DATA SHEET

DRAGEN™ Secondary Analysis

포괄적인 NGS 데이터 분석으로 정확하고 효율적인 변이 검출

illumina®

문서 번호: M-KR-00109 v2.0 KOR
원본 문서: M-GL-00680 v6.0

소개

연구와 의료의 진보를 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 유전체, 엑솜(exome), 전사체(transcriptome), 메틸롬(methylome) 연구를 포함한 다양한 응용 분야에서 NGS 데이터 분석 시 주로 발생하는 불편함을 해소하기 위해 개발된 제품입니다. DRAGEN 플랫폼은 NGS 데이터를 처리하고 더 심층적인 정보를 위한 3차 분석을 가능케 해 주는 2차 분석 소프트웨어 제품군입니다. 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 v3.7은 2020년 PrecisionFDA Truth Challenge V2(이하 PrecisionFDA V2)에서 All Benchmark Regions(전체 벤치마크 영역) 및 Difficult-to-Map Regions (매핑이 어려운 영역) 부문에서 우승을 차지하며 가장 정확한 Illumina 시퀀싱 데이터 분석 결과를 보여주었습니다.[1,2] 또한 Graph Genomes와 Illumina Machine Learning(ML)의 혁신적인 기술이 결합된 DRAGEN 4.0 소프트웨어는 All Benchmark Regions 부문에서 F1 점수(정밀도 및 재현율 통합 측정치) 99.83%를 달성하여 모든 시퀀싱 기술에 걸쳐 우수한 데이터 정확도를 보여줍니다(그림 1).[1,2] 또한 DRAGEN 4.0과 Graph를 결합한 분석 방법(ML 기본 설정)은 주조직 적합성 복합체(major histocompatibility complex, MHC) 영역에서 가장 정확한 검출 결과를 제공하여 모든 PrecisionFDA V2 제출 건 중 F1 점수가 가장 높습니다.

포괄적인 분석

DRAGEN Secondary Analysis는 하나의 플랫폼으로 다양한 실험에 대한 포괄적인 커버리지(coverage)를 제공하여 광범위한 NGS 연구를 수행하는 랩의 요구 사항을 충족합니다. DRAGEN 파이프라인은 전장 유전체 시퀀싱(whole-genome

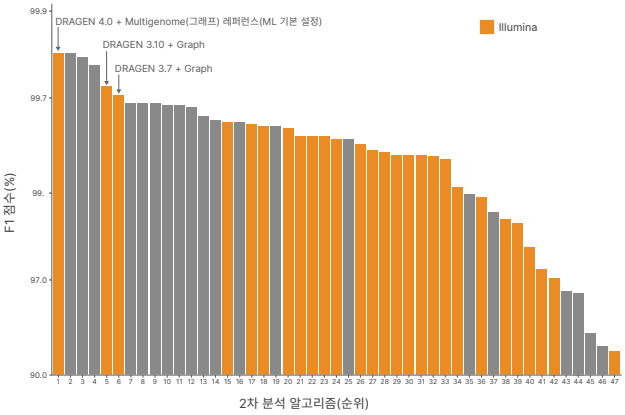

그림 1: PrecisionFDA Truth Challenge V2의 All Benchmark Regions 부문 제출 데이터 세트 대 DRAGEN 4.0 + Graph(ML 기본 설정)의 정확성 비교 — DRAGEN 4.0과 Graph를 결합한 분석 방법(ML 기본 설정)이 우수한 정확도를 보임. DRAGEN 3.10과 Graph를 결합한 분석 방법은 그래프 및 레퍼런스/ALT 콘티그 처리(reference/alt-contig handling) 능력을 향상하여 DRAGEN 3.7과 Graph를 결합한 분석 방법보다 우수한 성능을 보임. Y축의 F1 점수(%)는 진양성(true positive) 및 진음성(true negative) 결과를 전체 결과에 대한 비율로 계산한 값을 나타냄.[3,4]

sequencing, WGS), 인리치먼트 패널(enrichment panel), 단일세포 RNA-Seq(single-cell RNA-Seq), 단일세포 ATAC-Seq(assay for transposase-accessible chromatin with sequencing), 단일세포 멀티오믹스(multiomics) 연구, 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 소프트웨어는 변이 검출에 사용되는 다양한 variant caller(예: 반복 서열 확장(repeat expansion), 구조적 변이(structural variation, SV), 유전자 복제수 변이(copy number variation, CNV), ExpansionHunter 등의 variant caller와 *SMN*, *GBA*, *CYP2B6*, *CYP2D6*, *HLA* 등을 표적 검출하는 targeted caller 제공)를 포함하고 있어 폭넓은 유전체 커버리지를 지원합니다. 또한 DRAGEN Multigenome(그래프) 레퍼런스는 Illumina의 리드(read) 길이를 효과적으로 연장하고 복잡성이 낮은 영역(low-complexity region)에도 접근이 가능하므로 반복 시퀀스(repeat sequence)로 인해 분석이 어려운 유전체 영역도 분석할 수 있습니다. 이로써 의학적으로 잠재적인 연관성이 있는 유전자의 커버리지가 향상되고, Difficult-to-Map Regions에서 단일 염기서열 변이(single nucleotide variant, SNV), CNV 및 SV를 검출할 수 있습니다.

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

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 conversion	✓	✓	✓	✓	지원(커스텀만 해당)
DRAGEN ORA compression	✓	✓	✓		지원(커스텀만 해당)
DRAGEN FASTQ + MultiQC	✓	✓	✓	✓	✓
Whole genome	Germline + somatic	Germline만 지원 (Somatic은 곧 지원 예정)	Germline만 지원	Germline + somatic	Germline + somatic
Enrichment(엑솜 포함)	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic
DNA amplicon	✓		✓	✓	✓
RNA	✓	✓	✓	✓	✓
Single-cell RNA	✓		✓	✓	✓
Differential expression		✓	✓	✓	
NanoString GeoMx NGS			✓	✓	
RNA amplicon	✓			✓	곧 지원 예정
Methylation	✓	곧 지원 예정		✓	✓
Metagenomics				✓	
RNA pathogen detection				✓	
COVID	COVIDSeq. COVID lineage		COVIDSeq(클라우드만 지원)	COVIDSeq. COVID lineage	
TruSight Oncology 500	cfDNA 지원 (Solid 곧 지원 예정)			✓ 3.10에서 지원	✓
ScATAC-Seq	✓			✓	✓
Imputation	✓			✓	✓
PGx Star Allele Caller	✓			✓	✓
Illumina Complete Long Reads				✓	
DRAGEN secondary analysis for RPIP and UPIP	베타				베타

효율적인 분석

DRAGEN 소프트웨어는 랩이 NGS 데이터 세트를 최대한 활용하기 위해 확보해야 하는 데이터 분석 속도와 파일 옵션을 지원할 수 있도록 특별히 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 34× 커버리지의 전장 유전체에 대한 NGS 데이터를 약 30분 안에 처리할 수 있습니다. 일반적인 CPU 기반 시스템으로는 이러한 작업에 15시간이 넘는 시간이 소요됩니다.[7]

아울러 대용량 NGS 데이터 파일의 저장 공간을 절약할 수 있도록 DRAGEN Original Read Archive(ORA) 기술을 적용하여 FASTQ 파일을 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 압축할 수 있습니다. 또한 DRAGEN Secondary Analysis는 여러 단계에서 데이터 파일 입력을 지원하고 결과 파일을 생성하는 다양하게 활용이 가능한 파이프라인을 제공합니다(그림 2).

FPGA 및 하드웨어 가속

고도로 구성 가능한 FPGA는 베이스 콜(base call, BCL) 파일 변환, 매핑(mapping), 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.

맞춤형 레퍼런스

연구자는 DRAGEN Reference Builder를 이용해 해시 테이블(hash table)이라고 불리는 인간(human), 비인간(nonhuman) 또는 비표준(nonstandard) 레퍼런스(reference, 참조 유전체)를 만들 수 있습니다. 이렇게 만든 레퍼런스는 맞춤형 레퍼런스 파일을 지원하는 모든 DRAGEN 앱에 사용할 수 있습니다. BaseSpace™ Sequence Hub에서 DRAGEN Reference Builder를 실행하려면 FASTA 파일이 필요합니다. 대부분의 DRAGEN 파이프라인은 hg19, hg38(HLA 포함 또는 제외), GRCh37, hs37d5를 내장 지원합니다. 또 연구자는 DRAGEN Graph Toolkit을 이용해 그래프 레퍼런스 역량을 보다 다양한 인간 그래프 레퍼런스 파일에 확대 적용해 볼 수 있습니다.

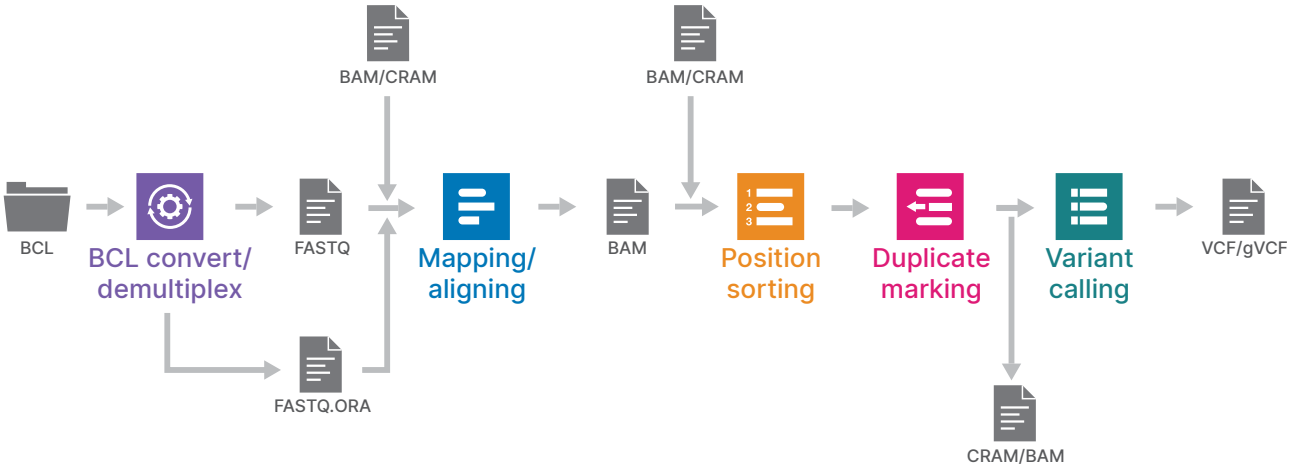

그림 2: DRAGEN 파이프라인의 유연성 — 각 DRAGEN 파이프라인은 정확하고 효율적인 분석에 필요한 구체적인 일련의 단계로 구성되어 있음. 다양한 파일 형식의 사용뿐 아니라 여러 가지 형식의 분석 결과 파일 생성도 지원하는 유연성을 갖추고 있는 DRAGEN 파이프라인은 연구자에게 맞춤화된 경험을 제공하여 개개인이 원하는 형식의 파일을 생성 가능함.

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 지원** — 기기에 내장되어 있는 온보드(onboard) DRAGEN은 1회의 런(run) 중 플로우 셀(flow cell)당 최대 네 가지 앱을 동시에 실행 가능합니다.

2. **버스트 용량(Burst Capacity)** — 샘플 수의 증가로 인해 작업량이 늘어난 경우, Illumina에서 제공하는 DRAGEN 소프트웨어 병렬 접근(parallel access) 옵션을 이용하여 추가 용량을 활용할 수 있습니다(그림 3).

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

4. **엑솜 시퀀싱에서 유전체 시퀀싱까지** — 전장 엑솜 시퀀싱(Whole-exome sequencing, WES)에서 WGS로 분석 방법을 바꾸면 데이터 생성량이 크게 증가합니다. DRAGEN 소프트웨어를 이용하면 하드웨어 인프라나 클라우드 기반의 솔루션을 추가하기 위해 큰 비용을 투자하지 않고도 엑솜에서 유전체로 범위를 확대해 분석을 수행할 수 있습니다.

5. **방대한 데이터 세트** — DRAGEN Secondary Analysis는 대규모 코호트(cohort) 분석을 위한 간소화된 워크플로우를 제공하며, 코호트 샘플링으로 크고 작은 변이를 정확하게 검출하는 데에 함께 사용되는 다양한 파이프라인을 갖추고 있습니다. DRAGEN 소프트웨어는 수천 개에서 수백만 개의 gVCF(genomic variant call format) 파일을 취합하고 지노타이핑(genotyping, 유전형 분석)을 수행하는 데 사용되며, 기존 배치(batch)를 다시 처리하지 않고 새로운 배치를 취합합니다. DRAGEN Joint Genotyping 파이프라인은 여러 유전체에서 변이를 한 번에 검출하는 조인트 콜링(joint calling)을 수행하고, 정확도 저하 없는 신속한 분석을 통해 대규모 코호트로 연구 규모를 확대할 수 있습니다.[8] DRAGEN Secondary Analysis로 1000 Genomes Project 데이터를 분석했을 때 다양한 샘플에 대한 정확한 대규모 변이 검출이 가능했으며, 커버리지 데이터가 균일하지 않거나 가정에서 벗어난 영역을 확인할 수 있었습니다.

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).

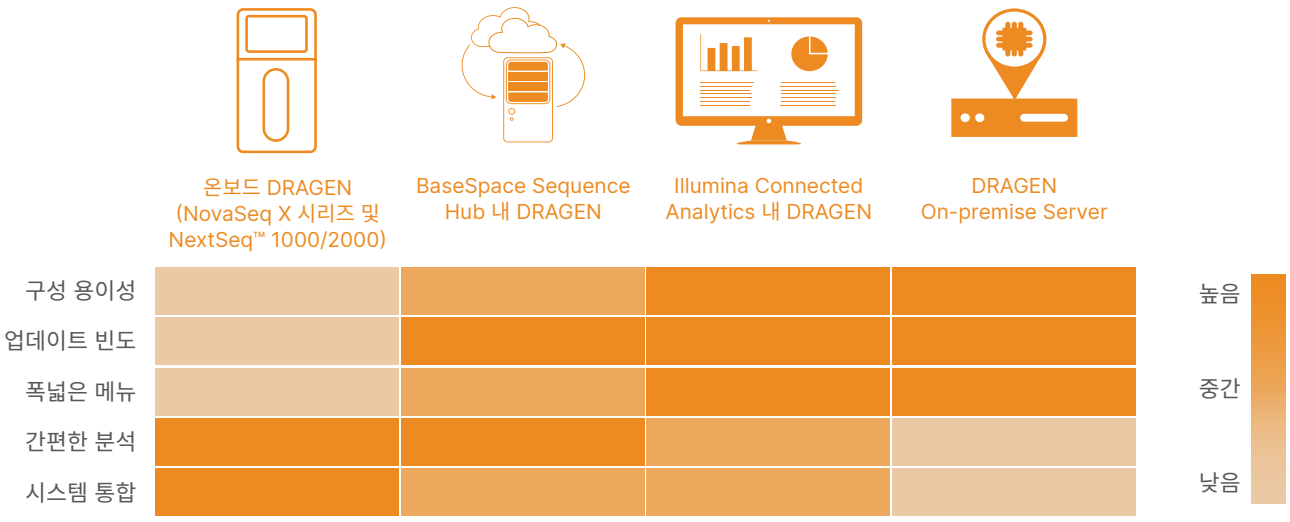

그림 3: 모든 랩의 NGS 분석 요건을 충족하는 기능을 제공하는 다양한 DRAGEN 파이프라인 접근 옵션

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.

- 다양한 커맨드 라인 인터페이스(command-line interface, CLI) 레벨 지원.
- 최대 30개의 기존 컴퓨트 인스턴스 대체.
- 34× 커버리지의 인간 유전체 1개에 대한 NGS 데이터를 약 30분 내 처리.

NovaSeq X 시리즈의 온보드 DRAGEN

NovaSeq X 시리즈에는 정확하고 포괄적이며 자동화된 2차 분석을 제공하는 가장 강력한 성능의 DRAGEN 소프트웨어가 내장되어 있습니다. 온보드 DRAGEN 소프트웨어 제품군은 BCL Convert, Germline, RNA 및 Enrichment를 포함하는 다양한 NGS 앱(표 1)을 통해 변이 검출 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.
- 지원되는 앱을 통해 최대 1/5 크기의 무손실 데이터 압축 및 분석 제공.
- 5년 이상 사용 시 절감되는 분석 비용이 NovaSeq X 시스템 구매 비용 능가.

NextSeq 1000 및 NextSeq 2000 시스템의 온보드 DRAGEN

NextSeq 1000 및 NextSeq 2000 시스템에는 신속하고 정확한 2차 분석을 제공하는 DRAGEN 소프트웨어가 내장되어 있습니다. 온보드 DRAGEN 소프트웨어는 가장 흔히 사용되는 NGS 앱(표 1)을 포함하는 파이프라인을 몇 가지 선별하여 제공하며, 전문가와 비전문가가 모두 신속하게 분석을 수행하고 분석 결과를 얻을 수 있도록 사용자 친화적인 인터페이스를 갖추고 있습니다.

온보드 DRAGEN은 다음과 같은 장점을 제공합니다.

- 엄선된 DRAGEN 인포매틱스(informatics) 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 Analytics에서 제공되는 DRAGEN Secondary Analysis는 종합적인 클라우드 기반의 바이오인포매틱스 플랫폼입니다. 연구자는 이를 이용해 안전하고 확장 가능하며 유연한 환경에서 방대한 양의 멀티오믹스 데이터를 관리, 분석 및 해석할 수 있습니다. 연구자는 Illumina Connected Analytics를 통해 다음과 같은 장점을 누릴 수 있습니다.

- DRAGEN 소프트웨어에 대한 완전한 액세스 제공. 사전 패키지(Prepackaged)된 파이프라인 또는 맞춤형 파이프라인에 사용할 개별 도구 선택 가능.
- 고속 대용량 분석을 요하는 최적화된 연구에 적합한 고도로 자동화된 워크플로우와 맞춤형 솔루션 지원.
- 건강 보험 양도 및 책임에 관한 법(Health Insurance Portability and Accountability Act, HIPAA)과 유럽 연합 일반 데이터 보호 규정(General Data Protection Regulation, GDPR) 원칙 준수를 위해 데이터 레지던시(data residency) 요구 사항을 충족하고, 통합 인증(single sign-on, SSO) 기능을 지원하며, 감사 로그(audit log)를 제공하여, 액세스가 제어되는 매우 안전한 환경에서 연구 가능.

요약

DRAGEN Secondary Analysis는 정확하고 포괄적이며 효율적인 NGS 데이터 분석에 사용되는 다양한 소프트웨어 도구를 하나의 패키지로 묶어 제공하는 제품입니다. 랩은 프로젝트의 유형과 규모에 따라 제공되는 DRAGEN 소프트웨어 옵션 중 가장 적합한 솔루션을 선택해 사용하면 됩니다. 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

Illumina Support Center의 DRAGEN Secondary Analysis 페이지

문의하기

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. Accessed March 14, 2022.
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. Accessed March 14, 2022.
3. 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 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. MIller NA, Farrow EG, Gibson M, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med*. 2015;7:100. doi: 10.1186/s13073-015-0221-8.
8. Illumina. Accurate and Efficient Calling of Small and Large Variants from PopGen Datasets Using the DRAGEN Bio-IT Platform. www.illumina.com/science/genomics-research/articles/popgen-variant-calling-with-dragen.html. Accessed March 14, 2022.

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

M-KR-00109 v2.0 KOR