NOTE TECHNIQUE

Améliorations de la précision de l'appel des variants germinaux dans l'analyse secondaire DRAGEN[MC]

Optimisation des performances d'appel des variants grâce à l'apprentissage automatique d'Illumina et à la cartographie multigénomique

illumina[MD]

Introduction

Tirer parti de la puissance du génome via le séquençage de nouvelle génération (SNG) est essentiel à la recherche biomédicale et à la médecine de précision. Pour optimiser les connaissances issues du SNG, les chercheurs ont besoin d'outils d'analyse de données pouvant traduire des données de séquençage brutes en résultats significatifs. L'analyse secondaire DRAGEN fournit une analyse secondaire précise, complète et efficace des données de SNG. L'utilisation de la technologie hautement reconfigurable du réseau prédiffusé programmable par l'utilisateur (FPGA, Field Programmable Gate Array) permet à l'analyse secondaire DRAGEN d'accroître la rapidité de l'analyse secondaire des données de SNG, y compris la cartographie, l'alignement et l'appel des variants. De plus, l'analyse secondaire DRAGEN est conçue pour relever les défis courants liés à l'analyse génomique, tels que les longues durées de traitement, les grands volumes de données et l'appel des variants dans les régions génomiques complexes.

L'analyse secondaire DRAGEN génère des résultats d'une précision exceptionnelle. Lors du 2020 PrecisionFDA Truth Challenge V2 (PrecisionFDA V2), l'analyse secondaire DRAGEN v3.7 a été déclarée la plus précise dans toutes les régions de référence et les régions difficiles à cartographier par rapport à d'autres solutions telles que Sentieon, Seven Bridges et BWA-GATK (figure 1)[1,2]. En seulement quatre ans, l'analyse secondaire DRAGEN v4.3 a apporté des améliorations significatives à cette performance déjà exceptionnelle, offrant une précision sans précédent de l'appel des petits variants avec un score F1 de 99,89 %, une mesure combinée de précision et de rappel, dans toutes les régions de référence avec des fonctionnalités nouvelles et percutantes.

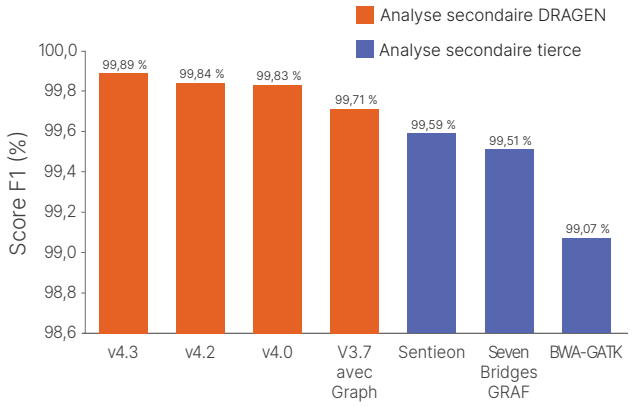

Figure 1 : Précision de l'analyse secondaire DRAGEN pour l'analyse de toutes les régions de référence de la FDA : le score F1 (%) est un calcul des résultats vrais positifs et vrais négatifs par rapport aux résultats totaux[5,6]. Des scores plus élevés indiquent une meilleure précision basée sur les données de référence.

Cette note technique décrit les améliorations récentes qui contribuent à la haute précision de l'analyse secondaire DRAGEN, y compris la fonction de cartographie multigénomique avec référence pangénomique, l'incorporation de l'apprentissage automatique (AA), l'appel des variants mosaïques, les paramètres d'appel spécialisés et la détection des variants structurels (VS) et des variants du nombre de copies (VNC).

Fonction de cartographie multigénomique avec référence pangénomique

La cartographie multigénomique, introduite pour la première fois dans l'analyse secondaire DRAGEN v3.7, permet d'améliorer la précision de l'appel des variants[3]. L'analyse secondaire DRAGEN v4.3 apporte des gains de précision significatifs, avec une réduction de 83 % des erreurs par rapport à la v3.6.3 et une réduction de 40 % des erreurs par rapport à la v4.2.7 (figure 2).

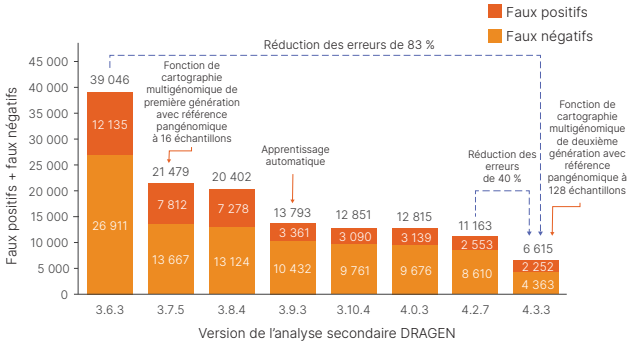

Figure 2 : L'innovation constante au cœur de l'analyse secondaire DRAGEN : les améliorations des taux de faux positifs et négatifs pour les polymorphismes mononucléotidiques et les indels à l'aide de l'échantillon HG002 du projet Genome in a Bottle, NIST v4.2.1[4] démontrent la réduction significative des erreurs qui a été obtenue en seulement quatre ans.

Pour mieux représenter une population spécifique, l'analyse secondaire DRAGEN v4.3 donne aux utilisateurs la possibilité de créer une référence pangénomique personnalisée pour améliorer davantage l'appel des variants dans leurs études. Les utilisateurs peuvent créer une référence pangénomique personnalisée à l'aide de leurs propres assemblages ou d'une sélection d'assemblages fournis par le Human Pangenome Reference Consortium (HPRC). Par exemple, une référence pangénomique personnalisée créée avec 44 assemblages HPRC représentant une population étudiée spécifique génère une plus grande précision de l'appel des variants par rapport aux versions précédentes de l'analyse secondaire DRAGEN, comme l'analyse secondaire DRAGEN v4.2 (figure 3). Cependant, la référence pangénomique par défaut incluse (basée sur 128 échantillons) dans la v4.3 devrait être plus performante pour les cas d'utilisation générale[4].

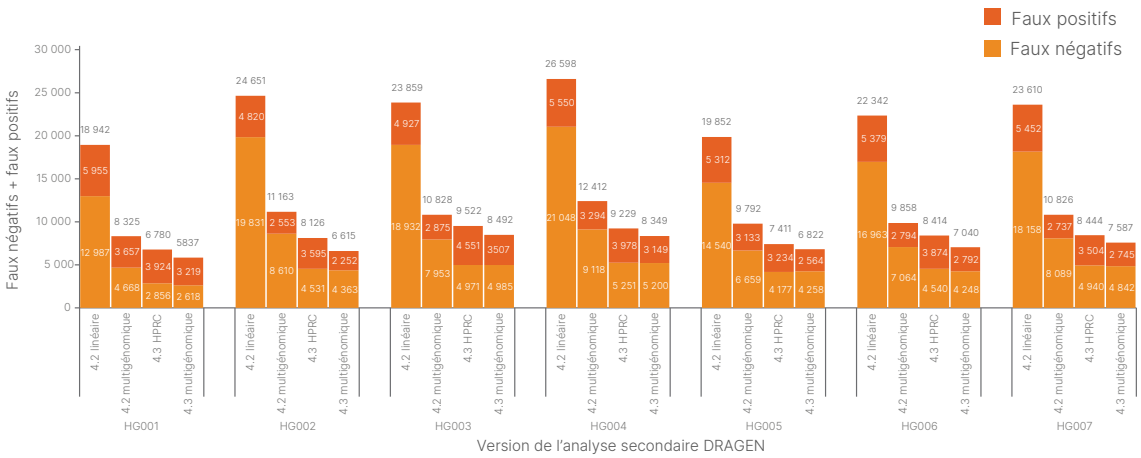

Figure 3 : Améliorations de la précision de l'appel des petits variants dans l'analyse secondaire DRAGEN avec des références personnalisées : la référence multigénomique basée sur l'HPRC de l'analyse secondaire DRAGEN v4.3 donne de meilleurs résultats de précision que la v4.2 lors de l'analyse des échantillons HG001 à HG007 du projet Genome in a Bottle[4]. La référence multigénomique par défaut (4.3 multigénomique), basée sur 128 échantillons, surpasse la référence 4.3 basée sur l'HPRC en utilisation générale.

Apprentissage automatique

Le module d'apprentissage automatique (AA), ajouté pour la première fois dans l'analyse secondaire DRAGEN v3.9 et amélioré dans la v3.10, utilise un modèle supervisé qui utilise des fonctionnalités contextuelles et basées sur les lectures extraites des paramètres d'appel des variants de l'analyse secondaire DRAGEN. La précision des petits variants est améliorée en réduisant les appels erronés grâce à la combinaison de la cartographie multigénomique et de l'AA pour obtenir les meilleurs résultats (figure 4). Des gains substantiels ont été systématiquement démontrés chez tous les sujets, notamment pour les données de test d'autres populations qui n'ont pas été utilisées pendant la formation.

Détection des variants mosaïques

L'analyse secondaire DRAGEN v4.3, prise en charge par un nouveau modèle d'AA, appelle désormais les variants mosaïques au sein du paramètre d'appel des petits variants germinaux. Lorsque le seuil de fréquence allélique est réduit à zéro, l'analyse secondaire DRAGEN peut détecter les variants dont les fréquences alléliques sont < 20 %.

L'analyse secondaire DRAGEN v4.3 détecte les variants mosaïques avec une plus grande exactitude et précision que les versions précédentes.

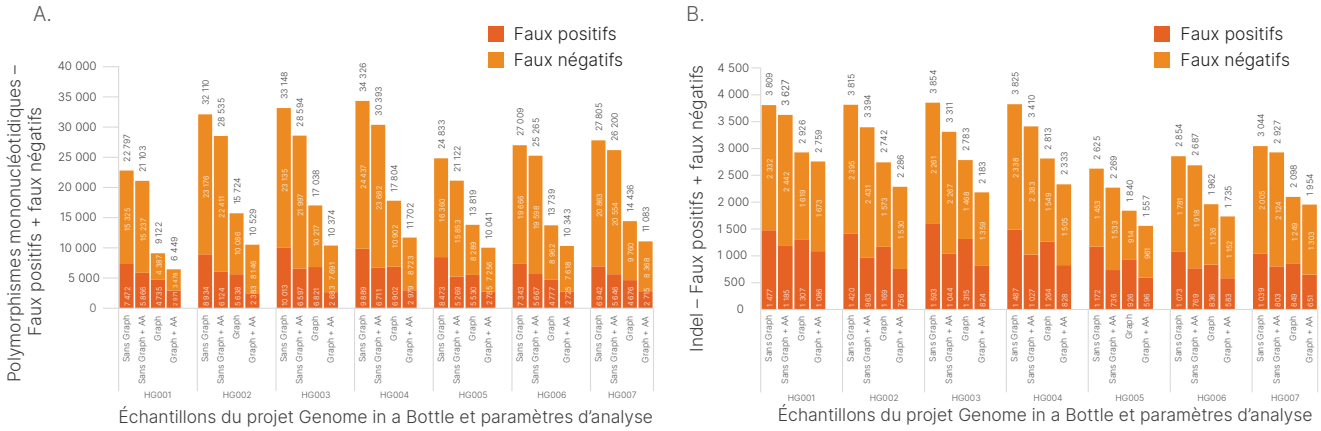

Figure 4 : L'AA et la cartographie multigénomique réduit les faux positifs et les faux négatifs : dans une analyse des échantillons HG001 à HG007 du projet Genome in a Bottle[4], l'AA permet une réduction des erreurs de 10 % avec la référence multigénomique (Graph) désactivée et une réduction des erreurs d'environ 30 % avec la référence multigénomique (Graph) activée. Lorsque la référence multigénomique et l'AA sont activés, les appels erronés sont réduits de 62 % pour les variants mononucléotidiques (A) et les indels (B).

Pour démontrer cela, quatre pipelines d'analyse secondaire DRAGEN ont été testés avec les données de l'ensemble représentatif de la réalité Mosaic du National Institute of Standards and Technology (NIST) : analyse secondaire DRAGEN v4.2, analyse secondaire DRAGEN v4.2 en mode sensibilité élevée (HSM, High-sensitivity Mode), analyse secondaire DRAGEN v4.3 et analyse secondaire DRAGEN v4.3 avec mode mosaïque activé. L'ensemble représentatif de la réalité Mosaic du NIST contient 73 variants mosaïques connus dans des données séquencées à une profondeur de 300×, qui n'ont pas été détectés par l'analyse secondaire DRAGEN v4.2 et v4.3, mais qui ont été détectés par l'analyse secondaire DRAGEN v4.2 en HSM et par l'analyse secondaire DRAGEN v4.3 en mode mosaïque. Cependant, l'analyse secondaire DRAGEN v4.3 en mode mosaïque a permis d'obtenir une plus grande précision de l'appel des variants mosaïques, avec 73 % de faux positifs en moins que l'analyse secondaire DRAGEN v4.2 en HMS (figure 5).

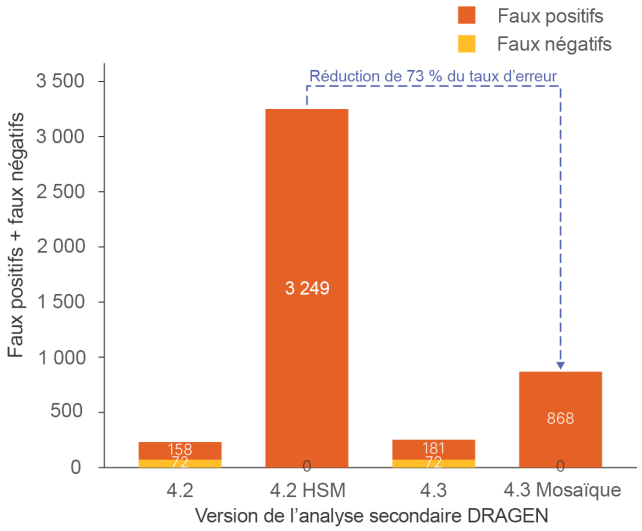

Figure 5 : Amélioration de l'exactitude et de la précision avec le mode de détection des variants mosaïques : une réduction de 73 % des erreurs est observée avec l'analyse secondaire DRAGEN v4.2 en mode sensibilité élevée (HSM) par rapport à DRAGEN v4.3 en mode de détection des variants mosaïques. Les données montrent également le nombre élevé de faux négatifs sans activer le mode HSM ou la détection des variants mosaïques.

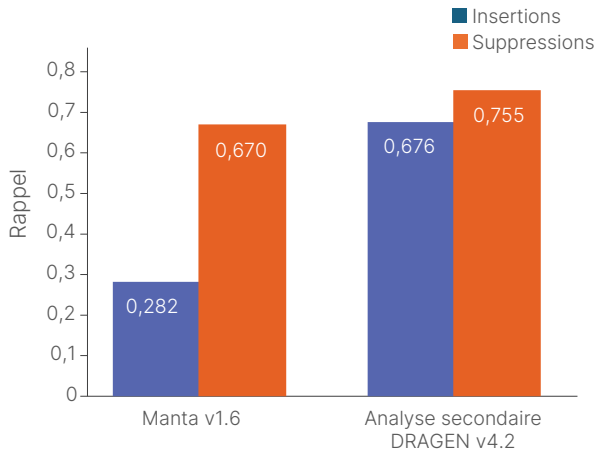

Figure 6 : Appel des VS très précis avec l'analyse secondaire DRAGEN : comparaison du rappel des indels de VS de l'analyse secondaire DRAGEN v4.2 et de Manta v1.6 évaluée avec les données de référence du projet Genome in a Bottle (GIAB SV v0.6)[7].

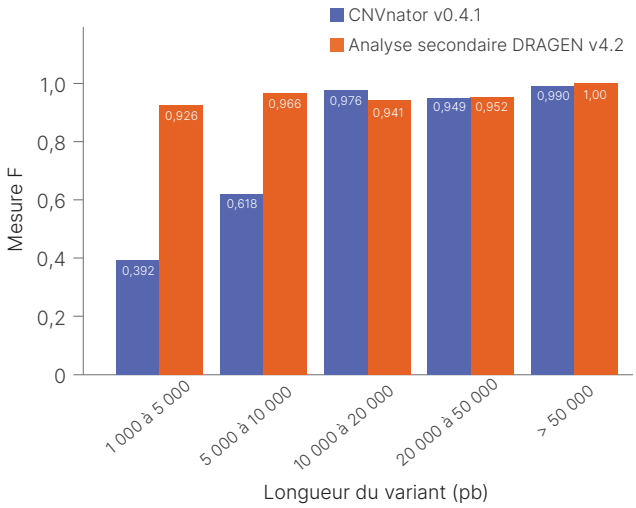

Figure 7 : Appel de VNC très précis avec l'analyse secondaire DRAGEN : appel de VNC par l'analyse secondaire DRAGEN v4.2 par rapport au CNVnator v1.6 pour différentes tailles de suppressions basées sur les données de référence du projet Genome in a Bottle (GIAB SV v0.6)[7].

Détection des VS et des VNC

Les variants structurels (VS) sont des altérations génomiques de 50 pb ou plus et les variants du nombre de copies (VNC) sont un type spécifique de VS où le nombre de copies d'une séquence génomique est réduit (suppressions) ou augmenté (insertions). L'analyse secondaire DRAGEN montre une plus grande précision pour l'appel des VS (figure 6) et l'appel des VNC (figure 7) par rapport aux solutions alternatives[7]. Les algorithmes avancés et les nouvelles approches adaptées aux régions génomiques complexes distinguent l'analyse secondaire DRAGEN des autres solutions.

Le paramètre d'appel des VS DRAGEN améliore les méthodes d'appel des variants structurels Manta et intègre les renseignements de la référence pangénomique, fournissant ainsi un filtrage plus précis et une précision améliorée dans la détection des VS. Il comprend un nouveau détecteur d'insertions d'éléments transposables pour l'identification des grandes insertions, des paramètres de paires optimisés pour un meilleur appel des grandes suppressions et un alignement de contigs affiné pour une découverte d'insertions accrue.

De plus, le logiciel DRAGEN introduit des améliorations dans les étapes d'assemblage, les calculs de probabilités sur la lecture et une meilleure gestion des couples qui se chevauchent et des bases coupées.

Le paramètre d'appel des VNC DRAGEN est principalement basé sur la profondeur de lecture et prend en charge divers modèles de segmentation et de notation pour s'adapter à plusieurs applications. En tirant parti du signal supplémentaire provenant de lectures discordantes et fractionnées, comme c'est le cas dans l'appel des VS, le paramètre d'appel des VNC améliore la sensibilité pour capturer des événements aussi petits que 1 kpb.

Le paramètre d'appel des VNC DRAGEN dispose également d'un module d'extension de duplication segmentaire, une fonctionnalité qui permet la détection des VNC dans les régions de duplication segmentaire du génome. Les régions de duplication segmentaire sont des régions du génome dont la similarité entre les séquences est > 90 %, représentant 5 % du génome. Elles ont une faible cartographiabilité, ce qui rend difficile la détection des variants dans ces régions. L'extension de duplication segmentaire récupère environ un million de bases de régions de VNC qui étaient auparavant exclues de l'analyse. Cela permet la détection des VNC dans 43 gènes pertinents sur le plan médical et améliore la précision globale de l'appel des variants.

Paramètres d'appel spécialisés et ciblés

Les paramètres d'appel ciblés prennent en charge le génotypage précis de gènes spécifiques difficiles à analyser en raison de facteurs tels qu'une similarité de séquence élevée avec les pseudogènes, des régions répétitives et des degrés élevés de polymorphisme. L'analyse secondaire DRAGEN relève ces défis en incorporant divers paramètres d'appel ciblés (tableau 1), permettant un génotypage précis des gènes pertinents sur le plan médical. Pour les renseignements pharmacogénomiques (PGx), le paramètre d'appel des allèles PGx Star appelle les allèles Star et le statut du métaboliseur pour 22 gènes (tableau 2).

Le paramètre d'appel des antigènes leucocytaires humains (HLA, Human Leukocyte Antigen) de DRAGEN permet un génotypage très précis des allèles HLA de classe I et II. Il permet d'aligner les lectures sur une base de données complète de plus de 9 000 allèles et peut aider dans des applications telles que la correspondance des greffes d'organes, l'immunogénétique et les études d'association de maladies.

Tableau 1 : Résumé des gènes couverts par les paramètres d'appel ciblés et spécialisés

Paramètre d'appel ciblé	Champ d'application de la recherche	Association de maladies
CYP21A2	Dépistage des porteurs	Hyperplasie surrénalienne congénitale (CAH, Congenital Adrenal Hyperplasia)
HBA	Dépistage des porteurs	Alpha-thalassémie
GBA	Dépistage des porteurs	Maladie de Gaucher, maladie de Parkinson
SMN	Dépistage des porteurs	Atrophie musculaire rachidienne
LPA	Maladies cardiovasculaires	Coronaropathie
RH	Typage sanguin	–
CYP2B6	PGx	–
CYP2D6	PGx	–
HLA	Correspondance des greffes, immunogénétique	–

Tableau 2 : Gènes présentant une pertinence sur le plan pharmacogénomique couverts par le paramètre d'appel des allèles PGx Star

Symbole du gène		
ABCG2	*CYP4F2*	*RYR1*
BCHE	*DPYD*	*SLCO1B1*
CACNA1S	*F5*	*TPMT*
CFTR	*G6PD*	*UGT1A1*
CYP2C19	*IFNL3*	*UGTB17*
CYP2C9	*MT-RNR1*	*VKORC1*
CYP3A4	*NAT2*	
CYP3A5	*NUDT15*	

L'analyse secondaire DRAGEN v4.3 introduit une nouvelle classe de paramètres d'appel qui permet la détection des variants *de novo* dans les régions avec des duplications segmentaires. Le paramètre d'appel de détection combinée multirégion (MRJD, Multiregion Joint Detection) met en œuvre un paramètre d'appel de petits variants *de novo* basé sur l'haplotype pour six gènes pertinents sur le plan médical dans les régions de duplication segmentaire (tableau 3).

Tableau 3 : Gènes couverts par le paramètre d'appel MJRD

Paramètre d'appel ciblé	Champ d'application de la recherche	Association de maladies
PMS2	Dépistage des cancers héréditaires	Syndrome de Lynch pour le cancer colorectal/ de l'endomètre
SMN1 (petits variants)	Dépistage des porteurs	Atrophie musculaire rachidienne
STRC	Dépistage des porteurs	Perte auditive non syndromique
NEB	Dépistage des porteurs	Myopathie à némaline
TTN	Dépistage des nouveau-nés, maladies rares	Cardiomyopathie
IKBKG	Dépistage des nouveau-nés	Incontinentia pigmenti, dysplasie ectodermique hypohidrotique

Résumé

L'analyse secondaire DRAGEN fournit une analyse secondaire très précise, complète et efficace pour les applications de SNG. Les améliorations continues offrent une précision accrue et une couverture étendue des régions difficiles du génome permettant la détection de variants complexes et pertinents sur le plan médical.

Annexe

Cartographie multigénomique avec référence pangénomique

En utilisant des haplotypes de population de variants mis en phase et en augmentant l'index de référence avec des contigs ALT dérivés de la population, l'analyse secondaire DRAGEN peut efficacement établir une cartographie par rapport à une référence pangénomique et améliorer la cartographie des lectures d'Illumina dans les régions difficiles. Cette nouvelle fonctionnalité étend efficacement la portée des lectures d'Illumina et permet une cartographie et un appel des variants précis dans les régions qui n'étaient pas accessibles auparavant.

Une fonction de cartographie multigénomique est une approche permettant de faciliter la cartographie avec des données de population où le contenu de séquence alternatif, observé dans la population, est représenté sous forme de divers chemins divergents et convergents (figure 8). Les lectures d'échantillons peuvent être alignées sur n'importe quel chemin correspondant le mieux par l'entremise de la fonction de cartographie multigénomique.

En savoir plus, The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3.

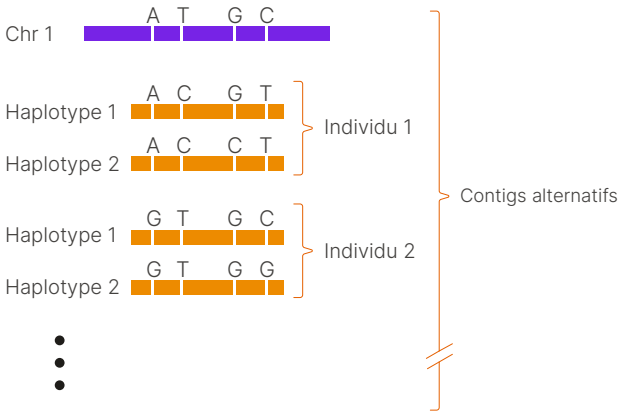

Figure 8 : Fonction de cartographie multigénomique avec référence pangénomique : dans une référence, le contenu de séquence alternatif enregistré dans une population est représenté sous forme de divers chemins divergents et convergents.

Masquage alternatif

Depuis la mise à jour du logiciel d'analyse secondaire DRAGEN v3.9, le logiciel DRAGEN comprend le masquage alternatif, une nouvelle approche pour gérer les contigs ALT de référence natifs, dans laquelle les positions stratégiques des contigs ALT sont masquées pour augmenter la précision. Cette approche est simple à définir, à gérer et à affiner au fil du temps.

En savoir plus, DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph

Apprentissage automatique

Le logiciel d'analyse secondaire DRAGEN v3.9 dispose désormais d'un pipeline de réétalonnage de l'AA puissant et efficace en option dans le flux de travail des petits variants germinaux. Il est activé par défaut dans le logiciel d'analyse secondaire DRAGEN v4.0. Lorsqu'il est activé, le pipeline exécute le modèle d'AA après l'appel de variants standard. Cette étape permet de réétalonner les champs QUAL et GQ qui sont générés dans le fichier VCF final. Dans certains cas, l'AA peut modifier le champ GT. Les valeurs avant l'utilisation de l'apprentissage automatique de ces champs sont conservées dans les champs DQUAL, DGT et DGQ pour ne perdre aucun renseignement. Cette étape ajoute environ cinq minutes au flux de travail standard pour une analyse germinale d'un séquençage du génome entier (WGS, Whole-genome Sequencing) à 30×, de sorte que les améliorations de précision ont un impact limité sur la durée totale de l'analyse.

Le modèle d'apprentissage automatique est généré à l'aide d'une formation hors ligne avec suivi. Le modèle traite un ensemble de fonctionnalités contextuelles et basées sur les lectures pour affiner la précision des scores de qualité du paramètre d'appel des petits variants. Les fonctionnalités utilisées pour entraîner le modèle comprennent la cartographiabilité, la fréquence allélique (FA), la qualité de l'appel des variants, la profondeur, la teneur en GC, les mésappariements et d'autres indicateurs internes de cartographie, d'alignement et d'appel des variants.

Calcul du score F1

$$F1 = 2 \times (Rappel \times Précision)/(Rappel + Précision)$$

$$F1_{parents} = \sqrt{F1_{HG003} \times F1_{HG004}}$$

Ligne de commande DRAGEN

Trouvez des formules de démarrage sur Formule DRAGEN – WGS des variants germinaux

Références

1. Food and Drug Administration. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. precision.fda.gov/challenges/10/results. Consulté le 19 septembre 2024.
2. Illumina. DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph. illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. Publié le 12 janvier 2022. Consulté le 19 septembre 2024.
3. Illumina. The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3. illumina.com/science/genomics-research/articles/second-gen-multigenome-mapping.html. Publié le 12 août 2024. Consulté le 30 septembre 2024.
4. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025. Publié le 7 juin 2016. doi:10.1038/sdata.2016.25
5. Illumina. DRAGEN wins at PrecisionFDA Truth Challenge V2 showcase accuracy gains from alt-aware mapping and graph reference genomes. illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html. Consulté le 19 septembre 2024.
6. Données internes. Illumina, Inc. 2022.
7. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. Publié en ligne le 25 octobre 2024. doi:10.1038/s41587-024-02382-1

Numéro sans frais aux États-Unis : + (1) 800 809-4566 | Téléphone : + (1) 858 202-4566
techsupport@illumina.com | www.illumina.com

M-GL-01016 FRA v3.0