TECHNICAL NOTE

DRAGEN™ Bio-IT Platform을 통한 생식세포 연구의 작은 변이 검출 정확도 향상

Illumina 머신 러닝과 Multigenome(그래프) 레퍼런스로 변이 검출 성능 최적화

illumina®

문서 번호: M-KR-00121 KOR
영어 원본: M-GL-01016 v1.0

소개

생물의학 연구와 정밀의료의 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 정확하고 효율적으로 시퀀싱 raw data를 의미 있는 결과로 해석할 수 있는 데이터 분석 도구가 필요합니다. Illumina DRAGEN(Dynamic Read Analysis for GENomics) Bio-IT Platform은 NGS 데이터의 정확하고 포괄적이며 효율적인 2차 분석을 지원합니다. DRAGEN 플랫폼은 고도로 재구성 가능한(highly reconfigurable) 필드 프로그래밍 가능 게이트 어레이(field programmable gate array, FPGA) 기술을 채택하여 매핑(mapping), 정렬(alignment), 변이 검출(variant calling) 등의 2차 NGS 데이터 분석 속도를 높입니다. DRAGEN 플랫폼의 근본적인 기능들은 유전체 분석 시 흔히 발생하는 긴 처리 시간, 방대한 양의 데이터, 분석이 어려운 영역에서의 변이 검출 등 난제 해결에 중점을 두고 있습니다.

본 Application Note는 연구 시 DRAGEN 플랫폼이 제공하는 향상된 생식세포(germline) 작은 변이 검출(small variant calling) 정확도를 진리 집합(truth set)과 비교하여 설명하며, DRAGEN v4.0 소프트웨어의 정확도를 Illumina 시퀀싱 데이터를 비롯한 PrecisionFDA Truth Challenge V2 제출된 다양한 데이터 세트를 분석한 DRAGEN v3.7 및 BWA GATK의 정확도와 비교하였습니다(그림 1).

DRAGEN의 정확도 향상

다음과 같은 세 가지 DRAGEN 기술 혁신이 광범위한 인구 집단 샘플에 걸쳐 인간 유전체 상당 부분에 대한 분석 정확도* 향상에 기여하였습니다.

- Multigenome(그래프) 레퍼런스로 분석이 어려운 영역의 매핑정확도 향상
- ALT 마스킹(Alt-masking)으로 매핑 모호성(ambiguity) 감소
- 머신 러닝(machine learning, ML)으로 작은 변이 검출 능력 개선

방법

PrecisionFDA Truth Challenge V2는 PrecisionFDA, Genome in a Bottle(GIAB) 컨소시엄 그리고 미국 국립 표준 기술 연구소(National Institute of Standards and Technology, NIST)의 주최로 진행되었습니다. 이 챌린지는 Difficult-to-Map Regions(매핑이 어려운 영역), 부분 중복(segmental duplication) 및 주조직 적합성 복합체(major histocompatibility complex, MHC)의 벤치마킹에 초점을 두고, 일반적인 기준틀에 대한 작은 변이 검출 파이프라인의 성능을 확인하기 위해 개최되었습니다.

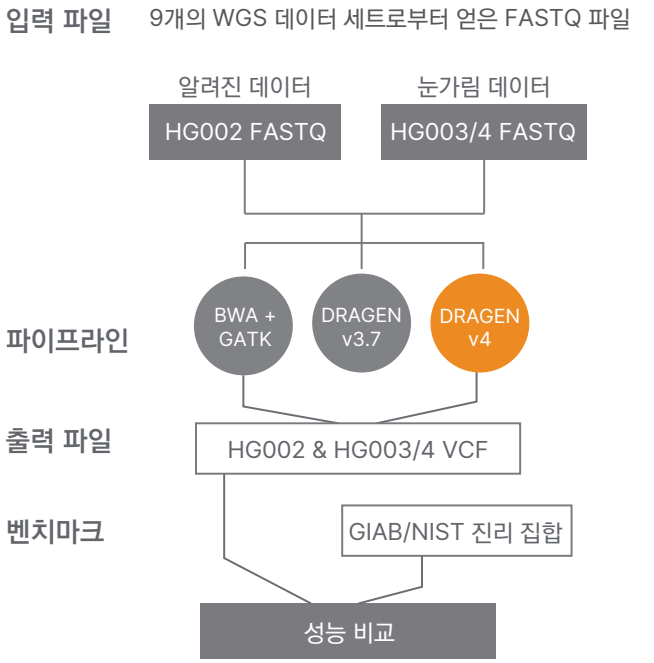

그림 1: PrecisionFDA Truth Challenge 개요도 — 세 가지 기술로 시퀀싱된 9개의 데이터 세트(HG002, HG003 및 HG004 샘플)로부터 얻은 FASTQ 파일을 다양한 분석 파이프라인으로 분석해 쿼리(query) VCF 파일을 생성한 후 GIAB/NIST 진리 집합과 비교.

GIAB 컨소시엄은 연결된 리드(linked read)와 롱 리드(long read)를 사용하여 의학적 관련이 있는 여러 유전자를 포함한 이전 진리 집합과 비교했을 때 유전체를 7% 더 커버하는 신뢰도* 높은 확장된 진리 콜(truth call) 집합을 개발하였습니다. 이렇게 확장된 진리 집합은 매핑률이 낮은 영역(low-mappability region)과 부분 중복 내 2억 7천만 개 이상의 염기(base)를 커버합니다.

이 새로운 진리 집합이 바로 세 가지 시퀀싱 기술이 적용된 제한적인 샘플 데이터에서 분석이 어려운 영역에 대한 가장 적합한 매핑 및 변이 검출 방법이 무엇인지를 결정하는 새로운 바이오인포매틱스(bioinformatics, 생명정보학) 챌린지의 기초가 되었습니다(그림 1). PrecisionFDA Truth Challenge V2 샘플, 진리 집합 그리고 결과는 최근 DRAGEN v4.0의 향상된 성능*을 벤치마킹하는 데 사용되었습니다.

* 연구 전용이며, 본 사양은 임상/진단 용도로 승인되지 않음.

결과

All Benchmark Regions에서 타 소프트웨어보다 높은 DRAGEN의 정확도

DRAGEN Bio-IT Platform은 매우 정확한 결과를 제공합니다. DRAGEN v3.7은 과거 2020 PrecisionFDA Truth Challenge V2에서 All Benchmark Regions(전체 벤치마크 영역) 및 Difficult-to-Map Regions 부문에서 가장 정확한 Illumina 시퀀싱 데이터 분석 결과를 보여 우승을 차지한 바 있습니다.

PrecisionFDA Truth Challenge 이후 Multigenome(그래프) 레퍼런스와 Illumina 머신 러닝의 지속적인 기술 혁신을 통해 개발된 DRAGEN v4.0은 모든 시퀀싱 기술이 적용된 데이터 세트에 걸쳐 단일 염기 다형성(single nucleotide polymorphism, SNP) 및 삽입/결실(insertion/deletion, Indel) 검출 정확도*의 새로운 기준을 제시하면서 All Benchmark Regions 부문에서 F1 점수 99.83%를 달성했습니다(그림 2, 표 1).

FP+FN으로 보는 DRAGEN의 정확도

DRAGEN v3.7은 이미 업계를 선도하는 다수의 인포매틱스 솔루션과 경쟁하는 위치에 있었지만, 여기에 몇 가지 새로운 기술 변경(부록 참조)을 적용해 탄생한 DRAGEN v4.0은 한층 더 향상된 정확도*를 갖추게 되었습니다. 또한 벤치마킹 비교 결과를 보면, 이러한 기술 향상을 토대로 DRAGEN v4.0이 해당 연구에서 분석된 모든 정확도 메트릭스(metrics)에서 일반적으로 널리 사용되는 다양한 분석 파이프라인 및 시퀀싱 기술에 걸쳐 매우 높은 정확도를 달성한 것을 알 수 있습니다(그림 2).

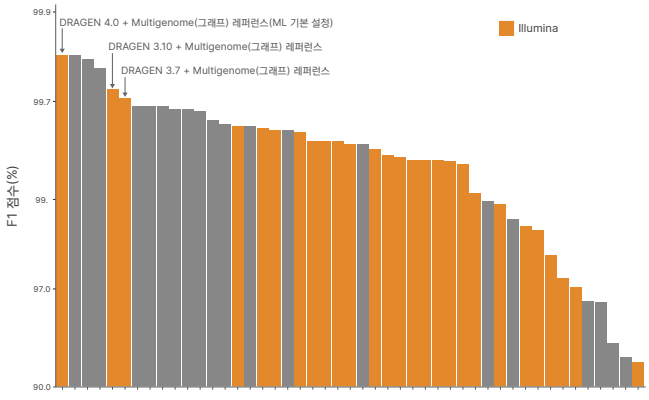

그림 2: 정확성의 새로운 기준을 제시하는 DRAGEN 플랫폼 — 참 양성(true positive) 및 참 음성(true negative) 결과를 전체 결과에 대한 비율로 계산한 F1 점수(%)로 DRAGEN v4.0 Mulitgenome(그래프) 레퍼런스(머신 러닝 기본 설정)의 뛰어난 정확도* 확인 가능.

표 1: PrecisionFDA Truth Challenge V2 벤치마킹 결과

순위	시퀀싱 기술	바이오인포매틱스 파이프라인	F1
1	Illumina	DRAGEN v4.0 & 그래프 (ML 기본 설정)[a]	0.9983
6	Illumina	DRAGEN v3.10 & 그래프[a]	0.9974
7	Illumina	DRAGEN v4.0 & 그래프[a]	0.9974
8	Illumina	DRAGEN v3.7 & 그래프	0.9971
39	Illumina	BWA-GATK(Genetalks)	0.9907

a. PrecisionFDA Truth Challenge V2의 일환으로 제출되지 않음.

Multigenome(그래프) 레퍼런스와 머신 러닝이 각각 전반적인 정확도에 어떠한 영향을 주는지 확인하기 위해, 다양한 벤치마킹 샘플에 걸친 거짓 양성(false positive, FP) 결과와 거짓 음성(false negative, FN) 결과를 Multigenome(그래프) 레퍼런스 활성/비활성 조건 및 머신 러닝 활성/비활성 조건으로 나누어 표로 정리했습니다(그림 3).* 머신 러닝의 경우 Multigenome(그래프) 레퍼런스 비활성 조건에서 10%, 활성화 조건에서는 약 30%의 오류 감소*가 관찰되었습니다. Multigenome(그래프) 레퍼런스와 머신 러닝이 모두 활성화된 조건에서는 시너지 효과로 인해 거짓 콜(false call)이 62% 감소*되었습니다.

* 연구 전용이며, 본 사양은 임상/진단 용도로 승인되지 않음.

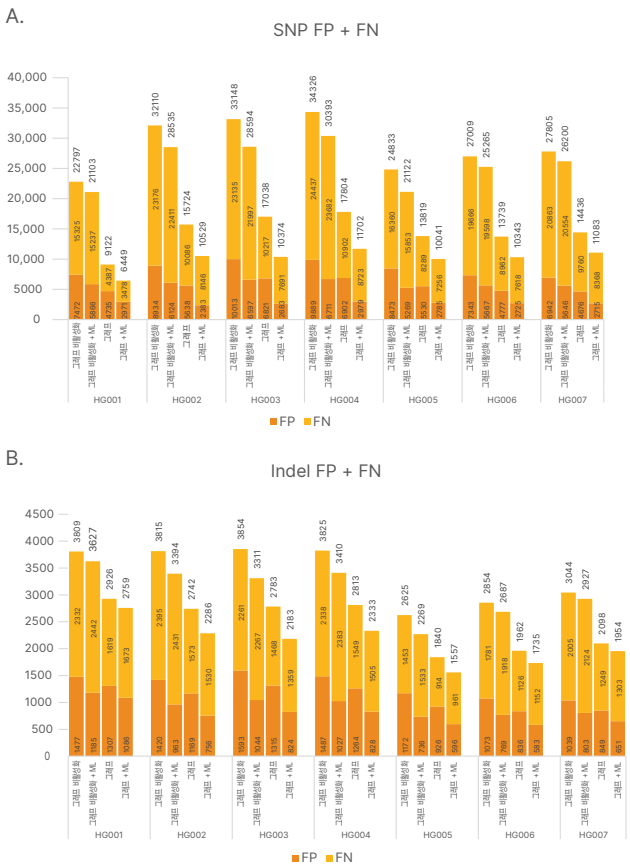

그림 3: FP 및 FN 결과를 줄이는 머신 러닝 및 Multigenome(그래프) 레퍼런스 — 머신 러닝이 활성화되었을 때 Multigenome(그래프) 레퍼런스 비활성 조건에서 10%, 활성화 조건에서는 약 30%의 오류 감소*가 관찰됨. Multigenome 레퍼런스와 머신 러닝이 모두 활성화되었을 때 (A) SNV와 (B) Indel의 거짓 콜은 62% 감소*됨.

분석이 어려운 영역에서도 높은 정확도

DRAGEN v4.0 소프트웨어는 All Benchmark Regions에서만 우수한 성능을 보이는 것이 아니라, 분석이 특히 어려운 MHC 영역에서도 매우 높은 SNP 및 Indel 검출 정확도*를 보입니다 (그림 4). 정확성, 포괄성, 효율성을 모두 갖춘 DRAGEN은 사용자가 NGS의 잠재력을 이끌어내어 유전체에 관해 최대한 많은 통찰을 얻을 수 있도록 해 줍니다.

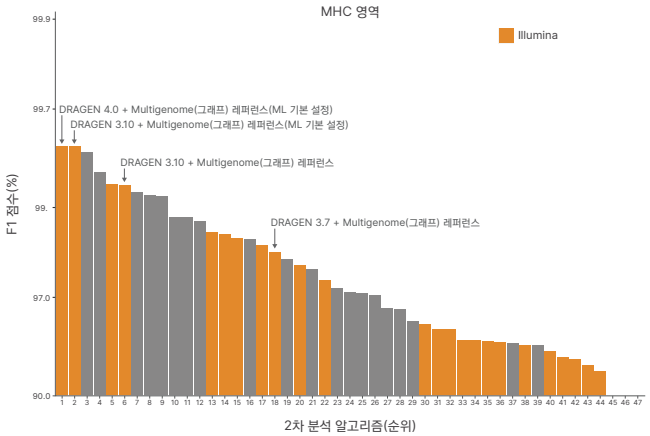

그림 4: MHC 영역에서 DRAGEN v4.0의 정확도 — DRAGEN v4.0는 다른 PrecisionFDA Truth Challenge V2 제출 건에 비해 MHC 영역에서 뛰어난 정확도(F1 점수 참조)를 보임.

결론

최근 DRAGEN의 매핑 및 생식세포 작은 변이 검출 성능이 향상되면서 DRAGEN 분석 기능을 활용하는 Illumina 시퀀싱 기술이 다른 시퀀싱 방법이나 분석 방법과 비교했을 때 더 높은 정확도를 제공할 수 있게 되었습니다. 정확성, 포괄성, 효율성을 모두 갖춘 DRAGEN은 연구자가 유전체학의 최대 잠재력을 이끌어낼 수 있도록 해 줍니다.

요약

DRAGEN 플랫폼은 매우 정확하고 포괄적이며 효율적인 규모에 맞는 2차 분석을 제공합니다. 지속적인 정확도의 향상과 분석이 어려운 유전체 지역까지의 영역 확장은 포괄적인 유전체 분석 솔루션에 매우 중요합니다. 이를 통해 분석이 어렵고 의학적 관련이 있는 변이의 검출이 가능해집니다. DRAGEN v4.0은 성능 향상에 힘입어 PrecisionFDA Truth Challenge V2에 제출된 시퀀싱 기술과 분석 방법에 걸쳐 가장 정확한 작은 변이 검출력을 보였습니다.

* 연구 전용이며, 본 사양은 임상/진단 용도로 승인되지 않음.

부록

Multigenome(그래프) 레퍼런스

DRAGEN 플랫폼은 페이징된 변이(phased variant)의 인구 집단 내 하플로타입(haplotype)을 활용하며 인구 집단에서 유래된 ALT 콘티그(ALT contig)로 참조 인덱스(reference index)를 증대시켜 Multigenome(그래프) 레퍼런스에 효과적으로 매핑하고 분석이 어려운 영역에서 Illumina 리드의 매핑을 향상시킬 수 있습니다. 이 새로운 기능은 Illumina 리드의 적용 범위를 효과적으로 확장해 주고 이전에는 분석이 어려웠던 영역에서의 정확한 매핑 및 변이 검출을 지원합니다.

PrecisionFDA Truth Challenge V2는 GIAB 컨소시엄이 자체적인 벤치마크를 확대한 주요 영역인 Difficult-to-Map Regions에 초점을 맞추었습니다. 이 영역에 대한 정확한 쇼트 리드(short read) 매핑이 어렵기 때문에, 이 영역에서는 쇼트 리드 데이터로 고품질의 변이 검출을 수행하기가 힘들고 오류도 발생하기 쉽습니다.

일반적으로 variant caller는 가장 가능성이 높은 본래의 시퀀스(sequence, 염기서열)를 확인하기 위해 특정 유전좌위(locus)에 매핑된 리드의 파일업(pileup)을 분석합니다. 다음의 영역에서는 매핑이 어려울 수 있습니다.

- 매우 다형적(polymorphic)이고 샘플 리드가 참조 유전체(reference genome)와 크게 다른 영역
- 매우 반복적(repetitive)이거나 부분 중복이 있어 샘플 리드는 상당히 매칭이 잘 되지만 특이도(specificity)가 낮은 영역

인구 집단 데이터의 경우에는 해당 인구 집단에서 관찰되는 교대(alternate) 시퀀스 콘텐츠가 다양한 갈라지고 합쳐지는 패스(path)로 표현되는데, Multigenome(그래프) 레퍼런스는 바로 이러한 데이터의 매핑에 유용합니다(그림 5). 샘플 리드는 참조 다중유전체(즉, multigenome reference)에 걸쳐 가장 잘 매칭이 되는 패스에 정렬될 수 있습니다.

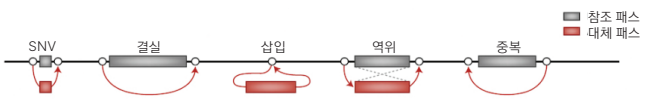

그림 5: Multigenome(그래프) 레퍼런스 예시 — 레퍼런스는 특정 인구 집단에서 관찰되는 교대 시퀀스 콘텐츠를 다양한 갈라지고 합쳐지는 패스로 표현함.

DRAGEN 플랫폼은 리드 매핑의 베이스라인으로 선형(linear) 레퍼런스를 사용하며, 다음과 같이 이러한 선형 레퍼런스를 효과적인 그래프 레퍼런스로 만들어주는 두 가지 기능을 제공하여 분석이 어려운 영역의 정확한 리드 매핑을 지원합니다.

- 선형 레퍼런스상 특정 인구 집단 내 알려진 SNV를 위한 다염기(multibase) IUPAC-IUB code 지원
- 특정 그래프 내 대체 패스인 ALT 콘티그(각각 선형 레퍼런스로 사전 정의된 리프트오버(lifeover) 정렬)를 위한 고급 ALT-aware 기능 지원

상세 정보: illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html

ALT 마스킹

DRAGEN 소프트웨어는 DRAGEN v3.9 소프트웨어 업데이트부터 원래의 참조 ALT 콘티그를 처리하는 새로운 방법인 ALT 마스킹 기능을 제공하고 있습니다. ALT 마스킹은 정확도 향상을 위해 ALT 콘티그의 전략적 위치를 감추는 기능으로, 시간이 지나도 정의, 유지 및 개선이 쉽습니다.

상세 정보: illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html

머신 러닝

DRAGEN v3.9 소프트웨어의 생식세포 작은 변이 검출 워크플로우에 강력하고 효율적인 머신 러닝 재보정(machine learning recalibration) 파이프라인이 옵션으로 추가되었습니다. 이 파이프라인은 DRAGEN v4.0 소프트웨어에서 기본적으로 설정되어 있으며, 일반적인 변이 검출 작업이 완료되면 머신 러닝 모델을 실행합니다. 이 단계에서는 최종 VCF 파일에 포함되는 QUAL 및 GQ 필드가 재보정됩니다. 경우에 따라서는 머신 러닝 모델이 GT 필드를 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 수행 시 표준 워크플로우를 약 5분 연장하므로 이 단계가 주는 정확도 향상이 전체 런 타임에 미치는 영향은 제한적이라 할 수 있습니다.

머신 러닝 모델은 오프라인 지도 학습을 통해 생성됩니다. 이 모델은 일련의 리드 기반 기능 및 맥락적(contextual) 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.

상세 연구 방법

입력 데이터 세트

본 분석에는 PrecisionFDA Truth Challenge V2에 사용된 세 가지 시퀀싱 기술을 이용해 제출된 9개의 기존 데이터 세트와 세 명의 개인이 활용되었습니다. 모든 데이터 세트는 PrecisionFDA 웹사이트를 통해 공개적으로 이용 가능합니다.

F1 점수 계산 및 벤치마크 설명

V4.2 벤치마크 진리 집합에 포함된 HG003 및 Hg004 샘플의 벤치마크값은 Wittyer를 사용하여 계산되었습니다. 최종 제출 결과의 평가에는 HG003 및 HG004 샘플의 SNV 및 INDEL F1 점수 합의 기하평균(geometric mean)이 사용되었습니다. 구체적인 계산 방법은 다음과 같습니다.

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

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

DRAGEN 커맨드 라인

```
/opt/edico/bin/dragen
        --output-directory=/output_folder/
        --events-log-file=/output_folder/events_log.csv
        --output-file-prefix=prefix_string
pipeline generated files
        --fastq-file1=/data_folder/fastq_file_1.fastq.gz
        --fastq-file2=/data_folder/fastq_file_1.fastq.gz
        --RGID=DRAGEN_RGID
        --RGSM=read_group_sample_name
        --ref-dir=/reference_genomes/ref_genome
//Initiate DRAGEN
//Specifies the output directory
//Specify log file location
//Outputs file name prefix for all
//Input FASTQ file 1
//Input FASTQ file 2
//Specifies read group ID
//Specifies read group sample name
//Specifies the directory containing the reference hash table
--enablehttp-server=true
--enable-metrics-json=true
--generate-sa-tags=true
--vc-enable-profile-stats=true
--enable-vcf-compression=true
--enable-save-bed-file=true
--enable-variant-caller=true
--enable-map-align=true
--enable-map-align-output=true
--enable-sort=true
--enable-duplicate-marking=true
```

상세 정보

DRAGEN Bio-IT Platform에 대한 자세한 정보는 sapac.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/dragen-bio-it-data-sheet-m-gl-00680/dragen-bio-it-data-sheet-m-gl-00680-kor.pdf에서 확인하실 수 있습니다.

illumina®

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

M-KR-00121 KOR