NOTA TECNICA

Miglioramento dell'accuratezza dell'identificazione di varianti della linea germinale nell'analisi secondaria DRAGEN™

Ottimizzazione delle prestazioni di identificazione di varianti con l'apprendimento automatico Illumina e la mappatura multigenomica

illumina®

Introduzione

Nei campi della ricerca biomedica e della medicina di precisione è fondamentale sfruttare al massimo le potenzialità del genoma attraverso il sequenziamento di nuova generazione (NGS, Next-Generation Sequencing). Per utilizzare al meglio le informazioni raccolte durante l'NGS, i ricercatori necessitano di strumenti di analisi dei dati in grado di tradurre i dati di sequenziamento non elaborati in risultati significativi. L'analisi secondaria DRAGEN fornisce un'analisi secondaria accurata, completa ed efficiente dei dati NGS. L'utilizzo della tecnologia completamente riconfigurabile delle matrici di porte logiche programmabili sul campo (FPGA, Field-Programmable Gate Array) consente all'analisi secondaria DRAGEN di accelerare l'analisi secondaria dei dati NGS, inclusi mappatura, allineamento e identificazione di varianti. Inoltre, l'analisi secondaria DRAGEN è progettata per affrontare le sfide comuni legate all'analisi genomica, come lunghi tempi di calcolo, volumi consistenti di dati e identificazione di varianti in regioni genomiche difficili.

L'analisi secondaria DRAGEN genera risultati straordinariamente accurati. La Precision FDA Truth Challenge V2 (PrecisionFDA V2) del 2020 ha premiato l'analisi secondaria DRAGEN v3.7 come la più accurata in tutte le regioni di riferimento e nelle regioni difficili da mappare rispetto ad altre soluzioni come Sentieon, Seven Bridges e BWA-GATK (Figura 1).[1,2] In soli quattro anni, l'analisi secondaria DRAGEN v4.3 ha apportato miglioramenti significativi a queste prestazioni già eccezionali, garantendo un'accuratezza senza precedenti nell'identificazione di varianti piccole con un punteggio F1 pari al 99,89%, misura combinata di precisione e richiamo, in tutte le regioni di riferimento con funzionalità nuove e di grande impatto.

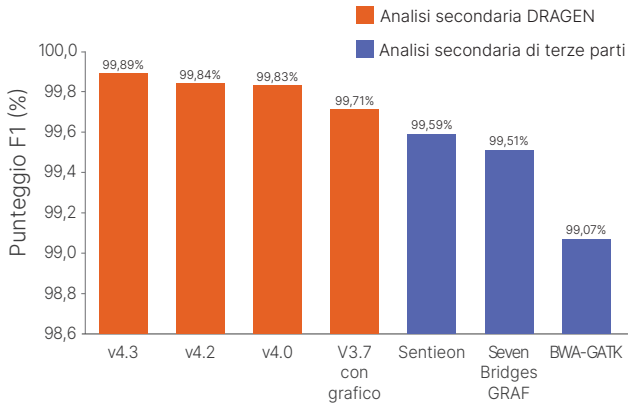

Figura 1: accuratezza dell'analisi secondaria DRAGEN per l'analisi di tutte le regioni di riferimento della FDA. Il punteggio F1 (%) corrisponde al calcolo dei risultati veri positivi e veri negativi come percentuale dei risultati totali.[5,6] Punteggi più alti indicano una migliore accuratezza in base ai dati di riferimento.

Questa nota tecnica descrive i recenti miglioramenti che contribuiscono all'elevata accuratezza dell'analisi secondaria DRAGEN, tra cui il mappatore multigenomico con riferimento del pangenoma, l'integrazione dell'apprendimento automatico (ML, Machine Learning), l'identificazione di varianti a mosaico, gli identificatori specializzati e il rilevamento di varianti strutturali (SV, Structural Variant) e di varianti del numero di copie (CNV, Copy Number Variant).

Mappatore multigenomico con riferimento del pangenoma

La mappatura multigenomica, introdotta per la prima volta nell'analisi secondaria DRAGEN v3.7, consente una migliore accuratezza nell'identificazione di varianti.[3] L'analisi secondaria DRAGEN v4.3 garantisce miglioramenti significativi in termini di accuratezza, con una riduzione dell'83% degli errori rispetto alla v3.6.3 e una riduzione del 40% degli errori rispetto alla v4.2.7 (Figura 2).

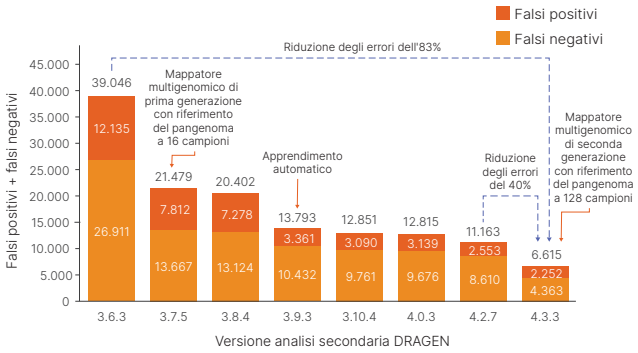

Figura 2: innovazione costante alla base dell'analisi secondaria DRAGEN. Le percentuali migliorate dei falsi positivi e negativi per polimorfismi a singolo nucleotide (SNP, single nucleotide poly-morphism) e indel utilizzando un campione Genome in a Bottle HG002, NIST v4.2.1[4] dimostrano la significativa riduzione degli errori ottenuta in soli quattro anni.

Per rappresentare meglio una popolazione specifica, l'analisi secondaria DRAGEN v4.3 offre agli utenti la possibilità di creare un riferimento del pangenoma personalizzato, migliorando ulteriormente l'identificazione di varianti all'interno dei loro studi. Gli utenti possono creare un riferimento del pangenoma personalizzato utilizzando i propri gruppi o una selezione di gruppi forniti dallo Human Pangenome Reference Consortium (HPRC). Ad esempio, un riferimento del pangenoma personalizzato realizzato con 44 gruppi HPRC che rappresentano una popolazione di ricerca specifica produce una maggiore accuratezza nell'identificazione di varianti rispetto alle precedenti versioni dell'analisi secondaria DRAGEN, come l'analisi secondaria DRAGEN v4.2 (Figura 3). Tuttavia, il riferimento del pangenoma predefinito incluso (basato su 128 campioni) nella v4.3 deve fornire le prestazioni migliori per i casi d'uso generale.[4]

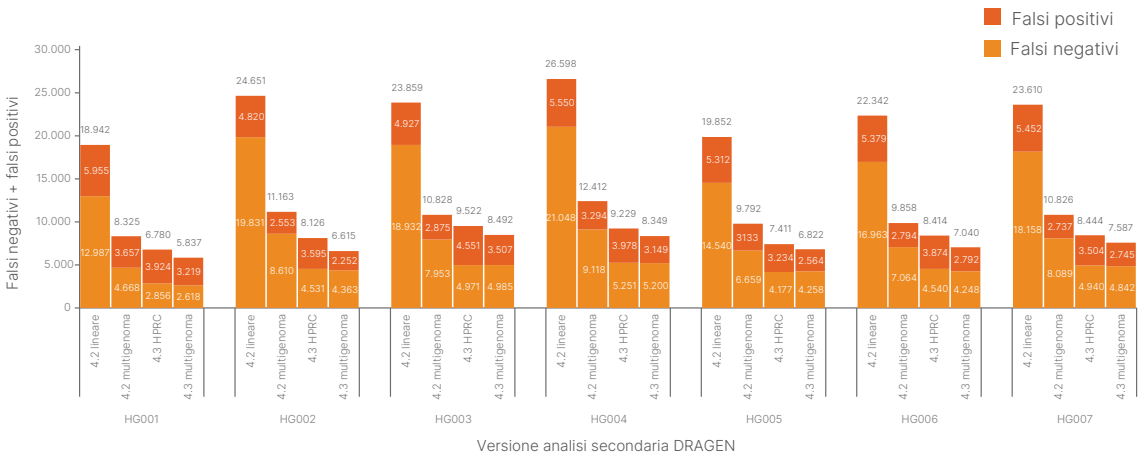

Figura 3: miglioramento dell'accuratezza dell'identificazione di varianti piccole dell'analisi secondaria DRAGEN con riferimenti personalizzati. Il riferimento del multigenoma basato sull'HPRC dell'analisi secondaria DRAGEN v4.3 fornisce risultati più accurati rispetto alla v4.2 nell'analisi di campioni Genome in a Bottle HG001-HG007.[4] Il riferimento del multigenoma predefinito (multigenoma 4.3), basato su 128 campioni, supera il riferimento basato su HPRC 4.3 nell'uso generale.

Apprendimento automatico

Il modulo ML, aggiunto per la prima volta nell'analisi secondaria DRAGEN v3.9 e migliorato nella v3.10, utilizza un modello supervisionato con caratteristiche contestuali e basate sulla lettura estratte dagli identificatori di varianti dell'analisi secondaria DRAGEN. L'accuratezza delle varianti piccole viene migliorata riducendo le false identificazioni grazie alla combinazione di mappatura multigenomica e ML per fornire i risultati migliori (Figura 4). I miglioramenti significativi sono stati dimostrati in modo coerente in tutti i soggetti, inclusi i dati dei test relativi ad altre popolazioni non utilizzate durante la formazione.

Rilevamento delle varianti a mosaico

L'analisi secondaria DRAGEN v4.3, supportata da un nuovo modello ML, ora identifica le varianti a mosaico all'interno dell'identificatore di varianti piccole della linea germinale. Grazie all'azzeramento della soglia di frequenza allelica, l'analisi secondaria DRAGEN è in grado di rilevare varianti con frequenze alleliche inferiori al 20%.

L'analisi secondaria DRAGEN v4.3 rileva le varianti a mosaico con maggiore accuratezza e precisione rispetto alle versioni precedenti.

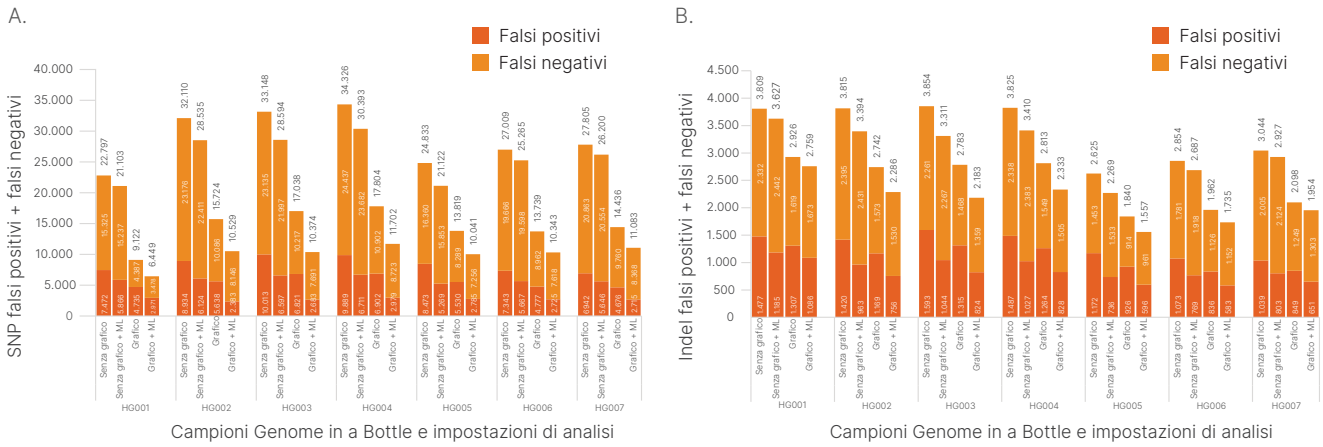

Figura 4: il ML e la mappatura multigenomica riducono i falsi positivi e i falsi negativi. In un'analisi di campioni Genome in a Bottle HG001-HG007,[4] ML fornisce una riduzione degli errori del 10% con il riferimento (grafico) multigenomico disabilitato e circa il 30% di riduzione degli errori con il riferimento (grafico) multigenomico abilitato. Quando sono abilitati sia il riferimento multigenomico sia ML, le false identificazioni vengono ridotte del 62% per (A) SNV e (B) indel.

Per dimostrarlo, le seguenti quattro pipeline di analisi secondaria DRAGEN sono state testate sui dati della serie Mosaic truth del National Institute of Standards and Technology (NIST): analisi secondaria DRAGEN v4.2, analisi secondaria DRAGEN v4.2 in modalità ad alta sensibilità (HSM, high-sensitivity mode), analisi secondaria DRAGEN v4.3 e analisi secondaria DRAGEN v4.3 con modalità Mosaic abilitata. La serie Mosaic truth del NIST contiene 73 varianti Mosaic note in dati 300×, non rilevate dall'analisi secondaria DRAGEN v4.2 e v4.3, ma rilevate dall'analisi secondaria DRAGEN v4.2 in HSM e dall'analisi secondaria DRAGEN v4.3 in modalità Mosaic. Tuttavia, l'analisi secondaria DRAGEN v4.3 in modalità Mosaic ha dimostrato una maggiore accuratezza nell'identificazione di varianti Mosaic, con il 73% di falsi positivi in meno rispetto all'analisi secondaria DRAGEN v4.2 in HSM (Figura 5).

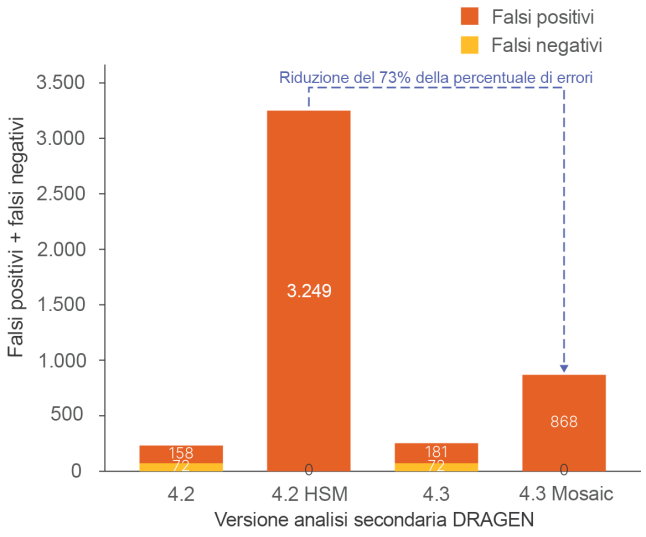

Figura 5: migliore accuratezza e precisione con la modalità di rilevamento Mosaic. Riduzione del 73% degli errori con l'analisi secondaria DRAGEN v4.2 in modalità ad alta sensibilità (HSM) rispetto a DRAGEN v4.3 in modalità di rilevamento Mosaic. I dati mostrano anche il numero elevato di falsi negativi senza la modalità HSM o con rilevamento Mosaic abilitato.

Rilevamento di SV e CNV

Le varianti strutturali (SV) sono delle alterazioni genomiche lunghe almeno 50 bp, mentre le varianti del numero di copie (CNV) sono un tipo specifico di SV in cui il numero di copie di una sequenza genomica è ridotto (delezioni) o aumentato (inserzioni). L'analisi secondaria DRAGEN mostra una maggiore accuratezza per l'identificazione di SV (Figura 6) e di CNV (Figura 7) rispetto alle soluzioni alternative.[7] Gli algoritmi avanzati e i nuovi approcci realizzati su misura per regioni genomiche complesse danno una marcia in più all'analisi secondaria DRAGEN rispetto alle altre soluzioni.

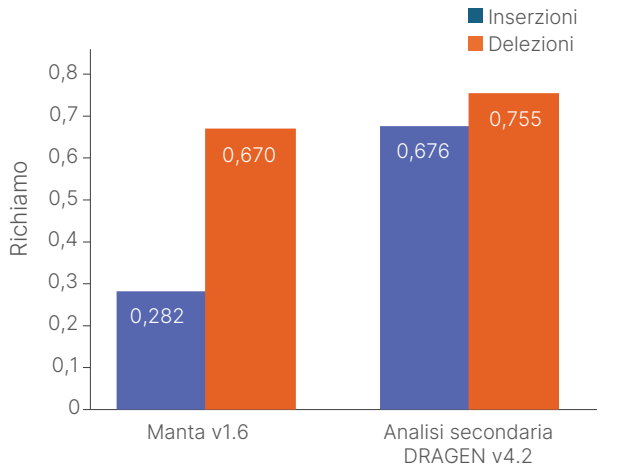

Figura 6: identificazione di SV altamente accurata con l'analisi secondaria DRAGEN. Confronto del richiamo indel SV dell'analisi secondaria DRAGEN v4.2 e Manta v1.6 valutato con i dati di riferimento Genome in a Bottle (GIAB SV v0.6).[7]

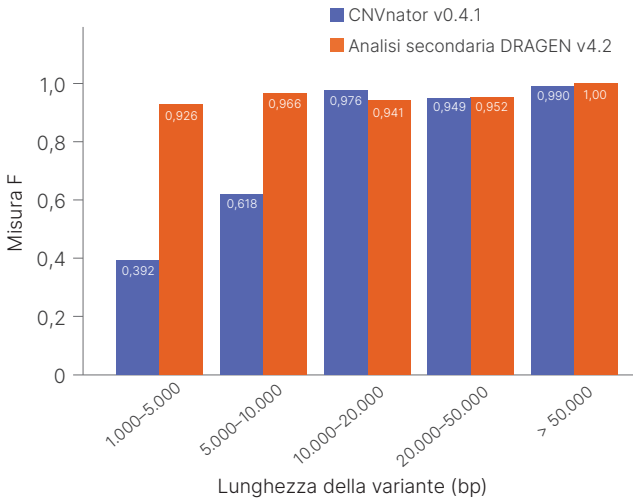

Figura 7: identificazione di CNV altamente accurata con l'analisi secondaria DRAGEN. Identificazione di CNV mediante l'analisi secondaria DRAGEN v4.2 rispetto a CNVnator v1.6 su diverse dimensioni di delezioni in base ai dati di riferimento Genome in a Bottle (GIAB SV v0.6).[7]

L'identificatore di SV DRAGEN migliora i metodi di identificazione di varianti strutturali Manta e incorpora le informazioni ottenute dal riferimento del pangenoma, consentendo un filtraggio più preciso e una migliore accuratezza nel rilevamento di SV. Ciò include un nuovo rilevatore di inserzione di elementi mobili per identificare inserzioni di grandi dimensioni, parametri di coppia ottimizzati per migliori identificazioni di ampie delezioni e un allineamento contig perfezionato per una migliore individuazione dell'inserzione. Inoltre, il software DRAGEN introduce miglioramenti nelle fasi di assemblaggio, nei calcoli della probabilità di lettura e nella gestione migliorata degli accoppiamenti sovrapposti e delle basi sottoposte a clipping.

L'identificatore di CNV DRAGEN è principalmente un identificatore basato sulla profondità di lettura, con il supporto di diversi modelli di segmentazione e punteggio per adattarsi a molteplici applicazioni. Utilizzando un segnale aggiuntivo proveniente da letture discordanti e suddivise, come avviene nell'identificazione di SV, l'identificatore di CNV migliora la sensibilità per acquisire eventi di appena 1 kbp.

L'identificatore di CNV DRAGEN dispone anche di un modulo di estensione della duplicazione segmentale, una funzione che consente il rilevamento di CNV nelle regioni di duplicazione segmentale del genoma. Le regioni di duplicazione segmentale sono regioni del genoma con una similarità della sequenza superiore al 90%, che rappresenta il 5% del genoma. Queste hanno una scarsa mappabilità, rendendo difficile il rilevamento delle varianti in queste regioni. L'estensione della duplicazione segmentale recupera circa un milione di basi di regioni CNV precedentemente escluse dall'analisi. Ciò consente il rilevamento di CNV su 43 geni clinicamente rilevanti e migliora l'accuratezza complessiva dell'identificazione di varianti.

Identificatori specializzati e mirati

Gli identificatori mirati supportano la genotipizzazione accurata di geni specifici difficili da analizzare a causa di fattori come l'elevata similarità delle sequenze agli pseudogeni, le regioni ripetitive e gli elevati gradi di polimorfismo. L'analisi secondaria DRAGEN affronta queste complessità incorporando diversi identificatori mirati (Tabella 1), consentendo una genotipizzazione precisa dei geni clinicamente rilevanti. Per informazioni farmacogenomiche (PGx, pharmacogenomics), PGx Star Allele Caller identifica gli alleli star e lo stato del metabolizzatore per 22 geni (Tabella 2).

L'identificatore degli antigeni leucocitari umani (HLA, Human Leukocyte Antigen) DRAGEN consente una genotipizzazione estremamente accurata degli alleli HLA di classe I e II. Allinea le letture a un database completo di oltre 9.000 alleli e può fornire un supporto per applicazioni come la corrispondenza nei trapianti d'organo, l'immunogenetica e gli studi di associazione con patologie.

Tabella 1: riepilogo dei geni interessati dagli identificatori mirati e specializzati

Identificatore mirato	Area di applicazione di ricerca	Associazione con patologie
CYP21A2	Screening dei portatori	Iperplasia surrenale congenita (ISC)
HBA	Screening dei portatori	α-talassemia
GBA	Screening dei portatori	Malattia di Gaucher, malattia di Parkinson
SMN	Screening dei portatori	Atrofia muscolare spinale
LPA	Patologia cardiovascolare	Coronaropatia
RH	Tipizzazione del sangue	–
CYP2B6	PGx	–
CYP2D6	PGx	–
HLA	Corrispondenza nei trapianti, immunogenetica	–

Tabella 2: geni con rilevanza PGx interessati dal PGx Star Allele Caller

Simbolo del gene		
ABCG2	*CYP4F2*	*RYR1*
BCHE	*DPYD*	*SLCO1B1*
CACNA1S	*F5*	*TPMT*
CFTR	*G6PD*	*UGT1A1*
CYP2C19	*IFNL3*	*UGTB17*
CYP2C9	*MT-RNR1*	*VKORC1*
CYP3A4	*NAT2*	
CYP3A5	*NUDT15*	

L'analisi secondaria DRAGEN v4.3 introduce una nuova classe di identificatori che consente il rilevamento di varianti *de novo* in regioni con duplicazioni segmentali. L'identificatore di rilevamento congiunto multiregione (MRJD, Multiregion Joint Detection) implementa un identificatore di varianti piccole *de novo* basato su aplotipo per sei geni clinicamente rilevanti nelle regioni di duplicazione segmentale (Tabella 3).

Tabella 3: geni interessati dall'identificatore MJRD

Identificatore mirato	Area di applicazione di ricerca	Associazione con patologie
PMS2	Screening del tumore ereditario	Sindrome di Lynch per il cancro del colon-retto/dell'endometrio
SMN1 (varianti piccole)	Screening dei portatori	Atrofia muscolare spinale
STRC	Screening dei portatori	Perdita dell'udito non sindromica
NEB	Screening dei portatori	Miopatia nemalinica
TTN	Screening neonatale, malattie rare	Cardiomiopatia
IKBKG	Screening neonatale	Incontinentia pigmenti, displasia ectodermica ipoidrotica

Riepilogo

L'analisi secondaria DRAGEN fornisce un'analisi secondaria notevolmente accurata, completa ed efficiente per le applicazioni NGS. I miglioramenti continui forniscono una maggiore accuratezza e l'ampliamento della copertura anche delle regioni difficili del genoma, consentendo il rilevamento di varianti difficili e clinicamente rilevanti.

Appendice

Mappatura multigenomica con riferimento del pangenoma

Utilizzando gli aplotipi di popolazioni delle varianti sottoposte a determinazioni delle fasi e aumentando l'indice di riferimento con contig alt derivati dalle popolazioni, l'analisi secondaria DRAGEN è in grado di eseguire una mappatura efficace rispetto a un riferimento del pangenoma e di migliorare la mappatura delle letture Illumina in regioni difficili. Questa nuova funzione amplia efficacemente la portata delle letture di Illumina e consente una mappatura accurata e l'identificazione di varianti in regioni a cui in precedenza non era possibile accedere.

L'approccio tramite mappatore multigenomico facilita la mappatura con i dati della popolazione in cui il contenuto di sequenze alternative, osservato nella popolazione, viene rappresentato come vari percorsi divergenti e convergenti (Figura 8). Le letture dei campioni possono essere allineate a qualsiasi percorso con corrispondenza ottimale attraverso il mappatore multigenomico.

Ulteriori informazioni: The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3.

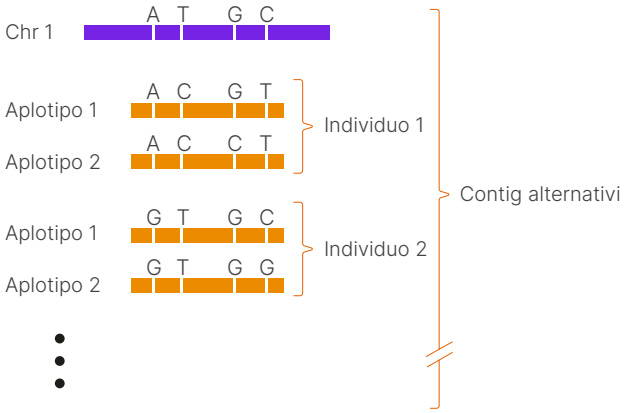

Figura 8: mappatore multigenomico con riferimento del pangenoma. In un riferimento, il contenuto di sequenze alternative registrato in una popolazione è rappresentato come vari percorsi divergenti e convergenti.

Alt-masking

Dall'aggiornamento del software per l'analisi secondaria DRAGEN v3.9, il software DRAGEN include l'Alt-masking, un nuovo approccio per gestire i contig ALT nativi di riferimento, in cui le posizioni strategiche dei contig ALT sono mascherate al fine di aumentare l'accuratezza. Questo approccio è semplice da definire, gestire e perfezionare nel tempo.

Ulteriori informazioni: DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph

Apprendimento automatico

Il software per l'analisi secondaria DRAGEN v3.9 ha aggiunto una pipeline di ricalibrazione ML performante ed efficiente come opzione all'interno del flusso di lavoro delle varianti piccole della linea germinale, abilitata per impostazione predefinita nel software per l'analisi secondaria DRAGEN v4.0. La pipeline esegue il modello ML dopo l'identificazione di varianti standard quando abilitata. Questo passaggio esegue la ricalibrazione dei campi QUAL e GQ inviati per il formato di identificazione delle varianti (VCF, variant call format) finale. In alcuni casi, il modello ML può modificare il genotipo (GT, genotype). I valori di apprendimento pre-automatico di questi campi sono conservati nei campi DQUAL, DGT e DGQ al fine di evitare la perdita di qualsiasi informazione. Questa fase aggiunge circa cinque minuti al flusso di lavoro standard per una corsa della linea germinale di sequenziamento dell'intero genoma (WGS, whole genome sequencing) 30×, limitando l'impatto del miglioramento dell'accuratezza sulle tempistiche totali della corsa.

Il modello ML viene generato utilizzando la formazione offline supervisionata. Il modello elabora una serie di funzioni basate sulla lettura e contestuali per perfezionare l'accuratezza dei punteggi qualitativi dell'identificatore delle varianti piccole. Le funzioni utilizzate per formare il modello includono mappabilità, AF, VC-Qual, DP, contenuto GC, mancate corrispondenze e altre metriche interne di mappatura, allineamento e VC.

Calcolo del punteggio F1

$$F1 = 2 \times (Richiamo \times Precisione)/(Richiamo + Precisione)$$

$$F1_{Genitori} = \sqrt{F1_{HG003} \times F1_{HG004}}$$

Riga di comando DRAGEN

Istruzioni iniziali disponibili su DRAGEN recipe-germline WGS

Bibliografia

1. Food and Drug Administration. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. precision.fda.gov/challenges/10/results. Consultato il 19 settembre 2024.
2. Illumina. DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph. illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. Pubblicato il 12 gennaio 2022. Consultato il 19 settembre 2024.
3. Illumina. The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3. illumina.com/science/genomics-research/articles/second-gen-multigenome-mapping.html. Pubblicato il 12 agosto 2024. Consultato il 30 settembre 2024.
4. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025. Pubblicato il 7 giugno 2016. doi:10.1038/sdata.2016.25.
5. Illumina. DRAGEN wins at PrecisionFDA Truth Challenge V2 showcase accuracy gains from alt-aware mapping and graph reference genomes. illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html. Consultato il 19 settembre 2024.
6. Dati interni in archivio. Illumina, Inc., 2022.
7. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. Pubblicato online il 25 ottobre 2024. doi:10.1038/s41587-024-02382-1.

illumina®

Numero verde 1.800.809.4566 (U.S.A.) | Tel. +1.858.202.4566
techsupport@illumina.com | www.illumina.com

M-GL-01016 ITA v3.0