DATA SHEET

DRAGEN™ Secondary Analysis

포괄적인 NGS 데이터 분석으로 정확하고 효율적인 변이 검출

illumina®

연구 전용입니다. 진단 절차에는 사용할 수 없습니다.

문서 번호: M-KR-00109 v3.0 KOR
원본 문서: M-GL-00680 v9.0

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 유전적인 정보를 최대한 활용하려면 정확하고 효율적으로 시퀀싱 raw data를 의미 있는 결과로 해석할 수 있는 데이터 분석 도구가 필요합니다. 또한 기관에서 NGS의 이점을 십분 활용하기 위해서는 다양한 사용자를 수용하면서 경제적 부담은 적고 기술적으로 도입이 쉬워 사용이 용이한 솔루션이 필요합니다.

Illumina DRAGEN(Dynamic Read Analysis for GENomics) Secondary Analysis는 전장 유전체(whole-genome), 엑솜(exome), 전사체(transcriptome), 메틸롬(methylome) 등의 연구를 포함한 광범위한 응용 분야에서 NGS 데이터 분석 시 주로 발생하는 불편함을 해소하기 위해 개발된 제품입니다. DRAGEN Secondary Analysis 소프트웨어는 NGS 데이터를 처리하고 더 심층적인 정보를 위한 3차 분석의 기반을 만들어 주는 다양한 앱을 제공합니다. 또한 다양한 도구로 구성된 매우 정확하고 포괄적이며 효율적인 솔루션을 지원하므로 랩에서 규모나 연구 분야와 관계없이 유전체 데이터를 한층 더 효과적으로 활용할 수 있습니다.

정확한 결과

DRAGEN Secondary Analysis는 매우 정확한 결과를 제공합니다. DRAGEN Secondary Analysis v3.7은 2020년 precisionFDA Truth Challenge V2(precisionFDA V2)에서 전체 벤치마크 영역(all benchmark regions) 및 매핑이 어려운 영역(difficult to map regions) 부문에서 우승을 차지하며 가장 정확한 Illumina 시퀀싱 데이터 분석 결과를 보여주었습니다.[1,2] DRAGEN은 머신 러닝(machine learning, ML) 및 DRAGEN Multigenome(그래프) 기술 등의 발전을 통해 이후 출시된 버전에서도 계속해서 정확성의 새로운 기준을 제시하고 있습니다. 최신 버전인 DRAGEN Secondary Analysis v4.3은 전체 벤치마크 영역에서 F1 점수(정밀도(precision) 및 재현율(recall) 통합 측정값) 99.89%를 달성하여 월등히 우수한 작은 변이 검출 정확도를 확인할 수 있었습니다(그림 1). 이러한 정확도는 Human Pangenome Reference Consortium(HPRC) 데이터에서 얻은 256개의 하플로타입(haplotype)이 있는 128개의 샘플을 기반으로 보다 높은 유전적 다양성을 보여주는 3세대 DRAGEN Multigenome(그래프) 레퍼런스를 통해 확보할 수 있었습니다. 또한 대립유전자의 빈도(allele frequency)가 최저 3%인 모자이크 변이를 검출할 수 있는 새로운 통합 mosaic caller도 정확도의 향상에 기여했습니다.

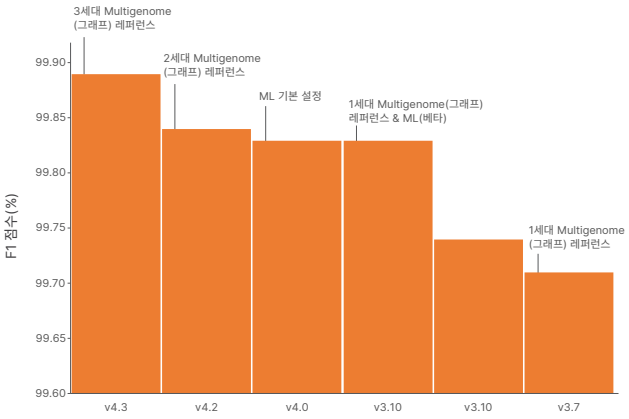

그림 1: DRAGEN Secondary Analysis의 정확성 — Y축의 F1 점수(%)는 진양성(true positive) 및 진음성(true negative) 결과를 전체 결과에 대한 비율로 계산한 값을 나타냄.[3,4]

포괄적인 분석

DRAGEN Secondary Analysis는 포괄적인 유전체 커버리지(coverage)와 광범위한 앱을 지원하여 NGS 연구를 수행하는 여러 랩의 다양한 요구 사항을 충족합니다. DRAGEN 파이프라인은 전장 유전체 시퀀싱(whole-genome sequencing, WGS), 전장 엑솜 시퀀싱(whole-exome sequencing, WES), 인리치먼트 패널(enrichment panel), 단일세포 RNA-Seq(single-cell RNA-Seq), 단일세포 ATAC-Seq(assay for transposase-accessible chromatin with sequencing), 벌크 RNA-Seq(bulk RNA-Seq), 메틸화 분석(methylation analysis) 등 여러 유형의 실험을 지원합니다(표 1). DRAGEN 소프트웨어의 다양한 기능 중 일부만을 재현한다 해도 30개가 넘는 오픈 소스 도구가 필요합니다.[3,4]

DRAGEN Secondary Analysis는 생식세포(germline) 분석을 위한 ExpansionHunter와 같은 다양한 variant caller와 *SMN*, *GBA*, *CYP2B6*, *CYP2D6*, *HLA* 등을 표적 검출하는 targeted caller를 포함하고 있습니다. DRAGEN v4.3은 새로운 specialized caller인 MRJD도 추가하여 *PMS2*, *SMN1*, *SMN2*, *STRC*, *NEB*, *TTN*, *IKBKG* 등 분절 중복(segmental duplication) 영역의 분석이 어려운 유전자에 대한 커버리지를 지원합니다. 이러한 도구는 확장된 유전체 영역에서 단일 염기서열 변이(single nucleotide variant, SNV), 삽입/결실(insertion/deletion, Indel), 반복 확장(repeat expansion), 구조적 변이 등 광범위한 유전자 변이의 분석에 사용할 수 있습니다. 또한 DRAGEN Multigenome(그래프) 레퍼런스는 매핑(mapping) 품질을 강화하여 변이 검출 정확도를 높여주고 시퀀스 복잡성으로 인해 연구가 어려운 유전체 영역도 분석할 수 있도록 해 줍니다. 이로써 의학적으로 잠재적인 연관성이 있는 유전자의 커버리지가 향상되고, 매핑이 어려운 영역에서 SNV, 작은 Indel, 유전자 복제수 변이(copy number variation, CNV), 구조적 변이(structural variant, SV)를 검출할 수 있습니다.

표 1: DRAGEN Secondary Analysis가 지원하는 광범위한 연구에 활용 가능한 2차 분석 앱[a]

연구용 앱	On-premise Server	Illumina 시퀀싱 시스템 내		Illumina 클라우드 플랫폼	
	DRAGEN Server	NovaSeq X 시리즈	NextSeq 1000 & NextSeq 2000 시스템	BaseSpace Sequence Hub	Illumina Connected Analytics
BCL Convert	✓	✓	✓	✓	커스텀만 해당
DRAGEN ORA compression	✓	✓	✓		커스텀만 해당
DRAGEN FASTQ + MultiQC	✓	✓	✓	✓	✓
Whole genome	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic
Enrichment(엑솜 포함)	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic
DRAGEN Amplicon	✓		DNA만 지원	✓	✓
RNA	✓	✓	✓	✓	✓
Single-cell RNA	✓		✓	✓	✓
NanoString GeoMx NGS			✓	✓	
Methylation	✓	✓		✓	✓
Metagenomics	✓[b]			✓	
RNA pathogen detection				✓	
COVID	COVIDSeq. COVID lineage		COVIDSeq(클라우드만 지원)	COVIDSeq. COVID lineage	
TruSight Oncology 500 portfolio	✓			✓[c]	✓
scATAC-Seq	✓			✓	✓
Imputation	✓			✓	✓
PGx Star Allele Caller	✓	✓	✓	✓	✓
Illumina Complete Long Reads				✓	✓
DRAGEN secondary analysis for RPIP and UPIP	✓			✓	베타

a. 요구되는 DRAGEN 소프트웨어 버전은 플랫폼마다 상이하므로 자세한 정보는 지역 담당자에게 문의.
b. Kmer 분류기를 기반으로 하는 메타유전체학 연구 앱으로, 앞으로 더 많은 도구 지원 예정.
c. Illumina Connected Analytics 구독 필요.

효율적인 분석

DRAGEN 소프트웨어는 랩이 NGS 데이터 세트 프로세싱의 효율성을 최적화하기 위해 확보해야 하는 데이터 분석 속도를 지원하도록 설계되었습니다. DRAGEN Secondary Analysis는 빠른 턴어라운드 시간을 달성하기 위해 하드웨어 가속(hardware acceleration) 및 필드 프로그래밍 가능 게이트 어레이(field programmable gate array, FPGA) 아키텍처를 이용하였습니다. DRAGEN 분석 알고리즘의 효율성은 두 가지 유전체 데이터 분석 세계 기록 수립에도 기여하였습니다.[5,6] 실제 적용 시 랩 내에서 직접 실행 가능한 온프레미스(on-premise) DRAGEN Secondary Analysis는 모든 caller에서 40× 커버리지의 전장 유전체에 대한 NGS 데이터를 약 35분 안에 처리할 수 있는 반면, 제한된 수의 변이형을 검출하는 일반적인 오픈 소스 방식으로는 이러한 작업에 8시간이 넘는 시간이 소요됩니다.[7]

아울러 대용량 NGS 데이터 파일을 보다 쉽게 저장, 관리 및 공유할 수 있도록 DRAGEN Original Read Archive(ORA) 기술을 적용하여 기존 fastq.gz 형식의 FASTQ 파일을 최대 1/5 크기로 무손실 압축(lossless compression)합니다. DRAGEN ORA는 무손실 압축 시 FASTQ 파일의 세부 정보는 그대로 유지하며 속도가 월등히 빨라 50~70 GB 크기의 FASTQ 파일을 약 8분 내에 압축하므로 일반적으로 연구되는 광범위한 종(species)을 지원할 수 있습니다. 또한 DRAGEN Secondary Analysis는 여러 단계에서 데이터 파일 입력을 지원하고 결과 파일을 생성하는 다양하게 활용이 가능한 파이프라인을 제공합니다(그림 2).

* DRAGEN v4.3에서 제공되는 MRJD, VNTR와 같은 새로운 specialized caller 없이 DRAGEN Server v4에서 HG001-HG007 표준 물질을 사용해 얻은 Illumina 내부 데이터에 근거함.

FPGA 및 하드웨어 가속

고도로 구성 가능한 FPGA는 베이스 콜(base call, BCL) 파일 변환, 매핑, 정렬(alignment), 분류(sorting), 중복 리드 표시(duplicate marking), 하플로타입 변이 검출(haplotype variant calling)과 같은 유전체 분석 알고리즘을 하드웨어 가속을 사용해 초고효율적으로 구현할 수 있도록 해 줍니다. Illumina는 FPGA의 유연성을 기반으로 다양한 DRAGEN 앱 파이프라인을 개발하였으며, 뛰어난 정확성, 포괄성 및 효율성을 제공하기 위해 꾸준히 파이프라인을 업데이트하고 추가하고 있습니다.

커스텀 레퍼런스

연구자는 DRAGEN Secondary Analysis를 이용해 인간(human), 비인간(nonhuman) 또는 비표준(nonstandard) 레퍼런스(reference, 참조 유전체)를 맞춤 생성할 수 있습니다. 이렇게 만든 레퍼런스는 커스텀 레퍼런스 파일을 지원하는 모든 DRAGEN 앱에 사용할 수 있습니다. 대부분의 DRAGEN 파이프라인은 hg19, hg38(HLA 포함 또는 제외), GRCh37, CHM13v2, hs37d5를 내장 지원합니다. 또 연구자는 DRAGEN 소프트웨어를 통해 다양한 집단과 특정 집단 모두에 표준 Multigenome(그래프) 레퍼런스 기능을 확대 적용해 볼 수 있습니다.

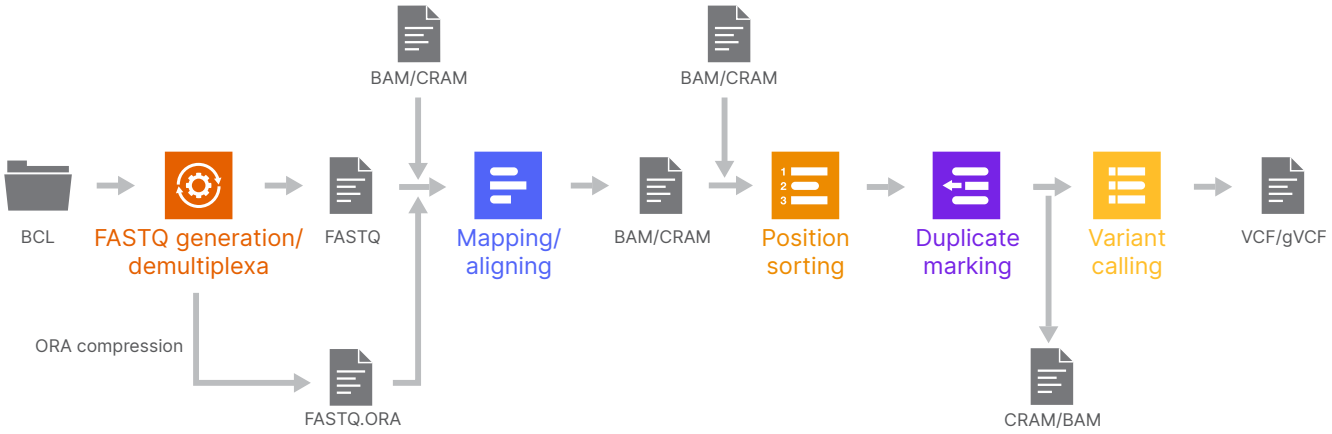

그림 2: DRAGEN Secondary Analysis 파이프라인의 유연성 — 각 DRAGEN 파이프라인은 정확하고 효율적인 분석에 필요한 구체적인 일련의 단계로 구성되어 있음. 다양한 파일 형식의 사용뿐 아니라 여러 가지 형식의 분석 결과 파일 생성도 지원하는 유연성을 갖추고 있는 DRAGEN Whole Genome Germline(예시) 파이프라인은 연구자에게 맞춤화된 경험을 제공하여 개개인이 원하는 형식의 파일을 생성 가능함.

a. BCL Convert는 독립 실행형 도구로도 사용 가능.

확장성

랩은 DRAGEN Secondary Analysis를 사용하여 적은 비용과 짧은 턴어라운드 시간을 유지하면서 작업 규모를 필요한 수준으로 확대할 수 있습니다. DRAGEN 소프트웨어는 다음을 통해 연구 역량의 강화에 기여합니다.

1. **NovaSeq™ X 시리즈, NextSeq™ 1000 및 NextSeq 2000 시스템 지원** — 기기에 내장되어 있는 온보드(onboard) DRAGEN은 1회의 런(run) 중 플로우 셀(flow cell)당 복수의 앱(BCL Convert 파이프라인 1개와 사용자가 선택한 파이프라인 3개까지 최대 4개의 앱 동시 사용)을 동시에 실행 가능합니다.

2. **버스트 용량(Burst Capacity)** — 샘플 수의 증가로 인해 작업량이 늘어난 경우, 랩에서는 Illumina Connected Analytics에서 DRAGEN Secondary Analysis를 이용하거나 BaseSpace Sequence Hub에서 DRAGEN 앱을 이용해 클라우드에서 추가 용량을 활용할 수 있습니다(그림 3).

3. **작업 규모 확대** — 하나의 DRAGEN 인스턴스(instance)로 다양한 DRAGEN 파이프라인과 지원되는 종류의 샘플을 사용할 수 있습니다. 포괄성 및 효율성을 갖춘 DRAGEN 소프트웨어를 이용하면 턴어라운드 시간이나 분석 결과 품질에 미치는 영향 없이 작업의 규모를 확대할 수 있습니다.

4. **유전체로의 연구 전환** — DRAGEN의 사전 구축된 파이프라인을 이용해 표적 패널에서 엑솜으로 또 유전체로 손쉽게 연구를 전환할 수 있습니다.

5. **대규모 집단 유전체학 이니셔티브** — DRAGEN Secondary Analysis는 대규모 코호트(cohort) 연구를 위한 간소화된 워크플로우를 제공하며, 높은 정확도로 유전자 변이를 검출하는 데에 함께 사용되는 다양한 파이프라인을 갖추고 있습니다. DRAGEN gVCF Genotyper는 수천 개에서 수백만 개의 gVCF(genomic variant call format) 파일을 취합하며 기존 배치(batch)를 다시 처리하지 않고 새로운 배치를 취합합니다. 또 ORA 압축으로 스토리지 비용을 절감해 줍니다.

6. **딥 시퀀싱 연구 지원** — DRAGEN Secondary Analysis는 우수한 효율성으로 유전체의 경우 300×, 엑솜의 경우 1000× 이상의 평균 커버리지를 위한 높은 뎁스의 시퀀싱 분석을 지원합니다. 이러한 딥 시퀀싱(deep sequencing) 역량은 종양학 연구 및 희귀 유전 질환 연구와 같은 연구 분야에 유용합니다.

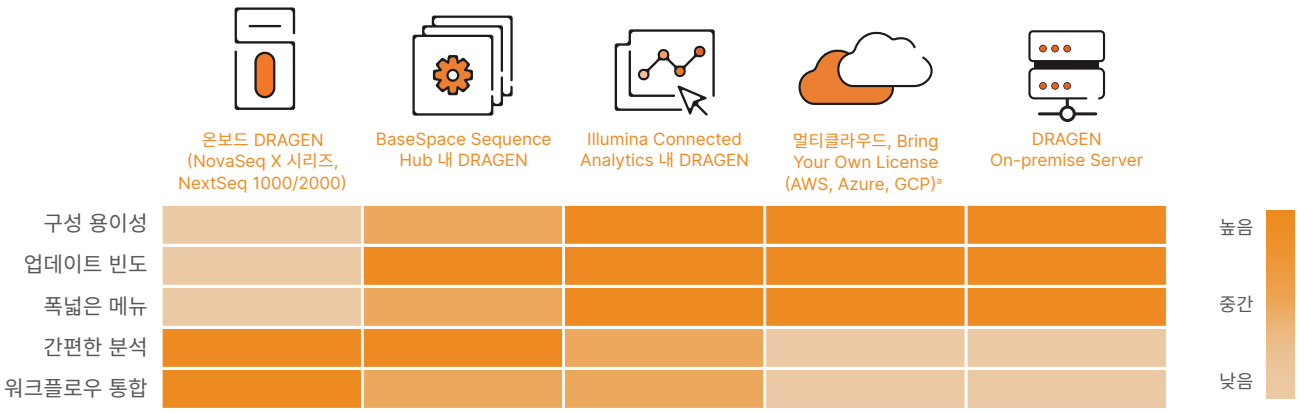

그림 3: 모든 랩의 NGS 분석 요건을 충족하는 기능을 제공하는 다양한 DRAGEN 파이프라인 액세스 옵션

a. Amazon Web Service(AWS), Azure 또는 Google Cloud Platform(GCP; 얼리 액세스)의 액세스 방법은 Illumina 담당자에게 문의.

멀티플랫폼 접근성

DRAGEN이 제공하는 다양한 파이프라인은 온프레미스, 클라우드 또는 기기에 내장된 솔루션을 통해 사용이 가능하므로 랩은 필요에 따라 가장 적합한 솔루션을 선택하면 됩니다(그림 3).

DRAGEN On-premise Server

DRAGEN On-premise Server는 로컬 스토리지 솔루션을 이용해 NGS 데이터를 수집하고 저장합니다. 시퀀싱이 완료되면 raw data는 로컬 네트워크 연결을 통해 시퀀싱 기기에서 로컬 스토리지로 전송된 후 다시 DRAGEN Server로 전송됩니다. 그 다음 DRAGEN Server는 연구자가 선택한 워크플로우를 실행합니다. 분석이 완료되면 소프트웨어는 생성된 분석 결과 파일을 지정된 로컬 스토리지 위치에 저장합니다. DRAGEN On-premise Server는 다음과 같은 장점을 제공합니다.

- 하나의 커맨드 라인 인터페이스(command-line interface, CLI)를 통해 DRAGEN 기능의 유연한 구성 지원
- 최대 30개의 기존 컴퓨트 인스턴스 대체
- 40× 커버리지의 인간 유전체 1개에 대한 NGS 데이터를 약 35분 내 처리

NovaSeq X 시리즈의 온보드 DRAGEN

NovaSeq X 시리즈는 NovaSeq X 시리즈가 생성하는 방대한 양의 데이터를 지원하도록 설계된 온보드 DRAGEN Secondary Analysis를 통해 정확하고 간소한 자동화된 분석을 제공합니다. 온보드 DRAGEN 소프트웨어 제품군은 BCL Convert, Germline, Somatic, RNA 및 Methylation을 포함하는 NGS 앱(표 1)을 통해 2차 분석 및 ORA 압축 기능을 제공합니다. 온보드 DRAGEN은 다음과 같은 장점을 가지고 있습니다.

- 복수의 2차 분석 파이프라인 동시 실행
- 1회의 런 중 플로우 셀당 최대 네 가지 앱 동시 실행 가능
- 최대 1/5 크기로 무손실 데이터 압축 및 스토리지 비용 절감
- 5년 이상 사용 시 절감되는 분석 비용이 NovaSeq X 시스템 구매 비용 능가

NextSeq 1000 & NextSeq 2000 시스템의 온보드 DRAGEN

NextSeq 1000 및 NextSeq 2000 시스템에는 신속하고 정확한 2차 분석을 제공하는 DRAGEN 소프트웨어가 내장되어 있습니다. DRAGEN 소프트웨어는 사용자 친화적인 그래픽 인터페이스를 갖추고 있어 전문가와 비전문가가 모두 신속하게 분석을 수행하고 분석 결과를 얻을 수 있습니다. 온보드 DRAGEN 소프트웨어는 흔히 사용되는 NGS 앱(표 1)을 포함하는 파이프라인을 선별해 제공하며, 수상을 통해 그 성능이 입증된 바 있는 ML 및 Multigenome(그래프) 레퍼런스 분석을 고품질 변이 검출에 적용하고 있습니다. 온보드 DRAGEN은 다음과 같은 장점을 제공합니다.

- 온보드 DRAGEN Secondary Analysis를 통해 벤치탑 시퀀싱 시스템 중 최고 수준의 정확도 확보
- 엄선된 DRAGEN 인포매틱스(informatics, 정보학) 파이프라인에 대한 액세스 제공
- 빠르면 2시간 내 분석 결과 도출 가능
- 직관적인 파이프라인 알고리즘의 사용으로 외부 인포매틱스 전문가에 대한 의존도 경감

BaseSpace Sequence Hub

BaseSpace Sequence Hub에서 이용 가능한 클라우드 기반의 DRAGEN 소프트웨어 제품군은 정확하고 효율적인 분석뿐만 아니라 안전한 생태계와 유연한 확장성도 제공합니다. 랩은 규모나 연구 분야에 상관없이 BaseSpace Sequence Hub에서 DRAGEN 소프트웨어를 이용해 간편하게 버튼 조작만으로 2차 분석을 수행할 수 있습니다. BaseSpace Sequence Hub는 Illumina 기기를 확장된 환경에서 사용할 수 있도록 해 줍니다. 기기에서 BaseSpace Sequence Hub로 암호화된 데이터가 전송되므로 연구자가 큐레이션(curation)을 거친 다양한 앱을 실행해 손쉽게 데이터를 관리하고 분석할 수 있습니다. Amazon Web Services(AWS) 기반의 BaseSpace Sequence Hub는 다음과 같은 장점을 가지고 있습니다.

- 간편하게 버튼 조작만으로 DRAGEN 분석을 실행하는 솔루션 제공
- 전문가와 비전문가가 모두 효율적으로 사용할 수 있도록 직관적인 그래픽 사용자 인터페이스(graphical user interface, GUI) 적용
- 추가 인프라에 투자하지 않고도 고성능 컴퓨팅 리소스 사용 가능

Illumina Connected Analytics

Illumina Connected Analytics는 종합적인 클라우드 기반의 바이오인포매틱스(bioinformatics, 생명정보학) 플랫폼입니다. 연구자는 이를 이용해 안전하고 확장 가능하며 유연한 환경에서 방대한 양의 멀티오믹스 데이터를 관리, 분석 및 해석할 수 있습니다. 연구자는 Illumina Connected Analytics에서 사전 패키지(prepackaged)된 파이프라인 또는 커스텀 파이프라인에 통합할 개별 도구로 DRAGEN Secondary Analysis 제품군을 사용할 수 있습니다.

요약

DRAGEN Secondary Analysis는 정확하고 포괄적이며 효율적인 NGS 데이터 분석에 사용되는 다양한 소프트웨어 도구를 하나의 패키지로 묶어 제공하는 제품입니다. 랩은 제공되는 DRAGEN 소프트웨어 배포 옵션 중 실제 프로젝트의 유형과 규모에 가장 적합한 솔루션을 선택해 사용하거나, 프로젝트에 요구되는 성능과 워크플로우에 가장 알맞은 몇 가지 배포 옵션을 결합해 사용할 수 있습니다. DRAGEN Secondary Analysis는 지속적인 NGS 기술의 발전에 발맞춰 기존 파이프라인이 최상의 성능을 유지할 수 있도록 신속한 업데이트를 제공하고 있으며, 앱이 개발될 때마다 새로운 파이프라인을 계속해서 추가하고 있습니다.

상세 정보

DRAGEN Secondary Analysis 지원 페이지

문의하기

참고 문헌

1. The food and Drug Administration. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. https://precision.fda.gov/challenges/10/results. Accessed March 14, 2022.
2. Catreux S, Jain V, Murray L, et al. DRAGEN Sets New Standard for Data Accuracy in PrecisionFDA Benchmark Data. Optimizing Variant Calling Performance with Illumina Machine Learning and DRAGEN Graph. Illumina website. illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. Published January 12, 2022 Accessed March 14, 2022.
3. Mehio R, Ruehle M, Catreux S, et al. DRAGEN Wins at PrecisionFDA Truth Challenge V2 Showcase Accuracy Gains from Alt-aware Mapping and Graph Reference Genomes. Illumina website. illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html. Published November 9, 2020. Accessed March 14, 2022.
4. Internal data on file. Illumina, Inc., 2022.
5. BioIT World. Children's Hospital Of Philadelphia, Edico Set World Record For Secondary Analysis Speed. bio-itworld.com/news/2017/10/23/children-s-hospital-of-philadelphia-edico-set-world-record-for-secondary-analysis-speed. Accessed March 14, 2022.
6. San Diego Union Tribune. Rady Children's Institute sets Guinness world record. https://www.sandiegouniontribune.com/95899028-132.html. Published February 12, 2018. Accessed March 14, 2022.
7. Betschart RO, Thiéry A, Aguilera-Garcia D, et al. Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. *Sci Rep*. 2022;12(1):21502. Published 2022 Dec 13. doi:10.1038/s41598-022-26181-3
8. Gross A, Maciuca S, Cox A, et al. Accurate and Efficient Calling of Small and Large Variants from PopGen data sets Using the DRAGEN Bio-IT Platform. Illumina website. www.illumina.com/science/genomics-research/articles/popgen-variant-calling-with-dragen.html. Published May 24, 2021. Accessed March 14, 2022.

illumina®

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

M-KR-00109 v3.0 KOR