illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx en NextSeq 550Dx

Guía del usuario de la aplicación

PROPIEDAD EXCLUSIVA DE ILLUMINA
N.º de documento 200025238 v00
Febrero de 2023
PARA USO DIAGNÓSTICO IN VITRO.

Historial de revisiones

Documento	Fecha	Descripción del cambio
200025238 v00	Febrero de 2023	Publicación inicial.

Índice

Descripción general

La aplicación DRAGEN for Illumina DNA Prep with Enrichment Dx (DRAGEN for IDPE Dx) se utiliza para planificar y realizar análisis secundarios de las bibliotecas IDPE Dx generadas para la secuenciación en el NextSeq 550Dx.

DRAGEN for IDPE Dx admite secuenciación para el análisis cuando se utiliza con la preparación de bibliotecas Illumina DNA Prep with Enrichment Dx, NextSeq 550Dx y Illumina DRAGEN Server for NextSeq 550Dx.

Métodos de análisis

DRAGEN for IDPE Dx realiza demultiplexado, generación de FASTQ, asignación de lecturas, alineación con un genoma de referencia y llamada de variantes pequeñas dependiendo de los flujos de trabajo seleccionados:

- Generación de FASTQ
- Generación de FASTQ y VCF Germline (germinal)
- Generación de FASTQ y VCF Somatic (somática)

NOTA La compresión ORA se puede utilizar con los tres flujos de trabajo. La compresión DRAGEN ORA es un software de compresión totalmente sin pérdidas que crea un archivo con una extensión de archivo de lectura original (*.ora). El formato ora es un formato de compresión basado en referencias para archivos FASTQ, y está diseñado para una compresión/descompresión muy rápidas y una elevada relación de compresión.

Generación de FASTQ

Las secuencias ensambladas se escriben en archivos FASTQ por muestra. Los archivos FASTQ son archivos de texto que contienen datos de secuenciación y puntuaciones de calidad para una sola muestra. Para cada muestra, se generan archivos FASTQ independientes por carril de la celda de flujo y por lectura de secuenciación. El nombre de la muestra especificado durante la configuración del experimento se incluye en el nombre del archivo FASTQ. Los archivos FASTQ incluyen los datos principales para la alineación. El primer paso para la generación de FASTQ es el demultiplexado. El demultiplexado asigna los grupos que pasan el filtro a una muestra comparando cada secuencia de lectura del índice con las secuencias de índices especificadas en el experimento. En este paso, no se considera ningún valor de calidad. Las lecturas del índice se pueden identificar por medio de los siguientes pasos:

- Las muestras están numeradas comenzando por el número uno de acuerdo con el lugar que ocupan en el experimento.

- La muestra número cero se reserva para los grupos que no se han asignado al experimento.
- Los grupos se asignan a una muestra siempre que la secuencia de índice coincida de forma exacta o cuando exista solamente una discrepancia por lectura del índice.

El software incluye compresión ORA para comprimir los archivos FASTQ. Este formato se puede habilitar opcionalmente. Cuando se utiliza el formato ORA (*.ora), la suma de comprobación md5 del contenido FASTQ se preserva tras un ciclo de compresión y descompresión para asegurar una compresión sin pérdida.

Cartografía y alineación de ADN

Después de generar FASTQ, las lecturas se cartografían y alinean con un genoma de referencia. La primera etapa de la cartografía es la generación de semillas a partir de la lectura, y a continuación la búsqueda de coincidencias exactas en el genoma de referencia. Estos resultados se refinan posteriormente ejecutando alineaciones de Smith-Waterman completas en las ubicaciones con la mayor densidad de coincidencias con las semillas. Este algoritmo bien documentado compara cada posición de la lectura con todas las posiciones candidatas de la referencia. Estas comparaciones corresponden a una matriz de posibles alineaciones entre lectura y referencia. Para cada una de estas posiciones de alineación candidatas, Smith-Waterman genera puntuaciones que se utilizan para evaluar si la mejor alineación que pasa por esa celda de la matriz la alcanza por una coincidencia o discrepancia de nucleótidos (movimiento diagonal), una deleción (movimiento horizontal) o una inserción (movimiento vertical). Una coincidencia entre lectura y referencia proporciona una bonificación en la puntuación, y una discrepancia o indel impone una penalización. La ruta global a través de la matriz con una mayor puntuación es la alineación elegida. El algoritmo se acelera mediante hardware en las tarjetas de array de puertas programable in situ (FPGA) DRAGEN. El genoma de referencia utilizado en la aplicación se crea a partir del FASTA de hg19 del UCSC con la opción DRAGEN para crear una tabla fragmentada ALT-Aware basada en liftover.

Llamada de variantes germinales de DRAGEN

El llamador de variantes pequeñas germinales de DRAGEN acepta como entradas lecturas de ADN cartografiado y alineado y llama polimorfismos de nucleótido único (SNP) e inserciones o deleciones (indels) mediante una combinación de detección por columnas y de ensamblado local *de novo* de haplotipos. Para habilitar el llamador de variantes pequeñas germinales de DRAGEN, seleccione el flujo de trabajo de variantes germinales.

La llamada de variantes germinales se utiliza normalmente para muestras germinales cuya ploidía se sabe que es de dos. Las regiones de referencia llamables se identifican por vez primera con una suficiente cobertura de alineación. En estas regiones de referencia, una exploración rápida de las lecturas ordenadas identifica las regiones activas, centradas en columnas de apilamiento con evidencias de una variante. Las regiones activas se rellenan con suficiente contexto para cubrir el contenido cercano significativo, no de referencia. Si hay indicios de indels, las regiones activas reciben un relleno adicional.

Las lecturas alineadas se recortan dentro de cada región activa y se ensamblan en un gráfico de De Bruijn. Los bordes de las lecturas recortadas se ponderan mediante recuentos de observación, con la secuencia de referencia como eje. Tras una cierta limpieza y simplificación del gráfico, todas las rutas de origen a destino se extraen como haplotipos candidatos. Cada haplotipo se alinea al genoma de referencia con Smith-Waterman para identificar las variantes que representa. Este conjunto de eventos puede ser aumentado por una detección basada en la posición. Para cada par lectura-haplotipo, la probabilidad P(r|H) de observar la lectura, asumiendo que el haplotipo es la verdadera muestra de inicio, se estima utilizando un modelo de Markov oculto (HMM) de pares.

Mediante una exploración por posiciones de referencia sobre la región activa, se forman genotipos candidatos a partir de combinaciones diploides de eventos de variantes (SNP o indels). Para cada evento (incluida la referencia), la probabilidad condicional P(r|e) de observar cada lectura solapada se estima como la P(r|H) máxima para los haplotipos que justifican el evento. Estos se combinan en la probabilidad condicional P(r|e1e2) para un genotipo (par de eventos) y se multiplican para obtener la probabilidad condicional P(R|e1e2) de observar el apilamiento de lecturas completo. Utilizando la fórmula de Bayes, se calcula la probabilidad posterior P(e1e2|R) de cada genotipo diploide, y se obtiene el ganador.

DRAGEN for IDPE Dx aplica el filtrado automático. Consulte *Anotaciones de los archivos VCF de flujos de trabajo germinales*, en la página 14 para obtener más información.

Llamada de variantes somáticas de DRAGEN

El llamador de variantes pequeñas somáticas de DRAGEN acepta como entradas lecturas de ADN cartografiado y alineado y llama SNV e indels mediante una combinación local *de novo* de haplotipos en una región activa. Para habilitar el llamador de variantes pequeñas somáticas de DRAGEN, seleccione una aplicación de variantes somáticas.

La llamada de variantes somáticas suele utilizarse para muestras de tumores. Con este flujo de trabajo, DRAGEN no realiza ninguna suposición de ploidía, lo que permite la detección de alelos de baja frecuencia. Para locus con una cobertura de hasta 100x en la muestra tumoral, DRAGEN tiene un umbral de detección en frecuencias de alelos de variantes del 5 %. El límite aumenta con una profundidad cada vez mayor en función de cada locus y se reduce a la mitad cada vez que la cobertura se duplica más allá de 100x. Las regiones de referencia llamables se identifican por vez primera con una suficiente cobertura de alineación. En estas regiones de referencia, una exploración de las lecturas ordenadas identifica las regiones activas, centradas en columnas de apilamiento con evidencias de una variante en las lecturas de tumores. Las regiones activas se rellenan con suficiente contexto para cubrir el contenido cercano significativo, no de referencia. Si hay indicios de indels, las regiones activas reciben un relleno adicional.

Las lecturas alineadas se recortan dentro de cada región activa y se ensamblan en un gráfico de De Bruijn. Los bordes de las lecturas recortadas se ponderan mediante recuentos de observación, con la secuencia de referencia como eje. Tras una cierta limpieza y simplificación del gráfico, todas las rutas de origen a destino se extraen como haplotipos candidatos. Cada haplotipo se alinea al genoma de

referencia con Smith-Waterman para identificar las variantes que representa. Para cada par lectura-haplotipo, la probabilidad P(r|H) de observar la lectura se estima utilizando un modelo de Markov oculto (HMM) de pares, asumiendo que el haplotipo es la verdadera muestra de inicio.

Para determinar la puntuación del límite de detección del tumor (TLOD), el llamador de variantes pequeñas somáticas de DRAGEN primero analiza por posición de referencia cada evento somático candidato, así como el evento de referencia sobre la región activa. La probabilidad condicional P(r|e) de observar cada lectura solapada se estima como la P(r|H) máxima para los haplotipos que justifican el evento. Estas se combinan en la probabilidad condicional P(r|E) para una hipótesis de evento, E, que implica una mezcla de los alelos somáticos candidatos y de referencia en un intervalo de posibles frecuencias alélicas, y se multiplican para obtener la probabilidad condicional P(R|E) de observar el apilamiento de lecturas completo. A partir de lo anterior, se calcula una puntuación TLOD como evidencia de que un alelo ALT está presente en la muestra tumoral en un locus dado.

DRAGEN for IDPE Dx aplica el filtrado automático. Consulte *Anotaciones de los archivos VCF de flujos de trabajo somáticos*, en la página 17 para obtener más información.

Crear un experimento planificado

Siga estos pasos para configurar un experimento en Illumina Run Manager, ya sea en el NextSeq 550Dx o utilizando un navegador en un ordenador conectado a la red. Utilice un navegador en un ordenador conectado a la red si desea importar datos de muestras. Consulte Guía del software Illumina Run Manager para NextSeq 550Dx (n.º de documento 200025239) para obtener instrucciones sobre el acceso al Illumina Run Manager desde un ordenador conectado a la red.

Hay dos maneras diferentes de crear un nuevo experimento planificado:

- **Import Run** (Importar experimento): utilice una hoja de muestras de un experimento existente como plantilla para otro nuevo. Consulte Guía del software Illumina Run Manager para NextSeq 550Dx (n.º de documento 200025239) para obtener información sobre cómo importar un experimento.
- **Create Run** (Crear experimento): introduzca manualmente los parámetros del experimento. En las instrucciones siguientes se describe la creación de un experimento.

NOTA Los campos de entrada obligatorios de la interfaz del usuario están marcados con un asterisco (*).

Aplicación

1. En la pestaña Planned (Planeados) de la pantalla Runs (Experimentos), seleccione **Create Run** (Crear experimento).
2. Seleccione la aplicación DRAGEN for Illumina DNA Prep with Enrichment Dx y, a continuación, seleccione **Next** (Siguiente).

Ajustes de configuración del experimento

1. En la pantalla Run Settings (Ajustes de configuración del experimento), introduzca un nombre único para el experimento. El nombre del experimento lo identifica desde la secuenciación hasta el análisis.
2. **[Opcional]** Escriba una descripción que ayude a identificar mejor el experimento.
3. Seleccione el kit o los kits de adaptadores indexados utilizados durante la preparación de la biblioteca.
4. Revise el valor de Longitud de lectura y modifíquelo si es necesario. Lectura 1 y Lectura 2 tienen un valor predeterminado de 151 ciclos. Índice 1 e Índice 2 tienen un valor fijo de 10 ciclos y no se pueden modificar.
5. **[Opcional]** Indique un ID de tubo de bibliotecas.
6. Seleccione **Next** (Siguiente).

Datos de la muestra

Los datos de la muestra son Sample ID (ID de la muestra), Well Position (Posición del pocillo) (posición del pocillo de la placa de índices) y Library Name (Nombre de biblioteca). Cuando se utiliza Índice A y B, el valor de Well Position (Posición del pocillo) también incluye Plate identifier (Identificador de la placa).

Hay dos maneras de introducir los datos de la muestra:

- **Import Samples** (Importar muestras): utilice un archivo de plantilla disponible para su descarga en la pantalla Sample Data (Datos de la muestra).
- **Manually** (Manualmente): introduzca los datos de la muestra directamente en la tabla de la pantalla Sample Data (Datos de la muestra).

Importación de las muestras

Al planificar un experimento de secuenciación mediante un navegador en un ordenador conectado a la red, se puede descargar un archivo de plantilla (`*.csv`) de la pantalla Sample Data (Datos de la muestra). El archivo de plantilla no está disponible para su descarga al acceder a Illumina Run Manager a través del software del sistema operativo de NextSeq 550Dx. Para introducir datos de una muestra utilizando la función Import Samples (Importar muestras), realice los siguientes pasos.

NOTA Siga los pasos de Run Settings (Ajustes de configuración del experimento) antes de continuar.

1. Seleccione **Download Template** (Descargar plantilla) para descargar un archivo CSV en blanco.
2. Desde el archivo de plantilla, introduzca los datos de la muestra y, a continuación, guarde el archivo. El nombre de la biblioteca es opcional.

NOTA Cuando se utiliza Índice A y B, los datos de la columna B deben incluir tanto la posición de la placa como la del pocillo (posición del pocillo de la placa de índices). Ejemplo: A-A01, A-A02, A-A03.

3. Seleccione **Import Samples** (Importar muestras) y vaya al archivo de plantilla que contenga los datos de la muestra del paso anterior.
4. Seleccione **Open** (Abrir), **Proceed** (Continuar) y, después, **Next** (Siguiente).

NOTA Cambiar el valor de Sample ID (ID de la muestra) antes de seleccionar Next (Siguiente) puede provocar un error. Finalice la configuración del experimento antes de realizar cambios para evitar errores.

Introducción de las muestras de forma manual

Utilice la tabla que aparece en la pantalla Sample Data (Datos de la muestra) para introducir manualmente los datos de la muestra.

1. Introduzca un ID de muestra único en el campo Sample ID (ID de muestra).
2. Utilice **Well Position** (Posición del pocillo) (Índice A o Índice B) o **Plate - Well Position** (Placa - Posición del pocillo) (Índice A y B) para seleccionar el índice asociado para las muestras.
   Los campos i7 Index (Índice i7), Index 1 (Índice 1), i5 Index (Índice i5) e Index 2 (Índice 2) se rellenan automáticamente.
3. **[Opcional]** Introduzca un nombre de biblioteca.
4. Añada filas y repita los pasos 1–3 según sea necesario hasta que se hayan añadido todas las muestras a la tabla. Puede añadir varias filas a la vez introduciendo primero el número de filas que se añadirán y, a continuación, seleccionando el icono +. También puede eliminar filas seleccionando la casilla situada junto al número de la fila y, a continuación, haciendo clic en el icono de la papelera.
5. Seleccione **Next** (Siguiente).

Configuración de los ajustes del análisis

1. Seleccione el flujo de trabajo de análisis que desee:
   • Generación de FASTQ
   • Generación de FASTQ y VCF para un flujo de trabajo germinal (se requiere archivo de manifiesto)
   • Generación de FASTQ y VCF para un flujo de trabajo somático (se requiere archivo de manifiesto)
2. **[Opcional]** **Generate ORA compressed FASTQs** (Generar FASTQ comprimidos con ORA) está habilitado de forma predeterminada. La compresión FASTQ con ORA comprime sin pérdidas los archivos FASTQ hasta 5x, en comparación con fastq.gz. Desmarque **Generate ORA compressed FASTQs** (Generar FASTQ comprimidos con ORA) si se prefieren datos sin comprimir (fastq.gz).
3. Para los flujos de trabajo germinales y somáticos se requiere un archivo de manifiesto. Utilice el menú desplegable **Manifest File Selection** (Selección de archivo de manifiesto) para seleccionar un archivo de manifiesto. El manifiesto es un archivo BED (`*.bed`) delimitado por tabuladores que especifica los nombres y ubicaciones de las regiones de referencia de interés. Consulte *Archivo de manifiesto*, en la página 9 para obtener más información.
4. **[Opcional]** En los flujos de trabajo somáticos, utilice el menú desplegable **Noise File Selection** (Selección de archivo de ruido) para seleccionar un archivo de ruido sistemático.
   Puede especificarse un archivo BED (`*.bed.gz`) con un nivel de ruido específico para el centro para filtrar el ruido sistemático. Para obtener más información, consulte *Filtrado de ruido (opcional)*, en la página 9.
5. Seleccione **Next** (Siguiente).

Experimento Revisión

1. En la pantalla Review (Revisión), revise la información de Run Settings (Ajustes de configuración del experimento), Sample Data (Datos de la muestra) y Analysis Settings (Configuración de los ajustes del análisis).
2. Seleccione **Save** (Guardar).
   El experimento se guarda en la pestaña Planned (Planeados) de la pantalla Runs (Experimentos).

Configuración

Para ver o cambiar los ajustes de la aplicación DRAGEN for IDPE Dx, primero seleccione el icono Applications (Aplicaciones) en la pantalla principal. A continuación, seleccione la aplicación que desee ver o cambiar. Se requiere una cuenta de administrador para cambiar los ajustes.

Configuración

La pantalla de configuración muestra los siguientes ajustes de la aplicación:

- **Library Prep Kits** (Kits de preparación de bibliotecas): muestra el kit de preparación de bibliotecas predeterminado para la aplicación. Este ajuste no se puede cambiar.
- **Index Adapter Kits** (Kits de adaptadores de índices): muestra el kit de adaptadores de índices predeterminado para la aplicación. Este ajuste no se puede cambiar.
- **Read lengths** (Longitudes de lectura): las longitudes de lectura se han establecido en 151 para la aplicación de forma predeterminada, pero pueden cambiarse durante la creación de un experimento.
- **Manifest and Noise Files** (Archivos de manifiesto y ruido): cargar y cambiar la configuración para los archivos de manifiesto y ruido.
  - Seleccione **Upload File** (Cargar archivo) para cargar archivos para utilizarlos en el análisis.
  - Seleccione el botón de opción **Default** (Predeterminado) para fijar el archivo como archivo de manifiesto o de ruido seleccionado por defecto durante la creación del experimento cuando se selecciona la aplicación.
  - Seleccione la casilla **Enabled** (Habilitado) para que el archivo se muestre en el menú desplegable durante la creación del experimento.

Permisos

Utilice las casillas de la pantalla Permissions (Permisos) para gestionar el acceso de los usuarios a la aplicación.

Archivo de manifiesto

Cuando se utiliza DRAGEN for IDPE Dx, se requiere la introducción de un archivo de manifiesto para los siguientes flujos de trabajo:

- Generación de FASTQ y VCF para un flujo de trabajo germinal
- Generación de FASTQ y VCF para un flujo de trabajo somático

El archivo de manifiesto es un archivo de texto delimitado por tabuladores que utiliza el formato BED (*.bed) y donde se especifican los nombres y ubicaciones de las regiones de referencia de interés. La sección principal del archivo de manifiesto es la sección Regions (Regiones), y debe contener las siguientes columnas de datos:

Columna	Descripción
Nombre	Nombre único especificado por el usuario para el objetivo
Cromosoma	Ubicación del cromosoma (p. ej., chr10, chr5, etc.)
Inicio	Índice en base 1 de la posición inicial del objetivo
Parada	Índice en base 1 de la posición de parada del objetivo
Longitud de la sonda previa	Longitud de la sonda previa. Para la aplicación DRAGEN for IDPE Dx, debe situarse en 0.
Longitud de la sonda siguiente	Longitud de la sonda siguiente. Para la aplicación DRAGEN for IDPE Dx, debe situarse en 0.

NOTA Se requiere un formato de archivo de manifiesto válido para el análisis. DRAGEN detendrá el análisis si el archivo de manifiesto no es válido.

Filtrado de ruido (opcional)

El filtro de ruido sistemático está disponible para la llamada de variantes somáticas, y se puede utilizar para reducir las llamadas de falsos positivos teniendo en cuenta el ruido específico del centro. El archivo de ruido sistemático se genera recogiendo primero aproximadamente 50 muestras normales (preferentemente específicas para el panel, la preparación de la biblioteca y el secuenciador) y, a continuación, la suma de las frecuencias alélicas inferiores al 30 % de cada centro con suficiente cobertura se divide por el número total de muestras (se considera que las frecuencias alélicas superiores al 30 % son variantes germinales, y no ruido). Una vez generados los valores de ruido, se filtrarán las variantes somáticas detectadas en ese centro.

El filtro puede utilizarse en modo Tumor-Normal, pero es especialmente útil en experimentos Tumor-Only (Solo tumor) en los que no está disponible una muestra normal emparejada. El archivo de ruido

sistemático debe utilizar un archivo BED que tenga una extensión de archivo (`*.bed.gz`), y debe incluir cuatro columnas: Cromosoma, Inicio, Final y los niveles de ruido específicos del centro para cada fila. El filtrado del ruido sistemático es opcional.

Resultados del análisis

Los experimentos que estén en curso se muestran en la pestaña Active (Activos). Los experimentos realizados se muestran en la pestaña Completed (Finalizado). DRAGEN for IDPE Dx crea una carpeta de análisis con un nombre único para cada análisis, que es independiente de la carpeta que contiene los datos de secuenciación. La carpeta de análisis guarda la información siguiente:

- Archivo de manifiesto utilizado
- Versión del software
- ID de la muestra
- Total de lecturas alineadas
- Porcentaje de lecturas alineadas por muestra
- Número de SNV llamadas por muestra
- Número de indels llamadas por muestra
- Estadísticas de cobertura

Archivos de resultados de análisis

La ubicación de la carpeta de análisis se especifica mediante el ajuste External Storage for Analysis Results (Almacenamiento externo para los resultados del análisis). Consulte Guía del software Illumina Run Manager para NextSeq 550Dx (n.º de documento 200025239) para obtener más información sobre el ajuste External Storage for Analysis Results (Almacenamiento externo para los resultados del análisis).

En la pantalla Run Details (Detalles del experimento), el campo External Location (Ubicación externa) proporciona la ruta para los datos de secuenciación. El nombre único de la carpeta de análisis figura en el campo Analysis Output Folder (Carpeta de resultados del análisis) en la pantalla Run Details (Detalles del experimento). Los archivos exactos generados dependen del flujo de trabajo de análisis que se utilice. La aplicación genera los siguientes archivos de resultados del análisis.

NOTA Si se produce un error de limitación de longitud máxima de la ruta de archivo al acceder a los archivos de resultados del análisis, intente mover el archivo a una ubicación de ruta más corta o utilice un método diferente para abrir el archivo.

Archivo de resultados	Descripción
Informe de resumen de variantes (`*.pdf`)	Contiene un resumen de la información del archivo, las versiones del software, la información de la muestra, las estadísticas de lectura, así como las SNV, las inserciones, las deleciones y los resúmenes de cobertura. Solo los flujos de trabajo germinales y somáticos dan lugar a informes de variantes.
FASTQ (`*.fastq.gz` o `*.fastq.ora`)	Son archivos intermedios que contienen las puntuaciones de calidad de las llamadas a las bases. Los archivos FASTQ son la entrada principal para el paso de alineación. Cuando se selecciona la compresión ORA, se utiliza la extensión de archivo `*.fastq.ora`.
Archivos de alineación BAM (`*.bam`)	Contiene lecturas alineadas para una muestra dada.
Archivos de genoma VCF (`*.gvcf.gz`)	Contiene el genotipo de cada posición, ya sea llamada como variante o como referencia.
Archivos VCF (`*.vcf.gz`)	Contiene variantes llamadas en cada posición.
Informe de mediciones del experimento (`*.csv`)	Contiene mediciones de calidad sobre el experimento, incluyendo el rendimiento total no indexado y la puntuación Q30.

Archivos FASTQ

El formato FASTQ (*.fastq.gz, *.fastq.ora) es un formato de archivo de texto que contiene las llamadas de bases y los valores de calidad por lectura. Cada archivo contiene la siguiente información:

- El identificador de la muestra
- La secuencia
- Una puntuación de calidad según la escala Phred en formato encriptado ASCII + 33

El identificador de la muestra presenta el siguiente formato:

```
@Instrumento:IDExperimento:IDCeldaFlujo:Carril:Placa:X:Y
NúmeroLectura:IndicadorFiltro:0:NúmeroMuestra
Ejemplo:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=
```

Archivos BAM

Un archivo BAM (*.bam) es la versión binaria comprimida de un archivo SAM (mapa de alineación de secuencias) que se utiliza para representar secuencias alineadas de hasta 128 Mb. Los archivos BAM utilizan el formato de nomenclatura de archivos `SampleName_S#.bam`. # es el número de muestra determinado por el orden en que aparecen las muestras para el experimento. En modo multinodo, S# se fija en S1, con independencia del orden de la muestra.

Los archivos BAM contienen una sección de encabezado y una sección de alineación:

- **Encabezado**: contiene información sobre todo el archivo, como el nombre de la muestra, la longitud de la muestra y el método de alineación. Las alineaciones en la sección de alineaciones están asociadas con información específica en la sección de encabezado.
- **Alineaciones**: contiene el nombre de la lectura, la secuencia de lectura, la calidad de lectura, la información de alineación y marcadores personalizados. El nombre de lectura incluye el cromosoma, la coordenada de inicio, la calidad de alineación y la cadena del descriptor de coincidencias.

La sección de alineaciones incluye la siguiente información para cada lectura o par de lectura:

- AS: Calidad de la alineación “Paired-end”.
- RG: Grupo de lectura, que indica el número de lecturas para una muestra específica.
- BC: Marcador de código de barras, que indica el ID de muestra demultiplexado asociado a la lectura.
- SM: Calidad de la alineación “Single-end”.
- XC: Cadena del descriptor de coincidencias.
- XN: Marcador de nombre del amplicón, que registra el ID del amplicón asociado con la lectura.

Los archivos de índice BAM (*.bam.bai) facilitan un índice del archivo BAM correspondiente.

Archivos VCF

Los archivos de formato de llamada de variantes (*.vcf) contienen información sobre las variantes que se encuentran en posiciones específicas de un genoma de referencia.

El encabezado del archivo VCF incluye la versión de formato del archivo VCF, la versión del llamador de variantes y detalla las anotaciones utilizadas en el resto del archivo. El encabezado del VCF incluye también el archivo del genoma de referencia y el archivo BAM. La última línea del encabezado contiene los encabezados de las columnas para las líneas de datos. Cada una de las líneas de datos del archivo VCF contiene información sobre una sola variante.

Tabla 1 Encabezados del archivo VCF

Encabezado	Descripción
CHROM	El cromosoma del genoma de referencia. Los cromosomas aparecen en el mismo orden que en el archivo FASTA de referencia.
POS	La posición de base individual de la variante en el cromosoma de referencia. Para las variantes de nucleótido único (SNV), esta posición es la base de referencia con la variante. Para las indels, esta posición es la base de referencia que precede inmediatamente a la variante.
ID	El número rs (SNP de referencia) para el SNP obtenido de `dbSNP.txt`, si es aplicable. Si existen varios números rs en esta ubicación, la lista se delimita con punto y coma. Si no existe una entrada dbSNP en esta posición, se utiliza un marcador de valor que falta ('.').
REF	El genotipo de referencia. Por ejemplo, una deleción de una sola T se representa como TT de referencia y T alternativa. Una variante de nucleótido único de A a T se representa como A de referencia y T alternativa.
ALT	Los alelos que difieren de la lectura de referencia. Por ejemplo, una inserción de una sola T se representa como A de referencia y AT alternativa. Una variante de nucleótido único de A a T se representa como A de referencia y T alternativa.
QUAL	Una puntuación de calidad en la escala Phred asignada por el llamador de variantes. Puntuaciones más altas indican mayor confianza en la variante y una menor probabilidad de errores. Para una puntuación de calidad de Q, la probabilidad estimada de un error es 10-(Q/10). Por ejemplo, el conjunto de llamadas Q30 tiene una tasa de errores del 0,1 %. Muchos llamadores de variantes asignan puntuaciones de calidad a partir de sus modelos estadísticos, que son altas con relación a la tasa de errores observada.

Tabla 2 Anotaciones de los archivos VCF de flujos de trabajo germinales

Encabezado	Descripción
FILTER (Filtro)	Si se superan todos los filtros, aparece PASS (Apto) en la columna del filtro. Las posibles entradas de Filtro incluyen: • **DRAGENSnpHardQUAL**: se aplica si la puntuación QUAL de una variante SNP no alcanza el umbral • **DRAGENIndelHardQUAL**: se aplica si la puntuación QUAL de una variante indel no alcanza el umbral • **LowDepth**: sitio filtrado porque la profundidad de cobertura no alcanza el umbral • **LowGQ**: sitio filtrado porque la calidad del genotipo no alcanza el umbral • **PloidyConflict**: la llamada de genotipo del llamador de variantes no es coherente con la ploidía de cromosomas • **base_quality**: sitio filtrado porque la mediana de la calidad de las bases de las lecturas alternativas en este locus no alcanza el umbral • **filtered_reads**: sitio filtrado porque una fracción de las lecturas demasiado grande se ha excluido. • **fragment_length**: sitio filtrado porque la diferencia absoluta entre la mediana de la longitud de los fragmentos de las lecturas alternativas y la mediana de la longitud de los fragmentos de las lecturas de referencia en este locus supera el umbral • **low_depth**: sitio filtrado porque la profundidad de lectura es demasiado baja • **low_frac_info_reads**: sitio filtrado porque la fracción de lecturas informativas está por debajo del umbral • **low_normal_depth**: sitio filtrado porque la profundidad de lectura de las muestras normales es demasiado baja • **long_indel**: sitio filtrado porque la longitud de indel es demasiado larga • **mapping_quality**: sitio filtrado porque la mediana de la calidad de la cartografía de las lecturas alternativas en este locus no alcanza el umbral • **multiallelic**: sitio filtrado porque más de dos alelos alternativos pasan el LOD tumoral • **non_homref_normal**: sitio filtrado porque el genotipo de la muestra normal no es una referencia homocigótica • **no_reliable_supporting_read**: sitio filtrado porque no existe una lectura somática de apoyo fiable • **panel_of_normals**: observado en al menos una muestra en el vcf del panel de normales • **read_position**: sitio filtrado porque la mediana de las distancias entre el inicio y el fin de la lectura en este locus está por debajo del umbral • **RMxNRepeatRegion**: sitio filtrado porque todo o parte del alelo de la variante es una repetición de la referencia • **strand_artifact**: sitio filtrado debido a cortes importantes en la cadena

Encabezado	Descripción
FILTER (Filtro)	• **str_contraction**: sitio filtrado debido a la sospecha de un error de PCR en donde el alelo alternativo es una unidad de repetición menor que la referencia • **too_few_supporting_reads**: sitio filtrado porque hay demasiadas pocas lecturas de apoyo en la muestra tumoral • **weak_evidence**: la puntuación de la variante somática no alcanza el umbral
INFO	Las posibles entradas de INFO incluyen: • **AC**: número de alelos en los genotipos para cada alelo alternativo, en el mismo orden en que aparecen. • **AF**: frecuencia alélica para cada alelo alternativo, en el mismo orden en que aparecen. • **AN**: número total de alelos en los genotipos llamados. • **DB**: miembro de dbSNP. • **FS**: valor de p en la escala Phred utilizando la prueba exacta de Fisher para detectar cortes en la cadena. • **QD**: confianza de la variante/calidad por profundidad. • **R2_5P_bias**: puntuación basada en el sesgo de emparejamiento (“mate bias”) y la distancia desde el extremo de cebado en 5. • **SOR**: cociente de posibilidades simétrico de una tabla de contingencia de 2x2 para detectar cortes en la cadena. • **DP**: profundidad aproximada de lectura (informativa y no informativa); algunas lecturas pueden haber sido filtradas según mapq, etc. • **END**: posición de parada del intervalo. • **FractionInformativeReads**: la fracción de lecturas informativas con respecto al número total de lecturas. • **MQ**: calidad de la cartografía RMS. • **MQRankSum**: puntuación Z de la prueba de suma de rangos de Wilcoxon de las cualidades de cartografía de las lecturas alternativas frente a las de referencia. • **ReadPosRankSum**: puntuación Z de la prueba de suma de rangos de Wilcoxon del sesgo de posición de las lecturas alternativas frente a las de referencia. • **SOMATIC**: al menos una variante en esta posición es somática.

Encabezado	Descripción
FORMAT (Formato)	La columna Formato muestra campos separados por dos puntos. Por ejemplo, GT:GQ. Los campos disponibles incluyen: • **AD**: profundidades alélicas (contando solo las lecturas informativas del total de lecturas) para los alelos de referencia y alternativos en el orden en que aparecen. • **AF**: fracciones de alelos para los alelos alternativos en el orden en que aparecen. • **DP**: profundidad aproximada de lectura (las lecturas con MQ = 255 o con malas parejas se filtran). • **F1R2**: número de lecturas en orientación de pares F1R2 que apoyan cada alelo. • **F2R1**: número de lecturas en orientación de pares F2R1 que apoyan cada alelo. • **GT**: genotipo. 0 corresponde a la base de referencia, 1 corresponde a la primera entrada en la columna ALT, etc. La barra inclinada (/) indica que no hay disponible información sobre la fase de hebra retrasada. • **MB**: estadísticas de los componentes por muestra para detectar el sesgo de emparejamiento. • **PS**: información de ID de fase de hebra retrasada física, en donde cada ID único en una muestra dada (pero no entre muestras) conecta los registros dentro de un grupo en fase de hebra retrasada. • **SB**: estadísticas de componentes por muestra que incluyen la prueba exacta de Fisher para detectar cortes en la cadena. • **SQ**: calidad somática.
SAMPLE (Muestra)	La columna Muestra da los valores especificados en la columna Formato.

Tabla 3 Anotaciones de los archivos VCF de flujos de trabajo somáticos

Encabezado	Descripción
FILTER (Filtro)	Si se superan todos los filtros, aparece PASS (Apto) en la columna del filtro. Las posibles entradas de Filtro incluyen: • **base_quality**: sitio filtrado porque la mediana de la calidad de las bases de las lecturas alternativas en este locus no alcanza el umbral • **filtered_reads**: sitio filtrado porque una fracción de las lecturas demasiado grande ha sido filtrada • **fragment_length**: sitio filtrado porque la diferencia absoluta entre la mediana de la longitud de los fragmentos de las lecturas alternativas y la mediana de la longitud de los fragmentos de las lecturas de referencia en este locus supera el umbral • **low_depth**: sitio filtrado porque la profundidad de lectura es demasiado baja • **low_frac_info_reads**: sitio filtrado porque la fracción de lecturas informativas está por debajo del umbral • **low_normal_depth**: sitio filtrado porque la profundidad de lectura de las muestras normales es demasiado baja • **long_indel**: sitio filtrado porque la longitud de indel es demasiado larga • **mapping_quality**: sitio filtrado porque la mediana de la calidad de la cartografía de las lecturas alternativas en este locus no alcanza el umbral • **multiallelic**: sitio filtrado porque más de dos alelos alternativos pasan el LOD tumoral • **non_homref_normal**: sitio filtrado porque el genotipo de la muestra normal no es una referencia homocigótica • **no_reliable_supporting_read**: sitio filtrado porque no existe una lectura somática de apoyo fiable • **panel_of_normals**: observado en al menos una muestra en el vcf del panel de normales • **read_position**: sitio filtrado porque la mediana de las distancias entre el inicio y el fin de la lectura en este locus está por debajo del umbral • **RMxNRepeatRegion**: sitio filtrado porque todo o parte del alelo de la variante es una repetición de la referencia • **strand_artifact**: sitio filtrado debido a cortes importantes en la cadena • **str_contraction**: sitio filtrado debido a la sospecha de un error de PCR en donde el alelo alternativo es una unidad de repetición menor que la referencia • **too_few_supporting_reads**: sitio filtrado porque hay demasiadas pocas lecturas de apoyo en la muestra tumoral • **weak_evidence**: la puntuación de la variante somática no alcanza el umbral • **systematic_noise**: sitio filtrado a partir de las pruebas de que hay ruido sistemático en las muestras normales.

Encabezado	Descripción
INFO	Las posibles entradas de INFO incluyen: • **DP**: profundidad aproximada de lectura (informativa y no informativa); algunas lecturas pueden haber sido filtradas según mapq, etc. • **END**: posición de parada del intervalo. • **FractionInformativeReads**: la fracción de lecturas informativas con respecto al número total de lecturas. • **MQ**: calidad de la cartografía RMS. • **MQRankSum**: puntuación Z de la prueba de suma de rangos de Wilcoxon de las cualidades de cartografía de las lecturas alternativas frente a las de referencia. • **ReadPosRankSum**: puntuación Z de la prueba de suma de rangos de Wilcoxon del sesgo de posición de las lecturas alternativas frente a las de referencia. • **AQ**: puntuación de ruido sistemático. • **hotspot**: sitio somático conocido, utilizado para aumentar la confianza en la llamada. • **SOMATIC**: al menos una variante en esta posición es somática.

Encabezado	Descripción
FORMAT (Formato)	La columna Formato muestra campos separados por dos puntos. Por ejemplo, GT:GQ. Los campos disponibles incluyen: • **AD**: profundidades alélicas (contando solo las lecturas informativas del total de lecturas) para los alelos de referencia y alternativos en el orden en que aparecen. • **AF**: fracciones de alelos para los alelos alternativos en el orden en que aparecen. • **DP**: profundidad aproximada de lectura (las lecturas con MQ = 255 o con malas parejas se filtran). • **F1R2**: número de lecturas en orientación de pares F1R2 que apoyan cada alelo. • **F2R1**: número de lecturas en orientación de pares F2R1 que apoyan cada alelo. • **GP**: probabilidades posteriores en la escala Phred para los genotipos según se define en la especificación VCF. • **GQ**: calidad de genotipos. • **GT**: genotipo. 0 corresponde a la base de referencia, 1 corresponde a la primera entrada en la columna ALT, etc. La barra inclinada (/) indica que no hay disponible información sobre la fase de hebra retrasada. • **MB**: estadísticas de los componentes por muestra para detectar el sesgo de emparejamiento. • **PL**: verosimilitudes normalizadas en la escala Phred para los genotipos según se define en la especificación VCF. • **PRI**: probabilidades anteriores en la escala Phred para los genotipos. • **PS**: información de ID de fase de hebra retrasada física, en donde cada ID único en una muestra dada (pero no entre muestras) conecta los registros dentro de un grupo en fase de hebra retrasada. • **SB**: estadísticas de componentes por muestra que incluyen la prueba exacta de Fisher para detectar cortes en la cadena. • **SQ**: calidad somática.
SAMPLE (Muestra)	La columna Muestra da los valores especificados en la columna Formato.

Archivos VCF de genoma

Los archivos VCF (*.gvcf.gz) de genoma siguen una serie de convenciones para representar todos los sitios dentro del genoma en un formato razonablemente compacto. Los archivos gVCF incluyen todos los sitios de la región de interés en un solo archivo para cada muestra. El archivo gVCF muestra ausencia de llamadas en las posiciones que no superan todos los filtros. Un marcador ./. de genotipo (GT) indica una ausencia de llamadas.

Volver a poner un análisis en cola

Es posible que necesite volver a poner en cola un análisis si lo ha detenido, en caso de que haya fallado o si quiere volver a analizar el experimento modificando los ajustes de configuración. Para volver a poner un análisis en cola, realice estos pasos:

1. En la pantalla Run (Experimento), seleccione la pestaña Completed (Finalizado) y, a continuación, seleccione el nombre del experimento que desee volver a analizar.
   Si anteriormente se ejecutó Requeue Analysis (Volver a poner un análisis en cola), seleccione el nombre del Parent Run (Experimento matriz).
2. En la pantalla Run Details (Detalles del experimento), después de Sequencing Information (Información de la secuenciación), seleccione **Requeue Analysis** (Volver a poner un análisis en cola).
3. Seleccione una opción:
   - Requeue analysis with no changes (Volver a poner el análisis en cola sin cambios)
   - Edit run settings and requeue analysis (Editar los ajustes de configuración del experimento y volver a poner el análisis en cola)
   - Requeue analysis with a different application (Volver a poner el análisis en cola con una aplicación diferente)
4. Confirme que la ubicación donde se encuentren en ese momento los datos de secuenciación figure en el campo **Sequencing data file path** (Ruta de archivo de datos de secuenciación).

NOTA La ruta hacia los datos de secuenciación debe coincidir con la que aparezca en el ajuste External Storage for Analysis Results (Almacenamiento externo para los resultados del análisis). Consulte Guía del software Illumina Run Manager para NextSeq 550Dx (n.º de documento 200025239) para obtener información sobre cómo cambiar la ruta de almacenamiento externo.

5. Introduzca una respuesta en Reanalysis Reason (Motivo de la repetición del análisis).
6. Seleccione **Requeue Analysis** (Volver a poner un análisis en cola).
7. Efectúe los cambios deseados en Run Settings (Ajustes de configuración del experimento), Sample Data (Datos de la muestra) y Analysis Settings (Configuración de los ajustes del análisis).
8. Seleccione **Save** (Guardar). El análisis se iniciará teniendo en cuenta los parámetros de análisis actuales.

Asistencia técnica

Si necesita asistencia técnica, póngase en contacto con el servicio de asistencia técnica de Illumina.

Sitio web:	www.illumina.com
Correo electrónico:	techsupport@illumina.com

Hojas de datos de seguridad (SDS): Disponibles en el sitio web de Illumina, support.illumina.com/sds.html.

Documentación del producto: Disponible para su descarga de support.illumina.com.

Illumina
5200 Illumina Way
San Diego, California 92122 (EE. UU.)
+1 800 809 ILMN (4566)
+1 858 202 4566 (fuera de Norteamérica)
techsupport@illumina.com
www.illumina.com

IVD

EC REP

Illumina Netherlands B.V.
Steenoven 19
5626 DK Eindhoven
The Netherlands

Promotor australiano
Illumina Australia Pty Ltd
Nursing Association Building
Level 3, 535 Elizabeth Street
Melbourne, VIC 3000
Australia

PARA USO DIAGNÓSTICO IN VITRO.

illumina®