NOTA TÉCNICA

Mejoras en la precisión de la llamada de variantes germinales en el análisis secundario de DRAGEN™

Optimización del rendimiento de la llamada de variantes con el aprendizaje automático y la asignación multigenómica de Illumina

illumina®

Introducción

Para la investigación y la medicina de precisión, es fundamental liberar el potencial del genoma mediante la secuenciación de nueva generación (NGS, next-generation sequencing). Con el fin de obtener el máximo de conocimientos a través de la NGS, los investigadores precisan herramientas de análisis de datos que permitan una traducción de los datos de secuenciación sin procesar en resultados reveladores. DRAGEN ofrece un análisis secundario preciso, completo y eficiente de los datos de NGS. El uso de la tecnología de array de puertas programables por campo (FPGA, field-programmable gate array) altamente reconfigurable, permite acelerar drásticamente el análisis secundario de DRAGEN de datos de NGS, en el que se incluyen las tareas de asignación, alineación y llamada de variantes. Además, el análisis secundario de DRAGEN está diseñado para abordar las complejidades comunes del análisis genómico, como tiempos de computación prolongados, volúmenes masivos de datos y llamada de variantes en regiones genómicas complejas.

El análisis secundario DRAGEN genera resultados excepcionalmente precisos. En el Precision FDA Truth Challenge V2 (PrecisionFDA V2) de 2020, el análisis secundario de DRAGEN v3.7 obtuvo la mayor precisión en todas las regiones de control y regiones difíciles de asignar frente a otras soluciones como Sentieon, Seven Bridges y BWA-GATK (figura 1).[1,2] En solo cuatro años, se han conseguido mejoras significativas en este rendimiento ya excepcional del análisis secundario de DRAGEN v4.3, proporcionando una precisión sin precedentes en la llamada de variantes pequeñas con una puntuación F1 del 99,89 %, una medida combinada de precisión y retirada, en todas las regiones de control con características nuevas e influyentes.

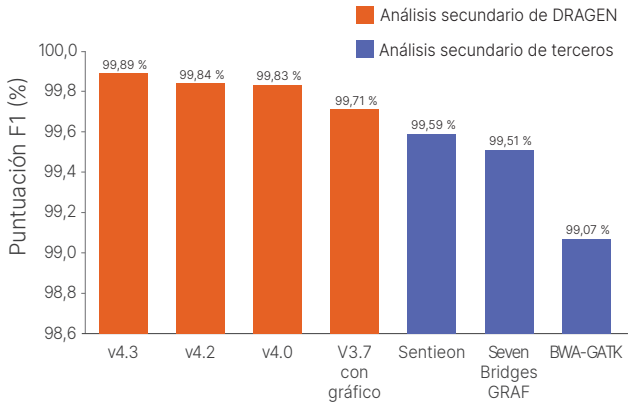

Figura 1: Precisión del análisis de DRAGEN para todas las regiones de control de la FDA: la puntuación F1 (%) es un cálculo de los resultados positivos verdaderos y negativos verdaderos como una proporción de los resultados totales.[5,6] Las puntuaciones más altas indican una mayor precisión según se ha establecido con los datos de control.

Esta nota técnica describe las mejoras recientes que contribuyen a la alta precisión del análisis secundario de DRAGEN, incluido el asignador multigenómico con control del pangenoma, la incorporación del aprendizaje automático (ML, machine learning), la llamada de variantes en mosaico, llamadores especializados, y la detección de variantes estructurales (SV, estructural variant) y variantes en el número de copias (CNV, copy number variant).

Asignador multigenómico con control del pangenoma

La asignación multigenómica, introducida por primera vez en el análisis secundario de DRAGEN v3.7, permite una mayor precisión de la llamada de variantes.[3] El análisis secundario de DRAGEN v4.3 aporta ganancias significativas en términos de precisión, con una reducción de los errores del 83 % en comparación con v3.6.3 y una reducción de los errores del 40 % en comparación con v4.2.7 (figura 2).

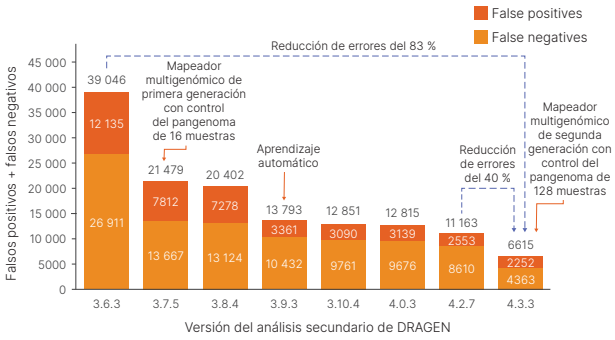

Figura 2: Innovación constante que impulsa el análisis secundario de DRAGEN: las mejoras en las tasas de falsos positivos y negativos para SNP e indels utilizando la muestra HG002 de Genome in a Bottle, NIST v4.2.1,[4] demuestran la reducción significativa de errores que se ha logrado en solo cuatro años.

Para una mejor representación de una población específica, el análisis secundario de DRAGEN v4.3 ofrece a los usuarios la opción de crear un control del pangenoma personalizado, mejorando aún más la llamada de variantes dentro de sus estudios. Los usuarios pueden crear un control del pangenoma personalizado utilizando sus propios conjuntos o utilizando una selección de conjuntos proporcionados por el Human Pangenome Reference Consortium (HPRC). Por ejemplo, un control del pangenoma personalizado, creado con 44 conjuntos del HPRC que representan una población de investigación específica, produce una mayor precisión de llamada de variantes en comparación con versiones anteriores del análisis secundario de DRAGEN, como la versión 4.2 (figura 3). Sin embargo, el control del pangenoma predeterminado (basado en 128 muestras) incluido en la v4.3 debe funcionar mejor para casos de uso general.[4]

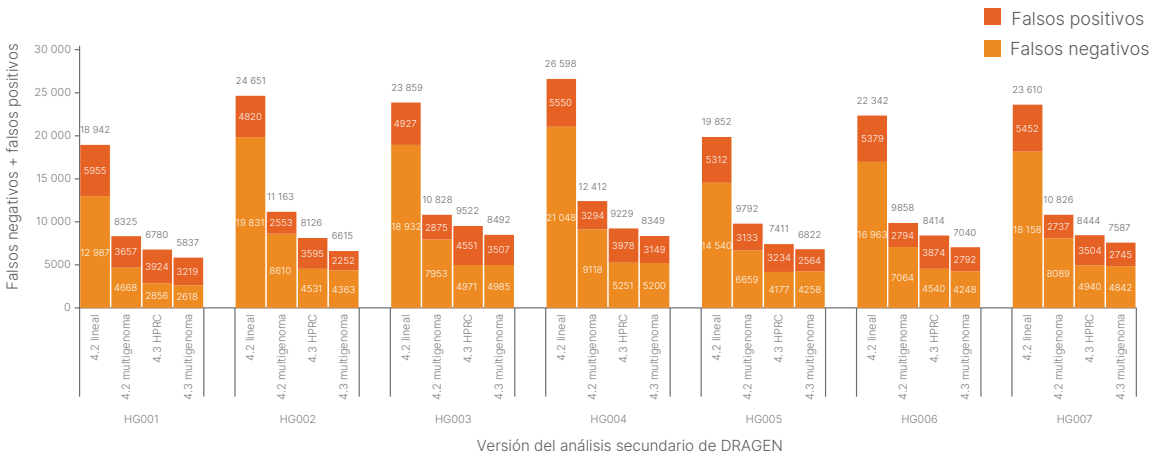

Figura 3: Mejoras en la precisión de la llamada de variantes pequeñas en el análisis secundario de DRAGEN con controles personalizados: el control del multigenoma de DRAGEN v4.3 procedente del HPRC produce mejores resultados de precisión que la v4.2 cuando se analizan las muestras HG001-HG007 de Genome in a Bottle.[4] El control del multigenoma predeterminado (multigenoma de v4.3), formado por 128 muestras, supera al control del HPRC de esta misma versión 4.3 en el uso general.

Aprendizaje automático

El módulo ML, añadido por primera vez en el análisis secundario de DRAGEN v3.9 y mejorado en la v3.10, emplea un modelo supervisado que utiliza funciones contextuales y correspondientes a lecturas extraídas de los llamadores de variantes del análisis secundario de DRAGEN. La precisión de las variantes pequeñas se mejora reduciendo las llamadas falsas con la combinación de la asignación multigenómica y el ML para ofrecer los mejores resultados (figura 4). Se demostraron ganancias considerables de forma sistemática en todos los sujetos, incluidos los datos de pruebas de otras poblaciones que no se utilizaron durante la formación.

Detección de variantes en mosaico

El actual análisis secundario de DRAGEN v4.3, respaldado por un nuevo modelo de ML, permite la llamada de variantes en mosaico dentro del llamador de variantes pequeñas de la línea germinal. Con el umbral de frecuencia de alelos reducido a cero, el análisis secundario de DRAGEN puede detectar variantes con frecuencias de alelos por debajo del 20 %.

El análisis secundario de DRAGEN v4.3 detecta variantes en mosaico con mayor exactitud y precisión que las versiones anteriores.

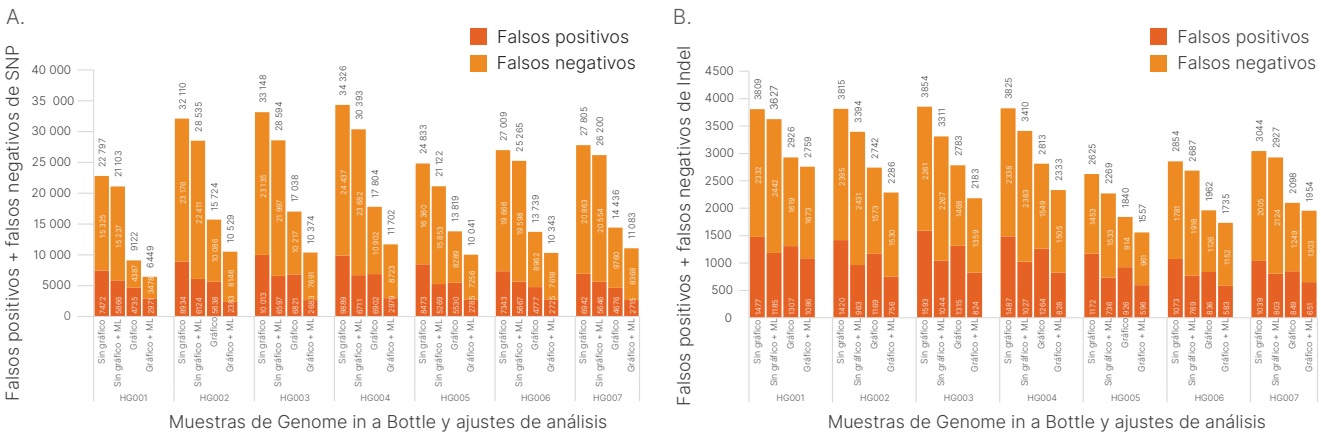

Figura 4: El ML y la asignación multigenoma reducen los falsos positivos y los falsos negativos: en un análisis de las muestras HG001-HG007 de Genome in a Bottle,[4] el ML produce una reducción del error del 10 % con el control del multigenoma (gráfico) desactivado y una reducción del error de aproximadamente un 30 % con el control del multigenoma (gráfico) activado. Cuando están habilitados tanto el control del multigenoma como el ML, las llamadas falsas se reducen en un 62 % para (A) las variantes de nucleótido único y (B) las indel.

Para demostrarlo, se probaron cuatro procesos de análisis secundario de DRAGEN en los datos del conjunto de verdad de mosaicos del National Institute of Standards and Technology (NIST): análisis secundario de DRAGEN v4.2, análisis secundario de DRAGEN v4.2 en modo de alta sensibilidad (HSM, high-sensitivity mode), análisis secundario de DRAGEN v4.3 y análisis secundario de DRAGEN v4.3 con el modo Mosaico habilitado. El conjunto de verdad de mosaicos del NIST contiene 73 variantes de mosaico conocidas en datos con una cobertura de 300×, que no se detectaron mediante el análisis secundario de DRAGEN v4.2 y v4.3, pero se detectaron mediante el análisis secundario de DRAGEN v4.2 en HSM y mediante el análisis secundario de DRAGEN v4.3 en modo Mosaico. Sin embargo, el análisis secundario de DRAGEN v4.3 en modo Mosaico logró una mayor precisión en la llamada de variantes de mosaico, con un 73 % menos de falsos positivos que el análisis secundario de DRAGEN v4.2 en HSM (figura 5).

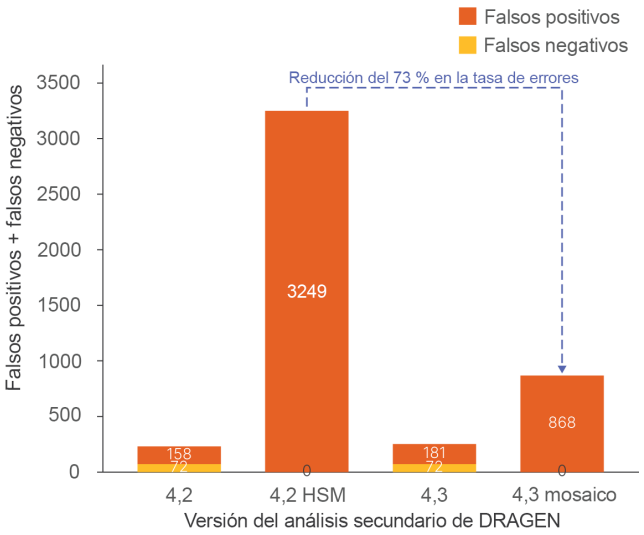

Figura 5: Precisión y exactitud mejoradas con el modo de detección de mosaicos: con el análisis secundario de DRAGEN v4.2 en modo de alta sensibilidad (HSM) se consigue una reducción del error del 73 % en comparación con DRAGEN v4.3 en modo de detección de mosaicos. Los datos también muestran el elevado número de falsos negativos sin el modo HSM o la detección de mosaico activados.

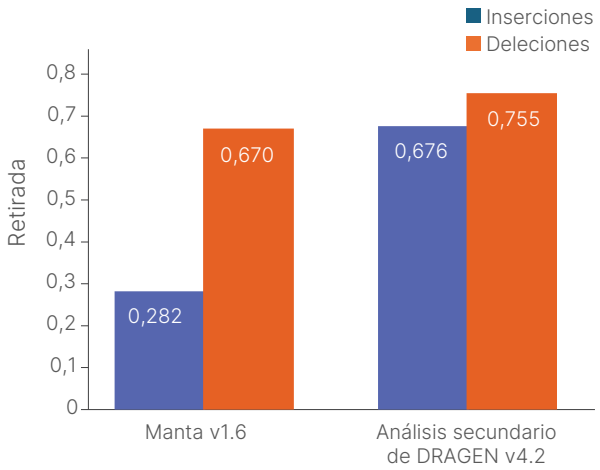

Figura 6: Llamada de SV de alta precisión con el análisis secundario de DRAGEN: comparación de la retirada de SV de tipo indel entre el análisis secundario de DRAGEN v4.2 y Manta v1.6 evaluado con los datos de control de Genome in a Bottle (GIAB SV v0.6).[7]

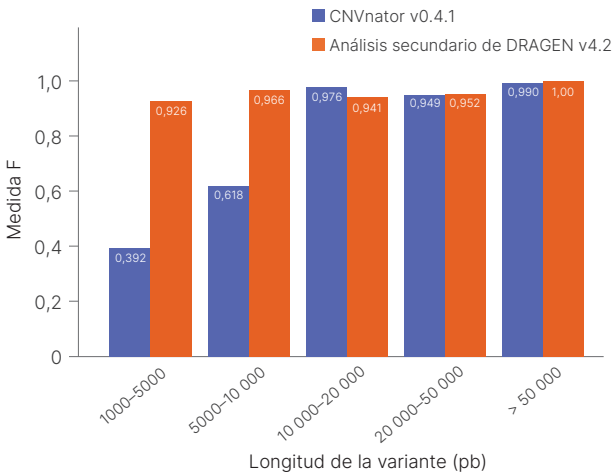

Figura 7: Llamada de CNV de alta precisión con el análisis secundario de DRAGEN: llamada de CNV mediante el análisis secundario de DRAGEN v4.2 en comparación con CNVnator v1.6 en diferentes tamaños de deleciones presentes en los datos de control de Genome in a Bottle (GIAB SV v0.6).[7]

Detección de SV y CNV

Las variantes estructurales (SV) son alteraciones genómicas de 50 pb o más y las variantes en el número de copias (CNV) son un tipo específico de SV en el que el número de copias de una secuencia genómica se reduce (deleciones) o aumenta (inserciones). El análisis secundario de DRAGEN muestra una mayor precisión para la llamada de SV (figura 6) y la llamada de CNV (figura 7) en comparación con soluciones alternativas.[7] Los algoritmos avanzados y los enfoques novedosos adaptados a regiones genómicas complejas diferencian el análisis secundario de DRAGEN de otras soluciones.

El llamador de SV de DRAGEN mejora los métodos de llamada de variantes estructurales de Manta e incorpora información del control del pangenoma, lo que conduce a un filtrado más preciso y a una mayor precisión en la detección de SV. Esto incluye un nuevo detector de inserción de elementos móviles para identificar inserciones grandes, parámetros de pares optimizados para llamadas de deleciones grandes y una alineación de cóntigos refinada para un mejor descubrimiento de inserciones.

Además, el software DRAGEN introduce mejoras en los pasos de ensamblaje, los cálculos de probabilidad de lectura y la manipulación mejorada de emparejamientos superpuestos y bases recortadas.

El llamador de CNV de DRAGEN es principalmente un llamador que emplea la profundidad de lectura, así como diversos modelos de segmentación y puntuación para adaptarse a múltiples aplicaciones. Al aprovechar la señal adicional de lecturas discordantes y divididas, como se hace en las llamadas de SV, el llamador de CNV mejora la sensibilidad para captar eventos de tan solo 1 kpb.

El llamador de CNV de DRAGEN también tiene un módulo de extensión de duplicación segmentaria, una función que permite la detección de CNV en regiones de duplicación segmentaria del genoma. Las regiones de duplicación segmentaria son regiones del genoma con una similitud de secuencia superior al 90 %, lo que representa el 5 % del genoma. La asignación en estas regiones es escasa, lo que dificulta la detección de variantes en ellas. La extensión de duplicación segmentaria recupera aproximadamente un millón de bases de regiones de CNV que previamente se habían excluido del análisis. Esto permite la detección de CNV en 43 genes médicamente relevantes y mejora la precisión general de la llamada de variantes.

Llamadores especializados y selectivos

Los llamadores selectivos favorecen el genotipado preciso de genes específicos que son difíciles de analizar debido a factores como la alta similitud de secuencias con pseudogenes, regiones repetitivas y altos grados de polimorfismo. El análisis secundario de DRAGEN aborda estos desafíos incorporando varios llamadores selectivos (tabla 1), lo que permite un genotipado preciso de genes médicamente relevantes. Para obtener información sobre la farmacogenómica (PGx), PGx Star Allele Caller llama a los subalelos y al estado del metabolizador para 22 genes (tabla 2).

El llamador de antígenos leucocitarios humanos (HLA, human leukocyte antigens) de DRAGEN permite un genotipado muy preciso de los alelos de HLA de clase I y II. Este alinea las lecturas con una completa base de datos de más de 9000 alelos y puede ser de ayuda en aplicaciones como el emparejamiento de trasplantes de órganos, la inmunogenética y los estudios de asociación de enfermedades.

Tabla 1: Resumen de genes abordados por llamadores específicos y especializados.

Llamador selectivo	Área de aplicación en investigación	Asociación de enfermedades
CYP21A2	Detección de portadores	Hiperplasia suprarrenal congénita
HBA	Detección de portadores	α-talasemia
GBA	Detección de portadores	Enfermedad de Gaucher, enfermedad de Parkinson
SMN	Detección de portadores	Atrofia muscular espinal
LPA	Enfermedades cardiovasculares	Arteriopatía coronaria
RH	Determinación del grupo sanguíneo	–
CYP2B6	PGx	–
CYP2D6	PGx	–
HLA	Emparejamiento de trasplantes, inmunogenética	–

Tabla 2: Genes con relevancia para PGx abordados por PGx Star Allele Caller

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

El análisis secundario de DRAGEN v4.3 introduce una nueva clase de llamadores que permite la detección de variantes *de novo* en regiones con duplicaciones segmentarias. El llamador de detección conjunta multirregional (MRJD, multiregion joint detection) implementa un llamador de variantes pequeñas *de novo* según el haplotipo para seis genes médicamente relevantes en regiones de duplicación segmentaria (tabla 3).

Tabla 3: Genes abordados por el llamador de MJRD

Llamador selectivo	Área de aplicación en investigación	Asociación de enfermedades
PMS2	Cribado del cáncer hereditario	Síndrome de Lynch para el cáncer colorrectal/ endometrial
SMN1 (variantes pequeñas)	Detección de portadores	Atrofia muscular espinal
STRC	Detección de portadores	Hipoacusia no sindrómica
NEB	Detección de portadores	Miopatía nemalínica
TTN	Cribado en recién nacidos, enfermedades raras	Miocardiopatía
IKBKG	Cribado en recién nacidos	Incontinencia pigmentaria, displasia ectodérmica hipohidrótica

Resumen

El análisis secundario de DRAGEN ofrece un análisis secundario muy preciso, completo y eficiente para aplicaciones de NGS. Las mejoras continuas proporcionan una mayor precisión y una cobertura ampliada de las regiones difíciles del genoma, lo que permite la detección de variantes complejas y médicamente relevantes.

Apéndice

Asignación multigenómica con control del pangenoma

Mediante el uso de haplotipos poblacionales de variantes de fase de hebra retrasada y el aumento del índice de control con cóntigos alternativos derivados de la población, el análisis secundario de DRAGEN puede hacer asignaciones eficaces en contraste con un control del pangenoma y mejorar la asignación de lecturas de Illumina en regiones difíciles. Esta nueva función amplía de forma eficaz el alcance de las lecturas de Illumina, además de permitir una asignación precisa y la llamada de variantes en regiones a las que antes no se podía acceder.

El enfoque centrado en un asignador multigenómico ayuda a la asignación con datos poblacionales en los que el contenido de secuencias alternativas, observadas en la población, se representa como varias rutas divergentes y convergentes (figura 8). Las lecturas de muestras se pueden alinear con cualquiera de las rutas que mejor coincida a través del asignador multigenómico.

Información adicional, 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.

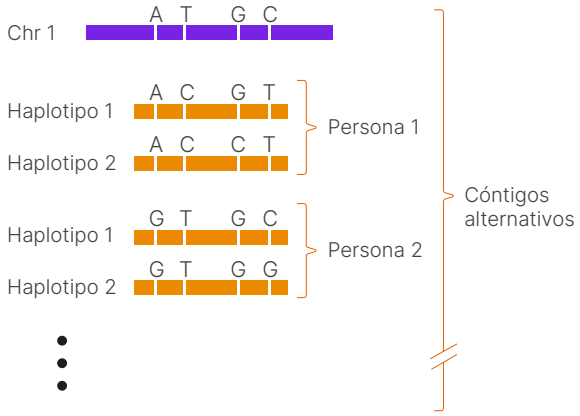

Figura 8: Mapeador multigenómico con control del pangenoma: se representa como varias rutas divergentes y convergentes en el contenido de una secuencia alternativa de control en una población.

Enmascaramiento de alternativas

Desde la actualización v3.9, el software de análisis secundario de DRAGEN incluye el enmascaramiento de alternativas; un nuevo enfoque para gestionar los cóntigos alternativos de control nativos, que permite enmascarar las posiciones estratégicas de los cóntigos alternativos para aumentar la precisión. Esta estrategia es fácil de definir, mantener y perfeccionar con el tiempo.

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

Aprendizaje automático

Se ha añadido al software de análisis secundario de DRAGEN v3.9 un proceso de recalibración ML potente y eficiente como opción una dentro del flujo de trabajo de variantes pequeñas germinales, que está habilitado de forma predeterminada en el software de análisis secundario de DRAGEN v4.0. Cuando está habilitado, este proceso ejecuta el modelo ML después de la llamada de variantes estándar. Este paso recalibra los campos QUAL y GQ que se generan en el archivo VCF final. En algunos casos, el ML puede cambiar GT. Los valores previos al ML de estos campos se conservan en los campos DQUAL, DGT y DGQ para que no se pierda información. Este paso añade aproximadamente cinco minutos al flujo de trabajo estándar para un experimento germinal de WGS de 30×, de modo que las mejoras de precisión tienen un impacto limitado en el tiempo total del experimento.

El modelo de ML se genera mediante formación fuera de línea supervisada. El modelo procesa un conjunto de funciones contextuales y fundamentadas en la lectura para refinar la precisión de las puntuaciones de calidad de los llamadores de variantes pequeñas. Las funciones utilizadas para formar el modelo incluyen la capacidad de asignación, AF, VC-Qual, DP, contenido de GC, discrepancias y otras métricas internas de asignación, alineación y VC.

Cálculo de la puntuación F1

$$F1 = 2 \times (Recalibración \times Precisión)/(Retirada + Precisión)$$

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

Línea de comandos DRAGEN

Encuentre fórmulas para empezar en DRAGEN recipe-germline WGS

Bibliografía

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. Fecha de consulta: 19 de septiembre de 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. Fecha de publicación: 12 de enero de 2022. Fecha de consulta: 19 de septiembre de 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. Fecha de publicación: 12 de agosto de 2024. Fecha de consulta: 30 de septiembre de 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. Fecha de publicación: 7 de junio de 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. Fecha de consulta: 19 de septiembre de 2024.
6. Datos internos archivados. Illumina, Inc. 2022.
7. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. Fecha de publicación en línea: 25 de octubre de 2024. doi:10.1038/s41587-024-02382-1

1 800 809 4566 (llamada gratuita, EE. UU.) | tel.: +1 858 202 4566
techsupport@illumina.com | www.illumina.com

M-GL-01016 ESP v3.0