NOTA TÉCNICA

Melhorias na precisão da identificação de variantes de linha genética no DRAGEN™ secondary analysis

Otimização do desempenho de identificação de variantes com aprendizado de máquina da Illumina e mapeamento multigenômico

illumina®

Introdução

Liberar o potencial do genoma por meio do sequenciamento de última geração (NGS) é fundamental para a pesquisa biomédica e para a medicina de precisão. Para maximizar as percepções do NGS, os pesquisadores precisam de ferramentas de análise de dados que possam traduzir dados brutos de sequenciamento em resultados significativos. O DRAGEN secondary analysis fornece uma análise secundária precisa, abrangente e eficiente dos dados do NGS. O uso da tecnologia de array de portas programáveis em campo (FPGA) altamente reconfigurável permite que o DRAGEN secondary analysis acelere a análise secundária de dados NGS, incluindo mapeamento, alinhamento e identificação de variantes. Além disso, o DRAGEN secondary analysis foi projetado para abordar desafios comuns na análise genômica, como tempos de computação longos, volumes enormes de dados e identificação de variantes em regiões genômicas desafiadoras.

O DRAGEN secondary analysis gera resultados excepcionalmente precisos. Na competição Precision FDA Truth Challenge V2 (PrecisionFDA V2) de 2020, o DRAGEN secondary analysis v3.7 foi mais preciso em todas as regiões de referência e regiões difíceis de mapear em comparação com outras soluções, como Sentieon, Seven Bridges, e BWA-GATK (Figura 1).[1,2] Em apenas quatro anos, o DRAGEN secondary analysis v4.3 fez melhorias significativas nesse desempenho já excepcional, fornecendo precisão na identificação de variantes pequenas sem precedentes com uma pontuação F1 de 99,89%, uma medida combinada de precisão e recall, em todas as regiões de referência com recursos novos e impactantes.

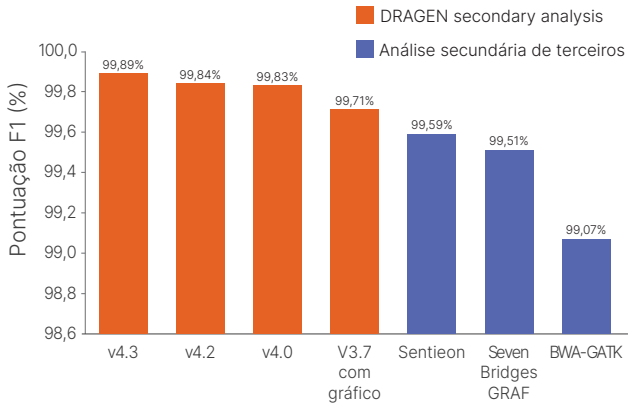

Figura 1: Precisão do DRAGEN secondary analysis para a análise de todas as regiões de referência da FDA: a pontuação F1 (%) é um cálculo de resultados positivos e negativos verdadeiros como proporção dos resultados totais.[5,6] Pontuações mais altas indicam melhor precisão com base nos dados de referência.

Esta nota técnica descreve melhorias recentes que contribuem para a alta precisão do DRAGEN secondary analysis, incluindo mapeador de múltiplos genomas com referência pangenômica, incorporação de aprendizado de máquina (ML), identificação de variantes em mosaico, identificadores especializados e detecção de variante estrutural (SV) e variante de número de cópias (CNV).

Mapeador multigenômico com referência pangenômica

O mapeamento multigenômico, introduzido pela primeira vez no DRAGEN secondary analysis v3.7, permite maior precisão na identificação de variantes.[3] O DRAGEN secondary analysis v4.3 traz ganhos significativos de precisão, com uma redução de 83% nos erros em comparação com a v3.6.3, e redução de 40% nos erros em comparação com a v4.2.7 (Figura 2).

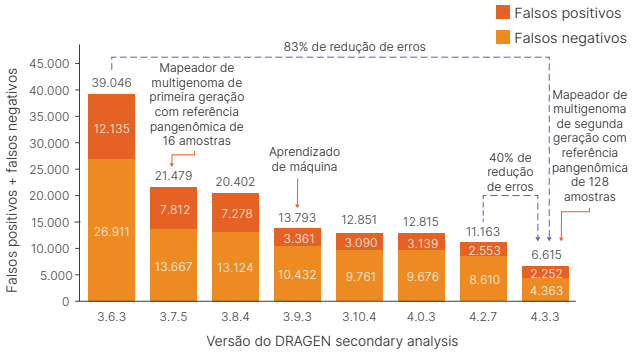

Figura 2: Inovação constante impulsionando o DRAGEN secondary analysis: melhorias nas taxas de falsos positivos e negativos para SNPs e indels usando uma amostra do Genome in a Bottle HG002, NIST v4.2.1[4] demonstram a redução significativa de erros que foi alcançada em apenas quatro anos.

Para representar melhor uma população específica, o DRAGEN secondary analysis v4.3 oferece aos usuários a opção de criar uma referência pangenômica personalizada, melhorando ainda mais a identificação de variantes em seus estudos. Os usuários podem criar uma referência de pangenoma personalizada usando seus próprios conjuntos ou usando uma seleção de conjuntos fornecidos pelo Human Pangenome Reference Consortium (HPRC). Por exemplo, uma referência pangenômica personalizada criada com 44 conjuntos de HPRC representando uma população de pesquisa específica produz maior precisão de identificação de variantes em comparação com versões anteriores do DRAGEN secondary analysis, como o DRAGEN secondary analysis v4.2 (Figura 3). No entanto, a referência de pangenoma padrão incluída (com base em 128 amostras) na v4.3 deve ter melhor desempenho para casos de uso geral.[4]

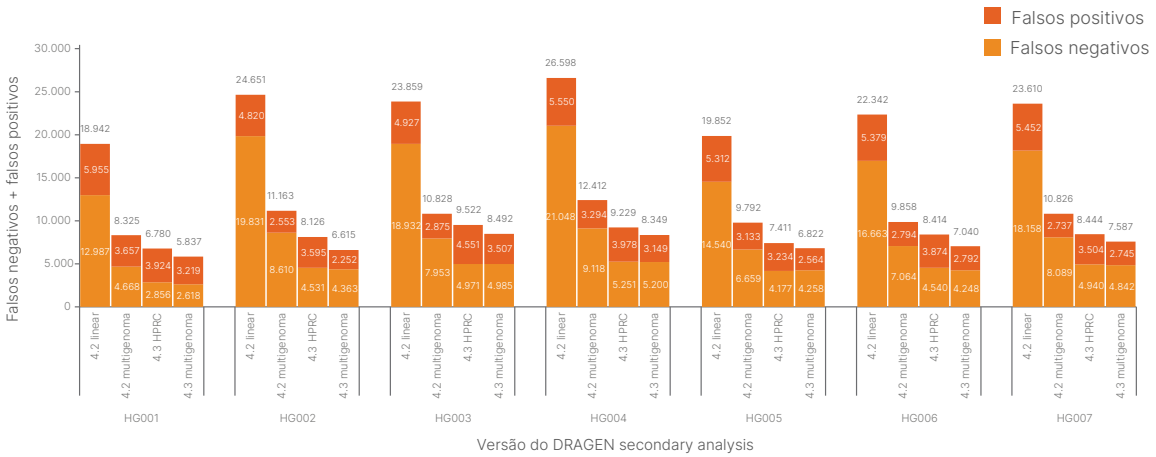

Figura 3: Melhorias de precisão na identificação de variantes pequenas com o DRAGEN secondary analysis com referências personalizadas: a referência multigênica baseada em HPRC do DRAGEN secondary analysis v4.3 produz resultados de precisão melhores do que a v4.2 ao analisar as amostras do Genome in a Bottle HG001–HG007.[4] A referência multigênica padrão (multigenoma 4.3), com base em 128 amostras, supera a referência baseada em HPRC 4.3 em uso geral.

Aprendizado de máquina

O módulo ML, adicionado pela primeira vez no DRAGEN secondary analysis v3.9 e aprimorado na v3.10, emprega um modelo supervisionado que usa recursos contextuais e baseados em leitura extraídos dos identificadores de variantes do DRAGEN secondary analysis. A precisão de variantes pequenas é melhorada pela redução de falsas identificações com a combinação de mapeamento multigenômico e ML para fornecer os melhores resultados (Figura 4). Ganhos substanciais foram consistentemente demonstrados em todos os participantes, incluindo dados de teste de outras populações que não foram usadas durante o treinamento.

Detecção de variantes de mosaico

O DRAGEN secondary analysis v4.3, suportado por um novo modelo de ML, agora identifica variantes de mosaico dentro do identificador de variantes pequenas da linha genética. Com o limite de frequência de alelos diminuído para zero, o DRAGEN secondary analysis pode detectar variantes com frequências de alelos < 20%.

O DRAGEN secondary analysis v4.3 detecta variantes de mosaico com maior precisão do que as versões anteriores.

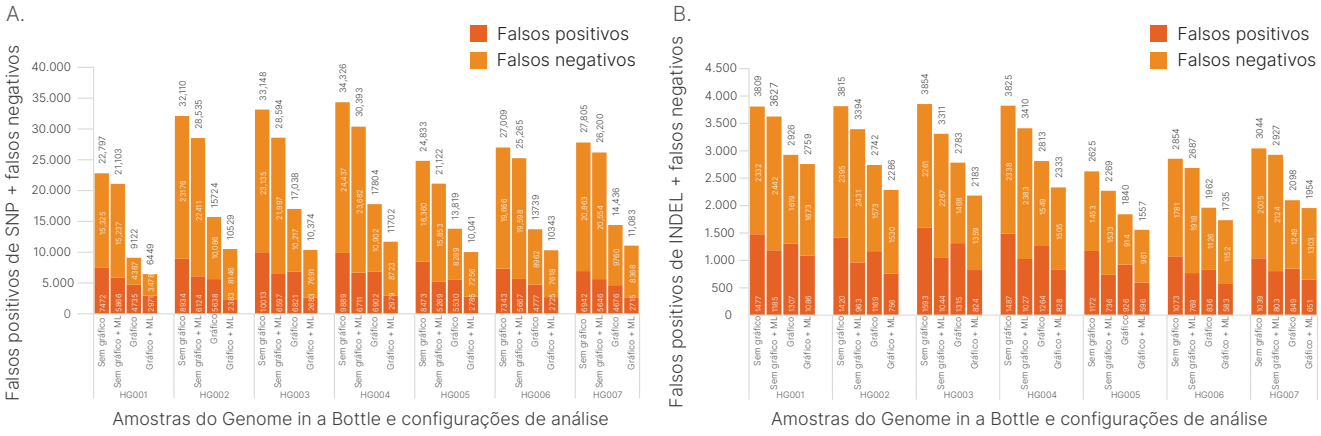

Figura 4: O uso de ML e mapeamento multigenômico reduz falsos positivos e falsos negativos. Em uma análise de amostras do Genome in a Bottle HG001–HG007,[4] o uso de ML reduz em 10% os erros com a referência multigenômica (gráfico) desativada, e reduz em ~30% os erros com a referência multigenômica (gráfico) ativada. Quando a referência multigenômica e o ML estão habilitados, as identificações falsas são reduzidas em 62% para (A) SNVs e (B) indels.

Para demonstrar isso, quatro pipelines do DRAGEN secondary analysis foram testados nos dados do conjunto de verdades do Mosaic do National Institute of Standards and Technology (NIST): DRAGEN secondary analysis v4.2, DRAGEN secondary analysis v4.2 no modo de alta sensibilidade (HSM), DRAGEN secondary analysis v4.3 e DRAGEN secondary analysis v4.3 com o modo Mosaic ativado. O conjunto de verdades do NIST Mosaic contém 73 variantes de mosaico conhecidas em dados de 300×, que não foram detectadas pelo DRAGEN secondary analysis v4.2 e v4.3, mas que foram detectadas pelo DRAGEN secondary analysis v4.2 no HSM e pelo DRAGEN secondary analysis v4.3 no modo Mosaic. No entanto, o DRAGEN secondary analysis v4.3 no modo Mosaic alcançou maior precisão na identificação de variantes do mosaico, com 73% menos falsos positivos do que o DRAGEN secondary analysis v4.2 no HSM (Figura 5).

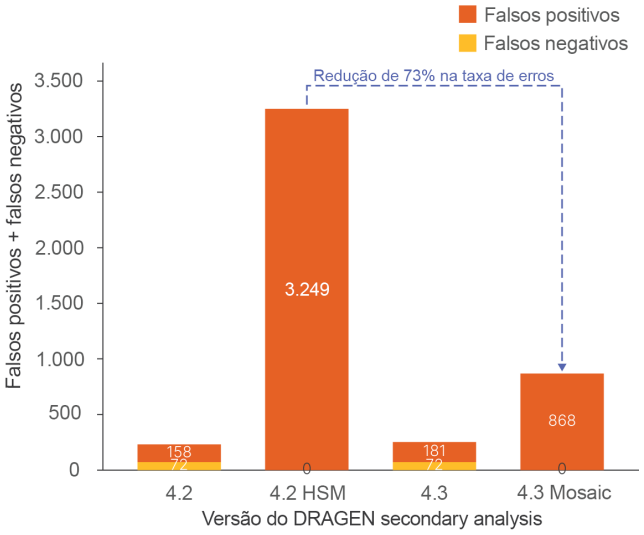

Figura 5: Precisão e exatidão aprimoradas com o modo de detecção de mosaico: há uma redução de erro de 73% do DRAGEN secondary analysis v4.2 no modo de alta sensibilidade (HSM) em comparação com o DRAGEN v4.3 no modo de detecção de mosaico. Os dados também mostram o alto número de falsos negativos sem a ativação do modo HSM ou da detecção de mosaico.

Detecção de SV e CNV

Variantes estruturais (SV) são alterações genômicas de 50 bp ou mais, e variantes de número de cópias (CNVs) são um tipo específico de SV em que o número de cópias de uma sequência genômica é reduzido (deleções) ou aumentado (inserções). O DRAGEN secondary analysis mostra maior precisão para identificação de SV (Figura 6) e identificação de CNV (Figura 7) em comparação com soluções alternativas.[7] Os algoritmos avançados e novas abordagens adaptadas para regiões genômicas complexas diferenciam o DRAGEN secondary analysis de outras soluções.

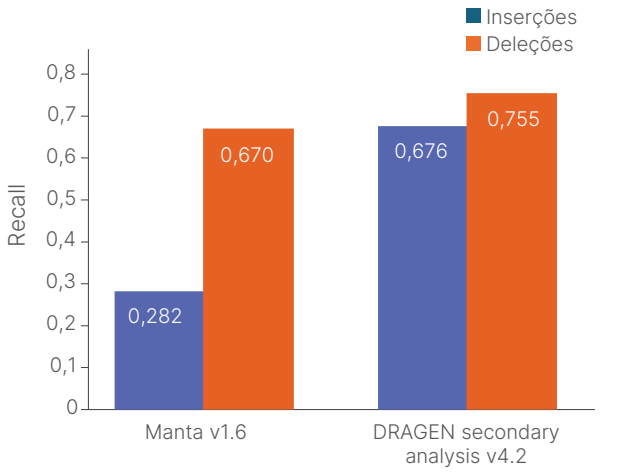

Figura 6: Identificação de SV altamente precisa com o DRAGEN secondary analysis: comparação de recall de indel de SV dos dados de referência do DRAGEN secondary analysis v4.2 e do Manta v1.6 avaliados com dados de referência do Genome in a Bottle (GIAB SV v0.6).[7]

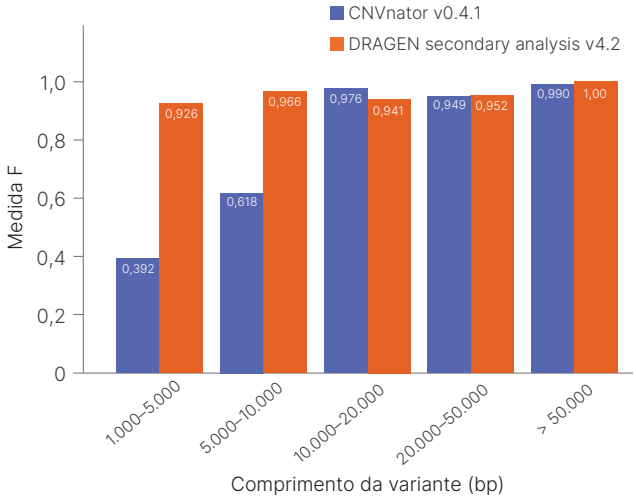

Figura 7: Identificação de CNV altamente precisa com o DRAGEN secondary analysis: identificação de CNV feita pelo DRAGEN secondary analysis v4.2 em comparação com o CNVnator v1.6 em diferentes tamanhos de deleções com base nos dados de referência do Genome in a Bottle (GIAB SV v0.6).[7]

O identificador do DRAGEN SV melhora os métodos de identificação de variantes estruturais Manta e incorpora informações da referência pangenômica, levando a uma filtragem mais precisa e maior precisão na detecção de SV. Isso inclui um novo detector de inserção de elemento móvel para identificar inserções grandes, parâmetros de par otimizados para melhor identificação de deleções grandes, e alinhamento contíguo refinado para melhor descoberta de inserções. Além disso, o software DRAGEN introduz melhorias nas etapas de montagem, nos cálculos de probabilidade de leitura e no manuseio aprimorado de pares sobrepostos e bases cortadas.

O identificador DRAGEN CNV é baseado principalmente na profundidade de leitura e oferece suporte a diversos modelos de segmentação e pontuação para diferentes aplicações. Ao aproveitar o sinal adicional de leituras discordantes e divididas, como é feito na identificação de SV, o identificador de CNV melhora a sensibilidade para capturar eventos tão pequenos quanto 1 kbp.

O identificador do DRAGEN CNV também tem um módulo de extensão de duplicação segmentar, um recurso que permite a detecção de CNV em regiões de duplicação segmentar do genoma. As regiões de duplicação segmentar são regiões do genoma com > 90% de similaridade de sequência, representando 5% do genoma. Elas têm baixa mapeabilidade, tornando desafiadora a detecção de variantes nessas regiões. A extensão de duplicação segmentar resgata aproximadamente um milhão de bases de regiões CNV que foram anteriormente excluídas da análise. Isso permite a detecção de CNV em 43 genes clinicamente relevantes e melhora a precisão geral da identificação de variantes.

Identificadores especializados e direcionados

Os identificadores direcionados suportam a genotipagem precisa de genes específicos que são difíceis de analisar devido a fatores como alta similaridade de sequência a pseudogenes, regiões repetitivas e altos graus de polimorfismo. O DRAGEN secondary analysis aborda esses desafios incorporando vários identificadores direcionados (Tabela 1), permitindo a genotipagem precisa de genes clinicamente relevantes. Para insights de farmacogenômica (PGx), o Identificador de alelos estrela PGx identifica alelos estrela e o status metabolizador para 22 genes (Tabela 2).

O identificador de antígenos leucocitários humanos (HLA) DRAGEN permite a genotipagem altamente precisa de alelos HLA de classe I e II. Ele alinha leituras a um banco de dados abrangente de mais de 9.000 alelos e pode ajudar em aplicações como correspondência de transplante de órgãos, imunogenética e estudos de associação de doenças.

Tabela 1: Resumo dos genes abordados por identificadores direcionados e especializados.

Identificador direcionado	Área de aplicação de pesquisa	Associação de condições
CYP21A2	Triagem do portador	Hiperplasia adrenal congênita (HAC)
HBA	Triagem do portador	α-talassemia
GBA	Triagem do portador	Doença de Gaucher, doença de Parkinson
SMN	Triagem do portador	Atrofia muscular espinhal
LPA	Doenças cardiovasculares	Doença arterial coronariana
RH	Tipagem sanguínea	–
CYP2B6	PGx	–
CYP2D6	PGx	–
HLA	Correspondência de transplantes, imunogenética	–

Tabela 2: Genes com relevância PGx abordados pelo Identificador de alelos estrela PGx

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

O DRAGEN secondary analysis v4.3 introduz uma nova classe de identificadores que permite a detecção de variantes *de novo* em regiões com duplicações segmentares. O identificador de detecção conjunta em várias regiões (MRJD) implementa um identificador de variante pequena *de novo* baseado em haplótipo para seis genes clinicamente relevantes em regiões de duplicação segmentar (Tabela 3).

Tabela 3: Genes abordados pelo identificador da MJRD

Identificador direcionado	Área de aplicação de pesquisa	Associação de condições
PMS2	Triagem de câncer hereditário	Síndrome de Lynch para câncer colorretal/endometrial
SMN1 (variantes pequenas)	Triagem do portador	Atrofia muscular espinhal
STRC	Triagem do portador	Perda auditiva não sindrômica
NEB	Triagem do portador	Miopatia nemalínica
TTN	Triagem de recém-nascidos, doenças raras	Cardiomiopatia
IKBKG	Triagem de recém-nascidos	Incontinentia pigmenti, displasia ectodérmica hipoidrótica

Resumo

O DRAGEN secondary analysis fornece análise secundária altamente precisa, abrangente e eficiente para aplicações de NGS. Melhorias contínuas fornecem maior precisão e cobertura expandida de regiões difíceis do genoma, permitindo a detecção de variantes desafiadoras e clinicamente relevantes.

Apêndice

Mapeamento multigenômico com referência pangenômica

Ao usar haplótipos populacionais de variantes em fases e aumentar o índice de referência com contigs alternativas derivadas da população, o DRAGEN secondary analysis pode mapear efetivamente contra uma referência pangenômica e melhorar o mapeamento de leituras da Illumina em regiões difíceis. Esse novo recurso amplia efetivamente o alcance das leituras da Illumina e permite precisão no mapeamento e identificação de variantes em regiões que anteriormente não podiam ser acessadas.

Um mapeador de múltiplos genomas é uma abordagem para ajudar no mapeamento com dados populacionais em que o conteúdo de sequência alternativa, observado na população, é representado como vários caminhos divergentes e convergentes (Figura 8). As leituras de amostras podem ser alinhadas a qualquer caminho mais compatível por meio do mapeador de múltiplos genomas.

Saiba mais: 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.

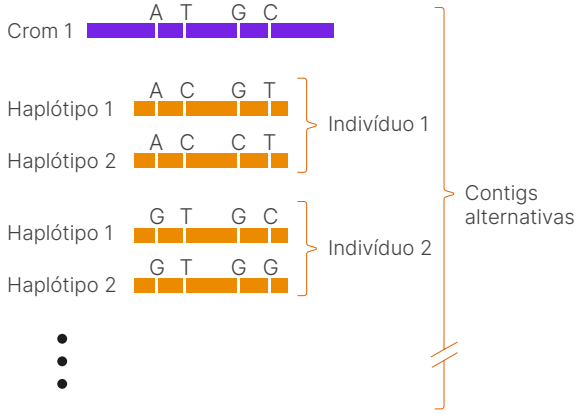

Figura 8: Mapeador de múltiplos genomas com referência pangenômica: Em uma referência, o conteúdo de sequência alternativa registrado em uma população é representado como vários caminhos divergentes e convergentes.

Mascaramento alternativo

Desde a atualização do software DRAGEN secondary analysis v3.9, o software DRAGEN inclui mascaramento alternativo, uma nova abordagem para lidar com contigs alternativas de referência nativa, em que as posições estratégicas das contigs alternativas são mascaradas para aumentar a precisão. Essa abordagem é simples de definir, manter e refinar ao longo do tempo.

Saiba mais, DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph

Aprendizado de máquina

O software DRAGEN secondary analysis v3.9 adicionou um pipeline de recalibração de ML poderoso e eficiente como uma opção dentro do fluxo de trabalho de variantes pequenas da linha genética. Ele fica ativado por padrão no software DRAGEN secondary analysis v4.0. Quando ativado, o pipeline executa o modelo de ML após a identificação de variantes padrão. Esta etapa recalibra os campos QUAL e GQ que são enviados para o VCF final. Em alguns casos, o ML pode alterar o GT. Os valores de aprendizado pré-máquina desses campos são preservados nos campos DQUAL, DGT e DGQ para que nenhuma informação seja perdida. Esta etapa adiciona aproximadamente cinco minutos ao fluxo de trabalho padrão para uma corrida de linha genética de 30× WGS, de modo que as melhorias de precisão têm um impacto limitado no tempo total da corrida.

O modelo de ML é gerado usando treinamento off-line supervisionado. O modelo processa um conjunto de recursos contextuais e baseados em leitura para refinar a precisão das pontuações de qualidade do identificador de variantes pequenas. Os recursos usados para treinar o modelo incluem mapeabilidade, AF, VC-Qual, DP, conteúdo GC, incompatibilidades e outras métricas internas de mapeamento, alinhamento e VC.

Cálculo da pontuação F1

$$F1 = 2 \times (Recal \times Precisão)/(Recall + Precisão)$$

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

Linha de comando do DRAGEN

Encontre receitas iniciais no DRAGEN recipe-germline WGS

Referências

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. Acessado em 19 de setembro de 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. Publicado em 12 de janeiro de 2022. Acessado em 19 de setembro de 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. Publicado em 12 de agosto de 2024. Acessado em 30 de setembro de 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. Publicado em 07 de junho de 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. Acessado em 19 de setembro de 2024.
6. Dados internos em arquivo. Illumina, Inc., 2022.
7. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. Publicado on-line em 25 de outubro de 2024. doi:10.1038/s41587-024-02382-1

+1 (800) 809-4566, ligação gratuita (EUA) | tel. +1 (858) 202-4566
techsupport@illumina.com | www.illumina.com

M-GL-01016 PTB v3.0