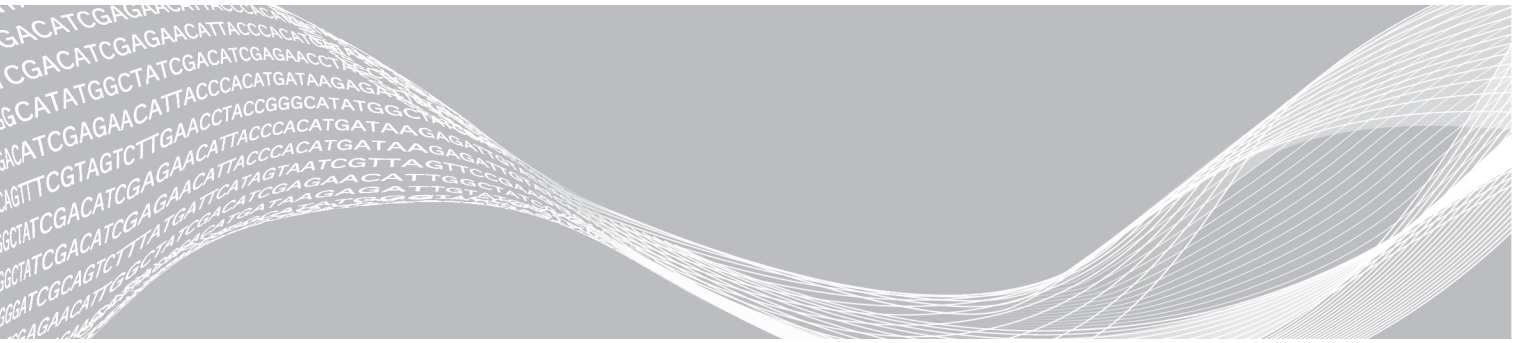
illumina®

Modulo di analisi Local Run Manager Somatic Variant

Guida al flusso di lavoro per MiSeqDx

PER USO DIAGNOSTICO IN VITRO



Documento n. 1000000012183 v01 ITA
Agosto 2021

PER USO DIAGNOSTICO IN VITRO

Descrizione generale

Il modulo di analisi Local Run Manager Somatic Variant è previsto per l'uso con il saggio TruSeq Custom Amplicon Kit Dx Illumina. Quando utilizzato con il modulo Somatic Variant, il saggio è destinato alla preparazione delle librerie utilizzate per il sequenziamento del DNA ottenuto da tessuto fissato in formalina e incluso in paraffina (Formalin-Fixed, Paraffin-Embedded, FFPE). Il saggio rileva le mutazioni somatiche di varianti a basse frequenze.

Il modulo di analisi valuta le regioni brevi di DNA amplificato, o ampliconi, rilevandone le varianti. Il sequenziamento mirato degli ampliconi consente l'elevata copertura di determinate regioni su un ampio numero di campioni. Il modulo di analisi esegue l'analisi secondaria e la generazione di report dai dati ottenuti dalle corse di sequenziamento utilizzando un approccio a doppio filamento che coinvolge i raggruppamenti in pool di oligonucleotidi forward e reverse. Vedere l'inserto della confezione di *TruSeq Custom Amplicon Kit Dx (documento n. 1000000029772)*.

Il modulo di analisi Somatic Variant richiede l'uso dei materiali di consumo per il sequenziamento di MiSeqDx Reagent Kit v3. Vedere l'inserto della confezione di *MiSeqDx Reagent Kit v3 (documento n. 1000000030849).*

Informazioni sulla guida

La presente guida fornisce istruzioni per l'impostazione dei parametri di una corsa per il sequenziamento e l'analisi sul modulo di analisi Somatic Variant. Per informazioni sul pannello di controllo e sulle impostazioni di sistema di Local Run Manager, vedere la *Guida di consultazione del software Local Run Manager (documento n. 1000000011880)*.

Immissione delle informazioni per la corsa

Impostazione dei parametri

1 Accedere a Local Run Manager.

2 Fare clic su **Create Run** (Crea corsa) e selezionare **Somatic Variant**.

3 Immettere un nome che identifichi la corsa dal sequenziamento fino all'analisi.
Utilizzare caratteri alfanumerici, spazi, trattini bassi o trattini.

4 [Facoltativo] Immettere una descrizione per identificare la corsa.
Utilizzare caratteri alfanumerici, spazi, trattini bassi o trattini.

5 Selezionare il numero di campioni e set di indici dall'elenco a discesa.

Importazione dei file manifest per la corsa

1 Assicurarsi che i file manifest da importare siano disponibili in una posizione di rete accessibile o su un dispositivo USB.

2 Fare clic su **Import Manifests** (Importa i file manifest).

3 Individuare il file manifest e selezionare i file manifest da aggiungere.

NOTA

Per far sì che i file manifest siano disponibili per tutte le corse utilizzando il modulo di analisi Somatic Variant, aggiungere i file manifest utilizzando la funzione Module Settings (Impostazioni modulo). Questa funzione richiede i permessi a livello di amministratore. Per maggiori informazioni, vedere la *Guida di consultazione del software Local Run Manager (documento n. 1000000011880)*.

Impostazione dei campioni per la corsa

Specificare i campioni per la corsa mediante una delle seguenti opzioni e procedure.

- **Immissione manuale dei campioni**: utilizzare la tabella vuota che si trova nella schermata Create Run (Crea corsa).
- **Importazione dei campioni**: individuare il file esterno nel formato con valori separati da virgola (*.csv). Dalla schermata Create Run (Crea corsa) è possibile scaricare un modello.

Dopo aver popolato la tabella dei campioni, è possibile esportare le informazioni dei campioni in un file esterno e utilizzare il file come riferimento quando si preparano le librerie o si importa il file per un'altra corsa.

Immissione manuale dei campioni

1. Immettere un ID campione univoco nel campo Sample Name (Nome campione).
   Utilizzare caratteri alfanumerici, trattini o trattini bassi.
   Il nome del campione popola automaticamente il pozzetto corrispondente nell'altro raggruppamento in pool.
2. [Facoltativo] Per i campioni di controllo positivi o negativi, fare clic con il pulsante destro del mouse e selezionare il tipo di controllo.
   Il campione di controllo di un campione popola automaticamente il pozzetto corrispondente nell'altro raggruppamento in pool assegnando lo stesso campione di controllo.
3. [Facoltativo] Immettere una descrizione del campione nel campo Sample Description (Descrizione del campione).
   Utilizzare caratteri alfanumerici, trattini o trattini bassi.
   La descrizione del campione popola automaticamente il pozzetto corrispondente nell'altro raggruppamento in pool.
   Le descrizioni del campione sono associate all'ID campione. Le descrizioni del campione sono sovrascritte se lo stesso ID campione viene utilizzato di nuovo in una corsa successiva.
4. Selezionare un adattatore indice 1 dall'elenco a discesa Index 1 (i7) (Indice 1 - i7).
5. Selezionare un adattatore indice 2 dall'elenco a discesa Index 2 (i5) (Indice 2 - i5).
6. Selezionare un file manifest dall'elenco a discesa Manifest (File manifest).
   I campioni in Pool A (Raggruppamento A) richiedono un file manifest diverso rispetto ai campioni in Pool B (Raggruppamento B).
7. Scegliere un'opzione per visualizzare, stampare o salvare il layout della piastra da utilizzare come riferimento al momento della preparazione delle librerie:
   - Fare clic sull'icona **Print** (Stampa) per visualizzare il layout della piastra. Selezionare **Print** (Stampa) per stampare il layout della piastra.
   - Fare clic su **Export** (Esporta) per esportare le informazioni sui campioni su un file esterno.

   Assicurarsi che il file manifest e le informazioni sul campione siano corretti. La presenza di informazioni errate possono influire sui risultati.

8 Fare clic su **Save Run** (Salva corsa).

Importazione dei campioni

1 Fare clic su **Import Samples** (Importa campioni) e andare alla posizione in cui si trova il file contenente le informazioni sui campioni. Possono essere importati due tipi di file.
   - Fare clic su **Template** (Modello) sulla schermata Create Run (Crea corsa) per creare un nuovo layout della piastra. Il file modello contiene le intestazioni di colonna corrette per eseguire l'importazione. In ciascuna colonna, immettere le informazioni sui campioni da analizzare nella corsa. Eliminare le informazioni di esempio nelle caselle non utilizzate, quindi salvare il file.
   - Utilizzare un file, contenente le informazioni sui campioni, che era stato esportato dal modulo Somatic Variant mediante la funzione Export (Esporta).
2 Fare clic sull'icona **Print** (Stampa) per visualizzare il layout della piastra.
3 Selezionare **Print** (Stampa) per stampare il layout della piastra da utilizzare come riferimento per la preparazione delle librerie.
4 [Facoltativo] Fare clic su **Export** (Esporta) per esportare le informazioni sui campioni in un file esterno. Assicurarsi che il file manifest e le informazioni sul campione siano corretti. La presenza di informazioni errate possono influire sui risultati.
5 Fare clic su **Save Run** (Salva corsa).

Modifica di una corsa

Per istruzioni su come modificare le informazioni della corsa prima del sequenziamento, vedere la *Guida di consultazione del software Local Run Manager (documento n. 1000000011880)*.

Metodi di analisi

Il modulo di analisi Somatic Variant esegue la seguente procedura di analisi, quindi scrive i file di output dell'analisi nella cartella Alignment (Allineamento).

- Esegue il demultiplex delle letture indici
- Genera i file FASTQ
- Esegue l'allineamento su un riferimento
- Identifica le varianti

Demultiplex

Il demultiplex confronta ogni sequenza Index Read (Lettura indici) sulle sequenze indici specificate per la corsa. In questa fase non vengono considerati i valori qualitativi.

Le letture indici sono identificate mediante le fasi successive:

- I campioni sono numerati a partire da 1 in base all'ordine in cui sono stati elencati per la corsa.
- Il numero campione 0 è riservato per i cluster che non sono stati assegnati per 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 Index Read (Lettura indici).

Generazione di file FASTQ

Al termine del demultiplex, il software genera i file dell'analisi intermedia in formato FASTQ, un formato di testo utilizzato per rappresentare le sequenze. I file FASTQ contengono le letture per ogni campione e i punteggi qualitativi associati. I cluster che non hanno superato il filtro sono esclusi.

Ogni file FASTQ contiene le letture per un solo campione e il nome di quel campione è incluso nel nome del file FASTQ. I file FASTQ rappresentano gli input principali per l'allineamento. Vengono generati due file FASTQ per campione, uno dal raggruppamento in pool A e uno dal raggruppamento in pool B.

Allineamento

Durante la fase di allineamento, l'algoritmo con matrice a banda di Smith-Waterman allinea i cluster ottenuti da ciascun campione sulle sequenze degli ampliconi specificati nel file manifest.

L'algoritmo Smith-Waterman con matrice a bande esegue gli allineamenti semi-globali delle sequenze per determinare regioni simili tra due sequenze. Piuttosto che confrontare la sequenza intera, l'algoritmo di Smith-Waterman confronta i segmenti di tutte le lunghezze possibili.

Ogni lettura paired-end viene valutata inizialmente in termini di allineamento con le sequenze sonda pertinenti per quella lettura.

- Read 1 (Lettura 1) è valutata rispetto al complemento inverso degli oligonucleotidi specifici per il locus a valle (Downstream Locus Specific Oligo, DLSO).
- Read 2 (Lettura 2) è valutata rispetto agli oligonucleotidi specifici per il locus a monte (Upstream Locus-Specific Oligo, ULSO).
- Se l'inizio di una lettura corrisponde a una sequenza sonda che presenta non più di tre differenze (mancate corrispondenze o variazioni causate da Indel anticipate), l'intera lunghezza della lettura viene allineata rispetto all'amplicone target per quella sequenza.
- Data la chimica del saggio, non si osservano Indel né nei DLSO né negli ULSO.

Gli allineamenti sono filtrati sui risultati degli allineamenti basati sulle percentuali di mancate corrispondenze sulla regione di interesse o sull'intero amplicone, in base alla lunghezza dell'amplicone. Gli allineamenti filtrati vengono scritti nei file di allineamento come non allineati e non vengono utilizzati per l'identificazione delle varianti.

Identificazione delle varianti

Sviluppato da Illumina, Pisces Variant Caller identifica le varianti presenti a bassa frequenza nel campione di DNA.

Pisces Variant Caller identifica i polimorfismi di singolo nucleotide (Single Nucleotide Polymorphism, SNP) in tre fasi:

- Considera ogni posizione nel genoma di riferimento separatamente
- Conteggia le basi in una data posizione per le letture allineate che sono sovrapposte in quella posizione
- Calcola un punteggio per la variante che misura la qualità dell'identificazione utilizzando il modello Poisson. Sono escluse le varianti con un punteggio qualitativo inferiore a Q30.

Le varianti vengono prima identificate per ogni raggruppamento in pool separatamente. Quindi, le varianti appartenenti a ogni raggruppamento in pool vengono confrontate e combinate in un singolo file di output. Se una variante soddisfa i criteri seguenti, la variante viene indicata come PASS (Superata) nel file di identificazione delle varianti (VCF):

- La variante è presente in entrambi i raggruppamenti
- Presenta una profondità complessiva di 900x (almeno 450x per raggruppamento)
- Presenta una frequenza della variante di ≥ 2,6% come riportato nel file unito VCF.

Pisces (Somatic Variant Caller)

Pisces esegue l'identificazione di varianti somatiche per identificare le varianti a bassa frequenza nei campioni di DNA. L'applicazione può essere eseguita su tutti i campioni e genera file VCF e gVCF.

Per maggiori informazioni, vedere la pagina Web github.com/Illumina/Pisces/wiki.

Visualizzazione della corsa e dei dati del campione

1. Dal pannello di controllo di Local Run Manager, fare clic sul nome della corsa.
2. Dalla scheda Run Overview (Panoramica corsa), rivedere le metriche della corsa di sequenziamento.
3. [Facoltativo] Fare clic sull'icona **Copy to Clipboard** (Copia negli appunti) per copiare il percorso della cartella degli output della corsa.
4. Fare clic sulla scheda Sequencing Information (Informazioni sequenziamento) per rivedere i parametri della corsa e le informazioni relative ai materiali di consumo.
5. Fare clic sulla scheda Samples and Results (Campioni e risultati) per visualizzare la posizione del report dell'analisi.
   - Se l'analisi è stata ripetuta, allargare l'elenco di controllo Select Analysis (Seleziona analisi) e selezionare l'analisi appropriata.
6. Fare clic sull'icona **Copy to Clipboard** (Copia negli appunti) per copiare il percorso della cartella Analysis (Analisi).

Per maggiori informazioni sulle schede Run Overview (Panoramica corsa) e Sequencing Information (Informazioni sequenziamento) e su come rimettere in coda l'analisi, vedere la *Guida di consultazione del software Local Run Manager (documento n. 1000000011880)*.

Report dell'analisi

I risultati dell'analisi sono riepilogati nella scheda Samples and Results (Campioni e risultati) e come un report aggregato nella cartella Alignment (Allineamento). È inoltre disponibile un report per ogni campione nel formato file PDF.

Informazioni sulla scheda Sample and Details (Campione e dettagli)

1 Fare clic su un campione nell'elenco per visualizzare il report per il campione.

Tabella 1 Tabella Run Information (Informazioni sulla corsa)

Intestazione colonna	Descrizione
Run Status (Stato corsa)	Indica se la corsa di sequenziamento è riuscita o non è riuscita.
Total Yield (GB) (Resa totale - GB)	Il numero di basi identificate nella corsa di sequenziamento. Mostra la soglia di superamento e lo stato di superato o non superato.
% ≥ Q30	La percentuale di letture nella corsa di sequenziamento con un punteggio qualitativo di 30 (Q30) o superiore. Mostra la soglia di superamento e lo stato di superato o non superato.
Sample Name (Nome del campione)	Il nome del campione fornito al momento della creazione della corsa.
Total PF Reads (Totale di letture che attraversano il filtro)	Il numero totale di letture che hanno attraversato il filtro.
Read 1% ≥ Q30 (Lettura 1 % ≥ Q30)	La percentuale di letture in Read 1 (Lettura 1) con un punteggio qualitativo di 30 (Q30) o superiore per il campione.
Read 2% ≥ Q30 (Lettura 2 % ≥ Q30)	La percentuale di letture in Read 2 (Lettura 2) con un punteggio qualitativo di 30 (Q30) o superiore per il campione.
Autosome call rate (Percentuale di identificazioni autosomiche)	Il numero di posizioni genomiche sugli autosomi (dal cromosoma 1 al 22) che soddisfano un valore di affidabilità prestabilito, diviso per il numero totale di posizioni genomiche autosomiche interrogate. La percentuale di identificazione viene descritta in base al singolo campione e riportata come percentuale calcolata come 1 meno (numero di posizioni autosomiche con identificazioni incomplete) diviso il numero totale di posizioni autosomiche sequenziate.

Tabella 2 Tabella Sample Reports (Report dei campioni)

Intestazione colonna	Descrizione
Sample (Campione)	Il nome del campione fornito al momento della creazione della corsa.
Report Date (Data report)	La data in cui è stato generato il report.
Sample Information (Informazioni sui campioni)	L'ID del campione fornito al momento della creazione della corsa, il numero totale delle letture che hanno attraversato il filtro nel campione, la percentuale di letture per un campione con un punteggio qualitativo di 30 (Q30) o superiore e una percentuale di identificazione autosomica.
Amplicon Summary (Riepilogo sull'amplicone)	Il numero totale di regioni degli ampliconi sequenziate e la lunghezza totale nelle coppie di basi degli ampliconi sequenziati nelle regioni target, per il campione in Pool A (Raggruppamento A) e in Pool B (Raggruppamento B) e il file manifest utilizzato in ciascun raggruppamento in pool. Il file manifest specifica le regioni del genoma di riferimento e del riferimento target utilizzate nella fase di allineamento.
Read Level Statistics (Statistiche del livello della lettura)	Il numero e la percentuale di letture per il campione che coprono ogni posizione nel riferimento per Read 1 (Lettura 1) e Read 2 (Lettura 2) in Pool A (Raggruppamento A) e in Pool B (Raggruppamento B).
Variants Summary (Riepilogo sulle varianti)	Il numero di SNV, inserzioni e delezioni per il campione che ha superato i valori suggeriti per determinare se i risultati di qualità rientrano in un intervallo accettabile.

Intestazione colonna	Descrizione
Coverage Summary (Riepilogo sulla copertura)	Il numero totale di basi allineate diviso per la dimensione della regione target e la percentuale delle regioni degli ampliconi con valori di copertura superiori alla soglia minima di copertura di 0,2* della copertura media dell'amplicone, per il campione in Pool A (Raggruppamento A) e in Pool B (Raggruppamento B).
Coverage Plots (Grafici della copertura)	Il grafico Coverage by Amplicon Region (Copertura per regione degli ampliconi) mostra la copertura sulle regioni degli ampliconi per il campione. Le regioni con valori di copertura inferiori alla soglia di copertura sono visualizzati in rosso. La media di tutti i valori è indicata da una linea arancione. Viene fornito un grafico per la copertura di Pool A (Raggruppamento A) e Pool B (Raggruppamento B).
Software Versions (Versioni software)	Le versioni del software quando è stato sequenziato il campione. Include la versione di MiSeq Operating Software, del software Local Run Manager, del software RTA e del modulo Somatic Variant.

File di output dell'analisi

I seguenti file di output dell'analisi vengono generati per il modulo di analisi Somatic Variant e forniscono i risultati dell'analisi per l'allineamento e l'identificazione delle varianti. I file di output dell'analisi si trovano nella cartella Alignment (Allineamento).

Nome file	Descrizione
Demultiplex (*.txt)	File intermedi che contengono un riepilogo dei risultati del demultiplex.
FASTQ (*.fastq.gz)	File intermedi che contengono le identificazione delle basi qualitativamente valutate. I file FASTQ rappresentano gli input principali per la fase di allineamento.
File di allineamento in formato BAM (*.bam)	Contiene le letture allineate per un dato campione.
File di identificazione delle varianti prima del raggruppamento in pool in formato VCF (*.vcf)	Contiene le varianti identificate in ogni posizione sia nel raggruppamento forward che nel raggruppamento reverse.
File di identificazione delle varianti per il genoma in formato VCF (*.genome.vcf.gz)	Contiene il genotipo per ogni posizione, sia che sia stato identificato come una variante o identificato come un riferimento.
File di identificazione delle varianti consenso in formato VCF (*.vcf.gz)	Contiene le varianti identificate in ogni posizione di entrambi i raggruppamenti in pool.
AmpliconCoverage_M1.tsv	Contiene le informazioni relative alla copertura per amplicone per campione per ogni file manifest fornito. M# rappresenta il numero del file manifest.

Formato file per il demultiplex

Il processo di demultiplex legge la sequenza d'indice collegata a ogni cluster per determinare il campione da cui è stato originato il cluster. La mappatura tra i cluster e il numero di campione viene scritto su un file di demultiplex (*.demux) per ogni tile della cella a flusso.

Il file di demultiplex è nel formato s_1_X.demux, dove X rappresenta il numero della tile.

I file di demultiplex iniziano con un'intestazione:

- Versione (valore intero di 4 byte), attualmente 1
- Conteggio dei cluster (valore intero di 4 byte)

Il resto del file consiste di numeri dei campioni per ogni cluster della tile.

Al termine della fase di demultiplex, il software genera un file di demultiplex chiamato

DemultiplexSummaryF1L1.txt.

- Nel nome del file, **F1** rappresenta il numero della cella a flusso.
- Nel nome del file, **L1** rappresenta il numero della corsia.
- I risultati di demultiplex in una tabella con una riga per tile e una colonna per campione, incluso campione 0.
- Le sequenze più frequenti nelle letture indici.

Formato file FASTQ

FASTQ è un formato file di testo che contiene le identificazioni delle basi e i valori qualitativi per ogni lettura. Ciascun record contiene quattro righe:

- L'identificatore
- La sequenza
- Un segno più (+)
- I punteggi qualitativi su scala Phred in un formato codificato ASCII + 33

L'identificatore è formattato come

@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#:<#<;<<<????#=
```

Formato file BAM

Un file BAM (*.bam) è una versione binaria compressa di un file SAM utilizzato per rappresentare le sequenze allineate fino a un massimo di 128 Mb. I formati SAM e BAM sono descritti nei dettagli in samtools.github.io/hts-specs/SAMv1.pdf.

I file BAM utilizzano il formato di denominazione dei file NomeCampione_S#.bam dove # è il numero del campione in base all'ordine in cui i campioni sono elencati per la corsa.

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 degli allineamenti sono associati con le informazioni specifiche contenute nella sezione di 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, le coordinate di avvio, la qualità dell'allineamento e la stringa del descrittore della corrispondenza.

La sezione degli allineamenti include le seguenti informazioni per ciascuna lettura o coppia di letture:

- **AS:** qualità dell'allineamento paired-end
- **BC:** tag del codice a barre, che indica l'ID del campione sottoposto a demultiplex associato con la lettura.
- **SM:** qualità dell'allineamento unidirezionale.
- **XC:** stringa del descrittore della corrispondenza.
- **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.

Formato file VCF

Il formato per l'identificazione delle varianti (Variant Call Format, VCF) è un formato file comune sviluppato dalla comunità scientifica nel campo della genomica. Contiene le informazioni sulle varianti identificate nelle posizioni specifiche in un genoma di riferimento. I file VCF terminano con il suffisso .vcf.

L'intestazione del file VCF include la versione del formato file VCF, la versione dell'identificatore delle varianti ed elenca le annotazioni utilizzate nella parte restante 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 dei dati del file VCF contiene le informazioni relative a una variante.

Intestazioni del file VCF

Intestazione	Descrizione
CHROM (Cromosoma)	Il cromosoma del genoma di riferimento. I cromosomi vengono visualizzati nello stesso ordine del file di riferimento FASTQ.
POS (Posizione)	La posizione della singola base della variante nel cromosoma di riferimento. Per gli SNP, questa posizione è la base di riferimento con la variante; per le Indel o le delezioni, questa posizione è la base di riferimento immediatamente prima della variante.
ID (Identificazione)	Il numero rs per la variante ottenuta da dbSNP.txt, se applicabile. Se sono presenti numeri rs multipli in questa posizione, l'elenco è delimitato da punti e virgole. Se non esistono voci dbSNP in questa posizione, viene usato un indicatore di valore mancante ('.').
REF (Riferimento)	Il genotipo di riferimento. Ad esempio, una delezione di T singolo è rappresentata come TT di riferimento e T alternato. Una variante di singolo nucleotide da A a T è rappresentata come A di riferimento e T alternato.
ALT (Alternato)	Gli alleli diversi dalla lettura di riferimento. Ad esempio, un'inserzione di T singolo è rappresentata come A di riferimento e AT alternata. Una variante di singolo nucleotide da A a T è rappresentata come A di riferimento e T alternato.
QUAL (Qualità)	Un punteggio qualitativo su scala Phred assegnato da Variant Caller. Punteggi elevati indicano un'affidabilità superiore nella variante e minore probabilità di errori. Per un punteggio qualitativo di Q, la probabilità di errore stimata è $10^{-(Q/10)}$. Ad esempio, il set di identificazioni con punteggio qualitativo Q30 ha una percentuale di errore di 0,1%. Diversi Variant Caller assegnano punteggi qualitativi in base ai propri modelli statistici, che sono molto elevati rispetto alla percentuale di errore osservata.

Annotazioni del file VCF

Intestazione	Descrizione
FILTER (Filtro)	Se vengono attraversati tutti i filtri, nella colonna Filter (Filtro) viene riportato PASS (Superato). • LowDP (Profondità bassa): applicato ai siti con profondità di copertura inferiore a 450x in entrambi i raggruppamenti in pool. Per le posizioni degli ampliconi coperte sia nella lettura forward che nelle lettura reverse, è equivalente a 900x della lettura unidirezionale. • LowGQ (Qualità genotipizzazione bassa): la qualità di genotipizzazione (GQ) è inferiore al valore di cutoff. • Q30:: punteggio qualitativo superiore a 30. • LowVariantFreq (Bassa frequenza variante): la frequenza delle varianti è inferiore alla soglia. • PB (Distorsione sonda): la distorsione della sonda per il raggruppamento. La variante non identificata o identificata con una bassa frequenza in uno o due raggruppamenti sulla sonda. • R3x6: il numero di ripetizioni adiacenti (con lunghezza di 1-3 bp) rispetto alle identificazioni delle varianti ≥ 6. • SB (Distorsione soglia): la distorsione del filamento supera la soglia data.
INFO (Informazioni)	Le possibili voci della colonna INFO (Informazioni) includono: • AC (Conteggio alleli): il conteggio degli alleli nei genotipi per ciascun allele ALT (Alternato), nello stesso ordine in cui sono elencati. • AF (Frequenza allelica): la frequenza allelica per ciascun allele ALT (Alternato), nello stesso ordine in cui sono elencati. • AN (Numero alleli): il numero totale di alleli nei genotipi identificati. • CD (Regione codificante): un indicatore per segnalare che l'SNP si trova nelle regione codificante di almeno una voce 1 RefGene. • DP (Profondità): la profondità (il numero delle identificazioni delle basi allineate su una posizione e utilizzato nell'identificazione delle varianti). • Exon (Esone): un elenco separato da virgola delle regioni esoniche lette da RefGene. • FC (Conseguenze funzionali): le conseguenze funzionali. • GI (Geni identificati): un elenco separato da virgola degli ID dei geni letti da RefGene. • QD (Qualità profondità): l'affidabilità/qualità della variante per la profondità. • TI (Trascritti identificati): un elenco separato da virgola degli ID del trascritto letti da RefGene.
FORMAT (Formato)	La colonna del formato elenca i campi separati da due punti. Ad esempio, GT:GQ. L'elenco dei campi fornito dipende dall'identificatore delle varianti utilizzato. I campi disponibili includono: • AD (Assi): le voci nel formato X,Y, dove X rappresenta il numero delle identificazioni di riferimento e Y il numero di identificazioni alternate. • DP (Profondità): la profondità approssimativa della lettura; letture con MQ=255 o con accoppiamenti non corretti sono filtrate. • GQ (Qualità genotipo): la qualità del genotipo. • GQX (Qualità genotipo X): la qualità del genotipo. GQX rappresenta il minimo del valore GQ (Qualità genotipo) e la colonna QUAL (Qualità). In generale, questi valori sono simili; se si prende il valore minimo, GQX diventa la misura più conservativa della qualità del genotipo. • GT (Genotipo): il genotipo. 0 corrisponde alla base di riferimento, 1 corrisponde alla prima voce nella colonna ALT (Alternato), e così via. Il simbolo barra in avanti (/) indica che non è disponibile alcuna informazioni sulla fase. • NC (Nessuna identificazione): la frazione delle basi che non sono state identificate o con una qualità di identificazione delle basi inferiore alla soglia minima. • NL (Livello rumore): il livello del rumore; il valore stimato del rumore dell'identificazione delle basi in quella posizione. • PB (Distorsione sonda): la distorsione della sonda per il raggruppamento. Valore prossimi allo 0 indicano una maggiore distorsione verso un raggruppamento sulla sonda e una minore affidabilità nell'identificazione della variante. • SB (Distorsione filamento): la distorsione del filamento in quella posizione. Valore negativi più alti indicano una distorsione inferiore; valori prossimi allo 0 indicano una maggiore distorsione. • VF (Frequenza variante): la frequenza della variante; la percentuale di letture che supportano l'allele alternato.
SAMPLE (Campione)	La colonna dei campioni fornisce il valore specificato nella colonna FORMAT (Formato).

File VCF del genoma

I file VCF del genoma (gVCF) sono file VCF v4.1 che seguono una serie di convenzioni per rappresentare tutti i siti entro il genoma in un formato ragionevolmente compatto. I file gVCF (*.genome.vcf.gz) includono tutti i siti entro la regione di interesse in un singolo file per ogni campione.

I file gVCF non mostrano alcuna identificazione nelle posizioni che non hanno superato tutti i filtri. Un indicatore di genotipizzazione (GT), **./.** indica che non è stata rilevata alcuna identificazione.

Per maggiori informazioni, vedere la pagina Web sites.google.com/site/gvcftools/home/about-gvcf.

File VCF per raggruppamento in pool e di consenso

Il flusso di lavoro di Somatic Variant genera due set di file di identificazione delle varianti.

- **File VCF per raggruppamento in pool**: contengono le identificazioni delle varianti nel raggruppamento in pool forward o nel raggruppamento in pool reverse. I file di raggruppamento in pool vengono scritti nella cartella VariantCallingLogs (Registri identificazioni delle varianti).
- **File VCF di consenso**: contengono le varianti identificate da entrambi i raggruppamenti in pool. I file di consenso vengono scritti nella cartella Alignment (Allineamento).

I file VCF per raggruppamento in pool e di consenso includono sia i file VCF (*.vcf) che i file gVCF (*.genome.vcf) e seguono la convenzione di denominazione seguente, dove S# rappresenta l'ordine in cui i campioni sono elencati per la corsa:

- **Report per tutti i siti**: NomeCampione_S#.genome.vcf
- **Report solo per le varianti**: NomeCampione_S#.vcf

Il software confronta i file VCF per raggruppamento in pool e combina i dati in ogni posizione per creare un file VCF di consenso per il campione.

Le identificazioni delle varianti da ogni raggruppamento in pool sono unite in file VCF di consenso utilizzando i seguenti criteri.

Criteri	Risultato
Una identificazione di riferimento in ciascun raggruppamento in pool	Identificazione di riferimento
Una identificazione di riferimento in un raggruppamento in pool e una identificazione della variante nell'altro raggruppamento in pool	Identificazione delle varianti filtrata
Le identificazioni delle varianti corrispondenti con frequenze simili in ciascun raggruppamento in pool	Identificazione delle varianti
Le identificazioni delle varianti corrispondenti con frequenze significativamente diverse in ciascun raggruppamento in pool	Identificazione delle varianti filtrata
Le identificazioni delle varianti prive di corrispondenza in ciascun raggruppamento in pool	Identificazione delle varianti filtrata

Le metriche da ciascun raggruppamento in pool sono unite utilizzando i seguenti valori.

Metrica	Valore
Profondità	Aggiunta di profondità da entrambi i raggruppamenti in pool
Frequenza delle varianti	Il conteggio totale delle varianti diviso per la profondità di copertura totale
Punteggio qualitativo	Il valore minimo di entrambi i raggruppamenti in pool

File di copertura dell'amplicone

Per ogni file manifest viene generato un file di copertura dell'amplicone. M# nel nome del file rappresenta il numero di file manifest come elencato nella tabella dei campioni per la corsa.

Ciascun file include una riga di intestazione che contiene l'ID campione associato con il file manifest. Sotto la riga dell'intestazione sono presenti tre colonne che elencano le seguenti informazioni:

- L'ID del target come elencato nel file manifest.
- La profondità di copertura delle letture che hanno attraversato il filtro.
- La profondità di copertura totale.

File di output supplementari

I seguenti file di output forniscono informazioni supplementari o riepilogano i risultati della corsa e gli errori dell'analisi. Sebbene questi file non siano richiesti per valutare i risultati dell'analisi possono essere utilizzati per la risoluzione dei problemi. Tutti i file si trovano nella cartella Alignment (Allineamento), se non diversamente indicato.

Nome file	Descrizione
AnalysisLog.txt	Il registro dell'elaborazione che descrive tutte le fasi che si sono verificate durante l'analisi della cartella della corsa attuale. Questo file non contiene messaggi di errore. Si trova nella cartella Alignment (Allineamento).
AnalysisError.txt	Il registro dell'elaborazione che elenca qualsiasi errore verificatosi durante l'analisi. Questo file sarà vuoto se non si è verificato alcun errore. Si trova nella cartella Alignment (Allineamento).
DemultiplexSummaryF1L1#.txt	Riporta i risultati di demultiplex in una tabella con una riga per tile e una colonna per campione. # rappresenta la corsia 1, 2, 3 o 4 della cella a flusso. Si trova nella cartella Alignment (Allineamento).
AmpliconRunStatistics.xml	Contiene un riepilogo delle statistiche specifiche per la corsa. Si trova nella cartella Alignment (Allineamento).

Cartella Analysis (Analisi)

La cartella dell'analisi contiene i file generati dal software Local Run Manager.

La relazione tra la cartella di output e la cartella dell'analisi sono riepilogati qui di seguito:

- Durante il sequenziamento, Real-Time Analysis (RTA) popola la cartella di output con i file generati durante l'analisi delle immagini, l'identificazione delle basi e il calcolo del punteggio qualitativo.
- RTA copia i file nella cartella dell'analisi in tempo reale. Dopo che RTA ha assegnato un punteggio qualitativo a ciascuna base per ciascun ciclo, il software scrive il file RTAComplete.txt in entrambe le cartelle.
- L'analisi viene avviata quando è presente il file RTAComplete.txt.
- Durante l'analisi, Local Run Manager scrive i file di output nella cartella dell'analisi, quindi copia di nuovo i file nella cartella di output.

Cartelle per l'allineamento

Ogni volta che un'analisi viene rimessa in coda, Local Run Manager crea una cartella Alignment (Allineamento) chiamata **Alignment_N**, dove N rappresenta un numero sequenziale.

Struttura della cartella

- **Alignment**: contiene i file *.bam, *.vcf, FASTQ e file specifici per il modulo di analisi.
  - **Date and Time Stamp**: il timbro data_ora dell'analisi espresso come AAAAMMGG_OOMMSS
    - AnalysisError.txt
    - AnalysisLog.txt
    - AmpliconRunStatistics.xml
    - Sample1.genome.vcf.gz
    - Sample1.coverage.csv
    - Sample1.report.pdf
    - Sample1.summary.csv
    - Sample1.vcf.gz
    - Sample1.bam
    - **FASTQ**
      - **Sample1**
      - **Stats**
        - DemuxSummaryF1L1.txt
        - FastqSummaryF1L1.txt
- **Data**
  - **Intensities**
    - **BaseCalls**
      - **L001**: contiene una sottocartella per ciclo, ciascuna contenente i file *.bcl.
    - **L001**: contiene i file *.locs, uno per ciascuna tile.
  - **RTA Logs**: contiene i file di registro ottenuti dal software di analisi RTA.
- **InterOp**: contiene i file binari utilizzati per riportare le metriche della corsa di sequenziamento.
- **Logs**: contiene i file di registro che descrivono le fasi eseguite durante il sequenziamento.
- RTAComplete.txt
- RunInfo.xml
- runParameters.xml

Assistenza tecnica

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

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

Numeri di telefono dell'Assistenza clienti Illumina

Area geografica	Gratuito	Regionale
Nord America	+1.800.809.4566	
Australia	+1.800.775.688	
Austria	+43 800006249	+43 19286540
Belgio	+32 80077160	+32 34002973
Cina	400.635.9898	
Danimarca	+45 80820183	+45 89871156
Finlandia	+358 800918363	+358 974790110
Francia	+33 805102193	+33 170770446
Germania	+49 8001014940	+49 8938035677
Hong Kong	800960230	
Irlanda	+353 1800936608	+353 016950506
Italia	+39 800985513	+39 236003759
Giappone	0800.111.5011	
Paesi Bassi	+31 8000222493	+31 207132960
Nuova Zelanda	0800.451.650	
Norvegia	+47 800 16836	+47 21939693
Singapore	+1.800.579.2745	
Spagna	+34 911899417	+34 800300143
Svezia	+46 850619671	+46 200883979
Svizzera	+41 565800000	+41 800200442
Taiwan	00806651752	
Regno Unito	+44 8000126019	+44 2073057197
Altri paesi	+44.1799.534000	

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

Documentazione dei prodotti: la documentazione dei prodotti in formato PDF può essere scaricata dal sito Web Illumina. Andare al sito support.illumina.com, selezionare un prodotto, quindi fare clic su **Documentation & Literature** (Documentazione e letteratura).

Documento n. 1000000012183 v01 ITA

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

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