP,GCP)). Les erreurs d'indels sont plus susceptibles de se produire lors de courtes répétitions en tandem (STR), et la probabilité d'une erreur (donc de GOP et GCP) peut varier selon la période et la longueur des STR. Le processus d'erreur peut être significativement différent d'un ensemble de données à un autre, selon des facteurs tels que l'amplification PCR. Pour obtenir une détection précise, il est important d'utiliser des paramètres MMC qui modélisent le processus d'erreur avec précision pour chacun des échantillons. Toutefois, les outils d'appel de variants habituels utilisent souvent des paramètres fixes ou des fonctions prédéterminées qui ne sont pas propres à l'échantillon, ce qui empêche la modélisation précise du processus d'erreur et produit un faible rendement en matière de détection.

Le calibrage automatique du MMC mis en place sur la plateforme DRAGEN v3 résout le problème ci-dessus en estimant les paramètres PCR directement à partir de l'ensemble de données traité. Cette opération est effectuée après le mappage et l'alignement, et avant l'appel des variants, sans connaître la réalité sur le terrain et sans utiliser de bases de données externes de mutations connues. Les paramètres varient selon la période des STR et la longueur des répétitions.

Pour une période et une longueur des STR données, un ensemble de N locus avec la période et la longueur souhaitées est choisi, et les algorithmes vérifient les empilements de lectures mappées à ces locus en dénombrant les indels observés à chaque locus.

L'idée principale est qu'en prenant en compte un nombre suffisant de locus, il est possible d'estimer avec précision les paramètres d'intérêt. Pour ce faire, nous déterminons les paramètres qui maximisent la probabilité de produire l'ensemble de N empilements observés. Si le nombre de paramètres nécessaires à la maximisation de cette probabilité est suffisamment petit, (p. ex., 2), une recherche exhaustive est possible. Dans la présente mise en œuvre de la plateforme DRAGEN v3, une optimisation est effectuée pour deux paramètres : GOP et alpha, lesquels indiquent la probabilité d'obtenir des variants d'indels de toute longueur autre que zéro. Pour chaque période et longueur des STR prises en compte, la recherche produit les valeurs GOP et alpha qui maximisent la probabilité de produire l'ensemble de N empilements observés, et ces valeurs sont utilisées à titre de données d'entrée pour MMC. Il est aussi possible d'étendre la recherche à plus de deux paramètres, ce qui apportera une plus grande amélioration.

Diminution de la qualité des bases (BQD)

Les outils classiques d'appel des variants sont conçus selon l'hypothèse que les erreurs de séquençage sont indépendantes d'une lecture à l'autre. D'après cette hypothèse, il est peu probable que plusieurs erreurs identiques se produisent dans un locus en particulier. Toutefois, après analyse des ensembles de données de séquençage nouvelle génération, il a été observé que des rafales d'erreurs sont beaucoup plus fréquentes que ce que l'hypothèse d'indépendance prédit, et ces rafales d'erreurs peuvent provoquer de nombreux faux positifs.

Heureusement, ces erreurs ont des caractéristiques propres, ce qui permet de les différencier des vrais variants. L'algorithme de la BQD (diminution de la qualité des bases) mis en place dans la plateforme DRAGEN v3 est un mécanisme de détection qui exploite certaines propriétés de ces erreurs (biais lié au brin, emplacement de l'erreur dans la lecture et faible qualité moyenne des bases au locus d'intérêt) et les intègre au calcul de la probabilité d'une manière simple et robuste, dans le génotypeur. De nouvelles hypothèses sur des génotypes candidats sont ajoutées à la liste existante des génotypes diploïdes (celles qui supposent l'indépendance des erreurs d'empilement). Par exemple, dans le cas d'un locus avec un allèle ALT, en plus de prendre en compte P(G00|R), P(G01|R), P(G11|R), les deux hypothèses P(G00,E1|R) et P(G11,E0|R), où les allèles E0 et E1 représentent l'allèle de référence et l'allèle ALT provient d'une erreur de séquençage, s'ajoutent. Les propriétés de ces erreurs, notamment le biais lié au brin, l'emplacement de l'erreur dans la lecture et la moyenne de la qualité des bases, sont prises en compte dans le calcul de P(G00,E1|R) et P(G11,E0|R). Le génotype gagnant est ensuite maximisé (max(P(G00|R), P (G00,E1|R)), P(G01|R), max(P(G11|R), P(G11,E0|R))).

La capacité de caractériser les erreurs de séquençage corrélées au cœur même de l'outil d'appel des variants procure un gain significatif en matière de précision puisque beaucoup d'appels de FP sont retirés. Cela permet aussi une plus grande sensibilité par la correction des erreurs de génotype.

Détection des lectures étrangères (FRD)

Les outils classiques d'appel des variants traitent les erreurs de mappage comme des événements d'erreur indépendants par lecture, ignorant le fait que de telles erreurs surviennent habituellement en rafale. Cela peut produire des appels de variants émis avec un score de confiance élevé malgré un mappage de faible qualité et/ou une fréquence allélique biaisée. Pour atténuer ce problème, les outils classiques d'appel des variants filtrent habituellement en amont les lectures obtenues lors de l'appel des variants, en fonction du seuil de qualité du mappage (les lectures ayant un mappage de qualité inférieure au seuil sont exclues du calcul). Toutefois, cela entraîne le rejet de données importantes obtenues de l'outil d'appel des variants, et les faux positifs ne sont pas éliminés de manière efficace.

La plateforme DRAGEN v3 a mis en place l'outil de détection des lectures étrangères (FRD), une extension de l'algorithme de génotypage existant, en faisant l'hypothèse supplémentaire que certaines lectures de l'empilement sont de nature étrangère (leur emplacement réel se situe ailleurs sur le génome de référence et/ou à l'extérieur du génome de référence [contamination d'échantillon]). L'algorithme exploite de multiples propriétés (fréquence allélique biaisée et mappage de faible qualité) et ajoute ces données au calcul de la probabilité selon une approche mathématique rigoureuse.

De nouvelles hypothèses sur des génotypes candidats sont ajoutées à la liste existante des génotypes diploïdes (celles qui supposent l'indépendance des erreurs d'empilement). Par exemple, dans le cas d'un locus avec un allèle ALT, en plus de prendre en compte P(G00|R), P(G01|R), P(G11|R), les deux hypothèses P(G00,F1|R) et P(G11,F0|R), où les allèles F0 et F1 représentent l'allèle de référence et l'allèle ALT provient d'une erreur de mappage, s'ajoutent. Les propriétés de ces erreurs, notamment la profondeur de l'allèle et la qualité du mappage, sont prises en compte dans le calcul de P(G00,F1|R) et P(G11,F0|R). Le génotype gagnant est ensuite maximisé (max(P(G00|R), P(G00,F1|R)), P(G01|R), max(P(G11|R), P(G11,F0|R))).

La récupération des FN, la correction des génotypes et la diminution du seuil de qualité du mappage pour les saisies de lectures dans l'outil d'appel des variants permettent d'accroître la sensibilité. Le retrait des FP et la correction des génotypes permettent d'accroître la précision.

La détection des lectures étrangères est un outil plus puissant que les approches mettant l'accent sur la filtration après VCF pour améliorer la mesure F, puisque, au lieu de simplement détecter des résultats douteux (p. ex., fondés sur la profondeur allélique ou les erreurs de lecture) après l'appel des variants, l'algorithme de détection intègre directement la présence de lectures étrangères par une détection rigoureuse en fonction du maximum de vraisemblance.

Chemin optimal du MMC et détection colonne par colonne

Les outils d'appel des variants, comme l'outil d'appel d'haplotypes GATK et la plateforme DRAGEN, utilisent le graphe de De Bruijn pour réassembler les lectures afin de déterminer les haplotypes candidats et identifier les sites comportant potentiellement des variants. Dans les régions du génomes comportant des répétitions en tandem, des variants structurels ou des amplifiats d'erreurs de séquençage, l'incapacité de la méthodologie d'assemblage du graphe à fournir une liste complète des haplotypes candidats et des sites comportant des variants peut produire une plus faible sensibilité.

La détection des événements colonne par colonne complémentent le graphe de De Bruijn en balayant chacune des colonnes d'une région active à la recherche de sites comportant potentiellement des variants (SPN et indels) et en complétant la liste des haplotypes candidats. Cela permet de rétablir la sensibilité dans les régions où le graphe a échoué.

Incidence de la détection de lectures étrangères et de la diminution de la qualité des bases sur QUAL/GQ/QD et sur le filtrage rigoureux après VCF

L'outil d'appel des variants DRAGEN v3 utilise deux algorithmes qui modélise les erreurs corrélées dans les lectures dans un empilement donné : l'algorithme de la détection de lectures étrangères pour détecter les lectures incorrectement mappées et celui de la diminution de la qualité des bases pour détecter les erreurs de définition des bases corrélées. En plus d'améliorer la précision et la sensibilité, ces deux algorithmes sont avantageux de deux façons :

Les valeurs du score de confiance (QUAL, GQ, QD) se retrouvent dans un intervalle réaliste de l'échelle Phred.

Les outils classiques d'appel des variants produisent habituellement des valeurs QUAL gonflées se situant dans un intervalle de quelques milliers de valeurs dans l'échelle Phred, ce qui ne fournit aucune information statistique utile. La modélisation des erreurs corrélées à l'aide de l'outil d'appel des variants replace ces valeurs dans un intervalle réaliste et significatif d'un point de vue statistique.

La dépendance sur le plan des règles de filtrage après VCF est considérablement réduite.

En raison de l'incapacité des outils classiques d'appel des variants à différencier les erreurs corrélées des vrais variants, des règles de filtrage rigoureuses après VCF étaient nécessaires lors de leur utilisation pour éliminer le nombre excessif d'appels de FP. Plusieurs annotations VCF (p. ex., QD, MQ, FS, MQRankSum) ont été comparées aux seuils appropriés pour signaler les appels de FP. Ces annotations pouvaient également être ajoutées à un algorithme d'apprentissage automatique et entraînées par rapport à un ensemble représentatif de la réalité, pour permettre le filtrage des faux positifs en fonction de l'entraînement.

Les algorithmes de la plateforme DRAGEN v3 ont été améliorés au cœur même de l'outil d'appel des variants, ce qui a permis de réduire considérablement la dépendance lors du filtrage après VCF. La règle de filtrage rigoureux par défaut de la plateforme DRAGEN v3 utilise simplement la valeur QUAL avec un seuil correspondant à la meilleure mesure F (meilleur compromis entre la sensibilité et la précision).

Méthodes détaillées

Ensembles de données d'entrée

Trois ensembles de données ont été choisis pour représenter les méthodes de préparation de librairies multiples, aussi bien avec que sans PCR (ADN Nano TruSeq, ADN sans PCR TruSeq et ADN Flex Nextera). Chaque ensemble de données a été généré à partir de l'échantillon d'ADN NA12878. Les librairies d'ADN préparées selon le guide de référence approprié[5–7] ont été séquencées en analyses à lectures appariées 2×150 sur le système NovaSeq™ 6000. Pour normaliser le nombre de lectures, chaque ensemble de données a été sous-échantillonné à un niveau de couverture de 30× avec la trousse d'outils FASTQ dans BaseSpace Sequence Hub. Les trois ensembles de données sont accessibles au public dans BaseSpace Sequence Hub, ce qui permet d'effectuer une évaluation indépendante des résultats.

Génome humain de référence

Le génome humain de référence hs37d5 a été utilisé dans l'application DRAGEN sur BaseSpace, et le génome de référence équivalent a été utilisé pour l'analyse locale de chaque pipeline évalué. Cette référence comprend des leurres[8].

Pipelines d'analyse secondaire

Trois pipelines d'analyse secondaire ont été comparés. Le premier est le pipeline DRAGEN v2 complet (DRAGEN est utilisé pour l'étape du mappage et de l'alignement, ainsi que pour l'appel des variants). Le second est le pipeline DRAGEN v3 complet. Le troisième pipeline utilise BWA-MEM pour l'étape du mappage et de l'alignement, et GATK4-HC pour l'appel des variants.

Aux fins de la comparaison, la même règle de filtrage rigoureux pour les trois pipelines a été suivie, ce qui consistait à tenir compte d'un seuil GQ pour les VCF préfiltrés. Le seuil a été établi de façon à ce qu'il soit près du point de la meilleure mesure F pour chaque pipeline (tableau 3).

Tableau 3 : Seuils optimaux de CQ de la mesure F

GQ pour la meilleure mesure F	SNP	Indel
DRAGEN v3.2.8	9	9
DRAGEN v2.5	2	8
BWA + GATK	1	2

La plateforme DRAGEN a été exécutée sur un serveur sur site et en nuage avec BaseSpace Sequence Hub. Bien que le temps de traitement soit légèrement plus long en nuage, les résultats de l'appel des variants sont identiques. Le pipeline BWA + GATK a été exécuté sur le même serveur sur site que celui de la plateforme DRAGEN, sur lequel l'infrastructure BCBIO a été installée[9].

L'infrastructure BCBIO exécute l'application BWA + GATK en fonction des pratiques exemplaires pour l'utilisation de l'outil GATK, et applique des améliorations supplémentaires dans le but d'accroître le parallélisme visant à augmenter la rapidité d'analyse. Pour l'analyse en nuage, le pipeline BWA + GATK a été exécuté sur Terra.

DRAGEN 3.3.0

Version de l'application DRAGEN :
Pipeline germinal DRAGEN v3.2.8
Logiciel hôte DRAGEN v05.011.281.3.2.8

BWA-Mem (0.7.17) + GATK4 (4.0.2)

Tableau 4 : Paramètres du fichier de configuration des algorithmes BCBIO

Paramètre	Valeur
align_split_size	5 000 000
aligner	BWA
coverage_depth	High (Élevée)
coverage_interval	Regional (Régional)
mark_duplicates	True (Vrai)
merge_bamprep	False (Faux)
platform	Illumina
quality_format	Standard
realign	False (Faux)
recalibrate	False (Faux)
tools_off	Vqsr (Réétalonnage du score de qualité du variant)
variantcaller	GATK-haplotype

analyse : variant2
ressources : gatk-haplotype

BWA + GATK sur Terra

Les fichiers d'analyse Bam prêts de l'application BWA-Mem (provenant des analyses BCBIO) ont été utilisés à titre de données d'entrée pour l'exécution de GATK sur Terra. En résumé, le flux de travail GATK4-germline-snps-indels (https://github.com/gatk-workflows/gatk4-germline-snps-indels) a été suivi en tenant compte de modifications apportées à des paramètres en particulier afin qu'ils correspondent aux paramètres des analyses BCBIO. Toutes les analyses ont été effectuées sur Terra à partir d'un compte d'essai gratuit.

La méthode WDL (langage de description du flux de travail) précise est disponible dans les données publiques de BaseSpace Sequence Hub

Configurations de la méthode WDL :
Image Docker de GATK : broadinstitute/gatk : 4.0.2.0
Docker GITC : broadinstitute/genomes-in-the-cloud : 2.3.1-1500064817

Référence, FASTA : hs37d5 (semblable aux autres pipelines)

Uniquement des fichiers VCF bruts ont été générés pour ce pipeline. Le post-filtrage a été effectué localement. Les fichiers VCF bruts sont disponibles dans les données publiques de BaseSpace Sequence Hub.

BaseSpace (janvier 2019) précise les caractéristiques de l'instance AWS F1 utilisée (AWS F1, type 4×)
Version de l'application BaseSpace Sequence Hub : 3.2.8

Tableau 5 : Serveur local (CentOS 7 x86_64, Supermicro 1029)

Composante	Nom complet du modèle	Remarques
Châssis	SYS-1029GQ-TNRT	1 support
Processeur	2 processeurs Intel(MD) Xeon(MD) Gold 6126, 2,60 GHz	24 cœurs, 48 fils
RAM	384 Go	DDR4, 2 666 MHz
Transfert	Intel SSDPE2KE020T7	NVMe, 2 To

Comparaison à un ensemble représentatif de la réalité (NIST)

L'analyse comparative des appels de variants requiert un génome de référence précis et un ensemble d'appels de variants qui représentent les « vraies réponses » pour ce génome. Ce type d'ensemble d'appels de variants a comme particularité qu'il peut être utilisé à titre de données représentatives de la réalité pour identifier avec précision les faux positifs et les faux négatifs. Dans le cadre de cette étude, l'ensemble de données représentatif de la réalité est fondé sur des appels de référence liés à la même source d'ADN (NA12878) établie par le National Institute of Standards and Technology (NIST). Genome in a Bottle (GIAB) est un consortium public, privé et académique, présenté par le NIST. Ce consortium a publié un ensemble de référence de petits variants et des appels de référence pour son génome pilote, NA12878, présentant un génotype de grande fiabilité pour environ 90 % de GRCh37 et GRCh38.

Les vrais positifs (TP) sont des appels de variants qui sont en accord avec les appels de référence de l'ensemble représentatif de la réalité du NIST. Les faux positifs (FP) sont des appels de variants qui n'existent pas dans l'ensemble représentatif de la réalité, et les faux négatifs (FN) sont des variants de l'ensemble représentatif de la réalité qui n'ont pas été appelés dans le fichier QUERY VCF.

L'outil d'évaluation de l'appel des variants (VCAT) a été utilisé pour comparer chaque fichier QUERY VCF à la version 3.3.2 de l'ensemble représentatif de la réalité du NIST. Il exécute hap.py en utilisant le moteur d'évaluation vcfeval RTG. Les TP, FP et FN ont été déterminés par les fichiers de sortie hap.py *roc.Locations.INDEL.csv et *roc.Locations.SNP.csv de TRUTH TP, QUERY.TP, QUERY.FP et TRUTH.FN.

Un niveau de rigueur de concordance de type « concordance du génotype » a été utilisé pour calculer les TP, les FP et les FN (cf. [1]), pour lequel uniquement les sites comportant des allèles et des génotypes correspondants sont pris en compte dans les TP. Cela signifie que les erreurs de génotype et la non-concordance entre les allèles sont prises en compte dans les FP et les FN.

Analyse comparative des indicateurs de l'évaluation

Pour comparer la rapidité, la durée d'analyse totale en secondes, des fichiers FASTQ jusqu'aux fichiers VCF, est issue de l'analyse des fichiers journaux et/ou des durées d'analyse présentées dans les rapports.

Pour comparer la précision des divers pipelines, les standards recommandés en matière d'indicateurs de rendement ont été utilisés (Tableau 6)1. La précision est l'indicateur représentant la précision analytique ou la capacité à identifier adéquatement l'absence de variants ou l'absence de faux positifs. Le rappel est l'indicateur représentant la sensibilité analytique ou la capacité à détecter les variants dont leur présence est connue, ou l'absence de faux négatifs.

Les définitions et les calculs relatifs aux indicateurs de précision et du nombre de rappels sont fondés sur des références.

Tableau 6 : Définitions et calculs relatifs aux indicateurs de précision et de rappel

Indicateur	Nom courant	Définition	Formule
TRUTH.TP	Vrais positifs (réalité)	Nombre d'appels représentatifs de la réalité pour lesquels l'appel provenant de la requête correspond à l'appel représentatif de la réalité et à son génotype.	
QUERY.TP	Vrais positifs (requête)	Nombre d'appels provenant de la requête pour lesquels l'appel représentatif de la réalité correspond à l'appel provenant de la requête et à son génotype.	
TRUTH.FN	Faux négatifs	Nombre d'appels représentatifs de la réalité pour lesquels aucun appel provenant de la requête ne correspond à l'appel représentatif de la réalité ni à son génotype.	
QUERY.FP	Faux positifs	Nombre d'appels provenant de la requête pour lesquels aucun appel représentatif de la réalité ne correspond à l'appel provenant de la requête ni à son génotype.	
METRIC.Recall	Rappel, sensibilité	Fraction des appels représentatifs de la réalité qui correspondent à l'appel d'un allèle et d'un génotype provenant de la requête dans les régions de confiance.	TRUTH.TP / (TRUTH.TP + TRUTH.FN)
METRIC.Precision	Précision, valeur prédictive positive	Fraction des appels provenant de la requête qui correspondent à l'appel d'un allèle et d'un génotype représentatifs de la réalité dans les régions de confiance.	QUERY.TP / (QUERY.TP + QUERY.FP)

Références

1. Krusche P, Trigg L, Boutros PC, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019;37(5):555-560.
2. GATK Best Practices. software.broadinstitute.org/gatk/best-practices/. Consulté le 9 mai 2019.
3. Le projet BaseSpace. basespace.illumina.com/s/3ExEZMIH8Lkq. Consulté le 15 mai 2019.
4. FireCloud propulsé par Terra. firecloud.terra.bio/. Consulté le 15 mai 2019.
5. Illumina (2017) Guide de référence de l'ADN sans PCR TruSeq. Consulté le 6 mars 2019.
6. Illumina (2017) Guide de référence de l'ADN Nano TruSeq. Consulté le 6 mars 2019.
7. Illumina (2018) Guide de référence pour la préparation des librairies ADN Flex Nextera. Consulté le 6 mars 2019.
8. Génome de référence hs37d5. ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/. Consulté le 9 mai 2019.
9. Bcbio-nextgen. Docs. bcbio-nextgen.readthedocs.io/en/latest/. Consulté le 9 mai 2019.

Illumina, Inc. • 1 800 809 4566 (numéro sans frais aux États-Unis) • tél. +1 858 202 4566 • techsupport@illumina.com • www.illumina.com

illumina®