Mejoras en la precisión de la llamada de variantes pequeñas de línea germinal con la plataforma DRAGEN™

Una serie de algoritmos para mejorar la precisión permiten detectar variantes pequeñas con gran sensibilidad y especificidad, a la vez que se mantienen los estándares de DRAGEN para la velocidad de computación.

Introducción

Gracias a los avances en la tecnología de secuenciación de próxima generación (NGS, next-generation sequencing), el volumen de datos de secuenciación generados continúa creciendo de manera exponencial. Este crecimiento conlleva la demanda de métodos analíticos rápidos y eficientes que mantengan unos estándares elevados en la precisión de la llamada de variantes. La plataforma de tecnología bioinformática DRAGEN (del inglés, Dynamic Read Analysis for Genomics) de Illumina ofrece análisis secundarios ultrarrápidos y de gran precisión de datos de NGS. La plataforma DRAGEN utiliza tecnología de matriz de puertas programable in situ (FPGA, field-programmable gate array) que se puede volver a configurar para acelerar drásticamente el análisis secundario de datos de NGS, en el que se incluyen las tareas de asignación, alineación y llamada de variantes.

Las características esenciales de la plataforma DRAGEN abordan los principales desafíos del análisis genómico, como los tiempos de computación prolongados y los volúmenes de datos masivos. La plataforma DRAGEN ofrece rapidez, flexibilidad, precisión y rentabilidad. La naturaleza reprogramable de la plataforma DRAGEN hace posible mejorar los algoritmos para adaptarla a aplicaciones de NGS nuevas. La velocidad de la plataforma permite a los desarrolladores iterar con rapidez en los diseños de algoritmos con métodos intensivos en computación que no resultan prácticos con los modelos tradicionales de solo software. Por lo tanto, la precisión de la plataforma DRAGEN ha mejorado continuamente con las versiones nuevas y, ahora, proporciona una solución excelente para las llamadas de variantes pequeñas en la secuenciación del genoma completo (WGS, whole-genome sequencing) de la línea germinal.

En esta nota de aplicación se describen las mejoras recientes de la plataforma de tecnología bioinformática DRAGEN de Illumina para el análisis secundario rápido; además, se demuestra su velocidad y precisión con tres conjuntos de datos de WGS publicados. Realizamos análisis comparativos de DRAGEN v3.2.8 frente a otros procedimientos, entre los que se incluyen BWA-MEM+GATK4 y DRAGEN v2 (figura 1). Los resultados de las llamadas de variantes de todos los procedimientos se compararon con un "conjunto de verdad" para las llamadas de referencia con el fin de identificar los falsos positivos (FP, false positives) y los falsos negativos (FN, false negatives).

Las métricas que se utilizaron para comparar los procedimientos son los tiempos totales de ejecución y las métricas de exactitud, como la exhaustividad, la precisión, y los FP y FN. La combinación de velocidad, exactitud y una amplia gama de aplicaciones disponibles posiciona a la plataforma DRAGEN en el punto idóneo para revolucionar el panorama del análisis genómico.

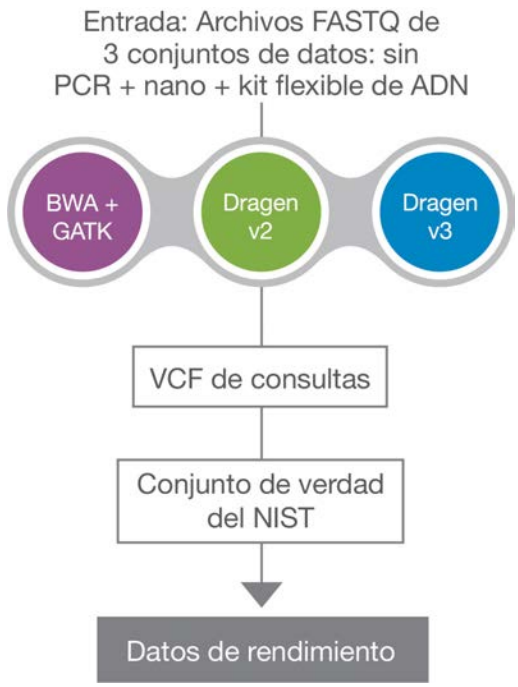

Figura 1: Diseño del estudio de análisis comparativo. Los archivos FASTQ de tres conjuntos de datos se ejecutaron a través de tres procesos de análisis para generar archivos VCF de consulta. A continuación, se utilizó la herramienta Variant Calling Assessment Tool (VCAT) para identificar los TP, FP y FN, en función de la comparación de llamadas de variantes con variantes de referencia en el conjunto de verdad del NIST.

Algoritmos de DRAGEN v3 para mejorar la precisión

DRAGEN v3 implementa las actualizaciones más recientes de algoritmos para detectar polimorfismos de nucleótido único (SNP, single-nucleotide polymorphisms) e inserciones/deleciones (indeles), que proporcionan mejoras en la precisión y la sensibilidad analítica.

Se hicieron mejoras en cuatro áreas para la llamada de variantes: modelo de error de indeles específico de la muestra, modelos matemáticos rigurosos de errores de pileup correlacionados, un enfoque optimizado para representar de forma exhaustiva un número exponencial de candidatos de haplotipos en regiones con muchas variantes y aumento por columnas de la lista de eventos generados por el conjunto de gráficos de De Bruijn. Estas mejoras dan como resultado una modesta aceleración de las ganancias, a la vez que elevan los estándares de exactitud en comparación con los procedimientos que se evalúan en este documento. Todas las mejoras de los algoritmos se describen con más detalle en el apéndice.

Métodos

Se siguieron cuidadosamente las prácticas recomendadas para el análisis comparativo.[1] Para demostrar la velocidad y la precisión con DRAGEN v3, se realizó un estudio comparativo con tres conjuntos de datos, de preparaciones de bibliotecas diferentes, generados a partir de la muestra NA12878 (figura 1). En resumen, el archivo FASTQ de cada conjunto de datos se utilizó como entrada para el análisis secundario de procedimientos independientes (DRAGEN v3.2.8, DRAGEN v2 y BWA + GATK[2]). Los archivos VCF resultantes de cada procedimiento (VCF de consultas) se cargaron en un proyecto en BaseSpace™ Sequence Hub. La herramienta Variant Calling Assessment Tool (VCAT v3.1.1 con la versión 0.3.10 de Hap.py) se utilizó para comparar cada archivo VCF de consulta con un "conjunto de verdad" de variantes de referencia con el fin de identificar llamadas de variantes verdaderas o falsas. Los resultados se recogieron y se colocaron en una tabla para realizar comparaciones entre los procesos. Todos los datos de entrada, los resultados de los análisis y las herramientas de evaluación se encuentran disponibles sin coste alguno en el proyecto de BaseSpace.[3] En el apéndice se pueden encontrar descripciones más detalladas de los métodos.

Resultados de los análisis comparativos

Los resultados tanto de los tiempos de ejecución como de las comparaciones de exactitud demuestran que DRAGEN proporciona una solución potente para el análisis secundario de los datos de NGS.

Exactitud de DRAGEN: FP+FN, exhaustividad y precisión

Aunque DRAGEN v2 ya era competitivo con una solución informática líder en el sector, DRAGEN v3 presenta varias modificaciones nuevas (descritas en la sección de métodos de algoritmos) que dan lugar a mejoras significativas en la precisión. Los resultados de este análisis comparativo también demuestran que las mejoras de DRAGEN v3 lo hacen superior en comparación con otros procesos de análisis populares, entre los que se incluye una versión anterior de DRAGEN, en todas las métricas de exactitud que se analizan en el estudio.

Cuando se evaluó la métrica de FP+FN para la detección de la variante de nucleótido único (SNV), DRAGEN v3 se desempeñó con una precisión significativamente mayor que los procesos de BWA + GATK4 y DRAGEN v2 para los tres conjuntos de datos (figura 2). Cuando se evaluó la métrica de FP+FN para la detección de indeles, DRAGEN v3 mostró un rendimiento mejor que el proceso de BWA + GATK4 para los tres conjuntos de datos; además, presentó una mejora adicional entre DRAGEN v3 y DRAGEN v2 (figura 3).

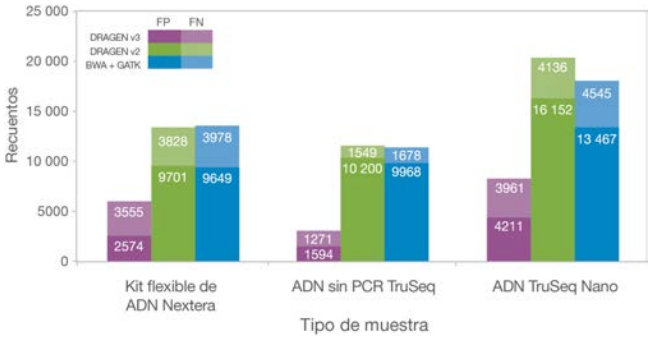

Figura 2: Falsos positivos y falsos negativos con detección de la SNV. Se analizaron archivos de incidencias (FASTQ) de tres conjuntos de datos en tres procesos independientes. Cada conjunto de datos (ADN sin PCR TruSeq, kit flexible de ADN Nextera y ADN TruSeq Nano) se generó a partir del ADN de la muestra NA12878, y las llamadas de variantes (VCF) de cada proceso de análisis se compararon con el conjunto de verdad del NIST (también basado en la muestra NA12878) para identificar los FP y FN.

Al evaluar las métricas de precisión y exhaustividad, la ventaja de las mejoras del algoritmo de DRAGEN v3 resulta evidente para la detección tanto de SNP como de indeles. Los valores tanto para la precisión como para la exhaustividad son sistemáticamente superiores al 99 % para todos los procesos y con cada conjunto de datos de detección de la SNV (tabla 1). En el caso de la detección de SNP, DRAGEN v2 se podía comparar con BWA + GATK4. Sin embargo, DRAGEN v3 muestra una mejora significativa tanto en la exhaustividad como en la precisión con respecto a los otros dos procesos. En el caso de la detección de indeles, DRAGEN v2 mostró una precisión mayor que BWA + GATK4, mientras que DRAGEN v3 arrojó una mejoría adicional sobre DRAGEN v2 tanto en lo que se refiere a la exhaustividad como a la precisión (tabla 2).

Velocidad de DRAGEN

Se recopiló la comparación de los tiempos de ejecución de DRAGEN para las soluciones que se encuentran tanto en la nube como en un entorno local. En el caso de la solución que se encuentra en un entorno local, se comparó DRAGEN v3 con BWA + GATK, con ambos procesos ejecutándose en el mismo servidor. En el caso de la solución alojada en la nube, DRAGEN v3 se ejecutó en BaseSpace Sequence Hub y se comparó con BWA + GATK, que se ejecutó en Terra.[4]

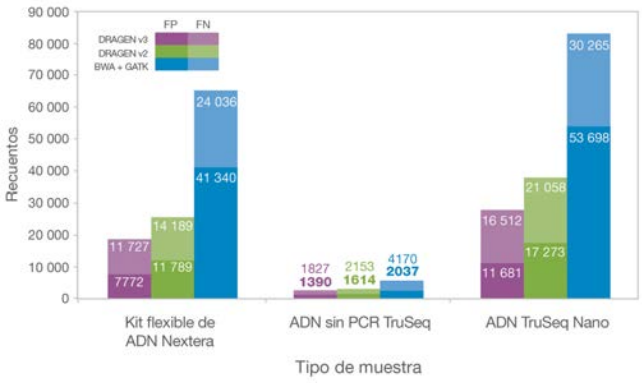

Figura 3: Falsos positivos y falsos negativos con detección de indeles. Se analizaron archivos de incidencias (FASTQ) de tres conjuntos de datos en tres procesos independientes. Cada conjunto de datos (ADN sin PCR TruSeq, kit flexible de ADN Nextera y ADN TruSeq Nano) se generó a partir del ADN de la muestra NA12878, y las llamadas de variantes (VCF) de cada proceso de análisis se compararon con el conjunto de verdad del NIST (también basado en el ADN de la muestra NA12878) para identificar los FP y FN.

DRAGEN acelera tanto el proceso de asignación como la llamada de variantes, que se pueden ejecutar de forma independiente. Aunque no se recoge en este estudio, vale la pena señalar que, antes del análisis secundario, DRAGEN también es compatible con la conversión acelerada de BCL2FASTQ, lo que mejora en gran medida la velocidad y la eficiencia, al mismo tiempo que produce FASTQ idénticos. También cabe destacar que DRAGEN produce automáticamente una lista exhaustiva de métricas de control de calidad, tanto en el nivel de asignación como en el de llamadas de variantes, con poca o ninguna repercusión en el tiempo de ejecución. Esto contrasta con otros procedimientos que dependen de herramientas de terceros que funcionan a un ritmo lento (por ejemplo, Samtools, Picard, etc.) a la hora de adquirir métricas de control de calidad, lo que repercute de forma significativa en el tiempo de ejecución.

Cuando se midieron las velocidades de ejecución con procesos que se ejecutaron en el mismo servidor local, DRAGEN v3 fue bastante más rápido que BWA + GATK, con incrementos de la velocidad en el rango de entre 16 y 18× (figura 4).

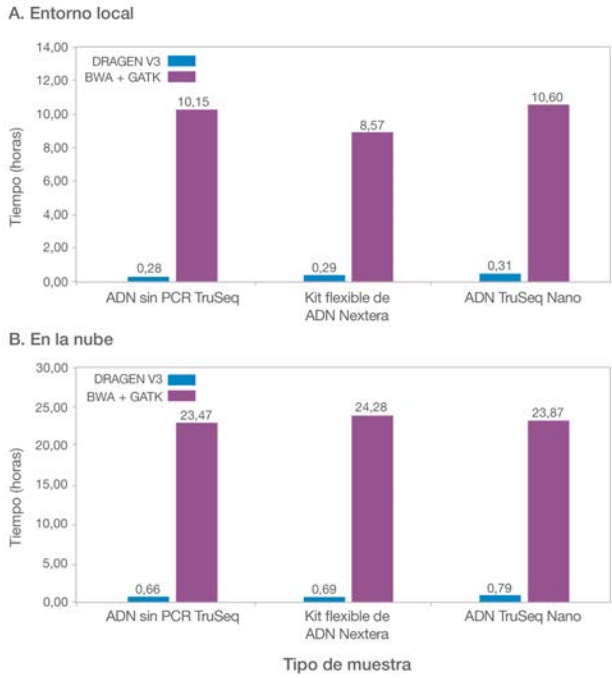

Figura 4: Comparaciones del tiempo de ejecución del análisis en el entorno local y en la nube. (A) DRAGEN v3 y BWA + GATK se ejecutaron en el mismo servidor local. (B) DRAGEN v3 se ejecutó en BaseSpace Sequence Hub y se comparó con BWA + GATK, que se ejecutó en Terra.

En el caso de los procesos que se ejecutaron en la nube, DRAGEN v3, que se ejecutó en BaseSpace Sequence Hub, fue bastante más rápido que BWA + GATK, que se ejecutó en Terra, con incrementos de la velocidad en el rango de entre 13 y 16×.

Tabla 1: Sensibilidad y especificidad de la detección de la SNV

	Precisión			Exhaustividad		
Conjuntos de datos	DRAGEN v3	DRAGEN v2	BWA + GATK	DRAGEN v3	DRAGEN v2	BWA + GATK
ADN sin PCR TruSeq	99,95 %	99,68 %	99,69 %	99,96 %	99,95 %	99,95 %
Kit flexible de ADN Nextera	99,92 %	99,70 %	99,70 %	99,89 %	99,88 %	99,88 %
ADN TruSeq Nano	99,87 %	99,50 %	99,58 %	99,88 %	99,87 %	99,86 %

Tabla 2: Sensibilidad y especificidad de la detección de indeles

	Precisión			Exhaustividad		
Conjuntos de datos	DRAGEN v3	DRAGEN v2	BWA + GATK	DRAGEN v3	DRAGEN v2	BWA + GATK
ADN sin PCR TruSeq	99,71 %	99,66 %	99,58 %	99,62 %	99,55 %	99,13 %
Kit flexible de ADN Nextera	98,37 %	97,54 %	91,53 %	97,56 %	97,05 %	95,01 %
ADN TruSeq Nano	97,56 %	96,39 %	89,37 %	96,57 %	95,63 %	93,71 %

Resumen

A medida que las aplicaciones genómicas avanzan hacia la caracterización precisa de regiones difíciles del genoma y la medición de llamadas de baja frecuencia de alelos a partir de muestras con un nivel de ruido elevado, DRAGEN demuestra ser la plataforma más adecuada para procesar los datos de NGS del futuro de forma eficiente y precisa.

La velocidad de DRAGEN no solo permite a los investigadores mantenerse al día con el creciente rendimiento de los instrumentos de NGS, sino que también permite iteraciones rápidas para la mejora continua de sus algoritmos con el fin de proporcionar una gran precisión.

Apéndice

Descripción detallada de los algoritmos nuevos

Modelo de error de PCR específico de la muestra

Uno de los retos en la llamada de variantes consiste en distinguir los errores de indeles de las variantes reales. Para ello, los llamadores de variantes suelen emplear un modelo oculto de Márkov (HMM, Hidden Markov Model), que modela el comportamiento estadístico de los errores de indeles, como parte del cálculo de probabilidad. Por lo general, el HMM presenta parámetros de entrada (penalización por apertura de brecha [GOP, Gap Open Penalty] y penalización por continuación de la brecha [GCP, Gap Continuation Penalty]), que están directamente relacionados con la tasa de error de indel (es decir, tasa de error de indel = f[GOP,GCP]). Los errores de indel son más probables en presencia de repeticiones cortas en tándem (STR, short tandem repeats) y la probabilidad de error (y, por lo tanto, de GOP y GCP) puede depender tanto del período como de la duración de la STR. El proceso de error puede variar mucho de un conjunto de datos a otro, en función de factores como la amplificación PCR. Si se desea una detección precisa, resulta fundamental utilizar parámetros de un HMM que modelen con precisión el proceso de error por muestra. Sin embargo, los llamadores de variantes habituales suelen utilizar parámetros fijos o funciones predeterminadas no específicas de la muestra que no modelan con precisión el proceso de error, lo que deriva en un rendimiento de detección deficiente.

La calibración automática del HMM implementada en DRAGEN v3 aborda el problema anterior estimando los parámetros de PCR directamente a partir del conjunto de datos que se está procesando. Esta operación se realiza tras las tareas de asignación y alineación, y antes de la llamada de variantes, sin conocimiento de la verdad del terreno y sin utilizar bases de datos externas de mutaciones conocidas. Los parámetros dependen tanto del período de STR como de la longitud de las repeticiones.

Para un período de STR y una longitud determinados, se selecciona un conjunto de N locus con el período y la longitud deseados, y el algoritmo analiza los pileups de lecturas asignadas a dichos locus y cuenta las indeles que se observan en cada locus. La idea clave es que, considerando un número suficiente de locus, es posible estimar con precisión los parámetros de interés. Lo hacemos encontrando los parámetros que maximizan la probabilidad de producir el conjunto de N pileups observados. Si el número de parámetros para maximizar la probabilidad es lo suficientemente pequeño (por ejemplo, 2), es posible realizar una búsqueda exhaustiva. En la implementación actual de DRAGEN v3, la optimización se lleva a cabo sobre dos parámetros: GOP y alfa, que indica la probabilidad de variantes de indel de cualquier longitud distinta de cero. Para cada período de STR y longitud considerados, la búsqueda genera GOP y alfa que maximizan la probabilidad de producir el conjunto de N pileups observados; dichos valores se utilizan como entrada al HMM. También se puede ampliar la búsqueda más allá de dos parámetros, lo que aportaría mejoras nuevas.

Caída de la calidad de la base (BQD)

Los llamadores de variantes convencionales se han diseñado con la suposición de que los errores de secuenciación son independientes en todas las lecturas; si se sigue esta suposición, es muy poco probable que se produzcan varios errores idénticos en un locus concreto. Sin embargo, tras analizar los conjuntos de datos de NGS, se observó que las ráfagas de errores son mucho más comunes de lo que se predeciría por la suposición de independencia; además, dichas ráfagas pueden dar como resultado muchos falsos positivos.

Por suerte, estos errores presentan características distintas que los diferencian de las variantes reales. El algoritmo de BQD (base quality drop off) que se ha implementado en DRAGEN v3 es un mecanismo de detección que extrae propiedades determinadas de esos errores (cortes en la cadena, localización del error en la lectura, calidad media de la base baja) en el locus de interés y las incorpora al cálculo de probabilidad de manera simple y sólida en el genotipo. Las hipótesis nuevas de candidatos a genotipo se añaden a la lista antigua de genotipos diploides (aquellos que suponen errores de pileups independientes). Por ejemplo, en el caso de un locus con un alelo ALT, además de considerar P(G00|R), P(G01|R) y P(G11|R), añadimos dos hipótesis más como P(G00,E1|R) y P(G11,E0|R), en las que los alelos E0 y E1 representan al alelo de referencia y al alelo ALT que procede de un error de secuenciación. Las propiedades de esos errores, como los cortes en la cadena, la localización del error en la lectura y la calidad media de la base, se incorporan en el cálculo de P(G00,E1|R) y P(G11,E0|R). Entonces, el genotipo ganador se hace cargo de max (max(P(G00|R), P(G00,E1|R)), P(G01|R) y max(P(G11|R), P(G11,E0|R)).

Ser capaz de caracterizar los errores de secuenciación correlacionados desde el núcleo del llamador de la variante se deriva en un aumento significativo de la especificidad porque se eliminan muchas de las llamadas de FP. También ayuda a la sensibilidad al corregir errores de genotipo.

Detección de lectura extraña (FRD)

Los llamadores de variantes convencionales tratan los errores de asignación como eventos de errores independientes por lectura e ignoran el hecho de que, por lo general, dichos errores se producen en ráfagas. Esto puede dar como resultado llamadas de variantes emitidas con puntuaciones de confianza muy altas, a pesar de tener una calidad de asignación (MAPQ, mapping quality) baja o un AF sesgado. Si se desea mitigar este problema, los llamadores de variantes convencionales suelen filtrar las lecturas antes de realizar la llamada, en función de un umbral de MAPQ (es decir, las lecturas con una MAPQ por debajo del umbral se excluyen del cálculo). Sin embargo, esto descarta una evidencia valiosa de dentro del llamador de variantes y lleva a cabo un trabajo pobre de suprimir falsos positivos.

DRAGEN v3 ha implementado la detección de lectura extraña (FRD, Foreign Read Detection), que es una extensión del algoritmo de genotipado antiguo, incorporando la hipótesis adicional de que algunas lecturas en el pileup son lecturas extrañas (es decir, su ubicación real se encuentra en otra parte del genoma de referencia o se originan fuera de este [es decir, contaminación de la muestra]). El algoritmo extrae varias propiedades (frecuencia alélica sesgada y MAPQ baja) e incorpora esta evidencia en el cálculo de probabilidad de una manera matemáticamente rigurosa.

Las hipótesis nuevas de candidatos a genotipo se añaden a la lista antigua de genotipos diploides (aquellos que suponen errores de pileups independientes). Por ejemplo, en el caso de un locus con un alelo ALT, además de considerar P(G00|R), P(G01|R) y P(G11|R), añadimos dos hipótesis más como P(G00,F1|R) y P(G11,F0|R), en las que los alelos F0 y F1 representan al alelo de referencia y al alelo ALT que procede de un error de asignación. Las propiedades de estos errores, como la profundidad del alelo y la MAPQ, se incorporan en el cálculo de P(G00,F1|R) y P(G11,F0|R). Entonces, el genotipo ganador se hace cargo de max (max(P(G00|R), P(G00,F1|R)), P(G01|R) y max(P(G11|R), P(G11,F0|R)).

Se mejora la sensibilidad desde el rescate de FN, la corrección de genotipos y la posibilidad de reducir el umbral de la MAPQ para las lecturas entrantes del llamador de variantes. Se perfecciona la especificidad a partir de la eliminación de FP y la corrección de genotipos.

La FRD se considera una herramienta más potente que los métodos de filtrado pos-VCF a la hora de mejorar la medida F porque, en lugar de limitarse a detectar resultados sospechosos (por ejemplo, en función de la profundidad del alelo o de los errores de lectura) después del llamador de la variante, el algoritmo de detección incorpora directamente la presencia de lecturas extrañas a través de una rigurosa detección de máxima probabilidad.

Detección por columnas y PDHMM

Los llamadores de variantes, como GATK HaplotypeCaller y DRAGEN, utilizan el gráfico de De Bruijn para volver a recopilar las lecturas con el fin de determinar los haplotipos candidatos e identificar las posibles posiciones de las variantes. En regiones del genoma con repeticiones en tándem, variantes estructurales o grupos de errores de secuenciación, la sensibilidad puede ser menor si la metodología de secuenciación de los gráficos no permite obtener una lista completa de posiciones de variantes y haplotipos candidatos.

La detección de eventos por columna complementa el gráfico de De Bruijn escaneando cada columna de una región activa en busca de posibles posiciones de variantes (SNP e indeles) y completando la lista de haplotipos candidatos. Esto restaura la sensibilidad en las regiones donde el gráfico falla.

Impacto de FRD/BQD en el filtrado duro de CAL/CG/QD y pos-VCF

El llamador de variantes DRAGEN v3 ha implementado dos algoritmos que modelan los errores correlacionados a través de las lecturas en un pileup determinado, la detección de lecturas extrañas (FRD) para detectar lecturas mal asignadas y el algoritmo de caída de calidad de base (BQD) para detectar errores de llamada de bases correlacionados. Además de mejorar la especificidad y la sensibilidad, estos dos algoritmos tienen un impacto o beneficio a dos niveles:

Los valores de la puntuación de confianza (CAL, CG, QD) se encuentran en un rango realista de la escala Phred.

Por lo general, los llamadores de variantes convencionales emiten valores CAL exagerados en la escala Phred en el rango de unos pocos miles que no tienen un significado práctico estadísticamente hablando. El modelado de errores correlacionados desde dentro del llamador de la variante devuelve estos valores a un rango estadísticamente realista y significativo.

La dependencia de las reglas de filtrado posteriores al VCF se reduce en gran medida.

En los llamadores de variantes convencionales, debido a la incapacidad de estos para distinguir entre errores correlacionados y variantes reales, se tuvieron que aplicar reglas de filtrado duras después del VCF para filtrar el número excesivo de llamadas de FP. Se compararon varias anotaciones del VCF (por ejemplo, QD, MQ, FS y MQRankSum) con umbrales ad-hoc para marcar las llamadas como FP. Como alternativa, esas anotaciones se podrían alimentar a un algoritmo de aprendizaje automático y entrenar frente a un conjunto de verdad; además, los falsos positivos se podrían filtrar en función del entrenamiento (por ejemplo, VQSR).

En DRAGEN v3, los algoritmos se mejoraron en el núcleo del llamador de las variantes y, por lo tanto, la dependencia del filtrado pos-VCF se redujo bastante. La regla de filtrado duro predeterminada de DRAGEN v3 utiliza tan solo el valor CAL con un umbral que se corresponde con los mejores Fmeas (mejor compensación entre sensibilidad y especificidad).

Métodos detallados

Conjuntos de datos de entrada

Se seleccionaron tres conjuntos de datos para representar varios métodos de preparación de bibliotecas, que incluían y excluían la PCR (ADN TruSeq Nano, ADN sin PCR TruSeq y kit flexible de ADN Nextera). Todos los conjuntos de datos se generaron con ADN de la muestra NA12878. Tras la preparación de la biblioteca de ADN de acuerdo con las respectivas guías de referencia, se secuenciaron de[5a7] bibliotecas resultantes en ejecuciones finales de 2 × 150 pares en el Sistema NovaSeq™ 6000. Para normalizar el número de lecturas, cada conjunto de datos se redujo a 30 veces la cobertura con el FASTQ Toolkit en BaseSpace Sequence Hub. Los tres conjuntos de datos se encuentran a disposición del público en BaseSpace Sequence Hub, para que se pueda realizar una evaluación independiente de los resultados.

Genoma humano de referencia

La referencia genómica utilizada fue Human hs37d5 en la aplicación DRAGEN BaseSpace y la referencia genómica equivalente se utilizó en el análisis local para cada proceso en evaluación. Esta referencia incluye los señuelos.[8]

Procesos de análisis secundarios

Comparamos tres procesos de análisis secundarios. El primer proceso es DRAGEN v2 de principio a fin (DRAGEN se utiliza tanto para la etapa de asignación y alineación como para la llamada de variantes). El segundo proceso es DRAGEN v3 de principio a fin. El tercer proceso utiliza BWA-MEM para la etapa de asignación y alineación, y GATK4-HC para la fase de llamada de variantes.

Para hacer una comparación justa, aplicamos la misma regla de filtrado duro para los tres procesos, que consistía en aplicar un umbral de CG a los VCF del prefiltro. El umbral se seleccionó para que estuviera cerca del mejor punto de Fmeas en cada proceso (tabla 3).

Tabla 3: Umbrales óptimos de control de calidad de los Fmeas

CG para los mejores Fmeas	SNP	Inserción y deleción
DRAGEN v3.2.8	9	9
DRAGEN v2.5	2	8
BWA + GATK	1	2

DRAGEN se ejecutó en un servidor local, además de en la nube con BaseSpace Sequence Hub. Aunque el tiempo de cálculo es ligeramente mayor en la nube, los resultados de las llamadas de variantes no difieren. El proceso de BWA + GATK se ejecutó en el mismo servidor local de DRAGEN, donde se instaló el marco de BCBIO.[9] BCBIO ejecuta el BWA + GATK siguiendo las directrices de mejores prácticas de GATK y aplica optimizaciones adicionales para mejorar el paralelismo y reducir el tiempo de ejecución. En el caso de los análisis en la nube, el proceso de BWA + GATK se ejecutó en Terra.

DRAGEN 3.3.0

Versión de la aplicación DRAGEN:
DRAGEN Germline Pipeline 3.2.8
DRAGEN Host Software Version 05.011.281.3.2.8

BWA-Mem (0.7.17) + GATK4 (4.0.2)

Tabla 4: Parámetros del archivo de configuración de los algoritmos de BCBIO

Parámetro	Valor
align_split_size	5 000 000
aligner	BWA
coverage_depth	Elev
coverage_interval	Regional
mark_duplicates	Verdadero
merge_bamprep	Falso
platform	Illumina
quality_format	Estándar
realign	Falso
recalibrate	Falso
tools_off	Vqsr
variantcaller	GATK-haplotype

análisis: variant2
recursos: gatk-haplotype

BWA + GATK en Terra

Los archivos Bam preparados para el análisis de BWA-Mem (de las ejecuciones de BCBIO) se utilizaron como entradas para ejecutar GATK en Terra. En resumen, seguimos el flujo de trabajo GATK4-germline-snps-indels (https://github.com/gatk-workflows/gatk4-germline-snps-indels) con modificaciones en parámetros específicos para que coincidan con los parámetros de las ejecuciones de BCBIO. Todas las ejecuciones se llevaron a cabo con una cuenta de prueba gratuita en Terra.

El método exacto de WDL se encuentra disponible en los datos públicos de BaseSpace Sequence Hub.

Configuraciones del método de WDL:
Imagen del cargador de GATK: broadinstitute/gatk:4.0.2.0
Cargador de GITC: broadinstitute/genomes-in-the-cloud:2.3.1-1500064817

FASTA de referencia: hs37d5 (igual que otros procesos)

En este proceso solo se generaron archivos VCF sin procesar. El posfiltrado se realizó de manera local. Los archivos VCF sin procesar se encuentran disponibles en los datos públicos de BaseSpace Sequence Hub.

Basespace (enero de 2019). Concrete la especificación de la instancia F1 de AWS (Amazon Web Services) que se ha utilizado (AWS, F1, 4×, grande).
Versión de la aplicación BaseSpace Sequence Hub: 3.2.8

Tabla 5: Servidor local (CentOS 7 x86_64, Supermicro 1029)

Pieza	Nombre completo del modelo	Notas
Chasis	SYS-1029GQ-TNRT	1 unidad de soporte
CPU	2 × Intel(R) Xeon(R) Gold 6126 CPU @ 2,60 GHz	24 núcleos, 48 hilos
RAM	384 GB	DDR4, 2666 MHz
Montaje	Intel SSDPE2KE020T7	2 TB NVME

Conjunto de verdad de análisis comparativo (NIST)

El análisis comparativo de las llamadas de variantes requiere un genoma de referencia específico y un conjunto asociado de llamadas que representen las "respuestas de verdad" de ese genoma. Estos conjuntos de llamadas cuentan con la propiedad de que se pueden utilizar como "verdad" para identificar con precisión los falsos positivos y negativos. Para este estudio, el conjunto de verdad utilizado se basó en llamadas de referencia basadas en la misma fuente de ADN (NA12878) que estableció el Instituto de estándares y tecnología de EE. UU., National Institute of Standards and Technology (NIST). El consorcio Genome in a Bottle Consortium (GIAB) es un consorcio académico público-privado que auspicia el NIST. El GIAB publicó un conjunto de referencia de variantes pequeñas y llamadas de referencia para su genoma piloto, NA12878, que caracteriza un genotipo de alta fiabilidad para aproximadamente el 90 % de los GRCh37 y GRCh38.

Los verdaderos positivos (TP) son llamadas de variantes que concuerdan con las llamadas de referencia del conjunto de verdad del NIST. Los falsos positivos (FP) son llamadas de variantes que no existen en el conjunto de verdad y los falsos negativos (FN), variantes del conjunto de verdad que no se llamaron en el VCF de consulta.

La herramienta Variant Calling Assessment Tool (VCAT) se utilizó para comparar el archivo VCF de consulta con el conjunto de verdad del NIST v3.3.2. Esta herramienta ejecuta hap.py con el motor de evaluación RTG vcfeval. Los TP, FP y FN fueron determinados por los archivos de salida de hap.py *roc.Locations.INDEL.csv y *roc.Locations.SNP.csv de TRUTH TP, QUERY.TP, QUERY.FP y TRUTH.FN.

El tipo de restricción utilizado para calcular los TP, FP y FN es "coincidencia de genotipo" (véase [1]), para el que solo se consideran TP las posiciones con alelos y genotipos que coinciden; lo que significa que los errores de genotipo y los desajustes de alelos se cuentan como FP y FN.

Métricas de análisis comparativos

Para las comparaciones de velocidad, el tiempo total de ejecución en segundos, desde FASTQ hasta VCF, se deriva de los archivos de registro de análisis o de los tiempos de análisis que se muestran en los informes.

Para realizar comparaciones de precisión entre varios procesos, utilizamos estándares recomendados en las métricas de rendimiento (tabla 6).[1] La precisión es la métrica que representa la especificidad analítica o la capacidad de identificar correctamente la ausencia de variantes o la "ausencia de falsos positivos". La exhaustividad es la métrica que representa la sensibilidad analítica o la capacidad de detectar variantes que se sabe que están presentes o la "ausencia de falsos negativos".

Las definiciones y los cálculos para las métricas relacionadas con la precisión y los números de exhaustividad se basan en la referencia.

Tabla 6: Definiciones y cálculos para métricas complicadas con precisión y exhaustividad

Métrica	Nombre común	Definición	Fórmula
TRUTH.TP	Positivos verdaderos (Truth = Verdadero)	Número de llamadas verdaderas para las que existe una llamada de consulta que es coherente con la llamada verdadera y su genotipo	
QUERY.TP	Positivos verdaderos (Query = Consulta)	Número de llamadas de consulta para las que existe una llamada verdadera que es coherente con la llamada de consulta y su genotipo	
TRUTH.FN	Falsos negativos	Número de llamadas verdaderas para las que no existe una llamada de consulta que sea coherente con la llamada verdadera y su genotipo	
QUERY.FP	Falsos positivos	Número de llamadas de consulta para las que no existe una llamada verdadera que sea coherente con la llamada de consulta y su genotipo	
METRIC.Recall	Exhaustividad, sensibilidad	Fracción de llamadas verdaderas que son coherentes con una llamada de genotipo y alelo de consulta dentro de las regiones seguras	TRUTH.TP / (TRUTH.TP + TRUTH.FN)
METRIC.Precision	Precisión, valor predictivo positivo	Fracción de llamadas de consulta que son coherentes con una llamada de genotipo y alelo verdadera dentro de las regiones seguras	QUERY.TP / (QUERY.TP + QUERY.FP)

Referencias

1. Krusche, P.; Trigg, L.; Boutros, P.C.; 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/. Acceso: 9 de mayo de 2019.
3. The BaseSpace project. basespace.illumina.com/s/3ExEZMlH8Lkq. Acceso: 15 de mayo de 2019.
4. FireCloud Powered by Terra. firecloud.terra.bio/. Acceso: 15 de mayo de 2019.
5. Illumina (2017). TruSeq DNA PCR-Free Reference Guide. Acceso: 6 de marzo de 2019.
6. Illumina (2017). TruSeq DNA Nano Reference Guide. Acceso: 6 de marzo de 2019.
7. Illumina (2018). Nextera DNA Flex Library Prep Reference Guide. Acceso: 6 de marzo de 2019.
8. hs37d5 Reference Genome. ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/. Acceso: 9 de mayo de 2019.
9. Bcbio-nextgen. Docs. bcbio-nextgen.readthedocs.io/en/latest/. Acceso: 9 de mayo de 2019.

Illumina, Inc. • 1.800.809.4566 (llamada gratuita, EE. UU.) • Tel.: +1.858.202.4566 • techsupport@illumina.com • www.illumina.com