DRAGEN™ Secondary Analysis를 통한 생식세포 연구의 변이 검출 정확도 향상

Illumina의 머신 러닝 및 multigenome mapping 기술로 변이 검출 성능 최적화

향상된 정확도

머신 러닝과 pangenome reference를 포함하는 multigenome mapper로 현격히 오류 감소

모자이크 변이 검출

머신 러닝 기반의 germline small variant caller로 정확하고 정밀하게 모자이크 변이 검출

확장된 커버리지

Specialized caller와 targeted caller로 매핑이 어려운 영역에 대한 폭넓은 커버리지 제공

문서 번호: M-KR-00121 v2.0 KOR
영어 원본: M-GL-01016 v4.0

소개

생물의학 연구와 정밀의료(precision medicine)의 발전을 위해서는 차세대 시퀀싱(next-generation sequencing, NGS)을 통해 유전체(genome)의 잠재력을 이끌어 내는 것이 매우 중요합니다. 연구자가 NGS를 통해 얻는 정보를 최대한 활용하려면 원시 시퀀싱 데이터를 유의미한 결과로 해석해 주는 데이터 분석 도구가 필요합니다. DRAGEN Secondary Analysis는 NGS 데이터의 정확하고 포괄적이며 효율적인 2차 분석을 제공합니다. 고도로 재구성 가능한(Highly reconfigurable) 필드 프로그래밍 가능 게이트 어레이(field programmable gate array, FPGA) 기술을 채택한 DRAGEN Secondary Analysis는 매핑(mapping), 정렬(alignment), 변이 검출(variant calling) 등의 2차 NGS 데이터 분석 속도를 높여 줍니다. 또한 DRAGEN Secondary Analysis는 유전체 분석 시 흔히 발생하는 긴 처리 시간, 방대한 양의 데이터, 분석이 어려운 영역에서의 변이 검출 등의 난제 해결에 중점을 두고 설계되었습니다.

DRAGEN Secondary Analysis는 매우 정확한 결과를 제공합니다. 그 예로 DRAGEN v3.7은 과거 2020 precision FDA Truth Challenge V2(이하 precisionFDA V2)에 참가해 All Benchmark Regions(전체 벤치마크 영역) 및 Difficult-to-Map Regions(매핑이 어려운 영역) 부문에서 가장 정확한 분석 결과를 보이며 우승을 차지한 바 있습니다(그림 1).[1,2] 이후 새롭고 강력한 기능을 기반으로 기존의 우수한 성능을 획기적으로 향상시킨 DRAGEN v4.4는 All Benchmark Regions 부문에서 99.90%라는 F1 점수(정밀도(precision) 및 재현율(recall)의 통합 측정 지표)를 받은 뛰어난 작은 변이 검출(small variant calling) 정확도를 보여 줍니다. 최근 Baylor College of Medicine에서 발표한 연구 결과에 따르면, DRAGEN Secondary Analysis는 단일 염기서열 변이(single nucleotide variation, SNV), 삽입 또는 결실(insertion/deletion, Indel), 짧은 연쇄 반복(short tandem repeat, STR), 구조적 변이(structural variation, SV), 유전자 복제수 변이(copy number variation, CNV)를 포함한 모든 변이 타입에서 현존하는 다른 최첨단 분석 방법들보다 우수한 속도와 정확도를 갖춘 것으로 확인되었습니다.[3]

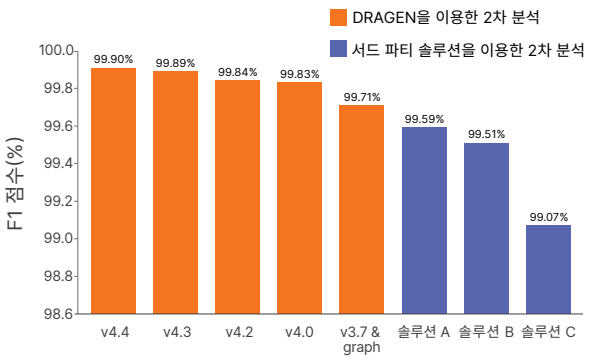

그림 1: All Benchmark Regions 부문에서 DRAGEN Secondary Analysis의 정확도

F1 점수(%)는 진양성(true positive) 및 진음성(true negative) 결과를 전체 결과에 대한 비율로 계산한 값을 나타냄.[5,6] 높아진 F1 점수는 레퍼런스 데이터 기준 향상된 정확도를 의미함

이 Technical Note는 pangenome reference를 포함하는 multigenome mapper, 머신 러닝(machine learning, ML)의 통합, 모자이크 변이(mosaic variant) 검출 기능, specialized caller, 향상된 SV 및 CNV 검출력 등 DRAGEN Secondary Analysis의 높은 정확도에 기여한 최신 개선 사항을 기술합니다.

Pangenome reference를 포함한 multigenome mapper

DRAGEN v3.7부터 도입되었던 pangenome reference를 활용하는 multigenome mapping 기능은 변이 검출 정확도를 한층 더 높여 줍니다.[4] DRAGEN v4.3에서는 더 많은 개인의 유전체 변이를 대표하는 pangenome reference까지 확장 가능한 기능을 탑재한 2세대 multigenome mapper를 처음 도입했습니다. 이번에 출시된 DRAGEN v4.4는 정확도가 크게 향상되어, 오류가 DRAGEN v3.6.3과 비교했을 때 87%, DRAGEN v4.2.7과 비교했을 때는 47% 감소했습니다(그림 2).

연구자는 이미 보유하고 있는 어셈블리(assembly)나 Human Pangenome Reference Consortium(HPRC)에서 제공하는 다양한 어셈블리를 사용해 맞춤화된 pangenome reference를 만들 수 있습니다. DRAGEN v4.3에서 처음 소개된 이 기능은 연구자가 특정 집단을 더 명확히 대표하는 레퍼런스를 만들 수 있도록 해주어, 혈통 편향(ancestry bias)을 줄이고 변이 검출 정확도는 높입니다. 맞춤화된 pangenome reference는 집단

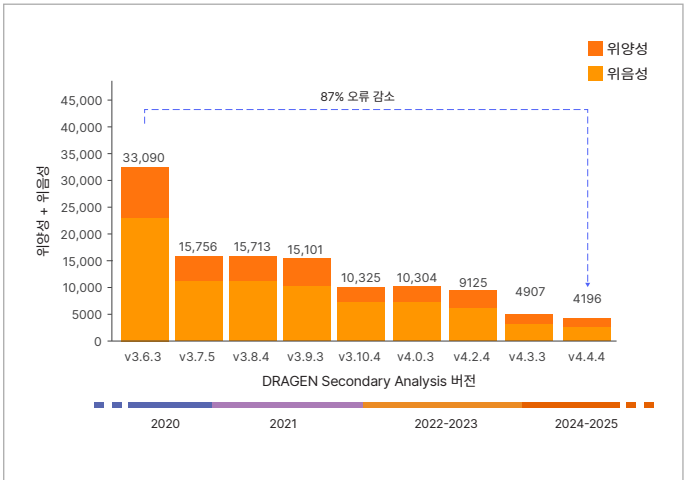

그림 2: DRAGEN Secondary Analysis에 적용된 지속적인 기술 혁신

Genome in a Bottle 샘플 HG002, NIST v4.2.1 사용 시[7] SNP 및 Indel의 위양성률(false positive rate)과 위음성률(false negative rate)이 개선되어 현격한 오류 감소뿐 아니라 Difficult-to-Map Regions 부문에서의 우수한 SNP 및 Indel 정확성이 확인됨

레벨의 하플로타입(haplotype)을 통합함으로써 다양한 코호트(cohort)에 걸쳐 매핑 성능을 강화하고 보다 공평한 유전체 분석을 지원합니다.

DRAGEN v4.4부터는 사용하는 샘플에 맞춰 레퍼런스를 만들어 주는 개인화(personalization)된 pangenome reference라는 특별한 기능이 도입되었습니다. 기존의 맞춤화된 레퍼런스가 집단 레벨의 맞춤화를 지원하는 반면, 개인화는 샘플 특이적 데이터를 사용하여 각 유전체에 대한 레퍼런스를 최적화합니다. 이 방법은 특히 매핑이 어려운 영역에서의 매핑 정확도와 변이 검출력을 향상시키고, 여러 DRAGEN 구성 중에서 가장 높은 정확도를 제공합니다(그림 3).

머신 러닝(ML)

DRAGEN v3.9에 최초로 추가된 후 v3.10에서 업그레이드된 ML 모듈은 DRAGEN Secondary Analysis의 다양한 variant caller에서 가져온 맥락적(contextual) 기능 및 리드(read) 기반 기능을 활용하는 지도 학습 모델(supervised model)을 채택했습니다. 최적의 결과를 제공하기 위해 multigenome mapping과 ML을 함께 적용해 거짓 콜(false call)을 줄임으로써 작은 변이 검출 정확도를 높였습니다(그림 2). 모델 학습 과정에 사용되지 않았던 타 집단의 테스트 데이터를 비롯한 모든 연구 대상에 걸쳐 현저히 향상된 정확도가 일관되게 관찰되었습니다.

모자이크 변이 검출

새로운 ML 모델이 지원되는 DRAGEN Secondary Analysis v4.3 이후 버전은 germline small variant caller 내에서 모자이크 변이를 검출합니다. 연구자는 경우에 따라 Mosaic Mode에 더 낮은 임곗값을 설정하여 20% 미만의 낮은 대립유전자 빈도(allele frequency)로 존재하는 모자이크 변이를 검출할 수 있으며, 이 값은 높은 뎁스(depth)의 샘플 분석 시 필요한 만큼 낮게(예: 1% 미만) 설정할 수 있습니다. DRAGEN v4.3 및 DRAGEN v4.4는 이전 버전보다 더 높은 정확도와 정밀도로 모자이크 변이를 검출합니다. 이를 입증하기 위해 National Institute of Standards and Technology(NIST)의 Mosaic 진리 집합(truth set) 데이터를 사용해 DRAGEN v4.2, HSM(high-sensitivity mode, 고민감도 모드)에서의 DRAGEN v4.2, DRAGEN v4.3, Mosaic Mode 활성화 상태에서의 DRAGEN v4.3, 이렇게 총 네 가지 DRAGEN 분석 파이프라인을 테스트했습니다. NIST의 Mosaic 진리 집합에는 300× 데이터에 존재하는 알려진 모자이크 변이 73개가 포함되어 있는데, 이 변이들은 DRAGEN v4.2 및 v4.3로는 검출되지 않았으나 HSM에서의 DRAGEN v4.2 및 Mosaic Mode에서의 DRAGEN v4.3로는 검출되었습니다. 그러나 Mosaic Mode에서의 DRAGEN v4.3는 HSM에서의 DRAGEN v4.2보다 위양성 수가 73% 적은 것으로 나타나, 모자이크 변이 검출 정확도가 더 높은 것을 확인할 수 있었습니다(그림 4).

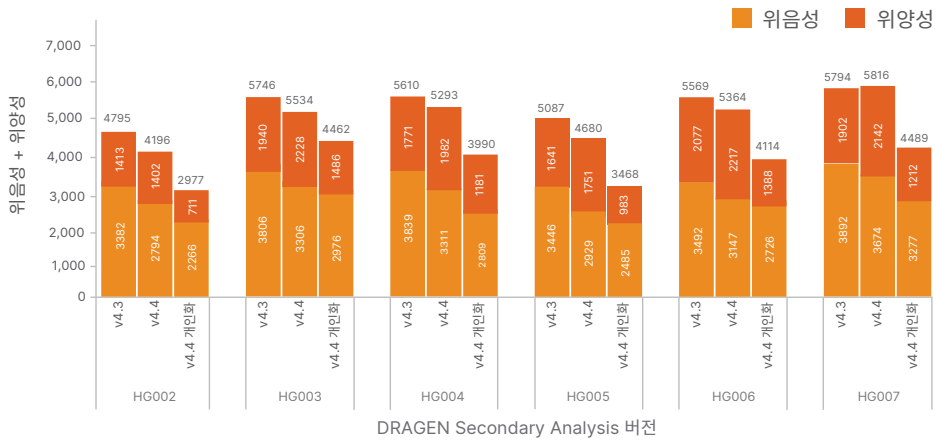

그림 3: 개인화된 pangenome reference를 사용하여 향상된 DRAGEN Secondary Analysis의 작은 변이 검출 정확도
매핑이 어려운 영역에 대한 NIST v4.2.1 진리 집합과 비교했을 때, 모든 DRAGEN 구성 중 개인화를 지원하는 DRAGEN v4.4가 가장 높은 정확성을 보임. DRAGEN v4.3와 DRAGEN v4.4의 성능을 능가하는 이러한 결과는 DRAGEN v4.3에서 베타로 출시되고 DRAGEN v4.4에서 정식 출시된 개인화 기능의 영향을 잘 보여줌

문서 번호: M-KR-00121 v2.0 KOR
영어 원본: M-GL-01016 v4.0

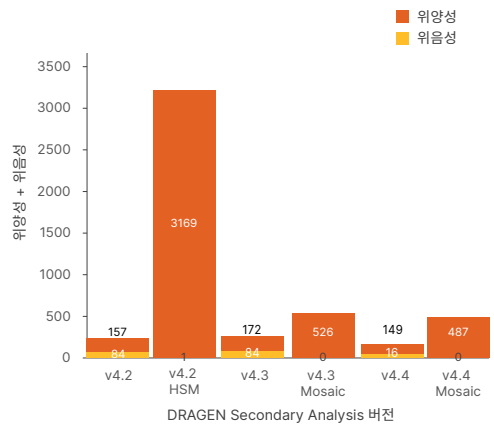

그림 4: Mosaic Detection Mode에서 향상된 정확도 및 정밀도

Mosaic Detection Mode에서 DRAGEN v4.4를 실행했을 때 HSM Mode에서 DRAGEN v4.2를 실행했을 때보다 오류가 약 85% 감소함. 상기 데이터를 보면 HSM Mode 또는 Mosaic Detection Mode를 활성화하지 않았을 때 위음성이 더 많은 것을 확인할 수 있음

SV 및 CNV 검출

SV는 길이가 50 bp 이상인 유전자 변이를 의미하며, CNV는 유전체 시퀀스의 복제수가 감소(결실) 또는 증가(삽입)한 특정 타입의 SV를 뜻합니다. DRAGEN Secondary Analysis는 정확한 SV 검출력(그림 5) 및 CNV 검출력(그림 6)을 제공합니다.[3] DRAGEN Secondary Analysis를 다른 솔루션과 차별화하는 요소는 바로 복잡한 유전체 영역에 맞춤화된 진보된 알고리즘과 새로운 접근법입니다.

DRAGEN v4.4는 SV caller에 pangenome reference를 확대 적용함으로써 SV 검출 정확도를 크게 향상시켰습니다. 26가지 혈통에 걸친 128개의 샘플을 포함하는 SV 집단 하플로타입이 DRAGEN pangenome reference에 통합되었습니다. F1 점수가 30% 넘게 향상된 DRAGEN Secondary Analysis는 쇼트 리드(short read)를 사용하여 기존에 달성하지 못했던 SV 검출 정확도를 보이고 있습니다. DRAGEN CNV caller는 기본적으로 리드 뎁스(read depth) 기반의 검출 도구로, 여러 애플리케이션에 적합한 다양한 분절화(segmentation) 및 채점(scoring) 모델을 지원합니다. DRAGEN CNV caller는 SV 검출과 마찬가지로 디스코던트 리드(discordant read)와 스플릿 리드(split read)의 추가적인 시그널을 활용하여 최소 1 kb의 이벤트까지 포착하도록 민감도를 높여 줍니다. 또한 DRAGEN CNV caller는 유전체의 분절 중복 영역(segmental duplication region)에서도 CNV 검출이 가능하도록 Segmental Duplication Extension 모듈을 지원하고 있습니다. 분절 중복 영역은 유전체에서 시퀀스 상동성(sequence similarity)이 90%를 초과하는 영역으로, 전체 유전체의 5%를 차지합니다. 이 영역에서는 매핑률(mappability)이 낮아 변이 검출이 어렵습니다. Segmental Duplication Extension 모듈은 이전에는 분석에서 제외되었던 CNV 영역 내 약 100만 개의 염기(base)를 다시 포함시킬 수 있습니다. 이 모듈은 의학적 관련이 있는 43개의 유전자에서 CNV 검출이 가능하도록 해주며 전반적인 변이 검출 정확도도 향상시킵니다.

Targeted Caller

Targeted caller는 유사 유전자(pseudogene)와의 높은 시퀀스 상동성, 반복적인 영역(repetitive region), 높은 빈도의 다형성(polymorphism)과 같은 요인으로 인해 분석이 어려운 특정 유전자에 대한 정확한 유전형 분석(genotyping)을 지원합니다. DRAGEN Secondary Analysis는 다양한 targeted caller(표 1)를 통합하여 의학적 관련이 있는 유전자의 정밀한 유전형 분석을 지원함으로써 이러한 어려움을 해소해 줍니다. 예를 들어 약물유전체학적(pharmacogenomics, PGx) 통찰력을 확보하려면 PGx Star Allele Caller를 사용해 22개의 유전자에 대한 별표 대립유전자(star(*) allele) 및 대사자 상태(metabolizer status)를 검출할 수 있습니다(표 2). 또 DRAGEN HLA caller를 사용하면 높은 정확도로 HLA(human leukocyte antigen, 인간 백혈구 항원) Class I 및 Class II 대립유전자의 유전형을 분석할 수 있습니다. DRAGEN HLA caller는 9천 개 이상의 대립유전자가 있는 포괄적인 데이터베이스에 리드를 정렬하며, 장기 이식 매칭(organ transplantation matching), 면역유전학(immunogenetics), 질병 연관성(disease association) 연구와 같은 분야에도 응용해 볼 수 있습니다.

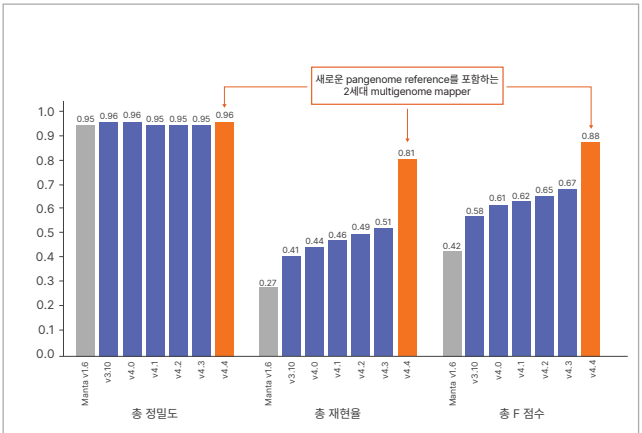

그림 5: DRAGEN Secondary Analysis v4.4의 향상된 SV 검출 정확도

HG002 NIST v1.011 벤치마크 데이터를 사용해 확인한 여러 DRAGEN 버전과 Manta v1.6의 SV 정밀도, 재현율 및 F1 점수 비교

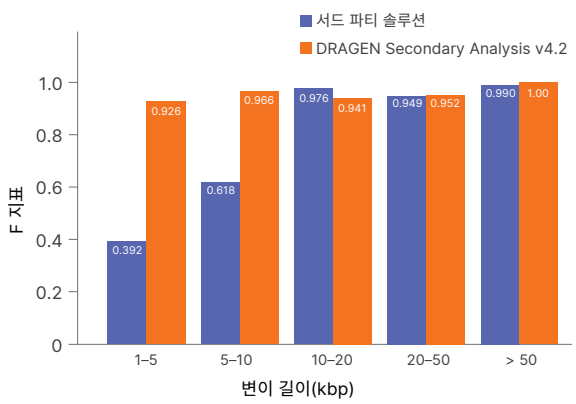

그림 6: DRAGEN Secondary Analysis의 매우 높은 CNV 검출 정확도

Genome in a Bottle(GIAB SV v0.6) 벤치마크 데이터를 사용해 각기 다른 크기의 결실에서 DRAGEN Secondary Analysis v4.2로 CNV를 검출한 결과[3]

표 1: Targeted caller로 검출 가능한 유전자의 요약

Targeted Caller	연구용 애플리케이션	질병 연관성 연구
CYP21A2	보인자 검사 (Carrier screening)	선천성 부신 과형성증 (Congenital adrenal hyperplasia, CAH)
HBA	보인자 검사	알파 지중해 빈혈 (α-Thalassemia)
GBA	보인자 검사	고셔병, 파킨슨병
SMN	보인자 검사	척추성 근위축증 (Spinal muscular atrophy, SMA)
LPA	심혈관 질환	관상 동맥 질환
RH	혈액형 검사	
CYP2B6	PGx	
CYP2D6	PGx	
HLA	이식 매칭, 면역유전학	자가면역 질환, 감염성 질환, 특정 암

DRAGEN Secondary Analysis v4.3부터 분절 중복이 있는 영역에서 새로운(*de novo*) 변이 검출을 지원하는 새로운 종류의 caller가 도입되었습니다. MRJD(Multi Region Joint Detection) caller는 분절 중복 영역에 존재하는 의학적 관련이 있는 유전자 6개에 대한 하플로타입 기반의 *de novo* small variant caller를 구현합니다(표 3).

표 2: PGx Star Allele Caller로 검출 가능한 PGx 관련 유전자

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

표 3: MRJD caller로 검출 가능한 유전자의 요약

유전자	연구용 애플리케이션	질병 연관성 연구
PMS2	유전성 암 검진	린치 증후군 (Lynch syndrome; 대장암 및 자궁내막암)
SMN1, SMN2	보인자 검사	SMA
STRC	보인자 검사	비증후군성 난청 (Nonsyndromic hearing loss)
NEB	보인자 검사	네말린 근병증 (Nemaline myopathy)
TTN	신생아 선별 검사 및 희귀 질환, ACMG 2차 소견 목록	심근 병증 (Cardiomyopathy)
IKBKG	신생아 선별 검사	색소 실조증(Incontinentia pigmenti), 저한성 외배엽 이형성증(hypohidrotic ectodermal dysplasia)

ACMG = American College of Medical Genetics and Genomics(미국의학유전학회)

문서 번호: M-KR-00121 v2.0 KOR
영어 원본: M-GL-01016 v4.0

요약

DRAGEN Secondary Analysis는 NGS 애플리케이션을 위한 매우 정확하고 포괄적이면서 효율적인 2차 분석을 지원합니다. 지속적인 기술 향상을 기반으로 더 높은 정확도뿐만 아니라 분석이 어려운 유전체 영역에 대한 확장된 커버리지를 제공하여, 이전에는 검출이 어려웠던 의학적 관련이 있는 변이의 검출도 가능케 해 줍니다.

상세 정보

DRAGEN Secondary Analysis

부록

Pangenome reference를 포함하는 multigenome mapping 기능

DRAGEN Secondary Analysis는 페이징된 변이(phased variant)의 집단 하플로타입을 활용하고 집단에서 유래된 alternate contig(ALT contig)로 레퍼런스 인덱스(reference index)를 증대시켜 pangenome reference에 효과적으로 매핑하고 분석이 어려운 영역에서 Illumina 리드의 매핑을 향상시킬 수 있습니다. 이 새로운 기능은 Illumina 리드의 적용 범위를 효과적으로 확장해 주고 이전에는 접근이 어려웠던 영역에서의 정확한 매핑 및 변이 검출을 가능하게 해 줍니다.

Multigenome mapper는 특정 집단에서 관찰되는 alternate sequence 콘텐츠를 다양하게 갈라지고 합쳐지는 패스(path)로 표현해 주는 집단 데이터의 매핑 도구입니다(그림 7). 샘플 리드는 multigenome mapper를 통해 최적의 매칭 패스에 정렬될 수 있습니다.

관련 논문 보기: 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

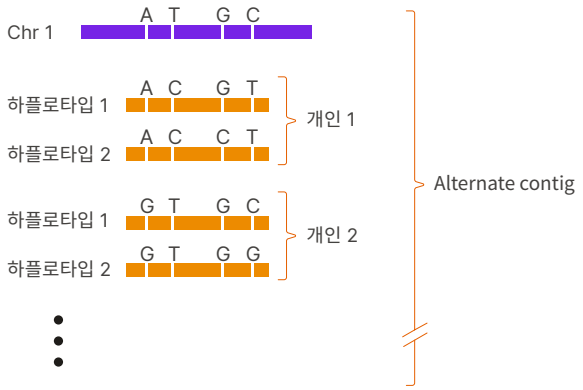

그림 7: Pangenome reference를 포함하는 multigenome mapper
Pangenome reference는 특정 집단에서 관찰되는 alternate sequence 콘텐츠를 다양하게 갈라지고 합쳐지는 패스로 표현함

Alt-masking 기능

DRAGEN 소프트웨어는 DRAGEN v3.9 업데이트부터 원래의 레퍼런스 ALT contig를 처리하는 새로운 방법인 Alt-masking 기능을 제공하고 있습니다. Alt-masking은 정확도 향상을 위해 ALT contig의 전략적 위치를 감추는 기능으로, 이 도구는 시간이 지나도 정의, 유지 및 개선이 쉽습니다.

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

Machine learning 모델

DRAGEN Secondary Analysis v3.9 소프트웨어는 생식세포 작은 변이 검출 워크플로우 내 하나의 옵션으로 강력하고 효율적인 ML 재보정(recalibration) 파이프라인을 추가했습니다. 이 파이프라인은 DRAGEN Secondary Analysis v4.0 이상 버전부터는 기본적으로 설정되어 있으며, 일반적인 변이 검출 작업이 완료되면 ML 모델을 실행합니다. 이 단계에서는 최종 VCF 파일에 포함되는 QUAL 및 GQ 필드가 재보정됩니다. 경우에 따라서는 ML 모델이 GT를 변경할 수 있습니다.

정보 손실을 방지하기 위해 세 필드의 머신 러닝 실행 이전 값은 DQUAL , DGT 및 DGQ 필드에 보존됩니다. 이 단계는 30× 전장 유전체 시퀀싱(whole-genome sequencing, WGS)을 기준으로 생식세포 런(run) 수행 시 표준 워크플로우를 약 5분 연장하므로 이 단계가 주는 정확도 향상이 전체 런 타임에 미치는 영향은 제한적이라 할 수 있습니다.

ML 모델은 오프라인 지도 학습을 통해 생성됩니다. 이 모델은 일련의 리드 기반 기능 및 맥락적 기능을 처리하여 small variant caller의 Q-Score(quality score, 품질 점수) 정확도를 개선합니다. 모델의 학습에 사용되는 기능으로는 매핑률, AF, VC-Qual, DP, GC 함량(GC content), 미스매치(mismatch) 및 기타 내부 매핑(internal mapping), 정렬 및 VC 메트릭스가 있습니다.

F1 점수 계산

$$F1 = 2 \times (Recall \times Precision)/(Recall + Precision)$$

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

DRAGEN 커맨드 라인

DRAGEN recipe-germline WGS에서 starter recipe를 확인하시기 바랍니다.

AF = allele frequency(대립유전자 빈도), DP = depth of coverage(커버리지 뎁스), GQ = Phred-scaled probability that the call is incorrect(콜이 부정확할 Phred 척도 확률), GT = genotyping(유전형 분석), QUAL = Phred-scaled probability that the site has no variant(부위에 변이가 존재하지 않을 Phred 척도 확률), VC-Qual = variant confidence quality(변이 신뢰도 품질)

참고 문헌

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. Accessed April 3, 2025.
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. Published January 12, 2022. Accessed April 3, 2025.
3. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. 2024. Published online ahead of print. doi:10.1038/s41587-024-02382-1
4. 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-genmultigenome-mapping.html. Published August 12, 2024. Accessed September 30, 2024.
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. Accessed April 3, 2025.
6. Internal data on file. Illumina, Inc., 2022.
7. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025. doi:10.1038/sdata.2016.25

무료 전화(한국) 080-234-5300
techsupport@illumina.com | www.illumina.com

M-KR-00121 v2.0 KOR

문서 번호: M-KR-00121 v2.0 KOR
영어 원본: M-GL-01016 v4.0