Accuratezza migliorata in Germline Small Variant Calling grazie alla piattaforma DRAGEN™

Diversi algoritmi che migliorano l'accuratezza consentono il rilevamento di varianti piccole con elevata sensibilità e specificità, mantenendo al contempo gli standard DRAGEN per la velocità di calcolo

Introduzione

Con i progressi della tecnologia di sequenziamento di nuova generazione (Next-Generation Sequencing, NGS), il volume dei dati di sequenziamento generati continua a crescere in maniera esponenziale. Assieme a questa crescita aumenta la richiesta di metodi di analisi veloci ed efficienti che siano in grado di mantenere elevati standard di accuratezza nell'identificazione delle varianti. La piattaforma DRAGEN (Dynamic Read Analysis for Genomics, analisi dinamica delle letture per genomica) Bio-IT Illumina fornisce l'analisi secondaria ultra veloce dei dati ottenuti dal sequenziamento NGS. La piattaforma DRAGEN utilizza la tecnologia altamente riconfigurabile delle matrici di porte logiche programmabili sul campo (Reconfigurable Field-Programmable Gate Array, FPGA) per accelerare drasticamente l'analisi secondaria dei dati NGS, inclusi mappatura, allineamento e identificazione delle varianti.

Le caratteristiche principali della piattaforma DRAGEN consentono di affrontare le sfide comuni dell'analisi genomica, come i lunghi tempi di calcolo e gli enormi volumi di dati. La piattaforma DRAGEN fornisce velocità, flessibilità, accuratezza ed efficacia in termini di costi. La natura riprogrammabile della piattaforma DRAGEN consente di migliorare gli algoritmi per andare incontro alle nuove applicazioni NGS. La velocità della piattaforma consente agli sviluppatori di iterare velocemente le progettazioni degli algoritmi utilizzando metodi di calcolo intensivi che sono improponibili con i modelli tradizionali basati solo su software. In questo modo l'accuratezza della piattaforma DRAGEN viene continuamente migliorata con nuove versioni e DRAGEN fornisce ora una soluzione eccellente per l'identificazione di varianti piccole nel sequenziamento dell'intero genoma (Whole-Genome Sequencing, WGS) germinale.

Questa nota sulle applicazioni descrive i recenti miglioramenti nella piattaforma DRAGEN Bio-IT Illumina per la rapida analisi secondaria e dimostra l'accuratezza e la velocità utilizzando tre set di dati WGS accessibili pubblicamente. Questo è lo studio comparativo di DRAGEN v3.2.8 rispetto ad altri software, inclusi BWA-MEM+GATK4 e DRAGEN v2 (Figura 1). I risultati dell'identificazione delle varianti generati da ciascun software sono stati confrontati con un "set vero" per le identificazioni di riferimento al fine di identificare i falsi positivi (False Positive, FP) e i falsi negativi (False Negative, FN). Le metriche utilizzate nei confronti tra i software sono metriche end-to-end dei tempi di elaborazione e di accuratezza come reidentificazione, precisione e FP+FN. La combinazione di velocità, accuratezza e un'ampia gamma di applicazioni disponibili consente alla piattaforma DRAGEN di rivoluzionare il campo dell'analisi genomica.

Algoritmi DRAGEN v3 per migliorare l'accuratezza

DRAGEN v3 implementa i più recenti aggiornamenti degli algoritmi per il rilevamento di polimorfismi di singolo nucleotide (Single-Nucleotide Polymorphism, SNP) e inserzioni/delezioni (Indel), che forniscono precisione e sensibilità analitica migliorate. Per l'identificazione delle varianti sono state migliorate quattro aree: modello di errori di Indel specifico per i campioni, rigorosi modelli matematici di errori pile-up correlati, un approccio ottimizzato per rappresentare in modo esaustivo un numero esponenziale di candidati per l'aplotipo in regioni di rumore o ricche di varianti e aumento in base a colonna dell'elenco di eventi generati dall'assembly del grafico De-Bruijn. Tali aggiornamenti offrono un modesto incremento nella velocità, elevando però gli standard in accuratezza rispetto ai software valutati in questo documento. L'appendice offre una descrizione dettagliata di ciascun miglioramento eseguito sull'algoritmo.

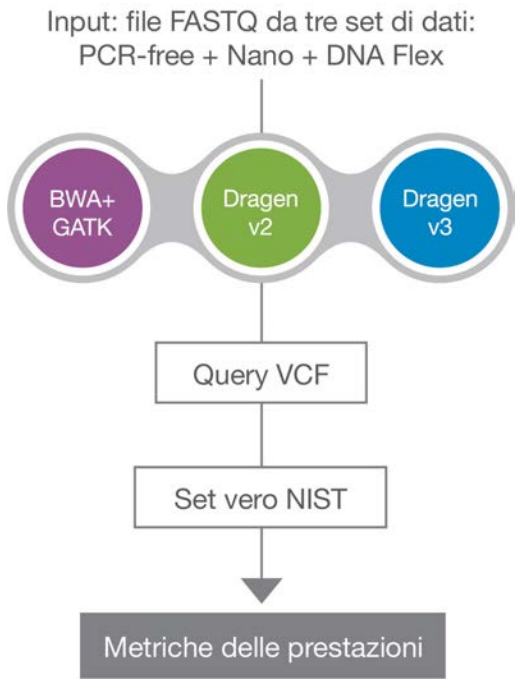

Figura 1: Progettazione dello studio comparativo: i file FASTQ generati da tre set di dati sono stati eseguiti su tre software di analisi per generare file di query in formato VCF. Variant Calling Assessment Tool (VCAT) è stato utilizzato per identificare TP, FP e FN in base al confronto delle identificazioni delle varianti rispetto alle varianti di riferimento contenute nel set vero NIST.

Metodi

Sono state seguite accuratamente le pratiche migliori raccomandate per lo studio comparativo.[1] Per dimostrare la velocità e l'accuratezza con DRAGEN v3, è stato eseguito uno studio di confronto utilizzando tre set di dati ottenuti da diverse preparazioni delle librerie e generati da un campione NA12878 (Figura 1). In sintesi, il file FASTQ ottenuto da ciascun set di dati è stato utilizzato come input per l'analisi secondaria eseguita da software indipendenti (DRAGEN v3.2.8, DRAGEN v2, e BWA+GATK[2]). I file VCF generati da ciascun software (QUERY VCF) sono stati caricati in un progetto in BaseSpace™ Sequence Hub. Variant Calling Assessment Tool (VCAT v3.1.1 con Hap.py versione 0.3.10) è stato utilizzato per confrontare ciascun file QUERY VCF rispetto al "set vero" di varianti di riferimento al fine di individuare le identificazioni delle varianti vere o false. I risultati sono stati raccolti e inseriti in tabelle per il confronto tra i software. Tutti i dati di input, i risultati dell'analisi e gli strumenti di valutazione sono disponibili gratuitamente nel progetto BaseSpace.[3] L'appendice descrive dettagliatamente i metodi.

Risultati dello studio comparativo

I risultati ottenuti dai confronti tra durate e accuratezza dimostrano che DRAGEN fornisce una soluzione efficace per l'analisi secondari dei dati NGS.

Accuratezza di DRAGEN: FP+FN, reidentificazione e precisione

Sebbene la piattaforma DRAGEN v2 fosse già competitiva rispetto a soluzioni informatiche leader nel settore, DRAGEN v3 presenta diverse nuove modifiche (descritte nella sezione dei metodi degli algoritmi) che offrono miglioramenti significativi nell'accuratezza. I risultati di questo studio comparativo dimostrano inoltre che i miglioramenti di DRAGEN v3 rendono la piattaforma superiore rispetto agli altri software di analisi più utilizzati (inclusa una precedente versione di DRAGEN) per tutte le metriche di accuratezza analizzate nello studio.

Quando è stata valutata la metrica FP+FN per il rilevamento delle SNV, DRAGEN v3 ha fornito un'accuratezza superiore rispetto a BWA+GATK4 e DRAGEN v2 per tutti e tre i set di dati (Figura 2). Quando è stata valutata la metrica FP+FN per il rilevamento delle Indel, DRAGEN v3 è risultato migliore rispetto a BWA+GATK4 per tutti e tre i set di dati, mostrando al contempo ulteriore miglioramento tra DRAGEN v3 e DRAGEN v2 (Figura 3).

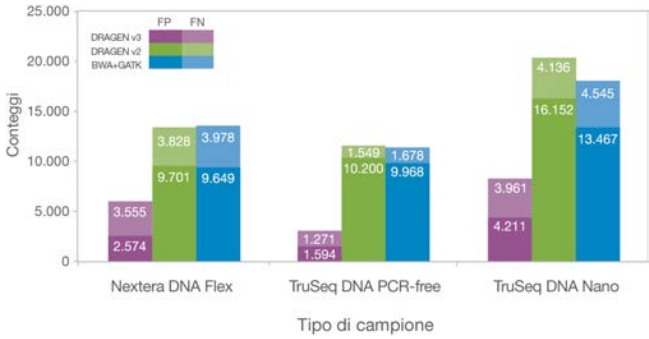

Figura 2: Falsi positivi e falsi negativi con il rilevamento di SNV: i file dei dati non elaborati (FASTQ) ottenuti da tre set di dati sono stati analizzati con tre software indipendenti. Ciascun set di dati (TruSeq DNA PCR-free, Nextera DNA Flex e TruSeq DNA Nano) è stato generato dal DNA del campione NA12878 e le identificazioni delle varianti (VCF) ottenute da ciascun software di analisi sono state confrontate con il set vero NIST (anch'esso basato sul campione NA12878) per identificare i FP e i FN.

Nella valutazione delle metriche di precisione e di reidentificazione, sono risultati evidenti i vantaggi dei miglioramenti degli algoritmi di DRAGEN v3 sia per il rilevamento degli SNP che delle Indel. I valori per la precisione e per la reidentificazione sono coerenti sopra il 99% per tutti i software e con ogni set di dati di rilevamento delle SNV (Tabella 1). Per l'identificazione di SNP, DRAGEN v2 era compatibile con BWA+GATK4. Ma DRAGEN v3 mostra significativi miglioramenti sia nella reidentificazione che nella precisione rispetto agli altri due software. Per l'identificazione di Indel, DRAGEN v2 ha mostrato un'accuratezza superiore rispetto a BWA+GATK4, mentre DRAGEN v3 ha dimostrato un ulteriore miglioramento rispetto a DRAGEN v2 sia nella reidentificazione che nella precisione (Tabella 2).

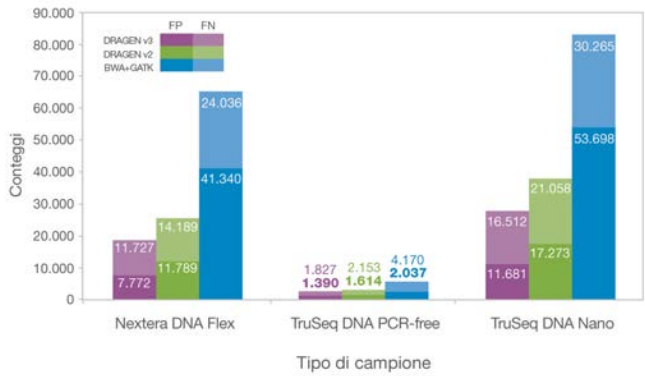

Figura 3: Falsi positivi e falsi negativi con il rilevamento di Indel: i file dei dati non elaborati (FASTQ) ottenuti da tre set di dati sono stati analizzati con tre software indipendenti. Ciascun set di dati (TruSeq DNA PCR-free, Nextera DNA Flex e TruSeq DNA Nano) è stato generato dal DNA del campione NA12878 e le identificazioni delle varianti (VCF) ottenute da ciascun software di analisi sono stati confrontati con il set vero NIST (anch'esso basato sul DNA del campione NA12878) per identificare i FP e i FN.

Tabella 1: Sensibilità e specificità nel rilevamento di SNV

	Precisione			Reidentificazione		
Set di dati	DRAGEN v3	DRAGEN v2	BWA+GATK	DRAGEN v3	DRAGEN v2	BWA+GATK
TruSeq DNA PCR-free	99,95%	99,68%	99,69%	99,96%	99,95%	99,95%
Nextera DNA Flex	99,92%	99,70%	99,70%	99,89%	99,88%	99,88%
TruSeq DNA Nano	99,87%	99,50%	99,58%	99,88%	99,87%	99,86%

Tabella 2: Sensibilità e specificità nel rilevamento di Indel

	Precisione			Reidentificazione		
Set di dati	DRAGEN v3	DRAGEN v2	BWA+GATK	DRAGEN v3	DRAGEN v2	BWA+GATK
TruSeq DNA PCR-free	99,71%	99,66%	99,58%	99,62%	99,55%	99,13%
Nextera DNA Flex	98,37%	97,54%	91,53%	97,56%	97,05%	95,01%
TruSeq DNA Nano	97,56%	96,39%	89,37%	96,57%	95,63%	93,71%

Velocità DRAGEN

Le durate delle corse ottenute con DRAGEN sono state confrontate e raccolte sia per la soluzione sul cloud che in laboratorio. Per la soluzione in laboratorio, DRAGEN v3 è stato confrontato con BWA+GATK con entrambi i software eseguiti sullo stesso server. Per la soluzione sul cloud, DRAGEN v3 è stato eseguito su BaseSpace Sequence Hub ed è stato confrontato con BWA+GATK eseguito su Terra.[4]

DRAGEN accelera sia il processo di mappatura sia l'identificazione delle varianti e possono essere eseguiti indipendentemente. Sebbene qui non indicato, vale la pena notare che a monte dell'analisi secondaria, DRAGEN supporta anche la conversione BCL2FASTQ accelerata, che migliora in modo significativo la velocità e l'efficienza generando file FASTQ identici. Vale la pena inoltre notare che DRAGEN genera automaticamente un elenco completo di metriche QC, sia per la mappatura che per l'identificazione delle varianti, senza ulteriore spreco di tempo. Al contrario di altri software che si affidano a lenti strumenti di terze parti (ad esempio, Samtools, Picard) per acquisire le metriche QC con significativo spreco di tempo.

Quando la velocità di esecuzione è stata misurata con i software eseguiti sullo stesso server in laboratorio, DRAGEN v3 è stato significativamente più veloce rispetto a BWA+GATK, con velocità nell'ordine di 16-18× superiore (Figura 4).

Nella misurazione della velocità di esecuzione con i software sul cloud, DRAGEN v3 eseguito su BaseSpace Sequence Hub è risultato significativamente più veloce rispetto a BWA+GATK eseguito su Terra, con velocità nell'ordine di 13-16× superiore.

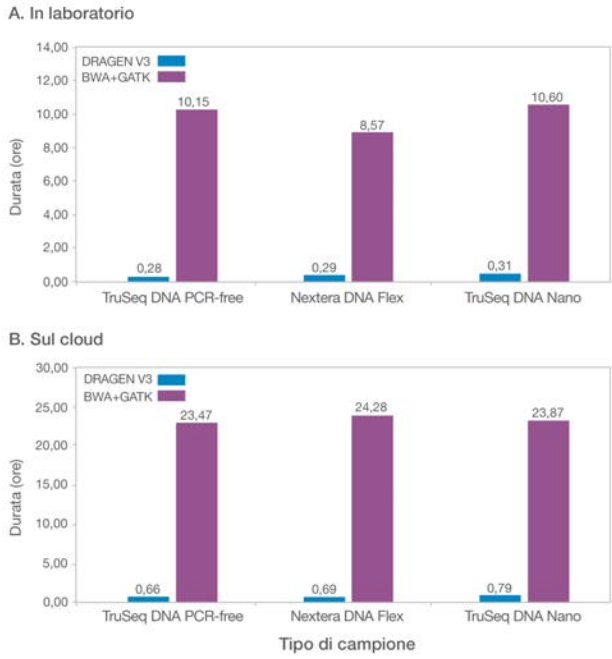

Figura 4: Confronti delle durate dell'analisi: (A) DRAGEN v3 e BWA+GATK eseguiti sullo stesso server in laboratorio. (B) DRAGEN v3 eseguito su BaseSpace Sequence Hub e confrontato con BWA+GATK eseguito su Terra.

Riepilogo

Le applicazioni di genomica mirano alla caratterizzazione precisa di regioni difficili del genoma e alla misurazione delle identificazioni della frequenza dell'allele minore da campioni con elevato livello di rumore. DRAGEN dimostra di essere la piattaforma più adatta per l'elaborazione dei futuri dati NGS sia per efficienza che per accuratezza.

La velocità di DRAGEN non solo consente ai ricercatori di affrontare la crescente quantità di dati generati dagli strumenti NGS, ma, altrettanto importante, consente anche la veloce iterazione per continui miglioramenti dei propri algoritmi per fornire elevata accuratezza.

Appendice

Descrizione dettagliata dei nuovi algoritmi

Modello di errore della PCR specifico per il campione

Una delle sfide nell'identificazione delle varianti è la possibilità di distinguere gli errori delle Indel da vere varianti. Per fare ciò, i Variant Caller utilizzano un Hidden Markov Model (HMM), che modella il comportamento statistico degli errori delle Indel, come parte del calcolo delle probabilità. Il modello HMM dispone solitamente di parametri di input, Gap Open Penalty (GOP) e Gap Continuation Penalty (GCP), che sono direttamente correlati alla percentuale di errore delle Indel [ossia, percentuale di errore delle Indel = f (GOP,GCP)]. Gli errori delle Indel sono probabilmente dovuti alla presenza di ripetizioni in tandem corte (Short Tandem Repeat, STR) e la probabilità di errore (quindi GOP e GCP) potrebbe dipendere sia dal periodo che dalla lunghezza delle STR. Il processo di errore potrebbe essere significativamente diverso fra un set di dati e l'altro, in base a fattori come l'amplificazione mediante PCR. Per il rilevamento accurato, è importante utilizzare i parametri HMM che modellano accuratamente il processo di errore in base al campione. Tuttavia, i Variant Caller spesso utilizzano parametri fissi o funzioni predeterminate non specifiche per il campione che non modellano accuratamente il processo di errore, fornendo quindi scarse prestazioni di rilevamento.

L'autocalibrazione HMM implementata in DRAGEN v3 affronta il problema sopra indicato stimando i parametri della PCR direttamente dal set di dati da elaborare. Questa operazione viene eseguita dopo la mappatura e l'allineamento e prima dell'identificazione delle varianti, senza conoscere dati empirici veri e senza utilizzare database esterni di mutazioni note. I parametri dipendono sia del periodo STR che dalla lunghezza delle ripetizioni.

Per un dato periodo e lunghezza di STR, viene selezionato un set di loci N con il periodo e la lunghezza desiderati e l'algoritmo esamina i pile-up delle letture mappate su questi loci, contando le Indel osservate in ciascun locus. L'idea alla base di questo approccio è che prendendo in considerazione un numero sufficiente di loci, è possibile stimare accuratamente i parametri di interesse. Per fare ciò vengono individuati i parametri che massimizzano la probabilità di generare un set di pile-up di N osservati. Se il numero di parametri che consentono di massimizzare la probabilità è sufficientemente basso (ad esempio, 2), è possibile eseguire una ricerca completa. Nell'attuale implementazione di DRAGEN v3, l'ottimizzazione viene eseguita su due parametri: GOP e alfa, che indicano la probabilità di varianti Indel di qualsiasi lunghezza non zero. Per ogni periodo e lunghezza di STR preso in considerazione, la ricerca genera output GOP e alfa che massimizzano la probabilità di generare il set di pile-up di N osservati; questi valori vengono utilizzati come input per HMM. È possibile allargare la ricerca a più di due parametri per ottenere ulteriori miglioramenti.

Calo della qualità della base (Base Quality Dropoff, BQD)

I Variant Caller convenzionali sono progettati con il presupposto che gli errori di sequenziamento sono indipendenti dalle letture; in base a questo presupposto, è poco probabile che più errori identici si verificheranno in un determinato locus. Tuttavia, dopo aver analizzato i set di dati NGS, è stato osservato che gli incrementi di errori erano più comuni di quanto previsto dal presupposto di indipendenza e questi incrementi possono risultare in lotti di falsi positivi.

Fortunatamente, questi errori hanno caratteristiche specifiche che li differenziano dalle varianti vere. L'algoritmo di calo della qualità della base (BQD) implementato in DRAGEN v3 è un meccanismo di rilevamento che sfrutta determinate proprietà di questi errori (distorsioni del filamento, posizionamento dell'errore nella lettura, bassa qualità media della base al locus di interesse) e le incorpora nel calcolo delle probabilità del genotipizzatore in modo semplice ed efficace. Vengono aggiunte ipotesi su nuovi genotipi candidati all'elenco di legacy dei genotipi diploidi (quelli che presumono errori di pile-up indipendenti). Ad esempio, nel caso di un locus con allele 1 ALT, oltre a prendere in considerazione P(G00|R), P(G01|R), P(G11|R), aggiungiamo altre due ipotesi come P(G00,E1|R) e P(G11,E0|R), dove gli alleli E0 e E1 rappresentano l'allele ALT proveniente dall'errore di sequenziamento. Le proprietà di questi errori, come le distorsioni del filamento, la posizione dell'errore nella lettura e la qualità media della base sono incorporate nel calcolo di P(G00,E1|R) e P(G11,E0|R). Quindi il genotipo rilevante viene preso su max (max(P(G00|R), P(G00,E1|R)), P(G01|R), max(P(G11|R), P(G11,E0|R))).

La possibilità di caratterizzare gli errori di sequenziamento correlati direttamente sui principali risultati ottenuti dal Variant Caller offre un significativo incremento nella specificità perché un gruppo di identificazioni FP viene rimosso. Contribuisce inoltre alla sensibilità in quanto corregge gli errori del genotipo.

Rilevamento di letture estranee (Foreign Read Detection, FRD)

I Variant Caller convenzionali trattano gli errori di mappatura come eventi di errore indipendenti per lettura, ignorando il fatto che tali errori di solito si verificano in gruppi. Di conseguenza le identificazioni delle varianti potrebbero essere generate con punteggi di affidabilità molto elevati nonostante valori MAPQ basso e/o AF non equilibrato. Per mitigare questo problema, i Variant Caller convenzionali di solito filtrano le letture a monte dell'identificazione delle varianti, in base a una soglia MAPQ (ossia, letture con soglia MAPQ< sono escluse dal calcolo). Tuttavia, queste esclusioni sono prove importanti generate dal Variant Caller e la soppressione dei fasi positivi è scarsa.

DRAGEN v3 ha implementato il rilevamento di letture estranee (FRD), un'estensione dell'algoritmo di genotipizzazione legacy, incorporando una ulteriore ipotesi in base alla quale alcune letture nel pile-up sono letture estranee [ossia, la loro effettiva posizione è altrove nel genoma di riferimento e/o sono originate al di fuori del

genoma di riferimento (ad esempio, contaminazione del campione)]. L'algoritmo sfrutta diverse proprietà (frequenza dell'allele non equilibrato e MAPQ basso) e incorpora questa prova nel calcolo delle probabilità in modo rigorosamente matematico.

Vengono aggiunte ipotesi su nuovi genotipi candidati all'elenco di legacy dei genotipi diploidi (quelli che presumono errori di pile-up indipendenti). Ad esempio, nel caso di un locus con l'allele 1 ALT, oltre a prendere in considerazione P(G00|R), P(G01|R), P(G11|R), aggiungiamo altre due ipotesi P(G00,F1|R) e P(G11,F0|R), in cui F0 e F1 rappresentano l'allele di riferimento e l'allele ALT proveniente da un errore di mappatura. Le proprietà di questi errori, come la profondità dell'allele e MAPQ sono incorporati nel calcolo di P (G00,F1|R) e P(G11,F0|R). Quindi il genotipo rilevante viene preso su max (max(P(G00|R), P(G00,F1|R)), P(G01|R), max(P(G11|R), P(G11,F0|R))).

La sensibilità viene migliorata recuperando FN, correggendo i genotipi e abbassando la soglia MAPQ per le nuove letture trasferite al Variant Caller. La specificità viene migliora rimuovendo i FP e correggendo i genotipi.

FRD è lo strumento più efficace nel filtraggio dopo la creazione dei file VCF al fine di migliorare la misurazione F, perché, piuttosto che rilevare semplicemente i risultati sospetti (ad esempio, basati sulla profondità dell'allele o dagli errori di lettura) dopo l'utilizzo di Variant Caller, l'algoritmo di rilevamento incorpora direttamente la presenza di letture estranee mediante un rilevamento rigoroso della massima verosimiglianza.

PDHMM e rilevamento in base a colonna

I Variant Caller, come GATK Haplotype Caller e DRAGEN, utilizzano il grafico Debruijn per riassemblare le letture al fine di determinare gli aplotipi candidati e identificare i potenziali siti delle varianti. Nelle regioni del genoma con ripetizioni in tandem, varianti strutturali o cluster degli errori di sequenziamento, si ottiene una sensibilità inferiore a causa di una metodologia errata di assemblaggio del grafico per ottenere un elenco completo di aplotipi candidati e siti di varianti.

Il rilevamento di eventi in base a colonna si integra nel grafico Debruijn eseguendo la scansione di ciascuna colonna nella regione attiva per individuare siti di potenziali varianti (SNP e Indel) e per completare l'elenco degli aplotipi candidati. Questo consente di ripristinare la sensibilità nelle regioni in cui il grafico ha fallito.

Impatto di FRD/BQD su QUAL/GQ/QD e filtraggio efficace in seguito alla creazione dei file VCF

DRAGEN v3 Variant Caller ha implementato due algoritmi che modellano gli errori correlati sulle letture in un dato pile-up: il rilevamento di letture estranee (FRD) per identificare le letture non mappate correttamente e l'algoritmo di calo della qualità della base (BQD) per rilevare gli errori delle identificazioni delle basi correlate. Oltre a migliorare la specificità e la sensibilità, questi due algoritmi incidono anche su due livelli nel rapporto impatto/beneficio:

Valori del punteggio di affidabilità (QUAL, GQ, QD) sono in una reale gamma su scala Phred.

I Variant Caller convenzionali di solito generano valori QUAL su scala Phred gonfiati (nell'intervallo di poche migliaia) che non hanno significato pratico dal punto di vista statistico. Il modellare gli errori correlati nel Variant Caller consente di riportare questi valori in un intervallo realistico e significativo dal punto di vista statistico.

Viene ridotta sostanzialmente la dipendenza da regole di filtraggio in seguito alla creazione di file VCF.

Poiché i Variant Caller convenzionali non sono in grado di distinguere tra errori correlati e varianti vere, è stato necessario applicare regole di filtraggio efficaci in seguito alla creazione di file VCF per filtrare il numero eccessivo di identificazioni FP. Diverse annotazioni VCF (ad esempio, QD, MQ, FS, MQRankSum) sono state confrontate con soglie ad-hoc, per indicare le identificazioni FP. In alternativa, quelle annotazioni potevano essere fornite a un algoritmo di apprendimento automatico e mirate su un set vero. I falsi positivi potevano essere quindi filtrati in base a quanto appreso dall'algoritmo (ad esempio, VQSR).

In DRAGEN v3, sono stati migliorati gli algoritmi alla base del Variant Caller ed è quindi stata sostanzialmente ridotta la dipendenza dal filtraggio dopo la creazione dei file VCF. La regola di filtraggio efficace predefinita di DRAGEN v3 utilizza semplicemente QUAL con una soglia corrispondente al Fmeas migliore (il migliore compromesso tra sensibilità e specificità).

Metodi dettagliati

Set di dati di input

Sono stati selezionati tre set di dati per rappresentare diversi metodi di preparazione delle librerie, includendo o escludendo la PCR (TruSeq DNA Nano, TruSeq DNA PCR-Free e Nextera DNA Flex). Ciascun set di dati è stato generato utilizzando il DNA del campione NA12878. In seguito alla preparazione delle librerie di DNA in base alle rispettive guide rapide,[5-7] le librerie ottenute sono state sequenziate in corse paired-end da 2×150 su un sistema NovaSeq™ 6000. Per normalizzare il numero di letture, ciascun set di dati è stato ricampionato a una copertura di 30× con FASTQ Toolkit in BaseSpace Sequence Hub. Tutti e tre i set di dati sono accessibili pubblicamente in BaseSpace Sequence Hub, in questo modo è possibile eseguire una valutazione indipendente dei risultati.

Genoma di riferimento umano

È stato utilizzato il genoma di riferimento umano Human hs37d5 nell'applicazione DRAGEN BaseSpace e l'equivalente genoma di riferimento è stato utilizzato nell'analisi in laboratorio per ciascun software valutato. Questo riferimento include decoy.[8]

Software per l'analisi secondaria

Abbiamo confrontato tre software per l'analisi secondaria. Il primo software è DRAGEN v2 end-to-end (DRAGEN utilizzato sia per la mappatura e l'allineamento sia per l'identificazione delle varianti). Il secondo software è DRAGEN v3 end-to-end. Il terzo software utilizza BWA-MEM per la mappatura e l'allineamento e GATK4-HC per l'identificazione delle varianti.

Per eseguire un confronto corretto, abbiamo applicato la stessa rigida regola di filtraggio a tutti e tre i software e consisteva nell'applicazione di una soglia GQ ai file VCF prima del filtraggio. La soglia è stata selezionata per essere la più prossima al miglior punto Fmeas per ciascun software (Tabella 3).

Tabella 3: Soglie QC Fmeas ottimali

GQ per Fmeas migliore	SNP	Indel
DRAGEN v3.2.8	9	9
DRAGEN v2.5	2	8
BWA+GATK	1	2

DRAGEN è stato eseguito su un server in laboratorio ed anche sul cloud utilizzando BaseSpace Sequence Hub. Sebbene la durata del calcolo sia leggermente più lunga sul cloud, i risultati dell'identificazione delle varianti non erano diversi. Il software BWA+GATK è stato eseguito sullo stesso server di DRAGEN in laboratorio, dove era installato il framework BCBIO.[9] BCBIO esegue BWA+GATK attenendosi alle linee guida per le pratiche migliori di GATK e applica inoltre ottimizzazioni aggiuntive per migliorare i parallelismi e accelerare la durata della corsa. Per l'analisi sul cloud, il software BWA+GATK è stato eseguito su Terra.

DRAGEN 3.3.0

Versione applicazione DRAGEN:
DRAGEN Germline Pipeline 3.2.8
DRAGEN Host Software versione 05.011.281.3.2.8

BWA-Mem (0.7.17) + GATK4 (4.0.2)

Tabella 4: Parametri del file di configurazione degli algoritmi BCBIO

Parametro	Valore
align_split_size (allinea_separazione_dimensione)	5000000
aligner (allineatore)	BWA
coverage_depth (profondità_copertura)	High (Elevato)
coverage_interval (intervallo_copertura)	Regional (Regionale)
mark_duplicates (marca_duplicati)	True (Vero)
merge_bamprep (accorpa_prepbam)	False (Falso)
platform (piattaforma)	Illumina
quality_format (qualità_formato)	Standard
realign (riallinea)	False (Falso)
recalibrate (ricalibra)	False (Falso)
tools_off (strumenti_non utilizzati)	Vqsr
variantcaller	GATK-haplotype

analisi: fonti
variant2: gatk-haplotype

BWA+GATK su Terra

I file Bam pronti per l'analisi ottenuti da BWA-Mem (dalle corse BCBIO) sono stati utilizzati come input per eseguire GATK su Terra. In sintesi, abbiamo seguito il flusso di lavoro GATK4-germline-snps-indels (https://github.com/gatk-workflows/gatk4-germline-snps-indels) modificando determinati parametri in modo che corrispondano a quelli delle corse BCBIO. Tutte le corse sono state eseguite con un account di prova gratuito su Terra.

L'esatto metodo WDL è disponibile nei dati accessibili pubblicamente di BaseSpace Sequence Hub.

Configurazioni del metodo WDL:
immagine docker GATK: broadinstitute/gatk:4.0.2.0
docker GITC: broadinstitute/genomes-in-the-cloud:2.3.1-1500064817

Fasta di riferimento: hs37d5 (lo stesso degli altri software)

In questo software sono stati generati solo i dati VCF non elaborati. Il filtraggio successivo è stato eseguito localmente. I file VCF non elaborati sono disponibili nei dati accessibili pubblicamente di BaseSpace Sequence Hub.

Basespace (gennaio 2019) Indicare le specifiche dell'istanza AWS F1 utilizzata (AWS F1 4× ampio).
Versione dell'applicazione BaseSpace Sequence Hub: 3.2.8

Tabella 5: Server in laboratorio (CentOS 7 x86_64, Supermicro 1029)

Parte	Nome completo modello	Note
Telaio	SYS-1029GQ-TNRT	1 unità rack
CPU	2 x Intel(R) Xeon(R) Gold 6126 CPU @ 2,60 GHz	24 core, 48 thread
RAM	384 GB	DDR4, 2.666 MHz
Staging	Intel SSDPE2KE020T7	NVME da 2 TB

Set vero (NIST) per lo studio comparativo

Lo studio comparativo delle identificazioni delle varianti richiede un determinato genoma di riferimento e un set di identificazioni associate che rappresentano le "risposte vere" per quel genoma. Tali set di identificazioni hanno la proprietà di poter essere utilizzati come "vero" per identificare accuratamente i falsi positivi e negativi. Per questo studio, il set vero utilizzato era basato sulle identificazioni di riferimento basate sulla stessa fonte di DNA (NA12878) stabilita dal National Institute of Standards and Technology (NIST). Genome in a Bottle Consortium (GIAB) è un consorzio pubblico-privato-accademico patrocinato da NIST. GIAB ha pubblicato un set di criteri comparativi di varianti piccole e identificazioni di riferimento per il genoma pilota, NA12878, caratterizzando un genotipo di elevata affidabilità per circa il 90% di GRCh37 e GRCh38.

I veri positivi (TP) sono identificazioni delle varianti che corrispondono alle identificazioni di riferimento del set vero NIST. I falsi positivi (FP) sono identificazioni delle varianti che non esistono nel set vero e i falsi negativi (FN) sono varianti presenti nel set vero ma non sono state identificate nel file QUERY VCF.

Tabella 6: Definizioni e calcoli per le metriche coinvolte nella precisione e nella reidentificazione

Metrica	Nome comune	Definizione	Formula
TRUTH.TP	Veri positivi (vero)	Numero di identificazioni vere per le quali esiste una identificazione query coerente con l'identificazione vera e il relativo genotipo	
QUERY.TP	Veri positivi (query)	Numero di identificazioni query per le quali esiste una identificazione vera coerente con l'identificazione query e il relativo genotipo	
TRUTH.FN	Falsi negativi	Numero di identificazioni vere per le quali non esiste una identificazione query coerente con l'identificazione vera e il relativo genotipo	
QUERY.FP	Falsi positivi	Numero di identificazioni query per le quali non esiste una identificazione vera coerente con l'identificazione query e il relativo genotipo	
METRIC.Recall	Reidentificazione, sensibilità	Frazione di identificazioni vere coerenti con un allele query e l'identificazione del genotipo nelle regioni affidabili	TRUTH.TP / (TRUTH.TP + TRUTH.FN)
METRIC.Precision	Precisione, valore predittivo positivo	Frazione di identificazioni query coerenti con un allele vero e l'identificazione del genotipo nelle regioni affidabili	QUERY.TP / (QUERY.TP + QUERY.FP)

Variant Calling Assessment Tool (VCAT) è stato utilizzato per confrontare ciascun file QUERY VCF rispetto al set vero NIST v3.3.2. Questo strumento esegue hap.py utilizzando il motore di valutazione vcfeval RTG. TP, FP e FN sono stati determinati dai file di output di hap.py *roc.Locations.INDEL.csv e *roc.Locations.SNP.csv di TRUTH TP, QUERY.TP, QUERY.FP e TRUTH.FN.

Il tipo di rigorosa corrispondenza utilizzato per il calcolo di TP, FP e FN è "corrispondenza genotipo" (cf. [1]), per il quale solo i siti con alleli e genotipi corrispondenti vengono conteggiati come TP. Questo significa che gli errori del genotipo e le mancate corrispondenze dell'allele vengono contate sia come FP che come FN.

Metriche di valutazione dello studio comparativo

Per il confronto della velocità, le durate totali in secondi, da FASTQ a VCF, vengono ottenute dai file di registro dell'analisi e/o dalle durate delle analisi mostrate nei report.

Per eseguire i confronti per l'accuratezza sui diversi software, nelle metriche delle prestazioni utilizziamo gli standard raccomandati (Tabella 6).[1] La precisione è la metrica che rappresenta la specificità analitica o la capacità di identificare correttamente l'assenza di varianti o "assenza di falsi positivi". La reidentificazione è la metrica che rappresenta la sensibilità analitica o la capacità di rilevare le varianti note per essere presenti o "assenza di falsi negativi".

Le definizioni e i calcoli delle metriche coinvolte nella precisione e il numero di reidentificazioni sono basati sul riferimento.

Bibliografia

1. Krusche P, Trigg L, Boutros PC, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019;37(5):555-560.
2. Pratiche migliori GATK. software.broadinstitute.org/gatk/best-practices/. Consultato il 9 maggio 2019.
3. Il progetto BaseSpace. basespace.illumina.com/s/3ExEZMlH8Lkq. Consultato il 15 maggio 2019.
4. FireCloud su Terra. firecloud.terra.bio/. Consultato il 15 maggio 2019.
5. Illumina (2017) TruSeq DNA PCR-Free Reference Guide (Guida rapida di TruSeq DNA PCR-Free). Consultato il 6 marzo 2019.
6. Illumina (2017) TruSeq DNA Nano Reference Guide (Guida rapida di TruSeq DNA Nano). Consultato il 6 marzo 2019.
7. Illumina (2018) Nextera DNA Flex Library Prep Reference Guide (Guida rapida alla preparazione delle librerie Nextera DNA Flex). Consultato il 6 marzo 2019.
8. Genoma di riferimento hs37d5. ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/. Consultato il 9 maggio 2019.
9. Bcbio-nextgen. Docs. bcbio-nextgen.readthedocs.io/en/latest/. Consultato il 9 maggio 2019.

Illumina, Inc. • Numero verde 1.800.809.4566 (U.S.A.) • Tel. +1.858.202.4566 • techsupport@illumina.com • www.illumina.com