TECHNICAL NOTE

DRAGEN™ Bio-IT Platform을 이용한 집단 유전학 데이터 처리

대규모 코호트 연구에 적합한 데이터 분석 및 변이 검출 방법

illumina®

소개

전장 엑솜 시퀀싱(whole-exome sequencing, WES)과 전장 유전체 시퀀싱(whole-genome sequencing, WGS) 및 후속 데이터 처리 비용이 계속해서 감소하면서, 전례 없는 규모의 집단 시퀀싱(population sequencing) 연구가 가능해지고 있습니다. 코호트(cohort) 수준의 변이(variation) 카탈로그는 가계(ancestry) 연구, 희귀 변이(rare variant)에 대한 통찰, 유전형(genotype)과 표현형(phenotype) 간의 연관성 발견 그리고 임상유전체학적 특징에 대한 어노테이션(annotation, 주석 처리)에도 활용되는 핵심적인 자원입니다. 때문에 높은 정확도의 코호트 콜 세트(cohort call set)를 확보하는 것이 중요하지만, 대량의 샘플로부터 데이터를 통합할 때 발생하는 정보학적, 분석적 문제는 여전히 해결되지 않고 있습니다.

집단 유전학 데이터 분석

일반적인 집단 유전학(population genetics, PopGen) 데이터 처리 워크플로우는 리드 매핑(read mapping) 및 변이 검출(variant calling) 단계에서 독립적으로 샘플을 분석하는 것으로 시작되며, 이 단계에서 검출된 변이는 gVCF 파일로 저장됩니다. 그다음 한 코호트 내 모든 샘플에 대한 gVCF 파일들이 취합되어 유전형과 연관된 신뢰도 메트릭스(confidence metrics)를 포함하는 하나의 개념적 매트릭스(matrix)가 만들어집니다(그림 1). 이 매트릭스는 multisample(다중 샘플) VCF(DRAGEN gVCF Genotyper), multisample gVCF(DRAGEN/Genome Analysis Toolkit(GATK) Combine gVCF) 또는 데이터베이스(GATK GenomicsDB, GLnexus RocksDB) 등의 형식으로 저장이 가능합니다. 어떤 형식이든 매트릭스의 목적은 전체 코호트에 걸친 지노타입 콜(genotype call)을 제공하는 변이 중심의 관점을 제시하는 것입니다. 이를 토대로 코호트의 정보를 이용하여 개별 샘플의 지노타입 콜을 향상시킬 수 있는데, 이러한 통계 모델을 조인트 지노타이핑(joint genotyping, 유전형 결합 분석)이라고 합니다. 단, 샘플 사이즈가 증가할수록 오류도 축적될 수 있기 때문에 주의해야 합니다.

조인트 지노타이핑이 실제 정확도에 어떠한 영향을 미치는지에 대한 정보는 부족한 실정입니다. 여기에는 여러 가지 이유가 있지만, 지금까지 조인트 지노타이핑 도구를 gVCF 취합 도구에서 분리하기가 쉽지 않았던 것을 한 가지로 이유로 들 수 있습니다. 대량의 샘플로부터 데이터를 취합할 때 코호트에 걸쳐 서로 다른 변이 표현을 통일하는 것은 특히 어렵습니다. 코호트 규모의 확대는 곧 다중 대립유전자 변이(multiallelic variant) 및 대체 대립유전자(alternative allele) 수의 증가를 의미하므로 반드시 gVCF의 전체 데이터 보존과 확장성 사이에서 균형을 찾아야 합니다. 더욱이 정립되어 있는 GATK의 데이터 처리 워크플로우들은 복잡하기 때문에 어려움은 더 가중됩니다.

DRAGEN Platform은 조인트 지노타이핑 전과 후 multisample VCF 형식의 파일을 생성하는 간소화된 코호트 분석 워크플로우를 제공하며(그림 1), 연구자는 이를 통해 조인트 지노타이핑 모델의 영향을 직접적으로 측정할 수 있습니다.

이 Technical Note는 대규모 집단 유전학 연구 프로젝트에서 흔히 DRAGEN Platform을 이용해 조인트 지노타이핑을 수행하는 세 가지 방법을 살펴보고 DRAGEN Platform의 성능을 평가합니다.

- 높은 커버리지(35x)의 WGS 샘플
- 낮은 커버리지(15×)의 WGS 샘플
- 높은 커버리지(50×)의 WES 샘플

먼저 1000 Genomes Project의 3단계 연구에서 수집한 샘플[1]을 최근 재시퀀싱(resequencing)한 데이터를 이용해 GATK로 생성한 콜 세트를 DRAGEN Platform으로 벤치마킹 비교(benchmarking comparison)한 결과를 제시합니다. 그다음 콜 세트의 정확도에 대한 각 워크플로우 단계의 기여도를 분석한 후, GATK Best Practices Workflow에 포함되어 있는 일부 방법들이 DRAGEN으로 생성한 데이터에는 유용하지 않을 것으로 예상되는 이유도 상세히 설명합니다. 마지막으로 즉시 분석이 가능한 변이를 획득하기 위해 DRAGEN Platform을 이용해 코호트 데이터를 처리하는 방법을 권장합니다.

방법

데이터 세트 입력

WGS 코호트 분석은 1000 Genomes Project의 코호트를 기반으로 했습니다.[2] 이 데이터 세트에는 NovaSeq™ 6000 시스템을 사용해 30×가 넘는 커버리지로 시퀀싱한 2,504개의 WGS 샘플이 포함되어 있습니다. 동일한 샘플을 GATK Workflow로 처리한 결과는 공개적으로 이용 가능하므로 결과를 재현해 볼 수 있습니다.[3,4] 한편 WES 코호트 분석은 CEPH 인구 집단(CEU)으로부터 얻은 여덟 개의 비혈연 관계 샘플과 Genome In a Bottle(GIAB) 컨소시엄으로부터 얻은 두 개의 트리오(trio, 부모와 자식) 샘플로 구성된 패널(총 10개의 샘플 포함)을 기반으로 했습니다.[5] 모든 샘플은 NovaSeq™ 6000 시스템으로 시퀀싱하였으며, 모든 분석에는 교대하는 콘티그(alternate contig)가 있는 인간 참조 유전체(reference genome)인 hg38이 사용되었습니다.

코호트 분석

WGS 분석의 경우, 코호트 연구에서 얻은 gVCF 파일들은 취합 후 DRAGEN Platform v3.5.7b를 이용해 조인트 지노타이핑을 진행하거나, GATK v3.5 Workflow에 따라 조인트 지노타이핑을 진행한 후 변이 품질 점수 재측정(variant quality score recalibration, VQSR)을 위한 처리 절차를 거쳤습니다(그림 1). 두 워크플로우 모두 염색체(chromosome)별로 multisample VCF(msVCF) 형식의 파일을 생성합니다.

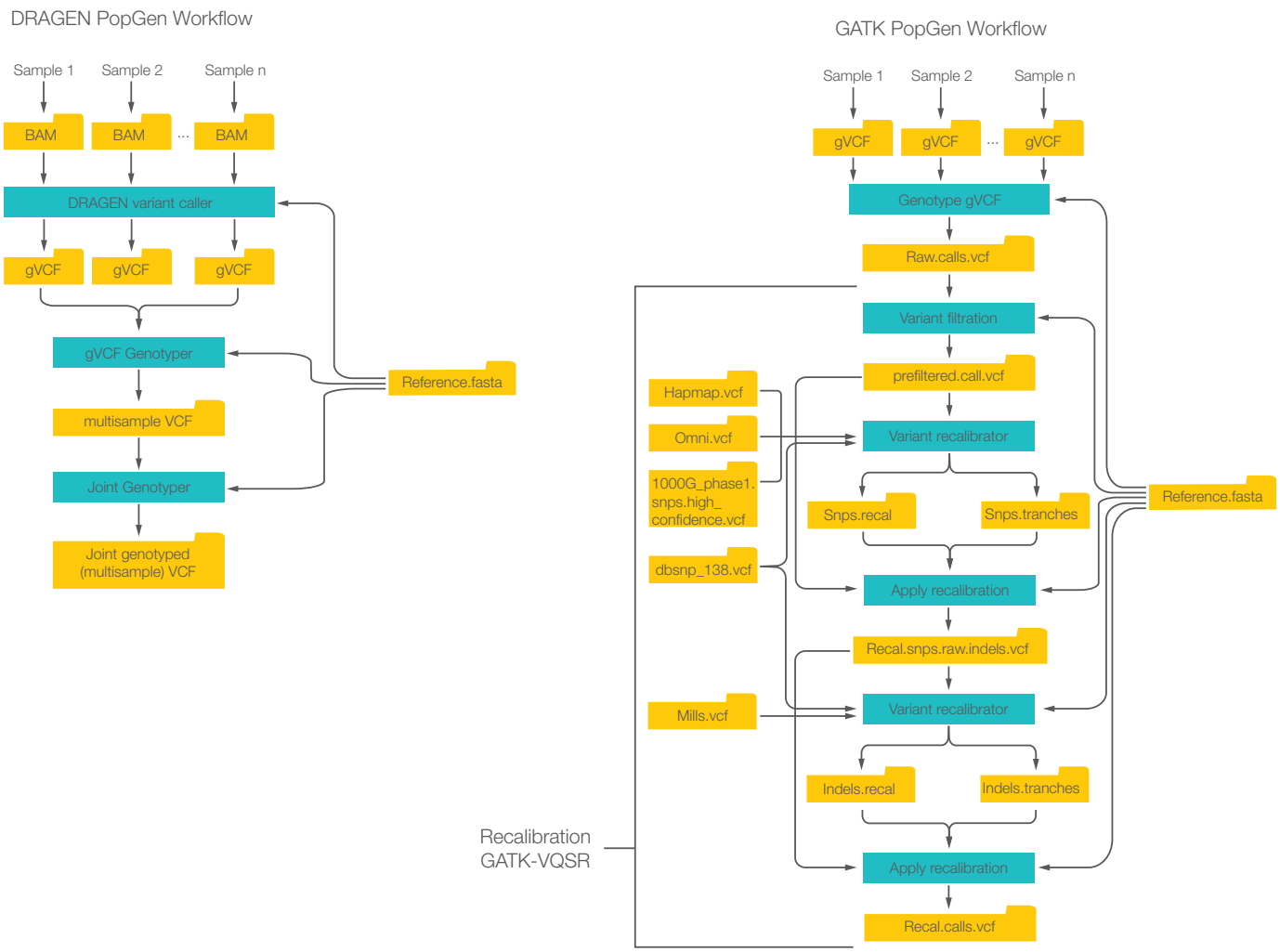


그림 1: DRAGEN Platform(좌)을 이용한 집단 유전학 데이터 처리 및 분석 워크플로우와 GATK Best Practices Workflow(우)[3] — gVCF Genotyper를 사용한 코호트에 걸친 gVCF 파일 취합 단계(DRAGEN-GG)와 Joint Genotyper를 사용한 조인트 지노타이핑 단계(DRAGEN-JG)의 두 단계로 구성되어 있는 DRAGEN PopGen Workflow. DRAGEN PopGen Workflow에는 재측정(recalibration) 단계가 포함되어 있지 않음.

높은 커버리지의 WGS

높은 커버리지의 WGS 샘플을 이용해 DRAGEN Platform의 성능을 확인하기 위해 DRAGEN Platform 콜 세트와 GATK 콜 세트의 정확도를 직접적으로 비교했습니다. DRAGEN Platform의 성능은 한 개의 특성이 확인된 샘플(NA12878)을 수신자 조작 특성(receiver operating characteristic, ROC) 메트릭스를 적용하고, 원래 코호트의 일부였던 GIAB 컨소시엄이 공개한 진리 변이(truth variant)들을 활용하여 측정했습니다. 연산 비용을 최소화하기 위해 17번 염색체만 분석에 포함했습니다.

ROC 곡선은 다양한 임계값에서의 위양성률(false positive rate) 대비 진양성률(true positive rate)을 그래프에 표시합니다. ROC 곡선 아래 면적은 변이 검출의 정확도를 나타내는 하나의 지표입니다.

결과

진리 샘플인 NA12878이 포함된 열(column)을 multisample VCF 파일에서 추출한 후 ROC 곡선을 적용하여 아래와 같이 총 네 개의 인구 집단 데이터 세트를 평가했습니다. 두 데이터 세트에는 GATK Workflow가 사용되었고, 나머지 두 데이터 세트에는 DRAGEN Platform이 사용되었습니다.

- 조인트 지노타이핑을 통과한 모든 변이(GATK-JG)*
- 조인트 지노타이핑을 통과하고 재측정 과정도 통과한 모든 변이(GATK-VQSR)
- gVCF Genotyper를 통과한 모든 변이(DRAGEN-GG)
- 조인트 지노타이핑 이후 모든 통과 변이(DRAGEN-JG)

전반적으로 워크플로우 구성에 상관없이 DRAGEN Platform이 GATK보다 우수한 성능을 보였으며, 특히 단일 샘플 SNP(single nucleotide polymorphism, 단일 염기 다형성)(그림 2A) 및 Indel(insertion/deletion, 삽입/결실)(그림 2B) 변이 검출 시 정확도가 뛰어난 것으로 확인되었습니다. 한 가지 예상치 못한 것은 조인트 지노타이핑 이후에 위양성 증가로 인해 DRAGEN Platform의 정확도가 감소하는 양상을 보였다는 점입니다(그림 2 및 그림 3). 종래의 조인트 콜링(joint calling) 방법은 DRAGEN single-sample (단일 샘플) gVCF 파일에 적용해도 아무런 이득을 가져다주지 못하며 불필요하게 높은 비용을 초래합니다. 그 이유는 DRAGEN Platform의 Genotyper가 PCR 유도 오류(PCR-induced error) 모델과 적체와 상관된 오류(pileup correlated error) 모델을 포함하고 있기 때문입니다.

Accuracy Improvements in Germline Small Variant Calling with the DRAGEN Platform Application Note 읽어보기

트리오에 대한 멘델리안 오류(Mendelian errors in trios)는 유전체에서 신뢰도가 높은 영역(high-confidence region) 내에 존재하는 변이에만 국한되지 않기 때문에 광범위한 정밀도 평가에 유용한 지표로 활용할 수 있습니다. 멘델리안 오류 수를 코호트의 트리오 중 최소 1명에게서 변이로 발견된 모든 부위의 수에 대해 평가[6]했을 때 이전 데이터와 일관된 결과를 확인할 수 있었습니다. 워크플로우와 상관없이 DRAGEN Platform이 더 높은 정확도를 보였으나, 조인트 지노타이핑 후에는 성능이 저하되는 것이 관찰되었습니다(표 1).

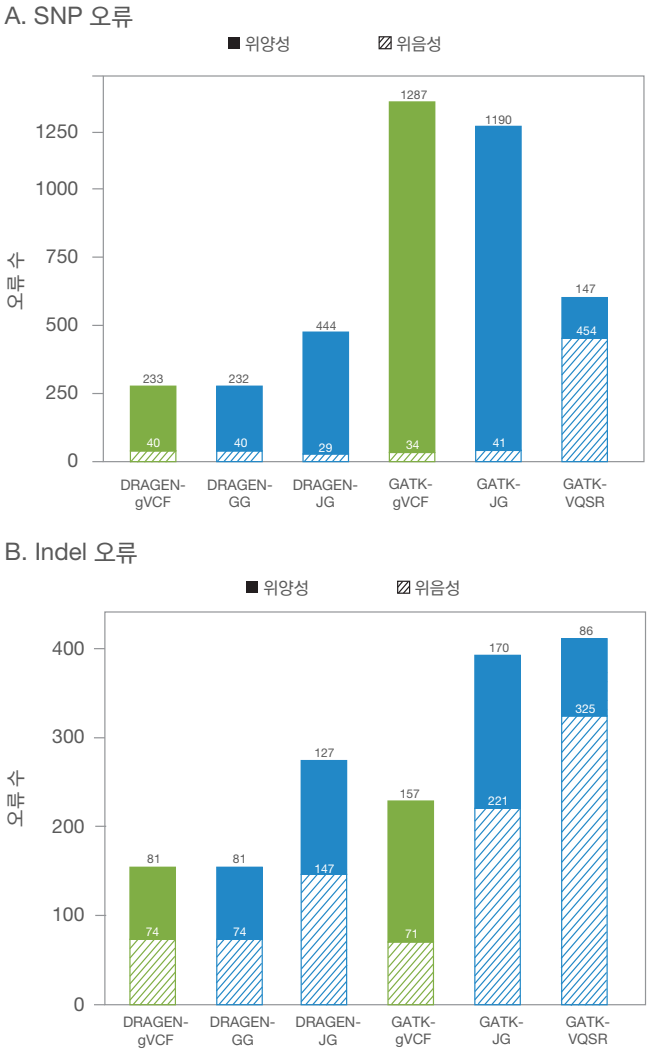


그림 2: 높은 커버리지의 WGS 데이터 세트에서 확인된 변이 검출 정확도 — DRAGEN Platform(GG, JG)과 GATK Workflow(JG, VQSR)를 이용해 집단 유전학 데이터 처리 후 single-sample gVCF 파일(초록색 막대)과 multisample VCF(파란색 막대) 파일에서 측정한 SNP(A)와 Indel(B)에 대한 위양성 및 위음성 변이 검출 결과를 표시한 그래프

* 조인트 지노타이핑 이전 GATK 취합 결과는 제공되지 않았음.

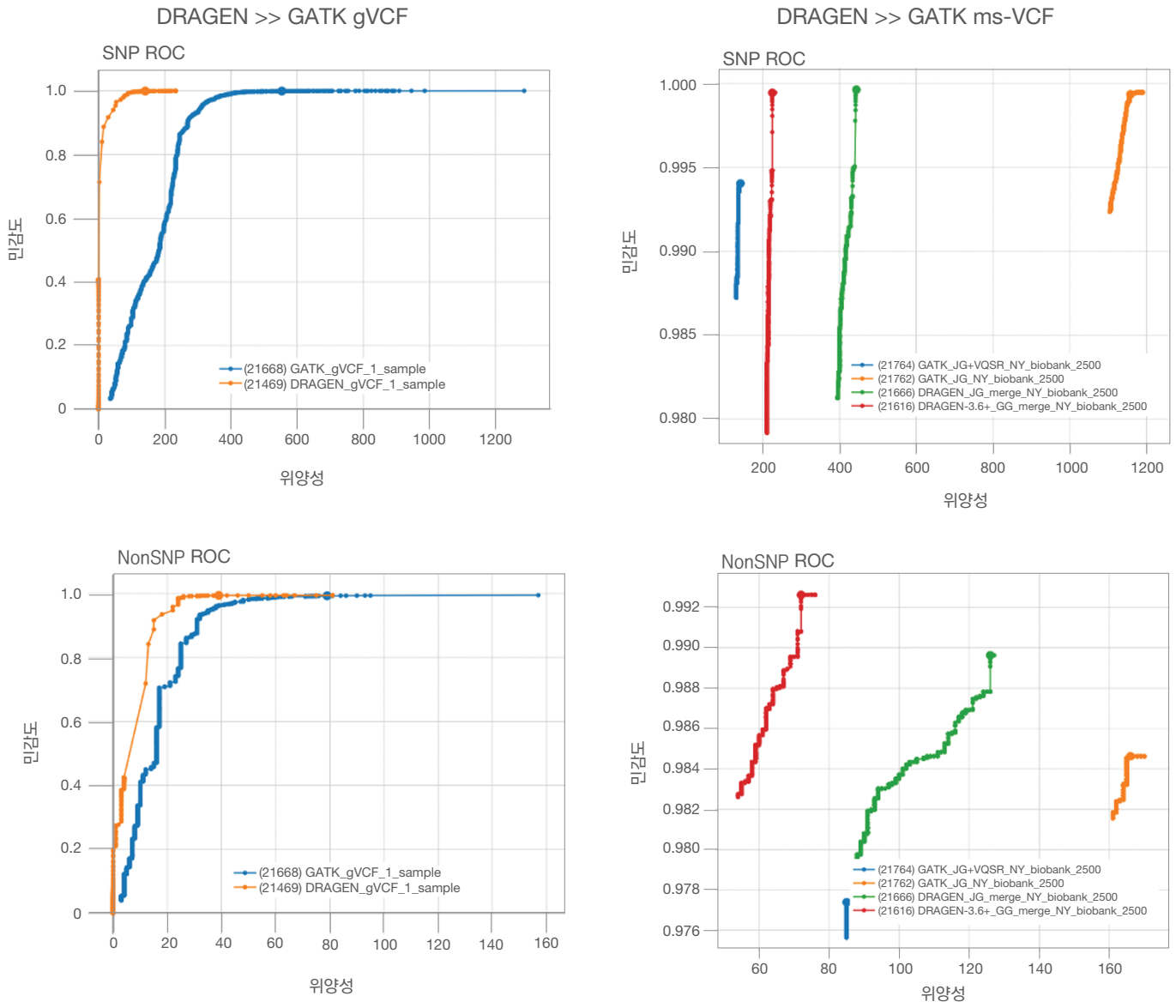


그림 3: 높은 커버리지의 WGS 샘플에 대한 코호트 분석 후 적용된 ROC 곡선 — 코호트 분석 워크플로우 후 생성된 single-sample gVCF(좌측 패널) 파일과 multisample VCF(우측 패널) 파일에 적용된 ROC 곡선

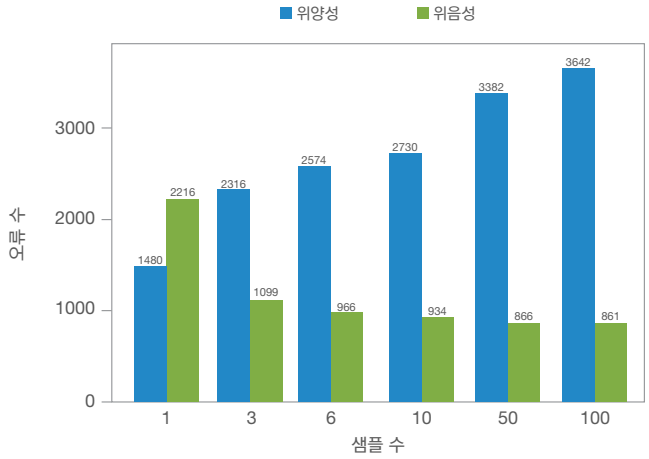
표 1: 높은 커버리지의 WGS 코호트에 존재하는 하나의 트리오에서 발견된 멘델리안 오류를 계산한 결과

멘델리안 오류	GATK Joint Genotyper	GATK VQSR	DRAGEN gVCF Genotyper	DRAGEN Joint Genotyper
신뢰 영역 내	1808/139,375 (1.30%)	833/133,195 (0.63%)	315/127,220 (0.25%)	385/127,667 (0.30%)
17번 전장 염색체	10,433/220,814 (4.72%)	5272/184,275 (2.86%)	4540/179,197 (2.53%)	5318/186,933 (2.84%)

샘플 사이즈가 코호트 분석에 미치는 영향

샘플 사이즈가 DRAGEN Platform의 조인트 지노타이핑 성능에 미치는 영향을 평가하기 위해 샘플의 수를 3개, 6개, 10개, 50개 그리고 100개로 늘려가며 유전체 전체에 걸친 정확도 메트릭스를 비교했습니다. 단일 샘플의 베이스라인 메트릭스와 비교했을 때, SNP의 위음성 수는 감소하고 위양성 수는 증가했으며(그림 4A), Indel은 두 메트릭스 모두 증가(그림 4B)했습니다. 앞서 설명한 바와 같이, 조인트 콜링 방법은 PCR 유도 오류 모델과 적체와 상관된 오류 모델을 포함하는 DRAGEN Platform의 single-sample gVCF 파일에는 이득을 가져다주지 못합니다.

A. SNP 오류



B. Indel 오류

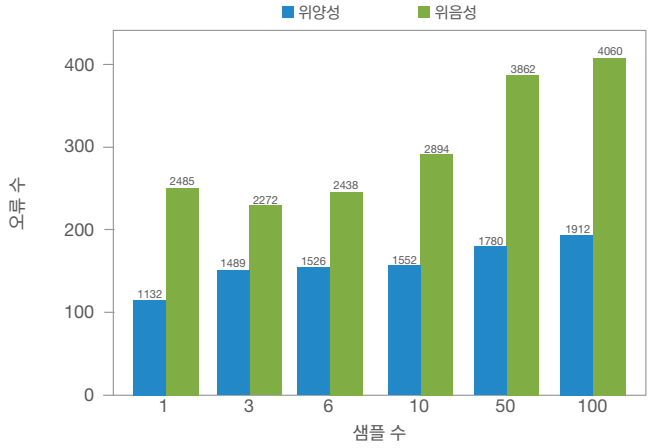


그림 4: 샘플 사이즈가 조인트 지노타이핑에 미치는 영향 — DRAGEN Platform을 이용해 높은 커버리지의 WGS 데이터 세트를 조인트 지노타이핑한 후 샘플 사이즈를 늘려가며 측정한 SNP(A)와 Indel(B)의 위양성 및 위음성 결과를 표시한 그래프

낮은 커버리지의 WGS

비교적 낮은 커버리지의 샘플에 대한 조인트 지노타이핑의 잠재적 이익을 파악하기 위해 1000 Genomes Project의 코호트 연구로부터 얻은 정렬(alignment) 데이터를 15×로 다운샘플링(downsampling)한 후 DRAGEN Platform으로 다시 처리했습니다. 이 분석에는 17번 염색체의 첫 10 Mbp로 이루어진 영역이 선택되었습니다. 다운샘플링된 데이터로부터 얻은 gVCF 파일들을 취합한 후 조인트 지노타이핑을 수행하였고, 진리 샘플인 NA12878에 ROC 메트릭스를 적용하여 성능을 측정했습니다.

결과

낮은 커버리지의 WGS 데이터 세트에 대한 성능을 측정하기 위해 gVCF Genotyper 및 Joint Genotyper에서 나온 multisample VCF 파일들로부터 진리 샘플인 NA12878을 포함하는 열을 추출한 후 오류 수를 그래프에 표시하였습니다. SNP 민감도(sensitivity)의 이득보다 특이도(specificity)의 손실이 더 커, 높은 커버리지의 데이터와 유사한 측정 결과를 보였고(그림 5A), Indel 검출의 경우 모든 메트릭스에서 퇴보를 보였습니다(그림 5B).

높은 커버리지의 WES

WES 데이터에 대한 DRAGEN Joint Genotyper의 성능을 측정하기 위해 CEU의 인구 집단으로부터 얻은 여덟 개의 비혈연 관계 샘플과 GIAB 컨소시엄의 트리오 샘플 중 두 자식의 샘플로 구성된 패널(총 10개의 샘플 포함)을 사용했습니다. 각각 1개, 3개, 4개, 6개, 8개 및 10개의 샘플로 구성된 여섯 가지 부분집합(subset)에 조인트 지노타이핑을 수행했습니다. 성능은 진리 샘플인 NA12878에 ROC 메트릭스를 적용하여 엑솜 캡처 영역(exome capture region) 내에서 측정했습니다.

결과

Multisample VCF 파일에서 진리 샘플인 NA12878을 포함하는 열을 추출한 후 그래프에 ROC 곡선을 적용하여 상기 부분집합으로 얻은 콜을 평가했습니다. 앞서 실시했던 분석들과 마찬가지로, 더 많은 수의 샘플을 조인트 지노타이핑했을 때 관찰된 뚜렷한 이득은 없었습니다(그림 6). 권장되는 DRAGEN PopGen Workflow는 gVCF Genotyper를 실행한 후 조인트 지노타이핑 단계는 생략하는 것입니다(그림 7).

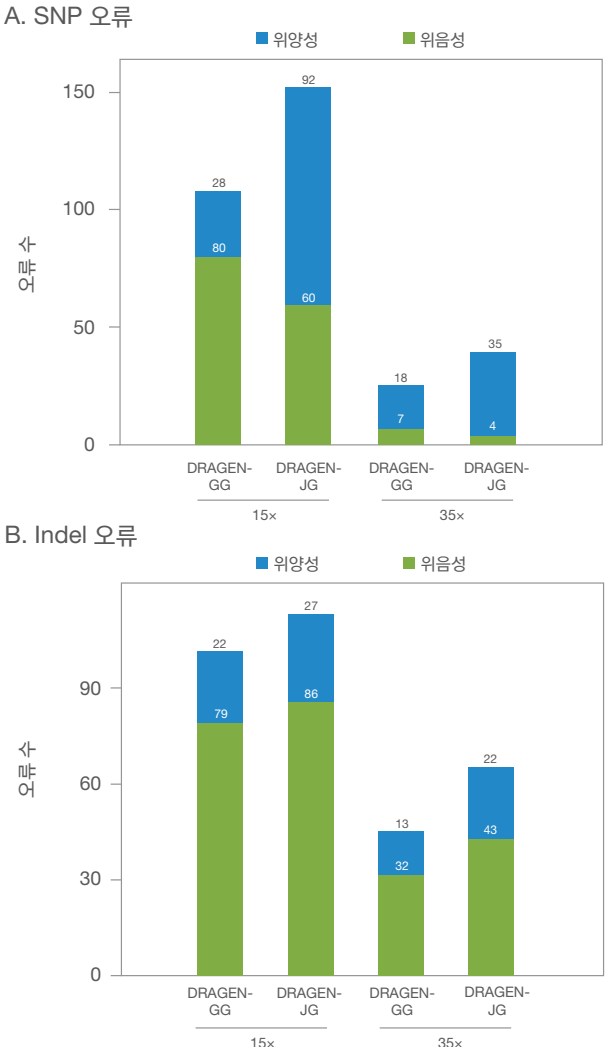


그림 5: 낮은 커버리지의 WGS 데이터 세트에서 확인된 변이 검출 정확도 — DRAGEN Platform(GG, JP)을 이용해 집단 유전학 데이터 처리 후 multisample VCF 파일에서 측정한 SNP(A)와 Indel(B)의 위양성 및 위음성 변이 검출 결과를 15× 대 35× 시퀀싱 커버리지에서 비교한 그래프

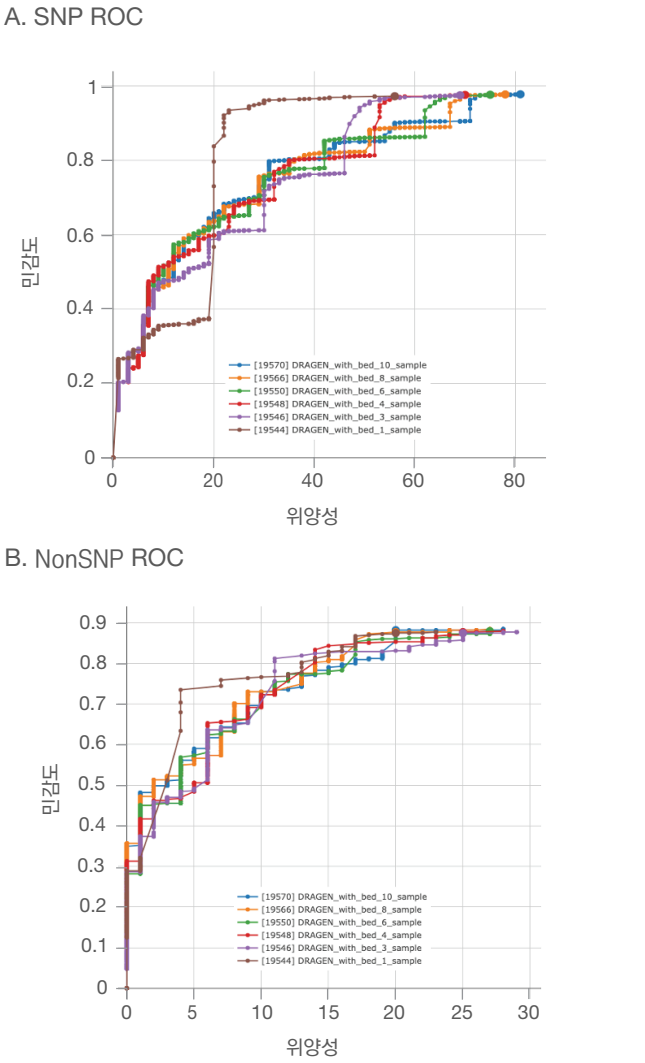


그림 6: 높은 커버리지의 WES에 대한 DRAGEN Joint Genotyper의 영향 — DRAGEN Platform을 이용한 조인트 지노타이핑 이후 샘플 수 증가에 대한 ROC 곡선을 표시한 그래프

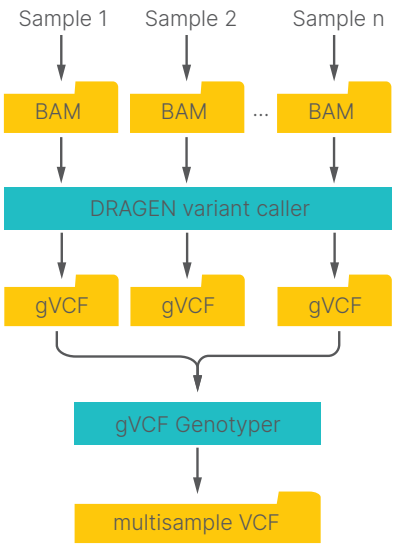


그림 7: 권장 DRAGEN PopGen Workflow

요약

코호트 데이터 처리 및 분석에 대해 정립되어 있는 GATK Best Practices Workflow에는 코호트의 정보를 이용하여 개별 샘플의 지노타입 콜을 향상시키는 조인트 지노타이핑 단계가 포함되어 있습니다. 하지만 본 Technical Note가 제시한 결과를 살펴보면, DRAGEN Platform과 대량의 높은 커버리지(최소 30× 커버리지)의 샘플을 사용하는 경우 GATK Workflow에 포함되어 있는 조인트 지노타이핑은 오류 발생 위험, 긴 연산 시간, 비용을 고려했을 때 적합하지 않습니다. 권장되는 DRAGEN PopGen Workflow는 gVCF Genotyper를 실행한 후 조인트 지노타이핑 단계는 생략하는 것입니다(그림 7). 이 방법을 사용하면 개별 gVCF 파일 취합이 완료된 후 즉시 분석이 가능한 변이를 포함하는 multisample VCF 파일이 생성됩니다. DRAGEN Platform을 사용하는 이 간소화된 워크플로우는 유연하고 효율적인 방식으로 정확도가 매우 높은 집단 유전학 콜 세트를 제공합니다.

참고 문헌

1. The 1000 Genomes Project Consortium; Auton A, Brooks LD, et al.. A global reference for human genetic variation. *Nature*. 2015;526:68–74. doi: 10.1038/nature15393.
2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061–73. doi: 10.1038/nature09534.
3. Intel, 2016. Infrastructure for Deploying GATK Best Practices Pipeline. intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf. Accessed December 01, 2020.
4. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–501.
5. Zook JM, McDaniel J, Olson ND, et al. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol*. 2019;37:561–6. doi: 10.1038/s41587-019-0074-6.
6. Roslin NM, Welli L, Paterson AD, Strug LJ. Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes. *bioRxiv*. 2016;078600–078600. doi: https://doi.org/10.1101.078600.

illumina®

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

M-GL-00561 v1.0 KOR