illumina®

DRAGEN for DNA Prep with Enrichment Dx su NextSeq 550Dx di Illumina

Guida per l'utente dell'applicazione

DI PROPRIETÀ DI ILLUMINA
Documento n. 200025238 v00
Febbraio 2023

PER USO DIAGNOSTICO IN VITRO.

Cronologia revisioni

Documento	Data	Descrizione della modifica
200025238 v00	Febbraio 2023	Versione iniziale.

Sommario

Descrizione generale

L'DRAGEN for DNA Prep with Enrichment Dx di Illumina applicazione (DRAGEN per IDPE Dx) viene utilizzata per pianificare ed eseguire l'analisi secondaria delle biblioteche IDPE Dx generate per il sequenziamento sul NextSeq 550Dx.

DRAGEN per IDPE Dx supporta il sequenziamento all'analisi quando utilizzato con la DNA Prep with Enrichment Dx di Illumina preparazione della libreria, NextSeq 550Dx e Illumina DRAGEN Server per NextSeq 550Dx.

Metodi di analisi

DRAGEN per IDPE Dx esegue il demultiplex, la generazione di file FASTQ, la mappatura delle letture, l'allineamento a un genoma di riferimento e l'identificazione di varianti piccole, a seconda del flusso di lavoro selezionato:

- Generazione di file FASTQ
- Generazione di file FASTQ e VCF per Germline
- Generazione di file FASTQ e VCF per Somatic

NOTA La compressione ORA è disponibile per l'uso con tutti e tre i flussi di lavoro. La compressione DRAGEN ORA è un software di compressione completamente senza perdita che crea un file con un'estensione Original Read Archive (*.ora). Il formato ora è un formato di compressione basato su riferimento per i file FASTQ ed è progettato per una compressione/decompressione molto veloce e un elevato rapporto di compressione.

Generazione FASTQ

Le sequenze assemblate vengono scritte in file FASTQ per ogni campione. I file FASTQ sono file di testo che contengono i dati di sequenziamento e i punteggi qualitativi di un solo campione. Per ogni campione, vengono generati file FASTQ separati per ogni corsia della cella a flusso e per ogni lettura di sequenziamento. Il nome del campione specificato durante l'impostazione della corsa è incluso nel nome del file FASTQ. I file FASTQ rappresentano gli input principali per l'allineamento. La prima fase della generazione di FASTQ è il demultiplex. Il demultiplex assegna i cluster che passano il filtro a un campione confrontando ogni sequenza di lettura indici con le sequenze degli indici specificate per la corsa. In questa fase non vengono considerati i valori qualitativi. Le letture indici vengono identificate tramite le seguenti fasi:

- I campioni sono numerati a partire da 1 in base all'ordine in cui sono stati elencati per la corsa.
- Il numero del campione 0 è riservato ai cluster non assegnati a un campione.

- I cluster sono assegnati a un campione quando la sequenza d'indice corrisponde esattamente o quando è presente una sola mancata corrispondenza per la lettura indici.

Il software include la compressione ORA per comprimere i file FASTQ. Questo formato può essere abilitato facoltativamente. Quando si utilizza il formato ORA (*.ora), il checksum md5 del contenuto del file FASTQ viene conservato dopo un ciclo di compressione e decompressione per garantire una compressione senza perdite.

Mappatura e allineamento del DNA

Dopo la generazione di file FASTQ, le letture sono mappate e allineate a un genoma di riferimento. La prima fase della mappatura consiste nel generare seed dalla lettura, per poi cercare le corrispondenze esatte nel genoma di riferimento. Questi risultati vengono poi perfezionati eseguendo allineamenti completi di Smith-Waterman sulle posizioni con la più alta densità di corrispondenze di seed. Questo algoritmo ben documentato agisce confrontando ogni posizione della lettura con tutte le posizioni candidate del riferimento. Questi confronti corrispondono a una matrice di potenziali allineamenti tra la lettura e il riferimento. Per ognuna di queste posizioni di allineamento candidate, lo Smith-Waterman genera dei punteggi che vengono utilizzati per valutare se il miglior allineamento che passa attraverso quella cella della matrice la raggiunge attraverso un match o un mismatch nucleotidico (movimento diagonale), una delezione (movimento orizzontale) o un'inserzione (movimento verticale). Una corrispondenza tra lettura e riferimento fornisce un bonus sul punteggio, mentre una mancata corrispondenza o un indel impongono una penalità. L'allineamento scelto è quello che ottiene il punteggio più alto nella matrice. L'algoritmo è accelerato dall'hardware sulle schede di matrici di porte logiche programmabili sul campo (FPGA) DRAGEN. Il genoma di riferimento utilizzato nell'app viene creato dall'UCSC hg19 FASTA con l'opzione DRAGEN per creare una tabella hash alt-aware basata sul liftover.

Identificazione della variante per Germline DRAGEN

L'identificatore di piccole varianti per Germline DRAGEN assume come input le letture di DNA mappate e allineate e richiama polimorfismi a singolo nucleotide (SNP) e inserzioni o delezioni (indel) attraverso una combinazione di rilevamento in colonna e assemblaggio locale *de novo* degli aplotipi. Per abilitare l'identificatore di piccole varianti per Germline DRAGEN, selezionare il flusso di lavoro della variante della linea germinale.

L'identificazione di varianti per linea germinale viene in genere utilizzata per i campioni di linea germinale in cui si sa che la ploidia è due. Vengono innanzitutto identificate le regioni di riferimento identificabili con una copertura di allineamento sufficiente. All'interno di queste regioni di riferimento, una rapida scansione delle letture ordinate identifica le regioni attive, che sono centrate sulle colonne di accumulo con evidenza di una variante. Le regioni attive sono riempite con un contesto sufficiente a coprire i contenuti significativi non di riferimento nelle vicinanze. Se c'è evidenza di indel, le regioni attive ricevono un riempimento aggiuntivo.

Le letture allineate vengono sottoposte a clipping all'interno di ciascuna regione attiva e assemblate in un grafico di De Bruijn. I margini delle letture sottoposte a clipping sono ponderati in base ai conteggi delle osservazioni, in cui la sequenza di riferimento è la struttura portante. Dopo una certa pulizia e semplificazione del grafico, tutti i percorsi source-to-sink vengono estratti come aplotipi candidati. Ogni aplotipo viene allineato con il Smith-Waterman al genoma di riferimento per identificare le varianti che rappresenta. Questo insieme di eventi può essere integrato da un rilevamento basato sulla posizione. Per ogni coppia lettura-aplotipo, la probabilità P(r|H) di osservare la lettura, presumendo che l'aplotipo sia il vero campione di partenza, viene stimata utilizzando un modello di Markov nascosto (HMM) a coppie.

Esaminando la regione attiva in base alla posizione di riferimento, i genotipi candidati sono formati da combinazioni diploidi di eventi varianti (SNP o indel). Per ogni evento (incluso il riferimento), la probabilità condizionata P(r|e) di osservare ogni lettura sovrapposta è stimata come il massimo P(r|H) per gli aplotipi che supportano l'evento. Queste vengono combinate nella probabilità condizionale P (r|e1e2) per un genotipo (coppia di eventi) e moltiplicate per ottenere la probabilità condizionale P (R|e1e2) di osservare l'intero accumulo di letture. Utilizzando la Formula di Bayes, si calcola la probabilità posteriore P(e1e2|R) di ciascun genotipo diploide e si definisce quello prescelto.

DRAGEN per IDPE Dx applica il filtraggio automatico. Per ulteriori informazioni, fare riferimento alle *Annotazioni del file VCF del flusso di lavoro della linea digerente* alla pagina 14.

Identificazioni di varianti per Somatic DRAGEN

L'identificatore di varianti per Somatic DRAGEN prende in input le letture di DNA mappate e allineate e identifica SNV e indel attraverso l'assemblaggio locale *de novo* di aplotipi in una regione attiva. Per abilitare l'identificatore di piccole varianti per Somatic DRAGEN, selezionare un'applicazione per varianti somatiche.

L'identificazione di varianti somatiche viene utilizzata generalmente per i campioni tumorali. Con questo flusso di lavoro, DRAGEN non fa alcuna ipotesi di ploidia, e ciò consente il rilevamento di alleli a bassa frequenza. Per i loci con copertura fino a 100x nel campione tumorale, DRAGEN ha una soglia di rilevamento alle frequenze alleliche varianti del 5%. Il limite aumenta con una profondità maggiore in base al locus e si dimezza ogni volta che la copertura raddoppia oltre 100x. Vengono innanzitutto identificate le regioni di riferimento identificabili con una copertura di allineamento sufficiente. All'interno di queste regioni di riferimento, una scansione delle letture ordinate identifica le regioni attive, che sono centrate intorno sulle colonne di accumulo con evidenza di una variante nelle letture tumorali. Le regioni attive sono riempite con un contesto sufficiente a coprire i contenuti significativi non di riferimento nelle vicinanze. Se c'è evidenza di indel, le regioni attive ricevono un riempimento aggiuntivo.

Le letture allineate vengono sottoposte a clipping all'interno di ciascuna regione attiva e assemblate in un grafico di De Bruijn. I margini delle letture sottoposte a clipping sono ponderati in base ai conteggi delle osservazioni, in cui la sequenza di riferimento è la struttura portante. Dopo una certa pulizia e semplificazione del grafico, tutti i percorsi source-to-sink vengono estratti come aplotipi candidati. Ogni

aplotipo viene allineato con il Smith-Waterman al genoma di riferimento per identificare le varianti che rappresenta. Per ogni coppia lettura-aplotipo, la probabilità P(r|H) di osservare la lettura viene stimata utilizzando un modello di Markov nascosto (HMM) a coppie, assumendo che l'aplotipo sia il vero campione di partenza.

Per determinare il punteggio di limite di rilevamento del tumore (TLOD), l'identificatore di piccole varianti per Somatic DRAGEN esegue prima la scansione in base alla posizione di riferimento per ogni evento somatico candidato, nonché per l'evento di riferimento sulla regione attiva. La probabilità condizionata P(r|e) di osservare ogni lettura sovrapposta è stimata come il massimo P(r|H) per gli aplotipi che supportano l'evento. Queste vengono combinate nella probabilità condizionale P(r|E) per un'ipotesi di evento, E, che coinvolge una miscela dell'allele somatico di riferimento e di quello candidato in un intervallo di possibili frequenze alleliche e moltiplicate per ottenere la probabilità condizionale P(R|E) di osservare l'accumulo dell'intera lettura. Da qui, viene calcolato un punteggio TLOD come prova della presenza di un allele ALT nel campione di tumore in un determinato locus.

DRAGEN per IDPE Dx applica il filtraggio automatico. Per ulteriori informazioni, fare riferimento alle *Annotazioni del file VCF del flusso di lavoro somatico* alla pagina 17.

Creare una corsa pianificata

Utilizzare i passaggi seguenti per impostare una corsa in Illumina Run Manager su NextSeq 550Dx o utilizzando un browser su un computer in rete. Utilizzare un browser su un computer in rete se si desidera l'importazione di dati campione. Fare riferimento a Guida al software Illumina Run Manager per NextSeq 550Dx (documento n. 200025239) per istruzioni sull'accesso a Illumina Run Manager da un computer in rete.

Vi sono due modi diversi per creare una nuova corsa pianificata:

- **Import Run** (Importa corsa): utilizzare un foglio campioni da una corsa esistente come modello per una nuova corsa. Fare riferimento a Guida al software Illumina Run Manager per NextSeq 550Dx (documento n. 200025239) per informazioni su come importare una corsa.
- **Create Run** (Crea corsa): immettere manualmente i parametri di esecuzione. Le seguenti istruzioni descrivono la modalità di creazione di una corsa.

NOTA I campi di immissione obbligatori nell'interfaccia utente sono contrassegnati da un asterisco (*).

Applicazione

1. Dalla scheda Planned (Pianificate) della schermata Runs (Corse), selezionare **Create Run** (Crea corsa).
2. Selezionare l'applicazione DRAGEN per DNA Prep with Enrichment Dx di Illumina, quindi selezionare **Next** (Avanti).

Impostazioni della corsa

1. Sulla schermata Run Settings (Impostazioni corsa), immettere un nome univoco per la corsa. Il nome della corsa identifica la corsa dal sequenziamento per tutta l'analisi.
2. **[Facoltativo]** Immettere una descrizione per identificare la corsa.
3. Selezionare il kit (o i kit) adattatore dell'indice utilizzato/i durante la preparazione della libreria.
4. Rivedere la Lunghezza Lettura e modificare se necessario. Lettura 1 e Lettura 2 hanno un valore predefinito di 151 cicli. Index 1 (Indice 1) e Index 2 (Indice 2) hanno un valore fisso di 10 cicli e non possono essere modificati.
5. **[Facoltativo]** Immettere l'ID di una provetta della libreria.
6. Selezionare **Next** (Avanti).

Dati del campione

I dati del campione includono l'ID campione, la posizione del pozzetto (posizione del pozzetto della piastra dell'indice) e il nome della libreria. Quando si utilizza l'indice A&B, la posizione del pozzetto include anche l'identificativo della piastra.

Vi sono due modi per inserire i dati del campione:

- **Import Samples** (Importa campioni): utilizzare un file modello disponibile per il download nella schermata Sample Data (Dati Campione).
- **Manually** (Manualmente): immettere i dati del campione direttamente nella tabella nella schermata Sample Data (Dati campione).

Importazione dei campioni

Un file modello (`*.csv`) è disponibile per il download nella schermata Sample Data (Dati campione) quando si pianifica una corsa di sequenziamento utilizzando un browser su un computer in rete. Il file modello non è disponibile per il download quando si accede Illumina Run Manager tramite il software del sistema NextSeq 550Dx operativo. Per immettere i dati del campione utilizzando la funzione Import Samples (Importa campioni), procedere come segue.

NOTA Completare i passaggi Run Settings (Impostazioni corsa) prima di procedere.

1. Selezionare **Download Template** (Scarica modello) per scaricare un file CSV vuoto.
2. Dal file modello, immettere i dati del campione, quindi salvare il file. Il nome della libreria è facoltativo.

NOTA Quando si utilizza l'indice A&B, i dati per la colonna B devono includere sia la piastra che la posizione del pozzetto (indice piastra posizione del pozzetto). Esempio: A-A01, A-A02, A-A03.

3. Selezionare **Import Samples** (Importa campioni) e passare al file modello contenente le informazioni sui dati del campione del passaggio precedente.
4. Selezionare **Open** (Apri), **Proceed** (Procedere), e poi **Next** (Avanti).

NOTA La modifica dell'ID campione prima di selezionare Next (Avanti) può causare un errore. Per evitare errori, terminare l'impostazione della corsa prima di apportare modifiche.

Immissione manuale dei campioni

Utilizzare la tabella della schermata Sample Data (Dati del campione) per immettere manualmente i dati del campione.

1. Immettere un ID campione univoco nel campo Sample ID (ID campione).
2. Utilizzare **Well Position** (Posizione pozzetto) (indice A o indice B) o **Plate - Well Position** (Piastra - Posizione pozzetto) (indice A&B) per selezionare l'indice associato per i campioni.
   I campi i7 Index (Indice i7), Index 1 (Indice 1), i5 Index (Indice i5) e Index 2 (Indice 2) vengono compilati automaticamente.
3. **[Facoltativo]** Immettere il nome di una libreria.
4. Aggiungere righe e ripetere i passaggi 1–3 secondo necessità fino a quando tutti i campioni sono stati aggiunti alla tabella. È possibile aggiungere più righe contemporaneamente inserendo prima il numero di righe da aggiungere, quindi selezionando l'icona +. È inoltre possibile rimuovere le righe selezionando la casella accanto al numero di riga, quindi facendo clic sull'icona del Cestino.
5. Selezionare **Next** (Avanti).

Impostazioni di analisi

1. Selezionare il flusso di lavoro di analisi desiderato:
   • Generazione di file FASTQ
   • Generazione di file FASTQ e VCF per un flusso di lavoro della linea germinale (è necessario un file manifest)
   • Generazione di file FASTQ e VCF per un flusso di lavoro somatico (è necessario un file manifest)
2. **[Facoltativo]** **Generate ORA compressed FASTQs** (Generare file FASTQ con compressione ORA) è abilitato per impostazione predefinita. La compressione ORA dei file FASTQ comprime senza perdita i file FASTQ fino a 5 volte rispetto a fastq.gz. Deselezionare **Generate ORA compressed FASTQs** (Generare file FASTQ con compressione ORA) se si preferiscono dati non compressi (fastq.gz).
3. Per i flussi di lavoro di linea germinale e somatica, è necessario un file manifest. Usare il menu a discesa **Manifest File Selection** (Selezione file manifest) per selezionare un file manifest. Manifest è un file BED (`*.bed`) delimitato da tabulazioni che specifica i nomi e le posizioni delle regioni di riferimento mirate. Per ulteriori informazioni, consultare *File manifest* alla pagina 8.
4. **[Facoltativo]**Per i flussi di lavoro somatici, utilizzare il menu a discesa **Noise File Selection** (Selezione file di rumore) per selezionare un file di rumore sistematico.
   È possibile specificare un file BED (`*.bed.gz`) con un livello di rumore specifico per il sito per filtrare il rumore sistematico. Per maggiori informazioni, consultare *Filtro del rumore (facoltativo)* alla pagina 9.
5. Selezionare **Next** (Avanti).

Corsa Revisione

1. Nella schermata Review (Revisione), rivedere le informazioni per Run Settings (Impostazioni corsa), Sample Data (Dati campione) e Analysis Settings (Impostazioni analisi).
2. Selezionare **Save** (Salva).
   La corsa viene salvata nella scheda Planned (Pianificate) della schermata Runs (Corse).

Impostazioni

Per visualizzare o modificare le impostazioni dell'applicazione DRAGEN per IDPE Dx, selezionare prima l'icona Applications (Applicazioni) dalla schermata principale. Quindi selezionare l'applicazione che si desidera visualizzare o modificare. Per modificare le impostazioni è necessario un account amministratore.

Configurazione

La schermata di configurazione visualizza le seguenti impostazioni dell'applicazione:

- **Library Prep Kits** (Kit di preparazione della libreria): visualizza il kit di preparazione della libreria predefinito per l'applicazione. Questa impostazione non può essere modificata.
- **Index Adapter Kits** (Kit adattatore indice): visualizza il kit adattatore indice predefinito per l'applicazione. Questa impostazione non può essere modificata.
- **Read lengths** (Lunghezze di lettura): le lunghezze di lettura sono impostate in modo predefinito su 151 per l'applicazione, ma possono essere modificate durante la creazione della corsa.
- **Manifest and Noise Files** (File manifest e rumore): carica e modifica le impostazioni per i file manifest e rumore.
  - Selezionare **Upload File** (Carica file) per caricare i file da utilizzare nell'analisi.
  - Selezionare il pulsante di opzione **Default** (Predefinito) per impostare il file come file manifest o rumore predefinito, selezionato durante la creazione della corsa quando l'applicazione è selezionata.
  - Selezionare la casella di spunta **Enabled** (Abilitata) per impostare la visualizzazione del file nel menu a discesa durante la creazione della corsa.

Permessi

Utilizzare le caselle di spunta della schermata Permissions (Autorizzazioni) per gestire l'accesso degli utenti all'applicazione.

File manifest

Quando si utilizza DRAGEN per IDPE Dx, è necessario un file manifest per i seguenti flussi di lavoro:

- Generazione del file FASTQ e VCF per il flusso di lavoro di una linea germinale
- Generazione del file FASTQ e VCF per il flusso di lavoro somatico

Il file manifest è un file delimitato da tabulazioni che utilizza il formato BED (*.bed), che specifica i nomi e le posizioni delle regioni di riferimento mirate. La sezione principale del file manifest è la sezione Regions (Regioni) e deve contenere le seguenti colonne di dati:

Colonna	Descrizione
Nome	Nome univoco specificato dall'utente per il target
Cromosoma	Posizione del cromosoma (ad es. chr10, chr5, ecc.)
Avvio	indice base 1 per la posizione iniziale del target
Arresto	indice base 1 per la posizione finale del target
Lunghezza sonda a monte	La lunghezza della sonda a monte. Per l'app DRAGEN per IDPE Dx, questo valore deve essere impostato su 0.
Lunghezza sonda a valle	La lunghezza della sonda a valle. Per l'app DRAGEN per IDPE Dx, questo valore deve essere impostato su 0.

NOTA Per l'analisi è necessario un formato di file manifest valido. DRAGEN interromperà l'analisi se il file manifest non è valido.

Filtro del rumore (facoltativo)

Il filtro del rumore sistematico è disponibile per l'identificazione di varianti somatiche e può essere utilizzato per ridurre le identificazioni false positive tenendo conto del rumore specifico del sito. Il file di rumore sistematico viene generato raccogliendo prima circa 50 campioni normali (preferibilmente specifici per il pannello, la preparazione della libreria e il sequenziatore); successivamente la somma delle frequenze degli alleli inferiori al 30% in ciascun sito con una copertura sufficiente è divisa per il numero totale di campioni (considerando le frequenze degli alleli superiori al 30% come varianti della linea germinale e non rumore). Una volta generati i valori di rumore, le varianti somatiche rilevate in quel sito verranno filtrate.

Il filtro può essere utilizzato in modalità tumore-normale, ma è particolarmente utile per le analisi solo tumorali, quando non è disponibile un normale abbinato. Il file di rumore sistematico deve utilizzare un file BED con estensione (`*.bed.gz`) e deve includere quattro colonne: livelli di rumore cromosomico, iniziale, finale e specifico del sito per ogni riga. Il filtraggio del rumore sistematico è facoltativo.

Output dell'analisi

Le corse in atto sono visualizzate nella scheda Active (Attive). Le corse completate vengono visualizzate nella scheda Completed (Completate). DRAGEN per IDPE Dx crea una cartella di analisi denominata in modo univoco per ogni analisi, separata dalla cartella contenente i dati di sequenziamento. La cartella di analisi include le seguenti informazioni:

- File manifest utilizzato
- Versione del software
- ID campione

- Letture totali allineate
- Percentuale di letture allineate per campione
- Numero di SNV identificate per campione
- Numero di indel identificati per campione
- Statistiche della copertura

File di output dell'analisi

La posizione della cartella di analisi è specificata dall'impostazione External Storage for Analysis Results (Memoria esterna per i risultati dell'analisi). Fare riferimento Analysis Guida al software Illumina Run Manager per NextSeq 550Dx (documento n. 200025239) per ulteriori informazioni sull'impostazione External Storage for Analysis Results (Memoria esterna per i risultati dell'analisi).

Nella schermata Run Details (Dettagli corsa), il campo External Location (Posizione esterna) fornisce il percorso per i dati di sequenza. Il nome univoco della cartella di analisi viene fornito nel campo Analysis Output Folder (Cartella di output dell'analisi) nella schermata Run Details (Dettagli corsa). I file esatti generati dipendono dal flusso di lavoro di analisi utilizzato. L'applicazione genera i seguenti file di output dell'analisi.

NOTA Se si verifica un errore di limitazione della lunghezza massima del percorso del file durante l'accesso ai file di output di analisi, occorre provare a spostare il file in una posizione con percorso più breve o utilizzare un metodo diverso per aprire il file.

File di output	Descrizione
Report di riepilogo varianti (`*.pdf`)	Contiene un riepilogo delle informazioni sui file, delle versioni del software, delle informazioni sui campioni, delle statistiche del livello delle letture e di SNV, inserimenti, delezioni e dei riepiloghi della copertura. Solo i flussi di lavoro della linea germinale e somatico producono un report delle varianti.
FASTQ (`*.fastq.gz` o `*.fastq.ora`)	File intermedi che contengono le identificazioni delle basi qualitativamente valutate. I file FASTQ rappresentano gli input principali per la fase di allineamento. Quando si seleziona la compressione ORA, viene utilizzata l'estensione del file `*.fastq.ora`.
File di allineamento BAM (`*.bam`)	Contiene le letture allineate per un determinato campione.
File del genoma VCF (`*.gvcf.gz`)	Contiene il genotipo per ogni posizione, sia che venga identificato come variante sia che venga identificato come riferimento.

File di output	Descrizione
File VCF (`*.vcf.gz`)	Contiene le varianti identificate in ogni posizione.
Relazione sulle metriche della corsa (`*.csv`)	Contiene le metriche qualitative della corsa, tra cui il rendimento totale e il punteggio Q30 senza indicizzazione.

File FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) è un formato file di testo che contiene le identificazioni delle basi e i valori qualitativi per ogni lettura. Ogni file contiene le informazioni seguenti:

- Identificatore del campione
- La sequenza
- I punteggi qualitativi su scala Phred in un formato codificato ASCII + 33

L'identificatore del campione è formattato nel seguente modo:

```
@Strumento:IDCorsa:IDCellaaflusso:Corsia:Tile:X:Y
NumLettura:IndicatoreFiltro:0:NumeroCampione
Esempio:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=
```

File BAM

Un file BAM (*.bam) è la versione binaria compressa di un file SAM (mappa di allineamento della sequenza) utilizzato per rappresentare sequenze allineate fino a 128 Mb. I file BAM utilizzano il formato di denominazione dei file `SampleName_S#.bam`. # è il numero del campione in base all'ordine in cui i campioni sono elencati per la corsa. In modalità multinodo, S# è impostato su S1, a prescindere dall'ordine del campione.

I file BAM contengono una sezione di intestazione e una sezione di allineamento:

- **Header** (Intestazione): contiene le informazioni sull'intero file, come il nome del campione, la lunghezza del campione e il metodo di allineamento. Gli allineamenti nella sezione allineamenti sono associati a informazioni specifiche nella sezione intestazione.
- **Alignments** (Allineamenti): contiene il nome della lettura, la sequenza della lettura, la qualità della lettura, le informazioni sull'allineamento e le tag personalizzate. Il nome della lettura include il cromosoma, la coordinata iniziale, la qualità dell'allineamento e la stringa del descrittore di corrispondenza.

La sezione degli allineamenti include le seguenti informazioni per ogni lettura o accoppiamento di letture:

- AS: Qualità dell'allineamento a estremità accoppiate.
- RG: Gruppo di lettura, che indica il numero di letture per un campione specifico.
- BC: Etichetta Barcode, che indica l'ID del campione demultiplexato associato alla lettura.
- SM: Qualità dell'allineamento a estremità singola.
- XC: Stringa del descrittore dell'accoppiamento.
- XN: Tag del nome dell'amplicone, che registra l'ID dell'amplicone associato con la lettura.

File indici BAM (*.bam.bai) forniscono un indice del corrispondente file BAM.

File VCF

I file Variant call format (*.vcf) contengono informazioni sulle varianti trovate in posizioni specifiche in un genoma di riferimento.

L'intestazione del file VCF include la versione del formato del file VCF e la versione dell'identificatore di varianti, ed elenca le annotazioni utilizzate nel resto del file. L'intestazione del file VCF include anche il file del genoma di riferimento e il file BAM. L'ultima riga dell'intestazione contiene le intestazioni delle colonne per le righe dei dati. Ogni riga di dati del file VCF contiene informazioni su una singola variante.

Tabella 1 Intestazioni dei file VCF

Intestazione	Descrizione
CHROM	Il cromosoma del genoma di riferimento. I cromosomi appaiono nello stesso ordine del file FASTA di riferimento.
POS	La posizione a singola base della variante nel cromosoma di riferimento. Per le varianti a singolo nucleotide (SNV), questa posizione è la base di riferimento con la variante. Per le indel, questa posizione è la base di riferimento immediatamente precedente la variante.
ID (Identificazione)	Il numero rs (riferimento SNP) per l'SNP ottenuto da `dbSNP.txt`, se pertinente. Se vi sono numeri rs multipli in questa posizione, l'elenco è delimitato da punti e virgole. Se non vi è una voce dbSNP in questa posizione, viene utilizzato un marcatore di valore mancante ('.').
REF	Il genotipo di riferimento. Ad esempio, una delezione di una singola T è rappresentata come TT di riferimento e T alternativa. Una variante da A a T a singolo nucleotide è rappresentata come A di riferimento e T alternativa.
ALT	Gli alleli che differiscono dalla lettura di riferimento. Ad esempio, l'inserimento di una singola T viene rappresentato come A di riferimento e AT alternativo. Una variante da A a T a singolo nucleotide è rappresentata come A di riferimento e T alternativa.

Intestazione	Descrizione
QUAL	Un punteggio di qualità con scala di Phred assegnato dall'identificatore della variante. Punteggi più alti indicano una maggiore affidabilità nella variante e una minore probabilità di errori. Per un punteggio qualitativo di Q, la probabilità di errore stimata è 10-(Q/10). Ad esempio, l'insieme delle identificazioni Q30 ha un tasso di errore dello 0,1%. Molti identificatori di varianti assegnano punteggi di qualità basati sui loro modelli statistici, che sono elevati in relazione al tasso di errore osservato.

Tabella 2 Annotazioni del file VCF del flusso di lavoro della linea digerente

Intestazione	Descrizione
FILTRO	Se tutti i filtri sono stati superati, nella colonna dei filtri viene scritto PASS (SUPERATO). Le possibili voci FILTER (FILTRO) includono: • **DRAGENSnpHardQUAL**: applicato se il punteggio QUAL della variante SNP non soddisfa la soglia • **DRAGENIndelHardQUAL**: applicato se il punteggio QUAL della variante indel non soddisfa la soglia • **LowDepth**: sito filtrato perché la profondità di copertura non soddisfa la soglia • **LowGQ**: sito filtrato perché la qualità del genotipo non soddisfa la soglia • **PloidyConflict**: identificazione di genotipo dall'identificatore di varianti non coerente con la ploidia cromosomica • **base_quality**: sito filtrato perché la qualità mediana delle basi delle letture alt in questo locus non soddisfa la soglia • **filtered_reads**: sito filtrato perché è stata filtrata una frazione troppo grande di letture • **fragment_length**: sito filtrato perché la differenza assoluta tra la lunghezza mediana dei frammenti delle letture alt e la lunghezza mediana dei frammenti delle letture ref a questo locus supera la soglia • **low_depth**: sito filtrato perché la profondità di lettura è troppo bassa • **low_frac_info_reads**: sito filtrato perché la frazione di letture informative è inferiore alla soglia • **low_normal_depth**: sito filtrato perché la profondità di lettura normale del campione è troppo bassa • **long_indel**: sito filtrato perché la lunghezza dell'indel è troppo lunga • **mapping_quality**: sito filtrato perché la qualità di mappatura mediana delle letture alt a questo locus non soddisfa la soglia • **multiallelic**: sito filtrato perché più di due alleli alt superano il LOD del tumore • **non_homref_normal**: sito filtrato perché il genotipo del campione normale non è omozigote di riferimento • **no_reliable_supporting_read**: sito filtrato perché non esiste una lettura somatica di supporto affidabile • **panel_of_normals**: osservato in almeno un campione del pannello delle normalità vcf • **read_position**: sito filtrato perché la mediana delle distanze tra l'inizio e la fine della lettura e questo locus è inferiore alla soglia • **RMxNRepeatRegion**: sito filtrato perché tutto o parte dell'allele della variante è una ripetizione del riferimento • **strand_artifact**: sito filtrato a causa di un grave bias del filamento • **str_contraction**: sito filtrato a causa di un sospetto errore di PCR in cui l'allele alt è inferiore di un'unità di ripetizione rispetto al riferimento • **too_few_supporting_reads**: sito filtrato perché ci sono troppo poche letture di supporto nel campione di tumore • **weak_evidence**: il punteggio della variante somatica non soddisfa la soglia

Intestazione	Descrizione
INFO	Le possibili voci INFO (INFORMAZIONI) includono: • **AC**: il conteggio degli alleli nei genotipi per ciascun allele ALT (Alternato), nello stesso ordine in cui sono elencati. • **AF**: la frequenza allelica per ciascun allele ALT (Alternato), nello stesso ordine in cui sono elencati. • **AN**: il numero totale di alleli nei genotipi identificati. • **DB**: appartenenza a dbSNP. • **FS**: valore p scalato con Phred mediante il test esatto di Fisher per rilevare il bias del filamento. • **QD**: l'affidabilità/qualità della variante per la profondità. • **R2_5P_bias**: punteggio basato sul bias di accoppiamento e sulla distanza dall'estremità principale 5. • **SOR**: rapporto di probabilità simmetrico della tabella di contingenza 2x2 per rilevare il bias del filamento. • **DP**: profondità di lettura approssimativa (informativa e non informativa); alcune letture possono essere state filtrate in base a mapq ecc. • **END**: posizione di arresto dell'intervallo. • **FractionInformativeReads**: la frazione di letture informative sul totale delle letture. • **MQ**: qualità della mappatura RMS. • **MQRankSum**: punteggio Z dal test della somma dei ranghi di Wilcoxon delle qualità di mappatura delle letture Alt rispetto a Ref. • **ReadPosRankSum**: punteggio Z del test di somma dei ranghi di Wilcoxon sulla polarizzazione delle posizioni di lettura Alt rispetto a Ref. • **SOMATIC**: almeno una variante in questa posizione è somatica.

Intestazione	Descrizione
FORMATO	La colonna Formato elenca i campi separati da due punti. Ad esempio, GT:GQ. I campi disponibili includono: • **AD**: profondità alleliche (contando solo le letture informative sul totale delle letture) per gli alleli ref e alt nell'ordine elencato. • **AF**: frazioni alleliche per gli alleli alt nell'ordine elencato. • **DP**: la profondità approssimativa della lettura (letture con MQ=255 o con accoppiamenti non corretti sono filtrate). • **F1R2**: conteggio delle letture in orientamento delle coppie F1R2 che supportano ciascun allele. • **F2R1**: conteggio delle letture in orientamento delle coppie F2R1 che supportano ciascun allele. • **GT**—Genotipo. 0 corrisponde alla base di riferimento, 1 corrisponde alla prima voce nella colonna ALT (Alternato), e così via. La barra in avanti (/) indica che non sono disponibili informazioni sul phasing. • **MB**: statistiche delle componenti per campione per rilevare i bias di accoppiamento. • **PS**: informazioni sull'ID fisico di phasing, dove ogni ID univoco all'interno di un dato campione (ma non tra i campioni) collega i record all'interno di un gruppo di phasing. • **SB**: statistiche dei componenti per campione che comprendono il test esatto di Fisher per rilevare il bias del filamento. • **SQ**: qualità di Somatic.
CAMPIONE	La colonna del campione fornisce i valori specificati nella colonna FORMAT (FORMATO).

Tabella 3 Annotazioni del file VCF del flusso di lavoro somatico

Intestazione	Descrizione
FILTRO	Se tutti i filtri sono stati superati, nella colonna dei filtri viene scritto PASS (SUPERATO). Le possibili voci FILTER (FILTRO) includono: • **base_quality**: sito filtrato perché la qualità mediana delle basi delle letture alt in questo locus non soddisfa la soglia • **filtered_reads**: sito filtrato perché è stata filtrata una frazione troppo grande di letture • **fragment_length**: sito filtrato perché la differenza assoluta tra la lunghezza mediana dei frammenti delle letture alt e la lunghezza mediana dei frammenti delle letture ref a questo locus supera la soglia • **low_depth**: sito filtrato perché la profondità di lettura è troppo bassa • **low_frac_info_reads**: sito filtrato perché la frazione di letture informative è inferiore alla soglia • **low_normal_depth**: sito filtrato perché la profondità di lettura normale del campione è troppo bassa • **long_indel**: sito filtrato perché la lunghezza dell'indel è troppo lunga • **mapping_quality**: sito filtrato perché la qualità di mappatura mediana delle letture alt a questo locus non soddisfa la soglia • **multiallelic**: sito filtrato perché più di due alleli alt superano il LOD del tumore • **non_homref_normal**: sito filtrato perché il genotipo del campione normale non è omozigote di riferimento • **no_reliable_supporting_read**: sito filtrato perché non esiste una lettura somatica di supporto affidabile • **panel_of_normals**: osservato in almeno un campione del pannello delle normalità vcf • **read_position**: sito filtrato perché la mediana delle distanze tra l'inizio e la fine della lettura e questo locus è inferiore alla soglia • **RMxNRepeatRegion**: sito filtrato perché tutto o parte dell'allele della variante è una ripetizione del riferimento • **strand_artifact**: sito filtrato a causa di un grave bias del filamento • **str_contraction**: sito filtrato a causa di un sospetto errore di PCR in cui l'allele alt è inferiore di un'unità di ripetizione rispetto al riferimento • **too_few_supporting_reads**: sito filtrato perché ci sono troppo poche letture di supporto nel campione di tumore • **weak_evidence**: il punteggio della variante somatica non soddisfa la soglia • **systematic_noise**: sito filtrato in base all'evidenza di rumore sistematico nei normali

Intestazione	Descrizione
INFO	Le possibili voci INFO (INFORMAZIONI) includono: • **DP**: profondità di lettura approssimativa (informativa e non informativa); alcune letture possono essere state filtrate in base a mapq ecc. • **END**: posizione di arresto dell'intervallo. • **FractionInformativeReads**: la frazione di letture informative sul totale delle letture. • **MQ**: qualità della mappatura RMS. • **MQRankSum**: punteggio Z dal test della somma dei ranghi di Wilcoxon delle qualità di mappatura delle letture Alt rispetto a Ref. • **ReadPosRankSum**: punteggio Z del test di somma dei ranghi di Wilcoxon sulla polarizzazione delle posizioni di lettura Alt rispetto a Ref. • **AQ**: punteggio del rumore sistematico. • **hotspot**: sito somatico noto, utilizzato per aumentare l'affidabilità nell'identificazione. • **SOMATIC**: almeno una variante in questa posizione è somatica.

Intestazione	Descrizione
FORMATO	La colonna Formato elenca i campi separati da due punti. Ad esempio, GT:GQ. I campi disponibili includono: • **AD**: profondità alleliche (contando solo le letture informative sul totale delle letture) per gli alleli ref e alt nell'ordine elencato. • **AF**: frazioni alleliche per gli alleli alt nell'ordine elencato. • **DP**: la profondità approssimativa della lettura (letture con MQ=255 o con accoppiamenti non corretti sono filtrate). • **F1R2**: conteggio delle letture in orientamento delle coppie F1R2 che supportano ciascun allele. • **F2R1**: conteggio delle letture in orientamento delle coppie F2R1 che supportano ciascun allele. • **GP**: probabilità posteriori scalate con Phred per i genotipi, come definito nella specifica del VCF. • **GQ**: qualità del genotipo. • **GT**—Genotipo. 0 corrisponde alla base di riferimento, 1 corrisponde alla prima voce nella colonna ALT (Alternato), e così via. La barra in avanti (/) indica che non sono disponibili informazioni sul phasing. • **MB**: statistiche delle componenti per campione per rilevare i bias di accoppiamento. • **PL**: probabilità normalizzate e scalate con Phred per i genotipi, come definito nelle specifiche del VCF. • **PRI**: probabilità antecedenti scalate con Phred per i genotipi. • **PS**: informazioni sull'ID fisico di phasing, dove ogni ID univoco all'interno di un dato campione (ma non tra i campioni) collega i record all'interno di un gruppo di phasing. • **SB**: statistiche dei componenti per campione che comprendono il test esatto di Fisher per rilevare il bias del filamento. • **SQ**: qualità di Somatic.
CAMPIONE	La colonna del campione fornisce i valori specificati nella colonna FORMAT (FORMATO).

File VCF del genoma

I file VCF del genoma (*.gvcf.gz) seguono una serie di convenzioni per rappresentare tutti i siti all'interno del genoma in un formato ragionevolmente compatto. I file gVCF includono tutti i siti all'interno della regione di interesse in un unico file per ciascun campione. Il file gVCF mostra le identificazioni non rilevate nelle posizioni che non passano tutti i filtri. Un tag genotipo (GT) tag di ./. indica un'identificazione non rilevata.

Requeue Analysis (Rimetti in coda l'analisi)

Un'analisi può essere rimessa in coda se l'analisi è stata arrestata, se l'analisi non è riuscita o se si desidera analizzare nuovamente una corsa con impostazioni diverse. Per rimettere in coda l'analisi, seguire la procedura riportata qui sotto:

1. Dalla schermata Run (Corsa), selezionare la scheda Completed (Completate), quindi selezionare il nome della corsa da rianalizzare.
   Se l'analisi era già stata rimessa in coda precedentemente, selezionare il nome della corsa principale.
2. Dalla schermata Run Details (Dettagli corsa), dopo le Sequencing Information (Informazioni sul sequenziamento), selezionare **Requeue Analysis** (Rimetti in coda l'analisi).
3. Selezionare un'opzione:
   - Rimettere in coda l'analisi senza modifiche
   - Modificare le impostazioni della corsa e rimettere in coda l'analisi
   - Rimettere in coda l'analisi con un'applicazione diversa
4. Verificare che la posizione in cui si trovano attualmente i dati di sequenziamento sia fornita nel campo **Percorso del file dei dati di sequenziamento**.

NOTA Il percorso verso i dati di sequenziamento deve corrispondere al percorso specificato in External Storage for Analysis Results (Memoria esterna per i risultati dell'analisi). Fare riferimento a Guida al software Illumina Run Manager per NextSeq 550Dx (documento n. 200025239) per informazioni sulla modifica del percorso di archiviazione esterno.

5. Immettere un motivo per la ripetizione dell'analisi.
6. Selezionare **Requeue Analysis** (Rimetti in coda l'analisi).
7. Modificare le modifiche desiderate a Run Settings (Impostazioni corsa), Sample Data (Dati campione) e Analysis Settings (Impostazioni di analisi).
8. Selezionare **Save** (Salva). L'analisi inizia utilizzando gli attuali parametri dell'analisi.

Assistenza tecnica

Per ricevere assistenza tecnica, contattare l'Assistenza tecnica Illumina.

Sito web: www.illumina.com
E-mail: techsupport@illumina.com

Schede dei dati di sicurezza (SDS): sono disponibili sul sito web Illumina all'indirizzo support.illumina.com/sds.html.

Documentazione sul prodotto: disponibile per il download all'indirizzo support.illumina.com.

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (fuori dal Nord America)
techsupport@illumina.com
www.illumina.com

IVD

EC REP

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

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

PER USO DIAGNOSTICO IN VITRO.

illumina®