illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx

Documentazione del prodotto per NovaSeq 6000Dx

DI PROPRIETÀ DI ILLUMINA
Documento # 200014776 v02
Settembre 2022
PER USO DIAGNOSTICO IN VITRO

Cronologia revisioni

Documento	Data	Descrizione della modifica
200014776 v02	Settembre 2022	Corretto il formato del file manifest da testo (*.txt) a BED (*.bed) nelle istruzioni per la creazione della corsa. Corretti i file VCF di consenso in file VCF nella sezione di output dell'analisi.
200014776 v01	Agosto 2022	Aggiunte: Sezione Impostazioni. Sezione Filtrazione sistematica del rumore. Aggiornate le istruzioni per la creazione della corsa per includere maggiori dettagli. Correzione di errori di battitura e grammaticali. Specificato che le istruzioni sono destinate all'applicazione quando viene utilizzata con lo strumento NovaSeq 6000Dx. Informazioni aggiornate sul contenuto del file di output VCF.
200014776 v00	Marzo 2022	Versione iniziale.

Sommario

Descrizione generale

L'applicazione DRAGEN™ per Illumina® DNA Prep with Enrichment Dx esegue il demultiplex, la generazione di FASTQ, la mappatura delle letture, l'allineamento a un genoma di riferimento e l'identificazione di varianti, a seconda del flusso di lavoro di analisi selezionato.

Metodi di analisi

DRAGEN for Illumina DNA Prep with Enrichment Dx esegue il demultiplex, la generazione di file FASTQ, la mappatura delle letture e l'allineamento a un genoma di riferimento, 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

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. 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

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.

I valori specifici scelti per i punteggi in questo algoritmo indicano come bilanciare, per un allineamento con molteplici interpretazioni possibili, la possibilità di una indel rispetto a uno o più SNP, o la preferenza per un allineamento senza clipping. I valori di punteggio predefiniti di DRAGEN sono ragionevoli per allineare letture di lunghezza moderata a un intero genoma umano di riferimento per applicazioni di identificazione di varianti. Qualsiasi insieme di parametri del punteggio di Smith-Waterman rappresenta un modello impreciso di mutazione genomica e di errori di sequenziamento. Valori del punteggio di allineamento regolati in modo diverso possono essere più appropriati per alcune applicazioni.

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 SNP e indel attraverso una combinazione di rilevamento in colonna e assemblaggio locale *de novo* degli aplotipi.

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 intorno alle 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 (Hidden Markov Model, 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.

Nella modalità gVCF, utilizzata per l'identificazione di varianti scalabili su più campioni, l'identificatore di piccole varianti per Germline DRAGEN può essere utilizzato per ogni campione per generare un file intermedio di identificazione di varianti genomiche (genomic Variant Call File, gVCF). Il gVCF può quindi essere utilizzato per un'efficiente genotipizzazione congiunta di più campioni, che consente una rapida elaborazione incrementale dei campioni e la scalabilità a coorti di grandi dimensioni.

Dal momento che l'identificatore di piccole varianti per Germline DRAGEN dispone di algoritmi che lo rendono in grado di distinguere efficacemente gli errori correlati dalle varianti vere, le regole di filtraggio sono molto semplici.

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.

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 alle 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 TLOD, l'identificatore di piccole varianti per Somatic DRAGEN esegue innanzitutto una scansione per posizione di riferimento per ogni evento somatico candidato e per l'evento di riferimento sulla regione attiva. La probabilità condizionale P(r|e) di osservare ogni lettura sovrapposta è stimata come la massima 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.

Creazione di una corsa

Utilizzare i passaggi seguenti per impostare una corsa in Illumina Run Manager su NovaSeq 6000Dx o utilizzando un browser su un computer in rete. I dati del campione possono essere inseriti manualmente o importando un foglio di campioni.

Impostazioni dell'applicazione e della corsa

1. Dalla schermata Runs (Corse), selezionare **Create Run** (Crea corsa).
2. Selezionare l'app DRAGEN for Illumina DNA Prep with Enrichment Dx, quindi selezionare **Next** (Avanti).
3. Sulla schermata Run Settings (Impostazioni corsa), immettere un nome per la corsa. Il nome della corsa identifica la corsa dal sequenziamento per tutta l'analisi.
4. **[Facoltativo]** Immettere una descrizione per identificare la corsa.
5. Assicurarsi che il kit di preparazione della libreria selezionato sia un kit di preparazione della libreria DNA Prep with Enrichment Dx di Illumina.
6. Selezionare il kit di adattatori per indici desiderato.
7. Immettere la lunghezza della lettura.
   Read 1 (Lettura 1) e Read 2 (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.
8. **[Facoltativo]** Immettere l'ID di una provetta della libreria.
9. Selezionare **Next** (Avanti).

Dati del campione

Utilizzare la tabella della schermata Sample Data (Dati del campione) per immettere manualmente le informazioni sul campione. In alternativa, selezionare **Import Samples** (Importa campioni) per caricare le informazioni sul campione. Per informazioni sull'importazione di informazioni sul campione, consultare *Importare campioni* alla pagina 5 (Importa campioni).

Immissione manuale dei campioni

1. Immettere un ID campione univoco nel campo Sample ID (ID campione).
2. Usare **Plate - Well Position** (Posizione piastra - pozzetto) per selezionare la posizione del pozzetto. 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.
5. Selezionare **Next** (Avanti).

Importare campioni

Un modello (*.csv) è disponibile per il download nella schermata Sample Data (Dati campione) quando si pianifica una corsa in Illumina Run Manager utilizzando un browser su un computer in rete.

1. Selezionare **Download Template** (Scarica modello) per scaricare un file CSV vuoto.
2. Dal file CSV, immettere le informazioni del campione e salvare il file.
   Il file CVS del foglio di esempio include le seguenti colonne di dati: Sample ID (ID campione), Plate - Well Position (Posizione piastra - pozzetto), **Optional** Library Name (Nome libreria facoltativo).
3. Selezionare **Import Samples** (Importa campioni) e aprire il percorso del file CSV.
4. Selezionare **Next** (Avanti).

Impostazioni di analisi

1. Selezionare il flusso di lavoro di analisi desiderato:
   - Generazione di file FASTQ
   - Generazione del file FASTQ e VCF per Germline per il flusso di lavoro di una linea germinale
   - Generazione del file FASTQ e VCF per Somatic per il flusso di lavoro somatico
2. **[Facoltativo]** Se lo si desidera, selezionare la casella di spunta **Generate ORA compressed FASTQs** (Generare file FASTQ con compressione ORA) per abilitare la compressione ORA del file FASTQ.
3. **[Flussi di lavoro per la generazione di VCF]** Usare il menu a discesa **Manifest File Selection** (Selezione file manifest) per selezionare un file manifest.
   Un file manifest è un input necessario per DRAGEN for Illumina DNA Prep with Enrichment Dx. Manifest è un file BED (*.bed) delimitato da tabulazioni che specifica i nomi e le posizioni delle regioni di riferimento mirate.
4. **[Flusso di lavoro per la generazione di file FASTQ e VCF per Somatic]** Utilizzare il menu a discesa **Noise File Selection** (Selezione file di rumore) per selezionare un file di rumore.
   È possibile specificare un file BED con un livello di rumore specifico per il sito per filtrare il rumore sistematico. Per maggiori informazioni, consultare *Filtrazione del rumore* alla pagina 6.
5. Selezionare **Next** (Avanti).

Corsa Revisione

1. Nella schermata Review (Revisione), rivedere le informazioni inserite nelle schermate 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

Selezionare l'applicazione nella schermata Applications (Applicazioni) per visualizzare le impostazioni correnti e modificarle.

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.

Filtrazione del rumore

La filtrazione sistematica del rumore è disponibile quando si utilizza il flusso di lavoro somatico. 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 rumore sistematico BED deve essere generato da campioni normali. Si raccomanda di creare file di rumore sistematico che siano specifici per la preparazione delle librerie, il sistema di sequenziamento e il pannello. Per la generazione di file di rumore, si consiglia di utilizzare circa 50 campioni normali.

Output dell'analisi

DRAGEN for Illumina DNA Prep with Enrichment Dx salva le seguenti informazioni nella cartella di analisi. Solo i flussi di lavoro della linea germinale e somatico producono un PDF.

- 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

L'applicazione genera i seguenti file di output. I file esatti generati dipendono dal flusso di lavoro di analisi utilizzato. I file di output si trovano nella cartella di analisi.

File di output	Descrizione
FASTQ (*.fastq.gz oo *.fastq.ora)	File intermedi che contengono le identificazioni delle basi qualitativamente valutate. I file FASTQ rappresentano gli input principali per la fase di allineamento. Se è stata selezionata la compressione ORA, il nome del file riflette questa scelta.
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 VCF (*.vcf.gz)	Contiene le varianti identificate in ogni posizione.

File di output	Descrizione
Relazione sulle metriche della corsa (*.csv)	Contiene le metriche qualitative della corsa, tra cui il rendimento totale e il punteggio Q30.

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
- Un segno più (+)
- 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=AAAAAAAAAA9#:<#<;<<<????#=
```

File BAM

Un file BAM (*.bam) è la versione binaria compressa di un file SAM (Sequence Alignment Map [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 ai file BAM indice (*.bam.bai) letti fornisce un indice del file BAM corrispondente.

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 (Single Nucleotide Variants, 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 dei file VCF

Intestazione	Descrizione
FILTRO	Se tutti i filtri sono stati superati, nella colonna dei filtri viene scritto PASS. Le possibili voci FILTRO del flusso di lavoro per Germline 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

Intestazione	Descrizione
FILTRO (segue)	• **weak_evidence**: il punteggio della variante somatica non soddisfa la soglia Le possibili voci FILTRO del flusso di lavoro per Somatic 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 voci INFO possibili del flusso di lavoro per Germline 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. Le possibili voci INFO del flusso di lavoro per Somatic 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 per il flusso di lavoro per Germline 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. I campi disponibili per il flusso di lavoro per Somatic 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.

Intestazione	Descrizione
FORMATO (segue)	• **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.

Visualizzazione dei risultati dell'analisi

Le corse in atto sono visualizzate nella scheda Active (Attive). Le corse completate sono visualizzate nella scheda Completed (Completate). Consultare Documentazione del prodotto NovaSeq 6000Dx (documento n. 200010105) per ulteriori informazioni sulla visualizzazione dei risultati.

Assistenza Tecnica

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

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

Numeri di telefono dell'Assistenza tecnica Illumina

Area geografica	Gratuito	Internazionale
Australia	+61 1800 775 688	
Austria	+43 800 006249	+43 1 9286540
Belgio	+32 800 77 160	+32 3 400 29 73
Canada	+1 800 809 4566	
Cina		+86 400 066 5835
Danimarca	+45 80 82 01 83	+45 89 87 11 56
Finlandia	+358 800 918 363	+358 9 7479 0110
Francia	+33 8 05 10 21 93	+33 1 70 77 04 46
Germania	+49 800 101 4940	+49 89 3803 5677
Hong Kong, Cina	+852 800 960 230	
India	+91 8006500375	
Indonesia		0078036510048
Irlanda	+353 1800 936608	+353 1 695 0506
Italia	+39 800 985513	+39 236003759
Giappone	+81 0800 111 5011	
Malesia	+60 1800 80 6789	
Paesi Bassi	+31 800 022 2493	+31 20 713 2960
Nuova Zelanda	+64 800 451 650	
Norvegia	+47 800 16 836	+47 21 93 96 93
Filippine	+63 180016510798	
Singapore	1 800 5792 745	
Corea del Sud	+82 80 234 5300	
Spagna	+34 800 300 143	+34 911 899 417

Area geografica	Gratuito	Internazionale
Svezia	+46 2 00883979	+46 8 50619671
Svizzera	+41 800 200 442	+41 56 580 00 00
Taiwan, Cina	+886 8 06651752	
Thailandia	+66 1800 011 304	
Regno Unito	+44 800 012 6019	+44 20 7305 7197
Stati Uniti	+1 800 809 4566	+1 858 202 4566
Vietnam	+84 1206 5263	

Schede dei dati di sicurezza (Safety Data Sheet, 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
Paesi Bassi

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

PER USO DIAGNOSTICO IN VITRO

illumina®